Lim and Jung *J Cheminform* (2021) 13:56
https://doi.org/10.1186/s13321-021-00533-z

Journal of Cheminformatics

RESEARCH ARTICLE

Open Access

MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning

Hyuntae Lim and YounJoon Jung*

Abstract

Recent advances in machine learning technologies and their applications have led to the development of diverse structure–property relationship models for crucial chemical properties. The solvation free energy is one of them. Here, we introduce a novel ML-based solvation model, which calculates the solvation energy from pairwise atomistic interactions. The novelty of the proposed model consists of a simple architecture: two encoding functions extract atomic feature vectors from the given chemical structure, while the inner product between the two atomistic feature vectors calculates their interactions. The results of 6239 experimental measurements achieve outstanding performance and transferability for enlarging training data owing to its solvent-non-specific nature. An analysis of the interaction map shows that our model has significant potential for producing group contributions on the solvation energy, which indicates that the model provides not only predictions of target properties but also more detailed physicochemical insights.

Introduction

The importance of solvation or hydration mechanisms and their accompanying free energy change has rendered in silico calculation methods for the solvation energy one of the most important applications in computational chemistry [3, 7, 12, 13, 15, 16, 18, 19, 21, 23, 33–37, 39, 40, 43, 44, 50, 52, 54, 57, 58, 65, 67, 71, 73, 79, 81]. Solvation free energy directly influences numerous chemical properties in condensed phases and plays a dominant role in various chemical reactions, such as drug delivery [18, 21, 51, 67], organic synthesis [53], electrochemical redox reactions [1, 30, 47, 72], etc.

Atomistic computer simulation approaches directly provide the microscopic structure of the solvent shell, which surrounds solute molecules [12, 21, 27, 36, 65, 81]. The solvation shell structure offers detailed physicochemical information, such as microscopic mechanisms on solvation or the interplay between the solvent and the solute molecules when using an appropriate force field and molecular dynamics parameters. However, the *explicit solvation* methods mentioned above require extensive numerical calculations as each individual molecular system must be simulated. Practical problems in the explicit solvation model restrict its applications to simulations of classical molecular mechanics [12, 65, 81] or to a limited number of QM/MM approaches [27, 36].

In classical mechanics approaches for macromolecules or calculations for small compounds at the quantum-mechanical level, the concept of *implicit solvation* enables calculation of the solvation free energy with feasible time and computational costs when one considers a given solvent as a continuous and isotropic medium, whose behavior is described by the Poisson–Boltzmann equation [16, 23, 33–35, 39, 40, 43, 54, 73]. Numerous theoretical advances have been made to construct the continuum solvation model, which involves parameterized solvent properties: the

*Correspondence: yjjung@snu.ac.kr
Department of Chemistry, Seoul National University, Seoul 08826, South Korea

polarizable continuum model (PCM) [43], the conductor-like screening model (COSMO) [35] and its variations [32, 34], generalized Born approximations, such as solvation model based on density (SMD) [39] or solvation model 6, 8, 12, etc. (SMx) [16, 40].

The structure–property relationship (SPR) via machine learning is rather a novel approach, which predicts the solvation free energy from a completely different point of view compared to computer simulation approaches with precisely defined theoretical backgrounds [11, 74]. Although we may not expect to obtain detailed chemical or physical insights other than the target property because this is a regression analysis in its nature, SPR has demonstrated significant potential in terms of transferability and outstanding computational efficiency [11, 74, 79]. Recent progress in machine learning (ML) techniques [59] and their implementation in computational chemistry [8, 79] are currently promoting broad applications of SPR in numerous chemical studies [1–4, 7, 9, 10, 15, 18, 20, 26, 28, 29, 50–52, 55, 56, 58, 60–62, 64, 66, 68, 71, 75–78, 82]. These studies show that ML guarantees faster calculations than computer simulations and more precise estimations than traditional SPR estimations; a considerable number of models showed accuracies comparable to ab initio solvation models in the aqueous system [79].

Previously, we introduced a novel artificial neural-network-based ML solvation model called *Delfos*, which predicts free energies of solvation for generic organic solvents [37]. The model not only has a significant potential for showing an accuracy comparable to the state-of-the-art computational chemistry methods [33, 40], but also offers information by which substructures play a dominant role in the solvation process. Herein we propose a novel approach to the ML model for the solvation energy estimation called *MLSolvA*, which is based on the group-contribution method. The key idea of the proposed model is the calculation of pairwise atomic interactions by mapping them into inner products of atomic feature vectors, while each encoder network for the solvent and the solute extracts such atomic features. We believe that the proposed approach presents a powerful tool for understanding solvation processes and is capable of strengthening various solvation models via computer simulations.

The paper is constructed as follows: in "Methods" section, we introduce the theoretical background of applied ML techniques and the overall architecture of our proposed model. "Results and discussion" section quantifies the model's prediction performance with 6239 data points, mainly focusing on pairwise atomic interactions and corresponding group contributions on the solvation free energy. "Conclusions" section summarizes and concludes our work.

Methods

Model architecture

In the proposed model, the linear regression task of calculating the solvation free energy between the given solvent and solute molecules starts with embedded atomistic vector representations [25, 37] of the solvent molecule consisting of $\mathbf{x}_\alpha$'s and the solute molecule consisting of $\mathbf{y}_\beta$'s, where α and β are the atom indices. Then we can describe the given molecule as a tensor, which is a collection (or a sequence) of atomistic vectors:

$$\mathbf{X} = \{\mathbf{x}_\alpha\} \quad \alpha \in \{1, \ldots, M_a\}, \tag{1a}$$

$$\mathbf{Y} = \left\{\mathbf{y}_\beta\right\} \quad \beta \in \{1, \ldots, M_b\}, \tag{1b}$$

where $\mathbf{x}_\alpha$ and $\mathbf{y}_\beta$ are the α-th row of $\mathbf{X}$ and the β-th row of $\mathbf{Y}$, respectively. Here, dimensions of two tensors are $M_a \times D$ for $\mathbf{X}$ and $M_b \times D$ for $\mathbf{Y}$, where M_a and M_b are the sizes of the given solvent and solute (by heavy atom count), and D is the embedding dimension. Then, the encoder function learns of their chemical structures and extracts feature tensors $\mathbf{P}$ for the solvent and $\mathbf{Q}$ for the solute,

$$\mathbf{P} = \{\mathbf{p}_\alpha\} = \text{Encoder}(\mathbf{X}), \tag{2a}$$

$$\mathbf{Q} = \left\{\mathbf{q}_\beta\right\} = \text{Encoder}(\mathbf{Y}). \tag{2b}$$

Dimensions of $\mathbf{P}$ and $\mathbf{Q}$ are $M_a \times N$ and $M_b \times N$, respectively. The numbers of rows are invariable because the encoder function should preserve the topological structure of the given molecule, however, the column dimension, D can differ with N, depending on the number of hidden units of the encoder. Rows of $\mathbf{P}$ and $\mathbf{Q}$, $\mathbf{p}_\alpha$ and $\mathbf{q}_\beta$ involve atomistic chemical features of atoms α and β, which are directly related to the target property, i.e. the solvation free energy in the present work. We calculate the un-normalized attention score (or *chemical similarity*) between the atoms α and β with Luong's dot-product attention [38],

$$\mathbf{I}_{\alpha\beta} = -\mathbf{p}_\alpha \cdot \mathbf{q}_\beta, \tag{3}$$

which is an element of $M_a \times M_b$ tensor of atomistic interactions, $\mathbf{I}$. Because our target quantity is the free energy of solvation, we expect such chemical similarity $\mathbf{I}_{\alpha\beta}$ to correspond to atomistic interactions between α and β, which includes both energetic and entropic contributions. Eventually, the free energy of solvation of the given solvent–solute pair, which is the final regression target, is expressed as a simple summation of atomistic interactions:

$$\Delta G^{\circ}_{\text{sol}} = \sum_{\alpha=1}^{M_a} \sum_{\beta=1}^{M_b} \mathbf{I}_{\alpha\beta}. \tag{4}$$

Certainly, one can also calculate the free energies of solvation from two molecular feature vectors, which represent the solvent properties **u** and the solute properties **v**, respectively:

$$\Delta G^{\circ}_{\text{sol}} = \mathbf{u} \cdot \mathbf{v} = \left(\sum_{\alpha=1}^{M_a} \mathbf{p}_{\alpha} \right) \cdot \left(\sum_{\beta=1}^{M_b} \mathbf{q}_{\beta} \right). \tag{5}$$

The inner-product relation between molecular feature vectors **u** and **v** has a formal analogy with the solvent-gas partition coefficient calculation method via the solvation descriptor approach [63, 70]. Figure 1 illustrates an overview of the architecture of the proposed ML solvation model.

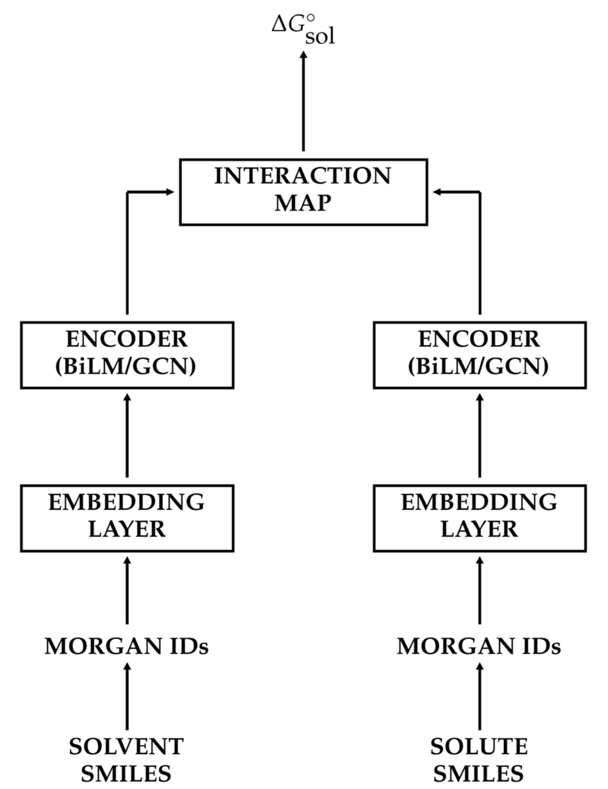

Fig. 1 Schematic of MLSolvA architecture. Each encoder network extracts atomistic feature vectors given pre-trained vector representations, and the interaction map calculates pairwise atomistic interactions from Luong's dot-product attention [38]

Encoder networks

We chose and compared two different neural network models to encode the input molecular structure and extract important structural or chemical features that are strongly related to solvation behavior. One is the bidirectional language model (BiLM) [49] based on the recurrent neural network (RNN), and the other is the graph convolutional neural network (GCN) [31] which explicitly handles the connectivity (bonding) between atoms with the adjacency matrix.

The detailed mathematical expressions of the BiLM, which is the first model, are given as follows [49]:

$$\overrightarrow{\mathbf{H}}^{(i+1)} = \overrightarrow{\text{RNN}}(\overrightarrow{\mathbf{H}}^{(i)}), \tag{6a}$$

$$\overleftarrow{\mathbf{H}}^{(i+1)} = \overleftarrow{\text{RNN}}(\overleftarrow{\mathbf{H}}^{(i)}). \tag{6b}$$

In Eq. 6, the right-headed arrow in $\overrightarrow{\text{RNN}}$ denotes a forward-directed recurrent unit that propagates from the leftmost to the rightmost sequence. The BiLM likewise involves a backward-directed recurrent neural network ($\overleftarrow{\text{RNN}}$) and propagates from the rightmost to the leftmost sequence as well. The superscript (i) in hidden layers $\mathbf{H}^{(i)}$ denotes the position at the stacked configuration. In the first layer, both forward and backward-directed RNN share the pre-trained sequence **X** as an input, $\overrightarrow{\mathbf{H}}^{(0)} = \overleftarrow{\mathbf{H}}^{(0)} = \mathbf{X}$.

Furthermore, more improved versions of RNNs, such as the gated recurrent unit (GRU) [14] or the long-short term memory (LSTM) [24] are more suitable when we consider cumulated numerical errors due to the deep-structured nature of RNNs [5],

$$\mathbf{H}^{(i)} = \overrightarrow{\mathbf{H}}^{(i)} + \overleftarrow{\mathbf{H}}^{(i)}. \tag{7}$$

Hidden layers from the forward and backward RNNs are then merged into a single sequence, as described in Eq. 7. Finally, we obtain the sequence of chemical feature vectors of the α-th atom in the given solvent with weighted summation of stacked RNN layers,

$$\mathbf{P} = \sum_{i} c_i \mathbf{H}^{(i)}, \tag{8}$$

where each weighing factor c_i is also a trainable parameter. The encoder function for solutes has an identical neural network architecture, which converts the pre-trained solute sequence **Y** into the feature sequence **Q**. In addition, each layer in the encoder must share the same number of hidden units N due to Eqs. 3, 8.

We consider the graph convolutional neural network (GCN), which is one of the most well-known algorithms in the chemical applications of neural networks [29, 31]. The GCN model represents the input molecule as a mathematical graph, instead of a simple sequence: each node corresponds to the atom, and each edge in the adjacency matrix **A** involves connectivity (or existence of bonding) between atoms:

$$\mathbf{H}^{(i+1)} = \mathrm{GCN}(\mathbf{H}^{(\mathrm{i})}, \mathbf{A}). \tag{9}$$

The role of the adjacency matrix in the GCN constrains convolution filters to the node itself and its nearest neighbors. Equation 10 describes a more detailed mathematical expression of the skip-connected GCN [31]:

$$\mathrm{GCN}(\mathbf{H}, \mathbf{A}) = \sigma\left(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{H}\mathbf{W}_1 + \mathbf{H}\mathbf{W}_2 + \mathbf{b}\right), \tag{10}$$

where **D** is the degree matrix, $\mathbf{W}_1$ and $\mathbf{W}_2$ are convolution filters, **b** is the bias vector, and σ denotes the activation function chosen as the hyperbolic tangent in the proposed model. The GCN encoder includes the stacked structure, and we obtain the feature sequence for each molecule in the same manner as described in Eq. 8.

Results and discussion

Computational setup and results

For the training and tests of the proposed neural network, we prepared 6239 experimental measures of free energies of solvation for 935 organic solvents and 146 organic solutes. A total of 642 experimentally measured values of the free energy of hydration are collected from the FreeSolv database [44, 45], and 5597 data points for non-aqueous solvents were collected with the Solv@TUM database version 1.0 [22, 23]. The compounds in the dataset comprise ten different kinds of atoms that are common in organic chemistry, viz. hydrogen, carbon, oxygen, sulfur, nitrogen, phosphorus, fluorine, chlorine, bromine, and iodine. The maximum heavy-atom count is 28 for the solute molecules and 18 for the solvent molecules.

At the very first stage, we perform the skip-gram pretraining process for 10,229,472 organic compounds, which are collected from the ZINC15 database [69], with Gensim 3.8.1 and Mol2Vec skip-gram model to construct the 128-dimensional embedding lookup table [25]. A total of 634 solutes 120 solvents in the FreeSolv/Solv@TUM combined dataset appear in the pretraining dataset. The pretraining process generates atomistic vector representations of the heavy atoms in different chemical environments distinguished by the Morgan identifiers [25, 46]. Although the skip-gram task does not guarantee a significant enhancement of the model's accuracy, we found that the pretrained model yielded more stable results in terms of RMSE variance (Additional file 1: Table S2). For the implementation of the neural network model, we mainly use TensorFlow 2.5.0 framework [41]. Each model has L2 regularization to prevent excessive changes on weights and to minimize the variance, and uses the RMSprop algorithm for minimization:

$$G_t = \rho G_{t-1} + (1-\rho)(\nabla_w L_t)^2, \tag{11a}$$

$$w_t = w_{t-1} + \frac{\eta}{\sqrt{G_t + \epsilon}}\nabla_w L_t, \tag{11b}$$

where L_t is the loss function, chosen as the mean squared error (MSE) in this work. G_t denotes a moving average of the squared gradient of L_t, and it scales update rates of the weight, w. The other parameters play the following roles: η is the initial learning rate, ρ is a discounting factor for the moving average, and ϵ prevents possible bursting of $1/\sqrt{G_t}$ for numerical stability. The selection of the optimized model for the target property is realized by an extensive Bayesian optimization process for tuning model hyperparameters [6] (Additional file 1: Table S1)

We employ five-fold nested cross-validation (CV) to evaluate the prediction accuracy of the chosen model. Nested CV incorporates two CV loops: the inner loop selects the best hyperparameters over the validation set, while the outer loop evaluates the model's final performance of prediction. This procedure prevents overlap and possible *information leakage* between the validation and test sets [48]. To evaluate the uncertainty of results taken from CV tasks, we take averages for all mean errors over eight independent nested CV runs, split from different random states. The results for the test run using nested CV tasks for the optimized models are shown in Fig. 2. We found that the BiLM encoder with LSTM layer performs slightly better than the GCN encoder, although their differences are not pronounced. The mean unsigned prediction error (MUE) for the BiLM/LSTM encoder model is 0.19 kcal/mol, while the GCN model results in MUE = 0.22 kcal/mol. Both values show that the proposed mechanism works efficiently and guarantees excellent prediction accuracies for well-trained chemical structures. We also perform the same CV procedure using the Direct Message-Passing Neural Network (D-MPNN) model [17], which is available at the chemprop package [80]. The prediction error of the D-MPNN model on the same dataset is MUE = 0.19 kcal/mol, which indicates our proposed model design yields a comparable accuracy with the deep-learning model in the current state-of-the-art (see Additional file 2 for the raw data).

Visualization of chemical similarity

The fundamental idea behind the proposed model is the encoder network, which maps complex chemical features

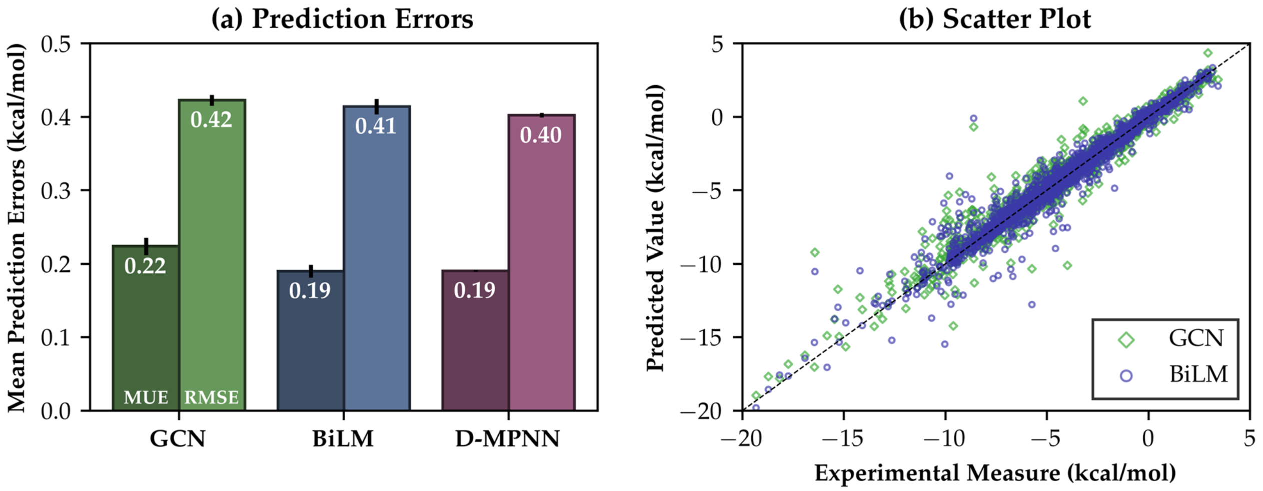

Fig. 2 **a** Prediction errors for BiLM and GCN models in kcal/mol, obtained by five-fold nested cross validation results. Results taken from the D-MPNN model [17, 80] are also depicted for comparison. **b** Scatter plot between experimental values and predicted values by the models. Green circles depict the BiLM model, while the GCN results are depicted by blue circles

into a vector representation. Because we aim for the free energy of solvation as the target property, geometries in the vector space must have a strong correlation with their solvation properties. We validate this point with t-Stochastic Neighbor Embedding (t-SNE) visualizations for pre-trained solute vectors **y**, and encoded molecular features **v** [48, 56]. The dimensions of those vectors must be reduced for visualization, because we use 128-dimensional vector representations, which cannot be directly drawn into a graph. Figure 3 presents the reduced geometries of **y** and **v** in two-dimensional space, which indicates that the proposed encoder neural network works as intended. Color shading depicts the predicted hydration free energies for 15,432 points, whose structures are randomly taken from the ZINC15 [69]; red and blue dots correspond to low and high hydration free energy cases, respectively. The significant correlation between reduced molecular feature vectors and predicted free energy values indicates how the proposed architecture extracts important molecular features and makes the prediction from them. Meanwhile, the pre-trained solute vectors from the skip-gram embedding model exhibit only weak correlations.

Advantage of model: transferability

Because our proposed neural network model is solvent-generic, as it considers both the solvent and solute structures as separate inputs, it exhibits a distinct and advantageous character when compared to other solvent-specific ML solvation models. Let us consider the

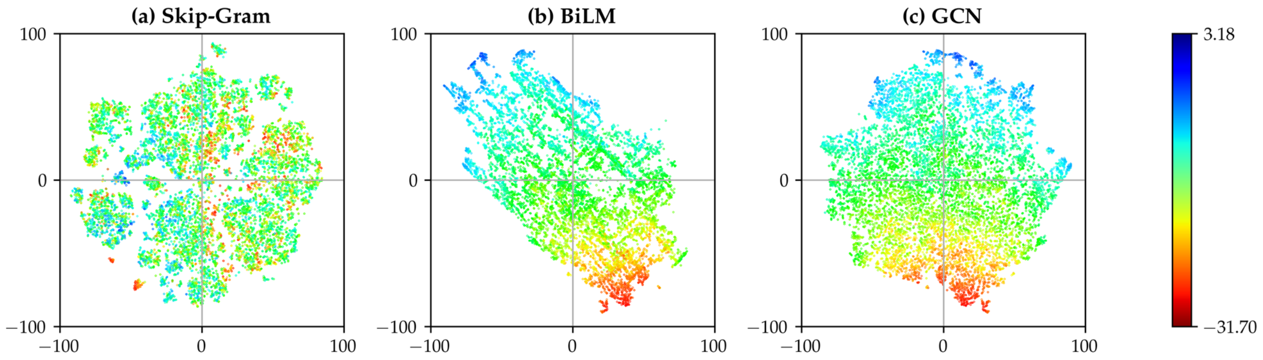

Fig. 3 Two-dimensional visualizations on **a** pre-trained vector from the skip-gram model $\sum_\beta \mathbf{y}_\beta$ and **b**, **c** extracted molecular feature vector **v** for 15,432 solutes. We reduce the dimensions of each vector using the t-SNE algorithm. The color representation denotes the hydration energy of each point

following possible situation where one wants to predict the solvation free energy of a solute compound $\mathcal{A}$ in a solvent $\mathcal{X}$, $\Delta G^{\circ}_{\mathcal{AX}}$. Since the model has been trained with varied kinds of solvents and solutes, the training database will likely involve solvation free energy measures for $\mathcal{A}$ in other different solvents, e.g. $\Delta G^{\circ}_{\mathcal{AY}}$, $\Delta G^{\circ}_{\mathcal{AZ}}$, and so on. Then the model would have already become aware of the structural features of $\mathcal{A}$, which could help the prediction of $\Delta G^{\circ}_{\mathcal{AX}}$ [37]; this mechanism would not happen if the model supports only one kind of solvent. Therefore, one of the largest advantages of our model is that we can easily enlarge the dataset for training, even in the scenario where we want to predict solvation free energies for a specific solvent. Figure 4 shows five-fold CV results for 642 hydration free energies (FreeSolv) from both BiLM and GCN models in two different situations. One uses only the FreeSolv [44, 45] database for training and tests, whereas the other uses both the FreeSolv and the Solv@TUM [22, 23] databases. Although the Solv@TUM database only involves non-aqueous data points, it enhances each model's accuracy by approximately 20% (BiLM) to 30% (GCN) in terms of the MUE. These results imply that there are possible applications of transfer learning to other solvation-related properties, such as aqueous solubilities [18] or octanol–water partition coefficients.

However, in some other situations, one may be concerned that the repetitive training for a single compound may cause overfitting by the model, and they could weaken the predictivity for the structurally new compound, which is considered an extrapolation. We investigate the model's predictivity for extrapolation situations with a *scaffold-based* split [20, 37, 42, 77]. Instead of the ordinary K-fold CV task with the random and uniform split method, the K-means clustering algorithm builds each fold with the Molecular ACCess system (MACCS) substructural fingerprint [77]. An extreme extrapolation situation can be simulated through CV tasks over

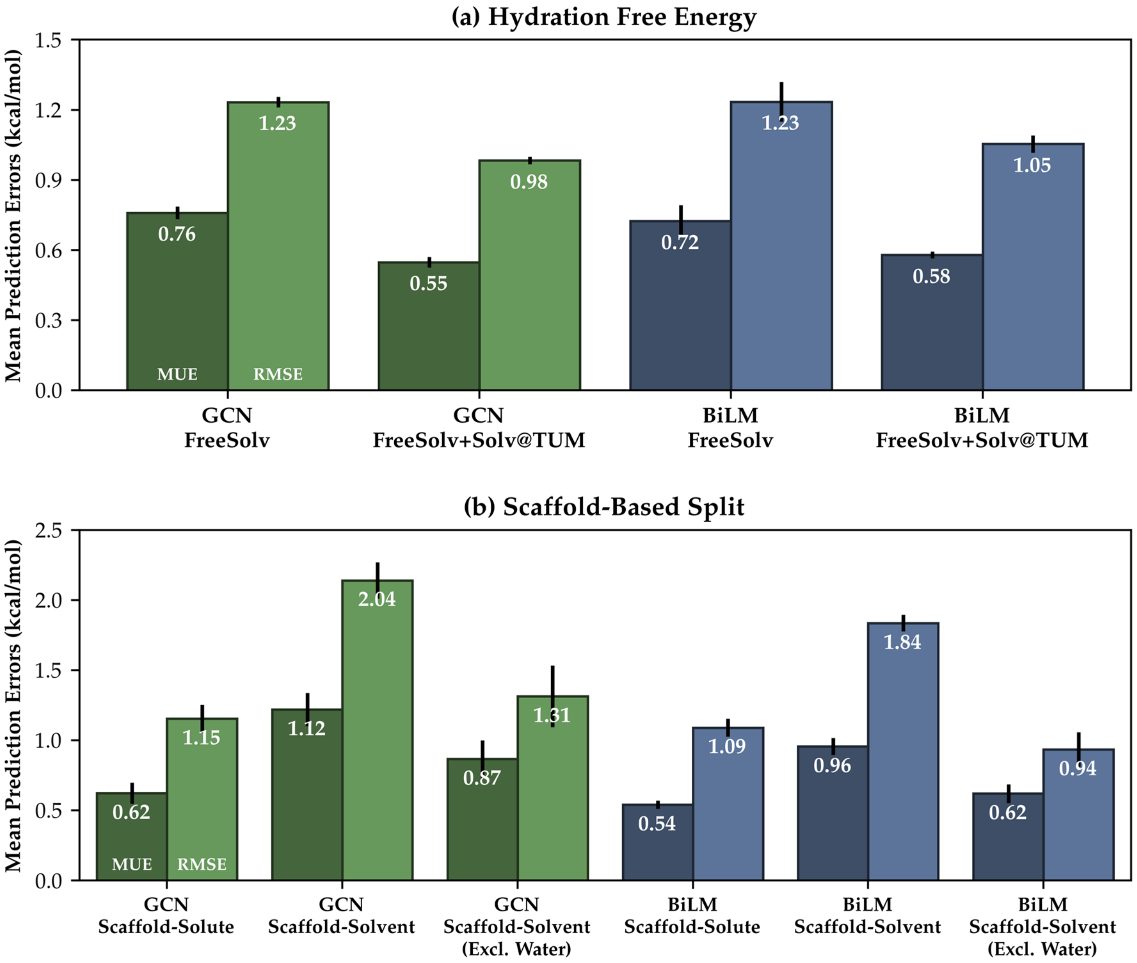

Fig. 4 **a** CV-results for FreeSolv hydration energies with two different training datasets. Deep-colored boxes depict CV results with augmented dataset with Solv@TUM database. **b** CV results for two scaffold-based split methods using K-means clustering algorithm

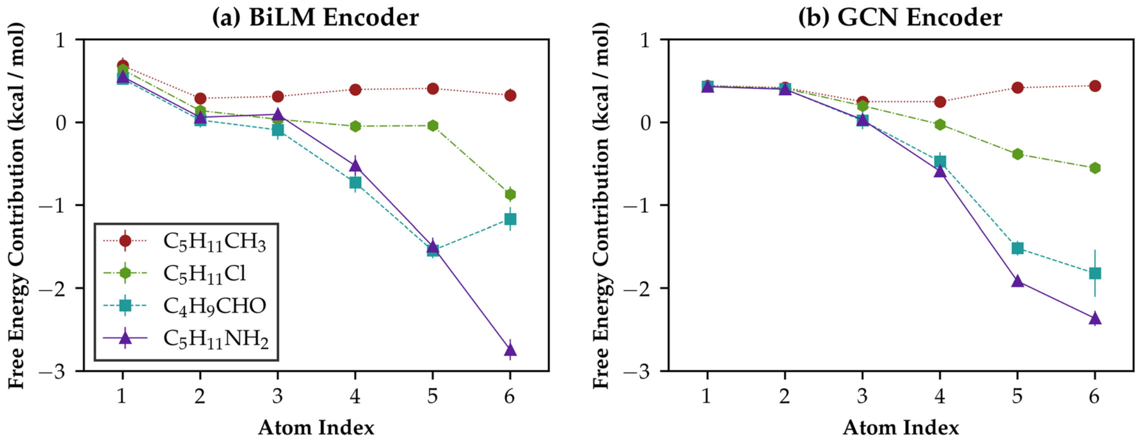

Fig. 5 ML-calculated atomistic group contributions for five small organic compounds with six heavy atoms (excluding the hydrogens). The atom index starts from the left-most point of the given molecule and only counts heavy atoms

the folds, which are constructed by the clustering on solvents or solutes. As shown in Fig. 4, the scaffold-based split on the solvents shows more degradation of prediction performances than the scaffold-solute-based split due to the limited kinds of solvent compounds in the dataset, although both results are still within an acceptable error level, given chemical accuracy of ∼ 1.0 kcal/mol (raw data is available in Additional file 2). A considerable part of degradation in the scaffold-solvent-based split arises from water solvent due to its unusually distinct physicochemical nature from other organic solvents [37]. Furthermore, the embedding scheme we use generates a unique Morgan identifier for the oxygen of water (864666390), which cannot be recognized or trained from the other hydroxyl oxygens such as alcohols (864662311).

Group contributions

Although we showed that the proposed NN model guarantees an excellent predictivity for solvation energies of various solute and solvent pairs, the main objective of the present study is to obtain the solvation free energy as the sum of decomposed interatomic interactions, as described in Eqs. 3 and 4. To verify the feasibility of the model's solvation energy estimation to decompose into group contributions, we define the sum of atomic interactions $\mathbf{I}_{\alpha\beta}$ over the solvent indices α as the group contributions of the β-th solute atom:

$$\mathbf{I}_\beta = \sum_\alpha \mathbf{I}_{\alpha\beta}. \tag{12}$$

Figure 5 shows hydration free energy contributions for five small organic solutes with six heavy atoms: *n*-hexane, 1-chloropentane, pentaldehyde, 1-aminopentane, and benzene. Both the BiLM and the GCN model exhibit a similar tendency in group contributions; the model estimates that atomic interactions between the solute atoms and water increase near the hydrophilic groups. It is obvious that each atom in benzene must have identical contributions to the free energy; however, the results in Fig. 5 clearly show that the BiLM model makes faulty predictions while the GCN model works well as expected. We believe that this malfunctioning of the BiLM model originates from the sequential nature of the recurrent neural network. Because the RNN considers that the input molecule is only a simple sequence of atomic vectors, and there are no explicit statements that involve bonding information, the model is not aware of the cyclic shape of the input compound [29, 51]. We conclude that it is inevitable to use explicitly bound (or connectivity) information when constructing a group-contribution based ML model, even though the RNN-based model provides good predictions in terms of their sum.

Conclusions

We introduced a novel approach for ML-based solvation energy prediction, which exhibits great potential to provide physicochemical insight on the solvation process. The novelty in our neural network model lies in the ability to calculate pairwise atomic interactions from the inner products of atomistic feature vectors [38]. This idea gives us more straightforward and interpretable information on intermolecular interactions between the solute and solvent molecules, and the model calculates the solvation free energy from the group-contribution-based prediction.

We quantified the proposed model's prediction performances for 6293 experimental data points of solvation energies, which were taken from the FreeSolv [44, 45] and Solv@TUM [22, 23] databases. We found a significant geometrical correlation between molecular feature vectors and predicted properties, which confirms that the proposed model successfully extracts chemical properties and maps them into vector representations. The estimated prediction MUEs from K-fold CV are 0.19 kcal/mol for the BiLM encoder and 0.23 kcal/mol for the GCN model.

The K-fold CV results from the scaffold-based split [77] showed that the prediction accuracy decreases by three times in extreme extrapolation situations; however, they nevertheless exhibit moderate performances, which was MUE = 0.60 kcal/mol. Moreover, we found that the solvent-generic structure of the proposed model is appropriate for enlarging the dataset size, i.e. experimental data points for a particular solvent are transferable to other solvents. We conclude that this transferability is the reason for our model's outstanding predictivity [37].

Finally, we examined pairwise atomic interactions obtained from the interaction map and found a clear tendency between hydrophilic groups and their contributions to the hydration free energy. Such results are obtained from a simple, graph-convolution based neural network instead of deep learning models in the current state-of-the-art [20, 62]. Despite the limitation of a simple model, the model showed a reliable performance with the concept of group contributions approach via neural networks. Thus, we expect that the suggested concept would have further developments with more progressed ML models or applications for molecular dynamics simulations [12, 13]. We believe that our model is capable of providing detailed information on the solvation mechanism, as well as the predicted value of the target property.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-021-00533-z.

Additional file 1. List of model hyperparameters for the Bayesian optimization process (Table S1) and influence of the pre-training task on the prediction results (Table S2).

Additional file 2. Raw prediction results file for the random CV and scaffold-based CV tasks.

Acknowledgements
Not applicable.

Authors' contributions
HL and YJ conceived of the presented idea. HL designed and implemented the methodology, analyzed the results, and drafted the manuscript. YJ supervised the current study. Both the authors discussed the results and contributed to the final manuscript. Both authors read and approved the final manuscript.

Funding
This research was supported by Samsung Display Co.,Ltd. (0414-20200051) and the Creative Materials Discovery Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT and Future Planning (NRF-2017M3D1A1039553).

Availability of data and materials
The source code is available on GitHub repository: https://github.com/ht0620/mlsolva. The data supporting the findings of this study are available within the article and its additional information files.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 4 February 2021 Accepted: 15 July 2021
Published online: 31 July 2021

References

1. Allam O, Cho BW, Kim KC, Jang SS (2018) Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries. RSC Adv 8(69):39414–39420. https://doi.org/10.1039/C8RA07112H
2. Arús-Pous J, Patronov A, Bjerrum EJ, Tyrchan C, Reymond JL, Chen H, Engkvist O (2020) SMILES-based deep generative scaffold decorator for de-novo drug design. J Cheminform 12(1):38. https://doi.org/10.1186/s13321-020-00441-8
3. Basdogan Y, Groenenboom MC, Henderson E, De S, Rempe SB, Keith JA (2020) Machine learning-guided approach for studying solvation environments. J Chem Theory Comput 16(1):633–642. https://doi.org/10.1021/acs.jctc.9b00605
4. Behler J (2017) First principles neural network potentials for reactive simulations of large molecular and condensed systems. Angew Chem Int Ed 56(42):12828–12840. https://doi.org/10.1002/anie.201703114
5. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166. https://doi.org/10.1109/72.279181. https://ieeexplore.ieee.org/document/279181/
6. Bergstra J, Yamins D, Cox DD (2013) Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th international conference on international conference on machine learning, Volume 28, ICML'13, JMLR.org, Atlanta, GA, USA, pp I–115–I–123
7. Borhani TN, García-Muñoz S, Vanesa Luciani C, Galindo A, Adjiman CS (2019) Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. Phys Chem Chem Phys 21(25):13706–13720. https://doi.org/10.1039/C8CP07562J
8. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature 559(7715):547–555. https://doi.org/10.1038/s41586-018-0337-2
9. Cao J, Pan Y, Jiang Y, Qi R, Yuan B, Jia Z, Jiang J, Wang Q (2020) Computer-aided nanotoxicology: risk assessment of metal oxide nanoparticles via nano-QSAR. Green Chem 22(11):3512–3521. https://doi.org/10.1039/D0GC00933D
10. Chen BWJ, Xu L, Mavrikakis M (2020) Computational methods in heterogeneous catalysis. Chem Rev. https://doi.org/10.1021/acs.chemrev.0c01060
11. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2014) QSAR modeling: where have you been? where are you going to? J Med Chem 57(12):4977–5010. https://doi.org/10.1021/jm4004285
12. Chong SH, Ham S (2011) Atomic decomposition of the protein solvation free energy and its application to amyloid-beta protein in water. J Chem Phys 135(3):034506. https://doi.org/10.1063/1.3610550
13. Chong SH, Ham S (2014) Interaction with the surrounding water plays a key role in determining the aggregation propensity of proteins. Angew Chem Int Ed 53(15):3961–3964. https://doi.org/10.1002/anie.201309317

14. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555 [cs]
15. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF (2017) Convolutional embedding of attributed molecular graphs for physical property prediction. J Chem Inf Model 57(8):1757–1772. https://doi.org/10.1021/acs.jcim.6b00601
16. Cramer CJ, Truhlar DG (2008) A universal approach to solvation modeling. Acc Chem Res 41(6):760–768. https://doi.org/10.1021/ar800019z
17. Dai H, Dai B, Song L (2016) Discriminative embeddings of latent variable models for structured data. In: International conference on machine learning, PMLR. pp 2702–2711. http://proceedings.mlr.press/v48/daib16.html
18. Delaney JS (2004) ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci 44(3):1000–1005. https://doi.org/10.1021/ci034243x
19. Duarte Ramos Matos G, Kyu DY, Loeffler HH, Chodera JD, Shirts MR, Mobley DL (2017) Approaches for calculating solvation free energies and enthalpies demonstrated with an update of the FreeSolv database. J Chem Eng Data 62(5):1559–1569. https://doi.org/10.1021/acs.jced.7b00104
20. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. arXiv:1704.01212 [cs]
21. Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, Kaus JW, Cerutti DS, Krilov G, Jorgensen WL, Abel R, Friesner RA (2016) OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J Chem Theory Comput 12(1):281–296. https://doi.org/10.1021/acs.jctc.5b00864
22. Hille C, Ringe S, Deimel M, Kunkel C, Acree WE, Reuter K, Oberhofer H (2018) Solv@TUM v 1.0. https://doi.org/10.14459/2018MP1452571. https://mediatum.ub.tum.de/1452571
23. Hille C, Ringe S, Deimel M, Kunkel C, Acree WE, Reuter K, Oberhofer H (2019) Generalized molecular solvation in non-aqueous solutions by a single parameter implicit solvation scheme. J Chem Phys 150(4):041710. https://doi.org/10.1063/1.5050938
24. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
25. Jaeger S, Fulle S, Turk S (2018) Mol2vec: unsupervised machine learning approach with chemical intuition. J Chem Inf Model 58(1):27–35. https://doi.org/10.1021/acs.jcim.7b00616
26. Jespers W, Esguerra M, Åqvist J, Gutiérrez-de Terán H (2019) QligFEP: an automated workflow for small molecule free energy calculations in Q. J Cheminform 11(1):26. https://doi.org/10.1186/s13321-019-0348-5
27. Jia X, Wang M, Shao Y, König G, Brooks BR, Zhang JZH, Mei Y (2016) Calculations of solvation free energy through energy reweighting from molecular mechanics to quantum mechanics. J Chem Theory Comput 12(2):499–511. https://doi.org/10.1021/acs.jctc.5b00920
28. Karpov P, Godin G, Tetko IV (2020) Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform 12(1):17. https://doi.org/10.1186/s13321-020-00423-w
29. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P (2016) Molecular graph convolutions: moving beyond fingerprints. J Comput-Aided Mol Des 30(8):595–608. https://doi.org/10.1007/s10822-016-9938-8
30. Kim J, Ko S, Noh C, Kim H, Lee S, Kim D, Park H, Kwon G, Son G, Ko JW, Jung Y, Lee D, Park CB, Kang K (2019) Biological nicotinamide cofactor as a redox-active motif for reversible electrochemical energy storage. Angew Chem Int Ed 58(47):16764–16769. https://doi.org/10.1002/anie.201906844
31. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907 [cs, stat]
32. Klamt A (2018) The COSMO and COSMO-RS solvation models: COSMO and COSMO-RS. Wiley Interdiscip Rev Comput Mol Sci 8(1):e1338. https://doi.org/10.1002/wcms.1338
33. Klamt A, Diedenhofen M (2015) Calculation of solvation free energies with DCOSMO-RS. J Phys Chem A 119(21):5439–5445. https://doi.org/10.1021/jp511158y
34. Klamt A, Eckert F, Arlt W (2010) COSMO-RS: an alternative to simulation for calculating thermodynamic properties of liquid mixtures. Ann Rev Chem Biomol Eng 1(1):101–122. https://doi.org/10.1146/annurev-chembioeng-073009-100903
35. Klamt A, Schüürmann G (1993) COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc Perkin Trans 2(5):799–805. https://doi.org/10.1039/P29930000799
36. König G, Pickard FC, Mei Y, Brooks BR (2014) Predicting hydration free energies with a hybrid QM/MM approach: an evaluation of implicit and explicit solvation models in SAMPL4. J Comput-Aided Mol Des 28(3):245–257. https://doi.org/10.1007/s10822-014-9708-4
37. Lim H, Jung Y (2019) Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. Chem Sci 10(36):8306–8315. https://doi.org/10.1039/C9SC02452B
38. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv:1508.04025 [cs]
39. Marenich AV, Cramer CJ, Truhlar DG (2009) 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 113(18):6378–6396. https://doi.org/10.1021/jp810292
40. Marenich AV, Cramer CJ, Truhlar DG (2013) Generalized born solvation model SM12. J Chem Theory Comput 9(1):609–620. https://doi.org/10.1021/ct300900e
41. Martín A, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, 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, Heng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. http://tensorflow.org/
42. Mayr A, Klambauer G, Unterthiner T, Hochreiter S (2016) DeepTox: toxicity prediction using deep learning. Front Environ Sci. https://doi.org/10.3389/fenvs.2015.00080
43. Mennucci B (2012) Polarizable continuum model: polarizable continuum model. Wiley Interdiscip Rev Comput Mol Sci 2(3):386–404. https://doi.org/10.1002/wcms.1086
44. Mobley DL, Guthrie JP (2014) FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput-Aided Mol Des 28(7):711–720. https://doi.org/10.1007/s10822-014-9747-x
45. DLMobleyMShirtsNLimJChoderaKBeauchampLee-Ping2018Mobleylab/Freesolv: Version 0.52Zenodo10.5281/ZENODO.1161245Mobley DL, Shirts M, Lim N, Chodera J, Beauchamp K, Lee-Ping (2018) Mobleylab/Freesolv: Version 0.52. Zenodo. https://doi.org/10.5281/ZENODO.1161245
46. Morgan HL (1965) The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Chem Doc 5(2):107–113. https://doi.org/10.1021/c160017a018
47. Park H, Lim HD, Lim HK, Seong WM, Moon S, Ko Y, Lee B, Bae Y, Kim H, Kang K (2017) High-efficiency and high-power rechargeable lithium-sulfur dioxide batteries exploiting conventional carbonate-based electrolytes. Nat Commun 8(1):14989. https://doi.org/10.1038/ncomms14989
48. 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 (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
49. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv:1802.05365 [cs]
50. Plante J, Werner S (2018) JPlogP: an improved logP predictor trained using predicted data. J Cheminform 10(1):61. https://doi.org/10.1186/s13321-018-0316-5
51. Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. Sci Adv 4(7):eaap7885. https://doi.org/10.1126/sciadv.aap7885
52. Rauer C, Bereau T (2020) Hydration free energies from kernel-based machine learning: compound-database bias. J Chem Phys 153(1):014101. https://doi.org/10.1063/5.0012230
53. Reichardt C, Welton T (2010) Solvents and solvent effects in organic chemistry: REICHARDT:SOLV.EFF. 4ED O-BK. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim. https://doi.org/10.1002/9783527632220
54. Ringe S, Oberhofer H, Hille C, Matera S, Reuter K (2016) Function-space-based solution scheme for the size-modified Poisson-Boltzmann equation in full-potential DFT. J Chem Theory Comput 12(8):4052–4066. https://doi.org/10.1021/acs.jctc.6b00435

55. Ryu S, Kwon Y, Kim WY (2019) A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. Chem Sci 10(36):8438–8446. https://doi.org/10.1039/C9SC01992H
56. Ryu S, Lim J, Hong SH, Kim WY (2018) Deeply learning molecular structure–property relationships using attention- and gate-augmented graph convolutional network. arXiv:1805.10988 [cs, stat]
57. Sato H (2013) A modern solvation theory: quantum chemistry and statistical chemistry. Phys Chem Chem Phys 15(20):7450. https://doi.org/10.1039/c3cp50247c
58. Scheen J, Wu W, Mey ASJS, Tosco P, Mackey M, Michel J (2020) Hybrid alchemical free energy/machine-learning methodology for the computation of hydration free energies. J Chem Inf Model 60(11):5331–5339. https://doi.org/10.1021/acs.jcim.0c00600
59. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003
60. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A (2017) Quantum-chemical insights from deep tensor neural networks. Nat Commun 8(1):13890. https://doi.org/10.1038/ncomms13890
61. Schütt KT, Kessel P, Gastegger M, Nicoli KA, Tkatchenko A, Müller KR (2019) SchNetPack: a deep learning toolbox for atomistic systems. J Chem Theory Comput 15(1):448–455. https://doi.org/10.1021/acs.jctc.8b00908
62. Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR (2018) SchNet—a deep learning architecture for molecules and materials. J Chem Phys 148(24):241722. https://doi.org/10.1063/1.5019779
63. Sedov IA, Salikov TM, Wadawadigi A, Zha O, Qian E, Acree WE, Abraham MH (2018) Abraham model correlations for describing the thermodynamic properties of solute transfer into pentyl acetate based on headspace chromatographic and solubility measurements. J Chem Thermodyn 124:133–140. https://doi.org/10.1016/j.jct.2018.05.003
64. Sels H, De Smet H, Geuens J (2020) SUSSOL-using artificial intelligence for greener solvent selection and substitution. Molecules 25(13):3037. https://doi.org/10.3390/molecules25133037
65. Shivakumar D, Williams J, Wu Y, Damm W, Shelley J, Sherman W (2010) Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. J Chem Theory Comput 6(5):1509–1519. https://doi.org/10.1021/ct900587b
66. Sifain AE, Lubbers N, Nebgen BT, Smith JS, Lokhov AY, Isayev O, Roitberg AE, Barros K, Tretiak S (2018) Discovering a transferable charge assignment model using machine learning. J Phys Chem Lett 9(16):4495–4501. https://doi.org/10.1021/acs.jpclett.8b01939
67. Skyner RE, McDonagh JL, Groom CR, van Mourik T, Mitchell JBO (2015) A review of methods for the calculation of solution free energies and the modelling of systems in solution. Phys Chem Chem Phys 17(9):6174–6191. https://doi.org/10.1039/C5CP00288E
68. Smith JS, Isayev O, Roitberg AE (2017) ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci 8(4):3192–3203. https://doi.org/10.1039/C6SC05720A
69. Sterling T, Irwin JJ (2015) ZINC 15–ligand discovery for everyone. J Chem Inf Model 55(11):2324–2337. https://doi.org/10.1021/acs.jcim.5b00559
70. Stolov MA, Zaitseva KV, Varfolomeev MA, Acree WE (2017) Enthalpies of solution and enthalpies of solvation of organic solutes in ethylene glycol at 298.15 K: prediction and analysis of intermolecular interaction contributions. Thermochim Acta 648:91–99. https://doi.org/10.1016/j.tca.2016.12.015
71. Subramanian V, Ratkova E, Palmer D, Engkvist O, Fedorov M, Llinas A (2020) Multisolvent models for solvation free energy predictions using 3D-RISM hydration thermodynamic descriptors. J Chem Inf Model 60(6):2977–2988. https://doi.org/10.1021/acs.jcim.0c00065
72. Takeda T, Taniki R, Masuda A, Honma I, Akutagawa T (2016) Electron-deficient anthraquinone derivatives as cathodic material for lithium ion batteries. J Power Sources 328:228–234. https://doi.org/10.1016/j.jpowsour.2016.08.022
73. Tomasi J, Mennucci B, Cammi R (2005) Quantum mechanical continuum solvation models. Chem Rev 105(8):2999–3094. https://doi.org/10.1021/cr9904009
74. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29(6–7):476–488. https://doi.org/10.1002/minf.201000061
75. Voršilák M, Kolář M, Čmelo I, Svozil D (2020) SYBA: Bayesian estimation of synthetic accessibility of organic compounds. J Cheminform 12(1):35. https://doi.org/10.1186/s13321-020-00439-2
76. Wang XY, Chen BB, Zhang J, Zhou ZR, Lv J, Geng XP, Qian RC (2021) Exploiting deep learning for predictable carbon dot design. Chem Commun. https://doi.org/10.1039/D0CC07882D
77. Winter R, Montanari F, Noé F, Clevert DA (2019) Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 10(6):1692–1701. https://doi.org/10.1039/C8SC04175J
78. Withnall M, Lindelöf E, Engkvist O, Chen H (2020) Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. J Cheminform 12(1):1. https://doi.org/10.1186/s13321-019-0407-y
79. Wu Z, Ramsundar B, Feinberg E, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9(2):513–530. https://doi.org/10.1039/C7SC02664A
80. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59(8):3370–3388. https://doi.org/10.1021/acs.jcim.9b00237
81. Zhang J, Tuguldur B, van der Spoel D (2015) Force field benchmark of organic liquids. 2. Gibbs energy of solvation. J Chem Inf Model 55(6):1192–1201. https://doi.org/10.1021/acs.jcim.5b00106
82. Zhang Z, Schott JA, Liu M, Chen H, Lu X, Sumpter BG, Fu J, Dai S (2019) Prediction of carbon dioxide adsorption via deep learning. Angew Chem Int Ed 58(1):259–263. https://doi.org/10.1002/anie.201812363

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.