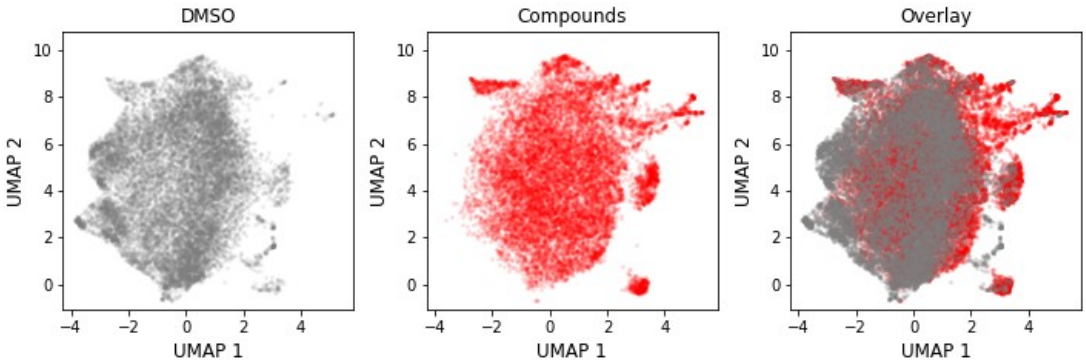
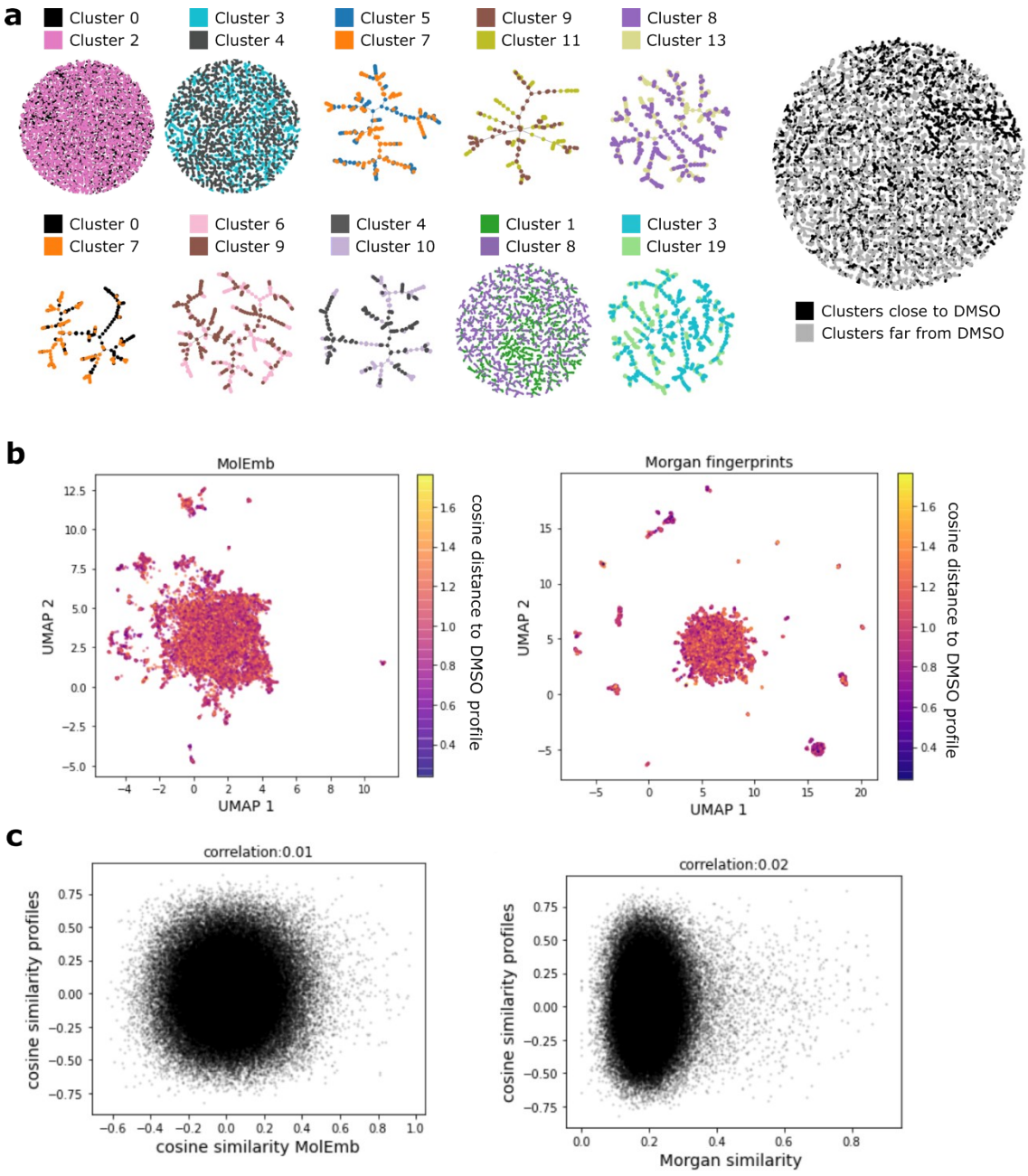
Electronic Supplementary Material (ESI) for Digital Discovery.

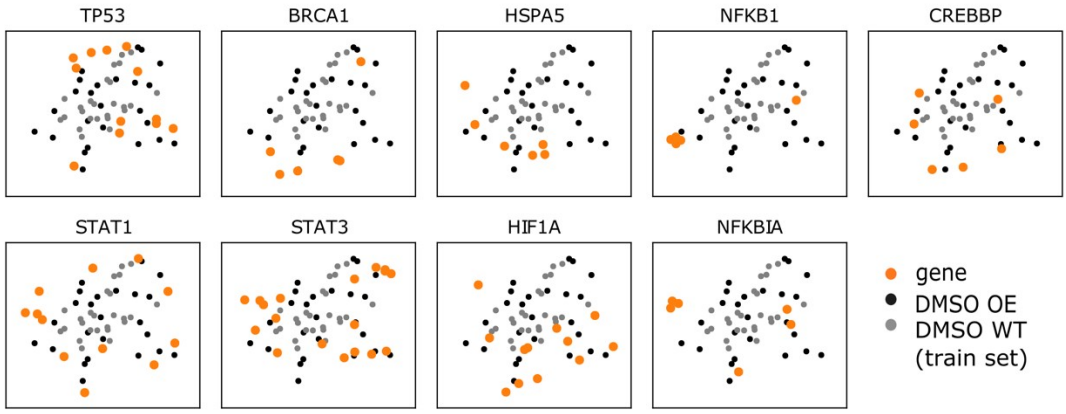
Supplementary Figures



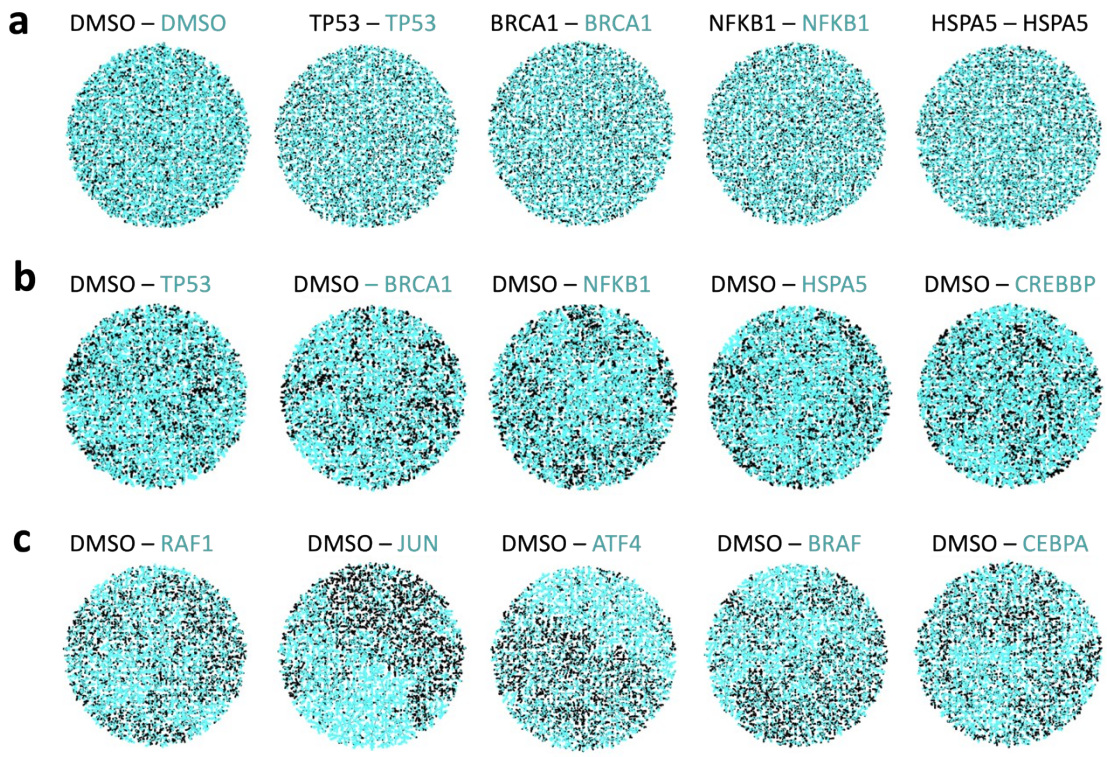
Supplementary Figure 1. Projection of training set morphological profiles. UMAP projections were computed based on 15,000 randomly selected profiles from compounds and DMSO (30,000 profiles in total).



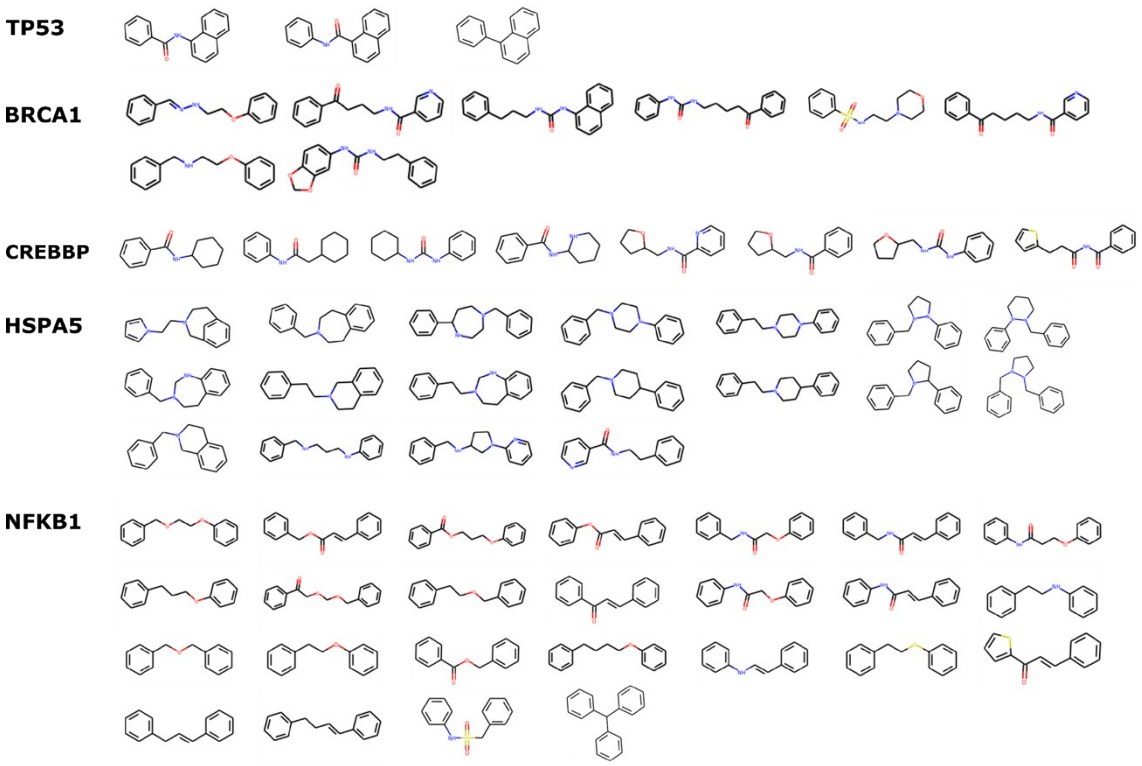
Supplementary Figure 2. Chemistry – morphology relationships in the training set. a t-map projections of Morgan fingerprints of training set molecules colored by their profiles' cluster (see Figure 2a). **b** UMAP projection of molecular embeddings (left) and Morgan fingerprints (right) of training set molecules colored by their morphological distance of their median profile to the median DMSO profile. **c** Correlation between morphological similarity and chemical similarity for training set molecules. 100,000 randomly selected pairwise similarities of median profiles vs their corresponding molecular embeddings cosine similarity (left) or dice similarity of Morgan fingerprints (right). Correlation values report Pearson correlation.



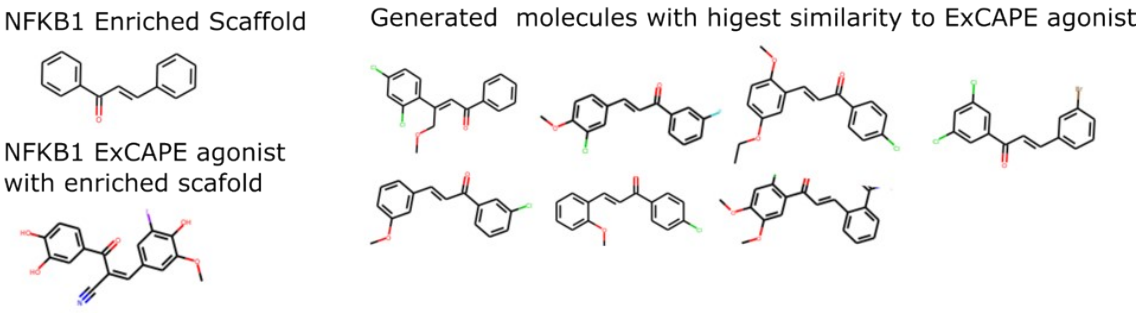
Supplementary Figure 3. Projection of overexpression profiles. UMAP projections of morphological profiles for the 9 overexpressed genes with available ExCAPE agonists and 50 randomly selected DMSO controls. DMSO WT: neutral controls from the training set. DMSO OE: empty vector controls from the overexpression dataset (test set). Only 3 out of the 5 fluorescent channels from the original dataset (BBBC037v1) are displayed. Hoechst: nucleus (blue), Phalloidin: Actin, Golgi and Plasma membrane (magenta), Concanavalin A: ER (green).



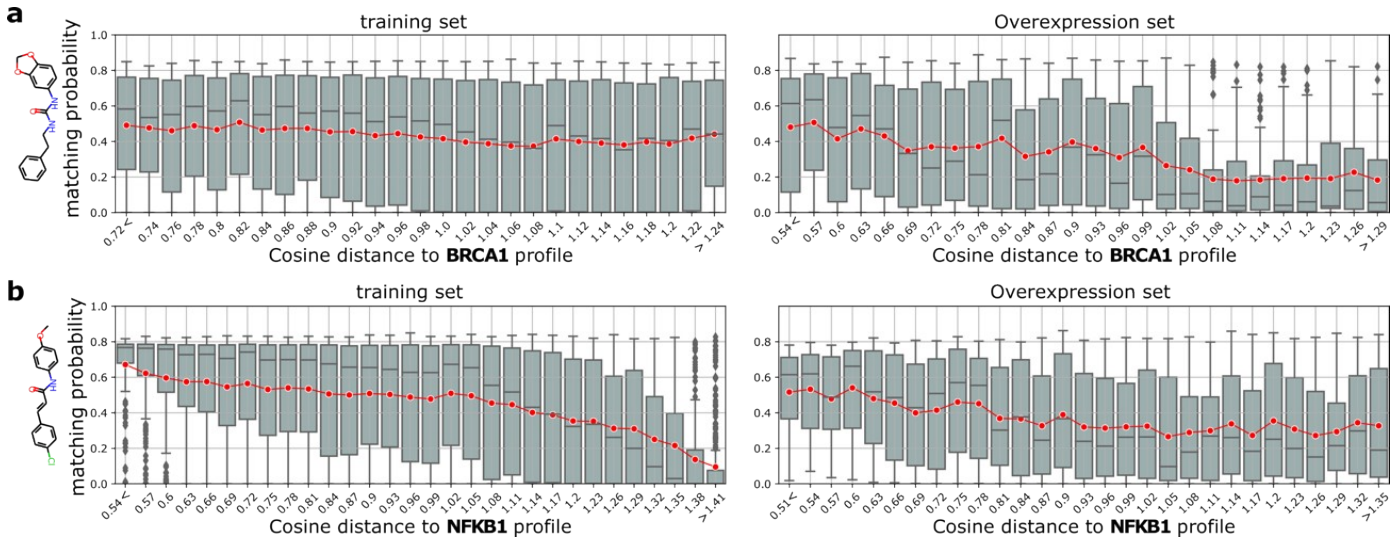
Supplementary Figure 4. Translation of phenotypic similarity to generated chemical similarity from over expression profiles. t-map projections of Morgan fingerprints of generated molecules, color-coded by the gene of their conditioning profile (5000 randomly selected samples per gene). **a** Intra-gene reference: two random sets of molecules generated with the same morphological conditioning. **b** Comparison of generated molecules conditioned on overexpressed genes with available ExCAPE agonists vs. negative controls. **c** Strong overexpression phenotype reference: Comparison of generated molecules conditioned on the top differentiable genes vs. negative controls. DMSO: empty vector controls from the overexpression dataset.



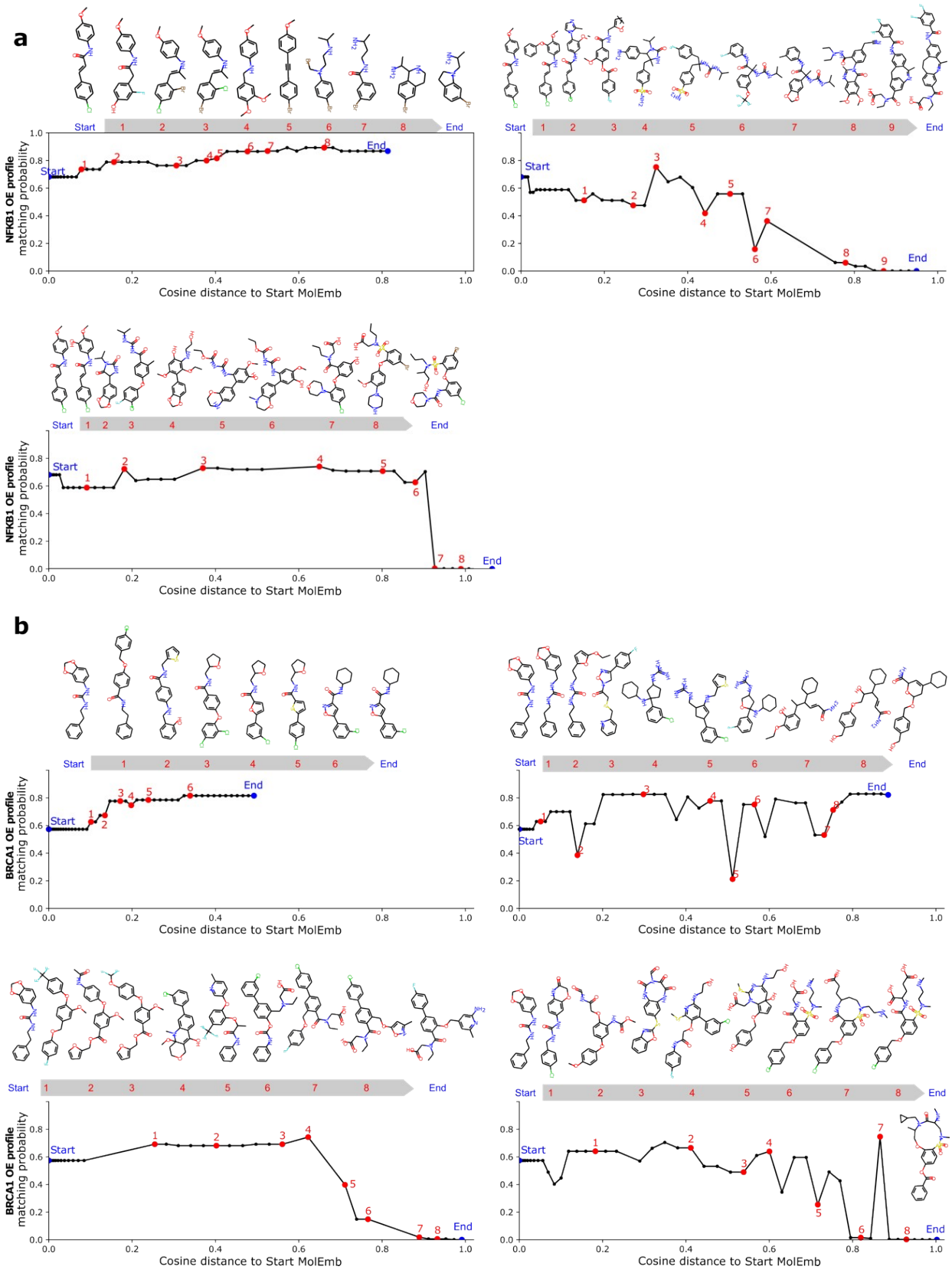
Supplementary Figure 5. Significantly enriched scaffolds on generated molecules conditioned on overexpression profiles. Scaffolds with significantly enriched counts relative to DMSO for the 5 overexpressed genes with most ExCAPE agonists. Significance was determined with a Fisher's exact test and a p-value of 0.01.



Supplementary Figure 6. Similar generated molecules to second NFKB1 ExCAPE agonist with a significantly enriched scaffold. Generated molecules conditioned on NFKB1 profiles with highest similarity to the specified NFKB1 agonist. All generated molecules displayed a dice similarity of at most 0.32 to the ExCAPE agonist.

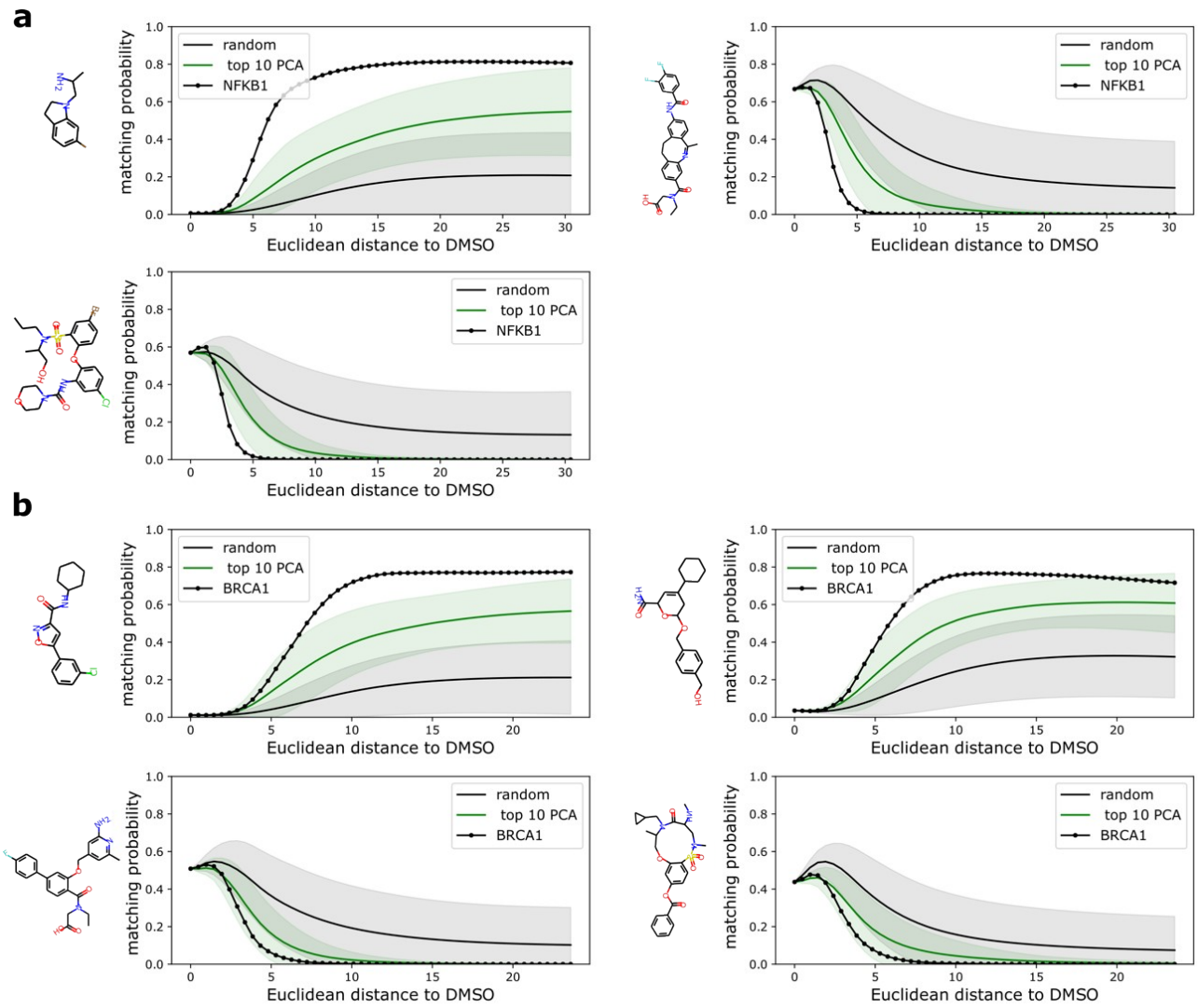


Supplementary Figure 7. Changes in the matching probability against selected ExCAPE agonists as a function of the distance to the overexpression profile of each agonist's gene. The condition network is used to compute the matching probability between the molecular embedding of the two ExCAPE agonists with enriched scaffolds and profiles in the training (per-compound median) and overexpression (per-gene median) datasets. Profiles are sorted and binned by increasing cosine distance to the overexpression profile of the respective agonist' gene. Histogram bins are equally spaced and capped at the lower and higher ends to ensure that edge bins would contain at least 100 samples. Red line shows the mean probability for each bin. **a** classification against BRCA1 ExCAPE agonist. **b** classification against NFKB1 ExCAPE agonist.

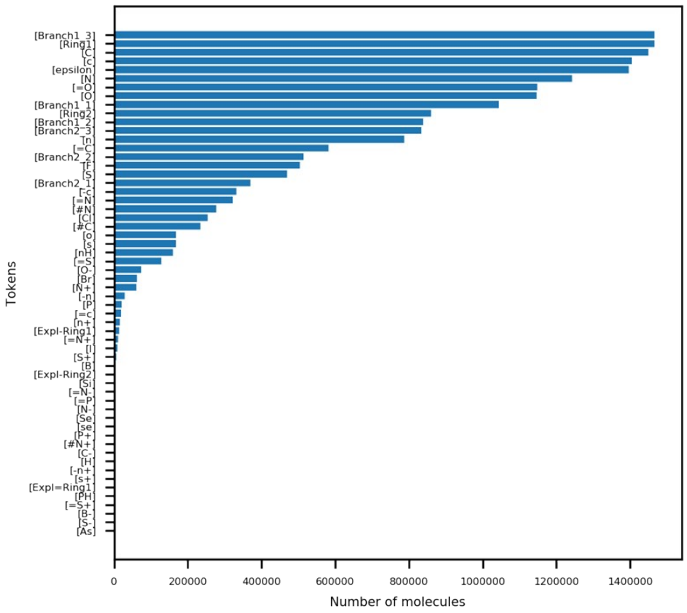


Supplementary Figure 8. Additional examples of molecular embedding interpolation and its effect on the condition match. The condition network is used to compute the matching probability between the median overexpression profile for a given gene and selected molecular embeddings along a linear interpolation trajectory between a start and end molecule. **a** Matching

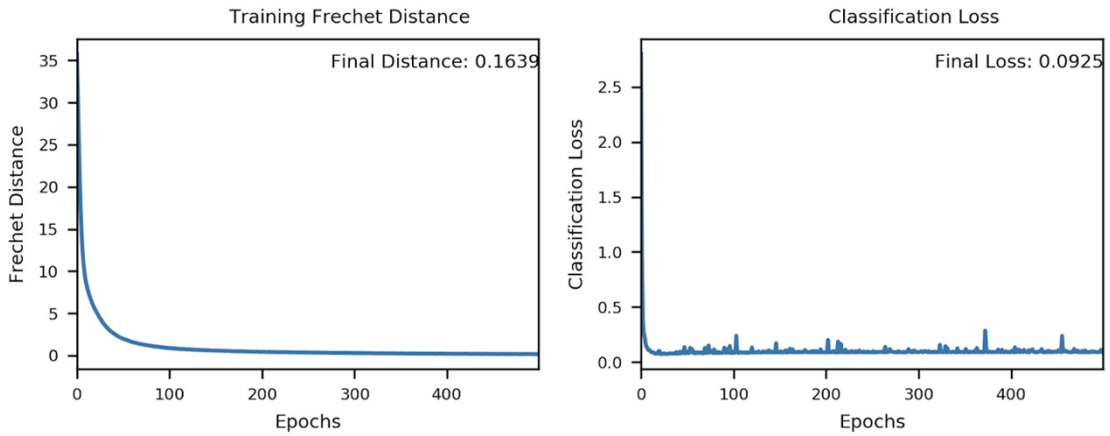
against NFKB1 profile. The start molecule is the ExCAPE agonist with a significant enriched scaffold for NFKB1. End molecules are other NFKB1 agonists from ExCAPE with higher matching probability than the start molecule (top-left) or random generated molecules with lower matching probability (top-right and bottom). **b** Matching against BRCA1 profile. The start molecule is the ExCAPE agonist with a significant enriched scaffold for BRCA1. End molecules are other BRCA1 agonists from ExCAPE with higher matching probability than the start molecule (top row) or random generated molecules with lower probability (bottom row).



Supplementary Figure 9. Additional examples of interpolation of the morphological space and its effect on the condition match. The condition network is used to compute the matching probability between the molecular embedding from a selected compound and morphological profiles along linear interpolation trajectories. **a** Interpolation curves between DMSO and NFKB1, 100 random directions, and the top 10 axes of variation determined by PCA. **b** Interpolation curves between DMSO and BRCA1. Continuous lines and shadows report mean and std, respectively. Selected compounds correspond to those used in Supplementary Figure 6 for chemical interpolation.



Supplementary Figure 10. **SELFIES Tokens included in the model**. Tokens that appear in at least 100 molecules from the 1.5 million contained in ChEMBL22.



Supplementary Figure 11. Monitoring model performance during adversarial training. Left: A molecular embedding was generated for each morphological profile in the training set at the end of each epoch. Embeddings were used to evaluate the similarity between the generated and real molecular representations using Fréchet distance. Right: classification loss of the pre-trained condition network after each epoch during adversarial training.

Supplementary Tables

Gene	ExCAPE		Generated				ExCAPE & Generated
	unique cpds	Unique scaff	unique cpds	% unique cpds	unique scaff	% unique scaff	Common scaff
TP53	12448	7252	19377 ± 14	96	15811 ± 39	81	234 ± 6
BRCA1	7886	4410	18641 ± 14	93	14265 ± 21	76	210 ± 5
HSPA5	591	515	19076 ± 14	95	16126 ± 38	84	47 ± 7
NFKB1	185	145	19094 ± 46	95	12355 ± 84	64	33 ± 2
CREBBP	80	72	19667 ± 22	98	17495 ± 74	88	0 ± 1
STAT1	78	70	19447 ± 36	97	16216 ± 18	83	14 ± 1
STAT3	63	52	19872 ± 12	99	16333 ± 73	82	10 ± 1
HIF1A	49	34	19812 ± 23	99	15538 ± 92	78	1 ± 0
NFKBIA	29	3	19119 ± 46	95	14983 ± 98	78	0 ± 0

Supplementary Table 1. Scaffold diversity of ExCAPE agonists and generated molecules conditioned on overexpressed genes. All genes from the overexpression dataset with at least 10 known agonists are reported. Values report the mean ± standard deviation among 3 inference repetitions, each generating 20,000 valid molecules passing custom physicochemical filters (see methods). cpds: compounds, scaff: Murcko scaffolds.