Solvation Free Energy Prediction from Pairwise Atomistic Interactions by Machine Learning

Hyuntae Lim
Seoul National University College of Natural Sciences https://orcid.org/0000-0003-0342-3482
YounJoon Jung (yjjung@snu.ac.kr)
Seoul National University https://orcid.org/0000-0002-9464-9999

Research article

Keywords: machine learning technologies, solvation free energy, ML-based solvation model, pairwise atomistic interactions

Posted Date: February 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-207945/v1

Version of Record: A version of this preprint was published at Journal of Cheminformatics on July 31st, 2021. See the published version at https://doi.org/10.1186/s13321-021-00533-z.

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 features calculates their interactions. The results of 6,493 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 provides not only predictions of target properties but also more detailed physicochemical insights.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the latest manuscript can be downloaded and accessed as a PDF.

Figures

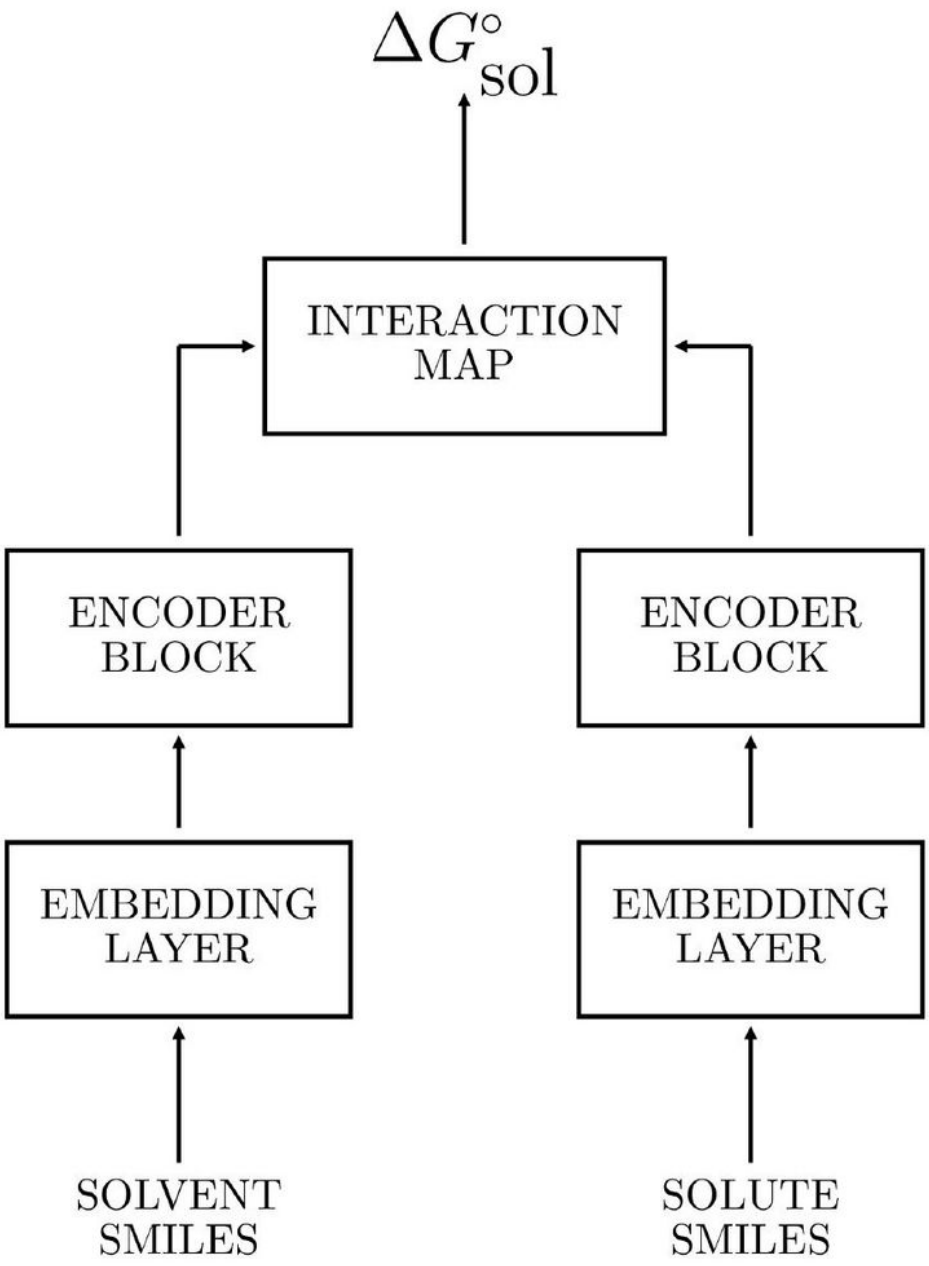

Figure 1

Schematic of MLSolv-A 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[36].

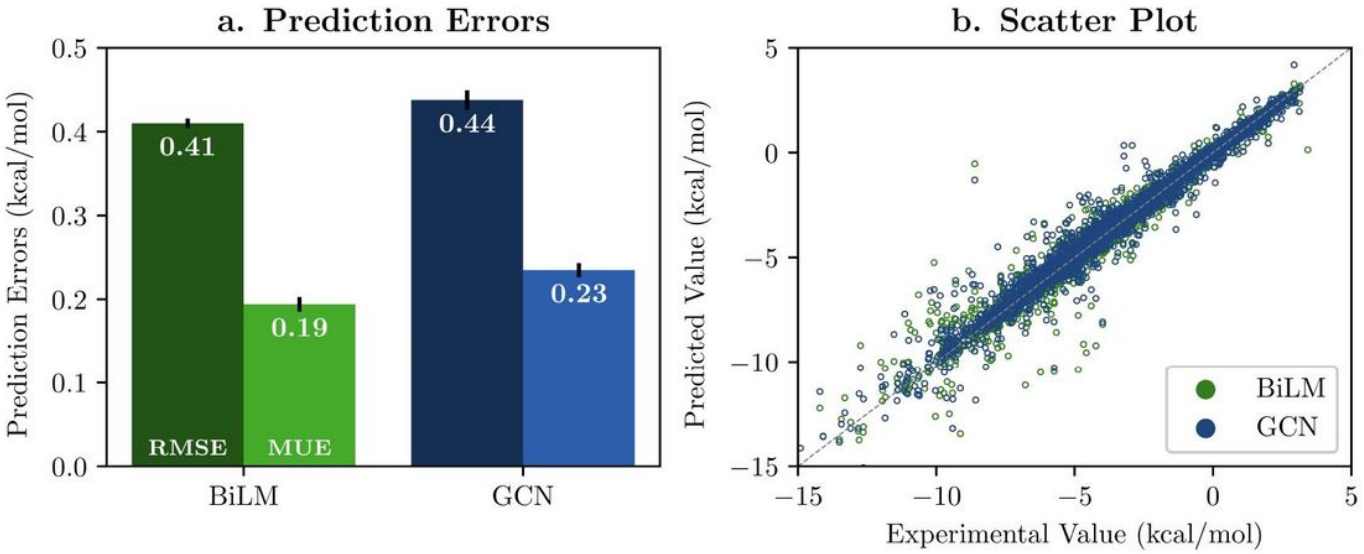

Figure 2

(a) Prediction errors for BiLM and GCN models in kcal/mol, obtained by five-fold cross validation results. (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.

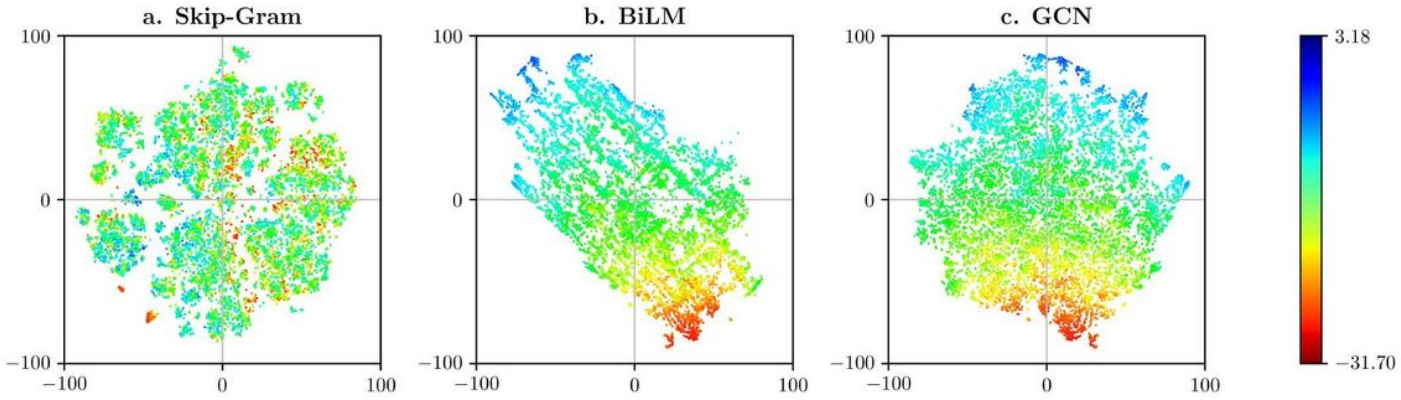

Figure 3

Two-dimensional visualizations on (a) pre-trained vector from the skip-gram model $\sum \beta y\beta$ 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.

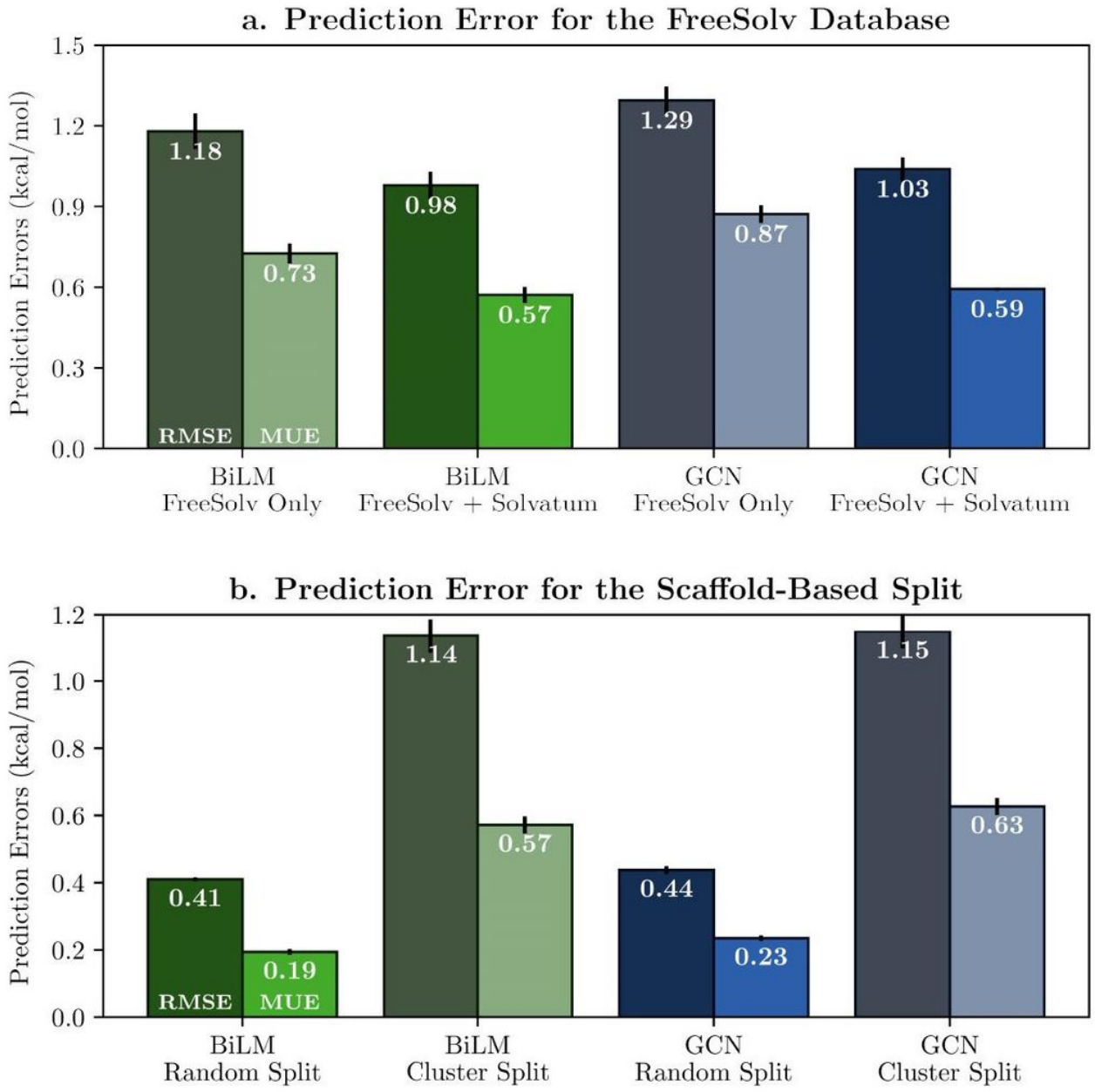

Figure 4

(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) Comparison between CV results with random-split and scaffold-based split (or cluster split).

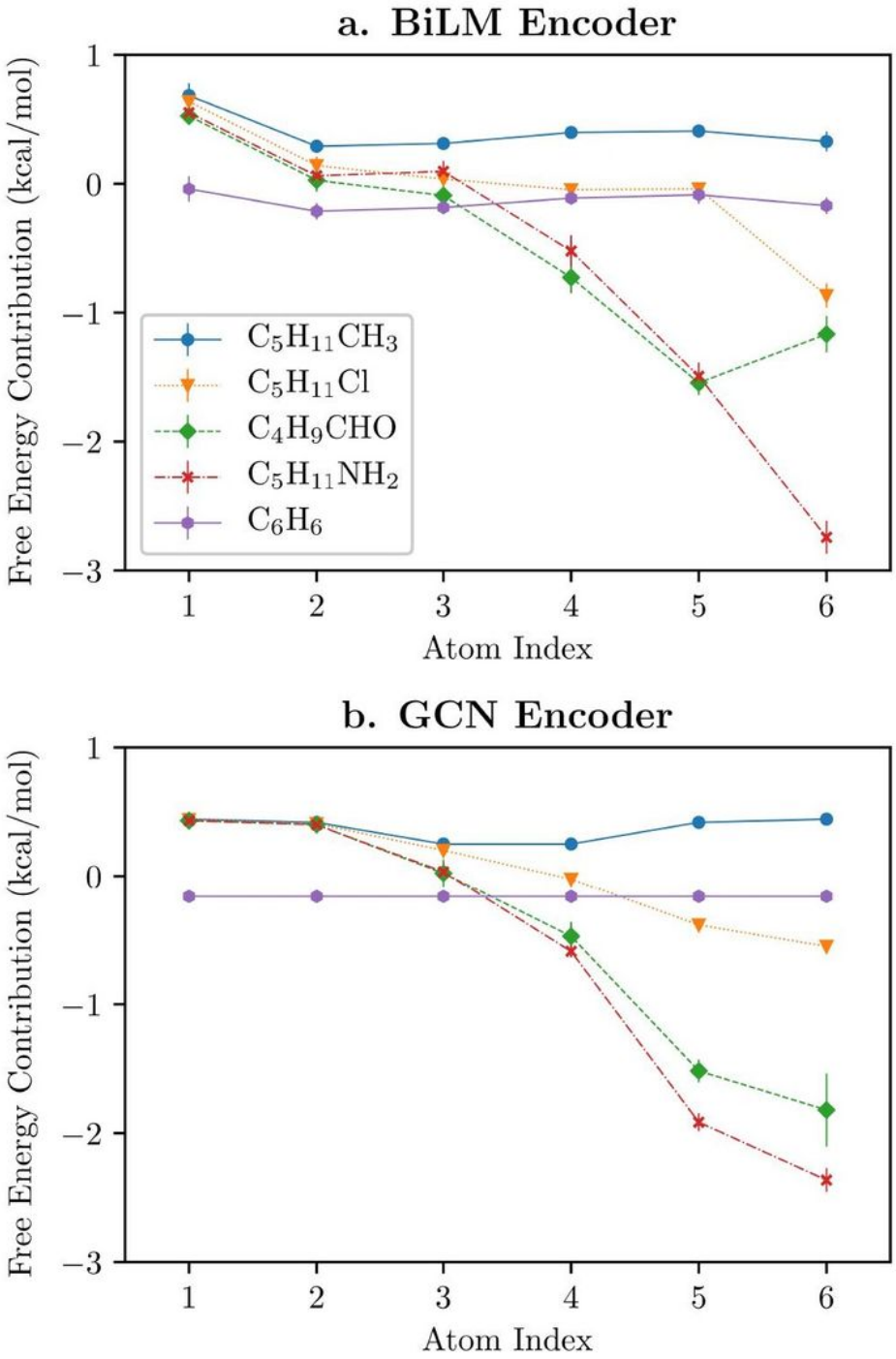

Figure 5

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- esi.pdf