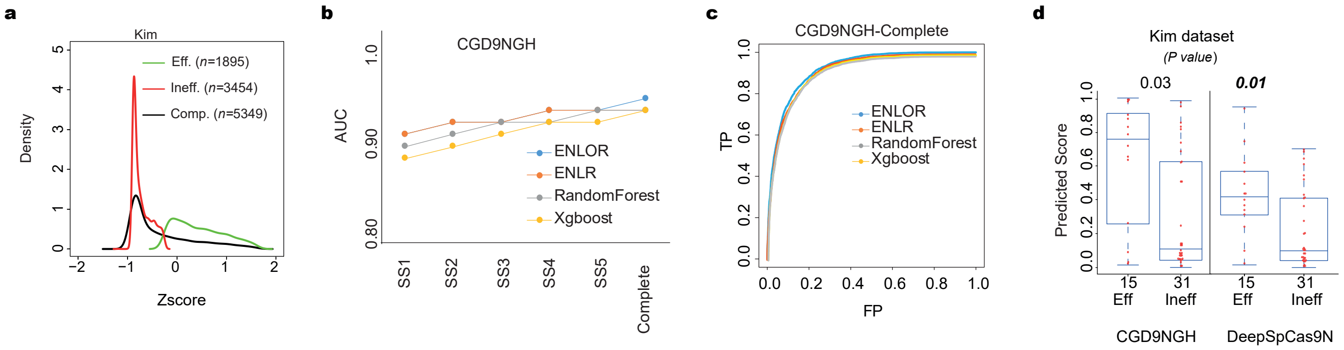
Supplementary Figure 5



Supplementary Figure 5: Model development for a CRISPR-Cas9 dataset generated using non-canonical PAMs. a. Representation of the gRNA distribution in the Cas9 dataset generated with non-canonical PAMs (NGH, where H = A,C,T). Green curve represents the number of efficient gRNAs in the dataset, red curve represents the number of inefficient gRNAs in the dataset, and black curve represents the total number of gRNAs in the dataset. **b**. The plot depicts AUC values for each algorithm applied to the CRISPR Cas9NGH system. AUC values for subsets as well as complete training data are shown. The *x-axis* denotes subset and complete training data, and the *y-axis* denotes AUC values. **c**. ROC curves depict the algorithms' performance on the complete training data. The *x-axis* denotes the average False Positive rate (FP) and the *y-axis* denotes the average True Positive rate (TP). **d**. Comparison of the performance of the CGD model with that of an existing deep learning algorithm (DeepSpCas9N) applied to independent test data using the KS test. The *x-axis* indicates the number of efficient (Eff) and inefficient (Ineff) gRNAs. The DeepSpCas9N values are rescaled between 0 to 1 on the *y-axis*.