(A) Single perturbation prediction error (B) Prediction error stratified by gene set size Replogle K562 Replogle RPE1 Adamson Replogle K562 Replogle RPE1 Adamson error 2 Prediction error  $(L_2)$ 10 prediction 3 2 Mean LM with G,P-from training LM with G,P from training GEARS-LM with G,P from training GEARS-LM with P from RPE1 Replogle scGPT scGPT GEARS scGPT LM with P from K562 Replogle scFoundation scFoundation LM with P from K562 Replogle scFoundation 100 1000 10 100 1000 10 100 1000 Number of genes (log-scale) **GEARS** LM with G,P from training LM with P from Replogle (C) Plugging different embeddings into the linear model GEARS vs. scGPT vs. scGPT vs. LM vs. LM vs. LM with G from scGPT LM with G from scFoundation LM with P from GEARS LM with P from Replogle LM with random P Prediction error (L2) Prediction error  $(L_2)$ Prediction error  $(L_2)$ Prediction error  $(L_2)$ Prediction error  $(L_2)$ 6 6 5 5 -5 Adamson Replogle K562 Replogle RPE1 Adamson Replogle K562 Replogle RPE1 Adamson Replogle K562 Replogle RPE1 Replogle K562 Replogle RPE1 Adamson Replogle K562 Replogle RPE1