

Annals of Stem Cell Research

Review Article

Machine Learning-Assisted Morphological Assessment of Stem Cell Colony Quality

Teixeira M^{1*}, Boer BD² and Kazlauskas M³

¹Department of Clinical Immunology, University of Coimbra, Coimbra, Portugal

²Department of Cell and Tissue Engineering, Radboud University, Nijmegen, Netherlands

³Department of Tissue Banking and Cell Therapy, Lithuanian University of Health Sciences, Kaunas, Lithuania

*Address for Correspondence: Mariana Teixeira, Department of Clinical Immunology, University of Coimbra, Coimbra, Portugal; E-mail: mariana.teixeira@universityofcoimbra.edu

Received: 1 March 2026; Accepted: 23 May 2026; Published: 4 September 2026; DOI: 10.5281/zenodo.22274906

Citation of this article: Teixeira M, Boer BD, Kazlauskas M (2026) Machine Learning-Assisted Morphological Assessment of Stem Cell Colony Quality. Ann Stem Cell Res, 9(1): 01-08.

Copyright: © 2026 Mariana Teixeira, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Automated morphological assessment of stem cell colonies is widely reported to exceed ninety per cent accuracy, but published figures rest on incompatible validation designs.

Objective: To read the field through the validation regime attached to each accuracy figure and ask what the reported numbers actually measure.

Methods: Narrative review of colony imaging studies, model benchmarks and inter-observer agreement data.

Findings: On 165,190 images across 32 cell lines and 93 batches, one model reached 99.9% accuracy under a batch-stratified split and 89.1% under a batch-separated split. A model trained on single-cell bright-field images reached 0.923 accuracy on the held-out 20% but 71.89-82.73% on separate experiments using inducers absent from training. Two experts agreed on 48% of 449 colony images.

Conclusions: The evidence supports reporting split construction with every accuracy and grouping splits by batch and donor; it does not support ranking published accuracies against one another.

Keywords: Induced pluripotent stem cells; Machine learning; Colony morphology; Image analysis; Model validation

Introduction

Pluripotent stem cell culture is still governed by looking. A technician examines a plate under phase contrast, decides which colonies are compact and phase-bright with defined borders and which have gone flat, loose or centrally dense, and marks the poor ones for removal. The judgement is fast, free, non-destructive and repeatable within one pair of hands. It is also undocumented, unquantified and unauditable, which is why it has become the obvious target for automation. Two decades of work now exist on extracting morphological descriptors from label-free images of stem cell colonies and mapping them onto quality labels, first with hand-crafted features and classical classifiers and more recently with convolutional networks and self-supervised representations [1,2,3].

The literature that has accumulated reports accuracies that look excellent. Headline figures above ninety per cent are routine, and a reader assembling them into a table would conclude that the problem is solved. It is not, and the reason is that the numbers are not measuring the same thing. An accuracy is a property of a model and a test set jointly, and the test sets in this field are constructed in at least three incompatible ways: by cross-validation within the training data, by holding out a split of the same experiment, and by evaluating on material the model has genuinely never encountered — a different batch, a different induction condition, a different donor, a different microscope. Papers rarely make the distinction load-bearing, and the resulting figures are stacked side by side in review tables as if they were commensurable. They are not. This review reads the field through that single lens: for every performance number quoted below, the validation regime is stated, and the gap between regimes is treated as the primary finding rather than as a caveat.

Grading colonies by eye

The case for automation rests on an assumption that manual grading is reproducible enough to be worth

reproducing. That assumption is weakly supported. The most direct measurement available compares two stem cell experts scoring 449 hESC (H9) colony images from three composite fields covering roughly 12 cm² on a five-point scale from poor to maximum pluripotency character; they agreed on 48% of colonies [4]. This figure must be reported with its provenance: it comes from a National Institute of Standards and Technology Information Systems Group dataset page, not from a peer-reviewed article, and the page itself carries no companion publication. It is the best number available for this task, and no journal article located in this survey measures inter-operator agreement on human pluripotent stem cell colony grading at all. That absence is itself a finding.

Adjacent imaging fields have done the measurement properly, and their results indicate that visual agreement is strongly feature-dependent rather than uniformly poor. In early embryo assessment, inter-observer median kappa reaches 0.80 or above for blastomere counting on days two and three, falls to 0.60-0.79 for cleavage-axis orientation, sits at 0.40-0.59 for most other parameters, and drops below 0.40 for pronuclear size and cytoplasmic halo, across 50 embryos scored by seven embryologists [5]. Discrete countable features are agreed upon; graded textural impressions are not. Colony quality is a graded textural impression.

A second embryo study makes the more uncomfortable point. On 245 images from 31 embryos, three embryologists reached 89.9% accuracy with a Fleiss kappa of 0.878 (95% CI 0.848-0.908), while ResNet-34 reached 78.8% accuracy with agreement kappa 0.932 and VGG16 reached 74.3% with kappa 0.940 [6]. Those accuracies come from a held-out test set within a single centre, not from an external cohort. The humans were more accurate and less consistent; the networks were less accurate and more consistent. Substituting a model for an expert therefore trades one kind of error for another, and a study

that reports only accuracy has measured only half of the exchange. The claim that a multi-class U-Net segmenting colony, single-cell, differentiated and dead regions is more accurate than skilled visual estimation of hiPSC culture status is made without an accompanying agreement statistic, which leaves the comparison unresolved [7].

Two decades of morphological models

The earliest systems were explicitly parametric. Automated scoring of hiPSC cultures from time-lapse bright-field imaging, using compaction, doubling time, colony-border spikiness and sensitivity to media change, reached 80% accuracy per imaging position and 89% per well against three-category expert ground truth, on a 50/50 split of 564 time-lapse datasets from 94 cultures [8]. Texture and pattern descriptors from the wndchrm framework, cross-validated on 60 colony images per line across nine lines, separated an iPSC line from a non-iPSC line at 0.87 and from mammary epithelial cells at 0.96 [9]. Parametric analysis of non-labelled live colonies took a different route, clustering 303 mature colonies drawn from 2,304 images into five morphological groups and validating those groups post hoc against OCT-3/4 and vimentin staining and single-colony microarray, where 1,454 probes differed significantly between the poor-quality cluster and the rest; poor subclones were enriched in that cluster at 27.3% versus 13.3% for the 201B7 parental line and 40.5% versus 4.8% for 253G1 [10]. Related feature-based classification of colony images and image-based irregularity detection across nominally identical culture processes established the same template [11,12], as did image-based pluripotency quantification software [13].

Deep learning changed the input rather than the question. Networks were trained to detect the earliest onset of differentiation from bright-field images [14], to identify iPSC-derived endothelial cells against CD31 immunofluorescence [15], to predict colony formation and morphology during reprogramming from time-lapse

sequences [16,17], and to read hematopoietic lineage choice from morphology up to three generations before molecular markers appear [18]. In mesenchymal stromal cells, morphological profiling of interferon-gamma-stimulated populations was shown to predict immunosuppressive capacity, extending the approach from identity to function [19,20]. Automated colony detection during reprogramming correlated with manual counting at $r = 0.877$ across three experiments under repeated random sub-sampling, with reliable detection from day 7 of a 20-24 day process [17].

Table 1: Representative colony and cell morphology models, with the validation regime that produced each reported figure [8,9,23,25,22,21].			
Study	Dataset size	Validation type	Reported performance
Maddah 2014	564 time-lapse datasets, 94 cultures	Held-out, 50/50	80% per position; 89% per well
Tokunaga 2014	60 images per line, 9 lines	Cross-validation	0.87 iPSC vs non-iPSC; 0.63-0.68 line vs line
Kavitha 2017	60 train, 30 test	Held-out plus 5-fold CV	95.5% held-out; above 90% CV
Vedeneeva 2023	150 colonies, 4,486 cells	Nested CV, no external	99 +/- 2%, AUC 0.997
Zhu 2021	149,428 single-cell images	Held-out 20%; separate experiment	0.923 held-out; 0.719-0.811 unseen inducers
Tong 2024	165,190 images, 93 batches	Batch-stratified vs batch-separated	99.9% vs 89.1% sequence-level

The validation regime sets the number

The clearest demonstration that regime, not architecture, drives the headline comes from cell line identification on an industrial bright-field dataset of 165,190 images spanning 32 cell lines and 93 batches. Under a batch-stratified split, where training and test images are drawn from all batches, CLANet reaches 99.9% +/- 0.1 sequence-level and 100.0% +/- 0.0 batch-level accuracy. Under a batch-separated split, with one batch per class for training and the remaining 61 batches held out, the same model on the same data reaches 89.1% +/- 2.2 sequence-level accuracy, 90.7% +/- 2.0 batch-level, and a

sequence-level F1 of 83.7% +/- 4.2 [21]. Nothing changed except whether batch membership leaked across the split. Comparator architectures fell much further under the batch-separated regime: a vision transformer on raw images to 51.2% +/- 2.8, and multiple-instance baselines to between 73.0% and 78.4%.

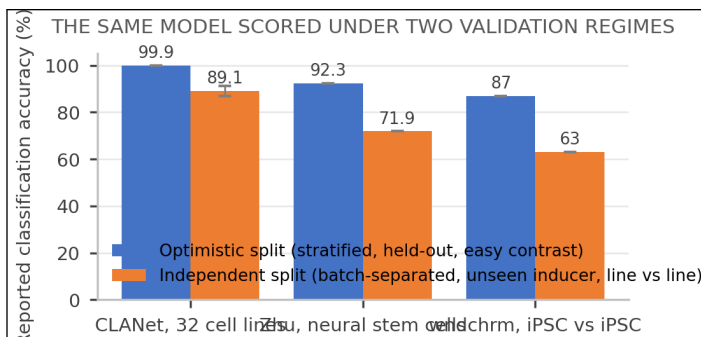


Figure 1: Paired accuracies for three studies in which the same method was scored under an optimistic and an independent validation regime; the left bar of each pair is not an independent estimate, since CLANet's 99.9% comes from a batch-stratified split of the same 165,190 images across 93 batches that yield 89.1% when batches are separated, the neural stem cell model's 92.3% is a held-out 20% of 149,428 single-cell images against 71.9% on the unseen LDH-NT3 inducer, and the 87% wndchrn figure separates iPSC from non-iPSC where line-versus-line separation falls to 63% on 60 colony images per line [21,22,9].

The second demonstration comes from neural stem cell differentiation. A bright-field model trained on 119,533 of 149,428 annotated single-cell images reached 0.923 overall accuracy on the held-out 20%, and a dark-field model reached 0.998. Applied to separate experiments using inducers absent from training, the same approach returned 81.09% for NGF, 82.73% for RA/SHH, 76.92% for CNTF, 76.79% for melatonin, 73.68% for NT3 and 71.89% for LDH-NT3 [22]. This is the same laboratory and the same cell source under different induction conditions, so it is a separate-experiment result rather than an external one in the regulatory sense — and it still costs eleven to twenty percentage points.

Cross-validation on small datasets produces the most extreme figures. A vector-based CNN on hiPSC colony images reported 95.5% accuracy from morphological features on a held-out test set of 30 images, with 60 for training, the authors noting the small training set as the

study's limitation [23]. On 150 colonies and 4,486 cells from three lines, four-parameter colony-level classification reached 74 +/- 6% under cross-validation and cell-level classification 69 +/- 3% [24]. Fusing cellular and colonial features on that same material under nested cross-validation with five inner and five outer folds raised random forest accuracy to 99 +/- 2% with an AUC of 0.997 [25]. That 99% has never met an independent sample. A CNN on hESC H9 phase-contrast images reported 89% on a 4:1 held-out split of 269 images from a single line [26]; transfer-learned ResNet-50 prediction of hMSC lineage reported 0.9572 accuracy and 0.9958 AUC on a 20% held-out split of 2,336 images [27]; and CNN staging of hepatocyte differentiation across three hiPSC lines reported 0.95 accuracy and 0.95 F1, with 592 of 614 test patches correct, on a patch-level 70/20/10 split of 1,331 images expanded to 6,972 patches [28]. Where the split is at patch level and patches are not grouped by source image, patches from one image can fall on both sides, which would inflate the estimate; the reachable record does not state whether grouping was applied.

The honest counterexamples sit in the same papers. Cross-validated discrimination of one iPSC line from another — the clinically relevant comparison — runs at 0.66, 0.68 and 0.63 against three different iPSC lines, against 0.87 and 0.96 for the easy contrasts [9]. Across 79 studies in a ten-year systematic scoping review, convolutional networks appeared in 33 (41%) and machine learning methods in 44 (55.6%), code was available for 29%, only two studies provided best-practice reports, 77% reported data-splitting practices and 67% documented hyperparameter tuning [1]. Reporting a split is not the same as reporting an independent one.

Batches, microscopes and laboratories

No study located here validates a stem cell colony-quality model at a second laboratory on a second microscope. The nearest evidence comes from adjacent imaging tasks and is discouraging. A systematic evaluation of

batch-correction methods across five scenarios of increasing heterogeneity — from one laboratory with 302 compounds and 13 batches, up to five laboratories, three microscope types, 82,412 compounds and 60 runs — concludes that in the hardest scenario none of the methods adequately removes the batch effects, with Harmony and Seurat RPCA among the top three by mean rank [29]. Self-supervised representations improve on engineered features by 29% in perturbation mean average precision ($p = 1.10 \times 10^{-6}$) and by 73% on cross-source perturbation retrieval, and recover 11 gene families against 3 for CellProfiler on unseen overexpression with adjusted mutual information 0.51 against 0.19 — yet on a genuinely external bioactivity task spanning 209 ChEMBL assays, mean AUROC is 0.72 against 0.73 for the best supervised CNN, that is, no advantage at all [30].

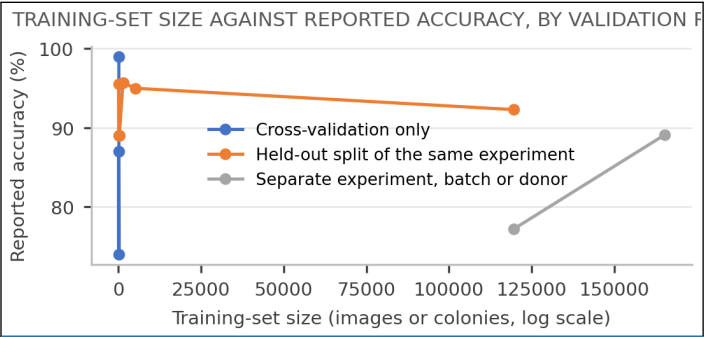


Figure 2: Reported accuracy against training-set size for eleven evaluations from eight studies, coloured by validation regime, showing that the two smallest datasets carry two of the highest figures and that the separate-experiment points (the mean of the six confirmed unseen-inducer accuracies, 77.2%, and the batch-separated 89.1%) sit below the held-out points at comparable scale [23,24,25,26,8,28,27,22,21,9].

Where the same model has been carried to new donors, the cost is measurable. Label-free quantitative bright-field prediction of iPSC-RPE function achieved $R^2 = 0.97$ with RMSE 70.6 ohm.cm2 for trans-epithelial resistance and 94% accuracy, 100% sensitivity and 90% specificity at the 400 ohm.cm2 threshold under held-out cross-validation across clone combinations, on 864 fields of view with segmentation checked against 12,750 hand-segmented cells. Transferred to eight iPSC clones from three age-related macular degeneration patients, R^2 was 0.92,

and across 18 cross-donor training combinations mean RMSE rose to 86.9 +/- 14.3 ohm.cm2, roughly 23% worse than the within-donor figure [31]. That is the best-behaved transfer in this literature, and it still degrades.

What the label is worth

Validation regime is only half of comparability; the other half is what the model was trained to reproduce. Where the label is one expert's three-category ranking of time-lapse movies [8], the ceiling is set by the 48% two-expert concordance described above [4]. Where the label is a single surface marker [15] or a protocol timepoint [28], the model learns the proxy, not the phenotype. Where the label is a genetic reporter, as in quantitative phase imaging of hematopoietic stem cells predicting Hlf-tdTomato intensity from 20 consecutive frames at $r = 0.73$ and mean squared error 0.02 on a separate held-out dataset, the ground truth is precise but unavailable in a clinical product [32]. Only direct functional measurement on the same wells escapes the problem [31].

Regulators set the terminal requirement. A potency assay must reflect the product's mechanism of action and ideally correlate with clinical response; the FDA mandates a validated quantitative functional potency assay for commercial release, while EU regulators permit a validated surrogate provided a functional assay exists for characterisation and the two correlate, and a matrix of methods is accepted for complex products [33]. A non-invasive imaging model is exactly the kind of surrogate that framework contemplates for release testing [34] — but only once the correlation has been demonstrated on material the model has never seen.

Table 2: What each model was trained against, and the known weakness of that ground truth [8,15,28,32,31,4].		
Study	Ground-truth label	Weakness of the label
Maddah 2014	Good / fair / poor culture	One rater; two experts agreed on 48% of 449 colonies

Study	Ground-truth label	Weakness of the label
Kusumoto 2018	CD31 immunofluorescence	One surface marker stands in for identity
Marzec-Schmidt 2023	Differentiation day	A schedule, not a phenotype
Yogo 2025	Hlf-tdTomato reporter	Requires a transgene; unavailable clinically
Schaub 2020	Trans-epithelial resistance, VEGF ratio	Measured function, but still not a clinical outcome

Conclusion

The reported accuracies in this field cannot be ranked against one another, and reviews that tabulate them as though they can are producing an artefact. The 99.9% and the 89.1% belong to one model on one dataset and differ only in whether batch identity leaked across the split; the 0.923 and the 0.719 belong to one model and differ only in whether the inducer was seen in training. Those two comparisons, rather than any single headline, are the state of the evidence. A 99 +/- 2% under nested cross-validation on 150 colonies from one laboratory and a 76% on unseen induction conditions from 149,428 images are not two points on the same scale. The practical requirements follow directly: report the split construction with every figure, group splits by source image, batch and donor rather than by tile, and treat cross-batch and cross-line evaluation as the minimum publishable standard rather than an appendix. Until a colony-quality model is evaluated at a second laboratory on a second microscope against a functional endpoint, it has not been tested.

References

1. Vo, Q.D., Saito, Y., Ida, T., Nakamura, K., Yuasa, S. (2024) The use of artificial intelligence in induced pluripotent stem cell-based technology over 10-year period: A systematic scoping review. *PLOS ONE*, 19(5): e0302537.

2. Deininger, L., Caldarelli, P., Zernicka-Goetz, M., Mikut, R. (2026) From pixels to patterns: the AI revolution in stem

cell-derived models. *Nature Methods*, 23(9): 1710-1723.

3. Coronello, C., Francipane, M.G. (2022) Moving Towards Induced Pluripotent Stem Cell-based Therapies with Artificial Intelligence and Machine Learning. *Stem Cell Reviews and Reports*, 18(2): 559-569.

4. National Institute of Standards and Technology, Information Systems Group. Pluripotency Scoring dataset. Web resource, <https://isg.nist.gov/deepzoomweb/data/pluripotencyScoring>. Not a peer-reviewed publication.

5. Paternot, G., Devroe, J., Debrock, S., D'Hooghe, T.M., Spiessens, C. (2009) Intra- and inter-observer analysis in the morphological assessment of early-stage embryos. *Reproductive Biology and Endocrinology*, 7: 105.

6. Kakulavarapu, R.K., Delbarre, E., Sharma, A., Jahanlu, D., Riegler, M.A., Haugen, T.B., et al. (2026) Investigating discrepancies in accuracy, agreement and interpretability for single-frame embryo classification tasks conducted by embryologists and deep learning models. *Frontiers in Reproductive Health*, 8: 1778326.

7. Piotrowski, T., Rippel, O., Elanzew, A., Niessing, B., Stucken, S., Jung, S., et al. (2021) Deep-learning-based multi-class segmentation for automated, non-invasive routine assessment of human pluripotent stem cell culture status. *Computers in Biology and Medicine*, 129: 104172.

8. Maddah, M., Shoukat-Mumtaz, U., Nassirpour, S., Loewke, K. (2014) A System for Automated, Noninvasive, Morphology-Based Evaluation of Induced Pluripotent Stem Cell Cultures. *SLAS Technology*, 19(5): 454-460.

9. Tokunaga, K., Saitoh, N., Goldberg, I.G., Sakamoto, C., Yasuda, Y., Yoshida, Y., et al. (2014) Computational image analysis of colony and nuclear morphology to evaluate human induced pluripotent stem cells. *Scientific Reports*, 4: 6996.

10. Kato, R., Matsumoto, M., Sasaki, H., Joto, R., Okada, M., Ikeda, Y., et al. (2016) Parametric analysis of colony morphology of non-labelled live human pluripotent stem cells for cell

- quality control. *Scientific Reports*, 6: 34009.
11. Joutsijoki, H., Haponen, M., Rasku, J., Aalto-Setälä, K., Juhola, M. (2016) Machine Learning Approach to Automated Quality Identification of Human Induced Pluripotent Stem Cell Colony Images. *Computational and Mathematical Methods in Medicine*, 2016: 3091039.
12. Nagasaka, R., Gotou, Y., Yoshida, K., Kanie, K., Shimizu, K., Honda, H., et al. (2017) Image-based cell quality evaluation to detect irregularities under same culture process of human induced pluripotent stem cells. *Journal of Bioscience and Bioengineering*, 123(5): 642-650.
13. Perestrelo, T., Chen, W., Correia, M., Le, C., Pereira, S.L., Rodrigues, A.S., et al. (2017) Pluri-IQ: Quantification of Embryonic Stem Cell Pluripotency through an Image-Based Analysis Software. *Stem Cell Reports*, 9(2): 697-709.
14. Waisman, A., La Greca, A., Mobbs, A.M., Scarafia, M.A., Santin Velazque, N.L., Neiman, G., et al. (2019) Deep Learning Neural Networks Highly Predict Very Early Onset of Pluripotent Stem Cell Differentiation. *Stem Cell Reports*, 12(4): 845-859.
15. Kusumoto, D., Lachmann, M., Kunihiro, T., Yuasa, S., Kishino, Y., Kimura, M., et al. (2018) Automated Deep Learning-Based System to Identify Endothelial Cells Derived from Induced Pluripotent Stem Cells. *Stem Cell Reports*, 10(6): 1687-1695.
16. Chu, S.-L., Sudo, K., Yokota, H., Abe, K., Nakamura, Y., Tsai, M.-D. (2022) Human induced pluripotent stem cell formation and morphology prediction during reprogramming with time-lapse bright-field microscopy images using deep learning methods. *Computer Methods and Programs in Biomedicine*, 229: 107264.
17. Fan, K., Zhang, S., Zhang, Y., Lu, J., Holcombe, M., Zhang, X. (2017) A Machine Learning Assisted, Label-free, Non-invasive Approach for Somatic Reprogramming in Induced Pluripotent Stem Cell Colony Formation Detection and Prediction. *Scientific Reports*, 7(1): 13496.
18. Buggenthin, F., Buettner, F., Hoppe, P.S., Ende, M., Kroiss, M., Strasser, M., et al. (2017) Prospective identification of hematopoietic lineage choice by deep learning. *Nature Methods*, 14(4): 403-406.
19. Klinker, M.W., Marklein, R.A., Lo Surdo, J.L., Wei, C.-H., Bauer, S.R. (2017) Morphological features of IFN-gamma-stimulated mesenchymal stromal cells predict overall immunosuppressive capacity. *Proceedings of the National Academy of Sciences*, 114(13): E2598-E2607.
20. Marklein, R.A., Klinker, M.W., Drake, K.A., Polikowsky, H.G., Lessey-Morillon, E.C., Bauer, S.R. (2019) Morphological profiling using machine learning reveals emergent subpopulations of interferon-gamma-stimulated mesenchymal stromal cells that predict immunosuppression. *Cytotherapy*, 21(1): 17-31.
21. Tong, L., Corrigan, A., Kumar, N.R., Hallbrook, K., Orme, J.P., Wang, Y., et al. (2024) CLANet: A comprehensive framework for cross-batch cell line identification using brightfield images. *Medical Image Analysis*, 94: 103123.
22. Zhu, Y., Huang, R., Wu, Z., Song, S., Cheng, L., Zhu, R. (2021) Deep learning-based predictive identification of neural stem cell differentiation. *Nature Communications*, 12: 2614.
23. Kavitha, M.S., Kurita, T., Park, S.-Y., Chien, S.-I., Bae, J.-S., Ahn, B.-C. (2017) Deep vector-based convolutional neural network approach for automatic recognition of colonies of induced pluripotent stem cells. *PLOS ONE*, 12(12): e0189974.
24. Krasnova, O.A., Gursky, V.V., Chabina, A.S., Kulakova, K.A., Alekseenko, L.L., Panova, A.V., et al. (2022) Prognostic Analysis of Human Pluripotent Stem Cells Based on Their Morphological Portrait and Expression of Pluripotent Markers. *International Journal of Molecular Sciences*, 23(21): 12902.
25. Vedeneeva, E., Gursky, V., Samsonova, M., Neganova, I. (2023) Morphological Signal Processing for Phenotype Recognition of Human Pluripotent Stem Cells Using Machine Learning Methods. *Biomedicines*, 11(11): 3005.

26. Mamaeva, A., Krasnova, O., Khvorova, I., Kozlov, K., Gursky, V., Samsonova, M., et al. (2023) Quality Control of Human Pluripotent Stem Cell Colonies by Computational Image Analysis Using Convolutional Neural Networks. *International Journal of Molecular Sciences*, 24(1): 140.
27. Mai, M., Luo, S., Fasciano, S., Oluwole, T.E., Ortiz, J., Pang, Y., et al. (2023) Morphology-based deep learning approach for predicting adipogenic and osteogenic differentiation of human mesenchymal stem cells (hMSCs). *Frontiers in Cell and Developmental Biology*, 11: 1329840.
28. Marzec-Schmidt, K., Ghosheh, N., Stahlschmidt, S.R., Kupperts-Munther, B., Synnergren, J., Ulfenborg, B. (2023) Artificial Intelligence Supports Automated Characterization of Differentiated Human Pluripotent Stem Cells. *Stem Cells*, 41(9): 850-861.
29. Arevalo, J., Su, E., Ewald, J.D., van Dijk, R., Carpenter, A.E., Singh, S. (2024) Evaluating batch correction methods for image-based cell profiling. *Nature Communications*, 15: 6516.
30. Kim, V., Adaloglou, N., Osterland, M., Morelli, F.M., Halawa, M., König, T., et al. (2025) Self-supervision advances morphological profiling by unlocking powerful image representations. *Scientific Reports*, 15: 4876.
31. Schaub, N.J., Hotaling, N.A., Manescu, P., Padi, S., Wan, Q., Sharma, R., et al. (2020) Deep learning predicts function of live retinal pigment epithelium from quantitative microscopy. *The Journal of Clinical Investigation*, 130(2): 1010-1023.
32. Yogo, T., Iwamoto, Y., Becker, H.J., Kimura, T., Ishida, R., Sugiyama-Finnis, A., et al. (2025) Quantitative phase imaging with temporal kinetics predicts hematopoietic stem cell diversity. *Nature Communications*, 16: 6496.
33. Salmikangas, P., Carlsson, B., Klumb, C., Reimer, T., Thirstrup, S. (2023) Potency testing of cell and gene therapy products. *Frontiers in Medicine*, 10: 1190016.
34. Tucker, B.A., Mullins, R.F., Stone, E.M. (2020) Autologous cell replacement: a noninvasive AI approach to clinical release testing. *The Journal of Clinical Investigation*, 130(2): 608-611.