scCMap: Connecting Genetic and Chemical Perturbations at Single-Cell Resolution

Yiming Li
Central South University
lym1998@csu.edu.cn

Min Zeng
Central South University
zengmin@csu.edu.cn

Min Li
Central South University
limin@mail.csu.edu.cn

Abstract

Systematic mapping of cellular perturbations is fundamental to biomedical research. Pioneering efforts like the Connectivity Map (CMap) established a landmark paradigm by linking compound and gene interventions through their resulting bulk cellular gene expression. More recent initiatives of Tahoe100M and X-Atlas/Orion demonstrated the feasibility of large-scale single-cell transcriptomic profiling of chemical and genetic perturbations, respectively. Yet they focus on single perturbation domains, and there is currently no comprehensive perturbation resource systematically integrates both domains at single-cell resolution. Therefore, we propose to construct **scCMap**, an experimentally harmonized single-cell transcriptomic perturbation map derived from both genetic and chemical screens, designed to enable cross-domain perturbation connection, analysis, and modeling. scCMap will capture complex multi-scale perturbation reagent information, extensive phenotypic matrices, and rich metadata. By providing a unified, high-quality data resource, scCMap will drive a range of AI tasks, including multi-modal molecular representation learning, cross-domain perturbation transfer, biological causal inference, universal single-cell foundation model construction, and virtual AI cell simulations. We anticipate scCMap to mark a new milestone in biomedical research and accelerate AI-powered precision drug discovery.

1 Motivation

Over the past decades, drug discovery has undergone paradigm shifts—from early phenotypic screening to target-based strategies, and now toward integrated approaches [1–3]. Artificial intelligence is uniquely positioned to accelerate this transition, as its powerful representation and prediction capabilities can simultaneously model drug–target molecular interactions and genetic–chemical perturbation phenotypic associations. This trend underscores the central role of cellular perturbation data and the need for richer and more comprehensive resources to enable AI-driven drug discovery.

2 Dataset Rationale

2.1 Limitation of existing large-scale perturbation resources

Low-resolution measurement Phenotypic profiling relied on bulk measurements before single-cell technologies widely adopted [4, 5]. Although projects like NCI-60 [6] and CMap [7, 8] conducted broad screenings across diverse compounds, genes, and cellular contexts, their averaged bulk profiles caused information collapse, obscuring cellular heterogeneity and rare subpopulation effects. Moreover, bulk data produce far fewer training samples than single-cell data, limiting the scale required for generative biological foundation model development [9].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

Low-content readouts Many large-scale perturbation screens, including GDSC [10, 11], CTRP [12, 13], gCSI [14], and DepMap [15–17], generate only single-value endpoint (e.g. drug sensitivity, gene dependency), which is limited to assess the relationships among diverse perturbations and to model complex biological mechanisms.

Single-domain focus Recent perturbation projects of Tahoe100M [18] and X-Atlas/Orion [19] have scaled single-cell perturbation profiling to unprecedented levels. However, each remains confined to a single perturbation domain (chemical or genetic), restricting cross-domain connection and modeling.

Non-molecular profiling Resources like RxRx3 [20] and JUMP-CP [21, 22] provide valuable high-dimensional morphological phenotype profiles across both genetic and chemical perturbations. However, these image-derived features lack direct molecular interpretability, limiting mechanistic and causal modeling.

2.2 Proposed dataset: scCMap

scCMap is a systematically harmonized single-cell perturbation map spanning both genetic and chemical screens, designed as a high-resolution evolution of previous bulk-level CMap project [8] (Table A1). The key characteristics of scCMap are as follows:

- **Data types and resolution:** Single-cell transcriptomic profiles derived from systematic genetic and chemical screens, providing consistent molecular readouts.
- **Scale and diversity:** Perturbations covering thousands of genes and compounds across multiple cellular contexts, enabling broad perturbation coverage and context-specific responses.
- **Metadata and annotations:** Detailed experimental metadata (i.e., perturbation reagent, dosage, treatment duration, time points, cell type, batch) to support analysis and reproducibility.

2.3 Technology scalability

The feasibility of scCMap is underpinned by recent advances in multiplexed perturbation and single-cell transcriptomic profiling technologies. For **chemical perturbations**, pooled cell line assays such as PRISM [23] and Mosaic [18] allow parallel measurement of drug responses across multiple cellular contexts within a single experiment. For **genetic perturbations**, pooled screening (e.g. CRISPR, RNAi) [24–27] allows systematic functional interrogation of thousands of genes in parallel. These complementary multiplexed strategies expand the diversity of perturbation reagent-context pairs across both domains, and single-cell RNA-seq platforms (e.g., sci-Plex [28], Perturb-seq [29]) further scale the throughput of generating molecular-interpretable post-perturbation profiles. Together, these innovations make cost-efficient, automatic, systematic, large-scale, and reproducible perturbation mapping both technically feasible and scalable.

3 Applications

Multi-domain perturbation phenotype analysis and modeling By integrating multi-domain perturbations in a unified single-cell representation space, scCMap enables connection and annotation of genetic and chemical perturbation effects, as well as transfer learning across perturbation domains.

Mechanistic and causal inference Large-scale perturbation-induced molecular readouts enable models to infer causal links from perturbations to phenotypes, facilitating discovery of drug mechanisms, gene functions, and context-specific cellular responses.

Single-cell foundation models and virtual cell simulations By covering diverse cellular contexts and data modalities, including large biomolecules, small molecules, and their perturbation phenotypes, scCMap provides a training ground for constructing universal single-cell foundation models and developing AI virtual cells [30].

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62225209 to M.L.) and the Graduate Student Innovation Research Program of Central South University (Grant No. 2023ZZTS0627 to Y.L.). The authors declare no competing interests.

References

[1] N. Aulner, A. Danckaert, J. Ihm, D. Shum, and S. L. Shorte, "Next-generation phenotypic screening in early drug discovery for infectious diseases," *Trends in Parasitology*, vol. 35, no. 7, pp. 559–570, 2019.

[2] F. Vincent, A. Nueda, J. Lee, M. Schenone, M. Prunotto, and M. Mercola, "Phenotypic drug discovery: recent successes, lessons learned and new directions," *Nature Reviews Drug Discovery*, vol. 21, pp. 899–914, Dec. 2022.

[3] C. R. C. Calado, "Bridging the gap between target-based and phenotypic-based drug discovery," *Expert Opinion on Drug Discovery*, July 2024.

[4] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells," *Cell*, vol. 161, pp. 1187–1201, May 2015.

[5] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets," *Cell*, vol. 161, pp. 1202–1214, May 2015.

[6] A. Monks, D. Scudiero, P. Skehan, R. Shoemaker, K. Paull, D. Vistica, C. Hose, J. Langley, P. Cronise, A. Vaigro-Wolff, M. Gray-Goodrich, H. Campbell, J. Mayo, and M. Boyd, "Feasibility of a High-Flux Anticancer Drug Screen Using a Diverse Panel of Cultured Human Tumor Cell Lines," *JNCI: Journal of the National Cancer Institute*, vol. 83, pp. 757–766, June 1991.

[7] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease," *Science*, vol. 313, pp. 1929–1935, Sept. 2006.

[8] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W.-N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, and T. R. Golub, "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles," *Cell*, vol. 171, pp. 1437–1452.e17, Nov. 2017.

[9] J. E. Rood, A. Hupalowska, and A. Regev, "Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas," *Cell*, vol. 187, pp. 4520–4545, Aug. 2024.

[10] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O'Brien, J. L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott, and C. H. Benes, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, pp. 570–575, Mar. 2012.

[11] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC):

a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Research*, vol. 41, pp. D955–D961, Jan. 2013.

[12] A. Basu, N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, M. L. Stewart, D. Ito, S. Wang, A. L. Bracha, T. Liefeld, M. Wawer, J. C. Gilbert, A. J. Wilson, N. Stransky, G. V. Kryukov, V. Dancik, J. Barretina, L. A. Garraway, C. S.-Y. Hon, B. Munoz, J. A. Bittker, B. R. Stockwell, D. Khabele, A. M. Stern, P. A. Clemons, A. F. Shamji, and S. L. Schreiber, "An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules," *Cell*, vol. 154, pp. 1151–1161, Aug. 2013.

[13] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, B. Alexander, A. Li, P. Montgomery, M. J. Wawer, N. Kuru, J. D. Kotz, C. S.-Y. Hon, B. Munoz, T. Liefeld, V. Dančík, J. A. Bittker, M. Palmer, J. E. Bradner, A. F. Shamji, P. A. Clemons, and S. L. Schreiber, "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset," *Cancer Discovery*, vol. 5, pp. 1210–1223, Nov. 2015.

[14] P. M. Haverty, E. Lin, J. Tan, Y. Yu, B. Lam, S. Lianoglou, R. M. Neve, S. Martin, J. Settleman, R. L. Yauch, and R. Bourgon, "Reproducible pharmacogenomic profiling of cancer cell line panels," *Nature*, vol. 533, pp. 333–337, May 2016.

[15] A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, R. M. Meyers, L. Ali, A. Goodale, Y. Lee, G. Jiang, J. Hsiao, W. F. J. Gerath, S. Howell, E. Merkel, M. Ghandi, L. A. Garraway, D. E. Root, T. R. Golub, J. S. Boehm, and W. C. Hahn, "Defining a Cancer Dependency Map," *Cell*, vol. 170, pp. 564–576.e16, July 2017.

[16] F. M. Behan, F. Iorio, G. Picco, E. Gonçalves, C. M. Beaver, G. Migliardi, R. Santos, Y. Rao, F. Sassi, M. Pinnelli, R. Ansari, S. Harper, D. A. Jackson, R. McRae, R. Pooley, P. Wilkinson, D. van der Meer, D. Dow, C. Buser-Doepner, A. Bertotti, L. Trusolino, E. A. Stronach, J. Saez-Rodriguez, K. Yusa, and M. J. Garnett, "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens," *Nature*, vol. 568, pp. 511–516, Apr. 2019.

[17] R. Arafeh, T. Shibue, J. M. Dempster, W. C. Hahn, and F. Vazquez, "The present and future of the Cancer Dependency Map," *Nature Reviews Cancer*, vol. 25, pp. 59–73, Jan. 2025.

[18] J. Zhang, A. A. Ubas, R. d. Borja, V. Svensson, N. Thomas, N. Thakar, I. Lai, A. Winters, U. Khan, M. G. Jones, V. Tran, J. Pangallo, E. Papalexi, A. Sapre, H. Nguyen, O. Sanderson, M. Nigos, O. Kaplan, S. Schroeder, B. Hariadi, S. Marrujo, C. C. A. Salvino, G. G. Olivares, R. Koehler, G. Geiss, A. Rosenberg, C. Roco, D. Merico, N. Alidoust, H. Goodarzi, and J. Yu, "Tahoe-100M: A Giga-Scale Single-Cell Perturbation Atlas for Context-Dependent Gene Function and Cellular Modeling," Feb. 2025.

[19] A. C. Huang, T.-H. S. Hsieh, J. Zhu, J. Michuda, A. Teng, S. Kim, E. M. Rumsey, S. K. Lam, I. Anigbogu, P. Wright, M. Ameen, K. You, C. J. Graves, H. J. Kim, A. J. Litterman, R. V. Sit, A. Blocker, and C. Chu, "X-Atlas/Orion: Genome-wide Perturb-seq Datasets via a Scalable Fix-Cryopreserve Platform for Training Dose-Dependent Biological Foundation Models," June 2025.

[20] M. M. Fay, O. Kraus, M. Victors, L. Arumugam, K. Vuggumudi, J. Urbanik, K. Hansen, S. Celik, N. Cernek, G. Jagannathan, J. Christensen, B. A. Earnshaw, I. S. Haque, and B. Mabey, "RxRx3: Phenomics Map of Biology," *bioRxiv*, p. 2023.02.07.527350, Feb. 2023.

[21] S. N. Chandrasekaran, J. Ackerman, E. Alix, D. M. Ando, J. Arevalo, M. Bennion, N. Boisseau, A. Borowa, J. D. Boyd, L. Brino, P. J. Byrne, H. Ceulemans, C. Ch'ng, B. A. Cimini, D.-A. Clevert, N. Deflaux, J. G. Doench, T. Dorval, R. Doyonnas, V. Dragone, O. Engkvist, P. W. Faloon, B. Fritchman, F. Fuchs, S. Garg, T. J. Gilbert, D. Glazer, D. Gnutt, A. Goodale, J. Grignard, J. Guenther, Y. Han, Z. Hanifehlou, S. Hariharan, D. Hernandez, S. R. Horman, G. Hormel, M. Huntley, I. Icke, M. Iida, C. B. Jacob, S. Jaensch, J. Khetan, M. Kost-Alimova, T. Krawiec, D. Kuhn, C.-H. Lardeau, A. Lembke, F. Lin, K. D. Little, K. R. Lofstrom, S. Lotfi, D. J. Logan, Y. Luo, F. Madoux, P. A. M. Zapata, B. A. Marion, G. Martin, N. J. McCarthy, L. Mervin, L. Miller, H. Mohamed, T. Monteverde, E. Mouchet, B. Nicke, A. Ogier, A.-L. Ong, M. Osterland, M. Otrocka, P. J. Peeters, J. Pilling, S. Prechtl, C. Qian, K. Rataj, D. E. Root, S. K. Sakata, S. Scrace, H. Shimizu, D. Simon, P. Sommer, C. Spruiell, I. Sumia, S. E. Swalley, H. Terauchi, A. Thibaudeau, A. Unruh, J. V. d. Waeter, M. V. Dyck, C. v. Staden, M. Warchoł, E. Weisbart, A. Weiss, N. Wiest-Daessle, G. Williams, S. Yu, B. Zapiec, M. Żyła, S. Singh, and

A. E. Carpenter, "JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations," *bioRxiv*, p. 2023.03.23.534023, Mar. 2023.

[22] S. N. Chandrasekaran, B. A. Cimini, A. Goodale, L. Miller, M. Kost-Alimova, N. Jamali, J. G. Doench, B. Fritchman, A. Skepner, M. Melanson, A. A. Kalinin, J. Arevalo, M. Haghighi, J. C. Caicedo, D. Kuhn, D. Hernandez, J. Berstler, H. Shafqat-Abbasi, D. E. Root, S. E. Swalley, S. Garg, S. Singh, and A. E. Carpenter, "Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations," *Nature Methods*, vol. 21, pp. 1114–1121, June 2024.

[23] C. Yu, A. M. Mannan, G. M. Yvone, K. N. Ross, Y.-L. Zhang, M. A. Marton, B. R. Taylor, A. Crenshaw, J. Z. Gould, P. Tamayo, B. A. Weir, A. Tsherniak, B. Wong, L. A. Garraway, A. F. Shamji, M. A. Palmer, M. A. Foley, W. Winckler, S. L. Schreiber, A. L. Kung, and T. R. Golub, "High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines," *Nature Biotechnology*, vol. 34, pp. 419–423, Apr. 2016.

[24] P. J. Paddison, J. M. Silva, D. S. Conklin, M. Schlabach, M. Li, S. Aruleba, V. Balija, A. O'Shaughnessy, L. Gnoj, K. Scobie, K. Chang, T. Westbrook, M. Cleary, R. Sachidanandam, W. Richard McCombie, S. J. Elledge, and G. J. Hannon, "A resource for large-scale RNA-interference-based screens in mammals," *Nature*, vol. 428, pp. 427–431, Mar. 2004.

[25] T. Wang, J. J. Wei, D. M. Sabatini, and E. S. Lander, "Genetic Screens in Human Cells Using the CRISPR-Cas9 System," *Science*, vol. 343, pp. 80–84, Jan. 2014.

[26] O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, and F. Zhang, "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells," *Science*, vol. 343, pp. 84–87, Jan. 2014.

[27] P. Datlinger, A. F. Rendeiro, C. Schmidl, T. Krausgruber, P. Traxler, J. Klughammer, L. C. Schuster, A. Kuchler, D. Alpar, and C. Bock, "Pooled CRISPR screening with single-cell transcriptome readout," *Nature Methods*, vol. 14, pp. 297–301, Mar. 2017.

[28] S. R. Srivatsan, J. L. McFaline-Figueroa, V. Ramani, L. Saunders, J. Cao, J. Packer, H. A. Pliner, D. L. Jackson, R. M. Daza, L. Christiansen, F. Zhang, F. Steemers, J. Shendure, and C. Trapnell, "Massively multiplex chemical transcriptomics at single-cell resolution," *Science*, vol. 367, pp. 45–51, Jan. 2020.

[29] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev, "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens," *Cell*, vol. 167, pp. 1853–1866.e17, Dec. 2016.

[30] C. Bunne, Y. Roohani, Y. Rosen, A. Gupta, X. Zhang, M. Roed, T. Alexandrov, M. AlQuraishi, P. Brennan, D. B. Burkhardt, A. Califano, J. Cool, A. F. Dernburg, K. Ewing, E. B. Fox, M. Haury, A. E. Herr, E. Horvitz, P. D. Hsu, V. Jain, G. R. Johnson, T. Kalil, D. R. Kelley, S. O. Kelley, A. Kreshuk, T. Mitchison, S. Otte, J. Shendure, N. J. Sofroniew, F. Theis, C. V. Theodoris, S. Upadhyayula, M. Valer, B. Wang, E. Xing, S. Yeung-Levy, M. Zitnik, T. Karaletsos, A. Regev, E. Lundberg, J. Leskovec, and S. R. Quake, "How to build the virtual cell with artificial intelligence: Priorities and opportunities," *Cell*, vol. 187, pp. 7045–7063, Dec. 2024.

[31] S. M. Corsello, R. T. Nagari, R. D. Spangler, J. Rossen, M. Kocak, J. G. Bryan, R. Humeidi, D. Peck, X. Wu, A. A. Tang, V. M. Wang, S. A. Bender, E. Lemire, R. Narayan, P. Montgomery, U. Ben-David, C. W. Garvie, Y. Chen, M. G. Rees, N. J. Lyons, J. M. McFarland, B. T. Wong, L. Wang, N. Dumont, P. J. O'Hearn, E. Stefan, J. G. Doench, C. N. Harrington, H. Greulich, M. Meyerson, F. Vazquez, A. Subramanian, J. A. Roth, J. A. Bittker, J. S. Boehm, C. C. Mader, A. Tsherniak, and T. R. Golub, "Discovering the anticancer potential of non-oncology drugs by systematic viability profiling," *Nature Cancer*, vol. 1, pp. 235–248, Feb. 2020.

[32] L. Xiang, Y. Wang, W. Shao, Q. Li, X. Yu, M. Wei, Y. Gui, S. Li, P. Qin, C. Hu, G. Zhang, X. Zhang, J. Wang, Y. Li, J. An, Y. Luo, Y. Liao, J. Deng, X. Tai, R. Y. Xu, L. Huang, D. Guo, G. Zhang, Z. Xie, Y. Deng, J. Xu, and D. Wang, "High-throughput profiling of chemical-induced gene expression across 93,644 perturbations," *Nature Methods*, pp. 1–10, Aug. 2025.

A Summary of Existing Large-Scale Perturbation Screen Datasets

Table A1: Summary of existing large-scale perturbation screen projects and datasets

Project/Dataset	Timeline/ Release Time	Perturbation Scale	Endpoint Readout	High-Res.[1]	High-Content[2]	Multi-Domain[3]	Mol-Profile[4]	Access Link
NCI-60 [6]	1990-Present	>50K compounds, 60 cell lines	Bulk-level cell growth inhibition	✗	✗	✗	✗	https://wiki.nci.nih.gov/spaces/NCIDTPdata/pages/147193864/NCI-60+Growth+Inhibition+Data
CMap [7, 8]	2006 & 2017	5075 genes, 19,811 compounds, 314 biologics, 9 core cell lines	Bulk-level gene expression	✗	✓	✓	✓	https://clue.io/
GDSC [11]	2012-2023	624 compounds, 978 cell lines	Bulk-level drug sensitivity	✗	✗	✗	✗	https://www.cancerrxgene.org/
CTRP [12, 13]	2013 & 2015	481 compounds, 823 cell lines	Bulk-level drug sensitivity	✗	✗	✗	✗	https://portals.broadinstitute.org/ctrp.v2/
gCSI [14]	2016	44 compounds, 569 cell lines	Bulk-level drug sensitivity	✗	✗	✗	✗	http://research-pub.gene.com/gCSI_GRvalues2019/
PRISM [23, 31]	2016 & 2020	4,518 compounds, 588 cell lines	Bulk-level drug sensitivity (cell line-pooled screen)	✗	✗	✗	✗	https://depmap.org/repurposing/
DepMap [15, 16]	2017-2022	∼1.8K genes, >1K cell lines	Bulk-level gene dependency	✗	✗	✗	✗	https://depmap.org/portal/
JUMP-CP [21]	2023	∼15K genes, ∼116K compounds, 2 cell lines	Image-based profile	✗→✓	✓	✓	✗	https://github.com/jump-cellpainting/datasets
RxRx3 [20]	2023	17,063 genes, 1,674 compounds, 1 cell line	Image-based profile	✗→✓	✓	✓	✗	https://www.rxrx.ai/rxrx3
Tahoe100M [18]	2025	379 compounds, 50 cell lines, >100M cells	Single-cell-level gene expression (cell line-pooled screen)	✓	✓	✗	✓	https://huggingface.co/datasets/tahoebio/Tahoe-100M
X-Atlas/Orion [19]	2025	18,903 genes, 2 cell lines, ∼8M cells	Single-cell-level gene expression (CRISPR-pooled screen)	✓	✓	✗	✓	https://doi.org/10.25452/figshare.plus.29190726
CIGS [32]	2025	13,221 compounds, 2 cell lines	Bulk-level gene expression	✗	✓	✗	✓	https://cigs.iomicscloud.com/
scCMap[5]	Future	>5K genes, >10K compounds, ≥9 core cell lines, >60M cells	Single-cell-level gene expression (CRISPR-pooled genetic & cell line-pooled chemical screens)	✓	✓	✓	✓	

[1] **High-Res.:** Indicates the assay provides single-cell-level phenotypic profiles. The microscopy images could generate single-cell profiles by using cell segment algorithms.
[2] **High-Content:** Indicates the assay measures thousands of features per profile, such as morphological or transcriptomic data.
[3] **Multi-Domain:** Indicates the dataset integrates both genetic (e.g., CRISPR knockout, CRISPRi, gene overexpression) and chemical (e.g., compound) perturbations.
[4] **Mol-Profile:** Indicates the assay provides molecular-level profiles, such as gene expression or protein abundance.
[5] The expected perturbation coverage and data scale of scCMap are referenced from CMap and the recent Tahoe100M and X-Atlas/Orion projects.