scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data

Tessa Durakis Green[1,*] **Stefan Peidli**[2,3,*] **Ciyue Shen**[1,4,5,6] **Torsten Gross**[7]
Joseph Min[1] **Samuele Garda**[3] **Jake P. Taylor-King**[7] **Debora S. Marks**[1,6]
Augustin Luna[1,4,5,6] **Nils Blüthgen**[2,3,x] **Chris Sander** [1,4,5,6,x]
[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA
[2]Institute of Pathology, Charité, Universitätsmedizin Berlin, Corporate Member
of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
[3]IRI Life Sciences, Humboldt-Universität zu Berlin, Berlin, Germany
[4]Department of Cell Biology, Harvard Medical School, Boston, MA, USA
[5]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[6]Broad Institute, Cambridge, MA, USA
[7]Relation Therapeutics, London, UK
tessa_green@fas.harvard.edu, stefan.peidli@charite.de
nils_bluethgen@charite.de, chris_sander@hms.harvard.edu
$*$: joint first authors x: joint senior authors

Recent biotechnological advances led to growing numbers of single-cell studies, which reveal molecular and phenotypic responses to large numbers of perturbations [1,2]. However, analysis across diverse datasets is typically hampered by differences in format, naming conventions, data filtering and normalization. To facilitate development and benchmarking of computational methods in systems biology, we collect a set of 44 publicly available single-cell perturbation-response datasets with molecular readouts, including RNA, proteins and chromatin accessibility (Figure Panel A). We apply uniform pre-processing and quality control pipelines and harmonize feature annotations. The resulting information resource enables efficient development and testing of computational analysis methods, and facilitates direct comparison and integration across datasets. 32 RNA datasets in this resource were perturbed using CRISPR and 9 were perturbed with drugs (Figure Panel B). We also include three scATAC datasets, as well as three CITE-seq datasets with protein and RNA counts separately downloadable. For each scRNA-seq dataset we supply count matrices, where each cell has a perturbation annotation, quality control metrics including gene counts and mitochondrial read percentage. Quality control plots for each dataset are also available on the scPerturb website. Notably, more than 8000 CRISPR perturbations are shared across multiple datasets. We anticipate this data resource being useful for developing machine learning models for perturbation responses across datasets and other tasks.

To compare and evaluate perturbations within each dataset, we demonstrate the application of a distance measure between high-dimensional point clouds to quantify perturbation similarity and strength. The E-distance (short for energy distance, [3,4]) contextualizes the notion that two such point clouds are distinguishable if they are far apart compared to the width of both clouds, providing intuition about the signal-to-noise ratio in a dataset.

Let $x_1, ..., x_N \in \mathbb{R}^d$ and $y_1, ..., y_N \in \mathbb{R}^d$ be samples from two distributions X, Y, corresponding to two sets of N and M cells, respectively.

36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, Louisiana, USA.

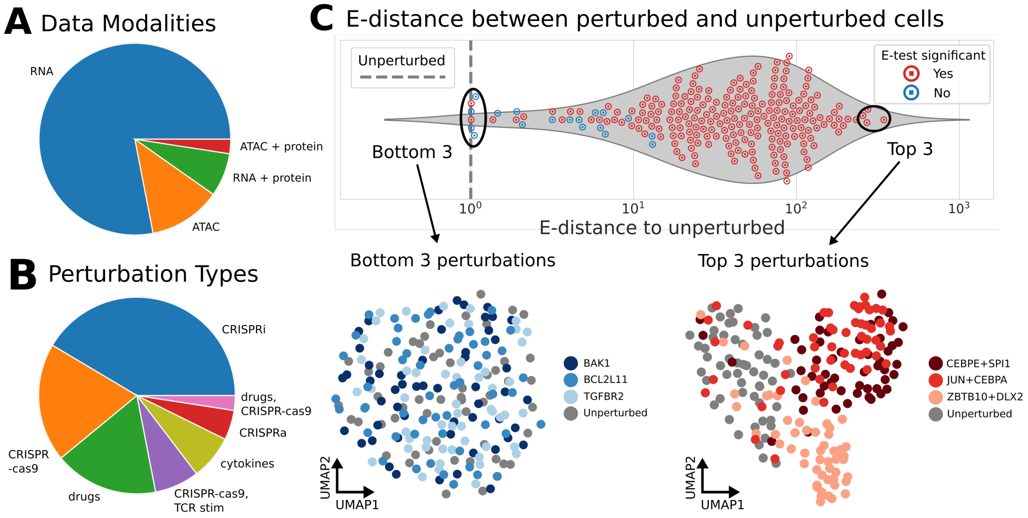

Figure 1: **Diverse single-cell perturbation-response datasets profiled with E-distances**. (A,B) The majority of included datasets are scRNA-seq data with CRISPR (DNA cut, inhibition or activation) perturbations using cell lines derived from various cancers. The studies performed on cells from primary tissues generally use drug perturbations. (C) E-distances between perturbed and unperturbed cells in [5], log+1 scaled. Color indicates significant difference to unperturbed cells according to E-test. UMAP embeddings of cells from top and bottom perturbations w.r.t. E-distance to unperturbed exemplify the connection between E-distance and similarity of point clouds

We define

$$\delta_{XY} = \frac{1}{NM}\sum_{i=1}^{M}\sum_{j=1}^{N}||x_i - y_j||$$

$$\sigma_X = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}||x_i - x_j||$$

and σ_Y defined accordingly. Intuitively, δ_{XY} is the mean distance between cells from the two distributions, while σ_X describes the mean distance between a cell from X to another cell from X. The E-distance between X and Y is defined as:

$$E(X,Y) := 2\delta_{XY} = \sigma_X - \sigma_Y$$

We computed the E-distance in PCA space between each group of perturbed cells and the unperturbed cells for our collected datasets. This results in a distribution of E-distances for each perturbation in a dataset, exemplarily shown for one selected dataset (Figure Panel C, upper portion). The more a perturbation is on the right side of this distribution, the more distinctive its effect on the cells is captured in the data. We also used the E-distance as a test statistic for a Monte Carlo permutation test ("E-test"), robustly testing whether the distinctiveness is statistically significant. To visualize what a high or low E-distance to unperturbed cells means, we select the three closest / furthest perturbations in the example dataset and calculate a UMAP embedding for the cells of these perturbations jointly with the unperturbed cells (Figure Panel C, lower portion). The top three furthest perturbations are easily distinguishable from the gray unperturbed cells, while the bottom three closest perturbations form a single, uniform cloud

This work provides an information resource and guide for researchers working with single-cell perturbation data and highlights conceptual considerations for new experiments. The scPerturb data collections is publicly available and is described in further detail in an associated preprint.

Acknowledgments and Disclosure of Funding

The data collection, scPerturb, is publicly available at scperturb.org. Further detail is available in our preprint (doi:10.1101/2022.08.20.504663).

1 Acknowledgements

- National Resource for Network Biology (NRNB, P41GM103504)
- Supported by the Wellcome Leap ΔTissue Program
- Deutsche Forschungsgemeinschaft (DFG, RTG2424 CompCancer)
- Einstein Stiftung Berlin (Einstein visiting fellow program)
- Computation was in part performed on the HPC for Research cluster of the Berlin Institute of Health.

References

[1] Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y. et al. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell. *167*.

[2] Dixit, A., Parnas, O., Li, B., and Chen, J. (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*.

[3] Replogle J.M., Saunders R.A., Pogson A.N., Hussmann J.A., Lenail A., Guna A., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell *185*.

[4] Rizzo, M.L. and Székely, G.J. (2016). Energy distance. WIREs Computational Statistics *8*.

[5] Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., and Weissman, J.S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science *365*.