A Tale of Two - different kinds of – Graphs

Helping to Mend the Disconnect Between Biological Research and Medicine

Disclosures

Financial

- Ben Busby is a full-time employee of -- and owns stock in – NVIDIA
- Ben owns stock in DNAnexus

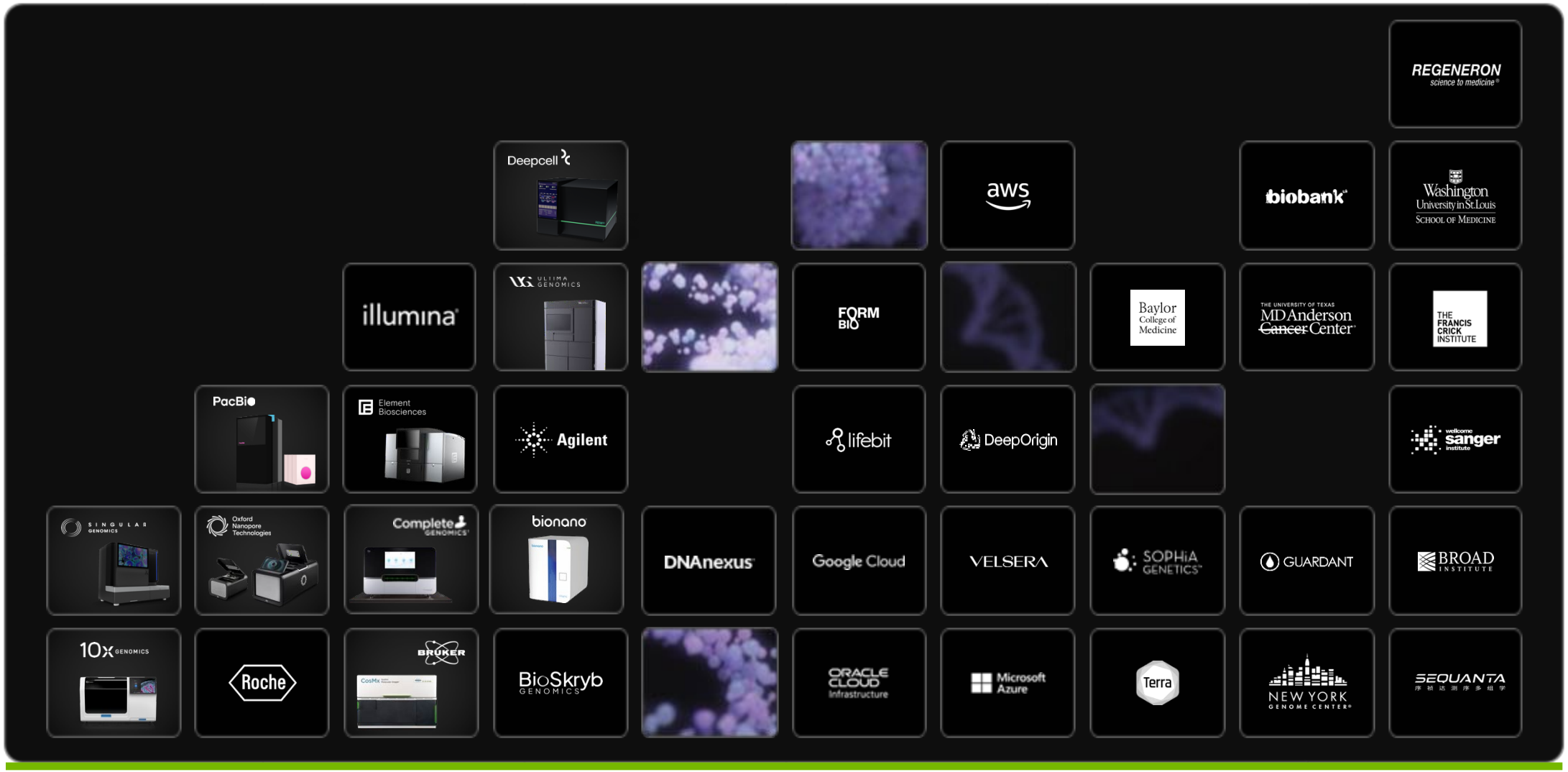
Non-Financial

- Ben also has various affiliations at Carnegie Mellon, Johns Hopkins, Stanford and the UK Biobank

Viewpoints

- Unless otherwise indicated, all opinions and predictions are my personal ones and do not necessarily reflect those of my employer or affiliated entities

Accelerated Genomics Ecosystem



Instruments

ISVs / Cloud Platforms

Industry and Research

Full-Stack Solutions to Accelerating Healthcare Breakthroughs

cuBLAS
cuFFT
NPP

Holoscan

Isaac

MONAI
cuCIMI

cuEquivariance
cuTensor

BioNeMo

Parabricks

RAPIDS
MSA-Search

RIVA
NeMo
ACE

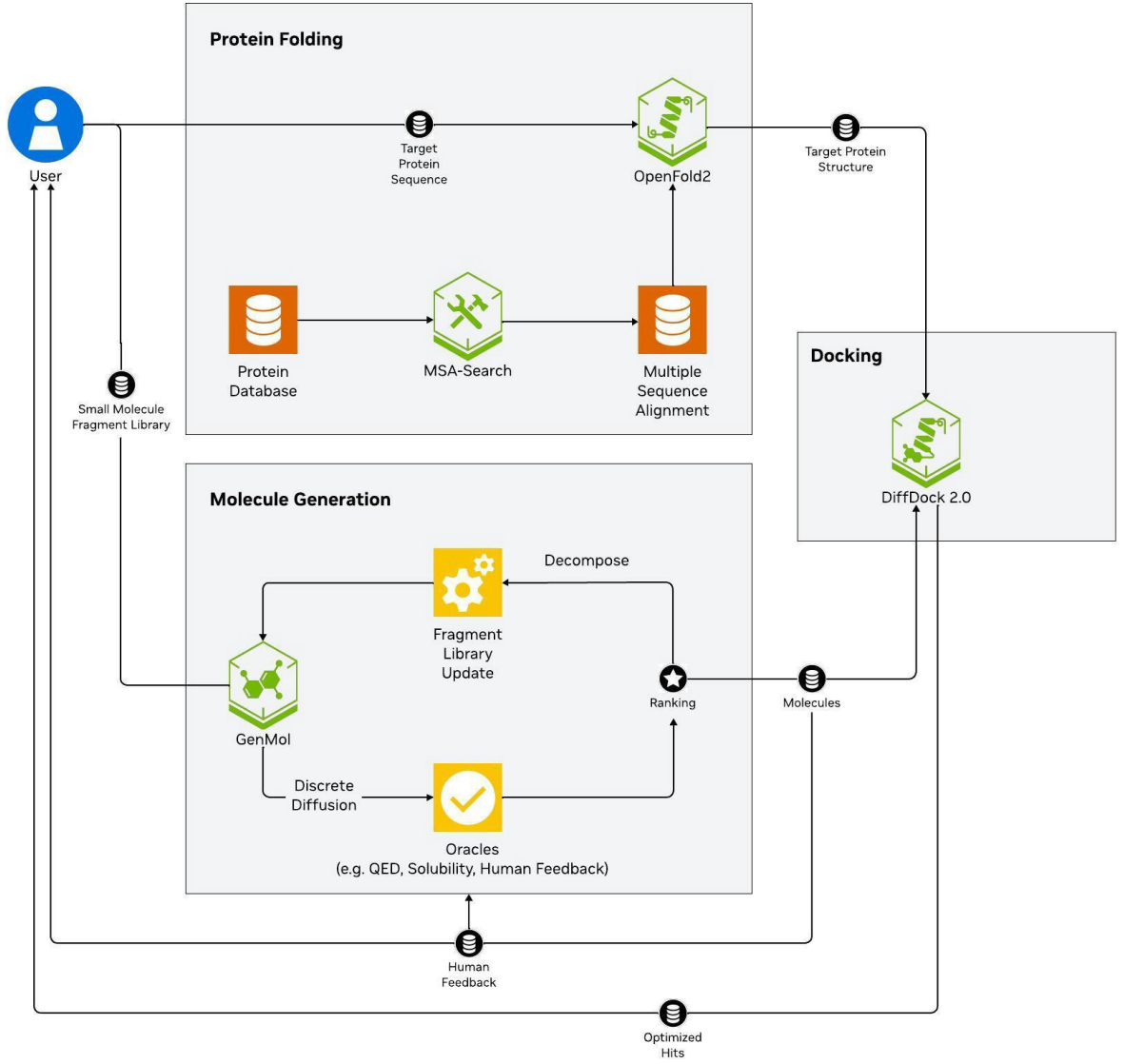
TensorRT-LLM
NCCL
Dynamo

BioNeMo Blueprints: Generative Virtual Screening

Models: MSA-Search, AlphaFold2, GenMol, DiffDock

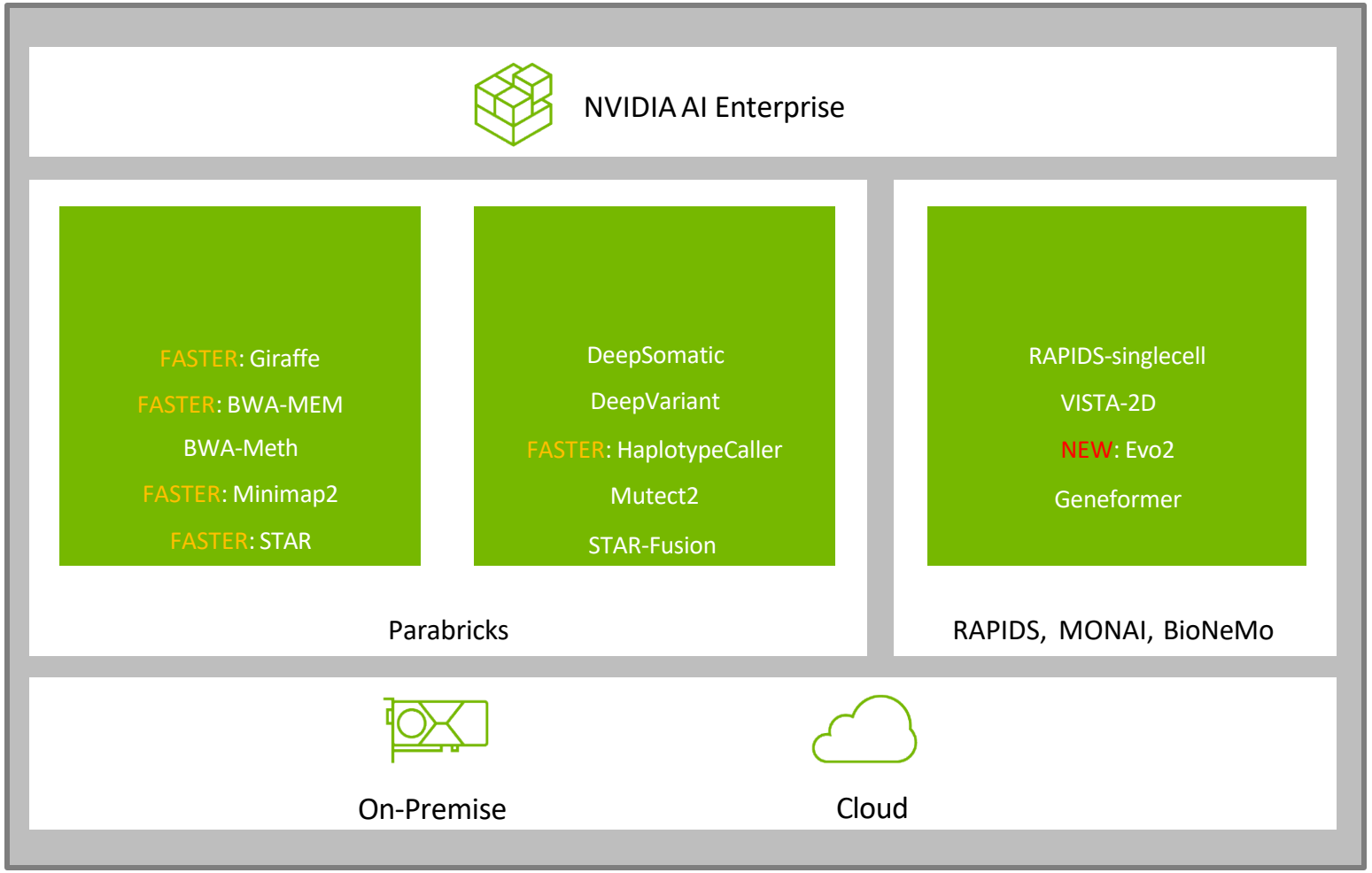
Benefits

- Use generative AI to more efficiently explore chemical space to optimize molecular designs for multiple features simultaneously
- Accelerated NIMs allow rapid evaluation of large molecule databases to identify better drug candidates faster
- Test fewer molecules to identify virtual hits, reducing the time and cost of drug development



The AI & GPU-Accelerated Software Suite for Omics Analysis

Higher Accuracy, Higher Speed, Lower Cost



Increase Speed

Experience **135x** faster analysis of WGS compared to CPU-only

Reduce Cost

Up to **50% lower** compute cost for WGS compared to CPU-only

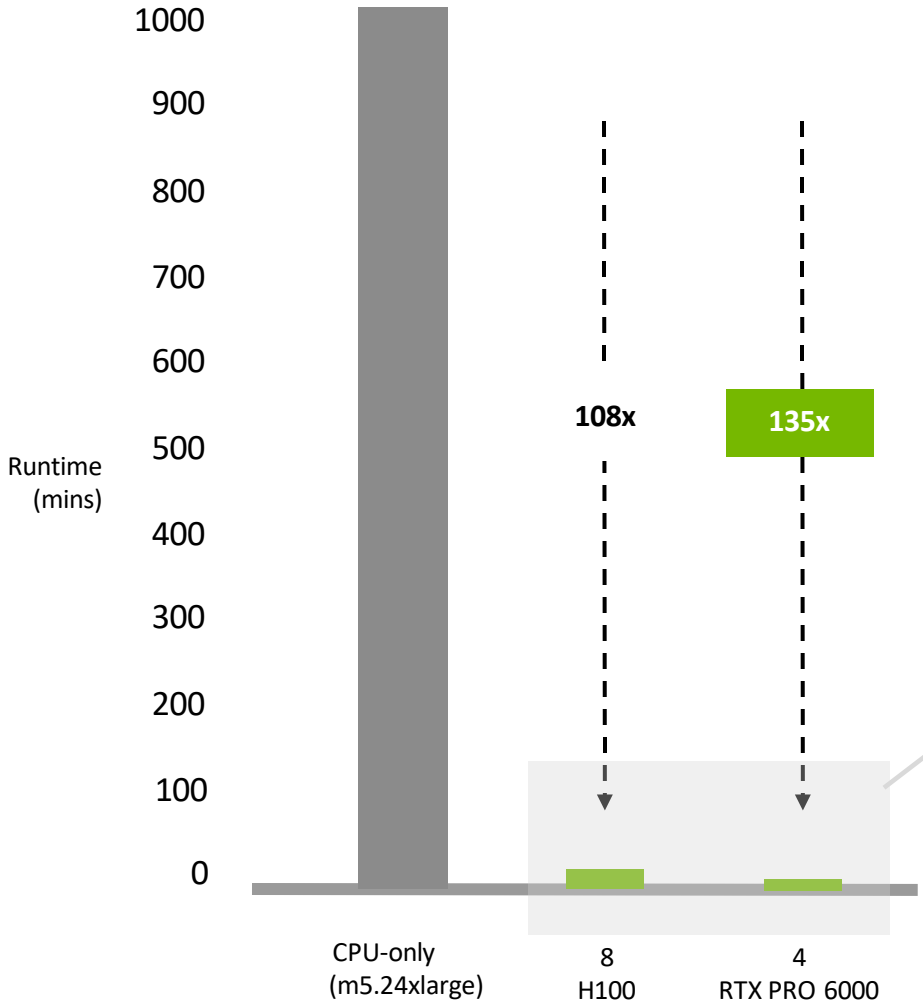
Boost Accuracy

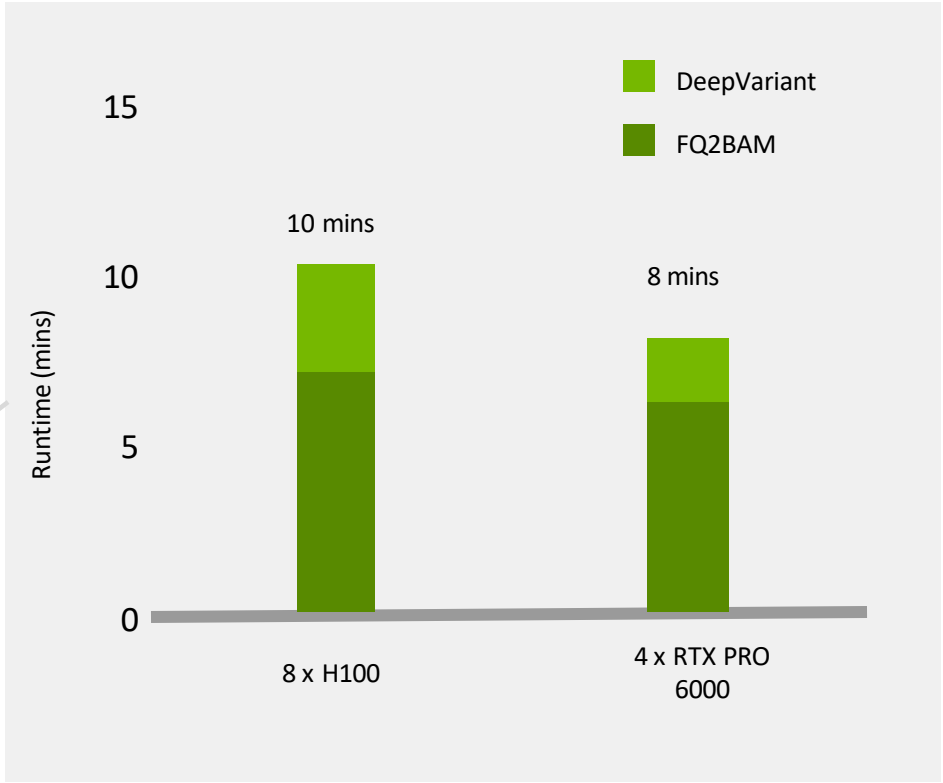
High accuracy deep learning and pangenome alignment

Higher Speed: Germline Analysis from 18 hours to 8 minutes

135x acceleration using RTX PRO 6000

Germline workflow runtime per whole genome
(HG002, 30x Illumina)

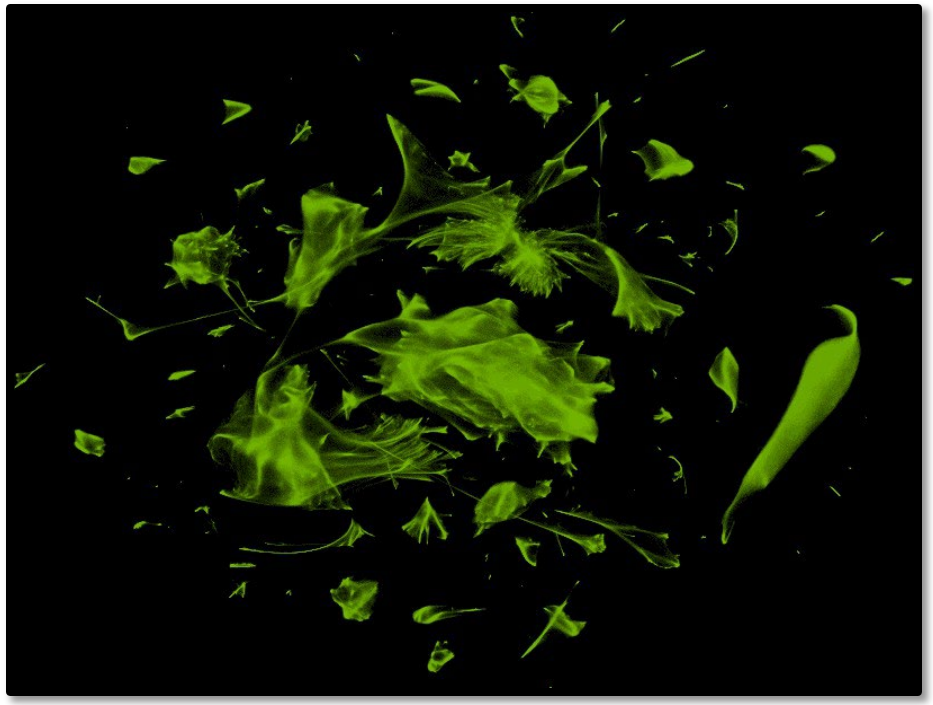




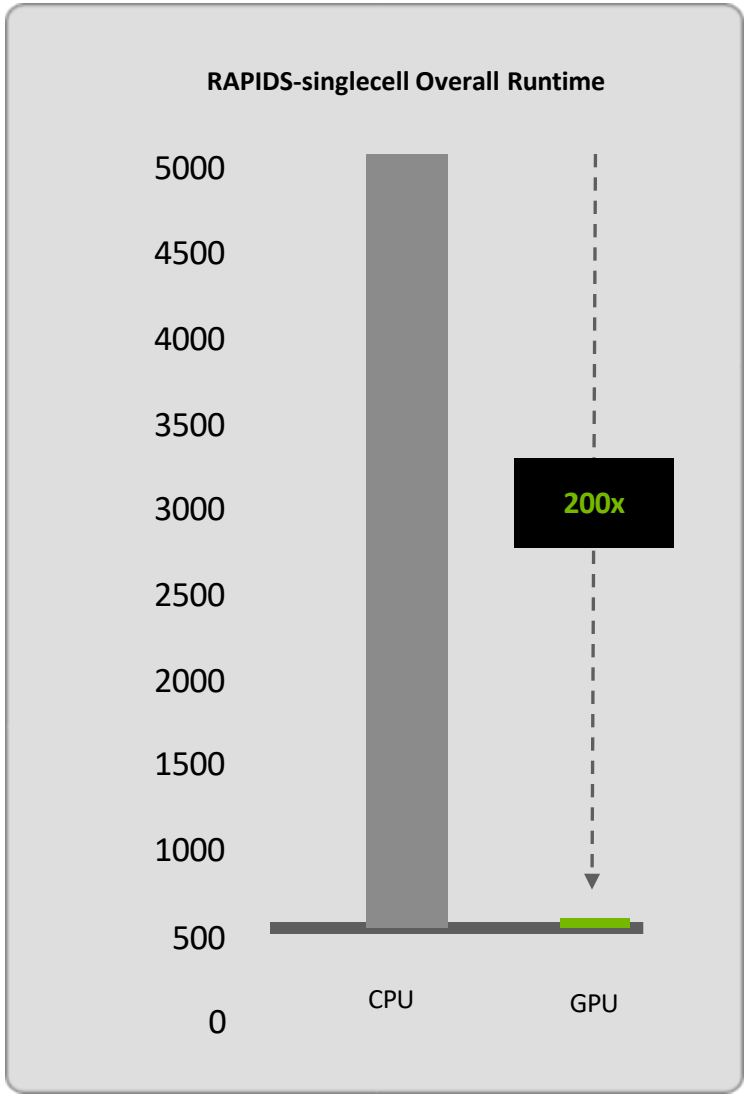
Analyze Orders of Magnitude More Data with RAPIDS-singlecell

Validate in real cells to enable scientific exploration and unlock biological insights

RAPIDS-singlecell
Introduces GPU-optimized versions of the
ScanPy library functions.



1 Million Cell Dataset
676x faster UMAP and **70x** faster PCA



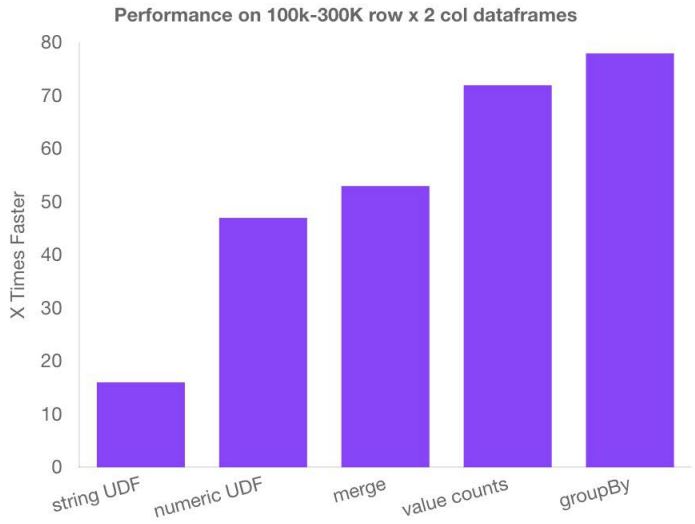
RAPIDS-singlecell developed by the scverse

GPU-ACCELERATED Data Science

FASTER PANDAS WITH CUDF

cuDF accelerates pandas with zero code changes and brings greatly improved performance.

Run this benchmark yourself ↗



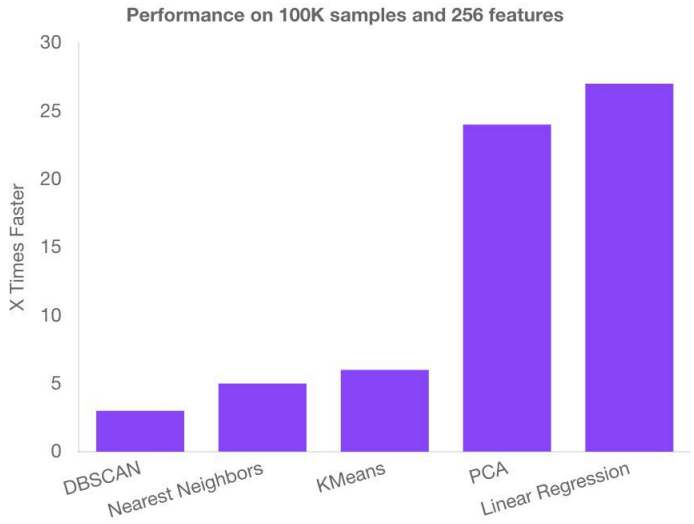
* Benchmark on AMD EPYC 7642 (using 1x 2.3GHz CPU core) w/ 512GB and NVIDIA A100 80GB (1x GPU) w/ pandas v1.5 and cuDF v23.02

Hash Maps

FASTER SCIKIT-LEARN WITH CUML

cuML brings huge speedups to ML modeling with an API that matches scikit-learn.

Run this benchmark yourself ↗



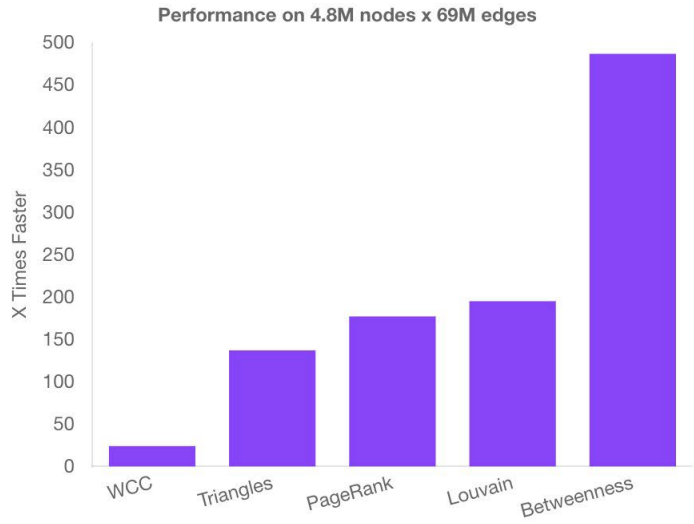
* Benchmark on AMD EPYC 7642 (using 1x 2.3GHz CPU core) w/ 512GB and NVIDIA A100 80GB (1x GPU) w/ scikit-learn v1.2 and cuML v23.02

Includes RF!

FASTER NETWORKX WITH CUGRAPH

cuGraph accelerates NetworkX with zero code changes for much greater performance at scale.

Run this benchmark yourself ↗



* Benchmark on Intel(R) Xeon(R) w9-3495X w/ 250 GB and NVIDIA A100 80GB (1x GPU) w/ NetworkX v3.4.1 and cuGraph/nx-cugraph v24.10; WCC = Weakly Connected Components; Betweenness = Betweenness Centrality with k=100

GNN + LLM

MONAI Multimodal

Data | Models | Agents

Data

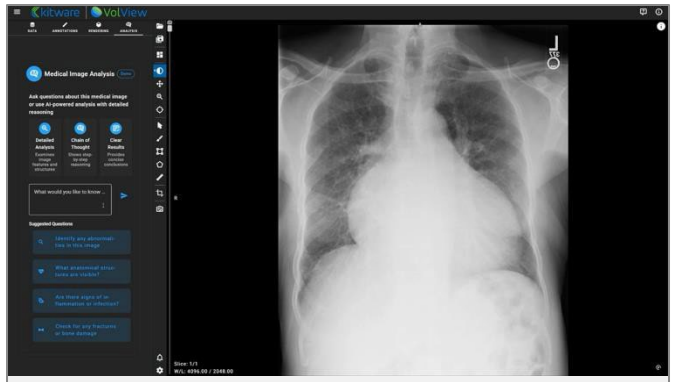
DICOM

TEXT

EHR

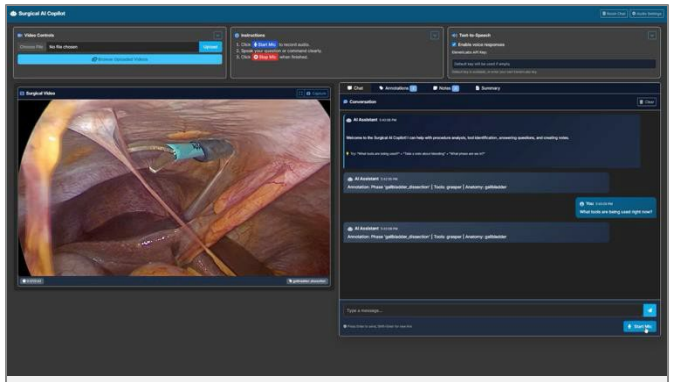
VIDEO

VOICE



Radiology Agent

Chain of Thought Reasoning



Surgical Agent

Information Retrieval & Notes

Agent Framework

Llama 3.2

Agents

Video Analysis

+

Task Models

Vision

- Patient has an early indication of something wrong from routine bloodwork
- Standard phenotypic workup is done in parallel to additional blood draw to be sent for WGS
- Pharmacological intervention is indicated by paths through graph(s), complementary to standard diagnosis criteria

The vast majority of diseases are multigenic and/or multifactorial

https://www.ncbi.nlm.nih.gov/books/NBK20363/

https://pmc.ncbi.nlm.nih.gov/articles/PMC9945947/

https://www.ncbi.nlm.nih.gov/books/NBK20363/

https://www.annualreviews.org/content/journals/10.1146/annurev-biodatasci-102022-120818

Variant Annotation is Common but Uncontextualized

JCO Clin Cancer Inform. 2020 Mar:4:310-317. doi: 10.1200/CCI.19.00132.

Integrated Informatics Analysis of Cancer-Related Variants

Kymberleigh A Pagel [1], Rick Kim [2], Kyle Moad [2], Ben Busby [3], Lily Zheng [1 4], Collin Tokheim [5], Michael Ryan [2], Rachel Karchin [1 6]

Affiliations + expand

PMID: 32228266 PMCID: PMC7113103 DOI: 10.1200/CCI.19.00132

Abstract

Purpose: The modern researcher is confronted with hundreds of published methods to interpret genetic variants. There are databases of genes and variants, phenotype-genotype relationships,

Multifactorial Disease

1. Polygenicity
2. Background genomic effects (cis- or trans-)
3. Environmental effects mediated by epigenetics
4. Direct environmental effects (immunological, receptor binding, etc)

Humans do not like thinking about multifactorial causation.

Baker, C. L., & Tenenbaum, J. B. (2023). How Occam's razor guides human decision-making. ***Proceedings of the National Academy of Sciences*****,** ***120*****(5), e2212351120.**

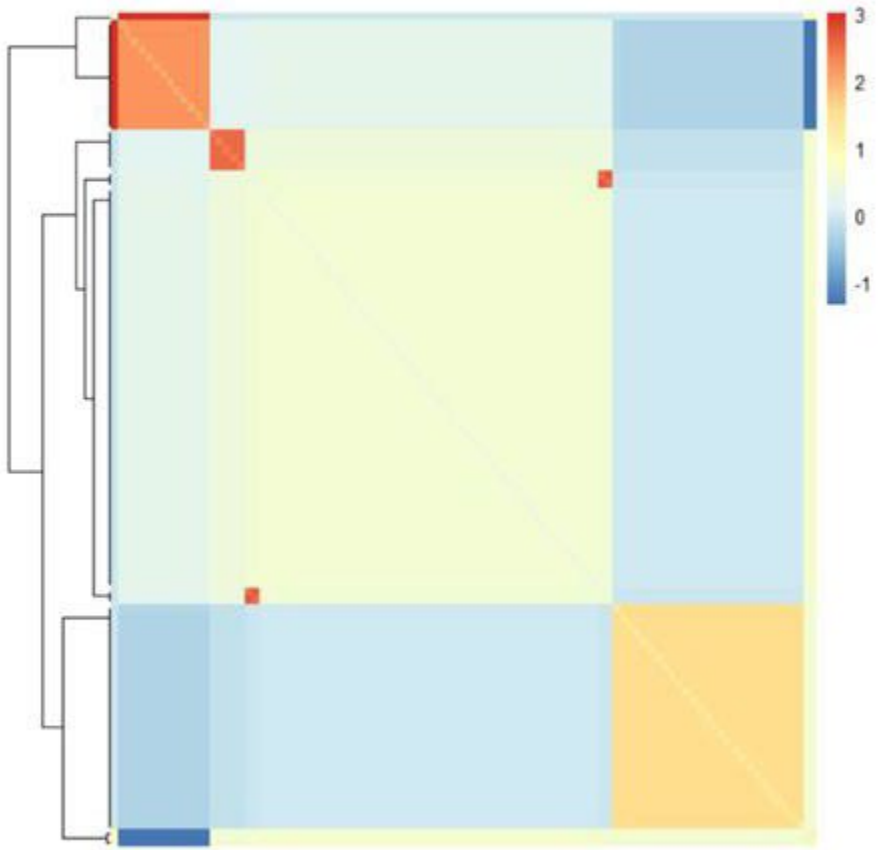
Duttle, K., & Inukai, K. (2015). Complexity Aversion: Influences of Cognitive Abilities, Culture and System of Thought. ***Economics Bulletin*****,** ***35*****(2), 846-855.**

Data

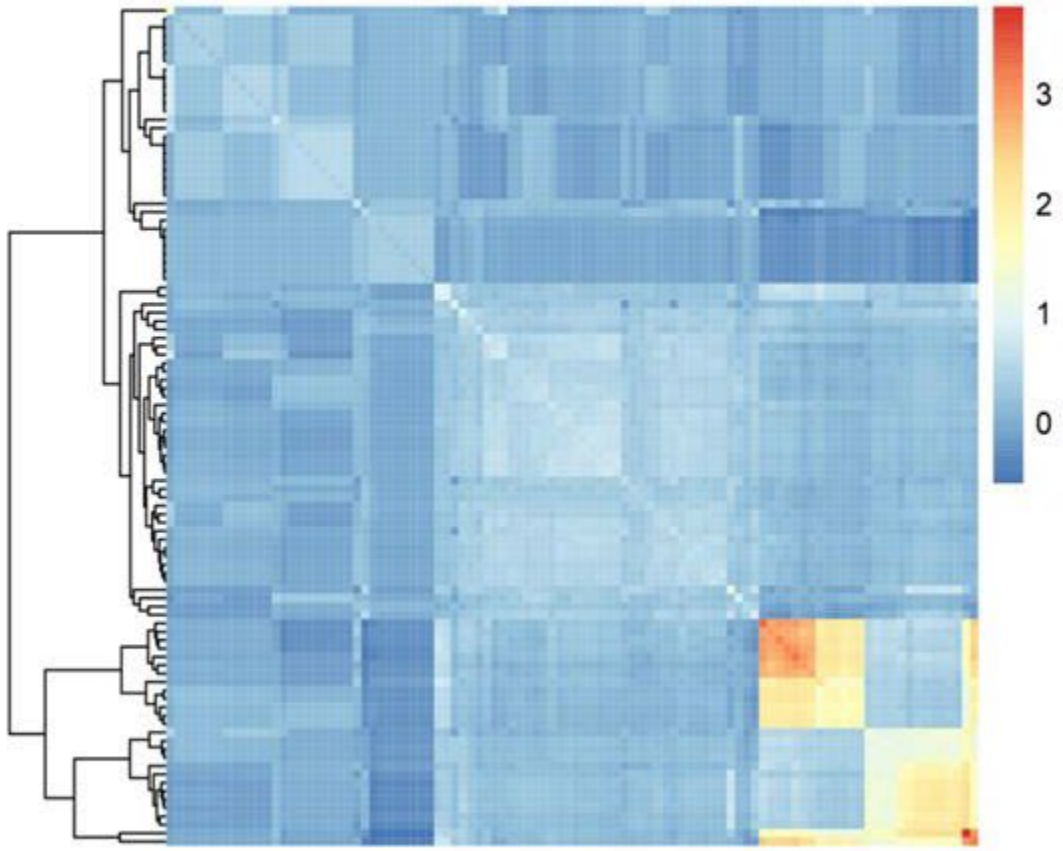
We downloaded phased VCFs for 3 populations from The 1000Genomes Project (The 1000 Genomes Project Consortium, 2015):

- British in England and Scotland (GBR): https://www.internationalgenome.org/data-portal/population/GBR
- Puerto Rican in Puerto Rico (PUR): https://www.internationalgenome.org/data-portal/population/PUR
- Chinese Dai in Xishuangbanna, China (CDX): https://www.internationalgenome.org/data-portal/population/CDX

Clustering for GRM from CDX in TNF and HLA-A

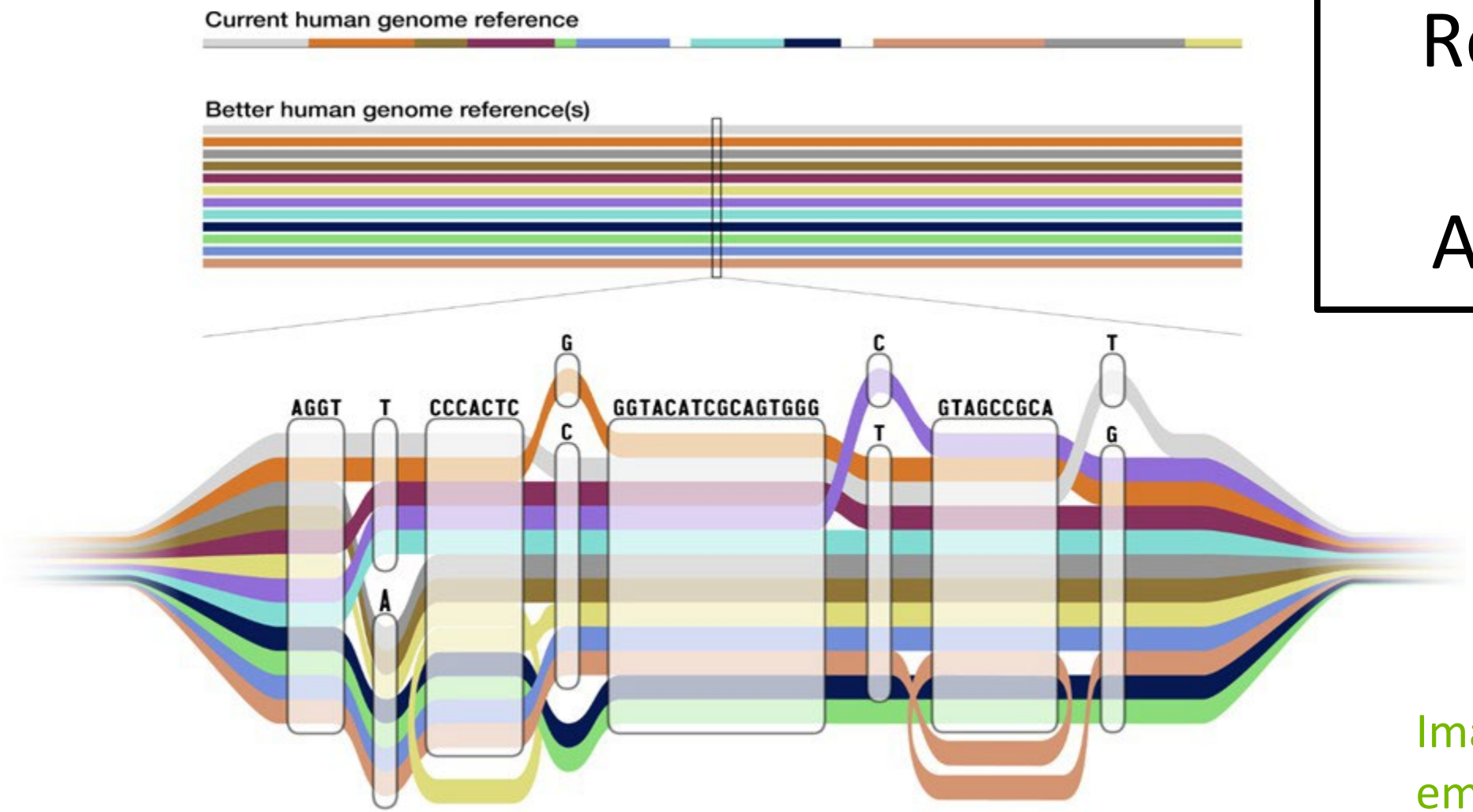


TNF



HLA-A

Genome graphs

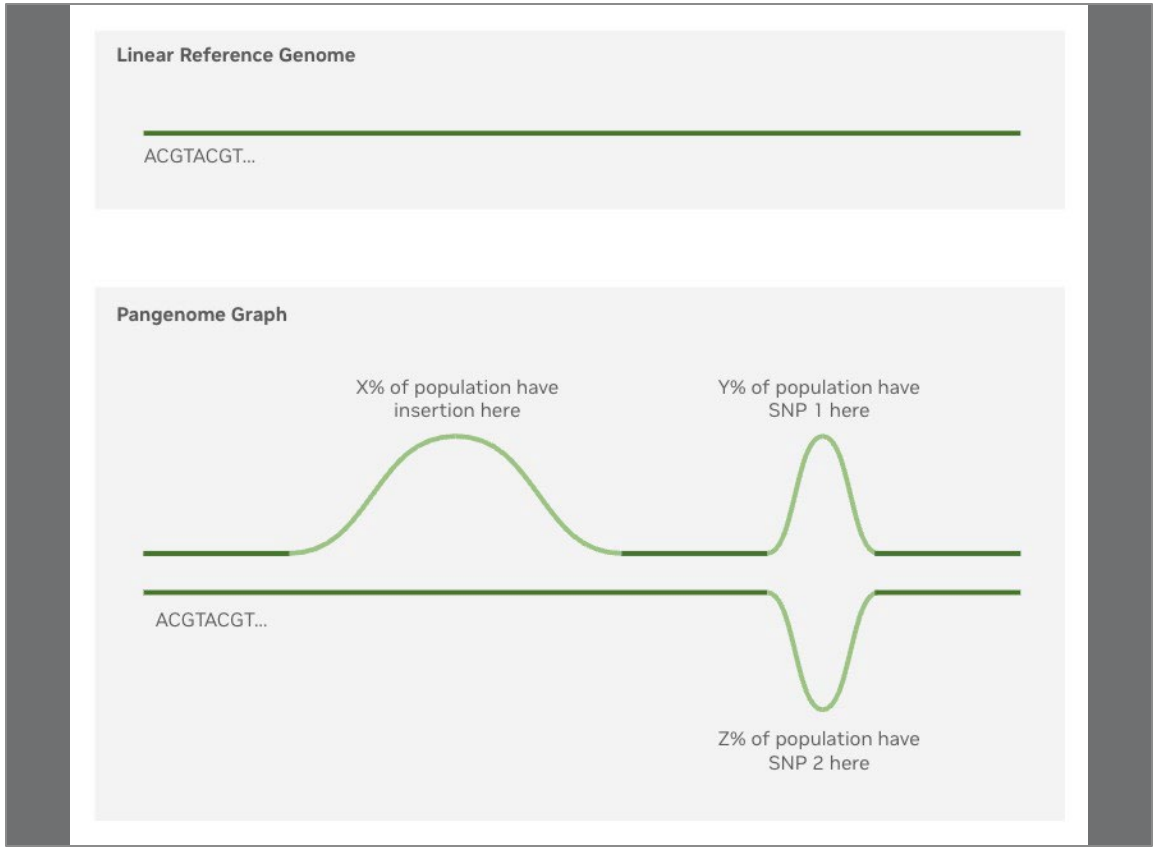


Reference Guided Assembly

Image from embl.org

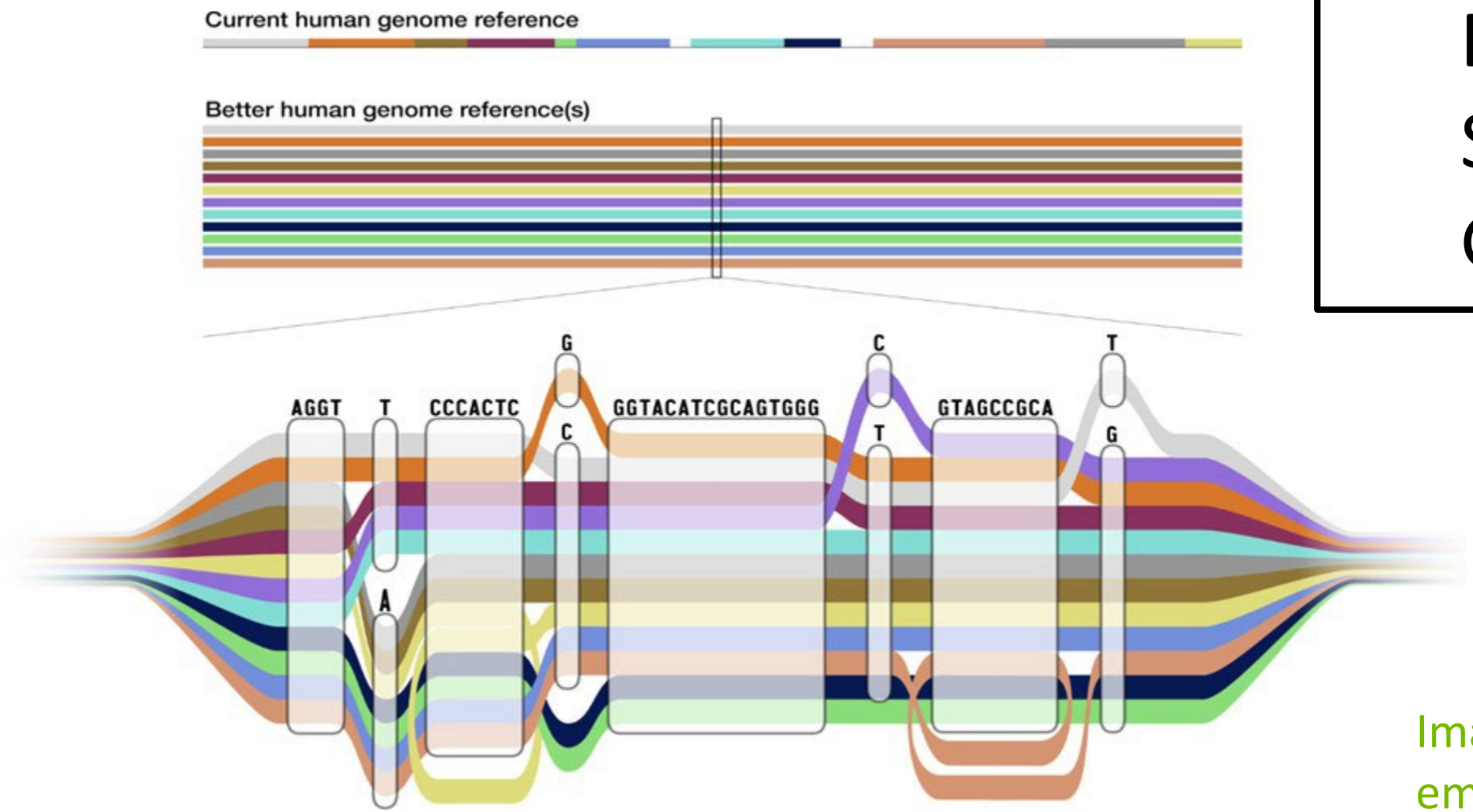
Accelerate Pangenome Alignment with Giraffe

Parabricks now supports UCSC's Giraffe



- Increase accuracy and improve variant calling—particularly across genetic variations and diverse populations
- GPU-accelerated Giraffe with single-end and pair-end support
- Equivalent results to open-source version of Giraffe

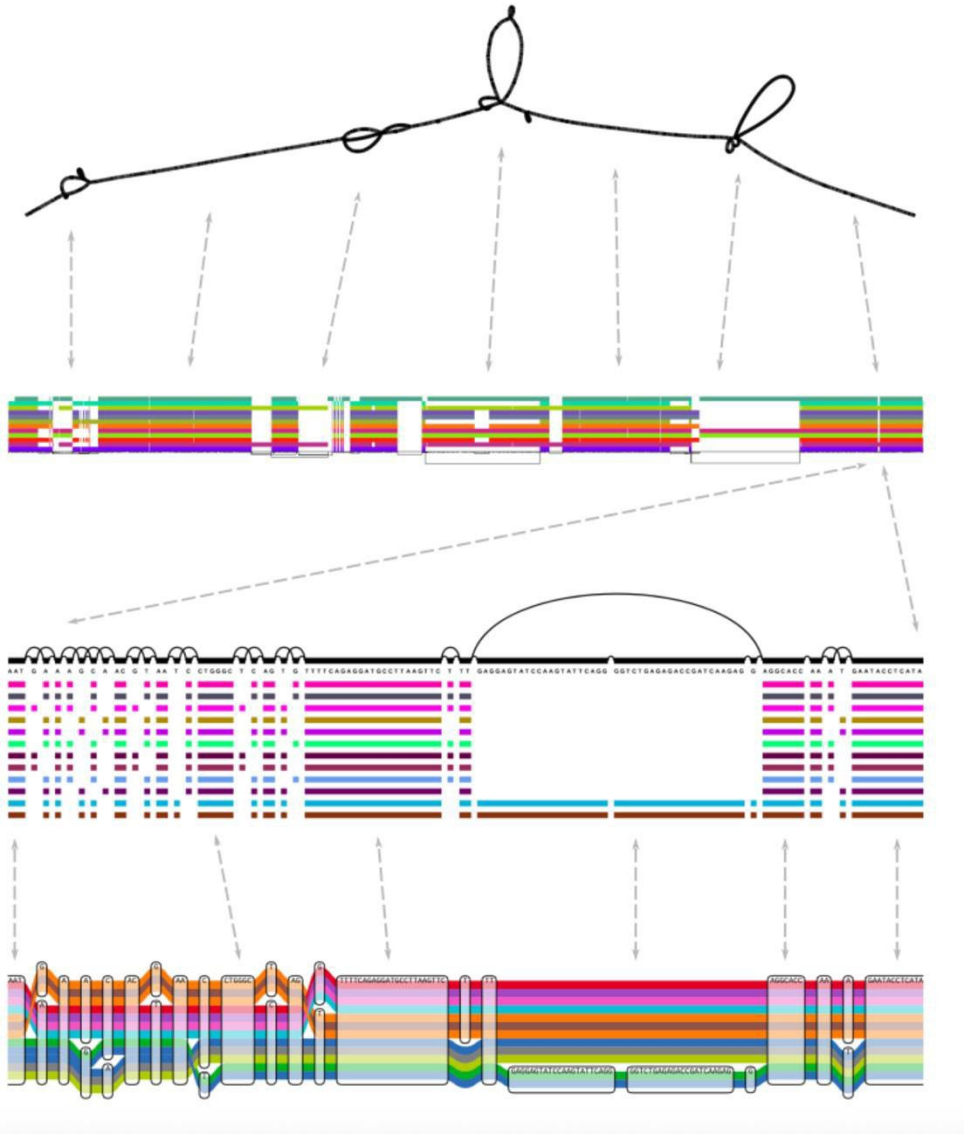
Genome graphs



Disease Specific Graphs

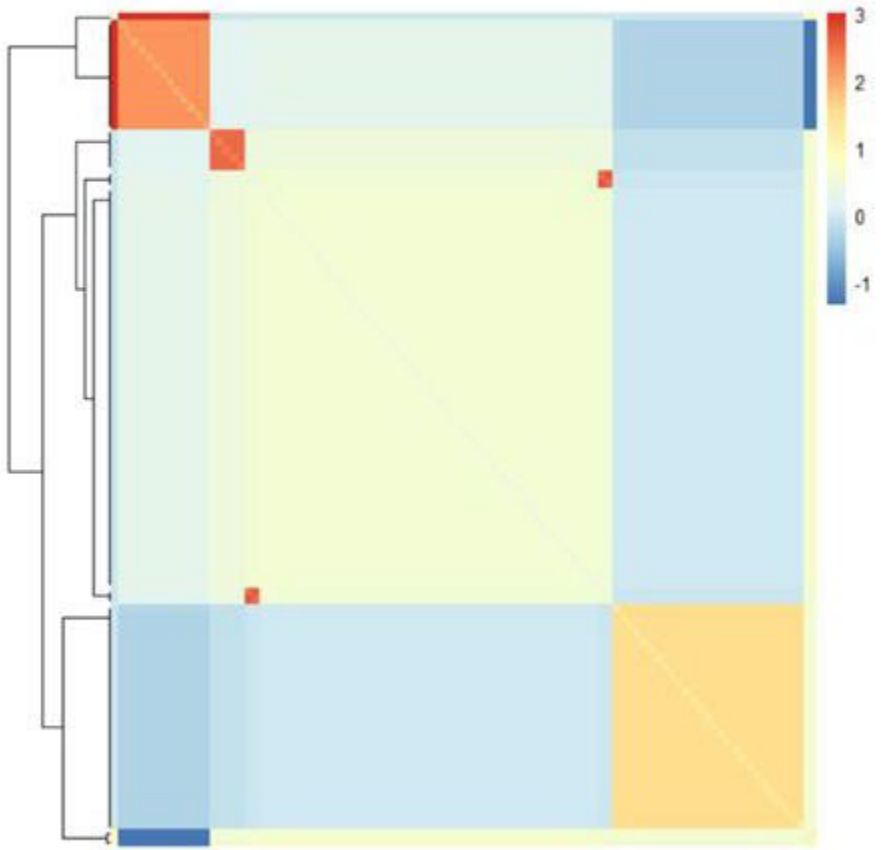
Image from embl.org

Genome graphs ➔ Hash Maps

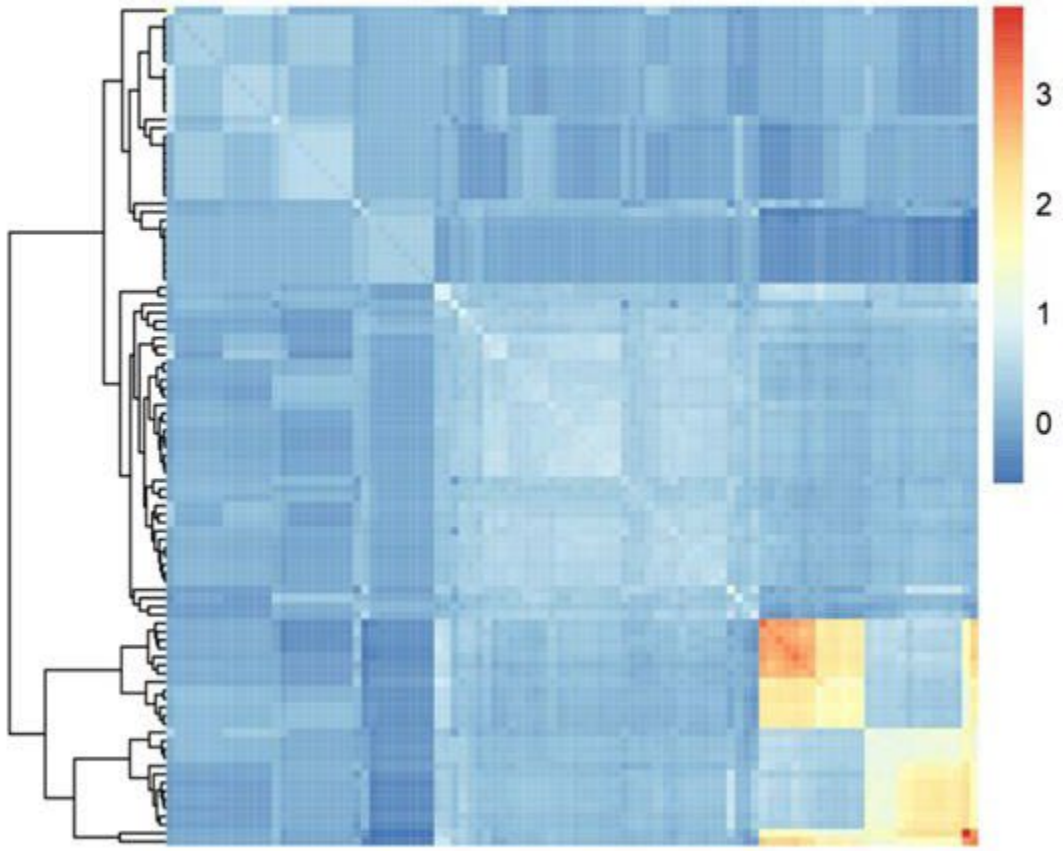


Credit:
pangenome.github.io

Clustering for GRM from CDX in TNF and HLA-A

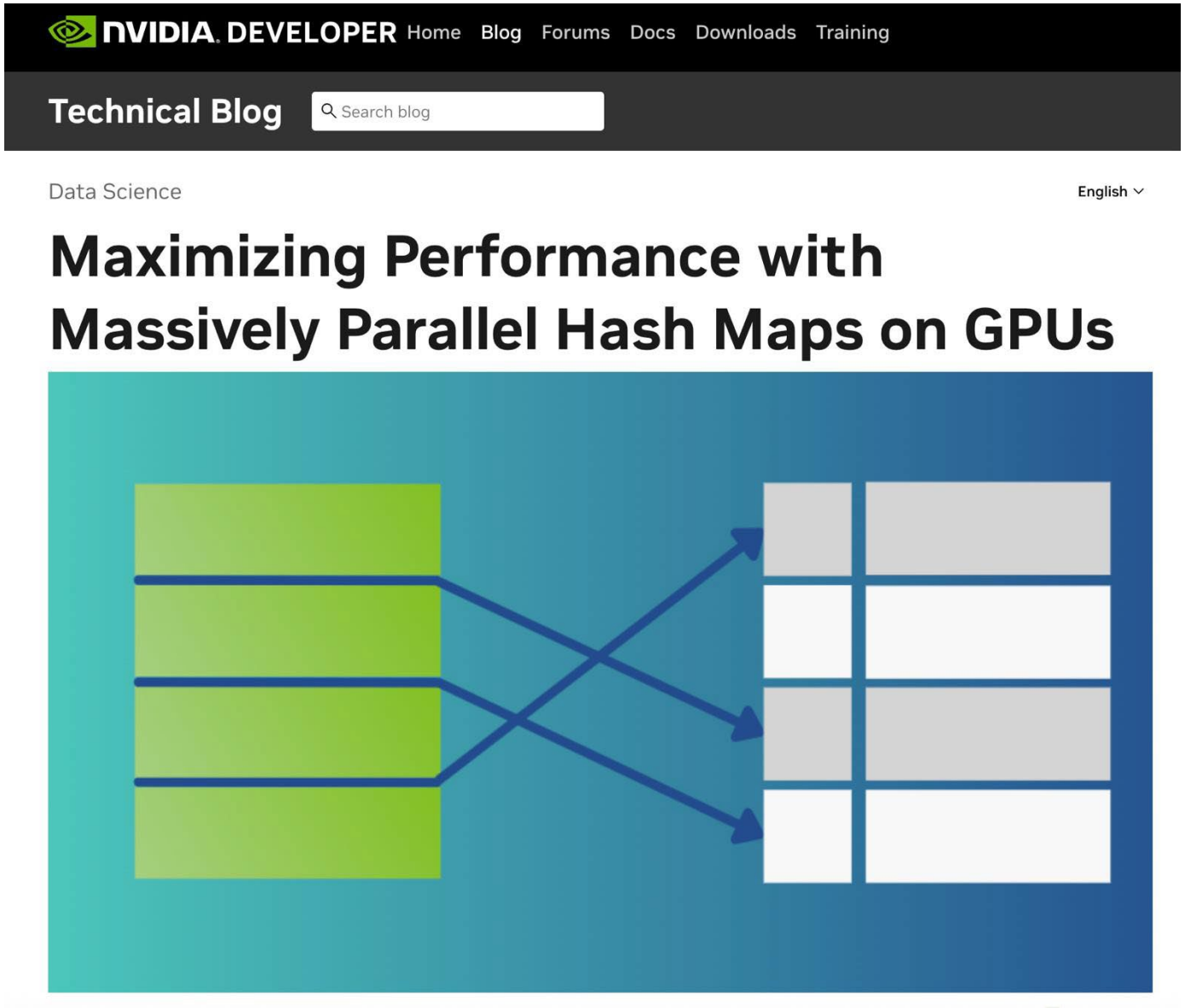


TNF

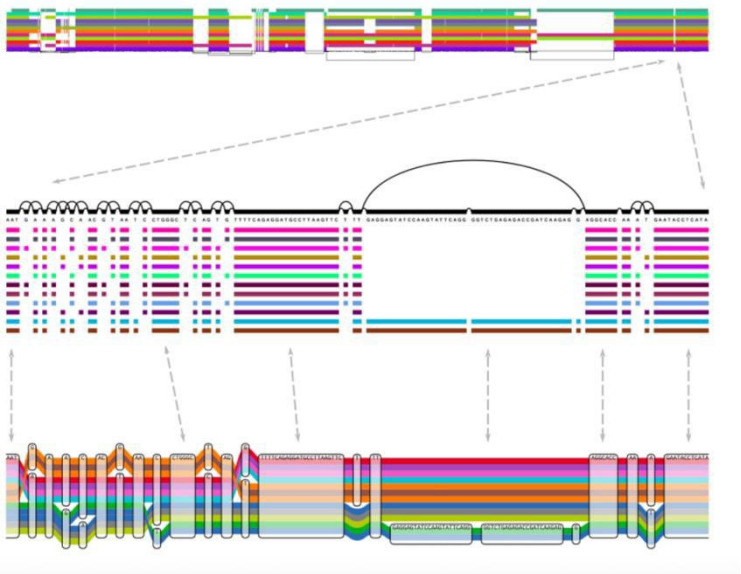


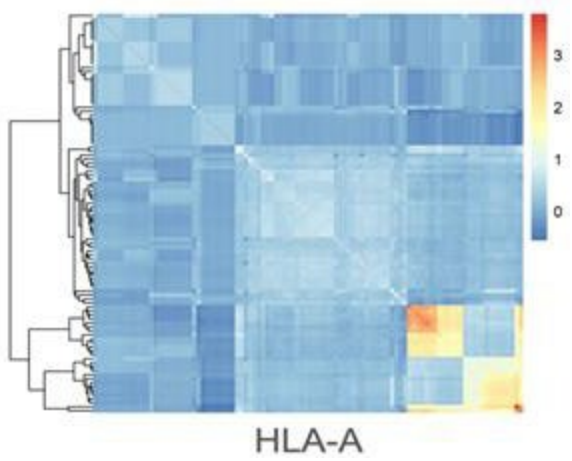
HLA-A

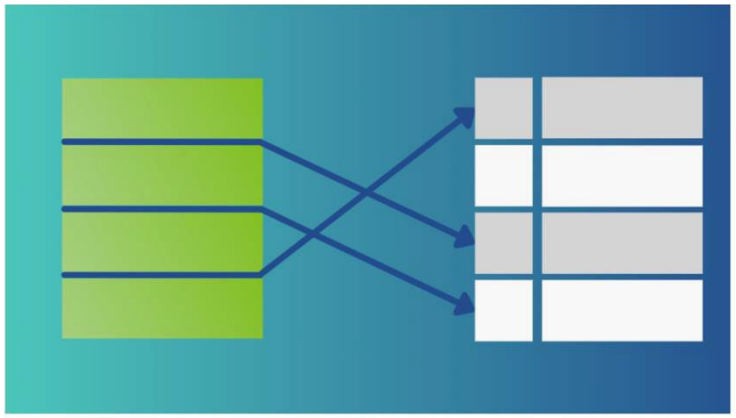
Genome graphs → Hash Maps



Genome graphs ➔ Hash Maps







Computational approaches are helping medical genomics

Accelerated Genomics (including graphs)
Single Cell and Proteomics
Agents and Models
Drug Prediction and Modification
Integration with imaging for testing
Federated Learning
Knowledge Graphs

Knowledge Graphs will allow

- Dynamic storage of data
- “Pruning” of garbage
- **More advanced clustering of disease subtypes**
  - Prediction of adequate pharmacology for these subtypes

Validating Subtype Specific Oncology Drug Predictions

Jędrzej Kubica[1,2], Emerson Huitt[3], Yusuke Suita[4], Amanda S. Khoo[5], Hyonyoung Shin[6], David Enoma[7], Nick Giangreco[8], and Ben Busby[9]

1 Laboratory of Structural Bioinformatics, Institute of Evolutionary Biology, Faculty of Biology, University of Warsaw, 00-927, Warsaw, Poland **2** Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, 00-927, Warsaw, Poland **3** Snthesis Inc., 331 W Main St STE 611, Durham, NC 27701, United States **4** Laboratory of Cancer Epigenetics and Plasticity, Therapeutic

This is the SNOMED CT graph for Breast Cancer.
SNOMED CT is a popular ontology in the US and the one that the UK health system uses.

```
graph TD
    A["Malignant neoplastic disease (disorder)"] --> B["Malignant neoplasm (morphologic abnormality)"]
    B --> C["Malignant neoplasm of breast (disorder)"]
    C --> D["Invasive malignant neoplasm of breast (disorder)"]
    C --> E["Non-invasive malignant neoplasm of breast (disorder)"]

    C -- "Is a" --> F["Malignant tumor of anatomical site (disorder)"]
    C -- "Finding site" --> G["Breast structure (body structure)"]
    C -- "Associated morphology" --> H["Malignant neoplasm, primary (morphologic abnormality)"]

    D -- "Is a" --> C
    E -- "Is a" --> C
```

```
graph TD
    subgraph Breast Cancer Hierarchy
        A["Malignant neoplastic disease (disorder)"] --> B["Malignant neoplasm (morphologic abnormality)"]
        B --> C["Malignant neoplasm of breast (disorder)"]
        C --> D["Invasive malignant neoplasm of breast (disorder)"]
        C --> E["Non-invasive malignant neoplasm of breast (disorder)"]
        C -- "Is a" --> F["Malignant tumor of anatomical site (disorder)"]
        C -- "Finding site" --> G["Breast structure (body structure)"]
        C -- "Associated morphology" --> H["Malignant neoplasm, primary (morphologic abnormality)"]
        D -- "Is a" --> C
        E -- "Is a" --> C
    end

    subgraph Molecular Subtypes
        C --> LUMA["Luminal A breast cancer (subtype)"]
        C --> LUMB["Luminal B breast cancer (subtype)"]
        C --> HER2E["HER2-enriched breast cancer (disorder)"]
        C --> TNBC["Triple-negative breast cancer (disorder)"]

        LUMA -- "Has receptor status" --> ER_POS["Estrogen receptor positive"]
        LUMA -- "Has receptor status" --> PR_POS["Progesterone receptor positive"]
        LUMA -- "Has receptor status" --> HER2_NEG["HER2 negative"]
        LUMA -- "Has proliferation index" --> KI67_LOW["Ki-67 low"]

        LUMB -- "Has receptor status" --> ER_POS
        LUMB -- "Has receptor status" --> PR_POS_VAR["Progesterone receptor positive (variable)"]
        LUMB -- "Has receptor status" --> HER2_POS_VAR["HER2 positive (variable)"]
        LUMB -- "Has proliferation index" --> KI67_HIGH["Ki-67 high"]

        HER2E -- "Has receptor status" --> HER2_POS["HER2 positive"]
        HER2E -- "Has receptor status" --> ER_NEG_VAR["Estrogen receptor negative (variable)"]
        HER2E -- "Has receptor status" --> PR_NEG_VAR["Progesterone receptor negative (variable)"]
```

```
graph TD
    subgraph Cancer Hierarchy
        A["Malignant neoplastic disease (disorder)"] --> B["Malignant neoplasm (morphologic abnormality)"]
        B --> CRC["Malignant colorectal neoplasm (disorder)"]
    end

    subgraph Colorectal Cancer Specifics
        CRC --> COLON_CA["Malignant neoplasm of colon (disorder)"]
        CRC --> RECTUM_CA["Malignant neoplasm of rectum (disorder)"]

        COLON_CA --> ADENO_COLON["Adenocarcinoma of colon (disorder)"]
        RECTUM_CA --> ADENO_RECTUM["Adenocarcinoma of rectum (disorder)"]

        CRC -- "Finding site" --> COLON_STRUCT["Colon structure (body structure)"]
        CRC -- "Finding site" --> RECTUM_STRUCT["Rectum structure (body structure)"]
        CRC -- "Associated morphology" --> MALIGNANT_MORPH["Malignant neoplasm, primary (morphologic abnormality)"]

        ADENO_COLON -- "Associated morphology" --> ADENO_MORPH["Adenocarcinoma (morphologic abnormality)"]
        ADENO_RECTUM -- "Associated morphology" --> ADENO_MORPH
    end
```

```
graph TD
    subgraph Cancer Hierarchy
        A["Malignant neoplastic disease (disorder)"] --> B["Malignant neoplasm (morphologic abnormality)"]
        B --> CRC["Malignant colorectal neoplasm (disorder)"]
    end

    subgraph Colorectal Cancer Specifics
        CRC --> COLON_CA["Malignant neoplasm of colon (disorder)"]
        CRC --> RECTUM_CA["Malignant neoplasm of rectum (disorder)"]

        COLON_CA --> ADENO_COLON["Adenocarcinoma of colon (disorder)"]
        RECTUM_CA --> ADENO_RECTUM["Adenocarcinoma of rectum (disorder)"]

        CRC -- "Finding site" --> COLON_STRUCT["Colon structure (body structure)"]
        CRC -- "Finding site" --> RECTUM_STRUCT["Rectum structure (body structure)"]
        CRC -- "Associated morphology" --> MALIGNANT_MORPH["Malignant neoplasm, primary (morphologic abnormality)"]

        ADENO_COLON -- "Associated morphology" --> ADENO_MORPH["Adenocarcinoma (morphologic abnormality)"]
        ADENO_RECTUM -- "Associated morphology" --> ADENO_MORPH
    end

    subgraph Consensus Molecular Subtypes (CMS)
        CRC --> CMS1["CMS1 (MSI Immune)"]
        CRC --> CMS2["CMS2 (Canonical)"]
        CRC --> CMS3["CMS3 (Metabolic)"]
        CRC --> CMS4["CMS4 (Mesenchymal)"]

        CMS1 -- "Has characteristic" --> MSI_H["Microsatellite instability-high (MSI-H)"]
        CMS1 -- "Has characteristic" --> BRAF_MUT["BRAF gene mutation"]
        CMS1 -- "Has characteristic" --> IMMUNE_ACT["Strong immune activation"]
```

A Glimpse into the Future

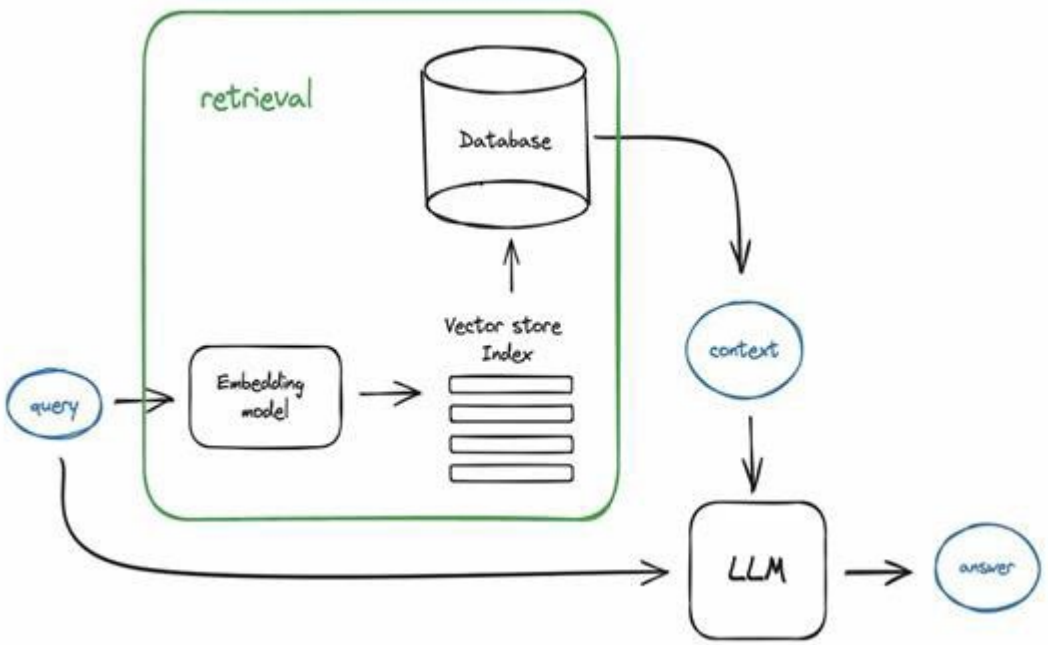
This image aims to capture a futuristic landscape, blending advanced technology with a harmonious environment, symbolizing progress and innovation.

A Glimpse into the Future

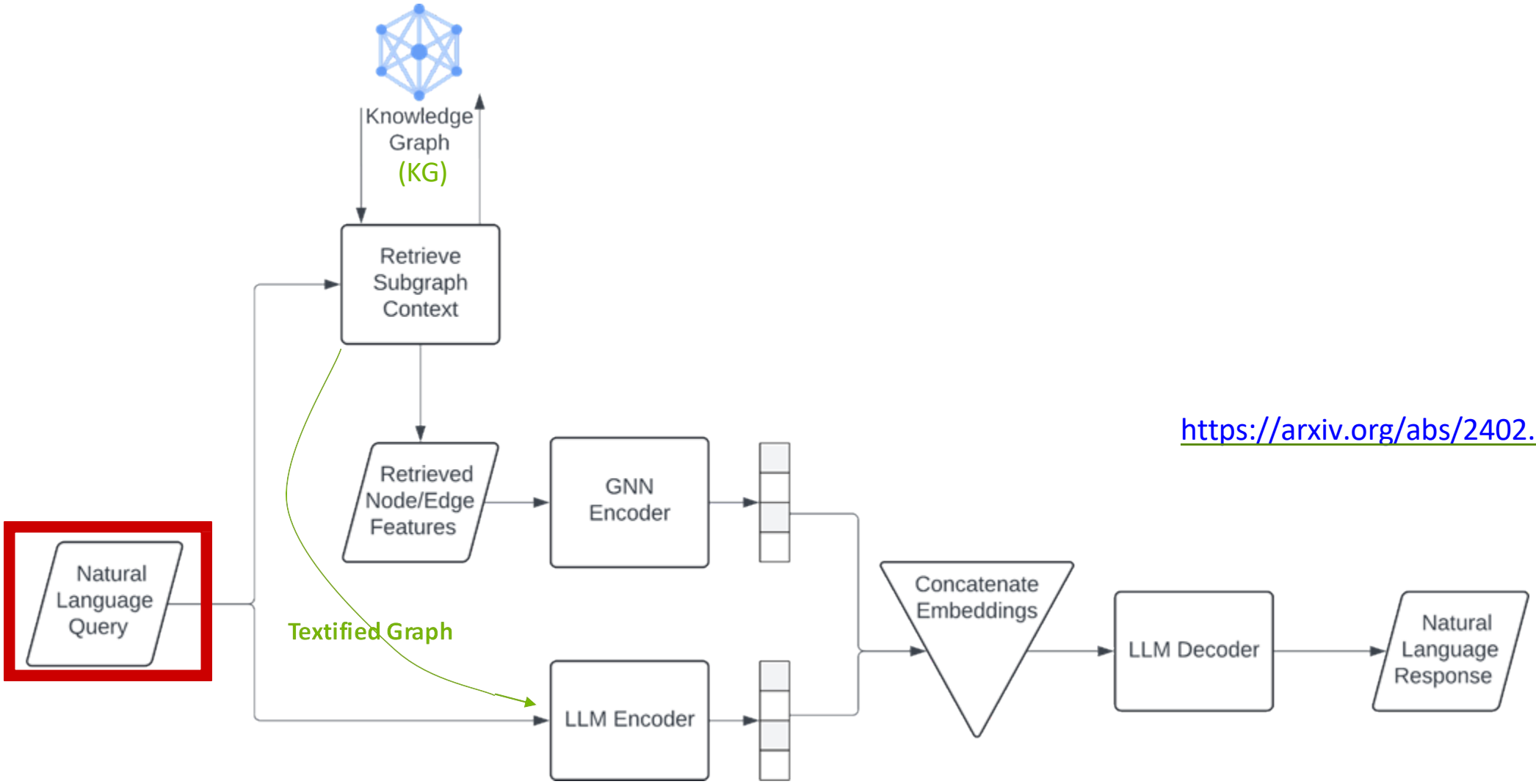
This image aims to capture a fu rmonious environment, symbolizing progress and innovation.

RAG: VectorRAG vs GraphRAG

- RAG = Retrieval Augmented Generation
- VectorRAG: retrieve top K relevant docs based on their embedding **vector**
  - Good enough when answer requires single doc
- GraphRAG: retrieve relevant sub**graph**
  - Good when answer requires multiple docs with related entities

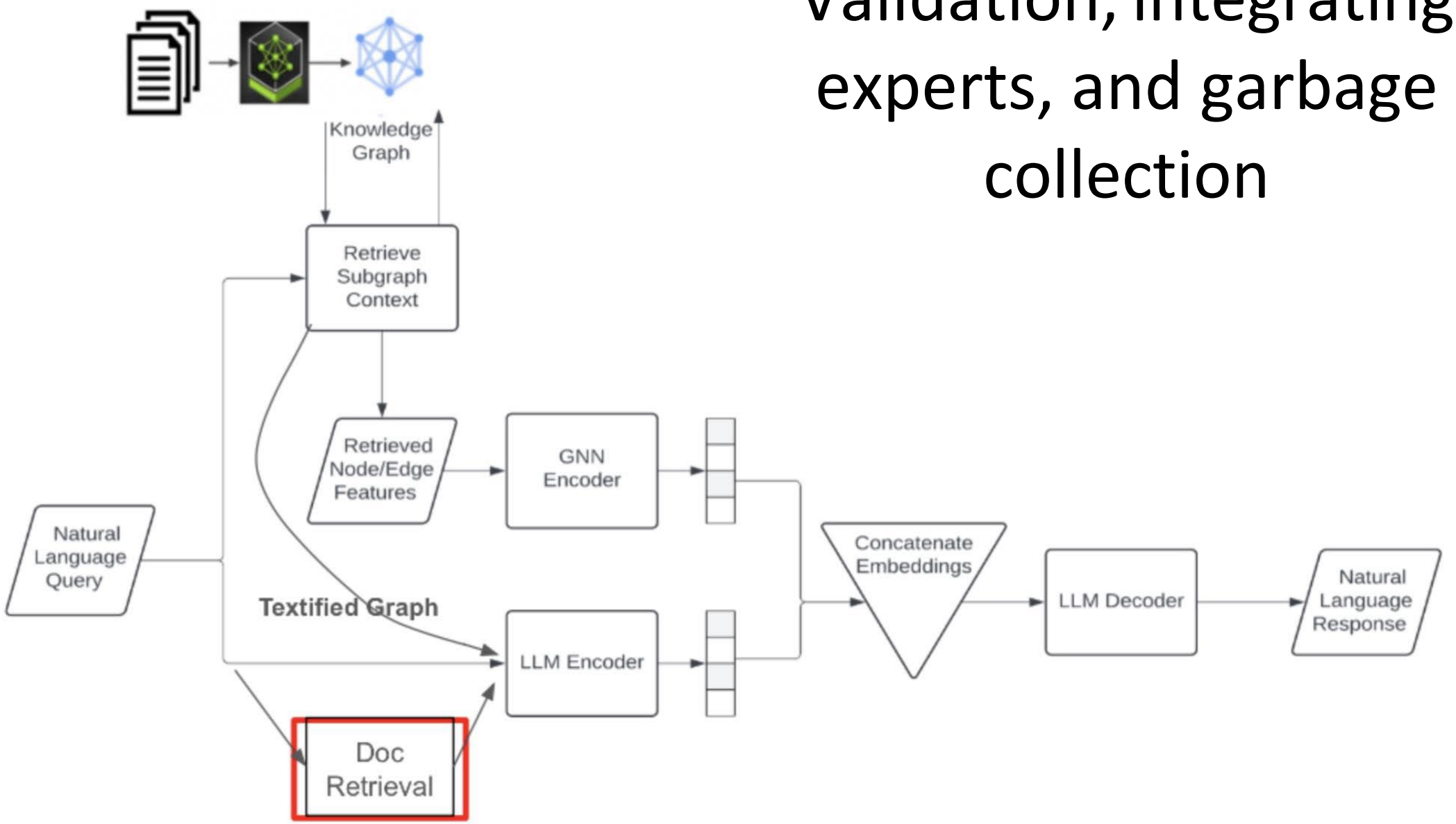


GNN+LLM Graph RAG (GNN Feeds LLM)



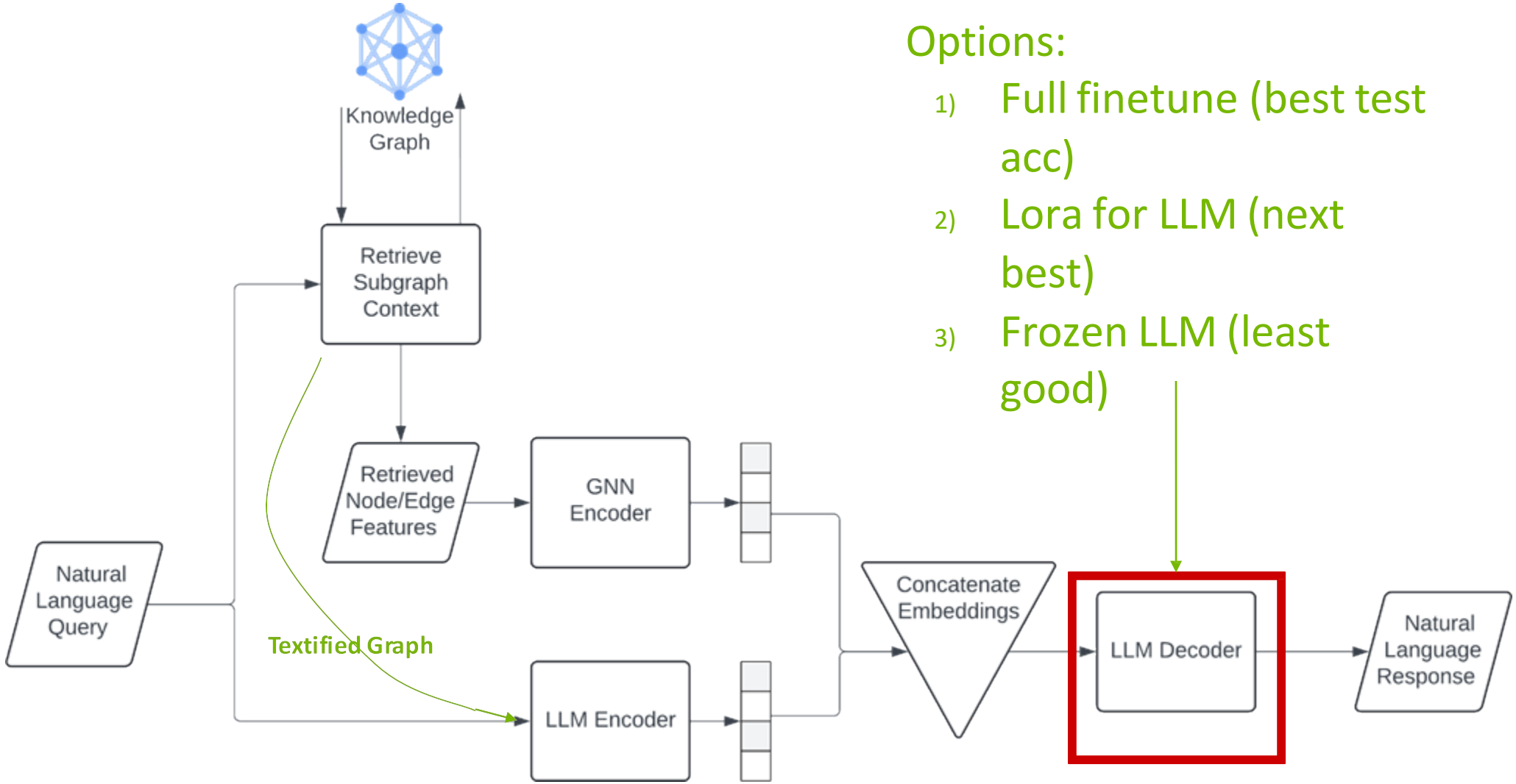
https://arxiv.org/abs/2402.07630

GraphRAG+VectorRAG



Validation, integrating experts, and garbage collection

GNN+LLM Graph RAG (GNN Feeds LLM)

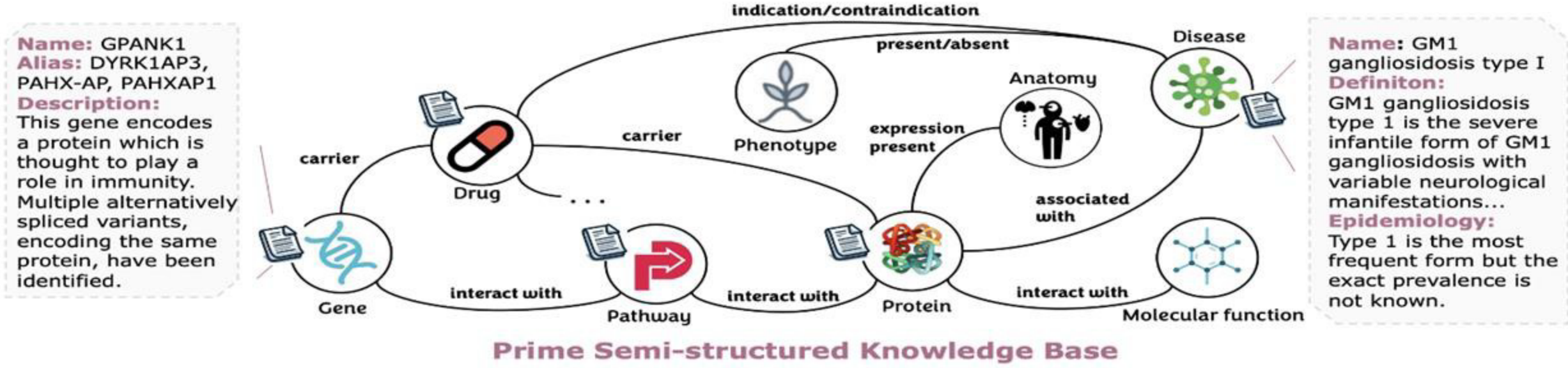


Options:

1) Full finetune (best test acc)
2) Lora for LLM (next best)
3) Frozen LLM (least good)

Neo4j Case Study

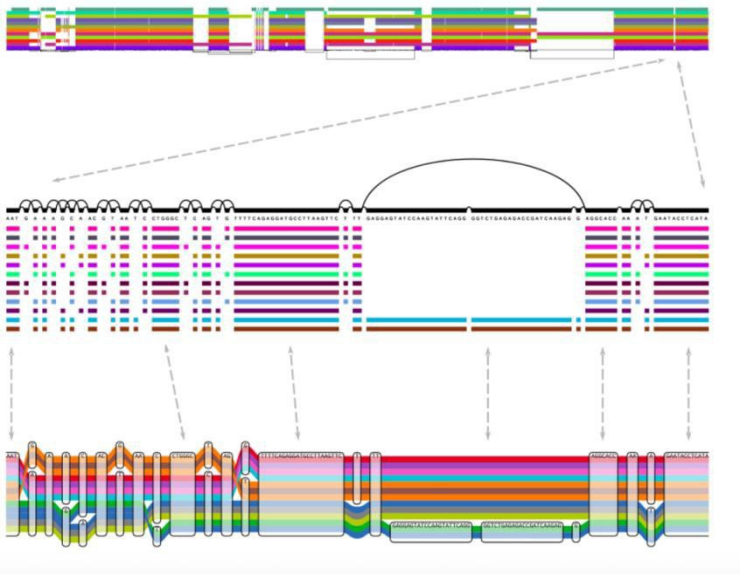
https://developer.nvidia.com/blog/boosting-qa-accuracy-with-graphrag-using-pyg-and-graph-databases/

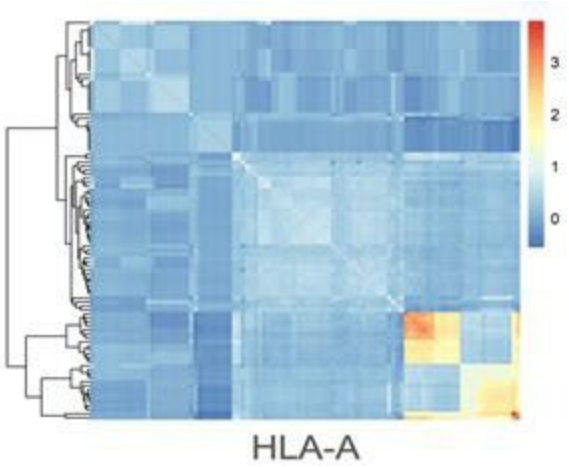


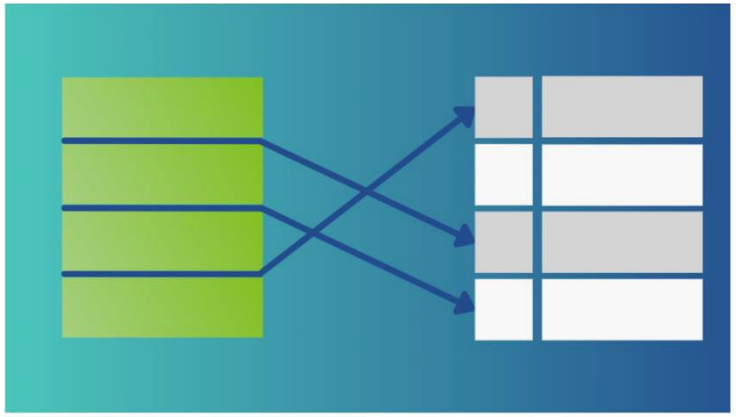
Computational approaches are helping medical genomics

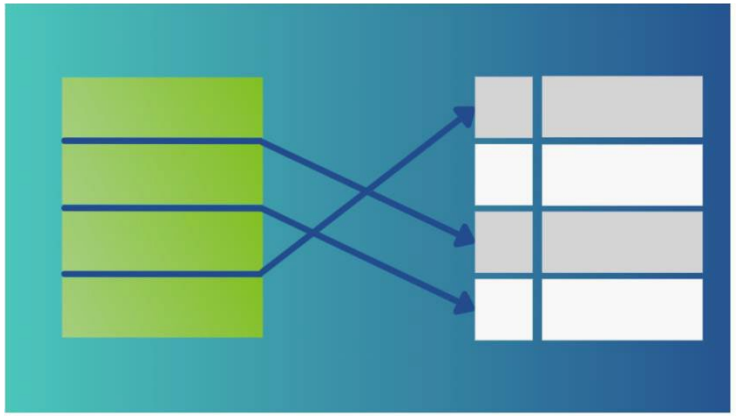
Accelerated Genomics (including graphs)
Single Cell and Proteomics
Agents and Models
Drug Prediction and Modification
Integration with imaging for testing
Federated Learning
Knowledge Graphs

Hash Maps ➔ Models

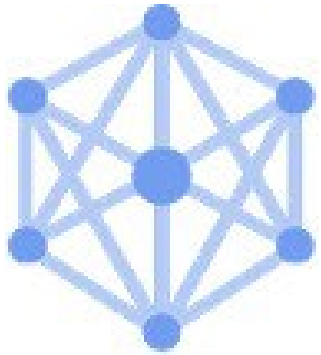




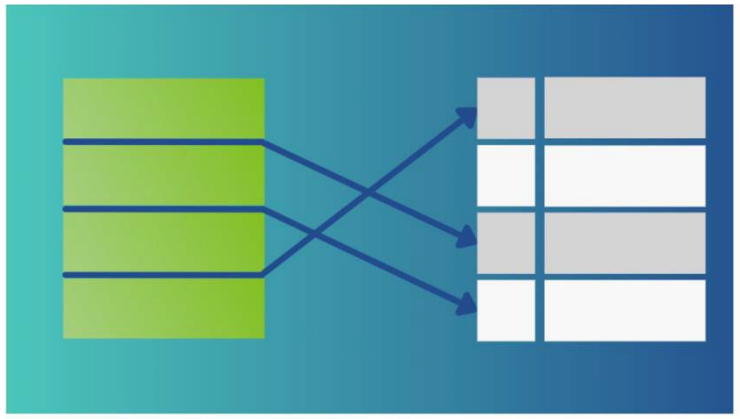




Phenotypic Knowledge Graph



Hash Maps ➔ Models



Phenotypic Knowledge Graph

Take Home Messages

- Changing genomic data structures will help analyze multifactorial genomic etiologies
- Building biological knowledge graphs of phenotypic data -- including derived phenotypes from other primary data sources -- will help to contextualize phenotypic information
- Putting effort into the above community initiatives is likely to help us deliver more precise treatments faster.

Acknowlegements!!

- **Jason Fenwick**
- **Chelsea Gomatam**
- TJ Chen
- Daniel Burkhardt
- Nick Venanzi
- Brad Genereaux
- Stephen Aylward
- **Rishi Puri**
- Yang Liu
- Xin Yu
- Danielle Short
- Eric Dawson
- Pankaj Vats
- Brian Welker
- **Jedrzej Kubica**
- **Maria Chikina**
- Halimat Chisom
- Rajarshi Mondal
- Li Chuin Chong
- **Jon Moller**
- **Hongsheng Lai**
- **Shijie Tang**
- Xirui Liu
- Zhiwen Bian
- Hairuo Wang
- Ali Saadat V
- Daniel Chang
- William Lu
- Emrah Kacar
- Avish Jha
- Francesco Andreace
- Minal Jamsandekar,
- Umran Yaman
- Eleni Mourouzidou
- Michael Olufemi
- Manasi Ghogare
- Shelby Kroeger
- Lisa Boatner
- Stanislaw Gizinski