Machine-Guided Representation for Accurate Graph-Based Molecular Machine Learning

Gyoung S. Na[1], Hyunju Chang[1], and Hyun Woo Kim[1]

[1]Korea Research Institute of Chemical Technology (KRICT)

1 Selected Atomic Features

To convert SMILES representation of the molecules into the attributed graphs, we used several atomic features represented as a numerical value. We chose the atomic features which can be roughly classified into four categories as shown in Table S1. The atomic features are given by the python mendeleev package[1].

Table S1: Selected atomic features and their description.

Category	Variable name	Comment	Unit
Size	atomic_volume	Atomic volume	cm^3/mol
	atomic_weight	Atomic weight	-
	atomic_weight_uncertainty	Atomic weight uncertainty	-
	atomic_radius	Atomic radius	pm
	atomic_radius_rahm	Atomic radius by Rahm et al.	pm
	covalent_radius_cordero	Covalent radius by Cerdero et al.	pm
	covalent_radius_pyykko	Covalent radius by Pyykko et al.	pm
	covalent_radius_slater	Covalent radius by Slater	pm
	vdw_radius	Van der Waals radius	pm
	vdw_radius_uff	Van der Walls radius from the UFF	pm
	vdw_radius_mm3	Van der Walls radius from the MM3 FF	pm
	vdw_radius_alvarez	Van der Waals radius according to Alvarez	pm
	density	Density at 295K	g/cm^3
	lattice_constant	Lattice constant	Angstrom
Heat	boiling_point	Boiling temperature	K
	melting_point	Melting temperature	K
	specific_heat	Specific heat at 20 C	J/(g mol)
	fusion_heat	Fusion heat	kJ/mol
	evaporation_heat	Evaporation heat	kJ/mol
	heat_of_formation	Heat of formation	kJ/mol
	thermal_conductivity	Thermal conductivity at 25 C	W/(m K)
Electronic	atomic_number	Atomic number	-
	electron_affinity	Electron affinity	eV
	period	Period in periodic table	-
	en_ghosh	Ghosh's scale of electronegativity	-
	en_pauling	Pauling's scale of electronegativity	-
	en_allen	Allen's scale of electronegativity	eV
	dipole_polarizability	Dipole polarizability	a.u.
	c6_gb	C_6 dispersion coefficient	a.u.
Abundance	abundance_crust	Abundance in the Earth's crust	mg/kg
	abundance_sea	Abundance in the seas	mg/L

[1]https://mendeleev.readthedocs.io/en/stable/

2 Implementation Details

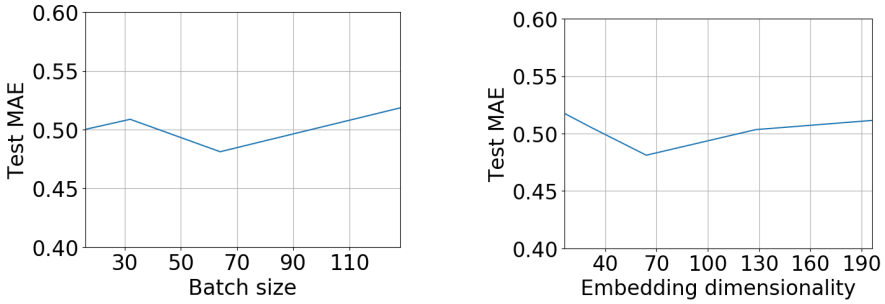
For the reproducibility of the experiments in the paper, we present the implementation details of GeDML for each dataset. Also, experiment scripts, source code of GeDML, and its embedding results are available at `https://github.com/KRICT-DATA/GeDML`. Table S2 shows hyperparameter settings of GeDML for each dataset.

Table S2: Hyperparameter settings of GeDML for each dataset. c = number of graph convolutional layer of GCN in GeDML; l = dimensionality of the vector representation created by GeDML; η = learning rate of Adam optimizer; λ = L_2 regularization coefficient; α = margin of the triplet loss.

Dataset	c	l	η	λ	α
ESOL	3	64	5e-5	1e-5	-
FreeSolv	3	64	5e-5	1e-5	-
Lipophilicity	3	64	5e-5	1e-5	-
QM7	3	64	5e-5	1e-5	-
BACE	3	64	1e-4	1e-5	0.7
BBBP	3	64	1e-4	1e-5	0.6

3 Hyperparameter Analysis

To evaluate the robustness of GeDML for its hyperparameters, we measured the prediction error of GeDML by changing the batch size and the dimensionality of the vector representation from GeDML (embedding dimensionality). We conducted the experiments on ESOL dataset[2] and changed the batch size and the embedding dimensionality in {16, 32, 64, 128} and {16, 32, 64, 128, 196}, respectively. Fig. 1 shows the experimental results.



(a) Prediction error of GeDML for various batch sizes

(b) Prediction error of GeDML for various embedding dimensionalities

Figure 1: Prediction error of GeDML on ESOL dataset for various hyperparameter values. The prediction errors were measured by changing two hyperparameters of GeDML (batch size and embedding dimensionality).

In the experiment, GeDML showed 0.4811∼0.5184 and 0.4811∼0.5176 MAE for changing batch size and embedding dimensionality, respectively. As a result, the difference between the lowest and highest errors was less than 10%.

[2] http://moleculenet.ai/datasets-1

4 RMSE

Table S3: RMSE in regression tasks for each molecular dataset. Note that the presented RMSEs are approximated values, and they were calculated based on the prediction results of each prediction model that was trained based on MAE criterion.

Dataset	GCN	GeDML+LR	GeDML+FNN	GeDML+XGB
ESOL	1.2028	0.7018	0.6955	0.6554
FreeSolv	1.7585	1.2105	1.1151	1.0131
Lipophilicity	0.9119	0.7041	0.7960	0.6933
QM7	0.7315	0.7713	0.7890	0.6771