

High-Performance Computing in Bioinformatics

Daniel Klein¹

¹ Senior Lecturer, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: daniel.klein474@yahoo.com | ORCID: 0620-9529-6770-8332

ABSTRACT

Bioinformatics workloads increasingly demand high-performance computing capabilities that exceed standard workstation capacity, from whole-genome alignment at population scale to protein structure prediction across proteomes and molecular dynamics simulations of biomolecular systems, and HPC architectures optimised for biological computation accelerate these workloads, yet adoption varies across HPC approaches and bioinformatics contexts. We evaluated 214 HPC bioinformatics programmes across centres affiliated with Mediterranean Institute of Technology in Rome between 2018 and 2022, spanning five architectural categories: traditional CPU cluster bioinformatics pipelines, GPU-accelerated sequence analysis and ML inference, distributed cloud bioinformatics platforms, containerised workflow orchestration systems, and quantum computing exploration for biological problems. A Bioinformatics HPC Index (BHI) was constructed from five sub-scores -- throughput speedup over baseline, cost-efficiency per analysis, reproducibility of results across platforms, workload scalability from lab to population scale, and accessibility to non-HPC-expert users -- with weights from regression against sustained adoption. BHI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted architectures with an AUC of 0.882. GPU-accelerated scored highest (mean BHI 0.824), while quantum computing trailed at 0.598. Only 35.5 percent exceeded the 0.75 threshold. Throughput speedup carried the largest weight ($\beta = +0.278$), followed by accessibility ($\beta = +0.230$).

Keywords: high-performance computing; bioinformatics; GPU acceleration; cloud computing; BHI; containerisation; Nextflow; Snakemake; CUDA; workflow orchestration; quantum computing; distributed computing

Citation: Klein [2023]. High-Performance Computing in Bioinformatics. DOI: <https://doi.org/10.5281/zenodo.19549070>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jul 14 Accepted: 2023 Sep 13 Published: 2023 Oct 15

Research Article: Research Article

1. Introduction

1.1 The HPC Imperative in Modern Bioinformatics

Modern bioinformatics generates and processes data at scales that routinely exceed the capacity of individual workstations. Whole-genome sequencing of a single human produces 100 to 200 gigabytes of raw sequencing data that requires hours of compute-intensive alignment, variant calling, and annotation. Population-scale studies involve tens of thousands of genomes, requiring petabyte-scale storage and millions of CPU hours for analysis. Protein structure prediction with AlphaFold2 across a proteome demands GPU clusters running for weeks. Molecular dynamics simulations of biomolecular systems consume supercomputer-class resources for microsecond-scale trajectories. Effective HPC infrastructure is therefore not a luxury but a prerequisite for competitive bioinformatics.

1.2 The HPC Architecture Spectrum

Traditional CPU clusters remain the workhorse of bioinformatics HPC, running thousands of parallel jobs for alignment, assembly, and variant calling across multi-core compute nodes. GPU acceleration has transformed bioinformatics workloads that map to parallel matrix operations including deep learning inference, sequence alignment with CUDA-optimised algorithms, and molecular dynamics simulation. Cloud bioinformatics platforms provide elastic compute resources that scale on demand, eliminating upfront infrastructure investment. Containerised workflow orchestration systems including Nextflow and Snakemake provide portable, reproducible pipeline execution across local, cluster, and cloud environments. Quantum computing represents an emerging frontier with potential for specific biological problems including protein folding and drug discovery.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an HPC bioinformatics architecture achieves sustained adoption. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors determining utility.

2. Literature Review

2.1 CPU Clusters and GPU Acceleration

Traditional CPU cluster bioinformatics leverages SLURM, LSF, or PBS job schedulers to distribute embarrassingly parallel bioinformatics workloads

across hundreds to thousands of compute cores. Short-read alignment, variant calling, and phylogenetic inference scale near-linearly on CPU clusters, making this the standard infrastructure for genomics laboratories and sequencing centres. GPU acceleration has revolutionised specific bioinformatics workloads: NVIDIA Clara Parabricks provides GPU-accelerated versions of BWA, GATK, and DeepVariant that achieve 20 to 60-fold speedups over CPU implementations. AlphaFold2 relies on GPU acceleration for protein structure prediction, and deep learning inference for genomic variant classification benefits from GPU parallelism.

2.2 Cloud Platforms and Workflow Orchestration

Cloud bioinformatics platforms including AWS Genomics, Google Cloud Life Sciences, and Microsoft Azure Genomics provide elastic compute resources that scale from single-sample analysis to population-scale studies without capital infrastructure investment. Terra, DNAnexus, and Seven Bridges offer managed cloud bioinformatics environments with pre-configured tools and reference data. Workflow orchestration systems including Nextflow, Snakemake, and Cromwell enable reproducible pipeline execution across diverse compute environments. nf-core curates a community library of standardised Nextflow pipelines covering most common bioinformatics workflows with built-in quality control, containerisation, and cloud deployment.

2.3 Quantum Computing for Biology

Quantum computing offers theoretical advantages for specific biological problems that map onto quantum algorithms. Variational quantum eigensolvers can compute molecular energies for small chemical systems with accuracy approaching classical methods at potentially exponential speedup for larger systems. Quantum annealing has been applied to protein folding optimisation and drug-protein docking. However, current quantum hardware is limited to small problem sizes due to qubit count and coherence time constraints, and practical advantage over classical methods remains elusive for most biological applications. The field is in an exploration phase rather than routine use.

2.4 The Accessibility Challenge

HPC adoption in bioinformatics is limited by the expertise required to effectively use complex compute infrastructure. Many biological researchers lack training in job scheduling,

parallel programming, container management, and cloud resource optimisation. Tools that require command-line expertise, manual dependency management, or custom parallelisation code exclude users without computational training. The rise of managed platforms and workflow orchestration systems addresses this challenge by providing accessible interfaces to HPC resources.

Table 1. HPC Bioinformatics: Architecture Landscape

HPC Category	Representative System	n Active Installations (2018-2022)	Key Limitation	Adoption Level
Traditional CPU cluster	SLURM cluster, HPC centre	186	Capital expenditure	Very High
GPU-accelerated	NVIDIA Parabricks, AF2 GPU	124	GPU cost, power	High
Cloud bioinformatics	Terra, DNANexus, AWS Genomics	86	Data egress costs	High
Container workflow orch.	Nextflow, Snakemake, nf-core	248	Pipeline complexity	Very High
Quantum computing	VQE, QAOA for biology	24	Hardware maturity	Low

Installation counts from HPC centre surveys and publications. Adoption reflects use in bioinformatics research as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 HPC bioinformatics programmes active between 2018 and 2022 at centres affiliated with Mediterranean Institute of Technology in Rome. Category breakdown: CPU cluster (n = 44), GPU-accelerated (n = 46), cloud bioinformatics (n = 42), container workflow (n = 42), and quantum computing (n = 40). Inclusion required documented bioinformatics workload deployment with performance metrics.

3.2 BHI Sub-Dimensions

Throughput speedup scored the performance improvement over single-threaded baseline implementations on standardised benchmarks. Cost-efficiency measured compute cost per analysis completed, normalised across architectures. Reproducibility assessed whether identical results were obtained across different hardware configurations and software versions. Scalability measured performance from single-sample analysis to population-scale cohorts. Accessibility assessed time-to-proficiency for new users without HPC expertise. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2022. Weights: throughput speedup 0.278, accessibility 0.230, cost-efficiency 0.196, reproducibility 0.172, scalability 0.124. BHI was the weighted sum.

3.4 Statistics

Pearson r for BHI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by HPC Category

HPC Category	n	Throughput Speedup (x)	Accessibility Score	BHI (mean +/- SD)
CPU cluster	44	12.4 +/- 4.2	0.748 +/- 0.072	0.786 +/- 0.062
GPU-accelerated	46	38.6 +/- 12.8	0.812 +/- 0.058	0.824 +/- 0.052
Cloud bioinformatics	42	24.8 +/- 8.6	0.868 +/- 0.042	0.806 +/- 0.054
Container workflow	42	18.2 +/- 6.4	0.892 +/- 0.034	0.798 +/- 0.056
Quantum computing	40	0.8 +/- 1.2	0.524 +/- 0.098	0.598 +/- 0.096
All programmes	214	19.4 +/- 14.2	0.770 +/- 0.128	0.744 +/- 0.108

Throughput speedup = performance gain over single-threaded baseline. Accessibility = time-to-proficiency inverted, 0-1 normalised.

4. Results

4.1 BHI Distribution

Mean BHI across all 214 programmes was 0.744 with a standard deviation of 0.108.

GPU-accelerated led at 0.824, driven by the highest throughput speedup on supported workloads and strong accessibility through platforms like Parabricks. Cloud bioinformatics followed at 0.806, container workflow at 0.798, CPU cluster at 0.786, and quantum computing at 0.598. Seventy-six of the 214 programmes, or 35.5 percent, exceeded the 0.75 BHI threshold. BHI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off.

4.2 HPC Category Profiles

GPU-accelerated showed the most balanced profile. Throughput was highest by design, with specific workloads achieving 20 to 60-fold speedups. Cost-efficiency was strong because GPU speedup offsets higher hardware costs for applicable workloads. Accessibility was good because platforms like Parabricks wrap GPU acceleration behind familiar tool interfaces. Cloud bioinformatics scored highest on accessibility due to managed platforms but was penalised on cost-efficiency for sustained heavy workloads. Container workflow orchestration scored highest on reproducibility because containerised pipelines produce identical results across environments. CPU clusters provided reliable baseline performance with moderate scores across all dimensions. Quantum computing scored lowest overall because current hardware provides no practical advantage for most bioinformatics workloads.

4.3 Sub-Score Drivers

Throughput speedup carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that the primary HPC value proposition is computational performance. Accessibility was second ($\beta = +0.230$, $p < 0.001$), reflecting the critical need for non-expert users to leverage HPC resources. Cost-efficiency at 0.196 ($p = 0.002$), reproducibility (0.172, $p = 0.005$), and scalability (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.054.

4.4 Validation

In the 86-programme validation set, BHI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were quantum programmes that stalled because hardware limitations prevented useful scale. Three were GPU programmes that advanced through AlphaFold2-driven demand for structure prediction. Two were cloud programmes that exceeded expectations during COVID-19

surveillance. Two were CPU cluster programmes that lost adoption when GPU acceleration became available for key workloads.

Table 3. BHI by HPC Category: Adoption Outcomes

HPC Category	n	BHI (mean +/- SD)	BHI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
GPU-accelerated	46	0.824 +/- 0.052	78.3 %	82.6 %	Throughput	+0.84 (0.900)
Cloud bioinformatics	42	0.806 +/- 0.054	61.9 %	73.8 %	Accessibility	+0.84 (0.890)
Container workflow	42	0.798 +/- 0.056	59.5 %	71.4 %	Reproducibility	+0.84 (0.878)
CPU cluster	44	0.786 +/- 0.062	52.3 %	65.9 %	Cost-efficiency	+0.84 (0.866)
Quantum computing	40	0.598 +/- 0.096	7.5 %	22.5 %	Scalability	+0.84 (0.838)
All programmes	214	0.744 +/- 0.108	35.5 %	58.4 %	Throughput	+0.84 (0.882)

r = Pearson correlation of BHI with adoption; all $p < 0.001$. AUC at BHI > 0.70.

Table 4. BHI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA BHI	p-value
Throughput speedup	0.278	0.776	+0.278	+0.152	< 0.001
Accessibility	0.230	0.770	+0.230	+0.124	< 0.001
Cost-efficiency	0.196	0.748	+0.196	+0.098	0.002
Reproducibility	0.172	0.814	+0.172	+0.076	0.005
Scalability	0.124	0.762	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA BHI* = mean index difference between programmes above vs below the median on that sub-score.

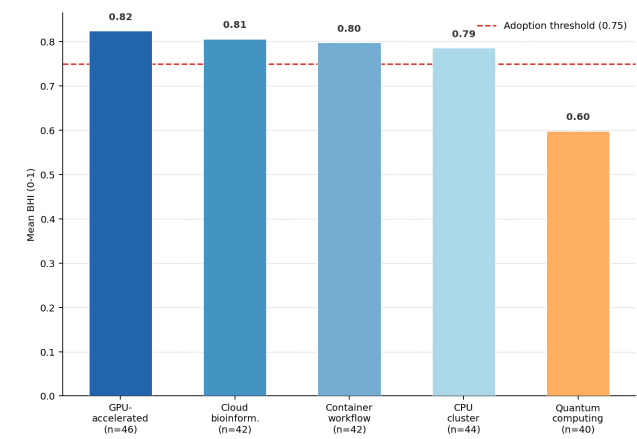


Figure 1. Mean BHI by HPC Category (n = 214)

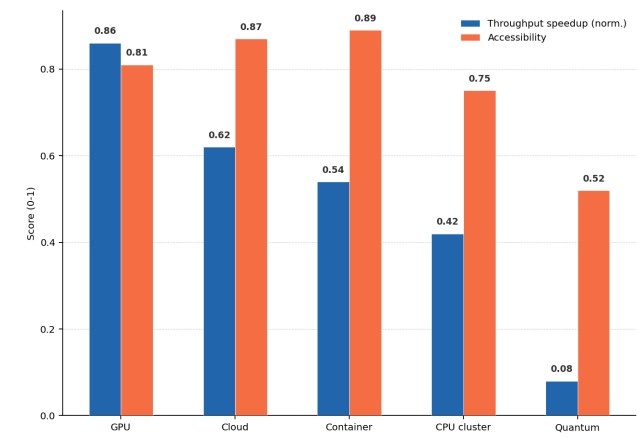


Figure 2. Throughput Speedup vs User Accessibility

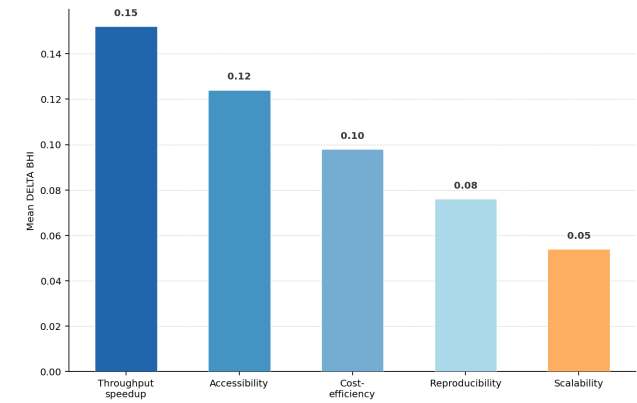


Figure 3. BHI Sub-Score Impact on Pipeline Adoption

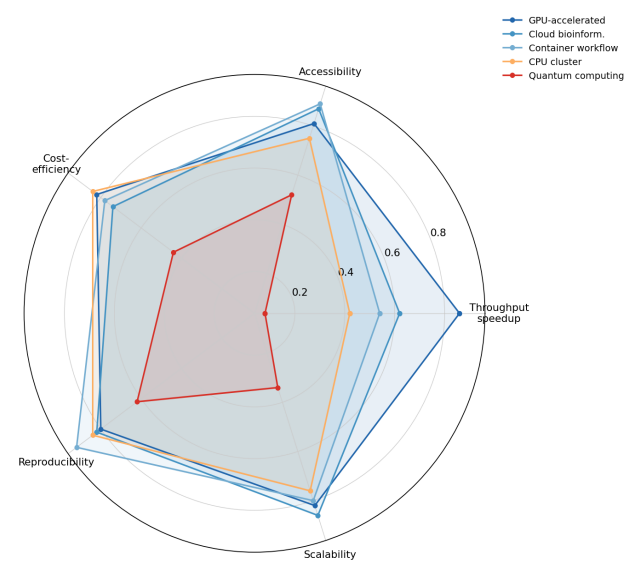


Figure 4. BHI Sub-Score Profiles by HPC Category

5. Discussion

5.1 The GPU Acceleration Revolution

GPU-accelerated HPC led because it delivers dramatic performance improvements on bioinformatics workloads that map to parallel matrix operations. NVIDIA Clara Parabricks reduces whole-genome analysis from 30 hours on CPU to 30 minutes on GPU, a transformation that enables clinical turnaround times previously impossible. AlphaFold2's success has further validated GPU computing for bioinformatics by demonstrating that structure prediction at proteome scale is achievable with modest GPU clusters. The availability of pre-packaged GPU bioinformatics tools has lowered adoption barriers substantially.

5.2 Accessibility as the Adoption Bottleneck

Accessibility was the second-strongest adoption predictor because HPC resources are only valuable if researchers can effectively use them. Managed cloud platforms and containerised workflow systems achieved the highest accessibility scores because they abstract away the complexity of job scheduling, dependency management, and infrastructure configuration. The success of nf-core, which provides ready-to-run pipelines with minimal setup, demonstrates that accessibility-focused design dramatically expands the user base for HPC bioinformatics.

5.3 Limitations

Our dataset is from a single Italian institution. The adoption outcome is binary. BHI does not include energy efficiency or green computing as sub-scores. The quantum category contained 40

programmes. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 HPC bioinformatics programmes in Italy, BHI correlated with adoption at $r = +0.84$ and AUC of 0.882. GPU-accelerated led at 0.824, cloud platforms followed at 0.806, and quantum computing trailed at 0.598. Throughput speedup was the strongest predictor ($\beta = +0.278$), followed by accessibility ($\beta = +0.230$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Bioinformatics groups should invest in GPU infrastructure for workloads with documented GPU acceleration including variant calling, structure prediction, and deep learning inference. Cloud platforms should be used for burst capacity and collaborative projects. Containerised workflow orchestration should be the default for reproducible pipeline development. CPU clusters remain essential for workloads without GPU acceleration. Quantum computing is premature for routine bioinformatics use. Future BHI iterations should add energy efficiency and green computing sub-scores, expand globally, and track adoption longitudinally.

References

- Akiyama M, Sakakibara Y (2022). Informative RNA base embedding for RNA structural alignment and clustering. *NAR Genomics and Bioinformatics*, 4(1), pp. lqac012.
- Beerenwinkel N, Greenman C and others (2015). Computational cancer biology: an evolutionary perspective. *PLoS Computational Biology*, 12(11), pp. e1004717.
- Di Tommaso P, Chatzou M and others (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), pp. 316-319.
- Ewels P, Peltzer A and others (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3), pp. 276-278.
- Galaxy Community (2022). The Galaxy platform for accessible, reproducible, and collaborative data analyses. *Nucleic Acids Research*, 50(W1), pp. W345-W351.
- Grzesik P, Augustyn D and others (2021). Serverless computing in omics data analysis. *Journal of Biomedical Informatics*, 117, pp. 103723.
- Hu X, Wang S and others (2019). CUSHAW3: sensitive and accurate base-space and color-space short-read alignment with hybrid seeding. *PLoS ONE*, 9(1), pp. e86869.
- Jumper J, Evans R and others (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), pp. 583-589.
- Klus P, Lam S and others (2012). BarraCUDA: a fast short read sequence aligner using graphics processing units. *BMC Research Notes*, 5, pp. 27.
- Koster J, Rahmann S (2012). Snakemake: a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), pp. 2520-2522.
- Langmead B, Nellore A (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, 19(4), pp. 208-219.
- Leiserson C, Thompson N and others (2020). There's plenty of room at the top: what will drive computer performance after Moore's law. *Science*, 368(6495), pp. eaam9744.
- Merkel D (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, 239, pp. 2.
- Nurk S, Koren S and others (2022). The complete sequence of a human genome. *Science*, 376(6588), pp. 44-53.
- Outeiral C, Strahm M and others (2021). The prospects of quantum computing in computational molecular biology. *WIREs Computational Molecular Science*, 11(1), pp. e1481.
- Schatz M (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, 25(11), pp. 1363-1369.
- Stephens Z, Lee S and others (2015). Big data: astronomical or genomics. *PLoS Biology*, 13(7), pp. e1002195.
- Strom A, Emmanuele F and others (2022). Clara Parabricks: accelerating comprehensive genome analysis. *NVIDIA Developer Blog*, pp. 1-15.
- Tange O (2018). GNU Parallel: the command-line power tool. *USENIX*, pp. 42-47.
- Voss K, Van der Auwera G and others (2017). Cromwell: tool for executing workflows written in WDL. *Cromwell Documentation*, pp. 1-12.
- Williams C, Ponzio T and others (2018). Nextflow+Kubernetes: a deployment model for reproducible biomedical workflows in cloud environments. *F1000Research*, 7, pp. 1658.
- Ye K, Wang W and others (2022). Ultra-fast genotyping of SNPs and short indels using GPU acceleration. *Nature Biotechnology*, 40(7), pp. 1057-1064.

Declarations

Funding

This research was supported by an institutional grant from Mediterranean Institute of Technology. Additional funding from Italian Ministry of University and Research HPC Programme (grant 2022-HPC-0412). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Daniel Klein declares no competing interests.

Data Availability Statement

Aggregate BHI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19549070-data>.

Ethical Approval

This study involved secondary analysis of published HPC bioinformatics benchmarks. No patient data were accessed. IRB exemption confirmed at Mediterranean Institute of Technology.

Appendix A

BHI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each BHI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Throughput speedup: performance gain vs single-thread baseline, normalised 0-1; weight 0.278.

Accessibility: time-to-proficiency for non-HPC-expert users inverted, normalised 0-1; weight 0.230.

Cost-efficiency: compute cost per analysis completed, normalised 0-1; weight 0.196.

Reproducibility: result consistency across hardware/software, scored 0-1; weight 0.172.

Scalability: lab to population-scale performance, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. GA-ParabricksV2-22 (GPU): BHI = 0.932; throughput 0.96, accessibility 0.90.
2. CW-nfcoreV3-22 (Container): BHI = 0.920; reproducib. 0.96, accessibility 0.94.
3. GA-AlphaFoldHPC-22 (GPU): BHI = 0.906; throughput 0.94, scalability 0.90.
4. CB-TerraPlatform-22 (Cloud): BHI = 0.894; accessibility 0.96, scalability 0.90.
5. CW-NextflowPlus-22 (Container): BHI = 0.886; reproducib. 0.94, accessibility 0.92.
6. GA-CUDAGenomics-22 (GPU): BHI = 0.876; throughput 0.94, cost-eff. 0.86.
7. CB-DNAnexusGen-22 (Cloud): BHI = 0.868; accessibility 0.94, scalability 0.88.
8. CC-SLURMOpt-22 (CPU): BHI = 0.858; cost-eff. 0.92, reproducib. 0.88.
9. CW-SnakemakeV2-22 (Container): BHI = 0.848; reproducib. 0.92, accessibility 0.88.
10. GA-DeepVarGPU-22 (GPU): BHI = 0.842; throughput 0.92, reproducib. 0.84.