
Supplementary information

A cell atlas foundation model for scalable search of similar human cells

In the format provided by the
authors and unedited

Supplementary Note 1: Analysis Task Parameters

Fig. 2a

Tissue name simplified
Disease name simplified

Fig. 2c

Tissue name simplified
Disease name simplified
Exclude training cells
prediction_nn_dist > 0.03
in_vivo = True

Fig. 3b-d

Test kidney data: CELLxGENE data a98b828a-622a-483a-80e0-15703678befd
Safelist to cell types in a98b828a-622a-483a-80e0-15703678befd

Fig. 3e

Training data: subsampled 100,000 cells from all training datasets, weighted by study size
Test data: subsampled 10,000 cells from each test dataset
TOSICA: default parameters
CellTypist: default parameters
scANVI training: use_layer_norm="both", use_batch_norm="none", encode_covariates=True, dropout_rate=0.2, n_layers=2
scANVI query update: max_epochs=100, weight_decay=0.0, check_val_every_n_epoch=10
Ontology aware classification correctness evaluation

Fig. 4b

prediction_nn_dist < 0.03
in_vivo = True
Limit cell types to monocytes and macrophages

Fig. 4c

Tissue name simplified
Disease name simplified
prediction_nn_dist < 0.03
in_vivo = True
FM similarity threshold for FM classification: 99 percentile

Fig. 5

Top 3000 genes by highest attribution for gene selection

500 positive cells from top 500 SCimilarity score cells

500 negative cells selected randomly from the full set of monocytes and macrophages

Fig. 6a

prediction_nn_dist < 0.03

in_vivo = False

Extended Data Fig. 1d

Tissue name simplified

Disease name simplified

Extended Data Fig. 2d

Top 100 signature score FM cells as centroid from GSE159354

Test data GSE128033

Downloaded trained models for scGPT and scFoundation

Embeddings for scGPT and scFoundation performed according to author tutorials.

Extended Data Fig. 2e

Prediction constrained to: ["macrophage", "alveolar macrophage", "classical monocyte", "non-classical monocyte", "conventional dendritic cell", "plasmacytoid dendritic cell", "mast cell", "CD4-positive, alpha-beta T cell", "regulatory T cell", "CD8-positive, alpha-beta T cell", "mature NK T cell", "natural killer cell", "B cell", "plasma cell", "type I pneumocyte", "type II pneumocyte", "club cell", "goblet cell", "ciliated cell", "basal cell", "neuroendocrine cell", "pulmonary ionocyte", "endothelial cell of vascular tree", "endothelial cell of lymphatic vessel", "pericyte", "smooth muscle cell", "fibroblast", "myofibroblast cell"]

Extended Data Fig. 3a

100 random samplings of 1000 cells

Extended Data Fig. 3e

Prediction constrained to: ['B cell', 'CD4-positive, alpha-beta T cell', 'CD8-positive, alpha-beta T cell', 'natural killer cell', 'classical monocyte', 'non-classical monocyte', 'conventional dendritic cell', 'plasmacytoid dendritic cell', 'megakaryocyte']

Extended Data Fig. 3g

scVI, scArches, Harmony and Scanorama all perform highly variable gene selection (n=2000)

scVI: n_layers=2, n_latent=30, gene_likelihood="nb"

scArches training data: subsampled 101,133 cells from training data, weighted by study size

scArches training: use_layer_norm="both", use_batch_norm="none", encode_covariates=True, dropout_rate=0.2, n_layers=2

scArches query update: max_epochs=100, weight_decay=0.0, check_val_every_n_epoch=10

Harmony: default parameters

Scanorama: default parameters

Extended Data Fig. 4c-h

Training data: subsampled 100,000 cells from all training datasets, weighted by study size

Kidney test data: CELLxGENE data a98b828a-622a-483a-80e0-15703678befd

Safelist to cell types in a98b828a-622a-483a-80e0-15703678befd when training all methods

TOSICA: default parameters

CellTypist: default parameters

scANVI training: use_layer_norm="both", use_batch_norm="none", encode_covariates=True, dropout_rate=0.2, n_layers=2

scANVI query update: max_epochs=100, weight_decay=0.0, check_val_every_n_epoch=10

Extended Data Fig. 5b-i

Training data: subsampled 100,000 cells from all training datasets, weighted by study size and constrained to cell types ['naive B cell', 'memory B cell', 'plasmablast', 'plasma cell', 'naive thymus-derived CD4-positive, alpha-beta T cell', 'CD4-positive, alpha-beta memory T cell', 'regulatory T cell', 'naive thymus-derived CD8-positive, alpha-beta T cell', 'CD8-positive, alpha-beta memory T cell', 'gamma-delta T cell', 'mucosal invariant T cell', 'mature NK T cell', 'CD16-negative, CD56-bright natural killer cell, human', 'CD16-positive, CD56-dim natural killer cell, human', 'CD14-positive monocyte', 'CD14-low, CD16-positive monocyte', 'conventional dendritic cell', 'plasmacytoid dendritic cell', 'neutrophil', 'hematopoietic stem cell', 'platelet', 'erythrocyte']

TOSICA: default parameters

CellTypist: default parameters

scANVI training: use_layer_norm="both", use_batch_norm="none", encode_covariates=True, dropout_rate=0.2, n_layers=2

scANVI query update: max_epochs=100, weight_decay=0.0, check_val_every_n_epoch=10

Extended Data Fig 6.

prediction_nn_dist < 0.03

in_vivo = True

Extended Data Fig. 7a

Tissue name simplified

Disease name simplified

prediction_nn_dist < 0.03

in_vivo = True

Fibroblast similarity threshold for fibroblast classification: 95 percentile

Extended Data Fig. 7b

Tissue name simplified

Disease name simplified

in_vivo = True

Fibroblast similarity threshold for fibroblast classification: 95 percentile

FM similarity threshold for FM classification: 95 percentile

Extended Data Fig. 7c

Top 100 signature score fibroblast and myofibroblast cells as centroid from GSE159354

Test data GSE128033

Downloaded trained models for scGPT and scFoundation

Embeddings for scGPT and scFoundation performed according to author tutorials.

Extended Data Fig. 7d

Prediction constrained to: ['B cell', 'CD4-positive, alpha-beta T cell', 'CD8-positive, alpha-beta T cell', 'natural killer cell', 'classical monocyte', 'non-classical monocyte', 'conventional dendritic cell', 'plasmacytoid dendritic cell', 'megakaryocyte']

Supplementary Note 2: Equations

Equation 1.
$$d(x, y) = ||f(x) - f(y)||_2^2$$

Equation 2.
$$L_{triplet} = \frac{\sum_i^N \max(d(x_i^a, x_i^p) - d(x_i^a, x_i^n) + \alpha, 0)}{N}$$

Equation 3.
$$L_{MSE} = \frac{\sum_i^N ||x_i^a - g(f(x_i^a))||_2^2}{N}$$

Equation 4.
$$L = (1 - \beta) * L_{MSE} + \beta * L_{triplet}$$

Equation 5.
$$||f(x_i^a) - f(x_i^p)||_2^2 < ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha$$

Equation 6.
$$||f(x_i^a) - f(x_i^p)||_2^2 > ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha$$

Equation 7.
$$||f(x_i^a) - f(x_i^n)||_2^2 - ||f(x_i^a) - f(x_i^p)||_2^2 < \alpha$$

Equation 8.
$$Importance_i(x, y) = \left| \max((x_i - y_i), 0) \times \int_{a=0}^1 \frac{\partial d(y + a \times (x - y), y)}{\partial x_i} \right|$$

Equation 9.
$$Signature_i(t) = \frac{1}{N} \sum_{c=1}^N Importance_i(t_c, b_c)$$

Equation 10.
$$AUC = \frac{U}{n_1 n_2}$$

Equation 11.
$$SCimilarity\ score = \frac{1}{1 - c_i \cdot c_j}$$

Equation 12.
$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Equation 13.
$$b(i) = \min_{j \neq i} \frac{1}{|C_j| - 1} \sum_{j \in C_j} d(i, j)$$

Equation 14.
$$Celltype(t) = \arg \max_t \left(\sum_{i \in t} \frac{1}{n} \right)$$

Equation 15.
$$Celltype(t) = \arg \max_t \left(\sum_{y \in t} \frac{1}{d(x, y)} \right)$$