A. Appendix

A.1. An Example of the Difficulty of Using Existing Methods

For this example, we are going to consider Integrated Gradients (IG) (Sundararajan et al., 2017) which produces local feature attribution explanations. Because IG is a supervised method, we start by training a classifier on top of the learned representation to get a multi-class classification model f that predicts which group a point belongs to. Because our goal is to explain the difference between Group A and Group B with IG, we average IG's explanation for each point in Group B relative to each possible baseline value of a point in Group A for f's Class B label. To be more precise:

$$\delta_{IG}(A \to B) = \frac{1}{|X_A|\,|X_B|} \sum_{x \in X_B} \sum_{a \in X_A} IG(x, \text{class} = B, \text{baseline} = a) \tag{11}$$

We will refer to this as 'group Integrated Gradients' or gIG.

Challenge 1: Comparing Explanation Types. Because IG produces feature attributions and TGT produces counterfactuals, there is no reliable metric in the literature to directly compare them. On the one hand, most feature attributions are the 'correct explanation' for their specific definitions for 'attribution' and the 'baseline' value; this has made measuring their quality challenging (Tomsett et al., 2019). On the other hand, we cannot treat a feature attribution as a transformation function/translation, so our metrics and other metrics for counterfactual explanations cannot be applied.

As a result, we compare TGT to gIG on the same synthetic dataset we used earlier. We found that gIG identifies the causal variables as being significant and ignores the noise variable, but that it also identifies the correlated variable as being significant. This indicates that it is likely to be unable to find sparse explanations as well as TGT can.

Challenge 2: Consistency of Aggregated Local Explanations. One of the reasons we chose IG as a baseline method to aggregate is because its attributions are symmetrical and transitive *with respect to a fixed class*. In other words, if all we cared about was explaining the differences between all of the groups of points with respect to a single reference group, say Group C, then IG would produce consistent explanations. However, explaining the features that separate Group A from Group B relative to Group C is not the problem we are trying to solve.

When we use Equation 11 to calculate $\delta_{IG}(i \to j)$ we found that the resulting explanations were not consistent. This does not violate the theory of IG because each $\delta_{IG}(i \to j)$ is calculated with reference Group j and so the assumption that we have a single reference group is not satisfied.

When we considered modifying Equation 11 to aggregate over the reference 'class/group' and potentially gain consistency that way, we either got uniform zero attributions (if we averaged over all reference groups) or inconsistent explanations (if we excluded any subset of $\{i, j\}$ from the averaging).

Conclusion. As suggested in Section 2, using existing explanation methods to find GCEs is going to be challenging because it is not what they are designed to do. We found that IG, a method that theoretically looked promising, was unable to be extended in a simple way to this setting.

A.2. Representation Function

Differentiability. TGT assumes that r is a differentiable function. Hidden in this assumption is the assumption that r is a function that we can evaluate on an arbitrary point. Although most methods for learning a low-dimensional representation satisfy this assumption, t-SNE does not. Fortunately there are parametric variations of t-SNE such as the one we used in our experiments (Ding et al., 2018). The assumption that r is differentiable can be relaxed by using a finite-difference optimization method, such as SPSA (Spall, 1998), at the expense of computational cost.

Learning Meaningful Structure. One assumption that every analysis (whether that is manual inspection, statistical testing, or interpretable ML) of the representation learned by r is that this function learned meaningful structure from the data. Because practitioners are already relying on these representations and, in some situations, have verified that they are meaningful, this concern is largely orthogonal to our work.

However, from an interpretable ML perspective, our goal is to explain r. So, if r identifies different structure when it is retrained or when it is trained with a different algorithm or structure, we expect TGT to produce different explanations since the embedding itself has changed.

Our experimental results show that the representation learned by (Ding et al., 2018) is stable to being retrained and to modifications to the dataset and that TGT produces stable explanations for these representations.

Identifying that Structure with Explanations. It is possible to have a model that learned the true structure of the data and to have an explanation that is technically true (as measured by some proxy metric for interpretability) about the model but that also fails to capture meaningful patterns. For example, adversarial examples (Szegedy et al., 2013) are technically local counterfactual explanations but they usually look like random noise and, as a result, do not tell a person much about the patterns the model has learned. TGT's design, which calculates the explanation between each pair of groups as if it were a compressed sensing problem but constrains those solutions to be symmetrical and transitive among all groups, was chosen as a prior to prevent this type of behavior.

A.3. Learned Representations

The learned representations and the corresponding groups of points for the datasets we studied are in Figures 10, 11, 12, 13, and 14.

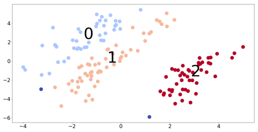

Figure 10: The learned representation and grouping for the UCI Iris dataset

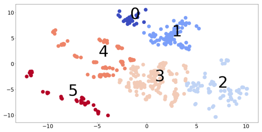

Figure 11: The learned representation and grouping for the UCI Boston Housing dataset

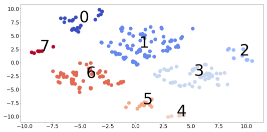

Figure 12: The learned representation and grouping for the UCI Heart Disease dataset

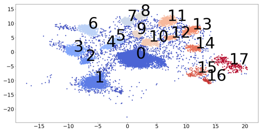

Figure 13: The learned representation and grouping for the single-cell RNA dataset

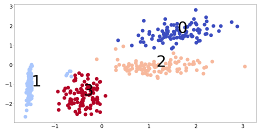

Figure 14: The learned representation and grouping for the synthetic dataset.

A.4. Pairwise Correctness and Coverage Plots

TGT is much better than DBM for finding 250-sparse explanations on the single-cell RNA dataset (Figures 15 and 16). On the synthetic dataset, TGT and DBM are equally effective explanations of the model (Figures 17 and 18). However, TGT only relied on the two causal variables while DBM included the correlated variable as well.

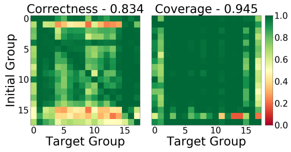

Figure 15: The pairwise metrics for TGT on the single-cell RNA dataset for 250-sparse explanations.

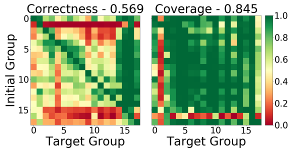

Figure 16: The pairwise metrics for DBM on the single-cell RNA dataset for 250-sparse explanations.

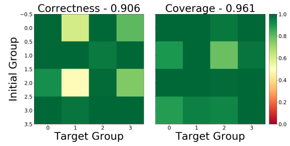

Figure 17: The pairwise metrics for TGT on the synthetic dataset with no sparsity constraint.

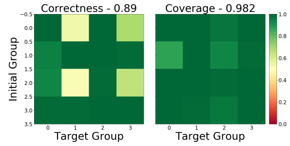

Figure 18: The pairwise metrics for DBM on the synthetic dataset with no sparsity constraint.

A.5. Qualitative Analysis of the UCI Datasets using the Labels

Although the representations we learned for these datasets were trained in an unsupervised manner, the groups that they find often have strong connections to the labels for the datasets; see Table 2, Figure 21, and Table 3. By using the connection between the groups and labels, we will be able to qualitatively assess whether or not TGT is finding real patterns in the data.

Iris Dataset. Looking at Table 2, we can see that the groups in this representation match very closely with the class labels. As a result, we would like to know whether or not the explanations TGT finds are consistent with a model trained directly to predict the labels. For this comparison, we used a simple decision tree, which is shown in Figure 19. Looking at TGT's explanations (Figure 20), we can see that they largely agree with the decision tree since both primarily use Petal Width to separate the classes/groups.

Table 2: The distribution of the labels per group for the UCI Iris dataset (classification).

Group \ Class	Iris Setosa	Iris Versicolour	Iris Virginica
0	0	5	38
1	0	44	12
2	48	0	0

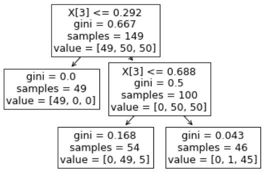

Figure 19: A small decision tree trained on this dataset. Notice that it relies on the Petal Width feature.

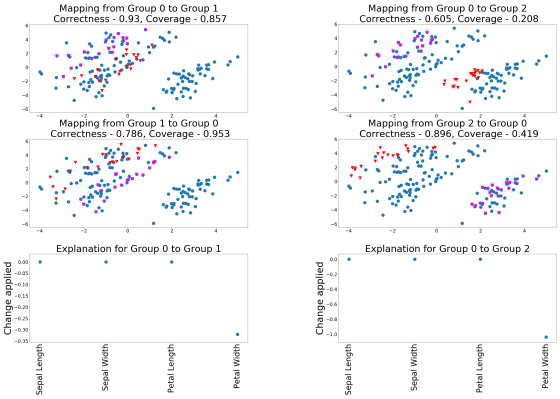

Figure 20: TGT's 1-sparse explanation of the difference between Group 0 and Group 1 (left) and Group 0 and Group 2 (right). Similar to the decision tree, they rely on the Petal Width feature.

Boston Housing Dataset. Looking at Figure 21, we can see that two comparisons between the groups stand out: Group 0 to Group 2, which shows a significant increase in the price, and Group 3 to Group 5, which has relatively little effect on the price. As a result, we would like to determine what the differences between these groups of houses are that influence their price.

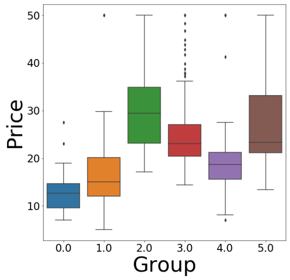

Figure 21: The distribution of the labels per group for the UCI Boston Housing dataset (regression).

Looking at Figure 22, we see that TGT found that the key differences between Group 0 and Group 2 appears to be the difference between a house being in an urban area vs being in a suburb: the proportion of land zoned for large residential lots and B[10] both increase while access to radial highways and tax rates both decrease. It also found that the key differences between Group 3 and Group 5 are, first, moving the house onto the Charles river and, second, decreasing B.

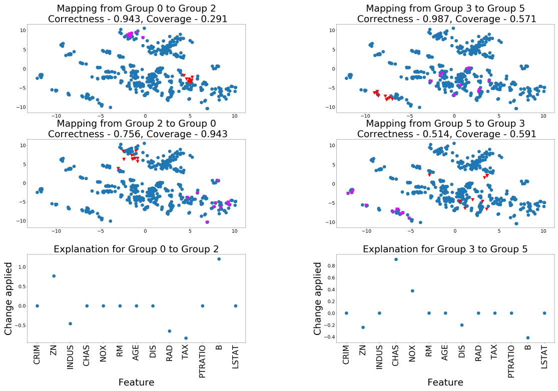

Figure 22: TGT's 5-sparse explanation for Group 0 to Group 2 (Left) and Group 3 to Group 5 (right).

Heart Disease Dataset. Looking at Figure 3, we see that there are three large groups that stand out: Group 1, which has a balanced risk of heart disease, Group 3, which has a relatively low risk, and Group 6, which has a relatively high risk. As a result, we would like to determine what the differences between these groups of subjects are that influence their risk of heart disease.

Table 3: The distribution of the labels per group for the UCI Heart Disease dataset (classification).

Group \ Class	No Heart Disease	Heart Disease
0	6	15
1	46	62
2	10	1
3	52	14
4	4	1
5	10	7
6	8	59
7	2	5

Looking at Figure 23, TGT found that the key differences between Group 1 and Group 3 are a moderate decrease in chest pain along with having exercised induced angina; these are subjects whose symptoms are explained by exercise induced angina rather than heart disease. It also found that the key difference between Group 1 and Group 6 is that Group 1 is made up of men while Group 6 is made up of women; this is consistent with the fact that heart disease is the leading cause of mortality in women (Bello & Mosca, 2004).

[10] This is a unusually defined feature that is related to the racial demographics of a town. Determining what it means to change this feature depends on a measurement that is not in the dataset.

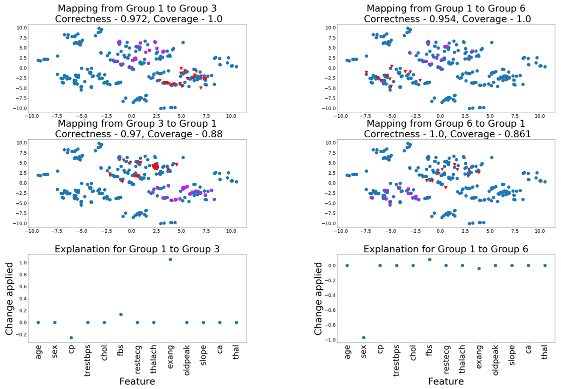

Figure 23: TGT's 3-sparse explanations for the difference between Group 1 and Group 3 (Left) and Group 1 and Group 6 (right).

A.6. Quantitative Analysis of Modified Versions of the UCI Datasets.

Because the UCI datasets are not synthetic datasets, we do not know the underlying process that generated the data and, as a result, it is difficult to quantitatively determine whether or not an explanation is "correct" in the way that we could with the synthetic dataset. Consequently, we performed a series of experiments on modified versions of the original datasets in order to answer two important questions:

- Does TGT correctly identify the modifications we made to the original dataset?
- Do TGT's explanations between the original groups change when the modified group is added to the dataset?

We found that TGT does identify the modifications we made and that, in doing so, it does not significantly change the explanations between the original groups. Importantly, this result remains true even if we retrain the learned representation on the modified dataset. These results are a strong indicator that TGT is finding real patterns in the data.

How do we modify the datasets? We create a modified version of the original dataset by: picking one of the groups of points in the original dataset, modifying that group of points in some way, and adding that new modified group of points to the original dataset. We will call the original dataset D and the modified dataset D', where $D' = D \cup G'$ and G' is the modified version of some group of points G. The critical choice to make during this process is to determine what modification to apply to G to get G'. We chose to add random noise to some of the features of the points in G and used the following two criteria when defining this modification for a particular dataset:

- G' should be approximately within the range/distribution of D.
- $r(G')$ should form it's own (approximately) distinct group. Intuitively, if $r(G')$ does not form its own group, then r thinks G' is similar to some other group in the dataset and, as a result, we would not expect TGT to be able to explain the differences between G' and that group of points.

The modifications we used are in Table 4.

Table 4: For each dataset, we chose a group of points to modify and modified it by applying these perturbations to the specified features.

Dataset	Group Modified	Feature	Perturbation Applied
Iris	0	Sepal Width	-0.4 + Uniform(-0.1, 0.1)
Housing	1	ZN	0.9 + Uniform(-0.1, 0.1)
		TAX	-0.5 + Uniform(-0.1, 0.1)
Heart	1	restecg	-0.9 + Uniform(-0.1, 0.1)
		exang	0.6 + Uniform(-0.1, 0.1)

Experimental Setup: We now have two versions of each dataset: D and D'. We also have the original learned representation r, which was trained on D, and a new learned representation r', which is trained on D'. As a result, we have three sets of explanations:

- **Original:** These explain r when applied to D
- **Modified:** These explain r when applied to D'
- **Retrained:** These explain r' when applied to D'

The visualization of the representation for the first setting is in the Appendix A.3 and the later two these settings is in Figures 24, 25, and 26. Note that applying r to D' looks the same as applying r to D except for the fact that there is an additional group from adding G' to D and that applying r' to D' often shows that r' has learned to separate G' from the other groups better than r did.

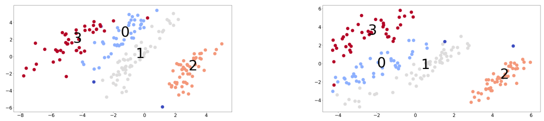

Figure 24: The learned representation for: r applied to D' (Left) and r' applied to D' (Right) for the Iris dataset

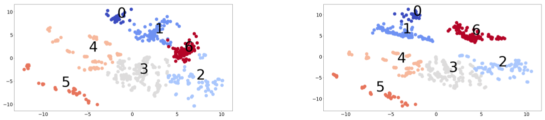

Figure 25: The learned representation for: r applied to D' (Left) and r' applied to D' (Right) for the Boston Housing dataset

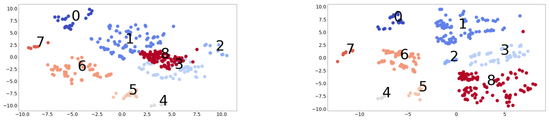

Figure 26: The learned representation for: r applied to D' (Left) and r' applied to D' (Right) for the Heart Disease dataset

Does TGT correctly identify the modifications we made to the original dataset? In Figure Figures 27, 28, and 29, we can see the explanations TGT found for the difference between G and G' for each of the datasets. If we compare the explanations to the modifications from Table 4, we can see that they identified which features we changed and, approximately, by how much. The error in the estimation of "by how much" is due to the l_1 regularization used to find a simple explanation.

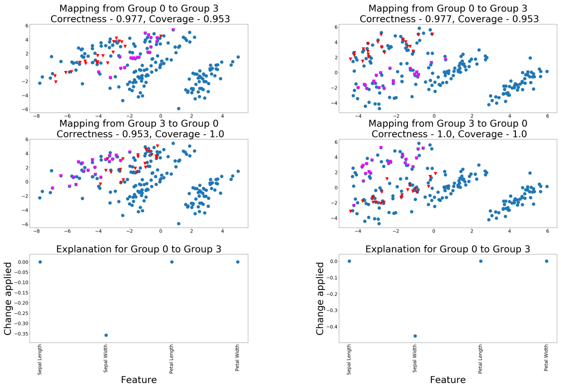

Figure 27: The Modified (Left) and Retrained (Right) explanation's explanation for the difference between G and G' on the Iris dataset.

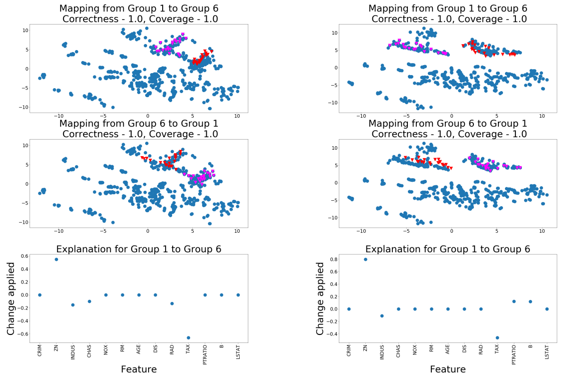

Figure 28: The Modified (Left) and Retrained (Right) explanation's explanation for the difference between G and G' on the Boston Housing dataset.

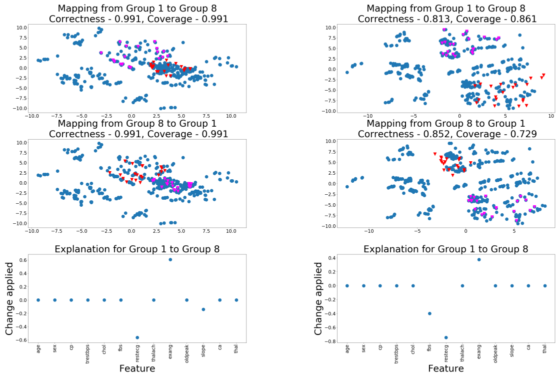

Figure 29: The Modified (Left) and Retrained (Right) explanation's explanation for the difference between G and G' on the Heart Disease dataset.

Do TGT's explanations between the original groups change when the modified group is added to the dataset? In Figures 30, 31, and 32, we can see a comparison of the explanations for the differences between the original groups for the Original vs the Modified and the Original vs the Retrained explanations. Adding G' to D did not cause TGT to find significantly different explanations between the groups in D. Explaining r' resulted in explanations that were generally similar, but adding another layer of variability (*i.e.,* training r') did add some noise.

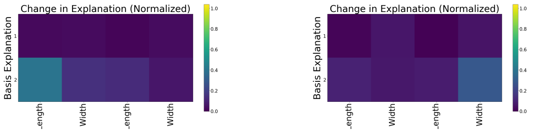

Figure 30: The absolute difference between the Modified (Left)/Retrained (Right) explanations and the Original explanations scaled relative to the Original explanations on the Iris dataset.

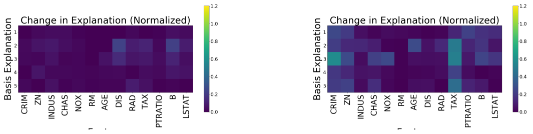

Figure 31: The absolute difference between the Modified (Left)/Retrained (Right) explanations and the Original explanations scaled relative to the Original explanations on the Boston Housing dataset.

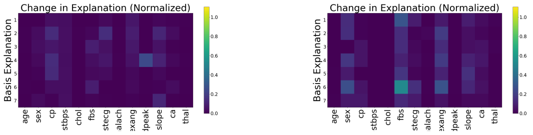

Figure 32: The absolute difference between the Modified (Left)/Retrained (Right) explanations and the Original explanations scaled relative to the Original explanations on the Heart Disease dataset.