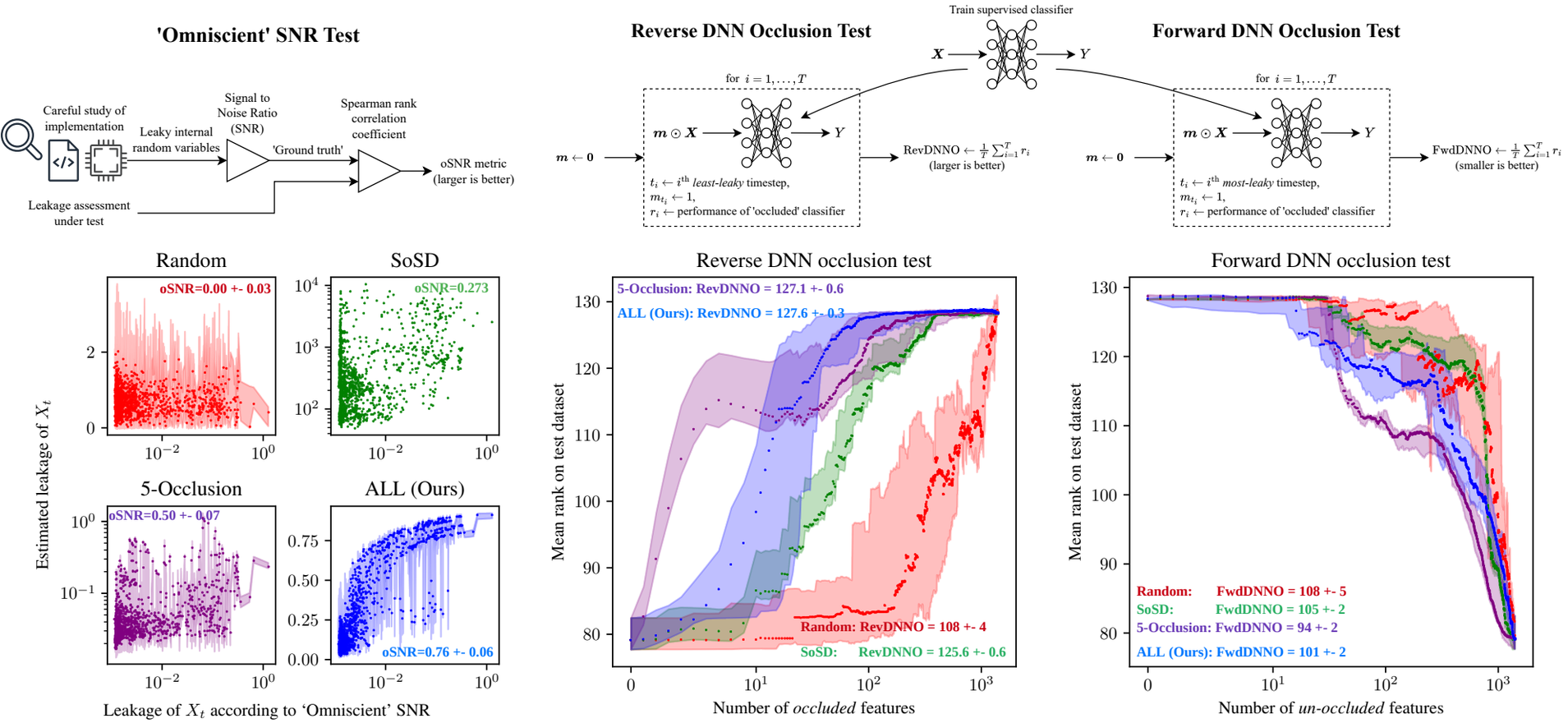


Figure 1: Overview of the performance evaluation metrics used in this paper. For illustration purposes, we include plots recorded on ASCADv1 (random) for ALL (ours), 5-occlusion (the *best-performing* neural net attribution method), SoSD (the *best-performing* first-order parametric method), and random guessing. (**left column**) The 'Omniscient' Signal to Noise Ratio (oSNR) metric. We compute 'ground truth'-like per-timestep leakiness measurements using the ASCADv1 implementation analysis of Egger et al. (2022), as well as knowledge of internal random variables which the leakage localization algorithms lack access to. Below we show plots with the 'ground truth' measurements on the horizontal axis and the measurements under test on the vertical axis; *closer to strictly-increasing is better*. The oSNR metric is given by the Pearson rank correlation coefficient between the 'ground truth' measurements and the measurements under test; *higher is better*. (**center column**) The reverse DNN occlusion (RevDNNO) test. To compute this metric, we first train a supervised DNN classifier to map emission traces to the sensitive variable. We then occlude all its inputs and incrementally un-occlude then from *least- to most-leaky* as estimated by the leakiness assessment under test, and at each step compute its performance (quantified by rank, lower is better) on the test dataset. The Rev-DNNO metric is given by the average value of these performance assessments (*higher is better*, because it indicates that claimed nonleaky features indeed have little utility to the classifier). If the two DNN occlusion metrics, this is more sensitive to *true/false negative* leakiness measurements. (**right column**) The forward DNN occlusion (FwdDNNO) test. This is identical to the RevDNNO test, except that we un-occlude inputs in the opposite order: from *most- to least-leaky*. The Fwd-DNNO metric is given by the average value of these assessments (*lower is better*, because it indicates that the claimed leaky features indeed have high utility to the classifier). This is more sensitive to *true/false positive* leakiness measurements.