Demo Guide

HCLS AI Factory

VCP/FTD Demo Walkthrough

Step-by-step guide for demonstrating the complete pipeline: from patient DNA to 100 ranked novel drug candidates.

NVIDIA DGX Spark | Parabricks | BioNeMo | Milvus | Claude

02/2026 | Version 1.0 | Apache 2.0 License

Author: Adam Jones

Demo Overview

Parameter	Value
Demo Duration	15-20 minutes (live walkthrough)
Pipeline Mode	demo (pre-configured VCP/FTD)
Hardware	NVIDIA DGX Spark (GB10, 128 GB unified)
Target Gene	VCP — Frontotemporal Dementia
End Result	100 ranked novel drug candidates

What the Audience Will See

1. Raw DNA data (FASTQ) entering the platform
2. GPU-accelerated variant calling (Parabricks) completing in minutes
3. 11.7 million variants annotated and indexed in a vector database
4. Interactive Claude RAG chat identifying VCP as a drug target
5. BioNeMo generating 100 novel VCP inhibitors
6. Ranked candidates with docking scores and drug-likeness profiles
7. PDF report generated automatically

Pre-Demo Setup

Step 1: Verify Hardware

bash

```
nvidia-smi              # Verify GB10 GPU
uname -m                # Expected: aarch64
```

Step 2: Set Environment Variables

bash

```
cp .env.example .env
# Edit .env:
# ANTHROPIC_API_KEY=sk-ant-...   (for Claude RAG)
# NGC_API_KEY=nvapi-...          (for BioNeMo NIMs)
```

Step 3: Start All Services

bash

```
./start-services.sh
```

Services start in dependency order: infrastructure → Stage 1 → Stage 2 → Stage 3 → landing page.

Step 4: Verify All Services Healthy

Open http://localhost:8080 — all 10 services should show green status.

Service	Port	Expected
Parabricks Portal	5000	GREEN
Milvus Vector DB	19530	GREEN
RAG API	5001	GREEN
Streamlit Chat	8501	GREEN
MolMIM NIM	8001	GREEN
DiffDock NIM	8002	GREEN
Discovery UI	8505	GREEN
Grafana	3000	GREEN
Prometheus	9099	GREEN
DCGM Exporter	9400	GREEN

Demo Script

Opening (1 minute)

Show: Landing page at http://localhost:8080

- "This is the HCLS AI Factory — patient DNA to drug candidates in < 5 hours."
- "Everything runs on this single DGX Spark — a $3,999 desktop workstation."
- "Three stages: genomics, target identification, and drug discovery."

Stage 1: Genomics (3-4 minutes)

Launch: python run_pipeline.py --mode demo

Show: Genomics portal at http://localhost:5000

Talking Points

- ~200 GB FASTQ input from Illumina sequencer (30× WGS)
- Parabricks BWA-MEM2: 20-45 min on GPU (vs. 12-24 hours on CPU)
- DeepVariant: 10-35 min, >99% accuracy (CNN-based)
- Output: VCF with ~11.7 million variants
- Show GPU utilization spiking on Grafana (http://localhost:3000)

Stage 2: RAG/Chat (5-6 minutes)

Annotation Pipeline

- ClinVar: 4.1M clinical variants → 35,616 patient matches
- AlphaMissense: 71M AI predictions → 6,831 scored variants
- VEP: functional consequences (HIGH/MODERATE/LOW/MODIFIER)

Vector Database

Show: Attu UI at http://localhost:8000

- 3.5M variant embeddings (BGE-small-en-v1.5, 384-dim)
- IVF_FLAT index, COSINE metric, 17 fields per record
- "This enables natural language queries over genomic data."

Interactive Chat

Show: Streamlit chat at http://localhost:8501

Type: "What are the most promising drug targets for neurodegenerative disease?"

Claude identifies VCP with full evidence chain:

- rs188935092 — ClinVar Pathogenic, AlphaMissense 0.87
- HIGH impact missense variant in D2 ATPase domain
- Known target — CB-5083 (Phase I), druggability 0.92

- 201 genes, 13 therapeutic areas, 171 druggable (85%)

Stage 3: Drug Discovery (4-5 minutes)

Structure Retrieval

- VCP → UniProt P55072 → RCSB PDB query
- 4 structures: 8OOI, 9DIL, 7K56, 5FTK
- 5FTK selected — 2.3 Å X-ray with CB-5083 inhibitor bound

Molecule Generation

Show: Discovery UI at http://localhost:8505

- MolMIM generates 100 novel analogs from CB-5083 seed
- 98 pass RDKit chemical validity checks

Docking & Ranking

- DiffDock docks each candidate against VCP D2 domain
- 34 candidates score below -8.0 kcal/mol (excellent)
- Composite: 30% generation + 40% docking + 30% QED

Key Demo Table

Metric	CB-5083 (Seed)	Top Candidate	Improvement
Dock Score	-8.1 kcal/mol	-11.4 kcal/mol	+41% binding
QED	0.62	0.81	+31% drug-likeness
MW	487.2 Da	423.5 Da	-13% (better)
Composite	0.64	0.89	+39% overall

PDF report generated automatically via ReportLab with full provenance.

Closing (2 minutes)

- "< 5 hours, $3,999 desktop → raw DNA to 100 ranked drug candidates"
- "Collapses weeks/months to a single session"
- "Same Nextflow pipelines scale to DGX SuperPOD"
- "Open-source — Apache 2.0"

Phase	Hardware	Scale
Phase 1	DGX Spark ($3,999)	Proof build — what you just saw
Phase 2	DGX B200	Department — multiple concurrent patients
Phase 3	DGX SuperPOD	Enterprise — thousands, federated

Troubleshooting

Service Not Starting

bash

```
docker compose ps                    # Check status
docker compose logs <service-name>   # Check logs
docker compose restart <service>     # Restart
```

BioNeMo NIM Not Ready

NIMs require NGC API key and may take 2-5 minutes to initialize.

bash

```
curl http://localhost:8001/v1/health/ready   # MolMIM
curl http://localhost:8002/v1/health/ready   # DiffDock
```

GPU Out of Memory

DeepVariant peaks at ~60 GB. Monitor with:

bash

```
watch -n 1 nvidia-smi
```

Quick Reference

Action	Command / URL
Start services	./start-services.sh
Launch demo	python run_pipeline.py --mode demo
Landing page	http://localhost:8080
Genomics portal	http://localhost:5000
Chat interface	http://localhost:8501
Milvus UI	http://localhost:8000
Discovery UI	http://localhost:8505
Grafana	http://localhost:3000

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