We thank the reviewers for their effort. We are pleased that everyone appreciated the novelty of the idea and significance of the problem setting of building locally consistent transparent models. We now address your specific concerns.

R1 and **R2** *Long rules and interpretability*: Firstly, it is not true that the clauses would involve all the features, as mentioned in the paper, our Boolean clauses are 2k sparse given k sparse local explanations i.e. PPs and PNs. So the sparsity of our clauses depends on the sparsity of the explanations which a contrastive method provides, where our contribution does not add any additional complexity. For example, this is seen on the Sky Survey dataset in the paper where the dataset has 17 features but the rules contain only 9 features. Secondly, the issue of many conjunctions and redundant features etc could arise even with known greedy algorithms for learning such structures. In fact, Rudin (citation [23] in the paper) argues that even if conjunctive rules are pages long, a person looking for an explanation can relatively easily "locate by inspection" a sufficiently sparse explanation by navigating the formula. This is cited as the reason for general desirability of rule/tree based classifiers that may not be succinct.

R1, R2, R3 and **R4** *Scalability*: If the dataset is high dimensional, as mentioned above, our method being simply a function of the PPs and PNs which are typically sparse we should be able to scale. If the number of points is large one can do random sampling or using prototype selection methods to control size of F. More importantly though, this size is not a bottleneck since we suggest training simple models on it such as a small decision tree or L1-logistic regression which scale well even with many features and would end up choosing only a few of these rules. The most time consuming part is really obtaining PNs from a contrastive method as it is a non-convex optimization problem. Although, this too could be parrallelized across different (batches of) data points.

R2 *Correctness of description of Eq 1*: The description is correct since, δ_j always lies between x_j and b_j no matter the sign. So δ_j can be $< -|x_j|$ only if $|b_j| > |x_j|$ and, b_j and x_j have opposite signs i.e. one is positive and the other is negative. Here again δ_j would be closer to b_j than x_j is to it.

R2 *Clarity of binning*: Please see Figure 1 in supplement for a toy example depicting the whole process. We will move this into the main paper in the final version, since NeurIPS typically allows an extra page.

R2 *Reporting fidelity*: Yes, we will add a row in Table 1 showcasing fidelity of the transparent model to the black-box. Although, the last 5 rows in Table 1 which show Test accuracy hint towards what the fidelities might be.

R3 *Regarding the form of the bounds and how they are set*: The bounds L_j and U_j in Eq.(1) and Eq.(2) are set manually and its basically a bound on the domain we know. For example, features like "Age" cannot be negative and will have a lower bound of 0, while for gray scale images the values would be 0 and 1 for all pixels. There is no guarantee that one can find a PN vector for every data point using these contrastive methods. However, that is independent of our current contribution and our method can create rules even if we have no PNs. This can be confirmed by looking at the Magic and Diabetes datasets in Table 1 where we can still create rules eventhough no PNs were found for those datasets.

R3 *Why not Knowledge distillation*: Please look at Table 1 where we perform favorably to knowledge distillation.

R3 *Regarding Eq 3*: By the definition of PN, we are aiming to generate a point $n(x)$ that is of a different class from the original input x and thus equation is correct.

R3 *Only one PP used in Eq. 5*: Contrastive methods (citations [8] and [31] in the paper) output only a single PP and at most one PN explanation for each data point. We use this PP and PN in Eq 5 for discretization.

R3 *Regarding circular argument*: The local consistency metric just ensures that the predictions as well as the local explanations (PPs and PNs) for the black-box are consistent with the transparent model. This is independent of how one might obtain a locally consistent model. When trying to build a such a model one is free to use all the information that is available to them. So we do not believe there is any circular argument. In fact, the Augmentation baseline we compare with in Table 1 is more closely related to just directly using PPs and PNs, and we outperform it by a significant margin.

R3 *Dimension of PPs and PNs in Eq (5)*: $p(x_j)$ means the j-th coordinate of the PP vector corresponding to the point x. So there is no discrepancy between this and Listing 1. We will clarify.

R4 *Selecting between PNs*: Data is typically normalized where PPs and PNs are generated by another method and given as input to ours. So though you have a valid point that is not the focus of this submission, rather we assume whichever local contrastive method we use has already done that for us and provided a valid local explanation. Realistic explanations in previous works were generated using autoencoders etc. (see citation [8] in the paper).

R4 *Assumption that x and the base vector share the same class*: This is not true. There is only one base vector that is set for the entire dataset. This may be a zero vector or one where the base value for each feature is set based on domain knowledge (see citation [9] in the paper). Given this, the limitations you mentioned do not apply.

R4 *More sensible to look at $C(x, y)$-$C(\delta, y)$*: Interesting suggestion! Although, the issue you mention can be addressed (indirectly) in Eq. 1 by having a large enough κ, which controls the margin between y and the next most likely class.