Fig. S1

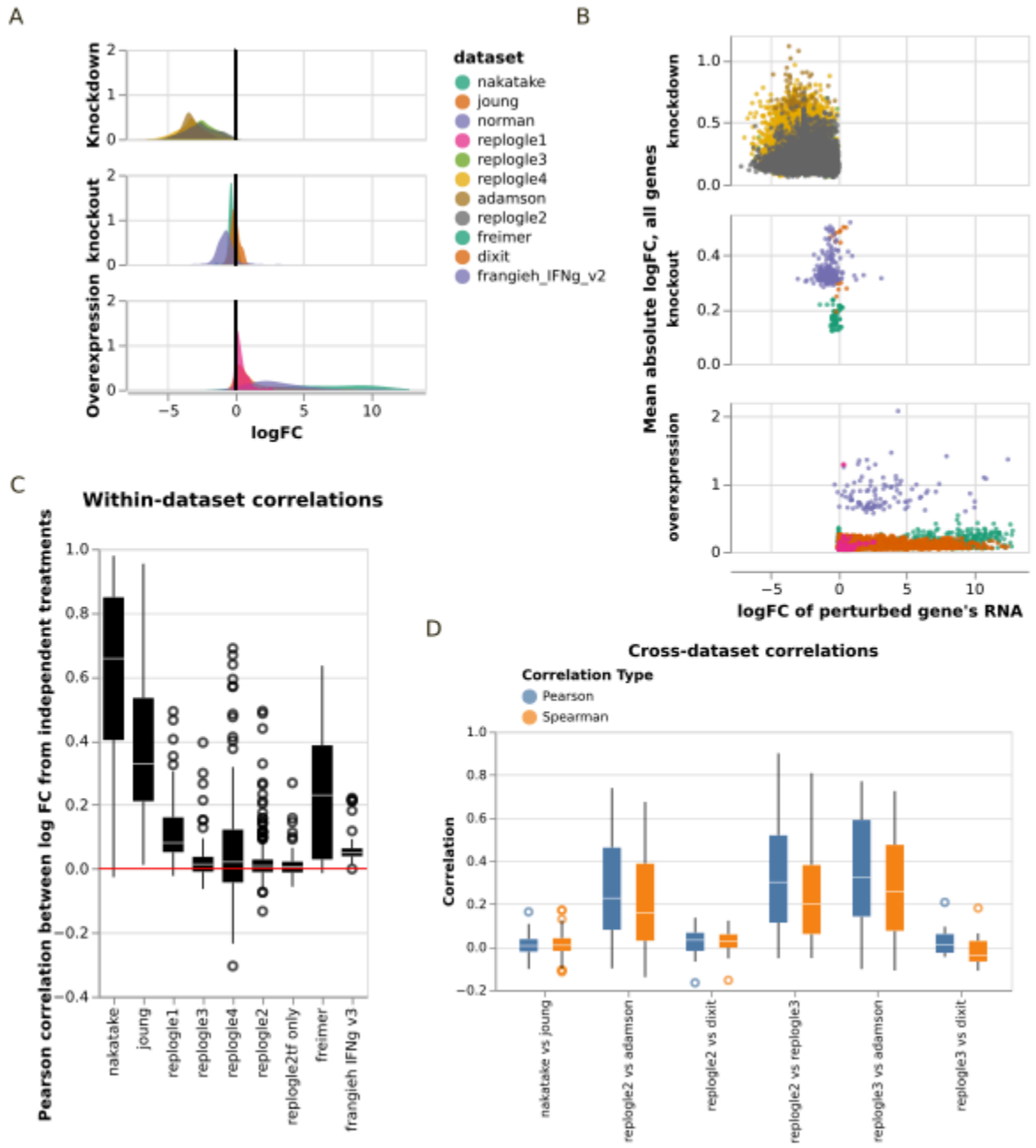

Fig. S1. Related to Fig. 1. Quality analysis of large perturbation datasets. All values shown in this figure are computed after filtering, aggregation, and normalization as described in the Methods.

A. Effect on total-count-normalized transcript levels of the perturbed gene (X-axis).

B. Effect on the perturbed gene (X-axis) and the whole transcriptome (Y-axis). Log fold change (logFC) over controls is estimated using a pseudocount of 1. For norman, dixit, and adamson, only 6,000 highly variable genes are included.
C. Pearson correlation between log fold change vectors for pairs of replicates (nakatake, joung), donors (freimer), or distinct guide RNA's (other datasets). Correlations are computed across all genes available in the dataset (**Methods**).

Pearson and Spearman correlation between log fold change vectors for each perturbation across pairs of datasets. Correlations are computed across all genes present in both datasets (**Methods**).

Fig. S2

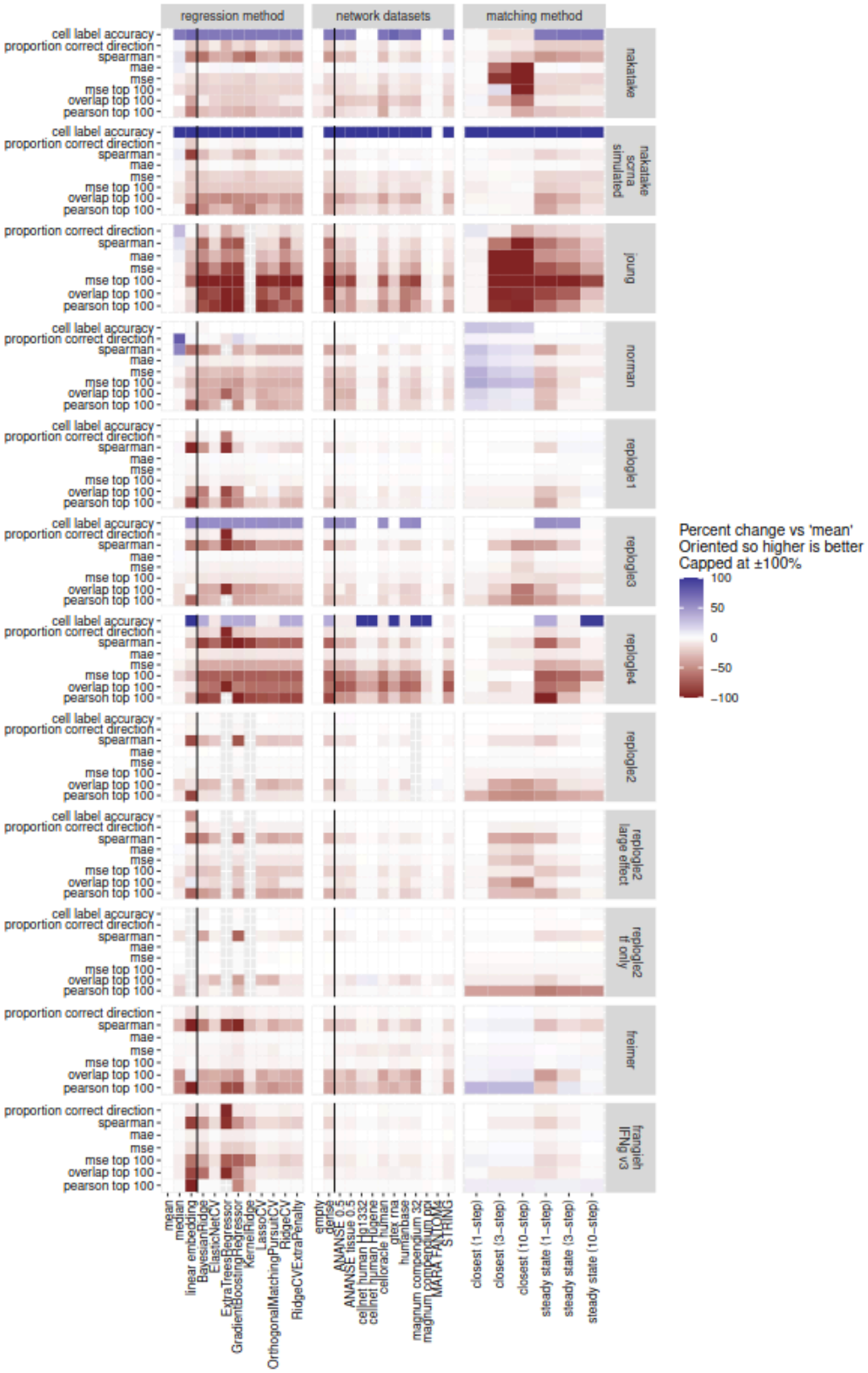
regression method
network datasets
matching method
cell label accuracy
proportion correct direction
spearman
mae
mse
mse top 100
overlap top 100
pearson top 100
nakatake
nakatake scrna simulated
joung
norman
replogle1
replogle3
replogle4
replogle2
replogle2 large effect
replogle2 tf only
freimer
frangieh IFNg v3
mean
median
linear embedding
BayesianRidge
ElasticNetCV
ExtraTreesRegressor
GradientBoostingRegressor
KernelRidge
LassoCV
OrthogonalMatchingPursuitCV
RidgeCV
RidgeCVExtraPenalty
empty
dense
ANANSE 0.5
ANANSE tissue 0.5
celnet human Hg1332
celnet human Hgeme
celloracle human
gtex rna
humantissue
magnum compendium 32
magnum compendium gtex
MARA FANTOM4
STRING
closest (1−step)
closest (3−step)
closest (10−step)
steady state (1−step)
steady state (3−step)
steady state (10−step)
Percent change vs 'mean'
Oriented so higher is better
Capped at ±100%
100
50
0
−50
−100

Fig. S2. Related to Fig. 2. Various performance metrics to assess prediction of held-out perturbations. All experiments within each dataset use the same data split and can be directly compared. The y axis shows different performance metrics as described in Table 3. The x axis labels describe which regression method was used (left), which network was used for feature selection (middle), or how models handle time (right). Black lines separate baselines within each panel. In the right column, “closest” indicates that features used to predict each perturbed profile are drawn from the closest control profile, except for the perturbed gene itself. “Steady state” indicates that features used to predict each perturbed profile are drawn from that same profile. ExtraTrees and KernelRidge are missing from certain panels due to memory requirements exceeding 64GB. The Ahlmann-Eltze baseline is missing from the replogle2 tf only panel because the entire training set consisted of genes that were perturbed but not measured in the original source data.

Fig. S3

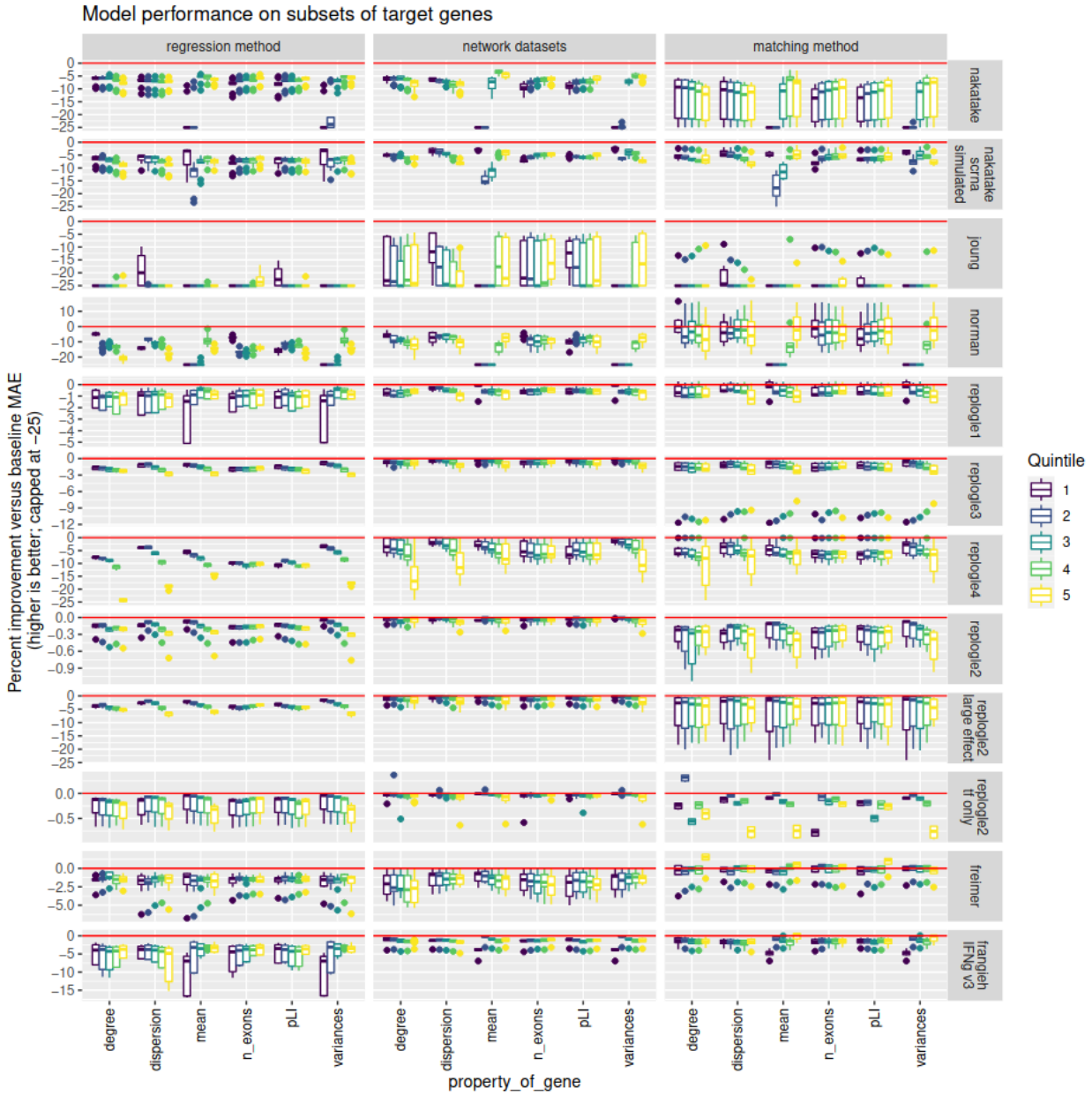

Fig. S3. Related to Fig. 2. Grouping target genes into quintiles using gene-level features reveals worse-than-baseline mae for most gene sets. The y axis shows the mae contribution of each gene set, expressed as percent change from the best baseline (inverted so that higher is better). The x axis indicates which feature we stratified genes on, and within each feature, the color indicates the quintile.

Fig. S4

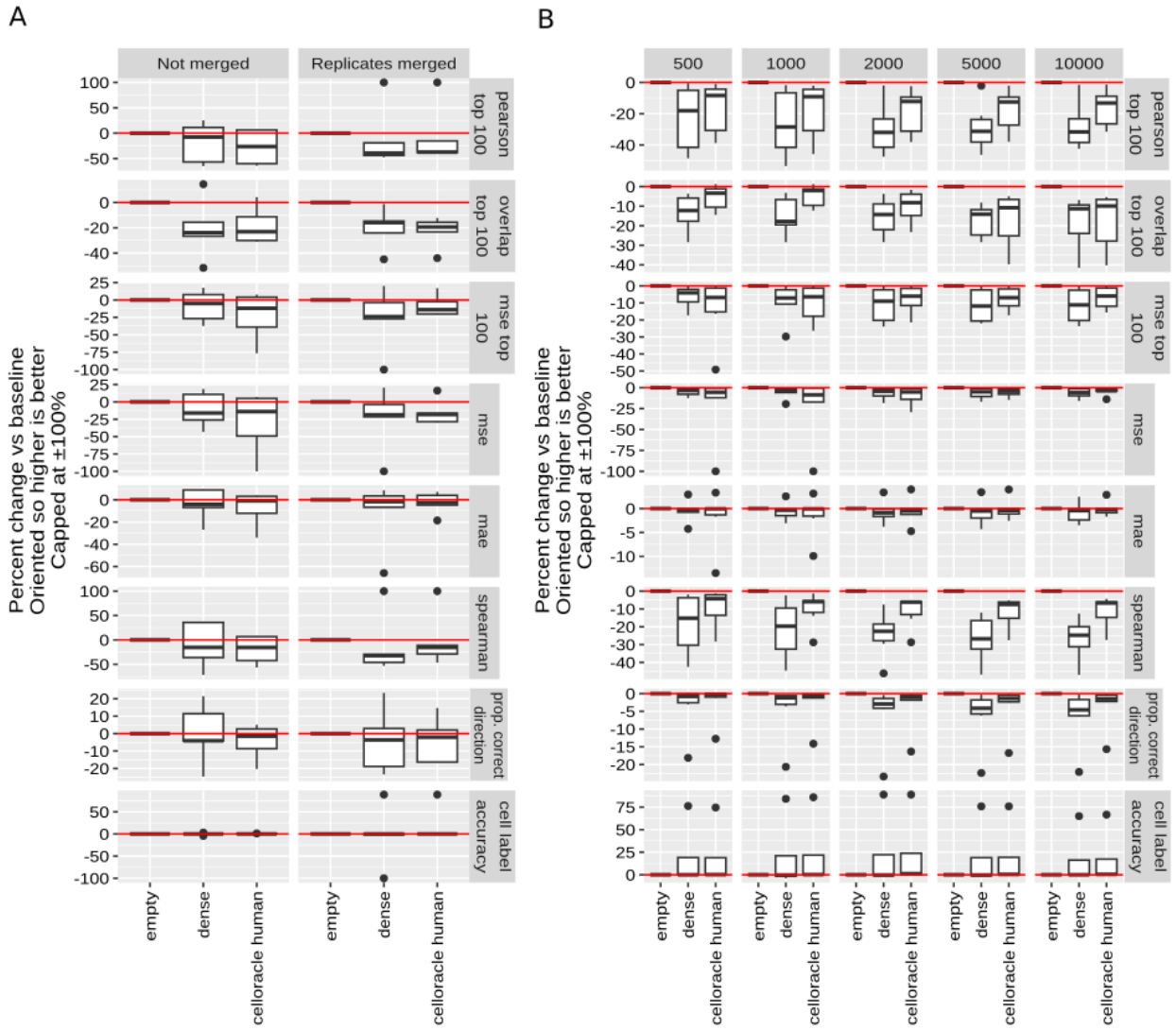

Fig. S4. Related to Fig. 2. Robustness checks.

A. Robustness to preprocessing. The x axis shows GRNs used. Each point represents performance of a single method on a single dataset according to a single metric. The y axis shows performance as percent improvement relative to the empty-network baseline, oriented so that higher numbers are better. The right margin shows different evaluation metrics and the top margin shows whether or not replicates or cells were averaged within each perturbation condition.

B. Robustness to the number of highly variable genes selected. The x axis shows GRNs used. Each point represents performance of a single method on a single dataset according to a single metric. The y axis shows performance as percent improvement relative to the empty-network baseline, oriented so that higher numbers are better. The right margin shows different evaluation metrics and the top margin shows how many variable genes were included in the analysis. The actual number of genes may vary from the number shown because all perturbed genes are always included.

Fig. S5

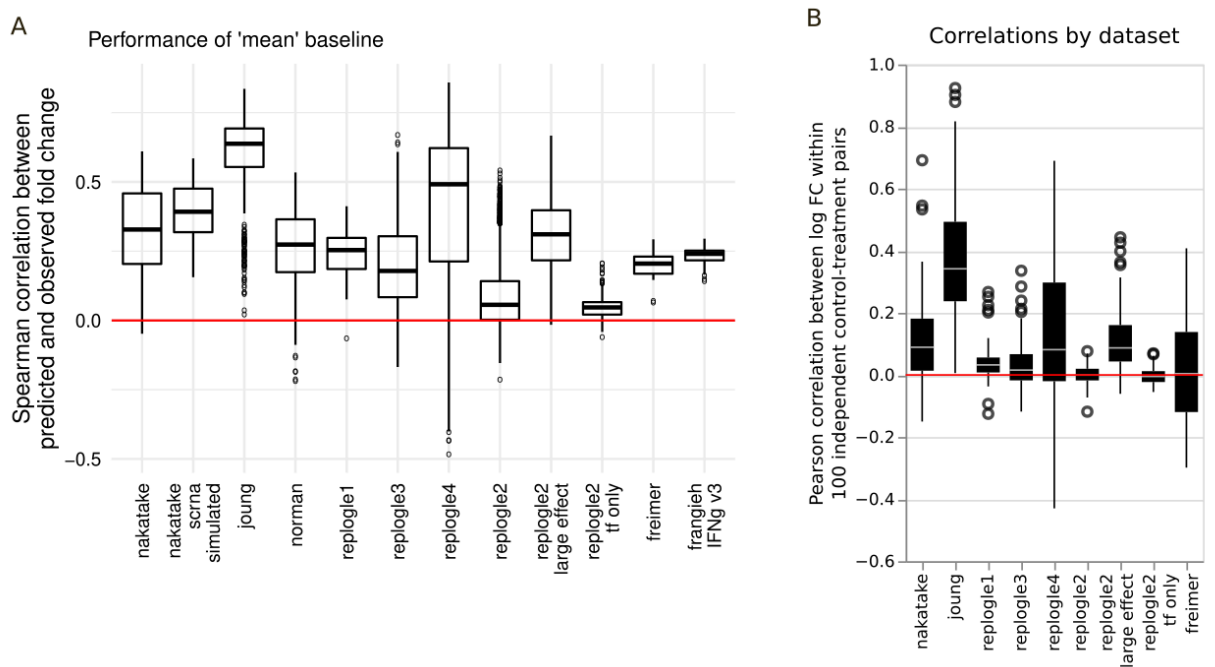

Fig. S5. Investigating strong performance of the “mean” baseline.

A. Spearman correlation between observed fold change over controls and fold change over controls predicted by the “mean” baseline (y axis) across various datasets (x axis).
B. Pearson correlation between observed fold change of treatment A over control A and treatment B over control B, where treatments A and B are different perturbations and controls A and B are separate replicates. Display includes 100 randomly selected pairs per dataset.

Fig. S6

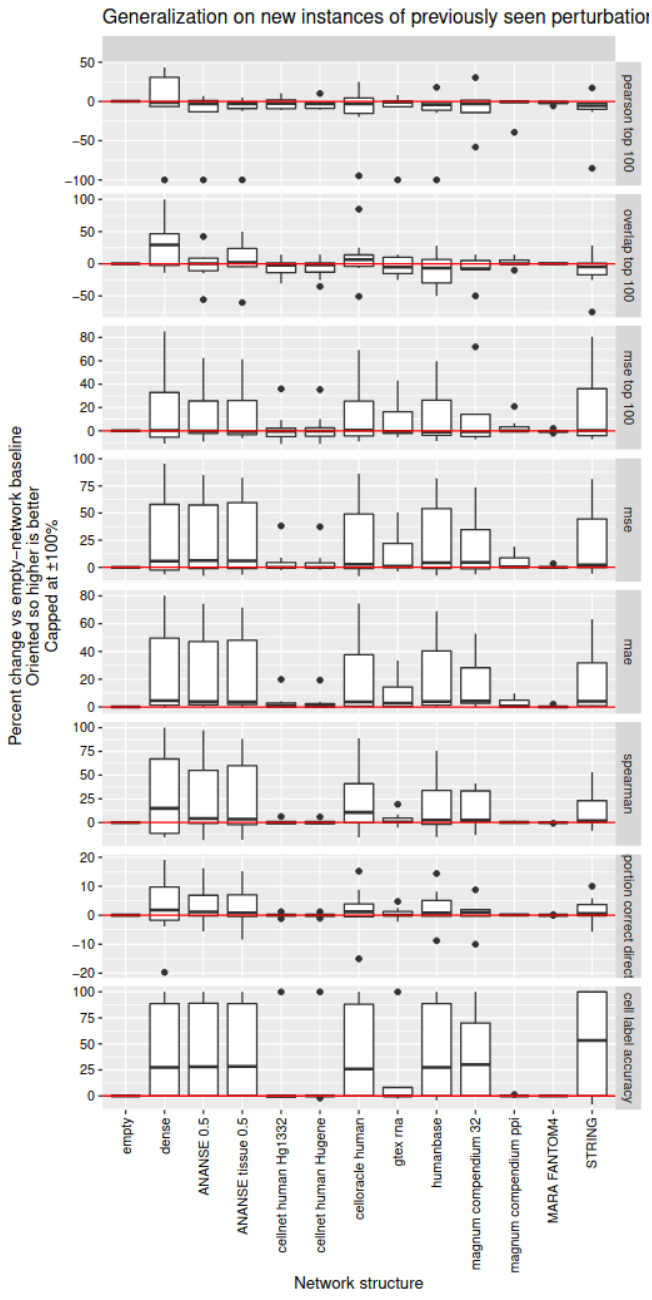

Fig. S6. Related to Fig. 2. Relative performance of different GRN structures on previously-seen observations. In this experiment, the data split guarantees that each perturbation condition present in the test data was seen at least once in the training data. Each point represents performance of a single method on a single dataset according to a single metric. The x axis shows networks used. The y axis shows performance as percent improvement relative to the

empty-network baseline, oriented so that higher numbers are better. The right margin shows different evaluation metrics.

Fig. S7

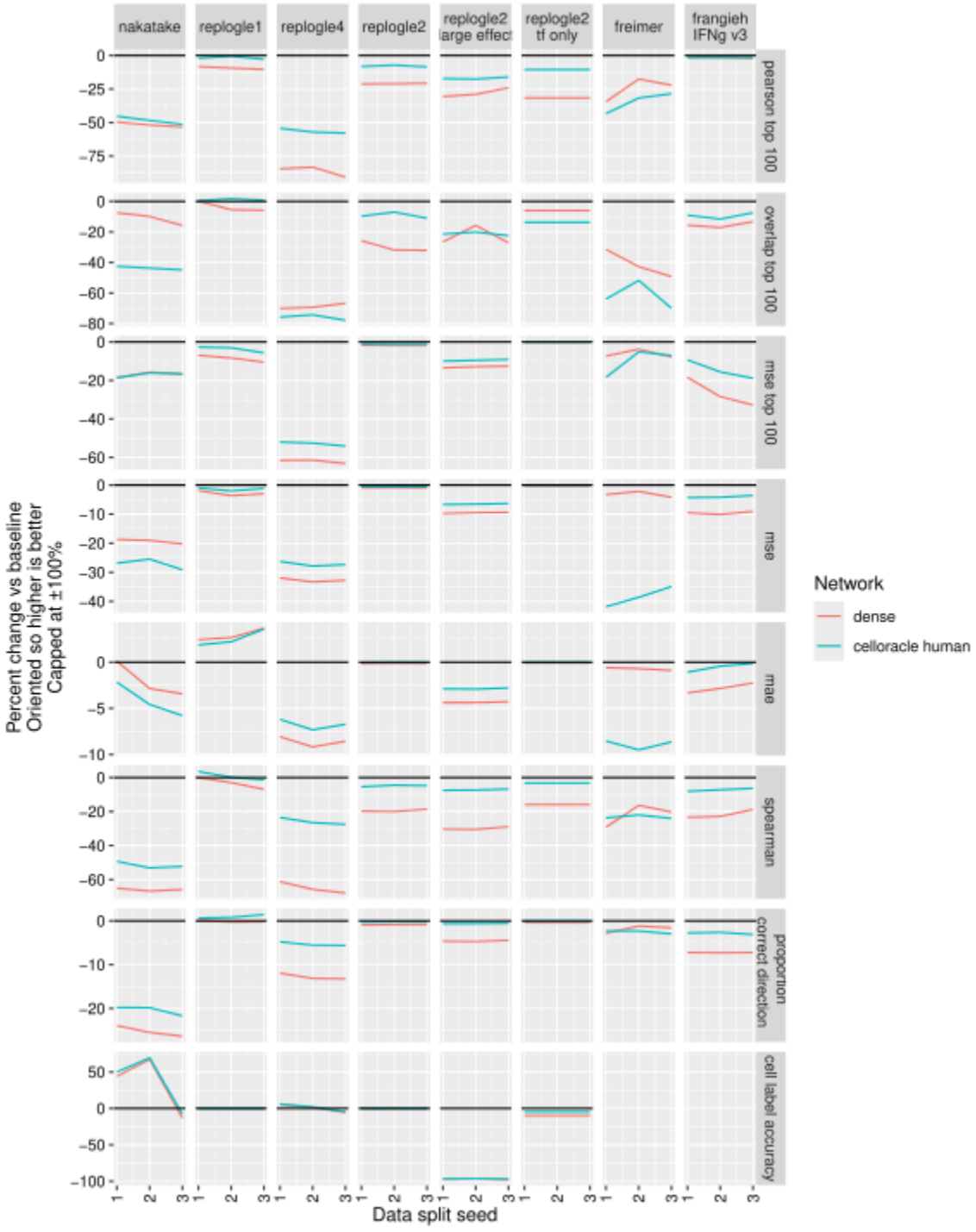

Fig. S7. Variability of various performance metrics (y axis) across data splits (x axis) in a comparison of ridge regression models using feature selection based on the default CellOracle human network, a dense (fully-connected) network, and an empty-network baseline (y=0).

Fig. S8

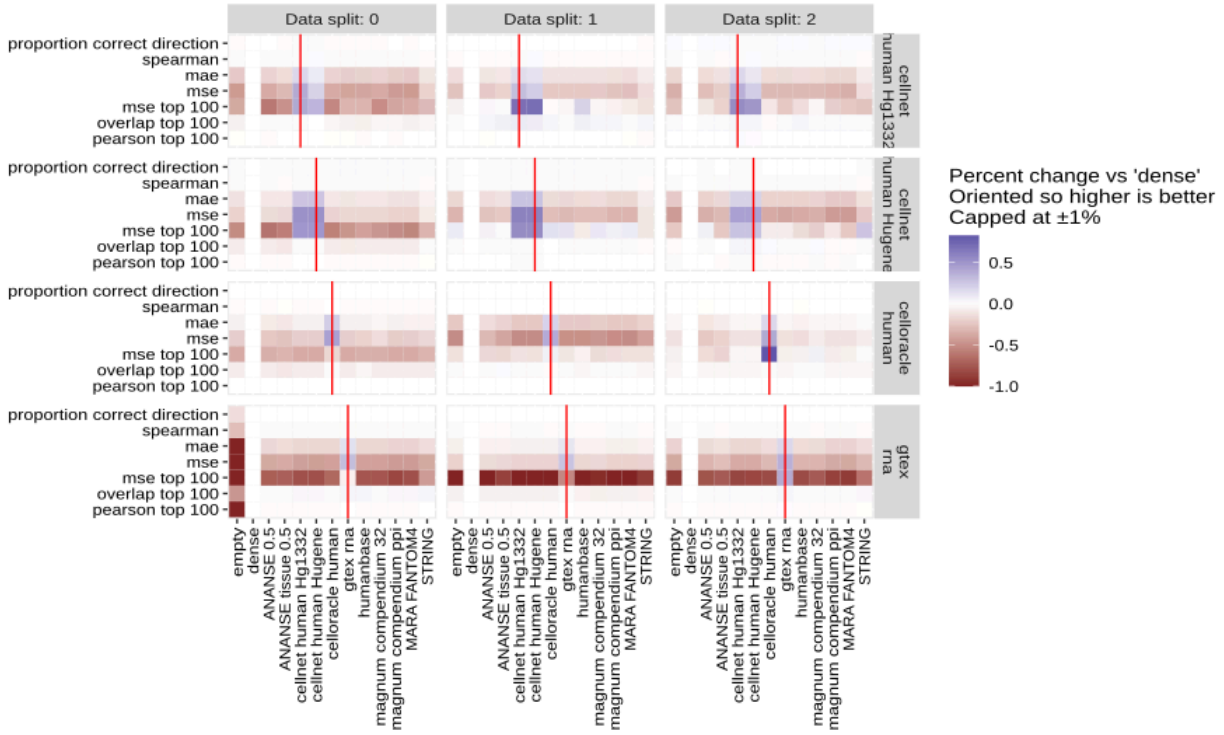

Fig. S8. Related to Fig. 2. Impact of regulatory network source (x axis) on expression forecasting performance (color scale). Evaluations use multiple train-test splits (top margin) of data generated from a specific network structure (right margin). Metrics are defined in Table 3. Data are generated from autoregressive models with random initial state that are run for a single time-step, with matched controls for each sample provided during training (Methods). Red lines indicate that the network used for inference matches the network used to generate the data.

Fig. S9

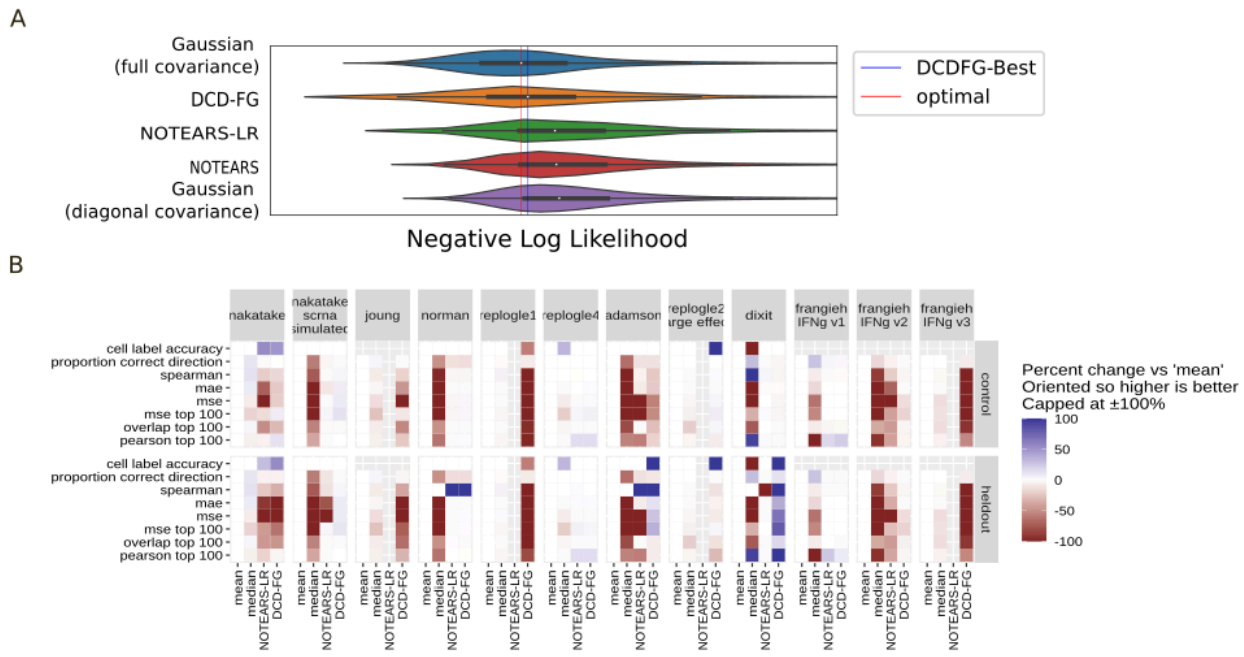

Fig. S9: Evaluations of DCD-FG.

a. Negative log likelihood on held-out perturbations in frangieh, performed using the original code for DCD-FG with Gaussian baselines added. The red line (“optimal”) indicates the best median performance across all methods. The blue line (“DCDFG-Best”) indicates the best median performance across all DCDFG-related methods.

b. Benchmarks of DCD-FG and baselines (x axis) on various datasets (top margin). We use as an initial state either the control or held-out expression of upstream genes (right margin). Evaluation metrics (y axis) are as defined in Table 3. Missing data indicate that DCD-FG or NO-TEARS did not converge, or that cell type labels were not available. Performance (colorscale) is shown as percent improvement over the ‘mean’ baseline, oriented so that higher numbers always indicate better performance. Percentages are capped at ±100%.

Fig. S10

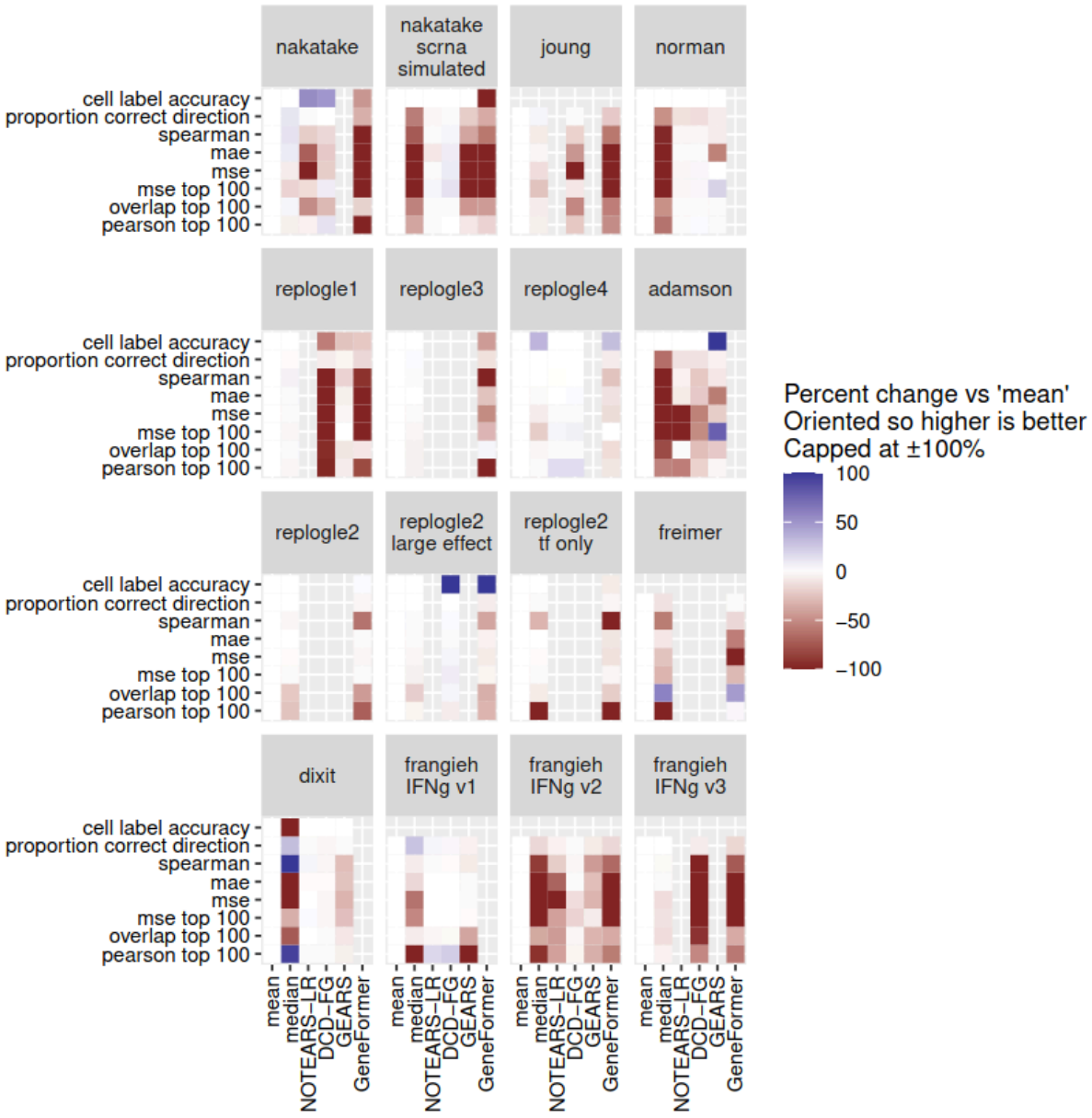

Fig. S10: Evaluations of published expression forecasting methods (x axis) versus simple baselines. Evaluation metrics (y axis) are as defined in Table 3. Performance (colorscale) is shown as percent improvement over the ‘mean’ baseline, oriented so that higher numbers

always indicate better performance. Percentages are capped at ±100%. Missing data indicate that metrics were not available, or methods did not converge, or methods cannot accept the available input data format.