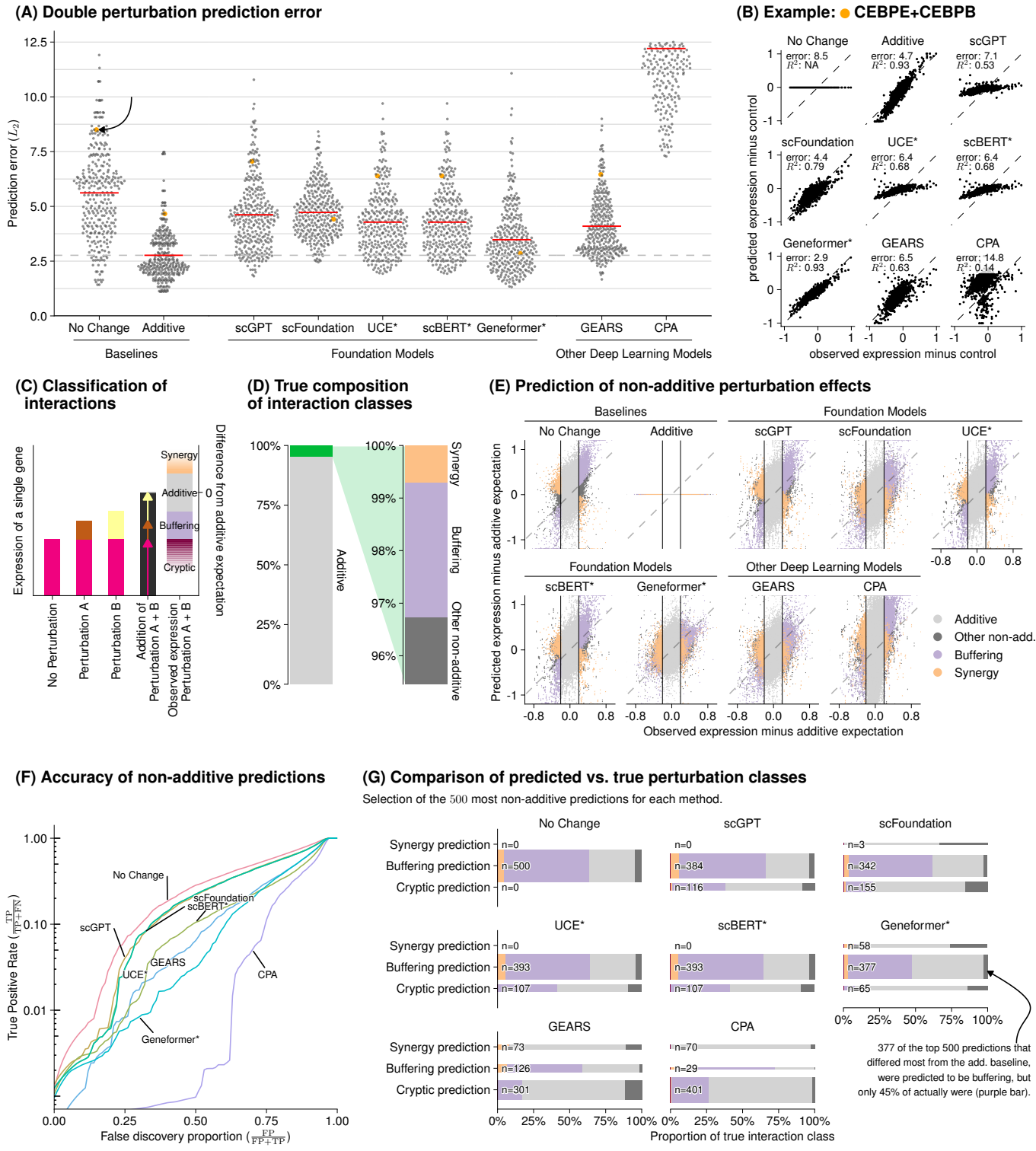
(A) Double perturbation prediction error
Prediction error (L_2)
12.5
10.0
7.5
5.0
2.5
0.0
No Change
Additive
scGPT
scFoundation
UCE*
scBERT*
Geneformer*
GEARS
CPA
Baselines
Foundation Models
Other Deep Learning Models
(B) Example: CEBPE+CEBPB
No Change
error: 8.5
R^2: NA
Additive
error: 4.7
R^2: 0.93
scGPT
error: 7.1
R^2: 0.53
scFoundation
error: 4.4
R^2: 0.79
UCE*
error: 6.4
R^2: 0.68
scBERT*
error: 6.4
R^2: 0.68
Geneformer*
error: 2.9
R^2: 0.93
GEARS
error: 6.5
R^2: 0.63
CPA
error: 14.8
R^2: 0.14
predicted expression minus control
observed expression minus control
(C) Classification of interactions
Expression of a single gene
Synergy
Additive
Buffering
Cryptic
Difference from additive expectation
No Perturbation
Perturbation A
Perturbation B
Addition of Perturbation A + B
Observed expression Perturbation A + B
(D) True composition of interaction classes
100%
75%
50%
25%
0%
Additive
99%
98%
97%
96%
Synergy
Buffering
Other non-additive
(E) Prediction of non-additive perturbation effects
Baselines
Foundation Models
Other Deep Learning Models
Predicted expression minus additive expectation
Observed expression minus additive expectation
Additive
Other non-add.
Buffering
Synergy
(F) Accuracy of non-additive predictions
True Positive Rate ($\frac{TP}{TP+FN}$)
False discovery proportion ($\frac{FP}{FP+TP}$)
No Change
scFoundation
scBERT*
scGPT
GEARS
UCE*
CPA
Geneformer*
(G) Comparison of predicted vs. true perturbation classes
Selection of the 500 most non-additive predictions for each method.
Synergy prediction
Buffering prediction
Cryptic prediction
Proportion of true interaction class
377 of the top 500 predictions that differed most from the add. baseline, were predicted to be buffering, but only 45% of actually were (purple bar).