Genome Variant Representation in Genome LMs and Multi-Omics AI

Efficiency, Scalability, and Medical Actionability

Neringa Jurenaite[1]

1 Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, TU Dresden

MOTIVATION

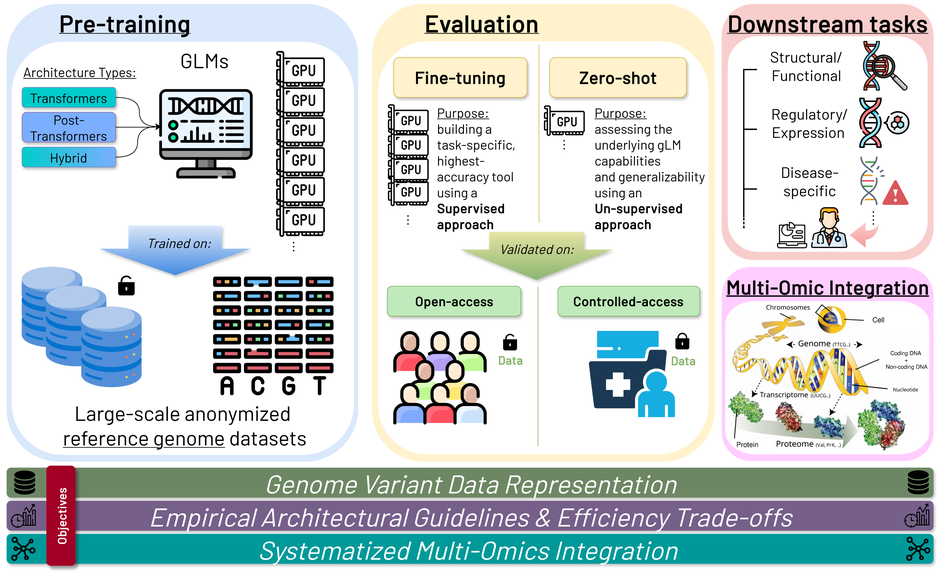
The recent rapid growth of large multi-omics datasets have provided unprecedented opportunities in medical and clinical research, in particular precision medicine. To solve the fundamental challenge of biological representation learning, we explore **Genome Language Models (gLMs)** for DNA sequence modeling and **Autoencoders** for broader multi-omics integration. Driven by their unsupervised approach to training and multi-task applicability, both frameworks convert complex, noisy, high-dimensional data into a representation that retains biological meaning for downstream medical and clinical actionability.

Pre-training gLMs on vast reference genome datasets requires significant hardware resources, after which the learned representations are **evaluated** on **downstream tasks** via supervised fine-tuning or (unsupervised) zero-shot inference. This work investigates and evaluates the applicability, scalability, and robustness of gLMs and autoencoders, in combination with other deep learning approaches, for genome variant representation and **multi-omics data integration** for oncology tasks.

KEY RESEARCH QUESTIONS

- How can gLMs **effectively model unstructured DNA sequence variants** for oncology tasks across coding and non-coding regions without relying on, e.g., manual feature selection or abstraction?
- How can **multi-omics representation learning be unified** into a standardized, reproducible framework which scales up to larger cohorts and out to various modalities?
- How can **architectural choices, hyperparameter tuning, and hardware-aware scaling** guidelines be combined to maximize the computational efficiency of gLMs and multi-omics frameworks on multi-GPU HPC infrastructures?

GLM OVERVIEW



1. GENOME VARIANTS AND GLMS

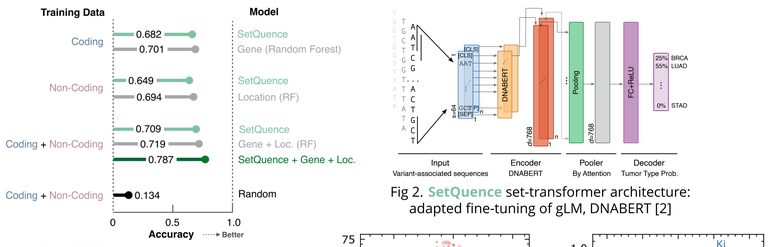


Fig 1. TCGA pan-cancer primary type multi-class classification, from variants

Feasibility study captures variant importance and variant-level attribution in oncology tasks using non-fixed whole genome variant-associated sequence sets with gLMs.

Ensemble model in Fig 1 combines **SetQuence** [4, 5] (Fig 2) and Random Forrest (RF) [1] for Non-synonymous frequencies (Gene) and Regional Mutation Density signatures (Loc).

Fig 2. **SetQuence** set-transformer architecture: adapted fine-tuning of gLM, DNABERT [2]

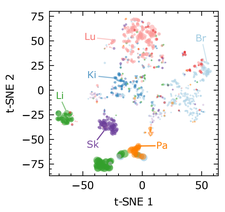


Fig 3. **Ensemble model** embedding space: patient clustering into 10 tumor classes

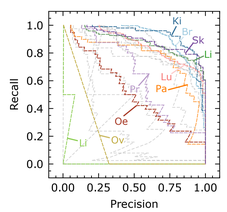


Fig 4. **Ensemble model** Precision-recall curves across tumor classes

2. GENOME VARIANTS IN MULTI-OMICS MODELS

SetOmic (Fig 5) extends the SetQuence framework into a multi-modal deep neural-network model (DNN):

- integrates unstructured, permutation-invariant sets of genome variant-associated sequences via gLM finetuning alongside complementary omics modalities (e.g., transcriptome expression counts),
- pools sequence-level and patient-level multi-modal features into a unified embedding for downstream prediction tasks, such as tumor classification,
- provides inherent interpretability at the nucleotide level together, with improved tumor type classification accuracy and robustness compared with single-omic approaches, including SetQuence and SetOmic-tr (transcriptome expression only DNN),
- achieves 95% multi-classification accuracy across 32 TCGA Pan-Cancer primary tumor types.

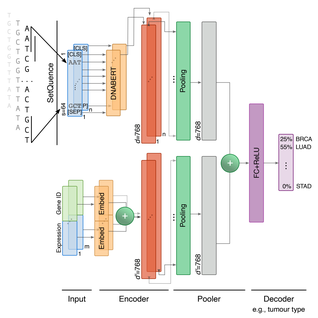


Fig 5. SetOmic architecture: late multi-omics fusion DNN for oncology tasks

Classifier	Precision	Recall	F1-score	Accuracy	ROC AUC
SetOmic	**0.945**	0.909	**0.921**	**0.950**	**0.997**
SetQuence	0.375	0.362	0.359	0.475	0.910
SetOmic-tr	0.876	0.887	0.880	0.923	0.976
OmiEmbed	0.932	**0.911**	0.914	0.942	**0.997**
Strat. Random	0.035	0.034	0.034	0.041	0.500
Prior Random	0.004	0.031	0.007	0.118	0.500

Table 1. Multi-class tumor type prediction (32 classes)

3. ARCHITECTURAL GUIDELINES & EFFICIENCY

Figures 6–8 present empirical investigations into model architecture, hyperparameter tuning, and computational scaling across training and fine-tuning. We leverage these results to analyze model performance vs efficiency trade-offs and evaluate model capabilities

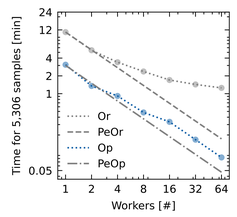


Fig 6. Scalability of **SetQuence** (one epoch) across A100s: **(Or)**iginal baseline vs **(Op)**timized together with their perfect linear scaling (**PeOr** and **PeOp**) [5]

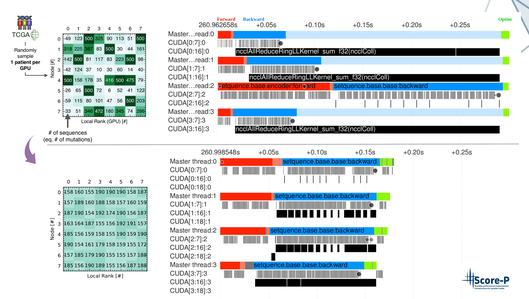


Fig 7. Diffusive load balancing of per-patient variant sequence input across A100s in **SetQuence** (original vs optimized) from the TCGA dataset [5]

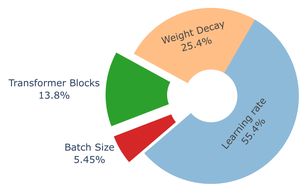


Fig 8. Hyperparameter importance in DNABERT [6, 2]

in zero-shot inference and multi-omics integration stages.

Some of the strategies include diffusive load balancing, knowledge distillation, multi-objective HPO and encoder freezing [3, 4, 5, 6].

4. STANDARDIZED MULTI-OMICS FRAMEWORK

AUTOENCODIX (Fig 9) [3] is a generalized, reproducible and versatile framework providing a **modular ecosystem** for an end-to-end processing of high-dimensional, multi-modal data through five major autoencoder architectures:

- *vanillix*: standard dense,
- *varix*: variational,
- *stackix*: hierarchical/stacked,
- *ontix*: ontology-driven, and
- *x-modalix*: cross-modal.

It unifies data preprocessing, model training, hyperparameter tuning, and evaluation into one standardized pipeline.

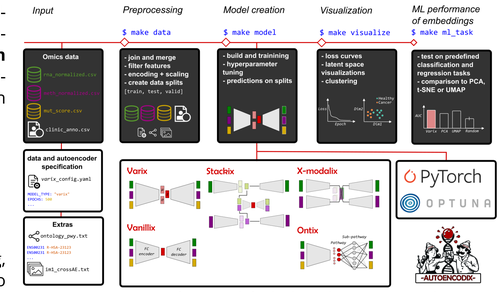


Fig 9. **AUTOENCODIX** framework for biological representation learning

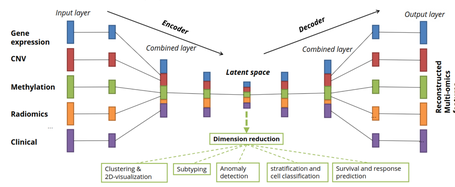


Fig 10. Multi-omics representation learning via *vanillix*

Empirical guidelines are provided for pan-cancer and single-cell datasets, detailing how the integration of different modalities impacts latent space representations.

Figure 10 demonstrates compression of data types into a latent embedding space, which is optimized via feature reconstruction and used for multiple downstream tasks.

REFERENCES

I sincerely thank Prof. Dr. Wolfgang E. Nagel, Prof. Dr. med. Jakob N. Kather, Dr. René Jäkel, Dr. Jan Ewald, Dr. Sunna Torge, Andrei Politov, Lalith Manjunath and others for their support and contributions.

[1] S. Chakraborty, A. Martin, Z. Guan, C. Begg, and R. Shen. Mining mutation contexts across the cancer genome to map tumor site of origin. *Nature Communications*, 12(1):3051, 2021.

[2] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021.

[3] M. J. Joas, **Jurenaite, N.**, D. Praščević, N. Scherf, and J. Ewald. AUTOENCODIX: A generalized and versatile framework to train and evaluate autoencoders for biological representation learning and beyond. *Nature Computational Science*, 6(1):96–108, 2026.

[4] **Jurenaite, N.**, D. León-Periñán, V. Donath, S. Torge, and R. Jäkel. SetQuence & SetOmic: Deep set transformer-based representations of cancer multi-omics. In *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–9. IEEE, 2022.

[5] **Jurenaite, N.**, D. León-Periñán, V. Donath, S. Torge, and R. Jäkel. SetQuence & SetOmic: Deep set transformers for whole genome and exome tumour analysis. *BioSystems*, 235:105095, 2024.

[6] **Jurenaite, N.** and A. Politov. Large genomic language models: Towards their hyperparameter optimization. In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, Atlanta, GA, USA, 2024.

CONCLUSION

- Self-attention mechanisms in gLMs learn contextual representations directly from raw nucleotide sequences, enabling modeling of long-range dependencies across coding and non-coding regions.
- Standardized multi-modal frameworks such as AUTOENCODIX and models such as SetOmic support reproducible multi-omics representation learning across pan-cancer cohorts.
- Hardware-aware scaling reduces memory bottlenecks in multi-GPU systems, improving computational throughput needed for broader model validation and supporting more accessible medical actionability.

FUTURE STEPS

- Comprehensive set of hardware-aware deployment guidelines for gLMs.
- Medical actionability and clinical deployment.
- Further cross-modal analysis.

CONTACT

TU Dresden

Neringa Jurenaite

neringa.jurenaite@tu-dresden.de

https://scads.ai

IN COOPERATION WITH

UNIVERSITÄT LEIPZIG

Dresden University of Technology

FUNDED BY

Federal Ministry of Research, Technology and Space

SACHSEN Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushalts.

The authors gratefully acknowledge the GWK support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden on HRSK-II.

SC26 Chicago, hpc matters