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 0.018, 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 = 249), SteMClass achieved 85.1% overall accuracy (Cohen’s K = 0.687) with a 12.9% rejection rate. Among classified samples (n = 217), accuracy was 97.7% (Cohen’s K = 0.93). 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, Kügelgen et al., 2025). 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
101 have been transformed by a classifier built on these principles, culminating in their formal
102 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 be leveraged to identify cells at defined stages of iPSC differentiation. Our method employs 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™ (Gibco) in Essential 8™ Medium supplemented with 10 µM Y-27632 (Selleck Chemicals). Cells were cultured for 2 days with daily medium changes using Essential 8™ Medium. After 2 days, the medium was replaced with neural induction medium composed of DMEM/F12 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™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™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.

272 Analysis of Publicly Available Datasets

273 A publicly available methylation array dataset was systematically curated and downloaded
274 from the following Gene Expression Omnibus (GEO) records as raw signal intensity files (IDAT
275 files):

276 GSE105093, GSE110544, GSE141521, GSE159479, GSE163322, GSE163324, GSE180402,
277 GSE196339, GSE197723, GSE207076, GSE207108, GSE236202, GSE247428, GSE253190,
278 GSE61461, GSE76372, GSE85828, GSE196339. The PRISMA (Preferred Reporting Items for
279 Systematic Reviews and Meta-Analyses; (Page et al., 2021)) flow chart illustrating the number

280 of datasets identified, included and excluded, and the reason for exclusions is available in
281 Supplementary Figure 6. All samples used in this study are outlined in Supplementary Table
282 5. GenomicMethylSet objects were generated from raw signal intensity files using the
283 readGEORawFile function of the GEOquery package (v2.62.1). (Davis & Meltzer, 2007).

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

291 Differential Methylation Analysis

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

304 Unsupervised Analysis

305 Principal component analysis (PCA) was performed on DNA methylation beta values to
306 visualise methylation differences across samples during iPSC to astrocyte differentiation
307 (n=35). For each CpG probe, the standard deviation across samples was calculated, and the
308 5,000 most variable probes were selected. The resulting beta value matrix was transposed
309 such that rows represented samples and columns represented CpG sites. PCA was conducted
310 using the prcomp function in R with mean centring and scaling enabled. The proportion of
311 variance explained by each principal component was calculated as the squared standard
312 deviation of each component divided by the total variance. The first two principal
313 components were used for plotting. Uniform Manifold Approximation and Projection (UMAP)
314 was performed on the top 50,000 most variable CpGs for each respective dataset using the R
315 package umap (v 0.2.10.0). UMAP projections were computed using the uwot R package (v

316 0.2.2). UMAP parameters were set to n_neighbors=18, min_dist=0.3 n_components=2, with
317 a Euclidean distance metric and a fixed random state for reproducibility. The resulting UMAP
318 layout was integrated with sample metadata for downstream analysis. A scatter plot of UMAP
319 coordinates was generated using ggplot2, with points representing individual samples, the
320 final plots depicting the distribution of samples and cell types based on the top 50,000 most
321 variable probes across samples.

322 Classifier development

323 A random forest classifier was built with ranger (v 0.16.0) within the tidymodels ecosystem
324 (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
325 v 1.2.1; yardstick v 1.3.0) under R 4.3.3. Parallel execution used BiocParallel (v 1.36.1) and
326 doParallel (v 1.0-17).

327 A nested cross-validation framework was used to obtain unbiased estimates of model
328 performance while preventing information leakage (5-fold outer cross-validation repeated
329 three times; 3-fold inner cross-validation).

330 **Nested cross validation, feature selection and hyperparameter tuning.**

331 Within each outer training split, CpG sites were first variance-filtered using the training data
332 only, retaining the top 50,000 CpGs ranked by standard deviation to reduce dimensionality.
333 Random forest hyperparameters were tuned exclusively in the inner cross-validation using a
334 20-point max-entropy (Latin hypercube) design over minimum node size (2-30) and number
335 of trees (300-2000), while mtry was fixed per split to the square root of the number of
336 retained predictors. Hyperparameters were selected by minimising the multiclass Brier score.
337 Using the selected hyperparameters, the model was refit on the full outer training split and
338 permutation importance was computed for all variance-filtered CpGs. CpGs were ranked by
339 permutation importance, and nested feature selection was conducted by fitting models on
340 increasing prefixes of this ranked list (2,000, 5,000, and 10,000 CpGs) using the same inner
341 cross-validation structure. The final feature count per outer split was determined using a one-
342 standard-error plateau rule on the inner-CV Brier score, selecting the smallest feature set
343 whose performance was within one standard error of the minimum. This conservative
344 criterion reduces the likelihood of selecting an overfit model driven by sampling variability.

345 Across outer splits, selected CpGs were aggregated by selection frequency. The final deployed
346 feature set comprised the top-ranked CpGs up to the median selected feature count across
347 splits. Consensus hyperparameters derived from the nested procedure were then used to refit
348 the model on all training data.

349

Probability score calibration and threshold identification.

For probability calibration, outer fold predictions were pooled, and a multinomial ridge regression model was trained in a cross-fitted manner: predictions for each outer fold were calibrated using a model trained only on predictions from the remaining outer folds. This ensured that no sample was used to fit either the base classifier or the calibrator responsible for its calibrated probabilities. Calibration performance was summarised using the multiclass Brier score.

To determine a rejection threshold, the macro-averaged Youden index was evaluated across candidate probability cutoffs using the pooled cross-fitted calibrated predictions. Uncertainty was quantified by nonparametric bootstrap resampling. The selected operating point was defined as the largest threshold on the macro-Youden plateau, that is, the highest cutoff before any further increase resulted in a statistically significant decrease in macro-Youden based on the 95 percent bootstrap confidence interval. Samples with maximum predicted class probability below this threshold were assigned to a “Not Classifiable” category.

Final model and variable importance

The final random forest model was trained on the entire training cohort using the final feature set (10,000 CpGs) and consensus hyperparameters (mtry= 100, min_n = 5, trees = 1192) with a calibrated probability score threshold of 0.5. Variable importance was extracted from the fitted ranger object using the impurity-based measure. The trained classifier was serialised and saved as a workflow object for downstream application.

Model evaluation

Performance was assessed using two independent datasets:

1. Internal validation cohort (n = 58).
2. External validation cohort (n = 249, publicly available methylation array data).

The pre-trained random forest classifier was loaded from the serialised workflow object and the underlying ranger model was extracted. The pre-processing recipe was reconstructed using the recipes package and applied to generate a numeric feature matrix and factor response vector. Feature-level contributions for each test sample and a training cohort bias vector were computed from the tidy forest object. Samples from the validation cohorts which received a classification discordant with their annotation were identified. For each such sample, the top CpG features with the most negative contribution toward the annotated class and the most positive contribution towards the SteMClass classification (i.e., features driving their discordant classification) were extracted for further analysis.

383 **SteMClass Web Application**

384 SteMClass was developed in R using the Shiny framework with shiny, shinyjs, shinycssloaders,
385 bslib, and thematic, employing a Bootstrap theme and custom CSS for interface rendering.
386 DNA methylation preprocessing relies on minfi, methylumi, wateRmelon and preprocessCore
387 for raw IDAT file handling and normalisation, with array manifests and annotation
388 (IlluminaMethylationManifest) loaded dynamically. Preprocessing includes noob
389 normalisation, conversion to a common 450K format where required, and genome mapping.
390 The pre-trained SteMClass random forest model (loaded from GitHub Releases
391 (https://github.com/pereze5/SteMClass_App/releases) is applied after aligning test data to
392 the training reference cohort and imputing missing CpGs using means learned during model
393 development. Predictions are output as class probabilities across eight reference
394 differentiation states, with a rejection option for samples not reaching the 0.6 probability
395 threshold. Visualisations include probability bar charts, UMAP projections generated with
396 uwot, and heatmaps generated with ComplexHeatmap and circlize. Global heatmaps
397 compare the test sample to reference profiles using the top 10,000 CpGs distinguishing a
398 selected differentiation state from iPSC, while gene-level heatmaps display CpGs annotated
399 to a specified gene (UCSC_RefGene_Name) ordered by genomic position and annotated by
400 gene region (UCSC_RefGene_Group). Results and figures are exportable as high-resolution
401 images or HTML outputs.

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

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

412 ## Nanopore Sequencing

413 Library preparation with barcode labelling was performed with ~200-400 ng input of genomic
414 DNA using the Rapid Barcoding Kit (SQK-RBK114.24, Oxford Nanopore Technologies, UK)
415 according to the manufacturer's instructions. During the library preparation, input DNA is
416 fragmented while simultaneously attaching barcodes using a time-efficient transposase-
417 based approach known as “tagmentation”. The final library was loaded onto an R10.4.1 flow
418 cell (FLO-MIN114, Oxford Nanopore Technologies, UK), and low coverage whole genome

419 sequencing (WGS) was performed for 6-24 hrs on a MinION Mk 1B device (Oxford Nanopore
420 Technologies, UK). POD5 files containing the raw data were obtained in real time using the
421 manufacturer's software MinKNOW (v.1.3.1-v.3.6.0) and transferred to a high-performance
422 computing (HPC) cluster for further analysis. Analysis was performed using the nanoDx
423 classification pipeline v0.6.2 (Kuschel et al., 2022). The SteMClass training cohort raw IDAT
424 files were processed and converted into an H5 file format for use as the training cohort for
425 the nanoDx pipeline using the R package rhdf5 (v2.50.2). The nanoDx classification and
426 analysis pipeline is publicly available at https://gitlab.com/pesk/nanoDx. The source code for
427 the ad hoc random forest implementation is available at
428 https://gitlab.com/pesk/nanoBenchmark.

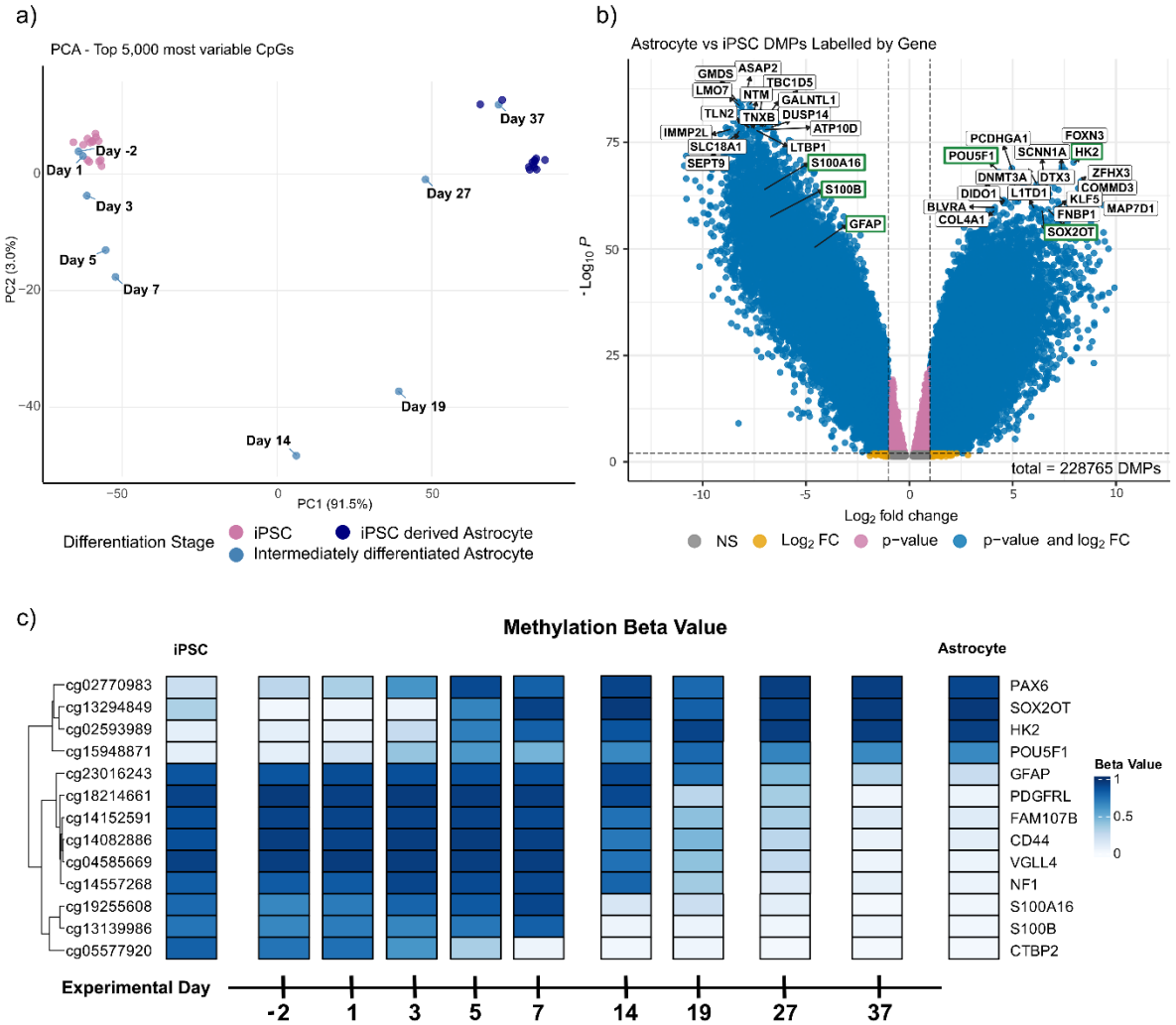
429 # Results

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

431 DNA methylation is well established as a marker of cell- and tissue-type identity in vivo. To
432 determine whether this specificity is preserved in isogenic cells in vitro, we analysed the DNA
433 methylation profiles of two parental iPSC lines differentiated to astrocyte-like cells, as the
434 length of the differentiation protocol (37 days) permitted longitudinal analysis across multiple
435 differentiation timepoints. For this we used a previously published protocol (TCW et al., 2017)
436 using a commercially available kit which results in *GLAST* positive cells by experimental day
437 37. iPSCs were cultured for two days in E8 medium (days -2 and -1) before starting the
438 differentiation protocol (day 0). The cells were purified by magnetic activated cell sorting
439 (MACS) at the end of the protocol (day 37). Principal component analysis (PCA) of the top
440 5,000 most variable CpG sites revealed a clear separation between iPSCs (n = 14) and their
441 post-differentiation astrocyte derivatives (n = 12) (Figure 1a). We further analysed samples
442 taken at nine time points during an astrocyte differentiation (time course experiment). These
443 intermediately differentiated astrocytes clustered along a continuum between the two end-
444 point differentiation states (Figure 1a). The vast majority of variation in the methylation data
445 was captured by PC1 (91.5% variance) which separated iPSCs from differentiated astrocytes
446 and intermediate states, capturing progression along the differentiation axis. Genome-wide
447 comparison of the iPSC and differentiated astrocytes uncovered 228,765 differentially
448 methylated positions (DMPs) (Figure 1b). In comparison to iPSC, differentiated astrocytes
449 displayed promoter hypomethylation of canonical astrocyte genes such as *GFAP*, *S100B*,
450 *S100A16*, and *SLC1A3* (Achicallende et al., 2022; Hernández-Ortega et al., 2024; Sturchler et
451 al., 2006) together with promoter hypermethylation of the pluripotency regulators *POU5F1*
452 and *HK2* (Figure 1b). Tracking the same CpGs during the time course differentiation
453 experiment showed that pluripotency loci (*POU5F1*, *HK2*) began to gain methylation by days
454 3-5, whereas astrocyte-specific genes such as *S100A16* and *S100B* started to lose methylation

455 by day 14 with de-methylation of *GFAP* occurring late in the differentiation (Figure 1c). Of
456 note, the canonical ectodermal marker *PAX6* became methylated in the gene body by days 3-
457 5, at the timepoint of ectodermal differentiation when the cells are expected to have *PAX6*
458 expression by FACS.

459 Together, these results demonstrate that global DNA methylation landscapes undergo
460 extensive and progressive remodelling during in vitro differentiation of isogenic iPSCs into
461 astrocytes, following a lineage-consistent trajectory across timepoints. The concordance of
462 these changes across genetically identical lines and their orderly progression with
463 differentiation stage indicate that the observed methylation differences are associated with
464 differentiation state, rather than arising from random variation or pre-existing epigenetic
465 differences.



466

467 **Figure 1: Identification of progressive DNA methylation changes during in vitro iPSC**
468 **differentiation into astrocytes.**
469 (a) Principal component analysis (PCA) plot of DNA methylation beta values from the 5,000
470 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.
(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 astrocyte samples (n = 12). Rows correspond to CpG sites, ordered by hierarchical clustering using Pearson correlation distance. CpG loci are annotated with CpG unique IDs and their associated gene names.

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

We next investigated whether the lineage-specific methylation changes observed during astrocyte differentiation generalise to other iPSC derived cell types. Early ectodermal, mesodermal and endodermal progenitor cultures were generated with at least two independent protocols per lineage from 11 independent iPSC lines (see Methods). For each of these cell types, we analysed homogeneous cell populations based on their conventional marker expression as analysed by FACS (Supplementary Figures 1 and2). Selection of populations in which at least 85% of the cells expressed the conventional differentiation markers was done to minimise intra-sample heterogeneity. Unsupervised analysis of the 50,000 most variable CpG sites across all samples was visualised with UMAP (Figure 2a). Four discrete clusters emerged: pluripotent iPSCs and three well-separated groups corresponding to the three germ-layers ectoderm, mesoderm and endoderm. Crucially, samples produced with different protocols converged on the same germ-layer cluster, and technical replicates from separate differentiation batches co-localised, demonstrating that the differentiation state-specific DNA methylation signatures are robust and reproducible between experiments.

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

513 Unsupervised analysis of the 50,000 most variable CpG sites across all samples, visualised with
514 UMAP, produced eight well-separated clusters that mapped one-to-one onto the designated
515 differentiation states (Figure 2b). The observed clustering of distinct differentiation states was
516 independent of potential batch effects such as parental cell line, donor sex, or DNA
517 methylation array (EPIC) version (Supplementary Figure 3a-d). To interpret the
518 differentiation-associated methylation patterns evidenced by our unsupervised analysis, we
519 performed pairwise differential methylation analyses comparing iPSCs against the 7
520 differentiation states (Supplementary Figure 4, Supplementary Table 4). Gene ontology
521 enrichment of the resulting DMPs confirmed that the methylation changes align with known
522 lineage-specific biological processes in each differentiation state (Supplementary Figure 4).

523 These findings demonstrate that genome-wide DNA methylation profiles provide a robust
524 read-out of differentiation state across a diverse set of iPSC lines and differentiation
525 derivatives, underscoring their utility for benchmarking and standardising in vitro
526 differentiation characterisation.

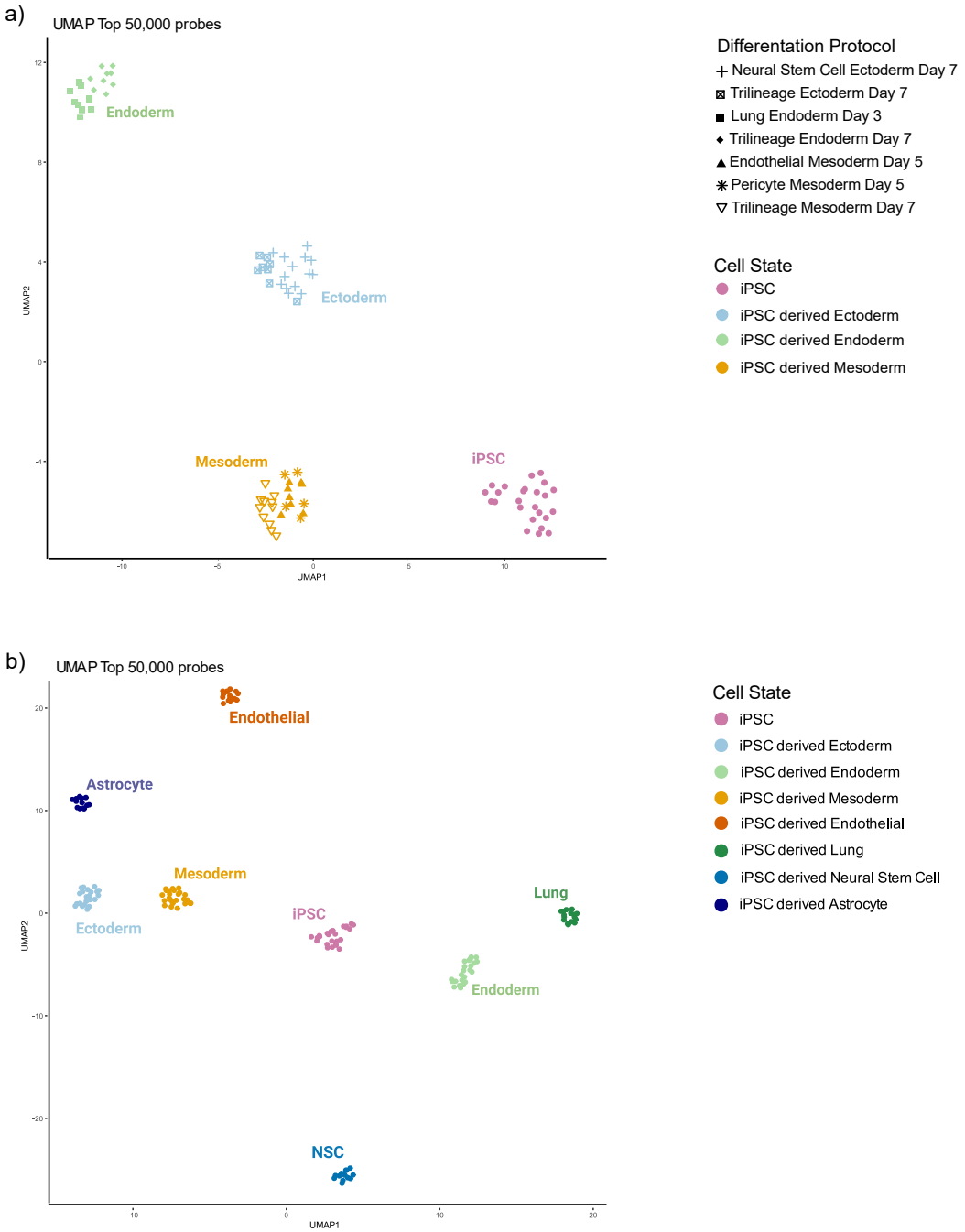


Figure 2: Unsupervised analysis indicates that differentiation state-specific DNA methylation signatures are reproducible across cell lines and protocols.

(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, 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

537 least two differentiation states per germ layer. Each point represents an independent technical and/or
538 biological replicate. Colours represent differentiation states; iPSCs are represented in pink,
539 endodermal differentiation states (endoderm and lung) are represented in green tones, ectodermal
540 differentiation states (ectoderm, neural stem cells and astrocytes) are represented in blue tones and
541 mesodermal differentiation states (mesoderm and endothelial cells) are represented in orange tones.

542 Building and Validating a DNA Methylation-Based Differentiation State 543 Classifier: SteMClass

544 On observing that the methylation differences between in vitro differentiation states were
545 reflective of differentiation processes, we leveraged this to train a random forest algorithm
546 for DNA methylation-based differentiation state classification. We chose to use a random
547 forest model due to its ability to implicitly handle feature collinearity, built-in feature
548 selection capabilities, and robustness in high-dimensional datasets, all of which are critical
549 properties when working with the correlated and high-dimensional structure of methylation
550 data. Model development followed a nested cross-validation framework (5-fold outer cross-
551 validation repeated three times; 3-fold inner cross-validation), enabling unbiased
552 performance estimation, hyperparameter tuning, feature selection, calibration, and
553 threshold determination. Within each outer split, CpGs were variance-filtered, ranked by
554 permutation importance, and the optimal feature-set size was selected using the one-
555 standard-error rule. Hyperparameters (minimum node size and number of trees; mtry was
556 fixed to the square root of the predictor count) were tuned in the inner loop using a 20-point
557 Latin-hypercube search, with model selection based on the multiclass Brier score (Methods,
558 Supplementary Figure 5a).

559 Cross-fitted calibrated probabilities obtained using a multinomial ridge regression calibrator
560 yielded a multiclass Brier score of 0.018 and a macro expected calibration error (ECE) of 0.028,
561 indicating well-calibrated predictions (Supplementary Figure 5b). To reduce overconfident
562 classifications, we evaluated candidate rejection thresholds between 0.10 and 0.95 using the
563 pooled cross-fitted calibrated predictions and the macro-averaged Youden index. The
564 selected operating point was a probability cut-off of 0.5 (macro-Youden = 0.978),
565 corresponding to the highest threshold before a statistically significant decline in macro-
566 Youden was observed (Figure 3). Lowering the threshold from 0.55 to 0.50 significantly
567 improved Macro-Youden ($p = 0.00899$), whereas lowering from 0.50 to 0.45 did not ($p =$
568 0.372). Samples with a maximum predicted class probability below 0.5 were therefore
569 rejected and labelled "Not Classifiable."

570 SteMClass was evaluated on an independent cohort of 58 samples representing the eight
571 distinct differentiation states included in the classifier and attained 96.5% accuracy, Cohen's

572 K = 0.959, and a 3% reject rate (Figure 3b). These validation samples were generated in the
573 same laboratory as the training cohort ("internally"), derived from the same iPSC lines, using
574 the same differentiation protocols. Despite these controlled conditions, the validation
575 samples represented independent experiments, reagent batches, and personnel.

576 To further evaluate SteMClass, we assembled an "external" validation cohort of 249 publicly
577 available DNA methylation array samples (Supplementary Figure 6a). Sample labels were
578 assigned based on the materials and methods described in the originating publications and
579 mapped to the differentiation states defined in this study. Importantly, these annotations
580 were generated across independent laboratories using diverse differentiation protocols and
581 variable levels of phenotypic and molecular characterisation and therefore do not constitute
582 a harmonised or uniformly defined ground truth. This heterogeneity reflects the current
583 landscape of in vitro differentiation studies and represents precisely the scenario SteMClass
584 was designed to address: objective, methylation-based state assignment across
585 experimentally variable systems. In this context, relative to literature-reported annotations
586 SteMClass demonstrated 85.1% agreement (Cohen's K = 0.687) with a 12.9% rejection rate
587 (Figure 3c). Among classified samples (n = 217), accuracy was 97.7% (Cohen's K = 0.93). To
588 visualise how many CpG probes drive the classifier decisions, we quantified the absolute
589 feature contribution of every probe for every test sample (internal and external combined)
590 and aggregated these values across samples (see Methods). The resulting cumulative-
591 contribution curves (Figure 3d) rise gradually: no class reaches 80% of its total weight before
592 1000 ranked probes and the remaining 20% is dispersed over several thousand additional
593 sites. Hence SteMClass recognises each lineage through a wide methylation footprint rather
594 than a handful of high-impact markers, making the model inherently robust to noise at
595 individual CpGs.

596 We additionally tested iPSC-derived samples from the same publicly available datasets in
597 cases where the sample metadata indicated a differentiation state outside the scope of
598 SteMClass's training. (External Out-of-Scope Validation Cohort (n=54); Supplementary Figure
599 6b, Supplementary Table 5). As expected for states outside the scope of SteMClass, most of
600 these samples were rejected. Six early-stage endoderm (< 3 days) and three embryoid body
601 samples were classified as iPSC. One embryoid body sample was classified as ectoderm and
602 one lung progenitor sample (at a differentiation stage not represented in SteMClass) was
603 classified as mesoderm; the remaining samples were not classifiable. The embryoid body
604 assignments to iPSC or ectoderm occurred in developmentally proximal states and are
605 therefore consistent with projection onto the closest reference lineage. Multiclass random

606 forests operate as closed-set classifiers. When a sample lies outside the training distribution,
607 vote fractions are distributed across all learned classes rather than withheld, and samples
608 near decision boundaries may receive a neighbouring lineage label instead of being rejected.
609 Across 303 external validation samples, only a single biologically discordant assignment (lung
610 progenitor classified as mesoderm) was observed. This pattern is consistent with expected
611 behaviour when projecting previously unseen differentiation states onto a finite reference
612 atlas and does not indicate systematic misclassification.

613 To evaluate the robustness of methylation-based classification across different platforms, a
614 subset of the internal validation cohort (n=13/58) was analysed using both EPIC array and
615 long read nanopore sequencing (Supplementary Figure 6c). Classification of the nanopore
616 sequencing data was performed using the previously published nanoDx pipeline (Euskirchen
617 et al., 2017) with an ad-hoc random forest model trained on the reference set generated in
618 this study. The classifier demonstrated high reproducibility, correctly classifying 12/13 cases
619 consistently between the nanopore sequencing and methylation array platforms. This serves
620 as a proof of principle that SteMClass can be implemented using nanopore sequencing.

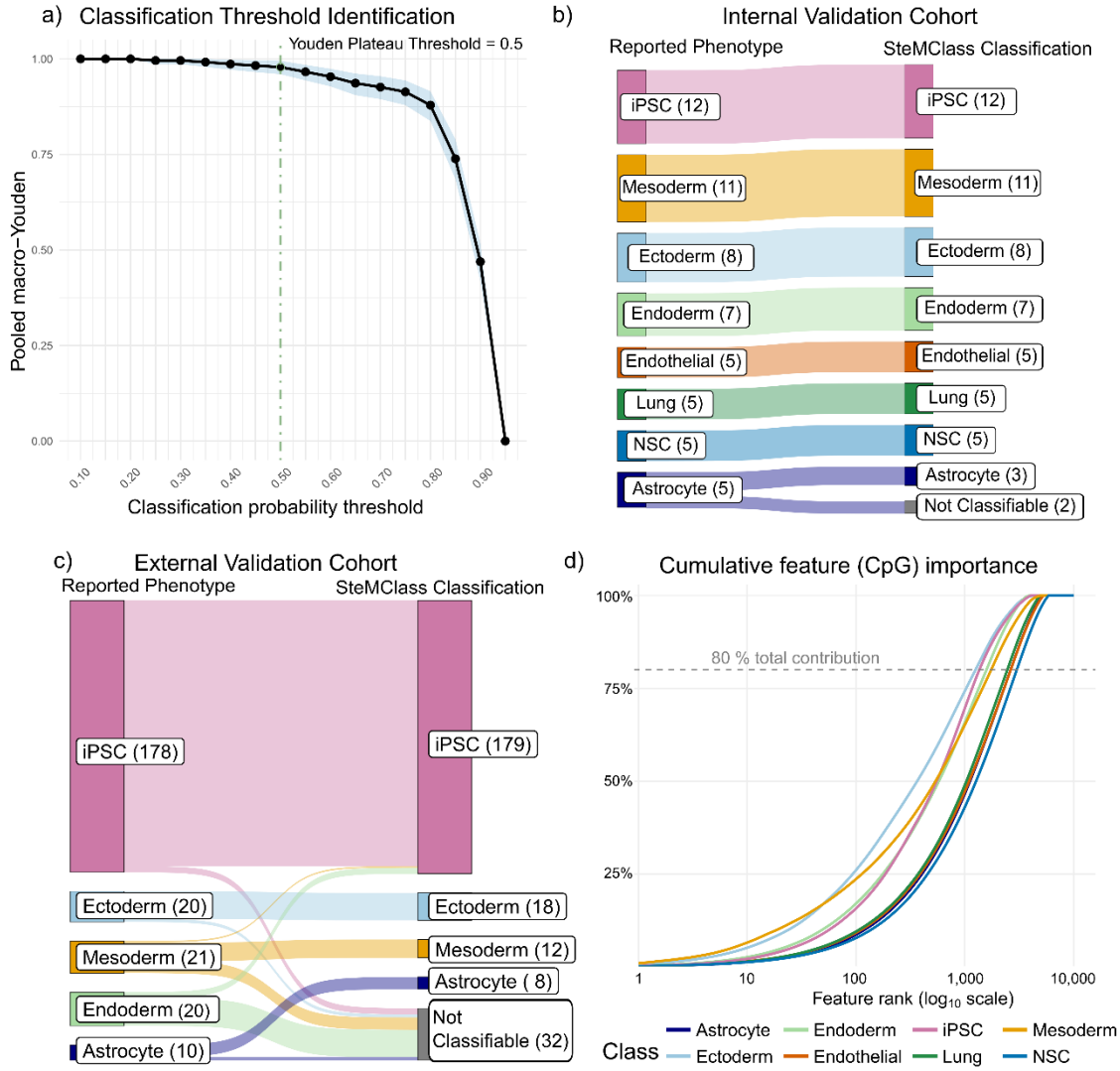


Figure 3. Integrated assessment of SteMClass: threshold tuning, independent validation, and feature-level signal distribution.

(a) Macro-Youden index as a function of classification probability threshold. Points denote empirical estimates computed from pooled cross-validated calibrated predictions and the ribbon shows the 95% percentile bootstrap confidence interval (1,000 resamples). Macro-Youden is defined as the mean one-vs-rest Youden's J across classes. The vertical dashed line indicates the largest threshold before a statistically significant improvement when lowering the cut-off. Significance was assessed using paired bootstrap differences between adjacent thresholds (one-sided test on ΔYouden).

(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. The classifier achieved 96.5% accuracy, Cohen's κ = 0.959, and a 3% reject rate. The internal cohort comprised: Astrocytes (n = 5), Ectodermal cells (n = 8), Endodermal cells (n = 7), Mesodermal cells (n = 11), Endothelial cells (n = 5), iPSCs (n = 12), Lung cells (n = 5), and NSCs (n = 5).

635 **(c)** Sankey diagram showing SteMClass predictions versus reported phenotype for the external
636 validation cohort. The external cohort comprises iPSCs (n = 178), iPSC-derived Ectoderm (n = 20),
637 Endoderm (n = 20), Mesoderm (n = 21) and Astrocytes (n=10) from multiple studies (Supplementary
638 Table 5).

639 **(d)** Cumulative absolute feature contribution per class. For each of the eight cell-state classes (colour-
640 coded), lines show how the summed absolute contribution to classification accumulates as CpG sites
641 are added in descending order of their mean absolute contribution across all 307 test samples (internal
642 + external). The x-axis is on a $\log_{10}$ scale; the dashed horizontal line marks 80% of the total
643 contribution. All classes reach this 80% level only after on the order of 10^3 CpGs, indicating that
644 SteMClass derives its decisions from a broad, distributed methylation signature rather than a small
645 handful of loci.

646 Integrative Broad and CpG-Level Profiling Explains Classification Outcomes.

647 To further interpret the classification results we looked beyond the scope of the classifier and
648 compared the broader DNA methylation profiles of the validation samples to the reference
649 cohort by investigating the top 20,000 CpGs that distinguish each differentiation state from
650 iPSC (see Methods; Supplementary Table 4). This allows a perspective beyond the 10,000 total
651 features used by SteMClass to investigate re-classified and "Not Classifiable" samples. Four of
652 the external validation cohort endoderm samples were re-classified as iPSC by SteMClass.
653 Hierarchical clustering of these samples with the reference cohort iPSC and endodermal cells
654 showed clustering with the iPSC samples based on their broader methylation profile (Figure
655 4a). Investigation of the top contributing variables to the classifier's decision in favour of iPSC
656 and against endoderm included the canonical pluripotency marker *POU5F1* and endodermal
657 marker *APC* (Figure 4b). Closer investigation of the *POU5F1* promoter region confirmed
658 methylation levels more consistent with reference iPSC in these test samples (Figure 4c).
659 Taken together, these results support the likelihood of inefficient differentiation experiments
660 rather than misclassifications.
661 Four iPSC samples from the external cohort were not classifiable. These four samples formed
662 a cluster distinct from the SteMClass reference iPSCs based on their 20,000 CpG broader
663 methylation profile (Figure 4d). To identify potential markers of iPSC undifferentiated state,
664 we interrogated the top 100 negatively contributing variables for iPSC common to the four
665 rejected samples (Figure 4e). Top hits included *PALLD* and *FGFR2* which were
666 hypermethylated in promoter CpGs in the four unclassifiable samples. Three of the samples
667 are more similar to reference iPSCs on the methylation profile, suggesting subtle cumulative
668 differences. One sample exhibited an especially divergent methylation profile, most notably
669 at the *POU5F1* gene promoter, suggesting a questionable undifferentiated state and
670 therefore compromised iPSC quality rather than a classification error for this sample (Figure
671 4f).

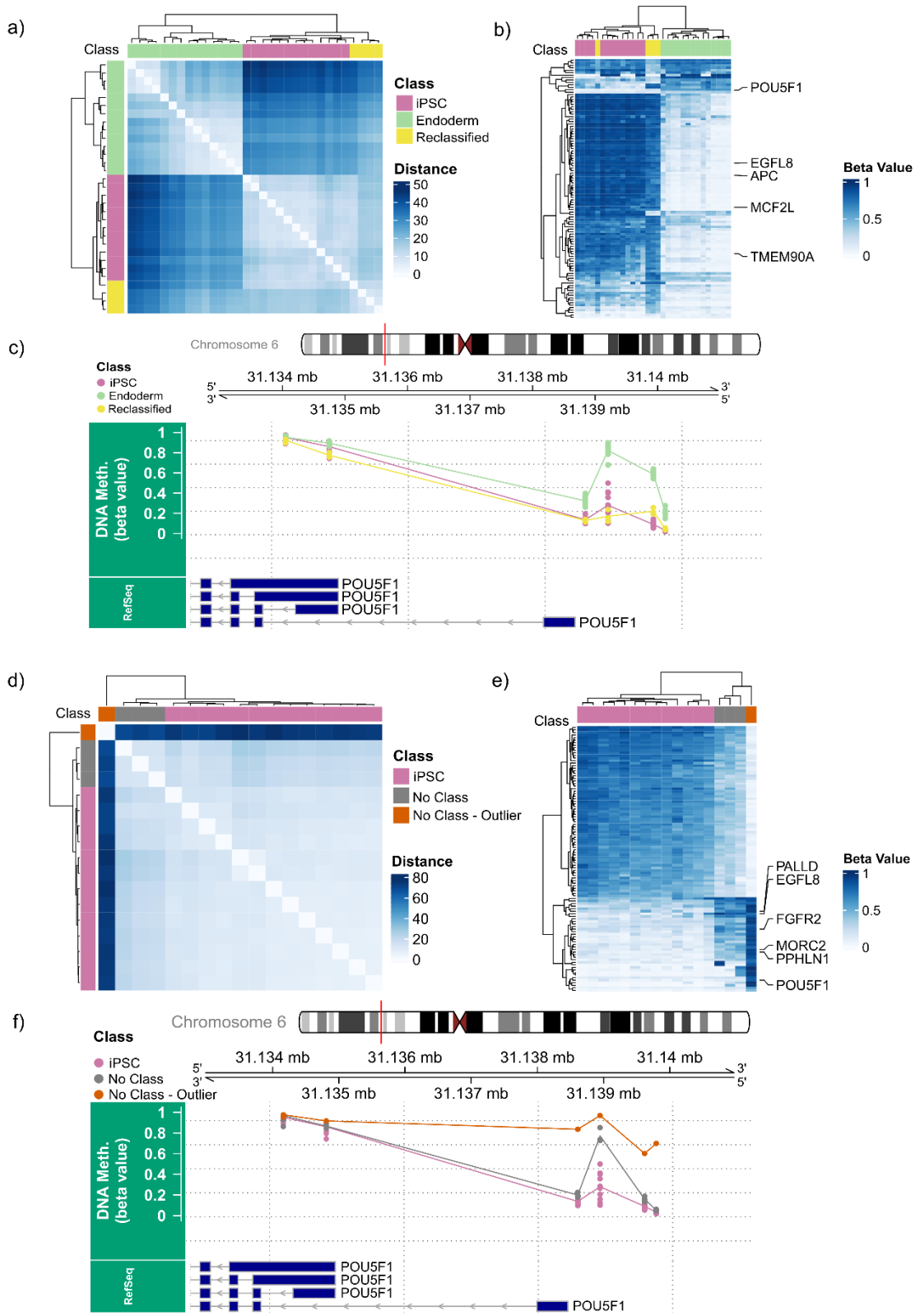


Figure 4. Interpreting SteMClass decisions via broad and narrow view methylation profiling

(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

677 top 20,000 CpGs most discriminating iPSC from all other cell types ("broad view"; Supplementary Table 4). The
678 samples reclassified by SteMClass cluster with reference iPSC. **(b)** Heatmap of the 100 CpGs with highest positive
679 importance for distinguishing iPSC vs. endoderm in the random forest model. Rows represent CpGs and columns
680 represent samples (reference and re-classified endoderm); both were hierarchically clustered using complete
681 linkage with Euclidean distance. Heatmap colours indicate beta value (white = low, blue = high). CpGs associated
682 with canonical markers such as *POU5F1* (pluripotency) and *APC* (endoderm) are labelled. Also, in this analysis
683 the reclassified samples cluster with reference iPSC. **(c)** Genomic view of beta values across the *POU5F1*
684 promoter region (chr6:31,134,000-31,140,000) for reference iPSC, reference endoderm, and the four re-
685 classified endoderm samples ("narrow view"). Tracks show methylation beta values for each sample (points),
686 confirming an iPSC-like profile in re-classified samples for this important pluripotency marker. **(d)** Hierarchical
687 clustering of four "Not Classifiable" test samples published as iPSC (grey and orange) with SteMClass reference
688 iPSC profiles (pink), using the same 20,000 CpG set most discriminating iPSC from all other cell types and
689 clustering parameters as in (a). Of the four rejected samples, three form a distinct clustering branch (grey),
690 indicating atypical iPSC methylation patterns, with one sample being a clear outlier from all other samples
691 (orange). **(e)** Heatmap of the 100 CpGs whose contributions to the iPSC class were most strongly negative across
692 all four not classifiable samples. (f) Genomic view of beta values at the POU5F1 promoter for the four non-
693 classifiable samples. The outlier sample (orange) displays markedly higher methylation, which does not conform
694 to the expected undifferentiated iPSC profile at this locus, while the three remaining samples (grey) remain
695 broadly consistent with iPSC status.

696 Seven of the eight non-classifiable/reclassified mesoderm samples (all from GSE85828)
697 clustered closer to iPSC and endoderm references than to mesoderm based on the top 20,000
698 CpG DNA methylation profiles, whereas one sample appeared as a clear outlier across all
699 differentiation states (Supplementary Fig. 7a). Three samples grouped directly with
700 endoderm references, and the remaining four localised between mesoderm and endoderm,
701 but with distance metrics indicating greater similarity to iPSC and endoderm than to
702 mesoderm. The CpGs most strongly contributing to rejection from mesoderm classification
703 included those associated with mesoderm regulators *MSGN1* and *WNT5B*, arguing against
704 technical misclassification and instead suggesting incomplete or divergent mesoderm
705 specification. Additional discriminatory loci included *APC* and *EFCAB2*. These CpGs were
706 hypomethylated in the test samples and in endoderm relative to iPSC and mesoderm
707 references (Supplementary Fig. 7b). Collectively, these findings indicate that the GSE85828
708 samples exhibit an epigenetic profile distinct from reference mesoderm and more likely an
709 intermediate "mesendoderm" differentiation state, consistent with the limited epigenetic
710 resolution between mesoderm and endoderm in this dataset which has been described
711 previously (Schmidt et al., 2022). Among the two non-classifiable ectoderm samples, one was
712 a clear outlier, while the remaining one resembled reference ectoderms in their global profiles
713 (Supplementary Figure 7d). Sixteen of twenty external validation cohort endoderm-
714 annotated samples failed to reach the 0.5 classification probability threshold and hierarchical
715 clustering confirmed that all non-classifiable endoderm samples grouped preferentially with

716 iPSC profiles rather than with the highly efficient (>90% $CXCR4^+/SOX17^+$) definitive endoderm
717 reference set (Supplementary Fig. 7e). This pattern suggests that many endoderm samples in
718 the external validation cohort retain epigenetic features more consistent with pluripotency
719 or an early transitional state than with fully specified definitive endoderm. Inspection of the
720 100 CpGs contributing most strongly to rejection from endoderm classification revealed
721 divergence at loci associated with lineage specification, including *APC*, (Supplementary Fig.
722 7f). Collectively, these findings point to substantial heterogeneity in endoderm differentiation
723 outcomes across protocols, potentially reflecting differences in timing, efficiency, or lineage
724 resolution rather than purely technical misclassification. These findings underscore the need
725 for a unified, quantitative assay for cellular differentiation state characterisation.

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

727 The CpG sites contributing to the rejected classification of the not classifiable astrocyte
728 samples from the internal validation cohort (n=2) and external validation cohort (n=2)
729 included the previously identified ectoderm-characteristic *PAX6* gene body region, possibly
730 indicating inefficiency in neural commitment in these experiments (Supplementary Figure 7h,
731 Figure 5a). In a limited comparison restricted to these two failed experiments (replicating this
732 is not feasible) compared to our successful reference and test sample experiments (total
733 samples n=65), this observed decrease in *PAX6* gene body methylation was evident only at
734 the day 37 astrocyte differentiation stage (Figure 5b). Improving the consistency of astrocyte
735 populations could involve incorporating SteMClass DNA methylation-based screening at a
736 later intermediate timepoint, enabling the selection of wells that are most closely progressing
737 toward the astrocyte state. Although it is beyond the scope of this study, as the SteMClass
738 reference set grows beyond this proof-of-concept classifier, our harmonised framework will
739 enable more robust interrogation of such epigenetic determinants of differentiation
740 propensity. To further explore how SteMClass predictions relate to the dynamic DNA
741 methylation patterns during differentiation, we examined changes in predicted
742 differentiation states during the transition from iPSCs to astrocytes (the same 9 time point
743 experiment displayed in Figure 1), as determined by SteMClass' predictions and probability
744 scores. The predicted cell identities and their associated probability scores at multiple
745 differentiation stages demonstrate how SteMClass adjusts as cells move from a pluripotent
746 state toward more lineage-restricted fates (Figure 5c). The progressive shift in probabilities
747 reflects the model's sensitivity to intermediate, transient states and its capacity to track subtle
748 changes in lineage commitment.

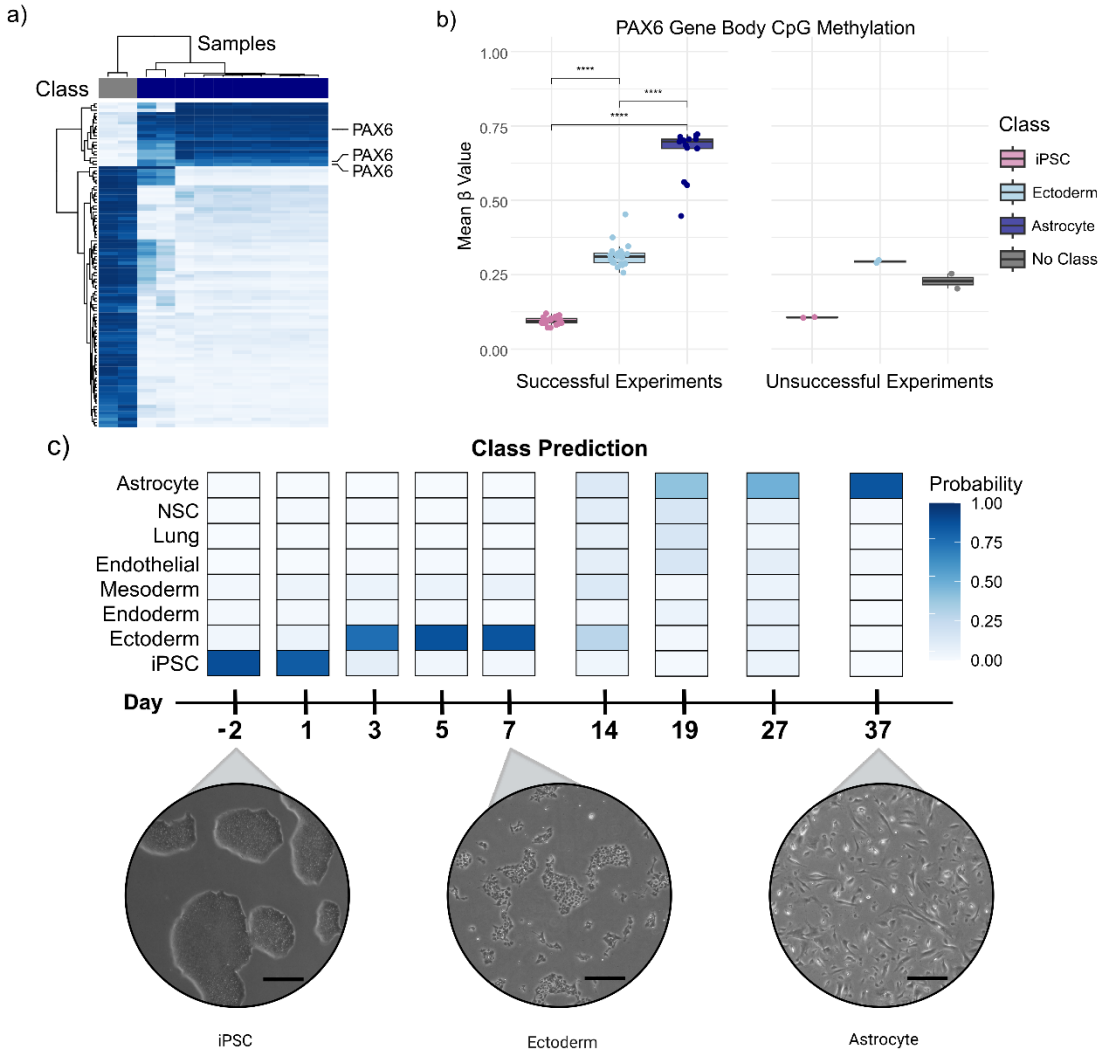


Figure 5. Epigenetic driver inference and dynamic tracking of astrocyte differentiation using SteMClass

(a) Heatmap of the 100 CpG sites with the strongest methylation differences (delta beta) in the two not classifiable internal validation samples (grey columns), compared to reference astrocytes (purple columns). Rows represent CpGs and columns represent samples; both were hierarchically clustered using complete linkage with Euclidean distance. Heatmap colour indicates beta value (white = low, blue = high); three *PAX6* gene body-associated CpGs rank among the top features, possibly implicating inefficient neural commitment in these two failed differentiations. **(b)** Boxplots show sample-level mean beta-values (0-1) for the *PAX6* CpGs opposing astrocyte classification across iPSC, ectoderm and astrocyte samples. Colours denote SteMClass-assigned differentiation states: iPSC (pink), ectoderm (light blue), astrocyte (dark purple) and "Not classifiable" (grey). Samples are split into panels by differentiation outcome: successful (n = 59) and unsuccessful (n = 6), defined by passing or failing the SteMClass 0.5 cutoff. In successful experiments, *PAX6* methylation increases during the differentiation from iPSC to ectoderm to astrocyte, whereas in failed experiments it decreases at the astrocyte stage. Since these CpGs are among the top features

765 contributing to ectoderm and astrocyte classification in SteMClass, monitoring the differentiation
766 using SteMClass at further intermediate stages may help identify at-risk differentiations earlier. All
767 pairwise comparisons within the successful group were significant (****, $p < 0.0001$; Wilcoxon rank-
768 sum). No statistical testing was applied to the unsuccessful group due to limited sample size. **(c)**
769 Temporal SteMClass classification shifts during the same iPSC to astrocyte nine time point
770 differentiation experiment shown in Figure 1c. Top panel: heatmap of predicted class probabilities (0-
771 1) at each harvest point for each class represented in SteMClass. Bottom panel: representative culture
772 images at day -2 (classified as iPSC), day 7 (classified as ectoderm), and day 37 (classified as astrocyte),
773 illustrating the morphological progression in line with classification shifts. Scale bars = 200 µm.

774 Discussion

775 A fundamental obstacle in stem cell research and regenerative medicine is the lack of
776 standardised frameworks to assess the quality and maturation stage of differentiated cells,
777 causing variability in both results and interpretation. While tools like the TaqMan hPSC
778 Scorecard Assay (Thermofisher Scientific) and hiPSCore (based on long-read nanopore
779 sequencing of 12 early trilineage -linked genes) help gauge pluripotency and trilineage
780 potential (Dobner et al., 2024; Lenz, 2012; Tsankov et al., 2015), most downstream
781 applications demand differentiation into more specialised cell types that are not covered by
782 these assays. Currently, evaluations of these more advanced stages vary widely and often rely
783 on a handful of molecular markers, genes or proteins whose expression may fluctuate based
784 on protocol specifics, laboratory conditions, and genetic backgrounds.

785 Our classifier prototype, "SteMClass", demonstrates strong predictive performance, with a
786 Brier score of 0.018 as calculated during the cross-validation procedure. Validation with one
787 internally generated independent validation cohort and one independent validation cohort
788 compiled from publicly available data from published studies reinforces these findings,
789 demonstrating StemClass' robustness to variations in experimental conditions, as well as the
790 classifier's ability to identify potentially problematic samples.

791 While these are promising results, further harmonisation requires data from multiple
792 research centres to address protocol, reagent, characterisation and data format
793 inconsistencies across institutions. To preliminarily explore its performance across different
794 institutions, we incorporated DNA methylation array data for which the published
795 differentiation protocols closely matched our training cohort, allowing for an initial
796 assessment of classifier applicability in varied contexts. External samples annotated as
797 mesoderm and endoderm that failed classification showed global epigenetic deviation from
798 the SteMClass references, suggesting biological divergence rather than technical

799 misassignment. In the study GSE85828, most mesoderm samples clustered closer to iPSC and
800 endoderm than to reference mesoderm, and rejection was driven by CpGs at canonical
801 mesoderm regulators (*MSGN1*, *WNT5B*) and additional loci including *APC*. These findings are
802 consistent with incomplete or unresolved lineage specification and align with prior reports
803 that iPSC-, mesoderm-, and endoderm-labeled samples in this dataset are not epigenetically
804 distinguishable by global DNA methylation profiling (Daily et al., 2017; Schmidt et al., 2022).
805 Notably, mesoderm identity in that study was assessed by flow cytometry for *Brachyury* and
806 *OCT4* expression. Brachyury, however, is characteristic of only very early mesodermal
807 progenitors, unlike the mature mesodermal cells used to train SteMClass, which were
808 validated at >90% efficiency using CD144/CD140b. This difference in differentiation stage
809 likely underlies the unexpected classification, whereas mesoderm samples generated through
810 standardised trilineage differentiation protocols in other studies classified as expected.
811 A similar explanation likely underlies the rejection of external dataset endoderm-annotated
812 samples to classify as endoderm. All SteMClass reference endoderm samples were validated
813 by flow cytometry, confirming high differentiation efficiency (>90% $SOX17^+/CXCR4^+$). Previous
814 studies often relied on semiquantitative marker assessments, which have been shown to yield
815 less distinct DNA methylation profiles (Daily et al., 2017; Dobner et al., 2024; Elsafi Mabrouk
816 et al., 2024; Schmidt et al., 2022). For example, GSE253190 reported heterogeneous *GATA6*
817 immunofluorescence and *SOX17* RT-qPCR without assessing residual pluripotency, while
818 GSE207076 showed persistent *OCT4* staining in the representative immunofluorescence
819 images. The Activin-based differentiation in GSE85828 also produced mesoderm and
820 endoderm populations that were not epigenetically resolved from each other or from iPSC by
821 methylation profiling. Collectively, these observations indicate that samples annotated as
822 "endoderm" often represent heterogeneous or incompletely specified states. Expanding the
823 training cohort to include rigorously validated, stage-resolved endoderm samples across
824 developmental windows may improve classifier utility and enable discrimination of epigenetic
825 subtypes, potentially resolving early regional identities (e.g., anterior versus posterior
826 definitive endoderm) rather than collapsing them into a single reference group as is the
827 current standard. In addition to our observation of differentiation outcome heterogeneity
828 across published studies, it has also been shown that iPSC-derived cardiomyocytes frequently
829 show inconsistent gene expression profiles and cell compositions across different laboratories
830 and methods (Grancharova et al., 2021) underscoring the need for more standardised and
831 robust assessment tools.

832 SteMClass is adaptable to new technologies, already integrating data from the latest EPICv2.0
833 methylation arrays and can also be adapted to nanopore sequencing based on our proof-of-

concept experiment. By adapting to nanopore sequencing, SteMClass can leverage direct, real-time analysis of native DNA without the need for bisulfite conversion, streamlining sample preparation and significantly reducing associated costs. Nanopore sequencing requires relatively low initial investment compared to array-based platforms, making it more accessible for laboratories with limited budgets. While single-cell DNA methylation profiling remains technically complex and expensive, SteMClass' use of bulk methylation data on relatively pure cell populations provides a practical alternative for cell-type-specific classification. This combination of cost-effective setup and simplified workflows enhances the utility and scalability of SteMClass for a broader range of research environments.

Looking ahead, the SteMClass prototype offers a conceptual foundation from which several future directions might be explored. Among the most promising possibilities would be an extension toward predicting differentiation potential, which would support the screening of iPSC lines for lineage-specific applications relevant to both regenerative medicine and disease modelling. The classifier's design might also inspire quality control frameworks for clinical-grade stem cell production. In organoid and co-culture systems, the classifier could track cell-state composition on sorted models, supporting studies on tissue modelling and cell interactions. Together, these applications position the classifier as a valuable starting point for advancing stem cell research, quality control in regenerative medicine, and precision medicine applications. In addition to the classification tool itself, the dataset is currently the most extensive DNA methylation data resource of iPSC lines and their isogenic differentiated derivatives across multiple stages of ectoderm, endoderm, and mesoderm, maintained and characterised under standardised conditions. This methylation dataset is a valuable research resource and can provide a wealth of information on the dynamics of DNA methylation during in vitro differentiation considering that cells derived from iPSCs in vitro differ inherently from their fully mature in vivo counterparts. Beyond classification, the DNA methylation data generated by both array-based and nanopore long-read sequencing methods can be used for additional quality control measures in iPSC maintenance and differentiation experiments such as CNV detection, biological sex prediction, and genotype verification without the need for additional testing (Capper et al., 2018; Daenekas et al., 2024; Kuschel et al., 2022).

Conclusion

SteMClass (https://pereze5-stemclass.share.connect.posit.cloud/) is a publicly available proof-of-concept DNA methylation-based classifier prototype that rapidly and reproducibly verifies iPSC and iPSC-derived differentiation states. It identifies and characterises distinct stages of iPSC differentiation, providing a standardised framework for classification that

enables consistent cross-laboratory comparisons. By supporting the development of harmonised reference datasets and interpretive criteria, SteMClass will enable standardisation in stem cell research and accelerates translation toward clinical applications.

References

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

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

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

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

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

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

Buckberry, S., Liu, X., Poppe, D., Tan, J. P., Sun, G., Chen, J., Nguyen, T. V., de Mendoza, A., Pflueger, J., Frazer, T., Vargas-Landín, D. B., Paynter, J. M., Smits, N., Liu, N., Ouyang, J. F., Rossello, F. J., Chy, H. S., Rackham, O. J. L., Laslett, A. L., … Lister, R. (2023).

900 Transient naive reprogramming corrects hiPS cells functionally and epigenetically.
901 *Nature*, *620*(7975), Article 7975. https://doi.org/10.1038/s41586-023-06424-7

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

906 Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm,
907 F., Chavez, L., Reuss, D. E., Kratz, A., Wefers, A. K., Huang, K., Pajtler, K. W., Schweizer,
908 L., Stichel, D., Olar, A., Engel, N. W., Lindenberg, K., … Pfister, S. M. (2018). DNA
909 methylation-based classification of central nervous system tumours. *Nature*,
910 *555*(7697), 469–474. https://doi.org/10.1038/nature26000

911 Cerneckis, J., Cai, H., & Shi, Y. (2024). Induced pluripotent stem cells (iPSCs): Molecular
912 mechanisms of induction and applications. *Signal Transduction and Targeted Therapy*,
913 *9*(1), 1–26. https://doi.org/10.1038/s41392-024-01809-0

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

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

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

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

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

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

940 Daily, K., Ho Sui, S. J., Schriml, L. M., Dexheimer, P. J., Salomonis, N., Schroll, R., Bush, S.,
941 Keddache, M., Mayhew, C., Lotia, S., Perumal, T. M., Dang, K., Pantano, L., Pico, A. R.,
942 Grassman, E., Nordling, D., Hide, W., Hatzopoulos, A. K., Malik, P., … Omberg, L. (2017).
943 Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell
944 lines and derivatives. *Scientific Data*, *4*(1), 170030.
945 https://doi.org/10.1038/sdata.2017.30

946 Danielsson, A., Nemes, S., Tisell, M., Lannering, B., Nordborg, C., Sabel, M., & Carén, H. (2015).
947 MethPed: A DNA methylation classifier tool for the identification of pediatric brain
948 tumor subtypes. *Clinical Epigenetics*, *7*(1), 62. https://doi.org/10.1186/s13148-015-
949 0103-3

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

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

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

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

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

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

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

980 Franzen, J., Georgomanolis, T., Selich, A., Kuo, C.-C., Stöger, R., Brant, L., Mulabdić, M. S.,
981 Fernandez-Rebollo, E., Grezella, C., Ostrowska, A., Begemann, M., Nikolić, M., Rath, B.,
982 Ho, A. D., Rothe, M., Schambach, A., Papantonis, A., & Wagner, W. (2021). DNA
983 methylation changes during long-term in vitro cell culture are caused by epigenetic
984 drift. *Communications Biology*, *4*(1), 1–12. https://doi.org/10.1038/s42003-021-
985 02116-y

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

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

994 Hajmousa, G., de Almeida, R. C., Bloks, N., Ruiz, A. R., Bouma, M., Slieker, R., Kuipers, T. B.,
995 Nelissen, R. G. H. H., Ito, K., Freund, C., Ramos, Y. F. M., & Meulenbelt, I. (2024). The
996 role of DNA methylation in chondrogenesis of human iPSCs as a stable marker of
997 cartilage quality. *Clinical Epigenetics*, *16*(1), 141. https://doi.org/10.1186/s13148-024-
998 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

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

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

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

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

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

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

1032 Kitayama, S., Zhang, R., Liu, T.-Y., Ueda, N., Iriguchi, S., Yasui, Y., Kawai, Y., Tatsumi, M., Hirai,
1033 N., Mizoro, Y., Iwama, T., Watanabe, A., Nakanishi, M., Kuzushima, K., Uemura, Y., &
1034 Kaneko, S. (2016). Cellular Adjuvant Properties, Direct Cytotoxicity of Re-
1035 differentiated Vα24 Invariant NKT-like Cells from Human Induced Pluripotent Stem
1036 Cells. *Stem Cell Reports*, *6*(2), 213–227. https://doi.org/10.1016/j.stemcr.2016.01.005

1037 Kramer, N., Rosner, M., Kovacic, B., & Hengstschläger, M. (2016). Full biological
1038 characterization of human pluripotent stem cells will open the door to translational
1039 research. *Archives of Toxicology*, *90*(9), 2173–2186. https://doi.org/10.1007/s00204-
1040 016-1763-2

1041 Kügelgen, N. von, Vallone, V. F., Ludwik, K. A., Babaeva, M., Diecke, S., Beule, D., &
1042 Stachelscheid, H. (2025). StemCNV-check: A pipeline for human pluripotent stem cell
1043 (hPSC) genomic integrity control using SNP array data and copy number variant scoring
1044 (p. 2025.12.01.691615). bioRxiv. https://doi.org/10.64898/2025.12.01.691615

1045 Kurdyukov, S., & Bullock, M. (2016). DNA Methylation Analysis: Choosing the Right Method.
1046 *Biology*, *5*(1), 3. https://doi.org/10.3390/biology5010003

1047 Kuschel, L. P., Hench, J., Frank, S., Hench, I. B., Girard, E., Blanluet, M., Masliah-Planchon, J.,
1048 Misch, M., Onken, J., Czabanka, M., Yuan, D., Lukassen, S., Karau, P., Ishaque, N., Hain,
1049 E. G., Heppner, F., Idbaih, A., Behr, N., Harms, C., … Euskirchen, P. (2022). Robust
1050 methylation-based classification of brain tumours using nanopore sequencing.
1051 *Neuropathology and Applied Neurobiology*. https://doi.org/10.1111/nan.12856

1052 Lanza, R., Russell, D. W., & Nagy, A. (2019). Engineering universal cells that evade immune
1053 detection. *Nature Reviews. Immunology*, *19*(12), 723–733.
1054 https://doi.org/10.1038/s41577-019-0200-1

1055 Lenz, M. (2012). PluriTest: Molecular diagnostic assay for pluripotency in human stem cells.
1056 In *Human Stem Cell Manual* (pp. 293–311). Elsevier. https://doi.org/10.1016/b978-0-
1057 12-385473-5.00018-7

1058 Liuyang, S., Wang, G., Wang, Y., He, H., Lyu, Y., Cheng, L., Yang, Z., Guan, J., Fu, Y., Zhu, J.,
1059 Zhong, X., Sun, S., Li, C., Wang, J., & Deng, H. (2023). Highly efficient and rapid
1060 generation of human pluripotent stem cells by chemical reprogramming. *Cell Stem*
1061 *Cell*, *30*(4), 450-459.e9. https://doi.org/10.1016/j.stem.2023.02.008

1062 Loyfer, N., Magenheim, J., Peretz, A., Cann, G., Bredno, J., Klochendler, A., Fox-Fisher, I., Shabi-
1063 Porat, S., Hecht, M., Pelet, T., Moss, J., Drawshy, Z., Amini, H., Moradi, P., Nagaraju, S.,
1064 Bauman, D., Shveiky, D., Porat, S., Dior, U., … Kaplan, T. (2023). A DNA methylation

1065 atlas of normal human cell types. *Nature*, *613*(7943), 355–364.
1066 https://doi.org/10.1038/s41586-022-05580-6

1067 Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C.,
1068 Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W.
1069 (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: A
1070 summary. *Neuro-Oncology*, *23*(8), 1231–1249.
1071 https://doi.org/10.1093/neuonc/noab106

1072 Ludwig, T. E., Andrews, P. W., Barbaric, I., Benvenisty, N., Bhattacharyya, A., Crook, J. M.,
1073 Daheron, L. M., Draper, J. S., Healy, L. E., Huch, M., Inamdar, M. S., Jensen, K. B., Kurtz,
1074 A., Lancaster, M. A., Liberali, P., Lutolf, M. P., Mummery, C. L., Pera, M. F., Sato, Y., …
1075 Mosher, J. T. (2023). ISSCR standards for the use of human stem cells in basic research.
1076 *Stem Cell Reports*, *18*(9), 1744–1752. https://doi.org/10.1016/j.stemcr.2023.08.003

1077 MacArthur, C. C., Pradhan ,Suman, Wetton ,Nichole, Zarrabi ,Aryan, Dargitz ,Carl, Sridharan
1078 ,Mahalakshmi, Jackson ,Stephen, Pickle ,Loni, & and Lakshmipathy, U. (2019).
1079 Generation and Comprehensive Characterization of Induced Pluripotent Stem Cells for
1080 Translational Research. *Regenerative Medicine*, *14*(6), 505–524.
1081 https://doi.org/10.2217/rme-2018-0148

1082 Maksimovic, J., Phipson, B., & Oshlack, A. (2017). A cross-package Bioconductor workflow for
1083 analysing methylation array data. *F1000Research*, *5*, 1281.
1084 https://doi.org/10.12688/f1000research.8839.3

1085 Moss, J., Magenheim, J., Neiman, D., Zemmour, H., Loyfer, N., Korach, A., Samet, Y., Maoz,
1086 M., Druid, H., Arner, P., Fu, K.-Y., Kiss, E., Spalding, K. L., Landesberg, G., Zick, A.,
1087 Grinshpun, A., Shapiro, A. M. J., Grompe, M., Wittenberg, A. D., … Dor, Y. (2018).
1088 Comprehensive human cell-type methylation atlas reveals origins of circulating cell-
1089 free DNA in health and disease. *Nature Communications*, *9*(1), 5068.
1090 https://doi.org/10.1038/s41467-018-07466-6

1091 Nishino, K., Takasawa, K., Okamura, K., Arai, Y., Sekiya, A., Akutsu, H., & Umezawa, A. (2021).
1092 Identification of an epigenetic signature in human induced pluripotent stem cells using
1093 a linear machine learning model. *Human Cell*, *34*(1), 99–110.
1094 https://doi.org/10.1007/s13577-020-00446-3

1095 Nishizawa, M., Chonabayashi, K., Nomura, M., Tanaka, A., Nakamura, M., Inagaki, A.,
1096 Nishikawa, M., Takei, I., Oishi, A., Tanabe, K., Ohnuki, M., Yokota, H., Koyanagi-Aoi, M.,
1097 Okita, K., Watanabe, A., Takaori-Kondo, A., Yamanaka, S., & Yoshida, Y. (2016).

1098 Epigenetic Variation between Human Induced Pluripotent Stem Cell Lines Is an
1099 Indicator of Differentiation Capacity. *Cell Stem Cell*, *19*(3), 341–354.
1100 https://doi.org/10.1016/j.stem.2016.06.019

1101 Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer,
1102 L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M.,
1103 Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., …
1104 McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance
1105 and exemplars for reporting systematic reviews. *The BMJ*, *372*, n160.
1106 https://doi.org/10.1136/bmj.n160

1107 Parry, A., Rulands, S., & Reik, W. (2021). Active turnover of DNA methylation during cell fate
1108 decisions. *Nature Reviews Genetics*, *22*(1), 59–66. https://doi.org/10.1038/s41576-
1109 020-00287-8

1110 Patsch, C., Challet-Meylan, L., Thoma, E. C., Urich, E., Heckel, T., O'Sullivan, J. F., Grainger, S.
1111 J., Kapp, F. G., Sun, L., Christensen, K., Xia, Y., Florido, M. H. C., He, W., Pan, W.,
1112 Prummer, M., Warren, C. R., Jakob-Roetne, R., Certa, U., Jagasia, R., … Cowan, C. A.
1113 (2015). Generation of vascular endothelial and smooth muscle cells from human
1114 pluripotent stem cells. *Nature Cell Biology*, *17*(8), 994–1003.
1115 https://doi.org/10.1038/ncb3205

1116 Phipson, B., Maksimovic, J., & Oshlack, A. (2016). missMethyl: An R package for analyzing data
1117 from Illumina's HumanMethylation450 platform. *Bioinformatics (Oxford, England)*,
1118 *32*(2), 286–288. https://doi.org/10.1093/bioinformatics/btv560

1119 Puri, D., Maaßen, C., Varona Baranda, M., Zeevaert, K., Hahnfeld, L., Hauser, A., Fornero, G.,
1120 Elsafi Mabrouk, M. H., & Wagner, W. (2023). CTCF deletion alters the pluripotency and
1121 DNA methylation profile of human iPSCs. *Frontiers in Cell and Developmental Biology*,
1122 *11*. https://doi.org/10.3389/fcell.2023.1302448

1123 Schmidt, M., Zeevaert, K., Elsafi Mabrouk, M. H., Goetzke, R., & Wagner, W. (2022). Epigenetic
1124 biomarkers to track differentiation of pluripotent stem cells. *Stem Cell Reports*.
1125 https://doi.org/10.1016/j.stemcr.2022.11.001

1126 Shi, Y., Inoue, H., Wu, J. C., & Yamanaka, S. (2017). Induced pluripotent stem cell technology:
1127 A decade of progress. *Nature Reviews Drug Discovery*, *16*(2), 115–130.
1128 https://doi.org/10.1038/nrd.2016.245

1129 Sturchler, E., Cox, J. A., Durussel, I., Weibel, M., & Heizmann, C. W. (2006). S100A16, a novel
1130 calcium-binding protein of the EF-hand superfamily. *The Journal of Biological*
1131 *Chemistry*, *281*(50), 38905–38917. https://doi.org/10.1074/jbc.M605798200

1132 Sultanov, R., Lebedeva, O., Arapidi, G., Lagarkova, M., & Kiselev, S. (2018). Methylation profile
1133 of induced pluripotent stem cells generated by integration and integration-free
1134 approaches. *Data in Brief*, *17*, 662–666. https://doi.org/10.1016/j.dib.2018.01.061

1135 Tanaka, Y., Furuhata, E., Maeda, S., Kishima, M., Suzuki, H., & Suzuki, T. (2023). A dataset of
1136 definitive endoderm and hepatocyte differentiations from human induced pluripotent
1137 stem cells. *Scientific Data*, *10*(1), 93. https://doi.org/10.1038/s41597-023-02001-9

1138 TCW, J., Wang, M., Pimenova, A. A., Bowles, K. R., Hartley, B. J., Lacin, E., Machlovi, S. I.,
1139 Abdelaal, R., Karch, C. M., Phatnani, H., Slesinger, P. A., Zhang, B., Goate, A. M., &
1140 Brennand, K. J. (2017). An Efficient Platform for Astrocyte Differentiation from Human
1141 Induced Pluripotent Stem Cells. *Stem Cell Reports*, *9*(2), 600–614.
1142 https://doi.org/10.1016/j.stemcr.2017.06.018

1143 Tesarova, L., Simara, P., Stejskal, S., & Koutna, I. (2016). The Aberrant DNA Methylation Profile
1144 of Human Induced Pluripotent Stem Cells Is Connected to the Reprogramming Process
1145 and Is Normalised During In Vitro Culture. *PLOS ONE*, *11*(6), e0157974.
1146 https://doi.org/10.1371/journal.pone.0157974

1147 Tompkins, J. D. (2022). Discovering DNA Methylation, the History and Future of the Writing
1148 on DNA. *Journal of the History of Biology*, *55*(4), 865–887.
1149 https://doi.org/10.1007/s10739-022-09691-8

1150 Tran, Q. T., Breuer, A., Lin, T., Tatevossian, R., Allen, S. J., Clay, M., Furtado, L. V., Chen, M.,
1151 Hedges, D., Michael, T., Robinson, G., Northcott, P., Gajjar, A., Azzato, E., Shurtleff, S.,
1152 Ellison, D. W., Pounds, S., & Orr, B. A. (2024). Comparison of DNA methylation based
1153 classification models for precision diagnostics of central nervous system tumors. *Npj*
1154 *Precision Oncology*, *8*(1), 1–12. https://doi.org/10.1038/s41698-024-00718-3

1155 Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., & Siegmund, K. D. (2013). Low-
1156 level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids*
1157 *Research*, *41*(7), e90. https://doi.org/10.1093/nar/gkt090

1158 Tsankov, A. M., Akopian, V., Pop, R., Chetty, S., Gifford, C. A., Daheron, L., Tsankova, N. M., &
1159 Meissner, A. (2015). A qPCR ScoreCard quantifies the differentiation potential of
1160 human pluripotent stem cells. *Nature Biotechnology*, *33*(11), Article 11.
1161 https://doi.org/10.1038/nbt.3387

1162 Vallone, V. F., Ludwik, K., Healy, L., Lefort, N., Onder, T., Pavinato, L., Okur, F. V. F., &
1163 Stachelscheid, H. (2025). Good practices in hPSC handling.
1164 https://www.protocols.io/view/good-practices-in-hpsc-handling-gdqmbs5u7
1165 Wang, S., Du, Y., Zhang, B., Meng, G., Liu, Z., Liew, S. Y., Liang, R., Zhang, Z., Cai, X., Wu, S.,
1166 Gao, W., Zhuang, D., Zou, J., Huang, H., Wang, M., Wang, X., Wang, X., Liang, T., Liu,
1167 T., … Shen, Z. (2024). Transplantation of chemically induced pluripotent stem-cell-
1168 derived islets under abdominal anterior rectus sheath in a type 1 diabetes patient.
1169 Cell, S0092867424010225. https://doi.org/10.1016/j.cell.2024.09.004
1170 Yang, H., Yang, Y., Kiskin, F. N., Shen, M., & Zhang, J. Z. (2023). Recent advances in regulating
1171 the proliferation or maturation of human-induced pluripotent stem cell-derived
1172 cardiomyocytes. *Stem Cell Research & Therapy*, *14*(1), 228.
1173 https://doi.org/10.1186/s13287-023-03470-w

1174 Yoshihara, M., Hayashizaki, Y., & Murakawa, Y. (2017). Genomic Instability of iPSCs:
1175 Challenges Towards Their Clinical Applications. *Stem Cell Reviews*, *13*(1), 7–16.
1176 https://doi.org/10.1007/s12015-016-9680-6

1177

1178 **Acknowledgements**

1179 This work was supported by The German Academic Exchange Service (DAAD) [E.P.] and the
1180 Berlin School of Integrative Oncology [E.P.]. The work was further supported by the German
1181 Cancer Consortium (DKTK), Partner Site Berlin [D.C.]. The authors thank Christine Sers,
1182 Gaetano Gargiulio, Frank Heppner and all members of the Heppner lab for helpful discussions
1183 and suggestions for this work.

1184 **Conflicting Interests**

1185 The authors have no conflicting interests to declare.

1186 **Author contributions**

1187 E.P. performed differentiation experiments, generated DNA methylation data, conducted
1188 bioinformatic analyses, prepared figures and tables and wrote the manuscript. V.F.-V.
1189 performed differentiation experiments and prepared figures and tables. D.T. and P.W.
1190 performed DNA extraction and EPIC array processing. C.S. and B.O. contributed to the data
1191 processing workflow. P.E. supported nanopore sequencing-based methylation profiling.
1192 D.C., V.F.-V. and H.S. conceived and supervised the study. All authors contributed to
1193 manuscript editing and approved the final version.
1194

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 GSE308134.