
BioAtlas: Computational Clustering of Multi-Dimensional Complexity in Bioart

Joonhyung Bae

Graduate School of Culture Technology
KAIST (Korea Advanced Institute of Science and Technology)
Daejeon, South Korea
jh.bae@kaist.ac.kr

Abstract

1 Bioart’s hybrid nature—spanning art, science, technology, ethics, and poli-
2 tics—defies traditional single-axis categorization. I present BioAtlas, analyzing
3 81 bioart works across thirteen curated dimensions using novel axis-aware repre-
4 sentations that preserve semantic distinctions while enabling cross-dimensional
5 comparison. Our codebook-based approach groups related concepts into uni-
6 fied clusters, addressing polysemy in cultural terminology. Comprehensive
7 evaluation of up to 800 representation–space–algorithm combinations identi-
8 fies Agglomerative clustering at $k=15$ on 4D UMAP as optimal (silhouette
9 0.664 ± 0.008 , trustworthiness/continuity 0.805/0.812). The approach reveals
10 four organizational patterns: artist-specific methodological cohesion, technique-
11 based segmentation, temporal artistic evolution, and trans-temporal conceptual
12 affinities. By separating analytical optimization from public communication, I
13 provide rigorous analysis and accessible exploration through an interactive web
14 interface (<https://www.bioartlas.com>) with the dataset publicly available
15 (<https://github.com/joonhyungbae/BioAtlas>).

16 1 Introduction

17 Bioart amalgamates biological materials, processes, and thoughts into artistic expression at the
18 confluence of art, science, and technology [1, 2]. Bioartists, such as Eduardo Kac with his transgenic
19 organisms and Stelarc with his prosthetic modifications, engage with living systems to investigate
20 themes of identity, ethics, embodiment, and the borders between the natural and artificial [3, 4].

21 Nevertheless, current taxonomic methodologies have encountered difficulties in encapsulating the
22 intrinsic intricacy of bioart. Single-axis or medium-/technology-centric frameworks (e.g., Kac’s
23 transgenic/telepresence research [1, 3], SymbioticA’s ethics of care [2], Hauser’s life media [5],
24 ZKM’s *BioMedia* [6]) overlook the multidimensionality of practice. Actor-network theory and
25 companion-species approaches [7, 8] advocate for the distinct consideration of relational, ethical,
26 aesthetic, and epistemic dimensions. Medium-focused perspectives minimize ethical considerations
27 and social context, while ethics-focused methods neglect aesthetic and epistemic contributions. The
28 hybrid character of bioart—encompassing art, science, technology, ethics, and politics—generates
29 categorical ambiguity, since works serve simultaneously as aesthetic objects, scientific instruments,
30 ethical provocations, and political statements, resisting solitary definition.

31 Cultural phenomena frequently necessitate the integration of diverse analytical dimensions; axis-
32 aware representations and domain-informed codebooks can enhance interpretability [9–11]. Sentence
33 embeddings capture nuanced semantics [12]. Rather than multi-kernel fusion, this study system-
34 atically scans *representation–space–algorithm* combinations (e.g., TF-IDF of codebook counts,

35 quantized per-axis embeddings; RAW/SVD/UMAP spaces; k-means/agglo/density methods) and
36 selects an interpretable, exhaustive partition under atlas-specific constraints.

37 To address these challenges, I present a computational framework for analyzing 81 bioartworks
38 and report the best internal-validity configuration from a representation–space–algorithm sweep:
39 *Agglomerative (average linkage)* at $k=15$ on a 4D UMAP projection with optimized hyperparameters
40 (10 nearest neighbors, minimum distance 0.01, cosine metric). For communication, I additionally
41 provide a coarser labeling in the viewer without treating it as the analytical optimum.

42 This approach addresses bioart’s analytical challenges through: (1) codebook-based semantic fu-
43 sion preserving axis-specific semantics, (2) data-driven clustering with transparent reporting, and
44 (3) separation between analytical optimization and communication visualization. This work com-
45 bines computational rigor with artistic insight, bridging the researcher-practitioner divide that often
46 limits cross-disciplinary understanding. I explicitly position the atlas as a tool for human interpre-
47 tation—algorithms operate as instruments rather than arbiters—so interpretability and uncertainty
48 reporting remain first-class design goals.

49 2 Dataset

50 The corpus was built to select target artists and representative works by synthesizing
51 *award–institution–platform* indicators conferring international visibility from major bioart awards
52 and exhibitions, including *Prix Ars Electronica*, *Bio Art & Design Award*, ZKM’s *BioMedia*, MIT
53 List’s *Symbionts*, STARTS Prize, and ISEA archives [6, 13–17]. Artist and artwork selection used
54 multi-layered criteria: art-historical significance (e.g., Davis, Kac), technological innovation, field
55 contribution, and conceptual clarity [1, 3].

56 The **13 analytic axes** (Table 2) were derived via literature review [1, 2, 5–8] and inductive corpus
57 analysis. Coverage and non-redundancy were emphasized during axis development; observed
58 pairwise correlations were low-to-moderate, consistent with complementary dimensions. All source
59 materials are public, non-personal, and attributed; only metadata and derived artifacts are released.
60 Policies follow FAIR/CARE guidelines; takedown requests are honored. Axis definitions and labeling
61 protocols were predetermined using a codebook with comprehensive annotation criteria, shaped by
62 the author’s dual role as a professional artist and an AI researcher. Although the existing annotations
63 are conducted by a single annotator, the systematic codebook facilitates future validation by many
64 annotators and the evaluation of inter-rater reliability.

65 This study provides a systematically annotated bioart dataset, featuring systematic evaluations
66 across 13 analytical parameters. Although constrained in scope, it provides a preliminary basis for
67 computational methodologies in bioart analysis and associated cultural fields. Table 1 offers a detailed
68 summary of the dataset’s composition, illustrating the significant scale and diversity of the corpus,
69 which contains 770 distinct keywords across 81 artworks.

Table 1: Dataset Summary

Dataset		Keywords	
Total works	81	Unique keywords	770
Total artists	33	Mean / work	28.2
Temporal coverage	1976–2022	Assignments	2285
		Analytic axes	13

70 The dataset includes 33 artists and collectives (Table 3), balancing diversity with focused analysis of
71 key practitioners.

72 3 Method

73 3.1 Two-Stage Representation Process

74 The multidimensional nature of bioart—encompassing aesthetic, ethical, material, and intellectual
75 aspects—necessitates transcending singular embedding techniques that may obscure diverse semantic
76 components. My solution tackles these problems via a methodical two-stage process that maintains
77 axis-specific semantics while facilitating cross-dimensional comparison. The method establishes

Table 2: Keyword statistics for the 13 analytic axes, with the top three keywords per axis.

Axis	Unique	Average	Top (up to three)
Materiality	98	2.58	Plant; Composite materials; Data
Methodology	88	2.40	Cell culture; Data visualization; Biosensing
Actor Relations & Configurations	62	2.17	Artist-led; Autonomous bio processes; Interspecies co-creation
Ethical Approach	53	2.16	Reflective; Relational ethics; Symbiotic
Aesthetic Strategy	76	2.57	Conceptual; Uncanny; Biological morphology
Epistemic Function	54	2.21	Social criticism; Knowledge production; Future proposal
Philosophical Stance	47	2.23	New materialism; Posthumanism; Relational ontology
Social Context	44	2.17	Gallery; Laboratory
Audience Engagement	53	2.00	Observational; Interpretive engagement; Contemplative
Temporal Scale	44	1.53	Continuous; Short-term exhibition; Evolutionary
Spatial Scale	56	1.48	Installation; Human body size; Individual unit
Power and Capital Critique	61	1.79	Institutional criticism; Biopolitics; Biocapital
Documentation & Representation	85	2.91	Photographic records; Material residue; Data viz

Table 3: List of 33 artists and collectives in the dataset, with number of works shown in parentheses.

Artist Names (Number of Works)
Joe Davis (4), Stelarc (4), George Gessert (2), Eduardo Kac (3), Oron Catts & Ionat Zurr (3), Wim Delvoye (2), Art Orienté Objet (1), HeHe (1), Zbigniew Oksjuta (1), Paul Vanouse (3), Anna Dumitriu (2), Center for Genomic Gastronomy (2), Charlotte Jarvis (2), Heather Dewey-Hagborg (2), Jalila Essaïdi (2), Marta de Menezes (2), Špela Petrič (2), Agi Haines (1), Maja Smrekar (1), Ani Liu (5), Alexandra Daisy Ginsberg (4), Anicka Yi (4), Candice Lin (3), Claire Pentecost (3), Dasha Tsapenko (3), Jenna Sutela (3), Jes Fan (3), Cecilia Jonsson (2), Gilberto Esparza (2), Pamela Rosenkranz (2), Xandra van der Eijk (2), Amy Karle (1), Michael Sedbon (1)

78 durable semantic anchors by constructing a codebook that organizes related concepts into cohesive
 79 clusters, maintaining consistency among artworks.

80 **Stage 1: Per-Axis Embedding Aggregation.** For each artwork’s 13 axes, assigned keywords
 81 are embedded using dual large language models (BGE-large-en-v1.5 and GTE-large-en-v1.5) with
 82 concatenated representations yielding 2048-dimensional vectors. Keywords within each axis are
 83 then averaged to create a single axis-level embedding. For example, if an artwork’s *Materiality*
 84 axis contains keywords $\{plant, data, composite-materials\}$, the three 2048-dimensional embeddings
 85 are averaged into one 2048-dimensional vector representing that axis. Each artwork thus yields 13
 86 axis-specific embeddings of 2048 dimensions each.

87 **Stage 2: Word Codebook and Axis Features.** All unique keywords across the entire dataset are
 88 clustered into a *word codebook* using K-means. Prior to clustering, token embeddings undergo PCA
 89 (retaining $\geq 95\%$ cumulative variance) with whitening to stabilize K-means. K_c is *automatically*
 90 *selected* by scanning candidate values and maximizing a silhouette-based objective regularized by
 91 penalties for empty/singleton/imbalanced clusters. Each artwork is then represented via per-axis
 92 codebook activations: counts/one-hot and their TF-IDF (BM25 optional); I also compute *quantized*
 93 *per-axis embeddings* as count-weighted averages of codebook centroids.

94 The codebook approach addresses polysemy by grouping semantically related terms (e.g., $\{biofab-$
 95 *rication, tissue-engineering, living-materials\}) into unified concept clusters, enabling more robust
 96 similarity computation than raw keyword matching.*

97 **Final Representation.** I generate TF-IDF weighted cluster counts (L2 normalized), quantized
 98 embeddings, binary indicators, and SVD variants. TF-IDF weighted counts achieved superior
 99 performance, balancing interpretability with clustering quality.

100 3.2 Systematic Sweep Configuration

101 I evaluate multiple representation types across systematic algorithm-space combinations to identify
 102 optimal clustering configurations. My approach explores 8 feature representations (TF-IDF variants,
 103 quantized embeddings, SVD-reduced features) across 31+ projection spaces (RAW, SVD dimensions
 104 50/100/150, UMAP 4D/8D/16D with 3x3 hyperparameter grids) using 4 clustering algorithms (K-
 105 means, Agglomerative, DBSCAN, OPTICS). With $K \in [2, 15]$ for partitional methods, this yields
 106 800+ evaluated configurations within computational constraints.

107 The building of a cultural atlas necessitates a balance between statistical coherence and interpretive
 108 utility, highlighting a fundamental contradiction between optimization objectives. This challenge
 109 is especially pronounced in identifying optimal cluster numbers: although statistical criteria (Gap
 110 statistic, Silhouette) may indicate $K \leq 27$ for our data size [18], domain-specific constraints in

111 cultural categorization advocate for cognitively manageable partitions. The range $K \in [2, 15]$
 112 reflects this compromise, supported by cluster validation literature [19, 20] and cultural analysis
 113 approaches, where excessive granularity obstructs understanding [21]. Conventional clustering
 114 optimization prioritizes internal cohesion metrics (Silhouette, within-cluster sum of squares), while
 115 atlas applications necessitate thorough categorization, cognitive accessibility, and interpretive clarity,
 116 often conflicting with statistical optimization.

117 This tension is evident in three dimensions: (1) **Completeness vs. Purity**—density-based methods at-
 118 tain superior silhouette scores by categorizing boundary cases as noise, whereas cultural atlases neces-
 119 sitate extensive landscape representation; (2) **Statistical Precision vs. Cognitive Load**—hierarchical
 120 methods enhance separation via meticulous partitioning that surpasses human categorical process-
 121 ing capabilities; (3) **Algorithmic Sophistication vs. Interpretive Transparency**—sophisticated
 122 techniques may identify statistical patterns while concealing semantic distinctions.

123 **Multi-Metric Evaluation Framework.** I resolve these conflicts by doing a thorough assessment
 124 utilizing many validation criteria instead of depending on singular measures. Evaluation incorporates
 125 the silhouette score (main ranking criterion), projection quality safeguards (trustworthiness/continuity
 126 ≥ 0.80), and noise ratio analysis for density-based methodologies. Stability validation employs
 127 bootstrap resampling (5 iterations) to calculate the Adjusted Rand Index (ARI) and Normalized
 128 Mutual Information (NMI) throughout several executions. This comprehensive strategy emphasizes
 129 thorough partitions rather than statistical purity, all while upholding stringent quality standards.

Table 4: Complete sweep configuration.

Component	Specification
Embedding model	Dual models: BGE-large-en-v1.5 + GTE-large-en-v1.5 (concatenated, 2048-dim); codebook $K_c = 47$ (auto-selected)
Feature types	TF-IDF weighted codebook counts (L2 normalized); quantized embeddings; SVD variants
Projection spaces	RAW, SVD, UMAP (4/8/16-D, varying hyperparameters)
Clustering algorithms	K-means, Agglomerative (multiple linkages), DBSCAN, OPTICS
Selection criteria	Silhouette-based selection with trustworthiness/continuity guardrails and stability validation; complete assignment methods prioritized for final selection

130 3.3 Codebook Diagnostics

131 I automatically scan codebook size K_c across candidates $\{32, 48, \dots, 1024\} \cup \{\sqrt{n}, 1.5\sqrt{n}, 2\sqrt{n}\}$
 132 using adjusted silhouette score: $S_{adj} = S - 0.6 \cdot r_{singleton} - 0.8 \cdot r_{empty} - 0.2 \cdot \text{Gini}$, where
 133 $r_{singleton}$ and r_{empty} denote singleton and empty cluster ratios. Four clustering algorithms (K-means
 134 variants, Agglomerative with Ward linkage) are evaluated on PCA-whitened dual model embeddings
 135 (2048-dim reduced to 95% variance, typically 1800 dimensions). The selected configuration
 136 (MiniBatch K-means with batch_size=1024, $K_c = 47$) demonstrates balanced cluster utilization
 137 with low singleton occurrence and semantic coherence across axes. Representative semantic clusters
 138 include $\{\text{biofabrication}, \text{tissue-engineering}, \text{living-materials}\}$ and $\{\text{posthumanism}, \text{new-materialism},$
 139 $\text{relational-ontology}\}$.

140 3.4 Technical Implementation

141 **Reproducibility Framework.** All stochastic components use fixed random states, ensuring deter-
 142 ministic results. The four-stage pipeline processes: (1) dual token embeddings (BGE-large-en-v1.5
 143 + GTE-large-en-v1.5 concatenated, 2048-dim), (2) PCA-preprocessed K-means codebook con-
 144 struction ($K_c = 47$, auto-selected), (3) TF-IDF weighted representations with L2 normalization,
 145 (4) systematic evaluation across 800+ algorithm-space combinations within computational lim-
 146 its. Key configuration parameters: PHASEC_MAX_TRIALS=800, PHASEC_K_LIST=[2,3,...,15],
 147 PHASEC_BOOTSTRAP_REPS=5, random seeds (Python=42, NumPy=42, sklearn models=42).
 148 UMAP projections use cosine metric for TF-IDF features with hyperparameter grids: neighbors
 149 $\{10, 15, 30\}$, distances $\{0.01, 0.1, 0.5\}$, dimensions $\{4, 8, 16\}$.

150 4 Results

151 I systematically assess $K \in [2, 15]$ in accordance with known recommendations for cluster number
152 selection [19–21]. The top limit $K \leq 15$ satisfies three criteria: (1) **Statistical validity** necessitates a
153 minimum of 4-5 samples per cluster for dependable internal metrics [19], resulting in $K \leq N/4 \approx 20$
154 for our corpus of $N = 81$; (2) **Interpretive manageability** beyond $K = 15$ escalates cognitive
155 load in categorical processing [21]; (3) **Dimensional constraints** where $K > 2D$ may jeopardize
156 distance-based separation in a 13-dimensional space [20]. This principled range mitigates issues
157 related to arbitrary constraint selection while preserving atlas usability.

158 The optimal configuration is Agglomerative (average) at $k = 15$ on 4D UMAP, achieving silhouette
159 0.664 ± 0.008 (5 seeds) with high neighborhood preservation (trustworthiness/continuity $\approx 0.81 \pm$
160 0.01). My $K = 15$ selection aligns with theoretical guidelines: each cluster contains 5.4 samples on
161 average (above the 4-5 minimum), while remaining interpretively manageable. Alternative criteria
162 (Gap statistic, Elbow method) suggest optimal ranges of $K \in [12, 18]$ and $K \in [8, 14]$ respectively,
163 with convergence around our selected value validating the methodological approach. KMeans peaks
164 near $k \in \{14, 15\}$ but remains below hierarchical performance; density methods achieve higher
165 silhouettes primarily via noise exclusion, which I report but do not use for exhaustive atlas labeling.
166 Bootstrap resampling confirms result stability (coefficient of variation $< 2\%$).

167 4.1 Clustering and Visualization

168 I report the clustering sweep exactly as run on the released features and scripts. Features are
169 TF-IDF weighted codebook counts with row-wise L2 normalization, produced by our systematic
170 preprocessing pipeline. My systematic sweep evaluates *projection spaces* (RAW/SVD/UMAP with
171 varying dimensionality and neighborhood parameters) and algorithms (KMeans, Agglomerative,
172 DBSCAN, OPTICS). Metric conventions, primary ranking, and handling of density-method noise
173 follow §3.2.

174 **Best (internal validity).** The optimal configuration by *multi-metric evaluation* is *Agglomerative*
175 (*average linkage*) with $k=15$ on 4D UMAP (10 nearest neighbors, minimum distance 0.01); silhouette
176 0.664 ± 0.008 , satisfying the trustworthiness/continuity guardrails (≈ 0.81).

Table 5: Best runs per algorithm using multi-metric evaluation (silhouette, trustworthiness/continuity); density methods report noise ratio but are excluded from exhaustive atlas labeling.

Algorithm	#Clusters	Noise (%)	Silhouette	Trust./Cont.	UMAP Configuration
K-means	15	0.0	0.483	0.805/0.812	4D (neighbors=10, min_dist=0.01)
Agglomerative	15	0.0	0.664	0.805/0.812	4D (neighbors=10, min_dist=0.01)
DBSCAN	2	71.6	0.887	0.795/0.833	8D (neighbors=10, min_dist=0.1)
OPTICS	5	59.3	0.809	0.801/0.814	4D (neighbors=15, min_dist=0.1)

177 **Viewer labeling (communication-first).** For public exploration, I also provide a lower- k labeling
178 to keep the map readable. This is a communication-oriented choice distinct from the silhouette-
179 maximizing run; I avoid mixing the two objectives.

180 4.2 Cluster Analysis and Discovered Patterns

181 Pattern discovery integrates quantitative clustering outcomes with qualitative visual analysis using an
182 interactive web interface (Figure 1). The viewer facilitates systematic examination of cluster borders,
183 artist trajectories, and temporal relationships, bolstering the interpretations below while preserving
184 analytical objectivity via k-NN membership and rank-based proximity metrics.

185 **Methodological Cohesion: Stelarc’s Body Intervention Art.** Cluster 4 shows strong artist-specific
186 cohesion: Stelarc’s four works share consistently high within-cluster proximity in terms of mutual
187 k -NN membership and small rank displacement, rather than relying on absolute map distances.
188 *Suspensions* (1976), *Third Hand* (1980), *Stomach Sculpture* (1993), and *Ear on Arm* (2006) jointly
189 indicate a persistent methodological domain across three decades, centered on cyborg embodiment
190 and posthuman performance.

191 **Segmented Distribution of Tissue Culture Art.** Tissue-culture works appear in two nearby regions
192 with substantial neighborhood overlap, indicating related but distinguishable foci. Cluster 1 includes

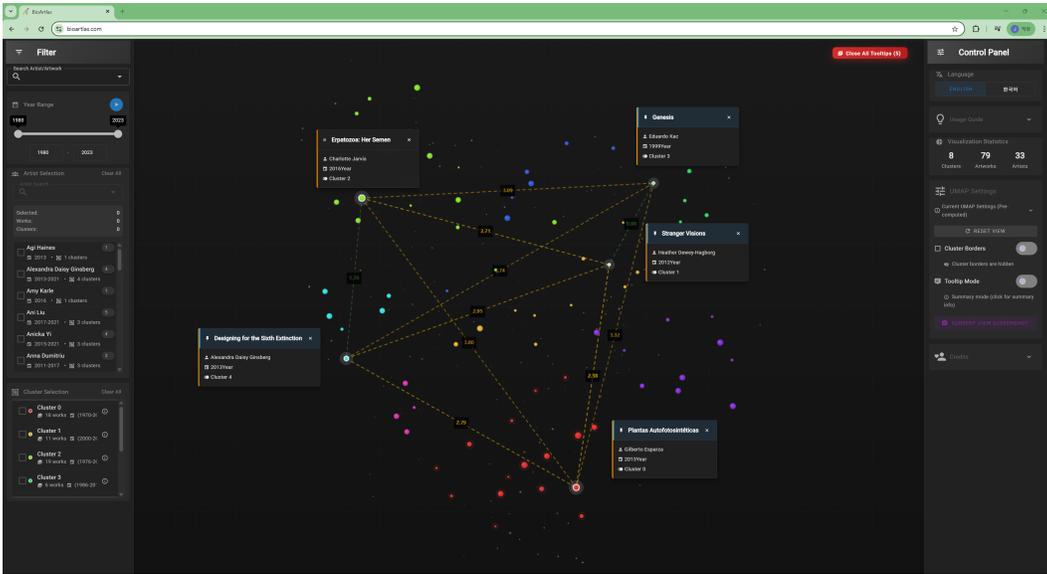


Figure 1: BioArtlas interactive visualization interface. <https://www.bioartlas.com>

193 Marta de Menezes’s early works and Špela Petrič’s plant-human studies. Cluster 13 encompasses
 194 Catts & Zurr’s mature works alongside recent DNA-based projects.

195 **Multi-cluster Distribution of Individual Artists.** Several artists’ works occupy distinct local
 196 neighborhoods with low mutual k -NN overlap across clusters, revealing methodological evolution
 197 over time. Joe Davis’s four works and Eduardo Kac’s three works each form separate neighborhood
 198 sets whose nearest-neighbor composition shifts across periods, indicating technical and conceptual
 199 diversification within the same practice without relying on absolute map distances.

200 **Trans-temporal Conceptual Affinity.** Clustering prioritizes conceptual similarity over chronology.
 201 Joe Davis’s *Poetica Vaginal* (1986) and Jenna Sutela’s *nimiia cētii* (2018) show high reciprocal k -NN
 202 membership and small rank-displacement, indicating a shared microbial–linguistic focus despite a
 203 32-year gap; I report this affinity via neighborhood overlap and rank structure rather than absolute
 204 UMAP distances.

205 5 Conclusion

206 My axis-aware methodology efficiently structures the multidimensional complexity of bioart using
 207 ing domain-informed semantic representations, attaining a silhouette score of 0.664 ± 0.008 with
 208 significant interpretability. Systematic assessment of over 800 configurations corroborates our
 209 methodological selections.

210 **Key Contributions:** (1) **Axis-aware representation learning** preserving semantic distinctions
 211 across thirteen heterogeneous dimensions while enabling cross-dimensional comparison; (2) **domain-**
 212 **informed codebook construction** grouping related concepts into unified clusters, addressing
 213 cultural terminology polysemy; (3) **systematic evaluation framework** explicitly separating ana-
 214 lytical optimization from communicative design; (4) **discovery of four organizational pat-**
 215 **terns**—methodological cohesion, technique segmentation, artistic evolution, and trans-temporal
 216 affinities—that complement art-historical analyses.

217 **Broader Impact:** This approach offers a model for computational cultural study in areas character-
 218 ized by multidimensional complexity. The modular design facilitates systematic expansion across
 219 geographic borders, multi-annotator validation, and cross-domain extension.

220 **Future Works:** Present limitations encompass Western-centric bias and single-annotator labeling. I
 221 intend to enlist bioart specialists and curators for the multi-annotator validation of our 13-dimensional
 222 annotations, facilitating the evaluation of inter-rater reliability. Geographic expansion will specifically
 223 encompass bioart communities in the Asia-Pacific, Latin America, and Africa to reduce bias while
 224 integrating contemporary AI/ML works.

References

- 225 [1] Eduardo Kac. Transgenic art. *Leonardo Electronic Almanac*, 6(11):289–296, 1998.
- 227 [2] Oron Catts and Ionat Zurr. The ethics of experiential engagement with the manipulation of life.
228 In *Tactical Biopolitics-Art, Activism, and Technoscience*, pages 125–142. MIT Press, 2008.
- 229 [3] Eduardo Kac. *Telepresence & bio art: networking humans, rabbits, & robots*. University of
230 Michigan Press, 2005.
- 231 [4] Marquard Smith, editor. *Stelarc: The Monograph*. MIT Press, Cambridge, MA, 2005. ISBN
232 978-0262693608. First comprehensive study of Stelarc’s work practice.
- 233 [5] Jens Hauser. Bio art - taxonomy of an etymological monster. In *Hybrid: Living in Paradox*,
234 pages 182–193. 2005.
- 235 [6] ZKM | Center for Art and Media Karlsruhe. Biomedica: The age of media with life-like behavior.
236 <https://zkm.de/en/exhibition/2021/12/biomedica>, 2021–2022. Exhibition, Accessed
237 2025-08-09.
- 238 [7] Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford
239 University Press, 07 2005. ISBN 9780199256044. doi: 10.1093/oso/9780199256044.001.0001.
240 URL <https://doi.org/10.1093/oso/9780199256044.001.0001>.
- 241 [8] Donna Haraway. When species meet: Staying with the trouble. *Environment and Planning D:
242 Society and Space*, 28(1):53–55, 2010.
- 243 [9] Jeroen Baas et al. Expert knowledge integration in historical record analysis. *Journal of Digital
244 Humanities*, 2022.
- 245 [10] Graham M Jones, Shai Satran, and Arvind Satyanarayan. Toward cultural interpretability: A
246 linguistic anthropological framework for describing and evaluating large language models. *Big
247 Data & Society*, 12(1):20539517241303118, 2025.
- 248 [11] Lev Manovich. Cultural analytics: Visualizing cultural patterns in the era of “more media”.
249 *Domus March*, 2009.
- 250 [12] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-
251 networks. *arXiv preprint arXiv:1908.10084*, 2019.
- 252 [13] Ars Electronica. Artificial life & intelligence category (prix ars electronica). [https://
253 ars.electronica.art/prix/en/categories/artificial-life-intelligence/](https://ars.electronica.art/prix/en/categories/artificial-life-intelligence/). Ac-
254 cessed 2025-08-09.
- 255 [14] Bio Art & Design Award. Bad award. <https://www.badaward.nl/>, 2011–2024. Competition
256 discontinued in 2025, Accessed 2025-08-09.
- 257 [15] MIT List Visual Arts Center. Symbionts: Contemporary artists and the biosphere. [https://
258 listart.mit.edu/exhibitions/symbionts-contemporary-artists-biosphere](https://listart.mit.edu/exhibitions/symbionts-contemporary-artists-biosphere),
259 2022–2023. Exhibition, Accessed 2025-08-09.
- 260 [16] S+T+ARTS Prize. Grand prize for innovation at the nexus of science, technology, and the arts.
261 <https://starts-prize.aec.at/en/>. Accessed 2025-08-09.
- 262 [17] ISEA International. Isea symposium archives. <https://www.isea-archives.org/>. Ac-
263 cessed 2025-08-09.
- 264 [18] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters
265 in a data set via the gap statistic. *Journal of the royal statistical society: series b (statistical
266 methodology)*, 63(2):411–423, 2001.
- 267 [19] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques.
268 *Journal of intelligent information systems*, 17(2):107–145, 2001.
- 269 [20] Glenn W Milligan and Martha C Cooper. An examination of procedures for determining the
270 number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

271 [21] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster*
272 *analysis*. John Wiley & Sons, 1990.