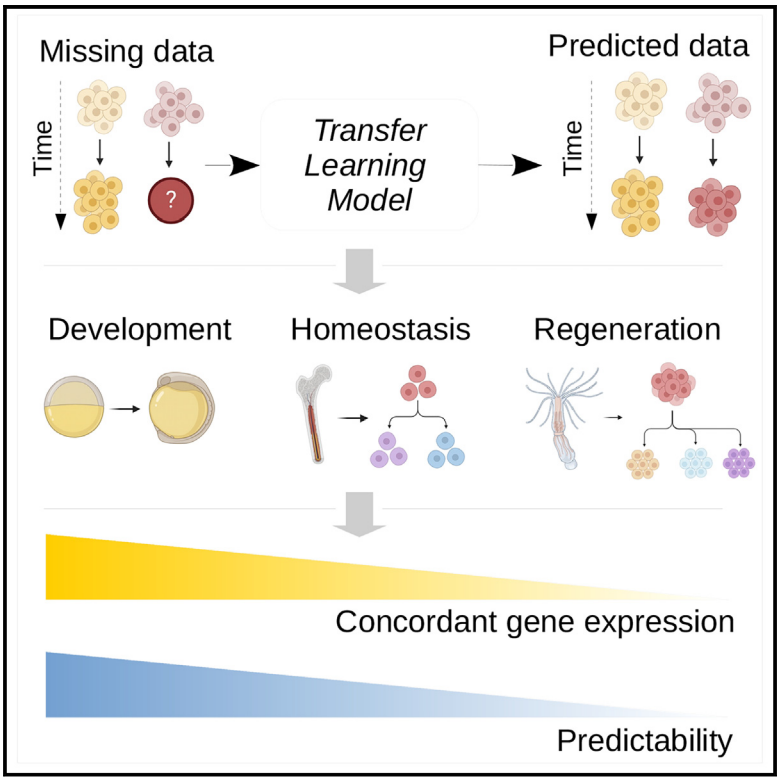
Cell Systems

Article

Inference of differentiation trajectories by transfer learning across biological processes

Graphical abstract



Authors

Gaurav Jumde, Bastiaan Spanjaard, Jan Philipp Junker

Correspondence

bastiaan.spanjaard@charite.de (B.S.), janphilipp.junker@mdc-berlin.de (J.P.J.)

In brief

Jumde et al. use transfer learning to quantify predictability of cellular differentiation trajectories. They observe that non-linear methods outperform linear approaches, and they achieve the best predictions with a variational autoencoder that models changes in transcriptional variance. They find that predictability is based on concordant transcriptional processes along embryonic differentiation.

Highlights

- Transfer learning to predict cellular differentiation trajectories
- Custom variational autoencoder explicitly models changes in transcriptional variance
- High predictability of differentiation trajectories in embryonic development
- Predictability is based on concordant transcriptional changes across biological processes

Jumde et al., 2024, Cell Systems *15*, 75–82
January 17, 2024
https://doi.org/10.1016/j.cels.2023.12.002

CellPress

Cell Systems

CellPress
OPEN ACCESS

Article

Inference of differentiation trajectories by transfer learning across biological processes

Gaurav Jumde,[1,2] Bastiaan Spanjaard,[1,3,*] and Jan Philipp Junker[1,3,4,*]
[1]Max Delbrück Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, 10115 Berlin, Germany
[2]Humboldt Universität zu Berlin, Faculty of Life Sciences, Department of Biology, 10115 Berlin, Germany
[3]Charité Universitätsmedizin Berlin, 10117 Berlin, Germany
[4]Lead contact
*Correspondence: bastiaan.spanjaard@charite.de (B.S.), janphilipp.junker@mdc-berlin.de (J.P.J.)
https://doi.org/10.1016/j.cels.2023.12.002

SUMMARY

Stem cells differentiate into distinct fates by transitioning through a series of transcriptional states. Current computational approaches allow reconstruction of differentiation trajectories from single-cell transcriptomics data, but it remains unknown to what degree differentiation can be predicted across biological processes. Here, we use transfer learning to infer differentiation processes and quantify predictability in early embryonic development and adult hematopoiesis. Overall, we find that non-linear methods outperform linear approaches, and we achieved the best predictions with a custom variational autoencoder that explicitly models changes in transcriptional variance. We observed a high accuracy of predictions in embryonic development, but we found somewhat lower agreement with the real data in adult hematopoiesis. We demonstrate that this discrepancy can be explained by a higher degree of concordant transcriptional processes along embryonic differentiation compared with adult homeostasis. In summary, we establish a framework for quantifying and exploiting predictability of cellular differentiation trajectories.

INTRODUCTION

Cellular differentiation, the process by which cells acquire a specialized state, is ubiquitous and necessary during development, homeostasis, and regeneration. During development, the zygote gives rise to the three germ layers, which generate the various cell types of the animal and eventually create a fully formed organism. Cell differentiation can be investigated with several conceptually different approaches: (1) studying the activity of the signaling pathways that instruct patterning and differentiation,[1] (2) constructing Waddingtonian-like dynamical landscapes from quantitative data,[2] (3) building mechanistic models based on reconstruction of gene regulatory networks,[3] and (4) using single-cell transcriptome profiling to measure cell identity changes.[4–6] Specifically, the transcriptomic profiles of single cells can be arranged in trajectories that describe the evolution of cells during development in, for example, zebrafish,[7,8] *Xenopus tropicalis*,[9] and mouse.[10–13] Single-cell trajectories can similarly be employed to describe how stem cells in the adult body give rise to different cell types through differentiation processes during homeostasis or regeneration. Examples of this are hematopoiesis,[14–18] homeostasis of the small intestinal epithelium,[19] and axolotl limb regeneration.[20]

A major goal in developmental biology is to predict the future states of a cell. However, the predictive power of the approaches described above is typically limited and, in particular, inferred differentiation trajectories have generally been used for describing and analyzing transcriptomic changes rather than for prediction of future cell states. Optimal transport-based latent space models have been applied to interpolate between different time points.[21,22] Here, we hypothesize that separate trajectories are governed by similar rules that act on populations of cells and that similar models could be used to extrapolate differentiation processes. We use transfer learning to transfer these rules between trajectories. Although transfer learning in single-cell data has previously been implemented in cFit[23] to enhance lower-quality data, we believe our design is complementary: cFit is designed around the common (biological) denominator of multiple datasets, while we focus on differences between trajectories. We evaluate these rules by training models with increasing complexity and then using them to predict differentiation.

Neural networks are often used for transfer learning tasks. A variational autoencoder[24] is a type of neural network that is trained to generate realistic high-dimensional predictions from a low-dimensional latent space that encodes informative features in the data. In single-cell transcriptomics, they have been used to predict drug and general perturbation response.[22,25–28] Predictions can be generated by operations within the latent space. Here, we use a combination of variational autoencoders and latent space normalizing flows[29] to predict gene expression changes in embryonic development and adult hematopoiesis. We show that latent space techniques to modulate gene expression and variability allow us to successfully predict differentiation

trajectories across germ layers in early zebrafish and mouse development at the single-cell level, but we find that predictive power is somewhat lower in adult hematopoiesis and is almost completely lost in the invertebrate hydra. Our results suggest that concordant patterns of mean gene expression and gene expression variability lead to a high degree of predictability during early development, while highly specific differentiation trajectories in adult homeostasis lead to lower predictability.

RESULTS

To investigate the predictability of cellular differentiation processes, we focused on a single-cell RNA sequencing (RNA-seq) dataset covering the first 12 h of zebrafish development.[7] This time window includes the emergence of cell-type diversity at gastrulation and therefore provides a powerful test case for inference of transcriptional states (Figures 1A and S1). Specifically, we asked to what degree we could predict transcriptional changes in the ectoderm between 5.3 and 12 hpf (hours post fertilization) based on the transcriptional profiles of mesoderm and endoderm at the same time points. The two lineages have already undergone a transcriptional split at 5.3 hpf, but diversify further into multiple transcriptional states by 12 hpf. We started our analysis with a simple linear model based on vector arithmetic in gene expression space. We computed the vector of mean gene expression change in mesoderm/endoderm between 5.3 and 12 hpf and added this vector to the single-cell gene expression values of the ectoderm (Figure 1B) (STAR Methods). We found that this approach led to a remarkably high correlation between the real and predicted data, with R^2 = 0.98 for mean gene expression and R^2 = 0.67 for the standard deviation. Of note, the procedure described here is a translation in gene expression change, i.e., each single-cell expression pattern is shifted by the same constant vector. Consequently, this approach is unable to account for any changes in gene expression variance. This is a major limitation, as, in embryonic development and to a similar degree in adult stem cell systems, a population of relatively homogeneous stem cells gives rise to a range of differentiated cell types. Differentiation thereby leads to a continuous increase in transcriptomic diversity.

Next, we wanted to investigate whether the accuracy of the prediction could be increased further by applying a non-linear machine learning approach. We used the scGen framework, a variational autoencoder that was successfully used to infer cellular perturbation response in single-cell RNA-seq data across cell types.[25] Similar to the linear approach, we again performed vector arithmetic, but now in the latent space created by the variational autoencoder (STAR Methods). However, this only led to a minimal improvement of the predictions (Figure 1C). In particular, we noted that the variational autoencoder failed at transferring the observed increase in gene expression variance between the two lineages. We therefore set out to develop deep cell predictor (DCP), a machine learning approach that explicitly models changes in transcriptional variance using a combination of variational autoencoders and normalizing flows (Figure S2A; STAR Methods; Table S1). After parameter selection and training (Figures S2B and S2C), we found that our DCP model performed better at predicting transcriptional variance, albeit still with lower correlations than for mean gene expression (Figure 1D). To visualize the information transferred, we decoded latent space predictions after vector arithmetic and after normalizing flows and performed principal-component analysis (PCA) for visualization (Figure S3). The predicted distribution of cells already started resembling the target distribution after vector arithmetic due to our regularization approach (STAR Methods), in contrast to the vector arithmetic performed in scGen (Figure 1C). The normalizing flows allowed us to further approximate the target distribution. In summary, we found that transfer learning allows successful inference of differentiation trajectories across germ layers at a stage of pronounced transcriptomic diversification.

Next, we sought to explore the predictability of developmental cell differentiation in more depth by swapping test and training data and by contrasting forward (5.3–12 hpf) and backward (12–5.3 hpf) predictions (Figure 2A). For this, we focused the analysis on our DCP model, which had performed best in our previous analysis and also led to better prediction of the variance here (Figures S4 and S5). We again found very high predictability of mean gene expression ($R^2 > 0.94$) (Figure 2B) and somewhat lower accuracy for inference of standard deviation (R^2 between 0.66 and 0.76) (Figure 2C). Although forward and backward inference generally performed equally well, we noticed that we consistently underpredicted the expansion of gene expression variance in forward predictions (Figures 2C and 2D), again highlighting the challenges associated with correctly capturing transcriptional spread.

We compared the gene expression variability with and without normalizing flows for forward predictions in zebrafish development and found that normalizing flows lead to a strong increase in predicted expression variability, especially for genes that increase in variability over time (Figures S6A–S6C). Temporal changes in variance are very important to understand the emergence of cellular heterogeneity. We next zoomed in on examples of genes with a high variance increase and compared their predictability (Figure S6D) with their expression patterns (Figure S6F) on a UMAP (uniform manifold approximation and projection) of the training and test data (Figure S6E). One clearly observable class of genes are genes that are ubiquitously expressed at 12 hpf and where the expression variability is due to continuous differences in expression level. This class contains genes like prothymosin alpha b (PTMAB), the β-tubulin TUBB2B, as well as many ribosomal genes such as RPLP1, RPL39, and RPL34. For these genes, normalizing flows outperform a model without normalizing flows, but the predicted variance is still lower than the observed variance. For some other genes, the variability comes from on/off expression patterns. If only few cells express these genes, for example, the lipid carrier gene APOEB or the hematopoietic development gene LMO2, the performance of normalizing flows is lacking, presumably due to lack of training data. However, if enough cells express the genes, for example, in prothymosin alpha a (PTMAA) and the developmental transcription factor homeobox protein CDX4, normalizing flows still outperform a model without normalizing flows.

We found that the predicted single-cell transcriptomes integrated well with the real data, suggesting that our approach can generate realistic single-cell transcriptomes (Figure S7; STAR Methods). However, for forward predictions, we also observed that the predicted data did not accurately capture

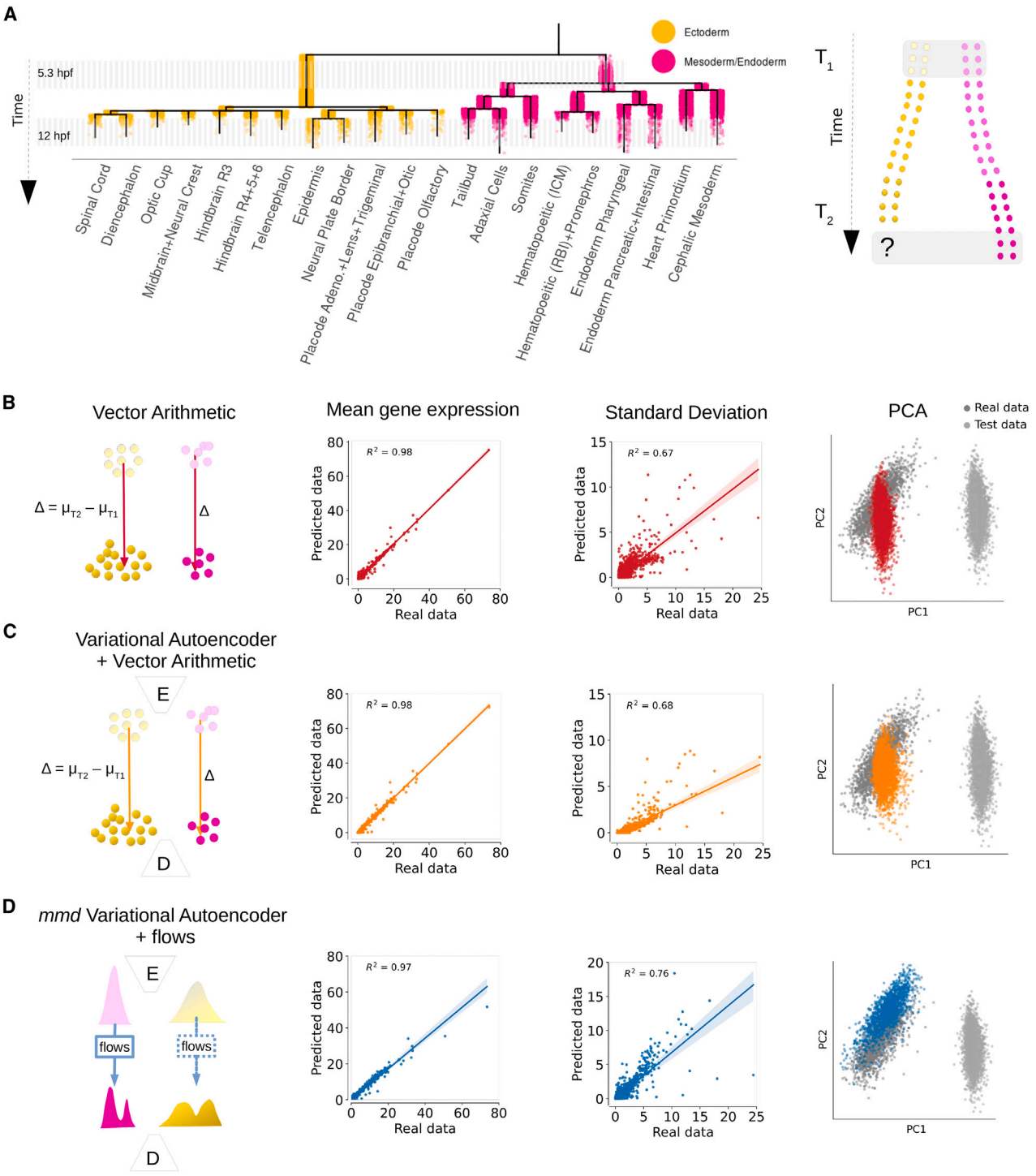


Figure 1. Variational autoencoder with explicit modeling of gene expression variability improves single-cell expression predictions in zebrafish development

(A) Transcriptome-based differentiation trajectories of zebrafish development. The single-cell RNA-seq data were analyzed with the diffusion-based computational trajectory reconstruction method URD, as in the original publication.[7] In this figure, gene expression in ectoderm at 12 hpf is predicted from mesoderm/endoderm at 5.3 and 12 hpf as well as ectoderm at 5.3 hpf.

(B–D) Predicted mean gene expression and gene expression variability across all genes, as well as first two principal components across all cells, using transcriptome vector arithmetic (B), latent space vector arithmetic (C), and latent space normalizing flows (D). Latent space normalizing flows capture the change in gene expression variability. Fit lines in (B)–(D) are linear regressions with zero intercept. Data from Farrell et al.,[7] GEO: GSE106587.

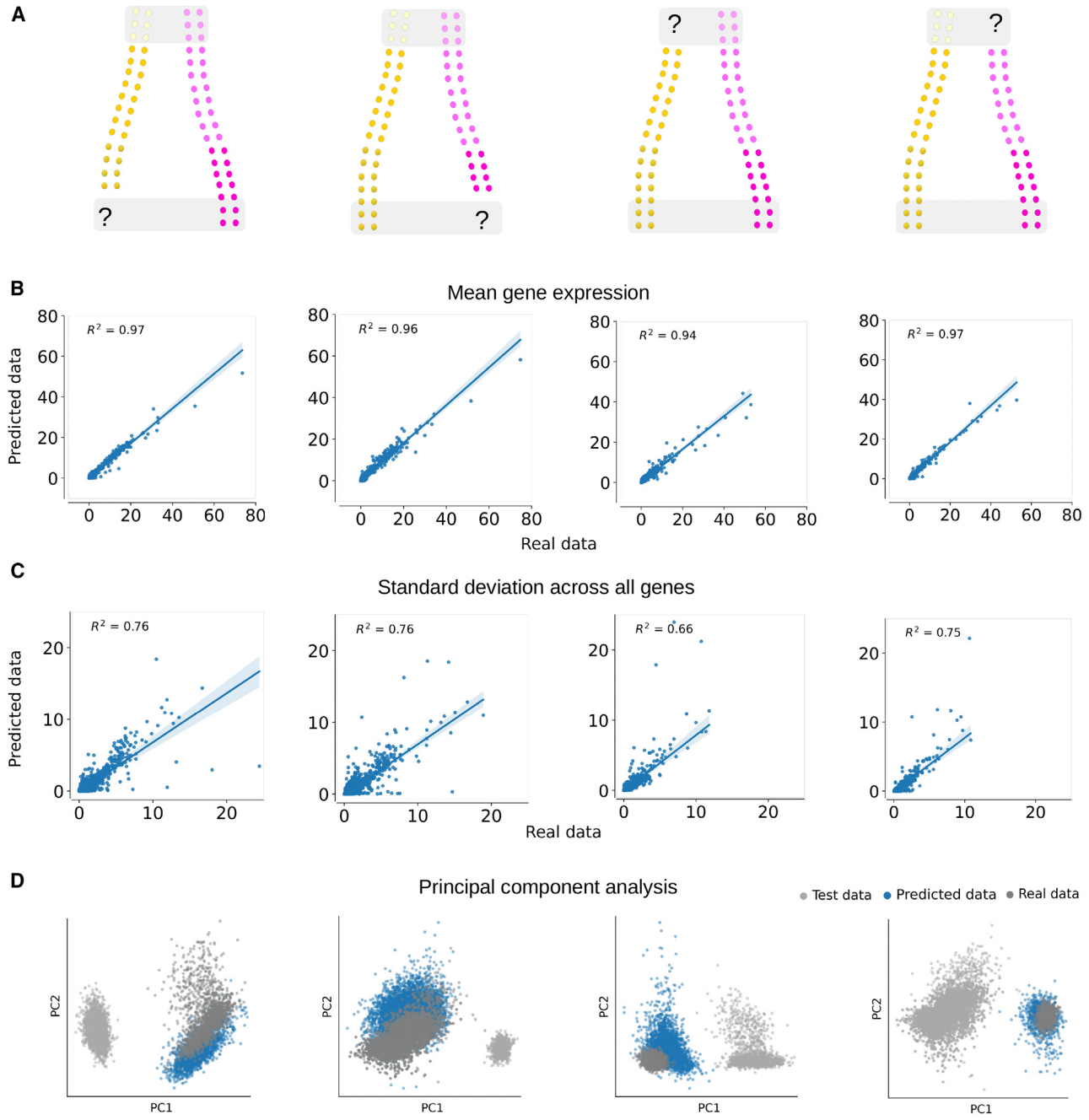


Figure 2. DCP prediction of mean gene expression and gene expression variability in zebrafish development

(A) We tested the predictive power of DCP in four different scenarios: forward (first two rows) and backward (last two rows) predictions of ectoderm and mesoderm, respectively.

(B) Mean gene expression across all genes is highly accurate in all scenarios.

(C) Gene variability predictions across all genes have correlations between 0.66 and 0.78, with slightly higher values obtained for forward predictions.

(D) Principal-component analysis across all cells corroborates accuracy of gene expression mean and variability. Fit lines in (B) and (C) are linear regressions with zero intercept. Data from Farrell et al.,[7] GEO: GSE106587.

the emerging structure of cell-type clusters at 12 hpf (Figure S7). This observation suggests that, despite accurate prediction of mean and average gene expression, our approach cannot correctly determine the covariance structure of cell-type-specific genes. Inspired by this finding, we hypothesized that the high predictability of the developmental trajectories might at least in part be explained by the fact that, for the analysis in Figures 1 and 2, we lumped together various cell types from the ectoderm and mesendoderm, respectively, thereby reducing the impact of cell-type-specific genes. When training on

individual cell types, we indeed observed a slight decline in correlation between real and predicted data (Figure S8). However, overall the correlations remained very high ($R^2 \geq 0.93$) for mean gene expression, suggesting that the observed high predictability is not caused by pooling of different cell types.

We next hypothesized that high predictability of early development might be specific to the zebrafish and could, for instance, be related to maternal mRNA decaying over time at the developmental stages investigated. We therefore decided to apply our approach to a published dataset covering embryonic development in the mouse[10] between E6.5 and E8.5. Training on ectoderm and predicting mesendoderm and vice versa, we found equally high predictability of gene expression changes as in zebrafish (Figure S9). Although more datasets will be needed to prove the generality of our findings, these results indicate that predictability of early development across lineages is not limited to specific systems.

The observed high predictability in developmental differentiation prompted us to investigate whether transfer learning also allows successful inference in adult stem cell systems. We decided to focus on mouse hematopoiesis, using a published dataset[18] in which we focused on the trajectories of monocyte and neutrophil differentiation (Figures 3A and S10). In contrast to the developmental data, prediction of mean gene expression largely failed (Figures 3B and 3C). With the exception of backward prediction of monocytes, all correlations of mean gene expression between real data and predictions were $R^2 < 0.3$. We next set out to understand which biological factors explain these pronounced differences in predictability. Taking the example of the forward prediction of monocytes, we found that our transfer learning strategy erroneously predicted upregulation of neutrophil-specific genes in the monocytes (Figure 3B). Because our approach is based on shared transcriptional signatures between test and training data without any further information, it is unavoidable that lineage-specific processes like upregulation of *S100a9* and *Ngp* are also transferred from neutrophils to monocytes. We therefore hypothesized that a small number of highly expressed lineage-specific genes might have a detrimental effect on overall predictability. Indeed, correlations increased when removing outlier genes (Figures 3C, S11, and S12). However, predictability still remained slightly lower than for the embryonic dataset, suggesting that additional differences between the analyzed datasets are responsible for the observed behavior.

We reasoned that the degree of predictability in transfer learning would ultimately be based on the fraction of genes that change in a concordant manner in the test and the training data. We therefore compared mean gene expression fold changes between the two time points across all genes for zebrafish development and mouse hematopoiesis and computed the mutual information (MI). We found a high concordance of fold changes between zebrafish ectoderm and mesoderm/endoderm for mean gene expression as well as for the standard deviation (MI = 0.90 and MI = 0.71 for $\log_2$ fold change of mean and standard deviation, respectively) (Figure 3D; Table S1). By contrast, the MI was considerably lower between monocytes and neutrophils (MI = 0.49 for mean and MI = 0.33 for standard deviation of $\log_2$ fold change). Hence, we concluded that the high predictability of the zebrafish developmental differentiation trajectories is based on a large set of concordant gene expression changes between lineages, which does not exist in the adult dataset. The concordant changes in zebrafish ectoderm and mesoderm/endoderm are enriched in genes that are involved in metabolism and gene expression (Datasets S1 and S2; STAR Methods), suggesting that the observed predictability is based on fundamental cellular processes that change globally across cell lineages in early zebrafish development.

To evaluate the information content of the latent space, we propose a consistency check based on two successive predictions (Figure S13). In this consistency check, we generate a prediction B2′ from A1, A2, and B1. We then use B2′ together with A1 and B1 to generate prediction A2″ and compare this to the ground truth A2. Importantly, we can generate these double predictions without any knowledge of B2. This provides a way to reveal whether the latent space captures the transcriptomic diversity of the dataset well.

To further validate the correspondence between concordant gene expression changes and predictability, and in particular the use of MI as a predictive metric, we identified a single-cell RNA-seq dataset of adult hydra as a counterexample[30]: because hydra polyps continually renew all of their cells using three separate stem cell populations, this invertebrate represents a hybrid case between a developmental and an adult system. Interestingly, we found that the gene expression fold changes between the ectoderm and endoderm lineages displayed only minimal concordance in hydra (MI = 0.11 and MI = 0.06 for $\log_2$ fold change of mean and standard deviation, respectively), leading to poor predictions of mean gene expression (Figure S14).

DISCUSSION

In summary, we found that developmental differentiation trajectories can be inferred remarkably well based on transfer learning. Although mean gene expression can be predicted successfully with a linear model by performing vector arithmetic in gene expression space, we found that our new approach based on a variational autoencoder and normalizing flows led to better prediction of transcriptional variance. This is an important aspect, because an increase in transcriptional diversity is a defining feature of developmental and adult stem cell differentiation processes. Our main goal here was to determine the extent and limits of transcriptional predictability, and we developed DCP specifically to address this question. Additional analysis and benchmarking would be required to evaluate whether DCP can also be applied in other scenarios, e.g., as an improved tool for perturbation modeling.

Our findings suggest that, during development, the distribution of single-cell transcriptomes changes its mean and variability in an at least partially predictable fashion, ultimately based on concordant gene expression programs between different lineages. One of the first steps in the analysis of single-cell transcriptomics data is typically the selection of highly variable genes. Although this is a useful and necessary step for dimensionality reduction, our work also shows that the focus on genes that distinguish the individual cell types can lead us to underestimate the degree of shared transcriptomic changes across lineages, which ultimately reflects concordant cellular processes that are discarded when selecting highly variable genes.

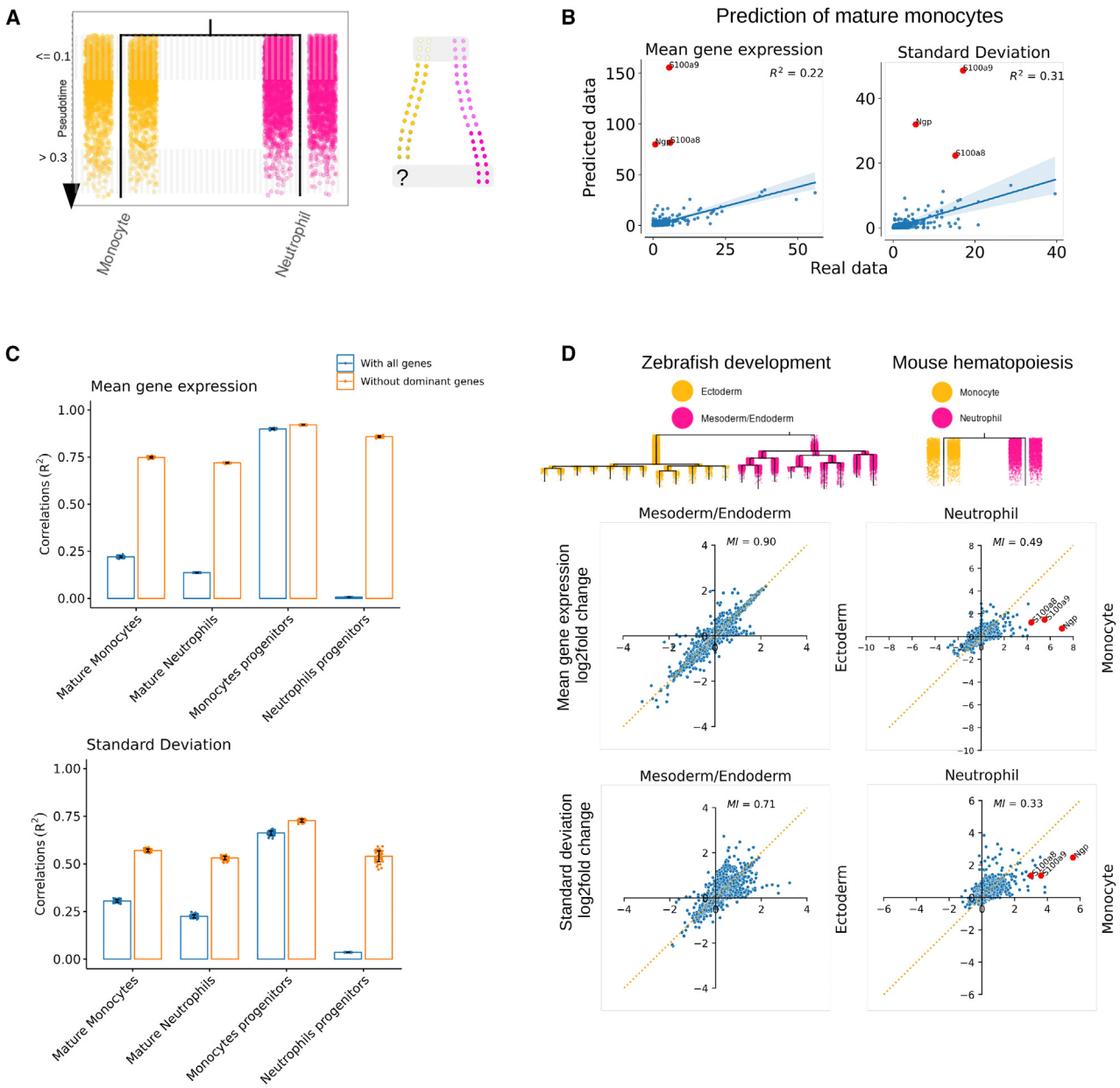


Figure 3. DCP fails to predict lineage-specific genes in hematopoietic differentiation

(A) Transcriptome-based lineage tree of single-cell hematopoietic differentiation data. Predictions in this figure are done between neutrophils and monocytes during their transition from a precursor to a differentiated state.

(B) DCP predictions of mature monocyte mean gene expression and expression variability is hindered by high expression of three neutrophil-specific genes. Fit line is regression with zero intercept.

(C) Removal of three neutrophil-specific genes strongly increases predictive power in hematopoietic differentiation.

(D) URD transcriptomic tree of zebrafish development and mouse hematopoiesis representing distinct cell-type lineages. Log-fold changes of mean gene expression and expression variability in mouse hematopoiesis show a stronger decoupling of lineages than in zebrafish development. Dotted yellow line of slope 1 indicates perfect correspondence in log-fold changes. Data from Farrell et al.[7] (GEO: GSE106587) and Weinreb et al.[18] (GEO: GSE140802). Error bars in (C) were determined by taking a 95% confidence interval of the bootstrap distribution.

We found a higher degree of concordant transcriptional changes in development compared with the two adult systems we studied, mouse hematopoiesis and hydra, which suggests that shared global regulation of processes related to, e.g., gene expression and metabolism, is more widespread in early development. As expected, our approach was unable to correctly predict the expression of genes that are highly lineage specific. An interesting question for future research will be to determine when and how developmental predictability breaks down. We anticipate that, in the future, transfer learning can be

combined with mechanistic modeling based on gene regulatory networks or inferred ligand-receptor interactions to improve predictions. Applications include inference of missing data points in longitudinal analysis and approximative identification of internal nodes in lineage trees.

Limitations of the study

Our approach requires a labeled differentiation process, i.e., the analysis is based on differentiation trajectories that are inferred by other methods. Furthermore, if used without ground truth information, it is difficult to evaluate how closely the predictions will match the real situation. In general, we recommend using our methodology in situations where a considerable amount of concordant transcriptional changes can be expected. As we have shown, our approach does not allow correct prediction of highly lineage-specific genes (Figures 3B and 3C). Furthermore, even if the predictions correctly capture changes in mean and standard deviation of gene expression, the approach may still not reproduce all aspects of the specific cell types (Figure S7). Hence, despite the observed high degree of predictability, especially in developmental differentiation, predictions obtained purely based on transfer learning should still be treated with caution.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Predicting single-cell transcriptomic data
  - The DCP algorithm
  - Variational Autoencoder
  - Normalizing flows
  - URD
  - Gene ontology enrichment analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2023.12.002.

ACKNOWLEDGMENTS

We acknowledge support by the MDC/BIMSB core facility for bioinformatics. Work in J.P.J.'s laboratory was funded by Helmholtz innovation and networking grant sparse2big and by Helmholtz AI grant SC-SLAM-ATAC.

AUTHOR CONTRIBUTIONS

B.S. and J.P.J. conceived the research plan with support from G.J. G.J. developed the computational model and performed analysis. B.S. and J.P.J. supervised the analysis. All authors wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 13, 2023
Revised: July 28, 2023
Accepted: December 6, 2023
Published: December 20, 2023

REFERENCES

1. Schier, A.F., and Talbot, W.S. (2005). Molecular genetics of axis formation in zebrafish. Annu. Rev. Genet. *39*, 561–613.
2. Sáez, M., Blassberg, R., Camacho-Aguilar, E., Siggia, E.D., Rand, D.A., and Briscoe, J. (2022). Statistically derived geometrical landscapes capture principles of decision-making dynamics during cell fate transitions. Cell Syst. *13*, 12–28.e3.
3. Janssens, J., Aibar, S., Taskiran, I.I., Ismail, J.N., Gomez, A.E., Aughey, G., Spanier, K.I., De Rop, F.V., González-Blas, C.B., Dionne, M., et al. (2022). Decoding gene regulation in the fly brain. Nature *601*, 630–636.
4. Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev. Cell *18*, 675–685.
5. Grün, D., Muraro, M.J., Boisset, J.C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. Cell Stem Cell *19*, 266–277.
6. Olsson, A., Venkatasubramanian, M., Chaudhri, V.K., Aronow, B.J., Salomonis, N., Singh, H., and Grimes, H.L. (2016). Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature *537*, 698–702.
7. Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science *360*. eaar3131–eaar3115.
8. Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science *360*, 981–987.
9. Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., and Klein, A.M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science *360*. eaar5780–eaar5717.
10. Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. Nature *566*, 490–495.
11. He, P., Williams, B.A., Trout, D., Marinov, G.K., Amrhein, H., Berghella, L., Goh, S.T., Plajzer-Frick, I., Afzal, V., Pennacchio, L.A., et al. (2020). The changing mouse embryo transcriptome at whole tissue and single-cell resolution. Nature *583*, 760–767.
12. Mittnenzweig, M., Mayshar, Y., Cheng, S., Ben-Yair, R., Hadas, R., Rais, Y., Chomsky, E., Reines, N., Uzonyi, A., Lumerman, L., et al. (2021). A single-embryo, single-cell time-resolved model for mouse gastrulation. Cell *184*, 2825–2842.e22.
13. Qiu, C., Cao, J., Martin, B.K., Li, T., Welsh, I.C., Srivatsan, S., Huang, X., Calderon, D., Noble, W.S., Disteche, C.M., et al. (2022). Systematic reconstruction of cellular trajectories across mouse embryogenesis. Nat. Genet. *54*, 328–341.
14. Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell *163*, 1663–1677.
15. Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human

haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281.

16. Giladi, A., Paul, F., Herzog, Y., Lubling, Y., Weiner, A., Yofe, I., Jaitin, D., Cabezas-Wallscheid, N., Dress, R., Ginhoux, F., et al. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. Nat. Cell Biol. *20*, 836–846.
17. Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. Nature *553*, 212–216.
18. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*, eaaw3381.
19. Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., et al. (2017). A single-cell survey of the small intestinal epithelium. Nature *551*, 333–339.
20. Gerber, T., Murawala, P., Knapp, D., Masselink, W., Schuez, M., Hermann, S., Gac-Santel, M., Nowoshilow, S., Kageyama, J., Khattak, S., et al. (2018). Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. Science *362*, eaaq0681.
21. Huguet, G., Magruder, D.S., Tong, A., Fasina, O., Kuchroo, M., Wolf, G., and Krishnaswamy, S. (2022). Manifold interpolating optimal-transport flows for trajectory inference. Adv. Neural Inf. Process. Syst. *35*, 29705–29718.
22. Tong, A., Huang, J., Wolf, G., van Dijk, D., and Krishnaswamy, S. (2020). TrajectoryNet: a dynamic optimal transport network for modeling cellular dynamics. Proc. Mach. Learn. Res. *119*, 9526–9536.
23. Peng, M., Li, Y., Wamsley, B., Wei, Y., and Roeder, K. (2021). Integration and transfer learning of single-cell transcriptomes via cFIT. Proc. Natl. Acad. Sci. USA *118*, e2024383118.
24. Kingma, D.P., and Welling, M. (2022). Auto-encoding variational Bayes. Preprint at arXiv.
25. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721.
26. Amodio, M., van Dijk, D., Montgomery, R., Wolf, G., and Krishnaswamy, S. (2019). Out-of-sample extrapolation with neuron editing. Preprint at arXiv.
27. Lotfollahi, M., Naghipourfar, M., Theis, F.J., and Wolf, F.A. (2020). Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics *36* (*Suppl 2*), i610–i617.
28. Yeo, G.H.T., Saksena, S.D., and Gifford, D.K. (2021). Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. Nat. Commun. *12*, 3222.
29. Rezende, D.J., and Mohamed, S. (2016). Variational inference with normalizing flows. Preprint at arXiv.
30. Siebert, S., Farrell, J.A., Cazet, J.F., Abeykoon, Y., Primack, A.S., Schnitzler, C.E., and Juliano, C.E. (2019). Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. Science *365*, eaav9314.
31. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.
32. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: a review for statisticians. J. Am. Stat. Assoc. *112*, 859–877.
33. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.
34. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (Association for Computational Linguistics), pp. 10–21.
35. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., and Winther, O. (2016). Ladder variational autoencoders. Preprint at arXiv.
36. Zhao, S., Song, J., and Ermon, S. (2018). InfoVAE: information maximizing variational autoencoders. Preprint at arXiv.
37. Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. J. Mach. Learn. Res. *22*, 1–49.
38. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. (2020). gprofiler2–an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Res. *9*, ELIXIR-709.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
zebrafish embryogenesis scRNA-seq data	Farrell et al.[7] (GEO accession GSE106587)	https://portals.broadinstitute.org/single_cell/study/single-cell-reconstruction-of-developmental-trajectories-during-zebrafish-embryogenesis
mouse hematopoiesis dataset	Weinreb et al.[18] (GEO accession GSE140802)	https://github.com/AllonKleinLab/paper-data/tree/master/Lineage_tracing_on_transcriptional_landscapes_links_state_to_fate_during_differentiation
mouse gastrulation single cell data	Pijuan-Sala et al.[10] (ArrayExpress accession E-MTAB-6967)	https://github.com/MarioniLab/EmbryoTimecourse2018
hydra stem cell differentiation single cell dataset	Siebert et al.[30] (GEO accession GSE121617)	https://github.com/cejuliano/hydra_single_cell
Software and Algorithms		
Deep Cell Predictor	This paper https://github.com/02infi/DCP	https://doi.org/10.5281/zenodo.10116010
Pytorch	https://pytorch.org/	https://doi.org/10.5281/zenodo.10116010
Scanpy	Wolf et al.[31]	https://scanpy.readthedocs.io/en/stable/
URD	Farrell et al.[7]	https://github.com/farrellja/URD
Python version 3.10	Python Software Foundation	https://www.python.org/downloads/release/python-3100/
Seaborn 0.11.2	https://seaborn.pydata.org/archive/0.11/index.html	N/A
SciPy version 1.8.1	https://pypi.org/project/scipy/1.8.1/	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jan Philipp Junker (janphilipp.junker@mdc-berlin.de).

Materials availability

This study did not generate new materials.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We downloaded the publicly available URD object (https://singlecell.broadinstitute.org/single_cell/study/SCP162/single-cell-reconstruction-of-developmental-trajectories-during-zebrafish-embryogenesis) of zebrafish embryogenesis scRNA-seq data consisting of 38,731 cells across 12 developmental time points from 3.3–12 hours post fertilization.[7] Here, we use the URD tree information to split the data into training and test data for the DCP model.

The mouse hematopoiesis dataset from Weinreb et al.[18] was obtained from https://github.com/AllonKleinLab/paper-data/tree/master/Lineage_tracing_on_transcriptional_landscapes_links_state_to_fate_during_differentiation. The scRNA-seq dataset shows an in vitro differentiation time course of hematopoietic progenitor cells to nine mature cell types: Erythrocytes (Er), megakaryocytes (Mk), basophils (Ba), mastcells (Ma), eosinophils (Eos), neutrophils (Neu), monocytes (Mo), dendritic cells (plasmocytoid pDC;

migratory migDC) and lymphoid precursors (Ly). The dataset was sampled at three time points during culture (days 2, 4 and 6). After filtering the cells, we randomly downsampled all cell types to a maximum of 5000 cells for each timepoint.

On this dataset, we applied URD to illustrate the transition from an early progenitor state to specified lineages across a pseudotime axis. To construct the URD tree, we identified the undifferentiated cells from day 2 as the root cells for the development tree and specified cell types at day 6 as tips. URD uses the R package *destiny* to build the kNN graph from the transcriptomic distance over all cells. URD incorporates transition probabilities computed over the kNN graph to calculate the diffusion pseudotime ordering of cells. For each cell type cluster (tips) at day 6, trajectories are identified by simulating random walks that are biased toward transitioning to cells younger or equal in pseudotime starting from each tip to the root cells backwards in developmental time. To build the URD tree, trajectories are joined at the point where they contain cells that are reached from multiple tips. We validated our URD tree by identifying the expression of marker genes at the root, intermediate and tips level which reflects the progenitors, intermediate and mature cells states. For our calculations, we focused on the monocyte and neutrophil lineages in the URD tree. We used pseudotime to split the data into undifferentiated (pseudotime <= 0.1) and differentiated cell states (pseudotime > 0.3) and used those for training and inference.

The mouse gastrulation single cell data from Pijuan-Sala et al.[10] were obtained from https://github.com/MarioniLab/EmbryoTimecourse2018. We excluded all the blood and extraembryonic cells from the dataset. We randomly downsampled the remaining data to a maximum of 3000 cells for each cell type. We then ran URD on this dataset to reconstruct the transcriptomic tree during mouse gastrulation. We considered epiblast from day 6.5 as the root cells and specified cell types in day 8.5 as tips for the transcriptomic tree. For our calculations, we focused on the ectoderm and mesoderm lineages in the reconstructed URD tree. We considered cells before day 7.5 as the progenitors and cells at day 8.5 as mature cells, which we then further used as training and test data.

The hydra stem cell differentiation single cell dataset[10] was obtained from https://github.com/cejuliano/hydra_single_cell. The endoderm and ectoderm URD transcriptomic tree was already calculated and further classified into differentiated (tips of URD tree) and undifferentiated cells (pseudotime <0.25) in the original publication. We used this dataset as the input to DCP for training and inference.

METHOD DETAILS

Predicting single-cell transcriptomic data

In this paper, we compare three different methods to predict single-cell transcriptomic data: a linear model, scGen[25] and the method developed in this paper, DeepCellPredictor (DCP). Datasets are preprocessed and normalized using scanpy,[31] with the normalization being performed by dividing each observation by the median of the total counts.

In the linear model, we use vector arithmetic on the gene expression data from a training dataset to calculate a vector δ by taking the difference between the average expression $\underline{X}$ of cells at timepoints T_1 and T_2 ($\delta_X = \underline{X_{T_2}} - \underline{X_{T_1}}$). To predict the transcriptomic cell states of the test dataset at time point T_2, we add the calculated vector δ_X to the gene expression of the test cells at time point T_1. Additionally, we regularize negative counts in the gene expression data by setting them to 0.

We further apply scGen, a deep learning perturbation model based on Variational Autoencoders (VAEs) and vector arithmetic in latent space. scGen trains the VAE to obtain the latent space representation of training and test data. Then, scGen calculates a latent vector by taking the difference between the average latent representation $\underline{Z}$ of cells in the training dataset at timepoints T_1 and T_2 ($\delta_T = \underline{Z_{T_2}} - \underline{Z_{T_1}}$). To predict the transcriptomic cell states at time point T_2, scGen first applies the calculated vectorδ_T to the latent representation of cells in the test data at time point T_1. The transformed latent representation is then decoded back into the gene expression space to obtain the predicted cell states at time point T_2.

Finally, DeepCellPredictor (DCP) is a transfer learning framework that combines an extension of Variational Autoencoders with normalizing flows. The next section contains a detailed explanation. Intuitively, DCP can be understood as a model that performs a mean shift, using vector arithmetic, and a shape modification, using normalizing flows, of a distribution of single cell profiles. These operations are learnt from training data and are performed in latent space, and we use an extension of Variational Autoencoders to map single cell profiles from gene space to latent space and back. The extension of Variational Autoencoders, negative binomial-based maximum mean discrepancy Variational Autoencoders (*nb*-mmd-VAE), is made using two modifications: by requiring a single cell-appropriate distribution, the negative binomial, as target distribution for the decoder, and by using maximum mean divergence (MMD) in the loss function to maximize the mutual information between latent space distribution and the data.

Training DeepCellPredictor consists of three steps. First, we train the nb-mmd variational autoencoder to find the lower dimensional representation Z_i of the gene features of single cell data X_i from the training dataset. As a second step, we estimate the latent time vector δ_T by calculating the difference between the average latent representation $\underline{Z}$ of cells at time point T_1 and T_2 ($\delta_T = \underline{Z_{T_2}} - \underline{Z_{T_1}}$). The time vector captures the changes in mean gene expression between the two timepoints and is used to mean shift the latent representation of cells at time point T_1 by adding the time vector δ_T to the latent representation of cells at time point T_1 from the training data ($Z_{T_2}{}^{\delta} = \delta_T + Z_{T_1}$).

The last training step is to learn the transformation from $P_{T_2}{}^{\delta}(Z)$ to $P_{T_2}(Z)$,the latent representations of $Z_2{}^{\delta}$ and Z_{T_2}. We estimate the probability density functions $P_{T_2}{}^{\delta}(Z)$ and $P_{T_2}(Z)$ by fitting a Gaussian kernel, and then use these as the input to the planar flows. The planar flows learn the transport map T_λ in the latent space that maps $P_{T_2}{}^{\delta}(Z)$ to $P_{T_2}(Z)$.

To predict the transcriptome cell states at time point T_2 from cells at time point T_1 in the test dataset, we first transform the latent representation of cells at time point T_1 by adding the learned time vector δ_T and then apply the transport map T_λ on the transformed representations. Finally, the decoder network maps the predicted latent representations to gene expression space to obtain the predicted transcriptomic cell states.

The DCP algorithm

The DCP algorithm consists of two components: negative binomial-based maximum mean discrepancy Variational Autoencoders, an extension of Variational Autoencoders, and normalizing flows. In the below sections, we describe both separately.

Variational Autoencoder

For completeness, we here first describe Variational Autoencoders,[24] adapting the description found in the literature (Blei et al.[32] and https://jaan.io/what-is-variational-autoencoder-vae-tutorial/), and then discuss the extension we have developed for this paper.

A Variational Autoencoder (VAE) consists of two key components: an encoder network with parameters φ and decoder network with parameters θ. The encoder network maps input data X to a latent space distribution $Q(Z|X,\varphi)$, and the decoder network maps latent variables Z to a distribution $P(X|Z,\theta)$. For our application, the samples $\{X_i,....,X_N\}$ are N single cell transcriptomes, and $\{Z_i,....,Z_N\}$ are their latent space representations.

A good latent space representation Z_i is one that the decoder maps to the given transcriptomic profiles X_i: the distribution $P(Z|X,\theta)$. Unfortunately, this distribution cannot be calculated directly. According to Bayes' rule, we can write the posterior $P(Z|X,\theta) = P(X|Z,\theta) \times P(Z|\theta)/P(X|\theta)$. The distribution $P(X)$ can be calculated by integrating over the latent variables $P(X) = \int P(X|Z)P(Z)dZ$, but this computation requires exponential time. Instead, VAEs approximate the real posterior $P(Z|X,\theta)$ with a family of distributions $Q(Z|X,\varphi)$.

A good approximation is characterized by a low Kullback-Leibler divergence (KL) between $Q(Z|X,\varphi)$ and $P(Z|X,\theta)$. The KL divergence can be written as:

$$KL(Q(Z|X,\varphi)||P(Z|X,\theta)) = logP(X) + E_Q[log\ Q(Z|X,\varphi)] - E_Q[log\ P(X,Z)],$$

where $E_Q[.]$ denotes the expectation over posterior distributions Q. Since $P(X)$ does not depend on φ and θ, we can minimize the KL divergence by maximizing $ELBO = E_Q[log\ P(X,Z)] - E_Q[log\ Q(Z|X,\varphi)]$, the Evidence Lower Bound.

Since the *ELBO* can be calculated as:

$$ELBO = E_Q[log\ P(X|Z,\theta)] - KL(Q(Z|X,\varphi)\ ||\ P(Z,\theta)),$$

we can train a VAE by minimizing

$$Loss(\theta,\varphi) = -\ ELBO = -E_Q[log\ P(X|Z,\theta)] + KL(Q(Z|X,\varphi)\ ||\ P(Z,\theta)),$$

summed over all the data points X_i. Here, the first term is the expected log-likelihood of the data or reconstruction loss and the second term is Kullback-Leibler divergence between the approximate posterior $Q(Z|X)$and the prior $P(Z)$.

Our adaptation of these general VAEs to single cell data consists of two separate steps. First, single cell transcriptomic can be described well by a negative binomial distribution. We therefore let the decoder output the parameters μ (mean expression of a gene in a cell) and ζ (dispersion of the gene expression over the cells) of the negative binomial distribution. At first, the decoder outputs the mean proportion $\rho_{c,g}$of transcripts expressed across all genes using softmax activation function at the last layer.[33] Then we multiply the mean proportions $\rho_{c,g}$ with library size to generate cell counts $\mu = library\ size \times \rho_{c,g}$. The dispersion parameter ζ for each gene across all the cells are considered constant and optimized during training.

Second, it has been found that the KL divergence term in VAE loss function is quite restrictive[34,35] and can lead to uninformative latent representations. Also, KL regularization is not strong enough compared to the reconstruction term and tends to overfit the data. To overcome these limitations, we use maximum mean discrepancy (MMD)[36] as regularization term instead of Kullback-Leibler (KL) divergence. MMD calculates the difference between the moments of two distributions. Unlike KL divergence, MMD maximizes the mutual information between the latent code and data.

In summary, we propose the *nb*-mmd Variational Autoencoder (*nb*-mmdVAE) which combines variance-based reconstruction loss and MMD as regularizer. The overall loss function of the *nb*-mmdVAE is:

$$Loss_i(\theta,\varphi) = -E_Q[log\ P(X|Z,\theta)] + MMD(Q(Z|X,\varphi)\ ||\ P(Z,\theta)),$$

where $P(X_i|Z_i)$ is modeled with negative binomial distribution $NB(X;\mu,\zeta)$.

Normalizing flows

Cell differentiation occurs when specific genes are upregulated or downregulated, leading to complex changes in distribution of gene expression patterns that are captured by single cell data. Linear methods can capture the shift in the mean, but are unable to capture changes in the variance between the two cell states. To achieve this, we implemented a transformation function based on planar flows, applied to latent space representations obtained from training a VAE and adding the latent time vector δ_T as discussed above.

Planar flows are a specialized case of normalizing flows,[29,37] neural networks that learn a reversible transformation between an initial distribution $P(X)$ and a target distribution $\underline{P}(Y)$. This reversible transformation $T : X \rightarrow Y$ is composed of K smooth and invertible functions $T^i_{\lambda_i} : Y_{i-1} \rightarrow Y_i$ (where $X = Y_0$ and $Y = Y_K$) that are parametrized by $\lambda = \{\lambda_1, ..., \lambda_K\}$:

$$T_\lambda = T^K_{\lambda_K} \circ T^{K-1}_{\lambda_{K-1}} \circ ... \circ T^1_{\lambda_1}.$$

A trained planar flow should minimize the KL divergence between its output distribution $P(Y;\lambda)$ and the target distribution $\underline{P}(Y)$:

$$Loss(\lambda) = KL(P(Y;\lambda) \,\|\underline{P}(Y)) = E_{P(Y;\lambda)}[log\, P(Y;\lambda) - log\, \underline{P}(Y)].$$

A change of variables introduces the determinant of the Jacobian. We use the composition of T_λ to calculate that specifically,

$$log\, P(Y;\lambda) = log\left(P(Y_{K-1};\lambda)\left|det \frac{\partial T^K_{\lambda_K}(Y_K)}{\partial Y_K}\right|^{-1}\right) = log\, P(Y_{K-1};\lambda) - log\left|det \frac{\partial T^K_{\lambda_K}(Y_K)}{\partial Y_K}\right| Y_{K-1};\lambda) - log\left|det \frac{\partial T^K_{\lambda_K}(Y_K)}{\partial Y_K}\right|$$

$$= log\, P(Y_{K-2};\lambda) - log\left|det \frac{\partial T^K_{\lambda_K}(Y_K)}{\partial Y_K}\right| - log\left|det \frac{\partial T^{K-1}_{\lambda_{K-1}}(Y_{K-1})}{\partial Y_{K-1}}\right|$$

$$= log\, P(X) - \sum_{i=1}^{K} log\left|det \frac{\partial T^i_{\lambda_i}(Y_i)}{\partial Y_i}\right|.$$

With this, the loss function becomes

$$Loss(\lambda) = E_{P(X)}\left[log\, P(X) - \sum_{i=1}^{K} log\left|det \frac{\partial T^i_{\lambda_i}(Y_i)}{\partial Y_i}\right| - log\underline{P}(T_\lambda(X))\right]$$

$$Loss(\lambda) = E_{P(X)}\left[- log\underline{P}(T_\lambda(X)) - \sum_{i=1}^{K} log\left|det \frac{\partial T^i_{\lambda_i}(Y_i)}{\partial Y_i}\right|\right]$$

after changing variables from Y to X. Normalizing flows are trained by sampling from the initial distribution $P(X)$ and calculating the loss through the transformation T and the determinant of the Jacobian.

We use a specific type of normalizing flow, called planar flow, which compresses and expands densities around a hyperplane. For planar flows, the parameter λ is specified as $\{W \epsilon R^d, U \epsilon R^d, B \epsilon R\}$ and T_λ takes the following form:

$$T(X) = X + U.H(W^T X + B)\; H(X) = tanh(X).$$

The absolute value of the determinant of the Jacobian can be calculated using the matrix determinant lemma $det(I + uv^T) = (1 + v^T u)$:

$$\left|det \frac{\partial T_\lambda(X)}{\partial X}\right| = det\left(I + UH'(W^T X + B)W^T\right)$$

$$= 1 + H'(W^T X + B)U^T W.$$

Here, we use $k = 32$ which represents the number of flows used to transform the initial distribution $P_X(X)$ to the target distribution $P_Y(Y)$.

URD

URD is a simulated diffusion based computational tool to reconstruct the developmental trajectories of differentiation processes in biological systems.[7] It takes single cell RNA sequencing data as the input and provides a reconstructed transcriptomic tree in the form of a 2D dendrogram that reveals cellular and temporal dynamics.

Gene ontology enrichment analysis

To identify which biological processes underlie the observed predictability of embryonic differentiation, we performed a GO analysis of genes that change in a concordant manner in ectoderm and mesoderm/endoderm between 5.3 and 12 hpf. We only considered

genes that change in the same direction across time points with a log2fold change > 1 or < -1 (Dataset S1). The GO analysis was performed using the R package *gProfiler2*[38] by querying concordant genes between lineages in zebrafish development and mouse hematopoiesis. The result of this analysis is shown in Dataset S2.

QUANTIFICATION AND STATISTICAL ANALYSIS

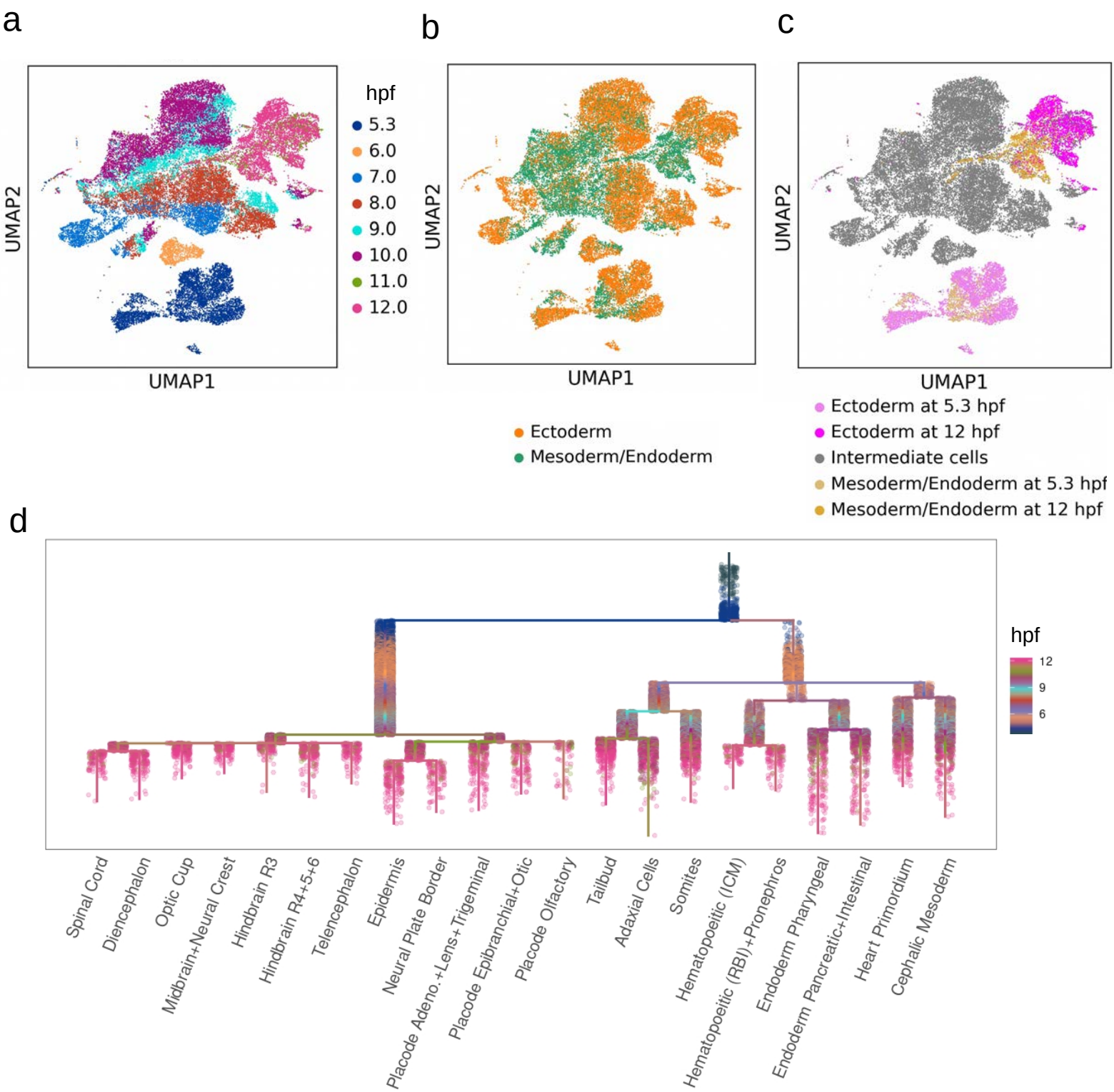
For statistical analysis of our predictions of single cell gene expression profiles, we repeated the analysis and sampled the data points 50 times, and we calculated the Pearson correlation between the real and predicted data for mean gene expression and standard deviation. The error bars shown in Figures 3C, S4A, and S4B were calculated by considering one standard deviation of uncertainty from the mean. We used SciPy and seaborn to perform correlation analysis and error calculation.

Cell Systems, Volume *15*

Supplemental information

Inference of differentiation trajectories by transfer learning across biological processes

Gaurav Jumde, Bastiaan Spanjaard, and Jan Philipp Junker



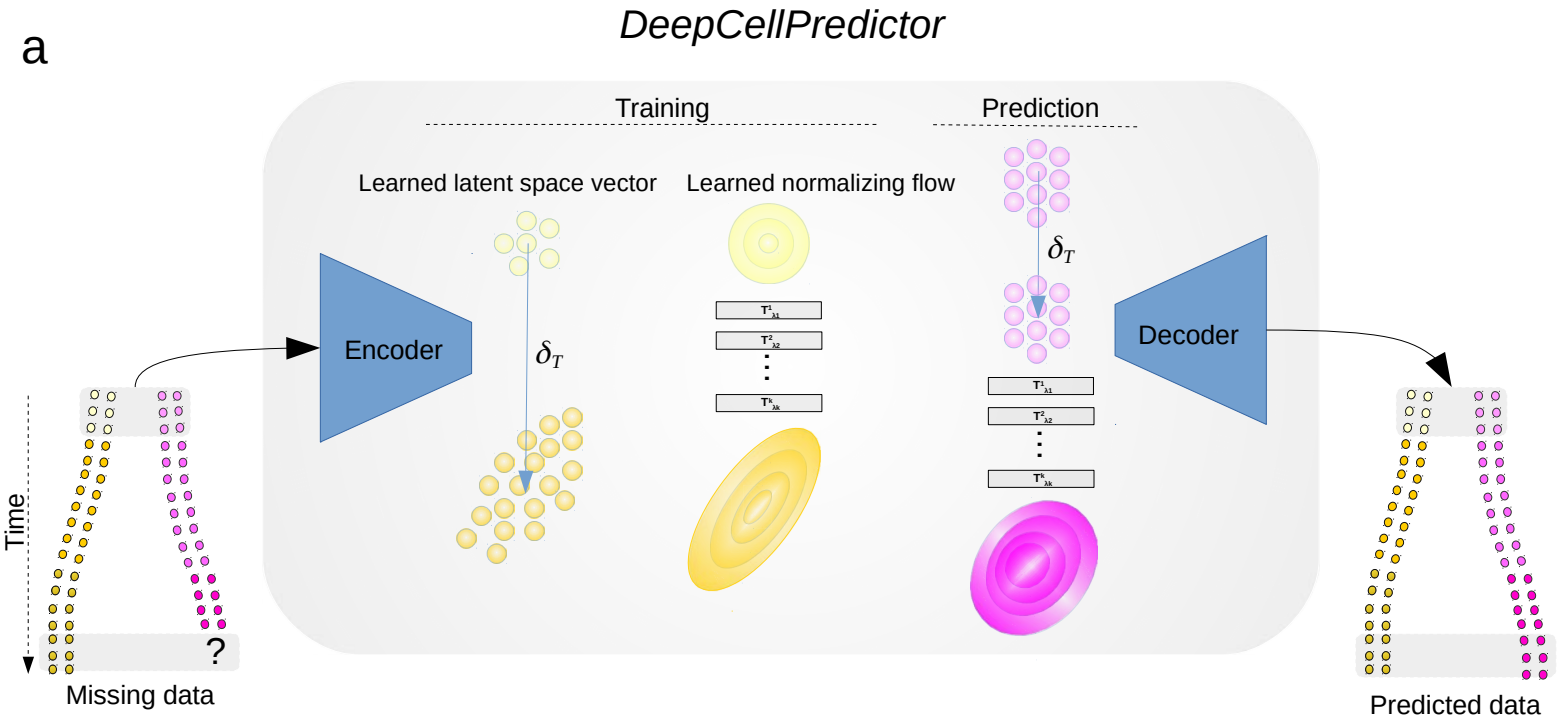
Supplementary Figure 1 | Developmental zebrafish dataset (related to Fig. 1).

a-c) UMAP plot of the developmental zebrafish dataset, indicating hours post fertilization (hpf) **(a)**, germ layer **(b),** and progenitor and mature cells at 5.3hpf and 12hpf of Ectoderm and Mesoderm/Endoderm germ layers **(c)**.

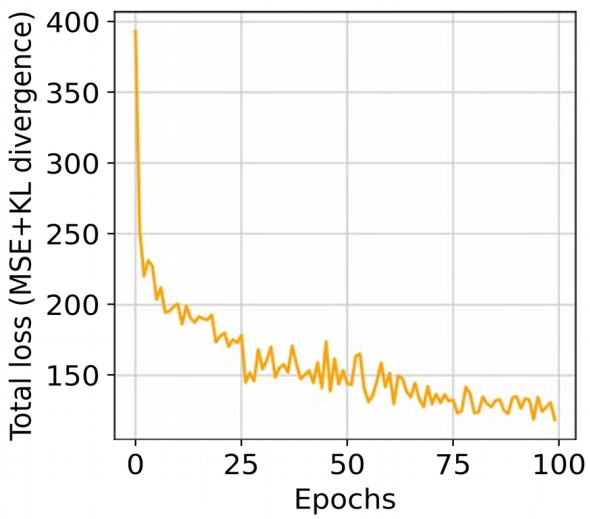
d) URD of the same dataset, indicating transcriptome-based inferred lineage splits. URD is the diffusion-based computational trajectory reconstruction method that was used in the original publication to analyze this dataset. Reproduced from original publication[7].

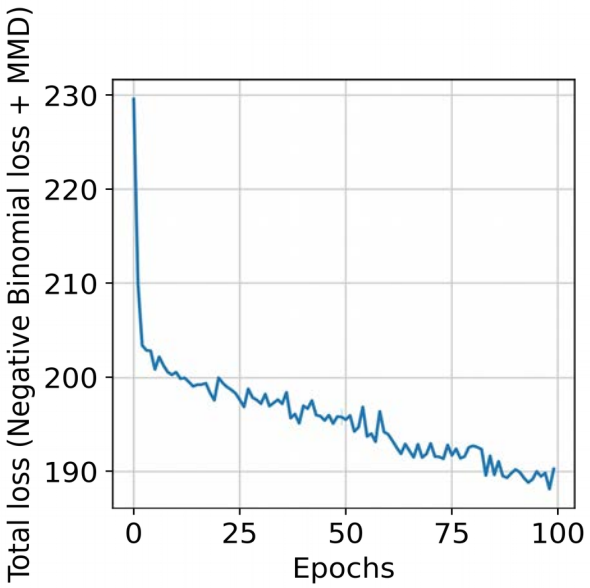
Data from Farrell et al., 2018[7], GEO accession GSE106587.

a

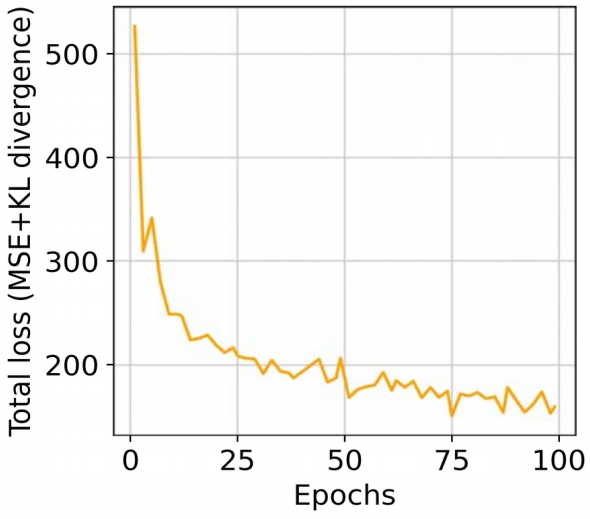

DeepCellPredictor
Training
Prediction
Learned latent space vector
Learned normalizing flow
Encoder
Decoder
δ_T
δ_T
Time
?
Missing data
Predicted data

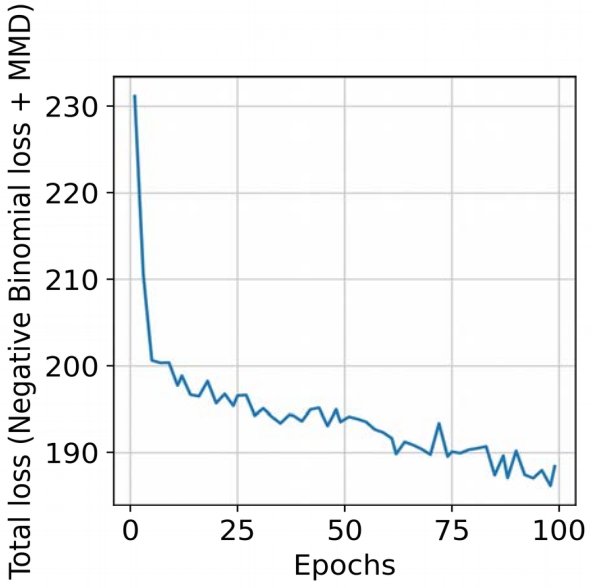
b


Total loss (MSE+KL divergence)
400
350
300
250
200
150
0
25
50
75
100
Epochs


Total loss (Negative Binomial loss + MMD)
230
220
210
200
190
0
25
50
75
100
Epochs

c


Total loss (MSE+KL divergence)
500
400
300
200
0
25
50
75
100
Epochs

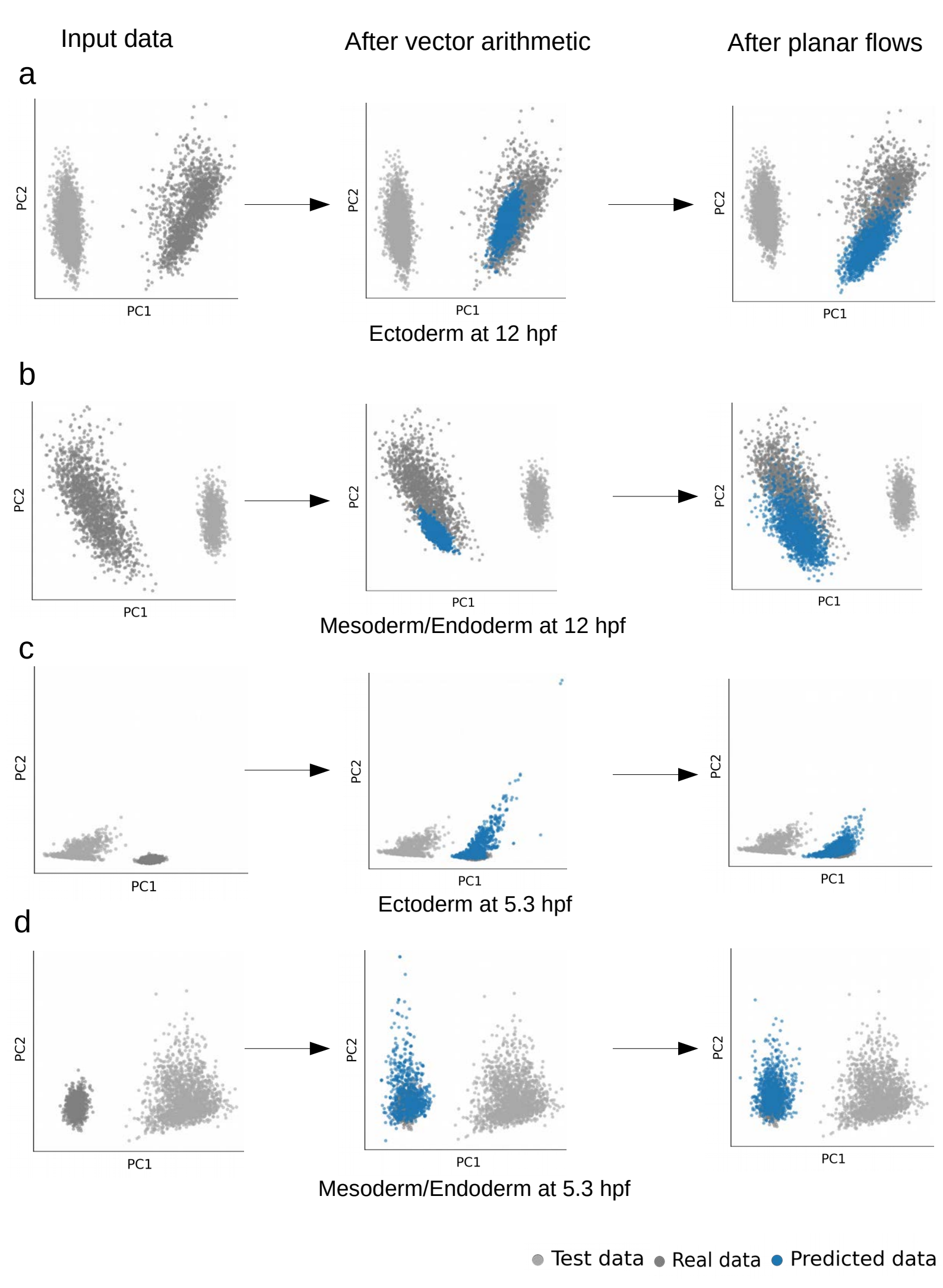

Total loss (Negative Binomial loss + MMD)
230
220
210
200
190
0
25
50
75
100
Epochs

Supplementary Figure 2 | Training convergence of neural network models (related to Fig. 1).

a) Schematic of the DeepCellPredictor (DCP) model.

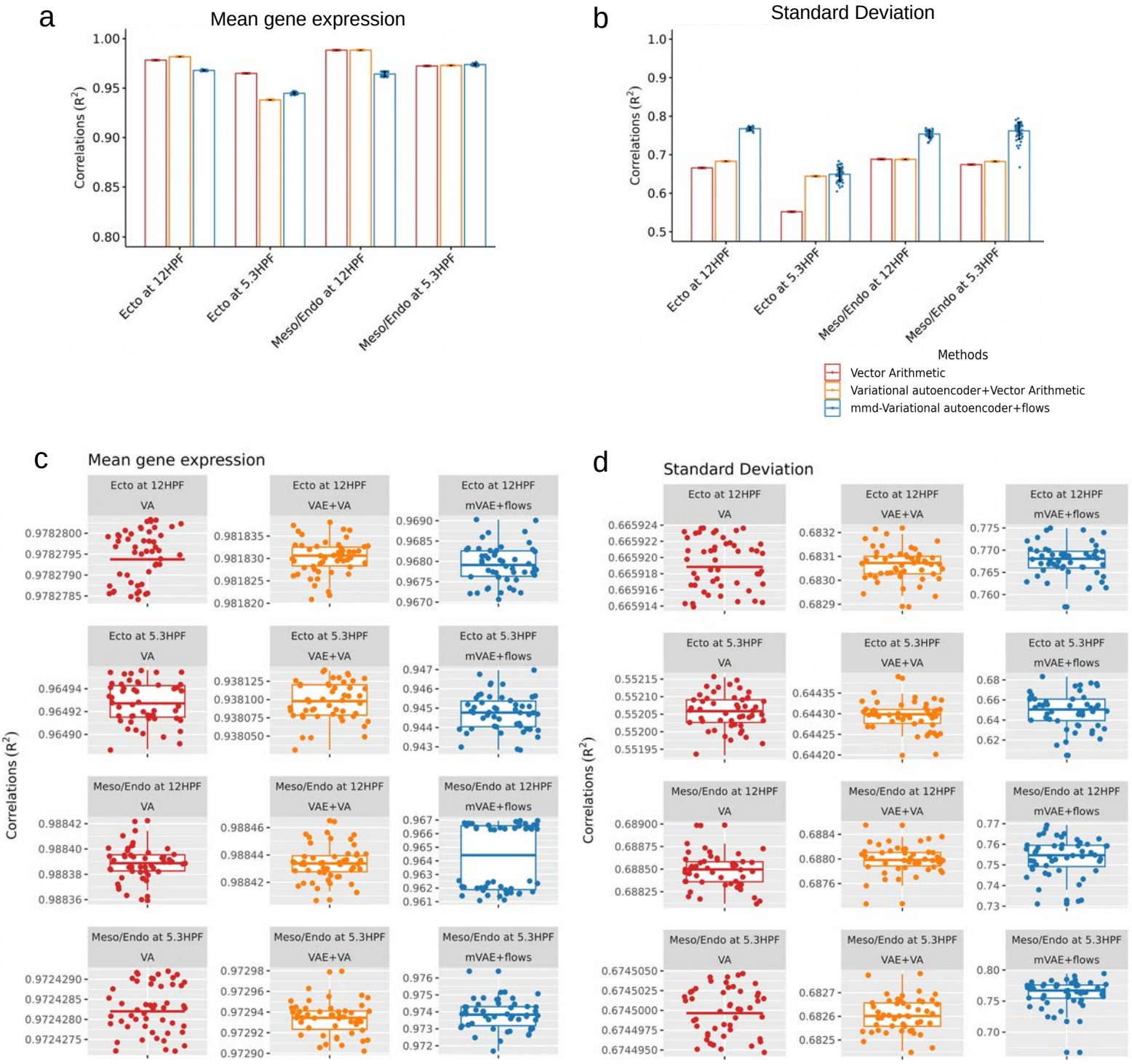
b-c) Total loss over epochs for regular variational autoencoder (left) and DCP model (right) trained on mesoderm/endoderm **(b)** and ectoderm data **(c).**

Data from Farrell et al., 2018[7], GEO accession GSE106587.


Input data
After vector arithmetic
After planar flows
a
PC2
PC1
Ectoderm at 12 hpf
b
Mesoderm/Endoderm at 12 hpf
c
Ectoderm at 5.3 hpf
d
Mesoderm/Endoderm at 5.3 hpf
Test data
Real data
Predicted data

Supplementary Figure 3 | Disentangling the effects of vector arithmetic and normalizing flows in DCP (related to Fig. 1).

To visualize the information transferred, we decoded latent space predictions after vector arithmetic and after normalizing flows and performed PCA for visualization. The predicted distribution of cells already starts resembling the target distribution after vector arithmetic due to our regularization approach, in contrast to the vector arithmetic performed in scGen (Fig. 1c). The normalizing flows allow us to further approximate the target distribution.
Data from Farrell et al., 2018[7], GEO accession GSE106587.

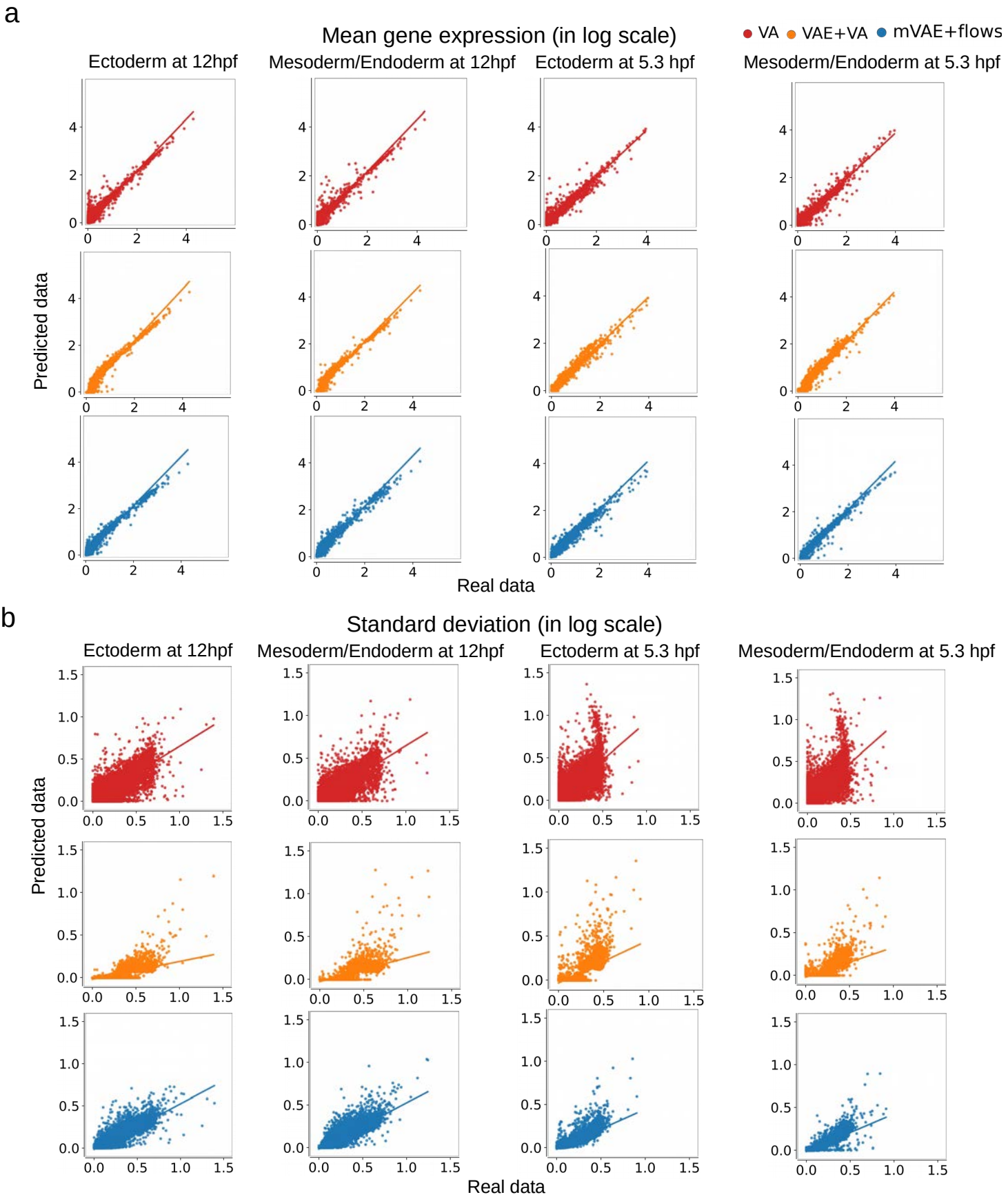


Supplementary Figure 4 | Comparison of prediction algorithms (vector arithmetic (VA), variational autoencoders with vector arithmetic (VAE+VA), and mmd-variational autoencoders with flows (mVAE+flows)) in the developing zebrafish (related to Fig. 2).

a-b) Correlation between real and predicted mean expression and expression variability. Error bars were determined by considering one standard deviation of uncertainty from the mean after 50 repeats.

c-d) Zoom-in box plots highlighting variance between repeats.

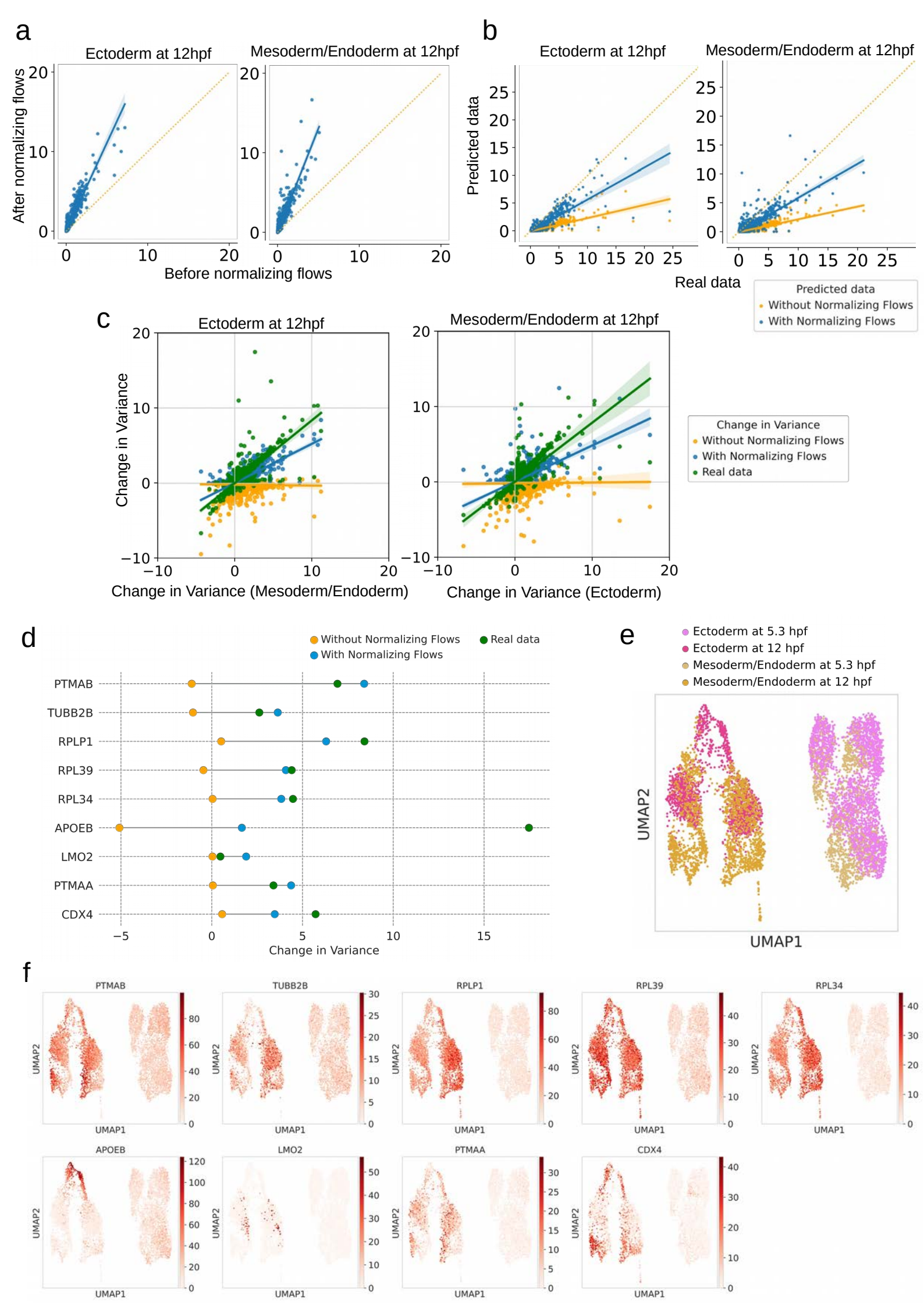
Data from Farrell et al., 2018[7], GEO accession GSE106587.



Supplementary Figure 5 | Comparison of prediction algorithms (vector arithmetic (VA), variational autoencoders with vector arithmetic (VAE+VA), and mmd-variational autoencoders with flows (mVAE+flows)) in the developing zebrafish (related to Fig. 2).

a-b) Scatter plots of predicted mean gene expression and expression variability across all genes in log(x + 1) scale with prediction algorithms in the developing zebrafish.

Data from Farrell et al., 2018[7], GEO accession GSE106587.


a
Ectoderm at 12hpf
Mesoderm/Endoderm at 12hpf
After normalizing flows
Before normalizing flows
b
Ectoderm at 12hpf
Mesoderm/Endoderm at 12hpf
Predicted data
Real data
Predicted data
Without Normalizing Flows
With Normalizing Flows
c
Ectoderm at 12hpf
Mesoderm/Endoderm at 12hpf
Change in Variance
Change in Variance (Mesoderm/Endoderm)
Change in Variance (Ectoderm)
Change in Variance
Without Normalizing Flows
With Normalizing Flows
Real data
d
Without Normalizing Flows
With Normalizing Flows
Real data
PTMAB
TUBB2B
RPLP1
RPL39
RPL34
APOEB
LMO2
PTMAA
CDX4
Change in Variance
e
Ectoderm at 5.3 hpf
Ectoderm at 12 hpf
Mesoderm/Endoderm at 5.3 hpf
Mesoderm/Endoderm at 12 hpf
UMAP2
UMAP1
f
PTMAB
TUBB2B
RPLP1
RPL39
RPL34
APOEB
LMO2
PTMAA
CDX4
UMAP2
UMAP1

Supplementary Figure 6 | Normalizing flows increase predictive value at high temporal gain of gene variance (related to Fig. 2).

a) Comparison of predicted gene expression variance with and without normalizing flows.

b) Comparison of predicted variance with and without normalizing flows to real variance.

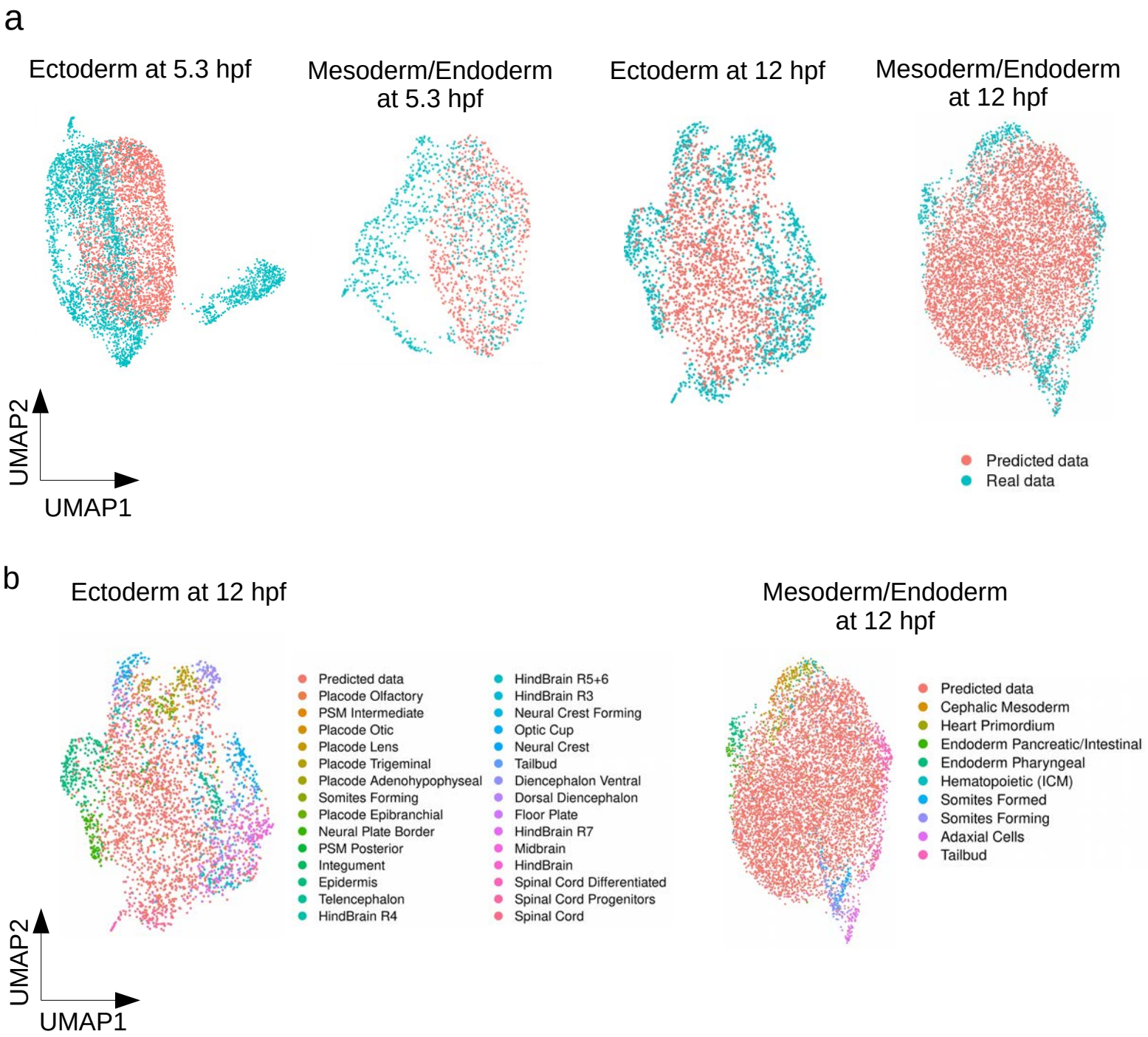
c) Change of expression variance in real and predicted data compared to the training data (in log scale). **a-c)**: predicting to 12hpf ectoderm (left) and 12hp mesoderm/endoderm (right).

d) Change of expression variance in real and predicted data compared to the training data for selected genes with high variance gain (in log scale).

e) UMAP of test and training data.

f) Expression of genes selected in **d)** on the UMAP shown in **e)**.

Data from Farrell et al., 2018[7], GEO accession GSE106587.

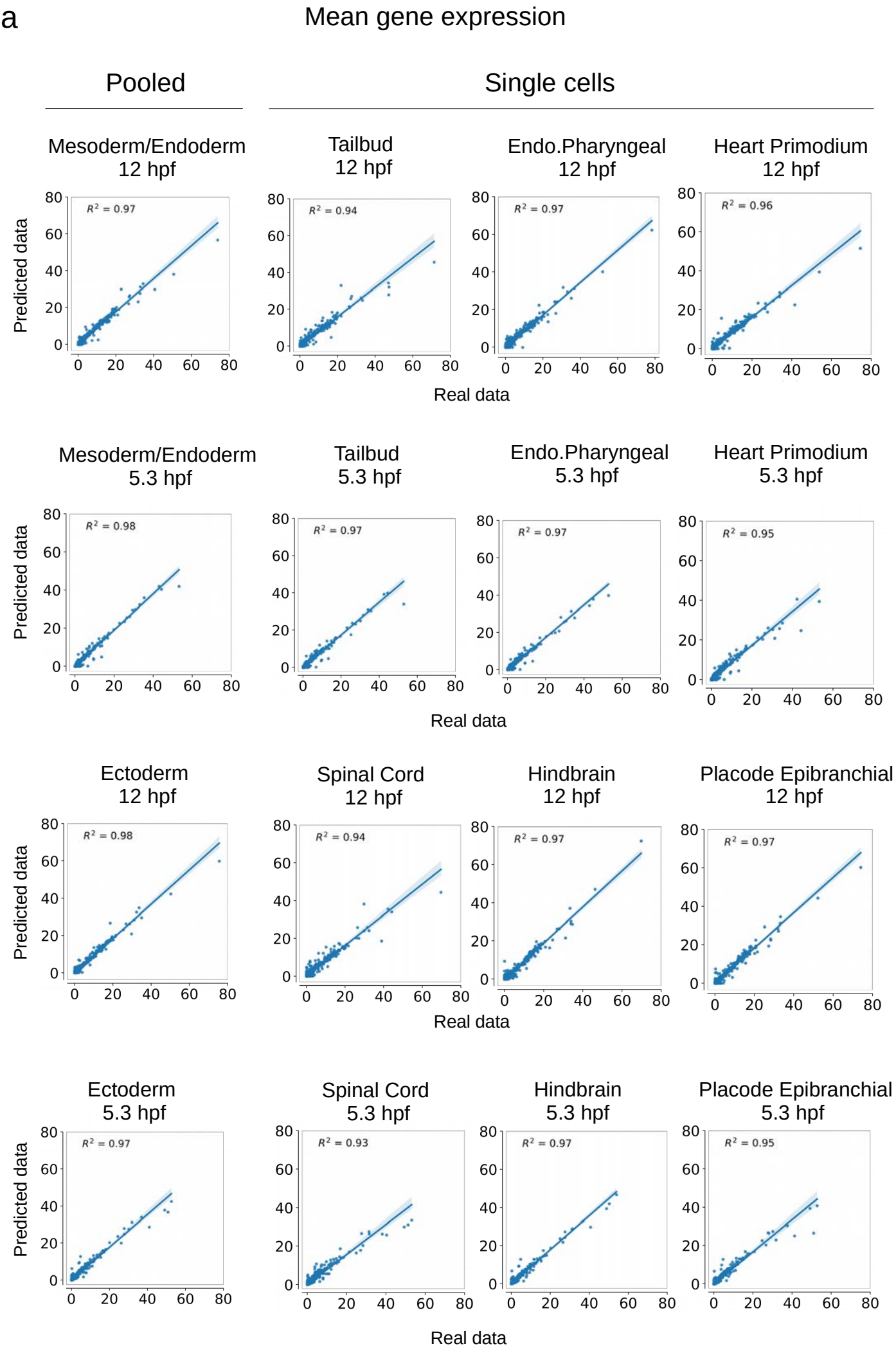


Supplementary Figure 7 | DCP generates realistic single-cell transcriptomes but does not accurately predict cell type clusters at 12 hpf (related to Fig. 2).

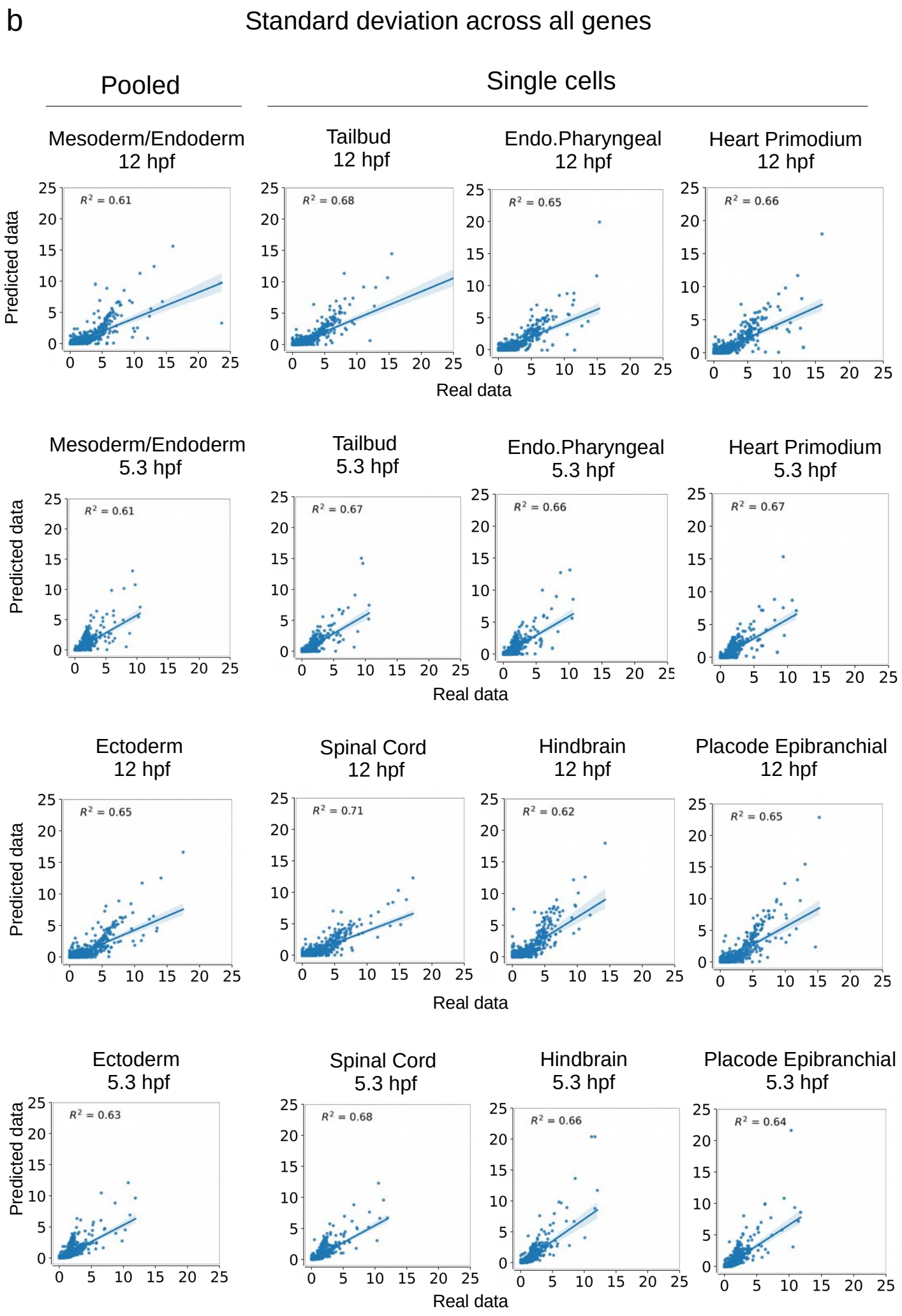
a-b) UMAP of integrated real and predicted data of zebrafish ectoderm and mesendoderm at 5.3 hpf and 12 hpf.

Data from Farrell et al., 2018[7], GEO accession GSE106587.

a

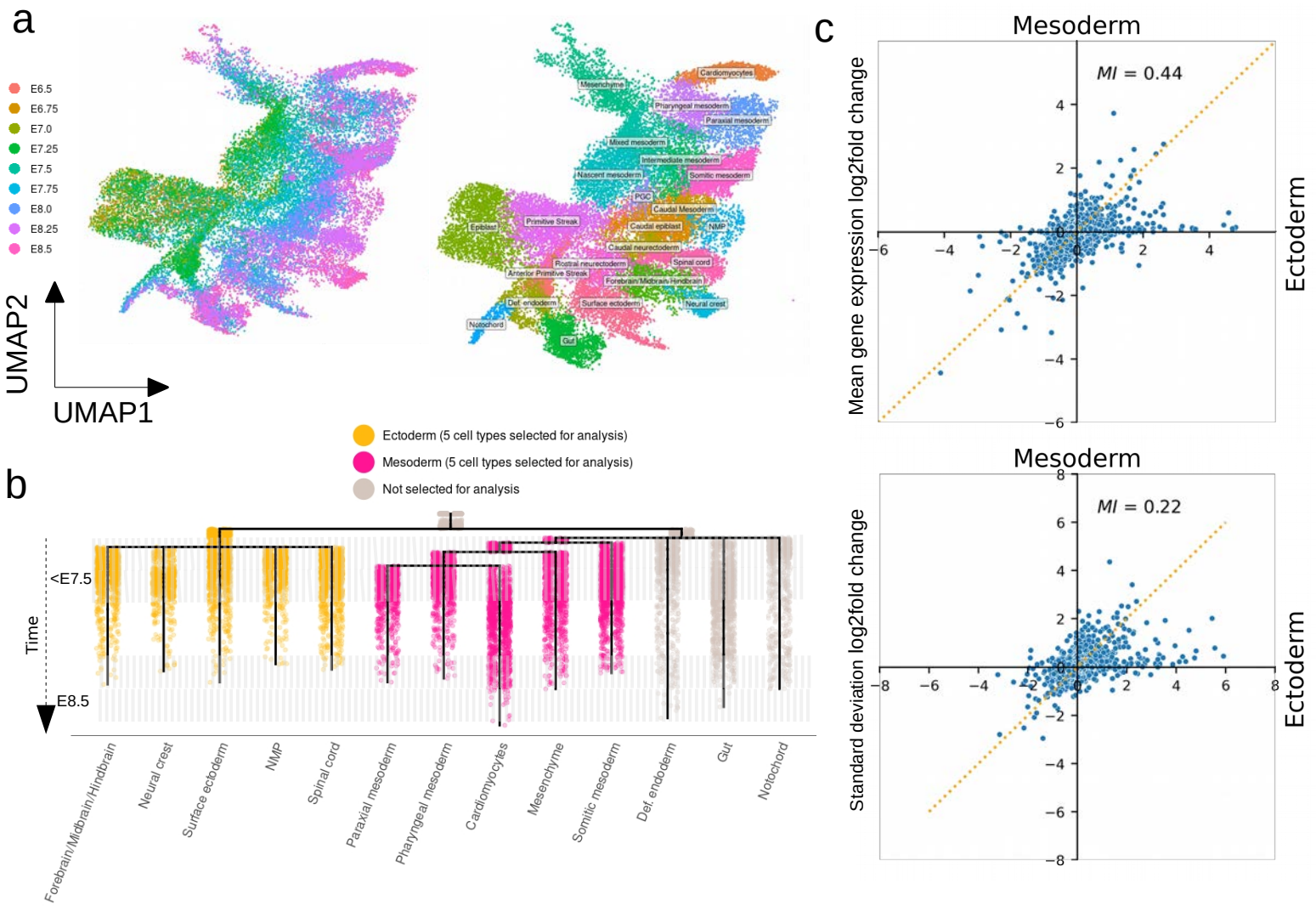

Mean gene expression
Pooled
Single cells
Mesoderm/Endoderm 12 hpf
Tailbud 12 hpf
Endo.Pharyngeal 12 hpf
Heart Primodium 12 hpf
Mesoderm/Endoderm 5.3 hpf
Tailbud 5.3 hpf
Endo.Pharyngeal 5.3 hpf
Heart Primodium 5.3 hpf
Ectoderm 12 hpf
Spinal Cord 12 hpf
Hindbrain 12 hpf
Placode Epibranchial 12 hpf
Ectoderm 5.3 hpf
Spinal Cord 5.3 hpf
Hindbrain 5.3 hpf
Placode Epibranchial 5.3 hpf
Predicted data
Real data
R² = 0.97
R² = 0.94
R² = 0.97
R² = 0.96
R² = 0.98
R² = 0.97
R² = 0.97
R² = 0.95
R² = 0.98
R² = 0.94
R² = 0.97
R² = 0.97
R² = 0.97
R² = 0.93
R² = 0.97
R² = 0.95
0
20
40
60
80

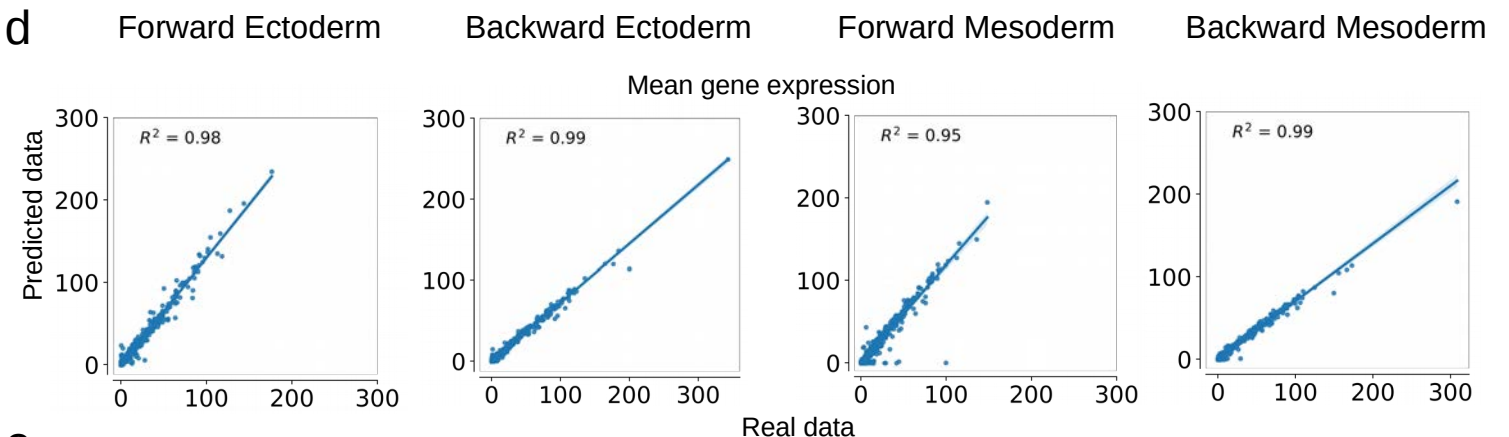
b

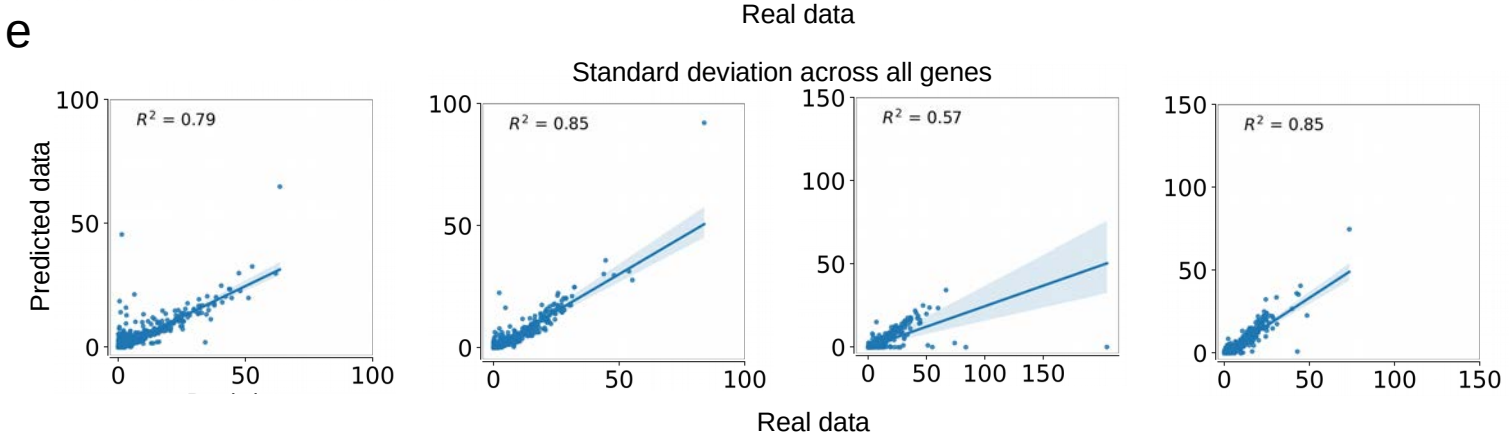

Standard deviation across all genes
Pooled
Single cells
Mesoderm/Endoderm 12 hpf
Tailbud 12 hpf
Endo.Pharyngeal 12 hpf
Heart Primodium 12 hpf
Mesoderm/Endoderm 5.3 hpf
Tailbud 5.3 hpf
Endo.Pharyngeal 5.3 hpf
Heart Primodium 5.3 hpf
Ectoderm 12 hpf
Spinal Cord 12 hpf
Hindbrain 12 hpf
Placode Epibranchial 12 hpf
Ectoderm 5.3 hpf
Spinal Cord 5.3 hpf
Hindbrain 5.3 hpf
Placode Epibranchial 5.3 hpf
Predicted data
Real data
R² = 0.61
R² = 0.68
R² = 0.65
R² = 0.66
R² = 0.61
R² = 0.67
R² = 0.66
R² = 0.67
R² = 0.65
R² = 0.71
R² = 0.62
R² = 0.65
R² = 0.63
R² = 0.68
R² = 0.66
R² = 0.64

Supplementary Figure 8 | DCP prediction of pooled and single cell types in zebrafish development (related to Fig. 2).

a-b) Mean gene expression and variability were estimated for single cell types and compared to pooled cell types. For this analysis we upsampled the number of cells for three ectodermal cell types (spinal cord, hindbrain, placode epibranchial) and three mesendodermal cell types (tailbud, endo.pharyngeal, heart primordium) to 500 cells. We trained and tested each set of paired ectoderm-mesendoderm cells in both pooled and single type dataset format.
Data from Farrell et al., 2018[7], GEO accession GSE106587.

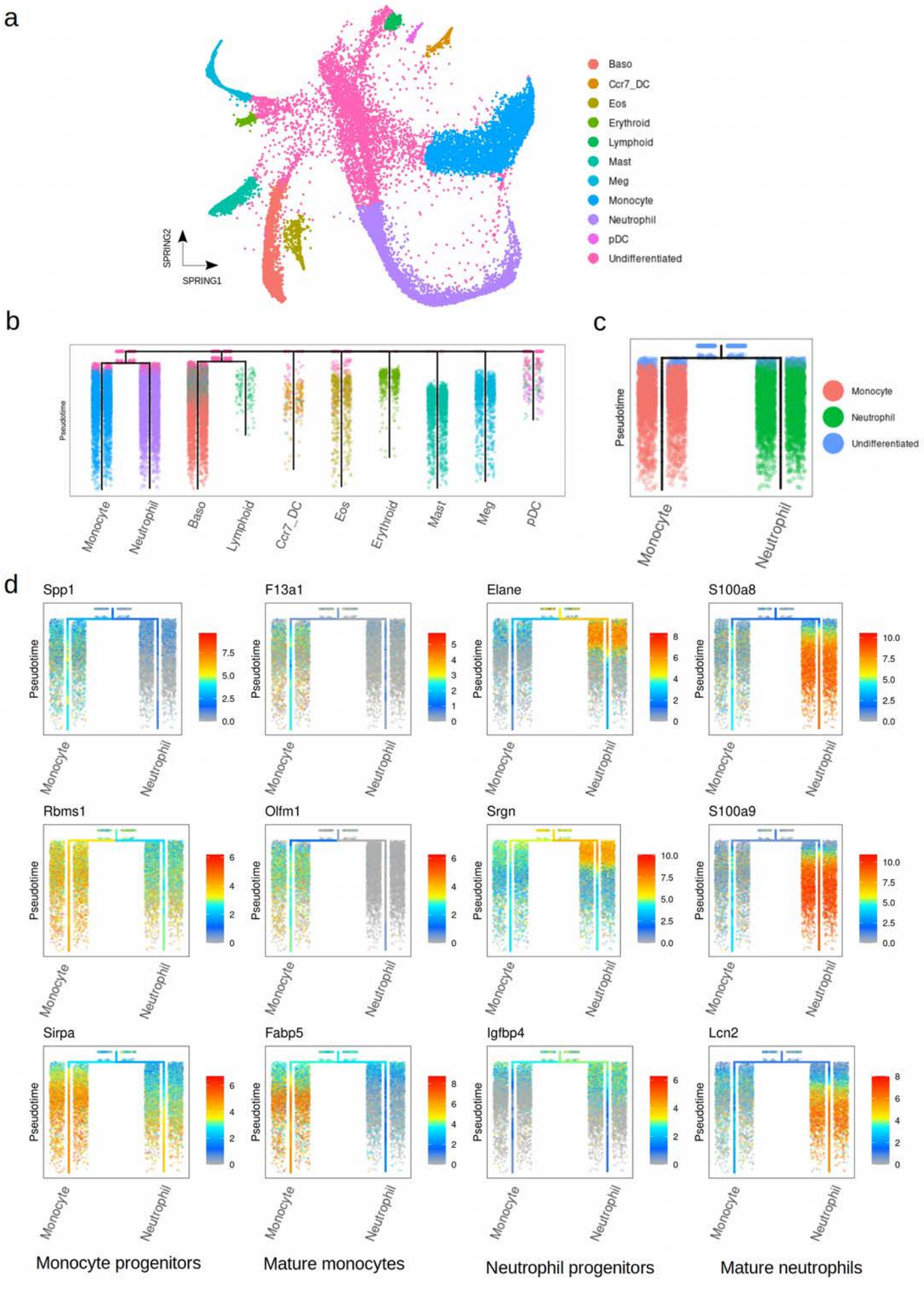

a
E6.5
E6.75
E7.0
E7.25
E7.5
E7.75
E8.0
E8.25
E8.5
UMAP2
UMAP1
Cardiomyocytes
Mesenchyme
Pharyngeal mesoderm
Paraxial mesoderm
Mixed mesoderm
Intermediate mesoderm
Nascent mesoderm
Somitic mesoderm
PGC
Caudal Mesoderm
Epiblast
Primitive Streak
Caudal epiblast
NMP
Caudal neurectoderm
Rostral neurectoderm
Spinal cord
Anterior Primitive Streak
Forebrain/Midbrain/Hindbrain
Def. endoderm
Surface ectoderm
Neural crest
Notochord
Gut
b
Ectoderm (5 cell types selected for analysis)
Mesoderm (5 cell types selected for analysis)
Not selected for analysis
Time
<E7.5
E8.5
Forebrain/Midbrain/Hindbrain
Neural crest
Surface ectoderm
NMP
Spinal cord
Paraxial mesoderm
Pharyngeal mesoderm
Cardiomyocytes
Mesenchyme
Somitic mesoderm
Def. endoderm
Gut
Notochord
c
Mesoderm
MI = 0.44
Mean gene expression log2fold change
Ectoderm
Mesoderm
MI = 0.22
Standard deviation log2fold change
Ectoderm


d
Forward Ectoderm
Backward Ectoderm
Forward Mesoderm
Backward Mesoderm
Mean gene expression
Predicted data
R² = 0.98
R² = 0.99
R² = 0.95
R² = 0.99
Real data


e
Standard deviation across all genes
Predicted data
R² = 0.79
R² = 0.85
R² = 0.57
R² = 0.85
Real data

Supplementary Figure 9 | Mouse gastrulation dataset (related to Fig. 2).

a) UMAP representation of the mouse gastrulation dataset, indicating embryonic day (E). Reproduced from original publication after excluding blood and extraembryonic cells[10].
b) Transcriptome-based tree of same dataset constructed using URD. Similar to our analysis of the zebrafish development dataset, we selected five ectodermal and five mesodermal cell types for the analysis (forward and backward predictions).
c) Mutual information of log2 fold changes between ectoderm and mesoderm for mean and standard deviation of gene expression.
d) Correlation between real and predicted data for mean gene expression.
e) Correlation between real and predicted data for standard deviation of gene expression.
Data from Pijuan-Sala et al., 2019[10] (ArrayExpress accession E-MTAB-6967).


a
SPRING2
SPRING1
Baso
Ccr7_DC
Eos
Erythroid
Lymphoid
Mast
Meg
Monocyte
Neutrophil
pDC
Undifferentiated
b
Pseudotime
Monocyte
Neutrophil
Baso
Lymphoid
Ccr7_DC
Eos
Erythroid
Mast
Meg
pDC
c
Pseudotime
Monocyte
Neutrophil
Undifferentiated
d
Spp1
F13a1
Elane
S100a8
Rbms1
Olfm1
Srgn
S100a9
Sirpa
Fabp5
Igfbp4
Lcn2
Monocyte progenitors
Mature monocytes
Neutrophil progenitors
Mature neutrophils

Supplementary Figure 10 | Mouse hematopoiesis dataset (related to Fig. 3).

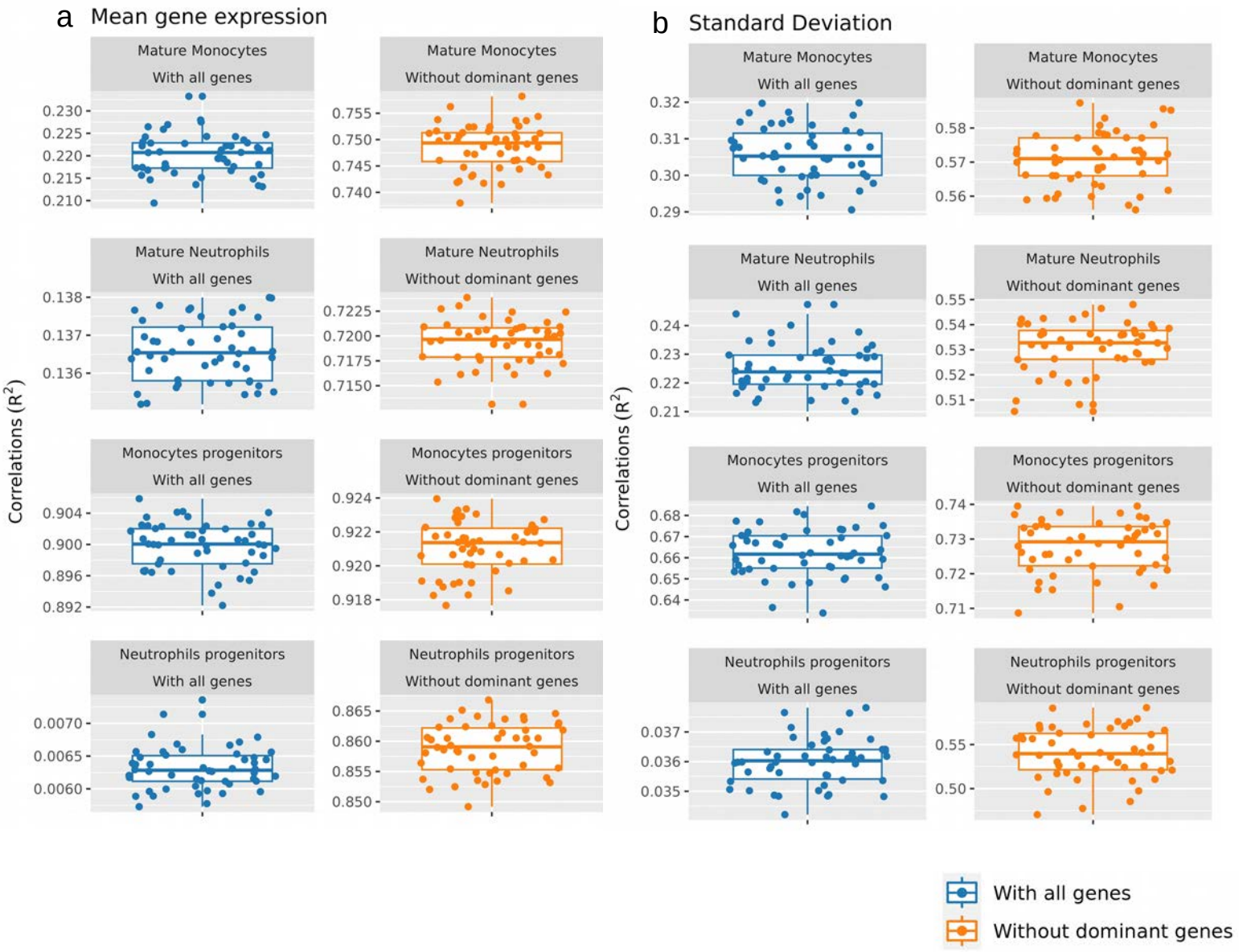
a) t-SNE of dataset.

b) URD of dataset showing branching lineage structure.

c) Subset used in our study: monocytes and neutrophils in a continuum between progenitor and mature states.

d) Representation of selected marker genes to validate that the URD analysis correctly separates mature and progenitor states of monocytes and neutrophils. Color scale represents gene expression in log scale.

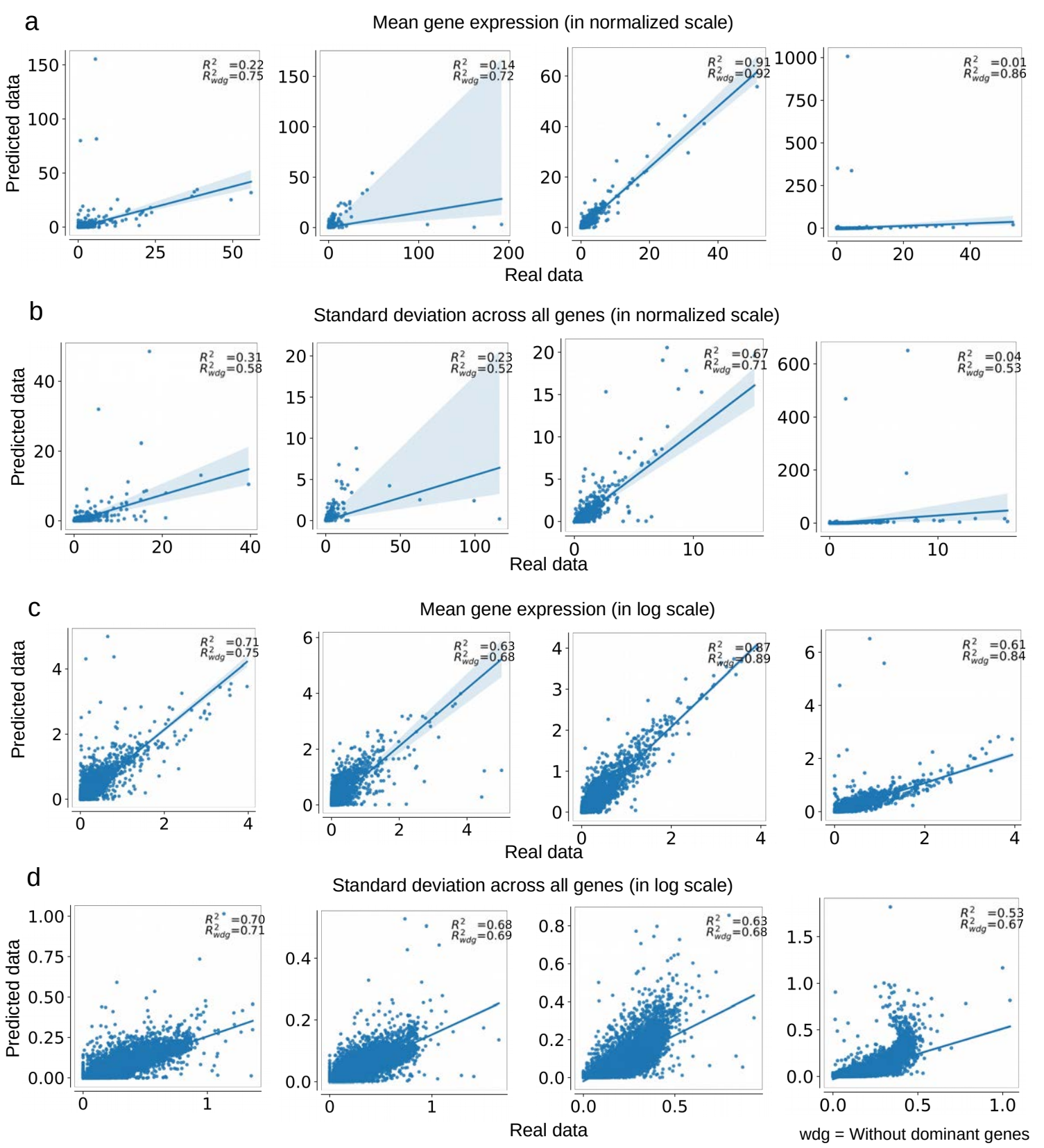
Data from Weinreb et al., 2020[18] (GEO accession GSE140802).



Supplementary Figure 11 | Uncertainty in DCP prediction of mature and progenitor cell states of monocyte and neutrophil cell lineages of mouse hematopiesis (related to Fig. 3).

a-b) Box plots representing distribution of correlation values of predicted mean gene expression and expression variability with all genes and without dominant genes.

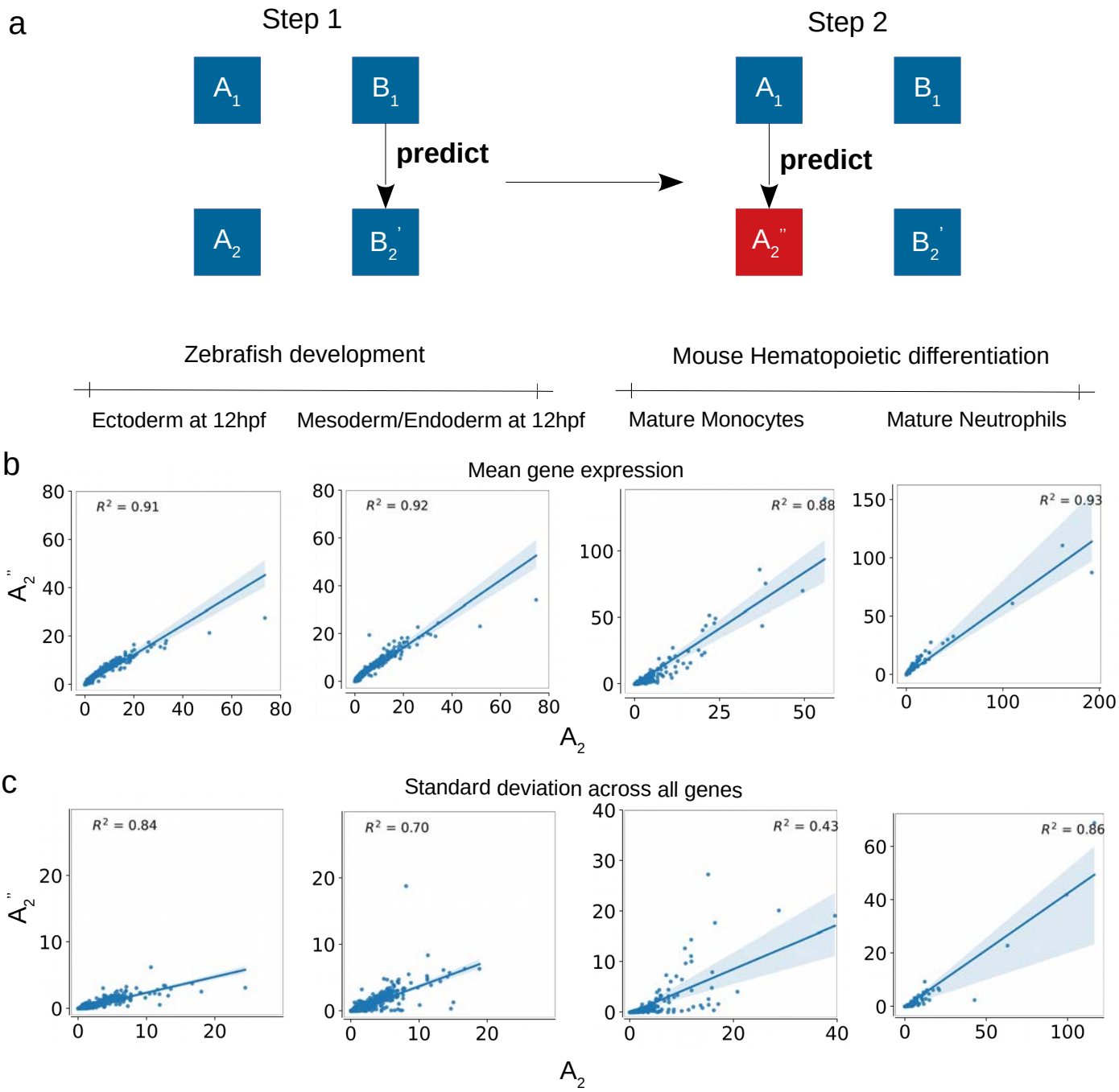
Data from Weinreb et al., 2020[18] (GEO accession GSE140802).


Mature Monocytes
Mature Neutrophils
Monocytes Progenitors
Neutrophils Progenitors
a
Mean gene expression (in normalized scale)
Predicted data
R^2 =0.22
R^2_{wdg}=0.75
R^2 =0.14
R^2_{wdg}=0.72
R^2 =0.91
R^2_{wdg}=0.92
R^2 =0.01
R^2_{wdg}=0.86
Real data
b
Standard deviation across all genes (in normalized scale)
Predicted data
R^2 =0.31
R^2_{wdg}=0.58
R^2 =0.23
R^2_{wdg}=0.52
R^2 =0.67
R^2_{wdg}=0.71
R^2 =0.04
R^2_{wdg}=0.53
Real data
c
Mean gene expression (in log scale)
Predicted data
R^2 =0.71
R^2_{wdg}=0.75
R^2 =0.63
R^2_{wdg}=0.68
R^2 =0.87
R^2_{wdg}=0.89
R^2 =0.61
R^2_{wdg}=0.84
Real data
d
Standard deviation across all genes (in log scale)
Predicted data
R^2 =0.70
R^2_{wdg}=0.71
R^2 =0.68
R^2_{wdg}=0.69
R^2 =0.63
R^2_{wdg}=0.68
R^2 =0.53
R^2_{wdg}=0.67
Real data
wdg = Without dominant genes

Supplementary Figure 12 | DCP prediction of mature and progenitor cell states of monocyte and neutrophil cell lineages of mouse hematopoiesis (related to Fig. 3).

a-d) Scatter plots of predicted mean gene expression and expression variability across all genes, **a,b)** normalized scale**, c,d)** log(x + 1) scale.

R^2 and R^2_{wdg} are correlation coefficients calculated with all genes and without dominant genes, respectively. Dominant genes are colored black, other genes are colored blue. Fit lines are lineage regressions with zero intercept. Data from Weinreb et al., 2020[18] (GEO accession GSE140802).

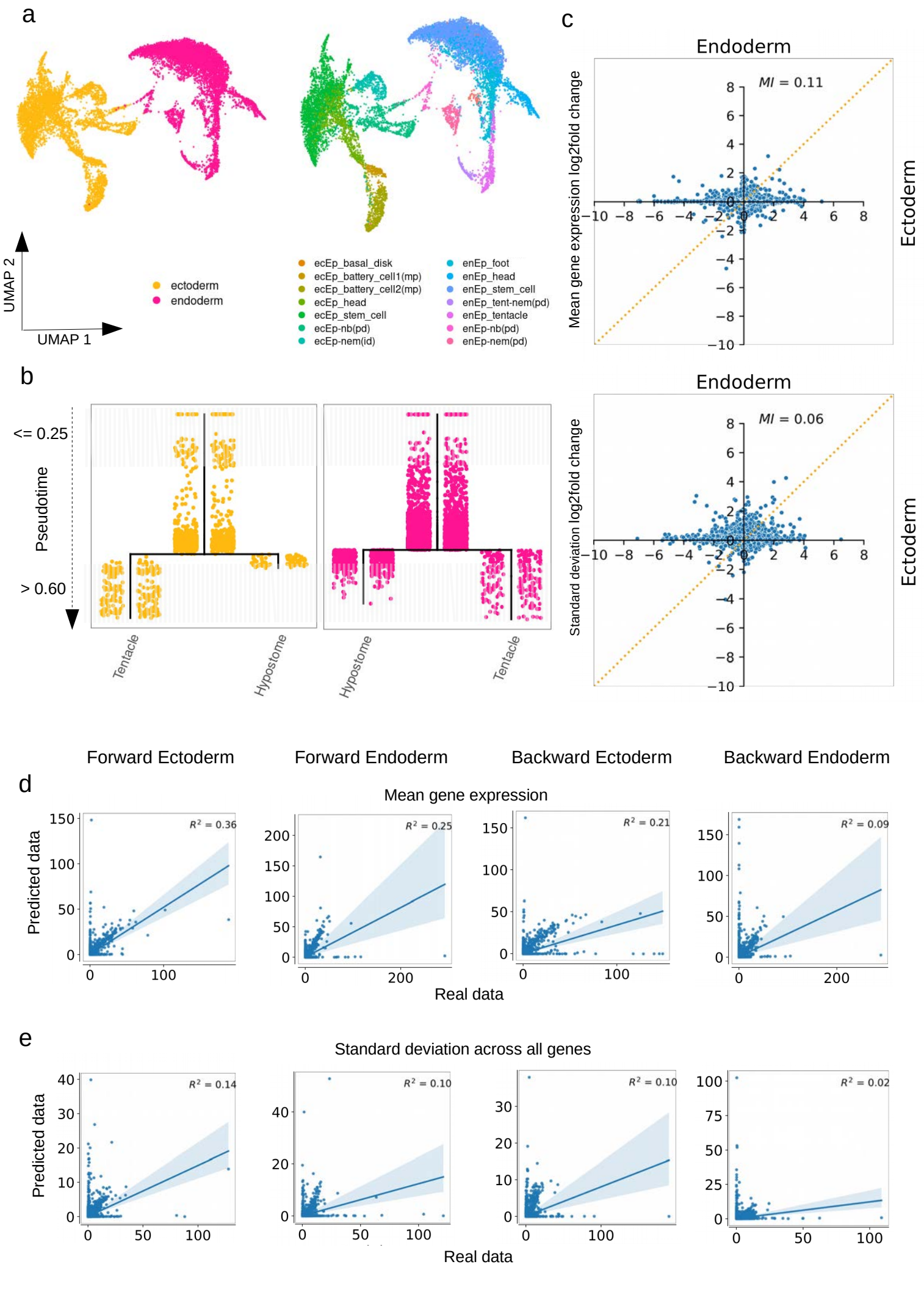


Supplementary Figure 13 | Successive predictions assess the information content of the latent space (related to Fig. 3).

a) To generate a double prediction, we first predict B2' from A1, A3 and B1. We then generate a new prediction for A2, A2'', from A1, B2 and B2'.

b-c) Zebrafish (left) and mouse hematopoiesis (right) comparison of double predictions with ground truth on mean gene expression **(b)** and expression standard deviation **(c)**.

Data from Farrell et al., 2018[7], GEO accession GSE106587 and Weinreb et al., 2020[18] (GEO accession GSE140802).


a
UMAP 2
UMAP 1
ectoderm
endoderm
ecEp_basal_disk
ecEp_battery_cell1(mp)
ecEp_battery_cell2(mp)
ecEp_head
ecEp_stem_cell
ecEp-nb(pd)
ecEp-nem(id)
enEp_foot
enEp_head
enEp_stem_cell
enEp_tent-nem(pd)
enEp_tentacle
enEp-nb(pd)
enEp-nem(pd)
c
Endoderm
MI = 0.11
Mean gene expression log2fold change
Ectoderm
b
<= 0.25
Pseudotime
> 0.60
Tentacle
Hypostome
Hypostome
Tentacle
Endoderm
MI = 0.06
Standard deviation log2fold change
Ectoderm
Forward Ectoderm
Forward Endoderm
Backward Ectoderm
Backward Endoderm
d
Mean gene expression
Predicted data
R² = 0.36
R² = 0.25
R² = 0.21
R² = 0.09
Real data
e
Standard deviation across all genes
Predicted data
R² = 0.14
R² = 0.10
R² = 0.10
R² = 0.02
Real data

Supplementary Figure 14 | Hydra stem cell differentiation dataset (related to Fig. 3).

a) UMAP representation of ectodermal and endodermal cells from the Hydra dataset30.

b) Transcriptome-based tree of same dataset constructed using URD.

c) Mutual information of log2 fold changes between ectoderm and endoderm for mean and standard deviation of gene expression.

d) Correlation between real and predicted data for mean gene expression.

e) Correlation between real and predicted data for standard deviation of gene expression.

Data from Siebert et al., 2019[30] (GEO accession GSE121617).