We want to thank all reviewers for their time and effort. We address the raised points below.

- **R1** *hard to learn large number of densities*: We used a threshold on the number of samples at each leaf, to decide whether LearnSPN or a fully factorised density is used. This drastically reduces the computational effort to learn the leaves. In general, GeDTs/GeFs allow for much freedom, as any density estimator can be used at the leaves.
- **R1** *additional computational overhead*: The asymptotic costs of learning and inference are comparable to standard RFs (see supp. material). In general, the overhead depends on the particular choice of leaf estimators.
- **R1** *when to use SPNs or fully factorised leaves*: A key consideration is overfitting, and using factorised leaves induces a conditional (on the DT decisions) independence assumption which is more robust for few samples.
- **R1** *prediction goes to 0 as X 2 approaches E(X2/X1) instead of 1*: This is indeed a typo: Prediction will be (w.h.p.) $Y = 0$ (the argumentation remains valid as presented in the paper). Thanks for the catch!
- **R1** *Fig.2 is slightly puzzling*: We see the point, and will revise the figure.
- **R3** *compare against the most related approaches (e.g., DT+KDE, DT+NBC, CNets)*: Note that GeDTs/GeFs leaves can be equipped with any density model. Therefore, when we equip them with KDEs or NBCs, they simply subsume DT+KDE and DT+NBC in terms of classification. The crucial difference in our paper is, that these previous approaches did not follow through with the interpretation of the **whole** DT as a joint distribution and its advantages (missing data, outlier detection). We will include a comparison to CNets.
- **R5** *practical interest beyond handling missing values, properties of GeDT/GeRF*: We focused on classification under missing data since it is the most compelling advantage: DT/RF practitioners do not need to change the structure learning algorithm, but can now treat missing inputs within the same model, and with guaranteed backwards compatibility (provided $p(Y, \mathbf{X}) = p(Y)p(\mathbf{X})$ in all leaves). We will include more thorough experiments on outlier detection in the revised version. Our main contribution is that we link two separate communities, PCs and DTs, which will likely lead to ample cross-fertilisation of ideas.
- **R5** *RF hyper-parameters*: Our experiments only compare models with the same underlying structure, so variations in the RF hyper-parameters would most likely only change the basis for comparison but not the overall results. As our goal is not to find the most accurate RF/GeF, we use literature defaults and defer tuning to future work (non-tuned results are already very good). That said, we did experiment with a few hyper-parameters. For instance, pruning has not proved very useful (not uncommon in RFs), see right side of Figure below.
- **R5** *LearnSPN on the full dataset*: We experimented with a SOTA implementation of LearnSPN and results are quite poor. For the sake of space we omitted it in the submission (see left side of Figure below), but will include it given acceptance.
- **R5** *explanations on density model part*: GeDTs/GeFs can be used with any density estimator, which is subject to exploration in the future (we use SPNs as cited). Also, we will explain fully factorised leaves in the paper. $P(X_i)$ is a marginal (multinomial/normal) trained on data matching the leaf.
- **R5** *Friedman method*: We will clarify differences in the paper; results do not automatically apply, but it may be possible to extend them. Note that Friedman's method ignores the actual values of explanatory variables, while fully factorised leaves do compute $p(\mathbf{x})$ to handle missing values.
- **R5** *MissForest, MERCS, Isolation Forest*: We agree that claiming GeFs are the new SOTA for treating missing values would require further experimentation, but we do not make such a claim. In our experiments, we focus on 'built-in' methods as they are closer to ours in nature, and also compared to KNN because it is the most common imputation method and the closest to ours in computational cost. We will cite and discuss the works pointed out.

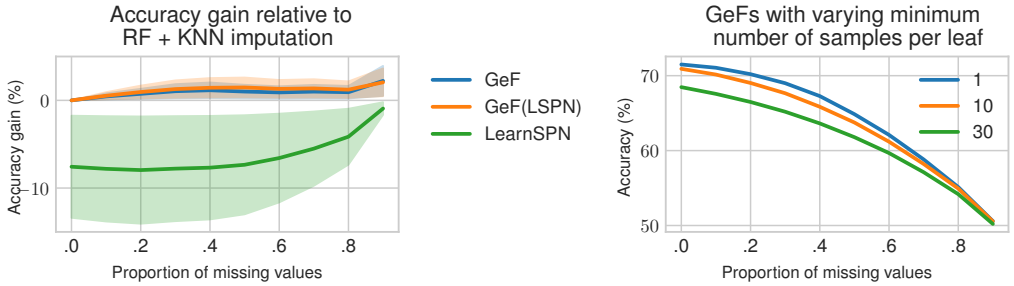

Figure 1: Average accuracy gain relative to RF+KNN imputation (left) and average accuracy on test data of GeF models with different minimum number of samples per leaf (right). Both averages are across the datasets considered.