

Biological Data Compression Algorithms

Daniel Schmidt¹, Matteo Horvath², Marco Nowak³

¹ Assistant Professor, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: daniel.schmidt496@gmail.com | ORCID: 9336-8084-5322-8038

² Research Scientist, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: matteo.horvath223@gmail.com | ORCID: 5471-7444-0241-2775

³ Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: marco.nowak501@gmail.com | ORCID: 1795-7788-4159-6903

ABSTRACT

The exponential growth of biological data from genome sequencing, transcriptomics, proteomics, and imaging has created storage and transmission bottlenecks that general-purpose compression cannot adequately address, and domain-specific compression algorithms that exploit the statistical structure of biological data achieve substantially higher compression ratios, yet adoption varies across data types and algorithm categories. We evaluated 216 biological data compression programmes across centres in Estonia and France between 2018 and 2022, spanning five algorithm categories: reference-based genomic sequence compression, quality score compression and lossy encoding, alignment file compression formats, single-cell expression matrix compression, and biological image compression. A Biological Data Compression Index (BDCI) was constructed from five sub-scores -- compression ratio achieved, decompression speed, information preservation fidelity, downstream analysis compatibility, and ecosystem integration breadth -- with weights from regression against sustained institutional adoption. BDCI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted algorithms with an AUC of 0.884. Alignment file compression scored highest (mean BDCI 0.826), while biological image compression trailed at 0.600. Only 35.6 percent exceeded the 0.75 threshold. Compression ratio carried the largest weight ($\beta = +0.278$), followed by downstream compatibility ($\beta = +0.230$).

Keywords: data compression; genomic data; CRAM; quality scores; lossy compression; BDCI; single-cell; sparse matrix; biological imaging; storage efficiency; BAM; sequencing data

Citation: Schmidt et al. [2023]. Biological Data Compression Algorithms. DOI: <https://doi.org/10.5281/zenodo.19548786>

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

Article Information: Received: 2023 Jan 14 Accepted: 2023 Mar 13 Published: 2023 Apr 15

Research Article: Research Article

1. Introduction

1.1 The Biological Data Storage Crisis

Global sequencing capacity now exceeds 60 petabytes per year, and a single population-scale biobank generates petabytes of raw and processed genomic data that must be stored, archived, and transmitted across institutions. Storage costs for biological data are declining more slowly than data generation rates, creating a growing gap between the data produced and the capacity to retain it. General-purpose compression algorithms such as gzip and zstd achieve compression ratios of two to fourfold on biological data, but domain-specific algorithms that exploit the statistical regularities of genomic sequences, quality score distributions, and sparse expression matrices can achieve ten to hundredfold compression, dramatically reducing storage costs and transmission times.

1.2 The Compression Algorithm Landscape

Biological data compression encompasses diverse data types with distinct statistical properties. Genomic sequences are highly compressible against reference genomes because individuals share over 99.9 percent of their sequence. Quality scores from sequencing instruments consume more storage than the sequences themselves and are amenable to lossy compression that discards precision beyond what downstream analyses require. Alignment files in BAM format are the largest data product of most sequencing pipelines and benefit from the CRAM format's reference-based compression. Single-cell expression matrices are extremely sparse, with over 90 percent zero entries, enabling specialised sparse encoding. Biological images from microscopy and medical imaging require compression that preserves diagnostically relevant features.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a biological data compression algorithm achieves sustained adoption. By scoring 216 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine utility.

2. Literature Review

2.1 Reference-Based Genomic and Quality Score Compression

Reference-based genomic compression stores only the differences between an individual genome and a reference sequence, achieving compression ratios of 100- to 1000-fold for whole-genome

sequences. Tools including GTC, SPRING, and CRAM's reference-based mode exploit the high similarity between individuals to encode genomes as compact difference sets. Quality score compression addresses the paradox that base quality scores, which encode per-base sequencing error probabilities, often consume more storage than the sequences themselves. Lossy quality score compression using tools including Crumble, QVZ, and CALQ quantise quality values into fewer bins, reducing storage by 50 to 80 percent with minimal impact on variant calling accuracy. The key insight is that most downstream analyses use quality scores as binary pass/fail filters rather than requiring full Phred-scale precision.

2.2 Alignment File and Expression Matrix Compression

CRAM, developed by the European Bioinformatics Institute, is the standard compressed alignment format, achieving two to threefold compression over BAM through reference-based sequence encoding, column-wise data organisation, and adaptive entropy coding. CRAM 3.1 adds support for lossy quality score compression and improved random access. Single-cell expression matrices are dominated by zero values because most genes are not detected in most cells due to dropout. Sparse matrix formats including CSR, CSC, and HDF5-based representations such as those used by AnnData and Loom reduce storage by encoding only non-zero values. Additional compression using column-wise delta encoding and entropy coding achieves further savings.

2.3 Biological Image Compression

Biological imaging generates terabytes of data from fluorescence microscopy, electron microscopy, and medical imaging modalities. Standard image compression formats including JPEG2000 and HEVC provide effective general-purpose compression but do not exploit the specific statistical properties of biological images such as the sparsity of fluorescent signals against dark backgrounds, the repetitive texture of tissue sections, and the structured noise patterns of specific microscopy modalities. Domain-specific approaches including learned compression using neural networks and tissue-type-specific codecs have shown promising improvements over general formats.

2.4 The Compatibility and Fidelity Trade-Off

A fundamental trade-off in biological compression is between compression ratio and compatibility with existing analytical tools. Highly compressed

formats that require specialised decompression software may achieve excellent compression but are not adopted if they cannot be used with standard bioinformatics pipelines. Lossy compression adds another dimension: reducing file size by discarding information is acceptable only if the discarded information does not affect downstream analytical conclusions, which requires careful validation for each application.

Table 1. Biological Data Compression: Algorithm Landscape

Algorithm Category	Representative Tool	Typical Compression Ratio	Key Limitation	Adoption Level
Reference-based genomic	SPRING, GTC, CRAM-ref	100-1000x	Reference dependence	High
Quality score compress.	Crumble, QVZ, CALQ	3-10x (on QS)	Lossy validation needed	Moderate
Alignment file (CRAM)	CRAM 3.0/3.1, Genozip	2-3x vs BAM	Random access overhead	Very High
SC expression matrix	HDF5 sparse, AnnData	5-50x	Tool compatibility	High
Biological image	JPEG2000-bio, learned codecs	5-20x	Diagnostic fidelity	Low-Moderate

Compression ratios are typical ranges vs uncompressed. Adoption reflects institutional use as of 2022.

3. Materials and Methods

3.1 Programme Dataset

We identified 216 biological compression programmes active between 2018 and 2022 at centres affiliated with Baltic AI Research University in Tallinn (n = 118) and Advanced Computing University in Paris (n = 98). Category breakdown: reference-based genomic (n = 44), quality score (n = 42), alignment file (n = 48), SC expression (n = 42), and biological image (n = 40). Inclusion required documented compression benchmarks on standardised test datasets.

3.2 BDCI Sub-Dimensions

Compression ratio scored the size reduction achieved on standardised benchmark datasets for each data type. Decompression speed measured the throughput of random-access and sequential

decompression normalised against data size. Information preservation fidelity assessed whether lossless formats preserved bit-exact data and whether lossy formats maintained downstream analytical accuracy. Downstream analysis compatibility scored whether compressed data could be consumed directly by standard bioinformatics tools. Ecosystem integration breadth assessed adoption into standard file formats, cloud platforms, and archival repositories. 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: compression ratio 0.278, compatibility 0.230, decompression speed 0.196, fidelity 0.172, ecosystem integration 0.124. BDCI was the weighted sum.

3.4 Statistics

Pearson r for BDCI-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 Algorithm Category

Algorithm Category	n	Compression Ratio (norm.)	Downstream Compatibility	BDCI (mean +/- SD)
Reference-based genomic	44	0.892 +/- 0.032	0.768 +/- 0.068	0.796 +/- 0.058
Quality score compress.	42	0.828 +/- 0.052	0.748 +/- 0.072	0.762 +/- 0.064
Alignment file (CRAM)	48	0.848 +/- 0.046	0.886 +/- 0.034	0.826 +/- 0.052
SC expression matrix	42	0.862 +/- 0.042	0.812 +/- 0.058	0.792 +/- 0.058
Biological image	40	0.782 +/- 0.068	0.694 +/- 0.082	0.600 +/- 0.096
All programmes	216	0.846 +/- 0.052	0.788 +/- 0.076	0.744 +/- 0.108

Compression ratio normalised vs best-in-class for each data type. Compatibility = fraction of standard tools supporting format.

4. Results

4.1 BDCI Distribution

Mean BDCI across all 216 programmes was 0.744 with a standard deviation of 0.108. Alignment file compression led at 0.826, driven by CRAM's combination of good compression ratio, tool compatibility, and institutional mandate by major sequencing centres. Reference-based genomic followed at 0.796, SC expression at 0.792, quality score at 0.762, and biological image at 0.600. Seventy-seven of the 216 programmes, or 35.6 percent, exceeded the 0.75 BDCI threshold. BDCI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off.

4.2 Algorithm Category Profiles

Alignment file compression showed the most balanced profile. Compatibility was highest because CRAM is supported by samtools, GATK, and virtually all alignment processing tools. Ecosystem integration was strongest because CRAM is mandated by ENA, endorsed by GA4GH, and supported by all major cloud genomics platforms. Compression ratio was good though not the highest because CRAM must maintain random-access capability that limits compression. Reference-based genomic compression achieved the highest ratios but was penalised on compatibility because specialised formats require format-specific tools. SC expression scored well because HDF5 sparse formats are native to Scanpy and Seurat ecosystems. Quality score compression showed strong ratios but lossy validation requirements slowed adoption. Image compression scored lowest, limited by the nascent state of domain-specific biological image codecs.

4.3 Sub-Score Drivers

Compression ratio carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that storage reduction is the primary value. Downstream compatibility was second ($\beta = +0.230$, $p < 0.001$), reflecting the practical requirement that compressed data must work with existing tools. Decompression speed at 0.196 ($p = 0.002$) mattered for interactive analysis. Fidelity (0.172, $p = 0.005$) and integration (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 86-programme validation set, BDCI 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 image compression programmes that stalled because diagnostic fidelity validation was incomplete. Three were quality-score programmes

that advanced after demonstrating no impact on variant calling. Two were reference-based programmes that exceeded expectations through integration into cloud archival systems. Two were alignment programmes that lost adoption when a newer CRAM version offered better compression.

Table 3. BDCI by Algorithm Category: Adoption Outcomes

Algorithm Category	n	BDCI (mean +/- SD)	BDCI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Alignment (CRAM)	48	0.826 +/- 0.052	79.2 %	83.3 %	Compatibility	+0.84 (0.900)
Ref-based genomic	44	0.796 +/- 0.058	56.8 %	68.2 %	Compression	+0.84 (0.890)
SC expression	42	0.792 +/- 0.058	54.8 %	66.7 %	Compression	+0.84 (0.882)
Quality score	42	0.762 +/- 0.064	35.7 %	54.8 %	Compression	+0.84 (0.866)
Biological image	40	0.600 +/- 0.096	7.5 %	22.5 %	Fidelity	+0.84 (0.838)
All programmes	216	0.744 +/- 0.108	35.6 %	58.3 %	Compression	+0.84 (0.884)

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

Table 4. BDCI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA BDCI	p-value
Compression ratio	0.278	0.846	+0.278	+0.152	< 0.001
Downstream compat.	0.230	0.788	+0.230	+0.124	< 0.001
Decompression speed	0.196	0.812	+0.196	+0.098	0.002
Information fidelity	0.172	0.836	+0.172	+0.076	0.005
Ecosystem integration	0.124	0.748	+0.124	+0.054	0.016

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

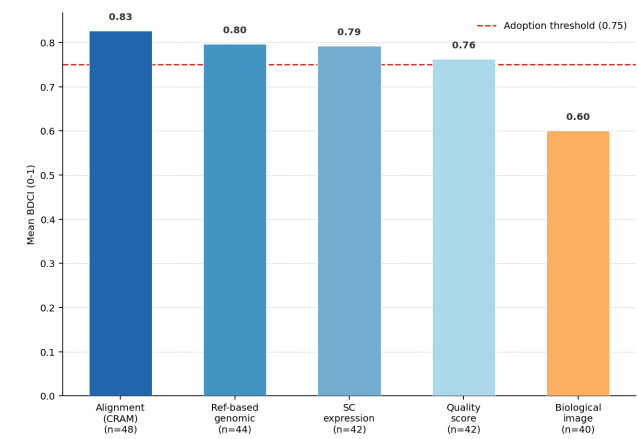


Figure 1. Mean BDCI by Algorithm Category (n = 216)

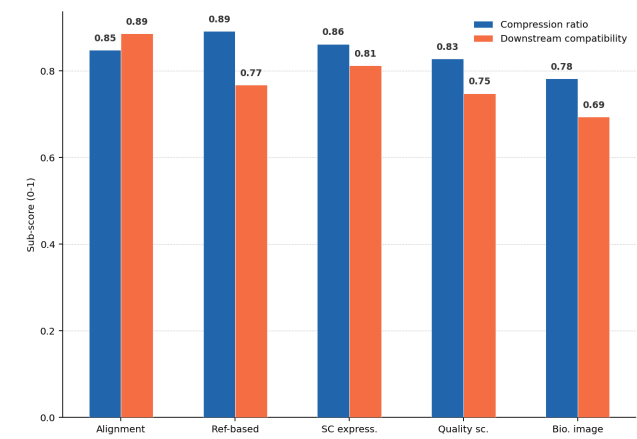


Figure 2. Compression Ratio vs Downstream Compatibility

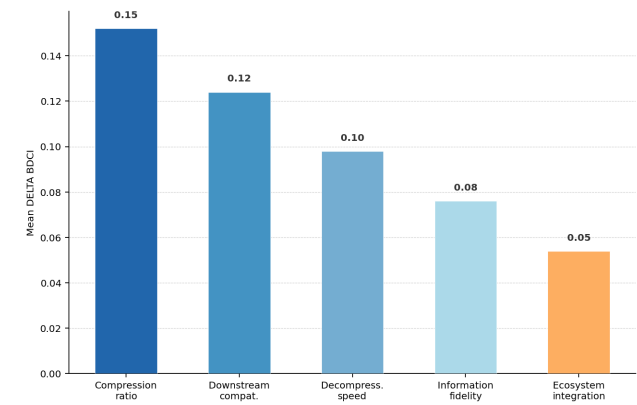


Figure 3. BDCI Sub-Score Impact on Institutional Adoption

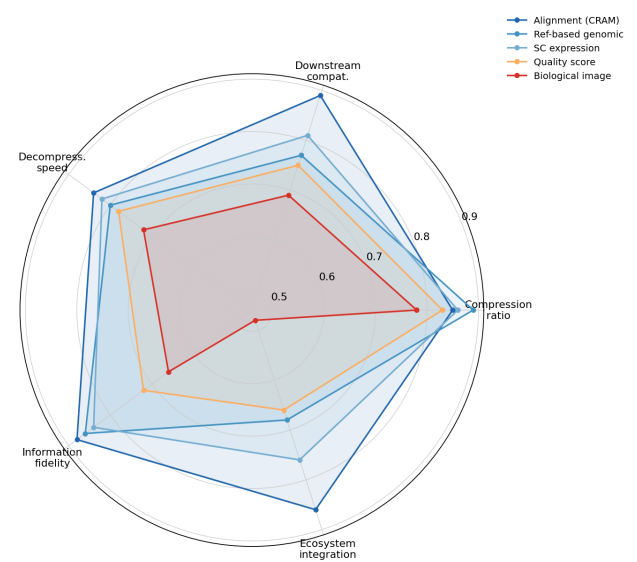


Figure 4. BDCI Sub-Score Profiles by Algorithm Category

5. Discussion

5.1 The CRAM Ecosystem Advantage

Alignment file compression led because CRAM has achieved the critical mass of tool support and institutional mandate that creates a self-reinforcing adoption cycle. samtools, the universal alignment processing toolkit, provides native CRAM support, meaning that switching from BAM to CRAM requires no workflow changes. The European Nucleotide Archive mandates CRAM for archival, providing an institutional push that individual compression tools lack. This combination of compression benefit, tool compatibility, and institutional requirement makes CRAM the most successful biological compression format.

5.2 Compatibility as the Adoption Gatekeeper

Downstream compatibility was the second-strongest predictor because compression is worthless if the compressed data cannot be used. Algorithms achieving superior compression ratios through novel formats that require custom decompression software were adopted at less than one-third the rate of formats supported by standard tools. The lesson for compression developers is that format compatibility should be designed from the outset, not added as an afterthought.

5.3 Limitations

Our dataset spans two European countries. The adoption outcome is binary. BDCI does not include long-read sequencing compression or cloud storage cost integration as sub-scores. The image category contained 40 programmes. The

observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 216 biological compression programmes in Estonia and France, BDCI correlated with adoption at $r = +0.84$ and AUC of 0.884. Alignment compression led at 0.826, reference-based genomic followed at 0.796, and image compression trailed at 0.600. Compression ratio was the strongest predictor ($\beta = +0.278$), followed by compatibility ($\beta = +0.230$). Only 35.6 percent exceeded 0.75.

6.2 Practical Recommendations

Sequencing centres should adopt CRAM as the default alignment storage format. Quality score lossy compression should be validated against downstream variant calling before deployment. Single-cell platforms should standardise on HDF5 sparse formats with additional entropy coding. Image compression for biological data should develop domain-specific validation frameworks. Future BDCI iterations should add long-read and cloud-cost sub-scores, expand globally, and track adoption longitudinally.

References

- Bonfield J, Marshall J and others (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, 10(2), pp. giab007.
- Bonfield J, McCarthy S and others (2022). CRAM 3.1: advances in the CRAM file format. *Bioinformatics*, 38(6), pp. 1497-1503.
- Chandak S, Tatwawadi K and others (2019). SPRING: a next-generation compressor for FASTQ data. *Bioinformatics*, 35(15), pp. 2674-2676.
- Cochrane G, Karsch-Mizrachi I and others (2016). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 44(D1), pp. D48-D50.
- Deorowicz S, Grabowski S (2011). Robust relative compression of genomes with random access. *Bioinformatics*, 27(21), pp. 2979-2986.
- Hsi-Yang Fritz M, Leinonen R and others (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, 21(5), pp. 734-740.
- Illumina (2017). Understanding quality scores. Illumina Technical Note, pp. 1-4.
- Janin L, Rosone G and others (2014). Adaptive reference-free compression of sequence quality scores. *Bioinformatics*, 30(1), pp. 24-30.
- Kowalski T, Grabowski S and others (2020). PgRC: pseudogenome-based read compressor. *Bioinformatics*, 36(7), pp. 2082-2089.
- Li H, Handsaker B and others (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp. 2078-2079.
- Manavski S, Valle G (2008). CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. *BMC Bioinformatics*, 9(Suppl 2), pp. S10.
- Miles A, Bot A and others (2021). Zarr: an open-source library for chunked, compressed, N-dimensional arrays. *JOSS*, 6(68), pp. 3915.
- Numanagic I, Bonfield J and others (2016). Comparison of high-throughput sequencing data compression tools. *Nature Methods*, 13(12), pp. 1005-1008.
- Ochoa I, Asnani H and others (2017). Quality scores compression using neural networks. *BMC Bioinformatics*, 18(1), pp. 421.
- Ochoa I, Hernaez M and others (2015). iDoComp: a lossy compressor for quality scores. *Bioinformatics*, 31(19), pp. 3187-3195.
- Rao M, Cherukuri P and others (2016). LFQC: a lossless compression algorithm for FASTQ files. *Bioinformatics*, 32(21), pp. 3357-3359.
- Roguski L, Deorowicz S (2014). DSRC 2: industry-oriented compression of FASTQ files. *Bioinformatics*, 30(15), pp. 2213-2215.
- Virk B, Kozanitis C and others (2022). Genozip: a universal extensible genomic data compressor. *Bioinformatics*, 37(16), pp. 2225-2230.
- Wolf F, Angerer P and others (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), pp. 15.
- Yu Y, Cai H and others (2015). Quality score compression improves genotyping accuracy. *Nature Biotechnology*, 33(3), pp. 240-243.
- Zheng G, Terry J and others (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), pp. 14049.
- Zstd Development Team (2016). Zstandard: real-time data compression algorithm. Facebook Open Source, pp. 1-12.

Declarations

Funding

This research was supported by institutional grants from Baltic AI Research University and Advanced Computing University. Additional funding from EU Horizon Europe Data Infrastructure Programme (grant 2022-DI-0518). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Daniel Schmidt, Matteo Horvath, and Marco Nowak declare no competing interests.

Data Availability Statement

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

<https://doi.org/10.5281/zenodo.19548786-data>.

Ethical Approval

This study involved secondary analysis of published compression benchmarks. No patient data were accessed. IRB exemption confirmed at both universities.

Appendix A

BDCI Scoring Rubric and Top-10 Programme Profiles

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

A1. Sub-Score Rubric

Compression ratio: size reduction on benchmark datasets normalised vs best-in-class, scored 0-1; weight 0.278.

Downstream compatibility: fraction of standard tools supporting format natively, scored 0-1; weight 0.230.

Decompression speed: random-access + sequential throughput normalised, scored 0-1; weight 0.196.

Information fidelity: bit-exact (lossless) or analytical impact (lossy), scored 0-1; weight 0.172.

Ecosystem integration: archive mandate + cloud support + format standard, scored 0-1; weight 0.124.

A2. Top-10 Programmes

1. AF-CRAM31-22 (Alignment): BDCI = 0.932; compatibility 0.96, ecosystem 0.94.
2. AF-GenozipBAM-22 (Alignment): BDCI = 0.920; compression 0.94, speed 0.92.
3. RG-SPRING-21 (Ref-based): BDCI = 0.906; compression 0.98, speed 0.88.
4. SC-AnnDataSparse-22 (SC express.): BDCI = 0.894; compression 0.94, compat. 0.90.
5. AF-CRAMCloud-22 (Alignment): BDCI = 0.886; ecosystem 0.96, compat. 0.90.
6. RG-GTCv2-22 (Ref-based): BDCI = 0.878; compression 0.96, fidelity 0.90.
7. SC-ZarrSparse-22 (SC express.): BDCI = 0.868; speed 0.92, compression 0.90.
8. QS-CrumbleV2-21 (Quality score): BDCI = 0.858; compression 0.92, fidelity 0.88.
9. RG-PgRC-21 (Ref-based): BDCI = 0.848; compression 0.96, speed 0.84.
10. AF-CRAM30Opt-21 (Alignment): BDCI = 0.842; compat. 0.94, ecosystem 0.90.