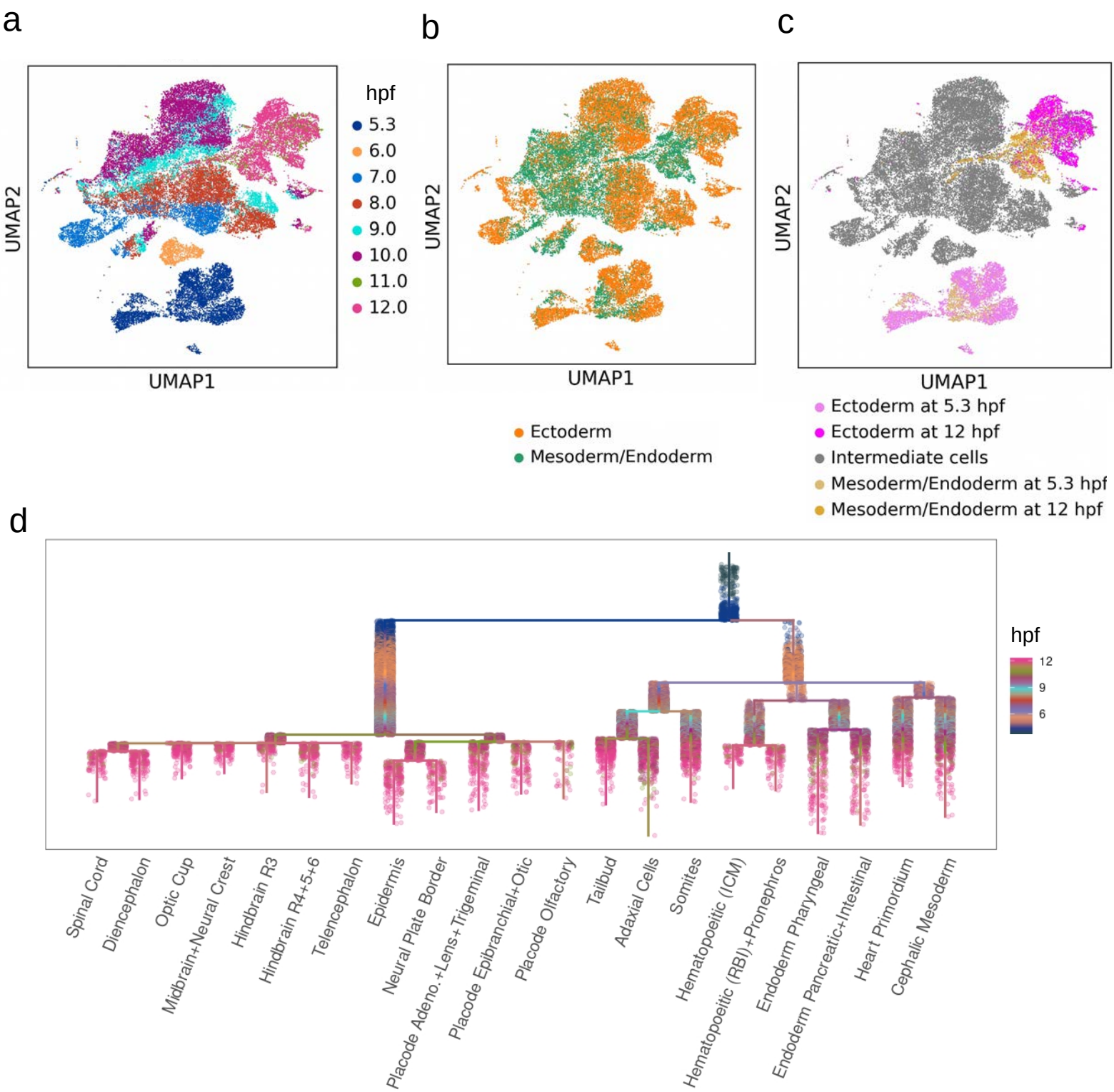
Cell Systems, Volume *15*

Supplemental information

Inference of differentiation trajectories by transfer learning across biological processes

Gaurav Jumde, Bastiaan Spanjaard, and Jan Philipp Junker



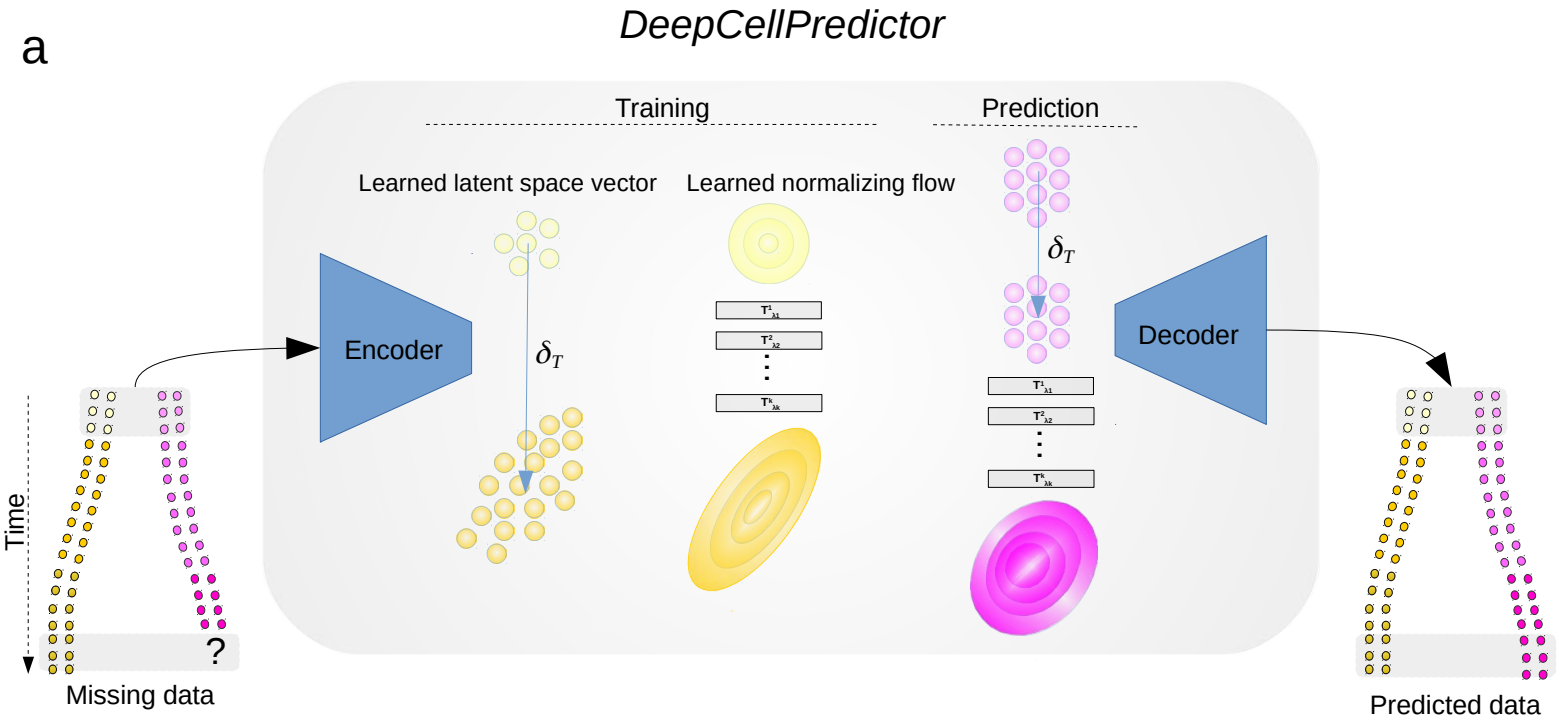
Supplementary Figure 1 | Developmental zebrafish dataset (related to Fig. 1).

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)**.

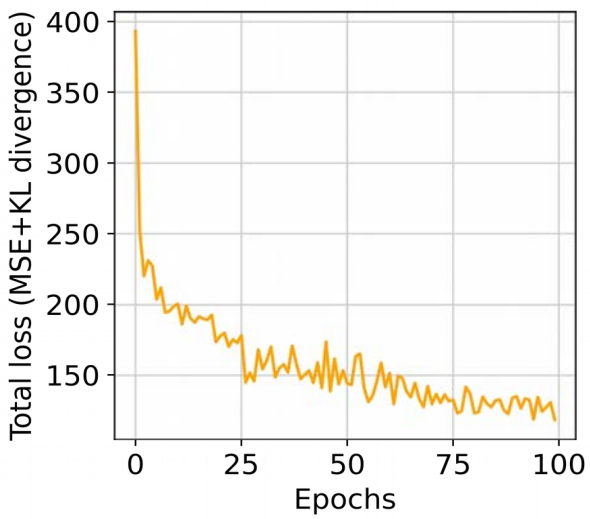
d) URD of the same dataset, indicating transcriptome-based inferred lineage splits. URD is the diffusion-based computational trajectory reconstruction method that was used in the original publication to analyze this dataset. Reproduced from original publication[7].

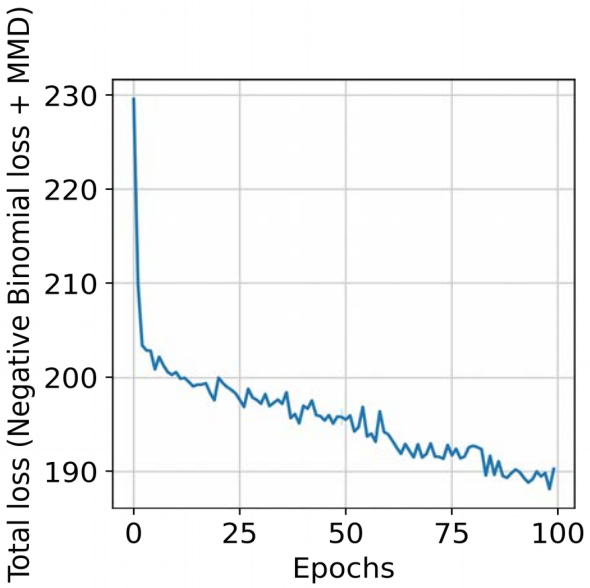
Data from Farrell et al., 2018[7], GEO accession GSE106587.

a

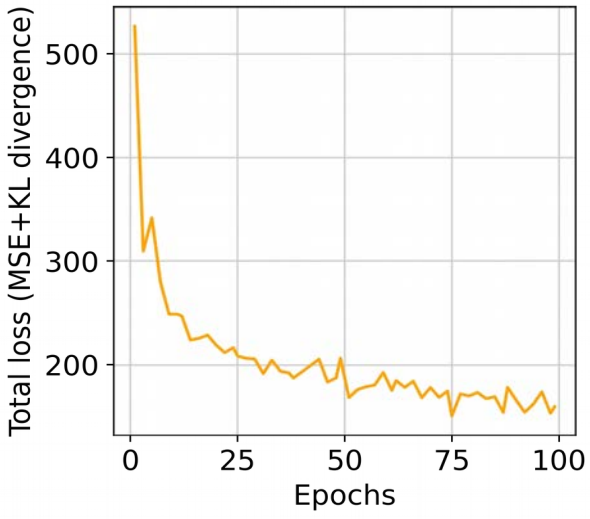

DeepCellPredictor
Training
Prediction
Learned latent space vector
Learned normalizing flow
Encoder
Decoder
δ_T
δ_T
Time
?
Missing data
Predicted data

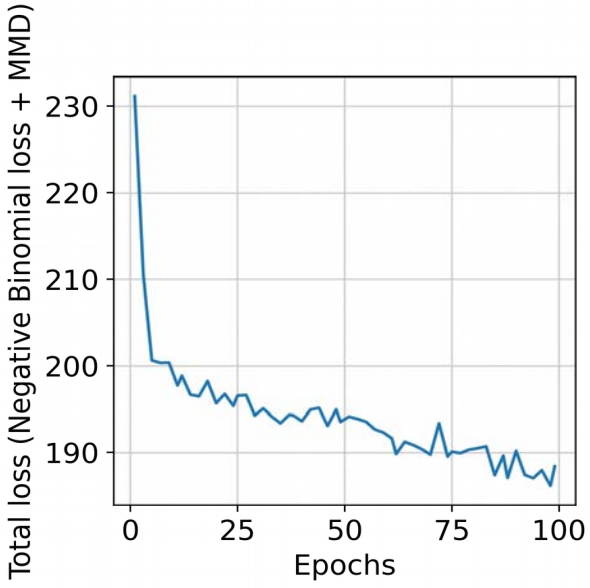
b


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

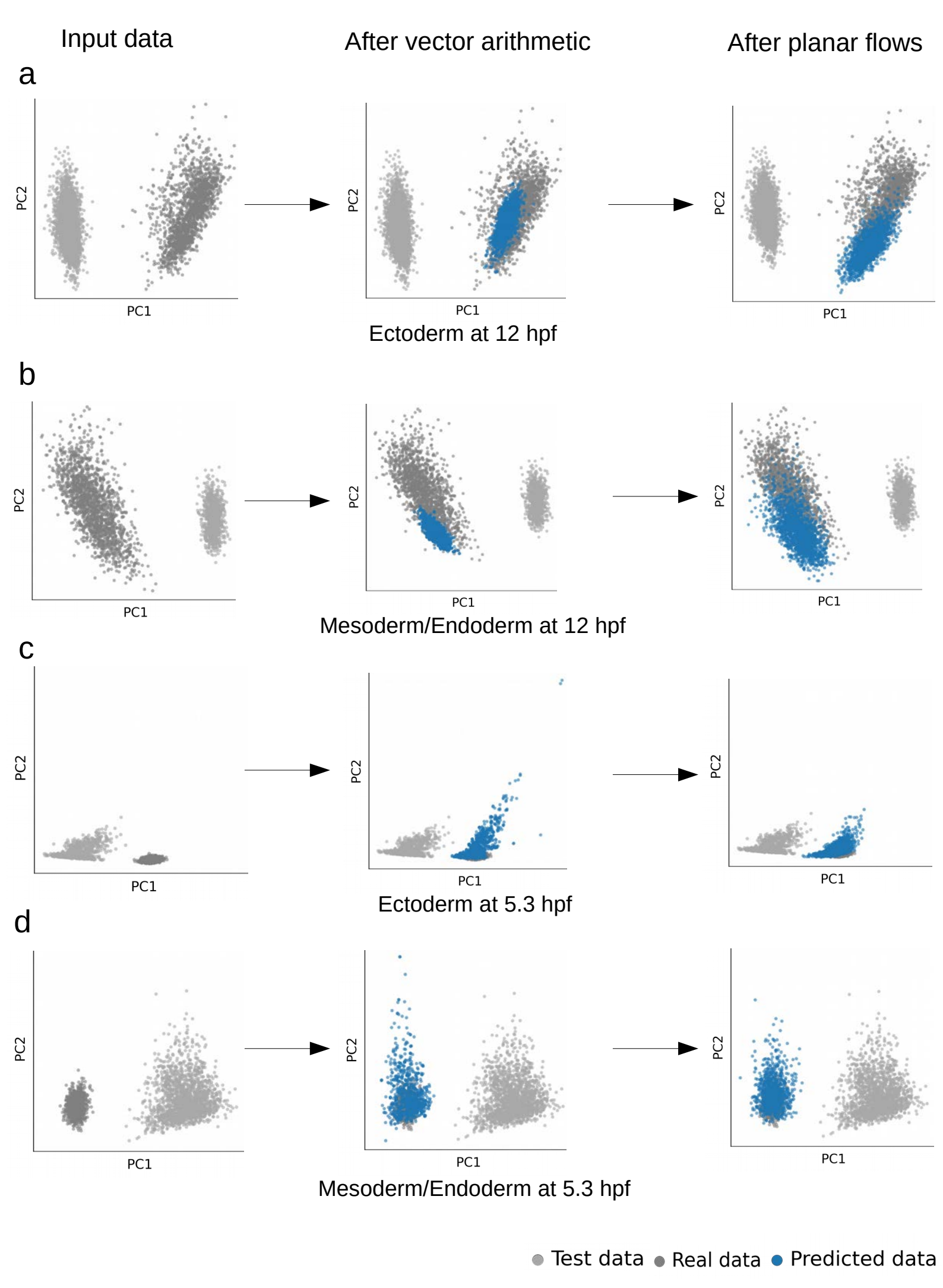

Total loss (Negative Binomial loss + MMD)
230
220
210
200
190
0
25
50
75
100
Epochs

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

a) Schematic of the DeepCellPredictor (DCP) model.

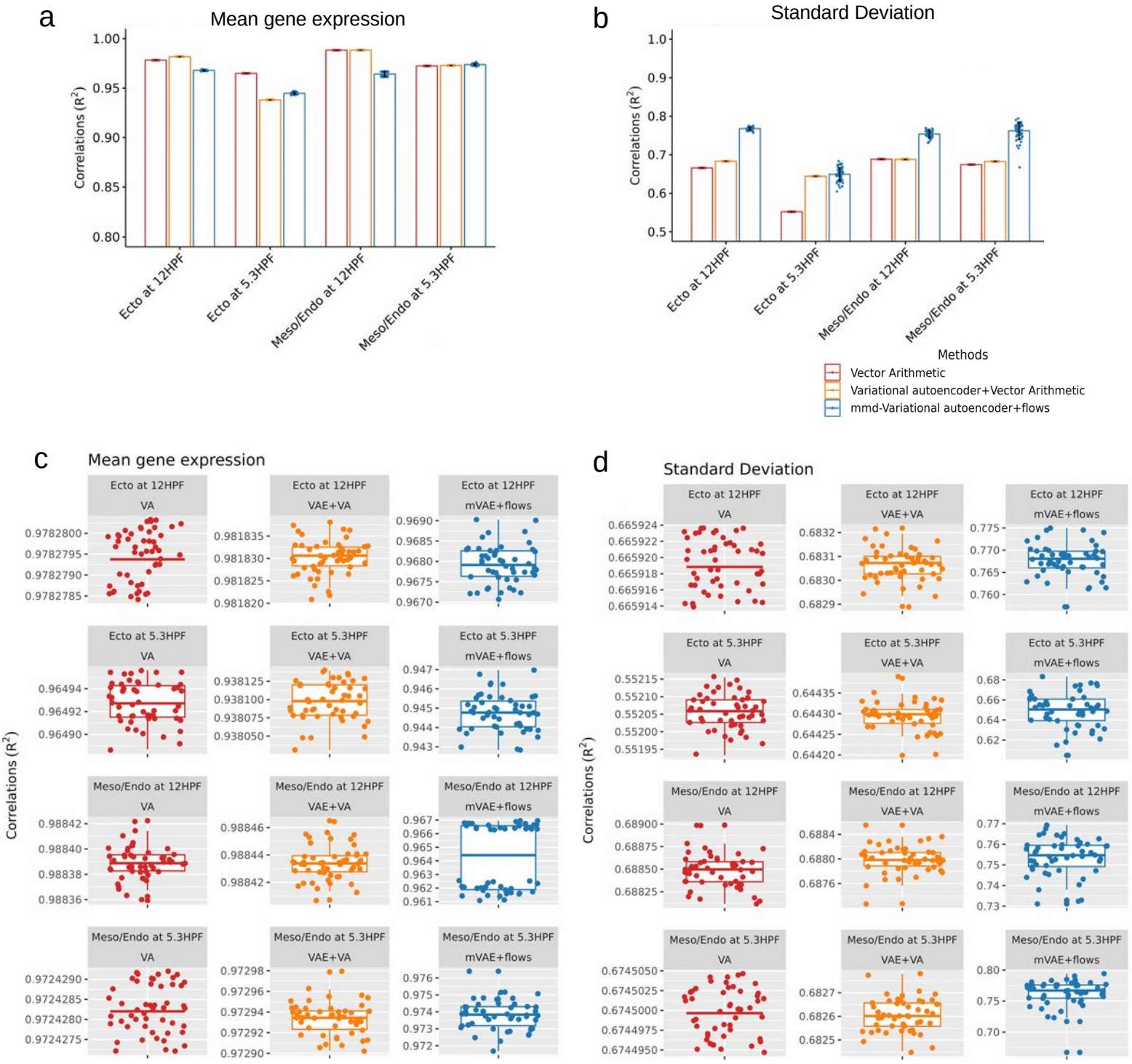
b-c) Total loss over epochs for regular variational autoencoder (left) and DCP model (right) trained on mesoderm/endoderm **(b)** and ectoderm data **(c).**

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.

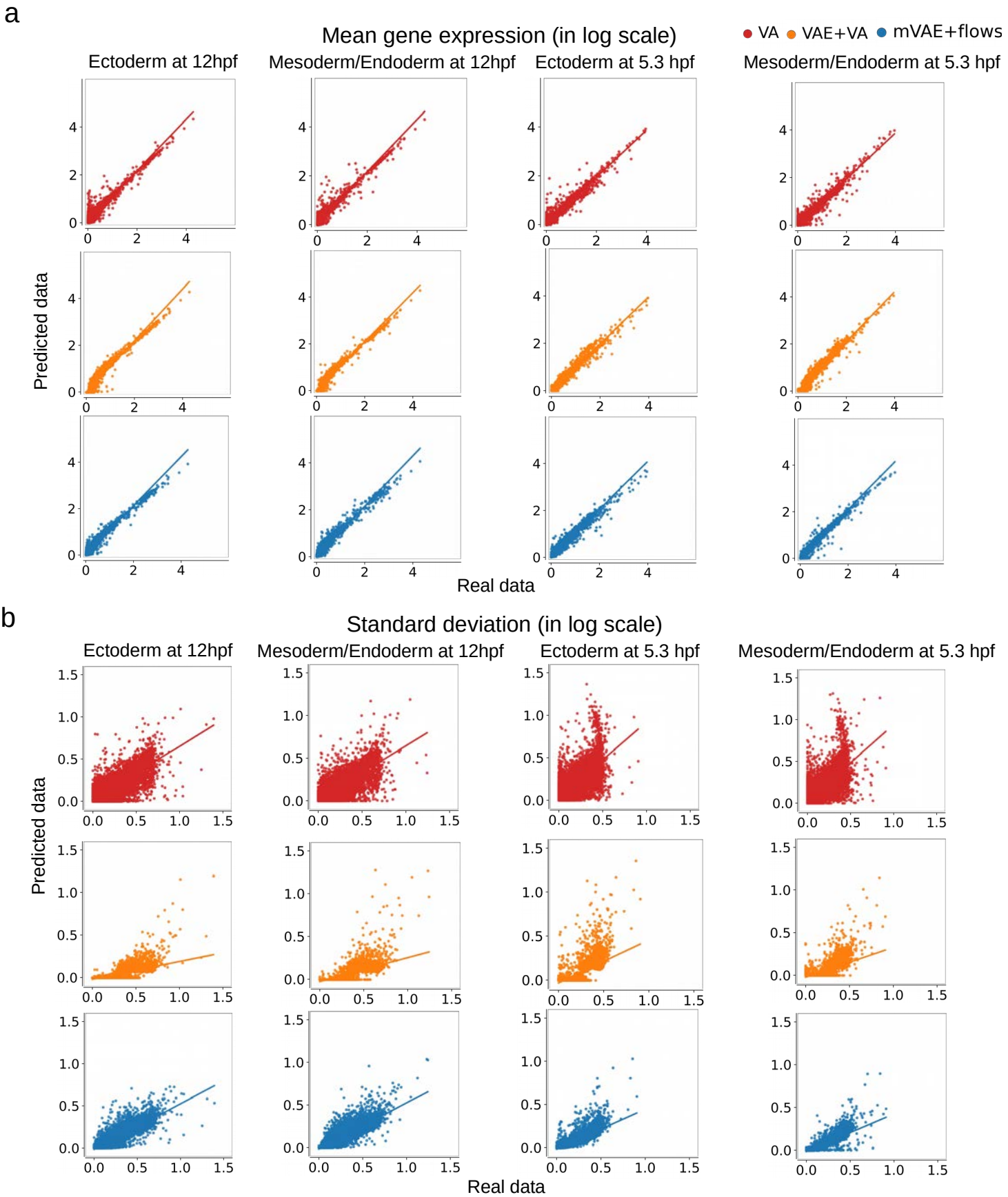


Supplementary Figure 4 | Comparison of prediction algorithms (vector arithmetic (VA), variational autoencoders with vector arithmetic (VAE+VA), and mmd-variational autoencoders with flows (mVAE+flows)) in the developing zebrafish (related to Fig. 2).

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.

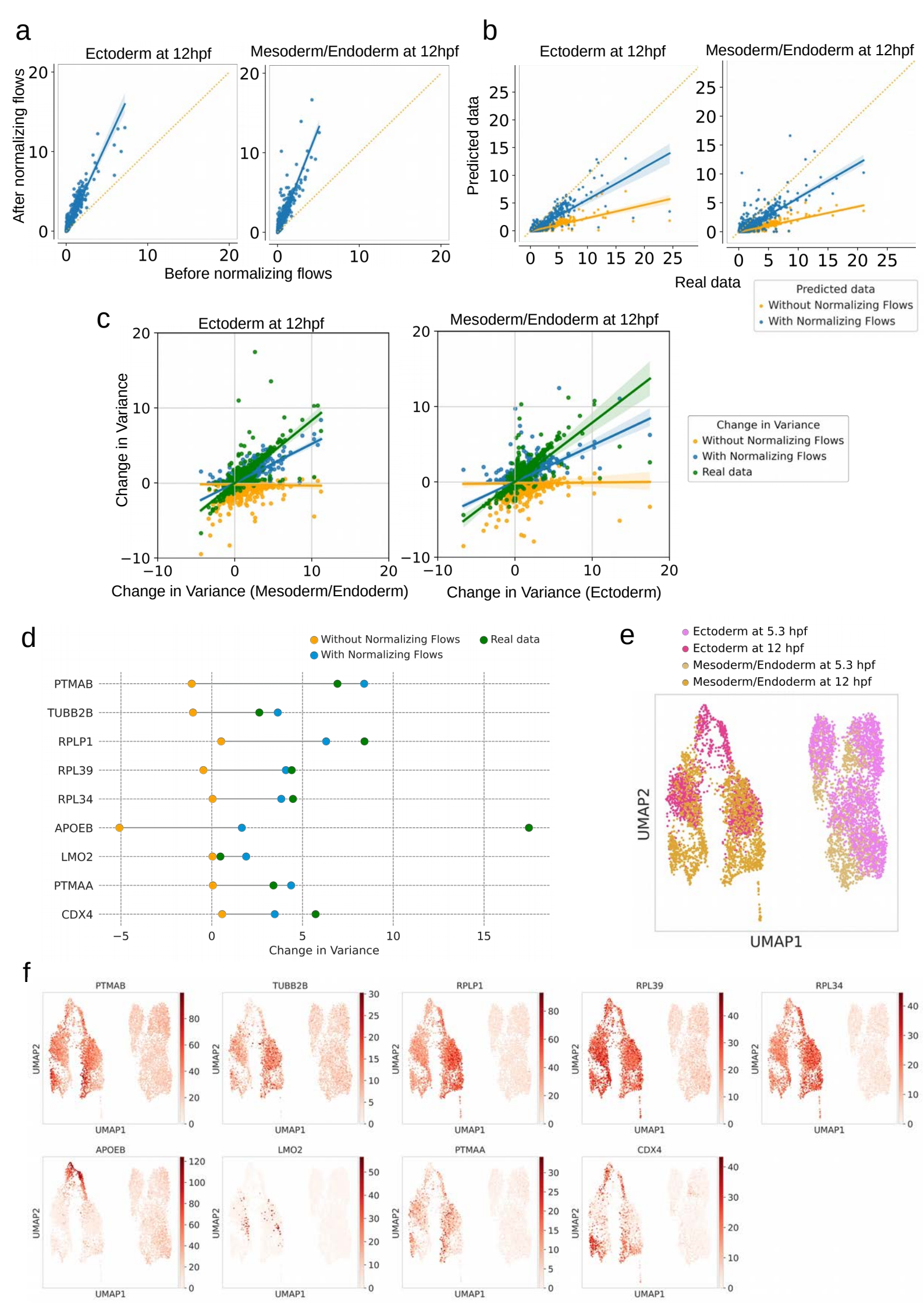
Data from Farrell et al., 2018[7], GEO accession GSE106587.



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.

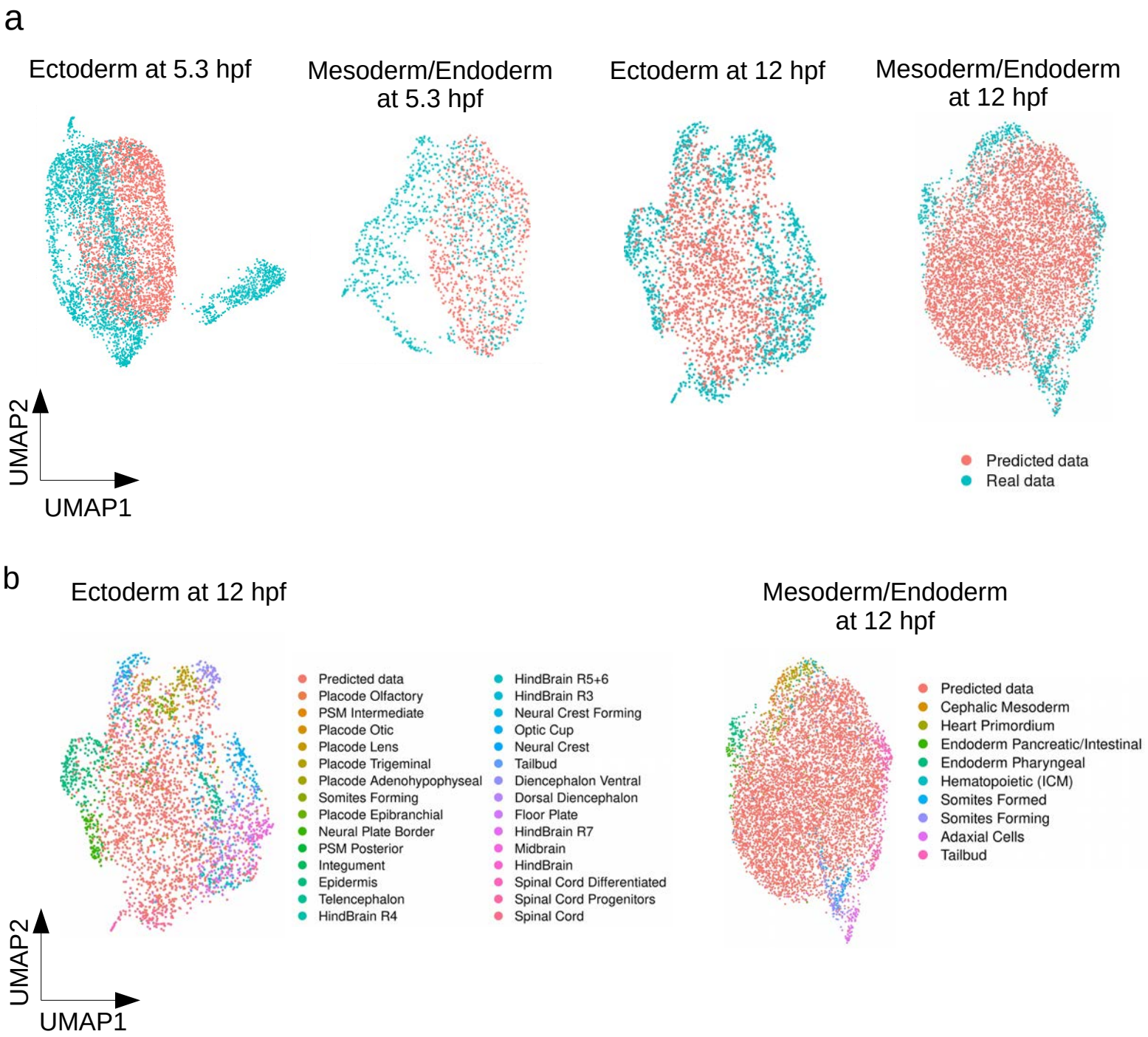
c) Change of expression variance in real and predicted data compared to the training data (in log scale). **a-c)**: predicting to 12hpf ectoderm (left) and 12hp mesoderm/endoderm (right).

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.

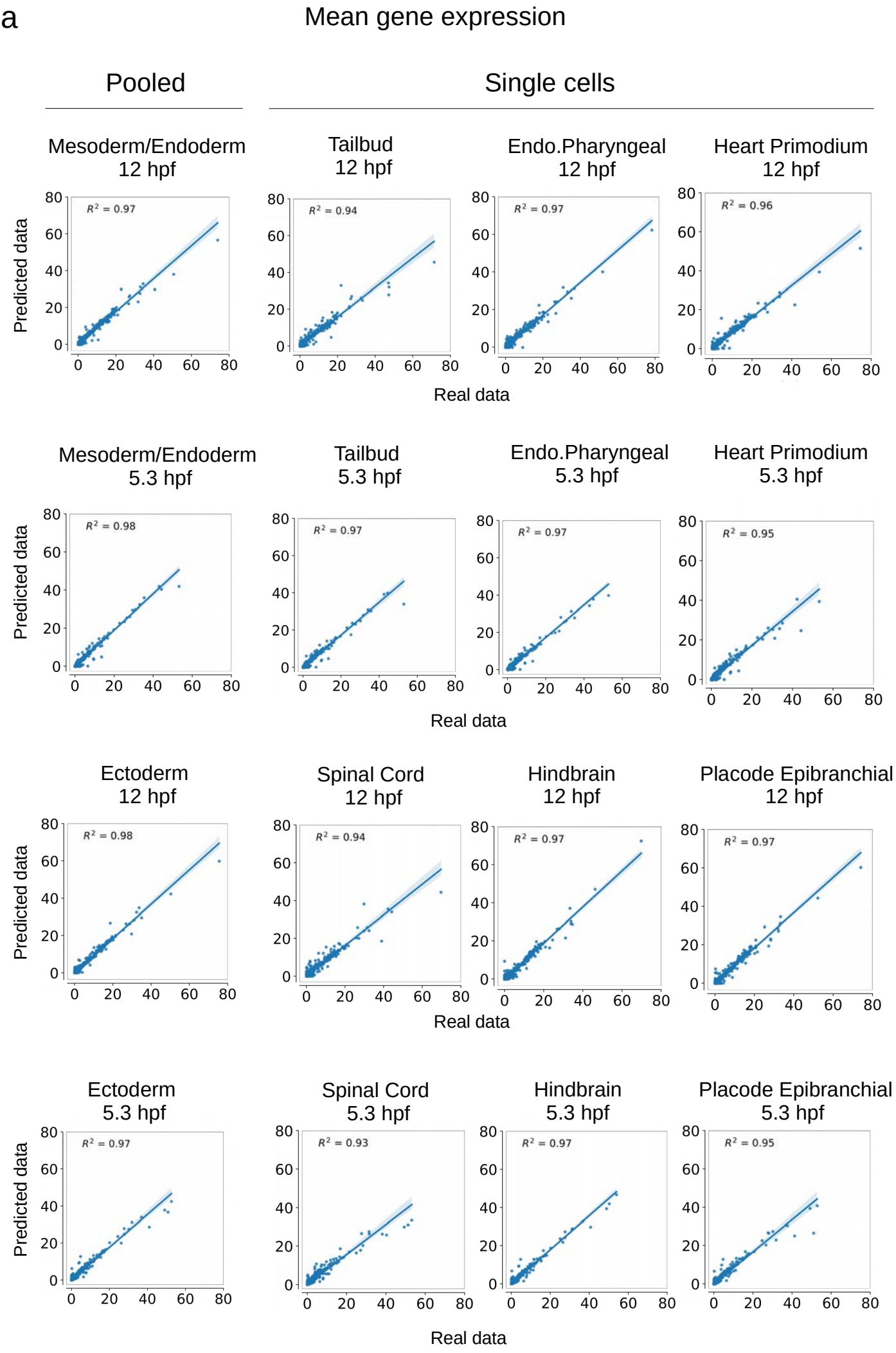


Supplementary Figure 7 | DCP generates realistic single-cell transcriptomes but does not accurately predict cell type clusters at 12 hpf (related to Fig. 2).

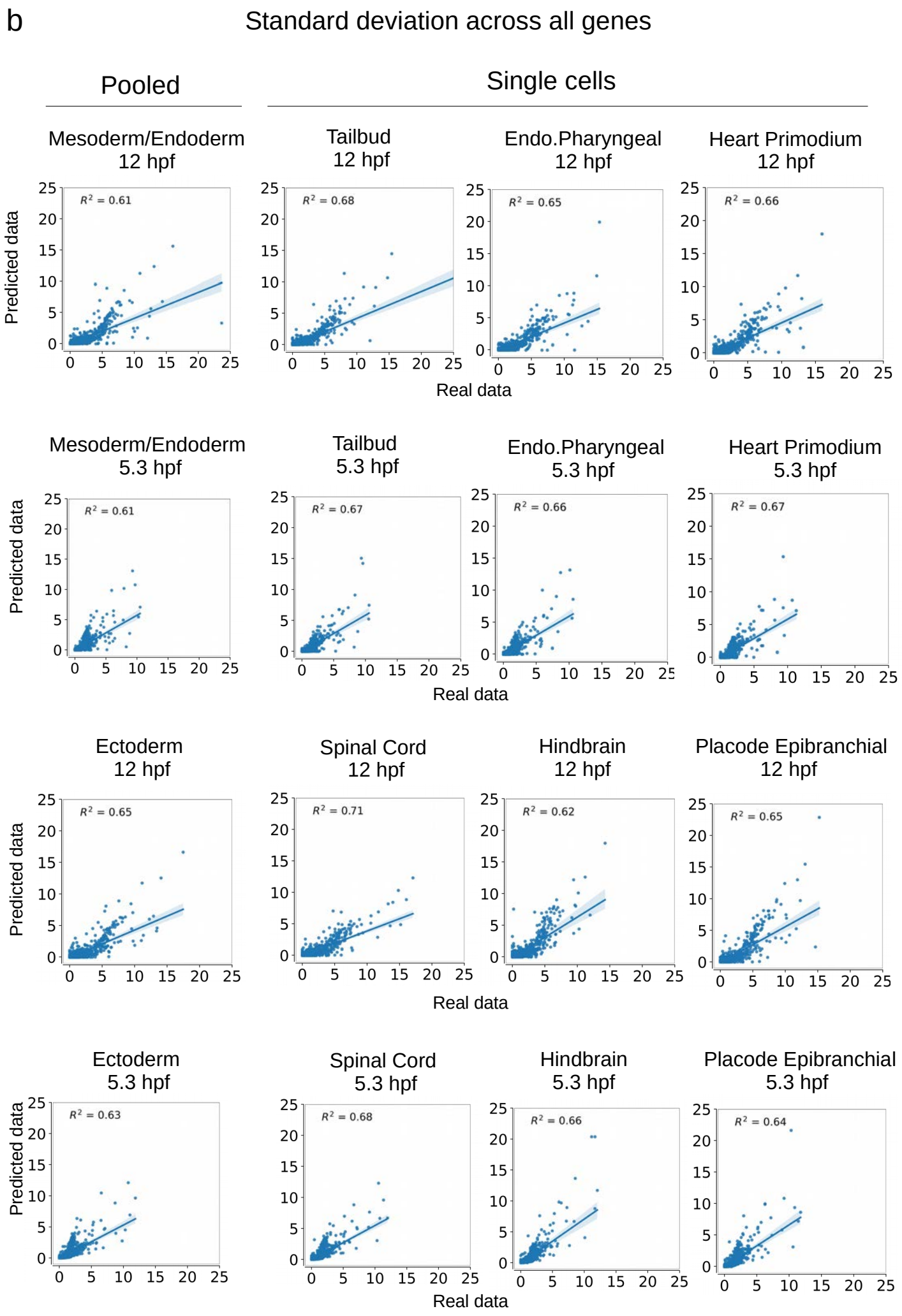
a-b) UMAP of integrated real and predicted data of zebrafish ectoderm and mesendoderm at 5.3 hpf and 12 hpf.

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

a

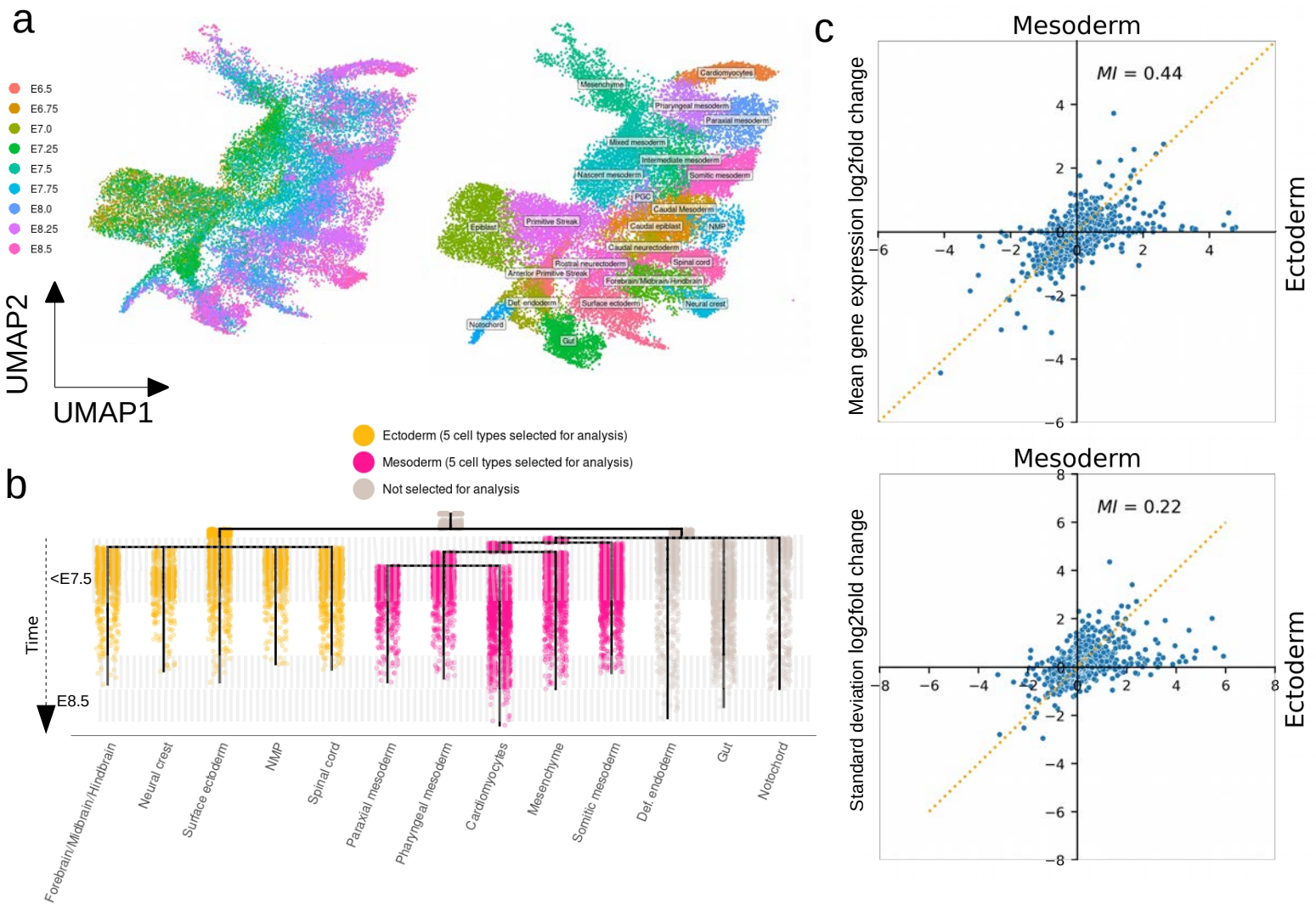

Mean gene expression
Pooled
Single cells
Mesoderm/Endoderm 12 hpf
Tailbud 12 hpf
Endo.Pharyngeal 12 hpf
Heart Primodium 12 hpf
Mesoderm/Endoderm 5.3 hpf
Tailbud 5.3 hpf
Endo.Pharyngeal 5.3 hpf
Heart Primodium 5.3 hpf
Ectoderm 12 hpf
Spinal Cord 12 hpf
Hindbrain 12 hpf
Placode Epibranchial 12 hpf
Ectoderm 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

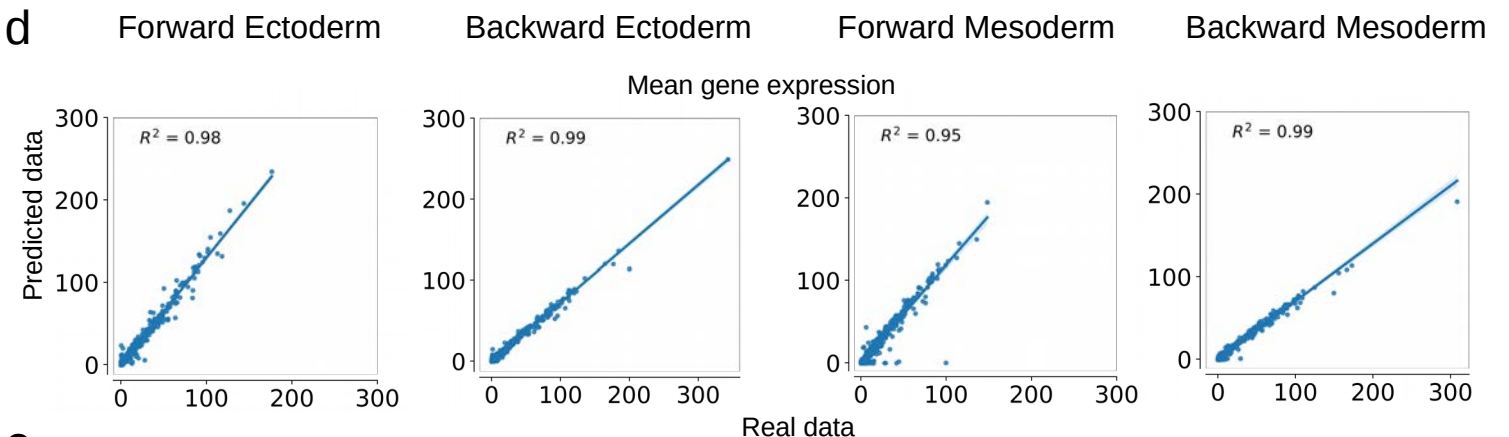
b

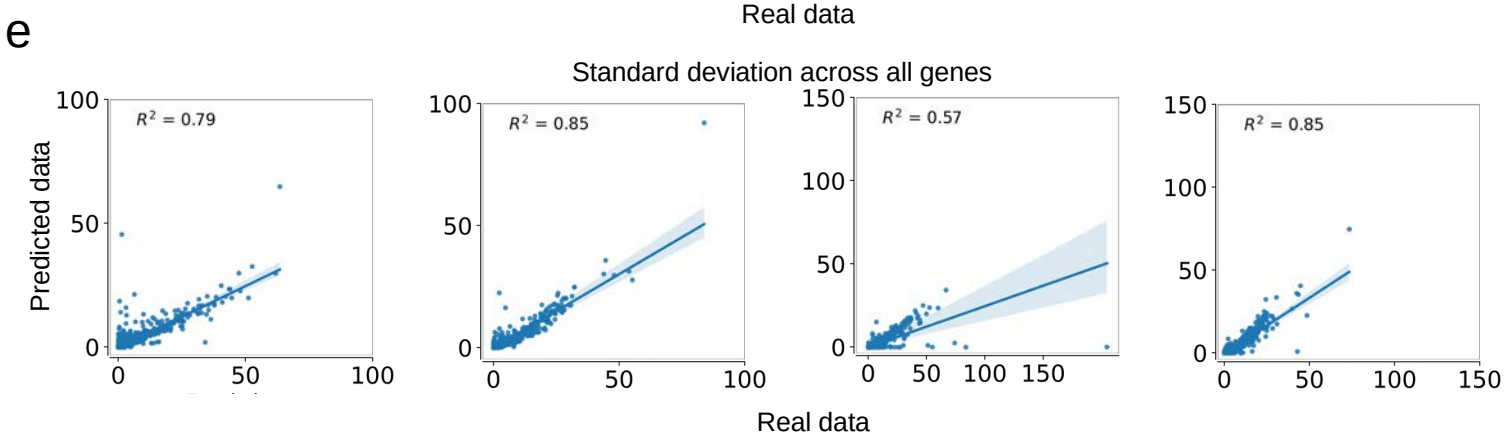

Standard deviation across all genes
Pooled
Single cells
Mesoderm/Endoderm 12 hpf
Tailbud 12 hpf
Endo.Pharyngeal 12 hpf
Heart Primodium 12 hpf
Mesoderm/Endoderm 5.3 hpf
Tailbud 5.3 hpf
Endo.Pharyngeal 5.3 hpf
Heart Primodium 5.3 hpf
Ectoderm 12 hpf
Spinal Cord 12 hpf
Hindbrain 12 hpf
Placode Epibranchial 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.

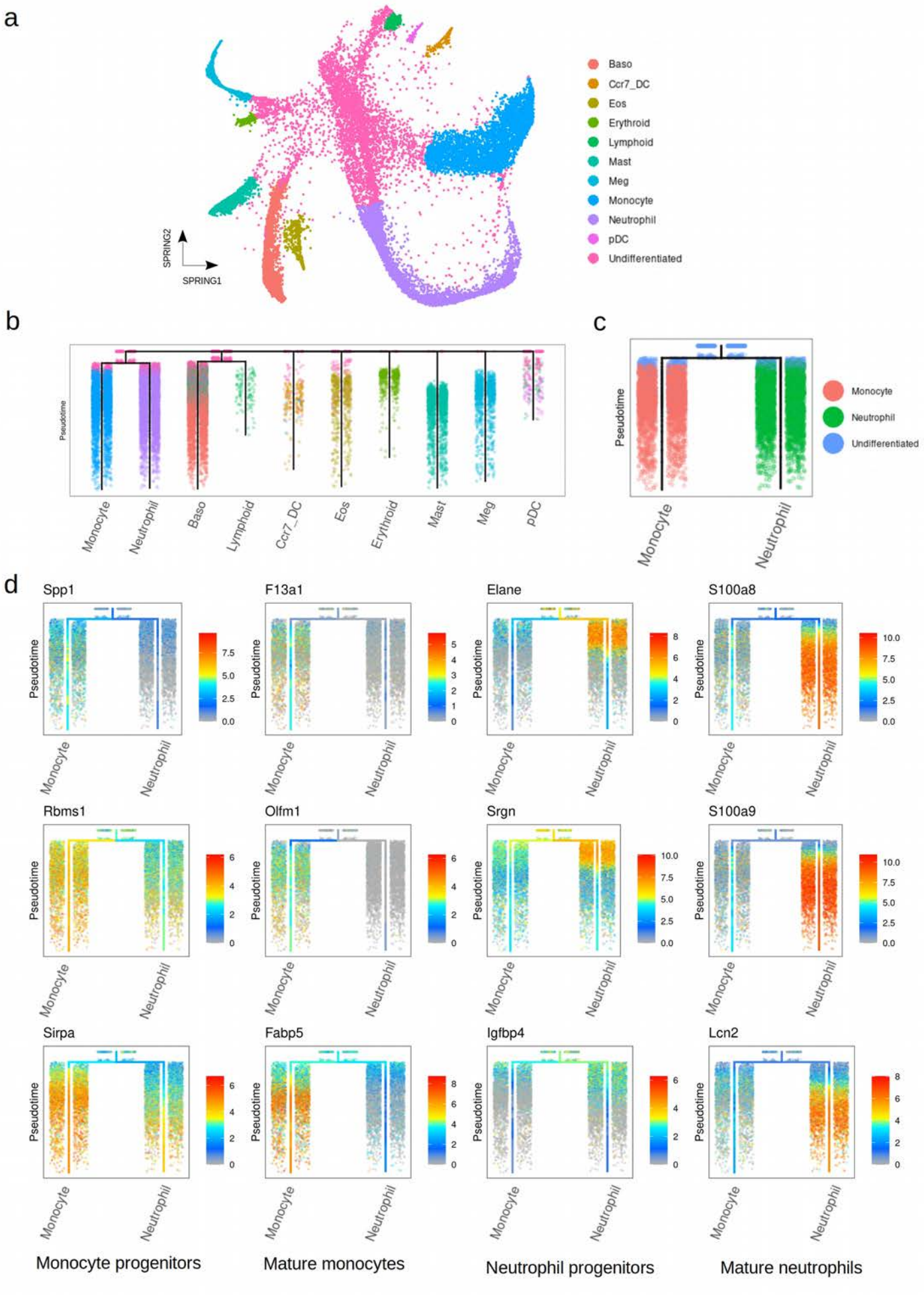

a
E6.5
E6.75
E7.0
E7.25
E7.5
E7.75
E8.0
E8.25
E8.5
UMAP2
UMAP1
Cardiomyocytes
Mesenchyme
Pharyngeal mesoderm
Paraxial mesoderm
Mixed mesoderm
Intermediate mesoderm
Nascent mesoderm
Somitic mesoderm
PGC
Caudal Mesoderm
Epiblast
Primitive Streak
Caudal epiblast
NMP
Caudal neurectoderm
Rostral neurectoderm
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).

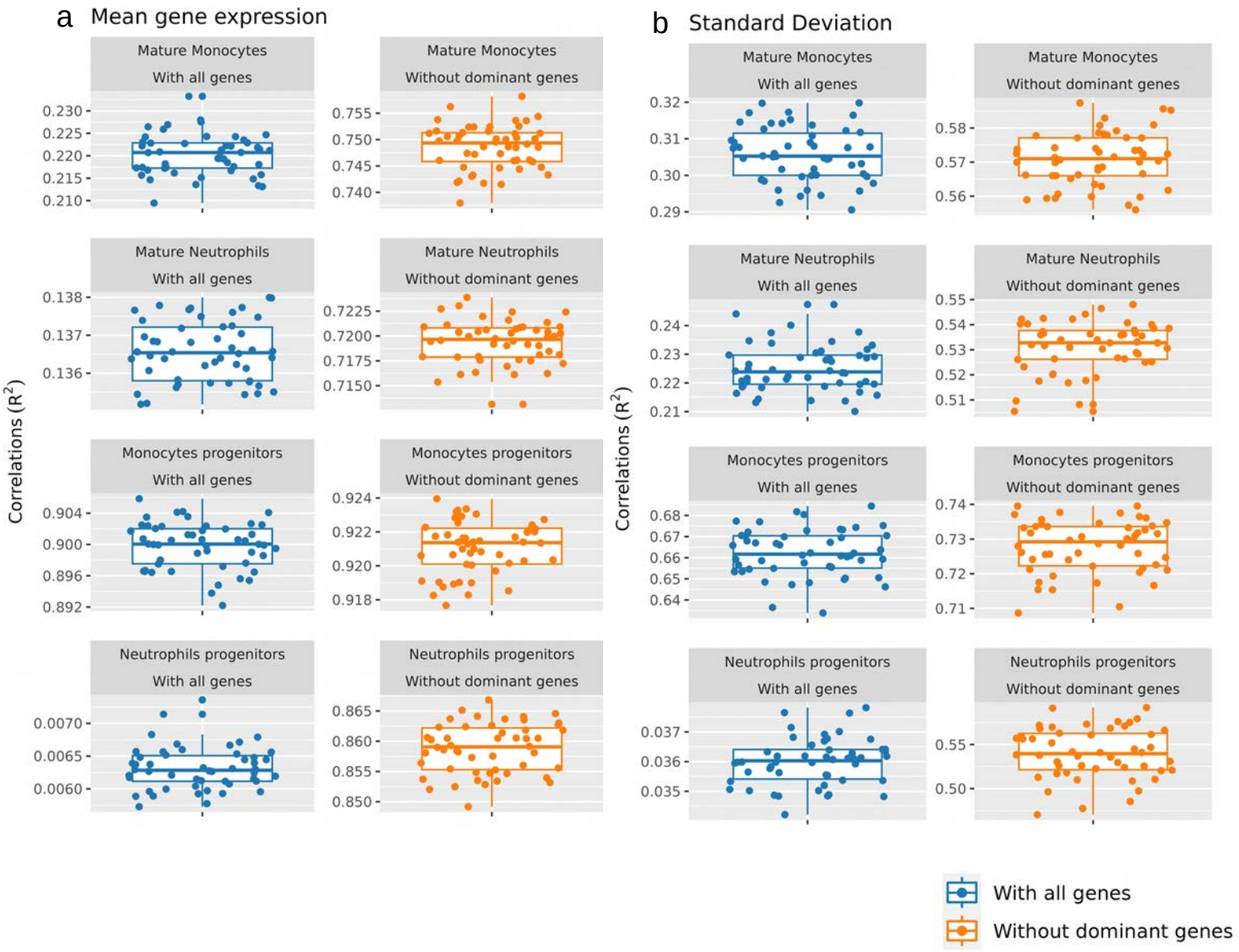
a) t-SNE of dataset.

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.

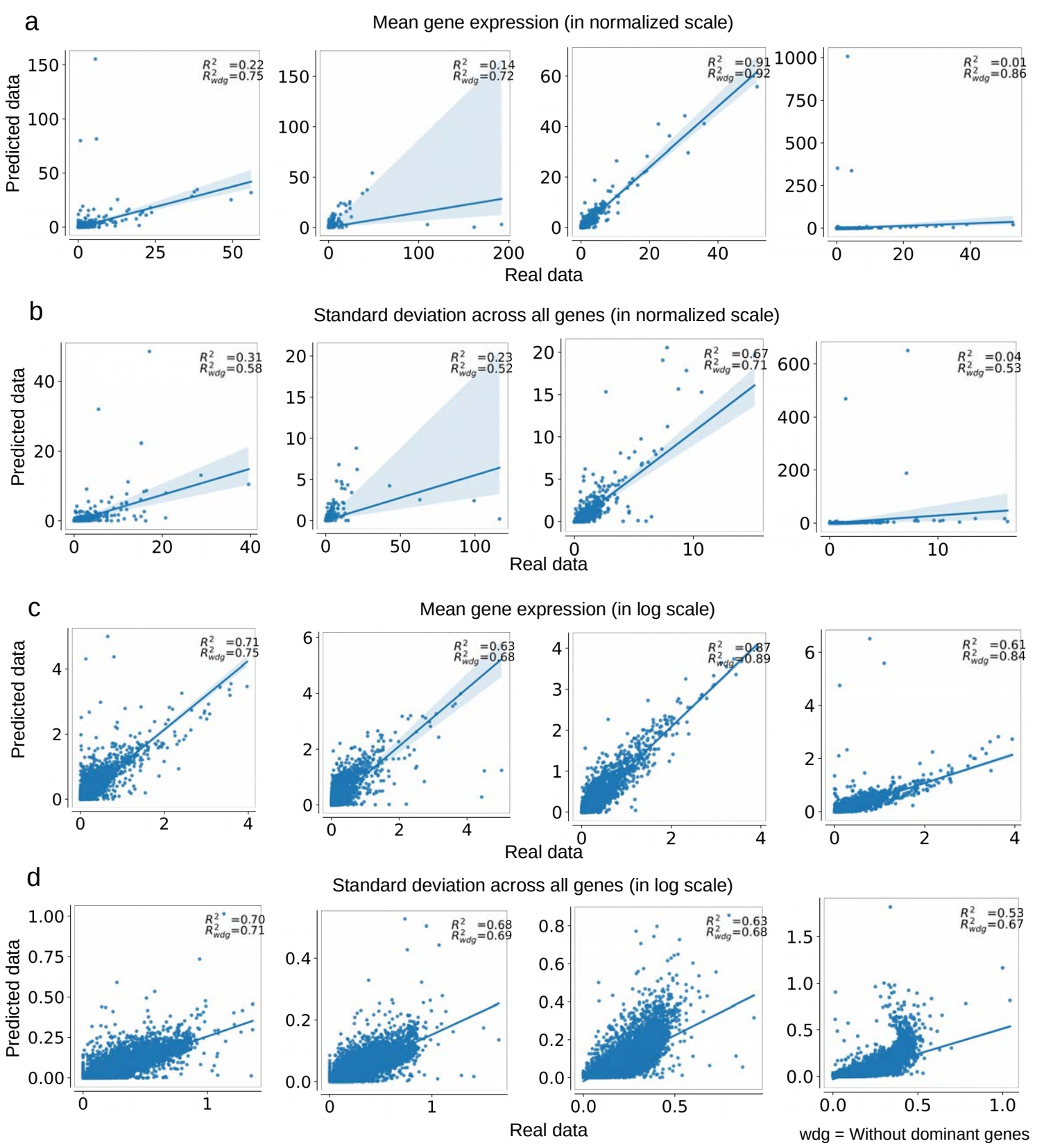
Data from Weinreb et al., 2020[18] (GEO accession GSE140802).



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.

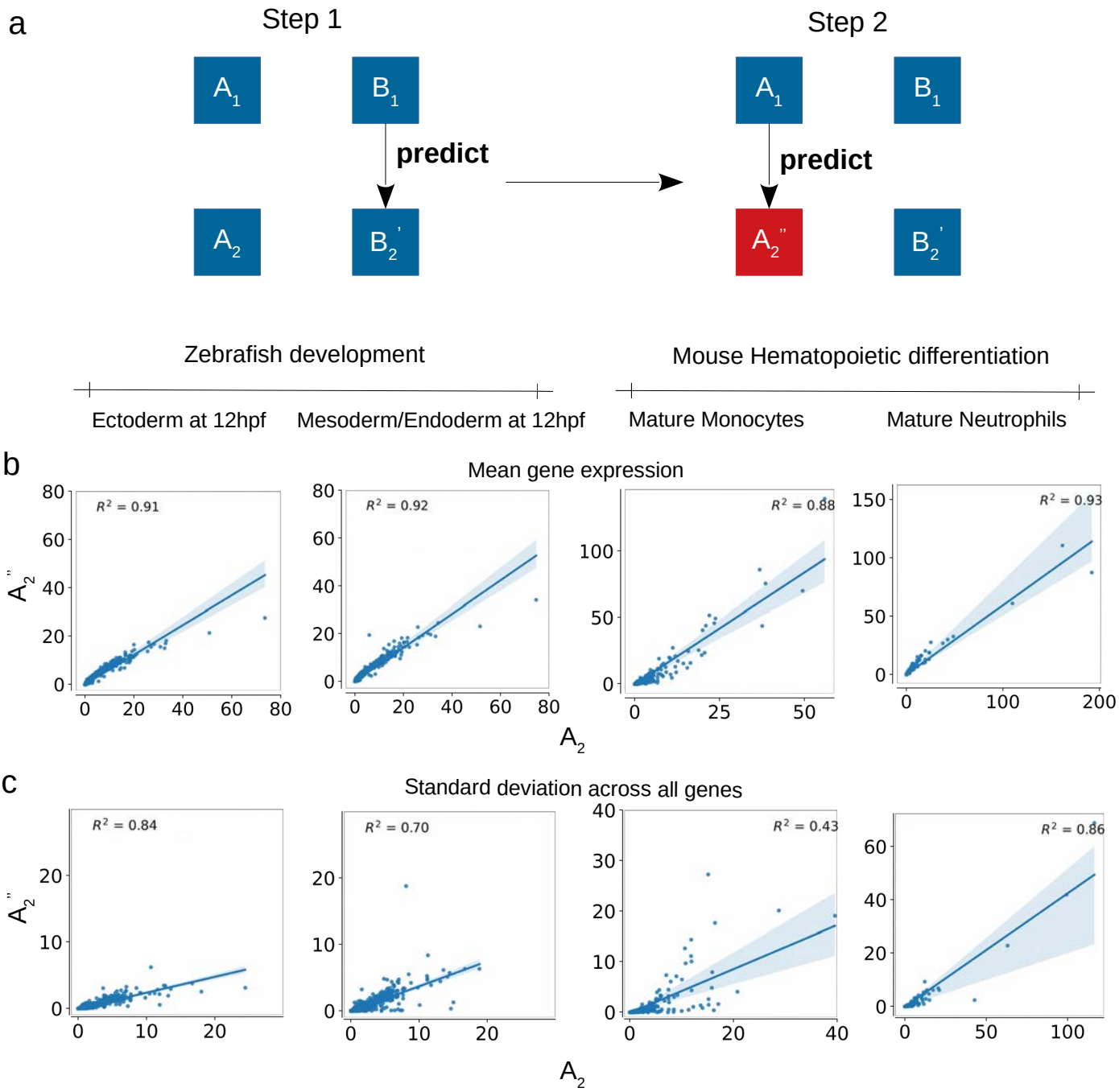
Data from Weinreb et al., 2020[18] (GEO accession GSE140802).


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).

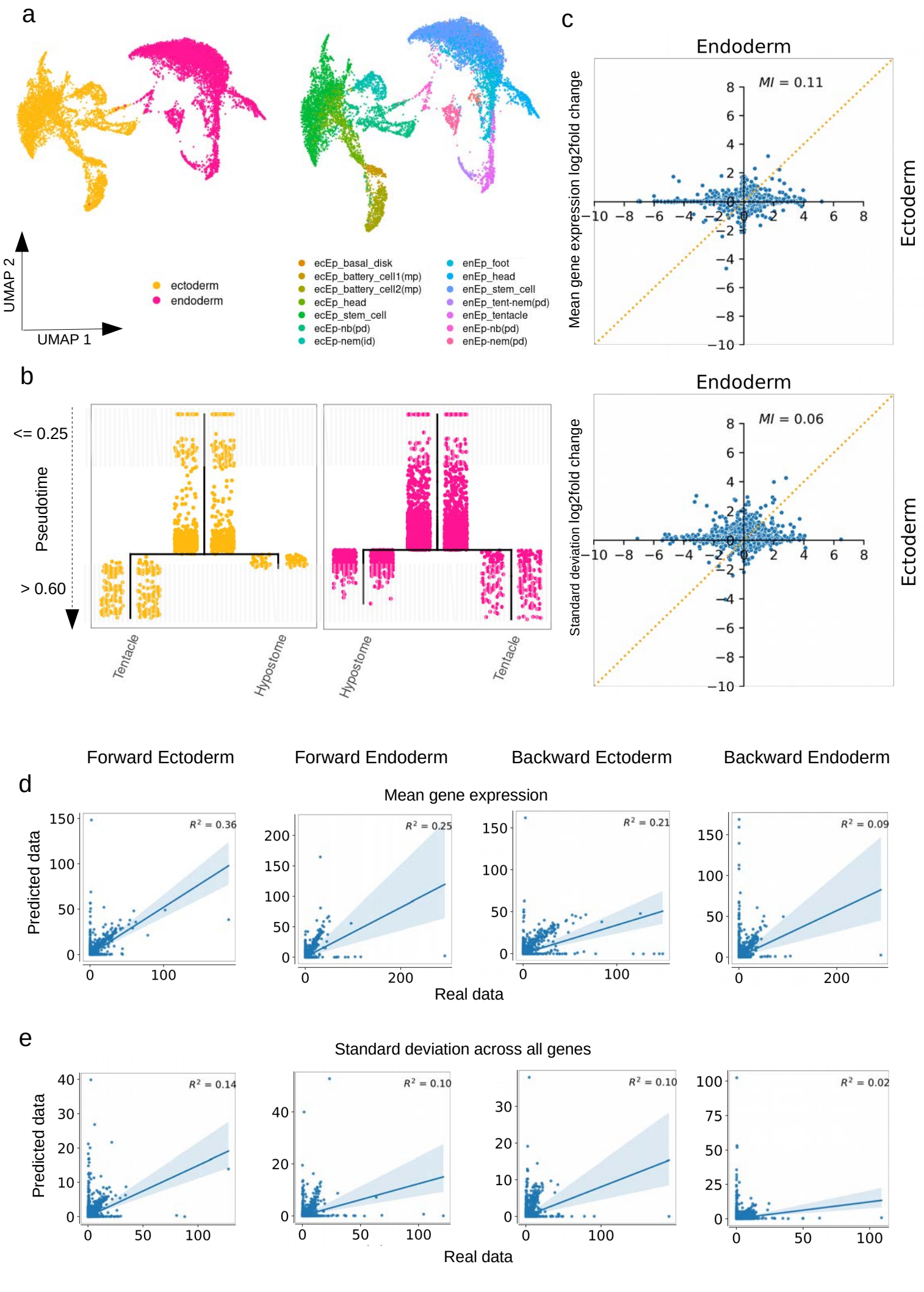


Supplementary Figure 13 | Successive predictions assess the information content of the latent space (related to Fig. 3).

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).