Table 1: Average consistency scores of different evaluation metrics (corresponding to Table 1 in the main paper).

	SHAPEGGEN	TR3	MUTAG	MNIST	BA3
GEF	0.800	0.734	0.800	0.934	0.734
SimOAR	0.800	0.867	0.800	0.934	0.934
OAR	0.867	0.934	1.000	0.934	1.000

Table 2: Correlation between metrics and Recall across explanatory subgraphs in four node classification datasets (corresponding to Figure 3 (a) in the main paper).

	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid
RM	0.312	0.321	0.295	0.411
DSE	0.406	0.377	0.343	0.384
OAR	**0.542**	**0.560**	**0.489**	**0.440**
SimOAR	0.527	0.535	0.471	0.431

Table 3: Correlation between metrics and Recall across explanatory subgraphs, where "+ L_1" represents employing the L_1-norm to constrain the size of the explanations (corresponding to Figure 3 (a) in the main paper).

	TR3	MUTAG	MNIST	BA3
SimOAR	0.867	0.800	0.934	0.934
SimOAR + L_1	**0.93**4 ↑	0.800	0.934	**1.000** ↑
OAR	0.934	1.000	0.934	1.000
OAR + L_1	0.934	1.000	**1.000** ↑	1.000

Table 4: The comparison between VGAE and spectrum (corresponding to Figure 3 (a) in the main paper). OAR represents the performance of our paradigm encapsulating VGAE as the OOD block, while "OAR+Spectra" represents the metric replacing VGAE with spectra.

	MUTAG	BA3	TR3	MNIST-sp
OAR	0.567	0.503	0.483	0.455
OAR + Spectra	0.511	0.459	0.449	0.433
Variation (%)	5.6 ↓	4.4 ↓	3.4 ↓	2.2 ↓

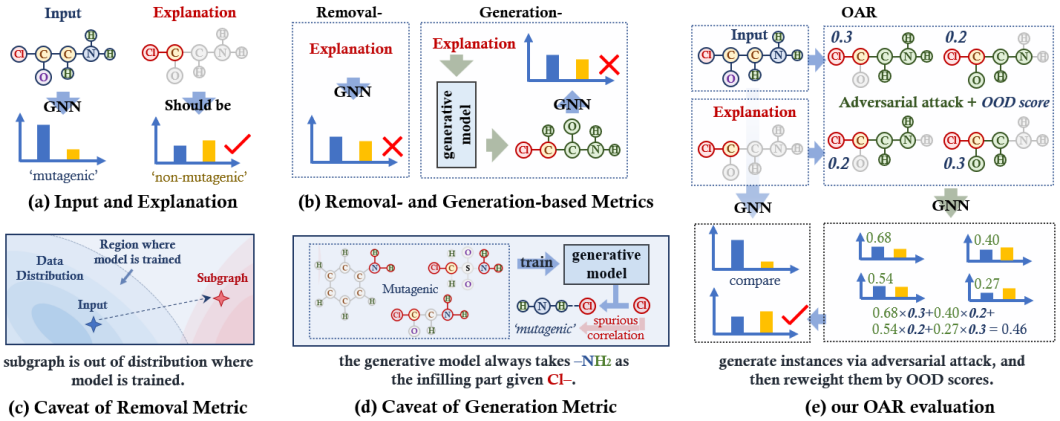

Figure 1: Pipelines and flaws of different evaluation methods. In the "Input" graph, -NH_2 is considered as the ground truth explanation for its mutagenicity. Best viewed in color.