Open-Source Project

Foundational Learning Guide

HCLS AI Factory

Introduction Level

How scientists use computers and AI to read DNA, find disease-causing changes, and design new medicines — all on a single desktop computer. No prior biology or computer science required.

NVIDIA DGX Spark | Parabricks | BioNeMo

02/2026 | Version 1.0 | Apache 2.0 License

Author: Adam Jones

Table of Contents

Chapter 1: What Is DNA?

Your Body’s Instruction Manual

Every cell in your body contains DNA — a long molecule shaped like a twisted ladder (the famous "double helix"). DNA is like an instruction manual written in a four-letter alphabet: A (adenine), T (thymine), C (cytosine), and G (guanine).

These four letters combine to form "words" called genes. Humans have about 20,000 genes, and each gene contains instructions for building a specific protein — the molecular machines that do almost everything in your body.

The Human Genome

Your complete set of DNA instructions is called your genome. It contains about 3.1 billion letter pairs, organized into 23 pairs of chromosomes (46 total). If you stretched out all the DNA in one cell, it would be about 6 feet long — but it’s coiled so tightly that it fits inside a cell nucleus smaller than the period at the end of this sentence.

What Are Variants?

No two people have exactly the same DNA (except identical twins). The differences between your DNA and a "reference" human genome are called variants. Most variants are harmless — they’re what make you unique. But some variants can cause diseases by changing how a protein works.

Key Vocabulary

DNA — the molecule that stores genetic instructions

Gene — a section of DNA that codes for one protein

Genome — your complete set of DNA (~3.1 billion letters)

Chromosome — a package of DNA (humans have 23 pairs)

Variant — a difference in your DNA compared to a reference

Chapter 2: Reading DNA — Sequencing

How Do Scientists Read DNA?

A machine called a DNA sequencer (made by companies like Illumina or Oxford Nanopore) reads your DNA by chopping it into millions of small pieces, reading each piece, and then using computers to put the puzzle back together.

Whole-Genome Sequencing (WGS)

Whole-genome sequencing reads your entire genome — all 3.1 billion letters. The version used in this platform is called "30× coverage," which means every position in your genome is read about 30 times. Reading it multiple times helps catch errors.

FASTQ Files

The sequencer produces huge files called FASTQ files. For a single person's whole genome at 30× coverage, the FASTQ files are about 200 gigabytes — that's roughly the same as 50 HD movies! The FASTQ file contains billions of short "reads" — each one about 250 letters long.

Key Numbers

Metric	Value
Human genome size	3.1 billion base pairs
Data per WGS run	~200 GB
Read length	250 letters (paired-end: two reads per fragment)
Coverage	30× (every position read ~30 times)

Chapter 3: Stage 1 — Finding Variants with GPU Power

The Computer Challenge

Once you have 200 GB of raw sequencing data, you need to:

1. Align each of the billions of short reads to the right position in the reference genome
2. Call variants — figure out where your DNA differs from the reference

On a regular computer, this takes 1-2 days. That's too slow for clinical use.

GPU Acceleration: NVIDIA Parabricks

A GPU (Graphics Processing Unit) is a special computer chip originally designed for video games. It turns out GPUs are also great at biology — they can process millions of DNA reads simultaneously.

NVIDIA Parabricks is software that uses a GPU to do both alignment and variant calling. On the NVIDIA DGX Spark (a desktop computer that costs $3,999), Parabricks completes the entire process in about 1-4 hours instead of 1-2 days. That's a 10-20× speedup!

BWA-MEM2: Alignment

BWA-MEM2 is the tool that aligns each short read to the reference genome. Think of it like finding where each puzzle piece goes in a 3.1-billion-piece puzzle. On the GPU, this takes 20-45 minutes.

Metric	Value
Duration	20-45 minutes
GPU Utilization	70-90%
Peak Memory	~40 GB
Output	Sorted BAM + BAI index

DeepVariant: Finding Differences

Google DeepVariant uses a type of AI called a convolutional neural network (CNN) — the same kind of AI used to recognize faces in photos — to identify variants. It looks at the aligned reads and determines which differences are real variants versus sequencing errors. DeepVariant is >99% accurate and takes 10-35 minutes on the GPU.

Metric	Value
Duration	10-35 minutes
GPU Utilization	80-95%
Peak Memory	~60 GB
Accuracy	>99% (CNN-based)

The VCF File

The output is a VCF (Variant Call Format) file containing every variant found. For a typical human genome, this includes about 11.7 million variants. After filtering for quality, about 3.5 million high-confidence variants remain.

Key Concepts

GPU — a computer chip that's very fast at parallel processing

Alignment — matching short reads to the reference genome

Variant calling — identifying where your DNA differs from the reference

VCF — a file listing all the variants found

Chapter 4: Stage 2 — Understanding What the Variants Mean

Not All Variants Are Equal

Of the 11.7 million variants found, most are harmless. The challenge is finding the few that actually cause disease. This is like finding a needle in a haystack — except the haystack has 11.7 million pieces of hay.

Three Annotation Databases

ClinVar

A public database maintained by the National Institutes of Health (NIH). It contains 4.1 million variants that scientists have studied and classified as:

- **Pathogenic —** known to cause disease
- **Likely pathogenic —** probably causes disease
- **VUS —** Variant of Uncertain Significance (we don't know yet)
- **Likely benign —** probably harmless
- **Benign —** known to be harmless

AlphaMissense

An AI tool from DeepMind (the same company that created AlphaFold, which won the Nobel Prize for predicting protein structures). AlphaMissense predicts how likely a variant is to cause disease, scoring each variant from 0 to 1. It covers 71 million missense variant predictions. A score above 0.564 means "likely pathogenic."

VEP (Variant Effect Predictor)

Tells you what the variant does to the protein. Does it change an amino acid? Does it break the protein? Does it have no effect? VEP classifies the impact as HIGH, MODERATE, LOW, or MODIFIER.

The Annotation Funnel

Starting with 11.7 million variants, the platform narrows down to the most important ones:

Stage	Variant Count	Filter
Raw VCF	~11.7M	—
Quality filter	~3.5M	QUAL > 30
ClinVar match	~35,616	Clinical significance annotated
AlphaMissense match	~6,831	AI pathogenicity predicted
High impact + pathogenic	~2,400	Actionable subset
In druggable genes	~847	Targetable by medicines

Vector Database: Milvus

To search through 3.5 million annotated variants quickly, the platform uses a vector database called Milvus. Each variant is converted into a list of 384 numbers (called an "embedding") using a model called BGE-small-en-v1.5. These embeddings capture the meaning of each variant, so you can search for similar variants using natural language questions.

Parameter	Value
Database	Milvus
Total embeddings	~3.5M
Embedding dimensions	384
Embedding model	BGE-small-en-v1.5
Distance metric	COSINE

Claude: AI-Powered Reasoning

Anthropic Claude is a large language model (like ChatGPT) that reads the variant evidence and the knowledge base to identify the best drug targets. It's "grounded" in the actual data — it can only cite variants and evidence that actually exist in the database.

The knowledge base contains 201 genes across 13 therapeutic areas (like neurology, oncology, and cardiovascular disease). Of these, 171 genes (85%) are known to be "druggable" — meaning scientists know how to design medicines that target them.

Key Vocabulary

- **Annotation —** adding information about what each variant means
- **Pathogenic —** disease-causing
- **Embedding —** a mathematical representation of text
- **Vector database —** a database that finds similar items by mathematical similarity
- **RAG —** Retrieval-Augmented Generation — feeding real data to an AI to ground its answers
- **Druggable —** a protein that can be targeted by a medicine

Chapter 5: Stage 3 — Designing New Medicines

From Gene to Drug

Once the AI identifies a disease-causing variant in a druggable gene, the next step is designing new medicines. This is normally a process that takes years and costs billions of dollars. The HCLS AI Factory does the first step — generating candidate molecules — in 8-16 minutes.

Step 1: Finding the Protein Structure

Before you can design a drug, you need to know what the target protein looks like in 3D. The platform queries the RCSB Protein Data Bank (PDB) — a public database of protein structures determined by X-ray crystallography and cryo-electron microscopy (Cryo-EM).

For the VCP protein (the demo target), four structures are available. The best one (called 5FTK) shows the protein with an existing drug (CB-5083) already bound to it. This tells us exactly where new drugs should attach.

Step 2: Generating New Molecules (MolMIM)

BioNeMo MolMIM is an AI from NVIDIA that generates new molecule designs. You give it a "seed" molecule (like CB-5083) and it creates 100 new molecules that are similar but different — like variations on a theme. It uses a technique called masked language modeling — the same approach that powers text AI, but applied to molecular structures instead of words.

Step 3: Checking if They Work (DiffDock)

BioNeMo DiffDock is another NVIDIA AI that predicts whether each new molecule will actually bind to the target protein. It uses a diffusion model (similar to AI image generators like DALL-E) to predict the 3D binding pose and calculate a docking score — a number that indicates how strongly the molecule binds.

A docking score below -8.0 kcal/mol is considered excellent. The best candidate in the VCP demo scores -11.4 kcal/mol — significantly better than the original CB-5083 drug (-8.1 kcal/mol).

Score (kcal/mol)	Interpretation
-12 to -8	Excellent binding affinity
-8 to -6	Good binding affinity
-6 to -4	Moderate binding affinity
> -4	Weak binding affinity

Step 4: Drug-Likeness (RDKit)

Not every molecule that binds to a protein would make a good drug. RDKit is a chemistry toolkit that checks whether each molecule has properties that would make it a practical medicine:

Lipinski's Rule of Five — a set of rules about molecular weight, fat-solubility, and other properties that predict whether a drug can be taken as a pill

QED — Quantitative Estimate of Drug-likeness — a single number (0-1) that combines multiple drug-like properties. Above 0.67 is considered drug-like

TPSA — a measure of how well the molecule can cross cell membranes

Step 5: Final Ranking

Each candidate gets a composite score based on:

30% how confident the AI is in the molecule design (generation score)

40% how well it binds to the protein (docking score)

30% how drug-like it is (QED score)

The final output is 100 ranked drug candidates with a PDF report.

Key Vocabulary

Protein structure — the 3D shape of a protein

Cryo-EM — a technique for determining protein structures using electron microscopes

Docking — predicting how a molecule fits into a protein's binding site

Lipinski's Rule of Five — rules for predicting if a molecule can be a pill

QED — a score measuring how drug-like a molecule is

Chapter 6: The VCP/FTD Demo

What Is VCP?

VCP (Valosin-Containing Protein, also called p97) is a protein that acts like a cellular recycling machine. It helps cells break down damaged or unwanted proteins through a process called the ubiquitin-proteasome system.

When VCP has a disease-causing mutation, it can cause:

Frontotemporal Dementia (FTD) — a brain disease that affects personality, behavior, and language

ALS — Amyotrophic Lateral Sclerosis — a disease that destroys motor neurons

IBMPFD — a condition affecting muscles, bones, and the brain

The Demo Variant

The variant rs188935092 (at position chr9:35065263, where G changes to A) is classified as Pathogenic by ClinVar and scores 0.87 on AlphaMissense (well above the 0.564 threshold for pathogenic).

Parameter	Value
Gene	VCP
Protein	p97 / Valosin-Containing Protein
UniProt	P55072
Variant	rs188935092 (chr9:35065263 G>A)
ClinVar	Pathogenic
AlphaMissense	0.87 (pathogenic, >0.564 threshold)
Diseases	FTD, ALS, IBMPFD
Seed Compound	CB-5083 (Phase I clinical VCP inhibitor)

Demo Results

The demo produces 100 novel VCP inhibitor candidates. The top candidate improves on the CB-5083 seed compound:

Metric	CB-5083 (Seed)	Top Candidate	Improvement
Composite score	0.64	0.89	+39%
Docking score	-8.1 kcal/mol	-11.4 kcal/mol	41% stronger binding
QED	0.62	0.81	31% more drug-like

Chapter 7: The Hardware — NVIDIA DGX Spark

A Supercomputer on Your Desk

The entire pipeline runs on the NVIDIA DGX Spark — a desktop computer that costs $3,999. Here are its specifications:

Component	Specification
GPU	NVIDIA GB10
Memory	128 GB unified LPDDR5x (CPU + GPU shared)
CPU	NVIDIA Grace ARM64, 144 cores
Storage	NVMe (fast storage for ~200 GB genomic data)
System RAM	128 GB
Price	$3,999

Why "Unified Memory" Matters

In most computers, the CPU and GPU have separate memory, and data must be copied back and forth. In DGX Spark, the CPU and GPU share the same 128 GB of memory. This eliminates copying overhead and makes everything faster.

Scaling Up

The same software that runs on a $3,999 DGX Spark can scale to larger systems:

Phase	Hardware	Price	Scale
1 — Proof Build	DGX Spark	$3,999	Single patient, desktop
2 — Departmental	DGX B200	$500K-$1M	Multiple patients simultaneously
3 — Enterprise	DGX SuperPOD	$7M-$60M+	Thousands of patients

Chapter 8: Why This Matters

Traditional vs. HCLS AI Factory

Step	Traditional	HCLS AI Factory
Sequence alignment	12-24 hours (CPU)	120-240 min (GPU)
Variant calling	8-12 hours (CPU)	10-35 min (GPU)
Annotation	Days (manual)	Minutes (automated)
Target identification	Weeks (literature review)	Minutes (Claude RAG)
Drug candidate generation	Months (medicinal chemistry)	8-16 min (BioNeMo AI)
Total	Weeks to months	< 5 hours

The Bigger Picture

This platform is part of the HCLS AI Factory — a broader ecosystem that also includes:

Imaging Intelligence Agent — AI analysis of CT scans, MRI, and X-rays

Cross-modal triggers — for example, a suspicious lung nodule found on a CT scan can automatically trigger genomic analysis to look for cancer-related variants

NVIDIA FLARE — technology that lets multiple hospitals train AI models together without sharing patient data

Open Source

The entire platform is released under the Apache 2.0 license — meaning anyone can use, modify, and share it for free. This is important because it means:

- Any hospital can run it
- Researchers can verify and improve the methods
- No expensive software licenses required

Glossary

Term	Definition
Alignment	Matching short DNA reads to the correct position in a reference genome
AlphaMissense	An AI tool that predicts whether a DNA variant is disease-causing
ARM64	A type of computer processor architecture (used in DGX Spark)

ClinVar	A public database of disease-related DNA variants
Chromosome	A package of DNA (humans have 23 pairs, 46 total)
Composite Score	A weighted combination of generation, docking, and QED scores
Cryo-EM	A technique for determining protein structures using frozen samples
DeepVariant	Google's AI for identifying DNA variants (>99% accuracy)
DGX Spark	NVIDIA's $3,999 desktop AI computer
DiffDock	NVIDIA's AI for predicting how molecules bind to proteins
DNA	Deoxyribonucleic acid — the molecule that stores genetic instructions
Docking	Predicting how a drug molecule fits into a protein
Druggable	A protein that can be targeted by a medicine
Embedding	A mathematical representation of text or data as a list of numbers
FASTQ	The file format for raw DNA sequencing data
FTD	Frontotemporal Dementia — a brain disease affecting personality
Gene	A section of DNA that codes for one protein
Genome	Your complete set of DNA (~3.1 billion letters)
GPU	Graphics Processing Unit — a chip for fast parallel computing
Lipinski	Rules predicting whether a molecule can be taken as a pill
Milvus	A vector database for fast similarity search
MolMIM	NVIDIA's AI for generating new molecule designs
Parabricks	NVIDIA's GPU-accelerated genomics software
Pathogenic	Disease-causing
QED	Quantitative Estimate of Drug-likeness (0-1 score)
RAG	Retrieval-Augmented Generation — grounding AI in real data
TPSA	Topological Polar Surface Area — predicts membrane permeability
Variant	A difference in your DNA compared to a reference genome
VCF	Variant Call Format — a file listing all detected DNA variants
VCP	Valosin-Containing Protein — the demo target gene
VEP	Variant Effect Predictor — classifies variant impacts
WGS	Whole-Genome Sequencing — reading the entire genome