adverSCarial: a tool for evaluating adversarial attacks on single-cell transcriptomics classifiers

Ghislain Fievet, Sébastien Hergalant

adverSCarial: a tool for evaluating adversarial attacks on single-cell transcriptomics classifiers

Ghislain Fievet and Sébastien Hergalant

UMRS INSERM 1256 NGERE, 9 Av. de la Forêt de Haye, 54500, Vandœuvre-lès-Nancy, France

Corresponding Author: ghislain.fievet@univ-lorraine.fr

In single-cell transcriptomics, machine learning techniques have been applied for automatic cell annotation [1], for the identification of cancer cell subpopulations [2], and for modelling the transcriptional dynamics that govern cellular fate and development [3-4]. These methods hold potential value for routine practice in clinical settings but must address critical challenges in the use – and misuse – of AI algorithms for reliability and interpretability [5]. The field of explainable AI addresses these concerns, among which the robustness to adversarial attacks, *i.e.* techniques designed to fool a machine learning model with deceptive and inaccurate data.

Here we present *adverSCarial*, an original R package that generates adversarial attacks on single-cell transcriptomics classifiers (https://github.com/GhislainFievet/adverSCarial). *adverSCarial* currently proposes four customizable functions to produce adversarial attacks. In this work, we define two types of methods: the minimal (min) and the maximal (max) change attacks. The min change attack finds the smallest possible perturbation in the input data leading to a change of classification. The max change attack finds the largest data modification which does not alter the initial classification.

On a reference peripheral blood mononuclear cells (PBMC) dataset of 2,700 cells and 22,042 genes [6], we further tested *adverSCarial* and compared the susceptibility of two published cell type classifiers to these attacks, the classification tree based CHETAH [7] and the marker based scType [8]. Both classifiers showed weaknesses to min and max adversarial attacks, especially to the max change method. Indeed, we found that it is possible to modify a significant proportion of genes without affecting the classification confidence. For a representative example, on the CD14 monocytes cluster of the studied PBMC dataset, replacing all 22,042 gene expression values by their last percentile did not modify CHETAH identification as CD14 cells. These results were generalized to all cell types and highlight the concern that machine learning algorithms may fail to detect even significant anomalies in the input data.

In conclusion, this work demonstrates the usefulness of such techniques in testing the robustness of classifier families and the extent of the modifications required to deviate their intended use. We believe that our approach and tool can guide the development and validation of more reliable models that could be used in clinical setups, and aim to extensively evaluate these tools on a wide variety of single-cell datasets.

References

[1] Xinlei Zhao, Shuang Wu, Nan Fang, Xiao Sun, Jue Fan. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Briefings in Bioinformatics*, pages 1581-1595, 2020.
[2] Dohmen, J., Baranovskii, A., Ronen, J. et al. Identifying tumor cells at the single-cell level using machine learning. *Genome Biol*, 2022.
[3] Kushagra Pandey, Hamim Zafar. Inference of cell state transitions and cell fate plasticity from single-cell with MARGARET. *Nucleic Acids Research*, 2022.
[4] Saelens, W., Cannoodt, R., Todorov, H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol*, 2019.
[5] de Hond, A.A.H., Leeuwenberg, A.M., Hooft, L. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digit. Med*., 2022.
[6] Zheng, G., Terry, J., Belgrader, P. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun,* 2017.
[7] Jurrian K de Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, Frank C P Holstege, CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research,* 2019.
[8] Ianevski, A., Giri, A.K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun,* 2022.