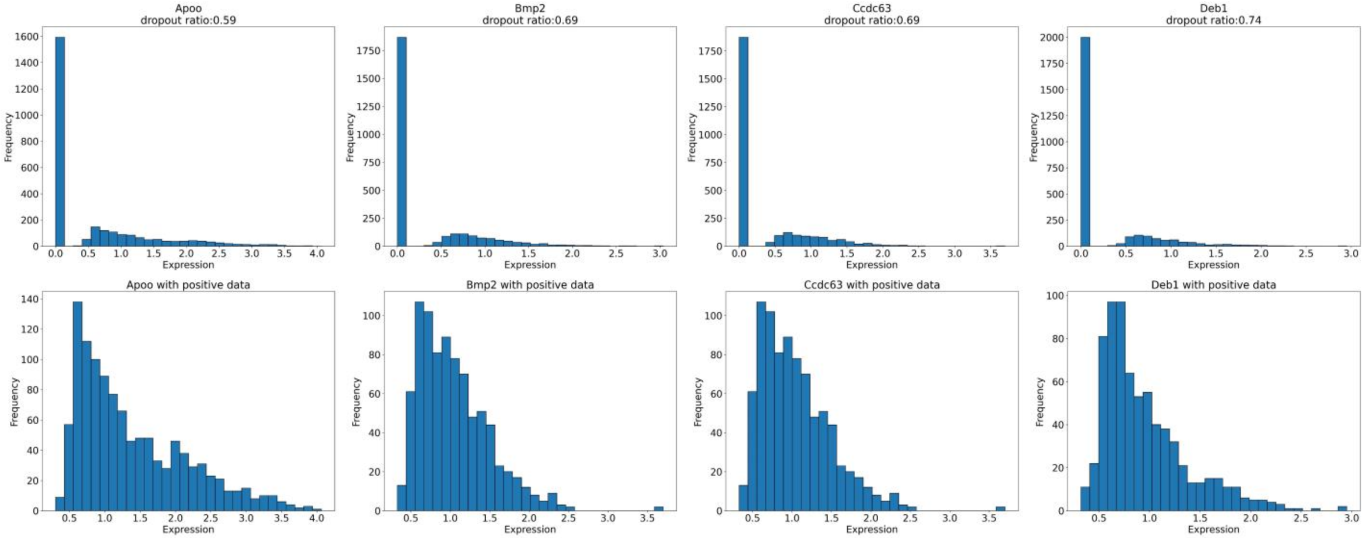
# AE-TPGG: A Novel Autoencoder-Based Approach for Single-cell RNA-seq Data Imputation and Dimensionality Reduction

**Shuchang ZHAO, Li ZHANG, Xuejun LIU**

Frontiers of Computer Science, DOI: 10.1007/s11704-022-2011-y

# Problems & Ideas

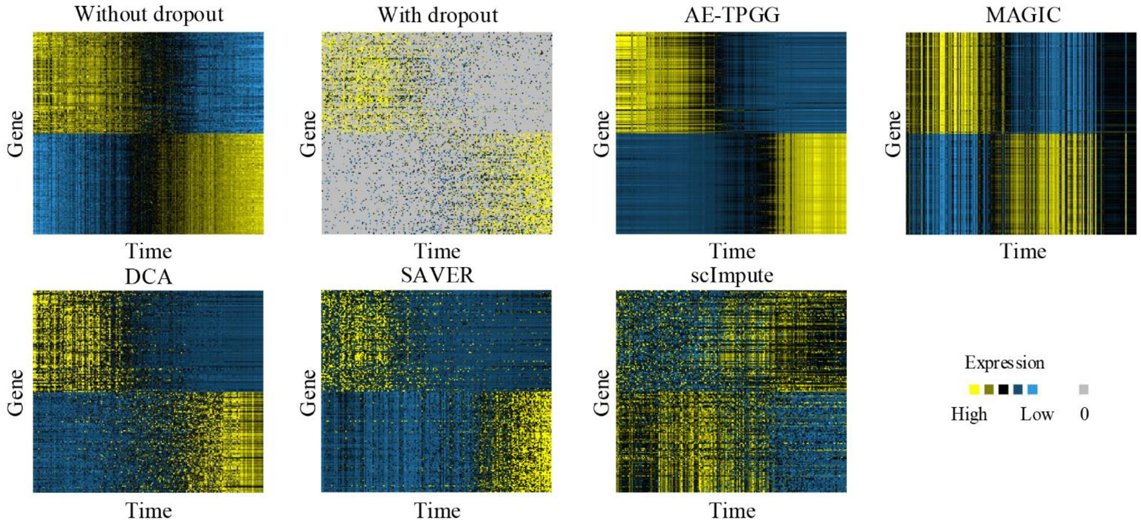
- Problems of conventional scRNA-seq data imputation and dimensionality reduction approaches:
  - The bimodal expression pattern and the right-skewed characteristic existed in normalized scRNA-seq data are still under-explored in current methods.
  - The scRNA-seq data imputation and dimensionality reduction usually need to be realized through two independent models.
- Ideas: The statistical model that accounts for the bimodal expression pattern and the right-skewed characteristic of normalized scRNA-seq data. A joint learning model that takes both scRNA-seq data imputation and dimensionality reduction into account.



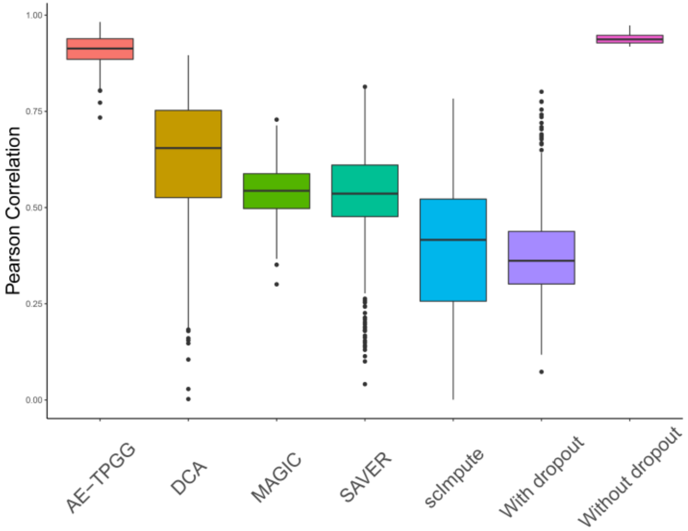
Histograms of the overall and positive expression distribution for Apoo, Bmp2, Ccdc63, and Deb1 genes in Klein dataset.

# Main Contributions

- Contributions:
  - The proposed TPGG distribution for modeling the gene expression that accounts for the bimodal expression mode and the right-skewed distribution of positive expression of of the normalized scRNA-seq data .
  - The proposed AE-TPGG based on autoencoder utilizes loss of gene-specific distribution of TPGG. Not only is the low dimensional representation of cells obtained, but the parameters of distribution can be inferred, leading to data imputation according to first order origin moment of TPGG distribution



Heatmaps of the top 200 highly variable genes that consists of 100 positive genes and 100 negative genes associated with time course within the dataset using expression data without dropout, with dropout, and the imputed data from AE-TPGG, MAGIC, DCA, SAVER, scImpute, respectively. Yellow and blue colors represent relative high and low expression levels, respectively. Zero values are colored grey.



Boxplots of Pearson correlation coefficients between gene expression and the known developmental pattern across the 500 most highly correlated genes within the dataset using the imputed data from AE-TPGG, DCA, MAGIC, SAVER, scImpute, and expression data with dropout, without dropout, respectively.