

Open-Source Project

White Paper

HCLS AI Factory on NVIDIA DGX Spark

From Patient DNA to Novel Drug Candidates in Under Five Hours — GPU-Accelerated Genomics,
RAG-Grounded Target Identification, and AI-Driven Drug Discovery

NVIDIA DGX Spark | Parabricks | BioNeMo

02/2026 | Version 1.0 | Apache 2.0 License

Author: Adam Jones

1. Executive Summary

The HCLS AI Factory follows a reusable pattern: identify a canonical artifact, build a persistent data model around it, and let agentic workflows operate continuously on that model. In genomics, the canonical artifact is the VCF — a structured record of every genomic variant identified in a patient’s DNA.

This white paper describes an end-to-end platform that processes raw DNA sequencing data through GPU-accelerated variant calling, RAG-grounded clinical reasoning, and AI-driven drug discovery — all on a single NVIDIA DGX Spark desktop workstation.

The platform transforms patient FASTQ files (~200 GB of raw sequencing data from a 30x whole-genome study) into 100 ranked novel drug candidates in under 5 hours. Three stages execute sequentially: NVIDIA Parabricks performs GPU-accelerated alignment and variant calling (120-240 min), producing ~11.7 million variants. A RAG pipeline annotates variants with ClinVar, AlphaMissense, and VEP, indexes 3.5 million high-quality variants in a Milvus vector database, and uses Anthropic Claude to identify druggable gene targets. Finally, BioNeMo NIM services (MolMIM and DiffDock) generate novel molecules, predict binding affinities, and rank candidates by a composite drug-likeness score.

The architecture is designed to run end-to-end on a \$3,999 DGX Spark for proof builds and scale to DGX SuperPOD for enterprise deployments. All components are open-source or NVIDIA-licensed, released under Apache 2.0 as part of the HCLS AI Factory.

2. The Precision Medicine Data Challenge

Precision medicine promises therapies tailored to an individual’s genetic profile. A single 30x whole-genome sequencing (WGS) run produces approximately 200 GB of raw data and 11.7 million genomic variants. The challenge is not generating this data — modern sequencers produce it reliably — but transforming it into actionable therapeutic hypotheses within a clinically relevant timeframe.

The Limits of Traditional Bioinformatics

Today’s genomic analysis pipelines assemble disconnected components: CPU-based alignment tools that take 12-24 hours, separate variant callers, annotation databases accessed through web APIs, and manual literature review for target identification.

This sequential, manual approach introduces three structural problems:

Compute bottleneck. CPU-based BWA-MEM alignment of a 30x WGS sample takes 12-24 hours on a 32-core server. DeepVariant on CPU adds another 8-12 hours. The genomics stage alone consumes 1-2 days of wall time.

Annotation fragmentation. Clinical variant databases (ClinVar), AI pathogenicity predictors (AlphaMissense), and functional annotation tools (VEP) exist as separate resources requiring bespoke ETL pipelines.

Target-to-drug gap. Even after identifying a pathogenic variant in a druggable gene, the path to a lead compound requires separate molecular modeling tools and medicinal chemistry expertise — typically a months-long process.

The GPU-Accelerated Opportunity

NVIDIA DGX Spark collapses the compute bottleneck. Its GB10 GPU accelerates BWA-MEM2 alignment from hours to 20-45 minutes and DeepVariant variant calling from hours to 10-35 minutes — a 10-20x speedup. More importantly, the same GPU that runs genomics can run vector similarity search (Milvus), molecular generation (MolMIM), and molecular docking (DiffDock).

3. Architecture Overview

Three-Stage Pipeline

| Stage | Technology | Duration | Input | Output |
|--------------------|---------------------------|-------------|-----------------|------------------------|
| 1 — Genomics | Parabricks 4.6 | 120-240 min | FASTQ (~200 GB) | VCF (~11.7M variants) |
| 2 — RAG/Chat | Milvus + BGE + Claude | Interactive | VCF | Target gene + evidence |
| 3 — Drug Discovery | MolMIM + DiffDock + RDKit | 8-16 min | Target gene | 100 ranked candidates |

Technology Stack

| Layer | Components |
|----------------|--|
| Compute | NVIDIA DGX Spark (GB10 GPU, 128 GB unified, 144 ARM64 cores) |
| Genomics | NVIDIA Parabricks 4.6.0-1, GRCh38 reference genome |
| Annotation | ClinVar (4.1M), AlphaMissense (71M), Ensembl VEP |
| Vector DB | Milvus 2.4, BGE-small-en-v1.5 (384-dim), IVF_FLAT |
| LLM | Anthropic Claude (claude-sonnet-4-20250514, temp=0.3) |
| Drug Discovery | BioNeMo MolMIM, BioNeMo DiffDock, RDKit |
| Orchestration | Nextflow DSL2, Docker Compose |
| Monitoring | Grafana, Prometheus, DCGM Exporter |

Service Architecture

The platform runs 14 services across 14 ports, organized by stage:

Orchestration: Landing page (8080) with 10-service health monitor

Stage 1: Genomics portal (5000)

Stage 2: Milvus (19530), Attu UI (8000), RAG API (5001), Chat (8501)

Stage 3: MolMIM NIM (8001), DiffDock NIM (8002), Discovery UI (8505), Portal (8510)

Monitoring: Grafana (3000), Prometheus (9099), Node Exporter (9100), DCGM (9400)

4. Stage 1 — GPU-Accelerated Genomics

NVIDIA Parabricks 4.6

Parabricks provides GPU-accelerated implementations of standard bioinformatics tools. On DGX Spark's GB10 GPU, it delivers 10-20× speedup over CPU implementations.

BWA-MEM2 Alignment (fq2bam)

Aligns paired-end reads against the GRCh38 reference genome. GPU-accelerated implementation achieves 70-90% GPU utilization, producing a sorted BAM file with index in 20-45 minutes.

Google DeepVariant

A CNN-based variant caller achieving >99% accuracy on the GIAB truth set. GPU-accelerated implementation achieves 80-95% GPU utilization, calling variants in 10-35 minutes. The deep learning approach outperforms traditional statistical callers (GATK HaplotypeCaller) on both SNPs and indels.

Input: HG002 Reference Standard

| Parameter | Value |
|-------------|--|
| Sample | HG002 (NA24385, GIAB reference standard) |
| Coverage | 30× whole-genome sequencing (WGS) |
| Read Length | 2×250 bp paired-end |
| File Size | ~200 GB (FASTQ pair) |
| Reference | GRCh38 (3.1 GB, pre-indexed) |

Output: Variant Call Format (VCF)

| Metric | Count |
|------------------------|----------|
| Total Variants | ~11.7M |
| High-Quality (QUAL>30) | ~3.5M |
| SNPs | ~4.2M |
| Indels | ~1.0M |
| Coding Region | ~35,000 |
| Multi-allelic Sites | ~150,000 |

5. Stage 2 — RAG-Grounded Target Identification

Variant Annotation

Stage 2 begins by annotating the 3.5 million high-quality variants with three complementary databases:

ClinVar (NCBI)

4.1 million clinical variant records mapping genomic positions to clinical significance classifications (Pathogenic, Likely pathogenic, VUS, Likely benign, Benign). Approximately 35,616 patient variants match ClinVar entries.

AlphaMissense (DeepMind)

71,697,560 AI-predicted pathogenicity scores for missense variants, derived from AlphaFold protein structure features. Thresholds: pathogenic >0.564, ambiguous 0.34-0.564, benign <0.34. Approximately 6,831 ClinVar-matched variants have AlphaMissense predictions.

Ensembl VEP

Functional consequence annotation mapping variants to genes, transcripts, and impact levels (HIGH, MODERATE, LOW, MODIFIER). Identifies missense variants, stop gains, frameshift variants, and splice site disruptions.

Vector Embedding and Indexing

Each annotated variant is transformed into a text summary and embedded using BGE-small-en-v1.5 (384 dimensions). The 3.5 million embeddings are indexed in Milvus 2.4 using IVF_FLAT (nlist=1024, COSINE metric) with 17 structured fields per record.

RAG-Grounded Reasoning with Claude

User queries are expanded using 10 therapeutic area keyword maps, embedded, and used for approximate nearest-neighbor search in Milvus (top_k=20). Retrieved variant contexts are assembled into a RAG prompt and processed by Anthropic Claude (claude-sonnet-4-20250514, temperature=0.3). Claude generates structured target hypotheses: gene name, confidence level, evidence chain, therapeutic area, and recommended action.

Knowledge Base: 201 Genes, 13 Therapeutic Areas

| Therapeutic Area | Genes | Examples |
|--------------------|-------|-------------------------------|
| Neurology | 36 | VCP, APP, PSEN1, MAPT, SOD1 |
| Oncology | 27 | EGFR, BRAF, KRAS, TP53, BRCA1 |
| Metabolic | 22 | GCK, PPARG, SLC2A2, PCSK9 |
| Infectious Disease | 21 | ACE2, CCR5, IFITM3, TLR4 |
| Respiratory | 13 | CFTR, SERPINA1, MUC5B |

| | | |
|-------------------------|----|-------------------------|
| Rare Disease | 12 | VCP, HTT, SMN1, DMD |
| Hematology | 12 | HBB, HBA1, F5, JAK2 |
| GI/Hepatology | 12 | HFE, ATP7B, NOD2 |
| Pharmacogenomics | 11 | CYP2D6, CYP2C19, CYP3A4 |
| Ophthalmology | 11 | RHO, RPE65, RS1, ABCA4 |
| Cardiovascular | 10 | LDLR, PCSK9, SCN5A |
| Immunology | 9 | HLA-B, TNF, IL6, JAK1 |
| Dermatology | 9 | FLG, MC1R, TYR, KRT14 |

Total: 201 genes, 171 druggable targets (85% druggability).

6. Stage 3 — AI-Driven Drug Discovery

10-Stage Drug Discovery Pipeline

| Stage | Process | Description |
|-------|---------------------|--|
| 1 | Initialize | Load target hypothesis, validate inputs |
| 2 | Normalize Target | Map gene → UniProt ID → PDB structures |
| 3 | Structure Discovery | Query RCSB PDB for structures |
| 4 | Structure Prep | Score by resolution, inhibitor, pockets |
| 5 | Molecule Generation | MolMIM generates novel SMILES from seed |
| 6 | Chemistry QC | RDKit validates chemical feasibility |
| 7 | Conformers | RDKit 3D conformer embedding (ETKDG) |
| 8 | Docking | DiffDock predicts binding poses and affinities |
| 9 | Ranking | 30% gen + 40% dock + 30% QED composite |
| 10 | Reporting | PDF report via ReportLab |

BioNeMo NIM Services

MolMIM (Port 8001)

A masked language model for molecular generation. Given a seed compound's SMILES string, it generates structurally novel analogs by masking and regenerating molecular tokens. Container:
nvcr.io/nvidia/clara/bionemo-molmim:1.0

DiffDock (Port 8002)

A score-based generative diffusion model for molecular docking. It predicts the 3D binding pose and affinity of a ligand in a protein binding site without requiring pre-defined binding pockets. Container: nvcr.io/nvidia/clara/difffdock:1.0

Drug-Likeness Scoring

Each candidate is evaluated against Lipinski's Rule of Five (MW≤500, LogP≤5, HBD≤5, HBA≤10), QED (>0.67 = drug-like), and TPSA (<140 Å² for oral bioavailability).

Composite Scoring Formula

The final ranking uses a weighted composite: 30% MolMIM generation confidence, 40% DiffDock binding affinity (normalized: max(0, min(1, (10 + dock_score) / 20))), and 30% QED score. This balances novelty, binding prediction, and drug-likeness.

7. VCP/FTD Demonstration

Target: Valosin-Containing Protein (VCP/p97)

The platform ships with a pre-configured demonstration targeting VCP — a AAA+ ATPase involved in the ubiquitin-proteasome pathway. Pathogenic VCP mutations cause Frontotemporal Dementia (FTD), ALS, and IBMPFD.

| Parameter | Value |
|----------------|--|
| Gene | VCP (UniProt P55072) |
| Variant | rs188935092 (chr9:35065263 G>A) |
| ClinVar | Pathogenic |
| AlphaMissense | 0.87 (pathogenic, >0.564 threshold) |
| Seed Compound | CB-5083 (Phase I clinical VCP inhibitor) |
| PDB Structures | 8O0I, 9DIL, 7K56, 5FTK |
| Binding Site | D2 ATPase domain (~450 Å ³ , druggability 0.92) |

Demo Results

The VCP/FTD demo produces 100 novel VCP inhibitor candidates. Typical results: 87 pass Lipinski's Rule of Five, 72 have QED >0.67, top 10 show docking scores from -8.2 to -11.4 kcal/mol, with composite scores ranging 0.68-0.89.

8. Cryo-EM Structure Evidence

Automated Structure Retrieval and Scoring

The drug discovery pipeline automatically queries RCSB PDB for protein structures and scores them by resolution, inhibitor presence (+3 bonus), druggable pocket count (+0.5 each), and experimental method (Cryo-EM +0.5).

| PDB | Resolution | Method | Key Feature |
|----------------------|------------|---------|---|
| 5FTK | 2.3 Å | X-ray | CB-5083 inhibitor-bound (highest score) |
| 7K56 | 2.5 Å | Cryo-EM | VCP complex |
| 800I | 2.9 Å | Cryo-EM | WT VCP hexamer |
| 9DIL | 3.2 Å | Cryo-EM | Mutant VCP |

The inhibitor-bound structure (5FTK) is preferred because it provides a pre-defined binding site and a reference ligand for molecular generation.

9. Orchestration and Monitoring

Nextflow DSL2 Orchestration

The pipeline is orchestrated by Nextflow DSL2, supporting five modes:

- full:** End-to-end (Stages 1→2→3)
- target:** From existing VCF (Stages 2→3)
- drug:** Known target to drug candidates (Stage 3 only)
- demo:** Pre-configured VCP/FTD demonstration
- genomics_only:** Variant calling only (Stage 1)

Six execution profiles (standard, docker, singularity, dgx_spark, slurm, test) adapt the pipeline to different infrastructure.

Monitoring Stack

Grafana dashboards (port 3000) visualize GPU utilization, memory pressure, pipeline progress, and service health. Prometheus (port 9099) collects metrics from DCGM Exporter (port 9400) for GPU telemetry and Node Exporter (port 9100) for system resources. The landing page (port 8080) provides a 10-service health grid with real-time status indicators.

10. Cross-Modal Integration

Imaging Intelligence Agent Integration

The HCLS AI Factory ecosystem includes an Imaging Intelligence Agent for CT, MRI, and X-ray analysis. Cross-modal triggers connect imaging findings to genomic analysis:

Imaging → Genomics: Lung-RADS 4B+ triggers FHIR ServiceRequest for tumor profiling

Genomics → Drug Discovery: Pathogenic variants trigger targeted molecule generation

Drug Discovery → Imaging: Candidates combined with imaging in clinical reports

NVIDIA FLARE for Federated Learning

Phase 3 deployments use NVIDIA FLARE for federated learning across institutions. Models train locally; only gradient updates are shared. Patient genomic data never leaves the originating institution.

11. Deployment Roadmap

Three-Phase Scaling

| Phase | Hardware | Orchestration | Scale |
|------------------|---------------------|----------------|---------------------------------|
| 1 — Proof Build | DGX Spark (\$3,999) | Docker Compose | Single patient, sequential |
| 2 — Departmental | 1-2× DGX B200 | Kubernetes | Multiple concurrent patients |
| 3 — Enterprise | DGX SuperPOD | K8s + FLARE | Thousands concurrent, federated |

Phase 1: DGX Spark Proof Build

A single DGX Spark runs the complete pipeline: GB10 GPU handles Parabricks, Milvus, MolMIM, and DiffDock sequentially. Docker Compose manages all 14 services. The 128 GB unified memory accommodates all stages. Total cost: \$3,999 hardware + API keys (Anthropic, NGC).

Phase 2: Departmental Scale

DGX B200 systems (8x B200 GPUs, 1-2 TB HBM3e) enable parallel processing of multiple patients, GPU-dedicated Milvus instances, and multiple BioNeMo NIM replicas. Kubernetes replaces Docker Compose.

Phase 3: Enterprise / Multi-Site

DGX SuperPOD with InfiniBand fabric, NVIDIA FLARE for federated learning, and institutional-scale variant databases. Thousands of patients processed concurrently with cross-institutional collaboration while maintaining data sovereignty.

12. Conclusion

The HCLS AI Factory demonstrates that the full precision medicine pipeline — from raw DNA to novel drug candidates — can run on a single desktop workstation. GPU acceleration collapses genomics from days to hours. Vector databases and LLM reasoning transform annotation from manual curation to interactive exploration. Generative chemistry and molecular docking automate the target-to-lead transition that traditionally takes months.

The three-stage architecture (Genomics → RAG/Chat → Drug Discovery) provides a reproducible, auditable, and scalable framework. The same Nextflow pipelines that run on a \$3,999 DGX Spark scale to DGX SuperPOD for enterprise deployments. All components are open-source or NVIDIA-licensed under Apache 2.0.

This is precision medicine as a continuous, computable workflow — not a disconnected collection of tools, but an integrated factory that transforms patient data into therapeutic hypotheses in a single session.

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