Dalle Molle Institute for Artificial Intelligence USI – SUPSI

Swiss Institute of Bioinformatics

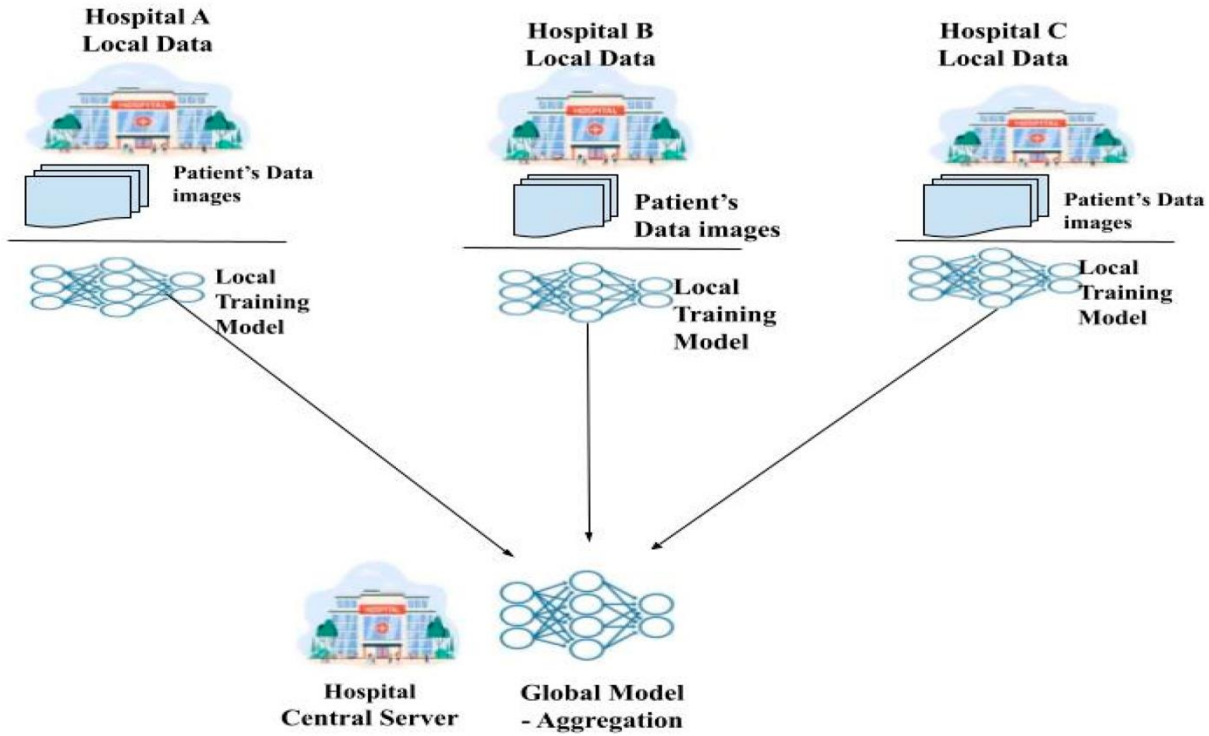
# Federated Learning in Bioinformatics: An Overview

Francesco Gualdi , PostDoc researcher, IDSIA - USI/SUPSI

SUPSI

# Why FL?

- Centralizing clinical records for deep learning raises concerns:
  - Privacy risks
  - Data ownership issues
  - Legal restrictions
- **Balancing the need for shared knowledge with protecting sensitive medical data**.
- **Privacy-preserving techniques** offer a solution to use multi-center data securely.
- **Federated Learning (FL):**
  - Trains large ML models across multiple centers
  - No need to share raw patient data
  - Preserves privacy and security

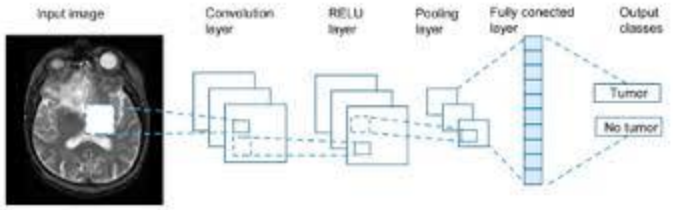


Dalle Molle Institute for Artificial Intelligence USI – SUPSI

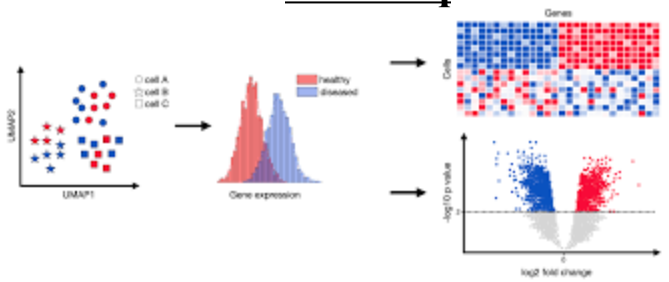
Swiss Institute of Bioinformatics

# Current applications in Bioinformatics

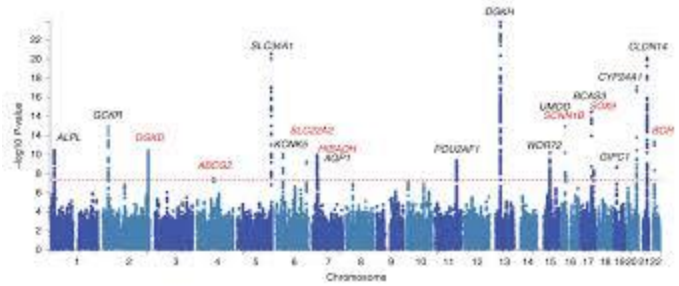
## Medical Imaging

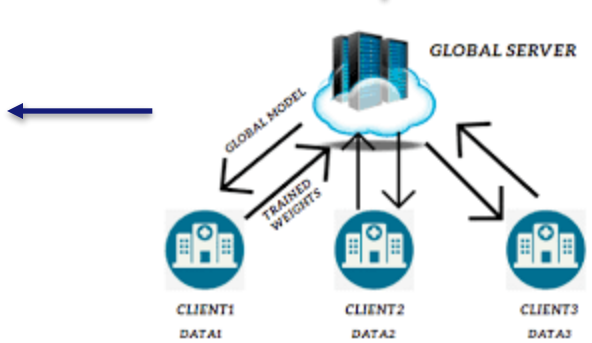



## RNA-seq

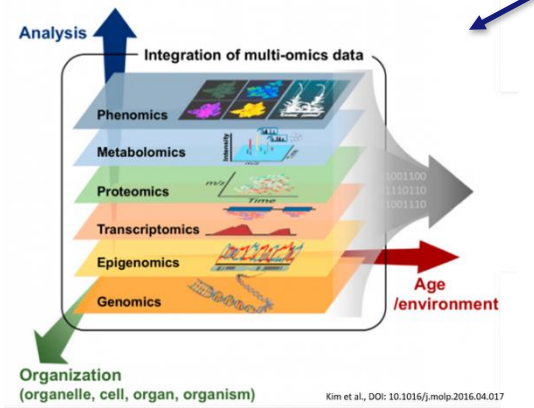



## Genome Wide association Studies (GWAS)

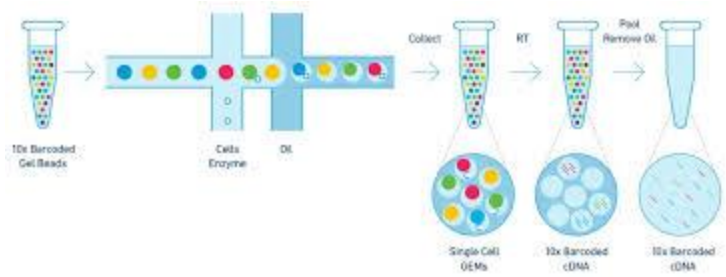



## Multi omics




## scRNA-seq



SUPSI

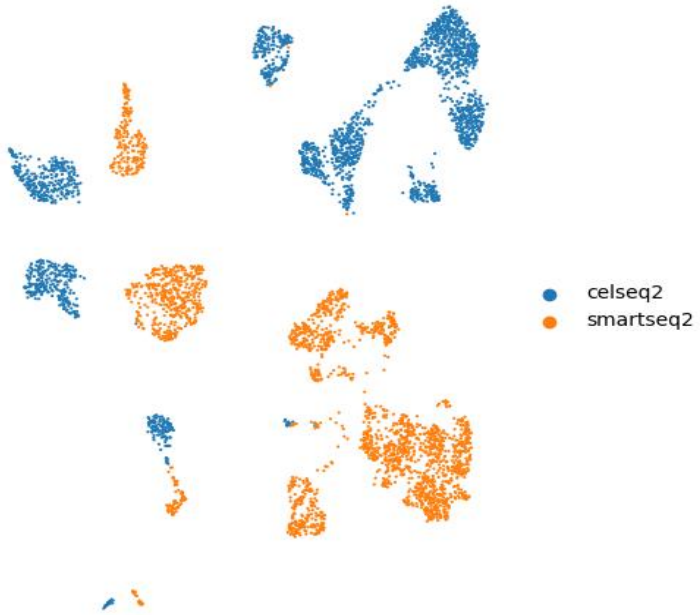
# Key Challenges in FL applied to Bioinformatics

- Data heterogeneity (different labs, protocols, platforms) -> Batch effect
- Scalability (large cohorts, omics data) -> Computational challenges
- Privacy and Data generation costs -> Difficulty in data managing

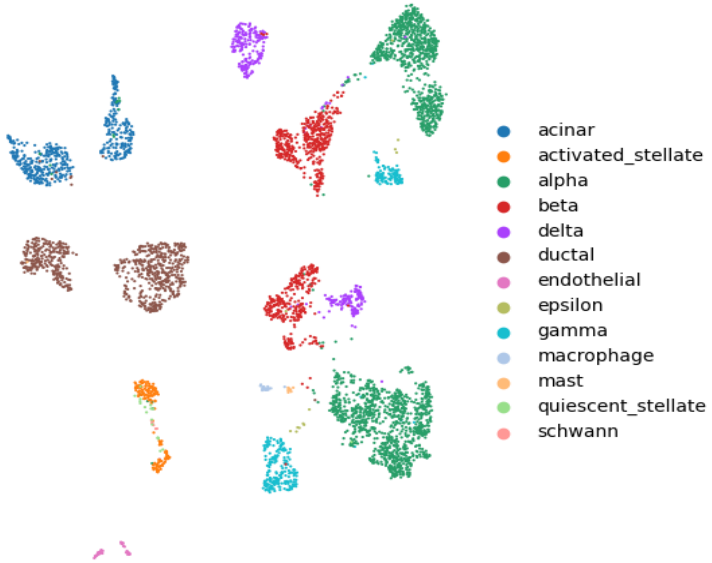
# Data Heterogeneity

## UMAP on scRNA-seq data

### Batch

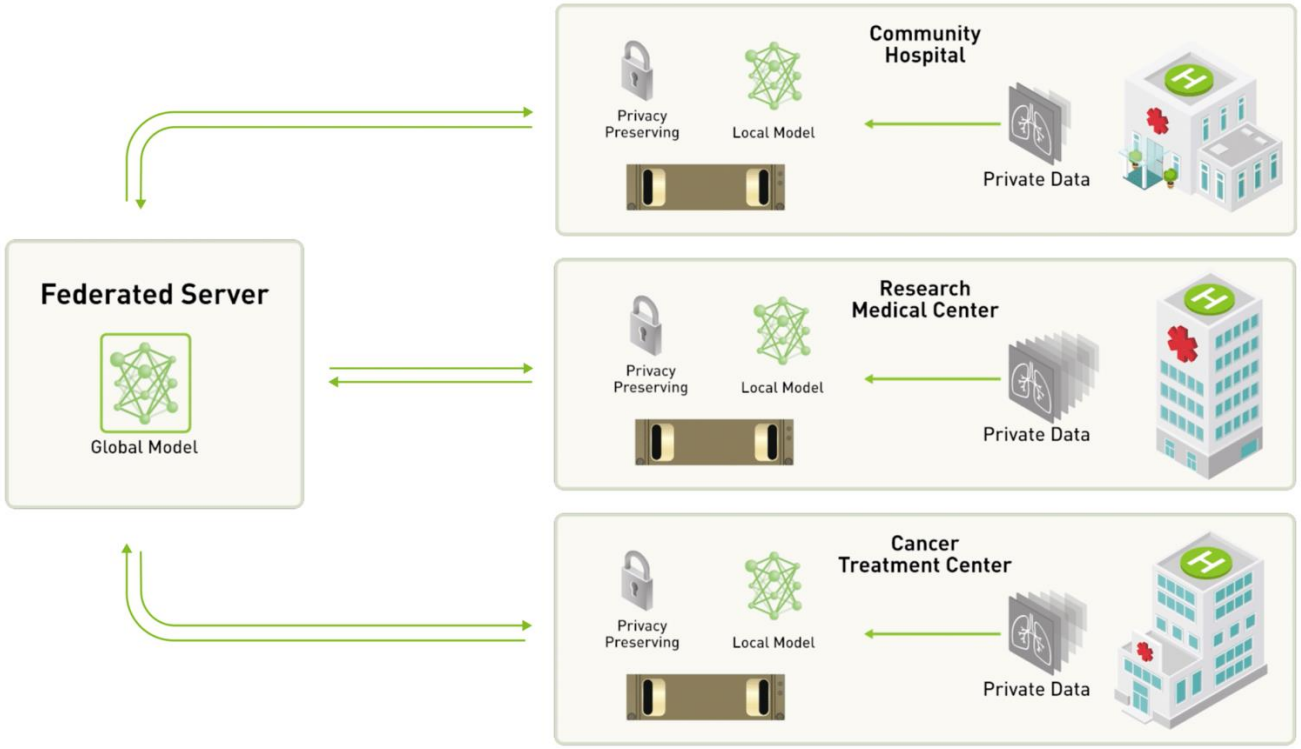



### Cell type



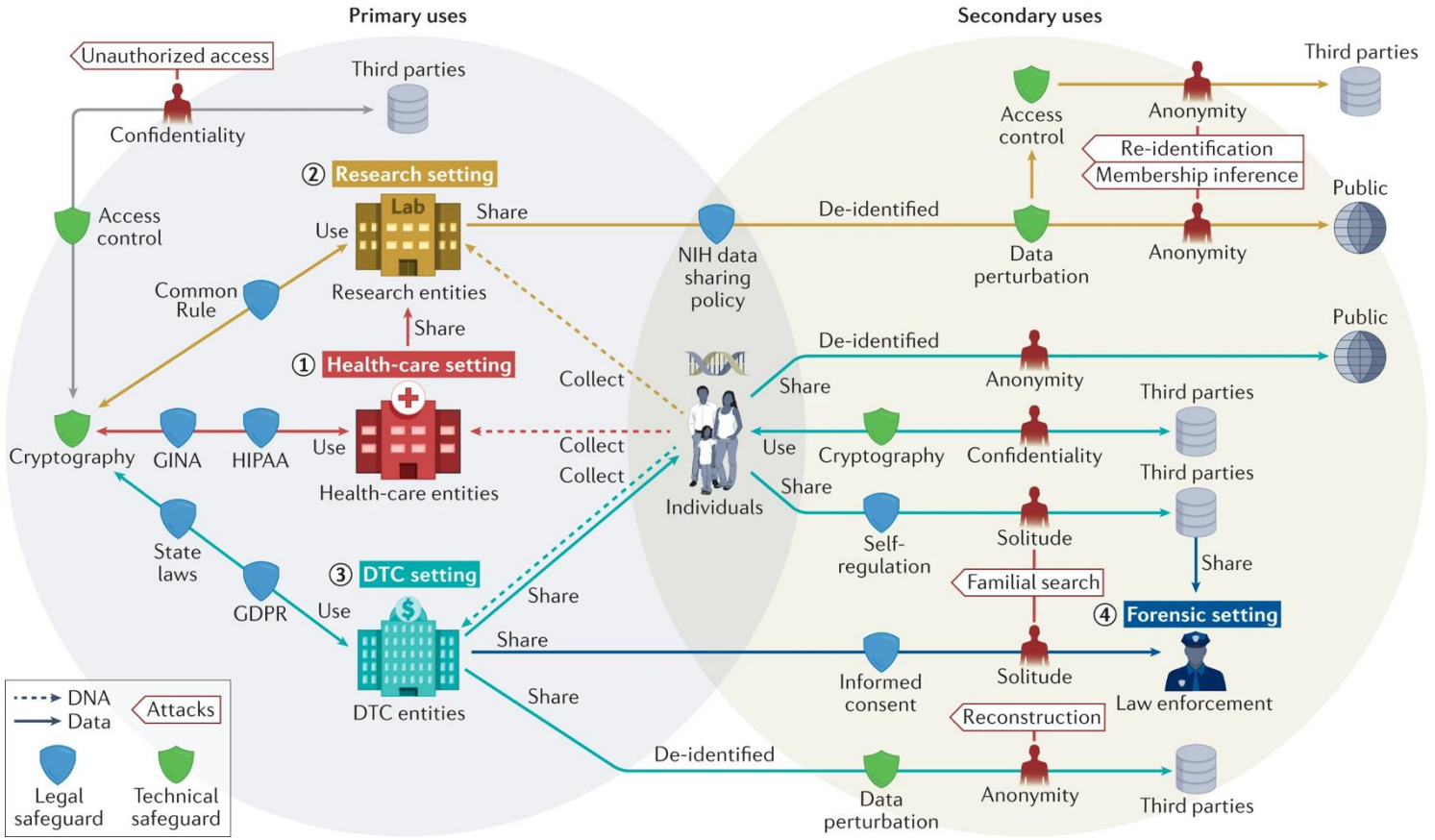

- Same Cell populations in different Clients have different distributions
- This can lead to biased results
- Need to batch mitigation strategies

# Scalability




- **Communication overhead:** Exchanging model updates across many clients increases bandwidth and latency costs.
- **System heterogeneity:** Clients differ in computational power, storage, and network reliability.
- **Server bottleneck:** A central aggregator may struggle to handle massive simultaneous client connections

# Privacy and data generation



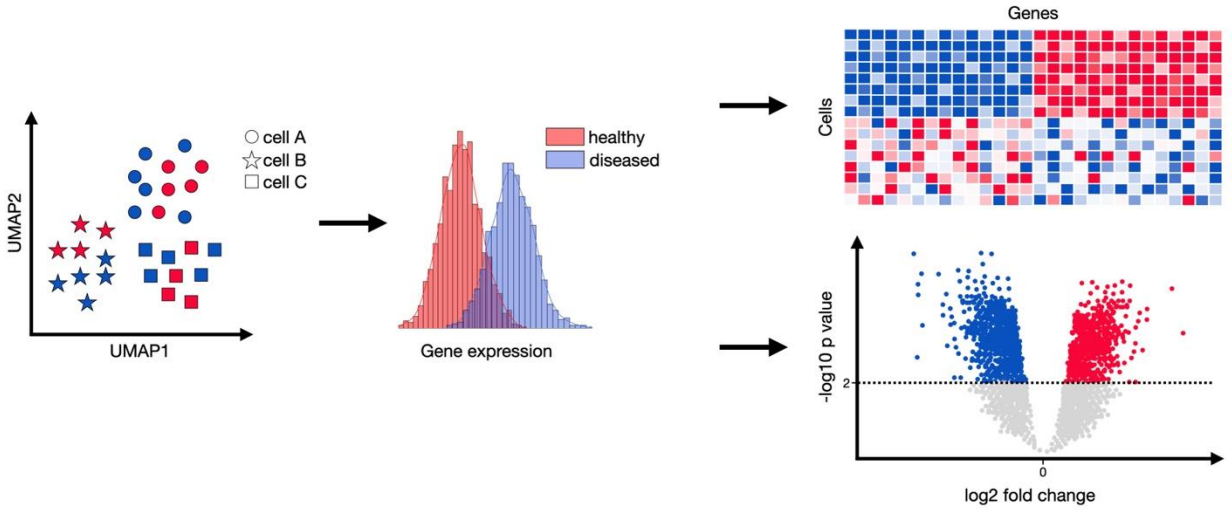

- Even with FL, model updates may leak private patterns (e.g., membership inference attacks, i.e. training an AI model in order to share sensible informations).
- Data generation has costs institutions are not keen in sharing it
- Companies using research data for profit in ways not aligned with participants' consent or public interest.

# Applications of FL in Bioinformatics

| Field | Main Task | Example FL Methods |
|---|---|---|
| Proteomics & Gene Expression | Differential expression, cell type classification | FedProt, HyFed (limma voom), DL (Flower, TFF) |
| GWAS | Association testing, scalable regression | FedGLMM, FedGMMAT, REGENIE (MPC/HE) |
| Single-cell RNA-seq | Cell type classification | scFed (ACTINN, SVM, XGBoost, GeneFormer) |
| Multi-omics | Diasgnostics | Vertical FL, adaptive neural networks |
| Medical Imaging | Classification, segmentation | Federated labelling, harmonised feature learning |

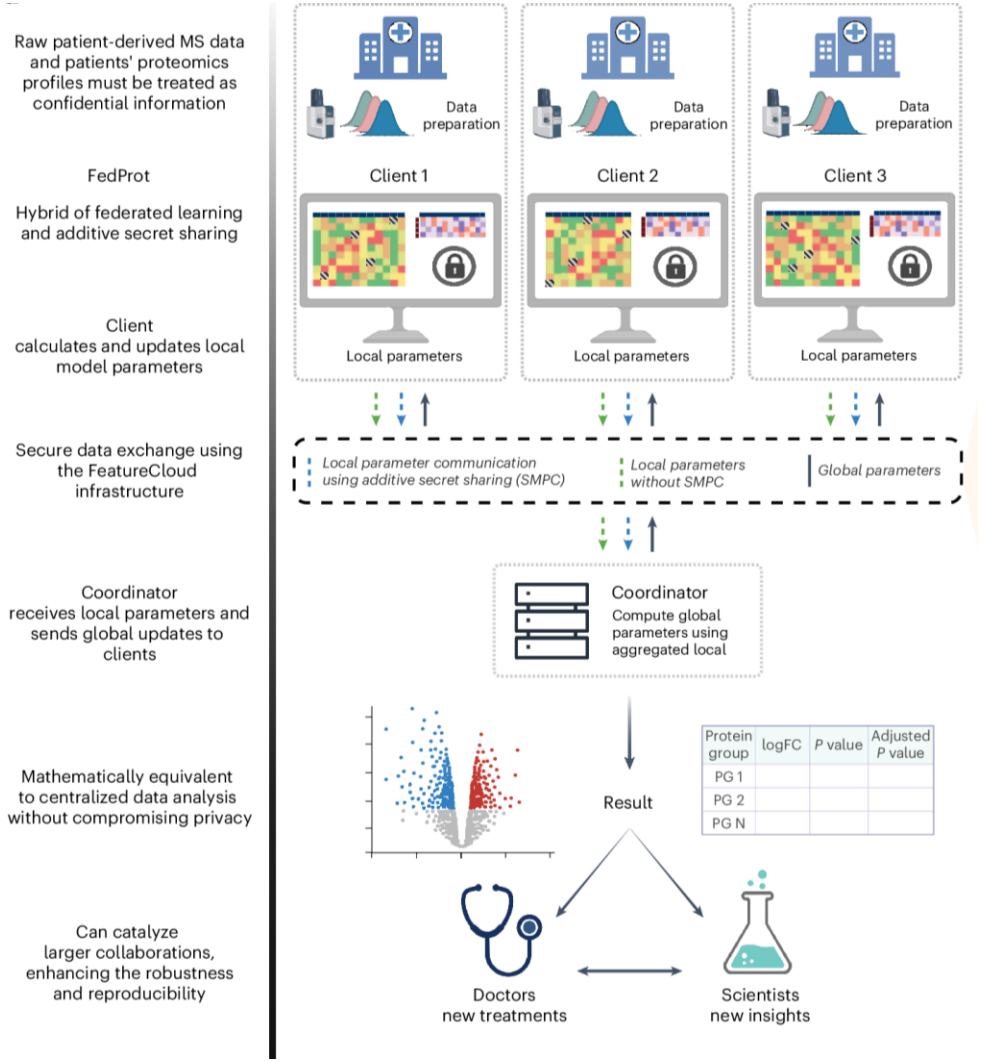
https://doi.org/10.48550/arXiv.2503.09649

# Differential Expression in Proteins and Genes




- Analysis of genes to identify those with significantly different expression levels between conditions (e.g., healthy vs. diseased tissue).
- Discover biomarkers, understand biological processes, and reveal molecular mechanisms of disease.

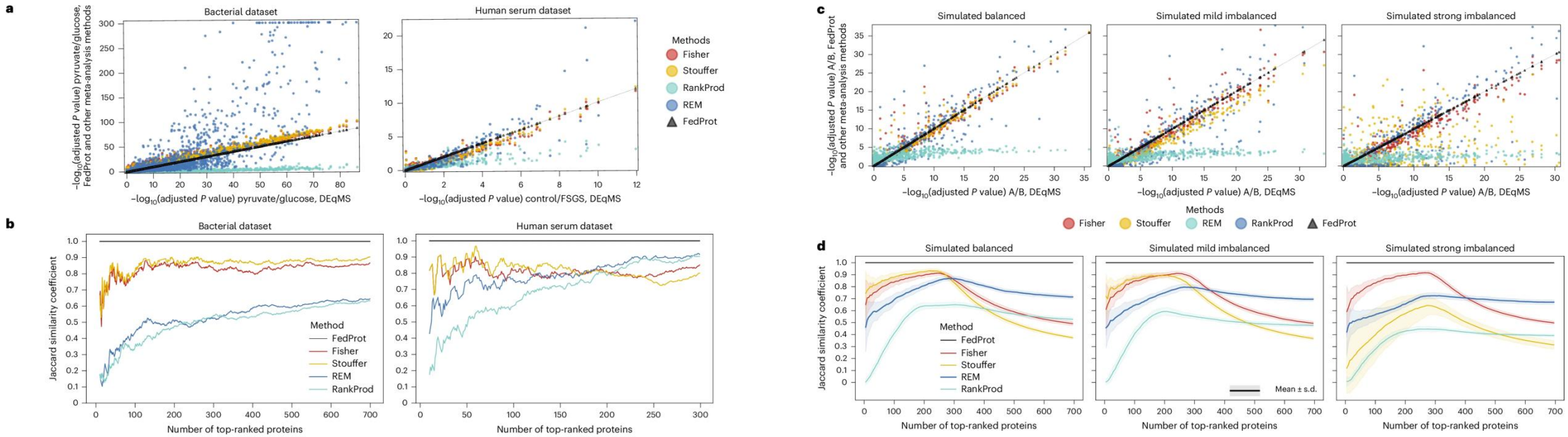
# FedProt




- Clients preprocess MS data; raw proteomics data stay private.
- Clients train a global model via parameter exchange, without sharing raw data.
- During client communication, privacy is maintained through secure multiparty computation
- Final output matches centralized analysis, with fold changes, confidence intervals, and adjusted p-values.

*Burankova, Y., Abele, M., Bakhtiari, M. et al. Privacy-preserving multicenter differential protein abundance analysis with FedProt. Nat Comput Sci 5, 675–688 (2025). https://doi.org/10.1038/s43588-025-00832-7*

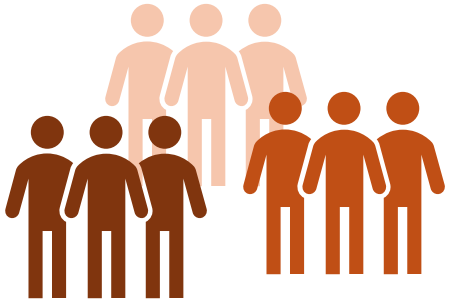
# FedProt



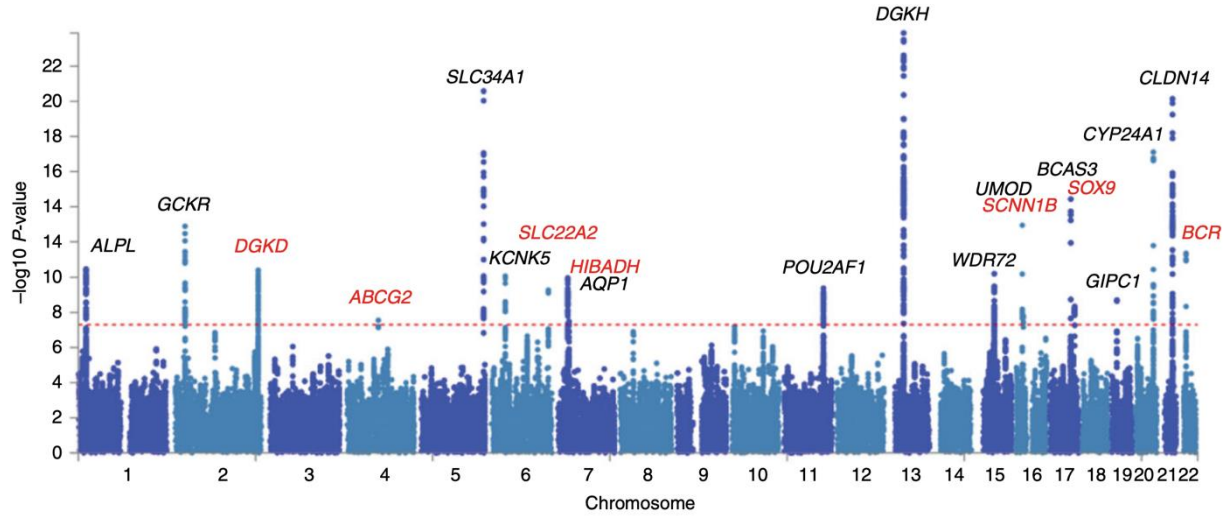

- **Comparison of P values**: −log10(BH-adjusted P) from FedProt/meta-analysis vs. centralized DEqMS (serum and simulated datasets).
- **Jaccard similarity**: Agreement in top-ranked proteins between centralized and decentralized approaches.

# Genome Wide Association Studies (GWAS)

Control Group



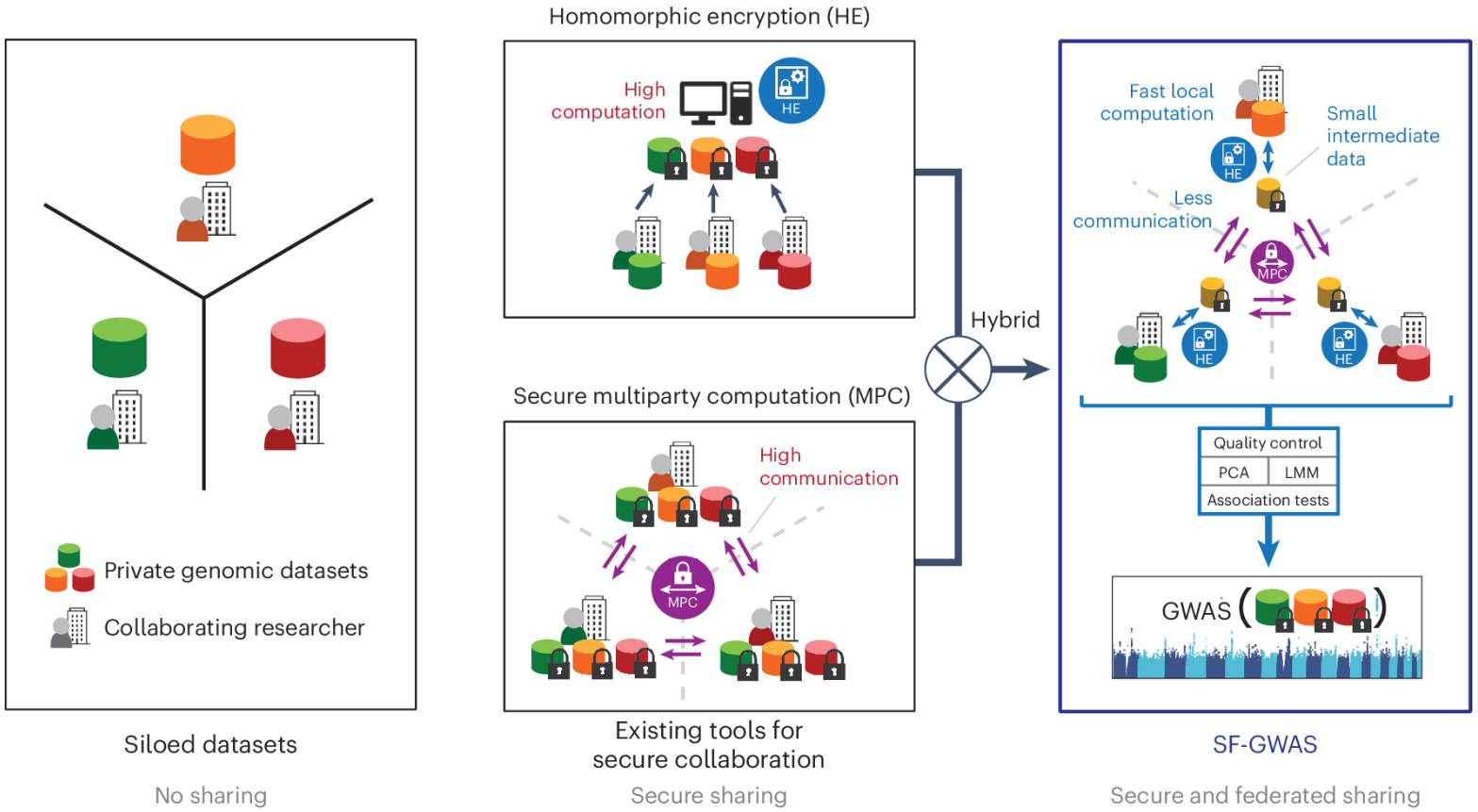
Case Group




*Howles, S.A., Wiberg, A., Goldsworthy, M. et al. Genetic variants of calcium and vitamin D metabolism in kidney stone disease. Nat Commun* ***10,*** *5175 (2019). https://doi.org/10.1038/s41467-019-13145-x*

- GWAS have successfully identified risk loci for a vast number of complex diseases
- Test association of every genotyped locus variant against a phenotype.
- The association is done by multiple regression using all subjects variants and phenotype.
- Regression allows to include other variables (covariates) on the equation.

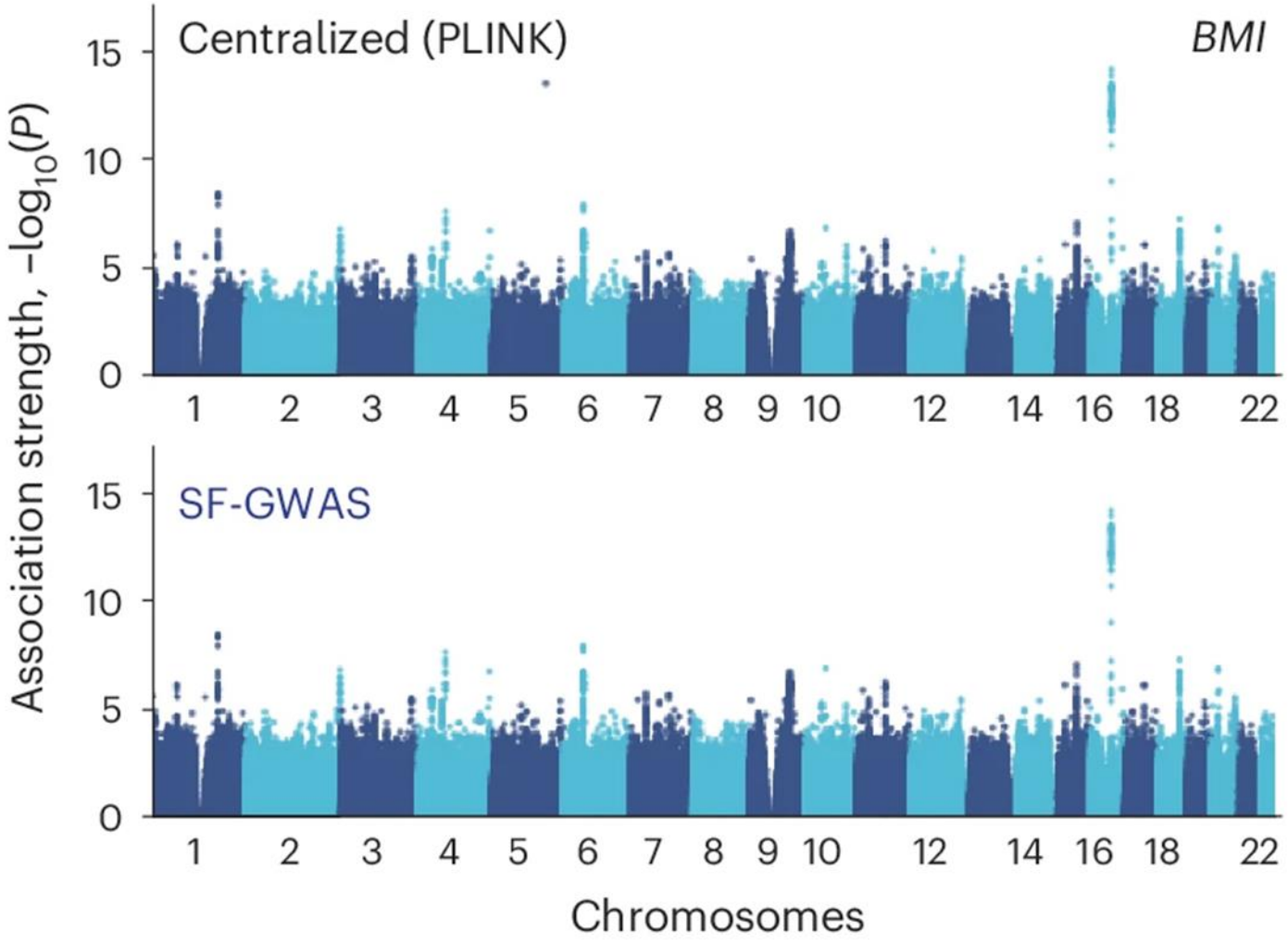
# Secure Federated (SF-GWAS)




- SF-GWAS is a secure, federated algorithm for multisite genome-wide association studies.
- Enables collaborative genomic studies with cryptographic privacy guarantees.
- Each site keeps its genomic dataset locally, reducing data transfers and computational costs.
- Combines MPC (multiparty computation) and HE (homomorphic encryption) in a hybrid framework.
- improved precision and efficiency.

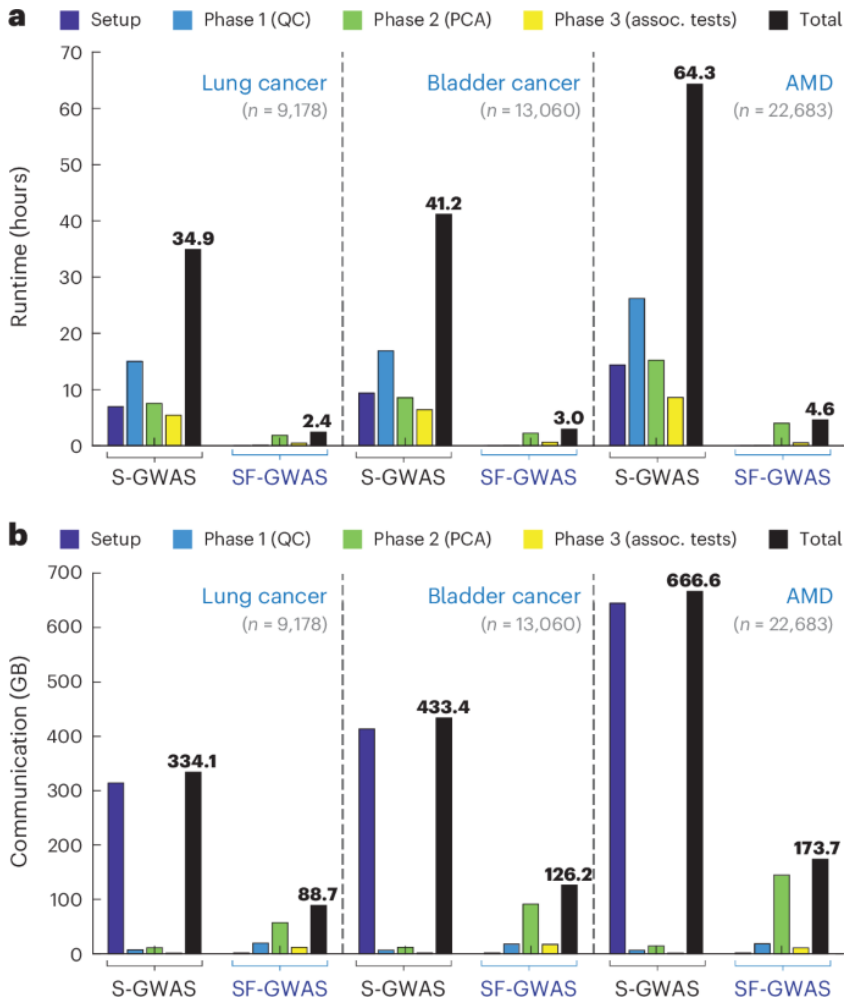
*Danek, Benjamin et al. "Federated Learning for multi-omics: a performance evaluation in Parkinson's disease." bioRxiv : the preprint server for biology 2023.10.04.560604. 12 Feb. 2024, doi:10.1101/2023.10.04.560604. Preprint.*

# SF - GWAS



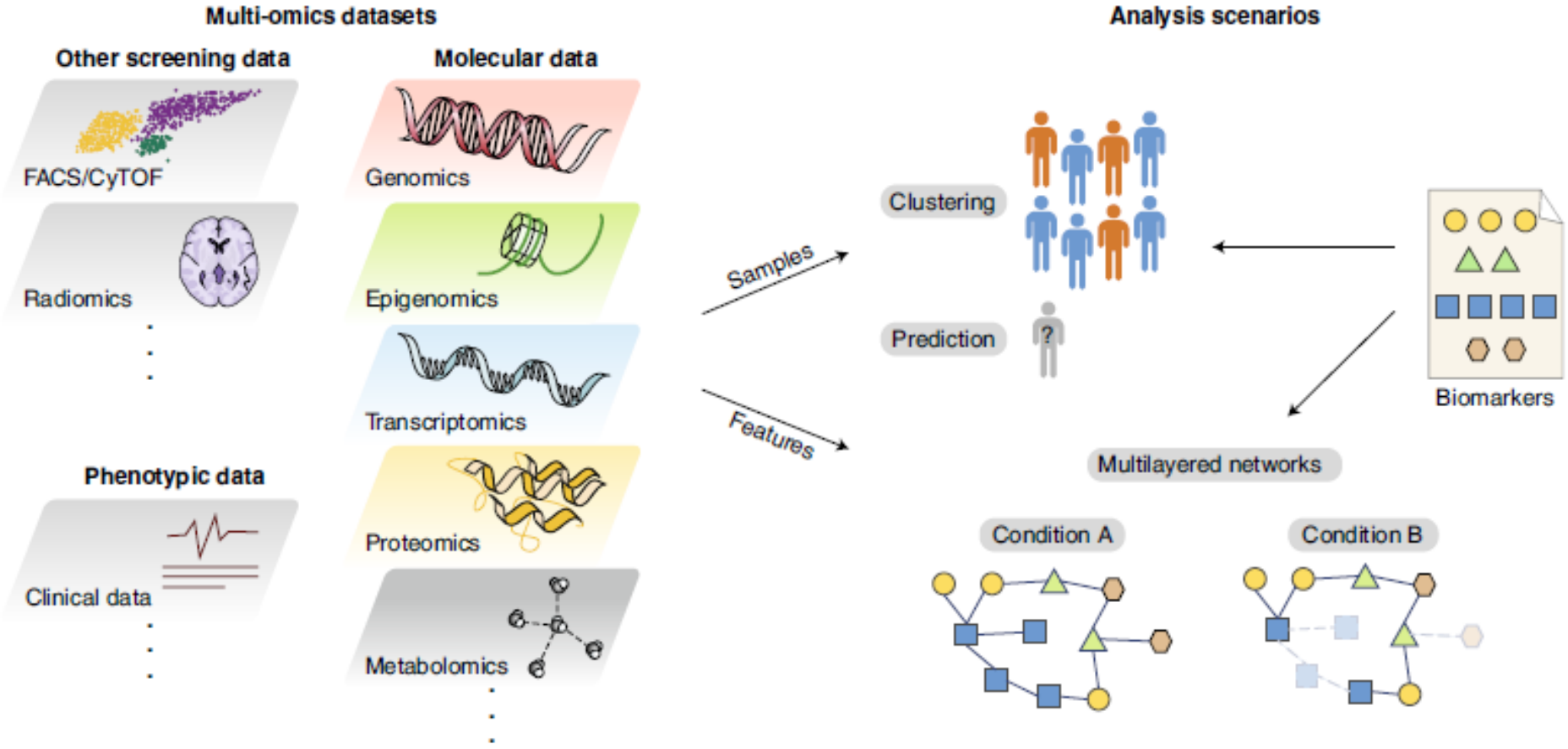

Replicates same results of centralized Dataset

# SF - GWAS



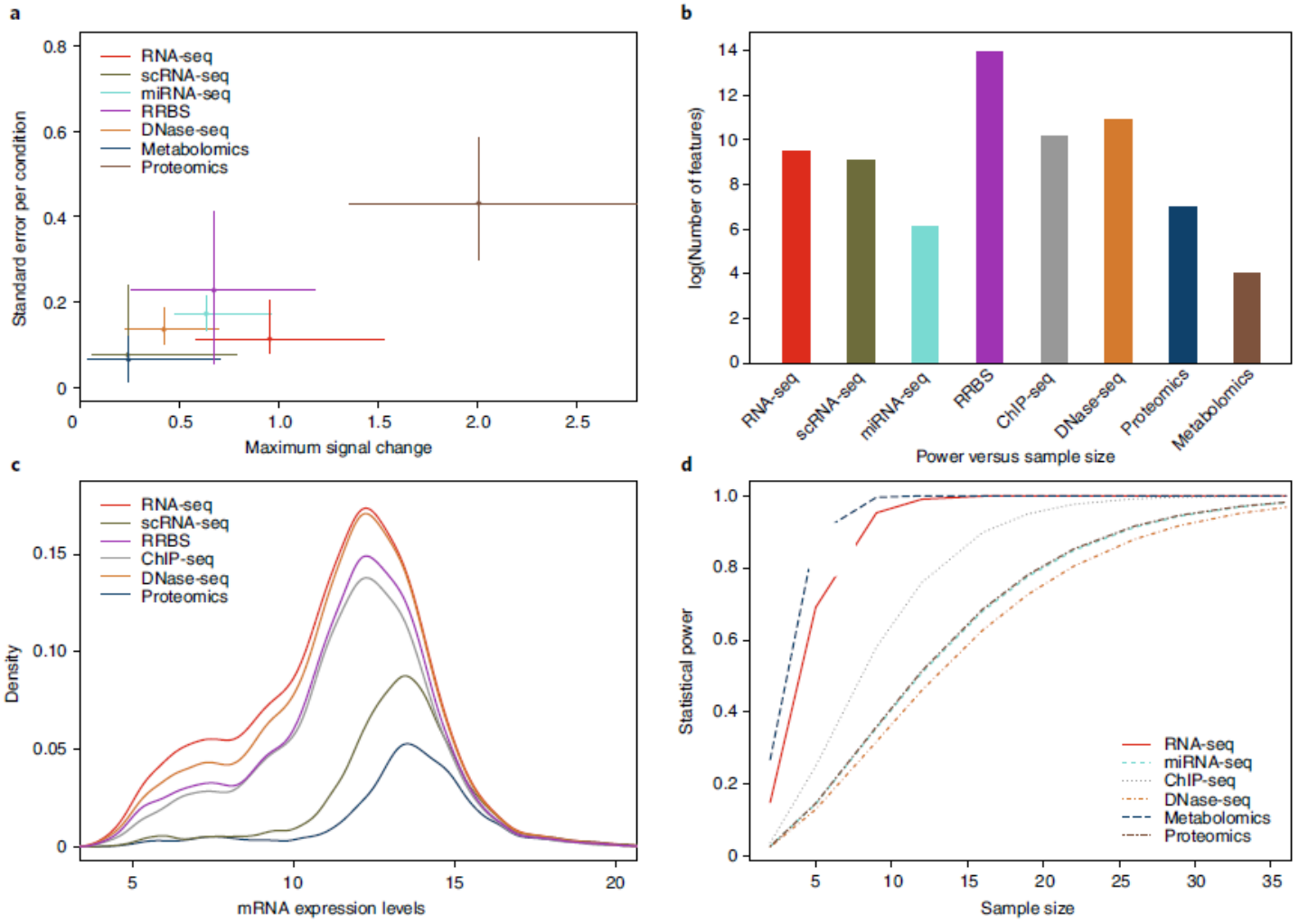

- Comparison with S-GWAS implementing cryptographic methods both implement full GWAS steps (QC, PCA, association tests).
- **Datasets**: Lung cancer, bladder cancer, and Age related Macular Degeneration; split across two machines to simulate collaboration.
- SF-GWAS avoids the encrypted data transfer required by S-GWAS, thanks to its federated design.
- **Results**: SF-GWAS is ~10 × faster and ~3.5 × more communication-efficient.

# Multi - Omics




*Tarazona, S., Arzalluz-Luque, A., & Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. Nature Computational Science, 1(6), 395–402. https://doi.org/10.1038/s43588-021-00086-z*

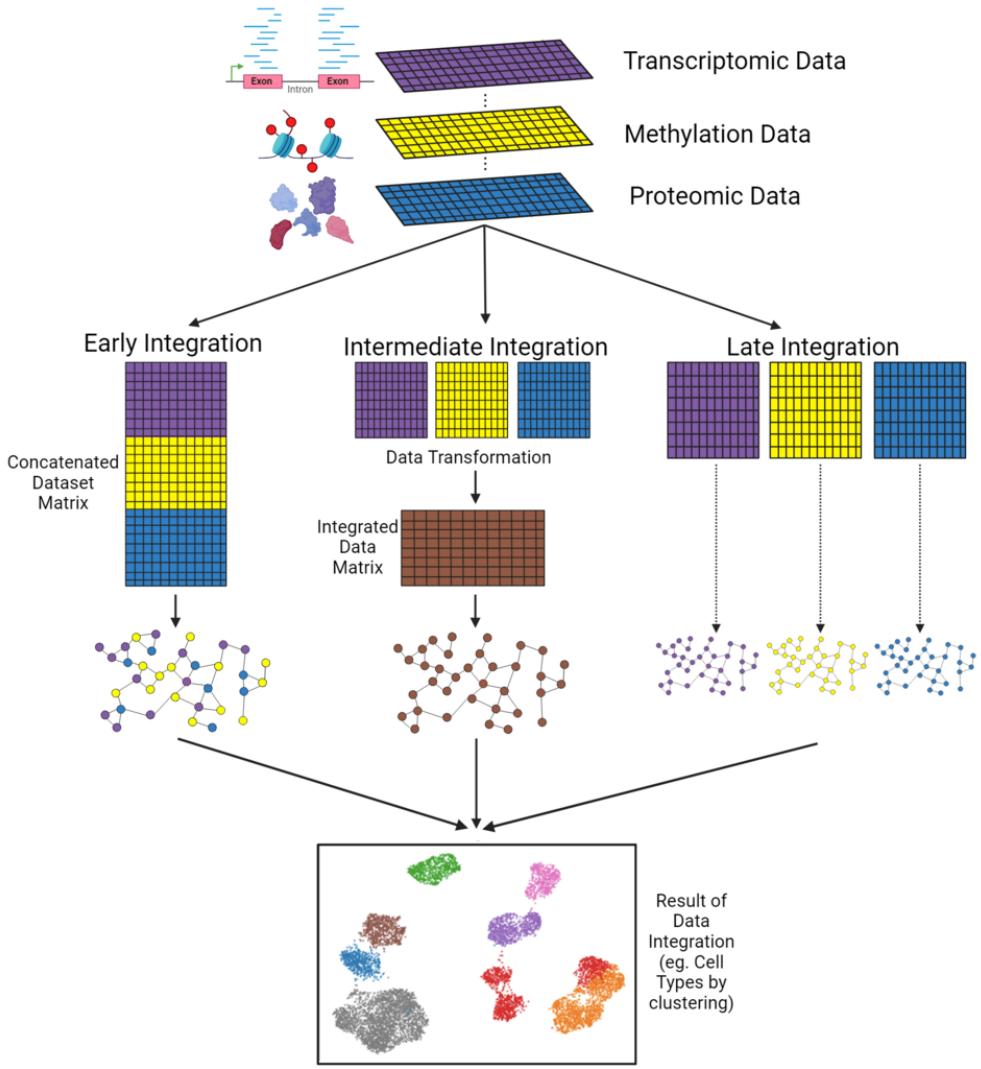
# Multi - Omics




**Comparison of the properties of omics data types.**

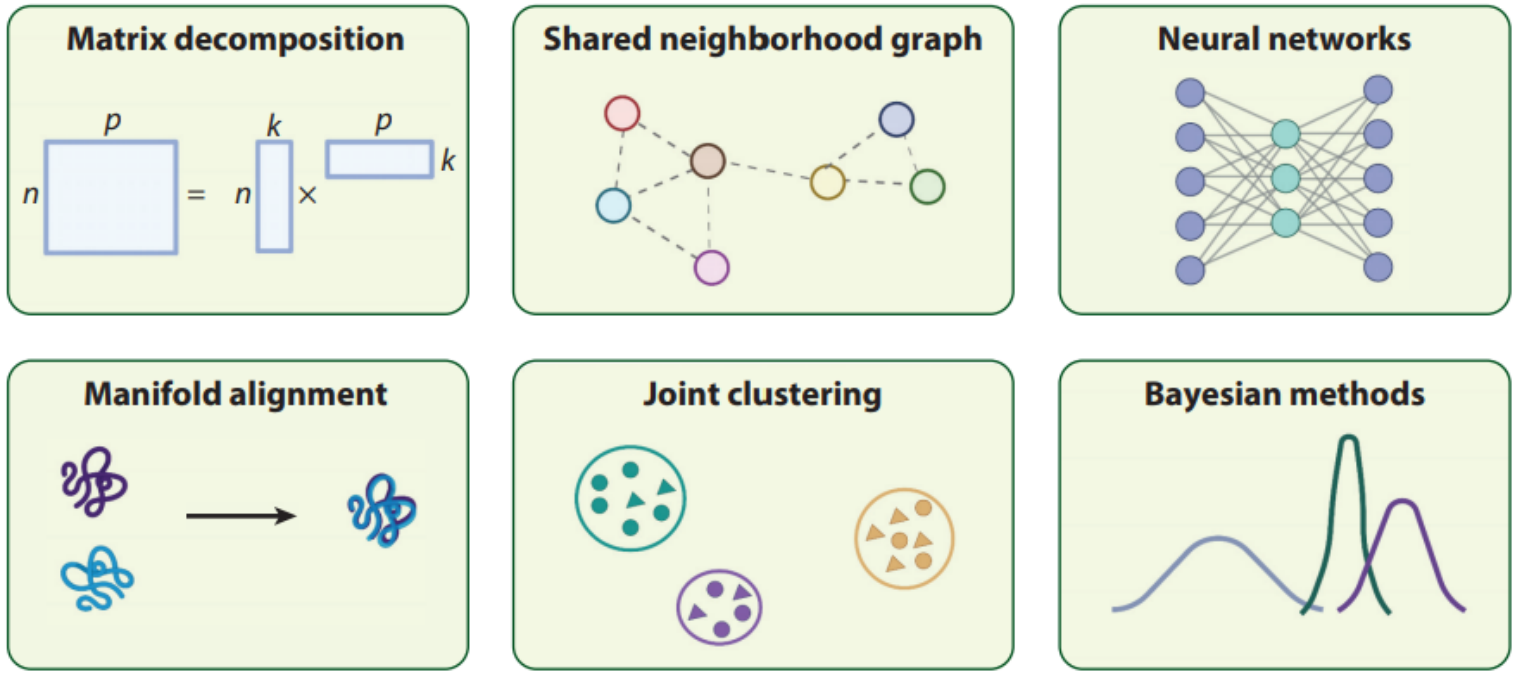
**A**: Signal-to-noise plot. Segments represent the interquartile ranges of the maximum signal change
**B**: Number of features detected by each technology.
**C** Differential coverage of the feature space. Each density line represents the distribution of the fraction of expressed genes captured by each method.
**D**: Statistical power curves across omics data types as a function of sample size

# Multi – Omics integration strategies



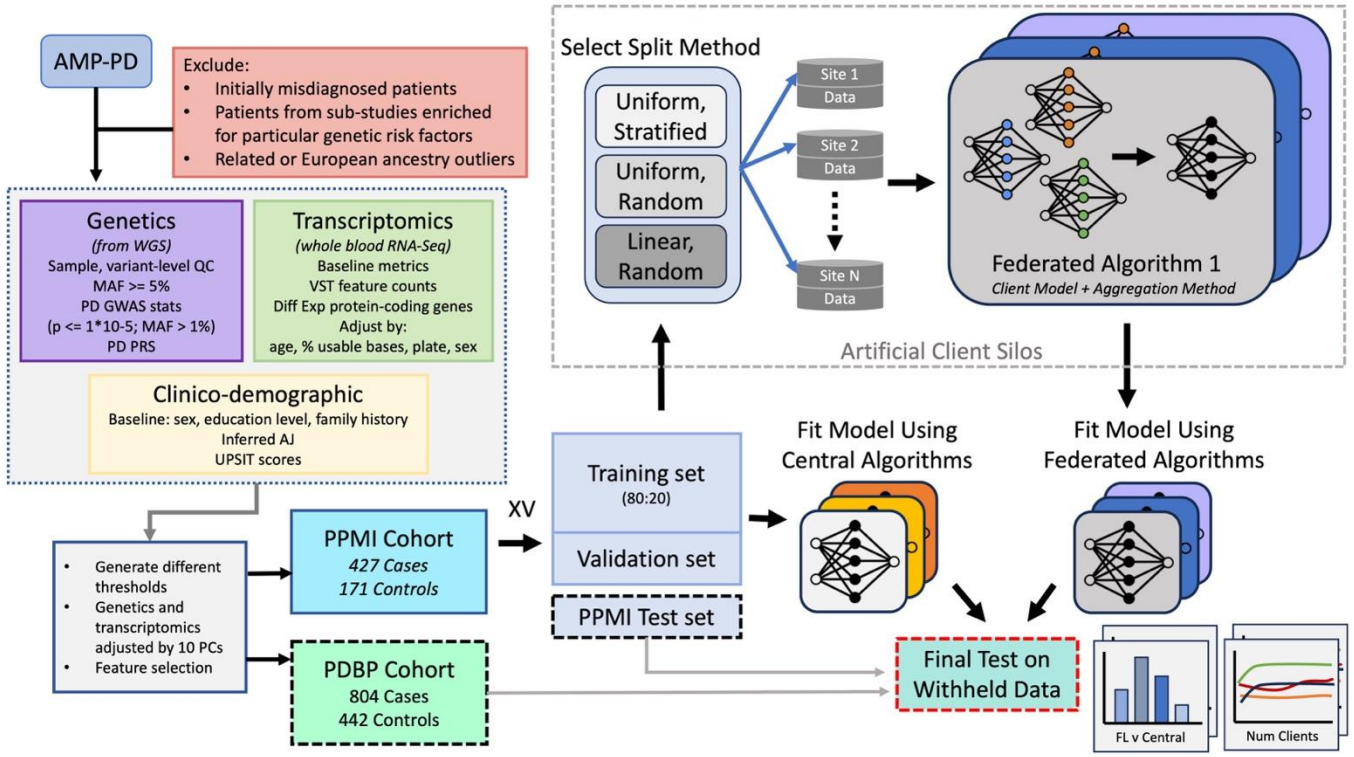

- **Early integration (concatenation-based)**
  - Combine all features into one matrix.
  - Apply machine learning/statistical models.
  - Simple, but risks overfitting and ignores data-type differences.
- **Intermediate integration (transformation-based)**
  - Extract features (e.g., latent factors via PCA, NMF, autoencoders).
  - Align data into a shared space.
  - Keeps omics-specific structure while enabling comparison.
- **Late integration (model-based)**
  - Analyze each omics layer separately.
  - Combine results at decision or model level Flexible but may miss cross-omics interactions.

# Multi – Omics – Integration strategies




- Active Research Field
- Many methdos allow to multi-omics integration
- Applications:
  - Patient stratification in cancer and rare diseases
  - Biomarker discovery for diagnostics
  - Understanding molecular mechanisms
  - Precision medicine and drug response prediction

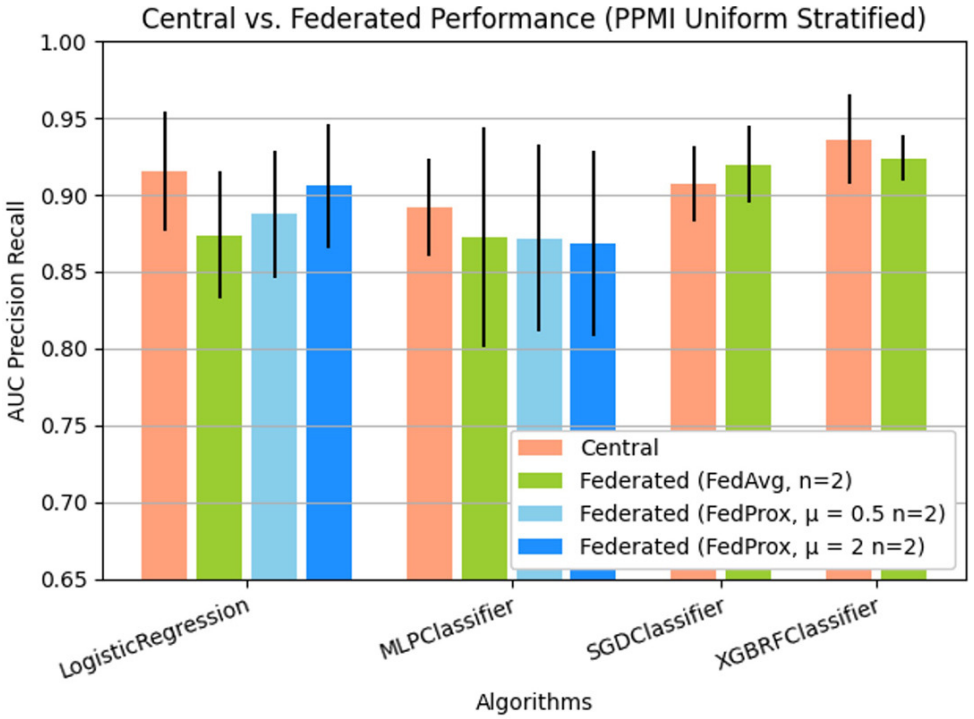
# Multi – Omics FL

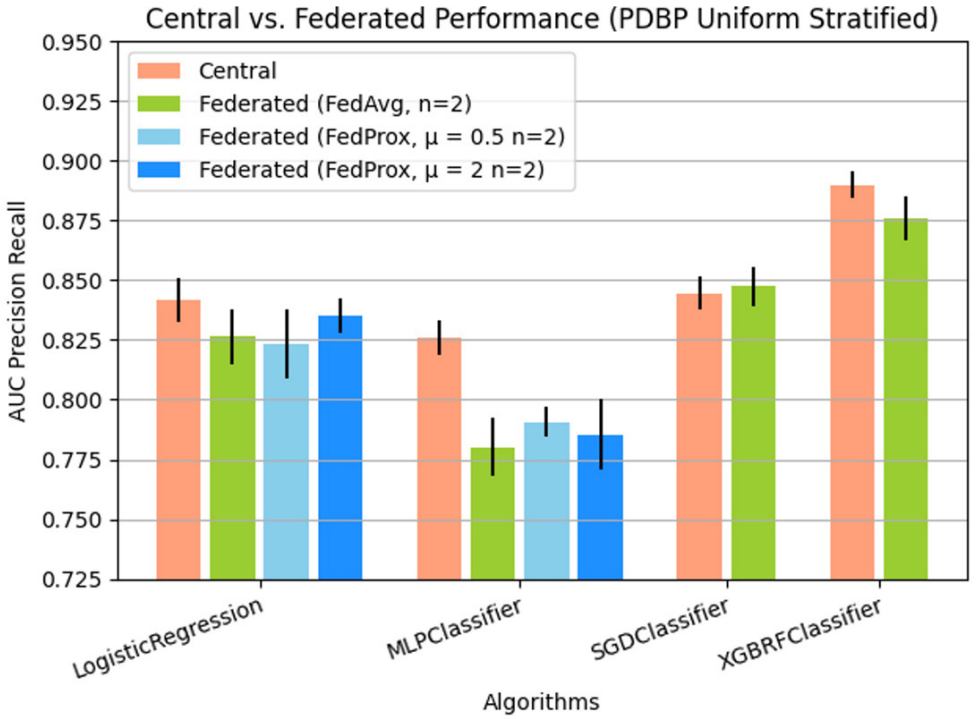



- The study used Multi Omics Alzheimer data to predict severity of the condition
- Comparison between Federated and Centralized Machine Learning models
- Early integration of different features

*Cho, H., Wu, D. & Berger, B. Secure genome-wide association analysis using multiparty computation. Nat Biotechnol 36, 547–551 (2018). https://doi.org/10.1038/nbt.4108*

# Multi – Omics FL

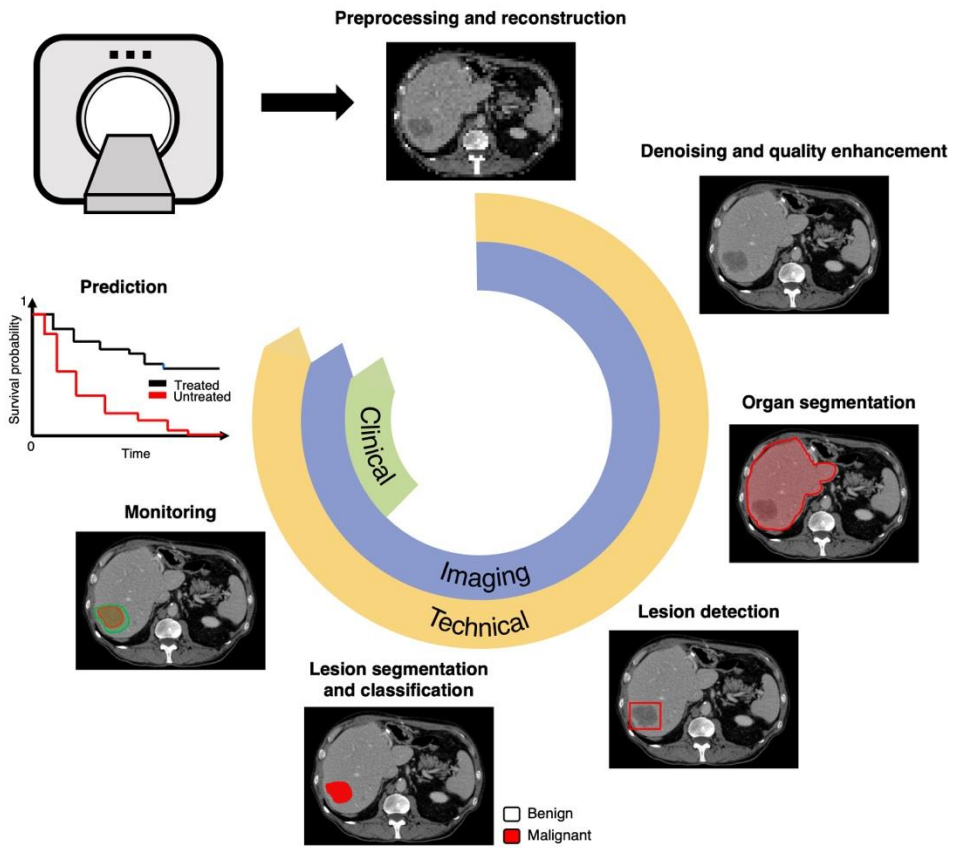



Federated and Centralized model have comparable performances

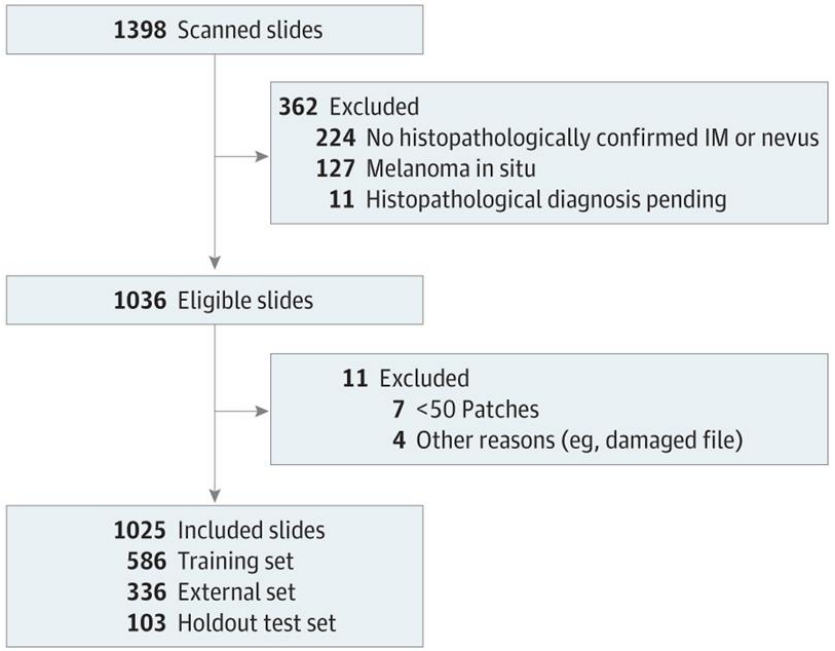
re traditional machine learning models that typically do not integrate the proximal regularization term used in FedProx. Their training procedures are distinct from neural network-based models, making the direct application of FedProx less straightforward or necessary

# Medical Imaging




- **Preprocessing & Reconstruction** – denoising, super-resolution, accelerating MRI/CT acquisition.
- **Segmentation** – automatic delineation of organs, tumors, or cells (e.g., U-Net in medical scans).
- **Classification** – labeling images or regions (healthy vs. diseased, cancer subtype).
- **Detection** – locating and identifying objects of interest (lesions, nodules, anomalies).
- **Feature Extraction & Radiomics** – quantifying shapes, textures, or biomarkers for predictive modeling.

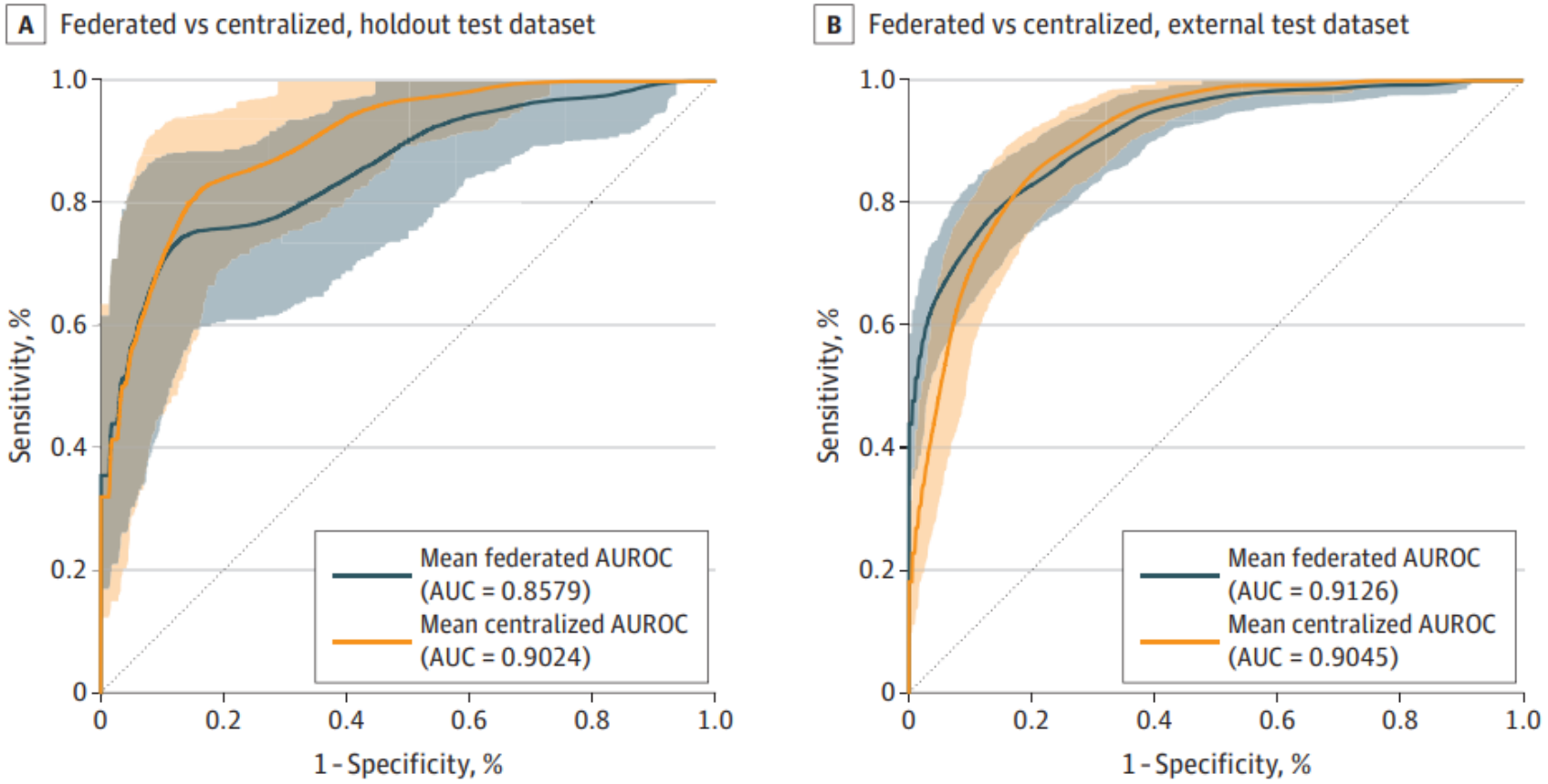
# Medical Imaging - FL




- FL for AI-based melanoma diagnostics.
- A federated model for melanoma-nevus classification was developed using histopathological whole-slide images from 6 German university hospitals
- FL can achieve comparable diagnostic performance to centralized learning while improving privacy protection

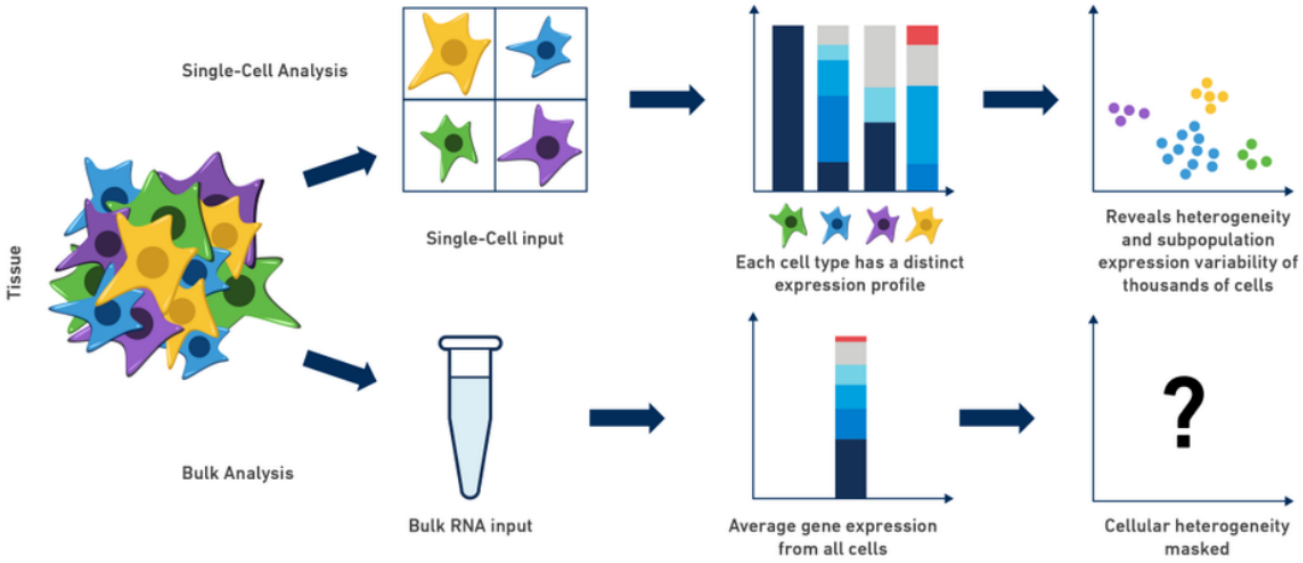
*Haggenmüller, Sarah et al. "Federated Learning for Decentralized Artificial Intelligence in Melanoma Diagnostics." JAMA dermatology vol. 160,3 (2024): 303-311. doi:10.1001/jamadermatol.2023.5550*

# Medical Imaging - FL




- Holdout (Data belonging to the participant facilities i.e. involved in the training)
- External (Data from an external hospital, not involved in the training)

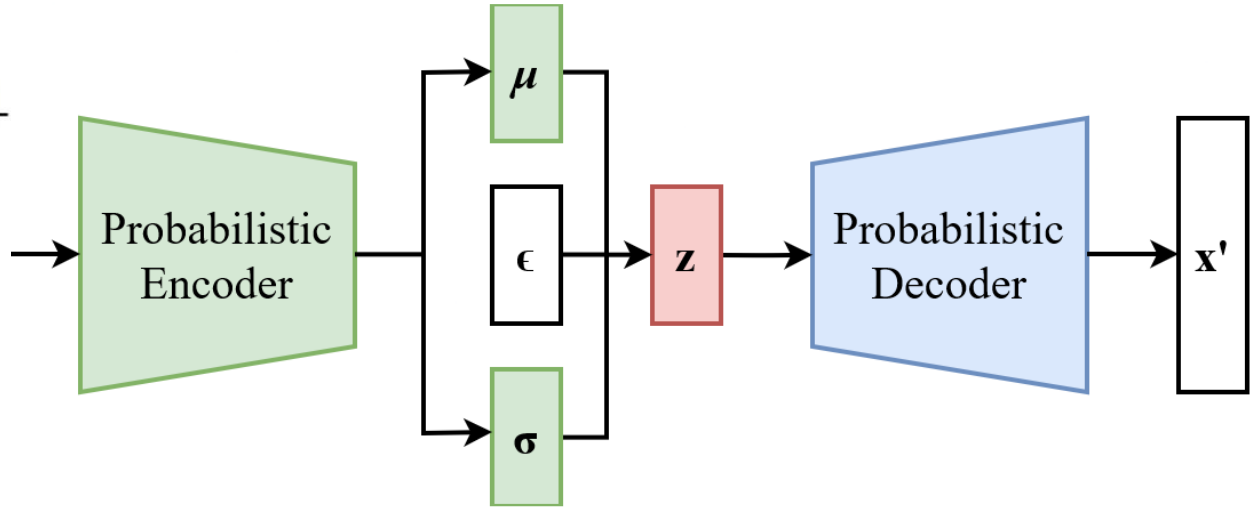
# Single – cell RNA-sequencing (scRNA-seq)

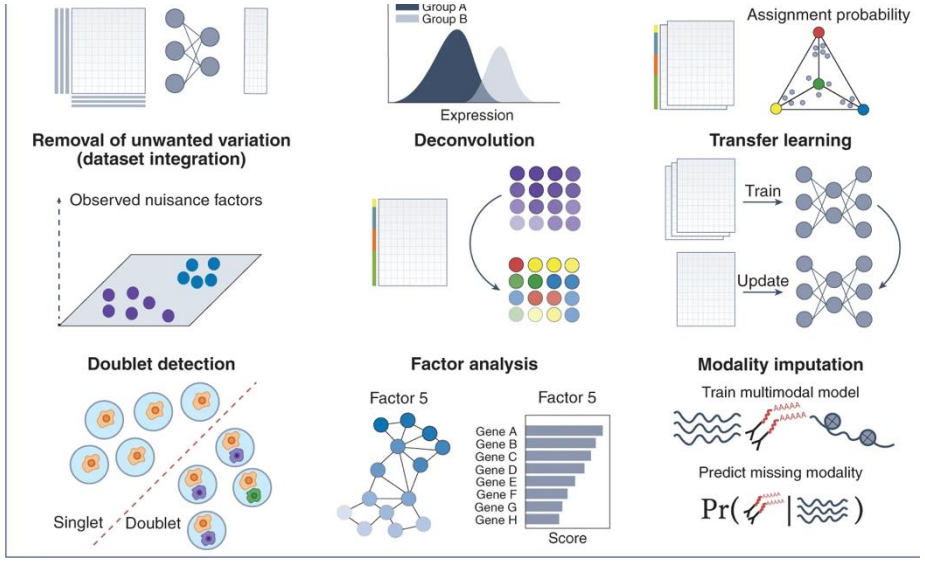



- Captures transcriptomes at the resolution of individual cells, revealing cell-to-cell variability hidden in bulk RNA-seq.
- Identifies **distinct cell types and subtypes** within a mixed population.
- Detects **cell states** (e.g., activation, stress, differentiation) within the same cell type.
- Quantifies **intratumoral heterogeneity** in cancer research, informing treatment resistance studies.

# scVI toolkit

| | Cell1 | Cell2 | ... | CellN |
|---|---|---|---|---|
| Gene1 | 3 | 2 | . | 13 |
| Gene2 | 2 | 3 | . | 1 |
| Gene3 | 1 | 14 | . | 18 |
| ... | . | . | . | . |
| ... | . | . | . | . |
| ... | . | . | . | . |
| GeneM | 25 | 0 | . | 0 |

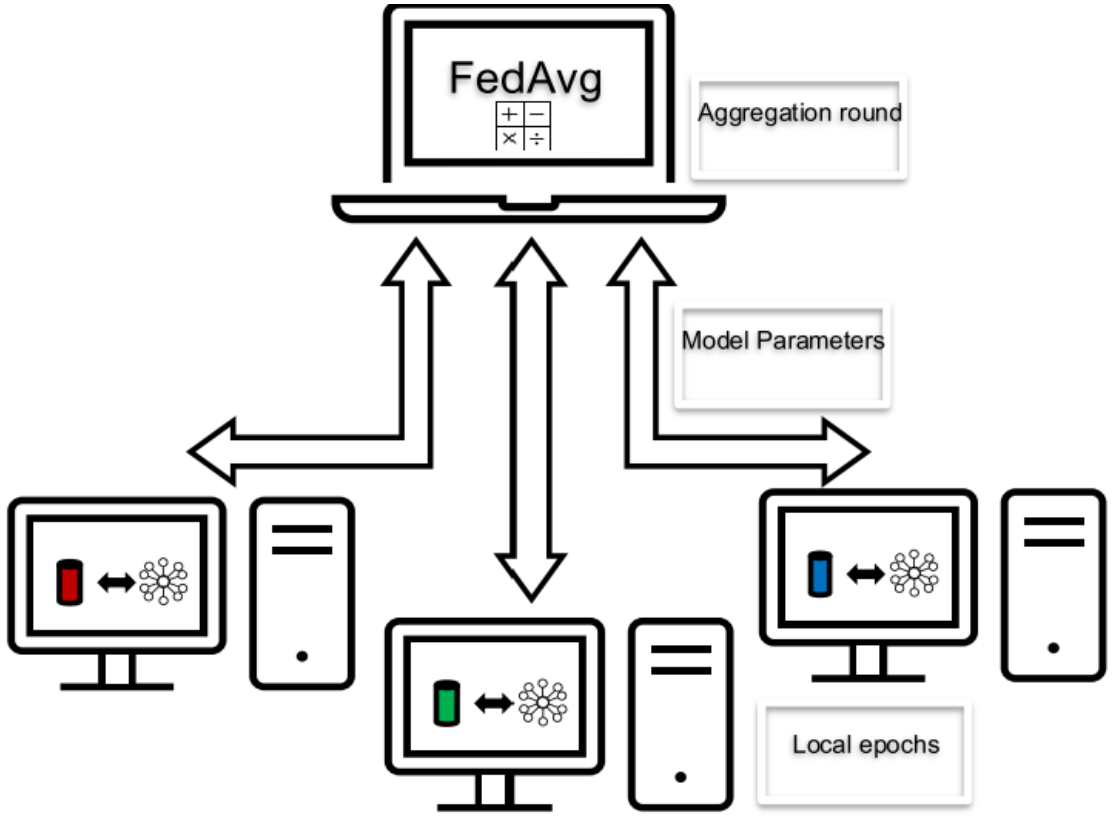



Downstream tasks:

- Batch effect Mitigation
- Reference Mapping
- DEG

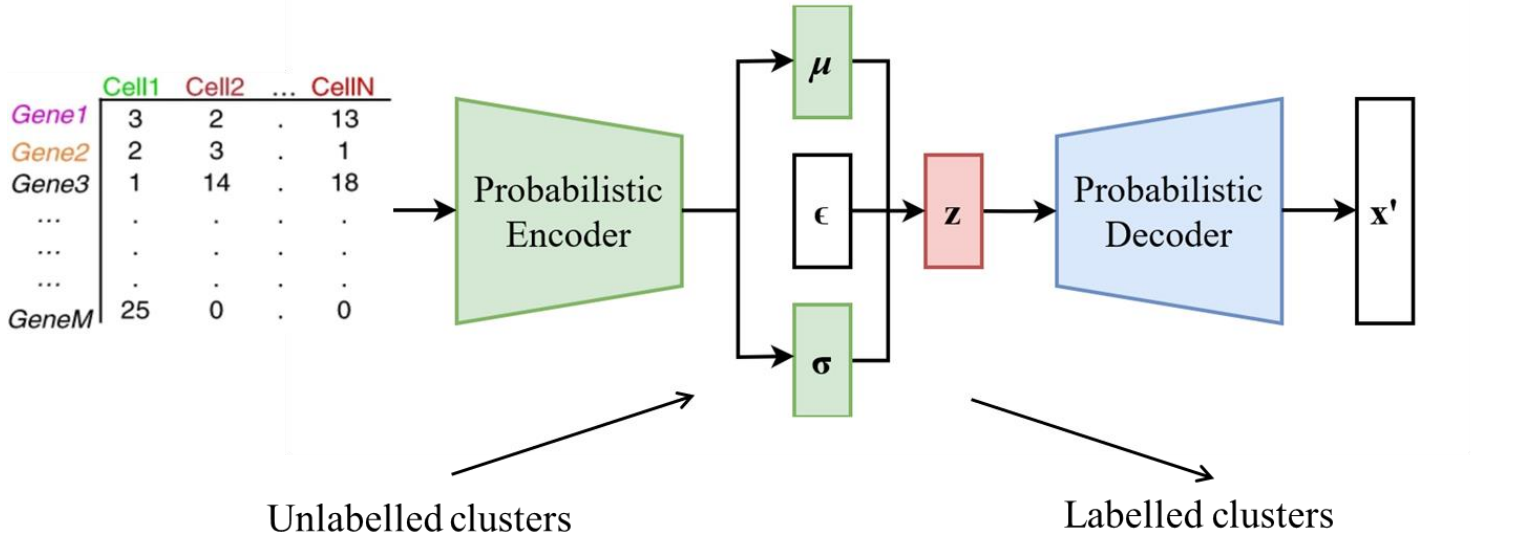
# Federated – scVIC

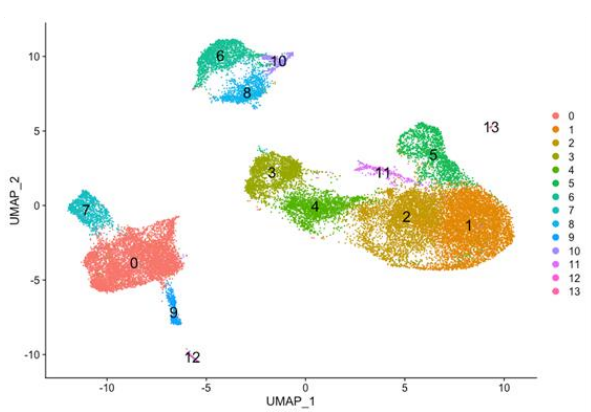


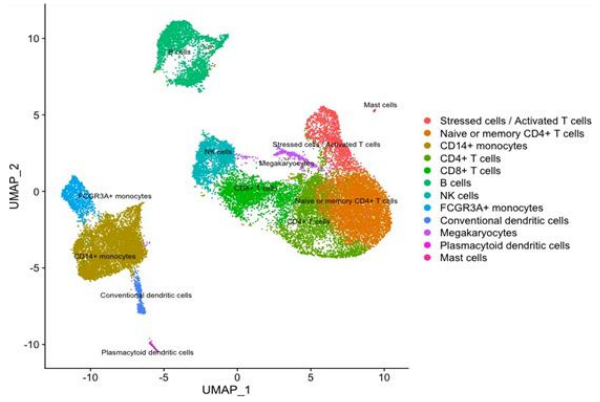

## Federated scVI

- **Objective**: learn a clean latent for cells.
- **Data**: count matrices + batch label (batch_key).
- **Training** : ELBO = Expected NB reconstruction – KL divergence regularization
- **Outputs**: batch-corrected latent for UMAP/clustering

# Task – Reference Mapping



# WILL IT WORK AS GOOD AS THE CENTRALIZED APPROACH?