

A Deep Learning Solvent-Selection Paradigm Powered by a Massive Solvent/Nonsolvent Database for Polymers

Anand Chandrasekaran,^{*,†} Chiho Kim,[†] Shruti Venkatram, and Rampi Ramprasad



Cite This: <https://dx.doi.org/10.1021/acs.macromol.0c00251>



Read Online

ACCESS |



Metrics & More

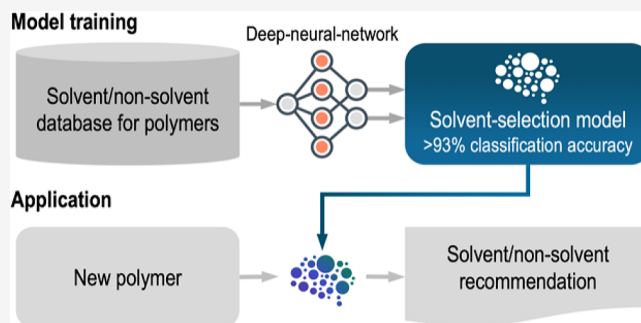


Article Recommendations



Supporting Information

ABSTRACT: Polymer solubility is critical for a variety of industrial and research applications such as plastics recycling, drug delivery, membrane science, and microlithography. For novel polymers, it is often an arduous process to find the appropriate solvents for polymer dissolution. Heuristic approaches, such as solubility parameters, provide only limited guidance with respect to solvent prediction and design. The present work highlights a novel data-driven paradigm for solvent selection in polymers. For this purpose, we utilize a deep neural network trained on a massive data set of over 4500 polymers and their corresponding solvents/nonsolvents. This deep-learning framework maps high-dimensional fingerprints/features to compact chemically relevant latent space representations of solvents and polymers. When these low-dimensional representations are visualized, we observe the spontaneous clustering of nonpolar, polar-aprotic, and polar-protic behavior. This large-scale data-driven approach possesses an overall classification accuracy of above 93% (on a hold-out set) and significantly outperforms existing methods to determine polymer/solvent compatibility such as the Hildebrand criteria.



INTRODUCTION

Polymers are an ubiquitous class of materials and find widespread applications in a number of technological and research endeavors. Because of their immense chemical and structural diversity, they are utilized in an unimaginable number of products ranging from bullet-proof vests (Kevlar) to intricate drug delivery systems.¹ However, this very same diversity renders it difficult to systematically search for polymers possessing specific properties. More recently, the nascent field of polymer informatics has shown promise in accelerating the rational design of novel functional polymers. High-throughput computations² in conjunction with an ever-increasing accumulation of experimental measurements has allowed us to use cutting-edge machine-learning (ML) algorithms^{3,4} to develop predictive models and discover hidden structure–property relationships and trends.

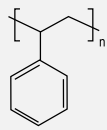
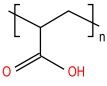
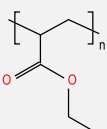
One important factor in the synthesis and processing of polymers is the selection of suitable solvents.⁵ The compatibility of polymers and solvents is of paramount importance for paint and coating formulations,⁶ plastics recycling,⁷ drug delivery,⁸ membrane science,⁹ and microlithography.¹⁰ In some cases, to precipitate a particular polymer from a mixture/blend, a nonsolvent is used. For a newly synthesized polymer, it is usually a very time-consuming process to find the right solvent/nonsolvent. Over the past decades, a number of quantitative (albeit heuristic) measures have been developed to aid in the extremely difficult process of solvent selection for polymers. The Hildebrand criteria,

introduced by Hildebrand and Scott, is likely the most well-established and well-known of the so-called solubility parameters.¹¹ Solubility parameters are based on the concept of “like dissolves like”, and when a polymer and solvent possess similar values of the Hildebrand solubility parameter, the solvent is designated as a good solvent for that particular polymer. The one-component Hildebrand parameter is derived from the cohesive energy density of the polymeric/molecular system, and it has been shown to perform reasonably well for nonpolar systems without hydrogen bonding.¹¹ However, this method completely fails in enabling suitable solvent predictions for polar systems with hydrogen bonding. The phenomenon of polymer/solvent compatibility is contingent on a multitude of chemical, thermodynamic, kinetic, and morphological factors. To attempt to capture the multidimensional nature of this problem, the three-dimensional Hansen solubility parameter system was developed by Hansen and Abbott.^{12,13} This approach involves the decomposition of the Hildebrand parameter into dispersion, polar, and hydrogen-bonding components. However, as was highlighted in our

Received: February 7, 2020

Revised: May 8, 2020

Table 1. Examples of Polymer and Solvents from the Data Set Used for Training the Neural Network Classifier

Polymer name	Structure	SMILES	Solvents	Non-solvents
Polystyrene		<chem>C(c1ccccc1)C</chem>	Chloroform, THF, benzene, toluene	Water, methanol
Polyacrylic acid		<chem>C(C(=O)O)C</chem>	Water, methanol, ethanol, dioxane	Benzene, chloroform, acetone, diethyl ether
Poly(ethyl methacrylate)		<chem>C(C(=O)OCC)CC</chem>	Acetone, benzene, THF, toluene	Methanol

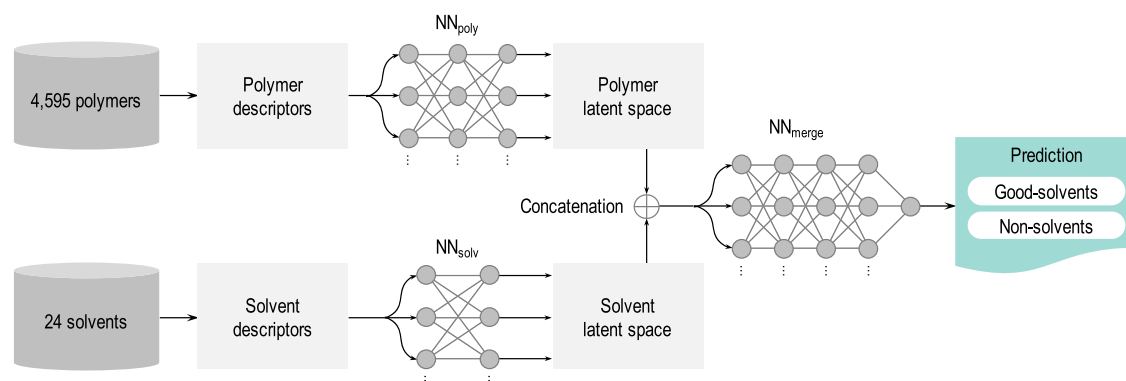


Figure 1. Neural network architecture for the prediction of good solvents and nonsolvents for polymers. The trained model is served on Polymer Genome (www.polymergenome.org). Once the user draws a new polymer, the algorithm iterates over the list of 24 solvents and provides a ranking of good solvents and nonsolvents.

recent work,¹⁴ the Hansen parameter is available only for a handful of polymers and offers only marginal improvement in prediction accuracy (relative to the Hildebrand approach). Quantifying such a complex phenomenon using a limited number of manually fitted parameters has proven to be an exceptionally challenging endeavor.¹⁵ There is, therefore, an immediate (and crucial) need for the development of novel theoretical/algorithmic frameworks to address the complex problem of polymer solubility.

In this work, we introduce a novel paradigm in which room-temperature polymer–solvent compatibility is determined using a purely data-driven framework in contrast to existing heuristic approaches based on solubility parameters. We first accumulate a massive data set of over 4500 polymers and their corresponding solvents/nonsolvents. The data set is used to train a deep neural network binary classifier, which takes as input the polymer and solvent descriptors and as output whether the solvent is a good solvent or a nonsolvent (for that particular polymer). The fingerprinting methodology, neural network architectures, and training protocols are detailed in the *Methods*. Our data-driven deep-learning framework is shown to overwhelmingly outperform the Hildebrand approach for all classes of polymers and solvents. In addition to the vastly improved prediction accuracy, this approach enables the visualization of solvents and polymers in terms of chemically relevant low-dimensional latent space vectors. It is

important to note that these latent space representations have been obtained intrinsically via the ML algorithm, without the need to input any empirically or manually derived parameters. Further discussions and details are elucidated in this Article. Finally, we also implement this method on our online polymer informatics platform, Polymer Genome.

RESULTS AND DISCUSSION

Our data set is comprised of 4595 polymers with an associated list of good solvents. A subset of this data set comprised of 3707 polymers also has an associated list of nonsolvents. A few examples from this data set are depicted in *Table 1*. There are a total of 11 958 polymer + solvent combinations and a total of 8469 polymer + nonsolvent pairs. In this work, we only consider 24 widely used solvents/nonsolvents (listed in *Table S1*) that can be broadly categorized as nonpolar, polar-aprotic, or polar-protic solvents. We randomly select 10% of the data set as a test-set (or hold-out set) and perform 5-fold validation for the rest of the data set as shown in *Figure S1*.

The chemical and structural features of the polymers are encoded in terms of a numerical vector (or fingerprint) as detailed in the *Methods*, and the solvents are represented using the so-called one-hot encoding technique. These numerical representations of the polymer and solvent are fed into a neural network whose architecture is schematically illustrated in *Figure 1*. The neural network, described in more detail in the

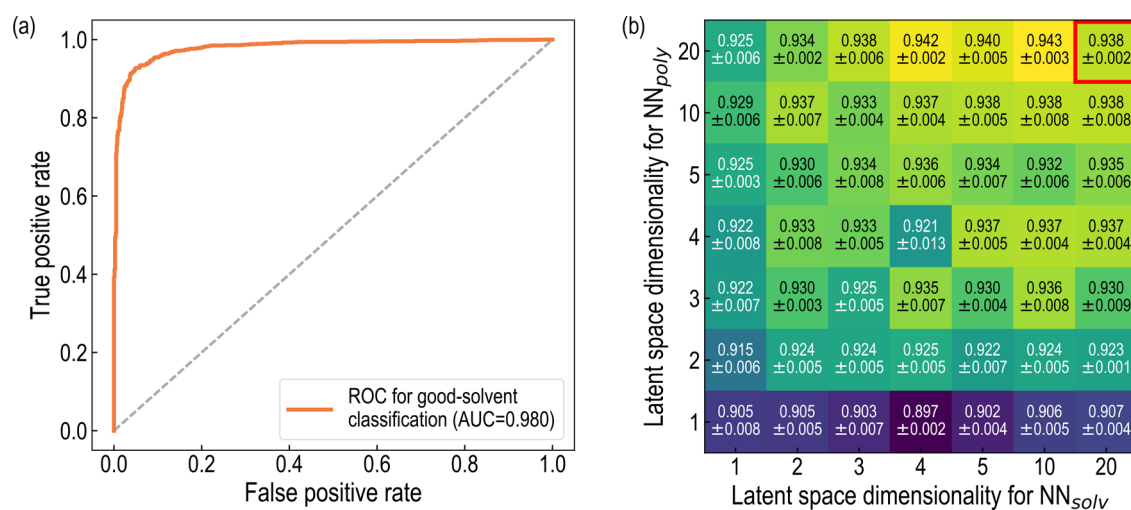


Figure 2. (a) ROC curve of the neural network model on the test set. (b) Accuracy of the neural network classification versus the latent space dimensionality of the polymer and solvent on the test set. The red square indicates the most accurate model obtained (based on the validation set shown in Figure S2).

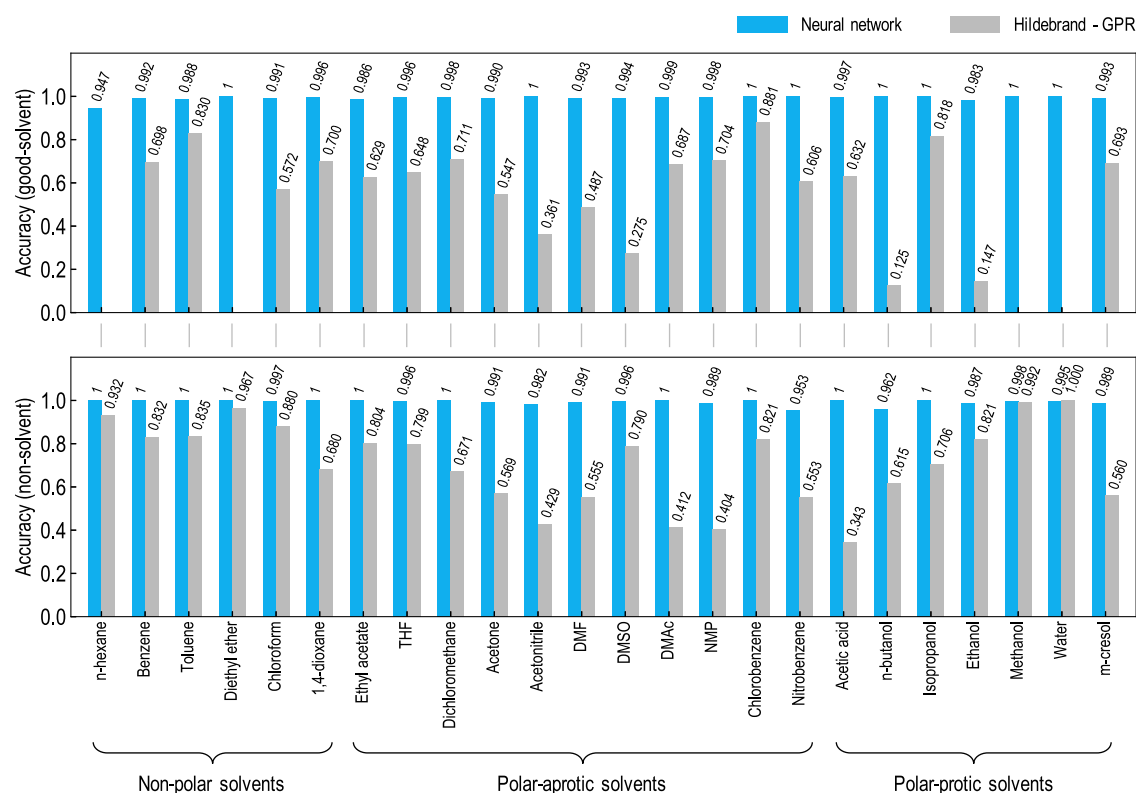


Figure 3. Solvent-wise prediction accuracy of soluble (top) and insoluble (bottom) polymers. We compare the accuracy of the neural network versus the Hildebrand criteria for nonpolar, polar-aprotic, and polar-protic solvents. With respect to good-solvent prediction, the Hildebrand parameter fails catastrophically for many solvents such as *n*-hexane, water, and methanol. Overall, the nonsolvent prediction accuracy of the Hildebrand parameter is higher as compared to that of good-solvent prediction due to a less stringent criteria of polymer insolubility.

Methods, functions as a binary classifier, which can simultaneously learn chemically relevant low-dimensional features of the polymers and solvents (as a result of the polymer latent space layer and the solvent latent space layer, respectively).

First, we evaluate the performance of the neural network approach in terms of simple classification accuracy. If the activation value of the final sigmoid neuron is greater than (or equal to) the threshold value of 0.5, the solvent is considered

as a good solvent for that particular polymer. If the activation value is less than 0.5, the solvent is considered as a nonsolvent for that polymer. Therefore, an activation value of 1 would imply an ideal good solvent and an activation value of 0 would imply an ideal nonsolvent.

To quantify the classification accuracy, we ask the following question for several cases of polymer–solvent combinations: what percentage of the time is the classifier able to correctly predict whether the polymer is soluble in that particular

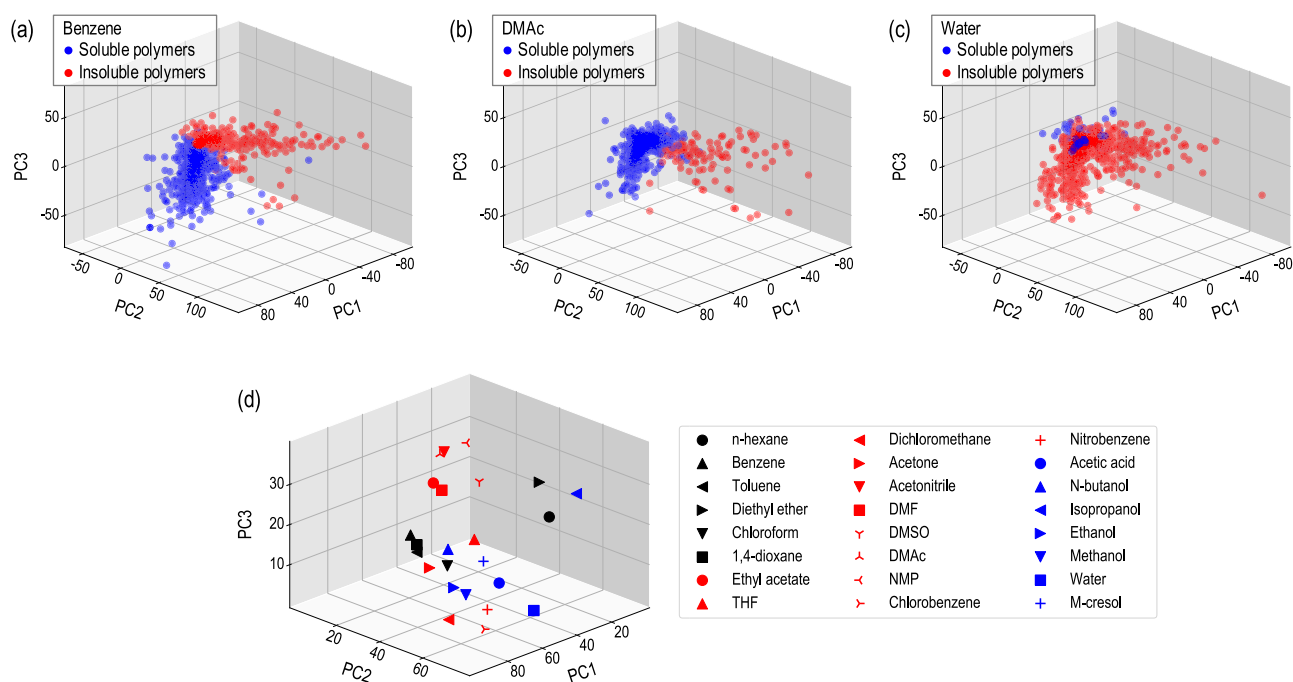


Figure 4. (a–c) The three principal components of the 20-dimensional latent space of polymers, which are soluble/insoluble in benzene, DMAc, and water, respectively. The blue “●” are polymers that are soluble, and the red “●” denote polymers that are insoluble in that particular solvent. (d) The distribution of nonpolar (black), polar-aprotic (red), and polar-protic (blue) solvents within the 20-dimensional solvent latent space (projected onto three principal components).

solvent? On the basis of the above criteria for classifying solvents/nonsolvents, we obtained an accuracy of 99.6% for the training set, an accuracy of 98.1% for the validation set, and an accuracy of 93.8% for the test set. The ROC (receiver operating characteristic) curve and associated AUC (area-under-curve) value for the good-solvent classification for the entire data set are shown in Figure 2a. We also show how the accuracy of the model is dependent on the polymer latent space dimensionality and solvent latent space dimensionality. As detailed in the Methods, the original dimensionality of the polymer fingerprint is 644, and the dimensionality of the solvent fingerprint is 24. Considering that the Hildebrand and Hansen criteria possess a dimensionality of one and three, respectively, determining the dimensionality of the ideal descriptor set is the next logical step. From Figure S2 we see that the validation accuracy saturates around 98.1. The most accurate model (highlighted in red in Figure 2b) is chosen as the final model for implementation and further analysis.

We compare the classification accuracy of the neural network against the well-established Hildebrand criteria. Although the Hildebrand parameter is available for all 24 solvents, only 100 or so polymers possess a corresponding solubility parameter that can be extracted from the existing literature. To overcome this limitation, we construct an ML model (from available data) to predict the Hildebrand parameter of all 4595 polymers in the data set. The construction and validation of the model are described in detail in the Methods.

In contrast to the neural network, the Hildebrand solubility criterion when applied to this large polymer data set displayed a good-solvent classification accuracy of only 50% and a nonsolvent classification accuracy of 70%. The nonsolvent classification accuracy is superior to that of the good-solvent classification because it is a simple criterion of exclusion. As

shown in Figure 3, the performance of the Hildebrand model is significantly inferior with respect to solubility prediction for polar-protic solvents such as water, methanol, and ethanol. However, surprisingly, it also performs poorly for nonpolar solvents, *n*-hexane and diethyl ether. The two solvents possess Hildebrand parameters of 14.9 and 15.4 MPa^{1/2}, respectively. On closer inspection of the histogram of the predicted Hildebrand parameter values shown in Figure S3c, we see that there are few polymers with Hildebrand parameters in the vicinity of these two solvents. This data set bias is reflected in the poor predictive capabilities for the solvents *n*-hexane and diethyl ether.

In Figure 4, we visualize the distribution of polymers and solvents in their corresponding latent spaces. To facilitate the visualization, we reduce the dimensionality of the polymer and solvent latent space using principal components analysis (PCA). Figure 4a–c clearly demonstrates the different regions of polymer latent space occupied by soluble (blue) and insoluble (red) polymers for benzene (a nonpolar solvent), dimethylacetamide or DMAc (polar-aprotic solvent), and water (polar-protic solvent), respectively. In Figure 4d, we see how nonpolar, polar-aprotic, and polar-protic solvents are clustered into different regions of the solvent latent space.

Finally, to demonstrate the sensitivity of the ML algorithm to the training set size, we plot the learning-curves of test accuracy and training accuracy versus the training set size in Figure S3. The convergence of the training and test accuracy is demonstrated as we systematically expand the training set size (starting from a training set size of 5% of the total data set), indicating that our learning algorithm is close to its optimal performance for the given data set size. The standard deviation of the accuracy during the 5-fold cross validation is shown as error bars.

In summary, we have developed a novel deep-learning framework capable of accurately predicting good solvents and nonsolvents for an enormous number and types of polymers. In a recent work,¹⁴ we critically examined the accuracy of the Hildebrand and Hansen criteria for a number of polymers and solvents (for which these criteria were available). The Hildebrand criteria were shown to have an accuracy of 60% for solvents and 76% for nonsolvents, whereas the Hansen criteria were shown to have an accuracy of 67% for solvents and 76% for nonsolvents.¹⁴ Our approach, therefore, not only outperforms the existing heuristic methods^{14,15} but is also applicable over a larger chemical space and enables the extraction of chemically meaningful representations for both solvents and polymers. The purely data-driven framework described in this work can be systematically improved as it is exposed to an even larger quantity and diversity of data. Moreover, the presence of a chemically meaningful low-dimensional latent space sets the stage for techniques such as transfer-learning,¹⁶ one-shot learning, and generative models.⁴

This deep learning-based solvent/nonsolvent prediction toolkit has been implemented in our online polymer informatics platform “Polymer Genome”. The user may draw the repeat unit of any arbitrary polymer and obtain the rankings of the best good solvents and nonsolvents for that polymer. The rankings are inferred from the relative magnitude of the activation of the final sigmoid neuron. The solvent predictions based on the Hildebrand approach are also provided (separately) in addition to predictions obtained from the deep-learning framework. Because Polymer Genome already predicts a variety of other properties such as the glass transition temperature (T_g), band gap, dielectric constant, refractive index, etc., the prediction of suitable solvents will significantly aid in the accelerated design of synthetically feasible/soluble polymers with a tailored set of properties. Going forward, additional descriptors for temperature-dependent solubility or for solubility involving mixtures of solvents would likely increase the utility of this approach.

METHODS

Data Set. Data were obtained from a number of publicly available resources, including handbooks such as the *Polymer Handbook*,¹⁷ *Handbook of Polymers*,¹⁸ *Properties of Polymers*,¹⁹ and *Polymer Data Handbook*.²⁰ Data were also accumulated from a number of literature sources and online repositories of data, including *Polymer Database*²¹ and *PolyInfo*.²² The chemical space of the polymers spans over a wide range but predominantly included the following atomic species: H, B, C, N, O, F, Si, P, S, Cl, Br, and I. Copolymers, polymer blends, polymers with additives, and cross-linked polymers are not considered in this study. We also limit this study to the investigation of room-temperature solubility and do not consider partial solubility or high-temperature solubility.

The polymers are represented using the simplified molecular-input line-entry system or SMILES. The SMILES strings of the polymers are constructed using the online draw tool implemented in the Polymer Genome platform,²³ and the terminal atoms of the polymer repeat unit are denoted by “*”.

Fingerprinting. The descriptors used for polymers have been described in detail in earlier works.^{23,24} In the current work, a total of 644 descriptors are utilized, and these descriptors are broadly categorized into four structural hierarchies. The first family of descriptors includes the so-called atomic-triple fingerprint,²⁵ and the second family of descriptors is obtained from the open-source cheminformatics package RDKit.²⁶ A total of 300 atomic-triple fingerprints are utilized, and 39 RDKit-based descriptors are used.²⁷ The third family of descriptors is obtained by searching for commonly

occurring substructures/blocks in this large polymer data set; 288 such descriptors are utilized. Finally, a set of 17 polymer-chain specific descriptors are included to encode certain higher level features such as the length of the side chain, number of rings in the main chain, etc. All of the descriptors are modified to take into account the one-dimensional periodicity of large molecular weight polymer chains. For instance, all of the fingerprint components are normalized with respect to the number of atoms in the polymer repeat unit.

In the current work, because we are considering only a limited number of solvents, we utilize one-hot encoding to represent each of the 24 solvents. One-hot encoding is a technique to represent categorical variables using a binary representation. In this particular case, the fingerprint vector for a particular solvent would have a dimensionality of 24, of which 23 components would be zero and only one component would possess the integer value of one.

Neural Network Details. A slightly modified version of the multilayer perceptron neural network algorithm is utilized for this particular binary classification (good solvent or nonsolvent) problem. The neural network, as depicted in Figure 1, consists of two input branches, one for the solvent fingerprints and the other for the polymer descriptors. The solvent branch of the neural network (NN_{sol}) consists of two hidden layers, each with 100 neurons, and the polymer branch of the neural network (NN_{poly}) consists of three hidden layers, each with 100 neurons. The final hidden layer of the solvent branch of the neural network consists of 20 neurons and is labeled as the “solvent latent space” in Figure 1. Similarly, the final layer of the polymer branch of the neural network consisted of 20 neurons and is labeled as the “polymer latent space”. The solvent latent space and polymer latent space hidden layers are merged using a concatenation operation and passed on to a final set of hidden layers (NN_{merge}). There are four hidden layers in NN_{merge} , each with 100 neurons. All of the hidden layers in the neural network are constructed with the parametrized rectified linear unit (PReLU) activation function. The final output of the neural network consists of a single neuron with the sigmoid activation function wherein an ideal good solvent is denoted by an activation value of “1” and an ideal nonsolvent is denoted by an activation value of “0”. The threshold to differentiate between a nonsolvent and a good solvent is set at 0.5.

A randomized 90%/10% train/test split is utilized, and 5-fold cross validation is further performed on the training set as shown in Figure S1. During the 5-fold cross validation, the validation set was utilized to determine the optimal number of epochs for each fold. The “Adam” gradient descent method is utilized for training the neural network weights along with a training batch-size of 100. The 10% test split represents the hold-out set, and the performance of the neural network on this set represents the generalizability of the model.

Gaussian Process Regression for the Hildebrand Parameter.

As mentioned earlier, we sought to benchmark the performance of our solvent prediction neural network against existing methods for determining solubility. For this purpose, we use the Hildebrand criteria, because it is not only widely utilized, but it is also the only solubility parameter that is available for a considerable number of polymers. As reported in recent works,^{14,23} we accumulated a data set of 113 polymer Hildebrand parameters from the data set sources. Once the polymer fingerprints are obtained for the polymers, Gaussian Process Regression using the radial basis function (RBF) kernel is utilized to map the polymers to their associated Hildebrand solubility parameters. Through this process, we are able to obtain a machine learning model to predict the solubility parameter of all of the polymers in our data set of 4595 polymers. A summary of the Hildebrand parameter polymer data set and the performance of the associate model are illustrated in Figure S1. For the case of the 24 solvents, the Hildebrand parameter was readily available from multiple sources in the literature.¹² Following the standard procedure,²⁸ if the absolute value of the difference in Hildebrand parameters of the solvent and polymer was less than $2 \text{ MPa}^{1/2}$, the solvent was classified as a good solvent for that particular polymer. If the difference was greater than $2 \text{ MPa}^{1/2}$, the solvent was classified as a nonsolvent for that polymer.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.macromol.0c00251>.

Table showing the number of soluble and insoluble polymers for each solvent along with the Hildebrand solubility parameter for the solvent; cross-validation and test split schema; heat-map of validation set accuracy as a function of latent space dimensionality; Hildebrand parameter of 113 polymers in the training set along with ML predicted Hildebrand parameter; and learning-curve of train and test accuracy (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Anand Chandrasekaran – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0002-2794-3717; Email: anandc88@gmail.com

Authors

Chiho Kim – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0002-1814-4980

Shruti Venkatram – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

Rampi Ramprasad – School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; orcid.org/0000-0003-4630-1565

Complete contact information is available at <https://pubs.acs.org/10.1021/acs.macromol.0c00251>

Author Contributions

[†]A.C. and C.K. contributed equally to this work.

Author Contributions

All authors contributed to the conception and design of the study, as well as writing of the manuscript. A.C. developed the data-curation/preprocessing pipeline and designed the neural network architecture. A.C. and C.K. wrote the Python-based training and prediction code. R.R. supervised the work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research through grants N00014-17-1-2656 and N00014-16-1-2580.

■ REFERENCES

- (1) Bajpai, A.; Shukla, S. K.; Bhanu, S.; Kankane, S. Responsive polymers in controlled drug delivery. *Prog. Polym. Sci.* **2008**, *33*, 1088–1118.
- (2) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A polymer dataset for accelerated property prediction and design. *Sci. Data* **2016**, *3*, 160012.
- (3) Staker, J.; Marshall, K.; Abel, R.; McQuaw, C. M. Molecular Structure Extraction from Documents Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1017–1029.
- (4) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559.
- (5) Miller-Chou, B. A.; Koenig, J. L. A review of polymer dissolution. *Prog. Polym. Sci.* **2003**, *28*, 1223–1270.

(6) Adamska, K.; Voelkel, A.; Héberger, K. Selection of solubility parameters for characterization of pharmaceutical excipients. *Journal of Chromatography A* **2007**, *1171*, 90–97.

(7) Kobayashi, S.; Uyama, H.; Takamoto, T. Lipase-catalyzed degradation of polyesters in organic solvents. A new methodology of polymer recycling using enzyme as catalyst. *Biomacromolecules* **2000**, *1*, 3–5.

(8) Breitskreutz, J. Prediction of intestinal drug absorption properties by three-dimensional solubility parameters. *Pharm. Res.* **1998**, *15*, 1370–1375.

(9) Koros, W.; Fleming, G. Membrane-based gas separation. *J. Membr. Sci.* **1993**, *83*, 1–80.

(10) Reichmanis, E.; Thompson, L. F. Polymer materials for microlithography. *Chem. Rev.* **1989**, *89*, 1273–1289.

(11) Barton, A. F. M. Solubility parameters. *Chem. Rev.* **1975**, *75*, 731–753.

(12) Hansen, C. *Hansen Solubility Parameters*; CRC Press: New York, 1999.

(13) Abbott, S.; Hansen, C. M. *Hansen Solubility Parameters in Practice*; Hansen Solubility, 2008.

(14) Venkatram, S.; Kim, C.; Chandrasekaran, A.; Ramprasad, R. Critical assessment of the Hildebrand and Hansen solubility parameters for polymers. *J. Chem. Inf. Model.* **2019**, *59*, 4188.

(15) Sanchez-Lengeling, B.; Roch, L. M.; Perea, J. D.; Langner, S.; Brabec, C. J.; Aspuru-Guzik, A. A Bayesian Approach to Predict Solubility Parameters. *Advanced Theory and Simulations* **2019**, *2*, 1800069.

(16) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, *5*, 1717.

(17) Brandrup, J.; Immergut, E. H.; Grulke, E. A. *Polymer Handbook*, 4th ed.; Wiley: New York, 2004.

(18) Wypych, G. *Handbook of Polymers*; Elsevier: New York, 2016.

(19) Van Krevelen, D.; Te Nijenhuis, K. *Properties of Polymers*; Elsevier: New York, 2009; pp 3–5.

(20) Mark, J. E. *Polymer Data Handbook*, 2nd ed.; Oxford University Press: New York, 2009.

(21) Polymer Database; <https://polymerdatabase.com/>.

(22) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 2011; pp 22–29.

(23) Kim, C.; Chandrasekaran, A.; Huan, T.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122*, 17575.

(24) Ramprasad, R.; Chandrasekaran, A.; Huan, T. D.; Kim, C. Patent Pending, U.S. Patent Application no. 62/834,428, 2019.

(25) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21*, 785.

(26) Landrum, G. RDKit: Open-source cheminformatics; <http://www.rdkit.org>.

(27) Nguyen, K.; Blum, L.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009**, *4*, 1803–1805.

(28) Burke, J. *Solubility Parameters: Theory and Application*; The Book and Paper Group of the American Institute for Conservation: Washington, DC, 1984.