4.1.6 Significance Tests

# Original

To compute whether the difference in reconstruction ability across datasets for each method was statistically significant, we used Friedman tests paired with Nemenyi post-hoc analysis as suggested in [12] as general performance is the main focus. Significance tests on each dataset were not used for the reasons described in [29], i.e. violations in the conclusions due to the increased probability of type-I errors, as well as the Friedman test making fewer assumptions.

To show the average reconstruction ability for each method, for each dataset, we present the average testing accuracy of a 10-fold cross-validation procedure, where each method gets the same train:test sets. The averages are also across the 3 black-box methods (so for each method, 30 runs are executed for each dataset, with the same random samples used across methods). The goal is for the extraction methods to be invariant to the black-box model used, hence the averaging.

# Condensed

To compute whether the difference in reconstruction ability across datasets for each method was statistically significant, we used Friedman tests paired with Nemenyi post-hoc analysis.

To show the average reconstruction ability for each method we present the average accuracy of a 10-fold cross-validation, where each method gets the same train-test sets. The averages are also across the 3 black-box methods therefore for each method, 30 runs are executed for each of the 20 datasets. The goal is for the extraction methods to be invariant to the black-box model used, hence the averaging.