Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model

Amin Alibakhshi[1,*], Bernd Hartke[1]

Theoretical Chemistry, Institute for Physical Chemistry, Christian-Albrechts-University, Olshausenstr. 40, 24118 Kiel, Germany

Corresponding Author's email: alibakhshi@pctc.uni-kiel.de

Abstract

Theoretical estimation of solvation free energy by continuum solvation models, as a standard approach in computational chemistry, is extensively applied by a broad range of scientific disciplines. Nevertheless, the current widely accepted solvation models are either inaccurate in reproducing experimentally determined solvation free energies or require a number of macroscopic observables which are not always readily available. In the present study, we develop and introduce the Machine-Learning Polarizable Continuum solvation Model (ML-PCM) for a substantial improvement of the predictability of solvation free energy. The performance and reliability of the developed models are validated through a rigorous and demanding validation procedure. The ML-PCM models developed in the present study improve the accuracy of widely accepted continuum solvation models by almost one order of magnitude with almost no additional computational costs. A freely available software is developed and provided for a straightforward implementation of the new approach.

Introduction

Free energy of solvation is one of the key thermophysical properties in studying thermochemistry in solution, where the majority of real-life chemistry happens. In theoretical studies of solution chemistry, estimation of free energies allows evaluation of reaction rates and equilibrium constants of physical or chemical reactions of interest. Nevertheless, direct evaluation of free energies in solution can be quite challenging since it sometimes requires appropriate sampling of phase space [1-3] and appropriate treatment of the non-covalent interactions between the solvent and solute, which can have a remarkable impact on electronic structures of both the solvent and solute and consequently on the microscopic and macroscopic observables [4,5].

Theoretical approaches for evaluating physical chemistry behind solvation free energy can be generally divided into two main categories, namely explicit solvent and implicit solvent approaches. In explicit solvent approaches, solvent molecules are treated explicitly, and the free energy is typically evaluated by analyzing the trajectory of time evolution of phase space obtained via molecular dynamics or Monte Carlo simulations. For that end, a number of efficient free energy estimators have been developed in the past decades such as thermodynamic integration, free-energy perturbation, and histogram analysis methods [11].

Despite obvious advantages of applying the explicit solvent methods such as retaining the physically proper picture of discrete solvent molecules, they suffer by a number of limitations when applied to free-energy estimation. For example, in case of applying methods which evaluate the free energy through alchemical transformations (e.g. thermodynamic integration or free energy perturbation), defining intermediate states and pathways between the endpoints appropriately can be quite tricky [13]. Also, necessity of employing appropriate force fields, which for many solute-solvent mixtures requires to develop or reparametrize a force field, and running the simulations and trajectory analyses can be laborious and time-taking tasks.

To overcome the mentioned limitations, the implicit solvent approach has been developed and is widely applied as standard method for studying solvent effects in computational chemistry. In implicit solvent approaches, the solvent molecules are treated implicitly as a continuous medium and the solute is placed in a cavity of this implicitly defined solvent. The solute-solvent interactions are then evaluated via considering the solvent polarization due to the solute charge distribution and its resulting potential field acting on the solute, known as the reaction field [5]. For a moderate level of theory and medium-sized molecules, implicit solvent approaches can yield a reasonable estimation of the solvation free energy in few seconds to few minutes on a normal desktop PC, while for explicit solvent approaches it might take from hours to days.

The most widely applied implicit solvent approaches are those based on the so-called polarizable continuum model (PCM) proposed by Tomasi and co-workers [14]. In polarizable continuum models, the solvation free energy is constructed by summing the contributions of electrostatic interactions including electronic, nuclear, and polarization interactions (ΔG_{ENP}), changes in free energy by solvent cavity formation, dispersion energy and local solvent structure changes (G_{CDS}), and corrections for differences in molar densities in the two phases compared with the standard state (ΔG°_{cons}). The contributions of electrostatic interactions are evaluated by iteratively solving the following relationship:

$$\Delta G_{ENP} = \langle \Psi^{(1)} \left| H + \frac{1}{2} V \right| \Psi^{(1)} \rangle - \langle \Psi^{(0)} | \, H \, | \Psi^{(0)} \rangle, \qquad (1)$$

which is known as the self-consistent reaction-field (SCRF) calculations [5]. Here, superscripts (0) and (1) refer to the gas and solution phases, respectively, and V is the potential energy operator resulting from the reaction field. Various constructions of the potential energy operator as well as G_{CDS} have resulted in different continuum solvation models. The parallel existence of several continuum solvation models is a good indicator that each of them has its own strengths and weaknesses, and choosing a single, optimal model is not trivial. It is totally impossible to provide a detailed overview here; a 2005 review of implicit solvation models [15] covered 95 pages and cited 936 references. In the present study, we only consider the most widely used PCM-based models.

One of simplest and yet successful continuum solvation models is CPCM which implements the conductor-like screening solvation boundary condition within the PCM framework. In CPCM, the following correction of the polarization charge densities by the scaling factor x is employed [16]:

$$\mathrm{f}(\varepsilon) = \frac{\varepsilon-1}{\varepsilon+\mathrm{x}}, \qquad (2)$$

where ε is the solvent dielectric constant. One main advantage of CPCM is its much simpler defined boundary conditions. More importantly, unlike more advanced PCM-based models which require the normal component of the solute electric field as input, CPCM only requires the solute electrostatic potential; for this reason it is much less affected by outlying charge errors (OCE) [17,18]. A more versatile model exploiting the conductor-like screening solvation boundary condition is COSMO-RS, developed by Klamt and co-workers [19,20], which although initially proposed in 1995, still is one of the most accurate available continuum solvation models. A more sophisticated treatment of the boundary condition is implemented in the integral equation formalism of PCM (IEF-PCM) taking into account apparent surface charge isotropic [25] or anisotropic[26] dielectric continuum solvation. Another extensively used continuum solvation model is the SMx family of methods which specifically focuses on more accurate estimation of the solvation free energy [4,5].

We already discussed the main advantages of continuum solvation models such as their efficiency in terms of computational cost. Nevertheless, it should be noted that all this has become possible for a considerable amount of assumptions and simplifications on the physics of the problem, such as overlooking the conformational entropy of solvent and solute which can have a significant contribution on the total free energy [35], neglecting the site-specific solute-solvent interactions and decoupling the polar and nonpolar components of free energies and considering them independent, linear and additive [36,37]. The inaccuracies resulting from such simplifications are commonly compensated for via incorporating additional macroscopic observables as well as adjustable parameters in the solvation models. In the CPCM model for example, this is achieved by implementing an ad hoc modification of the atomic radii via defining a number of adjustable parameters and empirical descriptors, such as the number of bonded hydrogens and the number of bonded active atoms [16]. In the COSMO-RS model, it is achieved by ad hoc modification of the interaction energies and effective contact area via some adjustable parameters [20].

In contrast, in the SMx family of methods, to provide a more accurate estimation of the solvation free energy, an ad hoc modification of the G_{CDS} term in (1) has been proposed. For that end, employing additional macroscopic observables in the model has been considered [4], including the refractive index, Abraham's hydrogen bond acidity and basicity of the solute, macroscopic surface tension of the solvent at the air/solvent interface at 298.15 K, the square of the fraction of solvent atoms that are aromatic carbon atoms, and the square of the fraction of solvent atoms that are F, Cl, or Br. Although these employed macroscopic observables indirectly introduce more physics into the model and hence provide the chance to make predictions of solvation free energies more universal, except for the last two they are not readily available for many new compounds and their experimental or theoretical evaluation is not straightforward.

In a number of recent studies, Machine Learning (ML) has been exploited to map the highly complicated relationship between solvation free energy and potentially relevant macroscopic or microscopic observables.

Wang et al. employed a pool of 30 molecular representations which all are either per atom reaction field energies or partial charges, as the input of the learning-to rank (LTR) machine learning algorithm, resulting in a root mean squared error (RMSE) of 1.05 kcal/mol [36]. Borhani et al. developed a QSPR model which requires 12 experimentally determined properties of solvent and 9 QM derived representations of solute as model input, yielding a Mean Unsigned Error (MUE) of 0.43 kcal/mol [38]. Hutchinson and Kobayashi proposed a structure property relationship for prediction of hydration free energy which yields a RMSE of 1.65 kcal/mol [39].

The most recent example is the kernel-based machine learning model of Rauer and Bereau which is developed to predict the free energy of solvating small organic molecules containing C, H, O, and N atoms in pure water via implicit-solvent molecular dynamics simulations [40]. For a 39-parameter model they reported a MUE of 1.06 kcal/mol.

A comparison of performance of state of the art ML models developed for solvation free energy prediction provided in table 2 reveals that none of the current ML models yield a remarkable improvement compared to the accuracy achievable by successful continuum solvation models, such as COSMO-RS.

In the present study, we propose a machine-learning-based PCM model, which similar to other conventional continuum solvation models is based on considering the solvent as a continuous medium and calculating the solvation energy components of a solute placed in the cavity of this medium by the SCRF procedure. Nevertheless, unlike the conventional PCM models which propose simple and ad hoc expressions to integrate and modify those calculated energy components, we employ machine learning for this purpose and show its efficiency in substantial improvements of the predictability of solvation free energy.

Methods

Dataset:

To benchmark our results, we used the solvation free energy data of 2493 binary mixtures of 435 neutral solutes and 91 solvents from diverse chemical families available in the Minnesota solvation database [4]. The full list of the studied samples can be found as supplementary material.

Computational details:

The performance of models is reported as mean unsigned error (MUE) and root mean squared error (RMSE) defined as:

$$MUE = \frac{1}{N}\sum\left(\left|y_i^{exp} - y_i^{pred}\right|\right), \qquad (3)$$

$$RMSE = \left(\frac{1}{N}\sum\left(\left(y_i^{exp} - y_i^{pred}\right)^2\right)\right)^{\frac{1}{2}}, \qquad (4)$$

where y_i^{exp} and y_i^{pred} are experimentally determined and predicted solvation free energies, respectively.

Prior to SCRF computations, all solute geometries were optimized in vacuo at the B3LYP/6-31G*level of theory. Using the optimized structures, the SCRF principal energy components listed in table 1 were computed for each compound at the B3LYP/6-31G* and DSD-PBEP86-D3/def2TZVP levels of theory. The latter method as a double hybrid has been shown to yield more precise charge distributions and energy estimations compared to lower-rung DFT or MP2 methods, for a cost comparable to that of the MP2 calculation [41].

Table 1- The components of the continuum solvation model

1	Solvation free energy calculated by the continuum solvation model
2	$<\Psi^{(0)} \mid \; H \; \mid \Psi^{(0)} >$
3	$< \Psi^{(0)} \mid H+V^{(0)}/2 \mid \Psi^{(0)}>$
4	$<\Psi^{(0)} \mid H+V^{(1)}/2 \mid \Psi^{(0)}>$
5	$<\Psi^{(1)} \mid \; H \; \mid \Psi^{(1)}>$
6	$<\Psi^{(1)} \mid H+V^{(1)}/2 \mid \Psi^{(1)}>$
7	Interaction energy of unpolarized solute and polarized solvent
8	Interaction energy of polarized solute and polarized solvent
9	Solute polarization energy
10	Total electrostatic interaction energy
11	Cavity surface area
12	Cavity volume
13	Total kinetic energy
14	Total potential energy
15	Sum of kinetic and potential energy

The SCRF energy components listed in table 1 were computed for two widely accepted polarizable continuum models, namely the IEF-PCM and CPCM, as implemented in Gaussian 16 [42]. For CPCM, the default value of zero is considered as the scaling factor x in relationship (2). However, a value of 0.5 has been shown to be a more reasonable choice for this scaling factor [17,43]. Therefore, in addition to the default implementation of CPCM in Gaussian 16, we also employed a CPCM model with a scaling factor of x=0.5 and denote it by CPCM$_{x=0.5}$. For that, we replaced the original dielectric constant of the solvent with an effective dielectric constant $\tilde{\varepsilon}(\varepsilon, x)$ calculated via:

$$\tilde{\varepsilon}(\varepsilon, x) = \frac{\varepsilon+x}{x+1}\,, \tag{5}$$

as suggested by Klamt et al.[17]. For comparison purposes, we also calculated the solvation free energy via the SMD approach.

We employed feed-forward neural networks to map the relationship between the solvation free energy and the calculated SCRF energy components, which in addition to the solvation free energy estimated by the applied continuum solvation model and to the dielectric constant of the solvent, comprised our model inputs.

The obtained pool of model inputs was further screened using the Minimum Redundancy and Maximum Relevance (MRMR) algorithm [44] resulting in various 8-16 membered combinations of those variables. MRMR is a highly efficient algorithms for selecting most effective sets of variables for developing robust machine-learning-based models [45]. For each number of selected variables, 25 different settings of the MRMR algorithm were applied, distinguished by the employed quantization level, level of dependency, forward or backward variable selection and considering pseudo-samples based on Bayesian statistics or not [44]. In many cases, this resulted in diversely selected set of variables, even for the same applied level of theory and continuum solvation model.

In the next step, various configurations of neural network models were set up and their reliability were examined with a demanding procedure based on the guidelines presented in a previous study [46]. Accordingly, we assigned large parts of the dataset for test (25%) and validation (15 %), and only 60% of the dataset compounds were used for training the models. We employed Levenberg-Marquardt backpropagation and Gradient descent backpropagation training algorithms, and hidden layer transfer functions of the logarithm-sigmoid and tangent-sigmoid types [47]. We only employed neural networks with one hidden layer and set the maximum number of neurons in such a way that the number of training samples be at least ten times the number of neural network weights and bias constants, as a crucially important consideration in developing reliable models [46,48]. For each neural network configuration, training was carried out for 60 randomly selected training, validation and test sets, and for each one 40 different initializations of weight and bias constants of the neural networks were made. Above all, to avoid getting misleading data affected by favorable or unfavorable division of dataset into training, validation and test sets, the post validation strategy proposed in a previous study [46] was carried out. Accordingly, during the initial training of the neural networks, for the models which yielded mean absolute percentage errors lower than 22%, the final optimized weights and bias constants of the neural network models were recorded. These recorded constants were used as the initial guess to train, validate and test the same neural network configurations but under 100 different randomly selected training, validation and test sets. The models for which in at least 80 out of 100 iterations their test and training sets errors had the same means and variances as evaluated by the two sample t-test method with 5% significance level were considered as reliably trained models. For them, the average of the test set results in all repeats were reported as the performance of that model. Setting up and running the neural network models were implemented in Matlab software. A freely available C++ code for practical use of our new ML-PCM models, with detailed user instructions, is provided as supplementary material.

All the computations were carried out on the High Performance Computing center clusters of the Christian-Albrechts-University of Kiel.

Results and discussions

After setting up and training the neural networks and screening the appropriately trained models via the post-validation strategy discussed in the previous section, the best results with MUE of 0.5871 and 0.5303 kcal/mol were observed for the computations at B3LYP/6-31G* and DSD-PBEP86-D3/def2TZVP levels of theory, respectively. The two models employed SCRF energy components and solvation free energy computed via PCM solvation model in both cases and are denoted by ML-PCM(B3LYP) and ML-PCM(DSD-PBEP86) hereafter, respectively. Details of the selected input variables and implementation instructions for all selected models are provided as supplementary materials. These results show a substantial improvement compared to the original continuum solvation model PCM, which for the same dataset yielded MUE of 2.9054 and 3.1569 kcal/mol, respectively.

In comparison to the SMD model, which for the same dataset and solvation free energy computations at B3LYP/6-31G* and DSD-PBEP86-D3/def2TZVP levels yields MUE of 0.78623 and 0.85396 kcal/mol, respectively, the obtained results still show a higher accuracy, without requiring additional solvent parameters needed in the SMD approach. Although in terms of MUE, the COSMO-RS model with 0.4214 kcal/mol still provides better results compared to the ML-PCM models, in terms of maximum unsigned error, the two ML-PCM models which yield maximum unsigned error of 6.1383 and 5.3934 kcal/mol, respectively, are more accurate than that of COSMO-RS for which this value is 6.8701 kcal/mol. For other continuum solvation models studied for the same dataset, the maximum unsigned error of the SMD, PCM, CPCM and $\text{CPCM}_{x=0.5}$ were 11.311, 12.75, 12.2, 12.6 kcal/mol for B3LYP/6-31G* and 11.311, 12.83, 12.31, 12.68 kcal/mol for DSD-PBEP86-D3/def2TZVP levels of theory, which are all substantially higher than those achievable by the ML based models.

The higher accuracy of the predicted solvation free energies by COSMO-RS model also motivated us to study neural networks which take SCRF energy components computed via PCM or CPCM models in addition to the solvation free energies predicted via COSMO-RS as neural network feeds. For these updates, the best results with MUEs of 0.3081 and 0.30208 kcal/mol and maximum unsigned errors of and 3.9195 and 3.3147 kcal/mol were obtained for energy components again calculated via PCM in both cases and computations at B3LYP/6-31G* and DSD-PBEP86-D3/def2TZVP levels of theory, respectively. These two models, which are denoted ML-PCM/COSMO-RS(B3LYP) and ML-PCM/COSMO-RS(DSD-PBEP86) hereafter, respectively, show a remarkable improvement in predicted solvation free energy compared to those obtained via the original implementation of COSMO-RS. This implies considerable flexibility of the proposed approach in improving accuracy of various solvation models.

The overall results obtained via newly developed ML models are compared with various other models proposed in the literature in table 2. Although a more informative comparison would be possible if different models are compared for the same dataset and, if applicable, the same level of theory, the larger size of the benchmark dataset used in the present study compared to most of the other works confirms the superior accuracy of the newly proposed method compared to the majority of the widely accepted ones. Furthermore, as can be seen in table 2, the results obtained via ML-PCM/COSMO-RS are the most accurate results ever reported for predicting solvation free energy of diverse solute and solvent mixtures.

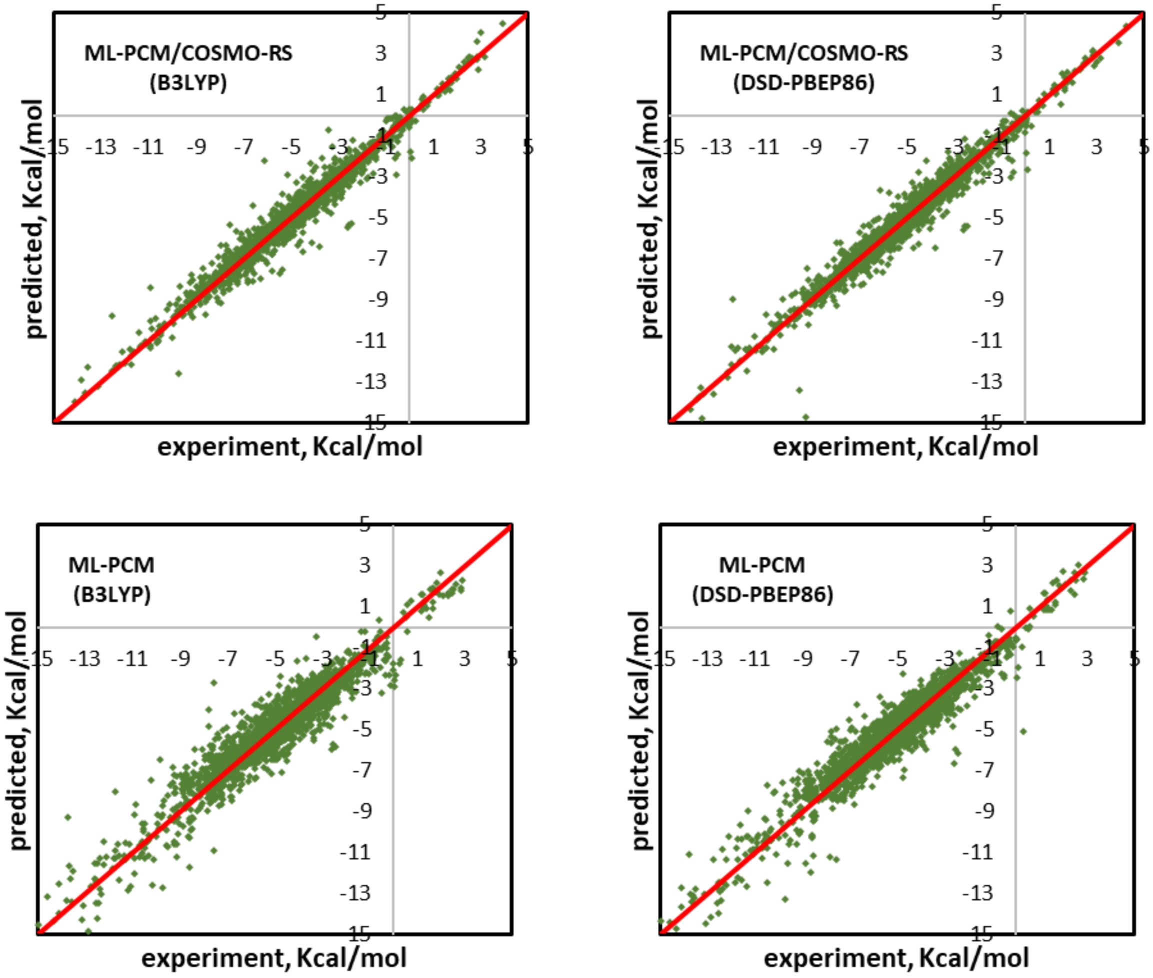

Figure 1 comparison of experimentally determined and predicted solvation free energies

Conclusion

In the present study, we demonstrated substantial improvements of continuum solvation models in evaluating solvation free energy with the help of machine learning. For that end, we proposed a more versatile machine learning assisted integration of the continuum solvation energy components calculated in SCRF computations which can be used to modify the predicted solvation free energy by various solvation models. It allowed us to achieve accurate predictions of solvation free energy with MUE as low as 0.30208 kcal/mol for a large dataset of 2493 binary mixtures of 435 neutral solutes and 91 solvents from diverse chemical families.

Table 2- Comparison of the results of the new method with other models

Method	Source	Nr. Samples	Nr. Solvents	Nr. Solutes	Deviation measure	Deviation (kcal/mol)
ML-PCM/COSMO-RS(DSD-PBEP86)	Present study	2224	88	300	MUE RMSE	0.30208 0.44279
ML-PCM/COSMO-RS(B3LYP)	Present study	2224	88	300	MUE RMSE	0.3081 0.46284
ML-PCM (DSD-PBEP86)	Present study	2488	91	435	MUE RMSE	0.53029 0.73558
ML-PCM (B3LYP)	Present study	2493	91	435	MUE RMSE	0.58705 0.82506
SMD (DSD-PBEP86)	Present study	2488	91	435	MUE RMSE	0.85396 1.3362
SMD (B3LYP)	Present study	2493	91	435	MUE RMSE	0.78623 1.1633
PCM (DSD-PBEP86)	Present study	2488	91	435	MUE RMSE	3.1569 3.6445
PCM(B3LYP)	Present study	2493	91	435	MUE RMSE	2.9054 3.3948
CPCM(DSD-PBEP86)	Present study	2488	91	435	MUE RMSE	2.9651 3.4426
CPCM(B3LYP)	Present study	2493	91	435	MUE RMSE	2.6942 3.1733
$\text{CPCM}_{x=0.5}$(DSD-PBEP86)	Present study	2488	91	435	MUE RMSE	3.1611 3.6466
$\text{CPCM}_{x=0.5}$(B3LYP)	Present study	2493	91	435	MUE RMSE	2.913 3.3985
COSMO-RS	Klamt and Diedenhofen [49]	2346	91	318	MUE RMSE	0.42145 0.69644
DCOSMO-RS	Klamt and Diedenhofen [49]	2346	91	318	MUE RMSE	0.6584 0.99724
SM12	Marenich et al. [50]	2403	91	352	MUE	0.5457- 0.6717
Structure-Property Relationship	Hutchinson and Kobayashi [39]	—	1 (water)	—	RMSE	1.65
atoms-in-molecules neural network	Zubatyuk et.al. [51]	—	—	414	MUE	1.1
kernel-based machine learning	Rauer and Bereau [40]	355	1 (water)	355	MUE	1.06
QSPR	Borhani et. al. [38]	1777	210	295	MUE RMSE	0.43 0.52
Feature Functional Theory	Wang et. al.[36]	668	1 (water)	668	RMSE	1.05

Data availability

All data produced in this study are available and can be provided by contacting the corresponding author.

Code availability

The source file of the C++ code developed for implementing the proposed method with detailed used instructions are available as supplementary material or can be provided by contacting the corresponding author.

Reference

1 Dittner, M. & Hartke, B. Globally Optimal Catalytic Fields–Inverse Design of Abstract Embeddings for Maximum Reaction Rate Acceleration. *Journal of chemical theory and computation* **14**, 3547-3564 (2018).

2 Gauthier, J. A., Dickens, C. F., Chen, L. D., Doyle, A. D. & Nørskov, J. K. Solvation effects for oxygen evolution reaction catalysis on IrO2 (110). *The Journal of Physical Chemistry C* **121**, 11455-11463 (2017).

3 Sakong, S. & Groß, A. The importance of the electrochemical environment in the electro-oxidation of methanol on Pt (111). *ACS catalysis* **6**, 5575-5586 (2016).

4 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. *The Journal of Physical Chemistry B* **113**, 6378-6396 (2009).

5 Cramer, C. J. & Truhlar, D. G. A universal approach to solvation modeling. *Accounts of chemical research* **41**, 760-768 (2008).

6 Grimme, S. Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chemistry–A European Journal* **18**, 9955-9964 (2012).

7 Jensen, J. H. Predicting accurate absolute binding energies in aqueous solution: thermodynamic considerations for electronic structure methods. *Physical Chemistry Chemical Physics* **17**, 12441-12451 (2015).

8 Ho, J. & Ertem, M. Z. Calculating free energy changes in continuum solvation models. *The Journal of Physical Chemistry B* **120**, 1319-1329 (2016).

9 Ho, J. Are thermodynamic cycles necessary for continuum solvent calculation of p K as and reduction potentials? *Physical Chemistry Chemical Physics* **17**, 2859-2868 (2015).

10 Chung, Y., Gillis, R. J. & Green, W. H. Temperature-dependent vapor–liquid equilibria and solvation free energy estimation from minimal data. *AIChE Journal* **66**, e16976 (2020).

11 Chipot, C. & Pohorille, A. *Free energy calculations*. (Springer, 2007).

12 Andreussi, O. *et al.* Solvent-aware interfaces in continuum solvation. *Journal of chemical theory and computation* **15**, 1996-2009 (2019).

13 Pohorille, A., Jarzynski, C. & Chipot, C. Good practices in free-energy calculations. *The Journal of Physical Chemistry B* **114**, 10235-10253 (2010).

14 Miertuš, S., Scrocco, E. & Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics* **55**, 117-129 (1981).

15 Tomasi, J. THEOCHEM 1999, 464, 211;(b) J. Tomasi, B. Mennucci, R. Cammi. *Chem. Rev* **105**, 2999 (2005).

16 Barone, V. & Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *The Journal of Physical Chemistry A* **102**, 1995-2001 (1998).

17 Klamt, A., Moya, C. & Palomar, J. A comprehensive comparison of the IEFPCM and SS (V) PE continuum solvation methods with the COSMO approach. *Journal of chemical theory and computation* **11**, 4220-4225 (2015).

18 Klamt, A. & Jonas, V. Treatment of the outlying charge in continuum solvation models. *The Journal of chemical physics* **105**, 9972-9981 (1996).

19 Klamt, A. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *The Journal of Physical Chemistry* **99**, 2224-2235 (1995).

20 Klamt, A., Jonas, V., Bürger, T. & Lohrenz, J. C. Refinement and parametrization of COSMO-RS. *The Journal of Physical Chemistry A* **102**, 5074-5085 (1998).

21 Zhan, C. *et al.* Specific ion effects at graphitic interfaces. *Nature communications* **10**, 1-8 (2019).

22 Fedele, L. *et al.* Disease-associated missense mutations in GluN2B subunit alter NMDA receptor ligand binding and ion channel properties. *Nature communications* **9**, 1-15 (2018).

23 Dhanker, R. *et al.* Large bipolaron density at organic semiconductor/electrode interfaces. *Nature communications* **8**, 1-7 (2017).

24 Wills, L. A. *et al.* Group additivity-Pourbaix diagrams advocate thermodynamically stable nanoscale clusters in aqueous environments. *Nature communications* **8**, 1-7 (2017).

25 Mennucci, B., Cammi, R. & Tomasi, J. Excited states and solvatochromic shifts within a nonequilibrium solvation approach: A new formulation of the integral equation formalism method at the self-consistent field, configuration interaction, and multiconfiguration self-consistent field level. *The Journal of chemical physics* **109**, 2798-2807 (1998).

26 Cances, E., Mennucci, B. & Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *The Journal of chemical physics* **107**, 3032-3041 (1997).

27 Wu, Y. *et al.* A multistage rotational speed changing molecular rotor regulated by pH and metal cations. *Nature communications* **9**, 1-10 (2018).

28 Wang, G. *et al.* Organocatalytic asymmetric N-sulfonyl amide CN bond activation to access axially chiral biaryl amino acids. *Nature communications* **11**, 1-10 (2020).

29 Ma, J. *et al.* Observation of dissociative quasi-free electron attachment to nucleoside via excited anion radical in solution. *Nature communications* **10**, 1-7 (2019).

30 Bag, S. *et al.* Palladium-catalyzed meta-C–H allylation of arenes: A unique combination of pyrimidine-based template and hexafluoroisopropanol. *Journal of the American Chemical Society* (2020).

31 Yang, Y. *et al.* Unusual KIE and dynamics effects in the Fe-catalyzed hetero-Diels-Alder reaction of unactivated aldehydes and dienes. *Nature communications* **11**, 1-10 (2020).

32 Murugesan, K. *et al.* Homogeneous cobalt-catalyzed reductive amination for synthesis of functionalized primary amines. *Nature communications* **10**, 1-9 (2019).

33 Guo, L. *et al.* Nickel-catalyzed Suzuki–Miyaura cross-couplings of aldehydes. *Nature communications* **10**, 1-6 (2019).

34 Bai, D.-C. *et al.* Palladium/N-heterocyclic carbene catalysed regio and diastereoselective reaction of ketones with allyl reagents via inner-sphere mechanism. *Nature communications* **7**, 11806 (2016).

35 Suárez, E., Díaz, N. & Suárez, D. Entropy calculations of single molecules by combining the rigid–rotor and harmonic-oscillator approximations with conformational entropy estimations from molecular dynamics simulations. *Journal of chemical theory and computation* **7**, 2638-2653 (2011).

36 Wang, B., Wang, C., Wu, K. & Wei, G. W. Breaking the polar-nonpolar division in solvation free energy prediction. *Journal of computational chemistry* **39**, 217-233 (2018).

37 Dzubiella, J., Swanson, J. M. & McCammon, J. Coupling hydrophobicity, dispersion, and electrostatics in continuum solvent models. *Physical review letters* **96**, 087802 (2006).

38 Borhani, T. N., García-Muñoz, S., Luciani, C. V., Galindo, A. & Adjiman, C. S. Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Physical Chemistry Chemical Physics* **21**, 13706-13720 (2019).

39 Hutchinson, S. T. & Kobayashi, R. Solvent-specific featurization for predicting free energies of solvation through machine learning. *Journal of chemical information and modeling* **59**, 1338-1346 (2019).

40 Rauer, C. & Bereau, T. Hydration free energies from kernel-based machine learning: Compound-database bias. *The Journal of Chemical Physics* **153**, 014101 (2020).

41 Kozuch, S. & Martin, J. M. DSD-PBEP86: in search of the best double-hybrid DFT with spin-component scaled MP2 and dispersion corrections. *Physical Chemistry Chemical Physics* **13**, 20104-20107 (2011).

42 Frisch, M. *et al.* (Gaussian, Inc. Wallingford, CT, 2016).

43 Cossi, M., Rega, N., Scalmani, G. & Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *Journal of computational chemistry* **24**, 669-681 (2003).

44 Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* **27**, 1226-1238 (2005).

45 Brown, G. in *Artificial intelligence and statistics.* 49-56.

46 Alibakshi, A. Strategies to develop robust neural network models: Prediction of flash point as a case study. *Analytica chimica acta* **1026**, 69-76 (2018).

47 Demuth, H. & Beale, M. Neural Network Toolbox For Use with Matlab--User'S Guide Verion 3.0. (1993).

48 Kline, R. B. *Principles and practice of structural equation modeling*. (Guilford publications, 2015).

49 Klamt, A. & Diedenhofen, M. Calculation of solvation free energies with DCOSMO-RS. *The Journal of Physical Chemistry A* **119**, 5439-5445 (2015).

50 Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Generalized born solvation model SM12. *Journal of Chemical Theory and Computation* **9**, 609-620 (2013).

51 Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science advances* **5**, eaav6490 (2019).

Acknowledgements

The Authors wish to thank Karsten Balzer and Simone Knief at the high performance computing center of Kiel University for their support and assistance in running the computations there.

Author Contributions

Amin Alibakhshi has contributed to method development, carried out the computations and contributed to writing the manuscript. Bernd Hartke supervised the project and contributed to method development and writing the manuscript.

Competing Interests

The authors declare no competing interests.

Supplementary Information

The details of studied samples are provided as supplementary information.