Database, 2025, 1–10
DOI: https://doi.org/10.1093/database/baaf064
Perspective/Opinion

Integrated data-driven biotechnology research environments

Rosalia Moreddu [1,2,*]

[1]School of Electronics and Computer Science, University of Southampton, University Road, SO17 1BJ, Southampton, United Kingdom
[2]Institute for Life Sciences, University of Southampton, University Road, SO17 1BJ, Southampton, United Kingdom

*Corresponding author. School of Electronics and Computer Science, University of Southampton, University Road, SO17 1BJ, Southampton, UK.
E-mail: r.moreddu@soton.ac.uk

Citation details: Moreddu, R. Integrated data-driven biotechnology research environments. Database (2025) Vol. 2025: article ID baaf064; DOI: https://doi.org/10.1093/database/baaf064

Abstract

In the past few decades, the life sciences have experienced an unprecedented accumulation of data, ranging from genomic sequences and proteomic profiles to heavy-content imaging, clinical assays, and commercial biological products for research. Traditional static databases have been invaluable in providing standardized and structured information. However, they fall short when it comes to facilitating exploratory data interrogation, real-time query, multidimensional comparison, and dynamic visualization. Integrated data-driven research environments aiming at supporting user-driven data queries and visualization offer promising new avenues for making the best use of the vast and heterogeneous data streams collected in biological research. This article discusses the potential of interactive and integrated frameworks, highlighting the importance of implementing this model in biotechnology research, while going through the state-of-the-art in database design, technical choices behind modern data management systems, and emerging needs in multidisciplinary research. Special attention is given to data interrogation strategies, user interface design, and comparative analysis capabilities, along with challenges such as data standardization and scalability in data-heavy applications. Conceptual features for developing interactive data environments along diverse life science domains are then presented in the user case of cell line selection for *in vitro* research to bridge the gap between research data generation, actionable biological insight, experimental design, and clinical relevance.

Biological data management

Biology, nanotechnology, and medicine are data-rich fields [1]. Over the last several decades, high-throughput technologies have revolutionized biology by generating massive datasets. These include genomic sequences, proteomics data, high-resolution imaging, long-term acquisitions, and clinical trial data.[2] On top of those, companies in the biotech industry have commercialized large amounts of biological models to be used in research, biotechnology, and pharmaceutical industries for *in vitro* research [3]. In response, the need for versatile and user-friendly resource and data management systems has grown dramatically [4]. Biological databases traditionally focused on cataloguing discrete pieces of information and statically showing them online (e.g. Cellosaurus for classifying cell lines) [5] or within private organizations (e.g. internal databases for storing laboratory equipment information). In some cases, they integrate simple search functions to facilitate retrieval of stored data and allow incremental data submission or periodic expansion by database curators [4]. Classic examples, such as GenBank [6] and the Protein Data Bank (PDB), offer comprehensive search and retrieval functions across standardized metadata fields (e.g. organism, gene name, accession number, and sequence features in GenBank). Such systems remain indispensable as reference sources, but they were largely designed to support data deposition, retrieval, and preservation, rather than interactive exploration or adaptive reuse. Their architecture typically centers around rigid schemas with limited user-driven comparison capabilities.

With the advent of high-throughput technologies, the volume and complexity of biological data expanded considerably [7]. In domains such as genomics, drug discovery, *in vitro* research, and personalized medicine, interactive and integrated platforms have the potential to transform the way we work with data, reducing time currently devolved to hypothesis testing and literature search, and facilitating discovery by designing meaningful workflows based on experimental objectives. This model would also enable scientists to focus their efforts on innovation and higher-end intellectual activities. Table 1 presents the comparison between traditional biological databases and the proposed approach. Figure 1 visualizes the potential of interactive and integrated data environments. The following sections guide the development of next-generation digital research platforms.

The concept of interactivity in database systems has been widely used to describe data portals or repositories with web or Application Programming Interface (API) access. Here, this concept is expanded to include dynamic, modular systems designed for bidirectional interaction, collaborative filtering, hypothesis generation, experimental planning, and feedback integration. This allows to incorporate experimental metadata, support multiscale comparative analysis, and integrate FAIR Digital Objects derived from user interaction. Although no universally accepted term yet exists for such environments, this conceptual framework lies at the intersection of intelligent decision-support platforms, collaborative data infrastructures, and multi-domain experimental design engines.

Received 3 April 2025; Revised 7 July 2025; Accepted 20 August 2025

Table 1. Comparison between traditional biological databases and next-generation interactive data environments.

Dimension	Classical biological databases	Interactive data environments
Access model	Read-only, query-based	Bidirectional and real-time
Data structure	Schema-defined	Flexible (relational, document, and graph)
User engagement	Individual	Multi-user
Update frequency	Curated (low frequency)	Real-time ingestion and user feedback
Knowledge generation	Initiated by the user	Embedded in dynamic workflows
Feedback	Limited to curation	Real-time FAIR integration
Use case focus	Archival and citation	Discovery and planning

This table summarizes key architectural, functional, and epistemological differences between classical repositories, designed primarily for data storage and retrieval, and the proposed integrated systems, which emphasize bidirectional data flow, real-time analytics, collaborative workflows, and integration of experimental design logic and user feedback.

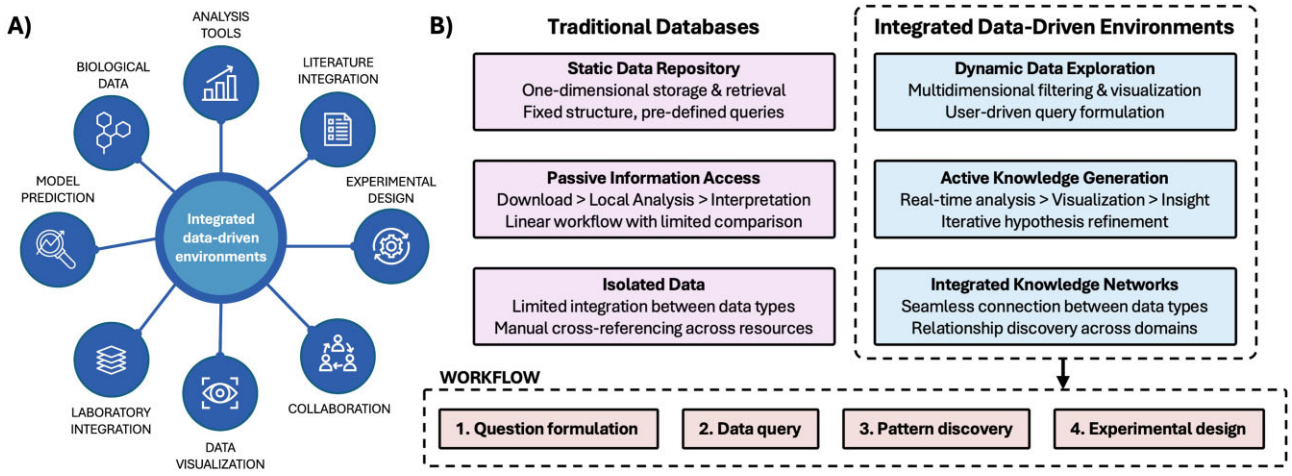

Figure 1. Integrated data environments. A) Overview of integrated data environments, needs, and features. B) Comparison between static repositories and interactive platforms across three key dimensions: data access methods, analysis workflows, and knowledge integration capabilities. The workflow at the bottom exemplifies the steps undertook by the user interfacing with an interactive data environment.

Interactive data environments

The concept of interactivity in database systems is closely related to developments in web technologies, artificial intelligence (AI), and data visualization techniques. Modern systems are capable to combine web technologies (e.g. JavaScript libraries for dynamic visualization) [8] and back-end data management solutions (e.g. NoSQL databases for unstructured data or graph databases for relationship modelling) [9]. Integrated research environments are not intended to replace domain-specific repositories such as GenBank, PDB, or Cellosaurus, which remain foundational for primary data submission and standardized archival. Instead, they could function as specialized integration and analysis layers that enable researchers to query, visualize, and analyse data across multiple existing repositories through unified interfaces. This architectural approach shares conceptual similarities with data lakes [10–12], yet extends beyond traditional data lake implementations in several crucial ways. Data lakes primarily provide infrastructure for storing heterogeneous data in native formats without imposing rigid schemas [13], whereas the research environments described here focus on active knowledge integration and experimental decision support through domain-specific analytical capabilities. Recent implementations such as genetic data lakes for drug discovery [10] store and process genetic data at scale, but typically lack the specialized comparative analysis capabilities and experimental design guidance that define the systems proposed hereby. General-purpose data lakes prioritize accommodating massive volumes of raw data [14], these biotechnology research environments implement domain-specific user interfaces and analytical workflows optimized for particular scientific tasks (e.g. cell line selection, pathway analysis, and biomarker identification). Furthermore, while data lakes typically operate as centralized repositories within organizational boundaries [15], the research environments envisioned here function as connection across the distributed ecosystem of existing biological repositories. They provide harmonized access layers that preserve the specialized governance and data submission workflows of underlying repositories while enabling cross-repository analyses not feasible through direct interaction with individual primary data sources.

Despite the dramatic development in computer science and web technologies, the life science domain still sees crucial gaps to enable smooth selection and dataset navigation [16, 17]. The need of transitioning towards these features is becoming evident through the growing complexity of biological questions, in parallel with the technological advances in other fields that make complex computations and visualizations feasible in real time and with less efforts from the user [18, 19]. The next subsections highlight selected desirable characteristics and their technological feasibility within interactive data environments for the life sciences domain. Cell line selection for *in vitro* research is presented as a possible implementation case.

Features

The heterogeneity of biological data, from structured clinical trial tables and semi-structured cell line annotations to unstructured experimental notes, poses a fundamental challenge for the development of integrated data environments [20]. Successfully integrating these different data types requires a strategic balance to ensure that the system accommodates evolving data landscapes without sacrificing analytical precision. At the core of this integration lies the concept of adaptive data modelling [21], where the choice of database schemes dictates both functionality and scalability. Relational models with a rigid table structure are indispensable for managing structured data such as genomic variants, patient demographics, and cell lines properties [22]. However, the dynamic nature of life sciences research sometimes demands schemeless architectures. In this context, document-oriented databases (e.g. MongoDB) [23] could be employed, allowing nested structures to capture variable data, for instance that associated with single-cell sequencing experiments [24]. For highly interconnected data, such as protein–protein interaction networks or metabolic pathways in cells, graph databases (e.g. Neo4j) offer the required traversal speed to enable real-time queries across millions of nodes and edges [25].

Data pipelines require automated workflows that analyse raw FASTQ files [26], screen online publications for experimental conditions, or obtain real-time sensor data from laboratory equipment [27, 28]. Tools with error-handling frameworks could standardize this process, e.g. Apache NiFi or custom Python scripts, but challenges exist [29]. For instance, inconsistencies in how labs report cell line contamination status require context-aware natural language processing models to normalize inputs [30, 31]. However, these models themselves may introduce noise or bias, especially when trained on incomplete or poorly annotated datasets. In addition to inconsistencies, metadata may be entirely missing or provided in minimal form, despite repository guidelines requesting rich contextual descriptors. Submitters may also inadvertently provide erroneous information due to lab tracking errors or manual entry mistakes. These factors further complicate data harmonization and highlight the need for robust validation mechanisms, error propagation awareness, and contributor-facing feedback loops within interactive platforms.

To bridge this gap, hybrid interfaces are gaining traction, e.g. Galaxy Project combining drag-and-drop workflows with Python scripting to allow users to transition from structured prompts (e.g. ‘show all breast cancer cell lines in the database having HER2 + status’) to programmatic analyses (e.g. suitability analysis for a given experiment evaluated using R libraries) [32]. Then, dynamic visualization could transform these raw query results into actionable insights. Modern systems integrate libraries to render interactive plots for genome-wide association studies, optimal cell profile to test a given technology, or 3D protein structures visualization. Examples of employed libraries are Plotly or D3.js [33]. However, interactivity implies that visualization must extend beyond passive observation. Interactive data environments could let users click on a graph or comparison plot to trigger a secondary query, e.g. extract all genes differentially expressed in a specific cluster, or group up all cells in different subclusters based on user-prompted features. Coupling visualization with analytical tools could enable this functionality, and assistive AI could amplify this interactivity and scope. All these functions drive the design of a suitable user interface and define user experience features.

Case study: biological cell line selection

Cell line selection represents an ideal case study for demonstrating the potential of interactive data environments in the life sciences. The complexity of choosing appropriate cell models from the thousands or commercially available lineages exemplifies why traditional static repositories are insufficient for modern research needs [34]. Currently, researchers often select cell lines based on convenience, tradition, or limited familiarity rather than comprehensive biological relevance, leading to potential experimental irreproducibility, translational failures, and wasted resources [34, 35]. The challenge lies in navigating multidimensional considerations simultaneously, spanning from genetic background, tissue origin, disease relevance, authentication status, growth characteristics, pathway activations, compatibility with experimental procedures, and more [34]. While valuable reference resources exist (Cellosaurus, ATCC catalogs, and LINCS) [36], they typically present information in isolation, making comparative analysis labour-intensive and prone to oversight of critical variables if run by humans.

An interactive data environment would transform cell line selection by enabling researchers to dynamically filter, compare, and visualize multiple cell lines across diverse parameters simultaneously. Such a system would integrate disparate data sources, including existing static repositories, literature outcomes, genomic profiles and user-contributed experimental metadata, creating a thoughtful support platform. The following subsections examine a potential envisioned architecture for implementation, structured into three interconnected layers: back-end data management systems for storing and processing diverse cell data, middleware and APIs facilitating integration and communication, and front-end technologies enabling intuitive exploration, comparison, and visualization. This is also schematized in Fig. 2.

Back-end data management systems

The back-end infrastructure forms the pillar of any interactive data environments [37, 38] where storage solutions are selected based on the inherent structure of biological information. Traditional relational database management systems, such as PostgreSQL and MySQL, build the foundations of many established biological repositories [39]. An example is Cellosaurus, a comprehensive repository of cell lines [5]. While invaluable as a reference, its traditional structure limits the utilization of the stored data. Currently, data can be visualized one by one for each cell line, making multidimensional comparisons across tens or hundreds of cells unfeasible. Document-oriented NoSQL databases offer significant advantages for cell line repositories that accumulate diverse experimental metadata [40]. MongoDB, for instance, can store cell lines as flexible JavaScript Object Notation (JSON) documents [23]. This allows to incorporate new characterization data of various nature, from morphological features to authentication profiles, without disruptive changes in the fundamental structure.

Network-oriented biological data presents another storage challenge that graph databases address. Systems such as

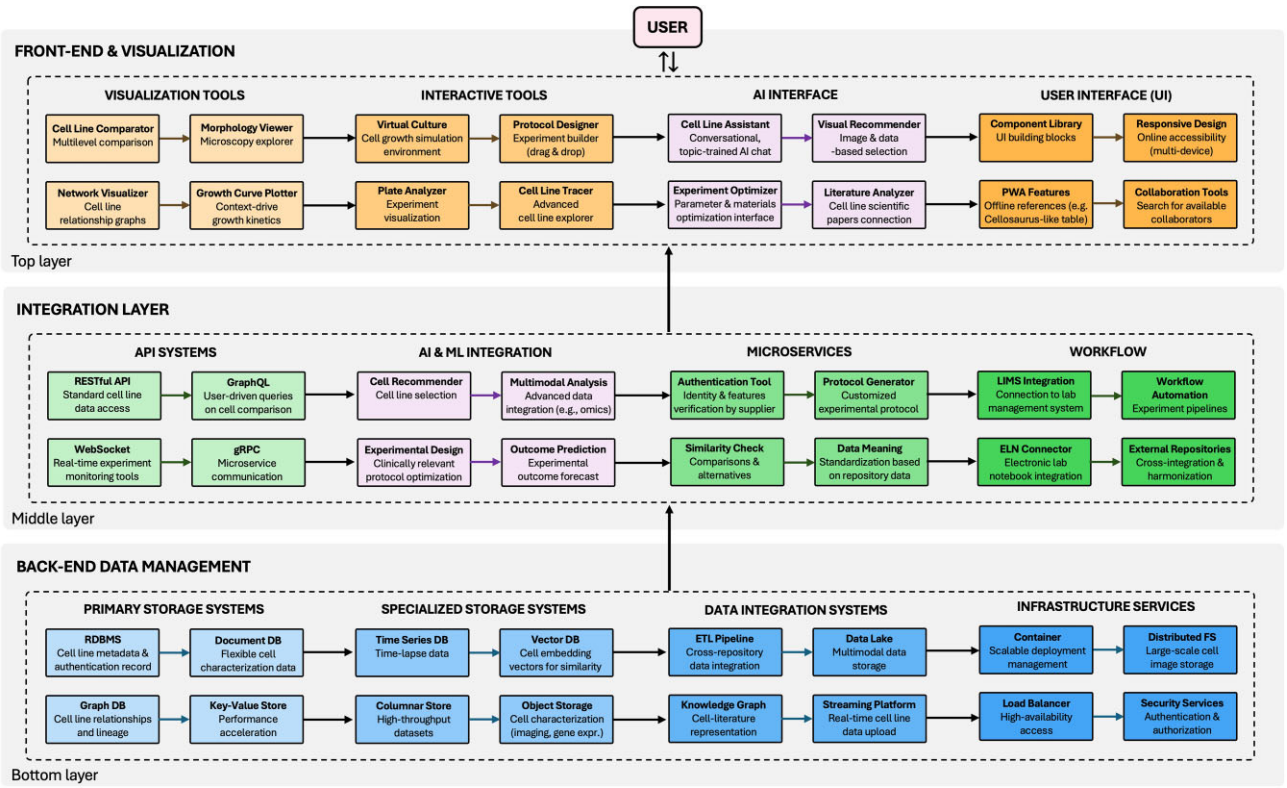

Figure 2. Technical architecture of integrated data environments: a case study on informed cell line selection. The schematic illustrates the three-tier structure comprising back-end data management systems (databases and storage solutions), middleware integration layer (APIs, microservices, and AI components), and front-end technologies (visualization tools and user interfaces), tailored to cell line selection as a sample case.

Neo4j could transform the way relationships between cell lines are accessed and understood [25]. In cancer research, one could explore the connections between patient-derived xenografts, immortalized cell lines and original tumour samples through intuitive graphs and user-driven multidimensional queries. This approach could reveal lineage relationships and experimental compatibility that remain obscured in conventional tabular repositories. Supplementary storage technologies could then drive performance considerations. Response times during comparative analyses could be dramatically reduced by employing key-value stores such as Redis to catch frequently accessed data, e.g. commonly requested cell lines or culture protocols. This hybrid storage approach would allow databases to maintain responsive performance even as users perform complex multiline comparisons simultaneously.

Middleware and APIs

The middleware layer in a database typically orchestrates communication between storage systems and user applications [41], in this case enabling to transform static cell line references into dynamic research tools. Instead of making separate requests for each cell line of interest, queries that directly compare multiple lines across selected parameters (growth kinetics, drug sensitivities, and genetic backgrounds) in a single operation could be constructed. Examples of tools to achieve this include GraphQL over traditional RESTful APIs [42]. AI represents one of the most powerful middleware integrations [43]. Machine learning microservices could analyse patterns across thousands of cell lines to recommend optimal models for specific research questions. Drug screening experiments could be informed by recommendation assistants that identifies cell lines most relevant to their target pathway based on expression profiles, previous experimental outcomes, and literature associations. Such systems transform passive cell line catalogues into active research planning tools, integrating them with the latest research findings. The latter could be, in turn, standardized over time by researchers themselves who engage with the interactive data environments.

The microservices architectural pattern can partition monolithic applications into independent, specialized components, enhancing system flexibility [44]. A modernized cell line database should separate authentication verification, experimental condition optimization, and cross-reference resolution into discrete services. When researchers upload new characterization data for a cell line, a validation microservice could automatically verify consistency with existing profiles, while another service updates recommended culture conditions based on combined experimental outcomes. Another key middleware component is workflow services, including laboratory information management systems [45], facilitating the link between information and action. An example is comparing metabolic profiles across hepatocyte cell lines to generate customized experimental protocols based on optimal culture conditions for each line, with reagent lists automatically adjusted for the specific metabolic properties of selected models.

Hybrid architecture and data integration framework

The proposed database architecture adopts a hybrid model combining a centralized repository for high-frequency primary data (e.g. cell line identifiers, validated traits) with federated integration of distributed secondary sources (e.g. genomic and literature databases) via standardized APIs. A

three-tiered mediation layer enables technical integration. The physical access layer could employ Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and bioinformatics standards such as Global Alliance for Genomics and Health (GA4GH) to standardize data harvesting. The middleware layer performs schema mapping for structured data and semantic normalization for unstructured content. At the top, a global identifier registry utilizes CURIEs and DOIs for cross-referencing across heterogeneous resources.

Three core interfaces are meant to ensure interoperability: (i) a unified query interface translating user requests into native database languages (SQL, Cypher, MongoDB); (ii) a synchronization interface maintaining consistency across relational, document, and graph stores; and (iii) a metadata exchange interface compliant with ISA-Tab for harmonizing biological experiment descriptions. These enable referential integrity across datasets and pathway graphs. A feedback integration pathway captures user interactions (e.g. annotations, experimental suitability tags) as FAIR Digital Objects with standardized metadata. A dedicated ingestion module validates and reintegrates these into the core database, creating a dynamic knowledge refinement loop. This approach might allow the system to evolve collaboratively, integrating expert input and computational insights to enhance both analytical functionality and data richness over time.

Front-end technologies

The front-end layer transforms static cell line catalogs into dynamic research platforms through intuitive interfaces for users [46]. This layer harmonically presents the features of the previous two layers through visualization and interactive features. Modern JavaScript frameworks enable sophisticated comparative visualization and interaction [46]. React components could render comparisons of cell lines, highlighting differences in morphology, growth characteristics, gene expression profiles, clinical relevance, usage, and much more, through interactive visualizations that respond instantly to parameter adjustments driven by the user. Visual comparison tools represent a simple yet potentially game-changing tool in cell line selection. Interactive matrices could display drug sensitivity patterns across multiple cell lines, with hierarchical clustering revealing unexpected relationships between seemingly unrelated models. This cross-check at multiple levels would likely produce precious novel insights based on fully exploiting and interpreting already-existing data. These comparisons could be filtered based on the most disparate parameters, needs or curiosities, based on specific mutations, tissue origins, or experimental conditions, transforming what would be weeks of literature review into minutes of interactive exploration [34].

AI assistants integrated directly into the front-end interface could provide tailored and specifically trained guidance through cell line selection processes based on experimental goals [47]. For example, they could highlight cell lines with relevant properties, flag potential authentication concerns, contradictory experimental findings from the literature, or suggest complementary models to strengthen experimental design. Finally, virtual cell culture simulators would be a transformative front-end addition to integrate mathematical models of cell behaviour with accumulated experimental data. This could facilitate the prediction around how different cell lines might respond to experimental manipulations before physical experiments begin, enabling to adjust culture conditions, test drug concentrations, or simulate time-course experiments through intuitive interfaces, with predictions based on historical data. To support collaborative data exploration and hypothesis generation, the front-end interface is designed to enable real-time multi-user interaction. This would enable to synchronously examine data views, annotate selections, and share insights within team environments. Examples of technologies are Firebase Realtime Database or Socket.IO, employed in collaborative sessions where selections and filters are mirrored across team members interfaces in real time. This setup might facilitate distributed but coordinated interpretation of complex biological datasets.

Selected applications of dynamic data software

The applications of interactive data environments span the entire spectrum of life sciences. In domains where relationships between entities are multidimensional and contextual, static tabular presentations fail to capture the complexity of biological systems. Interactive data environments are meant to integrate both structured and unstructured data. While structured data (genomic variants, protein structures, and clinical measurements) forms the foundation, unstructured data (scientific literature, clinical notes, and experimental protocols) provides crucial context at a given time. The following subsections examine four domains where interactive data environments are demonstrating and can further have particular impact: genomics, drug discovery, systems biology, and clinical research. Each case explores how the interactive paradigm can address domain-specific challenges and transforms research practices.

Genomics

The genomics field has been an early adopter of Interactive data environments approaches, driven by the complexity and volume of sequencing data which made this route inevitable [48]. Genome Aggregation Database (gnomAD) evolved from simple variant browsers to sophisticated interactive systems to explore allele frequencies across populations, visualize genomic contexts, and assess functional impacts of variants in real-time [48]. These capabilities have proven to be critical for rare disease research. Modern genomic interactive data environments integrate machine learning algorithms that predict variant pathogenicity while allowing users to adjust parameters based on domain knowledge [49]. For example, the ClinGen Pathogenicity Calculator enables clinicians to interactively apply American College of Medical Genetics and Genomics (ACMG) guidelines for variant classification while visualizing supporting evidence from multiple sources [49]. This represents a significant advance over static variant lists, enabling interpretation that adapts to evolving clinical knowledge.

Drug development

In pharmaceutical research, interactive data environments are revolutionizing multiple stages of the drug discovery and development pipeline. DrugBank systems allow to explore drug-target interactions across chemical and biological space [50]. Modern implementations combine structural databases with molecular docking algorithms, allowing users to interactively modify potential compounds and visualize predicted binding affinities in real-time. Virtual screening applications par-

ticularly benefit from interactive capabilities, where pharmacophore models or chemical similarity metrics can be adjusted to observe how these changes affect the ranking of potential hits. This interactivity significantly accelerates the iterative optimization process behind modern drug design. An example is Schrödinger LiveDesign integrating data from public and proprietary sources with interactive modelling tools that guide rational drug design while managing the complexity of structure–activity relationships [51].

Systems biology

Systems biology approaches benefit from interactive data environments that enable exploration of complex biological networks. Reactome and the Kyoto Encyclopedia of Genes and Genomes (KEGG) currently include interactive pathway browsers that enable to navigate from organism-level pathways down to molecular interactions, visualizing experimental data in context [52]. The ability to overlay multi-omics data (transcriptomics, proteomics, and metabolomics) onto these pathways in real-time provides insights that would be impossible to extract from static representations. Advanced systems biology databases incorporate simulation capabilities, where researchers can interactively perturb network components and observe predicted system-wide effects. For example, Cell Collective allows users to build and simulate logical models of biological networks, interactively testing hypotheses about regulatory relationships [53]. These interactive modelling approaches bridge static pathway maps and dynamic biological processes to facilitate *in silico* experimentation.

Clinical research

In clinical research, interactive data environments are transforming how patient cohorts are analysed and stratified [54]. Modern clinical trial databases allow to dynamically segment patient populations based on multiple clinical variables, biomarkers, treatment responses, and genomic profiles. These systems enable the identification of responder subgroups that might be missed in traditional aggregate analyses. This interactive approach is particularly valuable for precision medicine, where treatment decisions increasingly depend on complex combinations of biomarkers. In this context, cBioPortal for Cancer Genomics is used by clinicians to interactively explore relationships between genomic alterations and clinical outcomes across thousands of patients, and identify patterns that inform treatment selection for individual cases [54]. As these systems evolve, they increasingly incorporate natural language processing of clinical notes and AI-assisted pattern recognition to extract insights from unstructured clinical data.

The applications discussed across genomics, drug development, systems biology, and clinical research demonstrate different approaches to biological data interactivity. To better illustrate the current limitations and development gaps, Table 2 provides a systematic comparison of representative platforms in each domain, highlighting both implemented capabilities and features requiring further development.

Despite strong domain-specific performance, major limitations persist in cross-domain data integration. Systems must evolve to support standardized linking of genomic variants, pathway disruptions, drug targets, and clinical outcomes into unified, interactive analyses. In parallel, most platforms remain tailored for single-user interactions, lacking collaborative functions such as version control, permission management, or synchronized multi-user workspaces. Another limitation lies in the unidirectional flow of information. Current tools primarily serve as endpoints for querying existing knowledge rather than facilitating knowledge generation. Embedding structured annotation frameworks could enable users to contribute validated insights, fostering dynamic feedback loops between researchers and databases. Moreover, computational scalability remains a barrier: simple filtering and visualization are often responsive, but complex analyses are hindered by performance constraints. To overcome this, platforms will need distributed computing architectures optimized for biological data types. Finally, most systems offer only retrospective data exploration; integrating predictive modelling and simulation would enable hypothesis testing, allowing users to evaluate potential interventions prior to experimentation, significantly accelerating and improving research outcomes.

Discussion and challenges

Despite their transformative potential, interactive data environments in life sciences face substantial challenges. The most intuitive challenges span from unawareness among people, privacy and access, user experience, and intrinsically data-related challenges. People-related challenges revolve around the lack of awareness among part of life science researchers of what computing tools can offer to optimize, speed up, and improve the intellectual quality of life science research. This challenge reflects poor communication and limited exchange between life sciences and computational disciplines, which urgently need to be bridged. On the same line, user adoption represents a critical challenge. Interactive systems must accommodate diverse user groups with varying computational literacy while providing sufficient analytical depth to address complex biological questions.

Technical challenges are primarily about data standardization, performance, data quality, and accessibility. Data quality itself represents a central bottleneck, revolving around noisy or incomplete entries, inconsistent measurement protocols, and experimental bias, which can critically impair the interpretability and reproducibility of downstream analyses. Sophisticated analyses that provide meaningful biological insights often require computational resources incompatible with real-time interaction. This creates a challenging design space where analytical depth must be balanced against performance constraints. Privacy and access challenges are particularly relevant in clinical and patient-derived datasets, involving ethical concerns, consent frameworks, and jurisdictional restrictions. For example, the use and sharing of patient-level data must adhere to data protection regulations (e.g. GDPR and HIPAA), institutional review protocols, and evolving expectations around participant autonomy and trust. These constraints are essential for safeguarding rights and ethics, yet often introduce friction in integrating sensitive data across platforms. Initiatives such as the GA4GH are working to establish interoperable standards and policies to facilitate responsible data sharing while preserving privacy. Integrated data platforms intended to operate in this domain must be designed with embedded compliance layers and customizable permission systems.

Table 2. Comparison of existing integrated data environments across major medical biotechnology domains.

Category	Feature	Genomics (gnomAD)	Drug development (DrugBank)	Systems biology (reactome)	Clinical research (cBioPortal)
Data structure and storage	Multi-omics data	Variant-focused with limited transcriptomic data	Chemical-biological integration	Pathway-focused with limited multi-omics	Multi-omics with clinical data
	Unstructured data	Limited text mining	Chemical literature	Protocol documentation	Clinical notes processing
	Temporal data support	Static datasets	Limited reaction kinetics	Dynamic pathway simulation	Longitudinal patient data
Query and visualization	Real-time filtering of $>10^6$ records	Population-scale variant filtering	Limited to subsets of compound database	Pathway-limited scope	Sample-limited cohort selection
	Interactive cross-domain queries	Limited to genomic context	Drug-target-pathway connections	Within pathway boundaries	Genotype-phenotype correlations
	3D/spatial data visualization	Limited protein structure	Molecular structure viewers	Network topology	Limited anatomical context
Analytics and interactivity	Interactive statistical analysis	Population frequency tools	Structure-activity analysis	Enrichment analysis	Survival and correlation analysis
	Hypothesis testing framework	Limited to constraint metrics	Virtual screening	Pathway perturbation	Biomarker association testing
	Machine learning integration	Variant pathogenicity	Chemical similarity	Limited predictive models	Outcome prediction
Collaborative features	Multi-user simultaneous interaction	Single-user model	Single-user model	Single-user model	Limited sharing capabilities
	Version control of analyses	Download-only	Limited project saving	Export options	Study groups
	User annotation frameworks	None	Limited annotations	None	Basic study descriptions
Technical architecture	API extensibility	Comprehensive REST API	Basic REST endpoints	Limited API access	Comprehensive programmatic access
	Computational scalability	Cloud-based distributed computing	Limited analytical capacity	Server-based processing	Hybrid architecture
	FAIR data principles implementation	Partial implementation	Partial implementation	Partial implementation	Partial implementation

The table summarizes implemented features and highlights missing capabilities across representative platforms in genomics, drug development, systems biology, and clinical research, in relation to the proposed interactive model.

Data standardization

Biological data are produced by a wide variety of instruments and experimental methods, which often results in heterogeneous formats and varied quality. Standardizing data formats and ensuring data quality are fundamental challenges. Life science domains have developed specialized vocabularies that often overlap but use different terminologies for similar concepts. For example, cell lines may be described using inconsistent nomenclature across repositories (HeLa vs. HeLa S3 vs. Hela-S3), and the similarity across names could lead to misassignments. These misassignments, even if rare, could cause cascade problems. Ontology mapping has been already initiated, e.g. through OBO Foundry (Open Biological and Biomedical Ontologies) providing frameworks for ontology integration, but implementation remains challenging due to the evolving nature of biological knowledge [55]. Natural language processing models are increasingly employed to automatically map terms across vocabularies, but these systems require careful curation to validate mappings. The value of interactive queries depends fundamentally on the quality and completeness of underlying metadata, for instance, experimental details necessary for proper interpretation. To address this challenge, interactive data environments could implement validation systems that flag missing critical data and provide feedback to contributors about data quality. Some systems now employ data provenance indicators or reputation scores for data sources, allowing users to filter query results based on source trustworthiness and metadata completeness. Examples are BioThings Explorer and FAIRsharing [56, 57].

Performance

As databases grow in size and complexity, ensuring that query responses remain fast and accurate is crucial. Interactive medical biotechnology queries involve complex multidimensional parameters. For example, identification of cell lines with specific genetic mutations, protein expression patterns, and growth characteristics, alongside dynamic visualization of multiscale relationships across tens of cell types and culture. Advanced computational approaches addressing this challenge include bitmap indexing for genomic data and spatial indexing techniques adapted for multidimensional biological data [58].

Cloud-native database architectures that scale horizontally to handle compute-intensive queries are increasingly essential for interactive performance across large biological datasets. Interactive visualization of large biological datasets presents unique performance challenges. Modern interactive data environments address this through server-side aggregation, progressive loading techniques that refine visualizations incrementally, WebGL and GPU-accelerated rendering, and intel-

ligent sampling methods that preserve statistical properties while reducing data volume [59]. These techniques enable responsive exploration even for datasets too large to transmit in their entirety. This transition can be enabled step by step, handling datasets that are easier to manage first.

Data accessibility

Advanced interactivity can only be effective if users find the system intuitive and accessible. Interactive biological databases face a fundamental issue between analytical power and interface simplicity. Systems that expose the full complexity of underlying data models risk overwhelming non-computational users, while oversimplified interfaces may limit discovery potential. This challenge is particularly acute in multidisciplinary fields where users range from computational specialists to wet-lab biologists and clinicians. Adaptive interface approaches show promise in addressing this challenge through progressive disclosure of features based on user expertise, context-sensitive guidance, customizable workspaces, and natural language query capabilities for nontechnical users. Features such as automatic query history tracking (e.g. as implemented in National Center for Biotechnology Information (NCBI) resources) and computational notebooks (e.g. Jupyter) are now widely adopted tools that have proven effective in enhancing reproducibility, transparency, and user engagement when integrated into interactive database interfaces. Another issue related to data accessibility concerns the integration and retrieval of information from existing biological databases, which is often hindered by inconsistent formats, limited APIs, or restricted access policies.

Outlook

interactive data environments could represent a paradigm shift in how research data is stored, handle, and shared, offering significant advantages towards driving collective scientific progress meant for clinical translation and reliable fundamental results. Their emergence signals not merely a technological evolution but a fundamental shift in how biological knowledge is constructed, validated, and extended. This reconceptualizes the scientific process itself where the boundaries between hypothesis generation, data analysis, and experimental design become increasingly iterative, with a strong urge for data reproducibility and validation. Currently, different expertise in computational methods creates a gap between those who can and cannot effectively interrogate complex biological datasets. Integrated frameworks designed with intuitive interfaces could democratize access to advanced analytical capabilities, potentially shifting control over data interpretation and exploratory analysis from computational specialists to a broader range of scientists. The development trajectory of interactive data environments will inevitably be shaped by economic forces and institutional priorities that extend beyond purely scientific considerations. Commercial entities building such platforms face tensions between creating proprietary systems that generate revenue and contributing to open scientific models that maximize knowledge generation and sharing. These economic realities suggest that hybrid models combining open source models with commercial components may emerge.

Interactive data environments also have the potential to transform interdisciplinary collaboration by creating shared cognitive spaces where specialists from diverse backgrounds can explore complex biological questions. This potential extends beyond collaboration among human experts to include the integration of AI-driven tools that support data exploration and interpretation, while preserving human oversight and decision-making. Static repositories indirectly reinforce reductionist perspectives by presenting biological entities as discrete objects. Interactive systems that dynamically visualize multidimensional relationships could instead represent the existing interconnections between the most seemingly disparate domains, making full use of the acquired data across domains.

Conflicts of interest: None declared.

References

1. Martani A, Geneviève LD, Elger B *et al.* 'It's not something you can take in your hands'. Swiss experts' perspectives on health data ownership: an interview-based study. *BMJ Open* 2021;**11**:e045717. https://doi.org/10.1136/bmjopen-2020-045717
2. Perez-Riverol Y, Alpi E, Wang R *et al.* Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 2015;**15**:930–50. https://doi.org/10.1002/pmic.201400302
3. Seth A, Banyal A, Kumar P. Commercialization and technology transfers of bioprocess. In: Bhatt AK, Bhatia RK, Bhalla TC (eds), *Basic Biotechniques for Bioprocess and Bioentrepreneurship*. Academic Press, 2023, 455–69.
4. Jagadish HV, Olken F. Database management for life sciences research. *ACM SIGMOD Record* 2004;**33**:15–20. https://doi.org/10.1145/1024694.1024697
5. Bairoch A. The cellosaurus, a cell-line knowledge resource. *J Biomol Tech* 2018;**29**:25–38. https://doi.org/10.7171/jbt.18-2902-002
6. Clark K, Karsch-Mizrachi I, Lipman DJ *et al. Nucleic Acids Res* 2016;**44**:D67–72. https://doi.org/10.1093/nar/gkv1276
7. Thessen AE, Patterson DJ. Data issues in the life sciences. *ZooKeys* 2011;**150**:15–51. https://doi.org/10.3897/zookeys.150.1766
8. Jeong Y, Young L, Hicks D. Synchronized static and dynamic visualization in a web-based programming environment. In: *2016 IEEE 24th International Conference on Program Comprehension (ICPC)*. Springer, Cham. 2016. https://doi.org/10.1109/ICPC.2016.7503733
9. Meier A, Kaufmann M. NoSQL Databases. In: *SQL & NoSQL Databases*. Cham: Springer, 2019, 201–18. https://doi.org/10.1007/978-3-031-27908-9_7
10. Chatelain C, Lessard S, Klinger K *et al.* Building a human genetic data lake to scale up insights for drug discovery. *Drug Discov Today* 2025;**30**:104385. https://doi.org/10.1016/j.drudis.2025.104385
11. Wolski M, Woloszynski T, Stachowiak G *et al.* Bone data lake: a storage platform for bone texture analysis. *Proc Inst Mech Eng Part H J Eng Med* 2025;**239**:190–201. https://doi.org/10.1177/09544119251318434
12. Schneider M, Zolg DP, Samaras P *et al.* A scalable, web-based platform for proteomics data processing, result storage and analysis. *J Proteome Res* 2025;**24**:1241–49. https://doi.org/10.1021/acs.jproteome.4c00871
13. Fang H. Managing data lakes in big data era: what's a data lake and why has it became popular in data management ecosystem. In: *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. 2015.
14. Miloslavskaya N, Tolstoy A. Big data, fast data and data lake concepts. *Proc Comp Sci* 2016;**88**:300–305. https://doi.org/10.1016/j.procs.2016.07.439

15. Hai R, Geisler S, Quix C. Constance. In: *Proceedings of the 2016 International Conference on Management of Data*. 2016. https://doi.org/10.1145/2882903.2899389
16. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: errors, underlying themes, and potential solutions. *Proc Natl Acad Sci* 2018;**115**:2563–70. https://doi.org/10.1073/pnas.1708279115
17. Anderson NR, Lee ES, Brockenbrough JS *et al*. Issues in biomedical research data management and analysis: needs and barriers. *J Am Med Inform Assoc* 2007;**14**:478–88. https://doi.org/10.1197/jamia.M2114
18. Sheikh A, Anderson M, Albala S *et al*. Health information technology and digital innovation for national learning health and care systems. *Lancet Digital Health* 2021;**3**:e383–96. https://doi.org/10.1016/S2589-7500(21)00005-4
19. Szymkowiak A, Melović B, Dabić M *et al*. Information technology and Gen Z: the role of teachers, the internet, and technology in the education of young people. *Technol Soc* 2021;**65**:101565. https://doi.org/10.1016/j.techsoc.2021.101565
20. Birkland A, Yona G. BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinf* 2006;**7**:70. https://doi.org/10.1186/1471-2105-7-70
21. Joe Qin S. Recursive PLS algorithms for adaptive data modeling. *Comput Chem Eng* 1998;**22**:503–14. https://doi.org/10.1016/s0098-1354(97)00262-7
22. Biba M, Vajjhala NR. Statistical relational learning for genomics applications: a state-of-the-art review. In: Roy SS, Taguchi YH(eds), *Handbook of Machine Learning Applications for Genomics*, *Studies in Big Data*. vol. 103, Singapore: Springer, 2022, 31–42. https://doi.org/10.1007/978-981-16-9158-4_3
23. Gyorodi C, Gyorodi R, Pecherle G *et al*. A comparative study: mongoDB vs. MySQL. In: *2015 13th International Conference on Engineering of Modern Electric Systems (EMES)*. 2015. https://doi.org/10.1109/EMES.2015.7158433
24. Schadt EE, Linderman MD, Sorenson J *et al*. Computational solutions to large-scale data management and analysis. *Nat Rev Genet* 2010;**11**:647–57. https://doi.org/10.1038/nrg2857
25. Webber J. A programmatic introduction to Neo4j. In: *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*. 2012.
26. Frampton M, Houlston R. Generation of artificial FASTQ files to evaluate the performance of next-generation sequencing pipelines. *PLoS One* 2012;**7**:e49110. https://doi.org/10.1371/journal.pone.0049110
27. Brazdil P, van Rijn JN, Soares C *et al*. Automating workflow/pipeline design. In: *Metalearning*. Cognitive Technologies. Cham: Springer, 2022, 123–40. https://doi.org/10.1007/978-3-030-67024-5_7
28. Spjuth O, Bongcam-Rudloff E, Hernandez GC *et al*. Experiences with workflows for automating data-intensive bioinformatics. *Biol Direct* 2015;**10**:43. https://doi.org/10.1186/s13062-015-0071-8
29. Bindal PM,K. Quantum flow: enterprise data orchestration and processing suite. In: *IC3-2024: Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing*. 2024, 577–84. https://doi.org/10.1145/3675888.3676116
30. Horbach S, Halffman W. The ghosts of HeLa: how cell line misidentification contaminates the scientific literature. *PLoS One* 2017;**12**:e0186281. https://doi.org/10.1371/journal.pone.0186281
31. Capes-Davis A, Theodosopoulos G, Atkin I *et al*. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer* 2010;**127**:1–8. https://doi.org/10.1002/ijc.25242
32. Jalili V, Afgan E, Gu Q *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res* 2020;**48**:W395–402. https://doi.org/10.1093/nar/gkaa434
33. Sievert C. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. New York: Chapman and Hall/CRC, 2020. https://doi.org/10.1201/9780429447273
34. Dias D, Jones CF, Moreira AC, *et al*. Multidimensional classification framework for human breast cancer cell lines. arXiv:2502.15868, 2025.
35. Holliday DL, Speirs V. Choosing the right cell line for breast cancer research. *Breast Cancer Res* 2011;**13**:215. https://doi.org/10.1186/bcr2889
36. Laizé V, Rosa JT, Tarasco M *et al*. Status, challenges, and perspectives of fish cell culture—focus on cell lines capable of in vitro mineralization. In: Monzón IF, Fernandes JMO(eds), *Cellular and Molecular Approaches in Fish Biology*. Academic Press, 2022, 381–404. https://doi.org/10.1016/B978-0-12-822273-7.00004-5
37. DiFranzo D, Graves A, Erickson JS *et al*. The web is my back-end: creating mashups with linked open government data. In: Wood D(ed.), *Linking Government Data*. New York, NY: Springer, 2011, 205–19. https://doi.org/10.1007/978-1-4614-1767-5_10
38. Drucker J. The back end: infrastructure design for scholarly research. *J Mod Period Stud* 2017;**8**:119–33. https://doi.org/10.5325/jmodeperistud.8.2.0119
39. Dall'Alba G, Casa PL, Abreu FP *et al*. A survey of biological data in a big data perspective. *Big Data* 2022;**10**:279–97. https://doi.org/10.1089/big.2020.0383
40. Kumar KBS, Srividya, Mohanavalli S. A performance comparison of document oriented NoSQL databases. In: *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*. 2017. https://doi.org/10.1109/ICCCSP.2017.7944071
41. Patiño-Martinez M, Jiménez-Peris R, Kemme B *et al*. Middle-R. *ACM Trans Comput Syst* 2005;**23**:375–423. https://doi.org/10.1145/1113574.1113576
42. Burley SK, Berman HM, Kleywegt GJ *et al*. Protein Data Bank (PDB): the single global macromolecular structure archive. *Methods Mol Biol* 2017;**1607**:627–41. https://doi.org/10.1007/978-1-4939-7000-1_26
43. Wang J, Lu T, Li L *et al*. Enhancing personalized search with AI: a hybrid approach integrating deep learning and cloud computing. *J Adv Comput Syst* 2024;**4**:1–13. https://doi.org/10.69987/jacs.2024.41001
44. Marquez G, Astudillo H. Actual use of architectural patterns in microservices-based open source projects. In: *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*. 2018. https://doi.org/10.1109/APSEC.2018.00017
45. Prasad PJ, Bodhe GL. Trends in laboratory information management system. *Chemom Intell Lab Syst* 2012;**118**:187–92. https://doi.org/10.1016/j.chemolab.2012.07.001
46. Goh HA, Ho CK, Abas FS. Front-end deep learning web apps development and deployment: a review. *Appl Intell* 2023;**53**:15923–45. https://doi.org/10.1007/s10489-022-04278-6
47. Harrison Oke E, Regina Coelis K. Adebamigbe Alex, F. The future of software development: integrating AI and machine learning into front-end technologies. *Glob J Adv Res Rev* 2024;**2**:069–77. https://doi.org/10.58175/gjarr.2024.2.1.0031
48. Gudmundsson S, Singer-Berk M, Watts NA *et al*. Variant interpretation using population databases: lessons from gnomAD. *Hum Mutat* 2022;**43**:1012–30. https://doi.org/10.1002/humu.24309
49. Patel RY, Shah N, Jackson AR *et al*. ClinGen pathogenicity calculator: a configurable system for assessing pathogenicity of genetic variants. *Genome Med* 2017;**9**:3. https://doi.org/10.1186/s13073-016-0391-z
50. Knox C, Wilson M, Klinger CM *et al*. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res* 2024;**52**:D1265–75. https://doi.org/10.1093/nar/gkad976
51. May M. Computational tools take advantage of the data deluge. *Genet Eng Biotechnol News* 2023;**43**:42–44. https://doi.org/10.1089/gen.43.04.14
52. Nguyen H, Pham VD, Nguyen H *et al*. CCPA: cloud-based, self-learning modules for consensus pathway analysis using GO, KEGG and Reactome. *Brief Bioinf* 2024;**25**. https://doi.org/10.1093/bib/bbae222

53. Helikar T, Kowal B, McClenathan S *et al.* The cell collective: toward an open and collaborative approach to systems biology. *BMC Syst Biol* 2012;**6**:96. https://doi.org/10.1186/1752-0509-6-96
54. Gao J, Aksoy BA, Dogrusoz U *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1. https://doi.org/10.1126/scisignal.2004088
55. Smith B, Ashburner M, Rosse C *et al.* The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–55. https://doi.org/10.1038/nbt1346
56. Holik AZ, Law CW, Liu R *et al.* RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Res* 2017;**45**:e30. https://doi.org/10.1093/nar/gkw1063
57. Sansone SA, McQuilton P, Rocca-Serra P *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol* 2019;**37**:358–67. https://doi.org/10.1038/s41587-019-0080-8
58. Jong VL, Novianti PW, Roes KC *et al.* Selecting a classification function for class prediction with gene expression data. *Bioinformatics* 2016;**32**:1814–22. https://doi.org/10.1093/bioinformatics/btw034
59. Yu G, Liu C, Fang T *et al.* A survey of real-time rendering on Web3D application. *Virtual Real Intell Hardw* 2023;**5**:379–94. https://doi.org/10.1016/j.vrih.2022.04.002

Received 3 April 2025; Revised 7 July 2025; Accepted 20 August 2025