ISCI, Volume *7*

Supplemental Information

A Large-Scale Gene Expression Intensity-Based Similarity Metric for Drug Repositioning

Chen-Tsung Huang, Chiao-Hui Hsieh, Yen-Jen Oyang, Hsuan-Cheng Huang, and Hsueh-Fen Juan

Supplemental Figures

Figure S1. Frequencies of primary MoAs of LINCS compounds, related to Figure 1.
See Table S2 for detailed descriptions.

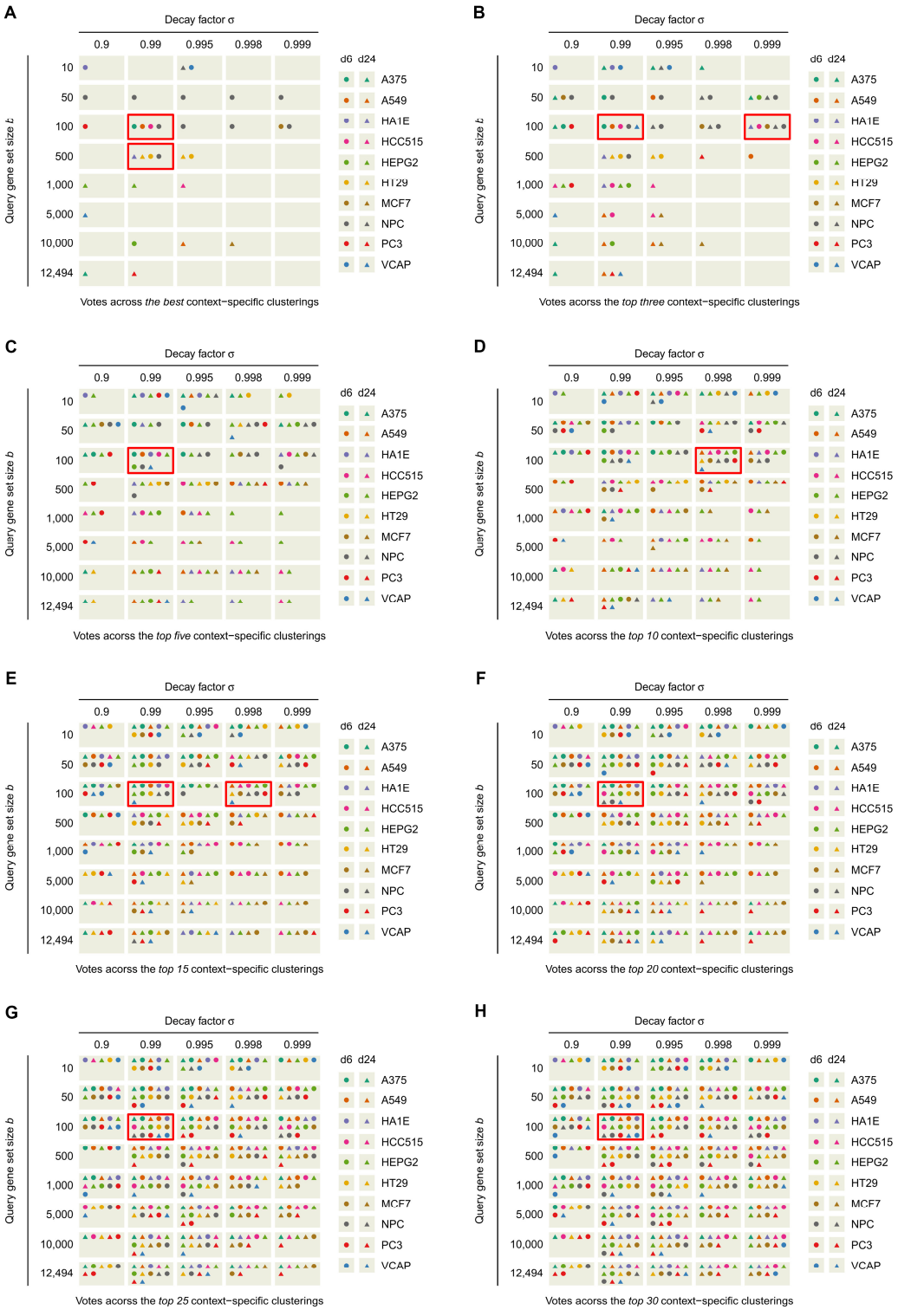

Figure S2. Votes for the best parameter set for the intensity-based metric across context-specific clusterings, related to Figure 2.

Shown are the votes among the best (A), top three (B), top five (C), top 10 (D), top 15 (E), top 20 (F), top 25 (G), or top 30 (H) context-specific intensity-based clusterings across all contexts, in which the parameter set (or sets) that received the most votes was boxed by a red rectangle.

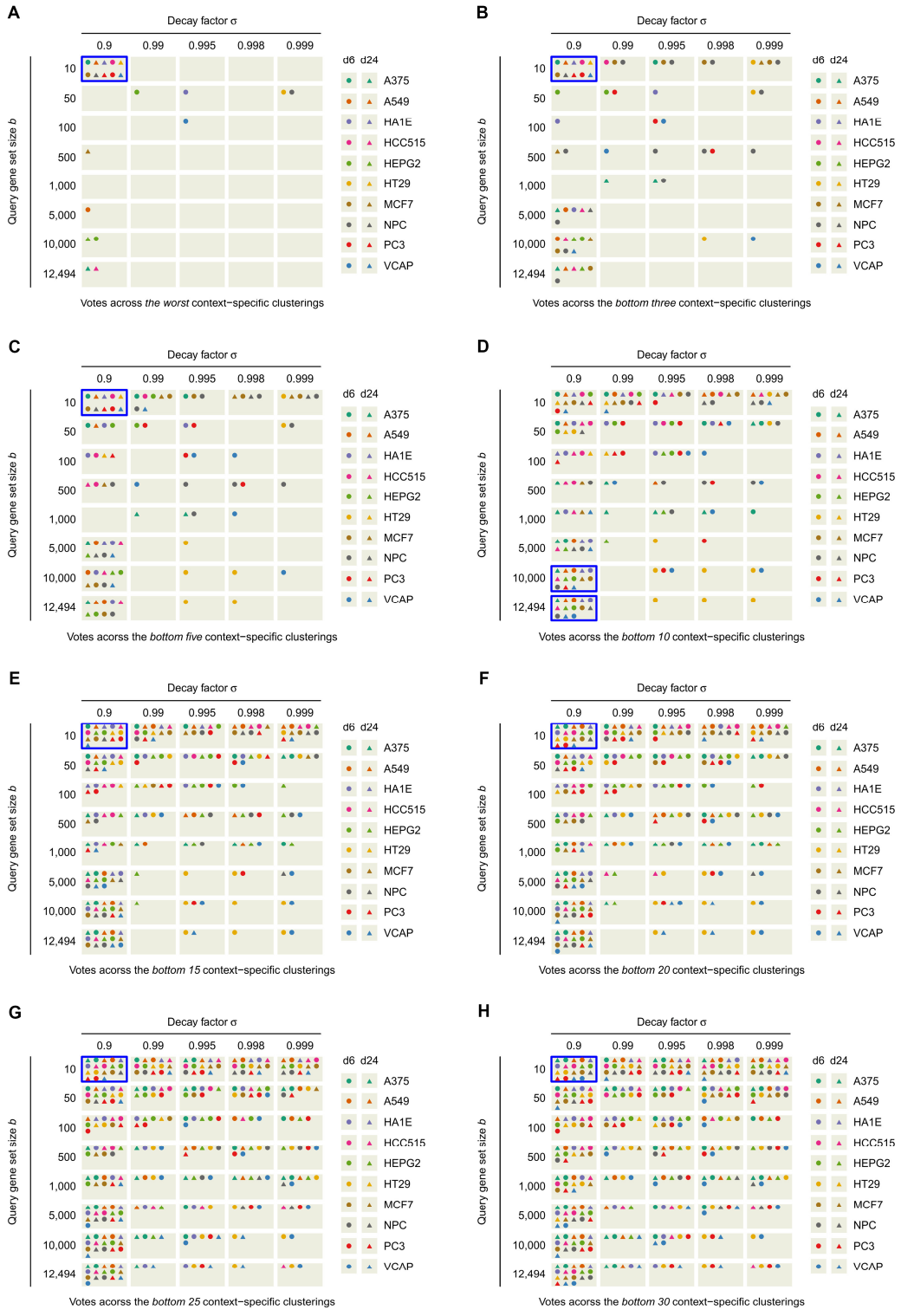

Figure S3. Votes for the worst parameter set for the intensity-based metric across context-specific clusterings, related to Figure 2.

Shown are the votes among the worst (A), bottom three (B), bottom five (C), bottom 10 (D), bottom 15 (E), bottom 20 (F), bottom 25 (G), or bottom 30 (H) context-specific intensity-based clusterings across all contexts, in which the parameter set (or sets) that received the most votes was boxed by a blue rectangle.

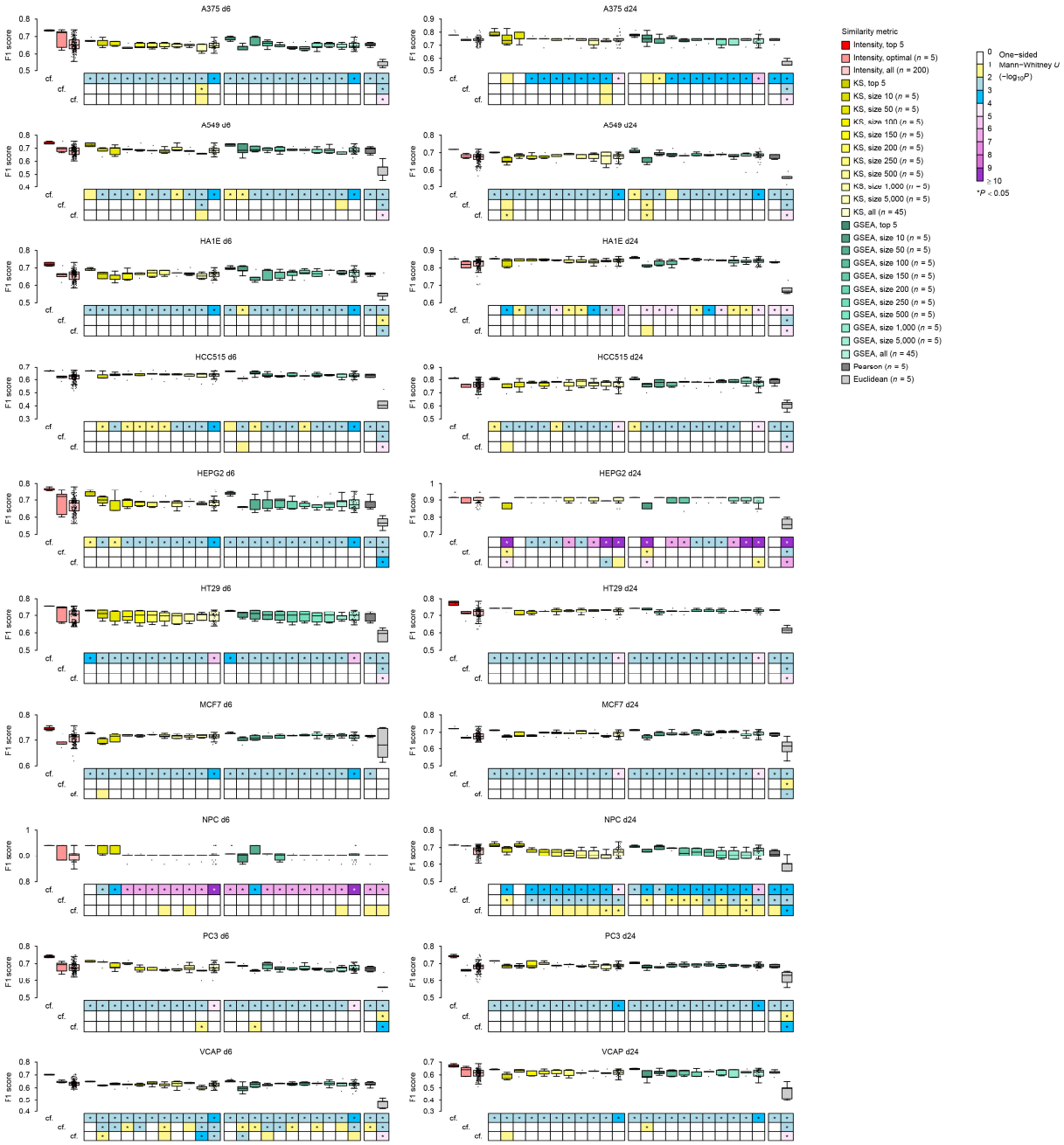

Figure S4. Comparison of F1 scores of intensity-based and common clusterings for each context, related to Figure 2.

With five biomedical clustering methods, F1 scores of intensity-based clusterings within the top five, using the optimal parameter set (n = 5), or across all parameter sets (n = 200) were compared with KS or GSEA-based clusterings within the top five, using a fixed query gene set size (n = 5), or across all gene set sizes considered (n = 45) and compared with common clusterings using the Pearson (n = 5) or Euclidean (n = 5) metric by one-sided Mann–Whitney U tests (with $-\log_{10}$(P-value) represented by the corresponding colored box at the bottom of each plot). Box plots depict the interquartile range (IQR) and whiskers represent 1.5× IQR. *P < 0.05.

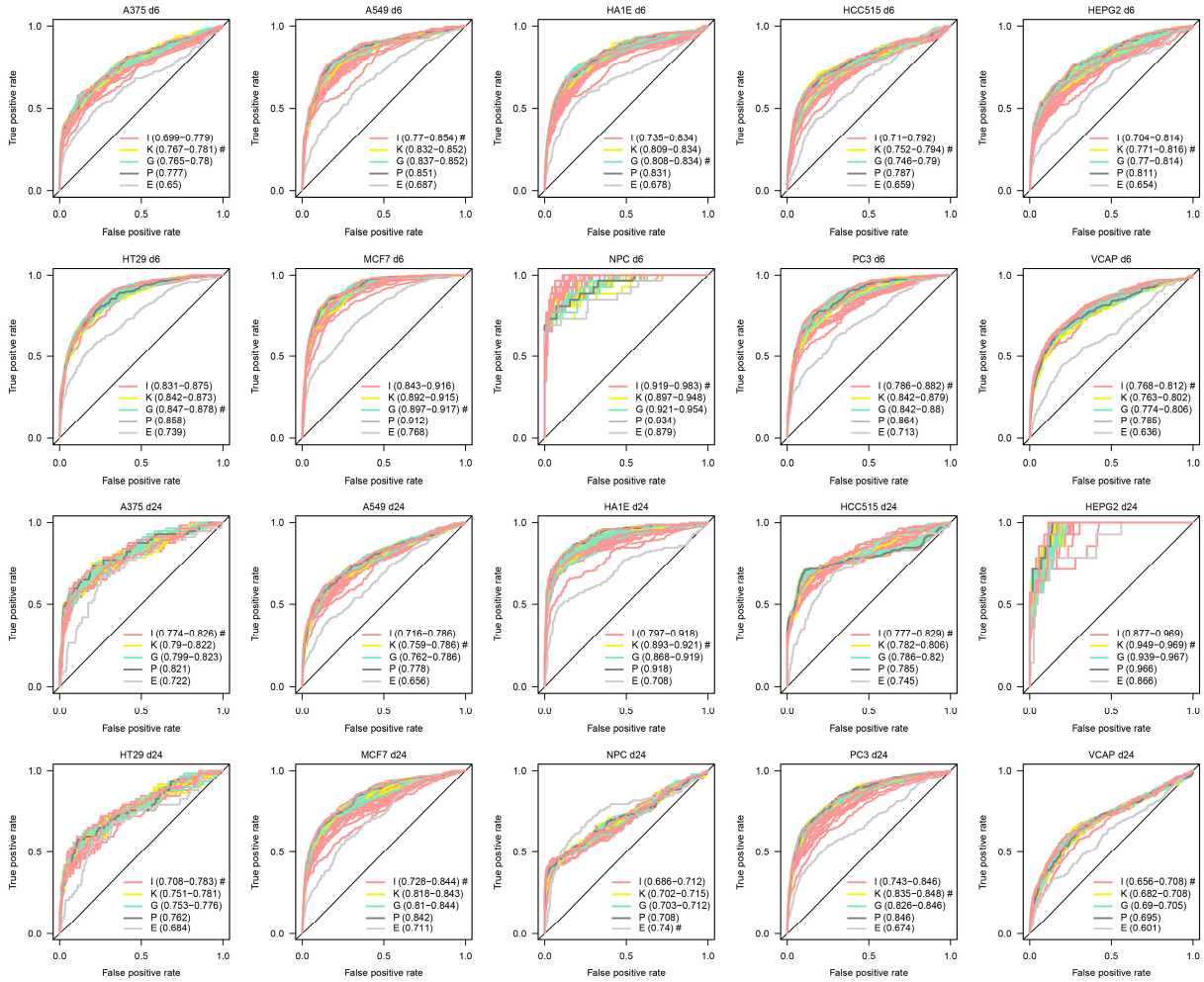

Figure S5. Comparison of AUROC performance among intensity-based and common metrics with respect to the gold-standard clustering, related to Figure 2.

For each context, an ROC curve was drawn using the intensity-based metric (I; light red lines) with all other parameter sets, KS (K; yellow lines) or GSEA (G; aquamarine lines) metric with all query gene set sizes considered, Pearson (P; dark gray line) metric, or Euclidean (E: light gray line) metric accompanied by a range of AUROC values in each parentheses, and the metric that achieved the best AUROC value was indicated by a hash (#).

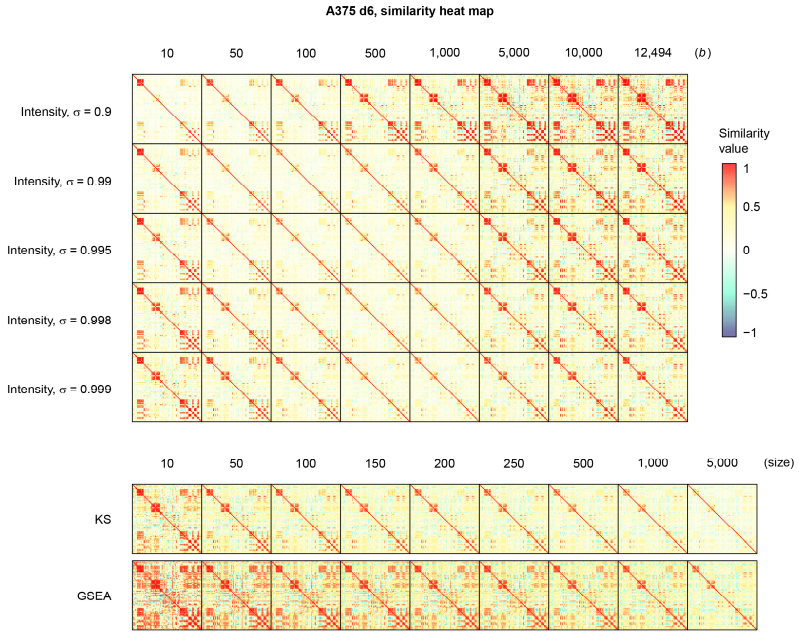

Figure S6. Similarity heat map of the gold-standard drugs for each intensity-based metric with a fixed parameter set (*b* and *σ*) and for each KS or GSEA metric with a fixed gene set size in the context of d6 perturbation in A375 cells, related to Figure 2.

For each similarity heat map, the drugs were ordered with MoAs being the same as in Figure 2A.

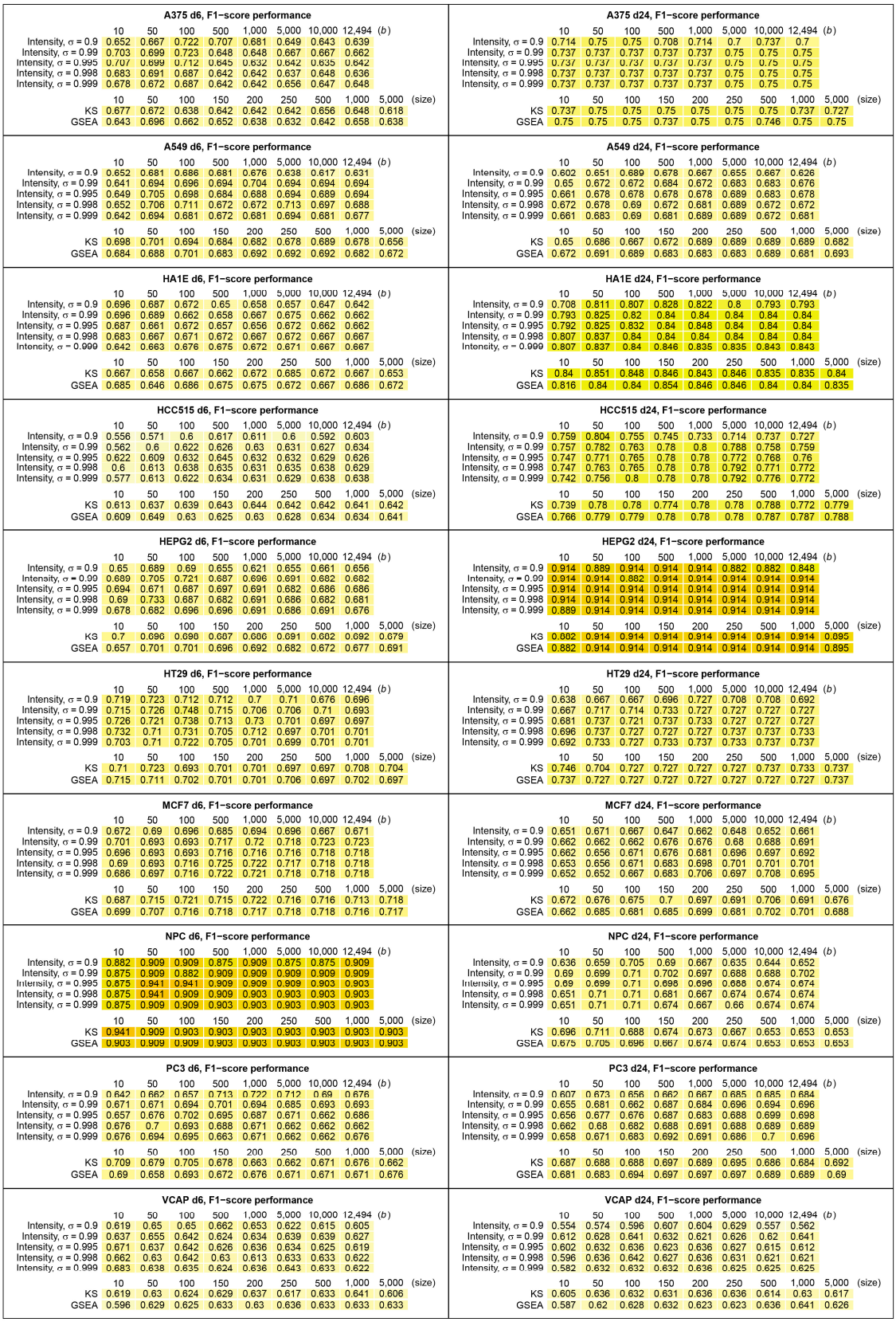

Figure S7. Robustness analysis of F1-score performance of the intensity-based, KS, and GSEA metrics to the gene set size, related to Figure 2.

For each context, the value in each cell represents the median F1 score among the clusterings produced by the five clustering methods for a given metric (an intensity-based metric with fixed *b* and *σ*, or a KS or GSEA metric with a fixed set size).

A375 d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.699	0.725	0.736	0.754	0.753	0.742	0.74	0.737	
Intensity, σ = 0.99	0.733	0.75	0.748	0.768	0.768	0.769	0.77	0.769	
Intensity, σ = 0.995	0.743	0.755	0.753	0.772	0.772	0.774	0.775	0.775	
Intensity, σ = 0.998	0.748	0.758	0.758	0.776	0.775	0.777	0.779	0.779	
Intensity, σ = 0.999	0.747	0.758	0.76	0.777	0.777	0.778	0.779	0.778	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.767	0.77	0.772	0.78	0.779	0.781	0.78	0.78	0.77	
GSEA	0.768	0.765	0.769	0.776	0.777	0.78	0.78	0.779	0.775	

A375 d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.826	0.791	0.804	0.789	0.808	0.794	0.803	0.803	
Intensity, σ = 0.99	0.814	0.774	0.787	0.78	0.797	0.812	0.816	0.816	
Intensity, σ = 0.995	0.813	0.778	0.787	0.784	0.799	0.815	0.822	0.821	
Intensity, σ = 0.998	0.81	0.784	0.79	0.792	0.804	0.821	0.824	0.825	
Intensity, σ = 0.999	0.807	0.791	0.793	0.799	0.81	0.822	0.825	0.825	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.79	0.801	0.813	0.811	0.817	0.815	0.814	0.82	0.822	
GSEA	0.799	0.809	0.818	0.815	0.815	0.813	0.817	0.823	0.821	

A549 d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.77	0.805	0.82	0.835	0.834	0.826	0.827	0.825	
Intensity, σ = 0.99	0.811	0.83	0.839	0.849	0.849	0.848	0.847	0.848	
Intensity, σ = 0.995	0.82	0.835	0.843	0.852	0.852	0.85	0.849	0.849	
Intensity, σ = 0.998	0.826	0.837	0.846	0.854	0.854	0.851	0.85	0.85	
Intensity, σ = 0.999	0.827	0.838	0.847	0.854	0.854	0.851	0.85	0.85	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.832	0.85	0.848	0.851	0.851	0.851	0.852	0.851	0.84	
GSEA	0.837	0.851	0.85	0.851	0.852	0.852	0.852	0.852	0.846	

A549 d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.716	0.742	0.755	0.764	0.769	0.761	0.75	0.747	
Intensity, σ = 0.99	0.748	0.76	0.767	0.78	0.783	0.779	0.779	0.779	
Intensity, σ = 0.995	0.755	0.764	0.772	0.784	0.785	0.781	0.78	0.781	
Intensity, σ = 0.998	0.76	0.768	0.777	0.786	0.784	0.78	0.779	0.78	
Intensity, σ = 0.999	0.763	0.771	0.779	0.786	0.783	0.778	0.778	0.778	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.759	0.778	0.784	0.785	0.786	0.785	0.783	0.78	0.764	
GSEA	0.762	0.782	0.786	0.785	0.785	0.785	0.783	0.782	0.774	

HA1E d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.735	0.774	0.78	0.799	0.797	0.793	0.793	0.79	
Intensity, σ = 0.99	0.763	0.808	0.815	0.827	0.828	0.83	0.829	0.828	
Intensity, σ = 0.995	0.771	0.816	0.82	0.828	0.83	0.832	0.831	0.83	
Intensity, σ = 0.998	0.779	0.822	0.825	0.83	0.832	0.834	0.831	0.83	
Intensity, σ = 0.999	0.783	0.825	0.827	0.832	0.833	0.833	0.831	0.83	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.809	0.831	0.829	0.828	0.828	0.832	0.833	0.834	0.82	
GSEA	0.808	0.828	0.83	0.828	0.828	0.831	0.831	0.834	0.827	

HA1E d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.797	0.845	0.86	0.875	0.882	0.883	0.879	0.879	
Intensity, σ = 0.99	0.859	0.874	0.884	0.897	0.903	0.904	0.904	0.905	
Intensity, σ = 0.995	0.87	0.88	0.89	0.902	0.907	0.908	0.909	0.91	
Intensity, σ = 0.998	0.88	0.887	0.895	0.907	0.912	0.914	0.915	0.916	
Intensity, σ = 0.999	0.883	0.891	0.898	0.909	0.915	0.917	0.918	0.918	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.893	0.907	0.912	0.913	0.914	0.914	0.919	0.921	0.916	
GSEA	0.868	0.897	0.904	0.908	0.909	0.911	0.916	0.919	0.919	

HCC515 d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.71	0.735	0.728	0.733	0.729	0.724	0.72	0.718	
Intensity, σ = 0.99	0.726	0.756	0.763	0.773	0.774	0.78	0.782	0.783	
Intensity, σ = 0.995	0.728	0.758	0.767	0.778	0.781	0.787	0.789	0.79	
Intensity, σ = 0.998	0.729	0.759	0.77	0.784	0.786	0.79	0.791	0.792	
Intensity, σ = 0.999	0.728	0.759	0.773	0.786	0.788	0.789	0.79	0.79	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.752	0.784	0.784	0.788	0.794	0.794	0.792	0.786	0.78	
GSEA	0.746	0.782	0.781	0.784	0.788	0.788	0.79	0.788	0.787	

HCC515 d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.829	0.829	0.822	0.804	0.794	0.785	0.785	0.785	
Intensity, σ = 0.99	0.821	0.797	0.789	0.778	0.777	0.784	0.784	0.784	
Intensity, σ = 0.995	0.815	0.789	0.783	0.777	0.777	0.785	0.785	0.785	
Intensity, σ = 0.998	0.809	0.783	0.779	0.777	0.78	0.786	0.786	0.785	
Intensity, σ = 0.999	0.806	0.779	0.778	0.779	0.782	0.787	0.786	0.785	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.806	0.793	0.79	0.792	0.791	0.791	0.791	0.789	0.782	
GSEA	0.82	0.794	0.789	0.788	0.788	0.789	0.79	0.789	0.786	

HEPG2 d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.704	0.728	0.733	0.756	0.759	0.764	0.762	0.762	
Intensity, σ = 0.99	0.731	0.759	0.773	0.793	0.795	0.8	0.8	0.8	
Intensity, σ = 0.995	0.739	0.766	0.781	0.801	0.803	0.807	0.808	0.808	
Intensity, σ = 0.998	0.747	0.773	0.788	0.808	0.81	0.813	0.813	0.813	
Intensity, σ = 0.999	0.751	0.776	0.791	0.81	0.813	0.814	0.814	0.814	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.771	0.797	0.805	0.811	0.814	0.814	0.816	0.813	0.81	
GSEA	0.77	0.787	0.797	0.807	0.812	0.814	0.814	0.812	0.812	

HEPG2 d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.877	0.929	0.941	0.95	0.95	0.943	0.939	0.95	
Intensity, σ = 0.99	0.928	0.966	0.953	0.955	0.957	0.953	0.954	0.955	
Intensity, σ = 0.995	0.937	0.968	0.951	0.957	0.96	0.956	0.955	0.956	
Intensity, σ = 0.998	0.94	0.969	0.95	0.959	0.96	0.963	0.963	0.962	
Intensity, σ = 0.999	0.939	0.967	0.948	0.959	0.96	0.966	0.967	0.967	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.954	0.949	0.952	0.96	0.963	0.959	0.963	0.969	0.965	
GSEA	0.947	0.941	0.939	0.949	0.953	0.955	0.959	0.965	0.967	

HT29 d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.831	0.852	0.863	0.872	0.871	0.863	0.863	0.867	
Intensity, σ = 0.99	0.855	0.868	0.871	0.874	0.875	0.87	0.87	0.873	
Intensity, σ = 0.995	0.861	0.872	0.873	0.873	0.873	0.867	0.866	0.868	
Intensity, σ = 0.998	0.865	0.875	0.874	0.871	0.869	0.861	0.86	0.861	
Intensity, σ = 0.999	0.866	0.875	0.874	0.869	0.865	0.856	0.855	0.856	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.859	0.873	0.873	0.87	0.867	0.863	0.859	0.854	0.842	
GSEA	0.862	0.878	0.875	0.871	0.87	0.868	0.863	0.858	0.847	

HT29 d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.708	0.751	0.747	0.751	0.749	0.754	0.761	0.767	
Intensity, σ = 0.99	0.745	0.761	0.762	0.769	0.767	0.779	0.778	0.783	
Intensity, σ = 0.995	0.748	0.763	0.763	0.768	0.768	0.777	0.775	0.777	
Intensity, σ = 0.998	0.747	0.768	0.766	0.767	0.766	0.77	0.769	0.77	
Intensity, σ = 0.999	0.745	0.77	0.769	0.767	0.764	0.764	0.762	0.763	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.751	0.779	0.781	0.772	0.771	0.764	0.765	0.756	0.754	
GSEA	0.753	0.775	0.776	0.773	0.771	0.771	0.767	0.762	0.754	

MCF7 d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.843	0.873	0.884	0.9	0.905	0.897	0.887	0.888	
Intensity, σ = 0.99	0.876	0.895	0.9	0.914	0.916	0.915	0.914	0.914	
Intensity, σ = 0.995	0.883	0.902	0.906	0.916	0.916	0.915	0.914	0.914	
Intensity, σ = 0.998	0.889	0.908	0.912	0.916	0.915	0.913	0.912	0.912	
Intensity, σ = 0.999	0.89	0.911	0.914	0.916	0.914	0.911	0.911	0.911	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.892	0.914	0.915	0.915	0.915	0.914	0.911	0.91	0.906	
GSEA	0.897	0.915	0.916	0.916	0.917	0.916	0.913	0.912	0.908	

MCF7 d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.728	0.757	0.772	0.811	0.819	0.835	0.836	0.833	
Intensity, σ = 0.99	0.769	0.779	0.789	0.818	0.826	0.841	0.842	0.842	
Intensity, σ = 0.995	0.786	0.791	0.799	0.823	0.83	0.842	0.844	0.844	
Intensity, σ = 0.998	0.805	0.805	0.812	0.83	0.835	0.843	0.844	0.844	
Intensity, σ = 0.999	0.817	0.814	0.821	0.834	0.838	0.843	0.844	0.844	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.818	0.831	0.837	0.84	0.841	0.843	0.842	0.842	0.838	
GSEA	0.81	0.83	0.837	0.839	0.84	0.843	0.843	0.844	0.844	

NPC d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.919	0.965	0.961	0.978	0.983	0.98	0.973	0.972	
Intensity, σ = 0.99	0.955	0.973	0.965	0.971	0.966	0.951	0.948	0.948	
Intensity, σ = 0.995	0.96	0.976	0.967	0.965	0.96	0.942	0.941	0.942	
Intensity, σ = 0.998	0.965	0.976	0.967	0.958	0.949	0.933	0.932	0.935	
Intensity, σ = 0.999	0.969	0.973	0.964	0.953	0.944	0.927	0.928	0.928	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.946	0.948	0.938	0.936	0.932	0.931	0.929	0.925	0.897	
GSEA	0.928	0.954	0.944	0.945	0.945	0.942	0.939	0.932	0.921	

NPC d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.7	0.694	0.693	0.7	0.701	0.695	0.689	0.686	
Intensity, σ = 0.99	0.693	0.696	0.702	0.711	0.71	0.711	0.71	0.71	
Intensity, σ = 0.995	0.695	0.699	0.705	0.711	0.71	0.71	0.709	0.71	
Intensity, σ = 0.998	0.696	0.703	0.71	0.71	0.708	0.709	0.709	0.71	
Intensity, σ = 0.999	0.697	0.706	0.712	0.709	0.708	0.708	0.709	0.71	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.709	0.715	0.713	0.711	0.71	0.709	0.708	0.706	0.702	
GSEA	0.703	0.712	0.709	0.708	0.71	0.71	0.71	0.709	0.706	

PC3 d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.786	0.809	0.811	0.835	0.835	0.841	0.837	0.833	
Intensity, σ = 0.99	0.805	0.848	0.856	0.879	0.878	0.876	0.873	0.871	
Intensity, σ = 0.995	0.808	0.858	0.865	0.882	0.88	0.874	0.872	0.87	
Intensity, σ = 0.998	0.811	0.866	0.871	0.881	0.877	0.868	0.866	0.865	
Intensity, σ = 0.999	0.814	0.869	0.874	0.878	0.873	0.863	0.861	0.861	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.842	0.877	0.879	0.875	0.872	0.87	0.865	0.859	0.842	
GSEA	0.842	0.873	0.88	0.875	0.875	0.873	0.871	0.866	0.852	

PC3 d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.743	0.774	0.79	0.81	0.819	0.82	0.821	0.818	
Intensity, σ = 0.99	0.786	0.816	0.82	0.836	0.842	0.844	0.842	0.841	
Intensity, σ = 0.995	0.796	0.824	0.825	0.839	0.844	0.845	0.842	0.841	
Intensity, σ = 0.998	0.804	0.832	0.832	0.843	0.845	0.845	0.843	0.842	
Intensity, σ = 0.999	0.808	0.836	0.836	0.844	0.846	0.846	0.844	0.843	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.835	0.843	0.846	0.846	0.848	0.848	0.846	0.846	0.838	
GSEA	0.826	0.842	0.843	0.844	0.846	0.846	0.846	0.846	0.843	

VCAP d6, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.768	0.785	0.793	0.807	0.803	0.799	0.79	0.79	
Intensity, σ = 0.99	0.791	0.803	0.807	0.811	0.808	0.8	0.794	0.794	
Intensity, σ = 0.995	0.796	0.808	0.81	0.808	0.804	0.795	0.79	0.79	
Intensity, σ = 0.998	0.798	0.812	0.811	0.803	0.798	0.788	0.785	0.785	
Intensity, σ = 0.999	0.797	0.812	0.81	0.799	0.794	0.784	0.781	0.781	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.789	0.802	0.799	0.794	0.794	0.793	0.785	0.781	0.763	
GSEA	0.792	0.803	0.806	0.802	0.801	0.8	0.792	0.788	0.774	

VCAP d24, AUROC performance

	10	50	100	500	1,000	5,000	10,000	12,494	(b)
Intensity, σ = 0.9	0.656	0.684	0.698	0.698	0.703	0.698	0.691	0.695	
Intensity, σ = 0.99	0.689	0.701	0.703	0.704	0.708	0.703	0.701	0.702	
Intensity, σ = 0.995	0.693	0.703	0.703	0.703	0.706	0.7	0.698	0.699	
Intensity, σ = 0.998	0.696	0.707	0.706	0.703	0.705	0.696	0.694	0.695	
Intensity, σ = 0.999	0.697	0.708	0.708	0.703	0.702	0.693	0.691	0.692	

	10	50	100	150	200	250	500	1,000	5,000	(size)
KS	0.7	0.708	0.7	0.697	0.697	0.697	0.695	0.692	0.682	
GSEA	0.69	0.705	0.704	0.704	0.703	0.703	0.701	0.697	0.69	

Figure S8. Robustness analysis of AUROC performance of the intensity-based, KS, and GSEA metrics to the gene set size, related to Figure 2.

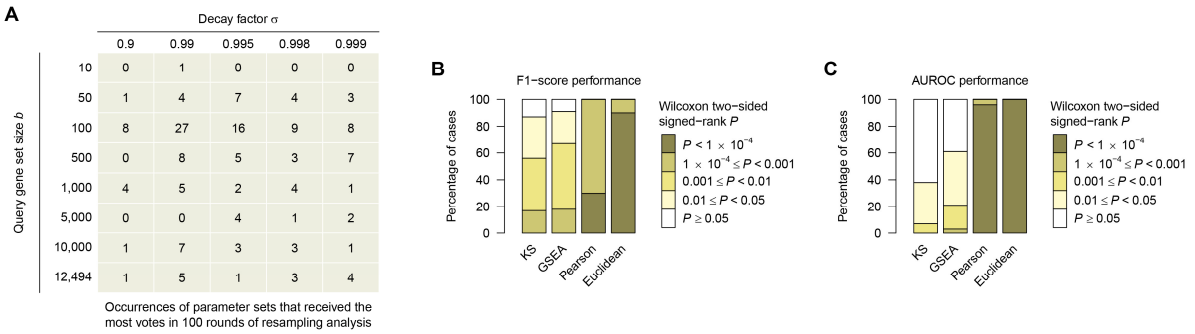

Figure S9. Resampling analysis for the parameter tuning and performance evaluation, related to Figure 2.

(A) Occurrences of parameter sets that received the most votes for the best set in a total of 100 rounds of the resampling analysis.

(B) Comparison of F1-score performance between the intensity-based metric with the KS, GSEA, Pearson-based, and Euclidean-based metrics in a total of 100 rounds of the resampling analysis (Wilcoxon two-sided paired signed-rank test).

(C) Comparison of AUROC performance between the intensity-based metric with the KS, GSEA, Pearson-based, and Euclidean-based metrics in a total of 100 rounds of the resampling analysis (Wilcoxon two-sided paired signed-rank test).

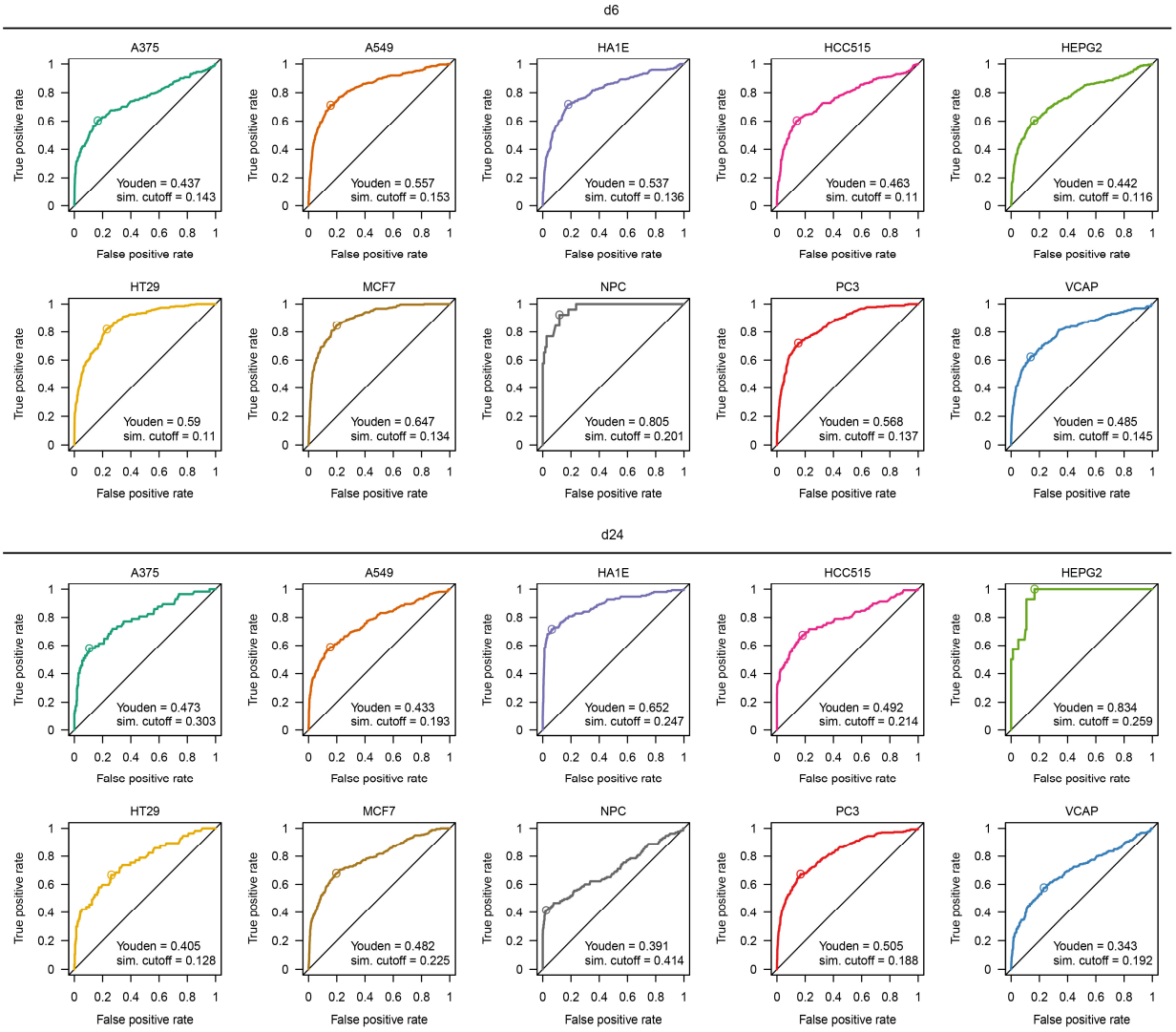

Figure S10. Similarity cutoffs defined by Youden's index using the optimal intensity-based similarity metric with respect to the gold-standard clustering, related to Figure 3.

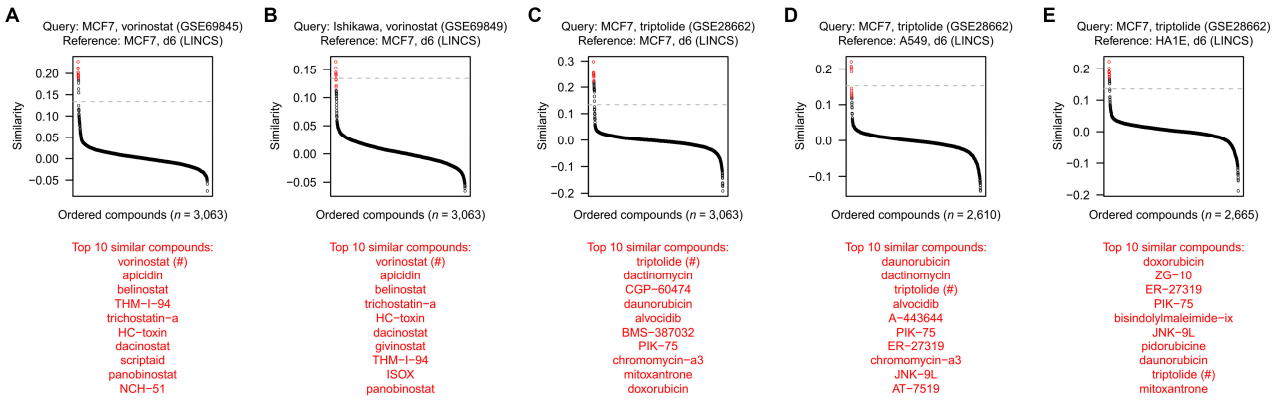

Figure S11. Validation of the intensity-based similarities using independently derived perturbation profiles to query the LINCS database, related to Figure 3.

Shown are the intensity-based similarities (using optimal intensity-based similarity metric) between a query chemical perturbation in a cell type (with a GEO accession number) and the reference chemical perturbations in a cell type (LINCS) as indicated. The top 10 similar reference compounds (red points) are listed below, with a hash (#) indicating the same reference compound as the query. For each plot, a horizontal dash gray line indicates the similarity cutoff defined by the Youden's index in the corresponding context as in Figure S10.

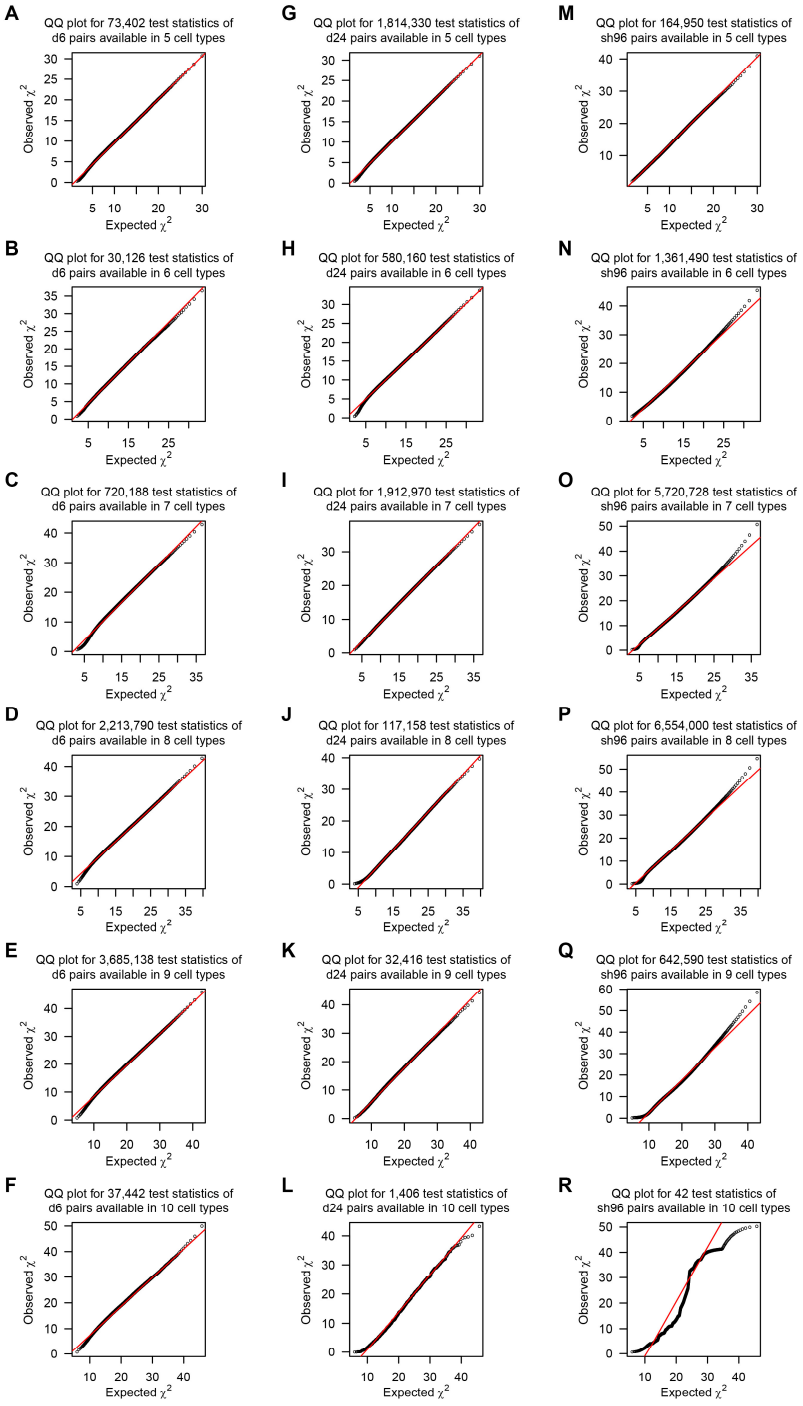

Figure S12. Quantile–quantile (Q–Q) plot for the chi-squared test statistics of the perturbation pairs used for RSS analysis, related to Figure 4.

For each of the d6 (A–F), d24 (G–L) and sh96 (M–R) perturbation types, we used the relative ranks of intensity-based similarities to generate a Q–Q plot for the chi-squared test statistics of the perturbation pairs available in n = 5 (A, G, M), 6 (B, H, N), 7 (C, I, O), 8 (D, J, P), 9 (E, K, Q), or 10 (F, L, R) cell types, respectively (corresponding to chi-squared distribution with 10, 12, 14, 16, 18, or 20 degrees of freedom, respectively). For each plot, we drew a red line that passes through the 100th and 900th permilles (i.e., 1000-quantiles).

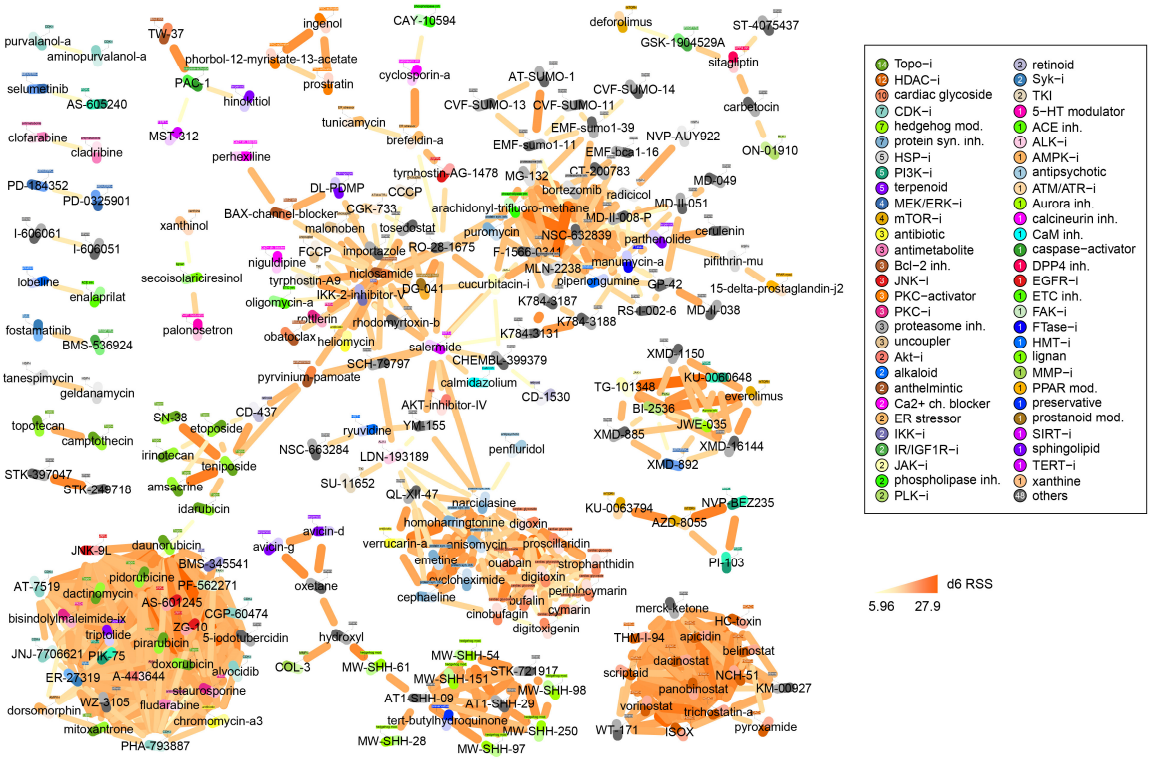

Figure S13. d6 RSS network, related to Figure 4.

Displayed are d6 chemical perturbation pairs that were recurrently similar across cell types (FDR < 0.001; Table S4; see Table S7 for network communities and their enrichment analyses). A primary MoA is assigned to each chemical perturbagen and color-coded on the darker side of each icon as well as in the text inset. The right box summarizes the occurrences of primary MoAs in the network.

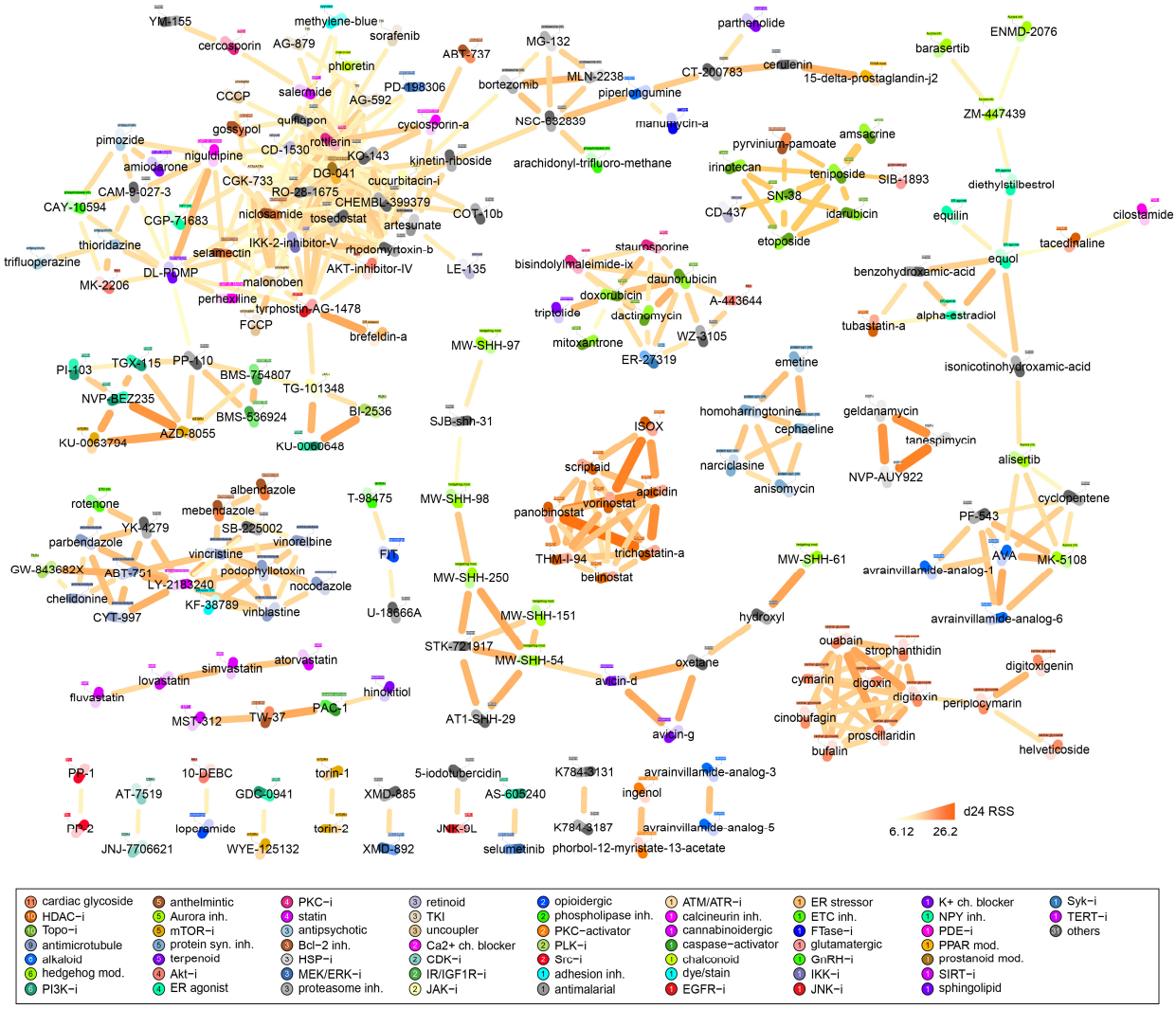

Figure S14. d24 RSS network, related to Figure 4.

Displayed are d24 chemical perturbation pairs that were recurrently similar across cell types (FDR < 0.001; Table S5; see Table S8 for network communities and their enrichment analyses). A primary MoA is assigned to each chemical perturbagen and color-coded on the darker side of each icon as well as in the text inset. The box below summarizes the occurrences of primary MoAs in the network.

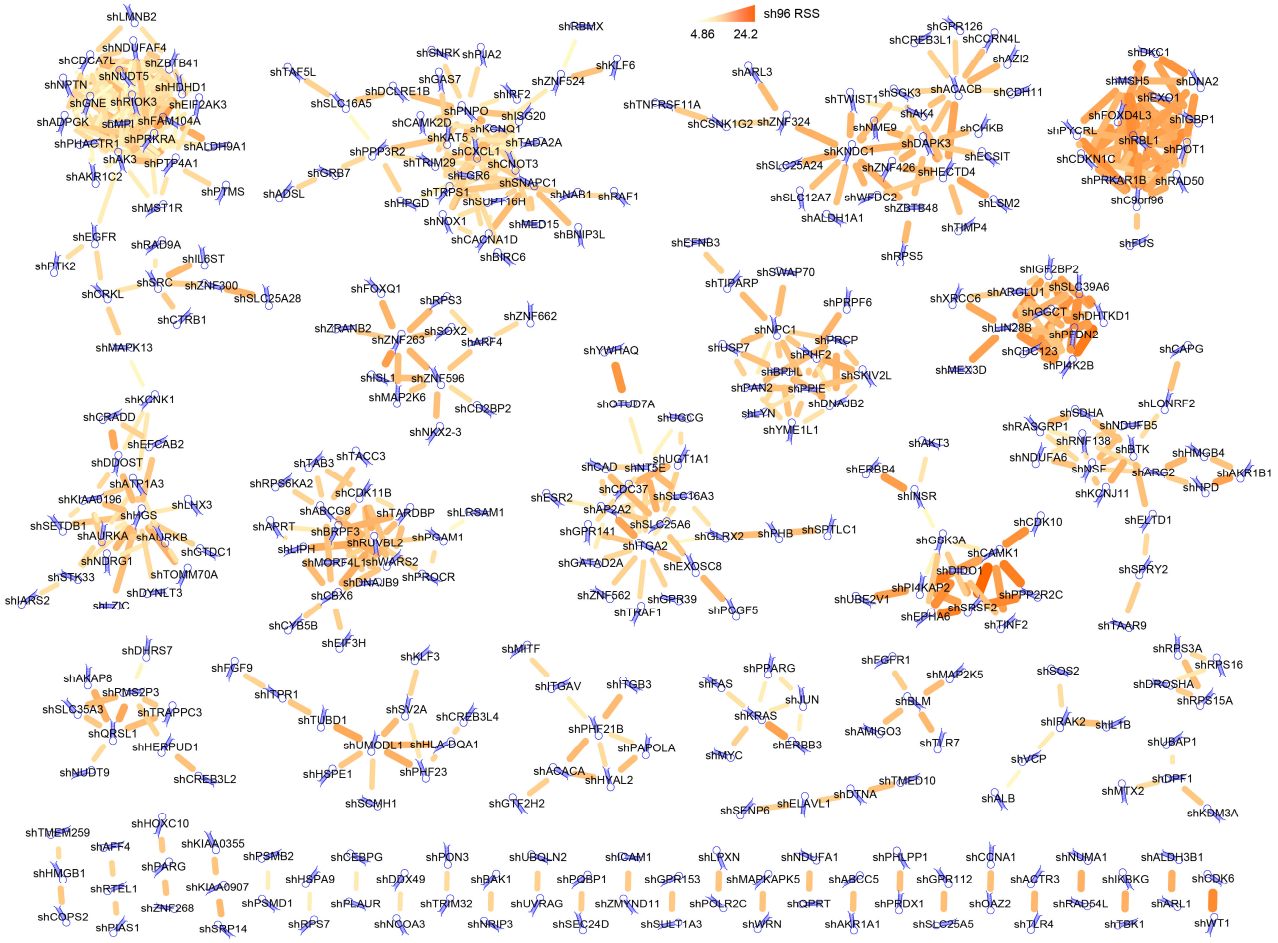

Figure S15. sh96 RSS network, related to Figure 4.

Displayed are sh96 genetic perturbation pairs that were recurrently similar across cell types (FDR < 0.001; Table S6; see Table S9 for network communities and their enrichment analyses).

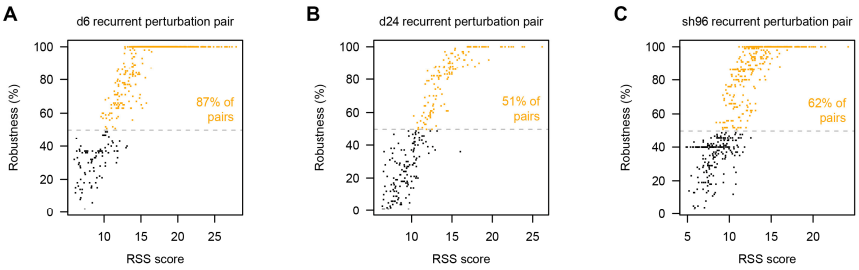

Figure S16. Resampling analysis for the recurrent perturbation pairs, related to Figure 4.
For recurrent perturbation pairs in each of the d6 (A), d24 (B), and sh96 (C) perturbation types, the percentages of cases in which they were recovered in the resampling analysis (as robustness; *y*-axis) were plotted against their real RSS scores (i.e., those from the RSS analysis when 100% of cell types were used; *x*-axis). For each plot, the robustness at 50% recovery is indicated by a horizontal dash gray line, above which the proportion of pairs (orange) is shown.

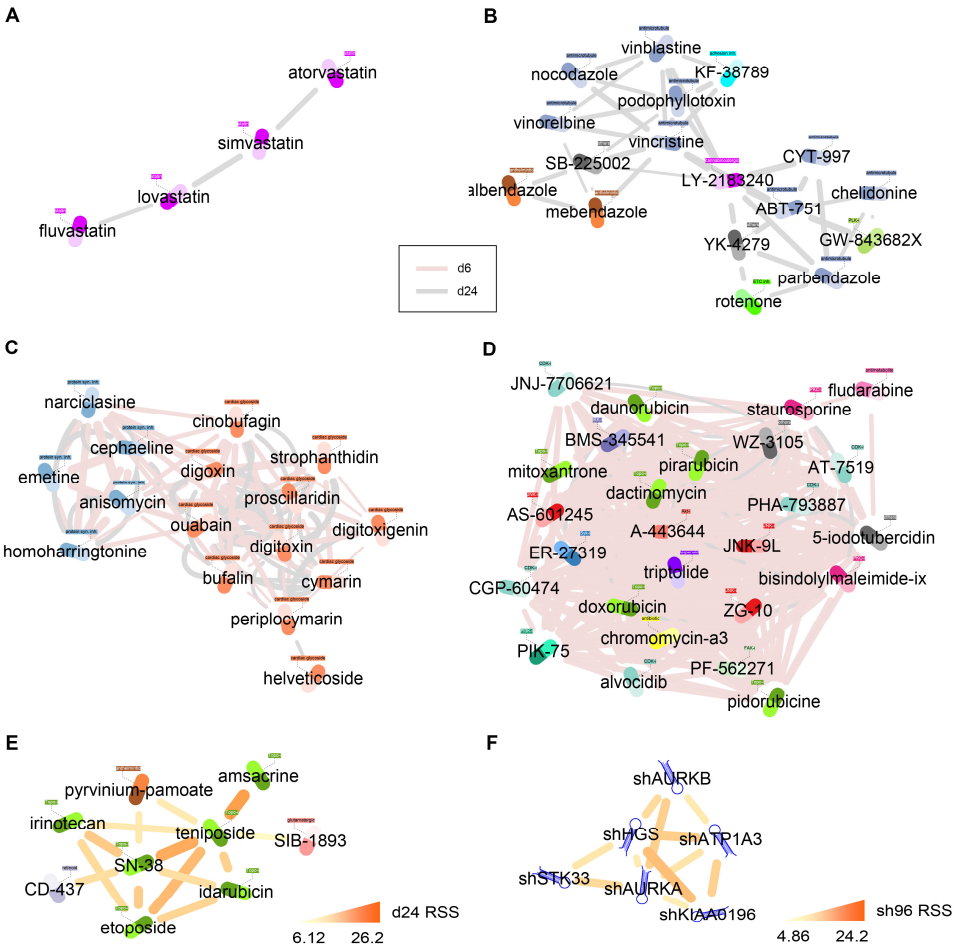

Figure S17. Subnetworks related to RSS analyses, related to Figure 4.

(A) A connected component involving lipid-lowering statins in the drug RSS network (merging d6 and d24 chemical perturbation networks).

(B) A connected component involving most antimicrotubules in the drug RSS network.

(C) A subnetwork spanning protein synthesis inhibitors and cardiac glycosides in the drug RSS network.

(D) A subnetwork of the PKC inhibitor staurosporine in the drug RSS network.

(E) A connected component involving the anthelmintic pyrvinium pamoate in the d24 RSS network.

(F) A subnetwork of the shAURKA perturbation in the sh96 RSS network.

Figure S18. Enrichments of hallmark gene sets for RSS network communities, related to Figure 4. Significant enrichments (corrected hypergeometric $P < 0.001$) of hallmark gene set collections from MSigDB for drug (A) and sh96 (B) RSS network communities (Tables S9 and S10). For each drug community, the number of drugs and the composition of MoAs are shown.

Figure S19. Enrichments of canonical pathways for RSS network communities, related to Figure 4. Significant enrichments (corrected hypergeometric $P < 0.001$) of canonical pathways from MSigDB for drug (A) and sh96 (B) RSS network communities (Tables S9 and S10). For each drug community, the number of drugs and the composition of MoAs are shown.

Figure S20. Enrichments of immunologic signatures for drug RSS network communities, related to Figure 4.
Shown are significant enrichments (corrected hypergeometric $P < 0.0001$) of immunologic signatures from MSigDB (Table S10) with the number of drugs and the composition of MoAs for each drug community indicated below.

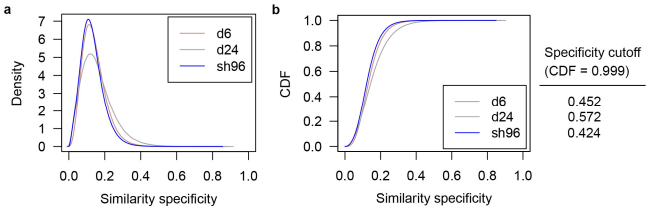

Figure S21. Similarity specificities of perturbation pairs, related to Figure 5.
Empirical density functions (A) and CDFs (B) against similarity specificities are shown for each perturbation type.

Transparent Methods

Data acquisition and preprocessing. We accessed LINCS (Subramanian et al., 2017) (Library of Integrated Network-based Cellular Signatures) L1000 gene expression datasets (level 3, normalized using invariant set scaling followed by quantile normalization; http://www.lincscloud.org/, now available at Gene Expression Omnibus; downloaded on 18 September 2014), which comprises 1,328,098 perturbation profiles for 77 cell types. All associated metadata were retrieved via Lincscloud API (http://api.lincscloud.org/, now replaced by https://clue.io/api) using R package 'rjson' (version 0.2.14). To facilitate similarity computations and to avoid redundancies, we reduced each probe-wise expression vector (dimension = 22,268; Affymetrix Human Genome U133A Array) to a gene-wise expression vector (dimension = 12,494) by taking the median of each gene with multiple probes in reference to R package 'hgu133a.db' (version 2.14.0).

We considered the following three types of perturbation: exposure to chemical drugs for 6 hours (abbreviated as d6); exposure to chemical drugs for 24 hours (d24); and exposure to short hairpin RNAs (shRNAs) for 96 hours (sh96). The following 10 selected cell types were analyzed: NPC (a human induced pluripotent stem cell-derived neural progenitor cell line); HA1E (a human kidney epithelial immortalized cell line); MCF7 (a human breast adenocarcinoma cell line); A549 and HCC515 (two human non-small cell lung adenocarcinoma cell lines); HT29 (a human colorectal adenocarcinoma cell line); HepG2 (a human hepatocellular carcinoma cell line); PC3 (a human prostate adenocarcinoma cell line); VCaP (a human metastatic prostate cancer cell line); and A375 (a human malignant melanoma cell line). All expression profiles were $\log_2$-transformed. The names of the perturbagens were set to those listed under the field 'pert_iname' in the metadata. For chemical perturbagens, we excluded names that were unclear in their meanings and might be unfamiliar to general users outside the Broad Institute; these names started with 'BRD-', 'SA-', 'BG-', 'ARG-', 'Broad-Sai-', 'FU-', 'JAS07-', 'KU-C', 'KUC', 'RAN-', 'SD-', 'ST-0', 'ST-2', 'ST-4', 'TL-', 'TUL', 'VU-', 'WZ-4-', 'DAC-', or 'WY-01'. The total number of non-redundant perturbagens analyzed in each cell type are summarized in Table S3, corresponding to a union of 3,332 chemical and 3,934 genetic perturbagens. We manually curated primary mechanisms of action (MoAs) for all chemical perturbagens with associated LINCS metadata and MoA tag descriptions (Figure S1 and Table S2).

Intensity-based similarity metric. We used perturbation intensity (i.e., the gene expression difference between perturbed and unperturbed conditions) and rank-based scoring functions with two tunable parameters to define an intensity-based similarity metric, inspired by the general observation from a biological perspective that a given perturbation should only affect a restricted panel of genes (Felix and Barkoulas, 2015). We therefore proposed that the most extremely perturbed genes compared to the unperturbed counterpart might be appropriate and sufficient to represent a given perturbation. For each perturbation u treated in a given cell type with N detection instances (one detection instance indicates one well of measurement in a 384-well plate for fluorescence detection

and the corresponding experiments may be performed in technical and biological replicates), we computed the 'perturbed profile' $\boldsymbol{d_u}$:

$$\boldsymbol{d_u} = \frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{r_{u,j}} - \boldsymbol{e_{u,j}}\right)$$

where $\boldsymbol{r}$ and $\boldsymbol{e}$ are expression vectors of the perturbation instance j and the corresponding control on the same plate of j, respectively. This perturbed profile represents the average perturbation intensity from available experiments in a way that reduces cross-plate batch effects, the extent of which became increasingly perceptible for some perturbations with multiple experimental replicates. Note that for most perturbations in a given cell type, experiments were performed in only one biological replicate with no more than three technical replicates, together with cross-plate batch effects, making it impractical to run standard differential expression analyses on LINCS data. Next, we derived an expression intensity rank vector $\boldsymbol{\rho_u}$, where the 1st most perturbed entry (regardless of the direction of gene regulation) is assigned 1, the 2nd most perturbed entry is assigned 2, and so on. This intensity rank allowed us to define a gene set size b that could be tuned to yield a query gene set B in which $|B| = b$. That is, the number of the top most perturbed genes for B could be fixed at $b = 10$, 50, 100, 500, 1,000, 5,000, 10,000, or 12,494 (the maximum size) with a corresponding query vector $\boldsymbol{\theta_u}$ for each sample u for which the selected entries (within B) are assigned 1 (up) or −1 (down) and the unselected (outside B) are assigned 0. Then, for any two perturbations u and v when v was used as a query, we computed the forward and reverse match scores $fwd_{u\leftarrow v}$ and $rev_{u\leftarrow v}$ (of v to u), respectively, as follows:

$$fwd_{u\leftarrow v} = \sum_{g\in G}\left[\!\left[sign\left(\boldsymbol{d_u}[g]\right) = \boldsymbol{\theta_v}[g] \right]\!\right] \times \sigma^{(\rho_u[g]-1)}$$

$$rev_{u\leftarrow v} = \sum_{g\in G}\left[\!\left[sign\left(\boldsymbol{d_u}[g]\right) = -\boldsymbol{\theta_v}[g] \right]\!\right] \times \sigma^{(\rho_u[g]-1)}$$

where G is the gene space ($|G| = 12{,}494$), $[\![\cdot]\!]$ returns 1 if the expression inside gives a TRUE value and returns 0 otherwise, $sgn(\cdot)$ is the sign function, $\boldsymbol{d_u}[g]$ is the entry of the perturbed profile $\boldsymbol{d_u}$ corresponding to the gene g, $\boldsymbol{\theta_v}[g]$ is the entry of the query vector $\boldsymbol{\theta_v}$ corresponding to the gene g, and the other tunable parameter decay factor σ to adjust the weights of ranks for σ fixed at 0.9, 0.99, 0.995, 0.998, or 0.999. Subsequently, we derived the directional similarity score of v to u as $sim_{u\leftarrow v} = (fwd_{u\leftarrow v} - rev_{u\leftarrow v})/S$, where S is the theoretical maximum match score (i.e., the sum of the geometric progression σ^0, σ^1,..., σ^{b-1} as the geometric series). Under this scenario, a +1 similarity score represents the perfect forward match whereas −1 represents the perfect reverse match. Similarly, we defined the

similarity score of u to v (when u was used as a query), where $sim_{u \leftarrow v} \neq sim_{v \leftarrow u}$. In this way, an undirected similarity score for any two perturbations v and u was obtained by averaging their directional similarity scores as $sim_{u \leftrightarrow v} = (sim_{u \leftarrow v} + sim_{v \leftarrow u})/2$.

Parameter optimization, comparison, and performance evaluation. We leveraged a ground truth clustering of a selected panel of chemical perturbagens with well-established primary MoAs ($n = 74$, but not all of these perturbagens were available in a given cell type at 6 or 24 h; thus, the exact number was variable; Table S1) to determine the optimal intensity-based similarity metric and facilitate the comparison among the intensity-based and all other similarity metrics.

For comparison, we considered four types of gene expression similarity metrics defined by the Kolmogorov–Smirnov (KS) statistic (Massey, 1951) (as the state of the art (Lamb et al., 2006; Iorio et al., 2010; Sirota et al., 2011; Kidd et al., 2016)), Gene Set Enrichment Analysis (GSEA) method (in which exponent = 1 was used for the standard enrichment score (Subramanian et al., 2005)), Pearson correlation, and Euclidean distance (two in common use (D'Haeseleer, 2005)). All these metrics were defined at the level of the 'perturbed' profiles $\boldsymbol{d_u}$ and $\boldsymbol{d_v}$ for any two perturbations u and v. For KS and GSEA metrics, we considered a range of set sizes at 10, 50, 100, 150, 200, 250, 500, 1,000 and 5,000 for each of the top- and bottom-ranked genes used for their matching algorithms; a metric was then defined by taking the difference of the matching scores between these two extremes ('top' minus 'bottom') over a value of 2 (the maximum of the difference when a perturbagen uses its ranked list to query itself, as did in those state-of-the-art studies (Iorio et al., 2010; Sirota et al., 2011; Kidd et al., 2016) except CMap (Lamb et al., 2006), which takes the largest value (≤ 2) among the absolute values of all possible differences produced when a gene signature is used to query the entire database).

For each similarity metric (intensity-based, KS, GSEA, Pearson, or Euclidean), the clustering performance was evaluated using the external cluster validity index F1 score based on a mapping approach to define true positives, true negatives, false positives, and false negatives (Wiwie et al., 2015). We explicitly considered the following five biomedical clustering methods: hierarchical clustering (HC) using average (HC-avg; using hclust in R base) or Ward (HC-ward; using agnes in R package cluster version 1.15.2) linkage; partitioning around medoids (PAM; using pam in R package cluster version 1.15.2); affinity propagation (AP; using apclusterK in R package apcluster version 1.4.3); and exemplar-based agglomerative clustering (EBC; using aggExCluster in R package apcluster version 1.4.3), with default function arguments. For each context (a perturbation type in a cell type) and each 'metric–method' setting (one metric plus one clustering method), we computed F1 scores for all possible clusters across a full range of number of clusters $k = 2,..., n - 1$ (where n is the number of drugs in the ground truth clustering) and determined an optimal clustering by picking the k with the best F1 score. We then prioritized these intensity-based clusters and all other clusters separately according to F1 scores and visualized these results with clusters of each clustering color-coded in a way that starting from the last row, the color consistency was maximized to the first row (i.e., the best of the intensity-based clusterings; Figure 2A provides an illustrative example of d6 perturbation

type in A375).

An optimal parameter set (b^*, σ^*) for the intensity-based similarity metric was determined by majority voting of occurrences of parameter sets among the top five intensity-based clusterings (including equal rank whenever possible) across 20 contexts (two perturbation types plus 10 cell types; Figure 2B). We also examined the voting results when varying the number of the top-scoring (Figure S2) or the bottom-scoring (Figure S3) intensity-based clusterings used for voting.

We first compared intensity-based clusterings with all other clusterings using these F1 scores. We extracted the highest F1 score for each context from clusterings using the intensity-based (across all parameter sets or using the optimal parameter set), KS or GSEA (across all gene set sizes or using a fixed gene set size), Pearson-based, or Euclidean-based metric and tested the statistical significance of their distributions across all contexts using Wilcoxon two-sided paired signed-rank tests (intensity-based over the others; Figure 2C). For each context, we also assessed whether the F1 scores of the intensity-based clusterings within the top five, using the optimal parameter set, or across all parameter sets were significantly higher than other clusterings using one-sided Mann–Whitney U tests (Figure S4). Next, we compared the intensity-based similarity metric with all other similarity metrics using a strategy based on the area under the receiver operating characteristic (ROC) curve (AUROC) with respect to the ground truth. For each context, we varied the threshold τ on a selected metric such that a perturbation pair with the similarity value beyond τ indicates that they are in the same cluster, whereas a value below τ not is in the same cluster; thus, we generated a ROC with an associated AUROC score. We aggregated the results using the optimal intensity-based metric across all contexts (Figure 2D) or each metric for each context (Figure S5, in which a hash (#) indicates the metric that achieved the best context-specific AUROC). We also extracted the highest AUROC score for each context achieved by the intensity-based, KS, GSEA, Pearson-based, or Euclidean-based metric and tested the statistical significance of their distributions across all contexts using Wilcoxon two-sided paired signed-rank tests (intensity-based over the others; Figure 2E).

For each context, we assessed the robustness of intensity-based, KS, and GSEA similarity metrics to the gene set size (i.e., the query gene set size b for the intensity-based metric with a fixed decay factor σ, or the set size for each of the top- and bottom-ranked genes for the KS or GSEA metric) based on the F1-score (Figure S7) and AUROC (Figure S8) performance. For the analysis using F1-score performance, we used the median F1 score among the clusterings produced by the five clustering methods for a given metric (an intensity-based metric with fixed b and σ, or a KS or GSEA metric with a fixed set size) in each context.

To examine whether the results of the analysis for parameter tuning and performance evaluation were robust against chemical perturbagens, we repeated the analysis with resampling of 75% of chemical perturbagens for 100 times. For parameter tuning for the intensity-based metric, in each round of the resampling analysis, votes for parameter sets were collected among the top five intensity-based clusterings across all contexts (as did in Figure 2B), and the parameter set that received the most votes was counted (more than one parameter set is possible if votes are equal in each round). The occurrence

of each parameter set that received the most votes in a total of 100 rounds of the resampling analysis is summarized in Figure S9A. For performance comparison, in each round of the resampling analysis, the best F1 scores for clusterings across all contexts (as did in Figure 2C) or the best AUROC scores across all contexts (as did in Figure 2E) achieved by the intensity-based metric were compared with those by the KS, GSEA, Pearson-based, or Euclidean-based metric using Wilcoxon two-sided paired signed-rank tests. The resulting *P*-values of the tests on the F1-score and AUROC performance from a total of 100 rounds of the resampling analysis are summarized in Figure S9B and S9C, respectively.

Similarity cutoffs. We used the optimal parameter set (b^*, σ^*) to compute pairwise intensity-based similarities for all available perturbations in LINCS. For each perturbation type (d6, d24, or sh96), we binned similarity values into bins at 0.001 resolution (i.e., rounded to the nearest three decimals) and plotted the empirical cumulative distribution function (CDF) of these bins within each cell type or across all cell types (Figure 3A). This overall CDF (across all cell types) was then used to derive a positive (sim^+_{d6}, sim^+_{d24}, or sim^+_{sh96}) and a negative (sim^-_{d6}, sim^-_{d24}, or sim^-_{sh96}) heuristic similarity cutoffs such that the positive value is the binned similarity value where CDF is greater than or approximately equal to 0.999 and the negative value is the binned similarity value where CDF is less than or approximately equal to 0.001. For each d6 or d24 perturbation type in a given cell type, we also determined a similarity cutoff defined by Youden's index (Youden, 1950) (i.e., sensitivity + specificity – 1) of the ROC curve by associating the intensity-based similarities with the ground truth clustering as a comparison to its corresponding positive heuristic similarity cutoff (Figure S10).

To explore the ability of the intensity-based similarity metric to recover small-molecule connections between LINCS and other resources, we computed the perturbation intensities of some chemical perturbagens independently derived from Gene Expression Omnibus (GEO) to query the LINCS database. We used the GEO datasets with the following accession numbers: GSE69845 (vorinostat at 6 h in MCF7 cells), GSE69849 (vorinostat at 6 h in Ishikawa cells), and GSE28662 (triptolide at 2, 4, and 6 h in MCF7 cells). For each dataset, we converted the probe set IDs into the gene symbols while taking the median expression values for each gene with multiple probes. The query results are shown in Figure S11.

Recurrent similarity score (RSS) analysis. We used a strategy combining the intensity-based similarities and a modified Fisher's method (Jacobsen et al., 2013) *ad hoc* to infer recurrent perturbation–perturbation relationships. Given a perturbation pair (j, μ) available in cell type c, we first fixed j and computed the relative rank of its similarities to all other perturbations containing μ as follows:

$$rr_{\mu,j,c} = \frac{r_{\mu,j,c}}{|L_{j,c}|} - \frac{1}{2|L_{j,c}|}$$

where $L_{j,c}$ is the ordered (from negative to positive) vector of similarities involving perturbation j (excluding itself) in cell type c and $r_{\mu,j,c}$ is the rank for perturbation μ in the vector $L_{j,c}$ such that $r_{\mu,j,c}$ is uniformly distributed on the interval (0, 1). Under the null hypothesis that no negative score exists between perturbations μ and j across all n cell types (n = 10), we would expect twice the negative natural logarithm of the product of these uniformly distributed relative ranks to approximately follow a chi-squared distribution with $2n$ degrees of freedom as follows (Figure S12):

$$H_0 : -2\sum_c \ln\left(rr_{\mu,j,c}\right) \sim X^2_{2n}$$

In this way, the chi-squared test gives a one-sided P-value called p_-. Alternatively, we could test the hypothesis at another side using the inverted rank ($|L_{j,c}| - r_{\mu,j,c} + 1$), giving another statistic called p_+. These two, one-sided P-values, p_- and p_+, were multiplied by 2 to adjust for multiple testing. Finally, for a given perturbation pair (j, μ) available in at least half of the cell types (≥ 5 in this case), we could define a recurrence score $REC_{j\to\mu}$ (when fix j and see μ) as the signed logarithm of these P-values as follows:

$$REC_{j\to\mu} = \begin{cases} \log_{10}\left(2\times p_-\right), \text{if } p_- < p_+ \\ -\log_{10}\left(2\times p_+\right), \text{if } p_+ < p_- \\ 0 \qquad\quad , \text{if } p_- = p_+ \end{cases}$$

Note that the rank of the similarity value of the perturbation pair (j, μ) in the context of j (i.e., fix j and see μ) should be different from that in the context of μ (i.e., fix μ and see j). Therefore, a recurrence score associating a perturbation pair (j, μ) in the context of μ could also be defined such that $REC_{\mu\to j} \neq REC_{j\to\mu}$. To resolve multiple testing problems, we used the Benjamini–Hochberg method to correct P-values as false discovery rates (FDRs) for all perturbation pairs. A recurrent similarity score (RSS) for any two perturbations j and μ was then defined as $RSS_{j,\mu} = (REC_{j\to\mu} + REC_{\mu\to j})/2$. A perturbation pair ($j$, μ) is considered recurrently similar if (1) both $FDR_{j\to\mu}$ and $FDR_{\mu\to j} < 0.001$, and (2) $sim_{j\leftrightarrow\mu} >$ the positive heuristic similarity cutoff (sim^+_{d6}, sim^+_{d24}, or sim^+_{sh96}) or < the negative heuristic similarity cutoff (sim^-_{d6}, sim^-_{d24}, or sim^-_{sh96}) in at least three cell types. We performed analyses of RSS scores separately for each perturbation type (d6, d24, or sh96), and all recurrently similar perturbation pairs are provided in Tables S4–S6 and visualized using Cytoscape (version 3.0.2) in Figures S13–S15 (d6 plus d24 were merged as in Figure 4).

To assess the robustness of the recurrent perturbation pairs against cell types, we resampled 60%

of cell types and repeated the RSS analysis for 100 times for each perturbation type (d6, d24, or sh96). A recurrent perturbation pair (j, μ) is recovered in one round of the resampling analysis if both $FDR_{j\to\mu}$ and $FDR_{\mu\to j} < 0.001$. The robustness of a given recurrent perturbation pair is then defined as the ratio of the number of cases in which they are recovered over the number of tested cases (Figure S16).

To detect perturbation communities in each recurrent network (d6, d24, sh96, or 'd6+d24'), we used a Markov cluster algorithm (MCL (Enright et al., 2002); using mcl in R package MCL version 1.0 with default function arguments while addLoops = TRUE), which simulates random walks on a given adjacency matrix. The resulting clusters were labeled according to the number of perturbagens in descending order. We then exploited a strategy based on hypergeometric tests to assess enrichments in each of Molecular Signature Database (MSigDB (Subramanian et al., 2005); version 5.0) gene set collections across all perturbation clusters. For each perturbation u (of d6, d24, or sh96), we defined an average perturbed profile $\bar{\boldsymbol{d}}_{\boldsymbol{u}}$ across l cell type-specific perturbed profiles available ($l \le n$ where $n = 10$ is the total number of cell types) as $\bar{\boldsymbol{d}}_{\boldsymbol{u}} = (\Sigma_{i=1}^{l} \boldsymbol{d}_{\boldsymbol{u}}^{(i)})/l$. For a chemical perturbagen with both d6 and d24 data in the recurrent drug network, we took the mean of its average perturbed profiles in d6 and d24. Next, for a given perturbation cluster, we computed the cluster average perturbed profile as the mean of all average perturbed profiles of involved perturbations, from which the most extremely perturbed genes (here, top 100) were used for the enrichment analysis. We then performed hypergeometric tests on enrichments for any MSigDB collection of interest across all perturbation communities in each recurrent network separately and adjusted the P-values using the Benjamini–Hochberg correction. All perturbation communities and their enrichment results for each recurrent network are provided in Tables S7–S10. To visualize enrichment results, we selected perturbation clusters and gene sets with at least one significant enrichment (corrected $P < 0.001$ in Figures S18 and S19, and $P < 0.0001$ in Figure S20 for presentation purposes) and then defined ubiquitous gene sets as having significant enrichments in at least 50% of considered perturbation clusters. The ubiquitous gene sets and perturbation clusters were ordered by the counts of valid enrichment P-values and, if counts were equal, they were ordered by the sum of magnitudes of $-\log_{10}(P\text{-value})$. The remaining non-ubiquitous gene sets were grouped according to which perturbation cluster had the most significant enrichment P-value. For each enriched chemical perturbation cluster, we showed the proportion of MoAs and indicated a predominant MoA (accounting for > 50% of MoAs in the cluster) when possible.

Highly specific similarity (HSS) analysis. For each perturbation pair available in at least half of n cell types ($n = 10$; ≥ 5 in this case), we defined a similarity specificity score as $specificity(\boldsymbol{X}) = 1 - (entropy(\boldsymbol{X})/\log_2(l))$ where $\boldsymbol{X}$ is the similarity vector $[s_1, s_2, ..., s_n]$ (a NA value was assigned to an entry without available experiments to define the similarity in that cell type), $entropy(\cdot)$ is the Shannon entropy such that $entropy(\boldsymbol{X}) = 1 - \Sigma p_i \times \log_2(p_i)$ in which $p_i = |s_i|/\Sigma|s_i|$ for $i = 1, ..., l$, and l is the number of available similarity values in $\boldsymbol{X}$ (i.e., non-NA entries; $n/2 \le l \le n$). We then used these similarity specificity values (binned at 0.001 resolution, i.e., rounded to the nearest three decimals) to generate an empirical CDF from which a similarity specific cutoff was derived as the binned value where CDF

is greater than or approximately equal to 0.999 (Figure S21). A perturbation pair is considered highly specific if the following conditions were met: (1) measurements occurred in at least half of cell types; (2) a similarity value was greater than the positive or less than the negative heuristic similarity cutoff in the dominant cell type as the HSS; and (3) its specificity was greater than the similarity specificity cutoff. We performed HSS analyses separately for each perturbation type (d6, d24, or sh96), and all highly specific perturbation pairs are summarized in Tables S11–S13 with a d24 HSS network visualized in Figure 5.

Experimental validation. *Cell culture and chemicals.* MCF7, A549, H1299, and H1975 cells were purchased from American Type Culture Collection (ATCC). Cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and incubated at 37 °C in humidified atmosphere with 5% CO_2 and routinely passaged when 90–95% confluent. Cells tested negative for mycoplasma. Pyrvinium pamoate (P0027), etacrynic acid (SML1083), and crystal violet (C3886) were purchased from Sigma-Aldrich. DMEM (12800), FBS (A3160601), and trypsin-EDTA (15400054) were obtained from Thermo Fisher. MTS (G1111) was ordered from Promega. Antibodies against γH2AX (9718, Cell Signaling), p53 (sc-126, Santa Cruz), TOP2A (ab74715, Abcam), p-TOP2A (Ser1106; ab75765, Abcam), and β-actin (MAB1501, Merck) were purchased from commercial vendors. Horseradish peroxidase-conjugated secondary antibodies were obtained from Abcam. Enhanced chemiluminescence reagent was ordered from Bio-Rad.

MTS assay. Cells (2×10^3 per well) were seeded in a 96-well plate for 24 h followed by indicated chemical treatments. At the endpoint, 20 μL MTS solution was added per well. The plate was incubated for 2 h at 37 °C with 5% CO_2 before measurement of the absorbance at 490 nm (A_{490}) using a ELISA reader (Bio-Rad). Data are represented by A_{490} values with background correction and normalized with the corresponding control group (DMSO) as 100%.

Clonogenic assay. Cells (3×10^4 per well) were seeded in a 6-well plate for 24 h followed by the indicated chemical treatments for 72 h. The remaining cells were then trypsinized and seeded in a new 6-well plate (500 per well) for 14 days. The colonies were fixed with 100% methanol overnight and stained with 1% crystal violet solution.

Western blot analysis. Cells were lysed in buffer containing 50 mM Tris-HCl (pH 7.4; Bioman), 150 mM NaCl, 1% NP40, 0.1% SDS (Bioshop), and 0.5% sodium deoxycholate (Sigma) with a protease inhibitor (16012540, Bioshop) and protein phosphatase inhibitor cocktail (FC0050-0001, Bionovas) immediately added before use. After sonication, protein concentrations were determined using the BCA method (23225, Thermo Fisher). Each sample with equal protein concentrations was subjected to SDS-PAGE and transferred to a PVDF membrane. Nonspecific binding was blocked with 5% nonfat milk. After immunoblotting with the desired primary antibodies, membranes were incubated with appropriate secondary antibodies. Protein bands were detected using ECL reagent.