Supporting Information

Pushing the boundaries of molecular representation for drug discovery with graph attention mechanism

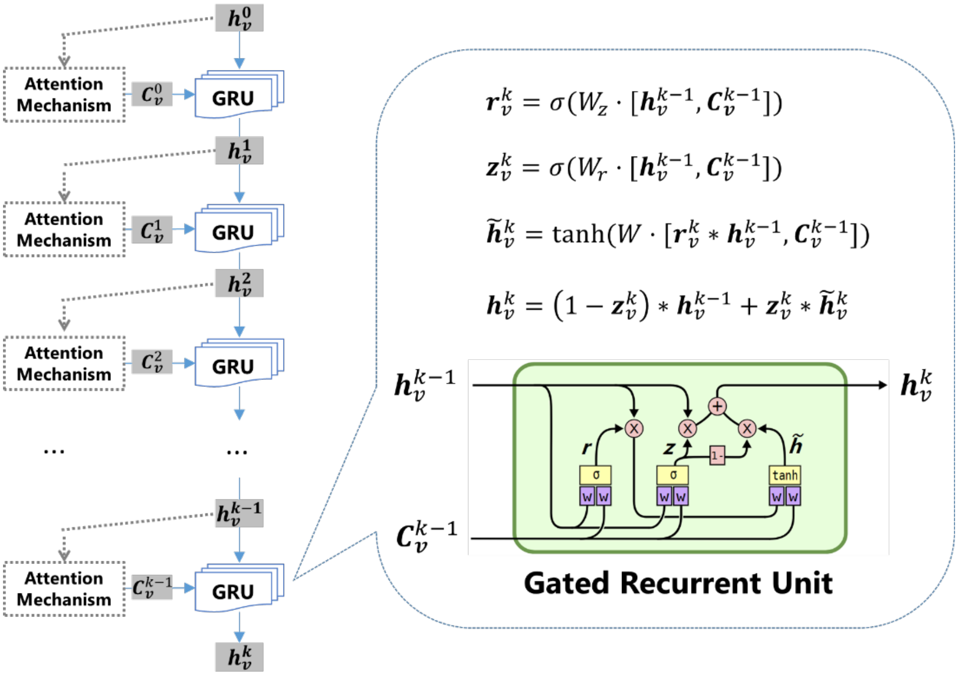
Zhaoping Xiong[1,2,3], Dingyan Wang[2,3], Xiaohong Liu[1,2], Feisheng Zhong[2,3], Xiaozhe Wan[2,3], Xutong Li[2,3], Zhaojun Li[2], Xiaomin Luo[2], Kaixian Chen[1,2], Hualiang Jiang[*,1,2], Mingyue Zheng[*,2]

[1]*Shanghai Institute for Advanced Immunochemical Studies, and School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China;*

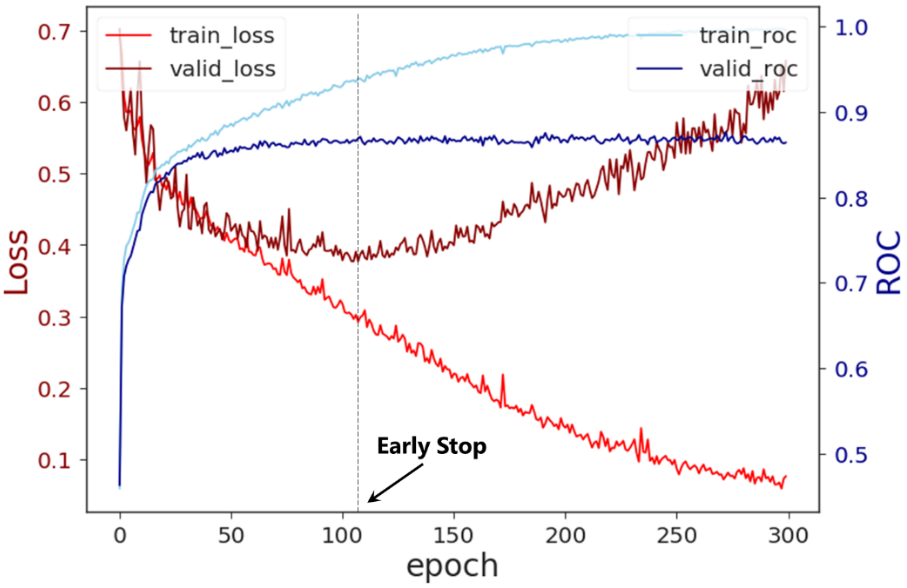
[2]*Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai, 201203, China;*

[3]*University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China*

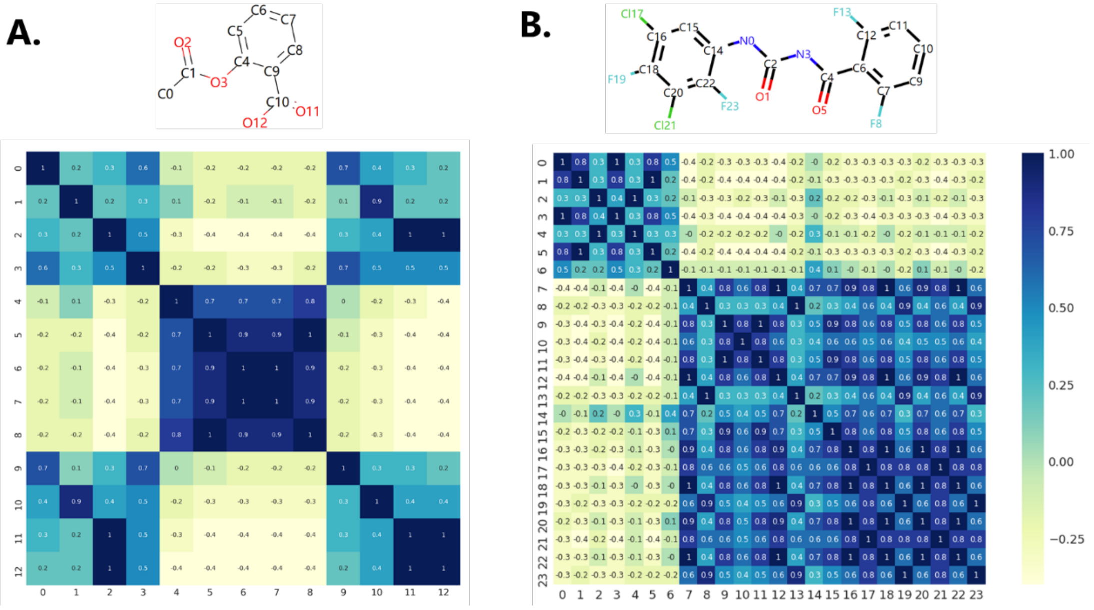
Contents



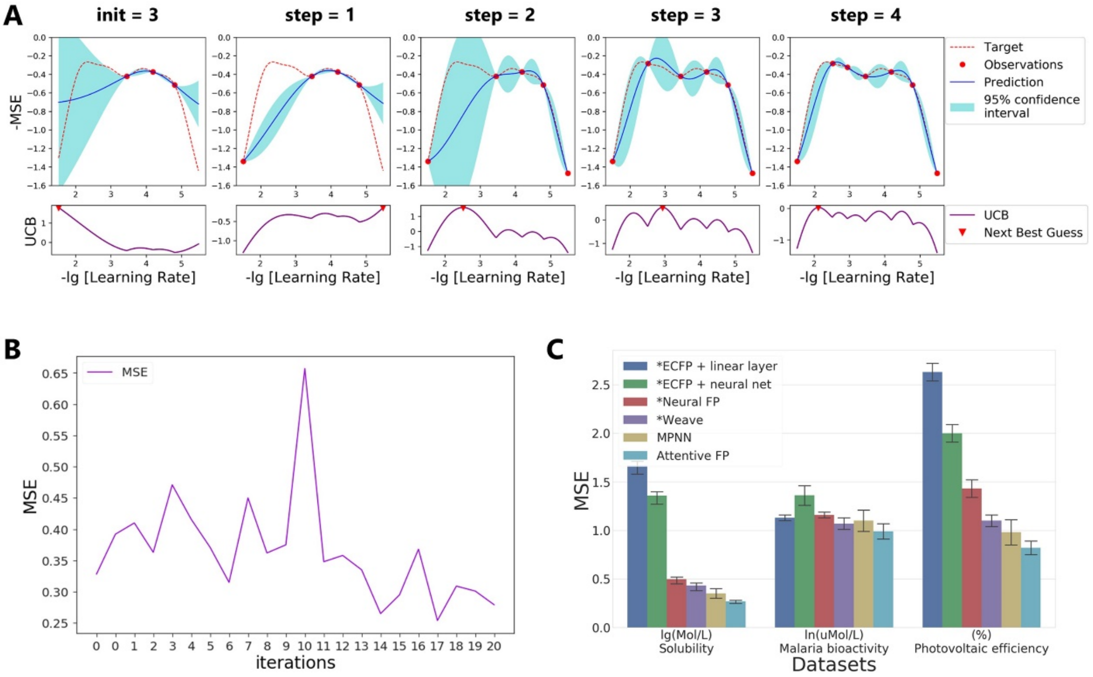
Supplementary Figure 1: Gated Recurrent Unit for node embedding renewal. $\boldsymbol{\sigma}$ is sigmoid nonlinear activation function. $\boldsymbol{r}_v^k$ is reset gate vector at time step $\boldsymbol{k}$, $\boldsymbol{z}_v^k$ is update gate vector.



Supplementary Figure 2: The Loss and ROC during training process on Tox21 tasks. Early stopping criterion for training are the Loss on validation set is no longer improving in 20 epochs and the ROC on validation set is no longer improving in 10 epochs. In this setting, it would stop at about epoch 120 to avoid overfitting, and get the best epoch 108. (This plot is not early stopped for demonstration)



Supplementary Figure 3: More example of learned hidden environment by Attentive FP model. A. Visualization of atom embedding indicates atom 9 is very different from atom 4-8 although they are in the benzene ring. An explanation is that atom 9 is connected with a potent electron-withdrawing carboxylic group, which leads to the diverse environment for atom 9. **B.** The linker atoms and two benzene rings are separated and clustered in two groups. Of note, atom 6 and atom 14 are the atoms in benzene rings that directly connect with the linkers. But they are allocated to different groups, which might be explained by the electron-drawing or electron-donating effects of their direct connected atoms.



Supplementary Figure 4: Proof of Concept Experiments. A. An example of using Bayesian Optimization for the learning rate. The target curve is generated by testing all the –log10 [Learning Rates] in the [1.5, 5.5] with a spacing of 0.02, which includes 201 points altogether. The prediction curve at the 95% confidence level was generated by the Gaussian process based on tested observations. The UCB function is used to predict the next point to test (Next Best Guess). **B.** The MSE value at each iteration during Bayesian Optimization for 6 hyper-parameters simultaneously. **C.** The predictive performances of Attentive FP models on three benchmarked datasets compared with previous models.

Supplementary Table 1: An overview of drug discovery relevant datasets

Datasets	#molecules	#Tasks	Model Types	Metrics	Description
QM9	133,885	12	Regression	MAE	DFT based quantum mechanical calculations
ESOL	1,128	1	Regression	RMSE	Water solubility
FreeSolv	643	1	Regression	RMSE	Solvation free energy
Lipop	4,200	1	Regression	RMSE	Lipophilicity
MUV	93,127	17	Classification	ROC	The Maximum Unbiased Validation (MUV) group selected from PubChem BioAssay
HIV	41,913	1	Classification	ROC	Inhibition to virus HIV replication
BACE	1,522	1	Classification	ROC	Inhibition to human β-secretase 1 (BACE-1)
BBBP	2,053	1	Classification	ROC	Blood-brain barrier penetration
Tox21	8,014	12	Classification	ROC	Qualitative toxicity measurements on 12 targets
ToxCast	8,615	617	Classification	ROC	Qualitative toxicity measurements over 617 experiment
SIDER	1,427	27	Classification	ROC	Adverse drug reactions for marketed drugs
ClinTox	1,491	2	Classification	ROC	Compounds approved or failed clinical trial for toxicity reasons

Supplementary Table 2: QM9 tasks for prediction

No.	Property	Unit	Description
1	mu	D	Dipole moment
2	alpha	Bohr^3	Isotropic polarizability
3	HOMO	Hartree	Energy of HOMO
4	LUMO	Hartree	Energy of LUMO
5	gap	Hartree	Gap ($\epsilon_{LUMO}-\epsilon_{HOMO}$)
6	R2	Bohr^2	Electronic spatial extent
7	ZPVE	Hartree	Zero point vibrational energy
8	U0	Hartree	Internal energy at 0 K
9	U	Hartree	Internal energy at 298.15 K
10	H	Hartree	Enthalpy at 298.15 K
11	G	Hartree	Free energy at 298.15 K
12	Cv	cal/(mol*K)	Heat capacity at 298.15 K

Supplementary Table 3: Algorithm pseudo-code and formulas for the Attentive FP neural network

Algorithm for the Attentive FP neural network

0. Given a molecule M, $\boldsymbol{v} \in Atom(M)$, $\boldsymbol{u} \in Neighbor(v)$

$\boldsymbol{A}_v \leftarrow AtomFeature(v)$; $\boldsymbol{B}_{vu} \leftarrow BondFeature(v,u)$

$\boldsymbol{i} = 0,1,\ldots,k$; $\boldsymbol{j} = 0,1,\ldots,t$

$\boldsymbol{s} \leftarrow virtual\ super\ node$ denotes whole molecule

1. Atom Embedding

```
while $\boldsymbol{i} = \boldsymbol{0}$:
    for each atom $\boldsymbol{v}$ in molecule M:
        $\boldsymbol{h}_v^0 \leftarrow \boldsymbol{relu}(W_{fc1} \cdot \boldsymbol{A}_v)$
while $\boldsymbol{i} \geq \boldsymbol{1}$:
    for each atom $\boldsymbol{v}$ in molecule M:
        for each atom $\boldsymbol{u}$ in $Neighbor(v)$:
            if $\boldsymbol{i} == \boldsymbol{1}$:
                $\boldsymbol{n}_u \leftarrow \boldsymbol{Concatenate}\ [\boldsymbol{A}_u, \boldsymbol{B}_{vu}]$
                $\boldsymbol{h}_u^0 \leftarrow \boldsymbol{relu}(W_{fc2} \cdot \boldsymbol{n}_u)$
            $\boldsymbol{e}_{vu}^{i-1} \leftarrow \boldsymbol{leaky\_relu}(W \cdot [\boldsymbol{h}_v^{i-1}, \boldsymbol{h}_u^{i-1}])$
            $\boldsymbol{a}_{vu}^{i-1} \leftarrow \boldsymbol{softmax}(e_{vu}) = \frac{\boldsymbol{exp}\ (\boldsymbol{e}_{vu}^{i-1})}{\sum_{u \in N(v)} \boldsymbol{exp}\ (\boldsymbol{e}_{vu}^{i-1})}$
            $\boldsymbol{C}_v^{i-1} \leftarrow \boldsymbol{elu}\left( \sum_{u \in N(v)} \boldsymbol{a}_{vu}^{i-1} \cdot W \cdot \boldsymbol{h}_v^{i-1} \right)$
        $\boldsymbol{h}_v^i \leftarrow \boldsymbol{GRU}(\boldsymbol{C}_v^{i-1}, \boldsymbol{h}_v^{i-1})$
```

2. Molecule Embedding

```
while $\boldsymbol{j} = \boldsymbol{0}$:
    $\boldsymbol{h}_s^0 \leftarrow \boldsymbol{Sum}(\boldsymbol{h}_v^k)$
while $\boldsymbol{1} \leq \boldsymbol{j} \leq$ t:
    for each atom $\boldsymbol{v}$ in molecule M:
        $\boldsymbol{e}_{sv}^{j-1} \leftarrow \boldsymbol{leaky\_relu}(\boldsymbol{W} \cdot [\boldsymbol{h}_s^{j-1}, \boldsymbol{h}_v^{j-1}])$
        $\boldsymbol{a}_{sv}^{j-1} \leftarrow \boldsymbol{softmax}(e_{sv}) = \frac{\boldsymbol{exp}\ (\boldsymbol{e}_{sv}^{j-1})}{\sum_{u \in N(v)} \boldsymbol{exp}\ (\boldsymbol{e}_{sv}^{j-1})}$
    $\boldsymbol{C}_s^{j-1} \leftarrow \boldsymbol{elu}\left( \sum_{v \in N(s)} \boldsymbol{a}_{sv}^{j-1} \cdot W \cdot \boldsymbol{h}_s^{j-1} \right)$
    $\boldsymbol{h}_s^j \leftarrow \boldsymbol{GRU}(\boldsymbol{C}_s^{j-1}, \boldsymbol{h}_s^{j-1})$
```

Supplementary Table 4: Bayesian optimization for solubility prediction task.

iteration	radius	T	fingerprint dimension	dropout	weight decay(L2)	learning rate	best epoch	best MSE
	Int([1,6])	Int([0,6])	Int([30,300])	[0,0.5]	10E-[0,6]	10E-[0,5]		
0	4	4	88	0.17	3.54	3.43	360	0.328
0	4	3	235	0.38	3.33	4.29	524	0.392
1	5	4	262	0.21	2.84	3.61	193	0.410
2	2	3	112	0.20	2.92	4.22	778	0.363
3	3	1	165	0.19	2.05	2.74	141	0.471
4	4	4	211	0.15	3.16	4.69	790	0.416
5	5	3	144	0.43	3.90	4.11	484	0.371
6	5	3	52	0.02	4.33	2.31	229	0.315
7	2	5	188	0.00	2.00	2.00	121	0.450
8	4	3	174	0.06	2.75	4.09	782	0.362
9	4	1	192	0.03	2.85	3.18	363	0.375
10	2	1	69	0.22	2.97	4.72	797	0.657
11	6	5	108	0.00	5.00	2.00	81	0.348
12	2	2	127	0.37	3.28	3.83	435	0.358
13	5	3	276	0.01	3.64	4.66	788	0.335
14	2	2	248	0.31	4.97	2.53	115	0.265
15	5	3	98	0.01	4.98	3.94	704	0.295
16	5	1	81	0.00	2.43	2.94	283	0.368
17	2	2	206	0.35	4.45	2.57	98	0.254
18	4	4	219	0.36	4.32	3.67	402	0.309
19	3	4	151	0.00	5.00	3.84	341	0.301
20	2	4	204	0.17	4.51	2.95	173	0.279

* weight decay and learning rate are rescaled by –log10.

Supplementary Table 5: Attentive FP performances on qm9 datasets (MAE)

Task	Sample MAD	Training	Validation	Test
mu	1.189±0.012	0.368±0.002	0.438±0.004	0.451±0.006
alpha	6.299±0.053	0.474±0.003	0.495±0.003	0.492±0.008
HOMO	0.016±0.00006	0.00315±0.00012	0.00356±0.00023	0.00358±0.00018
LUMO	0.039±0.00008	0.00388±0.00010	0.00418±0.00021	0.00415±0.00020
gap	0.040±0.003	0.00480±0.00022	0.00520±0.00012	0.00528±0.00015
R2	202.017±0.522	25.359±0.187	26.622±0.628	26.839±0.913
ZPVE	0.026±0.0010	0.00188±0.000018	0.00138±0.00015	0.00120±0.00016
U0	31.073±0.345	0.845±0.062	0.893±0.028	0.898±0.016
U	31.071±0.330	0.845±0.055	0.895±0.026	0.893±0.014
H	31.072±0.335	0.843±0. 052	0.855±0. 029	0.893±0.019
G	31.072±0.338	0.848±0. 073	0.845±0. 025	0.893±0.018
Cv	3.204±0.042	0.246±0.003	0.253±0.003	0.252±0.005

Supplementary Table 6: Predictive performance on bioactivities and properties for drug discovery.

Category	Datasets	#Mol.	#Tasks	Metrics	Training	Validation	Test	Average best epoch	radius	T	fingerprint dimension	dropout	weight decay(L2)	learning rate
Physical Chemistry	ESOL	1128	1	RMSE	0.290±0.065	0.496±0.032	0.503±0.076	83	2	2	200	0.3	5	2.5
	FreeSolv	643	1	RMSE	0.398±0.031	0.693±0.032	0.736±0.037	104	2	2	200	0.3	5	2.5
	Lipop	4200	1	RMSE	0.151±0.008	0.568±0.013	0.578±0.018	127	2	4	200	0.3	5	2.5
Bioactivity	MUV	93127	17	PRC	0.431±0.112	0.213±0.032	0.221±0.047	28	3	2	250	0.2	3.5	3.7
				ROC	0.951±0.012	0.846±0.015	0.843±0.012							
	HIV	41913	1	ROC	0.924±0.014	0.835±0.019	0.832±0.021	69	4	2	150	0.1	2.9	3.5
	BACE	1522	1	ROC	0.910±0.006	0.861±0.008	0.850±0.012	130	3	2	150	0.1	2.9	3.5
Physiology or Toxicity	BBBP	2053	1	ROC	0.96±0.012	0.912±0.014	0.920±0.015	198	3	2	150	0.1	2.9	3.5
	Tox21	8014	12	ROC	0.943±0.013	0.860±0.005	0.858±0.014	85	3	3	200	0.5	3	3.5
	ToxCast	8615	617	ROC	0.948±0.025	0.809±0.025	0.805±0.022	205	3	3	200	0.5	3	3.5
	SIDER	1427	27	ROC	0.869±0.012	0.612±0.015	0.637±0.017	45	3	3	200	0.5	3	3.5
	ClinTox	1491	2	ROC	0.961±0.008	0.942±0.009	0.940±0.018	70	3	3	200	0.5	3	3.5