FULL PAPER

Solution Processing

ADVANCED
THEORY AND
SIMULATIONS

www.advtheorysimul.com

A Bayesian Approach to Predict Solubility Parameters

Benjamin Sanchez-Lengeling,* Loïc M. Roch, José Darío Perea, Stefan Langner, Christoph J. Brabec, and Alán Aspuru-Guzik*

Solubility is a ubiquitous phenomenon in many aspects of material science. While solubility can be determined by considering the cohesive forces in a liquid via the Hansen solubility parameters (HSP), quantitative structure–property relationship models are often used for prediction, notably due to their low computational cost. Here, gpHSP, an interpretable and versatile probabilistic approach to determining HSP, is reported. Our model is based on Gaussian processes, a Bayesian machine learning approach that provides uncertainty bounds to prediction. gpHSP achieves its flexibility by leveraging a variety of input data, such as SMILES strings, COSMOtherm simulations, and quantum chemistry calculations. gpHSP is built on experimentally determined HSP, including a general solvents set aggregated from the literature, and a polymer set experimentally characterized by this group of authors. In all sets, a high degree of agreement is obtained, surpassing well-established machine learning methods. The general applicability of gpHSP to miscibility of organic semiconductors, drug compounds, and in general solvents is demonstrated, which can be further extended to other domains. gpHSP is a fast and accurate toolbox, which could be applied to molecular design for solution processing technologies.

1. Introduction

Solubility directly impacts a wide range of phenomena such as the miscibility of liquids, the compatibility and stability of

B. Sanchez-Lengeling, Dr. L. M. Roch, Prof. A. Aspuru-Guzik
Department of Chemistry and Chemical Biology
Harvard University
Cambridge, MA 02138, USA
E-mail: bsanchezlengeling@g.harvard.edu; beangoben@gmail.com

J. D. Perea, S. Langner, Prof. C. J. Brabec
Institute of Materials for Electronics and Energy Technology (i-MEET)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstrasse 7, 91058 Erlangen, Germany

Prof. C. J. Brabec
Bavarian Center for Applied Energy Research (ZAE Bayern)
Immerwahrstrasse 2, 91058 Erlangen, Germany

Prof. A. Aspuru-Guzik
Canadian Institute for Advanced Research
Toronto, Ontario M5G 1Z8, Canada

Prof. A. Aspuru-Guzik
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 3H7, Canada
E-mail: alan@toronto.ca

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adts.201800069

polymers, and the adsorption of solids on surfaces.[1] The fast and accurate prediction of solubility would benefit fields as diverse as organic semiconductors, paint coating, pharmaceuticals, food industry, and cosmetics. As a consequence, a deep understanding and modeling of solubility are of general interest. Predicting solubility is challenging since it depends on the interaction between solute and solvent, along with various other physical and chemical properties.

Current models combine physical description to reproduce experimental observations up to a certain degree. However, simulating all the several parameters involved in describing solubility is currently unfeasible. Therefore many approaches are data driven and come from the machine learning (ML) field, aiming at finding statistical relationships between information and experimental results to produce a predictive model. Herein we suggest gpHSP, a state-of-the-art Bayesian approach for modeling solubility.

Drug discovery has led efforts in the computational prediction of molecular solubility, notably because it is a determining physicochemical factor in the discovery process.[2,3] For example, an accurate prediction of the miscibility would guide the selection of new potential drug compounds, avoiding thus an exhaustive (virtual) screening.[4–6] Recently, with the growing interest in sustainable energy conversion, studies in the clean technology sector started to emerge. Kim et al. reported the usage of free energy of solvation as a proxy descriptor to predict the solubility of organic electrolytes of flow battery electrolyte solutions.[7] The vast field of organic electronics technology also greatly benefits from an improved accuracy in describing solubility, from organic light emitting diodes (OLEDs), organic transistors, and organic solar cells.[8] Recently, the impact of solubility on the performance of polymers and small molecule in solar cell devices was studied.[9,10] Miscibilties between mixture of materials can ultimately drive device performance.[11]

One physical theory that attempts to model solubility are solubility parameters. Hildebrand and Scott first introduced the Hildebrand solubility parameter as a numerical estimate to the degree of interaction between compounds. This concept was further refined by Hansen[1] with Hansen solubility parameters (HSP), where the Hildebrand value is divided into three components: i) dispersion forces, δd, ii) hydrogen bonding, δh, and iii) polarity, δp. The motivation for HSP was to quantify

ADVANCED
SCIENCE NEWS
www.advancedsciencenews.com

ADVANCED
THEORY AND
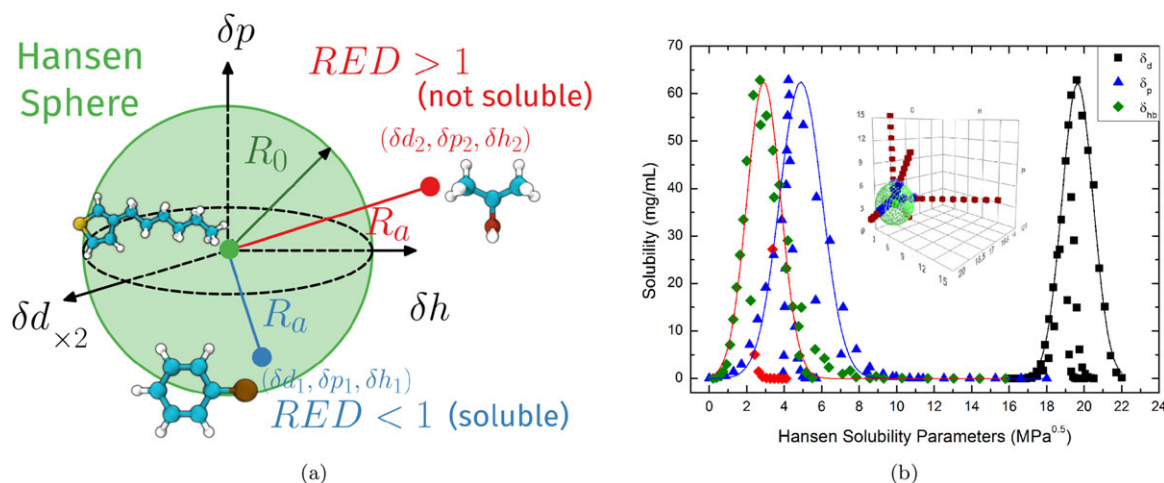**SIMULATIONS**
www.advtheorysimul.com

Figure 1. a) Schematic representation of the Hansen space and how distance in this 3D coordinate space can be used to determine solubility between a solvent and a solute. Note that R_a is the Euclidean distance in Hansen space, and R_0 the interaction radius experimentally measured for the dissolved substance. b) Solubility of oIDTBr as a function of dispersive, polar, and hydrogen bond parameters based on experimental measurements. Inset: Hansen sphere of oIDTBr as calculated by HSPiP, red voxels correspond to solvents outside of the sphere and blue inside and hence soluble.

similarity in solubility and nonsolubility patterns between materials. These three components define a 3D coordinate space called the Hansen space (see **Figure 1**a), where solubility between a solvent and a solute can be determined by the relative energy difference (RED) of the system.

$$RED = \frac{R_a}{R_0}, \quad R_a^2 = 4(\Delta\delta d)^2 + (\Delta\delta p)^2 + (\Delta\delta h)^2 \quad (1)$$

here, R_a is the Euclidean distance in Hansen space and R_0 the interaction radius, a value experimentally measured for the substance being dissolved. When $RED < 1$ there is high likelihood of both compounds to mix, and when $RED > 1$ they are unlikely to mix. Intuitively, RED captures the notion that the compound in Hansen space has to be contained in the Hansen sphere of the solute. Figure 1 illustrates this concept.

The performance of the HSP model has previously been demonstrated on large spectrum of technologies.[9,12–15] While it is possible to calculate HSP via molecular dynamics techniques and structure-interpolating group contribution methods (GCMs),[16,17] these methodologies explore discrete molecular interaction models and require knowledge of the corresponding interaction parameters to obtain thermodynamic properties. To address this limitation here we introduce gpHSP, a theoretical model based on Gaussian processes (GP) regression, which is an interpretable and probabilistic model. We demonstrate that gpHSP does really well in many applications where solubility is key. Notably, gpHSP includes molecular properties that impact solubility, such as the shape and size of molecules, the electrostatic forces, the molecular structure, and the σ-profile. The latter is the distribution function of a molecular surface segment having a specific charge density.[18] These profiles can be calculated by averaging the screening charge densities from surfaces created with the COSMO solvation model.[19] This profile represents a description of polarity properties in molecules. We also show its predictive capabilities when only partial information is included in the model. We show that the model outperforms common ML

baselines while also providing uncertainty estimates on the predictions.

In what follows, we begin with discussing the computational framework designed to predict the Hansen solubility parameters, and detail each of the components of the model. Then we present results obtained with gpHSP and benchmark it against three well-established machine learning algorithms. Before summarizing and drawing our conclusions, we enumerate domains where the model could be applied. Finally, the code is made available on GitHub.

2. Computational Framework for the HSP Prediction

From the structural information to the prediction of the HSP, a computational framework was designed to account for information from multiple sources. The overall workflow is depicted in **Figure 2**, and each of the four components (i.e., molecules, simulation, feature engineering, and predictive model) is briefly detailed hereafter.

2.1. Molecules

Four sets of molecules were used to train and validate gpHSP: 1) The solvents set contains 193 molecules with experimentally determined HSP from several sources.[20–23] The sizes of these molecules range from 2 to 85 heavy atoms. 2) The polymers set consists of 31 polymers and co-polymers with experimentally determined HSP used for model validation. The corresponding computational set utilizes oligomers with up to five monomer units. The chains are capped at a molecular weight of 1500. 3) The Hansen Solubility Parameters in Practice (HSPiP) set[24] contains around 8000 organic chemicals common to a wide array of contexts, 4) which is augmented with a drug-like set contains 64 drug-like molecules.[25,26] Section S1, Supporting Information
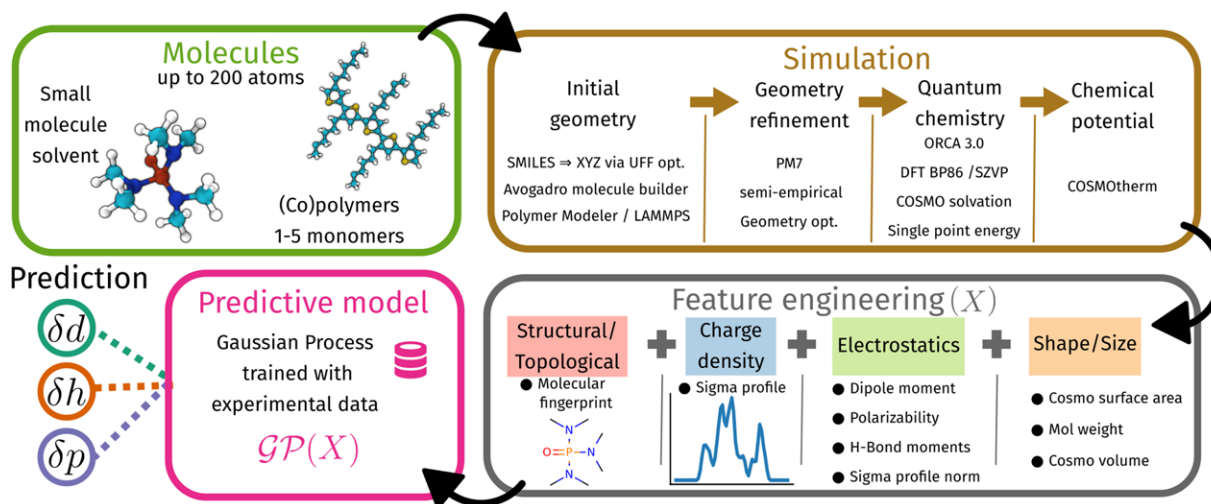
**ADVANCED
SCIENCE NEWS**
www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**
www.advtheorysimul.com

Figure 2. Schematic of the computational framework utilized for molecular simulation, building features, and predicting the Hansen solubility parameters.

outlines the format in which we provide these datasets and their source for experimentally determined data. Furthermore we will elaborate on the methodology for experimentally characterizing HSP.

2.1.1. Experimental Characterization

HSP are typically in MPa$^{0.5}$ units, and if not specified we will report them in order of δd, δh, and δp. Experimental HSP determination is usually done in two steps, measuring patterns of solubility and nonsolubility against a set of desired solvents and then Hansen coordinates inferred via software such as HSPiP or via statistical methods such as fitting Gaussian distributions. Experimental data can be collected by inverse gas chromatography, intrinsic viscosity, solubility measurements, or surface tension. Due to the optical properties of organic semiconductors in the range between ultraviolet and infrared, solubility determination via absorbance measurement is the preferred method of measurement. The foremost is the classical approach by Hansen[1] which assumes a set of solubility/nonsolubility assessments on a set of solvents. A more systematic approach is the binary solvent gradient method developed by Machui et al.,[9] which measures solubility/nonsolubility between the various degrees of mixtures of a set of solvents. The inset in Figure 1b shows measurements between a solute and four solvents (one for each line) at various degrees of mixing. Alternatively, another fast and reliable strategy is to measure the concentration of dissolved material in solvent mixtures after limited periods of several minutes.[27] Furthermore, visual inspection can be used to detect remained unsolved material particles in solution.[14,28]

Independent of the experimental methods, the determination of HSP from solubility values is usually done by one of the following two methods: 1) Use the HSPiP software to transfer the related solvents to the Hansen space and to separate the data into soluble and nonsoluble regions, which allows the user to create a solubility sphere.[1] New coordinates are determined in reference to other known solvents and their HSP.[29] 2) Use the

Gaussian behavior of solubility in the Hansen space to extract the HSP directly from solubility data.[30] For illustration purposes, Figure 1 b depicts the Gaussian fits to solubility measurement for the organic semiconductor oIDTBr. The choice of method can introduce variability, based on the same input data, method (1) leads to HSPs of $(19.54, 4.29, 2.87)\, MPa^{0.5}$, and method (2) to $(19.64, 4.88, 2.91)\, MPa^{0.5}$. Table S1, Supporting Information has eight different measurements for P3HT. We further explore the role of measurement uncertainty in Section 3.

2.2. Simulation

The simulation of the theoretical properties of the molecular systems were obtained at various levels of theory, including force field,[31] semi-empirical,[32–34] and ab initio methods.[35–40] Section S3, Supporting Information contains the statistics used to assess the quality of the geometry optimization methodologies used herein.

For the *solvent* set, the structures generated from a SMILES string were initially optimized with the UFF[31] force field, as implemented in RDKit.[41] Structures were then relaxed in the gas phase using the PM7 semi-empirical method as implemented in OpenMOPAC.[32] Then electronic structure was calculated at the BP86[36,37]/SVP[37] level of theory using the Orca quantum chemical package.[35] σ-profile of the relaxed structures were finally computed with the COSMOtherm13[42] software using BP86/SVP C30 1201 parameterization.[43]

For the *polymer* set, we generated initial structures using the Polymer Modeler in Nanohub,[44] and relax the geometries with LAMMPS.[45] From these initial structures we further refined the geometry via a semi-empirical quantum chemistry calculation. The rationale behind this approach is to balance computational cost and accuracy. To assess the level of theory to be used for optimization we compared relaxed structures obtained with PM7,[33] AM1,[46] DFTB+,[34] with high accuracy B97-D[38]/Def2-TZVPD[39] level of theory. We perform this test on five oligomers consisting of one up to five monomer units. The

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

Gaussian09 software package[40] was used for the DFT calculations. To assess the quality of each approach we look at the distribution of differences between bond lengths, bond angles, and dihedral angles, when compared to the higher level of theory. As shown in Table S2, Supporting Information, DFTB+ was found to be the method of choice for geometry optimization at a fraction of the QM computational cost, with PM7 in second. While DFTB+ was more accurate we experienced more convergence problems with the software and so we utilized PM7. Consequently all the polymers hereafter are optimized at that level. Once the structures of the polymers relaxed, single point energy calculations at the BP86/SVP level of theory (Orca) were performed to determine dipole moment, polarizability, and the COSMO cube file. This file was then processed by COSMOtherm to provide further physicochemical properties such as: σ-profile, hydrogen bond donor/acceptor ability, cavity surface, and volume. The σ-profile of the relaxed structures were computed with the COSMOtherm13[42] software using BP86/SVP C30 1201 parameterization.[43]

2.3. Feature Engineering

For HSP prediction we utilize information that we deemed to be physically relevant to this problem at hand, namely the shape and size of the molecules, the electrostatic forces, the σ-profile, and the molecular structure. Details on each of these features are discussed in this section.

2.3.1. Molecular Fingerprints

To construct the notion of similarity between molecules we utilize molecular fingerprints (FP). Although several varieties of FPs exist, we used Morgan[47] and MACCS. In a general sense, FPs are fixed length vectors that encode the absence or presence of specific molecular environments. We choose Morgan FP since they can represent a wide variety of molecular environments. We set the radius to 8 and the size to 2048. In Morgan FP, for a given atom molecular patterns up to connectivity distance of 8 (radius) are identified, indexed, and hashed to a vector of size 2048. These FP can be binary when only indicating presence or absence, or counted, when counting the number of occurrences of a pattern. Meanwhile, MACCS FP are more specific and concise, they are a 166-length binary vector that encodes mainstay molecular patterns such as the presence of S–S bond or rings of size 4. Computing FPs is straightforward with the RDKit chemoinformatics python module. In order to improve the effectiveness of this representation, we do dimensionality reduction of the fingerprints, utilizing a gradient boosted tree feature selection scheme, reducing the fingerprint to a lower dimension where most of the variability of the data is found.

2.3.2. σ-Profile

The σ-profile is the probability distribution of a molecular surface segment having a specific charge density.[18] These profiles can be calculated by averaging the screening charge densities from surfaces created with the COSMO solvation model.[19] The profile provides a description of polarity properties in molecules.

Previously σ-moments, the first Taylor expansion coefficients of the σ-profile, were utilized as a compact representation to predict HSP.[48] In this case we are utilizing the entire profile; in our datasets we found that all σ-profiles can be constrained to a 61-length continuous vector. Each vector can be interpreted as signal, as such we normalize each vector to unit length. Since most profiles have a similar structure, we subtract the mean signal of all profiles to contrast the differences between profiles. When the vector is normalized, the Euclidean distance between profiles can be interpreted as the cosine distance which indicates similarity among profiles. This formulation can be seen as the discrete version of COSMOtherm's measure of similarity for σ-profiles. We also tested a 6-component non-negative matrix factorization (NMF) as an alternative way to represent in a compact form the σ-profile. NMF finds a basis where most profiles can be reconstructed from 6D weight vectors.

2.3.3. Electrostatic

Additional electrostatic descriptors are obtained from the electronic structure calculations. In particular, the magnitude of the molecular dipole moment and the polarizability results from single point energy calculations in ORCA. From COSMOtherm we also use the hydrogen bonding moments, in particular the first moment for the donor and acceptor parts. These moments are descriptions of the σ-profile calculated only on hydrogen bonding acceptors and donors atoms. Additionally, we also utilize the norm of the σ-profile as a proxy for the magnitude of the polarity properties of a molecule.

2.3.4. Shape and Size

Since the shape and size of the molecules studied herein can vary significantly, in terms of number of atoms (from 5 to 85 heavy atoms per molecule), and in terms of spatial orientation, we looked into quantifying theses measures. Molecular weight relates to the size of a molecule: the bigger the molecules the more atoms it has. For shape we could have utilized more complex data such as 3D fingerprints,[49] but instead we decided on two summary statistics relating to the COSMO solvation surface. COSMO Surface Area and Volume provides indirect information about the shape of a molecule based on the space it occupies (volume) and its arrangement (area).

2.4. Gaussian Process Regression

Our model aims to be interpretable, not only predicting accurate solubilities, but also leveraging different types of data sources along with the experimental data on HSP. The model, referred to as gpHSP, is based on Gaussian processes regression.

GP regression is an established probabilistic framework in the field of machine learning to build flexible models, while
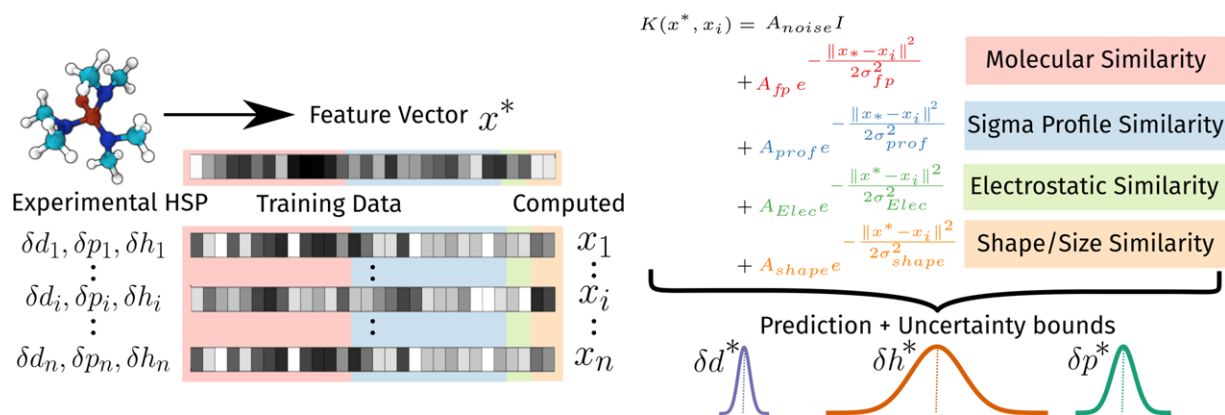
**ADVANCED
SCIENCE NEWS**
www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**
www.advtheorysimul.com

Figure 3. Schematic of prediction with Gaussian processes. On the left we combine several types of information to build our feature vectors in our training set, on the right we highlight the composition of the kernel function combining these different information sources. Each type of feature is highlighted with a color. Predictions with their uncertainty bounds represent a Gaussian distribution.

furnishing uncertainty bounds on predictions[50,51] and resilient to overfitting. A GP represents a family of smooth functions that are determined by a mean function and a covariance function. These two functions are optimized and fitted to a set of training data, and samples from a GP represent possible functions that could model the data at hand. Uncertainty predictions can be obtained when integrating all functions from a GP and calculating their spread, while the predictions will be the mean of all functions. GPs are distinct because of their associated covariance functions (i.e., the kernel). A covariance function is specified and optimized to learn a smooth function that utilizes the similarity of data points for prediction. A popular choice of covariance function is the sum of radial basis functions (RBF), also known as a squared exponential, along with a noise kernel as follows:

$$K(x^*, x) = A e^{-\frac{\|x^* - x\|^2}{2\ell^2}} + A_{\text{noise}} I \qquad (2)$$

This covariance function has three optimizable parameters—the noise variance A_{noise} which allows to quantify the inherent noise in the data, the RBF lengthscale ℓ, and variance A which control the spread of the Gaussian function. A large ℓ indicates the model is more smooth and global, while a smaller ℓ parameter indicates the model utilizes more local information. The $\|x^* - x\|$ part is used as a distance measure between two data points. This can be substituted for other distance or similarity measures. A good sanity metric when looking at GPs is to check that $A > A_{\text{noise}}$, which would indicate the model is based on signal rather than noise. Alternatively, predictions of GP can be interpreted as weighted averages of the training data, where the weights are probabilistic in nature. A schematic of how GP combine different types of information for prediction in the solvent case is outlined in **Figure 3**. The complete equation for the solvent covariance function is

$$K(x^*, x)_{\text{small}} = A_{\text{noise}} I + \sum_{i \in \text{features}} A_{\text{feature}} e^{-\frac{\|x^* - x\|^2}{2\ell^2_{\text{feature}}}} \qquad (3)$$

$$features = \{Mol_{\text{FP}}, \sigma\text{-profile}, \text{electrostatic}, \text{shape/size}\}$$

Equation (3) has nine hyperparameters, which control how distances between observations of data are interpolated and smoothed. These parameters are optimized with gradient descent to improve agreement with data and model flexibility. It is also important to note that GPs are inherently robust to overfitting since the training procedure penalizes more complex models (higher-rank kernels).

Due to different physics that underlie small molecules and polymers, we built an independent model for the polymers set. Due to the disparity between size of datasets between *solvent* and *polymer* sets, we also decided for a more simple model that leverages all types of data while the polymer model utilizes only two types of data sources, amounting to five hyperparameters. These last features were selected via automatic relevance detection (ARD) techniques on GP. In this step each feature is given an individual lengthscale and so when the model is optimized, more relevant features will have larger coefficients which can be selected against lower valued features. This methodology could easily be adopted for other applications where a model might need to incorporate additional information due to the underlying physical phenomena that is being modeled.

Due to the small size of both datasets and to avoid overfitting, we used leave-one-out cross-validation (LOO-CV), where we train our model on all data points except one, and predict its value. This is done for all data points. All reported predictions are always obtained on untrained data.

3. Results and Discussion

We assess the performance of our approach on the collected experimental values described in Section 2.1. We further compare the quality of the prediction against well-established methods, from simple and effective models to state-of-the-art models. In all the reported results, statistics on the error distributions are provided. Comparisons are made only on experimental values since they are the only source of ground truth.

The methods reported include: i) Lasso, which is an L_1-penalized form of linear regression, ii) Kernel Ridge regression (KernelRidge)[52] which is a form of L^2-penalized linear

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

Table 1. Comparison of multiple models predicting HSP. Shaded cells represent the best model for a given property and dataset. Predictions are compared against reported experimental HSP. All statistics are computed via LOO-CV.

Dataset	Target	Approach	MAE	AE σ	RMSE	r	R^2
Solvents $n = 193$	δd	gpHSP	0.68	0.76	1.02	0.84	0.69
		KernelRidge	0.80	0.92	1.22	0.75	0.56
		Lasso	0.91	1.11	1.44	0.63	0.39
		RGF	1.10	1.21	1.64	0.50	0.21
	δp	gpHSP	1.93	2.08	2.83	0.84	0.71
		KernelRidge	2.46	2.77	3.70	0.72	0.50
		Lasso	2.80	3.32	4.34	0.60	0.31
		RGF	2.26	2.37	3.27	0.79	0.61
	δh	gpHSP	1.57	1.83	2.41	0.91	0.83
		KernelRidge	2.25	2.22	3.16	0.84	0.70
		Lasso	2.66	2.38	3.57	0.79	0.62
		RGF	1.96	2.02	2.81	0.87	0.77
Polymers $n = 31$	δd	gpHSP	0.38	0.44	0.58	0.76	0.56
		KernelRidge	0.51	0.38	0.63	0.70	0.48
		Lasso	0.73	0.53	0.90	0.27	−0.06
	δp	gpHSP	1.82	2.03	2.72	0.76	0.58
		KernelRidge	2.58	2.36	3.50	0.58	0.30
		Lasso	2.71	2.31	3.55	0.56	0.28
	δh	gpHSP	1.88	1.93	2.69	0.85	0.62
		KernelRidge	2.22	2.26	3.17	0.69	0.47
		Lasso	2.88	2.97	4.13	0.37	0.10

regression utilizing kernels, and iii) Regularized Greedy Forest (RGF),[53] considered as a state-of-the-art method in ensemble and tree regression. Because of the low number of data points and high dimensionality of the feature vectors, we decided to not include neural networks.

To measure the effectiveness of each method we report on mean absolute error (MAE), standard deviation of absolute error (AE σ) root mean square error (RMSE), Pearson correlation coefficient (r), and the coefficient of determination (R^2). While MAE, AE σ, RMSE give an idea of the error distribution, r and R^2 inform on the quality of fit of each model. r is a measure of how linearly dependent the experimental and predicted values are, while R^2 is a measure of how well the model can capture the inherent variation of the experimental data. R^2 can range from $-\infty$ to 1, from an infinitely worse model to exact prediction. When $R^2 = 0$ the model can predict as well as the mean value of the data, therefore positive values are desired. All these statistics give an overall idea of the type of error that can be expected from each method. Results are summarized in **Table 1**.

Overall, gpHSP outperforms all other baseline methods in the regression metrics. In particular, R^2 is high in all settings ($R^2 > 0.5$), which indicates a strong goodness of fit. **Figure 4** displays a comparison between the experimental and predicted results for all HSP. It can be observed that the fits obtained with gpHSP are in good agreement with experiments, although there are cases where the prediction has a large error but on average the error is lower than the other baseline methods.

3.1. Interpretation of Parameters and Uncertainty

A key feature of gpHSP (and GPs in general) is the uncertainty predictions, which are represented as a gradient in the scatter plots (Figure 4). Points with the largest error tend to have a large associated uncertainty. There are also some cases of accurate predictions with high uncertainty. Intuitively, uncertainty is a measure of how different a data input is related to all previously observed inputs. When doing prediction on real-world data, by looking at the predicted uncertainty bounds and comparing them to the known bounds one can quantify how confident the model is and where more data might be more important to be measured.

By looking at the regression performance of different types of features we can interpret how correlated each parameter is to each source of information (see Tables S4 and S5, Supporting Information for a full tabulation of the statistics). In particular, we notice that molecular structure tends to outperform other forms of information. δd and δh can be predicted quite well just using MACCS information, while it performs quite poorly with δp. For δp, binary Morgan FP works best, which might indicate that more complex molecular patterns are required to correlate HSP with dipolar intermolecular forces.

The second most predictive feature is related to the electrostatics of each molecule. In particular for δp, the σ-profile performed the best, while with δd and δh the electrostatic properties perform better.

These trends can also be observed in the hyperparameters of each kernel. In all cases, structural properties tend to have higher variances and lengthscales. Something to note is that the kernel hyperparameters take into account the joint predictive power of all features combined, instead of independent predictive power as in our initial analysis.

From the hyperparameters, δd appeared to be more dependent on molecular structure, while δp depends on the electrostatic properties. For each target, the information provided by the σ-profile is diminished, indicating that electrostatic properties are preferred, for predictive power, over the σ-profile.

The size and shape feature shows a limited prediction ability. With δd, the variance of this kernel component is 0.0, indicating that it does not contribute to the overall prediction. The amount of non-negative entries in a fingerprint is typically related to the size of a molecule, so we hypothesize that shape/size features are redundant.

3.2. What is an Acceptable Error?

To determine the acceptable error in the $RED = \frac{R_a}{R_0}$ relation (see Equation (1) and **Figure 5**), one needs to have a close look at the solute–solvent distance, R_a, and at the Hansen radius, R_0.

R_a depends on the HSP of two materials. Since the latter are predicted values, and the distance will require a nonlinear combination of six values, the distribution of error in distances can be rather complicated. To better understand how this error is propagated, we computed all possible distances between solute and solvents in the general solvents set for experimental and predicted HSP. The difference between experimental and the
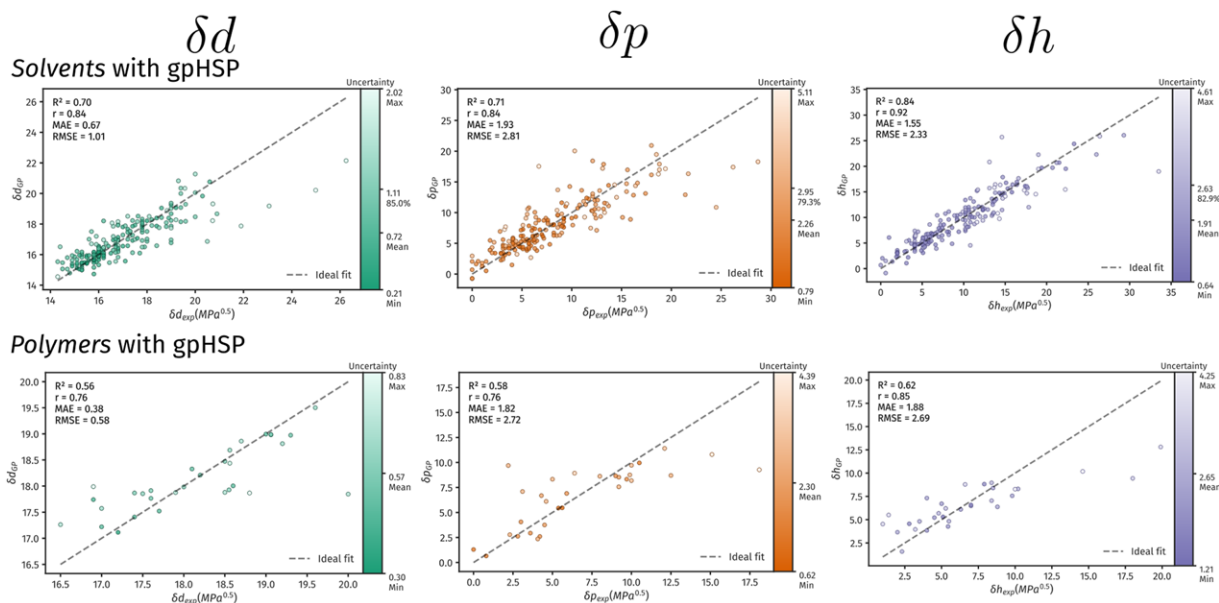
ADVANCED
SCIENCE NEWS
www.advancedsciencenews.com

ADVANCED
THEORY AND
SIMULATIONS
www.advtheorysimul.com

Figure 4. Prediction performance as scatter plots of experimental versus gpHSP predictions for *solvent molecules* (top row) and *polymers* (bottom row) on each parameter. The coloring matches the uncertainty bar. Error statistics are shown in the top right corner.
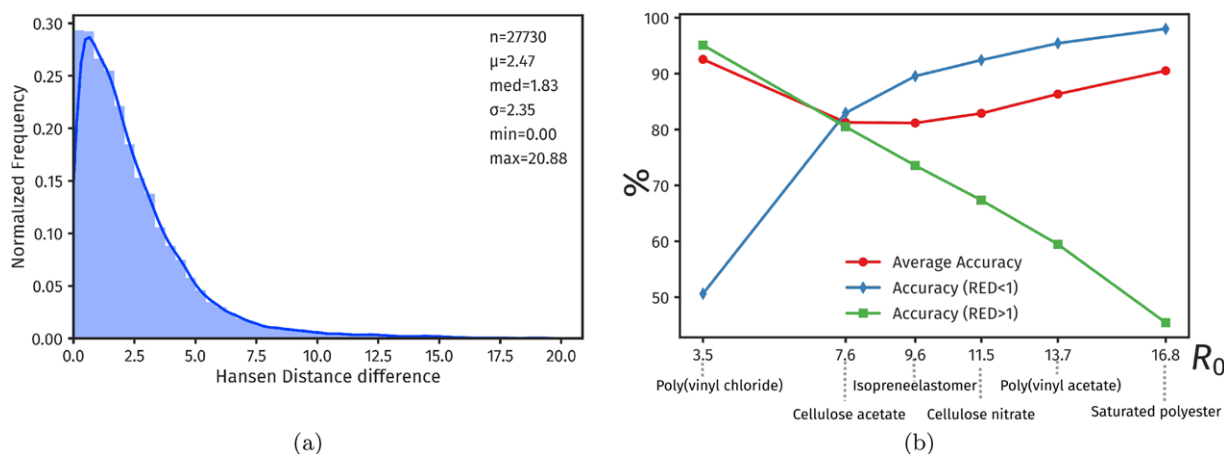

Figure 5. a) Histogram and a Kernel Density Estimation of the distribution of error between predicted and real Hansen distances. In the corner summary statistics are reported. b) Plots of expected average accuracy on each solubility task as a function of R_0. Exemplary compounds and values are shown below on the *x*-axis.

predicted distances is illustrated in Figure S9, Supporting Information, we can expect a mean error of 2.58 with standard deviation of 2.38. In the worst scenario the distance difference is 20.26.

To contextualize the aforementioned numbers, we look at R_0, which strongly depends on general solubility properties of the solute, especially in terms of crystallinity and molecular weight. Typically for compounds with identical HSP, a similar solubility trend to the surrounded solvents is expected, but the absolute solubility can differ enormously. R_0 is normally experimentally characterized, and reported values range from 3.5 to 16.8 and higher, depending on the solute.

It should be noted that there is also uncertainty in the experimental data. Results are significantly dependent on solvent selection, total number of solvents, and HSP accuracy of solvents. Additionally, the choice of experimental method and data evalua-

tion can influence the outcome of the experiment.[9,14,15,27,54] For example, for a single solute, the semiconducting polymer P3HT, several HSP can be found in the literature as illustrated in Table S1, Supporting Information. For P3HT, the reported HSP have a standard deviation of (0.42, 1.03, 1.54) MPa$^{0.5}$ which is comparable to the error obtained with our GP-based model.

By fixing a value of R_0 and applying the RED formula to true and predicted values of HSP, we can ask how likely are we to accurately report a pair as soluble or not soluble. As seen in Figure 5, we consider a range of R_0 values from commercial polymers and look at accuracy scores for each type of prediction. Overall we obtain an average accuracy above 80%, but the accuracy on each individual task (soluble or not) will vary greatly based on R_0. Lower values tend to be less soluble with a random solute and so accuracy for $RED < 1$ is low and $RED > 1$ is high. This trend is

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

inverted for higher values of R_0. This is also a symptom of large modeling error at the extreme values of each property.

Considering the average modeling error in distances is 2.58 we can expect most predictions used with RED to be accurate. Even so, there is still room for improvement to lower the modeling error. This is particularly true when comparing solvents with low and high values of R_0.

3.3. Domains of Application

In addition to the aforementioned comparisons with experimental values, we have applied gpHSP to several domains in order to create datasets where HSP might be useful while highlighting potential applications.

HSPiP has aggregated a set of around 10 000 compounds which are commonly used for solutes/solvents, with a subset of 8800 organic compounds. Each compound has an associated CAS-ID and common name. Using this information we retrieved smile strings from PubChem[55] and filtered on the subset of organic molecules. We make this data available in a comma separated file, including all relevant information needed to do predictions, including the σ-profile and calculated properties.

The upper bound to obtain the properties of a molecule in this set is 5 h on a single core, with an median time of 12 min. When employing only partial information such as FPs, gpHSP takes less than a second to compute the properties of a molecule. What is more, the performance of gpHSP is on par with the baseline methods trained on all the data. Therefore, our approach can be easily utilized for filtering in high throughput virtual screening framework in any material effort.

The most common use of HSP in a pharmaceutical context is in predicting how materials will interact when combined in multi-component formulations,[56] often many of these components are biological (e.g., metabolites, membranes) or material (e.g., keratin, nails).

Although HSPiP contains around 8000 drug-like molecules according to Lipinski's rule of five,[57] we examined drug molecules of interest from recent work. For example, HSP have been utilized to assess pharmacokinetic properties of drugs,[26] determining the most probable absorption sites for each drug and the corresponding metabolite by looking at miscibilities between these two. Hossin et al. looked at predicting nail–drug interactions by comparing drug and nail HSPs.[58]

One other domain of applicability for the HSPs is in the determination of relative miscibilities in organic materials blends,[59] where the interaction parameter of Flory–Huggins $\chi_{i,j}$ can be highly employed.[60–62] Following the Flory–Huggins theory, the interaction parameter for solute:solvent mixtures, can be estimated via the following relation.

$$\chi_{1,2} = \frac{v_{1,2}}{RT}(\delta T_1 - \delta T_2)^2 \tag{4}$$

where $\delta T_i^2 = \delta d_i^2 + \delta p_i^2 + \delta h_i^2$ is the total solubility parameter, R is the gas constant, T is the absolute temperature, and $v_{1,2}$ is the geometric mean of the solute:solvent molar volumes. In organic photovoltaic applications, the interaction parameters $\chi_{i,j}$ play a crucial role in controlling phase behavior

(miscibility) at the equilibrium microphase structure of at least two semiconducting materials with different energy level alignments that maximize charge separation, recombination, and transfer.[63–65] gpHSP could be an crucial tool in the optimization of organic photovoltaics and light emitting devices, which relies on controlling this nanoscale morphology based on the interaction parameter.[66–68] Indeed, gpHSP could be incorporated to a closed-loop approach[69] for a fast screening of potential candidates based on solubility before sending lead blend candidates to an autonomous platform for device optimization.[70] Also, gpHSP finds applications in the design of structural materials, such as optoelectronics, sensors, biocatalysis, and thermal and electromagnetic shielding.[71–73]

4. Conclusion

In conclusion, we presented gpHSP, a probabilistic and interpretable predictive model for HSPs. This model is trained on experimental and theoretical data, and is validated with regression metrics. Our work demonstrates higher predictive power over several baseline models, and can leverage several types of information. If using only topological information, prediction takes less than a second, significantly reducing the timeline of trial-and-error approaches to material synthesis and device fabrication of organic blends. Additionally, it avoids over-fitting and requires a low number of hyperparameters. We computed and predicted properties for several settings where gpHSP might be used: organic photovoltaic and semiconductor design, quantifying drug interactions, and polymer blends.

One important venue for improvement is the availability of large, high quality, open source HSP databases. Current data sources are scattered and as seen previously will have an inherent experimental error. New advances in automated measurements might pave the way for reduced error and high-throughput data collection.

Future directions of this work lie in improving the GP model by, for example, incorporating more relevant information (temperature/pressure, mixture blends), more powerful GP frameworks, and further development of how theory can be extended with measurable uncertainty bounds. On the modeling side, there are several new advances in GPs that could be leveraged such as utilizing heteroscedastic noise models, corregionalized kernels for correlating results from multiple data sources or the use of deep GPs, which offer more powerful Bayesian-based models, but require more data. Finally, the ability to associate predictions with not just point estimates but also uncertainty bounds could be incorporated into HSP theory to provide probabilistic estimates of solubility, and frame this theory in a probabilistic language.

All in all, the authors recommend gpHSP as a toolbox to predict solubility. The code is available and open source to be used and extended.[74]

Conflict of Interest

The authors declare no conflict of interest.

Keywords

[1] C. Hansen, J. Durkee, G. Kontogeorgis, C. Panayiotou, L. Williams, T. Poulsen, H. Priebe, P. Redelius, *Hansen Solubility Parameters*, CRC Press, Boca Raton, **2007**, pp. 1–26.

[2] R. E. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik, J. B. O. Mitchell, *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174.

[3] A. Llinàs, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model.* **2008**, *48*, 1289.

[4] M. A. Mohammad, A. Alhalaweh, S. P. Velaga, *Int. J. Pharm.* **2011**, *407*, 63.

[5] J. H. Fagerberg, E. Karlsson, J. Ulander, G. Hanisch, C. A. S. Bergström, *Pharm. Res.* **2014**, *32*, 578.

[6] A. Avdeef, *Pharm. Res.* **2018**, *35*, 2.

[7] S. Kim, A. Jinich, A. Aspuru-Guzik, *J. Chem. Inf. Model.* **2017**, *57*, 657.

[8] Y. Liang, Z. Xu, J. Xia, S.-T. Tsai, Y. Wu, G. Li, C. Ray, L. Yu, *Adv. Mater.* **2010**, *22*, E135.

[9] F. Machui, S. Langner, X. Zhu, S. Abbott, C. J. Brabec, *Sol. Energy Mater. Sol. Cells* **2012**, *100*, 138.

[10] B. Walker, A. Tamayo, D. T. Duong, X.-D. Dang, C. Kim, J. Granstrom, T.-Q. Nguyen, *Adv. Energy Mater.* **2011**, *1*, 221.

[11] J. D. Perea, S. Langner, M. Salvador, J. Kontos, G. Jarvas, F. Winkler, F. Machui, A. Görling, A. Dallos, T. Ameri, C. J. Brabec, *J. Phys. Chem. B* **2016**, *120*, 4431.

[12] D. T. Duong, B. Walker, J. Lin, C. Kim, J. Love, B. Purushothaman, J. E. Anthony, T.-Q. Nguyen, *J. Polym. Sci. B, Polym. Phys.* **2012**, *50*, 1405.

[13] S. D. Collins, N. A. Ran, M. C. Heiber, T.-Q. Nguyen, *Adv. Energy Mater.* **2017**, *7*, 1602242.

[14] F. Machui, S. Abbott, D. Waller, M. Koppe, C. J. Brabec, *Macromol. Chem. Phys.* **2011**, *212*, 2159.

[15] I. Burgués-Ceballos, F. Machui, J. Min, T. Ameri, M. M. Voigt, Y. N. Luponosov, S. A. Ponomarenko, P. D. Lacharmoise, M. Campoy-Quiles, C. J. Brabec, *Adv. Funct. Mater.* **2013**, *24*, 1449.

[16] M. Williams, N. R. Tummala, S. G. Aziz, C. Risko, J.-L. Brédas, *J. Phys. Chem. Lett.* **2014**, *5*, 3427.

[17] N. R. Tummala, C. Bruner, C. Risko, J.-L. Brédas, R. H. Dauskardt, *ACS Appl. Mater. Interfaces* **2015**, *7*, 9957.

[18] E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandler, C.-C. Chen, M. Zwolak, K. C. Seavey, *Ind. Eng. Chem. Res.* **2006**, *45*, 4389.

[19] A. Klamt, G. Schüürmann, *J. Chem. Soc. Perkin Trans. 2* **1993**, *0*, 799.

[20] S. Abbott, C. M. Hansen, H. Yamamoto, *Hansen Solubility Parameters in Practice*, 3rd ed., Hansen-Solubility.com, **2010**.

[21] P. Bustamante, M. Peña, J. Barra, *Int. J. Pharm.* **2000**, *194*, 117.

[22] C. Rey-Mermet, P. Ruelle, H. Nam-Trân, M. Buchmann, U. W. Kesselring, *Pharm. Res.* **1991**, *8*, 636.

[23] S. Verheyen, P. Augustijns, R. Kinget, G. Van den Mooter, *Int. J. Pharm.* **2001**, *228*, 199.

[24] HSPiP Datasets, https://www.hansen-solubility.com/HSPiP/datasets.php (accessed: March 2018).

[25] L. G. Martini, P. Avontuur, A. George, R. J. Willson, P. J. Crowley, *Eur. J. Pharm. Biopharm.* **1999**, *48*, 259.

[26] D. Obradović, F. Andrić, M. Zlatović, D. Agbaba, *J. Chemom.* **2018**, *32*, e2996.

[27] S. P. Carvalho, E. F. Lucas, G. González, L. S. Spinelli, *J. Braz. Chem. Soc.* **2013**, *24*, 1998.

[28] U. Vongsaysy, B. Pavageau, G. Wantz, D. M. Bassani, L. Servant, H. Aziz, *Adv. Energy Mater.* **2013**, *4*, 1300752.

[29] J. M. Hughes, D. Aherne, J. N. Coleman, *J. Appl. Polym. Sci.* **2012**, *127*, 4483.

[30] S. D. Bergin, Z. Sun, D. Rickard, P. V. Streich, J. P. Hamilton, J. N. Coleman, *ACS Nano* **2009**, *3*, 2340.

[31] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, W. M. Skiff, *J. Am. Chem. Soc.* **1992**, *114*, 10024.

[32] J. J. P. Stewart, *MOPAC (Molecular Orbital PACkage)*, http://OpenMOPAC.net, MOPAC2016, Stewart Computational Chemistry, Colorado Springs, CO, USA **2016**.

[33] J. J. P. Stewart, *J. Mol. Model.* **2012**, *19*, 1.

[34] B. Aradi, B. Hourahine, T. Frauenheim, *J. Phys. Chem. A* **2007**, *111*, 5678.

[35] F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *2*, 73.

[36] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098.

[37] J. P. Perdew, *Phys. Rev. B* **1986**, *33*, 8822.

[38] S. Grimme, *J. Comput. Chem.* **2006**, *27*, 1787.

[39] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

[40] Gaussian 09, Revision A.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, et al., Gaussian, Inc., Wallingford, CT **2016**.

[41] RDKit. http://www.rdkit.org

[42] COSMOtherm, Version C3.0, Release 17.01, COSMOlogic GmbH & Co KG, https://www.cosmologic.de

[43] A. Klamt, *COSMO-RS*, Elsevier, **2005**, pp. 83–107.

[44] B. P. Haley, N. Wilson, C. Li, A. Arguelles, E. Jaramillo, A. Strachan, "*Polymer Modeler*", https://nanohub.org/resources/polymod, **2018**, https://doi.org/10.4231/D37D2Q93C.

[45] S. Plimpton, *J. Comput. Phys.* **1995**, *117*, 1.

[46] E. Anders, R. Koch, P. Freunscht, *J. Comput. Chem.* **1993**, *14*, 1301.

[47] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742.

[48] G. Járvás, C. Quellet, A. Dallos, *Fluid Phase Equilib.* **2011**, *309*, 8.

[49] S. D. Axen, X.-P. Huang, E. L. Cáceres, L. Gendelev, B. L. Roth, M. J. Keiser, *J. Med. Chem.* **2017**, *60*, 7393.

[50] A. G. D. G. Matthews, M. Van Der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, J. Hensman, *J. Mach. Learn. Res.* **2017**, *18*, 1.

[51] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, **2006**.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.

[53] R. Johnson, T. Zhang, *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 942.

[54] F. Machui, P. Maisch, I. Burgués-Ceballos, S. Langner, J. Krantz, T. Ameri, C. J. Brabec, *ChemPhysChem* **2015**, *16*, 1275.

[55] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202.

[56] B. Hancock, *Int. J. Pharm.* **1997**, *148*, 1.

[57] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **2001**, *46*, 3.

[58] B. Hossin, K. Rizi, S. Murdan, *Eur. J. Pharm. Biopharm.* **2016**, *102*, 32.

[59] S. Ulum, N. Holmes, M. Barr, A. Kilcoyne, B. B. Gong, X. Zhou, W. Belcher, P. Dastoor, *Nano Energy* **2013**, *2*, 897.

[60] D. Leman, M. A. Kelly, S. Ness, S. Engmann, A. Herzing, C. Snyder, H. W. Ro, R. J. Kline, D. M. DeLongchamp, L. J. Richter, *Macromolecules* **2015**, *48*, 383.

[61] D. R. Kozub, K. Vakhshouri, L. M. Orme, C. Wang, A. Hexemer, E. D. Gomez, *Macromolecules* **2011**, *44*, 5722.

[62] J. A. Emerson, D. T. W. Toolan, J. R. Howse, E. M. Furst, T. H. Epps, *Macromolecules* **2013**, *46*, 6533.

[63] C. Zhang, S. Langner, A. V. Mumyatov, D. V. Anokhin, J. Min, J. D. Perea, K. L. Gerasimov, A. Osvet, D. A. Ivanov, P. Troshin, N. Li, C. J. Brabec, *J. Mater. Chem. A* **2017**, *5*, 17570.

[64] S. A. Dowland, M. Salvador, J. D. Perea, N. Gasparini, S. Langner, S. Rajoelson, H. H. Ramanitra, B. D. Lindner, A. Osvet, C. J. Brabec, R. C. Hiorns, H.-J. Egelhaaf, *ACS Appl. Mater. Interfaces* **2017**, *9*, 10971.

[65] N. Li, J. D. Perea, T. Kassar, M. Richter, T. Heumueller, G. J. Matt, Y. Hou, N. S. Güldal, H. Chen, S. Chen, S. Langner, M. Berlinghof, T. Unruh, C. J. Brabec, *Nat. Commun.* **2017**, *8*, 14541.

[66] J. D. Perea, S. Langner, M. Salvador, B. Sanchez-Lengeling, N. Li, C. Zhang, G. Jarvas, J. Kontos, A. Dallos, A. Aspuru-Guzik, C. J. Brabec, *J. Phys. Chem. C* **2017**, *121*, 18153.

[67] L. Ye, B. A. Collins, X. Jiao, J. Zhao, H. Yan, H. Ade, *Adv. Energy Mater.* **2018**, 1703058.

[68] R. A. Segalman, B. McCulloch, S. Kirmayer, J. J. Urban, *Macromolecules* **2009**, *42*, 9205.

[69] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* **2018**, *3*, 5.

[70] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, A. Aspuru-Guzik, *Science Robotics* **2018**, *3*, 19.

[71] A. Maiti, J. Wescott, P. Kung, *Mol. Simul.* **2005**, *31*, 143.

[72] J. W. Gooch, *Encyclopedic Dictionary of Polymers*, 2nd ed., Springer, New York **2011**.

[73] J. Xu, H. Liu, W. Li, H. Zou, W. Xu, *Macromol. Theory Simul.* **2008**, *17*, 470.

[74] Code repository to build models and predict, along with datasets can be found at https://github.com/aspuru-guzik-group/gpHSP.