# adverSCarial: a toolkit for exposing classifier vulnerabilities in single-cell transcriptomics

Ghislain Fievet, Julien Broséus, David Meyre, Sébastien Hergalant

▶ **To cite this version:**

Ghislain Fievet, Julien Broséus, David Meyre, Sébastien Hergalant. adverSCarial: a toolkit for exposing classifier vulnerabilities in single-cell transcriptomics. ISMB/ECCB2025, Jul 2025, Liverpool, United Kingdom. ⟨hal-05222861⟩

**HAL Id: hal-05222861**

**https://hal.univ-lorraine.fr/hal-05222861v1**

Submitted on 25 Aug 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

# adverSCarial: a toolkit for exposing classifier vulnerabilities in single-cell transcriptomics

Ghislain Fievet[1], Julien Broséus[1,2], David Meyre[1,3] and Sébastien Hergalant[1]

[1]INSERM U1256, NGERE lab, University of Lorraine, Nancy, France. [2]Department of Biological Hematology, University Hospital of Nancy. [3]Department of Molecular Medicine, University Hospital of Nancy.

## 1 Introduction

Single-cell RNA-sequencing (scRNA-seq) has transformed our ability to identify and characterize cell types with high granularity, thereby opening new frontiers in both biomedical research and precision medicine applications. Several machine learning (ML) algorithms dedicated to scRNA-seq data classification demonstrated good accuracy [1], although concerns regarding their vulnerability to adversarial attacks, well-targeted yet subtle input modifications that can corrupt classification, remain largely underexplored on black-box models [2,3]. Such weaknesses endanger biomedical research and clinical decision-support scenarios where undetectable gene expression modifications could lead to faulty conclusions. To bridge this gap, we developed *adverSCarial*, an R package that evaluates a transcriptomic classifier response to a range of subtle, medium and high perturbations in gene expression throughout multiple adversarial attack modes. In addition to revealing points of failure, these targeted modifications often expose the inner decision-making processes of the classifiers, hence contributing to explainable artificial intelligence (XAI) methods.

## 2 Materials and Methods

**Classifiers.** We evaluated five scRNA-seq classification tools: scType (marker-based) [4], CHETAH (hierarchical clustering) [5], scAnnotatR (support vector machine (SVM)) [6], scRF (random forests (RF)) and scMLP (multilayer perceptron (MLP) neural networks).

**Datasets.** Each classifier was trained on half of four hallmark datasets of the public domain, PBMC3k (peripheral blood, 3k cells), AXILLA10k (cancer metastases, 10k cells), LIVER10k (cancer metastases, 10k cells), and KIDNEY10k (embryology and tissue development, 10k cells), and subsequently tested on the remaining samples [7,8]. These datasets span multiple tissue origins and feature a diverse range of cell types and expression profiles.

## 3 Results

**Adversarial attack modes (Figure 1).** *Single-gene attacks* alter the expression of one gene at a time to produce misclassification. *Max-change attacks* attempt to modify as many genes as possible without changing the classification, thus exposing critical gene signatures. *Cluster-based Gradient Descent (CGD)* uses approximative gradient steps to iteratively push the prediction off the targeted cluster. *CGD* can simulate both subtle and large-scale perturbations by controlling the step size. With the entire cell x gene matrix defined as $X$, the subset matrix corresponding to a cluster of cells $m$ as $X_{clust\ m}$, the gene expression vector of a cell $i$ as $x_i$, $\hat{y}_{x_i}$ the cell type prediction of the cell $i$ , *ctype1* and *ctype2* respectively corresponding to the most likely predicted and the secondly most likely predicted cell type, and $X_{clustm}^{adv}$ being the resulting adversarial matrix :

$$\nabla_{j'}^{fd}\hat{y}_{x_i} = \hat{y}_{f_{j'}(x_i,\alpha,\epsilon)} - \hat{y}_{x_i}, \text{ with } f_{j'}(x_i,\alpha,\epsilon) = x_{ij}\left(1-\delta_{jj'}\right) + \left(\left(1+\alpha.sign\left(x_{ij}\right)\right)x_{ij} + \epsilon\right)\delta_{jj'}$$

$$slope = \left(\nabla_{j'}^{fd}\hat{y}_{x_i}\right)_{ctype2_{X_{clustm}^{n}}} - \left(\nabla_{j'}^{fd}\hat{y}_{x_i}\right)_{ctype1_{X_{clustm}^{n}}}, (x_i^{n+1})_j = x_{ij}\left(1-\delta_{jj'}\right) +$$

$$\left(\left(1+\alpha.sign\left(x_{ij}\right)sign(slope)\right)x_{ij} + \epsilon.sign(slope)\right)\delta_{jj'}, X_{clustm}^{adv} = (x_i^{n+1})_{i\in clustm}$$

**Modification types.** All attack functions allow various types of gene perturbation to mimic real-world technical or biological variability as well as malicious alterations: switch-on, switch-off, increase/decrease by decile of expression, aberrant values, additive and multiplicative step sizes, and custom modifications.

**Detectability of the modifications and performance metrics.** An essential aspect of adversarial attacks is the human perceptibility of the introduced changes [2]. Overall, in any classifier/dataset combination, *single-gene* attacks produce variably discernible alterations in the targeted gene's expression but preserve the visual patterns of the cell type, *max-change* obliterates the visual patterns of the targeted cell clusters but do not alter the classification, and *CGD* fools the classifiers when a sum of slight adjustments is reached. Classifiers show a decline in accuracy, F1-score, and area under the ROC curve (AUC) once enough adversarial modifications are introduced. The threshold of modification (the number of genes changed or the amplitude of perturbation) is critical: small adjustments can produce no visible misclassification, whereas large but selective changes may drastically shift predictions.

**Availability and Implementation:** *adverSCarial* is a freely available R package accessible from *Bioconductor*: https://bioconductor.org/packages/adverSCarial/.

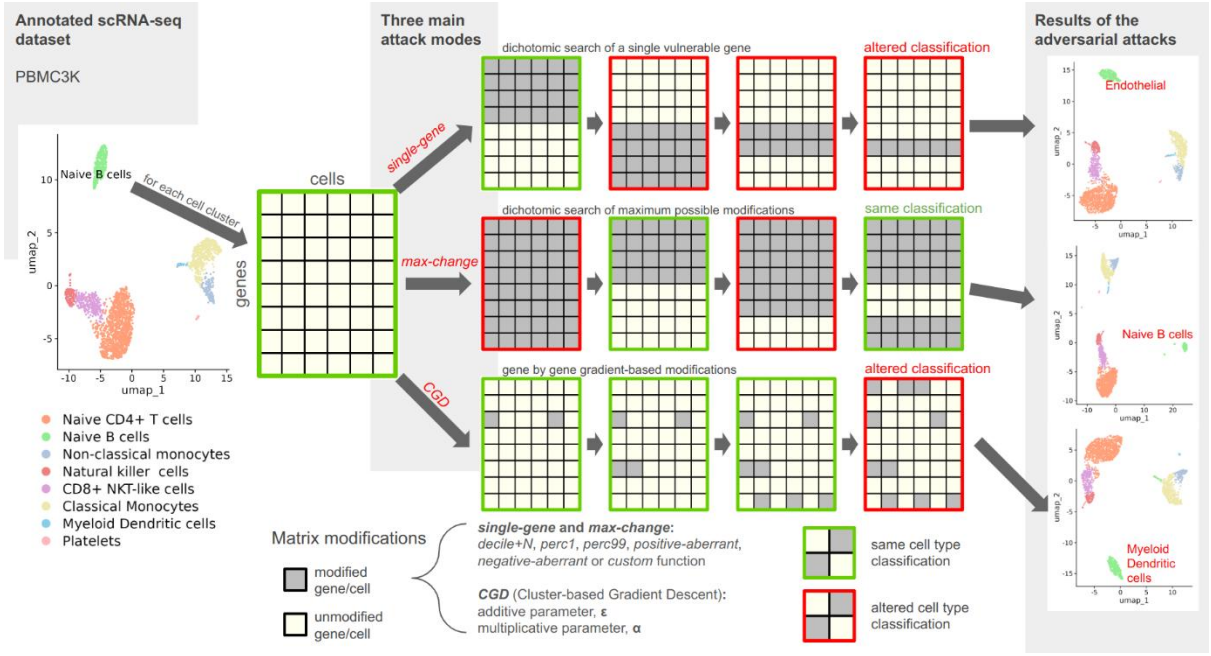


***Figure 1**: adverSCarial implementation and feature overview.*



## Discussion

Despite their distinct architecture, all five ML classifiers eventually failed with the right kind of perturbations, or sufficiently strong ones. However, the severity and nature of these failures depended on both the algorithm and the type of modification. For instance, tree-based methods like hierarchical clustering and random forests showed few responses to medium or high modifications, indicating a plateau in sensitivity due to a natural characteristic: once an input reaches the defined threshold of a tree node, further increases have no additional impact on the outcome. In contrast, MLPs were unresponsive below medium modifications and very vulnerable beyond with thousands of possible attacks: this effect stems from the neural network's use of multiplicative coefficients, which amplify sensitivity to outliers.
In conclusion, *adverSCarial* provides a framework to test scRNA-seq classifiers for robustness against adversarial attacks. By highlighting how different algorithms respond to various perturbation modes, it helps guide the design of more resilient and interpretable ML pipelines in single-cell transcriptomics. As scRNA-seq tools steadily move toward clinical applications, acknowledging adversarial vulnerabilities becomes increasingly important to ensure reliable results in precision medicine.

## References

[1] Zhao, X., Wu, S., Fang, N., et al. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Briefings in Bioinformatics,* 2020;21:1581-95. https://doi.org/10.1093/bib/bbz096
[2] Goodfellow, I. J., Shlens, J., Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint,* 2015. https://doi.org/10.48550/arXiv.1412.6572
[3] Sadria, M., Layton, A., Bader G.D. Adversarial training improves model interpretability in single-cell RNA-seq analysis. *Bioinformatics Advances*, 2023; vbad166, https://doi.org/10.1093/bioadv/vbad166
[4] 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;13:1246. https://doi.org/10.1038/s41467-022-28803-w
[5] de Kanter, JK., Lijnzaad, P., Candelli, T. *et al*. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, 2019;47(16):e95. https://doi.org/10.1093/nar/gkz543
[6] Nguyen, V., Griss, J. scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data. *BMC Bioinformatics,* 2022. 23(44). https://doi.org/10.1186/s12859-022-04574-5
[7] Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell*. 2020;181(2):236-249. https://doi.org/10.1016/j.cell.2020.03.053
[8] Stewart, B., Ferdinand, J., Young, M. *et al.* Spatiotemporal immune zonation of the human kidney, *Science,* 2019;365:1461-1466. https://doi.org/10.1126/science.aat5031