Supplementary materials

S1 Comparison with other fragmentation methods

FraGVAE and N-Gram Graph are existing fragment-based molecular property prediction models. Although these methods also breaks molecular graphs into fragments, these fragments are invalid in chemistry. Fig. S1 shows an example of breaking an molecule into fragments by FraGVAE and N-Gram Graph methods. It is obvious that the existing two fragment definitions will break an atomic group into small fragments that no longer represent a valid atomic group. Specifically, it will break the aromatic rings. And these small fragments cannot represent larger atomic groups or pharmacophores.

While our fragment definition is based on the common structure of atomic groups. It uses acyclic single bonds as breakable bonds, which can be considered as boundaries of atomic groups. As is shown in Fig. 1, the fragments generated by our method will always be valid in chemistry. And an atomic group will be preserved at least in one of the fragments. So, it makes it possible for the model to learn the latent relationship between functional groups and molecular properties.

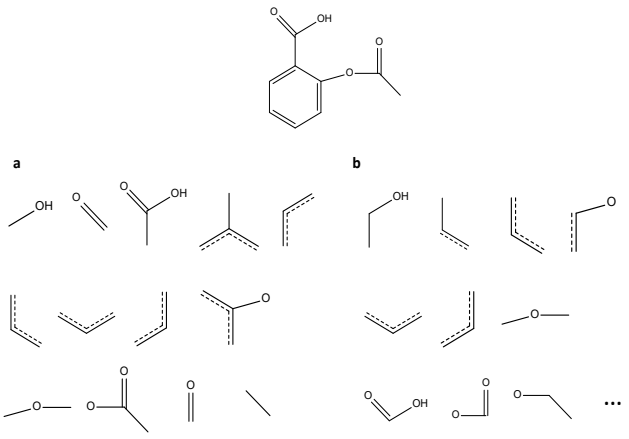

Figure S1: The fragments of aspirin extracted by (a) FraGVAE and (b) N-Gram Graph. Here for N-Gram Graph, n is set to be 3. Not all 3-gram fragments are listed.

S2 Statistical information of datasets

Table S1 shows the statistical information of datasets that we used in this work. In this table, n denotes the number of molecules in the dataset, N_{mean} denotes the average number of atoms, N_b^{mean} denotes the average number of breakable bonds, and N_b^{max} denotes the maximum number of breakable bonds. Table S2 shows the distribution of positive/negative samples in datasets with classification tasks. In this table, n_{pos} and n_{neg} denotes the number of positive/negative samples, respectively.

Table S1: Statistical information of benchmark datasets.

Category	Datasets	Tasks	n	N_{mean}	N_b^{mean}	N_b^{max}
Physio-chemical property	ESOL	1	1128	13.29	5.17	25
	FreeSolv	1	642	8.72	4.11	16
	QM9	12	133247	8.80	2.76	8
	Lipop	1	4200	27.04	8.80	51
	CEP	1	29978	27.66	1.41	4
Bioactivity	HIV	1	41127	25.51	8.71	161
	BACE	1	1513	34.09	13.74	62
	SHP2	1	865	29.41	8.14	19
	MUV	17	93087	24.23	7.69	23
	Malaria	1	9999	30.36	10.33	87
Physiology and Toxicity	BBBP	1	2050	24.06	8.29	52
	Tox21	12	7831	18.57	8.11	73
	SIDER	27	1427	33.64	15.21	365
	ClinTox	2	1484	26.16	10.67	87
	ToxCast	617	8597	18.78	8.09	76

Table S2: Distribution of pos/neg samples of datasets with classification tasks.

Dataset	Task	Total number	n_{pos}	n_{pos}/n_{neg}
Tox21	NR-AR	7265	309	0.044
	NR-AR-LBD	6758	237	0.036
	NR-AhR	6549	768	0.133
	NR-Aromatase	5821	300	0.054
	NR-ER	6193	793	0.147
	NR-ER-LBD	6955	350	0.053
	NR-PPAR-gamma	6450	186	0.030
	SR-ARE	5832	942	0.193
	SR-ATAD5	7072	264	0.039
	SR-HSE	6467	372	0.061
	SR-MMP	5810	918	0.188
	SR-p53	6774	423	0.067
ClinTox	FDA_APPROVED	1484	1390	14.787
	CT_TOX	1484	112	0.082
SIDER	SIDER1	1427	743	1.086
	SIDER2	1427	996	2.311
	SIDER3	1427	22	0.016
	SIDER4	1427	876	1.590
	SIDER5	1427	1151	4.170
	SIDER6	1427	997	2.319
	SIDER7	1427	1298	10.062
	SIDER8	1427	251	0.213
	SIDER9	1427	1024	2.541
	SIDER10	1427	727	1.039
	SIDER11	1427	376	0.358
	SIDER12	1427	1292	9.570
	SIDER13	1427	323	0.293
	SIDER14	1427	213	0.175
	SIDER15	1427	1108	3.473
	SIDER16	1427	885	1.633
	SIDER17	1427	1318	12.092
	SIDER18	1427	253	0.216
	SIDER19	1427	1006	2.390
	SIDER20	1427	1060	2.888
	SIDER21	1427	1016	2.472
	SIDER22	1427	911	1.766
	SIDER23	1427	125	0.960
	SIDER24	1427	659	0.858
	SIDER25	1427	988	2.251
	SIDER26	1427	1304	10.602
	SIDER27	1427	946	1.967
HIV	HIV	41127	1443	0.036
BACE	BACE	1513	691	0.841
BBBP	BBBP	2050	1567	3.244

S3 Information and the patents of SHP2 dataset.

The patents that we used to construct the SHP2 dataset is listed in Table S3. Among the molecules proposed in these patents, only the molecules of which IC50 values are not larger than 10 μM are considered to have good binding affinities and selected to build the SHP2 dataset.

Table S3: Information of the SHP2 dataset and the patents.

Patent Number	selected number of molecules
WO2015107493	29
WO2015107494	50
WO2015107495	88
WO2016203404	83
WO2016203405	193
WO2016203406	120
WO2017211303	5
WO2017216706	66
WO2018013597	4
WO2018057884	41
WO2018081091	107
WO2018136265	5
WO2018172984	21
WO2019067843	25
WO2019075265	28
Total	865

S4 Results of the experiments on QM9 dataset.

Table S4 shows the performance of models on different tasks of QM9 benchmark. The top-2 models are bolded. Comparing the results of FraGAT and Attentive FP, it can be figured out that the FraGAT model can achieve better performance on 8 of 12 tasks. The experiments on QM9 dataset in N-Gram Graph are **not** conducted in multi-task learning way, but training models seperately for each task. So that the performances of the N-Gram Graph model is relatively better on some tasks. However, the FraGAT model can still achieve better performance on 7 of 12 tasks. Values of Attentive FP and baselines are cited from (Xiong *et al.*, 2019)

Table S4: The performance on different tasks of QM9 benchmark.

Task	DTNN	GC	MPNN	Attentive FP	N-Gram XGB	FraGAT
mu	**0.244**	0.583	0.358	**0.451**	0.535	0.479
alpha	0.95	1.37	0.89	**0.492**	0.612	**0.446**
homo	0.00388	0.00716	0.00541	**0.00358**	0.005	**0.00356**
lumo	0.00513	0.00921	0.00623	**0.00415**	0.005	**0.00435**
gap	0.0066	0.0112	0.0082	**0.00528**	0.007	**0.00538**
r2	**17**	35.9	28.5	**26.839**	59.137	28.576
zpve	0.00172	0.00299	0.00216	0.00120	**0.000**	**0.00107**
u0	2.43	3.41	2.05	0.898	**0.427**	**0.658**
u298	2.43	3.41	2	0.893	**0.428**	**0.658**
h298	2.43	3.41	2.02	0.893	**0.428**	**0.658**
g298	2.43	3.41	2.02	0.893	**0.428**	**0.658**
Cv	0.27	0.65	0.42	**0.252**	0.334	**0.216**

S5 Influence of our data-augmentation method to the performance

During the evaluation step, the data-augmentation method is used for dealing with the additional randomness. To quantitively measure this randomness and the influence of our data-augmentation method to the performance of the model, we conduct a supplementary experiment. In this experiment, we test well-trained models on four datasets: ESOL, SHP2, ClinTox, SIDER. A fixed test set is used for evaluation. During the evaluation step, we no longer input all of the N_b samples, but input $\alpha * N_b$ samples, where α is a parameter to adjust the batch size of the augmented samples. Here 5 values of α are selected. And for each value, the evaluation is repeated for 50 times on the fixed test set.

The result of this experiment is shown in Fig. S2 and Table S5. The statistical information of N_b of the test sets of these four datasets is shown in Table S6. From Fig.S2 and Table S5, we can see that as the increase of α, the performance fluctuation decreases, which means the model is more stable. This indicates that by using the data-augmentation strategy, the performance uncertainty can be restrained. The slightly improved average index indicates that there might be some samples that the model cannot predict accurately. While using data-augmentation strategy, the influence of these difficult samples will be reduced. So that the average performance of the model can achieve a better level.

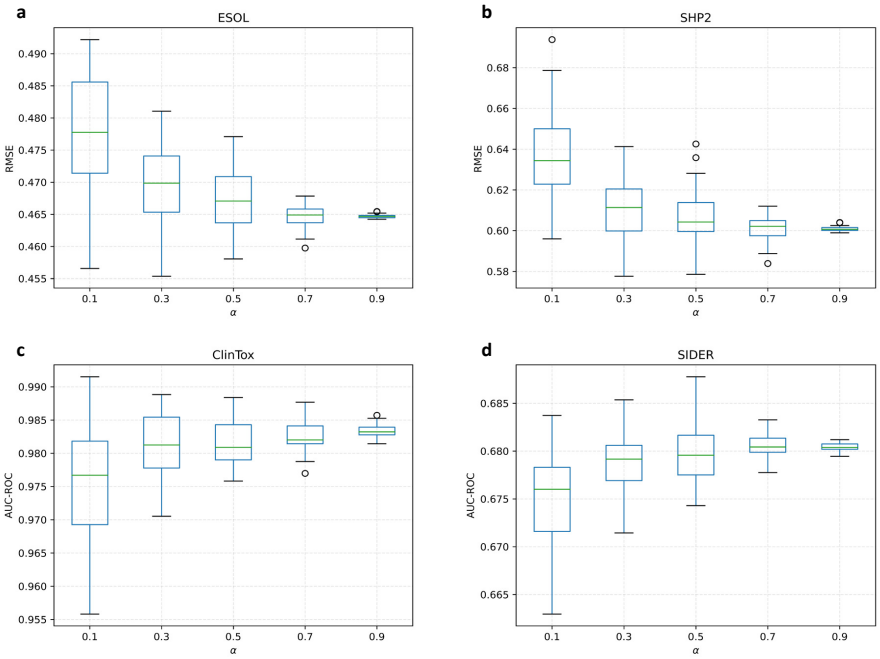

Figure S2: The result of the randomness experiments.

Table S5: The detailed result of the randomness experiments.

Dataset	α	P_{mean}	$P_{variance}$
ESOL	0.1	0.4780	0.0090
	0.3	0.4697	0.0061
	0.5	0.4674	0.0049
	0.7	0.4647	0.0018
	0.9	0.4647	0.0003
SHP2	0.1	0.6369	0.0203
	0.3	0.6101	0.0146
	0.5	0.6063	0.0139
	0.7	0.6008	0.0064
	0.9	0.6009	0.0011
ClinTox	0.1	0.9756	0.0081
	0.3	0.9810	0.0049
	0.5	0.9814	0.0031
	0.7	0.9825	0.0023
	0.9	0.9834	0.0009
SIDER	0.1	0.6748	0.0047
	0.3	0.6789	0.0029
	0.5	0.6796	0.0028
	0.7	0.6806	0.0012
	0.9	0.6805	0.0004

Table S6: Distribution of N_b of the test sets.

Dataset	N_b^{mean}	N_b^{max}
ESOL	5.36	20
SHP2	8.48	14
ClinTox	10.53	52
SIDER	11.39	70

S6 Number of learnable parameters of ablation models

The number of learnable parameters of M12 and FraGAT models are listed in Table S7. Those models that can achieve the best performance on each dataset are selected.

In fact, the numbers of learnable parameters of M12 and FraGAT are largely influenced by the length of the latent vector, F, and the structure of the classifier. These hyperparameters are determined automatically and they may be different in M12 and FraGAT. And the complexity of FraGAT is not always larger than that of M12. For example, for the M12 model trained on ESOL dataset, F=32 and the numbers of cells of each layer of the classifier are [32*3, 1]. While for the FraGAT model on ESOL, F=150 and the numbers of cells are [150*4, 512, 1]. So the total number of parameters of FraGAT on ESOL is about 55 times of that of M12. However, the situation is different on BBBP dataset. For M12, F=200 and the numbers of cells for each layer are [200*3, 128, 32, 2]. While for FraGAT, F=32 and the numbers are [32*4, 512, 2]. In this case, the total number of parameters of the M12 is almost 10 times of that of the FraGAT model.

Except for those special cases, compared with M12, FraGAT generally has more learnable parameters. As is shown in Table S7, the total number of parameters of M12 is about 0.6 to 0.8 times of that of the FraGAT model.

Table S7: The number of learnable parameters.

benchmark	M12	FraGAT	M12 / FraGAT
ESOL	30954	1729046	0.018
FreeSolv	1400705	1777656	0.788
HIV	504715	871569	0.579
BACE	1400738	2435003	0.575
BBBP	1181131	121041	9.758
Tox21	33185	57063	0.582
SIDER	36095	60933	0.592
ClinTox	31245	54483	0.573

S7 Influence of parameters of graph model to the performance

In this part, the influence of the parameters of graph model to the performance of FraGAT is tested. In our experiments, four datasets, ClinTox, BACE, ESOL and FreeSolv, are used. For each dataset, three parameters of the graph model, layers of Attentive FP for atom embeddings (denoted as k), layers of Attentive FP for molecule embeddings (denoted as T), and the length of the latent vector (denoted as F), are changed during the experiments. And the other parameters are set to be the same as those used in the benchmark experiments and remain unchanged. The model is trained and evaluated for 5 times for each parameter combination, and the mean of the metrics on the test set is reported. The results are shown in Fig. S3.

From Fig. S3, although we can see that the model will achieve an optimal performance on some specific parameter combination, the relationship of these parameters to the performance of the model is not obvious from these results. It is known that the number of layers of the graph model determines the distance that the information propagates in the graph. Thus, there must be some association among the radius of the graph, the number of the layers and the performance of the model. However, in our experiments, the radius of the molecular graphs and the fragments are diverse. So that the most appropriate number of layers for each graph might be different. Thus, it is not easy to find the concrete relationship between the parameter of the graph model and prediction performance. We intend to leave this issue for a future work.

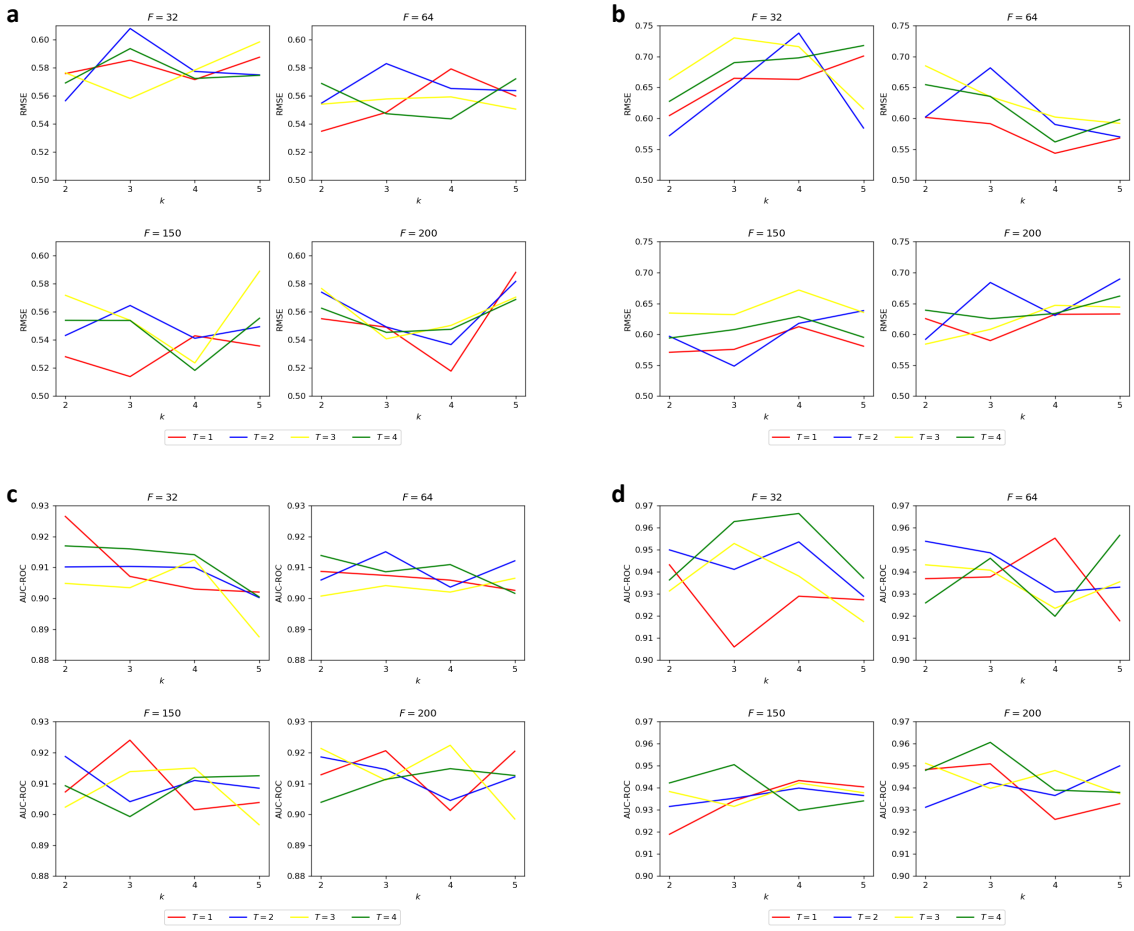

Figure S3: Influence of parameters of graph model to the performance. (a) ESOL (b) FreeSolv (c) BACE (d) ClinTox.

S8 Detailed results of case studies

Table S8 shows the detailed result of the interpretability experiment. The upper, middle and lower part of the table are the response of the model toward different samples of molecule a, b and c, respectively.

Table S8: Responses of the model toward different samples of three molecules.

bond number	y_i	l	absolute error	error ranking
0	0.159	0.064	0.095	5
1	0.040	0.064	0.025	4
2	0.056	0.064	0.009	2
3	0.043	0.064	0.021	3
4	0.302	0.064	0.238	6
5	0.058	0.064	0.006	1
0	0.409	0.024	0.385	11
1	0.039	0.024	0.015	1
2	-0.128	0.024	0.152	9
3	-0.039	0.024	0.063	3
4	-0.065	0.024	0.089	4
5	-0.065	0.024	0.089	4
6	-0.065	0.024	0.089	4
7	-0.079	0.024	0.103	7
8	0.330	0.024	0.306	10
9	-0.026	0.024	0.050	2
10	-0.080	0.024	0.104	8
0	0.163	0.003	0.160	8
1	0.006	0.003	0.003	1
2	0.150	0.003	0.147	7
3	0.043	0.003	0.040	5
4	-0.050	0.003	0.053	6
5	-0.009	0.003	0.012	2
6	-0.011	0.003	0.014	3
7	-0.011	0.003	0.014	4

S9 SHP099

SHP099 (Fortanet *et al.*, 2016) is a template molecule for studying drugs of target SHP2 in recent years. Researchers study to find molecules with better binding affinity by modify the structure of SHP099. The binding affinity of SHP099 is 0.07 μM, and its structure, which is shown in Fig. S4, consists of 3 parts: aryl, central ring and heterocycle.

According to the X-ray cocrystal analysis in (Fortanet *et al.*, 2016), the binding affinity is mainly contributed by two interactions: the PHE-113 interaction with the amino-group on the heterocycle and cationic-π stacking interaction between the aryl, central ring and ARG-111. In (Fortanet *et al.*, 2016), it is shown that the amino-group on the heterocycle will form ionic bonds with the SHP2 protein. And the bond energy of the ionic bond is so larger that this amino-group contributes major binding affinity to the molecule. For the interaction between SHP099 and ARG-111, it is revealed by (Fortanet *et al.*, 2016) that the ortho-chlorine on the Aryl effectively fill a hydrophobic pocket on the SHP2 protein, which is benifitial to binding. Besides, the amino-group on the central ring also form a hydrogen bond with GLU-250.

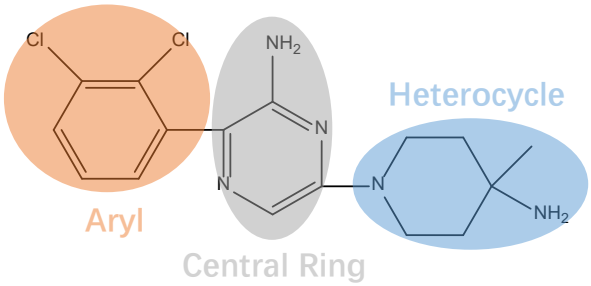

Figure S4: The structure of molecule SHP099.

S10 Molecular Docking Result

Fig. S5 shows the molecular docking result of molecule *c* in Sec. 3.3. Just like the discussion of SHP099 in Sec. S9, similar interactions are revealed in this figure, including the interaction between aryl, central ring and ARG-111, the interaction between amino-group on spirocycle and PHE-113 and the H-bond between amino-group on the central ring and GLU-250. Besides, the hydroxy on the spirocycle can form an extra hydrogen bond with GLU-249, which further increase the binding affinity.

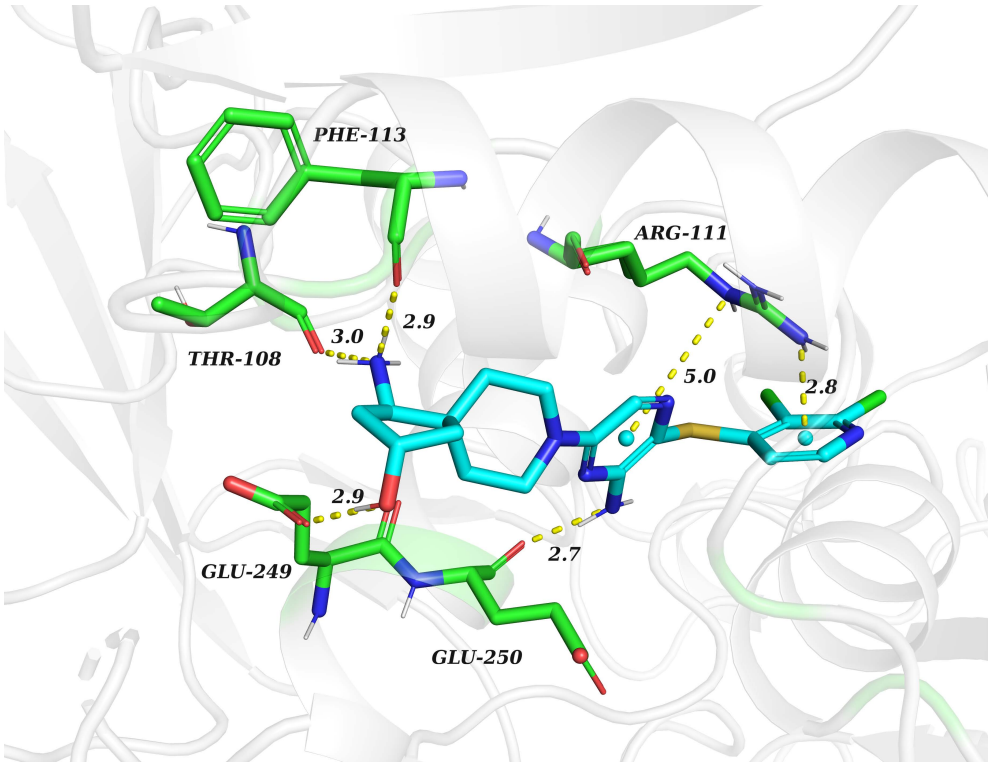

Figure S5: The docking result of molecule *c*.