ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

# Solubility Prediction from Molecular Properties and Analytical Data Using an In-phase Deep Neural Network (Ip-DNN)

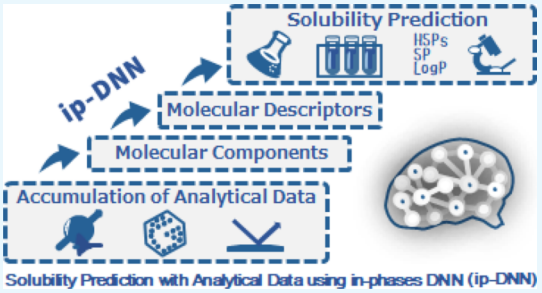
Atsushi Kurotani, Toshifumi Kakiuchi, and Jun Kikuchi*

Cite This: *ACS Omega* 2021, 6, 14278−14287

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Materials informatics is an emerging field that allows us to predict the properties of materials and has been applied in various research and development fields, such as materials science. In particular, solubility factors such as the Hansen and Hildebrand solubility parameters (HSPs and SP, respectively) and Log *P* are important values for understanding the physical properties of various substances. In this study, we succeeded at establishing a solubility prediction tool using a unique machine learning method called the in-phase deep neural network (ip-DNN), which starts exclusively from the analytical input data (e.g., NMR information, refractive index, and density) to predict solubility by predicting intermediate elements, such as molecular components and molecular descriptors, in the multiple-step method. For improving the level of accuracy of the prediction, intermediate regression models were employed when performing in-phase machine learning. In addition, we developed a website dedicated to the established solubility prediction method, which is freely available at "http://dmar.riken.jp/matsolca/".



Solubility Prediction with Analytical Data using in-phases DNN (ip-DNN)

## INTRODUCTION

In recent years, the application of data-driven models has been implemented in various research and development fields such as materials science, biorefinery, cosmetic chemistry, and drug discovery, especially at the industrial level. Sophisticated machine learning techniques are now becoming ubiquitous for the prediction of the physicochemical properties and engineering parameters. In materials science, the increasing availability of large amounts of data (both analytical and computational) has been recently used to advance the tools available for materials informatics (MI). It is known that a variety of indexes are commonly used to describe the solubility of substances. Among these, SP is defined by regular solution theory proposed by Hildebrand and Scott,[1] and Hildebrand solubility parameters (HSPs) are trinomial components proposed by Hansen[2] that correspond to the dispersion energy (dD), dipole interaction energy (dP), and energy of hydrogen bonding (dH) between molecules. Log *S* is the base 10 logarithm of the solubility *S* [mol/L] in water. Log *P* is the base 10 logarithm of the octanol−water partition coefficient that indicates octanol solubility and therefore lipophilicity. In particular, they are needed in various research and development fields where solubility information of substances such as materials, pharmaceuticals, and food is required.[3−5]

The calculation of the solubility values is mainly performed using the conventional group contribution method, although the machine learning method has also been attracting attention in recent years owing to the artificial intelligence boom along with the development of chemoinformatics and MI. In addition, simulation methods are often used as complementary techniques to the standard calculation of the solubility values.[6,7] The calculated solubility values by the group contribution method are based on the aggregation energy of the molecular structures (atoms, functional groups, etc.).[8] The group contribution method was developed in an early stage[9,10] and has been improved in recent years.[11−13] In addition, the application of the predicted Log *S* values to the group contribution method for drug delivery has also been reported.[14] The determination of the solubility values by machine learning methods relies on the prediction of these values by training known structural and physical properties on information related to the solubility as descriptors. As an example of prediction of Log *S* using machine learning, a report described how to calculate the desired value using a random forest to train the molecular descriptors of the CDK tool,[15] which is a chemoinformatic library in the Java language.[16] Another study predicted the Log *S*, Log *P*, melting point, and toxicity with a convolutional neural network (CNN) using the fingerprint of structural information as training data with SMILES strings.[17] Moreover, the prediction of SP, glass transition point, density, and so forth was performed by the Gaussian process regression (GPR) to train the molecular

**Received:** February 25, 2021
**Accepted:** April 28, 2021
**Published:** May 17, 2021





https://doi.org/10.1021/acsomega.1c01035
*ACS Omega* 2021, 6, 14278−14287

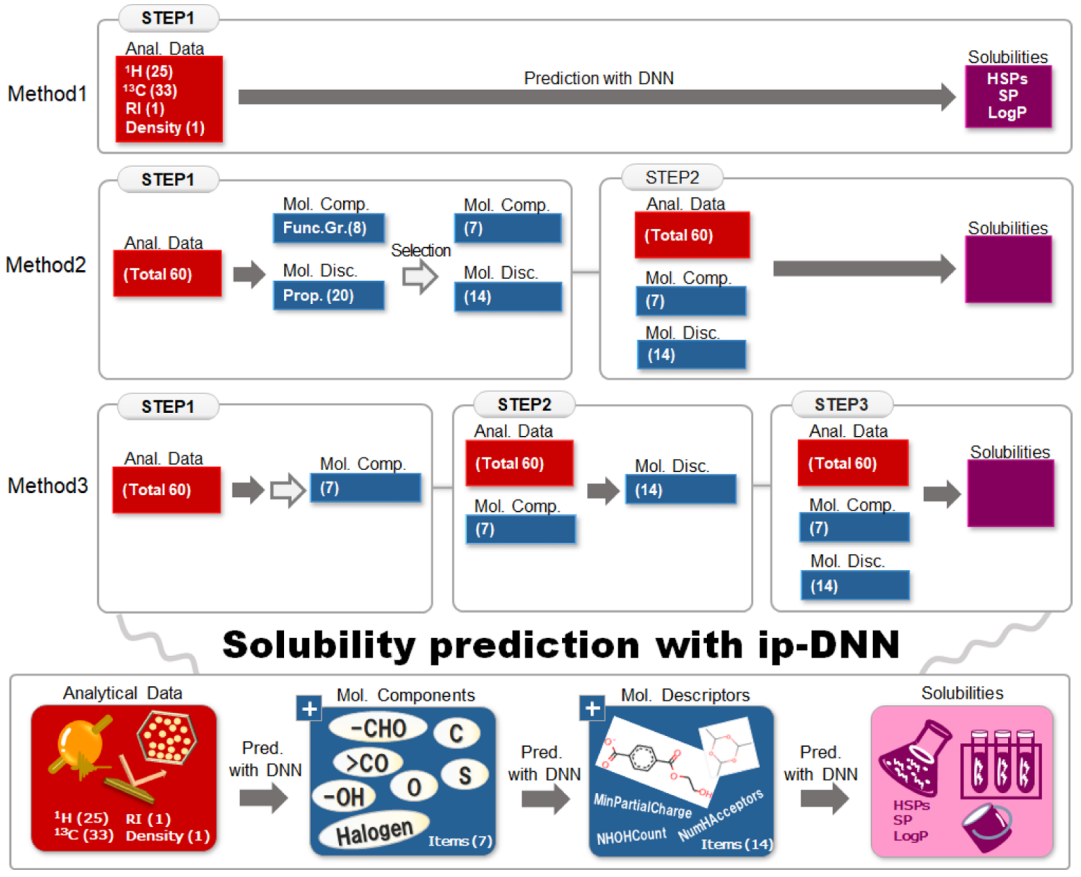



**Figure 1.** Solubility prediction methods from analytical data. Method1 allows us to predict the solubility values by simply starting from analytical data (shown as "Anal. Data") as input data using DNN. "RI" in "Anal. Data" means the refractive index. The numbers in parentheses show the number of the attributes for machine learning. Method2 is a 2-step DNN prediction method: In the first step, the molecular compositions (shown as "Mol. Comp.") and molecular descriptors (shown as "Mol. Disc.") are predicted from analytical data and are selected according to a defined threshold. Here, the molecular descriptors mean the data from RDKIT's descriptors. In the second step, the solubility values are predicted from the analytical data and selected molecular properties. Method3 is a 3-step DNN prediction method: In the first step, the molecular compositions are predicted from analytical data and selected by a defined threshold. In the second step, the molecular descriptors are predicted from the analytical data and selected molecular compositions. In the third step, the solubility values are predicted from the analytical data and selected molecular properties. This solubility prediction method from analytical data using intermediate molecular properties in phase was named as the "in-phase deep neural network (:ip-DNN)", and the image is shown at the bottom.

structure, quantitative structure−property relationship (QSPR)[18] descriptors that were obtained from the RDKIT tool,[19] and molecular morphological information, such as the side chain, distance between rings, and so forth.[20] HSPs were predicted using an improved MARS (multivariate adaptive regression splines[21]) method to train the QSPR molecular descriptors with the PaDEL tool[22] using SMILES strings.[23] HSPs were also predicted using GPR that trained the physical properties of compounds, such as the surface area, volume, and so forth, from molecular simulation data using SMILES string information.[24] As mentioned above, solubility-related predictions have been reported using various training data. However, the input data in these predicting methods require structure-related information, such as atoms, rings, bonds, functional groups, and molecular descriptors. The molecular descriptors can be obtained using chemoinformatic tools, such as RDKIT, CDK, and PaDEL, which demand at least one of the SMILES, SMARTS, sdf format, mol format, and so on. Therefore, when predicting the solubility of unknown substances with the abovementioned methods, structure-related information is required to be at least at the 2D level as input data.

In contrast, analytical data, such as NMR spectra, offer an enormous amount of information regarding the local structure and functional groups.[25,26] In particular, $^{1}$H and $^{13}$C chemical shifts can be used as information to predict the local structure or the entire molecular structure with the aid of chemo-informatics, even in the case of the primary stage analysis of a complex mixture. Such NMR spectral information along with the refractive index and density can potentially be obtained as primary-stage analytical data.[27−36]

Therefore, we developed a special solubility prediction tool using an in-phase DNN method, which is based exclusively on analytical data as input and allows us to improve the accuracy by regressing molecular information, including molecular composition and molecular descriptors, as intermediate data in a stepwise fashion (Figure 1 method3 and Figure S1b). In addition, we developed a web tool (http://dmar.riken.jp/matsolca/) to calculate mainly HSPs, SP, and Log *P* from the analysis data, including the NMR information, refractive index, and density, as input data. In addition, we confirmed the applicability of this prediction tool to polymer data whenever analytical data of a polymer are available. We believe that this tool may accelerate the creation of novel designs and


https://doi.org/10.1021/acsomega.1c01035
ACS Omega 2021, 6, 14278−14287

development of new materials since it allows us to predict the solubility from analytical data without the need for obtaining complete structural data.

## MATERIALS AND METHODS

**Dataset of Compounds, Solubility, and Analytical Data.** In this study, we prepared a dataset with 307 common low-molecular weight compounds. In this dataset, the number of C atoms in each compound ranged from 1 to 9, while the number of compounds containing N, S, Si, halogen (F, Cl, and Br), −OH, >CO, −CHO, −COOH, or aromatic groups was 48, 24, 4, 76, 33, 20, 11, 5, and 28, respectively (Table S1a−c). Information regarding the solubility, analytical data, molecular composition, and molecular descriptors of these compounds was collected. The solubility data included HSP, SP, and Log *P* values. The HSP values were obtained from the DIPPR database,[37] while the SP values were calculated from three literature HSP values according to the formula: $SP = \sqrt{dD^2 + dH^2 + dP^2}$.[38] The Log *P* values were derived using Crippen's computational Log *P*(s) also called as MolLogP,[39] which represents one of the molecular descriptors of RDKIT and can therefore be obtained with the RDKIT tool. The analytical data included 1D $^1$H NMR and 1D $^{13}$C NMR spectral data and refractive index and density values. It should be noted that the NMR spectral data were collected using the SBDB (spectral database of AIST[40]) and KnowItAll spectroscopy software (Bio-Rad Laboratories, Inc. 2018 version), while the refractive index and density values were obtained with the DIPPR database.[37] To simplify 1D $^1$H NMR and 1D $^{13}$C NMR spectral data, we converted the information regarding the peaks in the NMR spectra to the assignment information using the table of H/C-chemical shifts in organic compounds provided by Bruker.[41] The assignment information for 1D $^1$H NMR and 1D $^{13}$C NMR data is shown in Table S2a,b. Finally, we prepared 60 pieces of analytical data per compound, including 25 items of 1D $^1$H NMR, 33 items of 1D $^{13}$C NMR, a refractive index, and a density value.

**Dataset of Molecular Compositions.** We collected the conceivable general 11 items of molecular composition from chemical structural formula (H, C, N, S, Si, halogens, −OH, −CHO, >CO, −COOH, and aromatics), which are shown in Table S1b,c. Si, −COOH, and aromatics are excluded because Si and −COOH represent a small amount of data for training, and aromatics is included in the molecular descriptors of RDKIT. Therefore, we selected eight items (Table S3) of molecular composition as candidates for the feature value that correspond to the number of H and C, and the existence/absence of N, S, halogens, −OH, −CHO, and >CO is used as effective training data. In addition, using the item selection of eight molecular compositions from the DNN result, N is excluded owing to the lack of evaluation values (see also the DNN 2-step method in the Results and Discussion and Table S5b). The remaining seven items are included in the cascade in 2-step and 3-step DNN predictions as intermediate models.

**Dataset of Molecular Descriptors.** In this study, we use the molecular descriptors calculated with RDKIT derived from SMILES strings of each compound. Among a total of 200 molecular descriptors of RDKIT,[42] we selected 20 items (Table S4: Chi0n, Chi0v, Chi1v, HallKierAlpha, Kappa3, MaxPartialCharge, MinPartialCharge, MolMR, PEOE_VSA1, SlogP_VSA12, SMR_VSA5, SMR_VSA10, TPSA, VSA_EState9, NHOHCount, NumAromaticRings, NumHAcceptors, NumHDonors, NumHeteroatoms, and RingCount) as a candidate for the feature value, which are the top 70% between the highest and bottom level of the regression score, by calculating important factors[43] for the dD/dH/dP/SP regression model with random forest using the 200 molecular descriptors (Figure S2a−d). These items are generally used as either training data or objective variables for prediction. For the six molecular compositions based on −CHO, >CO, −OH, halogens, S, and N and the six molecular descriptors of RDKIT, that is, NHOHCount, NumAromaticRings, NumHAcceptors, NumHDonors, NumHeteroatoms and RingCount, we did not use their number but rather their presence or absence due to the fact that only few data correspond to more numbers higher than 1. This method based on the presence or absence of these items is indicated as presence/absence prediction, while the other is called as numerical prediction. In addition, using the item selection of 20 molecular descriptors considered from the DNN result, six items (PEOE_VSA1, Chi0v, Chi1v, MolMR, TPSA, and Kappa3) are excluded owing to the lack of evaluation values (see also the DNN 2-step method in the Results and Discussion and Table S5a). The remaining 14 items are included in the cascade in 2-step and 3-step DNN predictions as intermediate models.

**Adjustment of Calculation Values.** It should be noted that the values of the presence/absence prediction were adjusted to 0/1, while H and C were rounded to integers from the calculated value.

**Calculation with Machine Learning.** The training data were normalized with total data as preparation. DNN calculations with a fivefold cross-validation were performed using Keras-Tensorflow, which is a neural network library of python programs. The order of layers of the model is as follows: an input layer, hidden layer, activated layer, hidden layer, and output layer. The setting parameters at the time of the model calculation were the number of neurons of hidden layers (30−60), number of intermediate layers (fixed to 2), dropout rate (fixed to 0.5), activated layers (sigmoid, tanh, and relu), optimizer (adam and adagrad), learning rate (0.001−0.1), number of epochs (10−200), and batch size (32−64). The optimal values of the abovementioned parameters were determined using the Bayesian optimization method. For all other parameters reported as a range of values, the optimal items were determined using the all search (grid search) method. Random forest calculations were performed with a fivefold cross-validation using the caret package, which is a machine learning package of the R program.[44] XGBoost calculations were also performed with a fivefold cross-validation using Python's XGBRegressor library. The setting optimal parameters of XGBoost for the learning rate, max depth, subsample, and colsample by the tree were determined using the Bayesian optimization method.

**Test Data and Training Data for Machine Learning.** Among all 307 compounds, 31 compounds for the prediction evaluation test were randomly selected, which correspond to 1/10 of the total compounds, while the remaining 276 were used for training. In the first step of the 2-step DNN prediction method as descriptor selection, two more datasets, which are not duplicate in each set of prediction evaluation data, were prepared from the 307 compounds (Figure S3). The reason for preparing two more datasets in this case is to increase reliability in the descriptor selection and because the result of descriptor selection is used in the first step of the 3-step DNN


https://doi.org/10.1021/acsomega.1c01035
ACS Omega 2021, 6, 14278−14287

prediction method. The evaluation of descriptor selection was confirmed with a total of three sets.

**Model Performance Evaluation.** For the presence/absence prediction in descriptor selection, we checked the evaluation [e.g., positive predictive value (PPV), negative predictive value (NPV), recall, and specificity] to determine whether its minimum value is more than 50% of the cutoff. For numerical prediction in descriptor selection, we checked the evaluation of $R^2$, whether the value is more than 0.5 as the cutoff. For the model evaluation of solubility prediction, we checked $R^2$ and root mean squared error (RMSE).

**Confirming Exploration Performance for the Dataset.** As dataset evaluation of exploration, we tried leave-one-cluster-out cross-validation[45] (LOCO CV), for which the test data are selected by *k*-means clustering, while the training data are other clusters; the test and training data are changed *k* times, alternatingly. In this study, the *k* of *k*-means clustering was set to 5, and a random forest algorithm with fivefold cross-validation was used. We performed LOCO CV with shuffled and normalized 276 data, which is the same as the abovementioned training data, including 60 analytical data, seven molecular compositions, and 14 molecular descriptors as all explanatory variables in our model. Then, we compared model performance with each clustered test data and 31 test data, which is the same as the abovementioned test data used in our DNN model.

**Dataset of Polymer Compounds.** In this study, we tested our HSP prediction models against a total of 23 polymers belonging to seven different skeleton classes, with regard to density, refractive index, 1D $^1$H NMR, and 1D $^{13}$C NMR data. The polymers included six polyacrylates [poly-*n*-butylacrylate (PBA), polymethylmethacrylate (PMMA), polyethylmethacrylate (PEMA), poly-*n*-butylmethacrylate (PnBMA), polymethylacrylate (PMA), and polyethylacrylate (PEA)], six polyolefins [polyethylene (PE), polypropylene (PP), polybutadiene, polyisoprene, polychloroprene, and poly-1,1-dimethylethylene], four polyethers [polyethyleneoxide (PEO), polypropylene oxide (PPO), cellulose triacetate (CTA), and polyethersulfone (PES)], two polyesters [polyethyleneterephthalate (PET) and polycaprolactone (PCL)], two polyvinyls [polyvinylacetate (PVAc) and polyvinylchloride (PVC)], two polystyrenes [polystyrene (PS) and polybutadiene-*co*-styrene], and polysiloxane of polydimethylsiloxane (PDMS). In particular, we tested 22 polymers except polyethylene for dD and dH and 22 polymers except cellulose triacetate for dP based on the data available in the literature. Overall, the literature HSP values were obtained from the "Polymer Handbook"[46] and "PolyInfo Database",[47] while those for PnBMA and PET were obtained from other papers.[48,49] The analytical data relative to the refractive index and density were obtained from the "Polymer Handbook" and "PolyInfo Database", while the spectral 1D $^1$H NMR and 1D $^{13}$C NMR values were derived from the "Proton and Carbon NMR Spectra of Polymers"[50] and "PolyInfo Database".

## RESULTS AND DISCUSSION

**DNN Solubility Prediction 1-Step Method Using Analytical Data as Explanatory Variables.** Recently, solubility prediction tools were reported that used structural descriptors or molecular compositions and descriptors, such as RDKIT, CDK, and PaDEL, as training data.[20,23,24] Namely, the input data were based on the chemical formulas, SMILES strings, and so forth; thus, the molecular structure was mostly understood at the linear level. Therefore, this study aimed to predict the solubility (dD, dH, dP, SP, and Log *P*) of substances using only analytical data as input data (Figure 1 Method1, hereinafter called as the "1-step DNN method"). Subsequently, we tried to predict dD, dH, dP, SP, and Log *P* using the DNN with the analytical data as training. However, the results were not sufficiently accurate ranging from 0.35 to 0.53 in $R^2$ (Figure 2a).

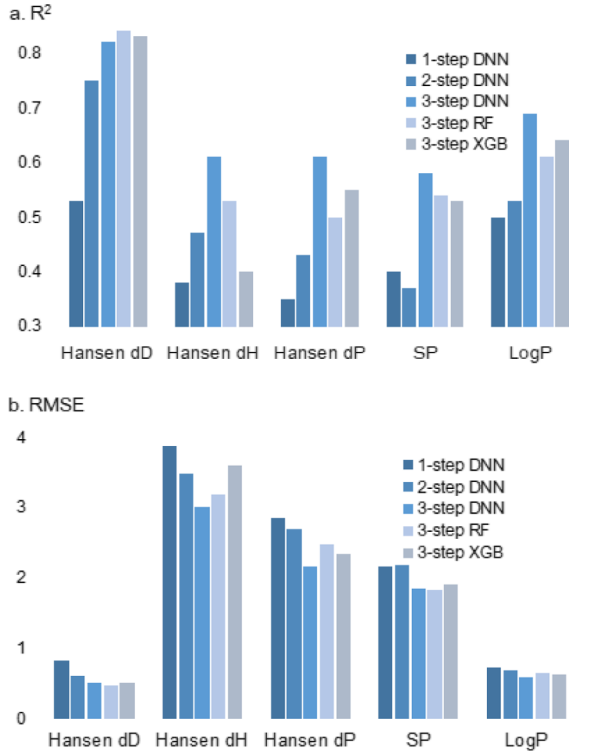



**Figure 2.** Results of $R^2$ and RMSE with test data of each prediction models. (a) Bar chart of each $R^2$ value of solubility predictions, which are for Hansen's solubility parameters (dD, dH, and dP), SP, and Log *P*, with the algorithms of the 1-step DNN method, 2-step DNN method, 3-step DNN method, 3-step random forest method, and 3-step XGBoost method. (b) Bar chart of each RMSE value.

**DNN Solubility Prediction 2-Step Method.** On the basis of previous studies, the prediction accuracy is expected to improve if the molecular information of a substance, such as the molecular composition and molecular descriptors, is used as training data. In this study, our aim was to predict the solubility using only analytical data as input data. Therefore, we attempted to develop a 2-step DNN solubility prediction method, which allows us to predict the solubility from analytical data and predicted intermediate data of molecular composition and molecular descriptors (Figure 1 Method2 and Figure S1a, hereinafter called as the "2-step DNN method"). Concretely, in the first step, we predicted a total of 28 items, namely eight items of molecular composition (described in the Materials and Methods) and 20 items of selected molecular descriptors (described in the Materials and Methods), using the analytical data as training. In these predictions, we used three datasets of test and validation data. One was the same dataset used in the 1-step DNN method. The others were two additional datasets prepared to avoid duplicates in the test set data (described in the Materials and Methods sections; see also Figure S3). Then, we validated a total of three sets in order to


https://doi.org/10.1021/acsomega.1c01035
ACS Omega 2021, 6, 14278−14287

ensure reliability for descriptor selection. According to these results, we extracted available items, for each of which the average value of $R^2$ in the three sets was higher than 0.5 for the numerical prediction, and the average of the lowest values of the PPV (%), NPV (%), recall (%), and specificity (%) in the three sets was more than 50% for the presence/absence prediction. As a result, a total of seven items, namely, a molecular composition N item and six molecular descriptor items (PEOE_VSA1, Chi0v, Chi1v, MolMR, TPSA, and Kappa3), were excluded from the training data in the next step of the prediction since they were below the cutoff value as defined above. On the other hand, the remaining 21 items, which comprise the seven molecular composition items, that is, H, C, S, halogens, −OH, −CHO, and >CO and the 14 molecular descriptor items, including NumHeteroatoms, Chi0n, MaxPartialCharge, MinPartialCharge, SlogP_VSA12, SMR_VSA5, SMR_VSA10, HallKierAlpha, VSA_EState9, NumAromaticRings, NumHAcceptors, NumHDonors, RingCount, and NHOHCount, were selected for use in the next step (Table S5a,b). In the second step, we predicted the solubility associated with dD, dH, dP, SP, and Log *P* using a combination of analytical data, the selected seven molecular compositions, and the selected 14 molecular descriptor items as explanatory variables. In this prediction, we used the same breakdown of the dataset of test and training data as that of the compounds used in the 1-step DNN method. Overall, the values of $R^2$ and RMSE were improved compared to those of the 1-step DNN method, although the values did no exhibit yet satisfactory accuracy except for dD, for which $R^2$ was 0.75 (Figures 2, S4).

**DNN Solubility Prediction 3-Step Method.** As shown in previous studies, solubility predictions based on molecular descriptors have already been investigated.[20,23,24] In this study, the prediction with a 2-step DNN method based on analytical data and predicted values of molecular compositions and molecular descriptors as training was found to be superior than that with the 1-step DNN method using only the analytical data as training. However, the prediction accuracy was not adequate. Therefore, we opted for an alternative 3-step DNN solubility prediction method (Figure 1 Method3 and Figure S1b, hereinafter called as the "3-step DNN method"). In the first step, we predicted the selected seven molecular composition items (described in the DNN solubility prediction 2-step method, Table S5a,b), including the number of H and C and the presence or absence of S, halogens, −OH, −CHO, and >CO, using these analytical data as training data. In the second step, we predicted the selected 14 molecular descriptor items (described in the DNN solubility prediction 2-step method, Table S5a,b) using a combination of analytical data and predicted molecular composition. In the third step, we predicted the solubility associated to dD, dH, dP, SP, and Log *P* with a combination of analytical data, predicted molecular composition, and 14 predicted RDKIT descriptor items. In this prediction, we used the same breakdown of the dataset of test and training data as that used in the 1-step DNN method. The results showed that the $R^2$ values for dD, dH, dP, SP, and Log *P* were 0.81, 0.61, 0.61, 0.58, and 0.69, respectively, which were enhanced values for all items compared to those of the 2-step DNN method (Figures 2, S4). The results of $R^2$ values for them with the random forest using the same 3-step method were 0.84, 0.53, 0.50, 0.54, and 0.61, respectively. In addition, the results of $R^2$ values for them with XGBoost using the same 3-step method were 0.83, 0.40, 0.55, 0.53, and 0.64, respectively. Hence, in this study, these results with the DNN were mostly better than those of random forest and XGBoost algorithms. As same as $R^2$, the results for the RMSE values improved for all items. In particular, the predicted dD, which indicates the dispersion energy, showed a relatively high accuracy. It was assumed that this was due to the use of the experimental refractive index value as training data, which is closely related to the weight per unit volume, density, and dD.[51] Actually, the refractive index is the most important factor in the case of the dD prediction (Figure 3,

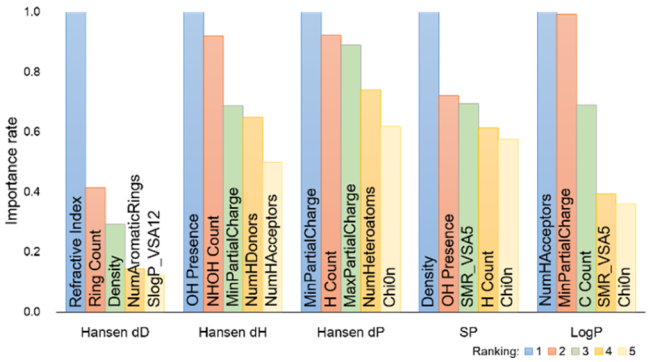



**Figure 3.** Importance of the solubility prediction. As the results of the determination of factor importance for certain attributes, the bar chart shows factors sorted by their importance ranking for each solubility. The checking calculations are performed using the random forest algorithm, which is the same program used in descriptor selection (see also the Materials and Methods section). The descriptors of NHOHCount, NumHAcceptors, NumHDonors, and NumHeteroatoms are the number of −NH and −OH, the number of hydrogen bond acceptors, the number of hydrogen bond donors, and the number of heteroatoms, respectively. The descriptors of Chi0n, MaxPartialCharge, MinPartialCharge, SlogP_VSA12, and SMR_VSA5 are the atomic valence connectivity index, maximum of molecular charge, minimum of molecular charge, MOE-type descriptor of Log *P* and surface area, and MOE-type descriptor of molar refractivity and surface area, respectively.

Table S6a). The dispersion energy dD is a weak intermolecular force that acts even for non-polar molecules, unlike the dipole moment dP. In general, larger molecules exhibit greater intermolecular forces. In other words, the greater the weight per unit volume, the stronger the intermolecular force. Therefore, it can be suggested that a strong relationship occurs between dD and the refractive index. Due to their importance for the dH prediction (Figure 3, Table S6b), the OH-, NH-, and H-related factors are at higher ranks. We believe that these results can be expected due to hydrogen bonding formation. In the case of the dP prediction (Figure 3, Table S6c), the partial charge, H, and number of heteroatoms are at higher ranks of importance. Since dP reflects the polarization rate, it can be assumed that the partial charge gives a large contribution to the dP prediction, and the lightest H atom and heteroatoms with unpaired electrons also have a great effect on the permanent dipole. As the accuracy of all $R^2$ and RMSE obtained with the 3-step DNN method is higher than that of the 2-step DNN method and the values of all $R^2$ are >0.5 (Figures 2, S4), it can be concluded that the solubility prediction of various substances using the 3-step DNN method based only on analytical data as input in the first step is effective. Although we prepared general compounds as a dataset, our models are built from a small dataset, and the


https://doi.org/10.1021/acsomega.1c01035
ACS Omega 2021, 6, 14278−14287

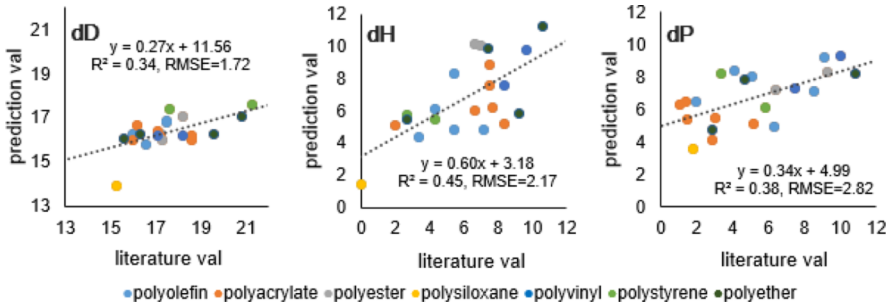



**Figure 4.** Scatter plots of the HSP literature and prediction values for various polymers. Application of our HSP solubility prediction models to 23 common polymers. Here, seven polymer classes are shown using different colors.

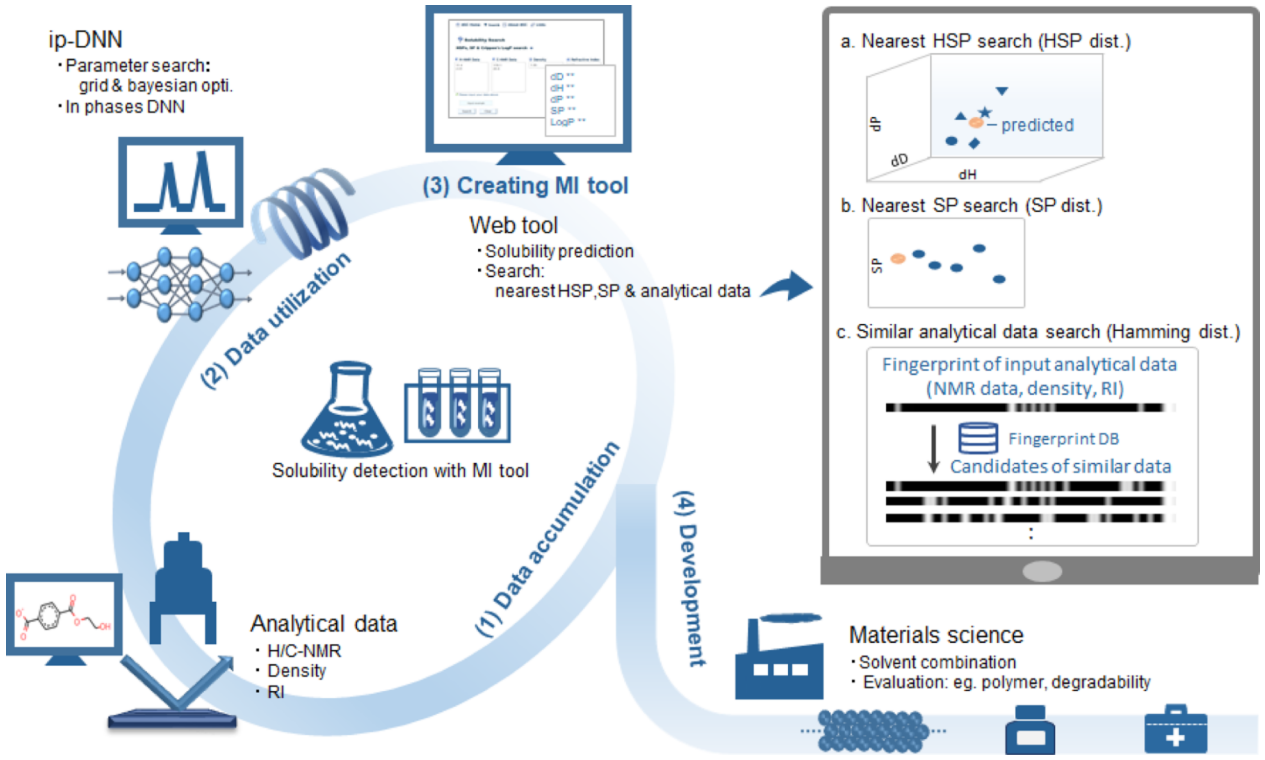



**Figure 5.** Positive cycle of solubility prediction in the materials science industry linked by the MI tool. A positive cycle in the materials science industry linked by an effective MI tool is performed as follows: (1) Due to the development of materials, measurement information of chemical substances is accumulated. (2) The accumulated measurement information is utilized for creating prediction models of chemical properties. (3) MI tools are created using the prediction models. (4) Using the prediction models or MI tools for the development of new materials and technologies, predictive technology is growing. Eventually, the development of predictive technology leads to the effective development of new materials and technologies and further accumulation of measurement information. (a) Visualization function of the "nearest HSP search" on the web tool. The orange circle is a predicted location with HSPs. Other blue symbols are literature locations with HSPs. Evaluation of the solubility among two substances uses the HSP distance. (b) Visualization function of the "nearest SP search" on the web tool. The orange circle is a predicted SP value. Other blue symbols are theoretical SP values. (c) Function of similar analytical data search on the web tool. First, the fingerprints of the analytical data of the user's input data (top) and literature data (database) are prepared. Second, the Hamming distances as evaluation of affinity among two substances are calculated. A substance having a low Hamming distance against the user's input data can be dissolved with a substance of the input data.

prediction performance of our models is not high. Therefore, we re-checked the entire dataset tendency using the LOCO CV method[45] (see the "Confirming Exploration Performance for Dataset" in the Materials and Methods section). Specifically, in this test, we confirmed the availability of our dataset for each solubility prediction using the random forest with the cross-validation method using cluster data as the test data prepared with the *k*-means method. As a result, the prediction performance using our test data was stable for clusters, as a whole; however, in some cases, there were lower values than clustered data (Figure S5). In particular, it seems that the prediction performance of dH is comparatively low. We consider that it is better to use these models to understand solubility tendency. In contrast, we tried creating solubility prediction models with only molecular descriptors, which are the same 14 items of RDKIT's descriptors in this study based on SMILES. The method used the same DNN described in the Materials and Methods section. The $R^2$ of dD, dH, dP, SP, and Log *P* is 0.82, 0.88, 0.91, 0.85, and 0.94, respectively (Figure S6), the performance of which is higher than that of the 3-step DNN for all models. Of note, this approach has been already reported[23,24] and requires SMILES information. As the


https://doi.org/10.1021/acsomega.1c01035
ACS Omega 2021, 6, 14278−14287

difference from our approach, which is prediction from analytical data, we consider that our models are more effective in the research stage such as without SMILES information. Moreover, in this study, creating prediction models step by step successfully increased the performance. This approach is similar to the intermediate supervision deep learning algorithm, which has been frequently used in the image-processing field in recent years.[52,53] Therefore, it is possible to adjust this method to our models. In addition, our stepwise method of DNNs in this study obtained models separately. Creating models with the all-in-one method, such as the abovementioned intermediate supervision deep learning, allows us to obtain an effective system and may improve model performance using interlocking models.

**Application of the HSP Prediction Model to Polymer Data.** The development of novel functional polymeric materials is an important research field that has been actively investigated from several viewpoints, such as the function, environment, and cost reduction. In recent years, a few reports have described solubility prediction approaches, such as machine learning methods using molecular structures, molecular descriptors, and so forth,[24] and molecular dynamics simulations.[6,7] On the other hand, our prediction model differs from other approaches as it exploits only four pieces of analytical data as input, that is, density, refractive index, and top values of the peaks in the 1D $^{1}$H NMR and 1D $^{13}$C NMR spectra. Therefore, it can also predict the solubility parameters from polymer data if these four pieces of analytical data and solubility values are available as inputs and objective variables, respectively. Therefore, we decided to apply our prediction model to polymer data. In this study, we decided to employ only the previously developed HSP (dD, dH, and dP) models as HSP parameters are the most commonly used factors to test the solubility of substances. We prepared a dataset of 23 common polymers including seven classes for testing, the details of which are mentioned in the Materials and Methods section. Upon predicting dD, dH, and dP, $R^2$ was found to be 0.34, 0.45, and 0.38, respectively (Figure 4, Table S7). The result obtained for dH was better than that of dD and dP. It was suggested that dH well reflected the chemical shift since the molecular composition and functional group features for dH were comparatively more important factors than for dD and dP (Figure 3, Table S6). In conclusion, the application of our prediction model to polymers is overall less accurate than for low-molecular weight compounds; however, we believe that it can offer a good estimate of solubility.

**Development of a Web tool and Potential Applications.** In order to allow for an effective use of our prediction models, we developed a freely accessible MI web tool (http://dmar.riken.jp/matsolca/) using the abovementioned regression models, which provides the calculated values of HSPs, SP, and Log *P* as solubility information and the calculated substances with approximate HSPs, SP, and analytical data as solubility-related information. In general, the closer the HSP, SP, and analytical data information among two substances is, the easier they are to dissolve. Therefore, this tool provides not only the solubility prediction values, but also three pieces of additional solubility-related information: (1) the nearest HSPs (Figure 5a), which is the information of the substances with literature HSPs close to the predicted HSPs using the HSP distance,[54] (2) the nearest SP (Figure 5b), which is the information of the substances with a theoretical SP close to the predicted SP using the SP distance that is the absolute value of the difference between two SP values, and (3) similar analytical data (Figure 5c), which is the information of the substances with a similar fingerprint of analytical data between their own database and the user's input data using the Hamming distance.[55]

Herein, we wish to discuss the versatility of this method since the solubility application range is wide. In this study, we succeeded in predicting the solubility features using only analytical information as input data. As for the process, it was not possible to obtain sufficient accuracy using only the analytical data as training data. However, the accuracy was improved using a 3-step DNN method, which utilizes selected and predicted molecular compositions and molecular descriptors in phase as intermediate data for training. Furthermore, we tried to apply this MI tool based on the HSP prediction models to polymer data. By judging from the $R^2$ and scatter plots, the results did not show high accuracy, but a good correlation occurred between literature and prediction values (Figure 4). Therefore, the use of low-molecular weight compounds as training data was sufficient to determine the tendency of solubility of polymers.

Furthermore, we created an efficient and user-friendly MI web tool as a solubility calculator based on our prediction models for users of several fields including industry dealing with solubility-related studies (Figure 5). Commonly, Log *P* is used as a hydrophobicity index for determining the solvent selection, bioaccumulation, and biodegradability,[56−58] while HSPs and SP are used for applications based on the solubility of two substances, such as solvent selection/combination, coating techniques, polymer research, and drug development.[49,59,60] Notably, although HSPs are convenient indices for establishing the degree of solubility between two components, the components can be used even in mixtures. For example, a study revealed that the solubility between an insecticide's solvent as a single component and a cockroach's body surface as a mixture could be evaluated according to the HSPs.[4] Thus, these solubility-related values are widely applicable. In addition, it can be expected that this solubility prediction tool will be used in the biorefinery area, such as biomass recycling, processing, and molding, and in the blue carbon field, including research and development of sea sediments composed of microalgae and seaweeds as a source of $CO_2$ absorption.[61−63] These land-based and water-based biomasses such as polysaccharides and lignin polymers are generally of low solubility;[64−66] therefore, a solubility prediction approach is useful to extend the industrial application in biorefinery processes. Recently, solubility predictions with several machine learning methods were developed and used.[17,23,24,67] However, in comparison with these predictions, our prediction models of solubility have an application advantage since they feature only analytical information as input data. Therefore, it can be expected that our models will find further application in several research and development fields where the solubility of compounds is important. In recent years, the accuracy of the NMR analysis and simplification of related measurements have been improved;[68−70] therefore, it can also be expected that more simple measurements will contribute to the prediction of physical properties such as solubility parameters. Furthermore, it is also anticipated that the creation of an efficient MI tool may lead to the sustainable development of the materials science industry via a positive cycle (ecosystem) including the accumulation of measurement data of chemical substances,


https://doi.org/10.1021/acsomega.1c01035
ACS Omega 2021, 6, 14278−14287

utilization of the data for creating an MI tool of chemical properties, and development of materials science with the data and then again accumulation of measurement data.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c01035.

Overview of the cascaded architecture; importance of the molecular descriptors; dataset for machine learning; results of scatter plots; confirming availability of datasets; scatter plots of prediction and the literature; overview of datasets; NMR assignment items; molecular composition items; total extracted molecular descriptors; predictions for descriptor selection; important factors of solubility predictions; and HSP literature and prediction values for various polymers (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Jun Kikuchi** − *RIKEN Center for Sustainable Resource Sciences, Yokohama, Kanagawa 230-0045, Japan; Graduate School of Medical Life Science, Yokohama City University, Yokohama, Kanagawa 230-0045, Japan; Graduate School of Bioagricultural Sciences, Nagoya University, Nagoya, Aichi 464-0810, Japan;* orcid.org/0000-0002-6809-394X; Phone: +81-45-508-7220; Email: jun.kikuchi@riken.jp

### Authors

**Atsushi Kurotani** − *RIKEN Center for Sustainable Resource Sciences, Yokohama, Kanagawa 230-0045, Japan*

**Toshifumi Kakiuchi** − *AGC Yokohama Technical Center, Yokohama, Kanagawa 230-0045, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c01035


### Author Contributions

This study was designed by all authors. A.K. implemented the models generated with DNN, analyzed the data, developed the web application, and drafted this paper. T.K. collected and analyzed the data. T.K. and J.K. wrote up the paper. All authors have given approval to the final version of the manuscript.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank T. Asakura, K. Ito, T. Matsumoto, and Y. Tsuboi (RIKEN) for their special support with the data collection and calculation and the examination of important factors for the solubility prediction.

## ■ REFERENCES


(1) Hildebrand, J. H.; Scott, R. L. Solutions of Nonelectrolytes. *Annu. Rev. Phys. Chem.* **1950**, *1*, 75.

(2) Hansen, C. M. 50 Years with solubility parameters - past and future. *Prog. Org. Coat.* **2004**, *51*, 77.

(3) Hossin, B.; Rizi, K.; Murdan, S. Application of Hansen Solubility Parameters to predict drug-nail interactions, which can assist the design of nail medicines. *Eur. J. Pharm. Biopharm.* **2016**, *102*, 32.

(4) Kato, Y.; Tsutsumi, S.; Fujiwara, N.; Yamamoto, H. Measurements of the Hansen solubility parameters of mites and cockroaches to improve pest control applications. *Heliyon* **2019**, *5*, No. e01853.

(5) Srinivas, K.; King, J. W.; Monrad, J. K.; Howard, L. R.; Hansen, C. M. Optimization of Subcritical Fluid Extraction of Bioactive Compounds Using Hansen Solubility Parameters. *J. Food Sci.* **2009**, *74*, No. E342.

(6) Faasen, D. P.; Jarray, A.; Zandvliet, H. J. W.; Kooij, E. S.; Kwiecinski, W. Hansen solubility parameters obtained via molecular dynamics simulations as a route to predict siloxane surfactant adsorption. *J. Colloid Interface Sci.* **2020**, *575*, 326.

(7) Chen, X.; Yuan, C.; Wong, C. K. Y.; Zhang, G. Molecular modeling of temperature dependence of solubility parameters for amorphous polymers. *J. Mol. Model.* **2012**, *18*, 2333.

(8) Wilson, G. M.; Deal, C. H. Activity Coefficients and Molecular Structure - Activity Coefficients in Changing Environments - Solutions of Groups. *Ind. Eng. Chem. Fundam.* **1962**, *1*, 20.

(9) Fedors, R. F. Method for Estimating Both Solubility Parameters and Molar Volumes of Liquids. *Polym. Eng. Sci.* **1974**, *14*, 147.

(10) Greenhalgh, D. J.; Williams, A. C.; Timmins, P.; York, P. Solubility parameters as predictors of miscibility in solid dispersions. *J. Pharm. Sci.* **1999**, *88*, 1182.

(11) Naef, R. A Generally Applicable Computer Algorithm Based on the Group Additivity Method for the Calculation of Seven Molecular Descriptors: Heat of Combustion, LogPO/W, LogS, Refractivity, Polarizability, Toxicity and LogBB of Organic Compounds; Scope and Limits of Applicability. *Molecules* **2015**, *20*, 18279.

(12) Stefanis, E.; Panayiotou, C. Prediction of Hansen Solubility Parameters with a New Group-Contribution Method. *Int. J. Thermophys.* **2008**, *29*, 568.

(13) Louwerse, M. J.; Maldonado, A.; Rousseau, S.; Moreau-Masselon, C.; Roux, B.; Rothenberg, G. Revisiting Hansen Solubility Parameters by Including Thermodynamics. *ChemPhysChem* **2017**, *18*, 2999.

(14) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266.

(15) Buonaiuto, M. A.; Lang, A. S. I. D. Prediction of 1-octanol solubilities using data from the Open Notebook Science Challenge. *Chem. Cent. J.* **2015**, *9*, 50.

(16) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Cherto, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.* **2017**, *9*, 33.

(17) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757.

(18) Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. A Practical Overview of Quantitative Structure-Activity Relationship. *Excli J.* **2009**, *8*, 74.

(19) Wozniak, M.; Wolos, A.; Modrzyk, U.; Gorski, R. L.; Winkowski, J.; Bajczyk, M.; Szymkuc, S.; Grzybowski, B. A.; Eder, M. Linguistic measures of chemical diversity and the "keywords" of molecular collections. *Sci. Rep.* **2018**, *8*, 7598.

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

(21) Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1.

(22) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466.

(23) Przybylek, M.; Jelinski, T.; Slabuszewska, J.; Ziolkowska, D.; Mroczynska, K.; Cysewski, P. Application of Multivariate Adaptive Regression Splines (MARSplines) Methodology for Screening of Dicarboxylic Acid Cocrystal Using 1D and 2D Molecular Descriptors. *Cryst. Growth Des.* **2019**, *19*, 3876.

(24) Sanchez-Lengeling, B.; Roch, L. M.; Perea, J. D.; Langner, S.; Brabec, C. J.; Aspuru-Guzik, A. A Bayesian Approach to Predict Solubility Parameters. *Adv. Theory Simul.* **2019**, *2*, 1800069.
(25) Komatsu, T.; Ohishi, R.; Shino, A.; Kikuchi, J. Structure and Metabolic-Flow Analysis of Molecular Complexity in a C-13-Labeled Tree by 2D and 3D NMR. *Angew. Chem., Int. Ed.* **2016**, *55*, 6000.
(26) Kikuchi, J.; Yamada, S. NMR window of molecular complexity showing homeostasis in superorganisms. *Analyst* **2017**, *142*, 4161.
(27) Blinov, K. A.; Smurnyy, Y. D.; Elyashberg, M. E.; Churanova, T. S.; Kvasha, M.; Steinbeck, C.; Lefebvre, B. A.; Williams, A. J. Performance validation of neural network based C-13 NMR prediction using a publicly available data source. *J. Chem. Inf. Model.* **2008**, *48*, 550.
(28) Chikayama, E.; Shimbo, Y.; Komatsu, K.; Kikuchi, J. The Effect of Molecular Conformation on the Accuracy of Theoretical H-1 and C-13 Chemical Shifts Calculated by Ab Initio Methods for Metabolic Mixture Analysis. *J. Phys. Chem. B* **2016**, *120*, 3479.
(29) Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminf.* **2019**, *11*, 2.
(30) Gerrard, W.; Bratholm, L. A.; Packer, M. J.; Mulholland, A. J.; Glowacki, D. R.; Butts, C. P. IMPRESSION - prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **2020**, *11*, 508.
(31) Hoffmann, F.; Li, D.-W.; Sebastiani, D.; Brüschweiler, R. Improved Quantum Chemical NMR Chemical Shift Prediction of Metabolites in Aqueous Solution toward the Validation of Unknowns. *J. Phys. Chem. A* **2017**, *121*, 3071.
(32) Ito, K.; Obuchi, Y.; Chikayama, E.; Date, Y.; Kikuchi, J. Exploratory machine-learned theoretical chemical shifts can closely predict metabolic mixture signals. *Chem. Sci.* **2018**, *9*, 8213.
(33) Ito, K.; Tsutsumi, Y.; Date, Y.; Kikuchi, J. Fragment Assembly Approach Based on Graph/Network Theory with Quantum Chemistry Verifications for Assigning Multidimensional NMR Signals in Metabolite Mixtures. *ACS Chem. Biol.* **2016**, *11*, 1030.
(34) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinf.* **2008**, *9*, 400.
(35) Wu, J.; Lorenzo, P.; Zhong, S.; Ali, M.; Butts, C. P.; Myers, E. L.; Aggarwal, V. K. Synergy of synthesis, computation and NMR reveals correct baulamycin structures. *Nature* **2017**, *547*, 436.
(36) Zhang, J.; Terayama, K.; Sumita, M.; Yoshizoe, K.; Ito, K.; Kikuchi, J.; Tsuda, K. NMR-TS: de novo molecule identification from NMR spectra. *Sci. Technol. Adv. Mater.* **2020**, *21*, 552.
(37) Thomson, G. H. The DIPPR(R) databases. *Int. J. Thermophys.* **1996**, *17*, 223.
(38) Blanks, R. F.; Prausnitz, J. M. Thermodynamics of Polymer Solubility in Polar + Nonpolar Systems. *Ind. Eng. Chem. Fundam.* **1964**, *3*, 1.
(39) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868.
(40) *SDBSWeb*. National Institute of Advanced Industrial Science and Technology (AIST). https://sdbs.db.aist.go.jp. 2018.07.18 updated version (accessed on 1 October 2020).
(41) *NMR Tables, $^{1}H/^{13}C$ Chemical Shifts in Organic Compounds*; Bruker Almanac Tables, 2013, T16. https://www.theresonance.com/wp-content/uploads/2013/02/Bruker_Almanac_Tables.pdf.
(42) The RDKit 2019.09.1 documentation."List of Available Descriptors" in "Getting Started with the RDKit in Python". https://www.rdkit.org/docs/. Greg Landrum (accessed 1 August 2020).
(43) Asakura, T.; Date, Y.; Kikuchi, J. Comparative analysis of chemical and microbial profiles in estuarine sediments sampled from Kanto and Tohoku regions in Japan. *Anal. Chem.* **2014**, *86*, 5425.
(44) Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Software* **2008**, *28*, 141475.
(45) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819.
(46) Brandrup, J.; Immergut, E. H.; Grulke, E. A. *Polymer Handbook*, 4th ed.; Wiley-Interscience, 1999.
(47) Polymer Database (PoLyInfo); National Institute for Materials Science. https://polymer.nims.go.jp/en/ 2020.5.18 updated version (accessed on 1 August 2020).
(48) Polymer Information Table Website; Diversified Enterprises. https://www.accudynetest.com/polytable_01.html (accessed on 1 October 2020).
(49) Paseta, L.; Potier, G.; Abbott, S.; Coronas, J. Using Hansen solubility parameters to study the encapsulation of caffeine in MOFs. *Org. Biomol. Chem.* **2015**, *13*, 2480.
(50) Pham, Q. T.; Pétiaud, R.; Waton, H.; Llauro-Darricades, M.-F. *Proton and Carbon NMR Spectra of Polymers*; Wiley, 2002.
(51) Fujiwara, N.; Nishida, T.; Yamamoto, H. Adaptation of Hansen solubility parameter in evaluating transparency of composite materials. *Heliyon* **2019**, *5*, No. e02833.
(52) Qiang, B. H.; Zhang, S. H.; Zhan, Y. S.; Xie, W.; Zhao, T. Improved Convolutional Pose Machines for Human Pose Estimation Using Image Sensor Data. *Sensors* **2019**, *19*, 718.
(53) Li, C.; Zia, M. Z.; Tran, Q.-H.; Yu, X.; Hager, G. D.; Chandraker, M. Deep Supervision with Intermediate Concepts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1828.
(54) Hansen, C. M. *Hansen Solubility Parameters: A User's Handbook*, 2nd ed.; CRC Press, 1999.
(55) Bookstein, A.; Kulyukin, V. A.; Raita, T. Generalized Hamming Distance. *Inf. Retr.* **2002**, *5*, 353.
(56) Bruce, L. J.; Daugulis, A. J. Solvent Selection-Strategies for Extractive Biocatalysis. *Biotechnol. Prog.* **1991**, *7*, 116.
(57) Knudsen, G. A.; Trexler, A. W.; Richards, A. C.; Hall, S. M.; Hughes, M. F.; Birnbaum, L. S. 2,4,6-Tribromophenol Disposition and Kinetics in Rodents: Effects of Dose, Route, Sex, and Species. *Toxicol. Sci.* **2019**, *169*, 167.
(58) Min, K.; Cuiffi, J. D.; Mathers, R. T. Ranking environmental degradation trends of plastic marine debris based on physical properties and molecular structure. *Nat. Commun.* **2020**, *11*, 727.
(59) Peña, M. A.; Reíllo, A.; Escalera, B.; Bustamante, P. Solubility parameter of drugs for predicting the solubility profile type within a wide polarity range in solvent mixtures. *Int. J. Pharma* **2006**, *321*, 155.
(60) Jhamb, S.; Enekvist, M.; Liang, X.; Zhang, X.; Dam-Johansen, K.; Kontogeorgis, G. M. A review of computer-aided design of paints and coatings. *Curr. Opin. Chem. Eng.* **2020**, *27*, 107.
(61) Macreadie, P. I.; Anton, A.; Raven, J. A.; Beaumont, N.; Connolly, R. M.; Friess, D. A.; Kelleway, J. J.; Kennedy, H.; Kuwae, T.; Lavery, P. S.; Lovelock, C. E.; Smale, D. A.; Apostolaki, E. T.; Atwood, T. B.; Baldock, J.; Bianchi, T. S.; Chmura, G. L.; Eyre, B. D.; Fourqurean, J. W.; Hall-Spencer, J. M.; Huxham, M.; Hendriks, I. E.; Krause-Jensen, D.; Laffoley, D.; Luisetti, T.; Marba, N.; Masque, P.; McGlathery, K. J.; Megonigal, J. P.; Murdiyarso, D.; Russell, B. D.; Santos, R.; Serrano, O.; Silliman, B. R.; Watanabe, K.; Duarte, C. M. The future of Blue Carbon science. *Nat. Commun.* **2019**, *10*, 3998.
(62) Saderne, V.; Geraldi, N. R.; Macreadie, P. I.; Maher, D. T.; Middelburg, J. J.; Serrano, O.; Almahasheer, H.; Arias-Ortiz, A.; Cusack, M.; Eyre, B. D.; Fourqurean, J. W.; Kennedy, H.; Krause-Jensen, D.; Kuwae, T.; Layery, P. S.; Lovelock, C. E.; Marba, N.; Masque, P.; Mateo, M. A.; Mazarrasa, I.; McGlathery, K. J.; Oreska, M. P. J.; Sanders, C. J.; Santos, I. R.; Smoak, J. M.; Tanaya, T.; Watanabe, K.; Duarte, C. M. Role of carbonate burial in Blue Carbon budgets. *Nat. Commun.* **2019**, *10*, 1106.
(63) Wei, F.; Ito, K.; Sakata, K.; Date, Y.; Kikuchi, J. Pretreatment and Integrated Analysis of Spectral Data Reveal Seaweed Similarities Based on Chemical Diversity. *Anal. Chem.* **2015**, *87*, 2819.
(64) Okushita, K.; Chikayama, E.; Kikuchi, J. Solubilization Mechanism and Characterization of the Structural Change of


https://doi.org/10.1021/acsomega.1c01035
*ACS Omega* 2021, 6, 14278−14287

Bacterial Cellulose in Regenerated States through Ionic Liquid Treatment. *Biomacromolecules* **2012**, *13*, 1323.
(65) Komatsu, T.; Kikuchi, J. Comprehensive Signal Assignment of C-13-Labeled Lignocellulose Using Multidimensional Solution NMR and C-13 Chemical Shift Comparison with Solid-State NMR. *Anal. Chem.* **2013**, *85*, 8857.
(66) Komatsu, T.; Kobayashi, T.; Hatanaka, M.; Kikuchi, J. Profiling Planktonic Biomass Using Element-Specific, Multicomponent Nuclear Magnetic Resonance Spectroscopy. *Environ. Sci. Technol.* **2015**, *49*, 7056.
(67) 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.
(68) Yu, P.; Xu, Y.; Wu, Z.; Chang, Y.; Chen, Q.; Yang, X. A low-cost home-built NMR using Halbach magnet. *J. Magn. Reson.* **2018**, *294*, 162.
(69) Yamada, S.; Kurotani, A.; Chikayama, E.; Kikuchi, J. Signal Deconvolution and Noise Factor Analysis Based on a Combination of Time-Frequency Analysis and Probabilistic Sparse Matrix Factorization. *Int. J. Mol. Sci.* **2020**, *21*, 2978.
(70) Cobas, C.; Iglesias, I.; Seoane, F. NMR data visualization, processing, and analysis on mobile devices. *Magn. Reson. Chem.* **2015**, *53*, 558.



https://doi.org/10.1021/acsomega.1c01035
*ACS Omega* 2021, 6, 14278−14287