Cell embeddings

# Protein language models

**Rosen, Y., Brbić, M., Roohani, Y. et al. Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. Nat Methods (2024).** [**https://doi.org/10.1038/s41592-024-02191-z**](https://doi.org/10.1038/s41592-024-02191-z)

Deep learning approach that integrates cross-species single-cell RNA-sequencing **(scRNA-seq**) datasets by coupling gene expression with **protein embeddings** generated by large protein language models. SATURN introduces a concept of macrogenes defined as groups of genes that share similar protein embeddings. The strength of associations of genes to macrogenes is learned to reflect this similarity, thereby allowing functionally related genes as captured by the protein embeddings to group together.

SATURN takes as input: (i) scRNA-seq count data from one or multiple species, (ii) protein embeddings generated by a large protein embedding language model like ESM2 (ref. 14), and (iii) initial within-species cell annotations (from cell-type assignments if available or obtained by running a clustering algorithm).

The language model takes a sequence of amino acids and produces a protein representation vector (Fig. 1a). Given gene expression and protein embeddings, SATURN learns an interpretable feature space shared between multiple species. We refer to this space as a macrogene space and it represents a joint space composed of genes inferred to be functionally related based on the similarity of their protein embeddings. The importance of a gene to a macrogene is defined by a neural network weight—the stronger the importance, the higher the value of the weight that connects the gene to the macrogene.

SATURN is composed of three modules:

* Macrogene initialization with Kmeans (scipy)
* Pretraining conditional autoencoder (scVI ZINB loss)
* Fine tuning cell clusters with weakly supervised metric learning

\*\* "Convert protein embeddings to gene embeddings by averaging the protein embeddings for each gene

Diagrama

Descripción generada automáticamente

# Embeddings RNA Secuence

# Embeddings seq-RNA

Universal Cell Embedding (UCE), a foundation model for single-cell gene expression that is designed to address questions in cell and molecular biology. UCE is uniquely able to generate representations of new single-cell gene expression datasets with no model finetuning or retraining while still remaining robust to dataset and batch-specific artifacts. Moreover, it does so while requiring no cell type annotation and no input dataset preprocessing, such as gene selection. UCE can be applied to any set of genes from any species, even if they aren’t homologs of genes seen during training. UCE learns a universal representation of cell biology that is intrinsically meaningful and can extend insights beyond the data that has been experimentally observed. The representations learned by UCE display an emergent organization of cell types that is consistent with known biology. These cell embeddings can be used to accurately predict cell types with no additional model retraining, showing improved performance in dataset integration against existing atlas-scale integration methods.

Integrating single-cell RNA sequencing (scRNA-seq) datasets is challenging for two primary reasons: scRNA-seq data does not always contain the same genes, or features, and those features are plagued by dataset-specific experimental artifacts or batch effects, which means models have to be built separately for each dataset. UCE overcomes these challenges by abstracting cells as ‘bags of RNA’ [31]. UCE (Fig. 1a) converts the RNA gene expression of a single cell into an expression weighted sample of its corresponding genes. Next, UCE represents the sample’s genes by their protein products, using a large protein language model. This allows UCE to meaningfully represent any gene, from any species, regardless of whether the species had appeared in the training data. Finally, after incorporating additional metadata about genes’ chromosomal locations, this representation is fed into a large transformer model [32]. UCE is able to map any cell, from any tissue, or any species, into one shared universal space, with no additional training.

In particular, UCE takes as an input (1) scRNA-seq count data and (2) the corresponding protein embeddings, generated by a large protein language model, ESM2 [33], for the genes in the dataset. The ESM2 protein language model takes amino acid sequences as an input and produces a numerical representation called a protein embedding. Given the expression count data for a cell, UCE takes a weighted and normalized sample, with replacement, of the cell’s genes. This sample can only contain genes which had non-zero expression, and can contain multiple copies of each gene. These genes are then tokenized by converting them to the protein embedding representation of the protein that they code for [34]. Genes belonging to the same chromosome are grouped together by placing them in between special tokens and are then sorted by genomic location. A special token representing the entire cell, the ‘CLS’ token, is appended to the beginning of the cell representation [35]. This combined representation is passed into a transformer neural network. The embedding of a cell is taken as the embedding of the CLS token at the final layer of the transformer

Diagrama

Descripción generada automáticamente