Supplementary: Explainability Methods for Graph Convolutional Neural Networks

Phillip E. Pope*
HRL Laboratories, LLC
pepope@hrl.com

Soheil Kolouri*
HRL Laboratories, LLC
skolouri@hrl.com

Mohammad Rostami
HRL Laboratories, LLC
mrostami@hrl.com

Charles E. Martin
HRL Laboratories, LLC
cemartin@hrl.com

Heiko Hoffmann
HRL Laboratories, LLC
hhoffmann@hrl.com

In the supplementary material, we provide additional information about the contrastivity and sparsity metrics, additional visualizations of explainability methods, and details on our substructure frequency analysis.

1. Contrastivity and Sparsity Metrics

An illustration of the contrastivity and sparsity metrics may be found in Figure 1.

2. Additional Visualizations of Explainability Methods

More extensive visualizations comparing the explainability methods may be found in Figures 2, 3 for Visual Genome datasets Indoor vs. Outdoor and Country vs. Urban, and Figures 4, 5, 6 for molecular datasets BBBP, BACE, and TOX21. Five random samples from each class are shown.

3. Substructure Frequency Analysis

Connected substructures often manifest in the saliency map of a graph obtained from an explanation method, naturally yielding candidate substructures for further analysis. We describe an automated method for counting the occurrence of salient substructures. In short, for each dataset, we count the frequency of each substructure observed in explanations. Further, we count the overall prevalence of a substructure in a class, which defines a notion of class-specificity.

To identify substructures, we took the connected components induced by the set of vertices with saliency value greater than some threshold $\tau \in [0, 1]$ (here, $\tau = 0$). We call these vertices activated. We collect the connected components induced by the activated vertices, and count their frequency. Counting subgraphs requires testing subgraphs for equivalence, the implementations of which are discussed in the following sections. We use GradCAM as the base explanation method.

More formally, let $\mathcal{G} = \{G_i\}_{i=1}^N$ be a collection of graphs with binary labels $\mathcal{Y} = \{y_i\}_{i=1}^N$. For each graph $G_i = (V_i, E_i)$, and for every vertex $v_j \in V_i$, let $a_j \in [0, 1]$ be the associated saliency value. We say that a vertex v_j is *activated* if for threshold τ, $a_j \geq \tau$. The set of activated nodes for graph G_i induces a subgraph S_i of G_i, possibly unconnected. Then, we say that each connected component c_{ij} of S_i where c_{ij} has more than one node, is a subgraph identified by the explanation method.

Let the collection of all identified subgraphs in $\mathcal{G}$ by the explanation method be denoted as $\mathcal{S}$. Next, define the counts associated to each identified substructure $s \in \mathcal{S}$ as N_e^s. Further, define N_p^s and N_n^s as the number of times a substructure s occurs in the positively labeled data, and the negatively labeled data respectively.

A substructure prevalent in the dataset may artificially show high prevalence in the collection of salient subgraphs. To account for this potential imbalance, we counted the occurrences of explanation-identified substructures in both positive and negative labeled data in the dataset. We used these counts to normalize the counts obtained from the explanations and construct three ratios: $R_e^s = \frac{N_e^s}{N_p^s + N_n^s}$, $R_p^s = \frac{N_p^s}{N_p^s + N_n^s}$, $R_n^s = \frac{N_n^s}{N_p^s + N_n^s}$. The ratio R_e^s measures the prevalence of a subgraph in explanations. The ratios R_p^s, R_n^s measure how prevalent a substructure occurs in positively and negatively labeled data respectively, and serve as a baseline for the first. Note that high R_p^s or R_n^s corresponds to high class specificity for salient subgraph s.

These ratios are sensitive to rare substructures. For instance if a substructure occurs only once in the explanations and the dataset, then it has $R_e^s = 1$. To mitigate this sensitivity, we report only substructures that occur more than 10

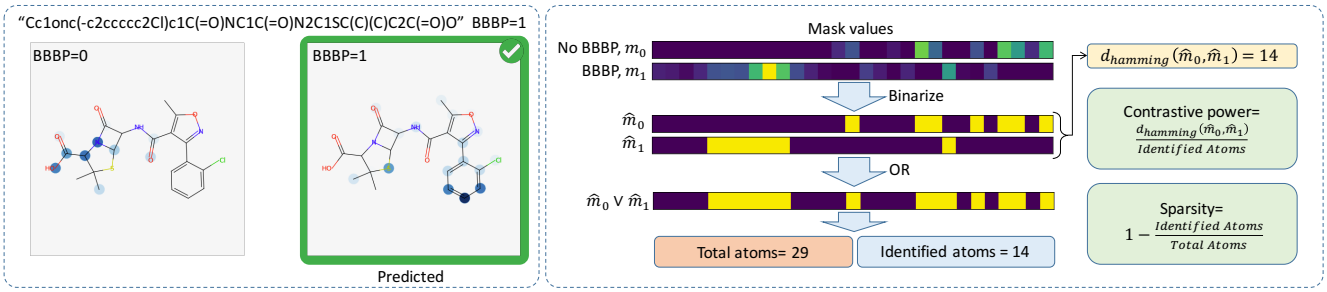

Figure 1: Visualization of the molecule "Oprea1_495871" with molecular formula "C19H17ClN3O5S", its corresponding SMILES representation, and the result of applying CAM to identify atoms that contribute to its BBBP characteristic (on the left), and the process of measuring contrastivity and sparsity of the method (i.e., CAM) for this molecule.

times in the dataset.

3.1. Visual Scene Graph Substructure Analysis

We carry above the analysis for the Visual Genome datasets, and report top 10 findings by R_e^s for each class in Figure 7.

We note that subgraphs with two vertices are the most predominant in the collection, however more complicated subgraphs exist in the saliency set.

For the indoor vs. outdoor dataset, (shelf, toiletries) is the top result for the indoor class, and the (blue sky, clouds) is the top result for the outdoor class.

For the country vs. urban dataset, (ground, straw) is the top result for the country class, and (pole, traffic lights) is the top result for the urban class.

By inspection of 7, we qualitatively confirm that the top subgraphs are consistent with the definition of each class. As this dataset was synthetically constructed, we do not place much value on these findings. However, they serve as proof-of-concept that the method returns consistent results in the case of visual scene graphs.

3.2. Molecular Substructure Analysis

We carry above the analysis for the molecular datasets, and report top 10 findings by R_e^s for each class in Figure 8.

In the case of molecules, subgraph analysis has a chemical interpretation in terms of *functional groups* that are relevant for a given molecular property, e.g., toxicity.

For graph matching we utilize the functionality found in the open source computational chemistry library RDKit.

Figure 8 shows the most prominent substructures according to our analysis. We note the identified substructures have high class specificity. In addition, we few patterns may be observed in these results: Tri-halogens (Cl,F,Br) are prevalent in explanations for BBBP, Amides are prevalent in explanations for BACE, and aromatic ring structures are prevalent for TOX21.

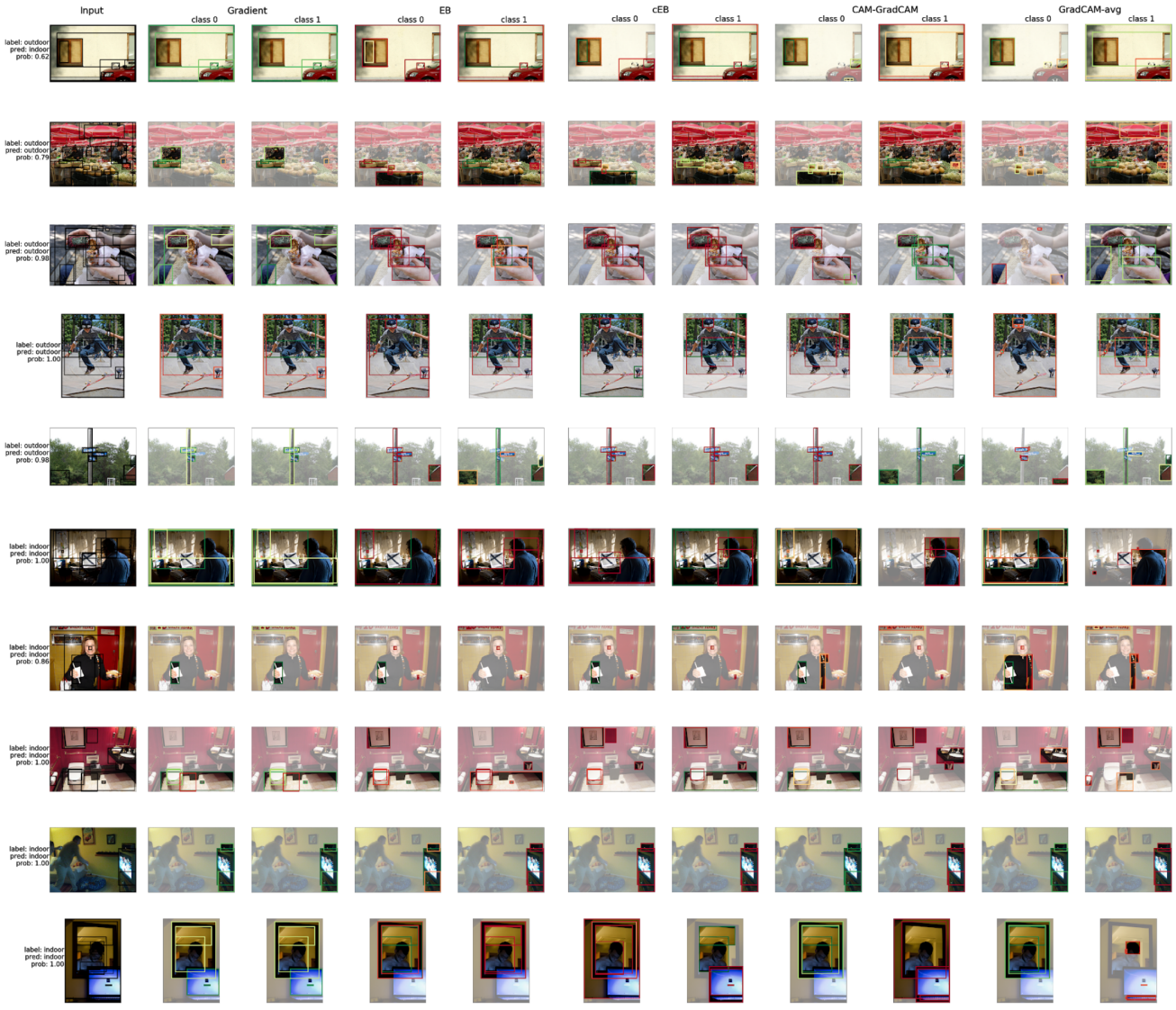

Figure 2: Additional examples for Visual Genome dataset Indoor vs. Outdoor.

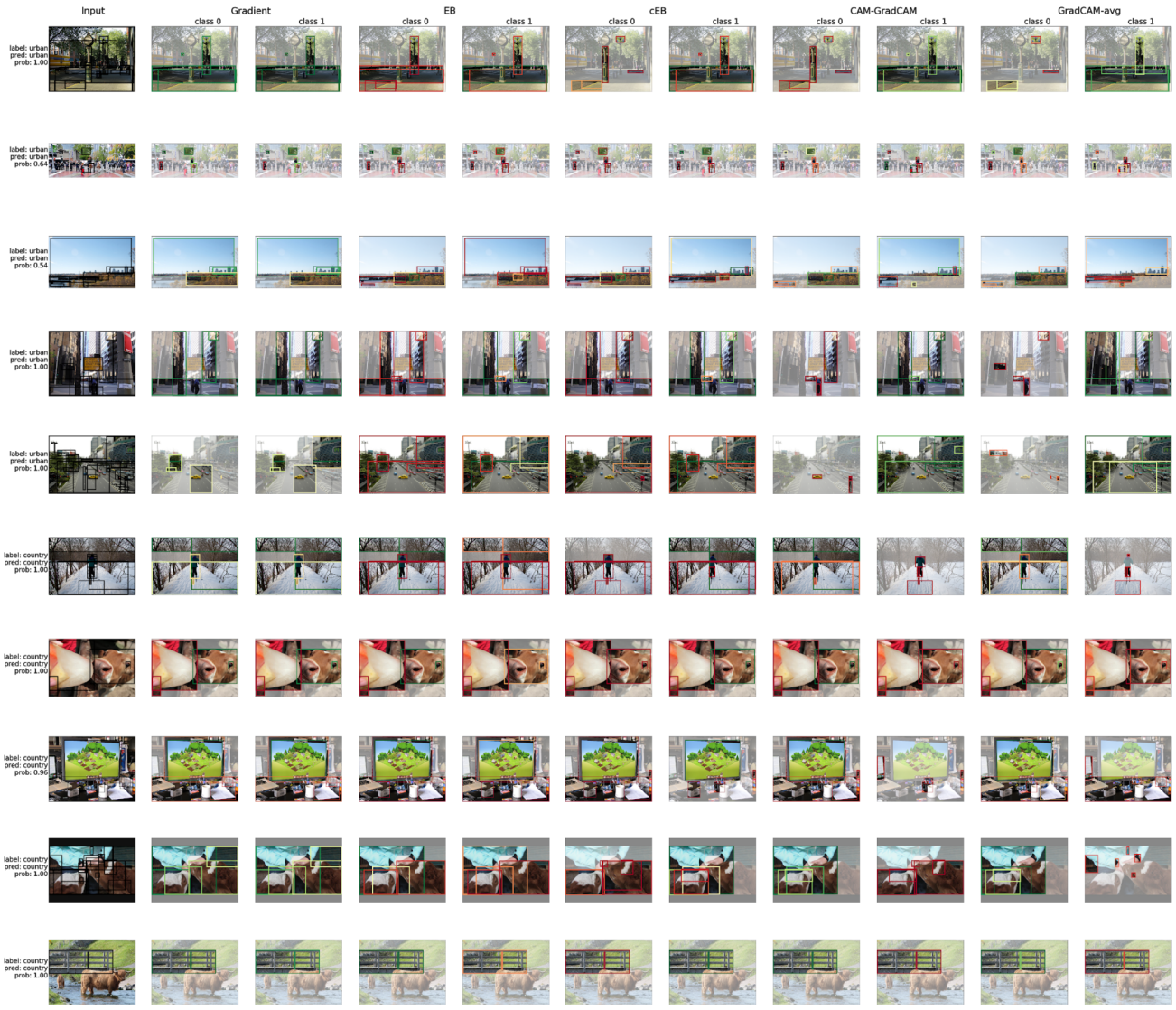

Figure 3: Additional examples for Visual Genome dataset Country vs. Urban.

Figure 4: Additional examples for molecular dataset BBBP.

Figure 5: Additional examples for molecular dataset BACE.

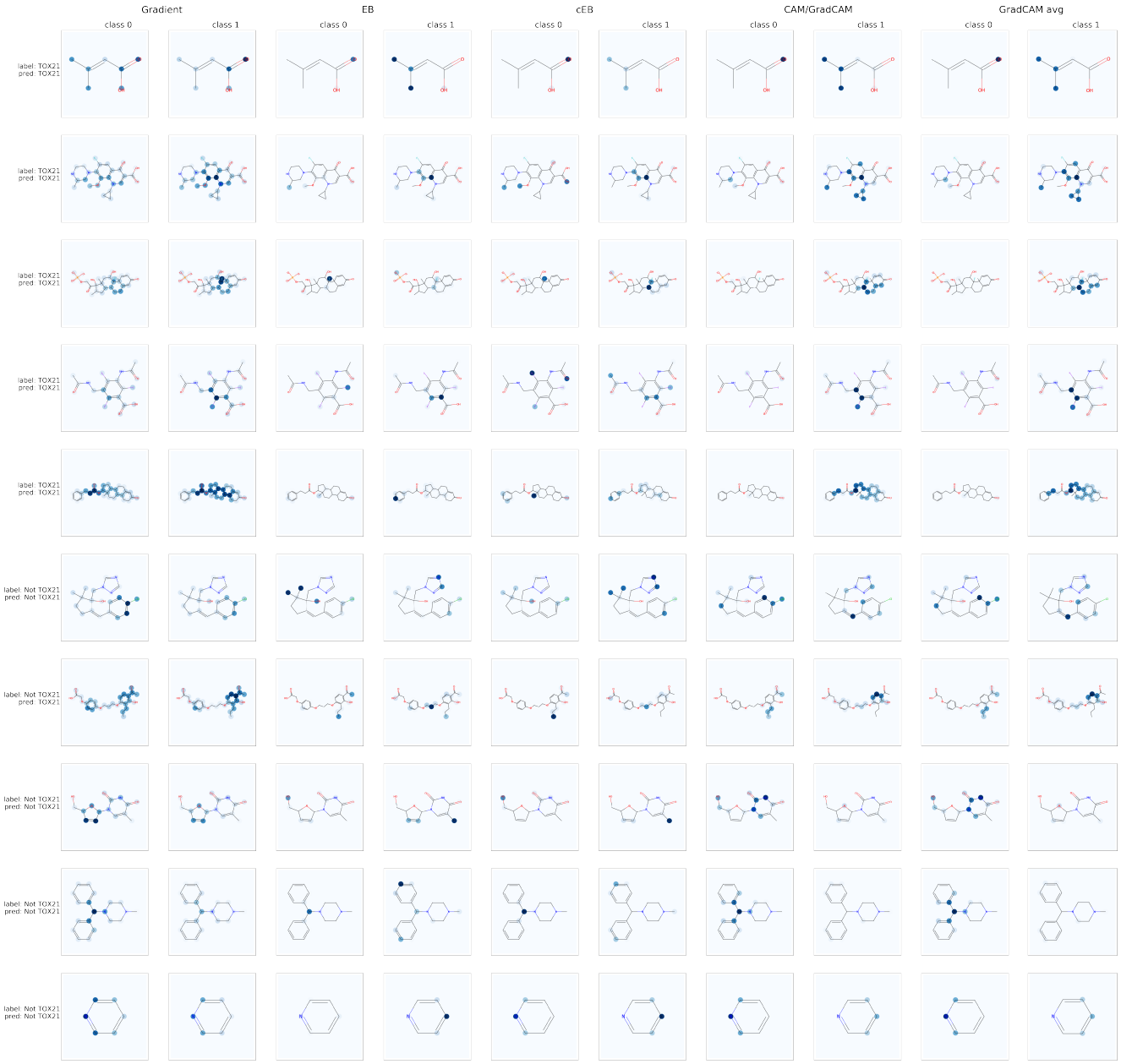

Figure 6: Additional examples for molecular dataset TOX21 (NR-ER).

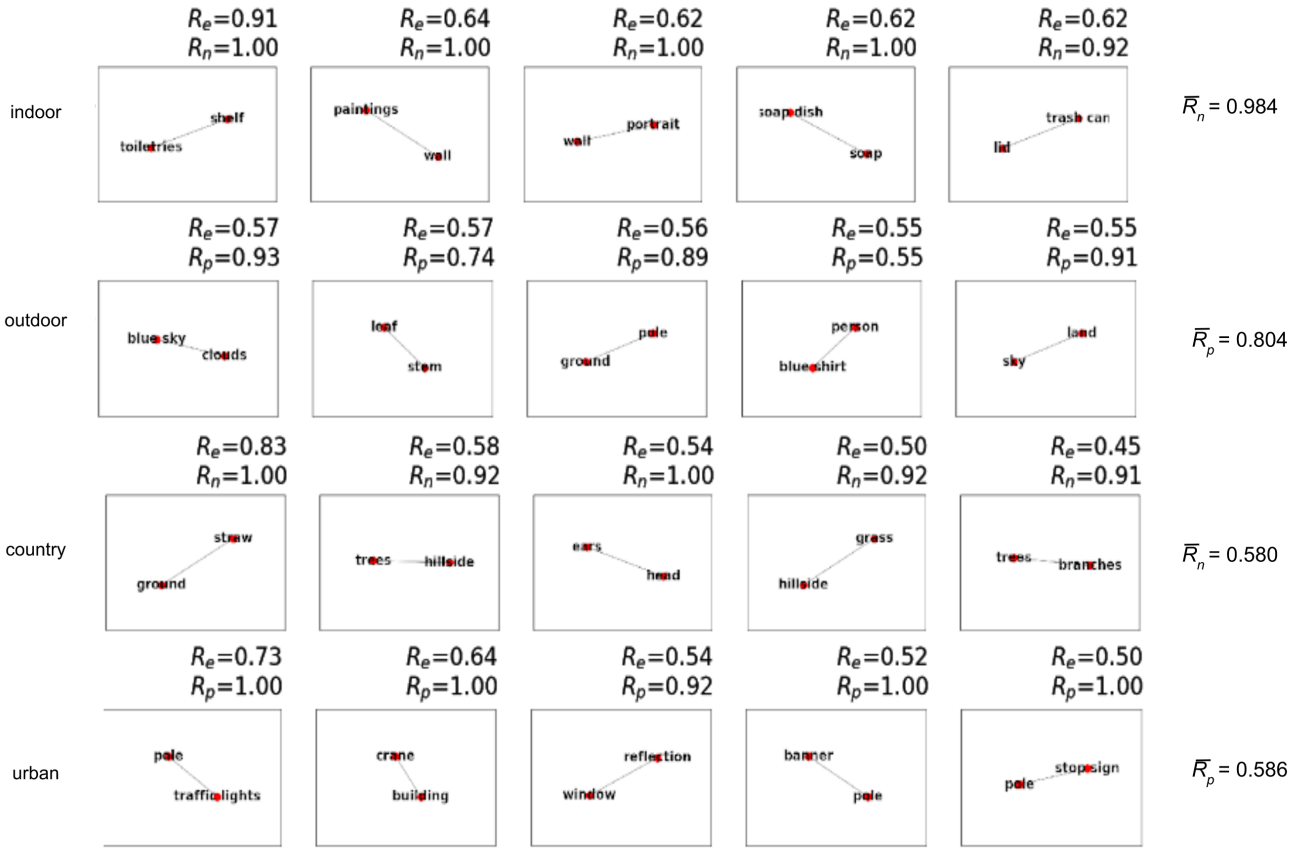

Figure 7: Top 5 most prevalent substructures for the VG dataset. We rank substructures by the ratio R_e, the number of times a substructure occurs in explanations over total occurrences in the dataset. For comparison, we also report the ratios R_p, R_n of how many times a substructure occurs in the positively and negatively labeled data respectively, over total occurrences. To account for rare structures, we report only substructures that occurred more than 10 times in the dataset.

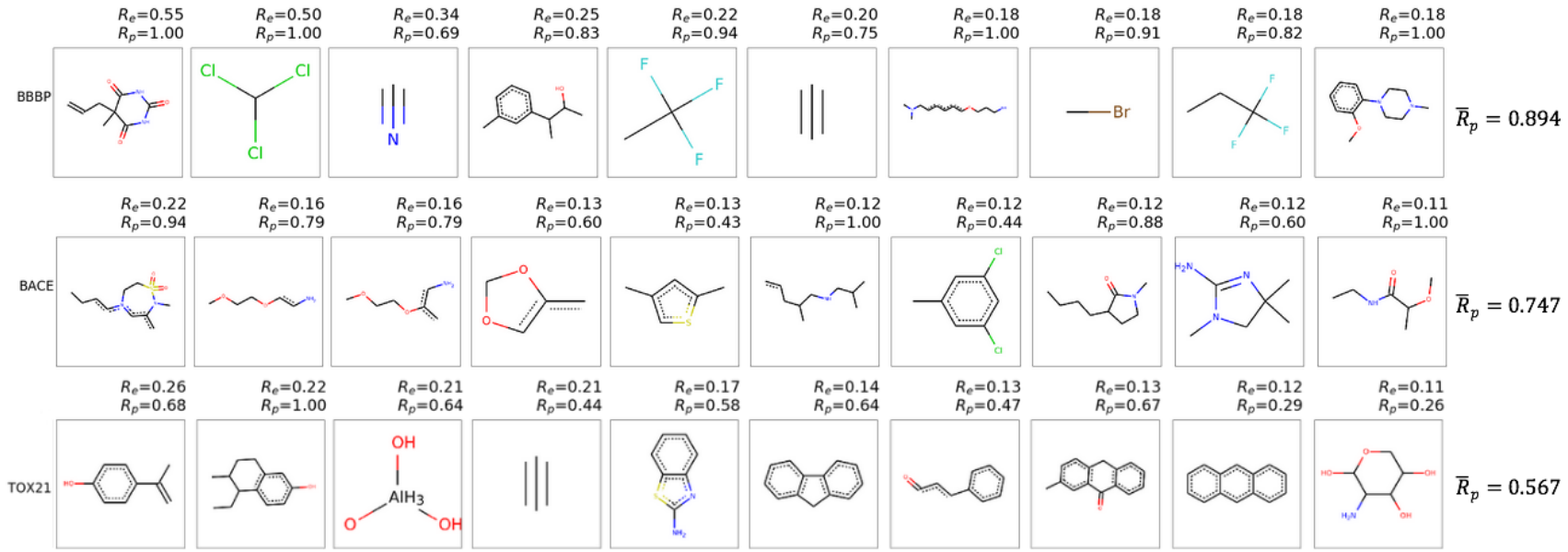

Figure 8: Top 10 most prevalent substructures for the molecular dataset. We rank substructures by the ratio R_e, the number of times a substructure occurs in explanations over total occurrences in the dataset. For comparison, we also report the ratio R_p of how many times a substructure occurs in the positively labeled set over total occurrences. To account for rare structures, we report only substructures that occurred more than 10 times in the dataset. The right-most column shows average R_p values.