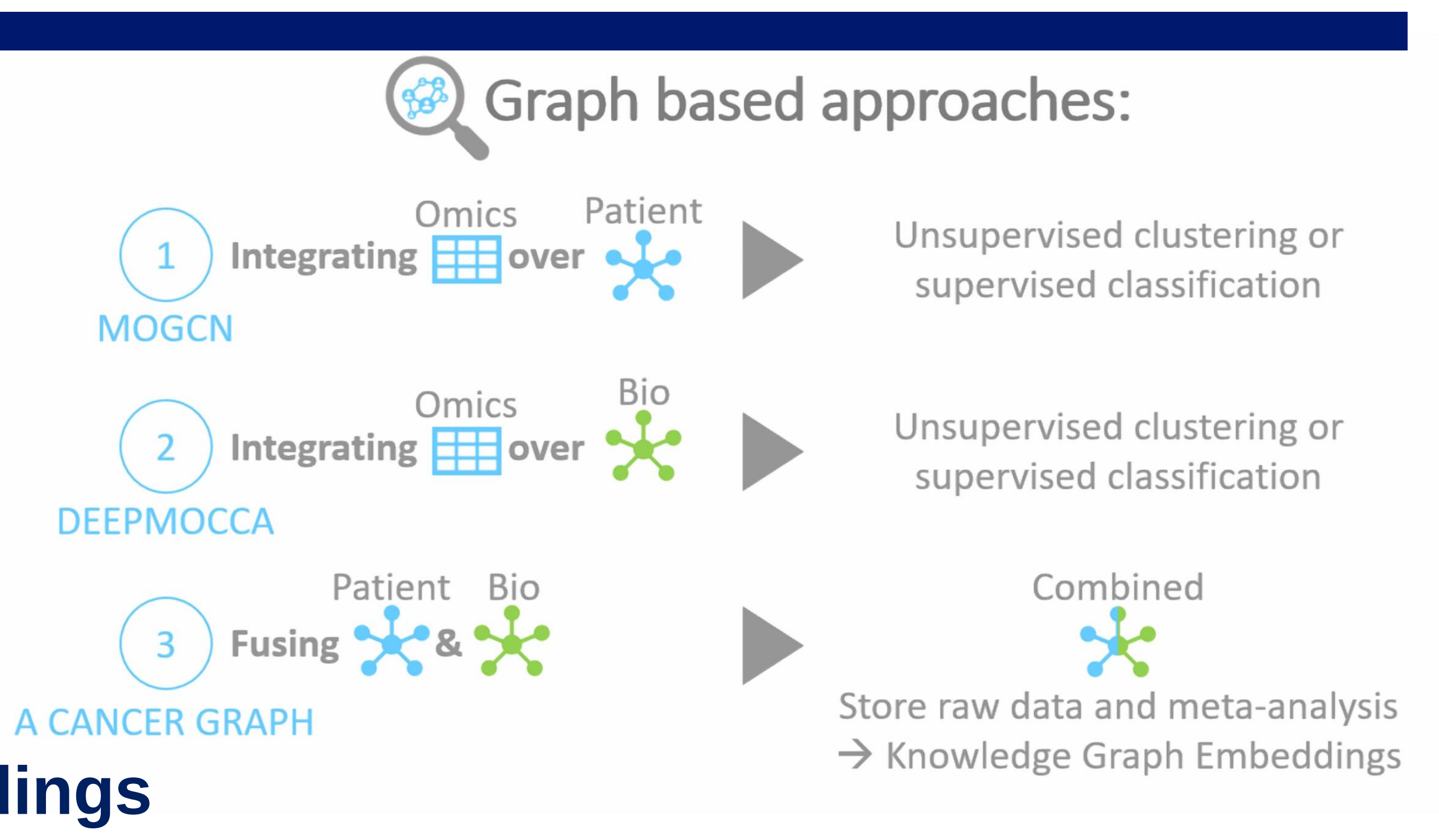
# Extracting knowledge from multi-omics & clinical datasets using graph machine learning

Ferdinand Popp[1,3], Nektarios A. Valous[1,2], Pornpimol Charoentong[2,4] and Inka Zörnig[2,4]

[1]Applied Tumor Immunity Clinical Cooperation Unit, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, 69120, Heidelberg, Germany
[2]Center for Quantitative Analysis of Molecular and Cellular Biosystems (Bioquant), Heidelberg University, Im Neuenheimer Feld 267, 69120, Heidelberg, Germany
[3]Faculty of Biosciences, Heidelberg University, Im Neuenheimer Feld 234, 69120, Heidelberg, Germany
[4]Department of Medical Oncology, National Center for Tumor Diseases (NCT), Heidelberg University Hospital (UKHD), Im Neuenheimer Feld 460, 69120, Heidelberg, Germany

## Overview & Objective

- Cancer is the result of multistage processes and events that incorporate multiscale information from the genome to the proteome, consequently interactions and synergistic effects are much better explored through multi-omics analysis. Multi-omics analysis, integrating clinical data, facilitates the discovery of hypothesis-generating biomarkers, and aid in uncovering mechanistic insights into clinical and microenvironmental processes.
- Graph machine learning offers a potentially reliable methodological toolset, for integrated multi-omics analysis, as a tangible alternative to cancer scientists and clinicians that seek ideas and implementation strategies for their data.
- Two common subtypes of lung cancer: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) have drastically different biological signatures, yet often they are treated similarly and classified together as non-small cell lung cancer (NSCLC) (Fang et. al., 2021).
- We are testing multiple graph machine learning architectures on publicly available NSCLC samples integrating multi-omics data and clinical information. This results in obtaining more granular patient subgroups, that we investigate in clinical and biological views.
- With the aggregation of biological and clinical knowledge graphs (KG) we aim to inspect a patient-KG-fused graph approach for storage, analysis and lookup of multi-omics data.

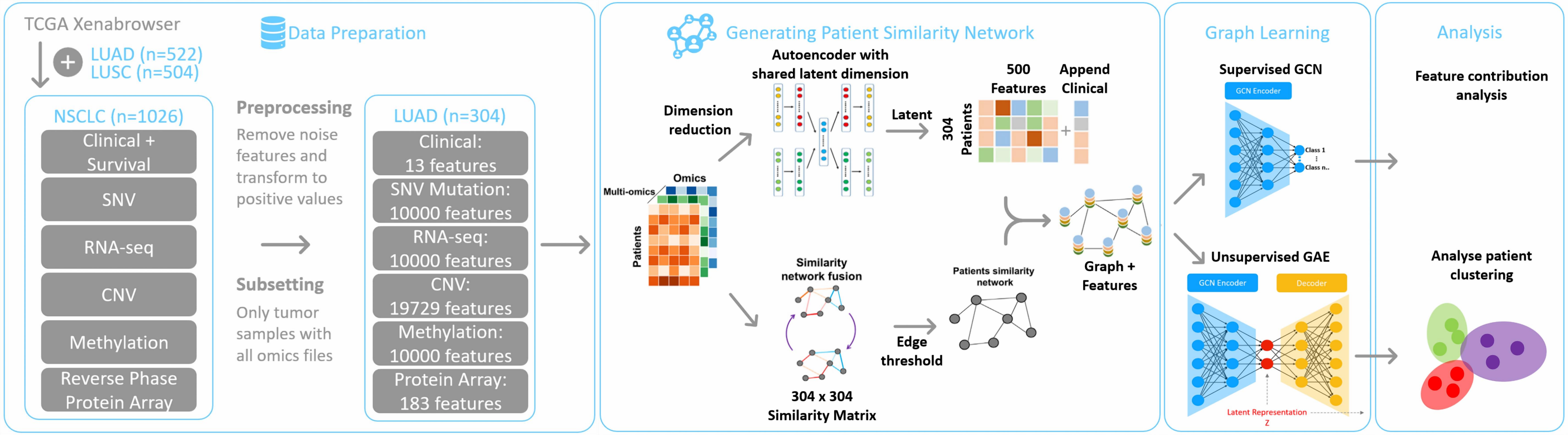

Graph based approaches:

1 Integrating Omics over Patient ▶ Unsupervised clustering or supervised classification
MOGCN

2 Integrating Omics over Bio ▶ Unsupervised clustering or supervised classification
DEEPMOCCA

3 Fusing Patient & Bio ▶ Combined
Store raw data and meta-analysis
→ Knowledge Graph Embeddings
A CANCER GRAPH

## MOGCN employs patients into latent embeddings


TCGA Xenabrowser
LUAD (n=522)
LUSC (n=504)

Data Preparation

NSCLC (n=1026)
Clinical + Survival
SNV
RNA-seq
CNV
Methylation
Reverse Phase Protein Array

Preprocessing
Remove noise features and transform to positive values

Subsetting
Only tumor samples with all omics files

LUAD (n=304)
Clinical: 13 features
SNV Mutation: 10000 features
RNA-seq: 10000 features
CNV: 19729 features
Methylation: 10000 features
Protein Array: 183 features

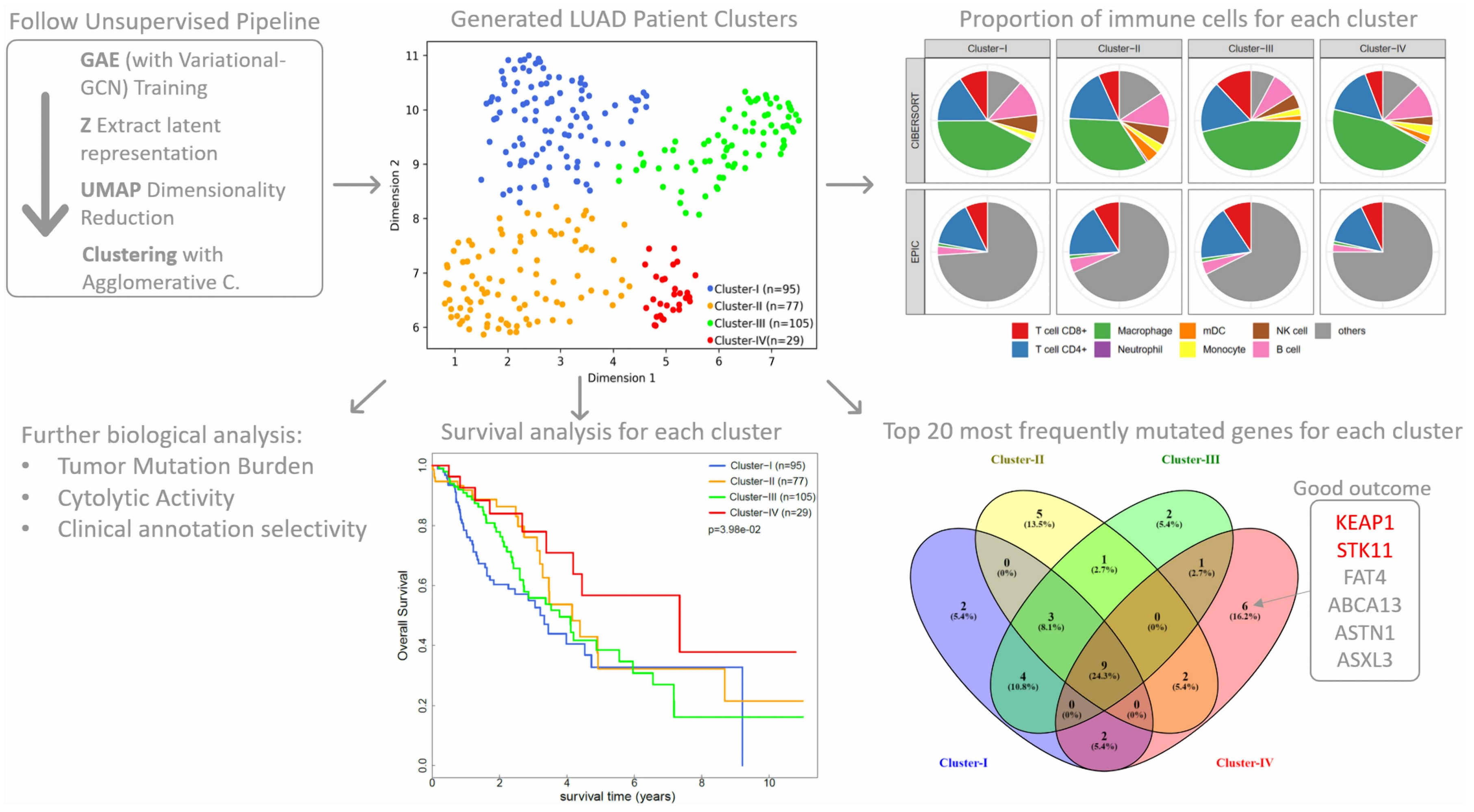
Generating Patient Similarity Network
Multi-omics
Omics
Patients
Dimension reduction
Autoencoder with shared latent dimension
Latent
304 Patients
500 Features
Append Clinical
Similarity network fusion
304 x 304 Similarity Matrix
Edge threshold
Patients similarity network
Graph + Features

Graph Learning

Analysis

The MOGCN pipeline creates a patient similarity graph with compressed input features adapted from (Li et. al., 2022). With a Graph Auto Encoder (GAE) latent patient embeddings can be induced in a data-driven way. In a supervised setting it can predict e.g., molecular subtype of a patient. The setup and implementation are inspired by (Fang et. al., 2021). Models are evaluated with mean silhouette score and AUC or F1 score respectively.

## Unsupervised clustering of LUAD patients


Follow Unsupervised Pipeline
**GAE** (with Variational-GCN) Training
**Z** Extract latent representation
**UMAP** Dimensionality Reduction
**Clustering** with Agglomerative C.

Generated LUAD Patient Clusters

Proportion of immune cells for each cluster

Further biological analysis:
- Tumor Mutation Burden
- Cytolytic Activity
- Clinical annotation selectivity

Survival analysis for each cluster

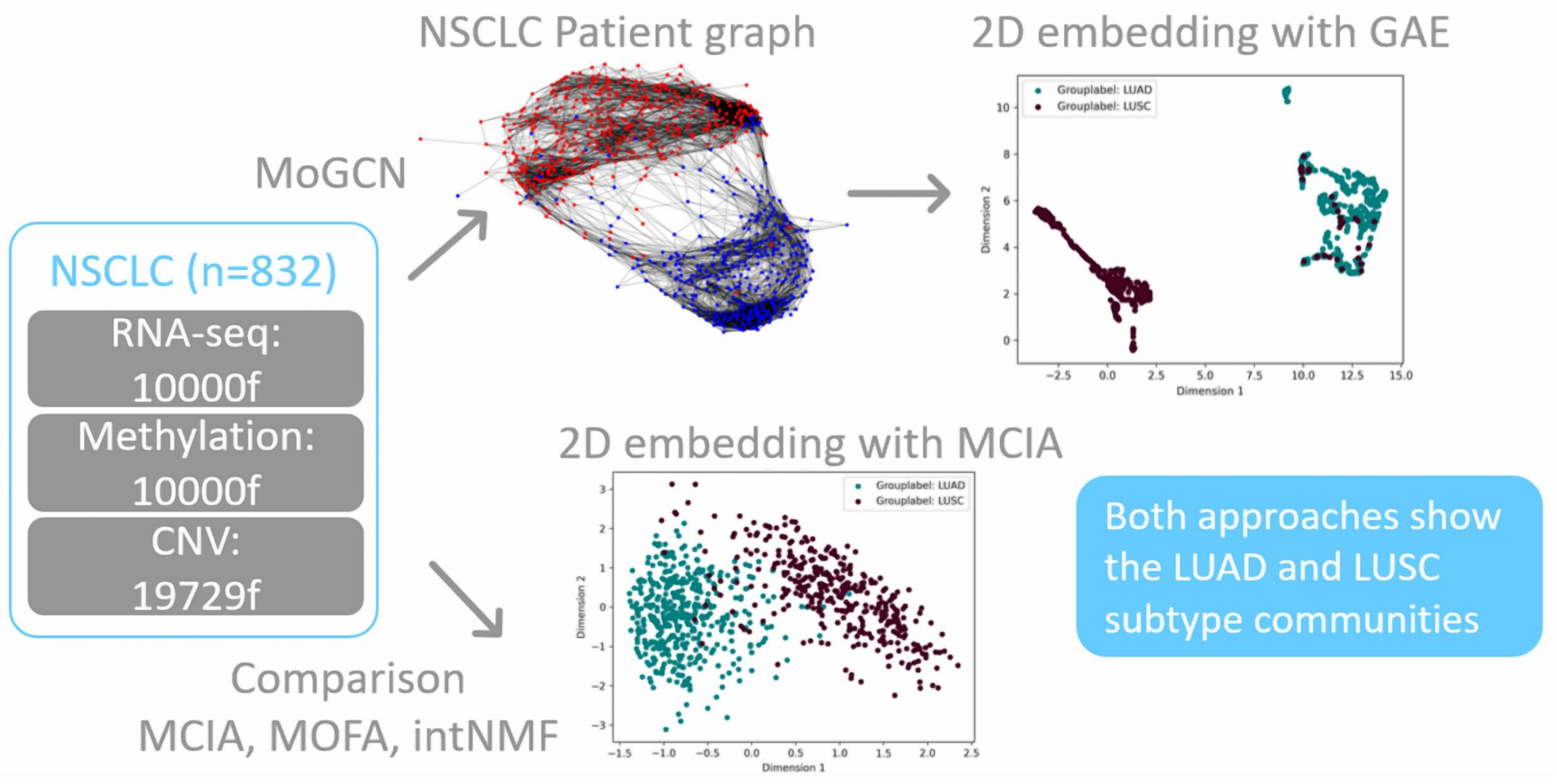
Top 20 most frequently mutated genes for each cluster

Good outcome
KEAP1
STK11
FAT4
ABCA13
ASTN1
ASXL3

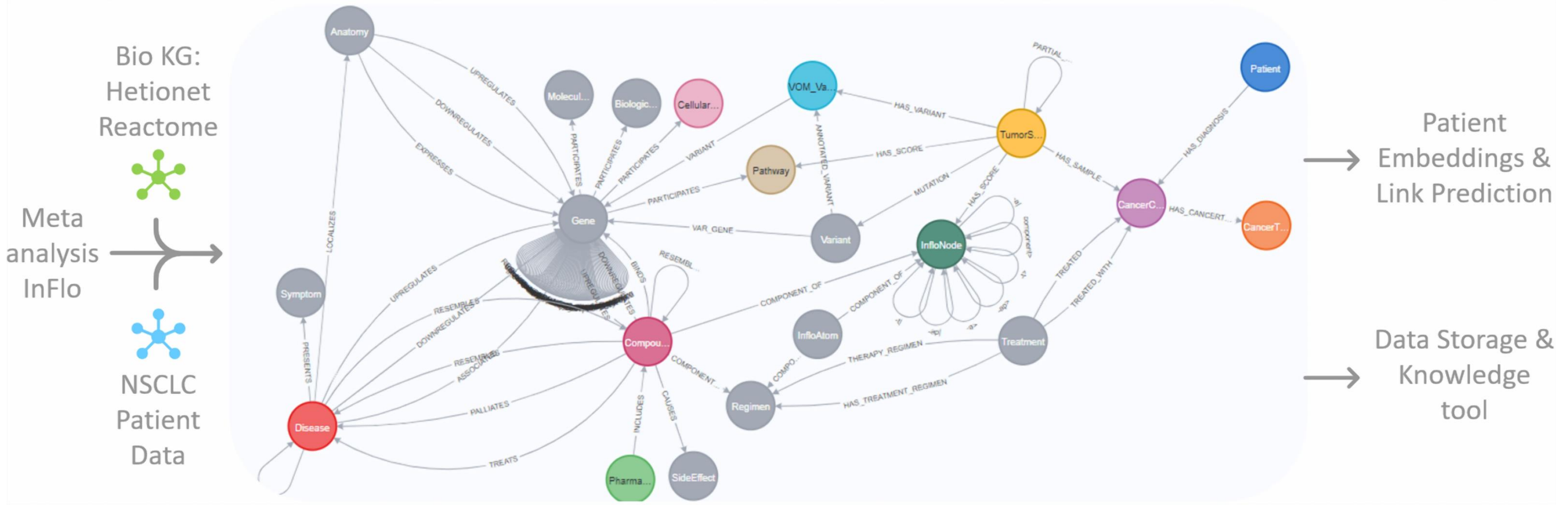
The unsupervised MOGCN pipeline generated 4 clusters for 304 LUAD patients. In survival analysis Cluster-IV patients had a significantly higher risk of survival than other clusters (p = 0.039). During the biological analysis the cytolytic activity, estimated using the expression of two genes Granzyme A (GZMA) and Perforin (PRF1) (Rooney et al., 2015), and the Tumor Mutation Burden (TMB) showed no significant change in clusters. We performed computational deconvolution (Li et al., 2020) to estimate the abundance of 9 tumor-infiltrating immune cell types (B cells, CD4 T cells, CD8 T cells, neutrophils, macrophages, monocytes, dendritic cells, NK cells and others). Several studies have associated mutations in Kelch-like ECH-associated protein 1 (KEAP1) and serine/threonine kinase 11 (STK11) genes, and their co-occurrence with different mutations, to poor outcomes in NSCLC patients treated with immune checkpoint inhibitors, addressing  this finding to the present of “cold” immune tumor microenvironment (de Lima et. al., 2022). From the MOGCN analysis we observed LUAD patients in cluster IV (with mutated KEAP1-STK11) had a significantly higher risk of survival than other clusters. The concurrent loss of function mutations in KEAP1 together with STK11 may promote tumor suppressor cells or induce apoptosis rate of tumor cells.

[1] Fang, et. al. (2021). DeePaN: deep patient graph convolutional network integrating clinico-genomic evidence to stratify lung cancers for immunotherapy. NPJ digital medicine, 4(1), 1-10.
[2] Li, et. al. (2022). MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. Frontiers in Genetics, 127.
[3] Rooney, et. al. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell, 160(1-2), 48–61.
[4] Li, et. al. (2020). Computational deconvolution of tumor-infiltrating immune components with bulk tumor gene expression data. In Bioinformatics for Cancer Immunotherapy (pp. 249-262). Humana, New York, NY.
[5] de Lima, et. al. (2022). STK11 and KEAP1 mutations in non-small cell lung cancer patients: Descriptive analysis and prognostic value among Hispanics (STRIKE registry-CLICaP). Lung Cancer, 170, 114-121.
[6] Tuck (2022). A cancer graph: a lung cancer property graph database in Neo4j. BMC Research Notes, 15(1), 1-4.
[7] Althubaiti, et. al. (2021). DeepMOCCA: A pan-cancer prognostic model identifies personalized prognostic markers through graph attention and multi-omics data integration. bioRxiv.

## Unsupervised clustering of NSCLC patients


NSCLC Patient graph
2D embedding with GAE
MoGCN
NSCLC (n=832)
RNA-seq: 10000f
Methylation: 10000f
CNV: 19729f
2D embedding with MCIA
Both approaches show the LUAD and LUSC subtype communities
Comparison
MCIA, MOFA, intNMF

## ‚A Cancer Graph‘ for NSCLC TCGA data adapted from (Tuck, 2022)


Bio KG:
Hetionet
Reactome
Meta analysis
InFlo
NSCLC Patient Data
Patient Embeddings & Link Prediction
Data Storage & Knowledge tool

## Conclusions and Perspectives

Graph-based approaches hold great potential to augment the integration and interrogation of biological data. NSCLC data gets unsupervised clustered into the LUAD and LUSC communities, while e.g. the LUAD community itself contains subgroups. One LUAD subgroup may have two mutations that favor higher survival risk. The MOGCN pipeline will be expanded to PanCancer analysis, and its clustering potential will be benchmarked on PAM50 Breast cancer classification. The KG approach allows the combined storage of molecular data and meta analysis and gives everyone access (e.g. Neodash) to answer related questions. We implemented a public data loader to make the upload of tabular data into graphs accessible (https://github.com/PRODYNA/capt-mifune). The integration over a biological net approach, e.g. over a PPI-network, is in progress by mapping the omics to proteins and applying a graph embedding as (Althubaiti et. al., 2021) used to do survival prediction via multi-omics data.

NCT NATIONAL CENTER FOR TUMOR DISEASES HEIDELBERG
supported by
German Cancer Research Center (DKFZ)
Heidelberg University Medical Center
Hospital for Thoracic Diseases
German Cancer Aid

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

dkfz. DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

BioQuant MODEL base of LIFE

UNIVERSITÄTS KLINIKUM HEIDELBERG

SCAN ME

f.popp@dkfz.de
Background in Molecular Biology, Programming/Data Science & Digital Health

Tools: neo4j PyTorch geometric

Created on 8.10.2022
Created as part of a Master Thesis @ UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386 & dkfz. DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

F. Popp

P. Charoentong

N. Valous

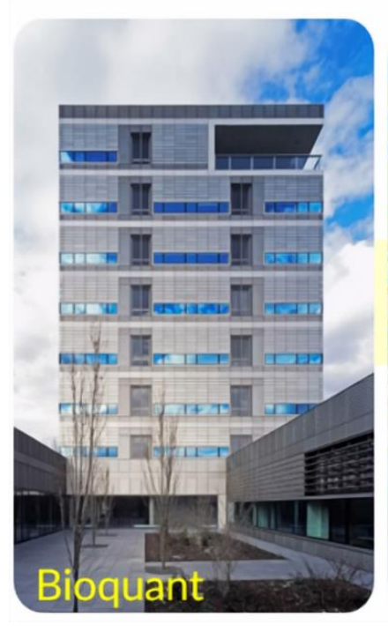
I. Zörnig

D. Jäger

AG Brors - DKFZ

Thank you to all collaborators for their efforts!

AG Wiemann - DKFZ


Bioquant

DKFZ

NCT