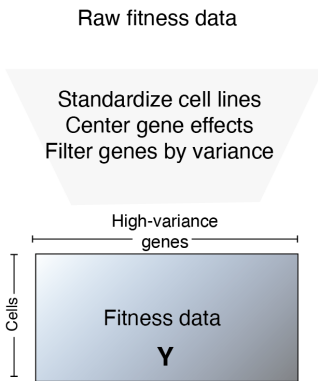


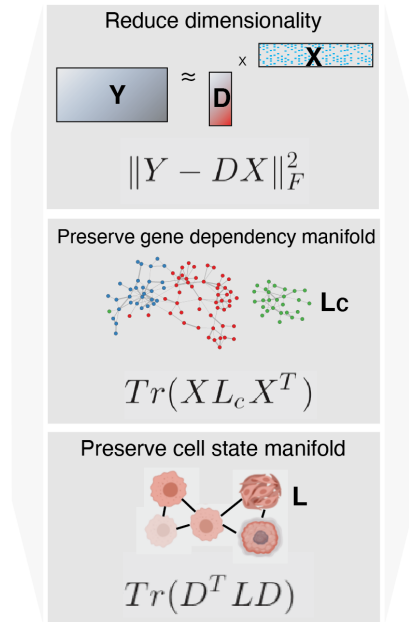
# Supplement

A

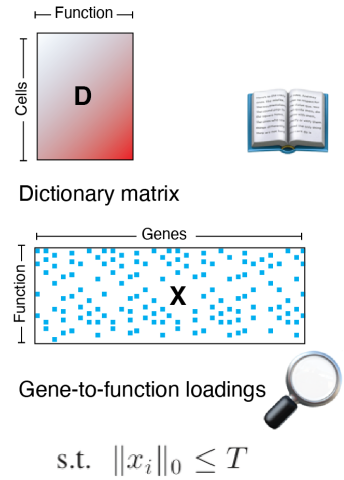
**Preprocessing**



**Graph-regularized dictionary learning**  
*Objectives*



**Output**



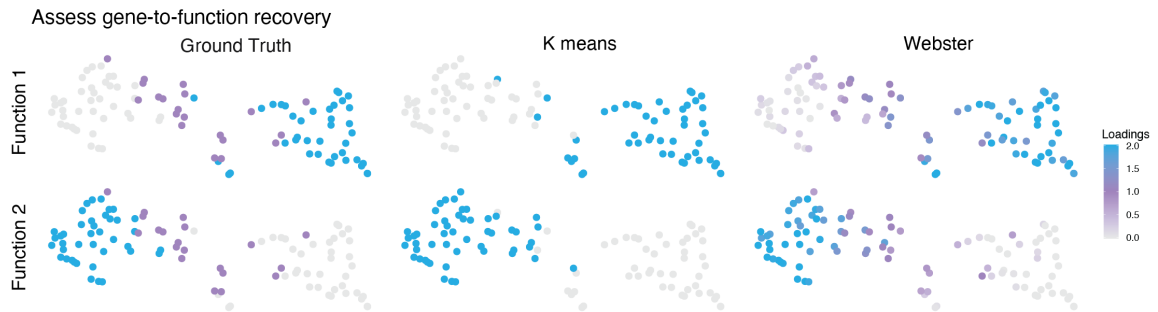
**Objective function**

$$\arg \min_{D, X} \|Y - DX\|_F^2 + \alpha Tr(D^T L D) + \beta Tr(X L_c X^T) \quad \text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i.$$

**Hyperparameters**

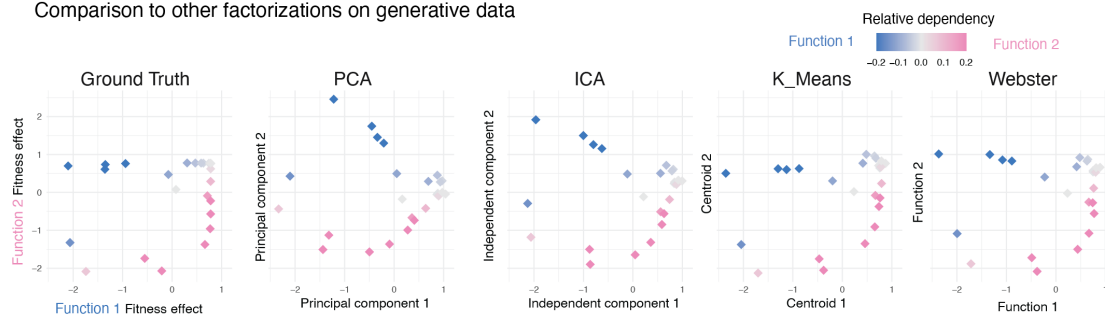
- k** = latent dimension size
- L** = cell Laplacian (num neighbors, metric)
- Lc** = gene Laplacian (num neighbors, metric)
- α** = weight of cell Laplacian
- β** = weight of gene Laplacian
- T** = sparsity

B



C

**Comparison to other factorizations on generative data**



## Figure S1: Methodological details of Webster. Related to Figure 1.

- A. Extended version of Figure 1A showing the objective function of graph-regularized dictionary learning (Yankelevsky and Elad, 2016). Given a raw fitness dataset, Webster first preprocesses the data by standardizing cell contexts (rows), then centering gene effects (columns). It then applies a simple selection threshold to automatically choose a set of high variance gene effects (columns) to compose the input data matrix  $Y$ . Webster factorizes  $Y$  into two low-rank matrices,  $D$  and  $X$ , by (1) minimizing the approximation error of the low-rank factorization, (2) preserving gene effect (column) similarity from  $Y$  across columns of  $X$ , and (3) preserving cell context (row) similarity from  $Y$  across rows of  $D$ . Besides the key parameters  $k$  and  $t$ , which controls the rank of the factorization and the number of non-zero entries per column of  $X$ , respectively, additional parameters include: the neighbor graphs used in the row and column graph-regularization (default: 5 nearest neighbors, chosen by cosine similarity); and the relative contributions of the graph regularization terms to the overall objective (default:  $\alpha = 0.2$  and  $\beta = 0.6$ , as explained in (Yankelevsky and Elad, 2016)).
- B. The input genes from Figure 1B are embedded in a 2D layout using UMAP, and the gene-to-function assignments for  $k$ -means and Webster are plotted as colors on each data point. While  $k$ -means and Webster capture the same latent variables from the data,  $k$ -means performs “hard clustering” that assigns pleiotropic genes to either function based on noise, while Webster performs “soft clustering” that accurately assigns pleiotropic genes to both functions.
- C. Comparison between Webster and other low-rank factorization methods commonly applied to biological data. Principal Components Analysis (PCA), Independent Components Analysis (ICA), and  $k$ -means were parameterized to recover two latent variables, using as input the data matrix described in Figure 1E. The recovered latent variables from each method are plotted in comparison with the ground truth described in Figure 1B (left) and the dictionary recovered by Webster described in Figure 1F (right). Both PCA and ICA are sensitive to global variance in the data and therefore capture outlier cells (those sensitive to both Function 1 and 2) in their first latent variable.  $k$ -means recovers nearly identical latent variables as Webster.

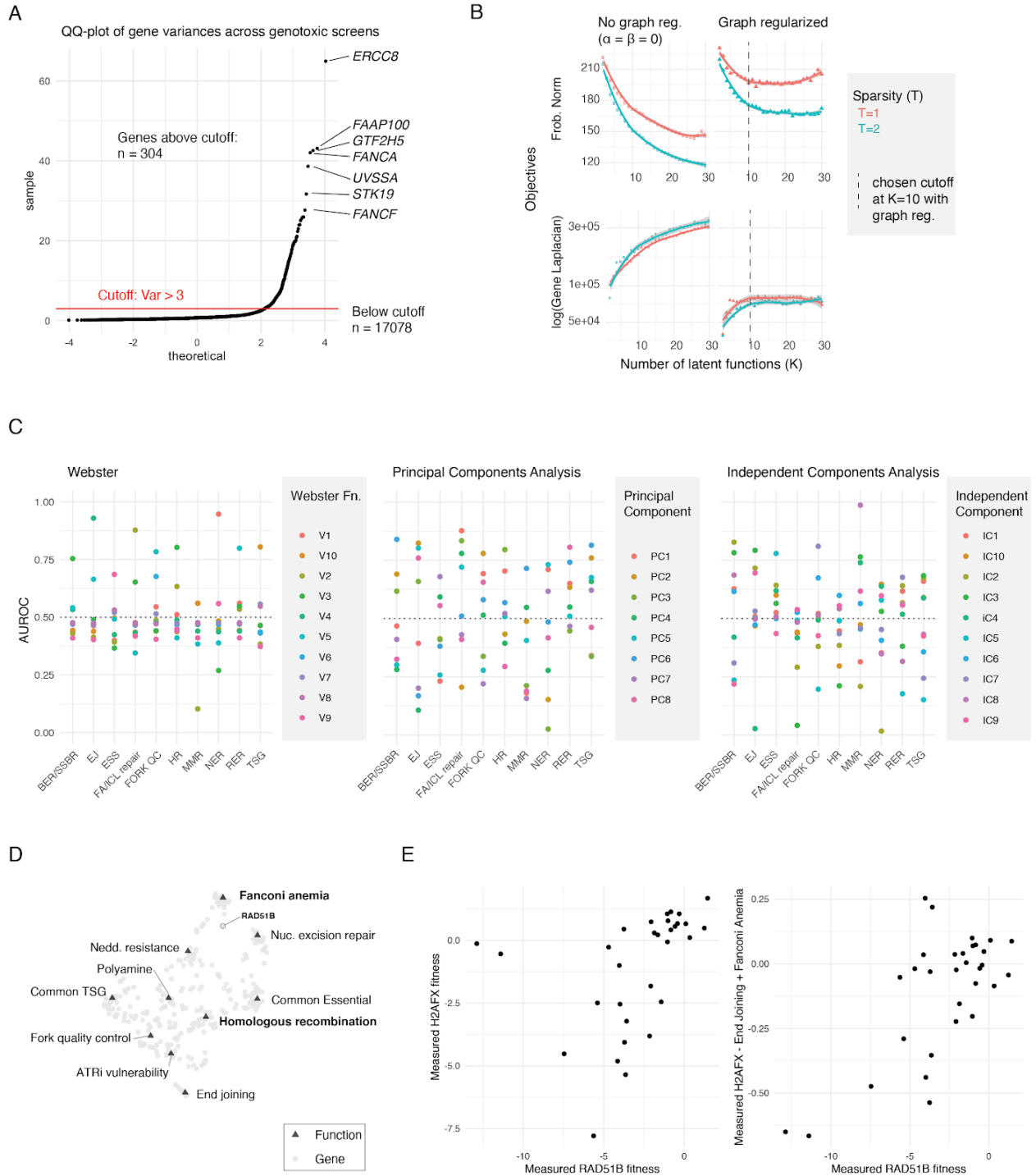


Figure S2: Assessment of Webster on genotoxic fitness data. Related to Figure 2.

A. High-variance gene selection. A quantile-quantile plot is shown for the observed fitness variances for 17,382 gene effect measurements (y-axis), in comparison to a theoretical normal distribution fitted to the distribution of these variances (x-axis). A threshold is drawn

to distinguish gene effects whose variance exceeds the theoretical normal distribution, resulting in 304 high-variance fitness genes chosen as input to Webster. In genome-scale screens, a large number of genes will be non-essential for fitness; such genes will exhibit fitness effects driven by experimental noise rather than biological signal. We assume that the variances of non-essential genes are normally distributed, and choose genes whose variances across treatments are positive outliers in this distribution.

- B. Webster parameter grid search. Using the same data as input, we applied Webster across many values of  $k$  and  $t$ , with and without graph-regularization. Diminishing returns for both reconstruction error (Frobenius norm) and gene similarity (Gene Laplacian) are reached with  $k = 10$  for both values of  $t$ . We chose  $t = 2$  in order to model pleiotropic effects in the genotoxic screening data.
- C. Interpretability of latent factors recovered from Webster, PCA and ICA. Using the literature annotations from (Olivieri et al., 2020) as ground truth, we calculated the Area Under the Receiver Operating Characteristic curve (AUROC) for each of ten genesets across each of the learned factors from all three models. The number of dictionary elements for Webster were chosen as described above; the number of PCA components was chosen with a standard elbow plot over PCA eigenvalues; the number ICA components was chosen according to (Kairov et al., 2017). The loadings for each gene over each component were used as the predictors for the AUROC metric. An AUROC  $> 0.5$  score indicates that positive loadings were predictive of the geneset, while an AUROC score  $< 0.5$  indicates that negative loadings were predictive of the geneset. A score of 0.5 in AUROC indicates a performance equivalent to random chance assignment. The imposed sparsity in Webster's gene loadings leads to interpretable latent variables mapping strongly to individual genesets.
- D. Joint UMAP embedding of gene and functional effects as in Figure 2G and 2H, with all functions labeled. The RAD51B gene effect is embedded between Fanconi Anemia and Homologous Recombination (bolded).
- E. Scatterplots comparing the measured fitness effect of RAD51B (x axis, both plots) with measured H2AFX (y-axis, left) and H2AFX - End Joining + Fanconi Anemia (y-axis, right).

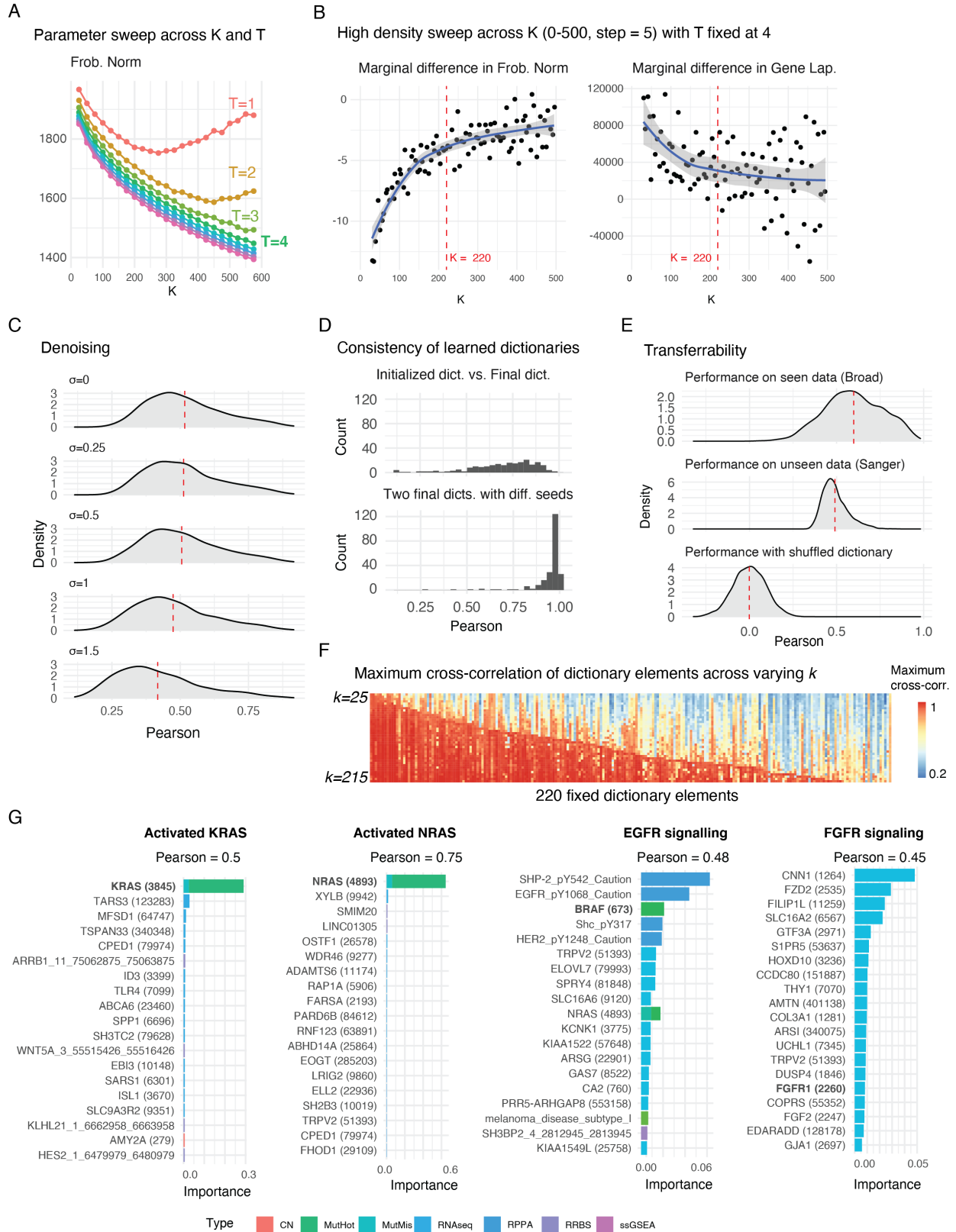


Figure S3: Assessment of Webster on cancer cell fitness data. Related to Figure 3.

- A. Webster parameter grid search. Using the cancer cell fitness data as input, we applied Webster across many values of  $k$  (25 to 600, with step size 25) and  $t$  (1 to 10). As  $t = 1:3$  performed poorly at large values of  $k$ , we chose  $t = 4$  for the factorization.
- B. Higher-density grid search. With  $t = 4$  fixed, we swept across  $k$  (25 to 600, with step size 5) with multiple random initializations with different seeds. Plotted are the marginal improvements seen in model objectives with each additional step size of  $k$ , averaged over random initializations. Diminishing returns in both objectives are observed around  $k = 220$ , which was chosen for the final factorization.
- C. Denoising properties of Webster. The starting fitness data was corrupted with different amounts of random noise. After splitting genes into training and test sets (3:1 split), we then applied Webster ( $k=220, t=4$ ) to learn a dictionary from the noisy training data. From this dictionary, we performed orthogonal matching pursuit to model the noisy test genes in terms of dictionary elements. We compared this reconstructed profile against the ground truth test gene profiles, which were unseen during model training. The Pearson correlation of the reconstructed test genes versus their ground truths are plotted as a distribution per noise level. The uppermost distribution ( $\sigma = 0$ ) corresponds to Webster's performance in the absence of noise. Dashed red lines mark the mean of each distribution.
- D. Dictionary learning metrics. Left: Initialized vs. final dictionaries. In our Webster implementation, we initialize dictionary learning using a dictionary of  $k$  initial gene effects chosen by  $k$ -medoids. Each column of the initial dictionary ( $k$ -medoids) was correlated to the corresponding column in the final dictionary after 20 algorithm iterations ( $k = 220, t = 4$ ). The resulting 220 Pearson correlation values are shown as a histogram. Right: Using the same  $k$ -medoids dictionary as a starting point, dictionary learning was performed using two different random seed initializations. The Pearson correlations of the corresponding columns from each dictionary are shown as a histogram.
- E. Transferability of Webster dictionary elements to unseen data. Parallel genome-scale screens were performed at Broad and Sanger Institutes for 150+ common cancer cell lines, using different CRISPR-Cas9 reagents and culturing strategies. We assessed the transferability of a Webster dictionary trained on Broad data (which used the Avana CRISPR-Cas9 guide library) to model gene effects captured by the Sanger Institute (which used the Sanger CRISPR-Cas9 guide library). We learned a Webster dictionary ( $k=220, t=4$ ) over the 675 cell lines screened by the Broad. We then subsetted the learned dictionary to a set of 150+ common cell lines, and used this smaller dictionary to model gene effects measured by Broad Institute or the Sanger Institute. The Pearson correlation of the reconstructed genes are plotted as a distribution. As a null comparison, we shuffled the rows of the dictionary and performed the same modeling using this shuffled dictionary. Dashed red lines mark the mean of each distribution.
- F. Reproducibility of dictionary elements learned at  $k=220$  over other values of  $k$ . Each column in the heatmap corresponds to one of the 220 dictionary elements reported in the paper ( $k=220, t=4$ ). Each row in the heatmap represents a dictionary that was learned at a smaller value of  $k$ , with  $t$  fixed ( $k=25, 30, 35, \dots, 215, t = 4$ ). Each cell in the heatmap is colored according to the maximum cross-correlation between all elements in the lower- $k$  dictionary (row) and a specific element in the finalized dictionary (column). Columns are

ordered according to the lowest  $k$  for which that element “appears” in the smaller dictionary (defined as Pearson cross-correlation  $> 0.9$ ).

- G. Biomarker analysis for SHOC2 functional effects. We performed a random forest regression on the fitness effect of each of the four underlying functions, using baseline -omics measurements across cancer cell lines as features (including RNA-seq bulk transcriptomic data, mutational hotspot data, protein abundance data, etc). The model performances (Pearson correlation) are shown next to barplots displaying the feature importances in the final models. Relevant biomarkers for each function are bolded. (Abbreviations; CN = copy number; MutHot = mutational hotspot; MusMis = missense mutation; RPPA = Reverse Phase Protein Array; RRBS = Reduced-representation bisulfite sequencing; ssGSEA = single sample gene set enrichment analysis)



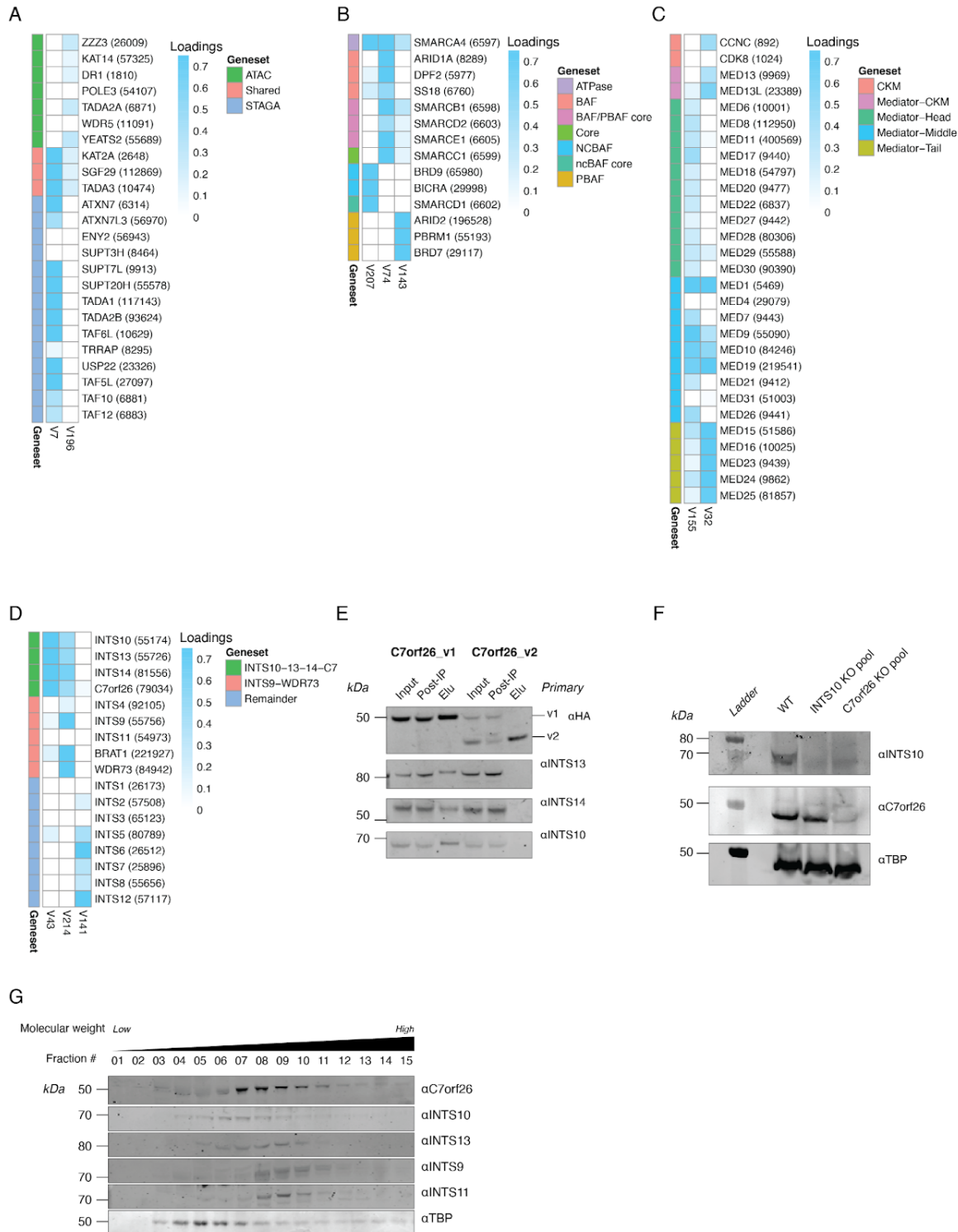
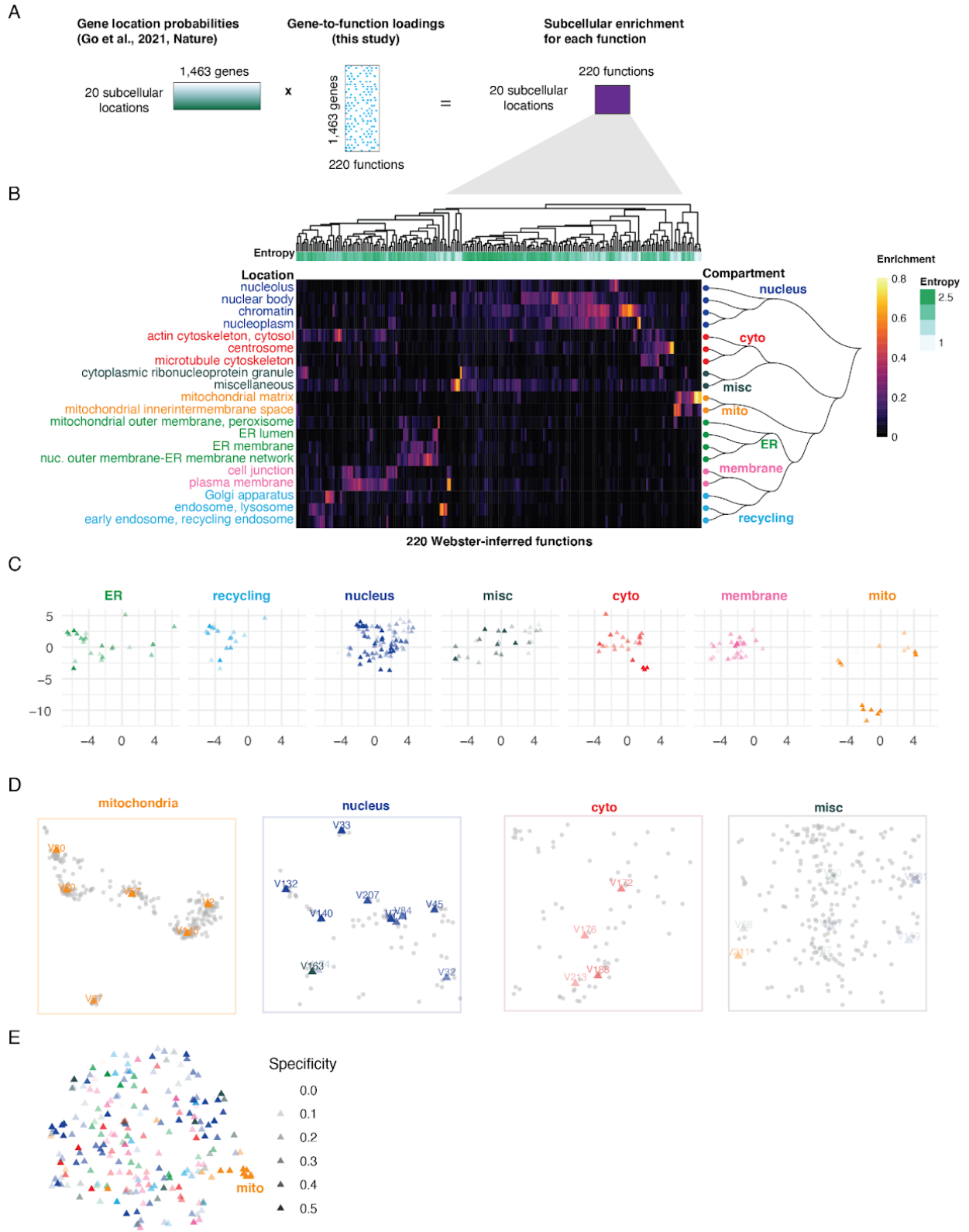


Figure S4: Modular pleiotropy in protein complexes from cancer fitness data. Related to Figure 4.

- A. Focus on STAGA/ATAC complexes. Subunits unique to STAGA, unique to ATAC or shared between both were taken from (Spedale et al., 2012). A heatmap is displayed where each row displays a subunit's loadings across selected Webster functions learned

from fitness data alone. Webster learned fitness effects for both complexes individually, and represented the fitness effect of shared subunits as a mixture of both (loaded onto both functions).

- B. Focus on SWI/SNF complexes. Subunit organization was taken from (Mashtalir et al., 2018). A heatmap is displayed where each row displays a subunit's loadings across selected Webster functions learned from fitness data alone. Webster learned fitness effects for ncBAF, cBAF and pBAF complexes individually, and the fitness effect of SMARCA4 as a mixture of all three.
- C. Focus on the Mediator complex. Subunit organization was taken from (Tsai et al., 2014). A heatmap is displayed where each row displays a subunit's loadings across selected Webster functions learned from fitness data alone. Webster learned a fitness effect for the Mediator Tail/CKM modules separately from the Mediator Head/Shoulder modules.
- D. Focus on the Integrator complex. Subunit organization was taken from (Pfleiderer and Galej, 2021; Sabath et al., 2020; Tilley et al., 2021; Zheng et al., 2020). A heatmap is displayed where each row displays a subunit's loadings across selected Webster functions learned from fitness data alone. Webster learned a fitness effect for the INTS10-13-14 module (Pfleiderer and Galej, 2021; Sabath et al., 2020), the WDR73-INTS9 module (Tilley et al., 2021), and the Backbone/Shoulder modules (Zheng et al., 2020) (designated above as Remainder). INTS11, the main catalytic subunit of the Integrator complex, is not loaded onto any of these functions, due to its status as a highly essential gene across all cancer cell lines.
- E. Biological replicate experiment of the immunoprecipitation shown in Figure 4E.
- F. Biological replicate experiment of the knockout experiment shown in Figure 4G.
- G. Density glycerol gradient ultracentrifugation on 293T nuclear extracts shows size separation of Integrator complex subunits across different molecular weights. TBP is shown as a non-Integrator complex control.

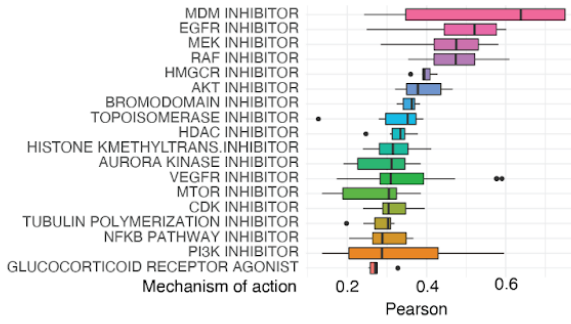


## Figure S5: Subcellular localization analysis from cancer fitness data. Related to Figure 5.

- A. Schematic overview of subcellular localization analysis. A set of 1,463 fitness genes were also profiled in a recent subcellular localization experiment (Go et al., 2021), which reports the localization probability of each gene product over 20 inferred subcellular locations. We performed a matrix multiplication between their localization probabilities and our fitness-inferred gene-to-function loadings. The resulting matrix of 220 functions x 20 locations represents the overall distribution of localization probabilities over the learned Webster functions.
- B. A heatmap of the matrix described in A. Both rows (locations) and columns (functions) are hierarchically clustered. Clustering rows results in seven hierarchically defined cell compartments: nucleus, mitochondria, endoplasmic reticulum (ER), recycling, membrane, cytoplasm and miscellaneous. The miscellaneous category is carried over from (Go et al., 2021). Because proximity labelling proteomics were used to define subcellular locations in that study, proteins that are part of large complexes were predominantly co-labeled with other protein complex subunits, thereby decreasing their ability to infer unique subcellular locations for these proteins.
- C. Facet plot of Figure 5B, in which only functions are plotted as data points in the embedding. Functions enriched for each of the seven compartments are plotted separately.
- D. Additional panels for Figure 5C, showing function-level insets for mitochondria, nucleus, cytoplasm and miscellaneous compartments.
- E. Accompanying figure for Figure 5E. Using only functional fitness effects (dictionary elements) in the global embedding ablates the compartmental structure observed in Figure 5B (in which genes and functions are co-embedded). This is because dictionary elements are relatively de-correlated from one another, a property known as *mutual incoherence*. The notable exception is the mitochondrial functions, which remain clustered in this setting due to the fact that a predominant confounder (media composition across cell lines) explains a portion of variance present in each of these dictionary elements (related to findings explored in (Rahman et al., 2021)).

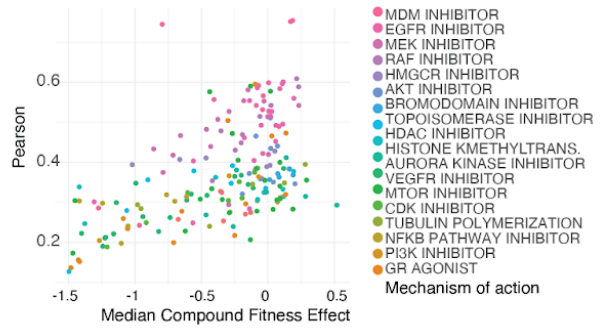
A

Approximation of compound sensitivity profiles in terms of dictionary elements learned from gene perturb. data



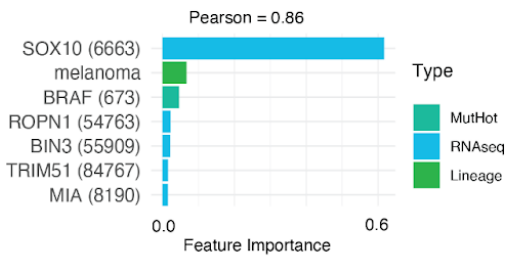
B

Scatterplot of approximation score (Pearson) and median compound fitness effect over cell lines



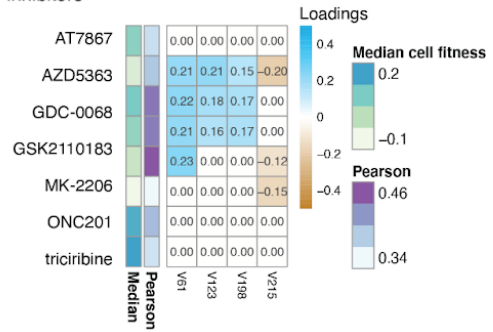
C

Biomarker analysis for the BRAF signaling function



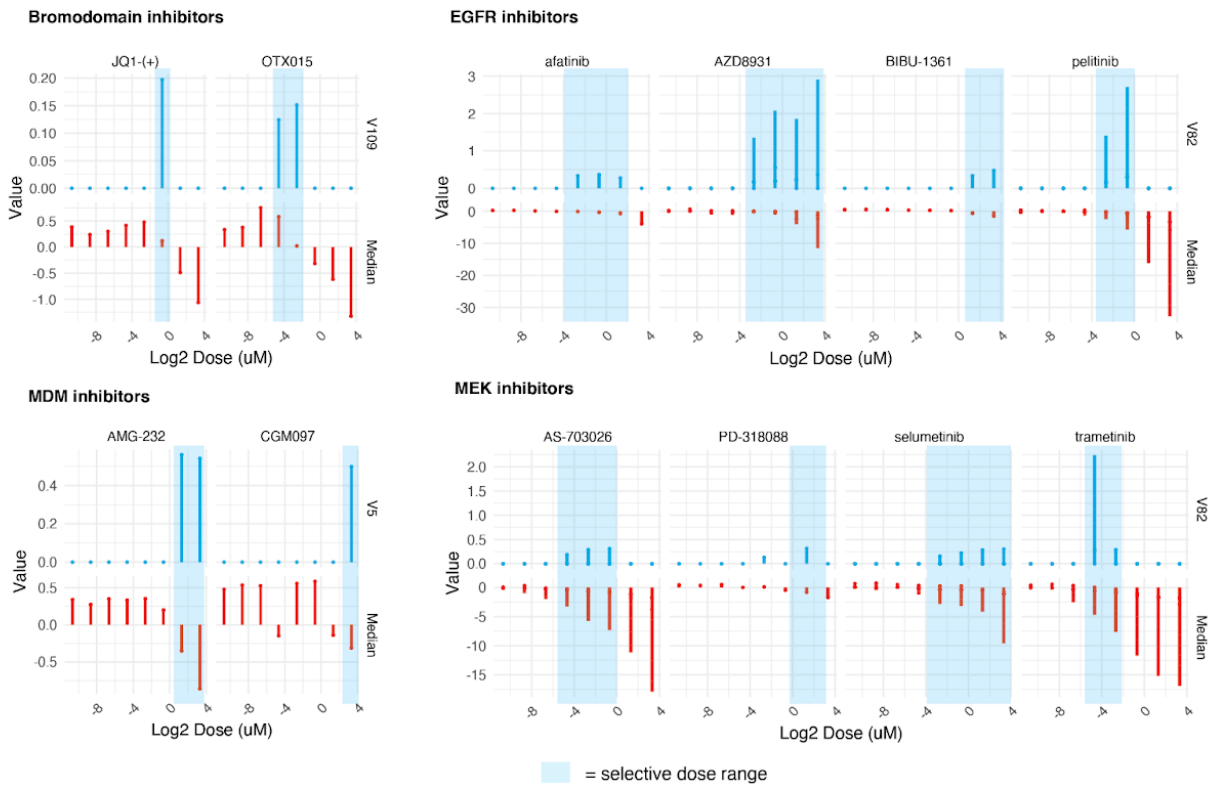
D

AKT inhibitors



E

Dose sensitivity of compound loadings



## Figure S6: Compound embedding results. Related to Figure 6.

- A. Compound sensitivity profiles over 360+ cancer cell lines were obtained from the PRISM Drug Repurposing dataset (Corsello et al., 2020). Each of these profiles was modeled as a sparse linear combination of four dictionary elements, using a dictionary trained on gene perturbation data (from Figure 3B). The quality of these approximations was assessed using a Pearson correlation to the original compound sensitivity profile. For each compound class, the distribution of Pearson correlations across individual drugs belonging to that class are shown in a box and whisker plot. Compound classes are ordered by their mean Pearson correlation.
- B. Each data point in the scatter plot represents one of the compounds from PRISM that was modeled in terms of gene functions. The X axis charts the median cell fitness of each compound, and the Y axis charts the Pearson correlation of the approximated profile to the measured profile.
- C. Same as Figure S3E, but for the BRAF Signaling function.
- D. A heatmap of compound-to-function loadings. Each row represents a compound sensitivity profile for an AKT inhibitor from the PRISM primary screen (2.5 uM dose), and each column represents a Webster function learned from genetic data. Loadings values are displayed in each cell of the heatmap. The first three gene functions model RICTOR/AKT, PIK3CA signaling and PTEN signaling, respectively. The last function displays a fitness effect specific to blood cell lines, and therefore captures a batch effect present in the original PRISM data (in which suspension and adherent cell lines display differing chemical sensitivity profiles). The median fitness effect across cells of that compound, as well as the Pearson correlation of the approximated profile to the measured profile, are also shown.
- E. Additional dose-sensitive loading plots accompanying Figure 6E.