SteMClass: A Novel DNA Methylation-Based Classifier for iPSC In Vitro Differentiation States

Eilís Perez[1,2]*, Valeria Fernandez Vallone[3], Daniel Teichmann[1], Philipp Euskirchen[1,4], Christin Siewert[4], Brendan Osberg[4], Peggy Wolkenstein[1], Harald Stachelscheid[3†], David Capper[1,4†]

[1] German Cancer Consortium (DKTK), Partner Site Berlin, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin School of Integrative Oncology (BSIO), Berlin, Germany

[3] Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Core Unit Pluripotent Stem Cells and Organoids, Berlin, Germany

[4] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institut für Neuropathologie, Berlin, Germany

†These authors contributed equally as co-lead authors.

*Correspondence: eilis.perez@charite.de

Abstract

Human induced pluripotent stem cells (iPSCs) hold great promise for regenerative medicine, disease modelling, and drug discovery, but most downstream applications require differentiation into specialised cell types not covered by current quality control assays. Here, we present “SteMClass”, a proof-of-concept DNA methylation-based classifier that standardises iPSC differentiation state identification across protocols with one test. We curated a reference cohort of 15 iPSC lines differentiated into seven distinct states (n = 97), performed array-based DNA methylation profiling, and trained a random forest model to classify the eight distinct differentiation states. In nested cross-validation, SteMClass achieved a Brier score of 0.0264, and on an independent cohort (n = 58) attained 96.5% accuracy (Cohen’s K = 0.959) with a 3% rejection rate. Applied to external data (n = 241), performance was 76.5% accuracy (Cohen’s K = 0.593) with a 16.6% rejection rate and enabled the detection of potentially inefficient differentiations. SteMClass is compatible with all Illumina methylation array versions, and accessible via an interactive web interface that supports classification and exploration of DNA methylation profiles. By providing a harmonised, single-assay framework for iPSC-derived differentiation state characterisation, SteMClass improves reproducibility and comparability across studies, paving the way for robust quality control standards and accelerating clinical translation.

Introduction

Due to their potential for self-renewal and to differentiate into all cell types of the human body, human induced pluripotent stem cells (iPSCs) hold immense promise for regenerative medicine, drug screening and disease modelling (Shi et al., 2017). Despite the potential of iPSCs, there are significant challenges to address before routine clinical use.

Key limitations include safety concerns regarding genomic stability, immunogenicity and the scalability and reproducibility of differentiation and characterisation methods. Many recent developments have been made in the field of iPSC production in order to combat these issues, such as the development of universal cells that are camouflaged from the immune system (Lanza et al., 2019), the use of small molecules to replace growth factors (Liuyang et al., 2023; Wang et al., 2024) and improved quality control of iPSC production such as genetic screening, and potency assessment (Dobner et al., 2024; Lenz, 2012, 2012; MacArthur et al., 2019; Tsankov et al., 2015; Yoshihara et al., 2017). Despite the advancements in iPSC production and standardisation, scalable and reproducible differentiation of iPSCs into standardised, well-defined cell types remains a key challenge.

In recent years, large cell and tissue atlases have been generated using gene expression and DNA methylation profiling to identify distinct cell populations in vivo (Loyfer et al., 2023; Moss et al., 2018). However, cells generated by the differentiation of iPSCs in vitro are inherently distinct from their terminally differentiated in vivo counterparts due to their developmentally immature state, in vitro cell culture associated changes, reprogramming artefacts, and the persistence of somatic cell-specific epigenetic signatures (Cerneckis et al., 2024; Diane et al., 2025; Edgar et al., 2022; Franzen et al., 2021). Conventional assessments of iPSC differentiation generally depend on measuring a limited panel of lineage-characteristic gene or protein markers. Because the choice and interpretation of these markers vary among experiments and laboratories, this strategy is susceptible to considerable variability across experiments and laboratories (Grancharova et al., 2021; Kramer et al., 2016). Dobner et al. (2024) recently demonstrated the limitations of conventional gene-expression markers in distinguishing germ-layer identity. Using long-read nanopore transcriptomics on directed trilineage-differentiated samples from 15 iPSC lines, they showed that even the markers endorsed by the International Society of Stem Cell Research (ISSCR) overlap significantly across differentiation states. Hence, the authors devised and validated a concise 12-gene signature, which they incorporated into a machine-learning classifier for assigning early lineage identities (Dobner et al., 2024; Ludwig et al., 2023).

68 DNA methylation is tightly regulated and varies according to cell type with dynamic
69 alterations playing a critical role in guiding cell fate and development. The epigenetic patterns
70 that define the gene expression profiles of differentiated cells are established during
71 development and maintained through subsequent cell divisions. As such, DNA methylation
72 serves as a more consistent and enduring marker of a cell's identity than gene expression,
73 which can be transient and subject to rapid changes in response to external stimuli (Bird,
74 2002; Gibney & Nolan, 2010; Jones & Taylor, 1980; Tompkins, 2022). The process of DNA
75 methylation, carried out by DNA methyltransferase (DNMT) enzymes, and the turnover of
76 methylation by TET enzymes are fundamental as cells exit pluripotency (Parry et al., 2021).
77 Therefore, understanding the dynamics of DNA methylation changes through differentiation
78 is key to improving differentiation efficiency and the standardisation of producing iPSCs and
79 their differentiation derivatives. Research indicates that the efficiency of iPSC differentiation
80 and the stability of differentiated cell populations are influenced by DNA methylation
81 (Buckberry et al., 2023; Hajmousa et al., 2024; Ideno et al., 2022; Tesarova et al., 2016; Yang
82 et al., 2023) and that DNA methylation patterns can serve as markers to identify iPSC lines
83 with varying differentiation capacities (Butcher et al., 2016; Choi et al., 2020; Edgar et al.,
84 2022; Nishizawa et al., 2016; Schmidt et al., 2022).

85 While whole-genome bisulfite sequencing (WGBS) provides a comprehensive overview of the
86 methylome, array-based methods, such as the Illumina Infinium assay, offer a cost-effective
87 alternative for large-scale studies, enabling quantitative analysis of hundreds of thousands of
88 curated CpG sites with high reproducibility and minimal inter-user variability (Bibikova et al.,
89 2006; Capper et al., 2018; Chirica et al., 2024; Kurdyukov & Bullock, 2016; Tran et al., 2024).
90 Leveraging this platform's scalability in conjunction with machine learning has led to the
91 development of robust classification tools for several types of tumours (Capper et al., 2018;
92 Danielsson et al., 2015; Dragomir et al., 2023; Hoang et al., 2024; Jurmeister et al., 2022; Tran
93 et al., 2024).

94 Within this framework of classification tool development, "harmonisation" encompasses the
95 standardisation of methodologies, reference datasets, analytical tools, and interpretive
96 criteria, ensuring the DNA methylation data are consistent and comparable across
97 laboratories, studies, and clinical settings. Such harmonisation, already successful in refining
98 protocols and establishing machine learning-based classification tools in the field of brain
99 tumours, ensures reliable classification, robust comparisons, and meaningful conclusions
100 while minimising technical bias (Capper et al., 2018). As a result, brain tumour diagnostics

have been transformed by a classifier built on these principles, culminating in their formal adoption by the WHO (Louis et al., 2021).

In this work, we bring these concepts to the iPSC research field by demonstrating that DNA methylation profiling combined with machine learning can accurately identify cells at defined stages of iPSC differentiation. Our method leverages a centrally curated, publicly available reference dataset used to train a random forest classification tool we call SteMClass.

Researchers can upload their own test data via an online interface (https://pereze5-stemclass.share.connect.posit.cloud/) to compare their test sample to the SteMClass reference data. This unified DNA methylation-based assay can classify multiple cell types, eliminating the need for numerous targeted tests for differentiation success assessment. This enhances efficiency and enables direct comparisons across iPSC studies and differentiation protocols with a single assay. Ultimately, this tool promotes harmonisation in iPSC research and elevates the reproducibility and interpretability of findings in this rapidly advancing field.

Keywords: DNA methylation, induced pluripotent stem cells, differentiation, epigenetics, regenerative medicine.

Materials and Methods

Cell Culture and Differentiation experiments

Culture of iPSCs

Human induced pluripotent stem cells (iPSCs) were maintained on growth factor reduced Geltrex (Gibco) in E8 or mTeSR (Stem Cell Technologies) media with daily media exchange. Cells were enzymatically clump-passaged when >70% confluency was reached, usually every 2–4 days using 0.5mM EDTA (UltraPure™ EDTA, Thermo Fisher). Supplementary Tables 1-3 compiles the iPSC lines used in this study, and their maintenance culture conditions, detailed protocol information is available on Protocols.io (Vallone et al., 2025). All iPSC lines were provided by Berlin Institute of Health Core Unit Pluripotent Stem Cells and Organoids (BIH-CUSCO) and have ethical approval for their use.

Differentiation Protocols

Cell type nomenclatures, iPSC lines and corresponding differentiation protocols used in this study are described in Supplementary Tables 1-3 and Supplementary Figure 1.

Differentiation to Ectoderm

iPSCs were dissociated into single cells using TrypLE (Gibco) and seeded at a density of 5×10^5 cells per well in a 6-well plate (5.2×10^4 cells/cm^2) coated with 1:120 dilution of Geltrex™

133 (Gibco) in Essential 8™ Medium supplemented with 10 µM Y-27632 (Selleck Chemicals). Cells
134 were cultured for 2 days with daily medium changes using Essential 8™ Medium. After 2 days,
135 the medium was replaced with neural induction medium composed of DMEM/F12
136 supplemented with 2.5 mM L-glutamine, 15 mM HEPES, 1× B-27 Supplement, 1× N-2
137 Supplement (all from Thermo Fisher Scientific), 10 µM dorsomorphin (Tocris Bioscience), and
138 10 µM SB431542 (Tocris Bioscience) (Supplemental Figure 1c). Medium was replenished
139 every 24 hours for 7 days, after which cells were characterised using fluorescence-activated
140 cell sorting (FACS) to assess the expression of the markers PAX6 (Miltenyi Biotec, 130-123-
141 328), Nestin (Miltenyi Biotec, 130-119-799), SOX1 (Miltenyi Biotec, 130-111-042), and SOX2
142 (Miltenyi Biotec, 130-120-721). Samples with >85% positive expression for these markers
143 were harvested and stored as cell pellets at -80 °C. These ectoderm day 7 were used as
144 starting points to differentiate Neural Stem Cells and Astrocytes.

145 **Differentiation to Neural Stem Cells**

146 iPSCs were differentiated into ectodermal cells as described above for 7 days, after which cells
147 were analysed by FACS for the markers PAX6 and SOX2 (Supplemental Figure 1c). Batches
148 with >80% marker expression were dissociated into single cells and seeded at a density of
149 approximately 15,000 cells/cm^2 onto plates coated with either Geltrex™/Matrigel® (1:120
150 dilution in DMEM), or vitronectin (1:100 dilution in PBS without Ca^{2+}, Mg^{2+}) in Neural
151 Induction Medium composed of DMEM/F12 supplemented with 2.5 mM L-glutamine, 15 mM
152 HEPES, 1× B-27 Supplement, 1× N-2 Supplement (all from Thermo Fisher Scientific), 10 µM
153 dorsomorphin (Tocris Bioscience), and 10 µM SB431542 (Tocris Bioscience). Medium was
154 replenished every 24 hours for 18 days, after which cells were assessed morphologically,
155 harvested and stored as pellets at -80 °C.

156 **Differentiation to Astrocytes**

157 Astrocyte progenitor cells were generated from iPSCs as previously described (TCW et al.,
158 2017) (Supplemental Figure 1c). Briefly, iPSCs were differentiated into ectodermal cells as
159 described above for 7 days, then dissociated into single cells and seeded at a density of
160 15,000 cells/cm^2 onto plates coated with Geltrex™/Matrigel® (1:120 dilution in DMEM), or
161 vitronectin (1:100 dilution in PBS without Ca^{2+}, Mg^{2+}) in Neural induction medium
162 supplemented with 10 µM Y-27632. Medium was changed every 24 hours for 48 hours,
163 followed by replacement with complete Astrocyte Medium (ScienCell Research Laboratories,
164 Cat# 1801). Medium was changed every 48 hours, and when cells reached 90-95% confluency
165 (every 6–7 days), they were passaged to the initial seeding density of 15,000 cells/cm^2. After
166 30 days, cells were dissociated, and GLAST-positive cells were isolated using the Anti-GLAST
167 (ACSA-1) MicroBead Kit and a MiniMACS™ Separator (Miltenyi Biotec, 130-095-825),

168 following manufacturer´s instructions. Isolated GLAST-positive cells were stored as pellets at
169 -80 °C.

170 **Differentiation to Endoderm**

171 Before endoderm differentiation iPSC lines which were not cultured in mTeSR^{TM}1 (Stem Cell
172 Technologies) on Geltrex (Gibco) coated plates, were adapted for at least 2 passages to these
173 conditions. 70-80% confluent iPSC were harvested as single cells using Accutase (Gibco).
174 Three million cells per well of 6WP were seeded on Geltrex coated surface in mTeSR^{TM}1
175 supplemented with 10µM Y-27632 (Stem Cell Technologies). After 24 hrs, endoderm
176 differentiation was started using the STEMdiff™ Definitive Endoderm Differentiation Kit (Stem
177 Cell Technologies) following manufacturer´s instructions. This day was considered day 0. At
178 day 3 cells were harvested using Accutase and immunostained for FACS analysis using CD184
179 (CXCR4) (Miltenyi Biotec, 130-120-708) and SOX17 (Miltenyi Biotec, 130-111-031). Definitive
180 endoderm differentiation with >90% efficiency were used to derive lung progenitors and a
181 subfraction of the cells were stored at -80°C as a pellet for DNA isolation and downstream
182 application.

183

184 **Differentiation to Lung Cells**

185 iPSC derived definitive endodermal cells were derived as described above and further
186 differentiated to lung progenitors according to Jacob et al. (Jacob et al., 2019) (Supplementary
187 Figure 1a). Briefly, day 3 endoderm cells harvested as clumps using Gentle Dissociation
188 reagent (Stem Cell Technologies) were seeded in 6WP as 1:4 and 1:6 ratios in cSFDM basal
189 medium (described in Jacob et al., 2019) supplemented with 10µM SB431542 (Sigma), 2µM
190 Dorsomorphin and 10µM Y-27632 (Stem Cell Technologies). After 24 hrs, medium was
191 changed removing Y-27632 (day 4). Anterior foregut differentiation continued another 48hs
192 without media change. At day 6, medium was changed to lung progenitor induction medium
193 conformed by cSFDM basal medium supplemented with 3µM CHIR (EZSolution Avantor),
194 10ng/ml hBMP4 (Prepotech) and freshly added 100nM retinoic acid (Sigma). Medium was
195 changed every 48 hrs until day 15. Lung progenitors were isolated based on Carboxypeptidase
196 M (CPM) expression using anti human-CPM antibody (WAKO Fujifilm, 014-27501) and
197 subsequent magnetic separation using Anti-mouse IgG2a+b Microbeads (Miltenyi Biotec,
198 130-047-201), MACS separation columns (Miltenyi Biotec, 130-042-201) and MiniMACS
199 Separator (Miltenyi Biotec). CPM positive cells were stored as cell pellet at -80°C.

200 **Differentiation to Mesoderm: Endothelial and Pericyte Progenitor Cells**

201 Mesoderm and subsequent endothelial cell differentiation was adapted from Patsch et al.
202 (Patsch et al., 2015) (Supplementary Figure 1b). Briefly, 70-80% confluent iPSC were
203 harvested as single cells using Accutase (Gibco). Four hundred thousand cells were seeded

204 per well of 6WP on Geltrex coated surface in E8 or mTeSR media (hiPSC line dependent)
205 supplemented with 10µM Y-27632 (Stem Cell Technologies). After 24 hrs, at day 0, medium
206 was exchanged for 4ml per well of Endo Priming medium composed by 50% DMEM/F12
207 (Thermo Fisher, 11320-033) and 50% Neurobasal medium (Thermo Fisher, 21103049)
208 supplemented with B27 without Vitamin A (Gibco), N-2 (Gibco), 50µM Beta-mercapto-
209 ethanol (Gibco), GlutaMAX (Gibco), 6µM CHIR (EZSolution Avantor), and 25ng/ml hBMP4
210 recombinant (Prepotech). No medium change was performed for 72 hrs. At day 3, medium
211 was exchanged for Endo Induction medium consisting of StemPro-34 Basal medium and
212 supplements (Gibco), GlutaMAX (Gibco), 200ng/ml human recombinant VEGF (A) 165
213 (Prepotech) and 2µM Forskolin (Abcam). Cells were cultured for 48 hrs. At day 5, cells were
214 harvested as single cells using Acutasse (Gibco). CD144 was used as surface marker to assess
215 differentiation efficiency (Miltenyi Biotec, 130-100-742) by FACS and to isolate endothelial
216 progenitors using magnetic MicroBeads (Miltenyi Biotec, 130-097-857), MACS separation
217 columns (Miltenyi Biotec, 130-042-201) and MiniMACS Separator (Miltenyi Biotec), following
218 manufacturer´s instructions. Cells belonging to the CD144 positive and negative fraction
219 (Mesoderm) were kept as pellets at -80°C. Cells belonging to the CD144 positive fraction were
220 used as the starting population for endothelial cell differentiation. Cells of the negative
221 fraction are considered pericyte progenitor cells.

222 **Differentiation to Endothelial Cells**

223 Endothelial Progenitor cells (CD144+) obtained as described above were seeded on Gelatin
224 1% (Millipore) coated T25 flasks in complete EGM™-2 Endothelial Cell Growth Medium-2
225 BulletKit™ (Lonza). Between day 12-14 when cells were 90% confluent, they were harvested
226 as single cells using Acutasse (Gibco). FACS for CD31 (Miltenyi Biotec, 130-092-652) and
227 CD144 (Miltenyi Biotec, 130-100-742) expression was performed. In most cases, >90% CD31+
228 were observed. Cultures with lower differentiation efficiency were subjected to CD31+ cells
229 magnetic enrichment using equipment described above and CD31 Microbeads (Miltenyi
230 Biotec, 130-091-935). CD31+ endothelial cells were stored as pellet at -80°C.

231 **Trilineage Differentiation**

232 70-80% confluent iPSCs were seeded and differentiated into Ectoderm, Endoderm and
233 Mesoderm using the StemMACS™ Trilineage Differentiation Kit (Miltenyi Biotec), following
234 manufacturer´s instructions. At day 7 of differentiation cells were harvested using Accutase
235 or TrypLE (both Gibco). FACS analysis was performed to characterise ectodermal (SOX2,
236 PAX6), endodermal (CXCR4, SOX17) and mesodermal (CD144, CD140b -Miltenyi Biotec, 130-
237 121-052-) cells. The remaining cells were collected as pellet and stored at -80°C.

238 ## Immunostaining for FACS

239 150,000-300,000 single cells were used per tube. Surface marker staining was performed in
240 FACS buffer: DPBS without Calcium and Magnesium (Gibco), BSA 0.5% (Miltenyi Biotec) and
241 2mM EDTA (Gibco) for 10 minutes at 4°C. After 2 rounds of washing cells were fixed and
242 permeabilized using FoxP3 Staining Buffer Set (Miltenyi Biotec), incubations were done for 30
243 minutes at 4°C. After washes using Perm/wash buffer cells were incubated with antibodies
244 against intracellular antigens for 60 minutes at 4°C. After final washes cells were resuspended
245 in 250µl FACS buffer. Flow cytometry acquisition was performed using MACSQuant VYB
246 (Miltenyi Biotec) and analysis using FCS Express 5 Plus software.

247 ## DNA Extraction

248 DNA extraction was performed on harvested cell pellets using the Maxwell® RSC Instrument
249 (Promega) with either the Maxwell® RSC Cultured Cells DNA Kit (Cat# AS1620) or the
250 Maxwell® RSC Blood DNA Kit (Cat# AS1400) for fresh and frozen cell samples, and the
251 Maxwell® RSC FFPE Plus DNA Purification Kit (Cat# AS1720) for paraformaldehyde (PFA)-fixed
252 cell samples, with no difference in DNA methylation analysis quality observed. Extracted DNA
253 quantities were measured using the Quantus system together with the "QuantiFluor® ONE
254 dsDNA" kit (Promega).

255 ## DNA Methylation Analysis

256 A cohort of n=155 iPSC derivates of eight differentiation states was generated in technical
257 (independent experiments) or biological (distinct cell line donors) replicates, then randomly
258 split into a training cohort of n=97 with a minimum of 8 samples per differentiation state, and
259 a validation cohort of n=58. DNA extraction was performed on each sample cell pellet for
260 subsequent DNA methylation profiling.

261 ## Array-Based DNA Methylation Profiling

262 Genomic DNA was bisulfite-converted using the EpiTect Fast Bisulfite Kit (Qiagen) following
263 the manufacturers' protocol. The bisulfite-converted DNA was subsequently analysed using
264 the Infinium MethylationEPIC BeadChip array (v1.0 or v2.0) and scanned on the iScan System
265 (Illumina). Raw DNA methylation data were pre-processed using R (v 4.3.3). Signal intensities
266 were obtained from IDAT files using the Bioconductor package minfi (v 1.40.0) (Aryee et al.,
267 2014). Dye bias correction was performed using the Noob (normal-exponential out-of-band)
268 method with the preprocessNoob function without "between-array normalisation" (Triche et
269 al., 2013). Sex prediction of samples was performed using the predictSex function of the
270 wateRmelon package (v 1.36.0). Beta values for DNA methylation were calculated from
271 normalised data using the minfi getBeta function.

Analysis of Publicly Available Datasets

A publicly available methylation array dataset was systematically curated and downloaded from the following Gene Expression Omnibus (GEO) records as raw signal intensity files (IDAT files): GSE110544, GSE163324, GSE105093, GSE141521, GSE159479, GSE163322, GSE180404, GSE197723, GSE207076, GSE207108, GSE236202, GSE247427, GSE247428, GSE247428, GSE253190, GSE253190, GSE61461, GSE85828. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; (Page et al., 2021)) flow chart illustrating the number of datasets identified, included and excluded, and the reason for exclusions is available in Supplementary Figure 6. All samples used in this study are outlined in Supplementary Table 5. GenomicMethylSet objects were generated from raw signal intensity files using the readGEORawFile function of the GEOquery package (v2.62.1). (Davis & Meltzer, 2007).

Data filtering steps included the removal of probes with detection p-values greater than 0.05 in more than 10% of samples, probes containing single nucleotide polymorphisms (SNPs) at the CpG site, probes located on sex chromosomes, and cross-reactive probes as previously described (Capper et al., 2018; Chen et al., 2013). Raw data from EPIC arrays were merged with publicly available 450K array datasets using the combineArrays function in minfi. The combined dataset underwent the same filtering and normalisation process as described above.

Differential Methylation Analysis

Differential methylation analysis on the CpG site and region level was performed in a one-to-one fashion between iPSCs and the respective differentiated cell states as previously described (Maksimovic et al., 2017). Differentially methylated positions (DMPs) were visualized as volcano plots using the package EnhancedVolcano (v 1.20.0). Methylation-specific gene ontology (GO) analysis of promoter-associated differentially methylated positions (DMPs) was conducted for each cell state using the gometh function within the missMethyl package (v 1.28.0) (Phipson et al., 2016). Investigation beyond the scope of the classifier i.e. the “broader profiles” was done by identifying the top 20,000 CpGs from our DMP analysis that distinguished each reference differentiated cell state from iPSC, for iPSC it was the top 20,000 CpGs that were common to all cell state vs iPSC comparisons. The annotated tables for the top 20,000 CpGs per cell state are available in Supplementary Table 4, full DMP analysis results tables are available on request.

Unsupervised Analysis

Principal component analysis (PCA) was performed on DNA methylation beta values to visualise methylation differences across samples during iPSC to astrocyte differentiation (n=35). For each CpG probe, the standard deviation across samples was calculated, and the

307 5,000 most variable probes were selected. The resulting beta value matrix was transposed
308 such that rows represented samples and columns represented CpG sites. PCA was conducted
309 using the prcomp function in R with mean centring and scaling enabled. The proportion of
310 variance explained by each principal component was calculated as the squared standard
311 deviation of each component divided by the total variance. The first two principal
312 components were used for plotting. Uniform Manifold Approximation and Projection (UMAP)
313 was performed on the top 50,000 most variable CpGs for each respective dataset using the R
314 package umap (v 0.2.10.0). UMAP projections were computed using the uwot R package (v
315 0.2.2). UMAP parameters were set to n_neighbors=18, min_dist=0.3 n_components=2, with
316 a Euclidean distance metric and a fixed random state for reproducibility. The resulting UMAP
317 layout was integrated with sample metadata for downstream analysis. A scatter plot of UMAP
318 coordinates was generated using ggplot2, with points representing individual samples, the
319 final plots depicting the distribution of samples and cell types based on the top 50,000 most
320 variable probes across samples.

321 Classifier development

322 A random forest classifier was built with ranger (v 0.16.0) within the tidymodels ecosystem
323 (tidymodels v 1.2.0; parsnip v 1.2.1; recipes v 1.0.10; workflows v 1.1.4; tune v 1.2.1; rsample
324 v 1.2.1; yardstick v 1.3.0) under R 4.3.3. Parallel execution used BiocParallel (v 1.36.1) and
325 doParallel (v 1.0-17).

326 Feature selection

327 Differential-methylation analyses were performed for each of the seven derivative cell states
328 against iPSCs (adjusted $P < 0.05$). For every CpG the smallest FDR across comparisons was
329 retained, the CpGs were ranked within cell state, and CpGs were sampled proportionally to
330 list length to assemble a panel of 10 000 sites. This ranked-allocation procedure ensured that
331 all cell states contributed to the predictor set proportionally, preserving the most significant
332 markers while reflecting the true DMP burden of each cell state.

333 Nested cross-validation and parameter tuning

334 For training the random forest model a nested cross-validation (CV) was used: An outer loop
335 with 5-fold stratified CV repeated 5 times for performance estimation and an inner loop with
336 3-fold stratified CV within each outer split for hyper-parameter tuning. A Latin-hypercube
337 search of 20 combinations was carried out over minimum node size (min_n) and number of
338 trees (trees). The number of variables sampled at each split (mtry) was fixed to the square
339 root of the predictor count, i.e. 100. Model selection used Brier score (brier_class). Because
340 the eight classes differed by at most a factor of two, no additional re-sampling or class
341 weighting was applied, repeated fold stratification was deemed sufficient.

342 **Threshold-based rejection**
343 Raw class probabilities from the outer folds were pooled and a probability cut-off was
344 optimised across thresholds 0.50-0.95 (step 0.05) using the macro-Youden index (sensitivity
345 + specificity - 1). Predictions with maximum class probability below this value were assigned
346 to a "Not Classifiable" category.

347 **Final model and variable importance**
348 The best hyper-parameters (min_n = 3, trees = 1570, mtry = 100) were used to fit the
349 definitive random forest model on the entire training cohort. Variable importance was
350 extracted from the fitted ranger object (impurity measure). The pre-trained random forest
351 classifier was saved as a serialised workflow object.

352 **Model evaluation**
353 Performance was assessed using two independent datasets:

354 1. Internal validation test cohort (n = 58)

355 2. Public test cohort (n = 241, publicly available methylation array data), where class
356 labels were predicted after harmonising feature columns.

357 The pre-trained random forest classifier was loaded from the serialised workflow object and
358 the underlying ranger model was extracted. The preprocessing recipe was reconstructed
359 using the recipes package and applied to generate a numeric feature matrix and factor
360 response vector. Test data were aligned to the training feature space by imputing missing
361 CpGs with the mean beta value calculated from the training cohort dataset. Feature-level
362 contributions for each test sample and a training set bias vector were computed from the tidy
363 forest object. Samples from the validation cohorts which received a classification discordant
364 with their annotation were identified. For each such sample, the top CpG features with the
365 most negative contribution toward the annotated class and the most positive contribution
366 towards the SteMClass classification (i.e., features driving their discordant classification) were
367 extracted for further analysis.

368 **SteMClass Web Application**
369 SteMClass was developed in R using the Shiny framework with shiny, shinyjs, shinycssloaders,
370 bslib, and thematic, employing a Bootstrap theme and custom CSS for interface rendering.
371 DNA methylation preprocessing relies on minfi, methylumi, wateRmelon and preprocessCore
372 for raw IDAT file handling and normalisation, with array manifests and annotation
373 (IlluminaMethylationManifest) loaded dynamically. Preprocessing includes noob
374 normalisation, conversion to a common 450K format where required, and genome mapping.
375 The pre-trained SteMClass random forest model (loaded from GitHub Releases
376 (https://github.com/pereze5/SteMClass_App/releases) is applied after aligning test data to

377 the training reference cohort and imputing missing CpGs using means learned during model
378 development. Predictions are output as class probabilities across eight reference cell states,
379 with a rejection option for samples not reaching the 0.6 probability threshold. Visualisations
380 include probability bar charts, UMAP projections generated with uwot, and heatmaps
381 generated with ComplexHeatmap and circlize. Global heatmaps compare the test sample to
382 reference profiles using the top 10,000 CpGs distinguishing a selected cell state from iPSC,
383 while gene-level heatmaps display CpGs annotated to a specified gene
384 (UCSC_RefGene_Name) ordered by genomic position and annotated by gene region
385 (UCSC_RefGene_Group). Results and figures are exportable as high-resolution images or
386 HTML outputs.

387 **Cumulative-contribution analysis of CpG features**

388 The tidy representation of the fitted forest (tidyRF) was queried with the featureContrib
389 function from the package tree.interpreter (v 0.1.1), producing a three-dimensional array P ×
390 D × N (P = 10 000 CpGs, D = 8 classes, N = number of test samples). For each CpG × class cell
391 we calculated the mean absolute contribution across the N samples, yielding a P × D matrix
392 of average importances. For every class separately, CpGs were then ranked in descending
393 order of this value, and the running sum was divided by the class-specific total to obtain a
394 cumulative-fraction vector. The resulting cumulative curves were combined into a long-
395 format data frame and plotted with ggplot2 on a log10 scale CpG-rank axis. An 80 % horizontal
396 reference line highlights how the contributions accumulate.

397 Nanopore Sequencing

398 Library preparation with barcode labelling was performed with ~200-400 ng input of genomic
399 DNA using the Rapid Barcoding Kit (SQK-RBK114.24, Oxford Nanopore Technologies, UK)
400 according to the manufacturer's instructions. During the library preparation, input DNA is
401 fragmented while simultaneously attaching barcodes using a time-efficient transposase-
402 based approach known as “tagmentation”. The final library was loaded onto an R10.4.1 flow
403 cell (FLO-MIN114, Oxford Nanopore Technologies, UK), and low coverage whole genome
404 sequencing (WGS) was performed for 6-24 hrs on a MinION Mk 1B device (Oxford Nanopore
405 Technologies, UK). POD5 files containing the raw data were obtained in real time using the
406 manufacturer's software MinKNOW (v.1.3.1-v.3.6.0) and transferred to a high-performance
407 computing (HPC) cluster for further analysis. Analysis was performed using the nanoDx
408 classification pipeline v0.6.2 (Kuschel et al., 2022). The SteMClass training set raw IDAT files
409 were processed and converted into an H5 file format for use as the training set for the nanoDx
410 pipeline using the R package rhdf5 (v2.50.2). The nanoDx classification and analysis pipeline
411 is publicly available at https://gitlab.com/pesk/nanoDx. The source code for the ad hoc
412 random forest implementation is available at https://gitlab.com/pesk/nanoBenchmark.

413 # Results

414 ## DNA methylation dynamics during in vitro Astrocyte differentiation.

415 DNA methylation is well established as a marker of cell- and tissue-type identity in vivo. To
416 determine whether this specificity is preserved in isogenic cells in vitro, we analysed the DNA
417 methylation profiles of two parental iPSC lines differentiated to astrocyte-like cells. For this
418 we used a previously published protocol (TCW et al., 2017) using a commercially available kit
419 which results in GLAST positive cells by experimental day 37. iPSCs were cultured for two days
420 in E8 medium (days -2 and -1) before starting the differentiation protocol (day 0). The cells
421 were purified by magnetic activated cell sorting (MACS) at the end of the protocol (day 37).
422 Principal component analysis (PCA) of the top 5,000 most variable CpG sites revealed a clear
423 separation between iPSCs (n = 14) and their post-differentiation astrocyte derivatives (n = 12)
424 (Figure 1a). We further analysed samples taken at nine time points during an astrocyte
425 differentiation (time course experiment). These intermediately differentiated astrocytes
426 clustered along a continuum between the two end-point differentiation states (Figure 1a).
427 The vast majority of variation in the methylation data was captured by PC1 (91.5% variance)
428 which separated iPSCs from differentiated astrocytes and intermediate states, capturing
429 progression along the differentiation axis. PC2 (3% variance) likely captures differences
430 between the two parental iPSC lines. Genome-wide comparison of the iPSC and differentiated
431 astrocytes uncovered 228,765 differentially methylated positions (DMPs) (Figure 1b). In
432 comparison to iPSC, differentiated astrocytes displayed promoter hypomethylation of
433 canonical astrocyte genes such as GFAP, S100B, S100A16, and SLC1A3 (Achicallende et al.,
434 2022; Hernández-Ortega et al., 2024; Sturchler et al., 2006) together with promoter
435 hypermethylation of the pluripotency regulators POU5F1 and HK2 (Figure 1b). Tracking the
436 same CpGs during the time course differentiation experiment showed that pluripotency loci
437 (POU5F1, HK2) began to gain methylation by days 3-5, whereas astrocyte-specific genes such
438 as S100A16 and S100B started to lose methylation by day 14 with de-methylation of GFAP
439 occurring late in the differentiation (Figure 1c). Of note, the canonical ectodermal marker
440 PAX6 became methylated in the gene body by days 3-5, at the timepoint of ectodermal
441 differentiation when the cells are expected to have PAX6 expression by FACS. Together, these
442 data show that global DNA methylation landscapes are extensively and progressively
443 remodelled in a lineage-consistent manner during in vitro differentiation of isogenic iPSCs to
444 their astrocyte derivatives, supporting that the observed changes reflect differentiation state
445 rather than inherent genetic differences.

446

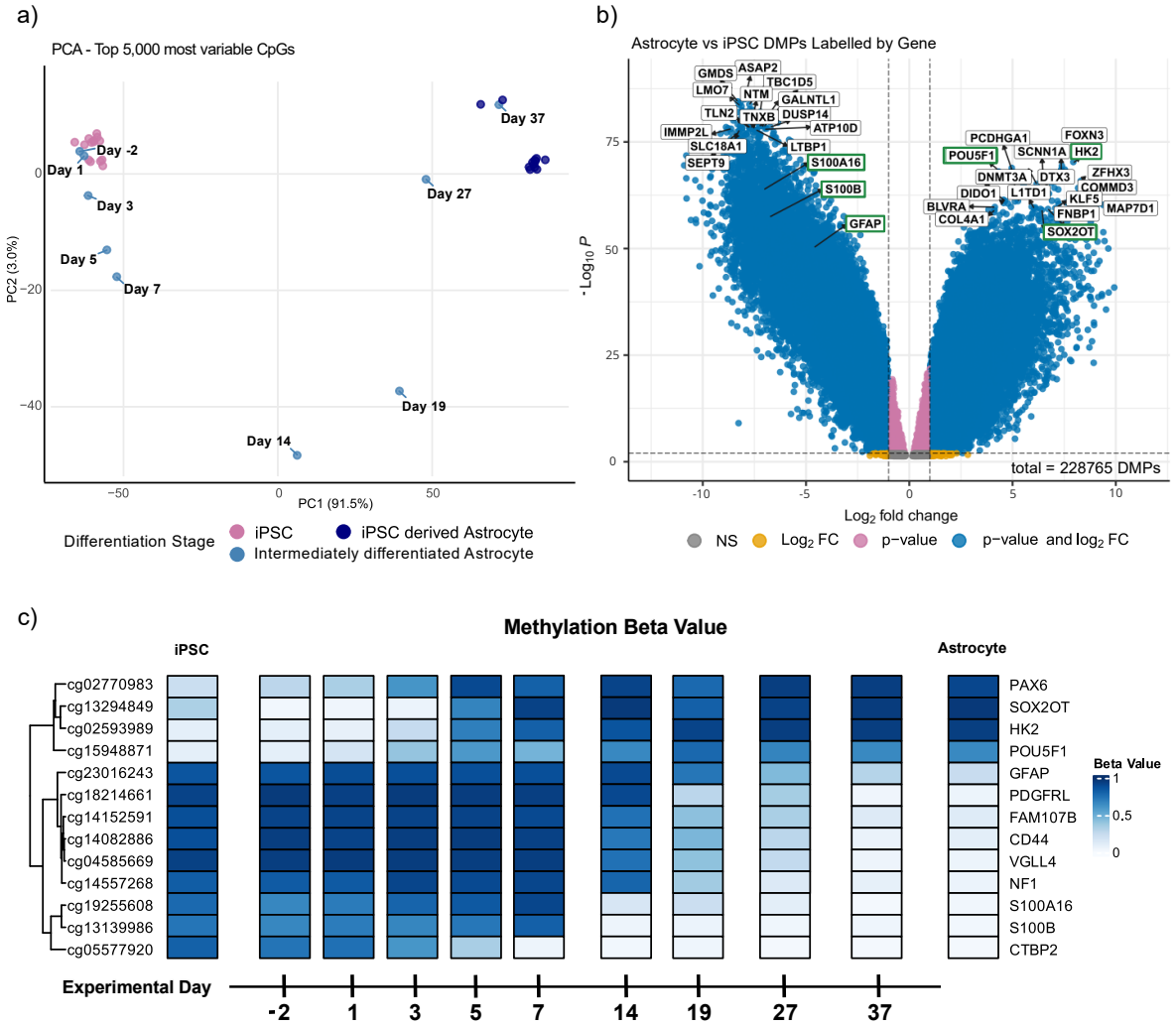


Figure 1: Identification of progressive DNA methylation changes during in vitro iPSC differentiation into astrocytes.

(a) Principal component analysis (PCA) plot of DNA methylation beta values from the 5,000 most variable CpG sites (selected by across-sample standard deviation) in iPSCs (n = 14) and iPSC-derived astrocytes (n = 12) (two parental iPSC lines), together with nine intermediate time-point samples from the differentiation time course experiment. Data were mean-centred and scaled prior to PCA. Each point denotes an independent technical replicate (separately cultured and processed). Colours indicate differentiation states. PC1 (91.5% variance) separates iPSCs from astrocyte derivatives and intermediate states, reflecting progression along the differentiation axis. PC2 (3% variance) likely captures differences between the two parental iPSC lines.

(b) Volcano plot displaying differentially methylated CpGs for the comparison between the iPSCs (n=14) and astrocytes (n=12). Top significant differentially methylated CpGs (False discovery rate-adjusted p-value < 0.05) are labelled by their associated gene, with canonical lineage-specific markers indicated in green.

(c) DNA methylation dynamics were evaluated during a 37-day iPSC derived astrocyte differentiation experiment. For visualisation purposes, the 13 CpG sites showing the largest differences in beta values between undifferentiated and differentiated states are depicted. The heatmap displays mean beta values across iPSC samples (n = 14), individual beta values at the nine intermediate differentiation experiment stages, and mean beta values in

468 astrocyte samples (n = 12). Rows correspond to CpG sites, ordered by hierarchical clustering
469 using Pearson correlation distance. CpG loci are annotated with CpG unique IDs and their
470 associated gene names.
471

472 **Unsupervised dimensionality reduction of DNA methylation data separates**
473 **iPSC-derived cells according to differentiation state.**

474 We next investigated whether the lineage-specific methylation changes observed during
475 astrocyte differentiation generalise to other iPSC derived cell types. Early ectodermal,
476 mesodermal and endodermal progenitor cultures were generated with at least two
477 independent protocols per lineage from 11 independent iPSC lines (see Methods). For each
478 of these cell types, we analysed homogeneous cell populations based on their conventional
479 marker expression as analysed by FACS (Supplementary Figure 1+2). Selection of populations
480 in which at least 85% of the cells expressed the conventional differentiation markers was done
481 to minimise intra-sample heterogeneity. Unsupervised analysis of the 50,000 most variable
482 CpG sites across all samples was visualised with UMAP (Figure 2a). Four discrete clusters
483 emerged: pluripotent iPSCs and three well-separated groups corresponding to the three
484 germ-layers ectoderm, mesoderm and endoderm. Crucially, samples produced with different
485 protocols converged on the same germ-layer cluster, and technical replicates from separate
486 differentiation batches co-localised, demonstrating that the differentiation state-specific DNA
487 methylation signatures are protocol-independent and reproducible between experiments.

488 To determine whether DNA methylation signatures can resolve a broader spectrum of
489 differentiation states, we expanded the dataset to a total of eight in vitro differentiation
490 states (undifferentiated iPSCs, ectodermal cells, mesodermal cells, endodermal cells, lung
491 cells, endothelial cells, neural stem cells, and astrocytes). Cell type nomenclatures and
492 corresponding differentiation protocols, along with the iPSC lines used in this study are
493 described in Supplementary Tables 1-3 and Supplementary Figure 1, and representative data
494 for marker expression and cell morphology are detailed in Supplementary Figure 2.

495 Unsupervised analysis of the 50,000 most variable CpG sites across all samples, visualised with
496 UMAP, produced eight well-separated clusters that mapped one-to-one onto the designated
497 cell states (Figure 2b). The observed clustering of distinct cell types was independent of
498 potential batch effects such as parental cell line, donor sex, or DNA methylation array (EPIC)
499 version (Supplementary Figure 3a-d). These findings demonstrate that genome-wide DNA
500 methylation profiles provide a robust, protocol-independent read-out of differentiation state
501 across a diverse set of iPSC derivatives, underscoring their utility for benchmarking and
502 standardising in vitro differentiation characterisation.

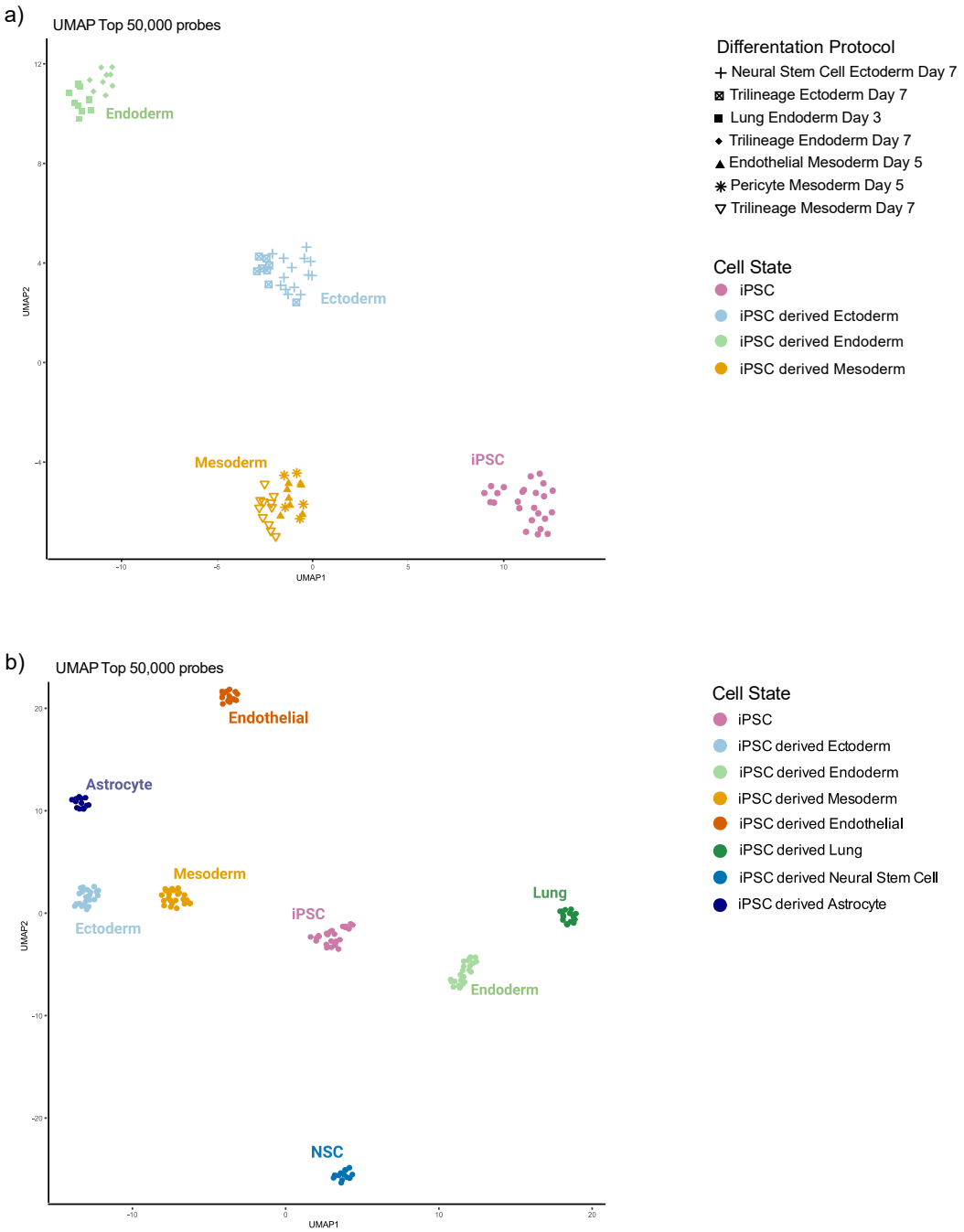


Figure 2: Unsupervised analysis indicates that differentiation state-specific DNA methylation signatures are protocol-independent and reproducible.

(a) UMAP plot of DNA methylation data (50,000 most variable CpGs) from 15 independent iPSC lines, 11 of which were differentiated into ectodermal, endodermal, and mesodermal lineages using at least two differentiation protocols per lineage with each point representing an independent technical and/or biological replicate. Colours represent differentiation states, with ectoderm, endoderm, and mesoderm forming distinct clusters, the differentiation protocol used is indicated by shape.

(b) UMAP plot of DNA methylation data (50,000 most variable CpGs) from 15 independent iPSC lines, 11 of which were differentiated into ectodermal, endodermal, and mesodermal lineages including at

513 least two differentiation states per germ layer. Each point represents an independent technical and/or
514 biological replicate. Colours represent differentiation states, with each differentiation state forming
515 distinct clusters. iPSCs are represented in pink, endodermal differentiation states (endoderm and lung)
516 are represented in green tones, ectodermal differentiation states (ectoderm, neural stem cells and
517 astrocytes) are represented in blue tones and mesodermal differentiation states (mesoderm and
518 endothelial cells) are represented in orange tones.

519

520 Differential methylation analysis identifies DNA methylation-based signatures 521 for cell state classification.

522 To isolate the differentiation-associated methylation patterns evidenced by our unsupervised
523 analysis, we performed pairwise differential methylation analyses comparing iPSCs against
524 the 7 differentiated cell states (Figure 3, Supplementary Figure 4, Supplementary Table 4).
525 Gene ontology enrichment of the resulting DMPs confirmed that the methylation changes
526 align with known lineage-specific biological processes in each cell state (Figure 3,
527 Supplementary Figure 4).

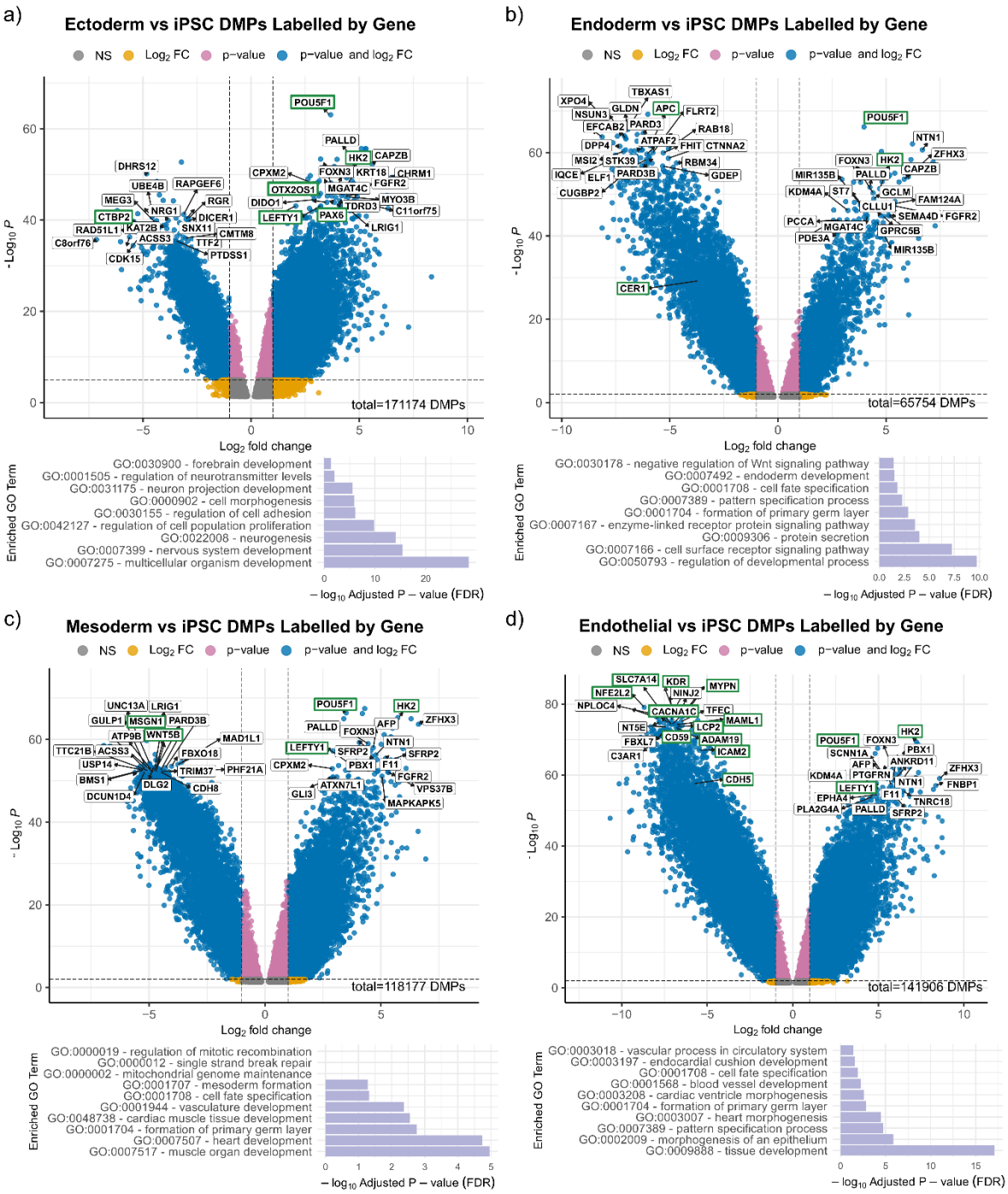


Figure 3. Differential DNA methylation identifies lineage-specific features for differentiation state classification.

(a–d) Volcano plots of differentially methylated positions (DMPs) comparing iPSCs to: (a) ectoderm, (b) endoderm, (c) mesoderm, and (d) endothelial cells. Each point is a differentially methylated CpG; blue points meet significance thresholds (FDR-adjusted $p < 0.05$ and $|\log_2 \text{fold change}| > 1.5$). The most significant CpGs are annotated with their associated gene. Canonical or candidate novel lineage markers are highlighted in green to emphasise key methylation shifts underlying differentiation. Bar plots summarise gene ontology enrichment among significant DMP-associated genes, showing overrepresented biological processes and pathways consistent with the expected differentiation programs. Overall, the methylation patterns support known lineage trajectories and provide differentiation-specific epigenetic signatures. NS, not significant.

Building and Validating a DNA Methylation-Based Cell State Classifier: SteMClass

These differentiation-specific DMPs allowed us to identify cell state-specific signatures. Because many CpGs were differentially methylated across multiple cell states, we assigned each CpG to the single cell state with which it had the most significant FDR-adjusted p-value. We then selected a total of 10 000 CpGs, sampling proportionally from each cell states ranked list of DMPs. Using these 10 000 CpGs, we built a Random Forest classifier ("SteMClass") in the R tidymodels framework with a ranger engine. In practice, once a CpG passes the feature selection step, its DNA methylation status is retained for all samples; any residual signal that overlaps multiple cell states remains available to the classifier. This strategy preserves cell state-specific methylation signatures while keeping feature dimensionality tractable.

A nested cross-validation (5-fold outer x 5 repeats, 3-fold inner) was applied to the model for unbiased performance estimation and hyperparameter tuning, exploring number of trees and minimum node size (min_n) via a Latin-hypercube search (Supplementary Figure 5a). The number of variables sampled per split (mtry) was fixed at 100 and model selection prioritised the Brier score. The final model, fitted on 97 training samples with min_n = 3, trees = 1570, mtry = 100, had a Brier score of 0.0264 and macro expected calibration error (ECE) of 0.029 (Supplementary figure 5b). These cross-validated out of bag metrics indicate SteMClass' robust calibration and overall discriminative power. To minimise overconfident calls, we pooled raw probabilities from all outer folds and optimised a rejection threshold between 0.50 - 0.95 (step 0.05) using the macro-Youden index and a cutoff of 0.6 (macro-Youden = 0.975) was chosen (Figure 4a). Subsequent test samples whose maximum class probability fell below 0.6 were rejected and labelled "Not Classifiable."

SteMClass was evaluated on an independent validation set of 58 samples representing the eight distinct cell states included in the classifier and attained 96.5% accuracy, Cohen's κ = 0.959, and a 3% reject rate (Figure 4b). The validation samples were generated in the same laboratory as the training set ("internally"), derived from the same iPSC lines, using the same differentiation protocols. Despite these controlled conditions, the test samples represented independent experiments, reagent batches, and personnel.

To further evaluate SteMClass' performance, a cohort of publicly available methylation array data was compiled (Supplementary Figure 6a). The dataset included 241 samples that, based on the materials and methods described in their respective publications, were deemed comparable to the cell states defined in this study. This external validation step was designed to test SteMClass' ability to generalise to data generated by independent laboratories and the

classifier achieved an overall accuracy (based on the sample annotation in the literature) of 76.5%, Cohen's κ = 0.593, and a 16.6% reject rate (Figure 4c). To visualise how many CpG probes drive the classifier decisions, we quantified the absolute feature contribution of every probe for every test sample (internal and external combined) and aggregated these values across samples (see Methods). The resulting cumulative-contribution curves (Figure 4d) rise gradually: no class reaches 80% of its total weight before 1000 ranked probes and the remaining 20% is dispersed over several thousand additional sites. Hence SteMClass recognises each lineage through a wide methylation footprint rather than a handful of high-impact markers, making the model inherently robust to noise at individual CpGs. iPSC derived samples from the same publicly available datasets for which the sample metadata indicated that the cell state was not captured by SteMClass were also tested (n=34; Supplementary Figure 6b, Supplementary Table 5). Four early stage (< 3 days) endoderm samples and three embryoid body samples were classified as iPSC, while none of the remaining samples were arbitrarily assigned a class, supporting the specificity of SteMClass to its defined states.

To evaluate the robustness of methylation-based classification across different platforms, a subset of the internal validation cohort (n=13/58) was analysed using both EPIC array and long read nanopore sequencing (Supplementary Figure 6c). Classification of the nanopore sequencing data was performed using the previously published nanoDx pipeline (Euskirchen et al., 2017) with an ad-hoc Random Forest model trained on the reference set generated in this study. The classifier demonstrated high reproducibility, correctly classifying 12/13 cases consistently between the nanopore sequencing and methylation array platforms.

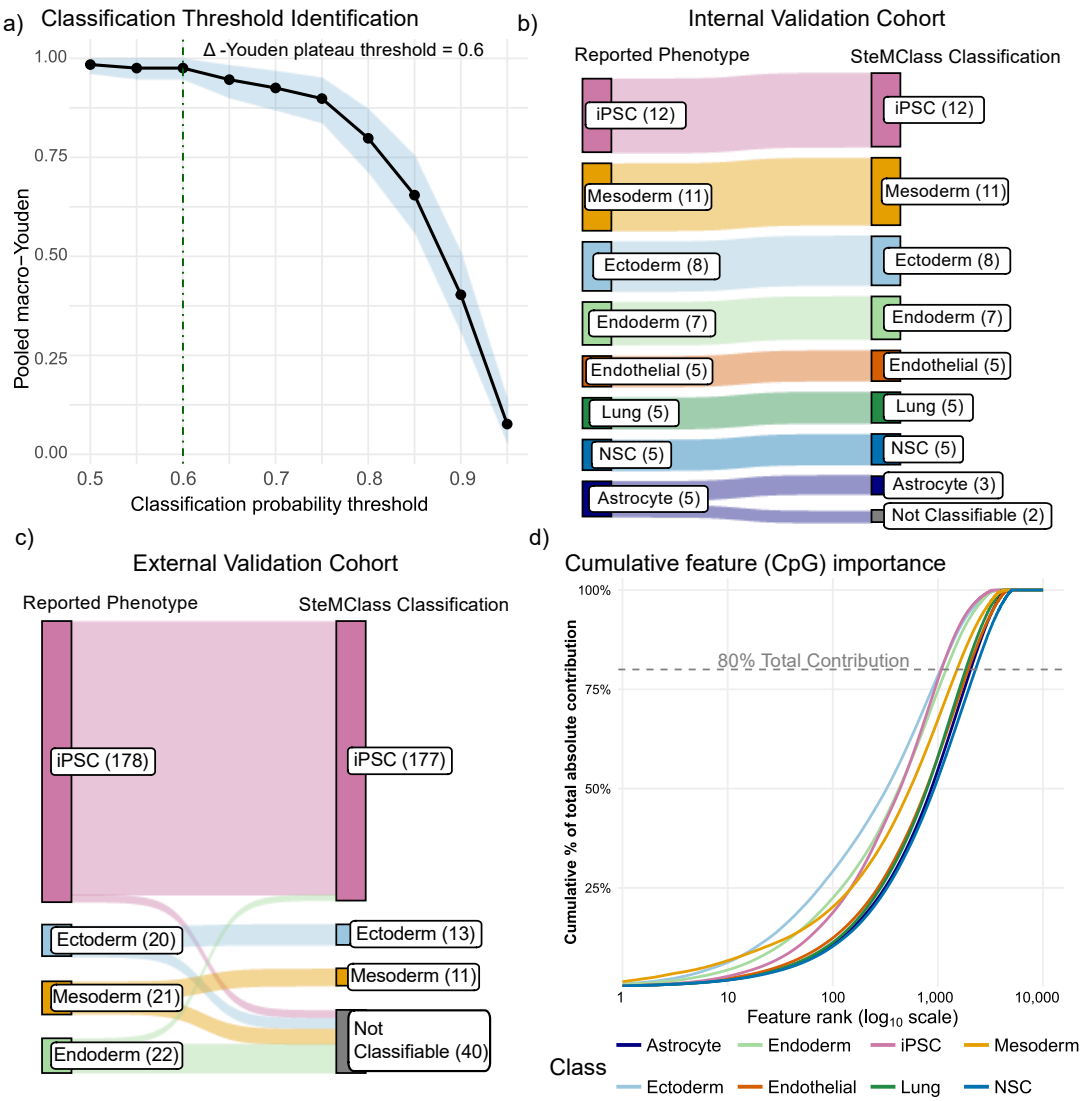


Figure 4. Integrated assessment of SteMClass: threshold tuning, independent validation, and feature-level signal distribution.

a) Identification of the classification probability cutoff via pooled macro-Youden and bootstrap inference. A line plot of pooled macro-Youden (y-axis) against candidate thresholds (x-axis), with ribbon showing the 95% bootstrap confidence interval and points the empirical estimates. A specificity constraint (mean macro-specificity ≥ 0.95) was imposed, the plateau threshold, defined as the highest threshold below which further lowering yields no statistically significant Youden gain, is indicated by the green dash line. Bootstrap analysis of adjacent threshold steps reveals that lowering from 0.65 to 0.60 gives a small but statistically significant improvement in macro-Youden (one-sided bootstrap p = 0.049), whereas further lowering from 0.60 to 0.55 does not (p = 1), supporting 0.60 as the threshold.

b) Sankey diagram showing classifier performance on the independent internal validation cohort (n = 58). Each flow connects the original sample annotation (reported phenotype) to the SteMClass predicted class (including "Not Classifiable" rejections). Despite being generated in the same laboratory and from the same iPSC lines and protocols as the training data, these samples derive from independent experiments, reagent lots, and personnel. The classifier achieved 96.5% accuracy, Cohen's κ = 0.959, and a 3% reject rate. The internal cohort comprised:

611 Astrocytes (n = 5), Ectodermal cells (n = 8), Endodermal cells (n = 7), Mesodermal cells (n = 11), Endothelial cells
612 (n = 5), iPSCs (n = 12), Lung cells (n = 5), and NSCs (n = 5).

613 **c)** Sankey diagram showing SteMClass predictions versus reported phenotype for the external validation cohort
614 compiled from publicly available methylation datasets (n = 241). Flows connect the original sample annotation
615 to the classifier's assigned class (including "Not Classifiable" rejections), highlighting agreement and
616 discordance. The external cohort comprises iPSCs (n = 178), iPSC-derived Ectoderm (n = 20), Endoderm (n = 22),
617 and Mesoderm (n = 21) from multiple studies: GSE105093 (Sultanov et al., 2018), GSE110544 (Banovich et al.,
618 2018), GSE141521 (Nishino et al., 2021), GSE159479 (Horii et al., 2021), GSE163322 (Tanaka et al., 2023),
619 GSE180402 (Cypris et al., 2022), GSE197723 (Crochemore et al., 2023), GSE207108 (Ahern et al., 2022),
620 GSE207076 (Schmidt et al., 2022), GSE247428 (Puri et al., 2023), GSE253190 (Elsafi Mabrouk et al., 2024),
621 GSE61461 (Johannesson et al., 2014), GSE76372 (Kitayama et al., 2016), GSE85828 (Daily et al., 2017).

622 **d)** Cumulative absolute feature contribution per class. For each of the eight cell-state classes (colour-coded),
623 lines show how the summed absolute contribution accumulates as CpG sites are added in descending order of
624 their mean absolute contribution across all 299 test samples (internal + external; n=299). The x-axis is on a $\log_{10}$
625 scale; the dashed horizontal line marks 80% of the total contribution. All classes reach the 80% level after
626 approximately 1,000 CpGs, indicating that SteMClass bases its decisions on a broad, distributed methylation
627 signature rather than being biased by a small set of marker loci.

628

629 **Integrative Broad and CpG-Level Profiling Explains Classification Outcomes.**
630 To further interpret the classification results we looked beyond the scope of the classifier and
631 compared the broader DNA methylation profiles of the test samples to the reference set by
632 investigating the top 20,000 CpGs that distinguish each cell state from iPSC (see Methods;
633 Supplementary Table 4). This allows a wider perspective beyond the 10,000 total features
634 used by SteMClass to investigate re-classified and "Not Classifiable" samples. Four of the
635 publicly available cohort endoderm samples were re-classified as iPSC by SteMClass.
636 Hierarchical clustering of these samples with the reference cohort iPSC and endodermal cells
637 showed clustering with the iPSC samples based on their broader cell state profile (Figure 5a).
638 Investigation of the top contributing variables to the classifier's decision in favour of iPSC and
639 against endoderm included the canonical pluripotency marker POU5F1 and endodermal
640 marker APC (Figure 5b). Closer investigation of the POU5F1 promoter region confirmed
641 methylation levels more consistent with reference iPSC in these test samples (Figure 5c).
642 Taken together, these results support the likelihood of inefficient differentiation experiments
643 rather than misclassifications.
644 Four iPSC samples from the external validation set were not classifiable. These four samples
645 formed a cluster distinct from the SteMClass reference iPSCs based on their 20,000 CpG
646 broader cell state profile (Figure 5d). To identify potential novel markers of iPSC
647 undifferentiated state, we interrogated the top 100 negatively contributing variables for iPSC
648 common to the four rejected samples (Figure 5e). Top hits included FGFR2, PALLD and HK2

649 which were hypermethylated in promoter CpGs in the four unclassifiable samples. Three of
650 the samples are more similar to reference iPSCs on the methylation profile, suggesting subtle
651 cumulative differences. One sample exhibited an especially divergent methylation profile,
652 most notably at the POU5F1 gene promoter, suggesting an unclear undifferentiated state and
653 therefore compromised iPSC quality rather than a classification error for this sample (Figure
654 5f).

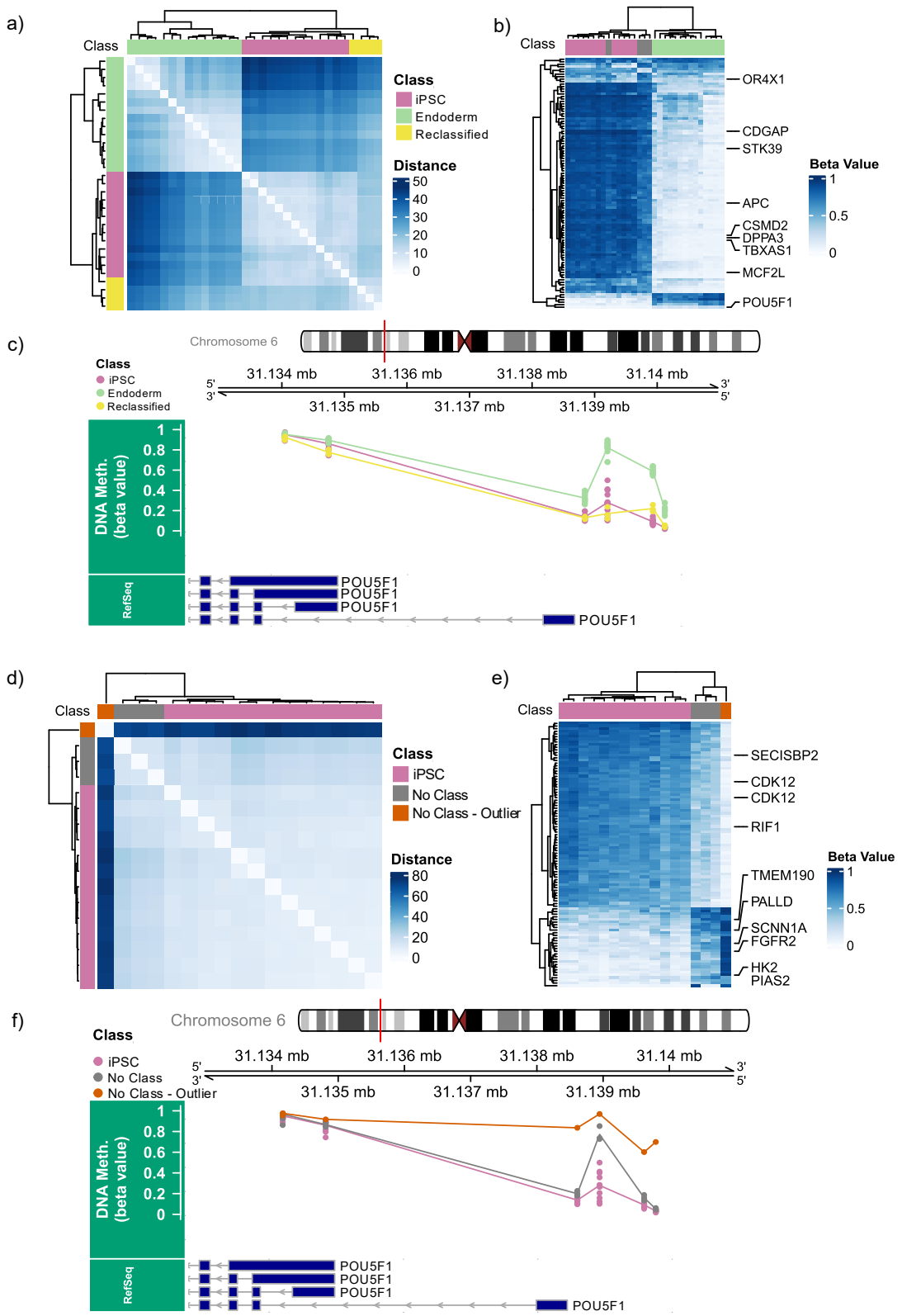


Figure 5. Interpreting SteMClass classification boundaries via broad and narrow view methylation analysis

(a) Hierarchical clustering of four samples published as having "endoderm-differentiation" that SteMClass re-classified as iPSC (yellow), together with reference iPSC (pink) and reference endoderm (green) profiles. Clustering was performed on a Euclidean-distance matrix (complete linkage) computed from beta values at the

660 top 20 000 CpGs most discriminating iPSC from all other cell types ("broad view"; Supplementary Table 4). The
661 samples reclassified by SteMClass cluster with reference iPSC.
662 **(b)** Heatmap of the 100 CpGs with highest positive importance for distinguishing iPSC vs. endoderm in the
663 random forest model. Rows represent CpGs and columns represent samples (reference and re-classified
664 endoderm); both were hierarchically clustered using complete linkage with Euclidean distance.
665 Heatmap colours indicate beta value (white = low, blue = high). CpGs associated with canonical markers such as
666 POU5F1 (pluripotency) and APC (endoderm) are labelled. Also, in this analysis the reclassified samples cluster
667 with reference iPSC.
668 **(c)** Genomic view of beta values across the POU5F1 promoter region (chr6:31,134,000–31,140,000) for reference
669 iPSC, reference endoderm, and the four re-classified endoderm samples ("narrow view"). Tracks show
670 methylation beta values for each sample (points), confirming an iPSC-like profile in re-classified samples for this
671 important differentiation marker.
672 **(d)** Hierarchical clustering of four "Not Classifiable" test samples published as iPSC (grey and orange) with
673 SteMClass reference iPSC profiles (pink), using the same 20 000 CpG set most discriminating iPSC from all other
674 cell types and clustering parameters as in (a). Of the four rejected samples, three form a distinct clustering
675 branch (grey), indicating atypical iPSC methylation patterns, with one sample being a clear outlier from all other
676 samples (orange).
677 **(e)** Heatmap of the 100 CpGs whose contributions to the iPSC class were most strongly negative across all four
678 not classifiable samples. Promoter CpGs in FGFR2, PALLD and HK2 show consistent hypermethylation relative to
679 reference iPSCs.
680 **(f)** Genomic view of beta values at the POU5F1 promoter for the four not classifiable samples, revealing that the
681 outlier sample (orange) exhibits markedly higher methylation, suggestive of incomplete reprogramming,
682 whereas the three samples (grey)remain broadly consistent with iPSC status in this region.

683

684 Six out of ten non-classifiable mesoderm samples clustered closer to iPSCs than to reference
685 mesoderm based on their methylation profile, with one sample being a clear outlier to both
686 cell states (Supplementary Figure 7a). The top CpGs driving rejection for classification as
687 mesoderm included those within the promoter of canonical mesoderm marker genes such as
688 MSGN1 and WNT5B consistent with incomplete mesoderm differentiation rather than
689 misclassification (Supplementary Figure 7b). Two samples fell within the reference mesoderm
690 clusters and scored for mesoderm but did not reach the 0.6 threshold for assignment to this
691 class, possibly implying either test sample heterogeneity or batch effects which could be
692 mitigated by including more protocol- and laboratory-diverse data to the reference set.
693 Among the seven rejected ectoderms, one was a clear outlier, while the remaining six
694 resembled reference ectoderms in their global profiles (Supplementary Figure 7d). However,
695 negatively contributing variables for ectoderm classification common to the rejected samples
696 highlighted PAX6 as one the most differentially methylated genes (Supplementary figure 7f)
697 indicating a fundamental difference compared to reference ectoderm in these six samples.
698 Two samples lacked the key differences and yielded borderline probability scores (~0.58),

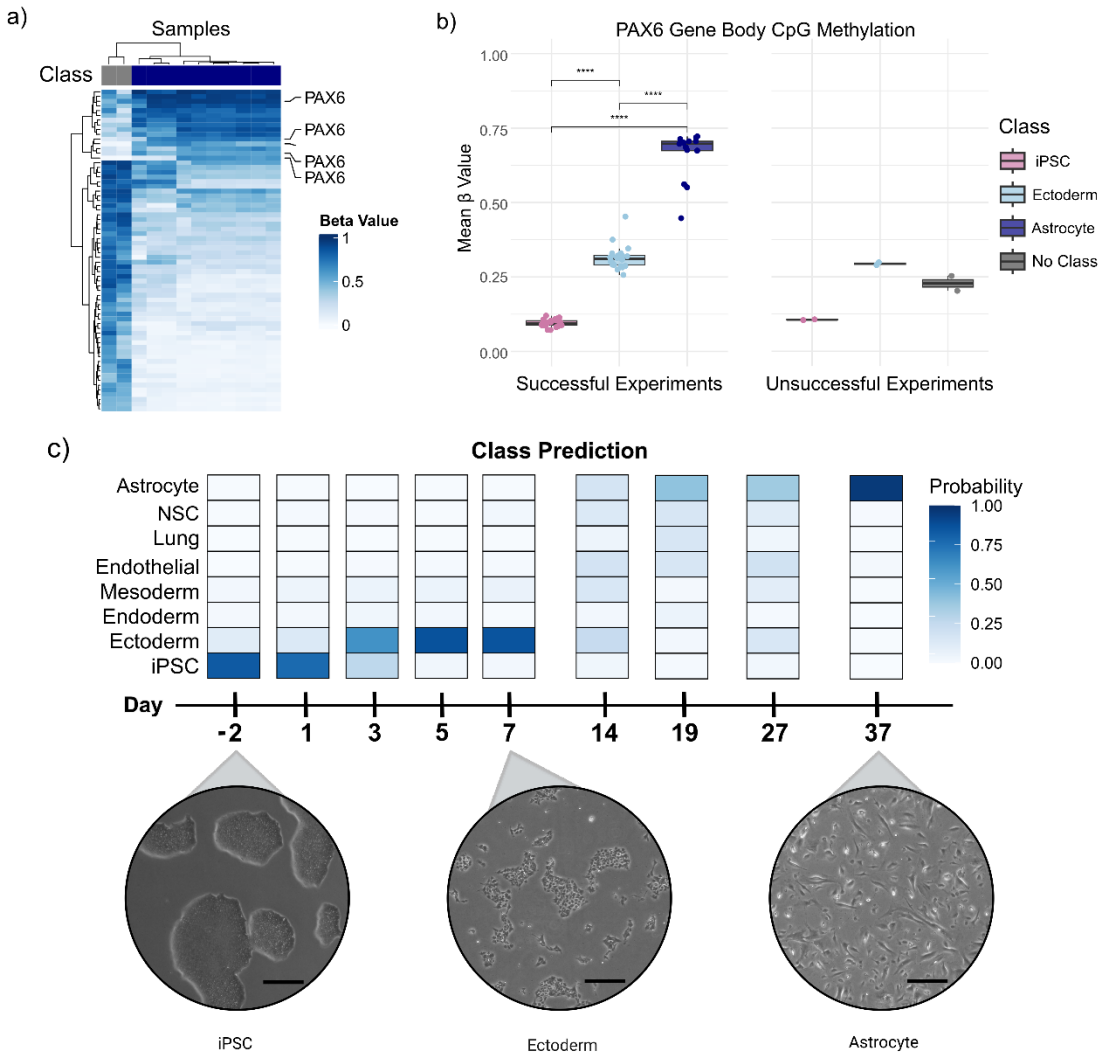
699 again possibly explained by the aforementioned batch effects. The other five samples, which
700 did show those DMP patterns, are more likely to represent incomplete ectoderm
701 differentiations.
702 Eighteen of 22 public endoderm samples failed to meet our 0.6 probability threshold and
703 instead clustered closer to reference iPSCs (Supplementary Fig. 7g), likely reflecting sample
704 heterogeneity, underrepresentation of certain protocols in the training set, and/or batch
705 effects from diverse data sources. Endodermal samples deviated more from reference
706 signatures than other lineages, suggesting greater variability in differentiation efficiency and
707 epigenetic outcome. Given the diversity of endoderm protocols and the critical role of timing
708 in patterning, it is likely that multiple biologically distinct subclasses are currently grouped
709 under the label "endoderm" in the literature. Therefore, expanding the reference set to
710 include broader protocol coverage and subclass definitions may improve classifier
711 performance. These findings underscore the need for a unified, quantitative assay for cellular
712 differentiation state characterisation.

713

714 **SteMClass Classification Mirrors Progressive DNA Methylation Shifts During Differentiation.**

715 The CpG sites contributing to the rejected classification of the two not classifiable astrocyte
716 samples from the internal validation set (Figure 4b) included the previously identified
717 ectoderm-characteristic PAX6 gene body region, possibly indicating inefficiency in neural
718 commitment in these experiments (Figure 6a). In a limited comparison restricted to these two
719 failed experiments (replicating this is not feasible) compared to our successful reference and
720 test sample experiments (total samples n=65), this observed decrease in PAX6 gene body
721 methylation was evident only at the day 37 astrocyte differentiation stage and also even
722 within the "successful experiment cohort" (Figure 6b). Improving the consistency of astrocyte
723 populations could involve incorporating SteMClass DNA methylation-based screening at a
724 later intermediate timepoint, enabling the selection of wells that are most closely progressing
725 toward the astrocyte state. Although it is beyond the scope of this study, as the SteMClass
726 reference set grows beyond this proof-of-concept classifier, our harmonised framework will
727 enable more robust interrogation of such epigenetic determinants of differentiation
728 propensity. To further explore how SteMClass predictions relate to the dynamic DNA
729 methylation patterns during differentiation, we examined changes in predicted cell states
730 during the transition from iPSCs to astrocytes (the same 9 time point experiment displayed in
731 Figure 1), as determined by SteMClass' predictions and probability scores. The predicted cell
732 identities and their associated probability scores at multiple differentiation stages
733 demonstrate how SteMClass adjusts as cells move from a pluripotent state toward more

734 lineage-restricted fates (Figure 6c). The progressive shift in probabilities reflects the model's
735 sensitivity to intermediate, transient states and its capacity to track subtle changes in lineage
736 commitment.



737

738 **Figure 6. Epigenetic driver inference and dynamic tracking of astrocyte differentiation using**
739 **SteMClass**
740 **(a)** Heatmap of the 100 CpG sites whose contributions most strongly oppose astrocyte classification
741 in the two not classifiable internal validation samples (grey columns), alongside reference astrocytes
742 (purple columns). Rows represent CpGs and columns represent samples; both were hierarchically
743 clustered using complete linkage with Euclidean distance. Heatmap colour indicates beta value (white
744 = low, blue = high); four PAX6 gene body-associated CpGs rank among the top features, possibly
745 implicating inefficient neural commitment in these two failed differentiations.
746 **(b)** Boxplots show sample-level mean beta-values (0-1) for the four PAX6 CpGs opposing astrocyte
747 classification across iPSC, ectoderm and astrocyte samples. Colours denote SteMClass-assigned
748 differentiation states: iPSC (pink), ectoderm (light blue), astrocyte (dark purple) and "Not classifiable"
749 (grey). Samples are split into panels by differentiation outcome: successful (n = 59) and unsuccessful

750 (n = 6), defined by passing or failing the SteMClass 0.6 cutoff. In successful experiments, PAX6
751 methylation increases during the differentiation from iPSC to ectoderm to astrocyte, whereas in failed
752 experiments it decreases at the astrocyte stage. Since these CpGs are among the top features
753 contributing to ectoderm and astrocyte classification in SteMClass, monitoring the differentiation
754 using SteMClass at further intermediate stages may help identify at-risk differentiations earlier. All
755 pairwise comparisons within the successful group were significant (****, $p < 0.0001$; Wilcoxon rank-
756 sum). No statistical testing was applied to the unsuccessful group due to limited sample size. **(c)**
757 Temporal SteMClass classification shifts during the same iPSC to astrocyte nine time point
758 differentiation experiment shown in Figure 1c. Top panel: heatmap of predicted class probabilities (0-
759 1) at each harvest point for each class represented in SteMClass. Bottom panel: representative culture
760 images at day -2 (classified as iPSC), day 7 (classified as ectoderm), and day 37 (classified as astrocyte),
761 illustrating the morphological progression in line with classification shifts. Scale bars = 200 µm.

762 # Discussion

763 A fundamental obstacle in stem cell research and regenerative medicine is the lack of
764 standardised frameworks to assess the quality and maturation stage of differentiated cells,
765 causing variability in both results and interpretation. While tools like the TaqMan hPSC
766 Scorecard Assay (Thermofisher Scientific) and hiPSCore (based on long-read nanopore
767 sequencing of 12 early trilineage -linked genes) help gauge pluripotency and trilineage
768 potential (Dobner et al., 2024; Lenz, 2012; Tsankov et al., 2015), most downstream
769 applications demand differentiation into more specialised cell types that are not covered by
770 these assays. Currently, evaluations of these more advanced stages vary widely and often rely
771 on a handful of molecular markers, genes or proteins whose expression may fluctuate based
772 on protocol specifics, laboratory conditions, and genetic backgrounds. In addition to our
773 observation of differentiation outcome heterogeneity across published studies, it has also
774 been shown that iPSC-derived cardiomyocytes frequently show inconsistent gene expression
775 profiles and cell compositions across different laboratories and methods (Grancharova et al.,
776 2021) underscoring the need for more standardised and robust assessment tools. Here we
777 describe the development of a proof-of-concept DNA methylation-based classifier for iPSC
778 and iPSC derived cells.

779 The classifier prototype, “SteMClass”, demonstrates strong predictive performance, with a
780 Brier score of 0.0264 as calculated during the cross-validation procedure. Validation with one
781 internally generated independent validation cohort and one independent validation cohort
782 compiled from publicly available data from published studies reinforces these findings,
783 demonstrating StemClass’ robustness to variations in experimental conditions, as well as the
784 classifier’s ability to identify potentially problematic samples.

785 While these are promising results, further harmonisation requires data from multiple
786 research centres to address protocol, reagent and data format inconsistencies across
787 institutions. To preliminarily explore its performance across different institutions, we
788 incorporated DNA methylation array data for which the published differentiation protocols
789 closely matched our training set, allowing for an initial assessment of classifier applicability in
790 varied contexts. Our data shows that endodermal cells especially deviate far more from
791 SteMClass reference signatures than samples of other lineages, implying that endoderm
792 differentiation protocols may exhibit greater variability in efficiency and epigenetic outcome.
793 All SteMClass reference endoderm samples were validated by flow cytometry, confirming
794 high differentiation efficiency (90–99% $SOX17^{+}/CXCR4^{+}$). Public datasets (e.g., GSE253190,
795 GSE163322, GSE207076) rarely report comparable quantitative measures, and
796 semiquantitative marker assessments have been shown to yield overlapping DNA
797 methylation profiles (Daily et al., 2017; Dobner et al., 2024; Elsafi Mabrouk et al., 2024;
798 Schmidt et al., 2022). Expanding the reference set to include diverse differentiation protocols
799 could improve classifier generalisability and enable assessment of how protocol-specific
800 efficiencies influence lineage specification. A larger dataset may also reveal methylation
801 subtypes of endoderm corresponding to early regional patterning (e.g., anterior, midgut,
802 posterior), shifting the classifier from a single endoderm signature toward a spectrum of
803 progenitor identities.

804 SteMClass is adaptable to new technologies, already integrating data from the latest EPICv2.0
805 methylation arrays and can also be adapted to nanopore sequencing based on our proof-of-
806 concept experiment. By adapting to nanopore sequencing, SteMClass can leverage direct,
807 real-time analysis of native DNA without the need for bisulfite conversion, streamlining
808 sample preparation and significantly reducing associated costs. Nanopore sequencing
809 requires relatively low initial investment compared to array-based platforms, making it more
810 accessible for laboratories with limited budgets. While single-cell DNA methylation profiling
811 remains technically complex and expensive, SteMClass' use of bulk methylation data on
812 relatively pure cell populations provides a practical alternative for cell-type-specific
813 classification. This combination of cost-effective setup and simplified workflows enhances the
814 utility and scalability of SteMClass for a broader range of research environments.

815 The SteMClass classifier offers a wide range of future applications that could significantly
816 advance stem cell research and clinical applications. By expanding to the prediction of
817 differentiation potential, it could help screen iPSC lines for lineage-specific applications,
818 facilitating regenerative medicine and disease modelling. It also holds potential as a blueprint

819 to develop quality control tools in clinical-grade stem cell production, ensuring reproducible
820 cell differentiation for therapeutic use. In organoid and co-culture systems, the classifier could
821 track cell-state composition on sorted models, supporting studies on tissue modelling and cell
822 interactions. Together, these applications position the classifier as a valuable starting point
823 for advancing stem cell research, quality control in regenerative medicine, and precision
824 medicine applications.

825 In addition to the classification tool itself, the dataset is currently the most extensive DNA
826 methylation data resource of iPSC lines and their isogenic differentiated derivatives across
827 multiple stages of ectoderm, endoderm, and mesoderm, maintained and characterised under
828 standardised conditions. This methylation dataset is a valuable research resource and can
829 provide a wealth of information on the dynamics of DNA methylation during in vitro
830 differentiation considering that cells derived from iPSCs in vitro differ inherently from their
831 fully mature in vivo counterparts. Beyond classification, the DNA methylation data generated
832 by both array-based and nanopore long-read sequencing methods can be used for additional
833 quality control measures in iPSC maintenance and differentiation experiments such as CNV
834 detection, biological sex prediction, and genotype verification without the need for additional
835 testing (Capper et al., 2018; Daenekas et al., 2024; Kuschel et al., 2022).

836 Conclusion

837 SteMClass (https://pereze5-stemclass.share.connect.posit.cloud/) is a publicly available
838 proof-of-concept DNA methylation-based classifier that rapidly and reproducibly verifies iPSC
839 and iPSC-derived differentiation states. It identifies and characterises distinct stages of iPSC
840 differentiation with high reproducibility, providing a standardised framework for classification
841 that enables consistent cross-laboratory comparisons. By supporting the development of
842 harmonised reference datasets and interpretive criteria, SteMClass advances standardisation
843 in stem cell research and accelerates translation toward clinical applications.

844

845

846 References

847 Achicallende, S., Bonilla-Del Río, I., Serrano, M., Mimenza, A., Lekunberri, L., Anaut-Lusar, I.,
848 Puente, N., Gerrikagoitia, I., & Grandes, P. (2022). GLAST versus GFAP as astroglial
849 marker for the subcellular study of cannabinoid CB1 receptors in astrocytes.
850 *Histochemistry and Cell Biology*, *158*(6), 561–569. https://doi.org/10.1007/s00418-
851 022-02139-4

852 Ahern, D. T., Bansal, P., Armillei, M. K., Faustino, I. V., Kondaveeti, Y., Glatt-Deeley, H. R.,
853 Banda, E. C., & Pinter, S. F. (2022). Monosomy X in isogenic human iPSC-derived
854 trophoblast model impacts expression modules preserved in human placenta.
855 *Proceedings of the National Academy of Sciences of the United States of America*,
856 *119*(40), e2211073119. https://doi.org/10.1073/pnas.2211073119

857 Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., &
858 Irizarry, R. A. (2014). Minfi: A flexible and comprehensive Bioconductor package for
859 the analysis of Infinium DNA methylation microarrays. *Bioinformatics (Oxford,*
860 *England)*, *30*(10), 1363–1369. https://doi.org/10.1093/bioinformatics/btu049

861 Banovich, N. E., Li, Y. I., Raj, A., Ward, M. C., Greenside, P., Calderon, D., Tung, P. Y., Burnett,
862 J. E., Myrthil, M., Thomas, S. M., Burrows, C. K., Romero, I. G., Pavlovic, B. J., Kundaje,
863 A., Pritchard, J. K., & Gilad, Y. (2018). Impact of regulatory variation across human
864 iPSCs and differentiated cells. *Genome Research*, *28*(1), 122–131.
865 https://doi.org/10.1101/gr.224436.117

866 Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang,
867 Y., Vollmer, E., Goldmann, T., Seifart, C., Jiang, W., Barker, D. L., Chee, M. S., Floros, J.,
868 & Fan, J.-B. (2006). High-throughput DNA methylation profiling using universal bead
869 arrays. *Genome Research*, *16*(3), 383–393. https://doi.org/10.1101/gr.4410706

870 Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*,
871 *16*(1), 6–21. https://doi.org/10.1101/gad.947102

872 Buckberry, S., Liu, X., Poppe, D., Tan, J. P., Sun, G., Chen, J., Nguyen, T. V., de Mendoza, A.,
873 Pflueger, J., Frazer, T., Vargas-Landín, D. B., Paynter, J. M., Smits, N., Liu, N., Ouyang,
874 J. F., Rossello, F. J., Chy, H. S., Rackham, O. J. L., Laslett, A. L., … Lister, R. (2023).
875 Transient naive reprogramming corrects hiPS cells functionally and epigenetically.
876 *Nature*, *620*(7975), Article 7975. https://doi.org/10.1038/s41586-023-06424-7

877 Butcher, L. M., Ito, M., Brimpari, M., Morris, T., Soares, F. A. C., Ährlund-Richter, L., Carey, N.,
878 Vallier, L., Ferguson-Smith, A. C., & Beck, S. (2016). Non-CG DNA methylation is a
879 biomarker for assessing endodermal differentiation capacity in pluripotent stem cells.
880 *Nature Communications*, *7*(1), 10458–10458. https://doi.org/10.1038/ncomms10458

881 Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm,
882 F., Chavez, L., Reuss, D. E., Kratz, A., Wefers, A. K., Huang, K., Pajtler, K. W., Schweizer,
883 L., Stichel, D., Olar, A., Engel, N. W., Lindenberg, K., … Pfister, S. M. (2018). DNA

884 methylation-based classification of central nervous system tumours. *Nature*,
885 *555*(7697), 469–474. https://doi.org/10.1038/nature26000

886 Cerneckis, J., Cai, H., & Shi, Y. (2024). Induced pluripotent stem cells (iPSCs): Molecular
887 mechanisms of induction and applications. *Signal Transduction and Targeted Therapy*,
888 *9*(1), 1–26. https://doi.org/10.1038/s41392-024-01809-0

889 Chen, Y., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., Gallinger,
890 S., Hudson, T. J., & Weksberg, R. (2013). Discovery of cross-reactive probes and
891 polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.
892 *Epigenetics*, *8*(2), 203. https://doi.org/10.4161/epi.23470

893 Chirica, M., Jurmeister, P., Teichmann, D., Koch, A., Perez, E., Schmid, S., Simon, M., Driever,
894 P. H., Bodden, C., van Tilburg, C. M., Hardin, E. C., Lavarino, C., Hench, J., Scheie, D.,
895 Cryan, J., Vicha, A., Buttarelli, F. R., Michiels, A., Haberler, C., … Capper, D. (2024). DNA
896 methylation-array interlaboratory comparison trial demonstrates highly reproducible
897 paediatric CNS tumour classification across 13 international centres. *Neuropathology*
898 *and Applied Neurobiology*, *50*(5), e13010. https://doi.org/10.1111/nan.13010

899 Choi, W.-Y., Hwang, J.-H., Cho, A.-N., Lee, A. J., Lee, J., Jung, I., Cho, S.-W., Kim, L. K., & Kim,
900 Y.-J. (2020). DNA Methylation of Intragenic CpG Islands are Required for
901 Differentiation from iPSC to NPC. *Stem Cell Reviews and Reports*, *16*(6), 1316–1327.
902 https://doi.org/10.1007/s12015-020-10041-6

903 Crochemore, C., Chica, C., Garagnani, P., Lattanzi, G., Horvath, S., Sarasin, A., Franceschi, C.,
904 Bacalini, M. G., & Ricchetti, M. (2023). Epigenomic signature of accelerated ageing in
905 progeroid Cockayne syndrome. *Aging Cell*, *22*(10), e13959.
906 https://doi.org/10.1111/acel.13959

907 Cypris, O., Franzen, J., Frobel, J., Glück, P., Kuo, C.-C., Schmitz, S., Nüchtern, S., Zenke, M., &
908 Wagner, W. (2022). Hematopoietic differentiation persists in human iPSCs defective
909 in de novo DNA methylation. *BMC Biology*, *20*(1), 141.
910 https://doi.org/10.1186/s12915-022-01343-x

911 Daenekas, B., Pérez, E., Boniolo, F., Stefan, S., Benfatto, S., Sill, M., Sturm, D., Jones, D. T. W.,
912 Capper, D., Zapatka, M., & Hovestadt, V. (2024). Conumee 2.0: Enhanced copy-
913 number variation analysis from DNA methylation arrays for humans and mice.
914 *Bioinformatics*, *40*(2), btae029. https://doi.org/10.1093/bioinformatics/btae029

915 Daily, K., Ho Sui, S. J., Schriml, L. M., Dexheimer, P. J., Salomonis, N., Schroll, R., Bush, S.,
916 Keddache, M., Mayhew, C., Lotia, S., Perumal, T. M., Dang, K., Pantano, L., Pico, A. R.,

Grassman, E., Nordling, D., Hide, W., Hatzopoulos, A. K., Malik, P., … Omberg, L. (2017). Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives. *Scientific Data*, *4*(1), 170030. https://doi.org/10.1038/sdata.2017.30

Danielsson, A., Nemes, S., Tisell, M., Lannering, B., Nordborg, C., Sabel, M., & Carén, H. (2015). MethPed: A DNA methylation classifier tool for the identification of pediatric brain tumor subtypes. *Clinical Epigenetics*, *7*(1), 62. https://doi.org/10.1186/s13148-015-0103-3

Davis, S., & Meltzer, P. S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, *23*(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

Diane, A., Mu-U-Min, R. B. A., & Al-Siddiqi, H. H. (2025). Epigenetic memory as crucial contributing factor in directing the differentiation of human iPSC into pancreatic β-cells in vitro. *Cell and Tissue Research*, *399*(3), 267–276. https://doi.org/10.1007/s00441-025-03952-8

Dobner, J., Diecke, S., Krutmann, J., Prigione, A., & Rossi, A. (2024). Reassessment of marker genes in human induced pluripotent stem cells for enhanced quality control. *Nature Communications*, *15*(1), 8547. https://doi.org/10.1038/s41467-024-52922-1

Dragomir, M. P., Calina, T. G., Perez, E., Schallenberg, S., Chen, M., Albrecht, T., Koch, I., Wolkenstein, P., Goeppert, B., Roessler, S., Calin, G. A., Sers, C., Horst, D., Roßner, F., & Capper, D. (2023). DNA methylation-based classifier differentiates intrahepatic pancreato-biliary tumours. *eBioMedicine*, *93*. https://doi.org/10.1016/j.ebiom.2023.104657

Edgar, R. D., Perrone, F., Foster, A. R., Payne, F., Lewis, S., Nayak, K. M., Kraiczy, J., Cenier, A., Torrente, F., Salvestrini, C., Heuschkel, R., Hensel, K. O., Harris, R., Jones, D. L., Zerbino, D. R., & Zilbauer, M. (2022). Culture-Associated DNA Methylation Changes Impact on Cellular Function of Human Intestinal Organoids. *Cellular and Molecular Gastroenterology and Hepatology*, *14*(6), 1295–1310. https://doi.org/10.1016/j.jcmgh.2022.08.008

Elsafi Mabrouk, M. H., Zeevaert, K., Henneke, A.-C., Maaßen, C., & Wagner, W. (2024). Substrate elasticity does not impact DNA methylation changes during differentiation of pluripotent stem cells. *Cytotherapy*, *26*(9), 1046–1051. https://doi.org/10.1016/j.jcyt.2024.03.485

Euskirchen, P., Bielle, F., Labreche, K., Kloosterman, W. P., Rosenberg, S., Daniau, M., Schmitt, C., Masliah-Planchon, J., Bourdeaut, F., Dehais, C., Marie, Y., Delattre, J.-Y., & Idbaih, A. (2017). Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathologica*, *134*(5), 691–703. https://doi.org/10.1007/s00401-017-1743-5

Franzen, J., Georgomanolis, T., Selich, A., Kuo, C.-C., Stöger, R., Brant, L., Mulabdić, M. S., Fernandez-Rebollo, E., Grezella, C., Ostrowska, A., Begemann, M., Nikolić, M., Rath, B., Ho, A. D., Rothe, M., Schambach, A., Papantonis, A., & Wagner, W. (2021). DNA methylation changes during long-term in vitro cell culture are caused by epigenetic drift. *Communications Biology*, *4*(1), 1–12. https://doi.org/10.1038/s42003-021-02116-y

Gibney, E. R., & Nolan, C. M. (2010). Epigenetics and gene expression. *Heredity*, *105*(1), 4–13. https://doi.org/10.1038/hdy.2010.54

Grancharova, T., Gerbin, K. A., Rosenberg, A. B., Roco, C. M., Arakaki, J. E., DeLizo, C. M., Dinh, S. Q., Donovan-Maiye, R. M., Hirano, M., Nelson, A. M., Tang, J., Theriot, J. A., Yan, C., Menon, V., Palecek, S. P., Seelig, G., & Gunawardane, R. N. (2021). A comprehensive analysis of gene expression changes in a high replicate and open-source dataset of differentiating hiPSC-derived cardiomyocytes. *Scientific Reports*, *11*(1), 15845. https://doi.org/10.1038/s41598-021-94732-1

Hajmousa, G., de Almeida, R. C., Bloks, N., Ruiz, A. R., Bouma, M., Slieker, R., Kuipers, T. B., Nelissen, R. G. H. H., Ito, K., Freund, C., Ramos, Y. F. M., & Meulenbelt, I. (2024). The role of DNA methylation in chondrogenesis of human iPSCs as a stable marker of cartilage quality. *Clinical Epigenetics*, *16*(1), 141. https://doi.org/10.1186/s13148-024-01759-y

Hernández-Ortega, K., Canul-Euan, A. A., Solis-Paredes, J. M., Borboa-Olivares, H., Reyes-Muñoz, E., Estrada-Gutierrez, G., & Camacho-Arroyo, I. (2024). S100B actions on glial and neuronal cells in the developing brain: An overview. *Frontiers in Neuroscience*, *18*. https://doi.org/10.3389/fnins.2024.1425525

Hoang, D.-T., Shulman, E. D., Turakulov, R., Abdullaev, Z., Singh, O., Campagnolo, E. M., Lalchungnunga, H., Stone, E. A., Nasrallah, M. P., Ruppin, E., & Aldape, K. (2024). Prediction of DNA methylation-based tumor types from histopathology in central nervous system tumors with deep learning. *Nature Medicine*, *30*(7), 1952–1961. https://doi.org/10.1038/s41591-024-02995-8

983 Horii, M., Morey, R., Bui, T., Touma, O., Nelson, K. K., Cho, H.-Y., Rishik, H., Laurent, L. C., &
984 Parast, M. M. (2021). Modeling preeclampsia using human induced pluripotent stem
985 cells. *Scientific Reports*, *11*(1), 5877. https://doi.org/10.1038/s41598-021-85230-5

986 Ideno, H., Imaizumi, K., Shimada, H., Sanosaka, T., Nemoto, A., Kohyama, J., & Okano, H.
987 (2022). Human PSCs determine the competency of cerebral organoid differentiation
988 via FGF signaling and epigenetic mechanisms. *iScience*, *25*(10), 105140.
989 https://doi.org/10.1016/j.isci.2022.105140

990 Jacob, A., Vedaie, M., Roberts, D. A., Thomas, D. C., Villacorta-Martin, C., Alysandratos, K.-D.,
991 Hawkins, F., & Kotton, D. N. (2019). Derivation of self-renewing lung alveolar epithelial
992 type II cells from human pluripotent stem cells. *Nature Protocols*, *14*(12), 3303–3332.
993 https://doi.org/10.1038/s41596-019-0220-0

994 Johannesson, B., Sagi, I., Gore, A., Paull, D., Yamada, M., Golan-Lev, T., Li, Z., LeDuc, C. A.,
995 Shen, Y., Yufeng Shen, Stern, S., Xu, N., Hong Ma, Ma, H., Hong Ma, Hong Ma, Kang,
996 E., Mitalipov, S., Sauer, M. V., … Egli, D. (2014). Comparable Frequencies of Coding
997 Mutations and Loss of Imprinting in Human Pluripotent Cells Derived by Nuclear
998 Transfer and Defined Factors. *Cell Stem Cell*, *15*(5), 634–642.
999 https://doi.org/10.1016/j.stem.2014.10.002

1000 Jones, P. A., & Taylor, S. M. (1980). Cellular differentiation, cytidine analogs and DNA
1001 methylation. *Cell*, *20*(1), 85–93. https://doi.org/10.1016/0092-8674(80)90237-8

1002 Jurmeister, P., Glöß, S., Roller, R., Leitheiser, M., Schmid, S., Mochmann, L. H., Payá Capilla,
1003 E., Fritz, R., Dittmayer, C., Friedrich, C., Thieme, A., Keyl, P., Jarosch, A., Schallenberg,
1004 S., Bläker, H., Hoffmann, I., Vollbrecht, C., Lehmann, A., Hummel, M., … Capper, D.
1005 (2022). DNA methylation-based classification of sinonasal tumors. *Nature*
1006 *Communications*, *13*(1), 7148. https://doi.org/10.1038/s41467-022-34815-3

1007 Kitayama, S., Zhang, R., Liu, T.-Y., Ueda, N., Iriguchi, S., Yasui, Y., Kawai, Y., Tatsumi, M., Hirai,
1008 N., Mizoro, Y., Iwama, T., Watanabe, A., Nakanishi, M., Kuzushima, K., Uemura, Y., &
1009 Kaneko, S. (2016). Cellular Adjuvant Properties, Direct Cytotoxicity of Re-
1010 differentiated Vα24 Invariant NKT-like Cells from Human Induced Pluripotent Stem
1011 Cells. *Stem Cell Reports*, *6*(2), 213–227. https://doi.org/10.1016/j.stemcr.2016.01.005

1012 Kramer, N., Rosner, M., Kovacic, B., & Hengstschläger, M. (2016). Full biological
1013 characterization of human pluripotent stem cells will open the door to translational
1014 research. *Archives of Toxicology*, *90*(9), 2173–2186. https://doi.org/10.1007/s00204-
1015 016-1763-2

1016 Kurdyukov, S., & Bullock, M. (2016). DNA Methylation Analysis: Choosing the Right Method.
1017 *Biology*, *5*(1), 3. https://doi.org/10.3390/biology5010003

1018 Kuschel, L. P., Hench, J., Frank, S., Hench, I. B., Girard, E., Blanluet, M., Masliah-Planchon, J.,
1019 Misch, M., Onken, J., Czabanka, M., Yuan, D., Lukassen, S., Karau, P., Ishaque, N., Hain,
1020 E. G., Heppner, F., Idbaih, A., Behr, N., Harms, C., … Euskirchen, P. (2022). Robust
1021 methylation-based classification of brain tumours using nanopore sequencing.
1022 *Neuropathology and Applied Neurobiology*. https://doi.org/10.1111/nan.12856

1023 Lanza, R., Russell, D. W., & Nagy, A. (2019). Engineering universal cells that evade immune
1024 detection. *Nature Reviews. Immunology*, *19*(12), 723–733.
1025 https://doi.org/10.1038/s41577-019-0200-1

1026 Lenz, M. (2012). PluriTest: Molecular diagnostic assay for pluripotency in human stem cells.
1027 In *Human Stem Cell Manual* (pp. 293–311). Elsevier. https://doi.org/10.1016/b978-0-
1028 12-385473-5.00018-7

1029 Liuyang, S., Wang, G., Wang, Y., He, H., Lyu, Y., Cheng, L., Yang, Z., Guan, J., Fu, Y., Zhu, J.,
1030 Zhong, X., Sun, S., Li, C., Wang, J., & Deng, H. (2023). Highly efficient and rapid
1031 generation of human pluripotent stem cells by chemical reprogramming. *Cell Stem*
1032 *Cell*, *30*(4), 450-459.e9. https://doi.org/10.1016/j.stem.2023.02.008

1033 Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-
1034 Porat, S., Hecht, M., Pelet, T., Moss, J., Drawshy, Z., Amini, H., Moradi, P., Nagaraju, S.,
1035 Bauman, D., Shveiky, D., Porat, S., Dior, U., … Kaplan, T. (2023). A DNA methylation
1036 atlas of normal human cell types. *Nature*, *613*(7943), 355–364.
1037 https://doi.org/10.1038/s41586-022-05580-6

1038 Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C.,
1039 Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W.
1040 (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: A
1041 summary. *Neuro-Oncology*, *23*(8), 1231–1251.
1042 https://doi.org/10.1093/neuonc/noab106

1043 Ludwig, T. E., Andrews, P. W., Barbaric, I., Benvenisty, N., Bhattacharyya, A., Crook, J. M.,
1044 Daheron, L. M., Draper, J. S., Healy, L. E., Huch, M., Inamdar, M. S., Jensen, K. B., Kurtz,
1045 A., Lancaster, M. A., Liberali, P., Lutolf, M. P., Mummery, C. L., Pera, M. F., Sato, Y., …
1046 Mosher, J. T. (2023). ISSCR standards for the use of human stem cells in basic research.
1047 *Stem Cell Reports*, *18*(9), 1744–1752. https://doi.org/10.1016/j.stemcr.2023.08.003

1048 MacArthur, C. C., Pradhan ,Suman, Wetton ,Nichole, Zarrabi ,Aryan, Dargitz ,Carl, Sridharan
1049 ,Mahalakshmi, Jackson ,Stephen, Pickle ,Loni, & and Lakshmipathy, U. (2019).
1050 Generation and Comprehensive Characterization of Induced Pluripotent Stem Cells for
1051 Translational Research. *Regenerative Medicine*, *14*(6), 505–524.
1052 https://doi.org/10.2217/rme-2018-0148

1053 Maksimovic, J., Phipson, B., & Oshlack, A. (2017). A cross-package Bioconductor workflow for
1054 analysing methylation array data. *F1000Research*, *5*, 1281.
1055 https://doi.org/10.12688/f1000research.8839.3

1056 Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz,
1057 M., Druid, H., Arner, P., Fu, K.-Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A.,
1058 Grinshpun, A., Shapiro, A. M. J., Grompe, M., Wittenberg, A. D., … Dor, Y. (2018).
1059 Comprehensive human cell-type methylation atlas reveals origins of circulating cell-
1060 free DNA in health and disease. *Nature Communications*, *9*(1), 5068.
1061 https://doi.org/10.1038/s41467-018-07466-6

1062 Nishino, K., Takasawa, K., Okamura, K., Arai, Y., Sekiya, A., Akutsu, H., & Umezawa, A. (2021).
1063 Identification of an epigenetic signature in human induced pluripotent stem cells using
1064 a linear machine learning model. *Human Cell*, *34*(1), 99–110.
1065 https://doi.org/10.1007/s13577-020-00446-3

1066 Nishizawa, M., Chonabayashi, K., Nomura, M., Tanaka, A., Nakamura, M., Inagaki, A.,
1067 Nishikawa, M., Takei, I., Oishi, A., Tanabe, K., Ohnuki, M., Yokota, H., Koyanagi-Aoi, M.,
1068 Okita, K., Watanabe, A., Takaori-Kondo, A., Yamanaka, S., & Yoshida, Y. (2016).
1069 Epigenetic Variation between Human Induced Pluripotent Stem Cell Lines Is an
1070 Indicator of Differentiation Capacity. *Cell Stem Cell*, *19*(3), 341–354.
1071 https://doi.org/10.1016/j.stem.2016.06.019

1072 Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer,
1073 L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M.,
1074 Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., …
1075 McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance
1076 and exemplars for reporting systematic reviews. *The BMJ*, *372*, n160.
1077 https://doi.org/10.1136/bmj.n160

1078 Parry, A., Rulands, S., & Reik, W. (2021). Active turnover of DNA methylation during cell fate
1079 decisions. *Nature Reviews Genetics*, *22*(1), 59–66. https://doi.org/10.1038/s41576-
1080 020-00287-8

1081 Patsch, C., Challet-Meylan, L., Thoma, E. C., Urich, E., Heckel, T., O'Sullivan, J. F., Grainger, S.
1082 J., Kapp, F. G., Sun, L., Christensen, K., Xia, Y., Florido, M. H. C., He, W., Pan, W.,
1083 Prummer, M., Warren, C. R., Jakob-Roetne, R., Certa, U., Jagasia, R., … Cowan, C. A.
1084 (2015). Generation of vascular endothelial and smooth muscle cells from human
1085 pluripotent stem cells. *Nature Cell Biology*, *17*(8), 994–1003.
1086 https://doi.org/10.1038/ncb3205

1087 Phipson, B., Maksimovic, J., & Oshlack, A. (2016). missMethyl: An R package for analyzing data
1088 from Illumina's HumanMethylation450 platform. *Bioinformatics (Oxford, England)*,
1089 *32*(2), 286–288. https://doi.org/10.1093/bioinformatics/btv560

1090 Puri, D., Maaßen, C., Varona Baranda, M., Zeevaert, K., Hahnfeld, L., Hauser, A., Fornero, G.,
1091 Elsafi Mabrouk, M. H., & Wagner, W. (2023). CTCF deletion alters the pluripotency and
1092 DNA methylation profile of human iPSCs. *Frontiers in Cell and Developmental Biology*,
1093 *11*. https://doi.org/10.3389/fcell.2023.1302448

1094 Schmidt, M., Zeevaert, K., Elsafi Mabrouk, M. H., Goetzke, R., & Wagner, W. (2022). Epigenetic
1095 biomarkers to track differentiation of pluripotent stem cells. *Stem Cell Reports*.
1096 https://doi.org/10.1016/j.stemcr.2022.11.001

1097 Shi, Y., Inoue, H., Wu, J. C., & Yamanaka, S. (2017). Induced pluripotent stem cell technology:
1098 A decade of progress. *Nature Reviews Drug Discovery*, *16*(2), 115–130.
1099 https://doi.org/10.1038/nrd.2016.245

1100 Sturchler, E., Cox, J. A., Durussel, I., Weibel, M., & Heizmann, C. W. (2006). S100A16, a novel
1101 calcium-binding protein of the EF-hand superfamily. *The Journal of Biological*
1102 *Chemistry*, *281*(50), 38905–38917. https://doi.org/10.1074/jbc.M605798200

1103 Sultanov, R., Lebedeva, O., Arapidi, G., Lagarkova, M., & Kiselev, S. (2018). Methylation profile
1104 of induced pluripotent stem cells generated by integration and integration-free
1105 approaches. *Data in Brief*, *17*, 662–666. https://doi.org/10.1016/j.dib.2018.01.061

1106 Tanaka, Y., Furuhata, E., Maeda, S., Kishima, M., Suzuki, H., & Suzuki, T. (2023). A dataset of
1107 definitive endoderm and hepatocyte differentiations from human induced pluripotent
1108 stem cells. *Scientific Data*, *10*(1), 93. https://doi.org/10.1038/s41597-023-02001-9

1109 TCW, J., Wang, M., Pimenova, A. A., Bowles, K. R., Hartley, B. J., Lacin, E., Machlovi, S. I.,
1110 Abdelaal, R., Karch, C. M., Phatnani, H., Slesinger, P. A., Zhang, B., Goate, A. M., &
1111 Brennand, K. J. (2017). An Efficient Platform for Astrocyte Differentiation from Human
1112 Induced Pluripotent Stem Cells. *Stem Cell Reports*, *9*(2), 600–614.
1113 https://doi.org/10.1016/j.stemcr.2017.06.018

1114 Tesarova, L., Simara, P., Stejskal, S., & Koutna, I. (2016). The Aberrant DNA Methylation Profile
1115 of Human Induced Pluripotent Stem Cells Is Connected to the Reprogramming Process
1116 and Is Normalised During In Vitro Culture. *PLOS ONE*, *11*(6), e0157974.
1117 https://doi.org/10.1371/journal.pone.0157974

1118 Tompkins, J. D. (2022). Discovering DNA Methylation, the History and Future of the Writing
1119 on DNA. *Journal of the History of Biology*, *55*(4), 865–887.
1120 https://doi.org/10.1007/s10739-022-09691-8

1121 Tran, Q. T., Breuer, A., Lin, T., Tatevossian, R., Allen, S. J., Clay, M., Furtado, L. V., Chen, M.,
1122 Hedges, D., Michael, T., Robinson, G., Northcott, P., Gajjar, A., Azzato, E., Shurtleff, S.,
1123 Ellison, D. W., Pounds, S., & Orr, B. A. (2024). Comparison of DNA methylation based
1124 classification models for precision diagnostics of central nervous system tumors. *Npj*
1125 *Precision Oncology*, *8*(1), 1–12. https://doi.org/10.1038/s41698-024-00718-3

1126 Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., & Siegmund, K. D. (2013). Low-
1127 level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids*
1128 *Research*, *41*(7), e90. https://doi.org/10.1093/nar/gkt090

1129 Tsankov, A. M., Akopian, V., Pop, R., Chetty, S., Gifford, C. A., Daheron, L., Tsankova, N. M., &
1130 Meissner, A. (2015). A qPCR ScoreCard quantifies the differentiation potential of
1131 human pluripotent stem cells. *Nature Biotechnology*, *33*(11), Article 11.
1132 https://doi.org/10.1038/nbt.3387

1133 Vallone, V. F., Ludwik, K., Healy, L., Lefort, N., Onder, T., Pavinato, L., Okur, F. V. F., &
1134 Stachelscheid, H. (2025). Good practices in hPSC handling.
1135 https://www.protocols.io/view/good-practices-in-hpsc-handling-gdqmbs5u7

1136 Wang, S., Du, Y., Zhang, B., Meng, G., Liu, Z., Liew, S. Y., Liang, R., Zhang, Z., Cai, X., Wu, S.,
1137 Gao, W., Zhuang, D., Zou, J., Huang, H., Wang, M., Wang, X., Wang, X., Liang, T., Liu,
1138 T., … Shen, Z. (2024). Transplantation of chemically induced pluripotent stem-cell-
1139 derived islets under abdominal anterior rectus sheath in a type 1 diabetes patient.
1140 Cell, S0092867424010225. https://doi.org/10.1016/j.cell.2024.09.004

1141 Yang, H., Yang, Y., Kiskin, F. N., Shen, M., & Zhang, J. Z. (2023). Recent advances in regulating
1142 the proliferation or maturation of human-induced pluripotent stem cell-derived
1143 cardiomyocytes. *Stem Cell Research & Therapy*, *14*(1), 228.
1144 https://doi.org/10.1186/s13287-023-03470-w

1145 Yoshihara, M., Hayashizaki, Y., & Murakawa, Y. (2017). Genomic Instability of iPSCs:
1146 Challenges Towards Their Clinical Applications. *Stem Cell Reviews*, *13*(1), 7–16.
1147 https://doi.org/10.1007/s12015-016-9680-6

Acknowledgements

This work was supported by The German Academic Exchange Service (DAAD) [E.P.] and the Berlin School of Integrative Oncology [E.P.]. The work was further supported by the German Cancer Consortium (DKTK), Partner Site Berlin [D.C.]. The authors thank Christine Sers, Gaetano Gargiulio, Frank Heppner and all members of the Heppner lab for helpful discussions and suggestions for this work.

Conflicting Interests

The authors have no conflicting interests to declare.

Author contributions

E.P. performed differentiation experiments, generated DNA methylation data, conducted bioinformatic analyses, prepared figures and tables and wrote the manuscript. V.F.-V. performed differentiation experiments and prepared figures and tables. D.T. and P.W. performed DNA extraction and EPIC array processing. C.S. and B.O. contributed to the data processing workflow. P.E. supported nanopore sequencing-based methylation profiling. D.C., V.F.-V. and H.S. conceived and supervised the study. All authors contributed to manuscript editing and approved the final version.

Data Availability

Web application source code is available at https://github.com/pereze5/SteMClass_App, all annotation, reference and test data required to reproduce the results reported here are available at https://github.com/pereze5/SteMClass_App/releases and raw data are deposited at [GEO Accession TBC].