

(Figure continues)

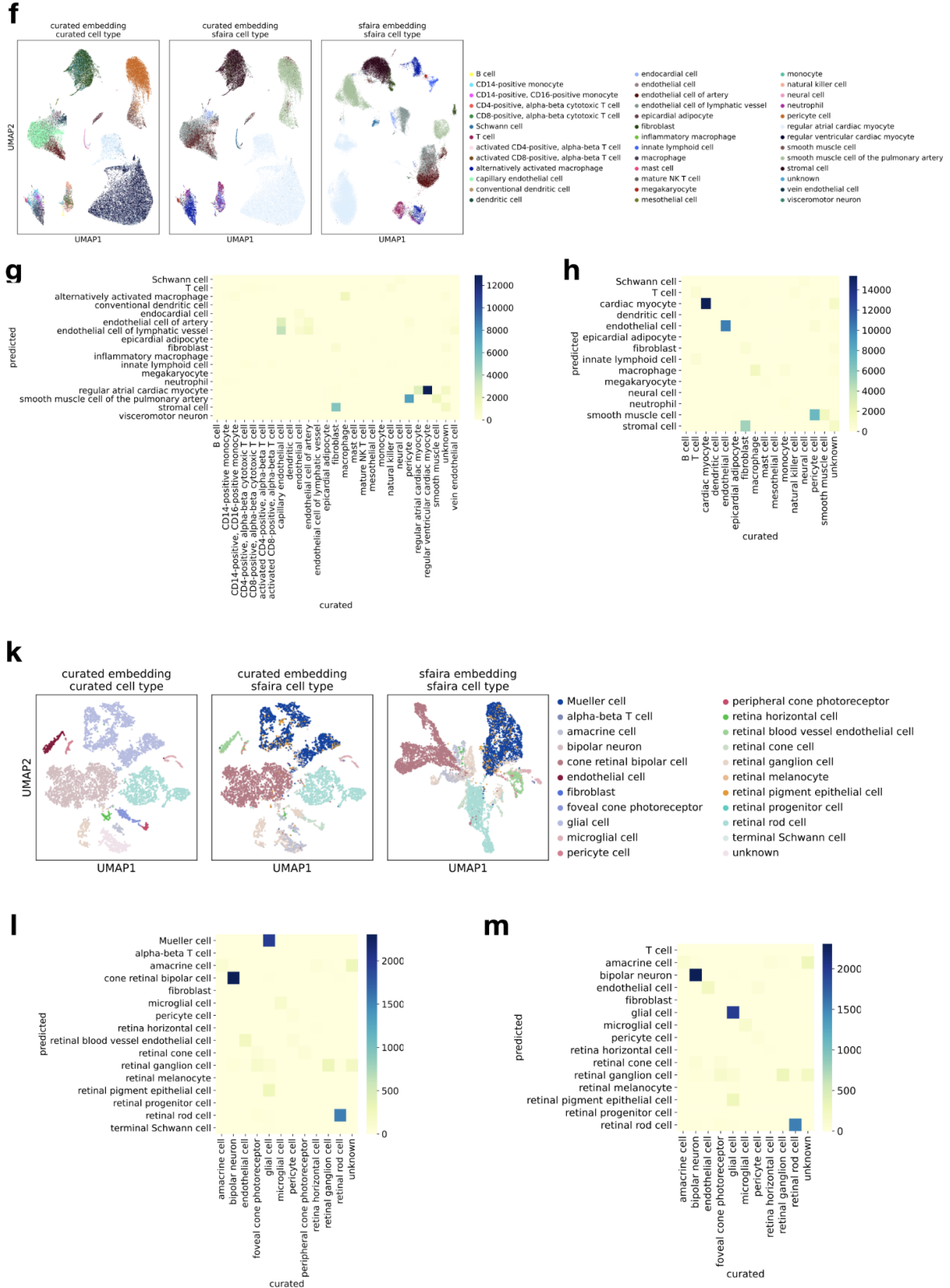


Fig. S1: Zero-shot analysis example cases. Shown are summaries of four zero-shot analysis scenarios which were not included in the training and testing data of the sfaira model zoo presented here (Methods): A pancreas data set (**a, b**), a kidney data set (**c-e**), a heart data set (**f-h**), and an eye data set (**k-m**). Shown are UMAPs of either the curated embedding published by the authors or a UMAP of a sfaira embedding model with curated cell types superimposed and with sfaira predicted cell types superimposed (**a,c,f,k**). Note that predicted and curated cell types are not always the same but often stand in an

ontological relationship, which derives from annotation coarsity differences between model training data and these new data sets. The relation between these predicted and observed cell types is further characterised in confusion matrices **(b,d,g,l)**, which were simplified to coarser cell ontology parent nodes in **(e,h,m)**. The analyses presented here were performed after the final fitting of the models that are also available on zenodo and thus represent the true behaviour of this system on new data. Minor inaccuracies in prediction are likely due to insufficient training data annotation or are related to issues in generalisation to the unseen domain effects in these test cases and can be resolved quickly in secondary analyses following this preliminary analysis.

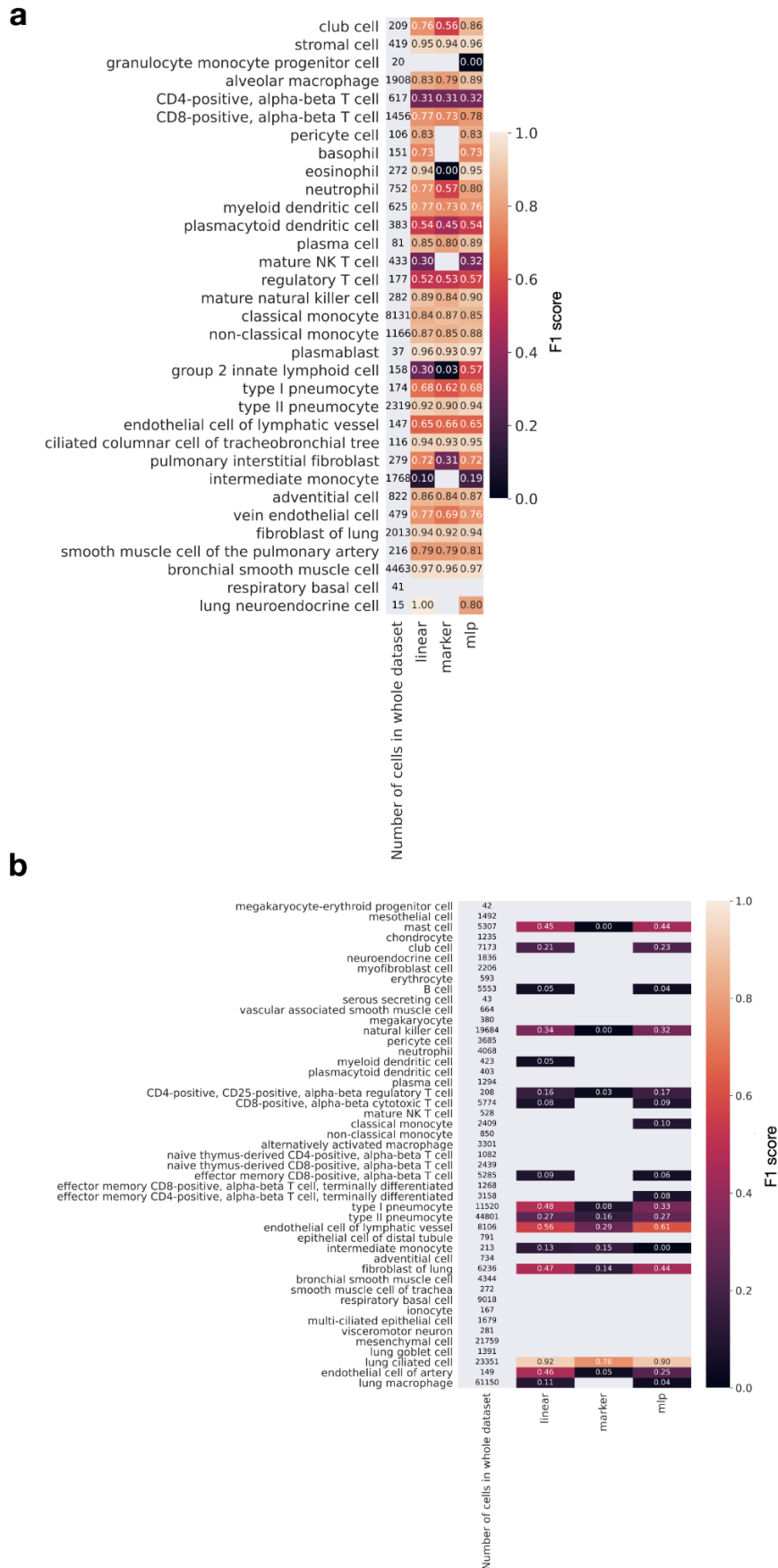


Fig. S2: Characterization of the cell type classification task. (a,b) Cell type wise accuracies for lung samples from humans (a) and mice (b).

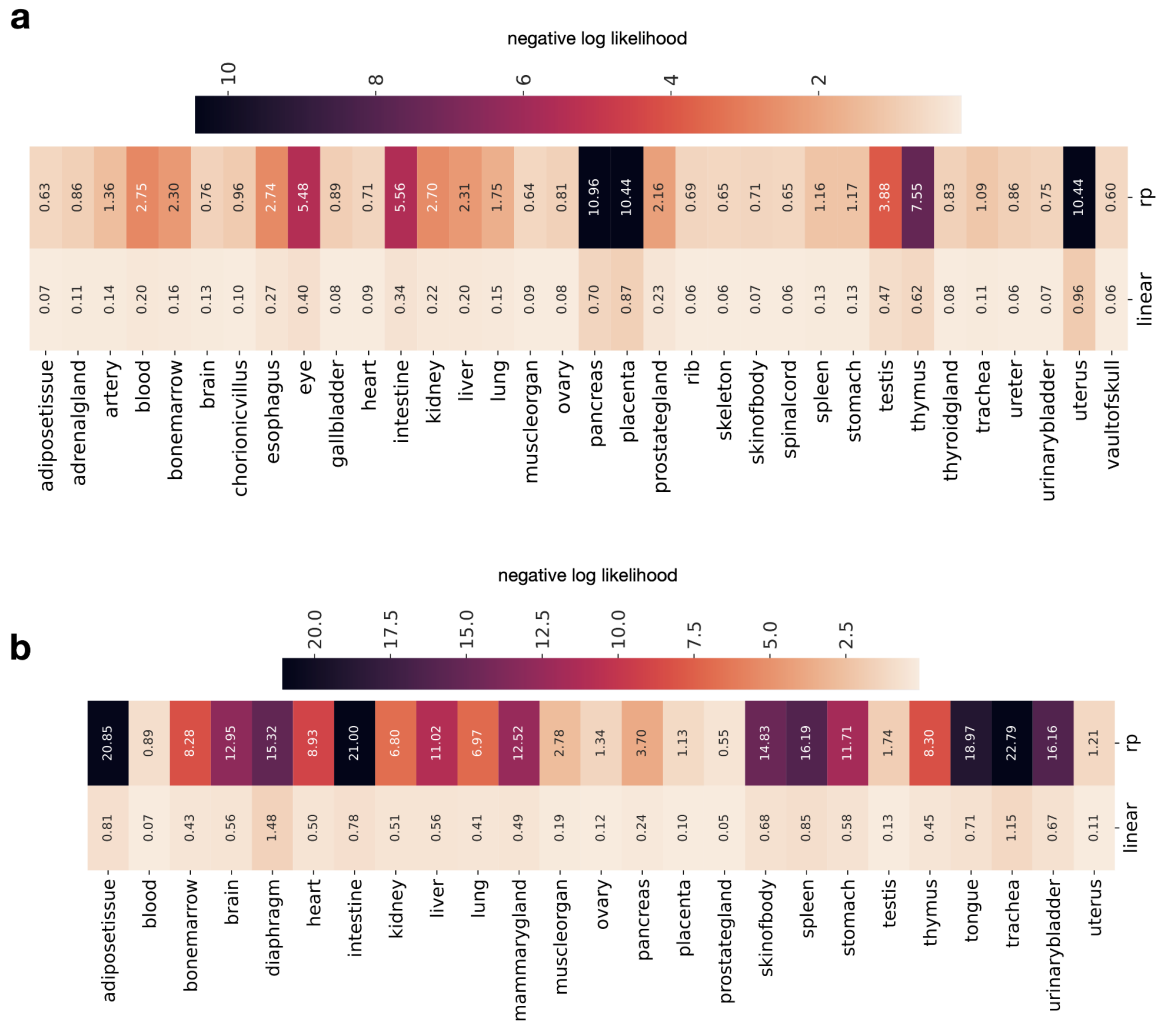


Fig. S3: Random projection models as baseline models for cell embedding models. Shown are the negative log likelihood on the held out test data of a linear embedding model as Fig. 5, and a random projection model (Methods) for human **(a)** and mouse **(b)**. Across all organs, the random projection model had significantly lower negative log likelihoods than the linear model when comparing the mean performance across cross validations in a one-sided paired t-test (human: $p=1.11e-05$, mouse: $p=2.11e-07$).

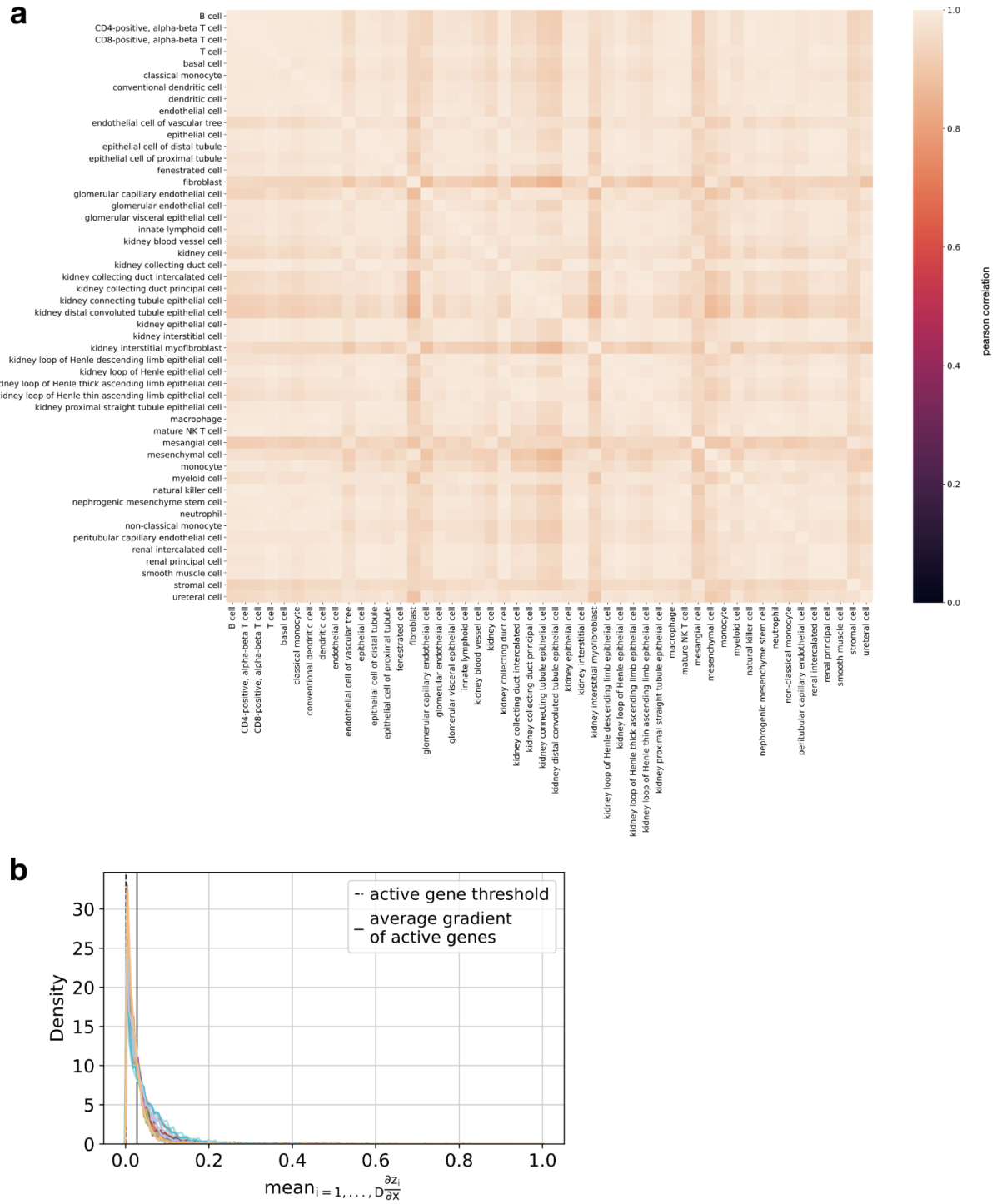


Fig. S4 Saliency-based interpretation of models trained on human kidney. Shown are results for an autoencoder. **(a)** Correlation of cell-type wise aggregated gradients of embedding with respect to input features. **(b)** Distribution of feature-wise aggregated gradients of embedding with respect to input features by cell type (color).