LUXEMBOURG INSTITUTE OF HEALTH

Multi-modal Data Integration: Classical and AI-based Approaches

Petr Nazarov & BioAI / MoDaS teams

LIH PI Retreat

20 - 21 June 2024, Mondorf-Les-Bains, Luxembourg

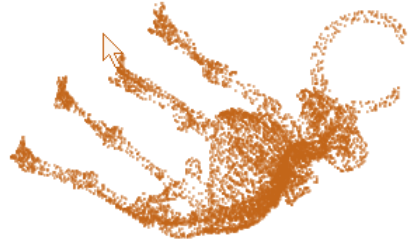
Why Data Integration

What biological object is this?

The same object from different perspectives

Modality A

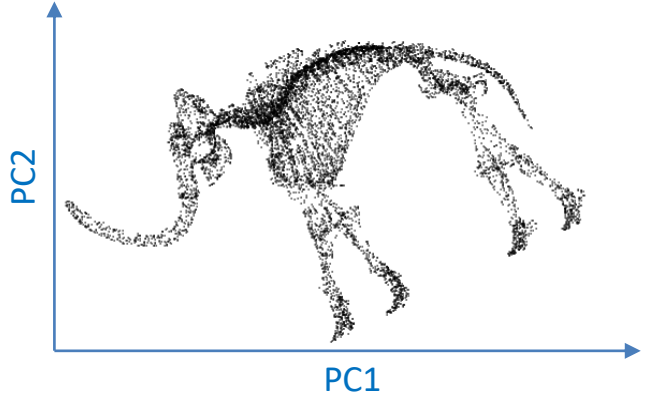
Modality B



Modality C

The object

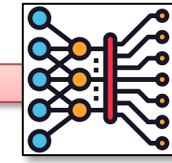
Integrated view



Data Modalities

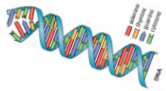
LUXEMBOURG INSTITUTE OF HEALTH

Structured Data → **Features** ← **Unstructured Data**



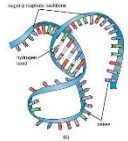
Deep learning: transforms data into embeddings

Genomics

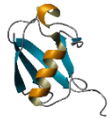


Epigenomics

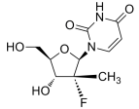
Transcriptomics



Proteomics

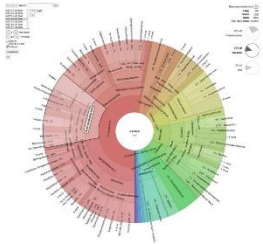


Metabolomics



Cytometry

Metagenomics

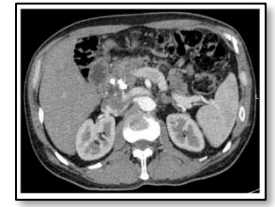


Structured Records

PatientID:	PX20C1793
Age:	58
Sex:	Male
History:	K86.1
Symptoms:	Ja/WL/AP
Results:	LRC179328
Diagnosis:	C25.3

Ideally, we should be able to combine all available data / features to characterize a patient

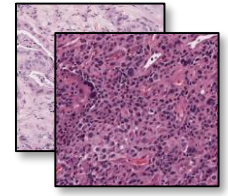
Radiology



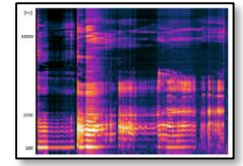
Free Text

The 58-year-old male with a history of chronic pancreatitis, presented with jaundice, unexplained weight loss, and abdominal pain radiating to the back. Laboratory findings revealed elevated levels of serum CA 19-9 and CEA. Imaging studies, including a CT scan, showed a hypodense lesion in the head of the pancreas, suggestive of a mass. Endoscopic ultrasound-guided fine-needle aspiration confirmed the diagnosis of PDAC. The patient's performance status and comorbidities are being evaluated for potential surgical resection as part of the treatment plan.

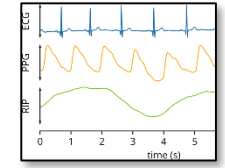
Histology



Voice



Sensors

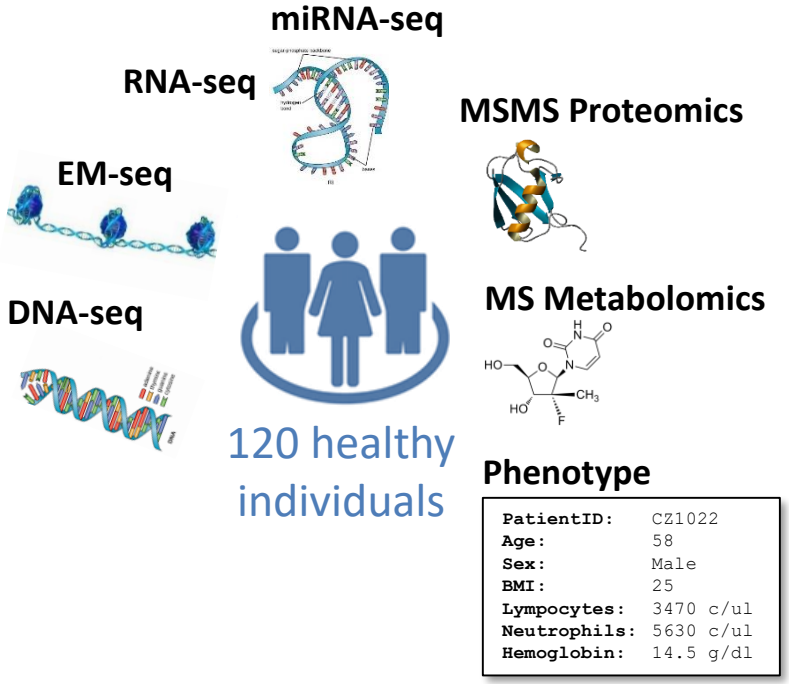


Example: EATRIS+ Data Integration

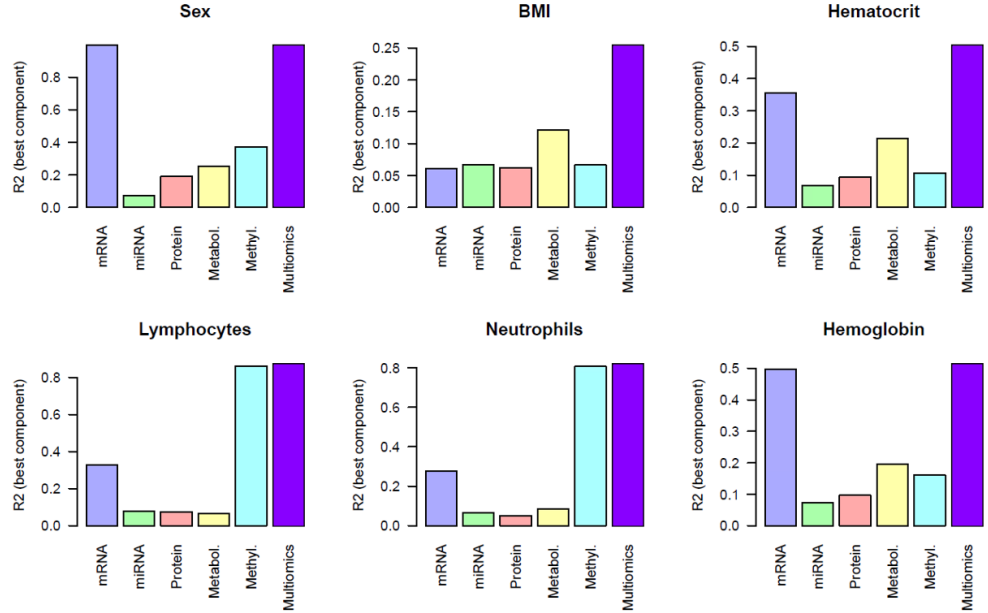
LUXEMBOURG INSTITUTE OF HEALTH

eatris+

EU project on building infrastructure for personalized medicine and research

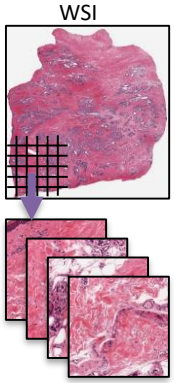


Predicting Phenotype (R^2)

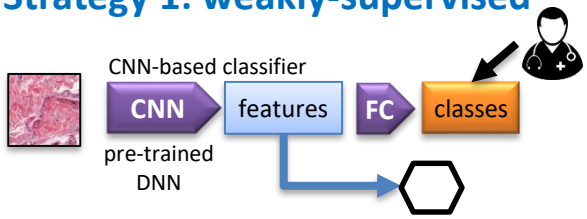


Combined multi-omics predictors always outperform!

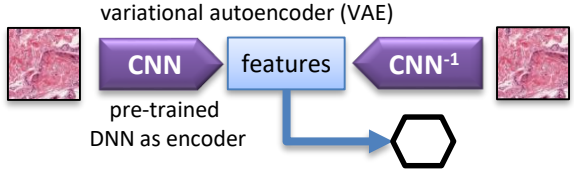
Feature Extraction in Unstructured Data



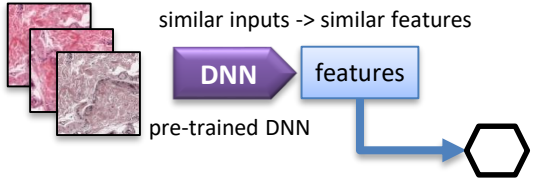
Strategy 1: weakly-supervised



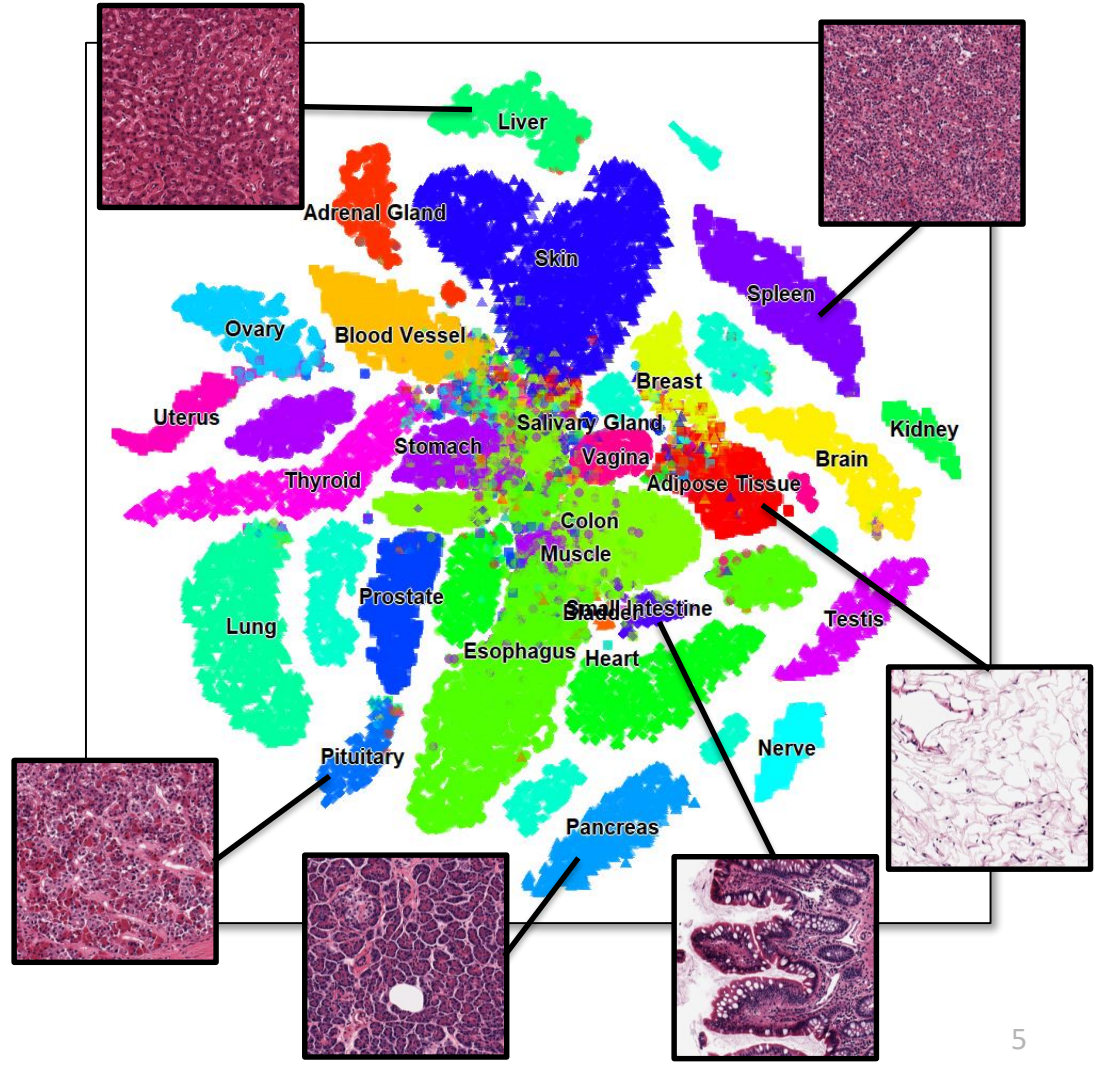
Strategy 2: unsupervised



Strategy 3: self-supervised



Strategy 4: use pre-trained models ☺

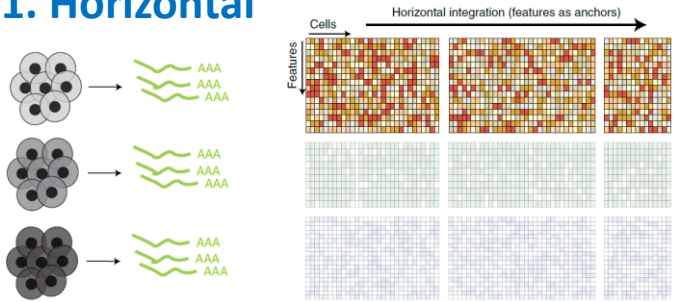


You can work with features/embeddings as with normal observations (e.g. gene expression)

Data Integration Tasks and Tools

LUXEMBOURG INSTITUTE OF HEALTH

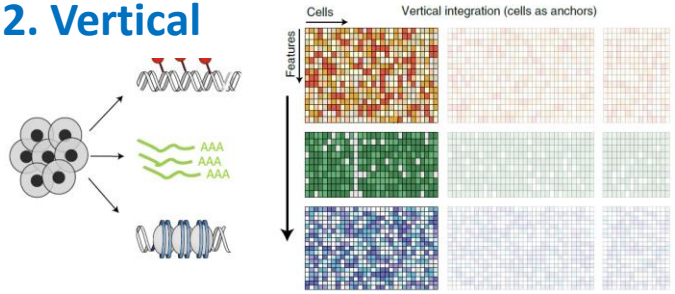
1. Horizontal



- Merging (with) public datasets
- Single-cell studies across different patients / technologies
- **MVD** – merging own data with LIH datasets

batch correction | ComBat | ICA | Seurat3 | Harmony | scVI

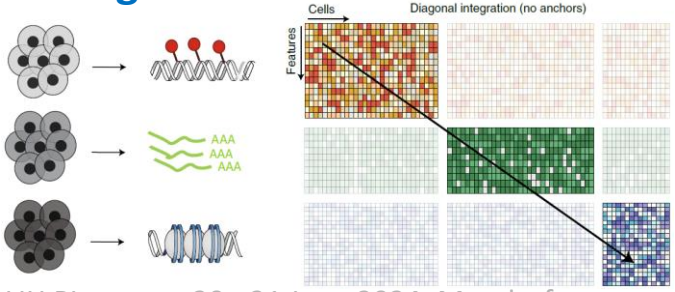
2. Vertical



- Multi-omics in a study
- Integrative analysis of **MVD**
- **CLINNOVA** multi-modal predictions

integration | CCA | PLS | ICA | MOFA | Seurat4 | scAI

3. Diagonal



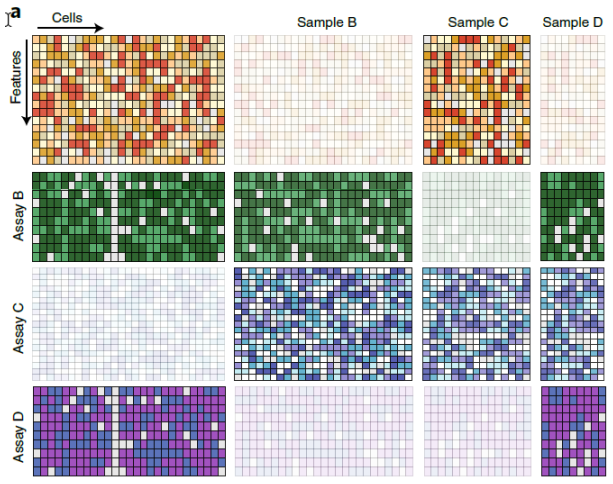
- Multi-omics in a single-cell study (e.g. RNA-seq + scATAC-seq)
- Combining scRNA-seq with CyTOF

integration of subpopulations | LIGER | network methods | SCIM

Argelaguet, Cuomo, Stegle, Nat.Biotech 2021

Data Integration Tasks and Tools

4. Mosaic

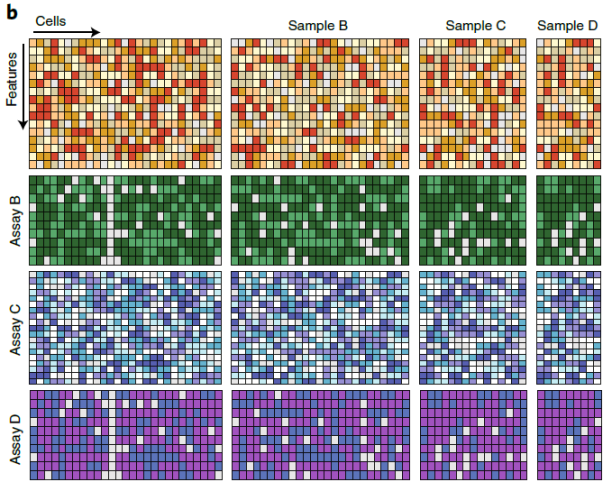


mosaic integration or imputation

UINMF

AI-based methods

various VAEs



- Different cells from the same samples and several studies united
- This is a realistic situation in any large project (including LIH-driven)

Although mosaic integration seems to be the most difficult, modern AI approaches (foundation models, t.b.c.) offer hope. ☺

Argelaguet, Cuomo, Stegle, Nat.Biotech 2021

Fouche, Zinovyev, Fron.Bioinformatics 2023

Data Integration via Matrix Factorization

LUXEMBOURG INSTITUTE OF HEALTH

ICA: independent runs for single-omics

Structured data

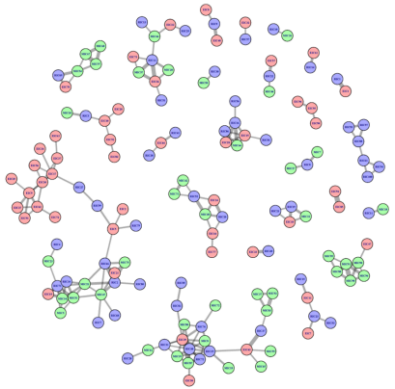
mRNA data → consICA → S x M

miRNA data → consICA → S x M

Methy-lation data → consICA → S x M

Integration (correlation)

Fonds National de la Recherche Luxembourg



Diagnostics

Prognostics

Capturing unknown factors

Biological signals

Nazarov et al, BMC Med Genomics 2019

MOFA: simultaneous analysis

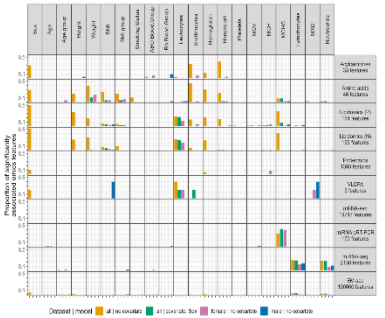
eatris+

mRNA data

miRNA data

Methy-lation data

MOFA → S x M



Diagnostics

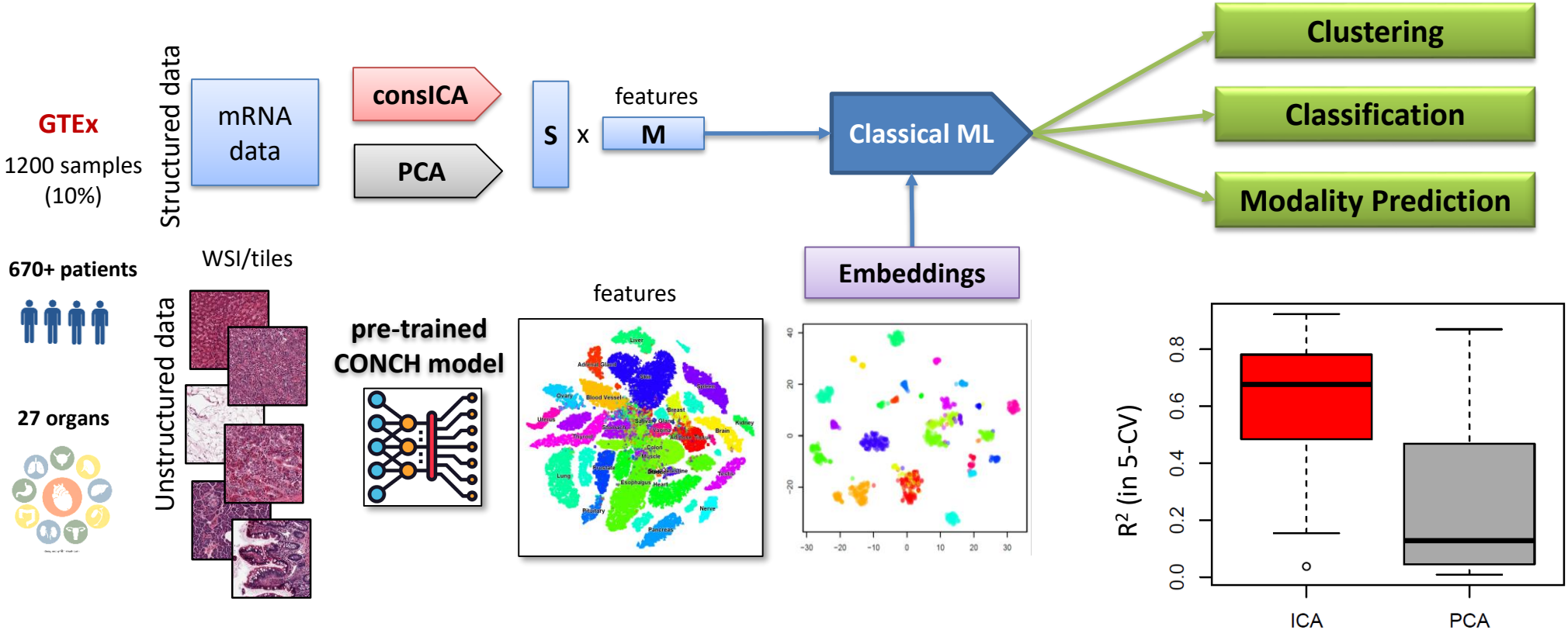
Prognostics

Capturing known factors

Biological signals

Argelaguet et al, Mol Syst Biol 2018

Mixed data: Histopathology and mRNA



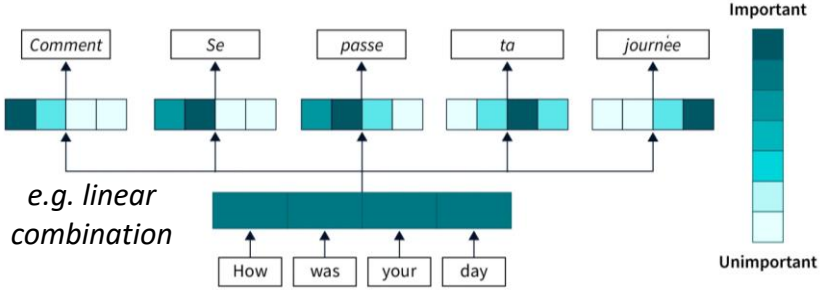
Ming Lu .. Faisal Mahmood , Nat Med, ***2024***
CONtrastive learning from Captions for Histopathology

ICA, which defects biological processes, shows a much better linkage to histopathology than PCA

Transformers and Attention Mechanism

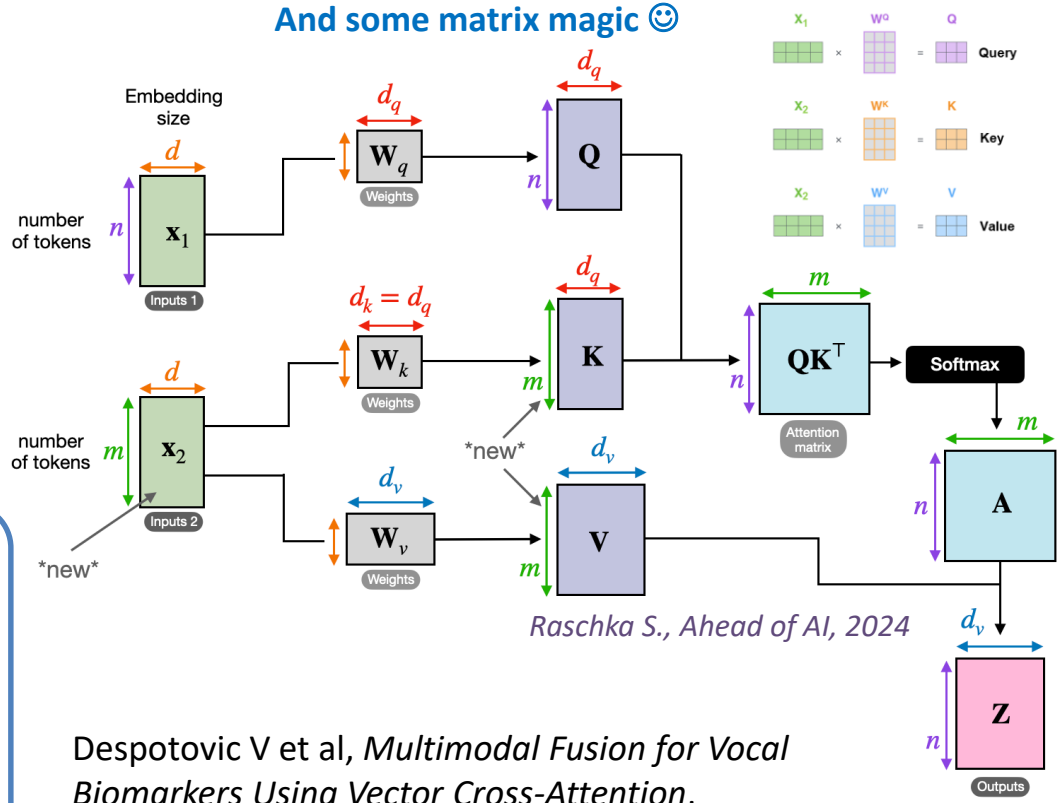
LUXEMBOURG INSTITUTE OF HEALTH

Attention mechanism



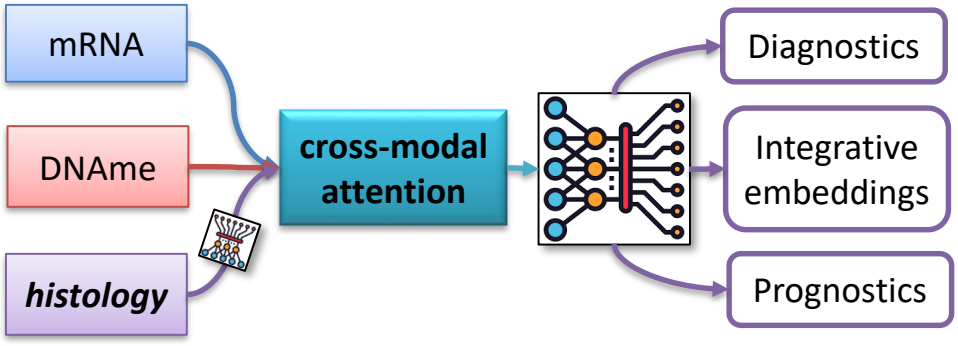
https://www.scaler.com/topics/deep-learning/attention-mechanism-deep-learning/

And some matrix magic ☺



Raschka S., Ahead of AI, 2024

Current Development in the Team

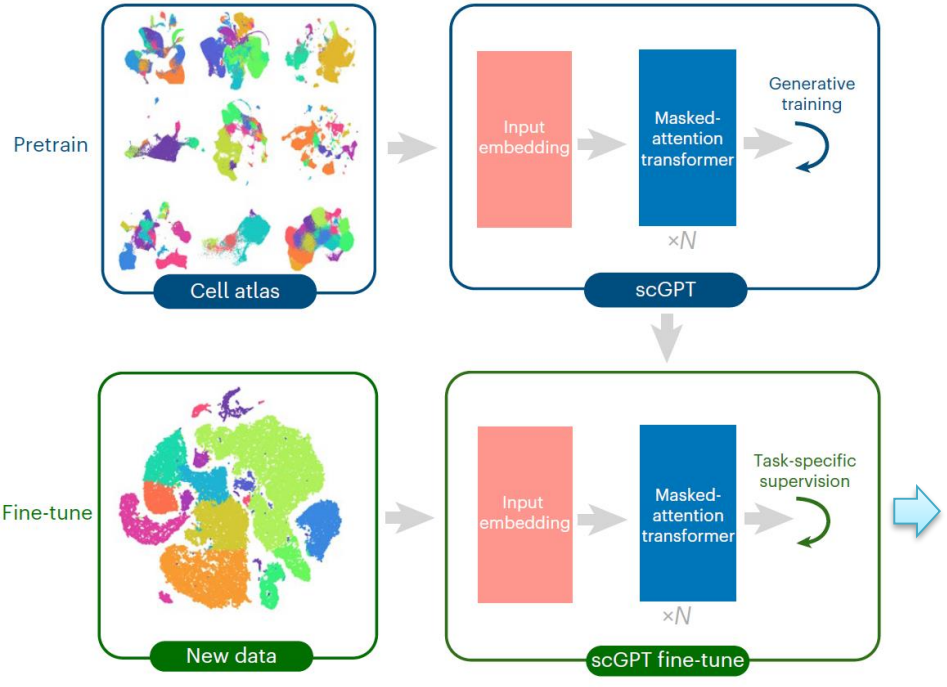


Despotovic V et al, *Multimodal Fusion for Vocal Biomarkers Using Vector Cross-Attention*, INTERSPEECH 2024

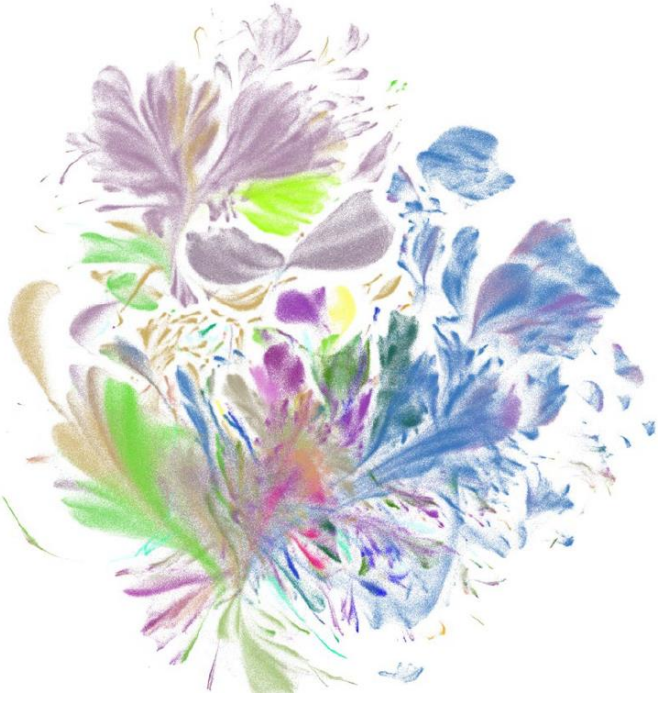
Modality	Accuracy [%]
Sustained vowel phonation	63.30 (1.54)
Reading	61.89 (3.13)
Early fusion	66.99 (2.26)
Single-head cross-attention	67.43 (2.76)
Multi-head cross-attention	**68.84 (2.53)**

Foundation Models for Data Integration

33M human cells



- Clustering
- Batch correction
- Cell type annotation
- Gene network inference
- Multi-omics integration
- Perturbation prediction



Team effort (DMI+DoCR): we are testing a local copy of scGPT on LIH data. So far results are very reasonable!

*Cui et al, **scGPT**, Nature Methods 2024*

Take Home Message

LUXEMBOURG INSTITUTE OF HEALTH

- Integration tasks, while generally solvable, present different levels of complexity. **'Horizontal' and 'vertical' integration are reasonably good**, while 'diagonal' and mosaic integration still pose challenges. But with the recent advancements in AI (foundation models), we hope to see a significant breakthrough in these areas.

- The standard mathematical approach to integration for structured data is **matrix factorization** (many methods exist!) We will use them in the frame of MVD

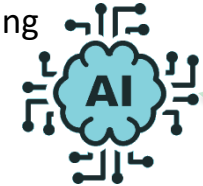
- Unstructured data require an AI (deep learning model) “layer” to generate features, which we call embeddings. You can use pre-trained models from large labs.

- We are now testing novel approaches at DMI / DoCR including pre-trained foundation models for the analysis of omics data, images, and text

The Team(s)

LUXEMBOURG INSTITUTE OF HEALTH

artificial intelligence
machine learning
deep learning
transformers



multi-omics
genomics
epigenomics
transcriptomics
etc

DEPARTMENT OF CANCER RESEARCH

Bioinformatics and AI Unit
@ Department of Medical Informatics

R.Toth*
P.Nazarov*

V.Despotovic*
S-Y.Kim
L.Zhang
T.Kaoma
O.Hunewald
A.Rauschenberger*

LUXGEN

S.Magni*

Genomics Platform

F.Stefanizzi*
M.Chepeleva
S. Akerblom (MSc st.)
D.Tabbakh

Multiomics Data Science
Research Group @ DoCR

NORLUX @ DoCR

B.Nosirov*
T.Lukashiv*

Biological research

(*) PhD

DEPARTMENT OF MEDICAL INFORMATICS

TRANSLATIONAL MEDICINE OPERATIONS HUB (TMOH)

DEPARTMENT OF INFECTION AND IMMUNITY

DEPARTMENT OF PRECISION HEALTH