



Figure 4-figure supplement 2. Model accuracy as a function of the size of the training set. A) Plotting model accuracy as a function of the training set's coverage of the total possible sequences for the indicated number of random nucleotides. Each point represents the prediction from a neural network model, trained on a set of sequences with a specific number of random nucleotides, of the activity of a different test set of sequences from the same library. The model's accuracy is defined as the R^2 value when comparing the model's predicted gene-regulatory activities for the test set and that set's measured activities. The coverage for that point is determined by dividing the number of sequences in the training set by the total possible number of sequences for that number of random nucleotides. The neural network model was tested on varying levels of coverage for 3 different libraries, with 8, 9, and 11 random nucleotides. B) Plotting model accuracy as a function of the number of sequences in the training set. The same set of trained models as in A, but plotted so that the X axis is the total number of sequences in the training set for a given model.