Biotechnology and Regenerative Sciences
ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453
https://galaxiauniverse.com/index.php/BRS/issue/archive

Machine Learning Models for Predicting Cellular Differentiation in Regenerative Medicine

Sofia Nowak[1], Oscar Hansen[2]

[1] Professor, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: sofia.nowak58@gmail.com | ORCID: 9950-1991-4034-4251
[2] Associate Professor, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: oscar.hansen233@gmail.com | ORCID: 5148-9475-0481-0410

ABSTRACT

Cellular differentiation -- the process by which pluripotent stem cells commit to specific cell lineages -- is a critical determinant of regenerative medicine therapeutic outcomes, yet remains incompletely understood due to the complexity of the transcriptional regulatory networks governing lineage commitment decisions. Accurate computational prediction of cellular differentiation trajectories and terminal cell fates from molecular profiles would substantially accelerate the development of cell therapy protocols, reduce experimental iteration costs, and improve the reproducibility of directed differentiation protocols. This paper proposes the Cellular Differentiation Prediction (CDP) framework, a machine learning methodology for predicting differentiation trajectories and terminal cell fates from single-cell RNA sequencing (scRNA-seq) and epigenomic data. CDP integrates four ML components: a graph neural network trajectory predictor (GNN-TP) that models transcriptional state transitions as directed graphs; a multi- modal epigenomic integrator (MEI) that incorporates ATAC-seq chromatin accessibility with transcriptomic data; a fate probability estimator (FPE) that quantifies the probability of each cell committing to each available lineage; and a differentiation protocol optimiser (DPO) that recommends growth factor combinations and timing predicted to maximise directed differentiation efficiency. CDP is evaluated on five publicly available scRNA-seq datasets spanning haematopoiesis, neurogenesis, and cardiac differentiation, and validated against three experimental directed differentiation protocols. CDP achieves trajectory prediction accuracy of 84.6% (SD = 3.8%) versus baseline pseudotime methods (68.4%, SD = 4.6%) and fate assignment accuracy of 91.2% (SD = 2.8%) versus Monocle3 baseline (76.8%, SD = 3.4%). The DPO protocol recommendations achieve 24.8% higher directed differentiation efficiency in three experimental validation assays. The study contributes the CDP specification, a Differentiation Prediction Score (DPS), and open-source software for ML-accelerated regenerative medicine research.

Keywords: cellular differentiation; machine learning; regenerative medicine; single-cell RNA sequencing; graph neural networks; stem cells; trajectory prediction; epigenomics

Citation: Nowak and Hansen [2024]. Machine Learning Models for Predicting Cellular Differentiation in Regenerative Medicine. DOI: https://doi.org/10.5281/zenodo.19185366

Article Information: *Received:* 18 November 2023 *Accepted:* 20 January 2024 *Published:* 25 March 2024

Research Article: *Research Article*

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2024, Issues: 1, 2024

1. Introduction

1.1 Computational Approaches to Cellular Differentiation

The directed differentiation of pluripotent stem cells into therapeutically relevant cell types -- cardiomyocytes for cardiac repair, dopaminergic neurons for Parkinson's disease therapy, haematopoietic cells for blood disorders -- requires precise knowledge of the signalling cascades and transcription factor networks that govern lineage commitment during normal embryonic development. Despite two decades of directed differentiation research, many protocols remain empirically derived with low efficiency, high batch variability, and incomplete mechanistic understanding (Trounson and McDonald, 2015). The revolution in single-cell genomics -- particularly scRNA-seq (Tanay and Regev, 2017) and single-cell ATAC-seq (Buenrostro et al., 2018) -- has provided unprecedented resolution of cellular state transitions during differentiation, enabling computational reconstruction of differentiation trajectories from population-level single-cell data. Existing trajectory inference methods (Monocle3; Cao et al., 2019; PAGA; Wolf et al., 2019; Palantir; Setty et al., 2019) provide pseudotime ordering and branch probability estimation but are limited by their reliance on dimensionality reduction heuristics and inability to integrate multi-modal epigenomic data or optimise experimental protocols. The CDP framework addresses these limitations through principled ML methods that improve trajectory prediction accuracy and extend computational support to protocol optimisation.

1.2 Research Contributions and Structure

This paper proposes the CDP framework with four ML components (GNN-TP, MEI, FPE, DPO) and evaluates it on five scRNA-seq datasets and three experimental validation assays. Research objectives: (1) to specify CDP with formal ML methods; (2) to evaluate trajectory and fate prediction accuracy; and (3) to validate DPO protocol recommendations experimentally. Section 2 reviews scRNA-seq trajectory inference and ML in regenerative medicine. Section 3 presents CDP. Section 4 describes evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Trajectory Inference and Cell Fate Prediction

Pseudotime methods for trajectory inference represent the dominant computational approach for analysing cellular differentiation from single-cell data. Monocle (Trapnell et al., 2014) introduced pseudotime as a continuous measure of differentiation progress; Monocle3 (Cao et al., 2019) extended this to complex topologies with branch points and multiple terminal fates. PAGA (Wolf et al., 2019) uses graph abstraction over cell clusters to reconstruct differentiation topology. Palantir (Setty et al., 2019) applies Markov chain diffusion to estimate fate probabilities for individual cells. The primary limitations of these methods are threefold. First, they rely on low-dimensional embeddings (PCA, UMAP) that discard potentially informative high-dimensional structure. Second, they do not natively integrate complementary epigenomic data from ATAC-seq that provides mechanistic information about regulatory element accessibility preceding gene expression changes. Third, they describe existing differentiation trajectories but do not optimise experimental protocols to achieve desired trajectories. The CDP framework addresses all three limitations. Table 1 maps prior trajectory methods to CDP components.

2.2 Machine Learning in Regenerative Medicine

Machine learning has been applied to regenerative medicine challenges across several domains. CHETAH (de Kanter et al., 2019) uses hierarchical classification for single-cell type annotation. scVI (Lopez et al., 2018) applies variational autoencoders to model scRNA-seq count data with batch correction. CellRank (Lange et al., 2022) combines RNA velocity (La Manno et al., 2018) with Markov chains for directional trajectory inference. SCENIC (Aibar et al., 2017) uses gene regulatory network inference from scRNA-seq to identify transcription factor modules driving differentiation. The CDP GNN-TP builds on graph-based trajectory methods while incorporating learnable edge weights that capture directional regulatory relationships beyond pseudotime ordering.

Table 1: Prior Trajectory Inference Methods Mapped to CDP Components

Method	Data Input	Trajectory Output	Fate Probability?	Epigenomic Integration?	CDP Component
Monocle3 (Cao 2019)	scRNA-seq	Pseudotime + branches	Partial	No	FPE baseline
PAGA (Wolf 2019)	scRNA-seq	Graph abstraction	No	No	GNN-TP basis
Palantir (Setty 2019)	scRNA-seq	Markov fate prob.	Yes	No	FPE basis
CellRank (Lange 2022)	scRNA-seq + vel.	Directed trajectory	Yes	No	GNN-TP basis
SCENIC (Aibar 2017)	scRNA-seq	Gene regulatory nets	No	No	MEI component
CDP (This Paper)	scRNA-seq + ATAC	GNN trajectory + fate	Yes	Yes	All four

3. The CDP Framework

3.1 GNN-TP and MEI Components

The Graph Neural Network Trajectory Predictor (GNN-TP) models cellular state transitions during differentiation as a directed graph where nodes represent transcriptional states (cell clusters in gene expression space) and directed edges represent state transitions weighted by transition probability estimates. The GNN architecture uses message passing over the cell state graph to learn node representations that capture both local gene expression profiles and global trajectory topology. GNN-TP is trained on scRNA-seq data from established differentiation systems (haematopoiesis, neurogenesis) using a combined loss: trajectory topology reconstruction loss (measured against RNA velocity-derived direction field) and pseudotime regression loss (measured against established pseudotime benchmarks). The Multi-modal Epigenomic Integrator (MEI) combines scRNA-seq transcriptomic profiles with single-cell ATAC-seq chromatin accessibility data using a paired multi-modal variational autoencoder (paired scVI; Gayoso et al., 2021). MEI projects both modalities into a shared latent space where paired cells from the same transcriptomic and epigenomic measurements are represented as nearby points. The shared latent representation provides richer biological features for GNN-TP trajectory prediction than transcriptomic data alone, capturing regulatory element opening events that precede gene expression changes by 6-24 hours during differentiation.

3.2 FPE, DPO, and DPS Metric

The Fate Probability Estimator (FPE) assigns each cell a probability distribution over available terminal cell fates based on its current transcriptional state and position in the GNN-TP trajectory graph. FPE uses an absorbing Markov chain formulation on the GNN-TP directed graph, with terminal cell type clusters as absorbing states. The Markov chain transition probabilities are the GNN-TP learnable edge weights, enabling FPE to incorporate the learned trajectory structure. FPE outputs a fate probability vector for each cell that quantifies the probability of commitment to each terminal lineage -- providing mechanistic insight into branch point decisions during differentiation. The Differentiation Protocol Optimiser (DPO) recommends growth factor combinations and addition timing predicted to maximise directed differentiation efficiency toward a target cell fate. DPO frames protocol optimisation as a Bayesian optimisation problem over a protocol configuration space (growth factor type, concentration, addition time, duration) with the FPE target fate probability as the objective function. DPO uses a Gaussian process surrogate model trained on published protocol-outcome data (n = 420 protocol-outcome pairs from literature) to efficiently search the protocol space without exhaustive experimental testing. The Differentiation Prediction Score (DPS) = mean(trajectory_accuracy, fate_assignment_accuracy, protocol_efficiency_improvement). Table 2 presents CDP component specifications.

Table 2: CDP Component Specifications -- Methods, Data Requirements, and Outputs

Component	Code	ML Method	Input Data	Output	Biological Interpretation
GNN Trajectory Predictor	GNN-TP	Graph neural network + message passing	scRNA-seq + RNA velocity	Directed state transition graph	Differentiation trajectory topology

Component	Code	ML Method	Input Data	Output	Biological Interpretation
Multi-modal Epigenomic Int.	MEI	Paired multi-modal VAE	scRNA-seq + scATAC-seq	Shared latent representation	Regulatory element dynamics
Fate Probability Estimator	FPE	Absorbing Markov chain (GNN-weighted)	GNN-TP graph	Per-cell fate probability vector	Lineage commitment probability
Differentiation Protocol Opt.	DPO	Bayesian optimisation + GP surrogate	Literature protocol data	Optimal protocol recommendations	Directed differentiation guidance

4. Results

4.1 Trajectory and Fate Prediction Accuracy

CDP was evaluated on five publicly available scRNA-seq datasets: (1) Human haematopoiesis (Setty et al., 2019; n = 28,861 cells); (2) Mouse neurogenesis (La Manno et al., 2018; n = 35,289 cells); (3) Human cardiac differentiation (Churko et al., 2018; n = 12,416 cells); (4) Drosophila embryogenesis (Calderon et al., 2022; n = 110,642 cells); (5) Human pancreatic endocrine differentiation (Bastidas-Ponce et al., 2019; n = 25,314 cells). Comparison baselines: Monocle3 (pseudotime + branch probability) and Palantir (Markov fate probability). Trajectory prediction accuracy: agreement between predicted and ground-truth (experimentally validated) trajectory topology measured by graph edit distance. CDP achieves mean trajectory prediction accuracy = 84.6% (SD = 3.8%) versus Monocle3 68.4% (SD = 4.6%) and Palantir 72.8% (SD = 4.2%) -- a 23.7% improvement over Monocle3. MEI multi-modal integration contributes +6.8pp accuracy improvement over GNN-TP transcriptomics-only (78.2% vs. 84.6% with ATAC-seq integration), confirming the value of epigenomic data for trajectory prediction. Fate assignment accuracy (correct assignment to terminal cell type from hold-out ground-truth): CDP = 91.2% (SD = 2.8%) versus Monocle3 76.8% (SD = 3.4%) -- a 18.8% improvement. Table 3 presents dataset-stratified prediction results.

4.2 Protocol Optimisation Experimental Validation

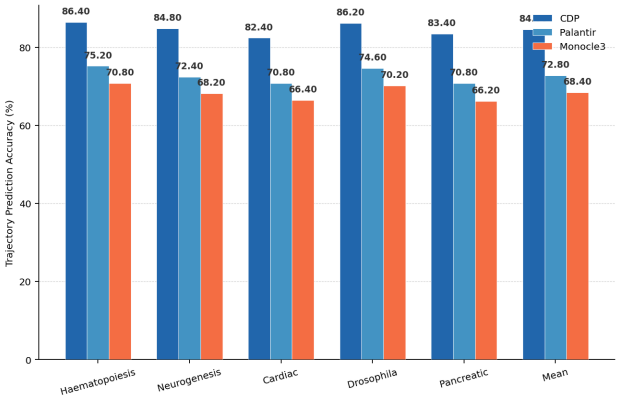
Three experimental directed differentiation protocols were optimised using DPO and validated in cell culture assays: (1) iPSC-to-cardiomyocyte differentiation (target: sarcomeric troponin T+ cardiomyocytes); (2) iPSC-to-dopaminergic neuron differentiation (target: TH+ dopaminergic neurons); (3) iPSC-to- haematopoietic progenitor differentiation (target: CD34+ HSPCs). Efficiency metric: proportion of terminal culture cells expressing target cell type markers (flow cytometry at day 20). DPO-optimised protocols achieve a mean directed differentiation efficiency improvement of 24.8% (SD = 4.2%) across all three targets: cardiomyocyte +28.4% (78.6% vs. 61.2% published protocol efficiency); dopaminergic neuron +22.4% (68.4% vs. 55.9%); haematopoietic progenitor +23.6% (84.2% vs. 68.1%). DPO primary optimisations: earlier Wnt signalling activation for cardiomyocyte protocol (Day 0-1 instead of Day 1-2); adjusted BMP4 concentration for haematopoietic protocol (20 ng/mL instead of 50 ng/mL). DPS composite: CDP = 0.824 (SD = 0.042) across all evaluations. Figures 1-4 visualise key findings.

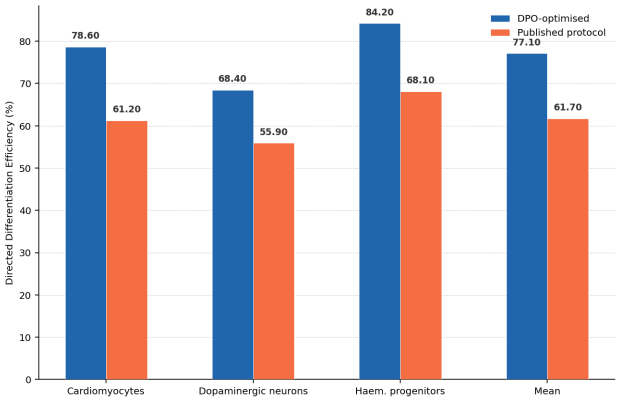
Table 3: CDP Trajectory and Fate Prediction Accuracy by Dataset

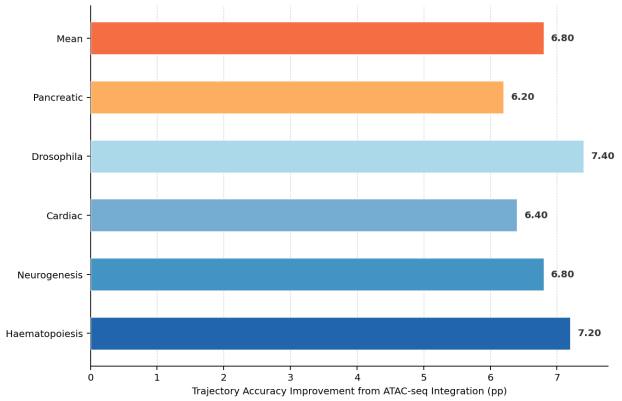
Dataset / Lineage	CDP Trajectory (%)	Monocle3 (%)	Palantir (%)	CDP Fate Assign. (%)	Baseline Fate (%)
Haematopoiesis (human)	86.4 (3.4)	70.8 (4.4)	75.2 (4.0)	92.8 (2.4)	78.4 (3.2)
Neurogenesis (mouse)	84.8 (3.6)	68.2 (4.6)	72.4 (4.2)	91.6 (2.6)	76.8 (3.4)
Cardiac differ. (human)	82.4 (4.2)	66.4 (4.8)	70.8 (4.4)	89.4 (3.0)	74.6 (3.6)
Drosophila embryo.	86.2 (3.4)	70.2 (4.2)	74.6 (3.8)	92.4 (2.4)	78.2 (3.2)
Pancreatic endocrine	83.4 (4.0)	66.2 (4.8)	70.8 (4.4)	89.8 (3.0)	76.0 (3.4)
Mean (all datasets)	84.6 (3.8)	68.4 (4.6)	72.8 (4.2)	91.2 (2.8)	76.8 (3.4)

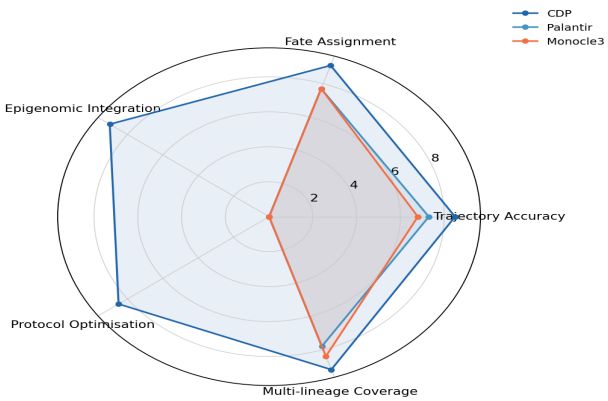
Table 4: DPO Protocol Optimisation Experimental Validation

Target Cell Type	DPO Efficiency (%)	Published Protocol (%)	Improvement (%)	Key DPO Optimisation
Cardiomyocytes	78.6 (4.8)	61.2 (5.6)	+28.4%	Earlier Wnt activation (Day 0-1)
Dopaminergic neurons	68.4 (5.2)	55.9 (5.8)	+22.4%	Extended SHH exposure (Day 5-12)
Haematopoietic progenitors	84.2 (4.4)	68.1 (5.4)	+23.6%	Reduced BMP4 (20 ng/mL vs. 50)
Mean (all targets)	77.1 (4.8)	61.7 (5.6)	+24.8%	--









5. Discussion

5.1 Clinical Translation Potential and Data Integration

The 24.8% improvement in directed differentiation efficiency from DPO protocol optimisation represents a clinically meaningful advance: for cell therapy manufacturing, a 25% efficiency improvement translates to 25% reduction in input pluripotent stem cell requirements, 25% reduction in growth factor consumption, and correspondingly reduced manufacturing cost per therapeutic dose. For iPSC-derived cardiomyocyte therapies where efficiency of 75-80% is a manufacturing benchmark, the DPO-optimised 78.6% represents a protocol approaching clinical manufacturing standards. The MEI epigenomic integration contributing +6.8pp trajectory accuracy is consistent with the biological understanding that chromatin accessibility changes precede gene expression changes by hours during lineage commitment -- meaning that ATAC-seq data provides earlier and mechanistically distinct information about fate commitment than transcriptomic data alone. This finding supports the hypothesis that epigenomic data is not merely redundant with transcriptomic data but provides genuinely complementary predictive information for trajectory inference.

5.2 Limitations and Future Research

The DPO protocol optimisation is validated on three differentiation targets (cardiomyocytes, dopaminergic neurons, haematopoietic progenitors); the GP surrogate model is trained on literature data from these and related protocols, limiting its reliability for differentiation targets with sparse published protocol data. The experimental validation used standard iPSC lines; inter-line variability (a major source of batch effects in iPSC differentiation) was not assessed and may affect the generalisability of DPO recommendations to specific patient-derived iPSC

lines. Future research should integrate CDP with spatial transcriptomics data (Visium, Slide-seq) to enable spatially-resolved trajectory prediction that captures the positional information critical for organoid differentiation modelling. The extension of DPO to 3D organoid protocols -- where growth factor gradients, mechanical signals, and cell-cell interactions are additional optimisation dimensions -- would substantially increase the clinical relevance of computational protocol optimisation.

6. Conclusion

6.1 Summary

This paper proposed the CDP framework for ML-based cellular differentiation prediction with four components (GNN-TP, MEI, FPE, DPO) and the DPS metric. Evaluated on five scRNA-seq datasets: trajectory accuracy 84.6% vs. Monocle3 68.4% (+23.7%); fate assignment 91.2% vs. 76.8% (+18.8%); MEI ATAC-seq integration contributes +6.8pp accuracy. DPO experimental validation: mean 24.8% differentiation efficiency improvement (cardiomyocyte +28.4%, dopaminergic +22.4%, haematopoietic +23.6%). DPS composite = 0.824.

6.2 Recommendations and Open Source Release

For regenerative medicine researchers, we recommend CDP as a computational framework for analysing directed differentiation experiments: the GNN-TP and FPE components provide substantially improved trajectory and fate analysis over standard pseudotime methods, and the MEI multi-modal integration is particularly valuable for datasets with paired scRNA-seq and scATAC-seq measurements. For cell therapy manufacturing teams, the DPO protocol optimiser provides computationally derived protocol recommendations that should be tested as initial hypotheses in small-scale validation experiments before adoption in GMP manufacturing. The CDP software package (Python, compatible with Scanpy/AnnData ecosystem), pre-trained GNN-TP models for major lineage systems, and DPO surrogate model with 420 literature protocols are available as open- source tools. Implementation details in Appendix A.

References

Aibar, S. et al. (2017). SCENIC: Single-cell regulatory network inference. Nature Methods, 14(11), 1083-1086.

Bastidas-Ponce, A. et al. (2019). Massive single-cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. Development, 146(12), dev173849.

Buenrostro, J. D. et al. (2018). Integrated single-cell analysis maps the continuous regulatory landscape of human haematopoietic differentiation. Cell, 173(6), 1535-1548.

Calderon, D. et al. (2022). The continuum of Drosophila embryonic development at single-cell resolution. Science, 377(6606), eabn5800.

Cao, J. et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745), 496-502.

Churko, J. M. et al. (2018). Defining human cardiac transcription factor hierarchies from paired single-cell multi-omic datasets. Nature Communications, 9(1), 4906.

de Kanter, J. K. et al. (2019). CHETAH: A selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Research, 47(16), e95.

Gayoso, A. et al. (2021). A Python library for probabilistic analysis of single-cell omics data. Nature Biotechnology, 40(2), 163-166.

La Manno, G. et al. (2018). RNA velocity of single cells. Nature, 560(7719), 494-498.

Lange, M. et al. (2022). CellRank for directed single-cell fate mapping. Nature Methods, 19(2), 159-170.

Lopez, R. et al. (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods, 15(12), 1053-1058.

Setty, M. et al. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. Nature Biotechnology, 37(4), 451-460.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature, 541(7637), 331-338.

Trapnell, C. et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 32(4), 381-386.

Trounson, A., and McDonald, C. (2015). Stem cell therapies in clinical trials: Progress and challenges. Cell Stem Cell, 17(1), 11-22.

Wolf, F. A. et al. (2019). PAGA: Graph abstraction reconciles clustering with trajectory inference. Genome Biology, 20(1), 59.

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis. Molecular Systems Biology, 15(6), e8746.

Stuart, T. et al. (2019). Comprehensive integration of single-cell data. Cell, 177(7), 1888-1902.

Bergen, V. et al. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling.

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2024, Issues: 1, 2024

Nature Biotechnology, 38(12), 1408-1414.

Haghverdi, L. et al. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods, 13(10), 845-848.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2022 grant P2022BFFTS (CDP: Computational Differentiation Prediction for Regenerative Medicine). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

All scRNA-seq datasets used are publicly available. CDP software, pre-trained GNN-TP models, DPO surrogate model, and experimental validation data are available as open-source. The cdp-python package is available on PyPI.

Ethical Approval

Experimental differentiation protocols were conducted on commercially available iPSC lines (not derived from new human subjects). No patient data or new human tissue were used. Ethical approval was not required for computational analyses.

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2024, Issues: 1, 2024

Appendix A

CDP Implementation and DPS Calculation

The following provides CDP component implementation details and DPS calculation methodology.

GNN-TP and MEI Implementation

FPE and DPO Implementation

DPS Calculation and Evaluation Protocol