HCLS AI Factory

From Patient DNA to Novel Drug Candidates in < 5 Hours

NVIDIA DGX Spark | Parabricks | BioNeMo | Milvus | Claude
02/2026 | Version 1.0 | Apache 2.0 License | Author: Adam Jones

What It Is

The HCLS AI Factory transforms patient DNA into ranked novel drug candidates in under 5 hours on a single NVIDIA DGX Spark ($3,999). Three GPU-accelerated stages run end-to-end with no manual intervention.

The Problem

- CPU-based genomics: 12-36 hours for a single 30× WGS sample
- Variant annotation fragmented across disconnected databases
- Variant-to-drug-lead gap: months of manual work
- No integrated platform connects genomics → reasoning → drug discovery

The Solution — Three Stages

Stage 1: GPU-Accelerated Genomics (120-240 min)

- NVIDIA Parabricks 4.6 — 10-20× faster than CPU
- BWA-MEM2 alignment: 20-45 min | DeepVariant: 10-35 min, >99% accuracy
- Input: ~200 GB FASTQ | Output: VCF (~11.7M variants)

Stage 2: RAG-Grounded Target Identification (Interactive)

- ClinVar (4.1M) + AlphaMissense (71M) + VEP annotation
- 3.5M variant embeddings in Milvus | BGE-small-en-v1.5 (384-dim)
- Claude RAG reasoning | 201 genes, 13 therapeutic areas, 171 druggable (85%)

Stage 3: AI-Driven Drug Discovery (8-16 min)

- BioNeMo MolMIM (generation) + DiffDock (docking) + RDKit (scoring)
- Composite: 30% generation + 40% docking + 30% QED
- Output: 100 ranked novel drug candidates + PDF report

Key Numbers

Metric	Value
Total Pipeline Time	< 5 hours
Input Data	~200 GB FASTQ (30× WGS)
Variants Called	~11.7 million
High-Quality Variants	~3.5 million (QUAL>30)
Knowledge Base	201 genes, 13 therapeutic areas
Druggable Targets	171 (85%)
Drug Candidates	100 (ranked by composite score)
Hardware Cost	$3,999 (DGX Spark)

VCP/FTD Demo Highlights

- Target: VCP gene — FTD, ALS, IBMPFD
- Variant: rs188935092 — ClinVar Pathogenic, AlphaMissense 0.87
- Seed: CB-5083 (Phase I VCP inhibitor)
- Top candidate: +39% composite improvement over seed
- Docking: -11.4 kcal/mol (vs. -8.1) | QED: 0.81 (vs. 0.62)
- All top 10 pass Lipinski's Rule of Five

Technology Stack

Layer	Technology
Hardware	NVIDIA DGX Spark (GB10, 128 GB unified, $3,999)
Genomics	Parabricks 4.6, DeepVariant (>99% accuracy)
Annotation	ClinVar, AlphaMissense, Ensembl VEP
Vector DB	Milvus 2.4, BGE-small-en-v1.5, IVF_FLAT
LLM	Anthropic Claude (RAG-grounded reasoning)
Drug Discovery	BioNeMo MolMIM, DiffDock, RDKit
Orchestration	Nextflow DSL2 (5 modes)
Monitoring	Grafana, Prometheus, DCGM Exporter
License	Apache 2.0 (fully open)

Deployment Roadmap

Phase	Hardware	Scale	Cost
1 — Proof Build	DGX Spark	1 patient, Docker Compose	$3,999
2 — Departmental	DGX B200	Multiple concurrent, K8s	$500K-$1M
3 — Enterprise	DGX SuperPOD	Thousands, FLARE federated	$7M-$60M+

Cross-Modal Integration

- Imaging → Genomics: Lung-RADS 4B+ triggers tumor profiling
- Genomics → Drug Discovery: Pathogenic variants trigger molecule generation
- NVIDIA FLARE: Federated learning (data stays local)

Competitive Differentiation

- Only platform: genomics-to-drug-candidates on a single desktop GPU
- End-to-end: No manual handoffs between stages
- < 5 hours total (vs. weeks/months traditional)
- $3,999 proof build (vs. $100K+ CPU infrastructure)
- Open-source: Apache 2.0, reproducible, auditable
- Scalable: Same pipelines from DGX Spark to SuperPOD

HCLS AI Factory — Apache 2.0 | Author: Adam Jones | February 2026