



SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Development of an AI-Enhanced Extended Reality Learning Environment for Biotechnology Laboratories: A Secondary-Data-Driven Design and Risk-Aware Evaluation

Rissa A. Lasap^{1*}, Luigi Carlo M. De Jesus²

^{1,2} Biological Science Department, Centro Escolar University, Manila, Metro Manila, Philippines. E-mail: rissalasap22@gmail.com, ORCID: 0009-0002-1029-3460

² Naveen Jindal Youth Fellowship Program, O.P. Jindal Global University, Sonapat, Haryana, India.

² Department of Computer, Electronics, and Electrical Engineering, De La Salle University, Manila, Metro Manila, Philippines.

² School of Engineering, Asia Pacific College, Makati City, Metro Manila, Philippines. E-mail: luigi_dejesus@dlsu.edu.ph | ORCID: 0000-0001-7889-6044

Abstract

Extended reality (XR) can provide repeatable, low-risk rehearsal of biotechnology laboratory procedures, yet existing virtual laboratories seldom combine adaptive artificial intelligence (AI), explicit safety control, process-aware analytics, protocol-level domain fidelity, and auditable governance. This study develops BioXR-AI, a safety-constrained XR information system, through a structured secondary evidence map and computational verification. Thirty-six peer-reviewed studies published from 2008 to 2025 were coded across ten capabilities spanning biotechnology context, immersion, virtual laboratories, adaptive AI, analytics, safety, procedural assessment, instructional accessibility, governance, and complementarity with physical laboratories. Evidence-weighted gap analysis identified safety scaffolding (weighted prevalence = 0.080), biotechnology specificity (0.125), governance and explainability (0.210), adaptive AI (0.236), and learning analytics (0.248) as high-priority design requirements. No included study combined biotechnology context, adaptive AI, immersive XR, explicit safety control, analytics, and governance in one learning environment. These gaps informed a layered architecture comprising a protocol digital twin, uncertainty-aware event interpretation, risk-sensitive feedback, grounded explanatory AI, learning analytics, accessibility support, and faculty-controlled governance. The architecture was evaluated using 7,200 synthetic protocol traces covering eight biotechnology procedures and ten deviation classes. On two held-out procedures, BioXR-AI achieved $F1 = 0.951$, critical-error recall = 0.904, and 0.46 unnecessary alerts per 100 benign traces, outperforming finite-state XR and linear-checklist baselines under explicitly stated detector assumptions. The contribution is a reproducible design-and-verification framework for safety-critical AI-XR laboratory learning. Human learning, usability, transfer, and real-world safety effects remain to be established through prospective trials.

Keywords: artificial intelligence; biotechnology laboratory; extended reality; human-computer interaction; learning analytics; protocol digital twin; safety-critical systems; virtual laboratory

1. Introduction

Biotechnology laboratory education requires learners to coordinate conceptual knowledge, procedural sequence, instrument handling, contamination control, time-sensitive decisions, technical language, and occupational safety. Unlike lecture content, laboratory competence is enacted through interdependent actions: a learner can understand polymerase chain reaction conceptually yet still compromise an experiment through incorrect pipetting, reagent substitution, unsafe centrifuge loading, premature disposal, or a timing error. Physical laboratories are therefore indispensable, but they are expensive to repeat, constrained by consumables and equipment, and unsuitable for unlimited rehearsal of hazardous or failure-prone scenarios. Virtual and remote laboratories can widen access and support preparation, while the strongest evidence favors deliberate complementarity with physical practice rather than replacement [1–3].

Biotechnology is a particularly promising domain for virtual laboratory design. Gamified simulations have been used to improve preparation and knowledge in biotechnology courses [4], and later case studies and reviews have documented their value for animal biotechnology, biology, and procedure familiarization [5–7]. However, most systems remain linear, scripted, and monolingual. They frequently record completion rather than interpret why an action is wrong, whether a variant is legitimate, how serious the deviation is, or when the learner should be interrupted. These limitations are consequential because a generic hinting mechanism can treat a harmless ordering variation and a contamination event as equivalent, while an unconstrained conversational model can generate plausible but unsafe laboratory advice.

Immersive XR adds embodied interaction, spatial awareness, and presence, but immersion is not itself a guarantee of learning. Reviews and meta-analyses report positive potential alongside heterogeneous designs and outcomes [8–11]. Controlled studies show that immersive representations can increase presence and motivation while also increasing extraneous cognitive load or reducing learning when guidance is insufficient [12–16]. The design problem is therefore not simply whether to use XR, but how to coordinate immersion, procedural reasoning, cognitive support, and safety.

AI and learning analytics offer mechanisms for adaptation and trace interpretation. Virtual-laboratory analytics increasingly use event logs for assessment and feedback [17], while laboratory-safety studies demonstrate the value of digital and 360-degree preparation [18–20]. Broader AI-in-education scholarship identifies personalization, assessment, prediction, and tutoring as major application families [21–24]. At the same time, human-centered reviews emphasize control, reliability, transparency, and stakeholder participation [25]. AI literacy and large-language-model research further warn that adaptive dialogue must be accompanied by critical evaluation, grounding, privacy protection, and safeguards against hallucination and overreliance [26–28].

An equally important issue is biotechnology-domain fidelity. XR systems must represent biological materials, aseptic boundaries, contamination pathways, instrument dependencies, timing windows, units, and disposal requirements accurately enough to support meaningful rehearsal. Generic STEM simulations may be immersive yet fail to capture the causal and safety structure of biotechnology procedures. Adaptive feedback in augmented reality must also be coordinated with cognitive-load management [29], while science-language and situated-learning research indicate that explanations should preserve disciplinary meaning and remain accessible to learners [30,31]. The common ground of this interdisciplinary study is therefore domain-informed AI-XR design: biological-science expertise anchors protocol validity and laboratory pedagogy, while computer, electronics, and electrical engineering expertise anchors sensing, AI reasoning, XR interaction, analytics, and system governance.

This paper develops BioXR-AI, a domain-informed, safety-constrained XR learning environment for biotechnology laboratories. The contribution is positioned within computer information systems, human-computer interaction, AI, information governance, biotechnology education, and industrial training. It combines (i) a transparent secondary-data evidence corpus, (ii) an evidence-weighted requirement-prioritization method, (iii) an information-system architecture that separates deterministic safety control from generative explanation, and (iv) a reproducible trace-level benchmark. The study makes no claim of measured student learning or real-laboratory safety improvement; those claims require prospective human-subject and transfer studies.

1.1 Research questions

RQ1: Which capabilities are well represented or underrepresented in the secondary evidence on biotechnology virtual laboratories, immersive learning, adaptive AI, procedural assessment, and laboratory safety?

RQ2: How can the resulting evidence gaps be translated into a domain-valid, safety-critical AI-XR information-system architecture for biotechnology laboratories?

RQ3: Under explicit synthetic detector assumptions, how does the proposed orchestration layer compare with linear-checklist and finite-state XR baselines on held-out biotechnology procedures?

RQ4: What governance, validation, and deployment conditions are required before educational or operational claims are justified?

Table 1. Claimed contributions and corresponding evidence

Contribution	Evidence supplied	Boundary of claim
Secondary evidence map	36 peer-reviewed studies with traceable bibliographic records; open study-level CSV; ten capability codes	Structured evidence mapping, not an exhaustive meta-analysis of learning effects
Requirement prioritization	Evidence-weighted prevalence, gap, importance, and priority scores	Priorities depend on transparent coding and prespecified importance weights
BioXR-AI architecture	Protocol digital twin, risk policy, instructional scaffold, analytics, governance	Design artifact; no claim of completed production deployment
Computational verification	7,200 traces, held-out procedures, baseline comparisons, latency and mutation analysis	Tests orchestration under simulated detector probabilities, not real learners or sensors

2. Materials and Methods

2.1 Secondary-data corpus and eligibility

The secondary corpus was assembled on 29 July 2026 through targeted searches of publisher and DOI-indexed records, backward and forward citation chaining from recent systematic reviews, and verification against bibliographic landing pages. Search concepts combined four blocks: (a) “virtual laboratory,” “virtual reality,” “augmented reality,” “mixed reality,” or “extended reality”; (b) “biotechnology,” “biology,” “biochemistry,” “laboratory,” or “laboratory safety”; (c) “artificial intelligence,” “adaptive feedback,” “learning analytics,” “chatbot,” or “digital twin”; and (d) “laboratory pedagogy,” “human factors,” “governance,” “accessibility,” or “explainability.” The purpose was to construct a design-relevant evidence map spanning the literatures that the proposed system must integrate. It was not presented as an exhaustive database-wide systematic review or effect-size meta-analysis. Records were eligible when they were peer-reviewed, had a resolvable DOI, and contributed empirical, review, or theoretical evidence to at least one design capability. Studies were retained across higher education, STEM laboratories, life sciences, safety training, AI in education, governance, and instructional accessibility because no

single literature contains all relevant system requirements. Codes were conservative: a capability received a value of 1 only when it was a central or explicit feature of the study; otherwise it received 0. Consequently, a zero denotes “not centrally reported in the coded evidence,” not proof that the feature was entirely absent from every implementation detail. Recent studies in the corpus extended the evidence base through life-science student perceptions, comparative virtual-learning synthesis, immersive-VR and VR/AR reviews, and distance science-laboratory analysis [32–36].

Table 2. Evidence coding fields and operational definitions

Capability	Operational coding rule
Biotechnology context	Study centrally addresses biotechnology, biology, biochemistry, physiology, or life-science laboratories.
Immersive XR	Head-mounted VR, AR, MR, 360-degree environments, or equivalent embodied immersive interaction is central.
Virtual laboratory	A simulated or remotely accessed laboratory environment is central.
Adaptive AI	The system or review centrally addresses intelligent tutoring, machine learning, adaptive feedback, or generative AI.
Learning analytics	Learner-event traces, dashboards, prediction, assessment analytics, or data-driven adaptation are central.
Safety scaffolding	Laboratory or occupational safety behavior is explicitly taught, detected, or governed.
Procedural assessment	Actions, sequences, skills, or performance processes are assessed rather than content knowledge alone.
Multilingual support	Multiple languages, translation, disciplinary-language access, or language learning is central.
Governance/explainability	Human control, transparency, ethics, privacy, reliability, auditability, or explainability is explicit.
Physical-lab complementarity	The study explicitly frames virtual experience as preparation for, complement to, or blend with physical practice.

2.2 Evidence weighting and requirement priority

Each study received an evidence weight w_i reflecting the inferential strength of its design for the present evidence-mapping purpose: meta-analysis = 1.00; systematic review = 0.95; controlled empirical study = 0.90; scoping review or strong review = 0.80–0.90; empirical case, survey, or design study = 0.70–0.80; and theoretical or design reflection = 0.60–0.70. These weights do not rank journals or claim a universal hierarchy of evidence. They provide a transparent sensitivity mechanism so that a capability reported only in low-strength evidence does not appear equivalent to one supported by multiple syntheses and controlled studies.

For capability j , evidence-weighted prevalence was calculated as $W_j = \sum_i (w_i x_{ij}) / \sum_i w_i$, where $x_{ij} \in \{0,1\}$. The evidence gap was $G_j = 1 - W_j$. A design-importance coefficient I_j between 0.65 and 1.00 was specified from the hazard and learning consequences of omitting the capability. The priority score was $P_j = I_j G_j$. The complete coding table, weights, equations, and derived summaries are supplied as CSV files so that readers can substitute alternative weights or coding decisions.

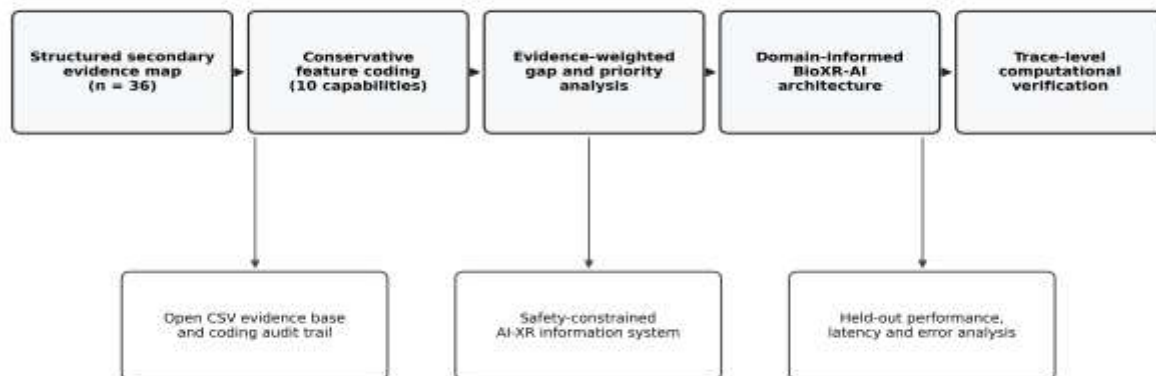


Figure 1. Secondary-data-to-artifact workflow. The evidence map informs requirements, architecture, and computational verification; it does not substitute for a human trial.

2.3 BioXR-AI information-system architecture

BioXR-AI is organized as a set of separable services rather than a single opaque AI agent. The XR interaction layer captures object manipulation, controller or hand events, gaze, voice, timing, and spatial context. A semantic event layer normalizes raw observations into tuples (action, object, parameter, location, time, confidence). The protocol digital twin represents each approved laboratory procedure as a directed attributed graph containing prerequisites, acceptable parameter intervals, conditional branches, recoverable variants, hazard classes, and evidence links.

For each event t , the policy estimates deviation probability $p(d_t|x_t,G)$, hazard severity s_t , observation confidence c_t , and recoverability r_t . The risk score is $R_t = p(d_t|x_t,G) \times s_t \times c_t \times (1 - r_t)$. Threshold bands map the score to silent observation, a noninterruptive cue, a contextual hint, a temporary pause, or a hard safety block. This formulation makes two design commitments explicit. First, the system should not interrupt merely because a learner differs from one canonical sequence when an approved variant is possible. Second, low-confidence perception should trigger clarification or abstention rather than an authoritative but uncertain safety instruction.

Generative AI is restricted to grounded explanation. It receives a bounded evidence packet containing the approved step, the detected deviation, confidence, hazard classification, allowed corrective actions, source identifiers, learner support preference, and explanation level. It cannot create a new reagent, parameter, or laboratory procedure outside the protocol graph. The deterministic policy therefore retains control over safety-critical actions, while the language model can vary wording, examples, and dialogic scaffolding within approved limits.

The instructional scaffold is not a free-form content-generation layer. It uses a faculty-approved domain terminology table, parallel plain-language micro-explanations, unit-preserving templates, and a terminology lock for reagents, organisms, equipment, and hazard labels. Optional localization may be added only after expert validation. The layer also records repeated requests for definitions or clarification so that conceptual uncertainty can be distinguished from procedural difficulty. This design links biotechnology-domain expertise with adaptive information-system behavior without allowing explanatory variation to alter the approved protocol.

Learning analytics store event sequences, confidence, feedback level, response time, recovery path, clarification requests, and mastery evidence. Governance services version protocols and instructional resources, enforce roles, log overrides, retain source provenance, and apply privacy and retention rules. Faculty can inspect why a feedback action occurred and can disable generative explanation without disabling deterministic safety checks.

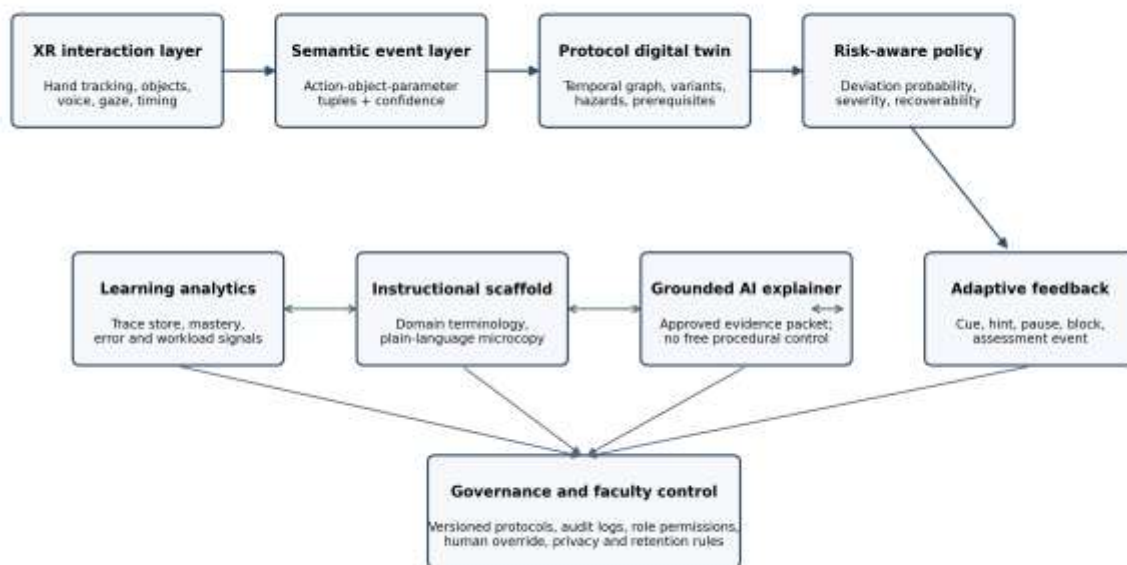


Figure 2. BioXR-AI architecture. Safety control remains deterministic and auditable; generative AI is limited to grounded explanation and instructional scaffolding.

2.4 Computational benchmark

Because no real multimodal interaction dataset was available for the proposed system, the architecture was evaluated with an explicit synthetic protocol-trace benchmark. Eight procedures were selected to cover liquid handling, aseptic technique, molecular biology, analytical measurement, and equipment safety. Each procedure generated 900 traces, producing $n = 7,200$. Ten trace classes were represented: omission, permutation, wrong parameter, wrong reagent, contamination, timing violation, equipment mismatch, personal-protective-equipment violation, disposal violation, and benign variant. The generator used fixed seed 20260714.

Table 3. Synthetic biotechnology protocol benchmark

Procedure	Nominal steps	Hazard checkpoints	Evaluation split
Micropipetting	18	4	Development
Aseptic streaking	20	6	Development
Serial dilution	22	5	Development
DNA extraction	30	7	Development
PCR setup	25	8	Development
Gel electrophoresis	28	7	Development
Spectrophotometry	18	4	Held out
Centrifuge balancing	16	6	Held out

Three orchestration models were compared. The linear checklist verified expected step occurrence and order but had weak access to context, parameters, hazards, or benign alternatives. The finite-state XR baseline tracked explicit states and common transitions, improving order and parameter checks but lacking uncertainty-aware graph relations and cross-cutting hazards. BioXR-AI used temporal graph reasoning, risk scoring, confidence-aware abstention, benign-variant acceptance, and hazard relations. Detector capabilities were specified as mutation-specific probabilities rather than learned from human data. The benchmark therefore evaluates how the orchestration logic behaves under stated detector assumptions, not whether a deployed sensor or AI model has already achieved those probabilities.

2.5 Outcomes and statistical analysis

Primary outcomes were F1 score, recall for critical errors, and unnecessary alerts per 100 benign traces. Secondary outcomes were accuracy, precision, recall, specificity, coverage, mutation-specific accuracy, and execution latency. BioXR-AI could abstain when observation confidence was below threshold; classification metrics were therefore reported together with coverage. Ninety-five percent Wilson confidence intervals were calculated for recall and mutation accuracy. Paired model differences were evaluated with McNemar tests on non-abstained comparable cases. Latency was measured over seven repeated runs and reported as median microseconds per event and the 95th percentile across runs. The complete trace-level dataset, predictions, analysis outputs, and executable code are provided in the supplementary package.

2.6 Ethics and reproducibility

The study used published secondary data and synthetic protocol traces. It did not involve human participants, personal data, biological samples, or physical laboratory experimentation; institutional ethics approval was therefore not required for this computational design study. All secondary-study records include DOI URLs, coding notes, and validation status. The CSV files are encoded in UTF-8 and can be opened directly in spreadsheet or statistical software.

3. Results

3.1 Evidence-corpus profile

The evidence corpus contained 36 studies published from 2008 to 2025. It deliberately combined virtual-laboratory, immersive-learning, AI, analytics, laboratory-safety, biotechnology, and language-accessibility streams. The median integration score was 3 of ten capabilities (interquartile range 2–4). No secondary study reported all ten capabilities. Only 1 study combined adaptive AI with immersive XR, only 1 combined multilingual support with XR, and none combined biotechnology context, adaptive AI, and immersive XR. These results support an integration gap rather than a claim that each individual component is novel in isolation.

Table 4. Distribution of the secondary evidence by design stream

Evidence stream	Studies	Mean capabilities reported
Virtual laboratories	5	3.40
Biotechnology / life science	5	4.20
Immersive learning	11	2.00
AI and analytics	5	3.40
AI governance	4	3.25
Safety training	3	4.00
AI–XR integration	1	5.00
Language accessibility	2	2.00



Figure 3. Capability prevalence within evidence streams. Percentages represent explicit central reporting under the conservative coding rule.

3.2 Evidence-weighted gaps and priorities

Procedural assessment was the most frequently represented capability, with an evidence-weighted prevalence of 0.822. Immersive XR (0.493) and virtual laboratories (0.381) also had substantial coverage. In contrast, capabilities important to a safety-critical biotechnology system were sparsely integrated: safety scaffolding had weighted prevalence 0.080, multilingual or language-access support 0.109, biotechnology context 0.125, governance/explainability 0.210, adaptive AI 0.236, and learning analytics 0.248. The resulting priority scores ranked safety scaffolding first (0.920), followed by accessibility support (0.802), governance/explainability (0.750), biotechnology specificity (0.743), and adaptive AI (0.687).

Table 5. Evidence-weighted capability coverage and design priority

Capability	Studies	Weighted prevalence	Importance	Priority
Safety scaffolding	3	0.080	1.00	0.920
Multilingual support	4	0.109	0.90	0.802
Governance / explainability	8	0.210	0.95	0.750
Biotechnology context	5	0.125	0.85	0.743
Adaptive AI	9	0.236	0.90	0.687
Learning analytics	9	0.248	0.85	0.640
Physical-lab complementarity	12	0.321	0.75	0.509
Virtual laboratory	14	0.381	0.65	0.402
Immersive XR	17	0.493	0.65	0.329
Procedural assessment	30	0.822	0.95	0.169

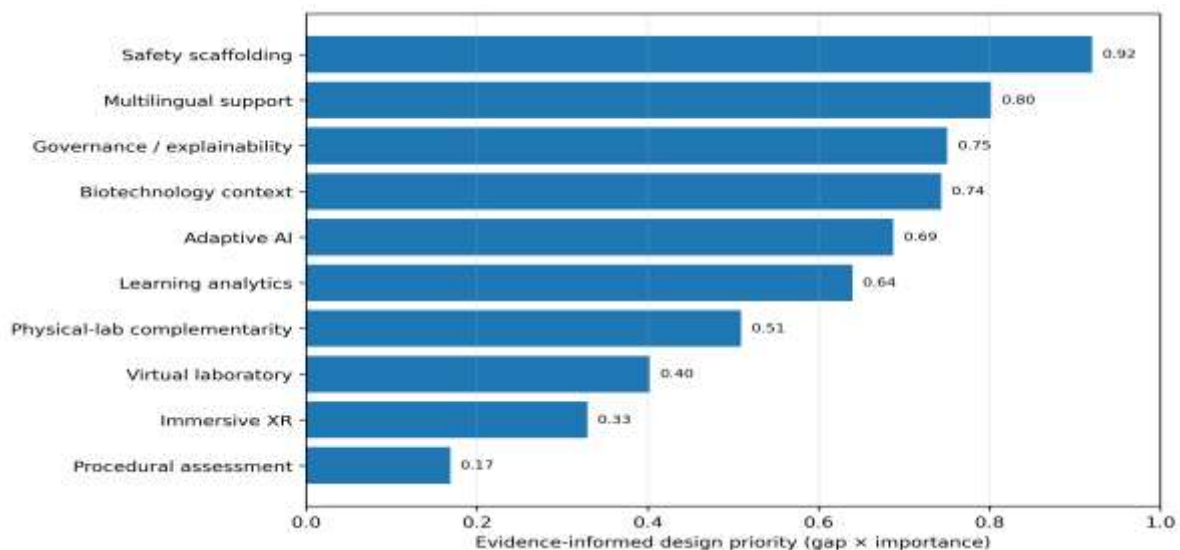


Figure 4. Evidence-informed design priorities. Higher values indicate capabilities that are important to the target system but weakly represented in the coded evidence.

3.3 Architecture requirements derived from the evidence

The evidence map changed the artifact in five material ways. First, safety was implemented as a first-class policy service rather than a feedback label because safety training was rare and validation practices were inconsistent. Second, instructional-accessibility support was placed between protocol semantics and explanation so that simplified or localized wording could not alter quantities, units, organism names, or hazard categories. Third, governance and explainability were implemented across all services through versioning, provenance, override, and retention controls. Fourth, adaptive AI was separated from the deterministic protocol controller to avoid allowing a generative model to improvise laboratory instructions. Fifth, the architecture preserved physical-laboratory complementarity by defining XR as preparation, rehearsal, and formative assessment rather than certification of tactile competence.

Table 6. Evidence-to-design traceability matrix

Evidence gap	BioXR-AI response	Verifiable design property
Sparse safety scaffolding	Risk score, hazard graph, interruption hierarchy, hard block	Every alert records deviation, severity, confidence, recoverability, and policy action
Sparse multilingual support	Curated English–Filipino terminology and plain-language microcopy	Translation version and terminology source stored with each explanation
Weak governance/explainability	Role-based control, protocol provenance, audit log, human override	Faculty can reconstruct and override every safety-relevant decision
Limited AI–XR integration	Semantic event layer and grounded explanatory AI	Generator receives only approved evidence packets; no direct equipment control
Limited biotechnology specificity	Protocol digital twins for eight representative procedures	Parameters, reagents, hazards, and variants encoded at procedure level
Need for physical complementarity	Preparation and rehearsal positioning; transfer validation gate	System does not issue physical-laboratory competency certification

3.4 Held-out computational performance

On the two held-out procedures (spectrophotometry and centrifuge balancing; $n = 1,800$ traces), BioXR-AI achieved coverage = 0.969, accuracy = 0.917, precision = 0.985, recall = 0.919, and F1 = 0.951. Critical-error recall was 0.904. The finite-state XR baseline achieved F1 = 0.806 and critical-error recall = 0.600; the linear checklist achieved F1 = 0.549 and critical-error recall = 0.287.

BioXR-AI also reduced unnecessary alerts to 0.46 per 100 benign traces, compared with 18.81 for finite-state XR and 49.08 for the linear checklist. This result is important for learning-system usability: a safety system that interrupts too frequently can train users to dismiss alerts, increase cognitive load, and obscure genuinely critical feedback.

Table 7. Performance on held-out procedures ($n = 1,800$ traces)

Model	Coverage	Accuracy	Precision	Recall	F1	Critical recall	False alerts/100 benign
Linear checklist	1.000	0.417	0.857	0.404	0.549	0.287	49.08
Finite-state XR	1.000	0.707	0.964	0.692	0.806	0.600	18.81
BioXR-AI	0.969	0.917	0.985	0.919	0.951	0.904	0.46

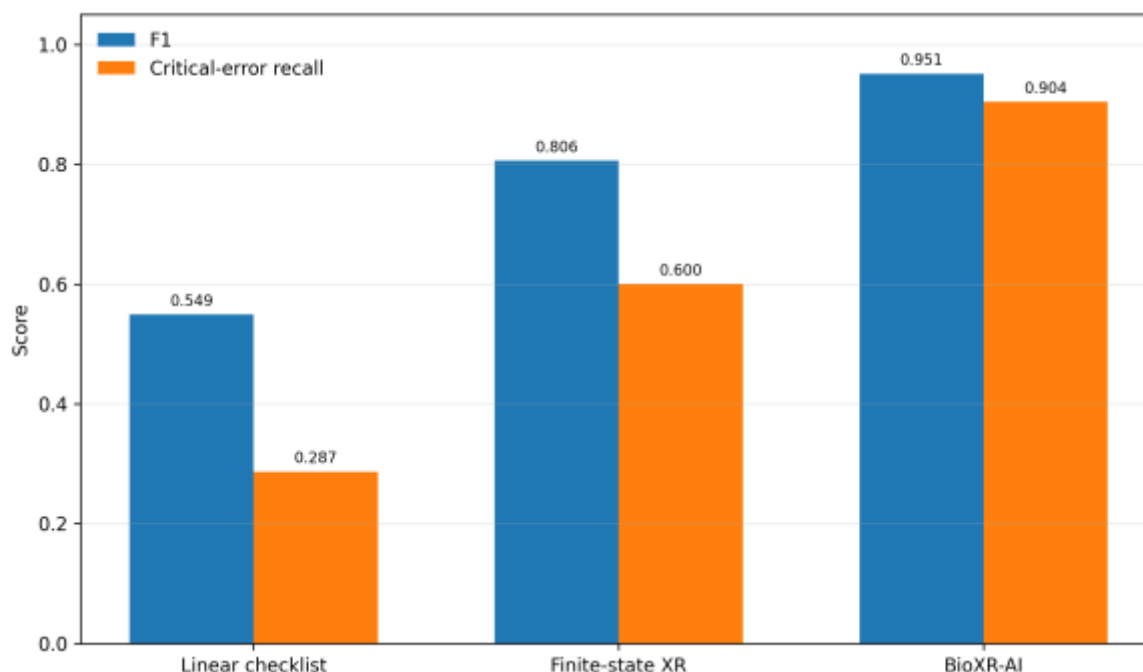


Figure 5. Held-out F1 score and critical-error recall under the stated synthetic detector assumptions.

3.5 Mutation-specific error profile and latency

Mutation analysis showed that the main advantage of BioXR-AI was not limited to ordinary omissions and permutations. Accuracy remained above 0.889 across all ten mutation classes, including contamination, timing, equipment, PPE, disposal, and benign variants. Finite-state XR performed relatively well on omission, permutation, and wrong reagent but was weaker on cross-cutting hazards and timing. The linear checklist performed poorly on contamination, timing, wrong parameter, equipment mismatch, PPE, and disposal because those deviations depend on contextual attributes not captured by simple sequence checking.

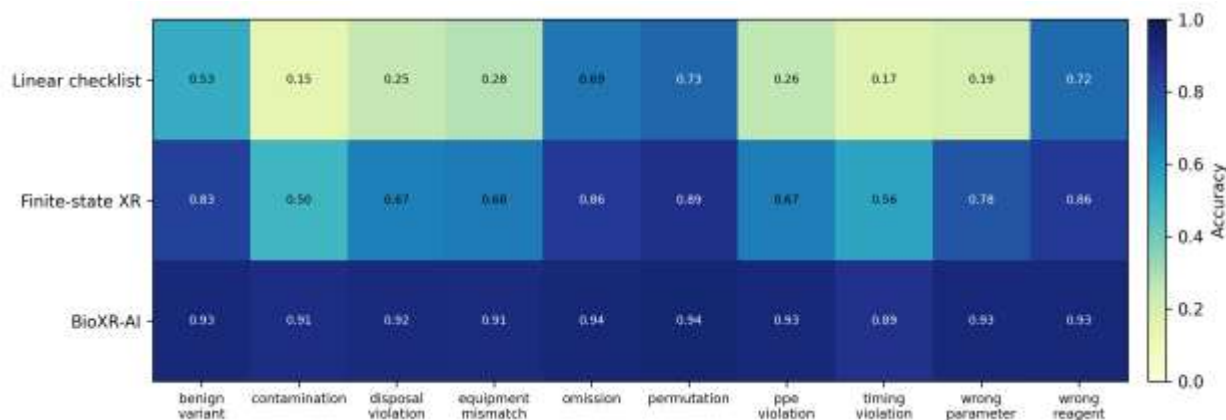


Figure 6. Mutation-specific accuracy across the complete 7,200-trace benchmark.

Execution latency remained negligible relative to XR interaction time. Median per-event latency was 0.866 microseconds for BioXR-AI, 0.255 microseconds for finite-state XR, and 0.155 microseconds for the linear checklist in the reference Python implementation. These figures measure only orchestration logic and exclude perception, rendering, networking, speech recognition, and language-model inference.

4. Discussion

4.1 Principal findings

The study found a fragmented evidence landscape. Virtual laboratories and immersive learning are mature enough to provide strong design foundations, while AI and learning analytics provide methods for adaptation and trace interpretation. However, biotechnology specificity, safety, instructional accessibility, and governance are usually treated in separate research streams. The scarcity of integrated studies is the main justification for the artifact. BioXR-AI does not claim novelty for virtual laboratories, adaptive feedback, or explanatory scaffolds individually; its contribution is the information architecture that makes these capabilities interoperable, constrained, and auditable in a safety-sensitive biotechnology setting.

The secondary data also challenge a technology-first design strategy. Immersive XR was relatively well represented, while safety, biotechnology specificity, and accessibility support were sparse. Investing first in visual fidelity would therefore optimize a capability that already has substantial research attention while leaving the highest-risk gaps unresolved. The priority analysis instead directs design resources toward protocol fidelity, hazard control, learner support, and governance before additional realism.

4.2 Contribution to computer information systems

From an information-systems perspective, BioXR-AI is a governed event-processing system. It transforms heterogeneous interaction data into semantically typed events, reasons over a versioned process model, applies a risk policy, generates role-appropriate feedback, and stores an auditable decision trace. This structure is relevant beyond education. Industrial biotechnology, pharmaceutical manufacturing, clinical laboratories, food processing, and quality-control facilities all require standard operating procedures, competency management, deviation logging, multilingual workforces, and controlled change management. The same architecture could support onboarding and recurrent training, provided domain protocols, regulations, translations, and sensors are validated for the target facility.

The protocol digital twin also provides a bridge between learning analytics and process management. Conventional analytics often describe clicks or time-on-task without representing the procedure being performed. By relating each event to prerequisites, hazards, and approved variants, analytics become process-aware. This enables questions such as whether a learner repeatedly bypasses sterility checks, whether clarification requests cluster around a specific term or step, or whether an alert policy causes avoidable interruption at a particular point in the procedure.

4.3 Biotechnology domain fidelity and instructional design

The biotechnology contribution depends on more than visual realism. Each protocol digital twin must encode scientifically valid materials, quantities, timing windows, aseptic boundaries, equipment constraints, acceptable variants, hazards, and disposal rules. Domain experts should verify both the nominal workflow and the failure modes because an XR sequence can appear plausible while omitting the causal relationships that determine contamination, sample integrity, or operator safety. The resulting model provides a formal interface between biological-science content and computational reasoning.

Implementation should be co-designed with biological-science faculty, laboratory instructors, safety officers, students, human-computer-interaction specialists, and accessibility experts. User testing should examine whether the system supports conceptual understanding, procedural fluency, cognitive-load management, and correct recovery from errors without encouraging automation bias. These questions define the authors' substantive common ground: biotechnology education and protocol validity on one side, and AI-XR information-system design and engineering validation on the other.

4.4 Safety, governance, and generative AI

Safety-critical educational AI should fail safely. BioXR-AI therefore treats the language model as an explanation service, not as the source of procedural truth. Approved protocols and hazard rules remain authoritative. High-risk decisions require deterministic evidence, and low-confidence observations should cause abstention or clarification. The architecture also supports faculty override, protocol rollback, explanation provenance, retention controls, and post-event audit. These mechanisms operationalize human-centered AI principles rather than relying on a general statement that "a human remains in the loop."

Governance is also necessary for domain content. An updated reagent label, revised institutional safety rule, changed equipment parameter, or corrected terminology entry should be versioned and propagated consistently. The system should record which protocol, hazard rule, and instructional-resource versions were shown during each session. Without this provenance, learning analytics may compare users who received materially different instructions, and safety investigations may be unable to reconstruct the information available at the time of an action.

4.5 Interpretation of the computational benchmark

The benchmark supports a narrow engineering conclusion: under the stated detector capability distributions, the graph-based, risk-aware orchestration policy is substantially more effective than sequence-only alternatives and produces far fewer nuisance alerts. It does not prove that real hand tracking, object recognition, speech recognition, or language models will reach the assumed probabilities. It also does not prove that learners will retain knowledge, transfer skills to physical laboratories, or behave more safely. The value of the benchmark is reproducibility: it exposes assumptions, provides mutation-level stress tests, and allows future detectors or policies to be substituted without changing the evaluation framework.

The held-out-procedure design offers a limited test of structural generalization. Spectrophotometry and centrifuge balancing were excluded from development, yet the same reasoning architecture was applied to them through new protocol graphs. This reflects the intended deployment model, where institutions author and validate protocol digital twins rather than retrain a monolithic model for every procedure. Future work should add adversarial traces, ambiguous sensor events, compound deviations, protocol revisions, and realistic temporal dependencies.

4.6 Limitations

Several limitations constrain the conclusions. First, the secondary corpus was a structured, design-oriented evidence map rather than an exhaustive database-wide systematic review. Its conservative binary codes simplify complex studies, and the evidence weights and importance coefficients are analytic choices. The open CSV permits

alternative coding, independent double-coding, sensitivity analysis, and expansion. Second, the benchmark traces were synthetic and the detector probabilities were assumed. Third, the architecture has not yet been evaluated for usability, accessibility, cybersickness, teacher workload, cybersecurity, cost, or interoperability with learning-management systems. Fourth, no student learning, procedural transfer, or real safety outcome was measured. Finally, the protocol models cover only representative procedures; expert validation against institution-specific biotechnology protocols remains necessary.

4.7 Prospective validation program

A staged validation program is recommended. Stage 1 should conduct expert verification of protocol graphs, biological assumptions, hazard rules, and instructional terminology. Stage 2 should evaluate usability, comprehension, cognitive load, and error recovery with representative learners using nonhazardous procedures. Stage 3 should compare desktop versus immersive delivery and conventional versus adaptive feedback in a preregistered study with delayed retention measures. Stage 4 should assess transfer to physical laboratory performance using blinded raters and predefined safety criteria. Stage 5 should evaluate institutional deployment, faculty workload, protocol governance, cybersecurity, and cost-effectiveness. Advancement between stages should require explicit safety and evidence gates.

5. Conclusions

This study developed an AI-enhanced XR learning environment for biotechnology laboratories by integrating secondary evidence with information-system design and computational verification. The 36-study evidence map showed that immersion and procedural assessment are comparatively well represented, whereas safety scaffolding, governance, biotechnology specificity, adaptive AI, and accessibility support remain weakly integrated. BioXR-AI addresses these gaps through a protocol digital twin, confidence-aware semantic event processing, risk-sensitive feedback, grounded generative explanation, instructional scaffolding, learning analytics, and faculty-controlled governance. In a reproducible 7,200-trace benchmark, the architecture outperformed linear-checklist and finite-state baselines on held-out procedures while sharply reducing nuisance alerts. The work provides a transparent foundation for implementation, but its educational and safety benefits must be established through expert review and prospective human studies. The strongest intended use is not replacement of physical laboratories, but safer preparation, repeated rehearsal, process-aware formative assessment, and domain-valid biotechnology training.

Acknowledgments

The authors acknowledge the research communities in biotechnology education, laboratory pedagogy, human-computer interaction, artificial intelligence, and information systems whose scholarship informed this interdisciplinary framework.

Funding Information

The authors received no specific funding for this work.

Author Contributions

Rissa A. Lasap: conceptualization; biotechnology-domain framing; laboratory pedagogy; secondary-evidence interpretation; writing-original draft; writing-review and editing. Luigi Carlo M. De Jesus: conceptualization; methodology; software; data curation; formal analysis; AI-XR system architecture; visualization; writing-original draft; writing-review and editing. Both authors are responsible for reviewing the final evidence coding, approving the submitted version, and ensuring the integrity of the work.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

All study-level secondary data, feature codes, derived summaries, synthetic protocol traces, predictions, statistical outputs, and analysis code are supplied as downloadable CSV and supplementary files accompanying this manuscript. Published-study content remains subject to the licenses of the original sources.

Ethics Statement

Ethics approval and informed consent were not required because the study used published secondary data and synthetic traces and did not involve human participants, personal data, biological samples, or physical laboratory experimentation.

Generative AI Use Disclosure

Generative AI tools assisted language editing, code drafting, and document preparation. The authors remain responsible for verifying the secondary-data coding, numerical results, references, domain assumptions, interpretations, and final manuscript before submission. No AI system is listed as an author.

References

1. de Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and Virtual Laboratories in Science and Engineering Education. *Science*, 340(6130), 305–308. <https://doi.org/10.1126/science.1230579>
2. Brinson, J. R. (2015). Learning Outcome Achievement in Non-Traditional (Virtual and Remote) versus Traditional (Hands-on) Laboratories: A Review of the Empirical Research. *Computers & Education*, 87, 218–237. <https://doi.org/10.1016/j.compedu.2015.07.003>
3. Potkonjak, V., Gardner, M., Callaghan, V., Mattila, P., Guetl, C., Petrović, V. M., & Jovanović, K. (2016). Virtual Laboratories for Education in Science, Technology, and Engineering: A Review. *Computers & Education*, 95, 309–327. <https://doi.org/10.1016/j.compedu.2016.02.002>
4. Bonde, M. T., Makransky, G., Wandall, J., Larsen, M. V., Morsing, M., Jarmer, H. Ø., & Sommer, M. O. A. (2014). Improving Biotech Education through Gamified Laboratory Simulations. *Nature Biotechnology*, 32(7), 694–697. <https://doi.org/10.1038/nbt.2955>
5. Yap, W. H., Teoh, M. L., Tang, Y. Q., & Goh, B. H. (2021). Exploring the Use of Virtual Laboratory Simulations before, during, and Post COVID-19 Recovery Phase: An Animal Biotechnology Case Study. *Biochemistry and Molecular Biology Education*, 49(5), 685–691. <https://doi.org/10.1002/bmb.21562>
6. Myburgh, P. H. (2022). Reflecting on the Creation of Virtual Laboratory Experiences for Biology Students. *Frontiers in Education*, 7, 796840. <https://doi.org/10.3389/feduc.2022.796840>
7. Mercado, J. C., & Picardal, J. P. (2023). Virtual Laboratory Simulations in Biotechnology: A Systematic Review. *Science Education International*, 34(1), 52–57. <https://doi.org/10.33828/sei.v34.i1.6>
8. Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A Systematic Review of Immersive Virtual Reality Applications for Higher Education: Design Elements, Lessons Learned, and Research Agenda. *Computers & Education*, 147, 103778. <https://doi.org/10.1016/j.compedu.2019.103778>
9. Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of Virtual Reality-Based Instruction on Students' Learning Outcomes in K-12 and Higher Education: A Meta-Analysis. *Computers & Education*, 70, 29–40. <https://doi.org/10.1016/j.compedu.2013.07.033>
10. Jensen, L., & Konradsen, F. (2018). A Review of the Use of Virtual Reality Head-Mounted Displays in Education and Training. *Education and Information Technologies*, 23(4), 1515–1529. <https://doi.org/10.1007/s10639-017-9676-0>
11. Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive Virtual Reality as a Pedagogical Tool in Education: A Systematic Literature Review of Quantitative Learning Outcomes and Experimental Design. *Journal of Computers in Education*, 8, 1–32. <https://doi.org/10.1007/s40692-020-00169-2>
12. Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding Immersive Virtual Reality to a Science Lab Simulation Causes More Presence but Less Learning. *Learning and Instruction*, 60, 225–236. <https://doi.org/10.1016/j.learninstruc.2017.12.007>
13. Parong, J., & Mayer, R. E. (2018). Learning Science in Immersive Virtual Reality. *Journal of Educational Psychology*, 110(6), 785–797. <https://doi.org/10.1037/edu0000241>
14. Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A Theoretical Research-Based Model of Learning in Immersive Virtual Reality. *Educational Psychology Review*, 33, 937–958. <https://doi.org/10.1007/s10648-020-09586-2>
15. Makransky, G., & Lilleholt, L. (2018). A Structural Equation Modeling Investigation of the Emotional Value of Immersive Virtual Reality in Education. *Educational Technology Research and Development*, 66, 1141–1164. <https://doi.org/10.1007/s11423-018-9581-2>
16. Makransky, G., Borre-Gude, S., & Mayer, R. E. (2019). Motivational and Cognitive Benefits of Training in Immersive Virtual Reality Based on Multiple Assessments. *Journal of Computer Assisted Learning*, 35(6), 691–707. <https://doi.org/10.1111/jcal.12375>
17. Elmoazen, R., Saqr, M., Khalil, M., & Wasson, B. (2023). Learning Analytics in Virtual Laboratories: A Systematic Literature Review of Empirical Research. *Smart Learning Environments*, 10, 23. <https://doi.org/10.1186/s40561-023-00244-y>
18. Viitaharju, P., Yliniemi, K., Nieminen, M., & Karttunen, A. J. (2021). Learning Experiences from Digital Laboratory Safety Training. *Education for Chemical Engineers*, 34, 87–93. <https://doi.org/10.1016/j.ece.2020.11.009>
19. Girmay, S., Yliniemi, K., Nieminen, M., Linnera, J., & Karttunen, A. J. (2024). Enhancing 360° Virtual Laboratory Safety Training with Linear Learning Pathway Design: Insights from Student Experiences. *Education for Chemical Engineers*, 47, 12–21. <https://doi.org/10.1016/j.ece.2024.02.002>
20. Stefan, H., Mortimer, M., & Horan, B. (2023). Evaluating the Effectiveness of Virtual Reality for Safety-Relevant Training: A Systematic Review. *Virtual Reality*, 27, 2839–2869. <https://doi.org/10.1007/s10055-023-00843-7>
21. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators?. *International Journal of Educational Technology in Higher Education*, 16, 39. <https://doi.org/10.1186/s41239-019-0171-0>
22. Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, Challenges, Roles and Research Issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. <https://doi.org/10.1016/j.caeai.2020.100001>

23. Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, 8, 75264–75278. <https://doi.org/10.1109/ACCESS.2020.2988510>
24. Roll, I., & Wylie, R. (2016). Evolution and Revolution in Artificial Intelligence in Education. *International Journal of Artificial Intelligence in Education*, 26(2), 582–599. <https://doi.org/10.1007/s40593-016-0110-3>
25. Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-Centred Learning Analytics and AI in Education: A Systematic Literature Review. *Computers and Education: Artificial Intelligence*, 6, 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
26. Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial Intelligence Literacy in Higher and Adult Education: A Scoping Literature Review. *Computers and Education: Artificial Intelligence*, 3, 100101. <https://doi.org/10.1016/j.caeai.2022.100101>
27. Kasneci, E., et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
28. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What If the Devil Is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education. *Smart Learning Environments*, 10, 15. <https://doi.org/10.1186/s40561-023-00237-x>
29. Sun, J., & Liao, T. (2025). Designing Adaptive Feedback Systems for Managing Cognitive Load in Augmented Reality. *Design Science*, 11, e49. <https://doi.org/10.1017/dsj.2025.10040>
30. Brown, B. A., & Ryoo, K. (2008). Teaching Science as a Language: A 'Content-First' Approach to Science Teaching. *Journal of Research in Science Teaching*, 45(5), 529–553. <https://doi.org/10.1002/tea.20255>
31. Parmaxi, A. (2023). Virtual Reality in Language Learning: A Systematic Review and Implications for Research and Practice. *Interactive Learning Environments*, 31(1), 172–184. <https://doi.org/10.1080/10494820.2020.1765392>
32. Griffin, C. S., et al. (2025). Virtual Laboratories Complement but Should Not Replace Face-to-Face Classes: Perceptions of Life Science Students at Dundalk Institute of Technology, Ireland. *Advances in Physiology Education*, 49(2), 314–330. <https://doi.org/10.1152/advan.00227.2024>
33. Santilli, T., Ceccacci, S., Mengoni, M., & Giaconi, C. (2025). Virtual vs. Traditional Learning in Higher Education: A Systematic Review of Comparative Studies. *Computers & Education*, 227, 105214. <https://doi.org/10.1016/j.compedu.2024.105214>
34. Stracke, C. M., et al. (2025). Immersive Virtual Reality in Higher Education: A Systematic Review of the Scientific Literature. *Virtual Reality*, 29, 64. <https://doi.org/10.1007/s10055-025-01136-x>
35. Sümer, M., & Vaněček, D. (2025). A Systematic Review of Virtual and Augmented Realities in Higher Education: Trends and Issues. *Innovations in Education and Teaching International*, 62(3), 811–822. <https://doi.org/10.1080/14703297.2024.2382854>
36. Faulconer, E. K., & Gruss, A. B. (2018). A Review to Weigh the Pros and Cons of Online, Remote, and Distance Science Laboratory Experiences. *International Review of Research in Open and Distributed Learning*, 19(2), 155–168. <https://doi.org/10.19173/irrodl.v19i2.3386>