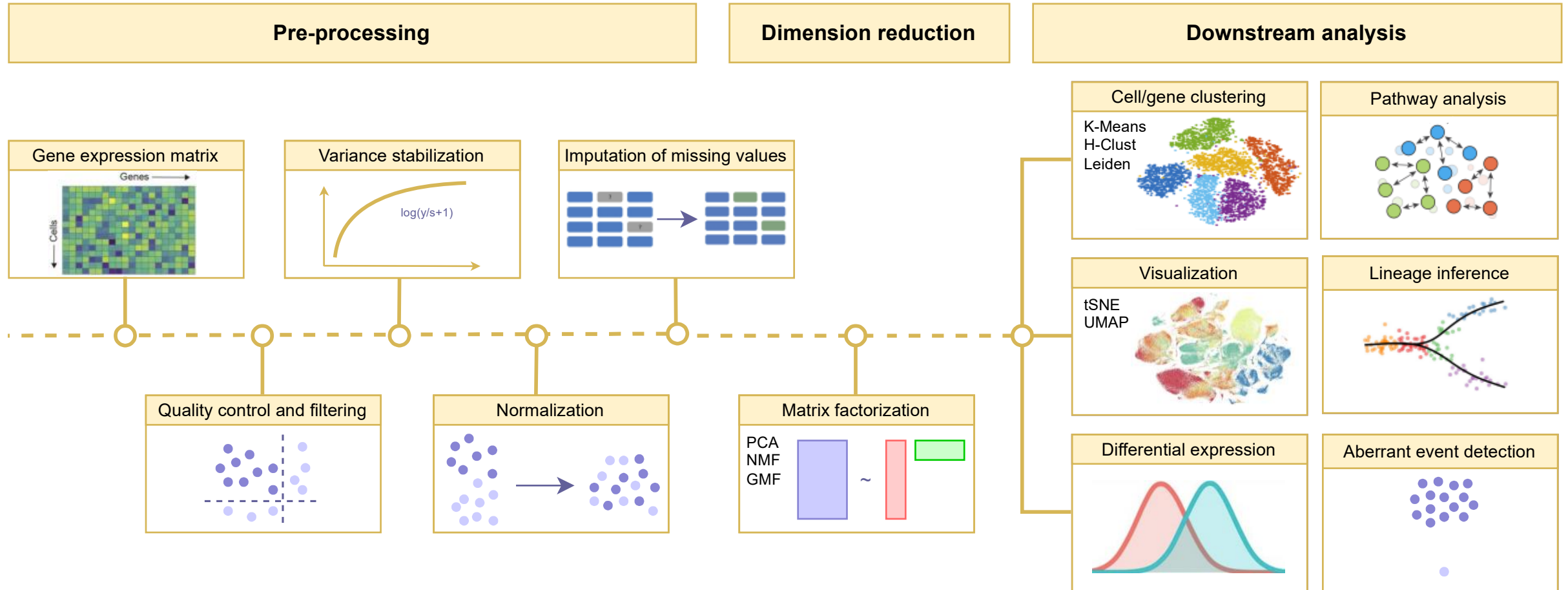


Dimensionality reduction of proteomics and transcriptomics data with omicsGMF

Alexandre Segers, Cristian Castiglione, Christophe Vanderaa,
Lennart Martens, Davide Risso and Lieven Clement
asegers1@its.jnj.com
14th of April 2026

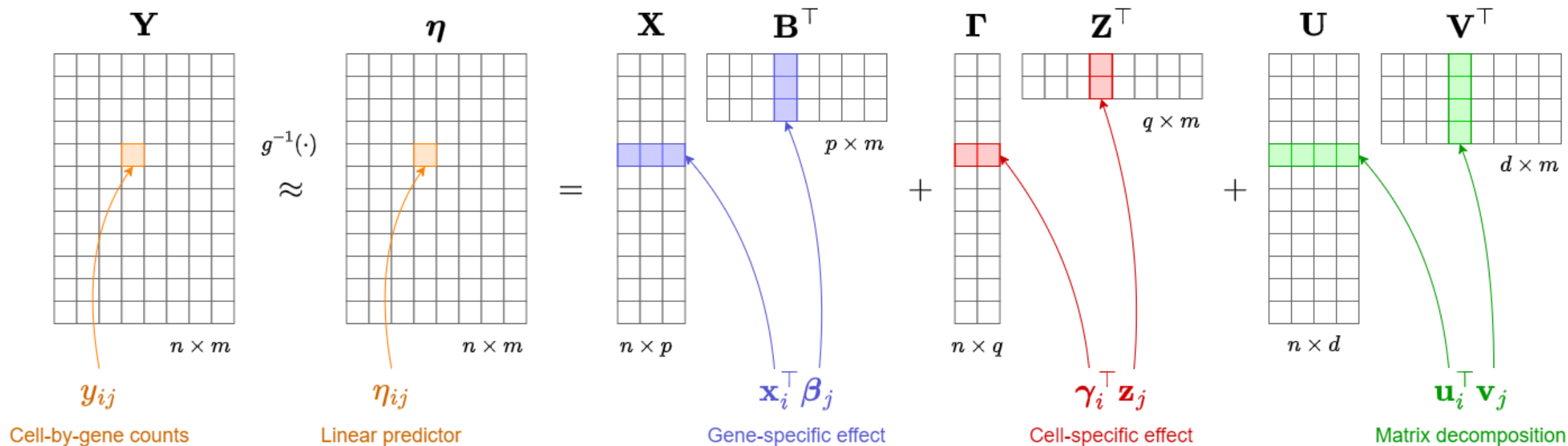
Conventional omics data analysis pipeline



sgdGMF/omicsGMF: Generalized matrix factorization

Model specification

$$y_{ij} \mid \vartheta_{ij}, \phi \sim \text{EF}(\vartheta_{ij}, \phi), \quad \mu_{ij} = b'(\vartheta_{ij}), \quad \nu_{ij} = \phi b''(\vartheta_{ij}), \quad g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{\gamma}_i^\top \mathbf{z}_j + \mathbf{u}_i^\top \mathbf{v}_j$$

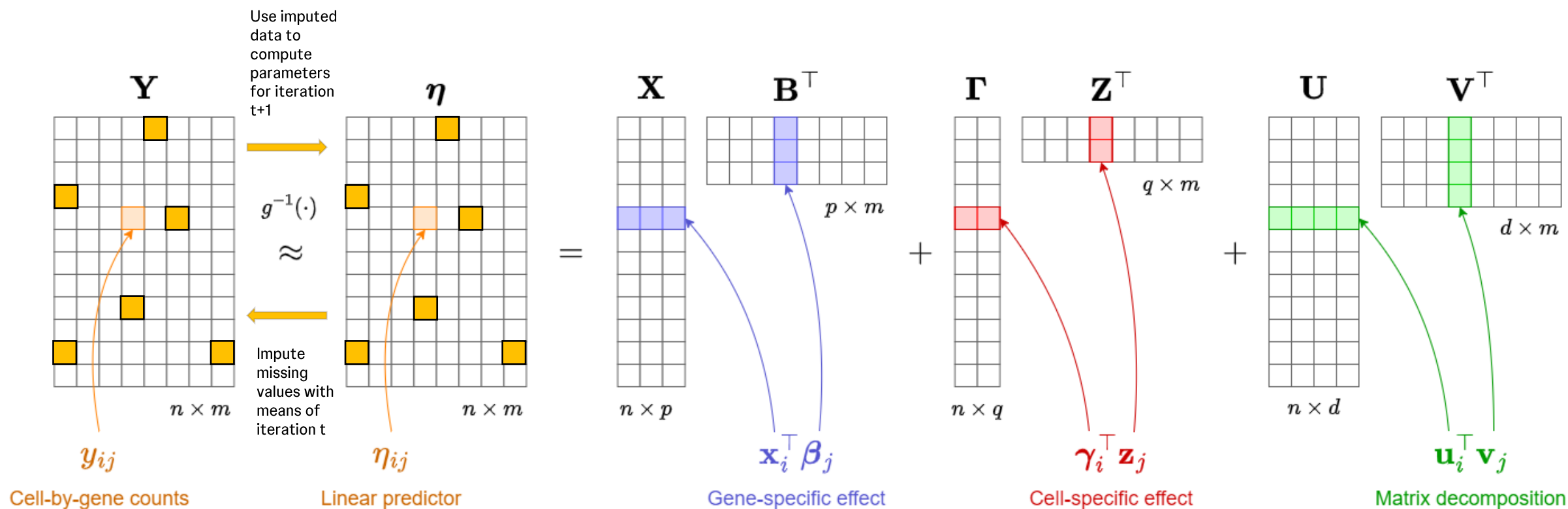


Example: $y_{ij} \mid \vartheta_{ij} \sim \text{Pois}(\vartheta_{ij}), \quad \vartheta_{ij} = \mu_{ij}, \quad \log(\mu_{ij}) = \eta_{ij}$

Imputation using expectation maximization

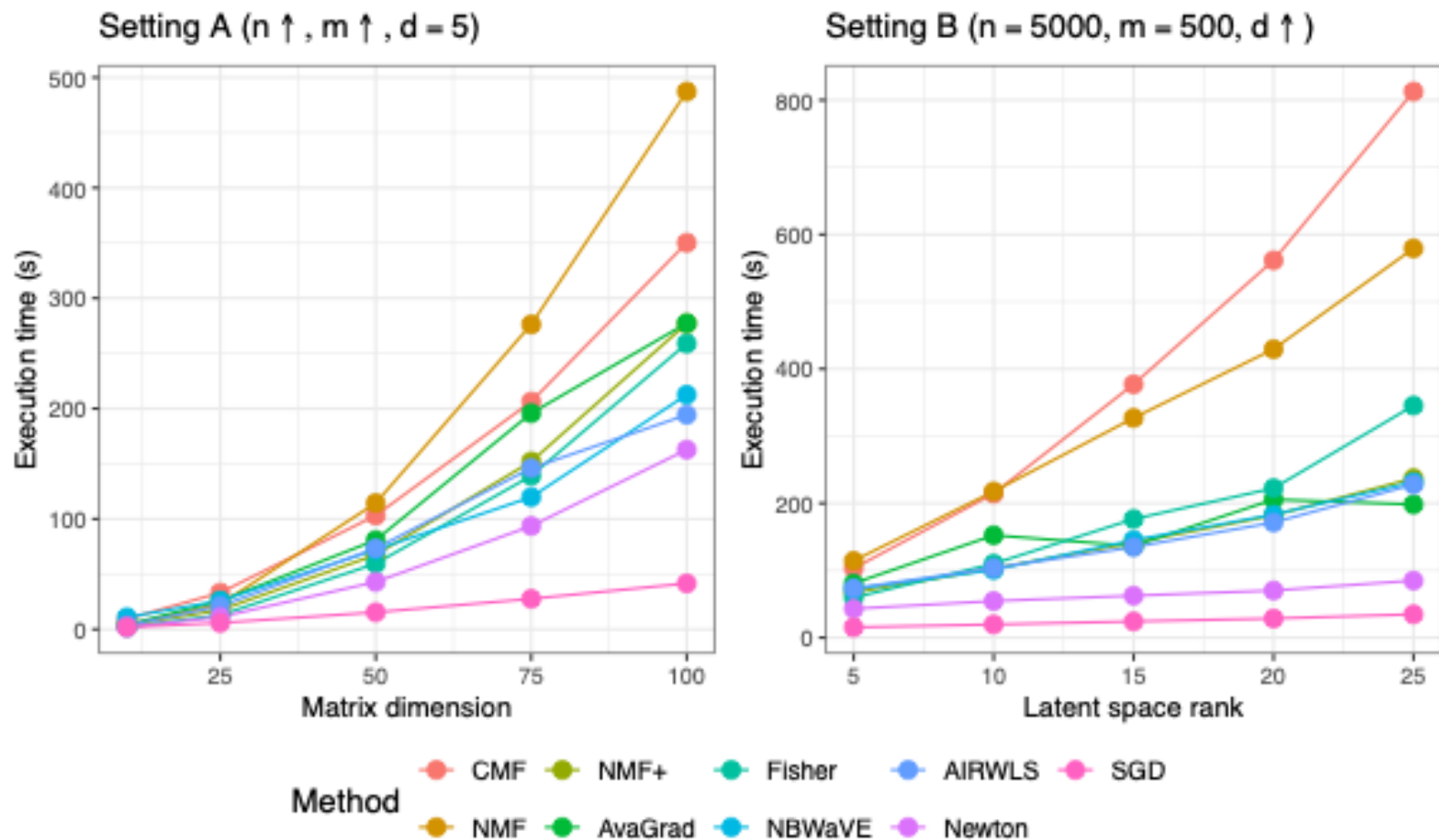
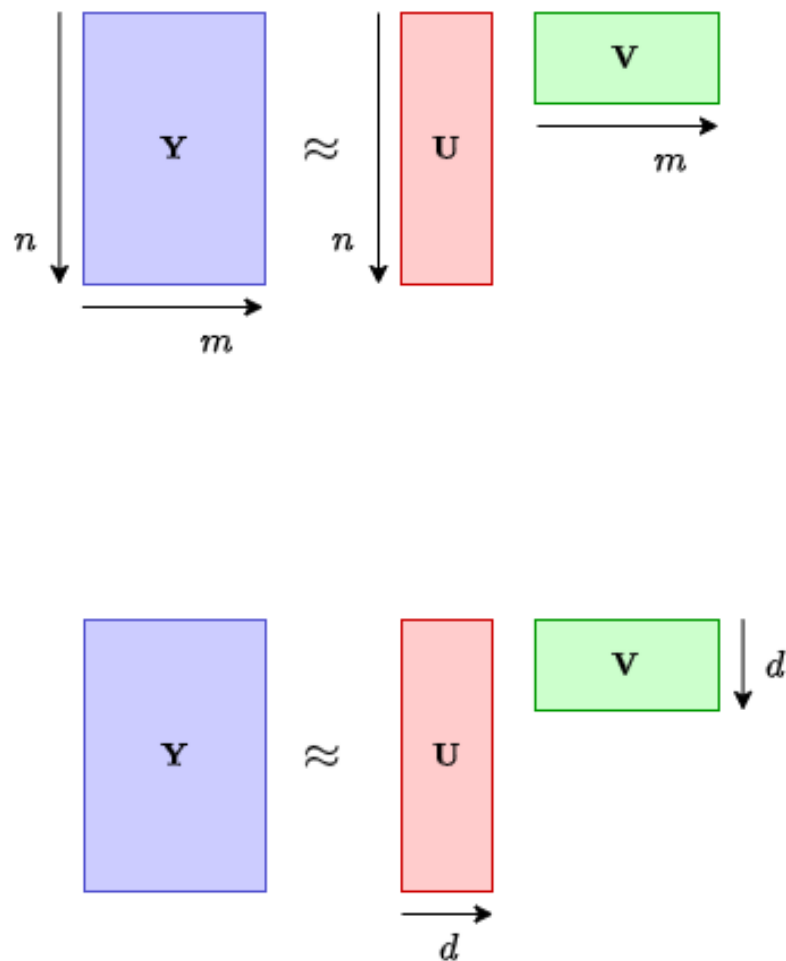
Model specification

$$y_{ij} \mid \vartheta_{ij}, \phi \sim \text{EF}(\vartheta_{ij}, \phi), \quad \mu_{ij} = b'(\vartheta_{ij}), \quad \nu_{ij} = \phi b''(\vartheta_{ij}), \quad g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{\gamma}_i^\top \mathbf{z}_j + \mathbf{u}_i^\top \mathbf{v}_j$$



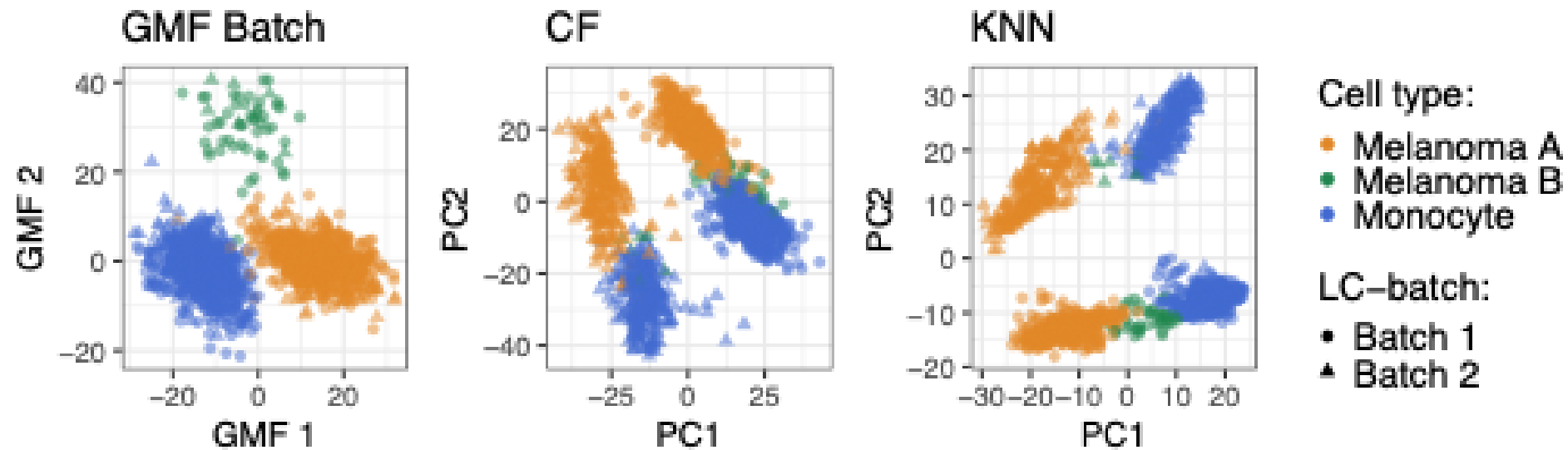
Example: $y_{ij} \mid \vartheta_{ij} \sim \text{Pois}(\vartheta_{ij}), \quad \vartheta_{ij} = \mu_{ij}, \quad \log(\mu_{ij}) = \eta_{ij}$

Scalability for Poisson likelihood

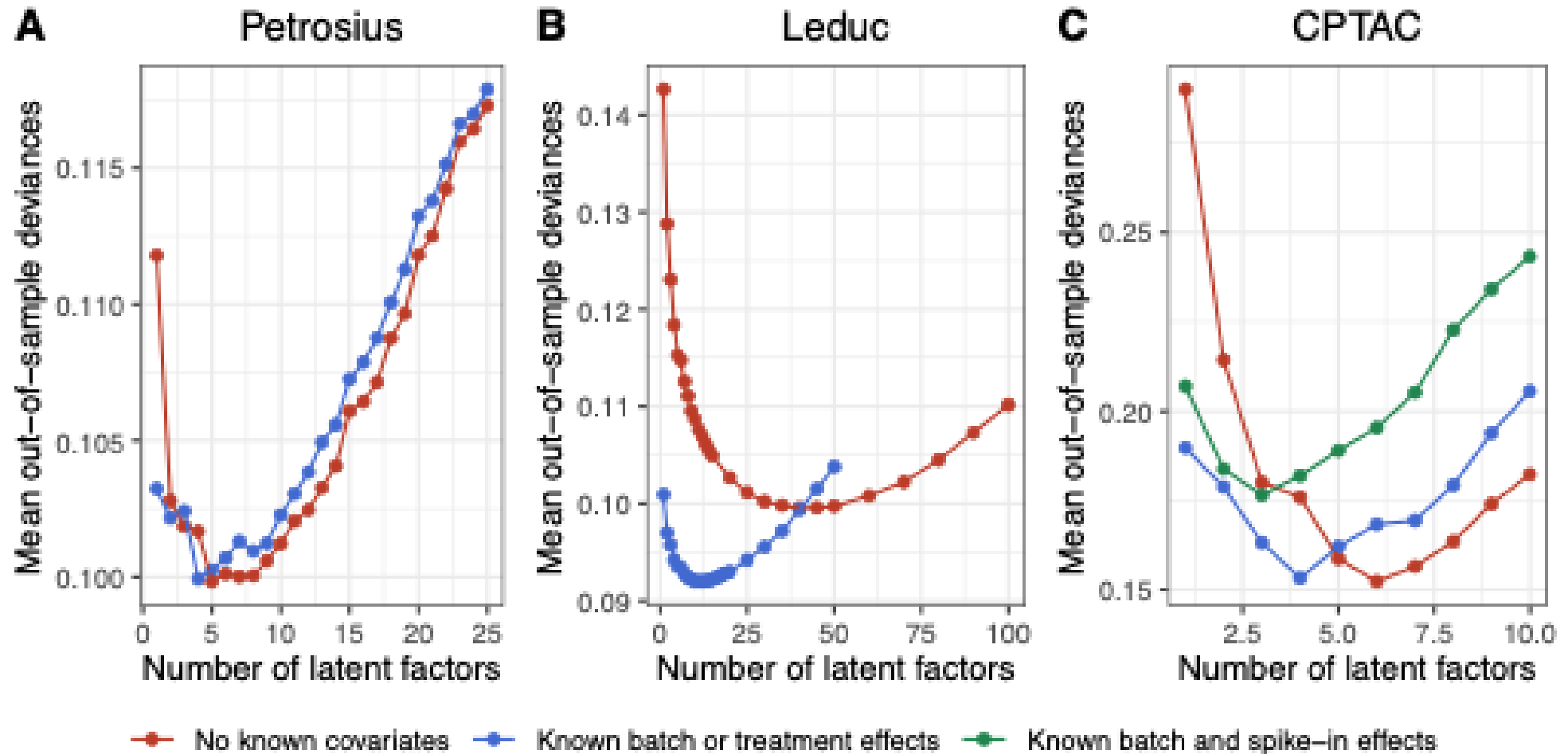


Single cell Proteomics: Leduc data

- 1508 cells from two cell lines: melanoma and monocytes
- 4435 peptides, 61.8% missing values
- TMT-multiplexing → up to 14 cells/MS-run → 142 run effects
- Change Liquid Chromatography column → LC batch effect

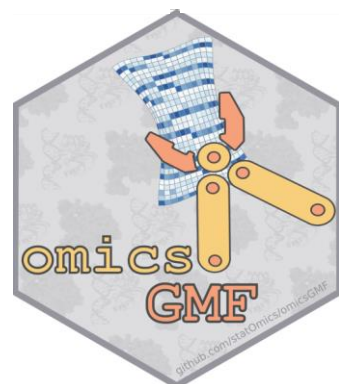


Model Selection via Cross-Validation



sgdGMF/omicsGMF package

A new R package for *stochastic gradient descent* estimation of *generalized matrix factorization*



Likelihood

- Gaussian
- Binomial
- Poisson
- Gamma
- Inverse-Gaussian
- Negative-Binomial
- Quasi-Likelihood

Link

- Identity
- Log
- Sqrt
- Logit
- Probit
- Cauchit
- cLogLog

Initialization

- GLM+SVD
- OLS+SVD
- Random
- Custom

Model selection

- Eigenvalue gap
- Cross-validation
- AIC / BIC

Estimation

- APRALS
- Newton
- SGD

Post-processing

- Fit refinement
- Matrix completion
- Residual analysis
- Prediction
- Screeplot
- Biplot

Thank you

- Castiglione, C., Segers, A., Clement, L., & Risso, D. (2026). Stochastic gradient descent estimation of generalized matrix factorization models with application to single-cell RNA sequencing data. *Biostatistics*, 27(1), kxag010.
- Segers, A., Castiglione, C., Vanderaa, C., Martens, L., Risso, D., & Clement, L. (2026). omicsGMF: a multi-tool for dimensionality reduction, batch correction and imputation in bulk-and single-cell proteomics. *Nature Communications*.

If you have more questions, please contact:

Alexandre Segers

asegers1@its.jnj.com