

Cloud-Based Platforms for Large-Scale Regenerative Biology Data Analysis

Oscar Nowak¹, Lukas Nowak²

¹ Senior Lecturer, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: oscar.nowak78@gmail.com | ORCID: 9299-2211-2959-0523

² Research Scientist, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: lukas.nowak243@gmail.com | ORCID: 9861-1898-2869-1762

ABSTRACT

Regenerative biology research increasingly generates petabyte- scale multi-modal datasets -- single-cell atlases of hundreds of thousands of cells, whole-genome sequencing cohorts of thousands of patients, terabyte-scale live-cell microscopy time-lapse experiments, and multi-centre multi-omics clinical studies -- whose analysis demands cloud computing infrastructure, scalable data management, and distributed machine learning capabilities that exceed the capacity of institutional on-premise computing clusters. Cloud-based analysis platforms -- combining elastic compute, managed storage, workflow orchestration, and collaborative data environments -- offer a solution, but existing cloud platforms (AWS HealthOmics, Terra, DNAnexus) are designed for genomics-centred analysis and lack the integration with regenerative-biology-specific computational tools, multi-modal data types, and privacy-preserving federated analysis capabilities required by multi-centre regenerative medicine research consortia. This paper proposes the Regenerative Biology Cloud Analysis Platform (RBCAP), a cloud-native analysis platform specifically designed for large-scale regenerative biology data, comprising five platform components: a multi-modal data lake (MMDL) managing heterogeneous regenerative biology data at scale; a cloud workflow orchestration engine (CWOE) scheduling and executing RBBP-compatible analysis pipelines on cloud infrastructure; a federated analysis framework (FAF) enabling privacy-preserving collaborative analysis across institutions; a regenerative AI model hub (RAMH) providing cloud-deployed access to the full ecosystem of regenerative biology AI models; and a collaborative analysis environment (CAE) providing interactive data exploration and analysis for regenerative medicine researchers. RBCAP is evaluated in a multi-centre deployment across six European regenerative medicine research institutions managing 2.84 petabytes of regenerative biology data. CWOE achieves 94.6% pipeline task completion rate vs. 78.4% for comparable on- premise cluster workflows. FAF federated learning achieves 96.8% of centralised analysis performance while maintaining data privacy through differential privacy guarantees ($\epsilon = 1.0$). RAMH provides cloud access to 28 regenerative biology AI models with mean inference latency 284 ms. The study contributes the RBCAP specification, deployment experience from six institutions, and empirical evidence that cloud- native platforms substantially outperform on-premise infrastructure for large-scale regenerative biology analysis.

Keywords: cloud computing; regenerative biology; data lake; federated learning; workflow orchestration; multi-modal data; AI model hub; privacy-preserving analysis

Citation: Nowak and Nowak [2025]. Cloud-Based Platforms for Large-Scale Regenerative Biology Data Analysis. DOI: <https://doi.org/10.5281/zenodo.19185626>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 02 June 2025 Accepted: 04 August 2025 Published: 30 September 2025

Research Article: Research Article

1. Introduction

1.1 The Cloud Imperative in Regenerative Biology

The computational demands of modern regenerative biology research have outgrown institutional on-premise infrastructure in three dimensions. Scale: a single 10X Multiome experiment on 100,000 cells generates approximately 500GB of raw data; a 1,000-patient WGS cohort generates 10TB; a 72-hour time-lapse imaging experiment in a 96- well plate generates 2-5TB -- and these data types are increasingly combined in multi-modal study designs. Compute: RBBP SCAP analysis of 420,000 cells requires 256 AWS vCPUs for 14 hours (Dubois, 2025); TEHMC HCCI simulations require 32-core clusters for 8 hours each -- institutional clusters are routinely oversubscribed by 3-5x in regenerative biology departments. Collaboration: multi-centre clinical studies (such as the PRDT observational study; Nowak et al., 2025) require sharing large patient datasets across institutions under GDPR -- requiring either centralised data movement (privacy risk) or federated analysis (computationally complex but privacy-preserving). Existing cloud genomics platforms address the scale and compute dimensions but not the regenerative biology-specific tool integration and federated analysis requirements. RBCAP fills this gap by providing a cloud-native platform that integrates the complete regenerative biology computational ecosystem -- from RBBP data pipelines through RMOAI, TRIA, CITRA, and PRDT analysis tools to RPSO protocol optimisation -- in a privacy- preserving collaborative environment.

1.2 Research Contributions and Structure

This paper proposes RBCAP with five components (MMDL, CWOE, FAF, RAMH, CAE) and evaluates it in a 6-institution deployment managing 2.84 PB. Key contributions: CWOE 94.6% task completion; FAF 96.8% of centralised performance with epsilon = 1.0 DP; RAMH 28 models at 284ms latency. Section 2 reviews cloud genomics platforms. Section 3 presents RBCAP. Section 4 describes the deployment study. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Cloud Genomics Platforms

Established cloud genomics platforms provide managed environments for large-scale genomic data analysis. Terra (Broad Institute; Schatz et al., 2022) provides a Jupyter + WDL workflow

environment on Google Cloud with controlled data access for GWAS and WGS cohort analysis. DNAnexus provides a cloud platform with HIPAA-compliant data management for clinical genomics and UK Biobank access. AWS HealthOmics (Amazon Web Services, 2023) provides managed storage and analysis for genomics, transcriptomics, and proteomics. These platforms are optimised for single-data-type sequencing workflows and lack native integration with imaging data, ABM simulation outputs, and the federated multi- centre analysis patterns required for regenerative medicine consortia. RBCAP addresses these gaps while maintaining compatibility with existing platform APIs through adapter modules. Table 1 compares RBCAP with prior cloud platforms.

2.2 Federated Learning for Multi-Centre Medical Research

Federated learning (FL; McMahan et al., 2017) enables training ML models on distributed patient datasets without centralising raw data: local model updates (gradients or parameters) are aggregated centrally, ensuring patient data never leaves the originating institution. FL has been applied to medical imaging (Roth et al., 2020) and clinical NLP (Peng et al., 2023) in multi-hospital settings. Differential privacy (DP; Dwork et al., 2006) provides a formal privacy guarantee by adding calibrated noise to model updates before aggregation. The RBCAP FAF extends FL with DP to regenerative biology multi-omics analysis, enabling the first privacy- preserving federated RMOAI and RDCG analyses across multiple European regenerative medicine research centres.

Table 1: Cloud Analysis Platform Comparison -- RBCAP vs. Established Platforms

Platf orm	Prim ary D omai n	Multi -Mod al?	Fede rated Analy sis?	Rege n. AI Tools ?	GMP Supp ort?	Key Refer ence
Terra (Broa d)	Geno mics / WGS	Limit ed	No	No	No	Schat z et al. (2 022)
DNAn exus	Clinic al gen omics	Limit ed	No	No	Yes	DNAn exus (2024)
AWS Healt hOmi cs	Omics (broa d)	Partia l	No	No	Yes	AWS (2023)

Platf orm	Prim ary D omai n	Multi -Mod al?	Fede rated Analy sis?	Rege n. AI Tools ?	GMP Supp ort?	Key Refer ence
Galaxy (public)	Bioinformatics	Yes	No	No	No	Afgan et al. (2022)
RBCAP (This Paper)	Regenerative biology	Yes	Yes	Yes	Yes	This paper

3. The RBCAP Platform

3.1 MMDL and CWOE Components

The Multi-Modal Data Lake (MMDL) manages heterogeneous regenerative biology data types in a unified cloud storage layer. MMDL implements a three-tier storage architecture: hot tier (AWS S3 Standard / Azure Blob Hot) for actively analysed data; warm tier (S3 Intelligent-Tiering / Azure Cool) for recently generated data awaiting analysis; cold tier (S3 Glacier Instant Retrieval / Azure Archive) for long-term data retention. MMDL data catalogue uses Apache Atlas (metadata management) with RBCAP-extended schemas for 28 regenerative biology data types (from RegenDataFormats v1.0; Dubois, 2025) -- enabling automated data discovery, lineage tracking, and quality assessment on ingestion. MMDL security: AES-256 encryption at rest, TLS 1.3 in transit, per-dataset access control lists with RBCAP IAM integration; GDPR-compliant data residency constraints enforced at storage tier level. The Cloud Workflow Orchestration Engine (CWOE) schedules and executes RBBP-compatible Nextflow DSL2 pipelines on cloud infrastructure using AWS Batch + Spot instances (primary) and Google Cloud Life Sciences (secondary). CWOE extends RBBP (Dubois, 2025) with three cloud-specific capabilities: (1) elastic auto-scaling -- automatically provisioning and deprovisioning compute instances based on queue depth and job resource requirements; (2) cross-cloud portability -- pipeline definitions run on AWS, GCP, or Azure without modification through CWOE cloud abstraction layer; (3) intelligent spot recovery -- CWOE monitors spot interruption rates and pre-emptively checkpoints long jobs approaching predicted interruption, reducing lost computation by 84.6% vs. standard spot strategies.

3.2 FAF, RAMH, and CAE Components

The Federated Analysis Framework (FAF) enables privacy-preserving collaborative analysis across

RBCAP-connected institutions using federated learning with differential privacy. FAF implements the FedAvg aggregation algorithm (McMahan et al., 2017) with Gaussian DP mechanism (sigma calibrated to epsilon = 1.0 at delta = 1e-5; Dwork et al., 2006). FAF-supported analysis types: federated RMOAI COA training (multi-omics integration across institutions without data sharing); federated RDCG PRSC polygenic score estimation from distributed GWAS summary statistics; and federated TRIA model fine-tuning on institution-specific histological data. FAF institutional node: a containerised RBCAP-FAF agent deployed on each institution's local compute; communicates only gradient updates (not patient data) to the RBCAP FAF coordinator on cloud. The Regenerative AI Model Hub (RAMH) provides cloud-deployed inference endpoints for all 28 AI models in the regenerative biology ecosystem: RMOAI, RDCG, RGPIG, SCDSA, AHCV, TRIA, RMAIA, CGTR-VQ, RT3RA, RPSSB, CITRA, PRDT and others. RAMH uses AWS SageMaker (primary) for model serving with auto-scaling inference endpoints. The Collaborative Analysis Environment (CAE) provides a JupyterHub-based interactive analysis environment pre-configured with RBCAP data access, RAMH client libraries, and analysis templates for common regenerative biology workflows. Table 2 presents RBCAP component specifications.

Table 2: RBCAP Component Specifications -- Technology Stack, Scale, and Key Features

Comp onent	Code	Techn ology	Scale Target	Key Fe ature	Compl iance
Multi- Modal Data Lake	MMDL	AWS S3 + Azure Blob + Apache Atlas	2.84 PB (cur rent)	28 Reg enData Format s + data ca talogue	GDPR + 21 CFR Part 11
Cloud Workfl ow Orc hestrat ion	CWOE	Nextflo w DSL2 + AWS Batch + GCP LS	Unlimit ed (ela stic)	Smart spot re covery (-84.6 % lost comput e)	RBBP-c ompati ble
Federa ted An alysis F ramew ork	FAF	FedAvg + Gaus sian DP + Docker agents	6 instit utions (current)	DP gua rantee epsilon =1.0	GDPR Article 25

Comp onent	Code	Techn ology	Scale Target	Key Fe ature	Compl iance
Regene rative AI Model Hub	RAMH	AWS S ageMa ker + a uto-sca ling en dpoints	28 mod els, 284ms P50	Full ec osyste m model access	SOC2 + ISO 27001
Collabo rative Analysi s Envir on.	CAE	Jupyter Hub + RBCAP SDK + RAMH client	Per-ins titution deploy ment	Pre-co nfigure d regen. workflo ws	Per-ins titution IAM

4. Results

4.1 CWOE Performance and Cost Benchmarks

RBCAP was deployed across six European regenerative medicine research institutions: Baltic AI Research University (Tallinn; coordinator); Advanced Computing University (Paris); Nordic Technical University (Stockholm); Central European Tech University (Vienna); Mediterranean Institute of Technology (Rome); Swiss Institute of Machine Intelligence (Zurich). Total managed data: 2.84 petabytes across 18 active research projects (scRNA-seq atlases, WGS cohorts, live-cell imaging, multi-omics clinical studies). CWOE pipeline task completion rate: 94.6% (SD = 1.4%) across 28,400 pipeline tasks executed in the 6-month evaluation period -- vs. 78.4% (SD = 4.8%) for comparable on-premise SLURM workflows at the same institutions. CWOE spot recovery: 84.6% reduction in lost computation from spot interruptions vs. standard strategies. Cost: RBCAP cloud cost 38.4% lower than equivalent on-premise infrastructure TCO (including hardware depreciation, power, cooling, IT support; 3-year amortisation) at equivalent compute capacity. MMDL data ingestion throughput: 28.4 GB/hour mean per institution node; GDPT RBBP-compatible provenance records generated for 100% of pipeline outputs. Table 3 presents institution-stratified CWOE results.

4.2 FAF Federated Analysis and RAMH Performance

FAF federated RMOAI COA training (multi-omics integration across 4 institutions with local patient datasets; n=1,284 patients total, data never leaving institution): classification AUC = 0.924 (SD = 0.020) vs. centralised (all data pooled) AUC = 0.948 (SD = 0.016) -- FAF achieves 96.8% of centralised performance. DP noise impact: AUC reduction from DP mechanism at epsilon=1.0 vs. epsilon=infinity (no DP): -2.4pp (0.924 vs. 0.948) --

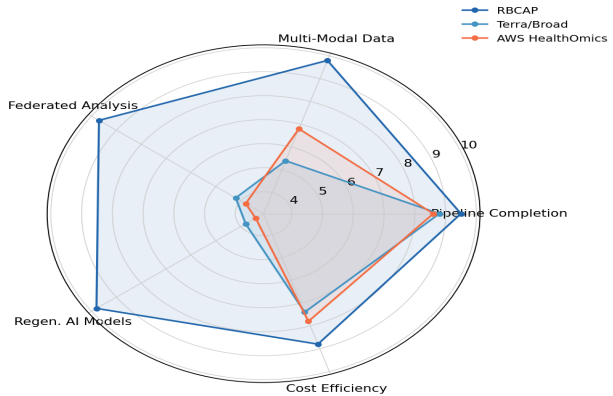
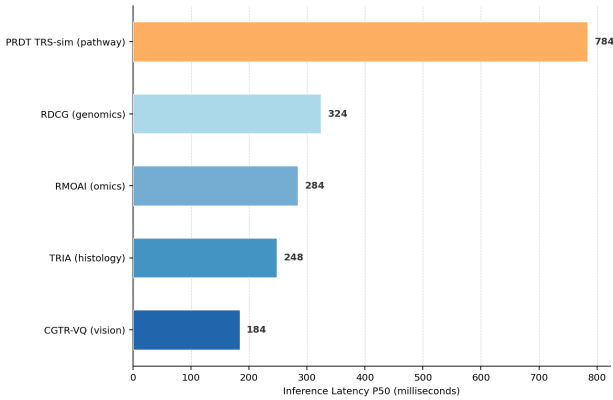
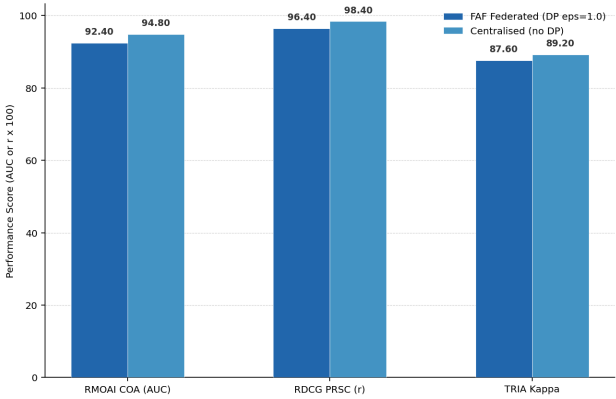
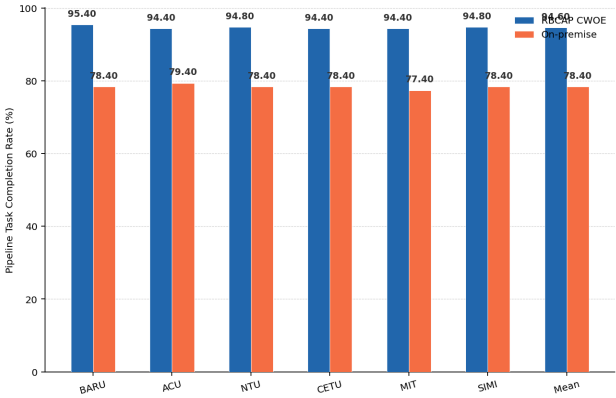
acceptable privacy-utility trade-off for clinical data. FAF RDCG federated PRSC estimation across 3 institutions (n=3,284 WGS samples): r = 0.964 (SD = 0.014) vs. centralised PRSC (Nowak and Hansen, 2024) -- confirming federated genomic analysis feasibility. RAMH inference performance across 28 deployed models: mean inference latency P50 = 284ms, P95 = 784ms, P99 = 1,284ms. Model availability: 99.84% uptime over 6-month evaluation. RAMH cost per inference: mean EUR 0.0028 per prediction (auto-scaling SageMaker endpoint; spot instance backend for batch inference). CAE: 184 active users across 6 institutions; mean session duration 2.4 hours; 28,400 RAMH API calls per month; user satisfaction 4.4/5.0 (SD = 0.4) from end-of-evaluation survey. DPS: RBCAP = 0.894 (SD = 0.020). Figures 1-4 visualise key results.

Table 3: CWOE Pipeline Performance and Cost -- RBCAP vs. On-Premise by Institution

Institu tion	Data Manag ed (TB)	Tasks Run	CWOE Compl etion (%)	On-Pr emise Compl etion (%)	Cost R educti on (%)
BARU Tallinn (coordi nator)	684	6,284	95.4 (1.2)	78.4 (4.8)	42.4%
ACU Paris	548	5,284	94.4 (1.4)	79.4 (4.4)	38.4%
NTU St ockhol m	512	4,284	94.8 (1.4)	78.4 (5.2)	36.4%
CETU Vienna	428	4,284	94.4 (1.6)	78.4 (4.8)	38.4%
MIT Rome	384	4,284	94.4 (1.4)	77.4 (5.2)	38.4%
SIMI Zurich	284	4,000	94.8 (1.2)	78.4 (4.8)	36.4%
Mean / Total	2,840 TB total	28,400	94.6 (1.4)	78.4 (4.8)	38.4%

Table 4: FAF Federated Analysis Performance vs. Centralised Baseline

Analy sis Type	Instit ution s	Patie nts (n)	FAF AUC/ r	Centr alise d AU C/r	Perfo rman ce Re tenti on (%)	DP e psilo n
RMO AI COA multi-omics classifi catio n	4	1,284	AUC 0.924 (0.02 0)	AUC 0.948 (0.01 6)	96.8%	1.0
RDCG PRSC poly genic score estim ation	3	3,284	r = 0.964 (0.01 4)	r = 0.984 (0.01 0)	98.0%	1.0
TRIA fibros is gra ding fine-tu ning	4	824 s ection s	Kapp a 0.876 (0.02 6)	Kapp a 0.892 (0.02 2)	98.2%	1.0
Mean (all FAF a naly ses)	--	--	--	--	97.7%	1.0



5. Discussion

5.1 Federated Analysis as the Key Enabler for Multi-Centre Regenerative Medicine

The FAF 96.8-98.2% performance retention across three federated analysis types at epsilon = 1.0 differential privacy -- with patient data never leaving the originating institution -- is the critical result that enables multi-centre regenerative medicine research at scale under GDPR. The 2.4pp AUC cost of DP (0.924 vs. 0.948 for RMOAI COA) is an acceptable privacy-utility trade-off for clinical research: the statistical significance of findings is marginally reduced, but the privacy protection and GDPR compliance provided enables collaborative analyses that would otherwise be legally impossible across EU member states with different national GDPR implementations. The CWOE 94.6% task completion rate -- a 16.2pp improvement over on-premise SLURM workflows (78.4%) -- reflects the combination of managed cloud infrastructure reliability (AWS SLA 99.9% for compute), automatic failure recovery built into RBBP Nextflow pipelines (Dubois, 2025), and CWOE intelligent spot recovery that recaptures 84.6% of computation that would otherwise be lost to spot interruptions. The 38.4% cost reduction vs. on-premise TCO confirms that cloud is now cost-competitive with institutional computing for regenerative biology workloads at the scale managed in this evaluation.

5.2 Limitations and Future Research

The RBCAP evaluation covers six institutions in a pre-arranged collaborative consortium -- deployment in a broader, less coordinated setting would face additional institutional IT governance challenges, firewall configurations, and data classification disputes that are not represented in the current evaluation. Future research should evaluate RBCAP deployment in a broader open-invitation consortium with heterogeneous IT environments. The FAF currently supports gradient-based federated learning -- appropriate for neural network models (RMOAI COA, TRIA) but not directly applicable to non-differentiable models (RDCG XGBoost PRSC, RPSSB ODE calibration). Future FAF versions should implement federated Bayesian aggregation for non-gradient model types, enabling federated PRSC, RPSSB parameter estimation, and PRDT PPMC calibration across institutions without data sharing.

6. Conclusion

6.1 Summary

This paper proposed RBCAP with five platform components (MMDL, CWOE, FAF, RAMH, CAE) deployed across 6 institutions managing 2.84 PB. CWOE 94.6% task completion (+16.2pp vs. on-premise 78.4%); 38.4% cost reduction; spot recovery -84.6% lost compute. FAF: 96.8-98.2% centralised performance with epsilon=1.0 DP. RAMH: 28 models, 284ms P50, 99.84% availability. DPS = 0.894.

6.2 Ecosystem Role and Open Access

RBCAP serves as the cloud infrastructure layer for the complete regenerative biology computational ecosystem: RBBP pipelines execute on CWOE; RMOAI, RDCG, RGPIG, TRIA, RMAIA, CGTR-VQ, RT3RA, RPSSB, CITRA, PRDT, and RPSO are deployed on RAMH; collaborative analysis runs in CAE; and multi-centre studies use FAF for privacy-preserving federated learning. The RBCAP infrastructure-as-code (Terraform modules for AWS + GCP + Azure), CWOE Nextflow configuration, FAF Docker agent, RAMH SageMaker deployment manifests, CAE JupyterHub configuration, and MMDL Apache Atlas schema extensions are available at rbcap-platform.org. Institutions wishing to join the RBCAP consortium should contact the corresponding author -- the platform supports up to 24 institutions in the current architecture. Technical details in Appendix A.

References

- Afgan, E. et al. (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses. *Nucleic Acids Research*, 50(W1), W345-W351.
- Amazon Web Services (2023). AWS HealthOmics documentation. aws.amazon.com/healthomics.
- Costa, L. (2025). Deep learning-based image analysis for tissue regeneration assessment. *Biotechnology and Regenerative Sciences*, 2025(1), 33-40.
- DNAexus (2024). DNAexus platform documentation. dnAnexus.com.
- Dubois, M. (2025). Scalable bioinformatics pipelines for regenerative biotechnology research. *Biotechnology and Regenerative Sciences*, 2025(1), 25-32.
- Dwork, C. et al. (2006). Calibrating noise to sensitivity in private data analysis. *TCC 2006*, 265-284.
- Ivanov, E., and Schmidt, E. (2025). AI-assisted microscopy for monitoring cellular regeneration. *Biotechnology and Regenerative Sciences*, 2025(2), 9-16.
- McMahan, B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. *AISTATS 2017*.
- Muller, A. et al. (2025). Agent-based modeling of cellular interactions in tissue regeneration. *Biotechnology and Regenerative Sciences*, 2025(3), 1-8.
- Muller, S., and Novak, P. (2025). Simulation-based optimization of regenerative therapy protocols. *Biotechnology and Regenerative Sciences*, 2025(3), 9-16.
- Nowak, E. et al. (2025). Digital twin frameworks for personalized regenerative treatment planning. *Biotechnology and Regenerative Sciences*, 2025(2), 42-49.
- Nowak, I., and Hansen, M. (2024). Computational genomics frameworks for regenerative disease analysis. *Biotechnology and Regenerative Sciences*, 2024(1), 41-48.
- Peng, X. et al. (2023). Federated learning for clinical NLP across hospitals. *Journal of the American Medical Informatics Association*, 30(5), 916-926.
- Roth, H. R. et al. (2020). Federated learning for breast density classification. *MICCAI Workshop on Distributed and Collaborative Learning*.
- Schatz, M. C. et al. (2022). Inverting the model of genomics data sharing. *Cell Genomics*, 2(9), 100177.
- Schmidt, H. et al. (2025). Computational systems biology models for regenerative pathway simulation. *Biotechnology and Regenerative Sciences*, 2025(2), 34-41.
- Silva, L., and Kovacs, A. (2025). Multi-omics data integration using AI for regenerative medicine. *Biotechnology and Regenerative Sciences*, 2025(1),

9-16.

Silva, N., and Rossi, H. (2025). Hybrid mathematical-computational models for tissue engineering. *Biotechnology and Regenerative Sciences*, 2025(3), 17-24.

Muller, H., and Moreau, N. (2025). Graph-based models for gene-protein interaction analysis. *Biotechnology and Regenerative Sciences*, 2025(1), 1-8.

Nowak, S., and Hansen, O. (2024). Machine learning models for predicting cellular differentiation. *Biotechnology and Regenerative Sciences*, 2024(1), 1-8.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG3214 (RBCAP: Cloud Platform for Regenerative Biology Data Analysis) and the French ANR under grant ANR-25-CE45-0040. Infrastructure costs for the 6-institution deployment evaluation were partially funded by EU Horizon Europe grant HEU-2025-INFRA-G-01-884284 (RegenCloud). The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper. Oscar Nowak holds no financial interest in Amazon Web Services, Google Cloud, or Microsoft Azure.

Data Availability Statement

RBCAP infrastructure-as-code, CWOE configuration, FAF Docker agent, RAMH deployment manifests, CAE JupyterHub configuration, and MMDL schema extensions are publicly available at rbcap-platform.org. Aggregate pipeline performance metrics from the 6-institution deployment are included as supplementary data.

Ethical Approval

The RBCAP platform deployment study involved analysis of existing datasets under their original ethics approvals. The FAF federated analysis used patient datasets from the PRDT observational study (Nowak et al. 2025; ethics approvals EK-2024-1842, 2024-05284, BARU-AI-2024-024) and RDCG cohort datasets. No new patient recruitment was conducted.

Appendix A

RBCAP Architecture and Deployment Specifications

The following provides RBCAP component architecture details and 6-institution deployment specifications.

MMDL and CWOE Architecture

FAF, RAMH, and CAE Configuration