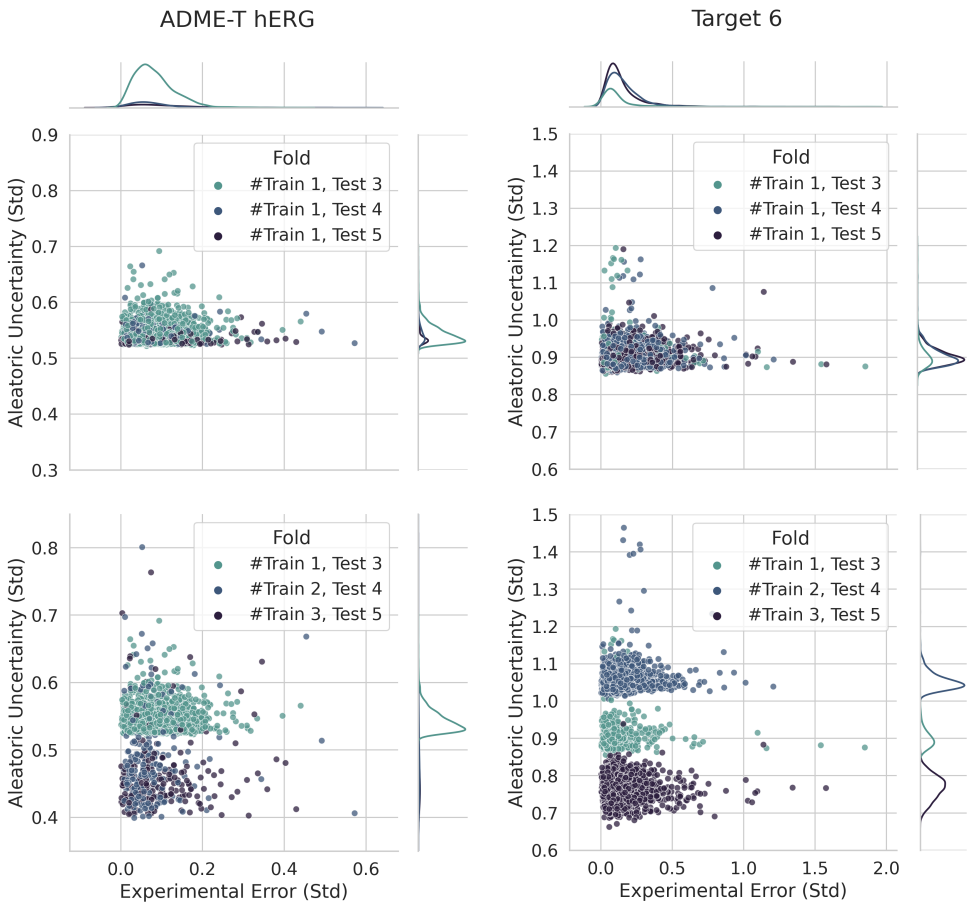

Figure 8: **Aleatoric Estimates by Gaussian Ensemble.** A practical illustration of the aleatoric uncertainty estimates by the Gaussian Ensemble on the two assays with the most available duplicate experiments in the test sets. Comparing the predicted aleatoric uncertainty to the standard deviation of duplicated experiments, i.e. experimental error. The top row shows prediction by only the model trained on the first fold. The bottom row shows each test set predicted by the respective, different models trained on all folds until the given test set.

optimized individually as explained in the model selection in Appendix B. Therefore, these models can have vastly different model architectures and be trained using different training procedures such as learning rate.

Crucially, none of these factors should have any effect on the aleatoric uncertainty as it is often categorized in literature as irreducible [2, 5, 6]. Despite this, we do see distinct differences between most of the distributions of aleatoric estimates by each of these different models in the bottom row of Fig. 8. Specifically, we see some instances where increasing amounts of training data result in generally lower aleatoric estimates. However, there is also one case where the opposite is true, between test folds 3 and 4 of the Target 6 assay. We can only assume this relation to be because of the differences in model architectures. Therefore, our empirical results raise questions about whether the explored models' supposed estimates of aleatoric uncertainty are really fully disentangled from the epistemic uncertainty. Similarly, recent theoretical work on deriving suitable measures for aleatoric, epistemic, and total uncertainty has found that the aleatoric and epistemic parts do not necessarily have to add up to the total uncertainty [59]. As such, the disentanglement of the sources of uncertainty should be considered an ongoing field of research that needs more work to fully determine how these estimates should be categorized and how they relate to the underlying noise in the data.

4 Conclusions

The low-data challenge in drug discovery is typically accompanied by additional, often overlooked, partial information from censored labels. Despite their potential value, censored labels have not yet been fully utilized due to the lack of