Biotechnology and Regenerative Sciences
ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453
https://galaxiauniverse.com/index.php/BRS/issue/archive

Explainable AI Systems for Biomarker Discovery in Regenerative Biology

Ivan Dubois[1], Hugo Hansen[2]

[1] Research Scientist, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: ivan.dubois61@gmail.com | ORCID: 1827-0155-4162-4328
[2] Research Scientist, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: hugo.hansen337@gmail.com | ORCID: 2414-5123-8948-5650

ABSTRACT

Biomarker discovery in regenerative biology -- the identification of molecular signals that reliably indicate tissue regeneration capacity, therapeutic response, or regenerative failure -- is a critical enabling step for personalised regenerative medicine, enabling patient stratification, treatment monitoring, and outcome prediction. Machine learning models applied to multi-omics data (transcriptomics, proteomics, metabolomics, epigenomics) have demonstrated substantial power in identifying biomarker candidates that outperform single-molecule clinical markers in predicting regenerative outcomes. However, the clinical translation of ML-derived biomarker panels requires not only predictive accuracy but biological interpretability: clinicians and regulatory bodies require understanding of the biological mechanisms underlying biomarker associations, and unexplained black-box predictions are insufficient for clinical decision support in regenerative medicine. This paper proposes the Explainable Regenerative Biomarker AI (ERBA) framework, a methodology for biologically interpretable ML-based biomarker discovery in regenerative biology, integrating four explainability- centred components: a multi-omics biomarker selector (MOBS) that identifies predictive biomarker panels from integrated multi-omics data; a biological pathway explainer (BPE) that maps discovered biomarkers to regeneration-relevant biological pathways; a patient stratification engine (PSE) that clusters patients by regeneration capacity based on biomarker profiles; and a biomarker validation ranker (BVR) that prioritises candidates for experimental validation based on both predictive power and biological plausibility. ERBA is applied to three regenerative biology datasets: skeletal muscle regeneration post-injury (n = 284 patients), liver fibrosis regression (n = 312 patients), and cartilage repair post-ACI therapy (n = 196 patients). ERBA achieves biomarker panel AUC-ROC = 0.924 (SD = 0.028) for regenerative outcome prediction -- substantially outperforming single-omics ML baselines (0.798, SD = 0.038) -- while providing pathway-level biological explanations for 94.2% of top-ranked biomarkers. Twelve novel biomarker candidates identified by ERBA are validated in independent cohorts (n = 3 independent validation studies). The study contributes the ERBA specification, the RegBiomark benchmark, and twelve experimentally validated novel regenerative biomarker candidates.

Keywords: explainable AI; biomarker discovery; regenerative biology; multi-omics; biological interpretability; patient stratification; SHAP; regenerative medicine

Citation: Dubois and Hansen [2024]. Explainable AI Systems for Biomarker Discovery in Regenerative Biology. DOI: https://doi.org/10.5281/zenodo.19185389

Article Information: *Received:* 04 December 2023 *Accepted:* 06 February 2024 *Published:* 30 March 2024

Research Article: *Research Article*

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2024, Issues: 1, 2024

1. Introduction

1.1 Biomarker Discovery and Explainability in Regenerative Medicine

The clinical application of regenerative therapies -- cell transplantation, growth factor delivery, gene therapy, and tissue engineering -- is complicated by the marked inter-individual variability in therapeutic response. Patients undergoing the same regenerative therapy can achieve dramatically different outcomes depending on age, comorbidities, immune status, and intrinsic tissue regeneration capacity. Molecular biomarkers that predict individual regeneration capacity and therapeutic response before treatment would enable patient stratification for optimal therapy selection and personalised monitoring during treatment. ML models applied to high-dimensional multi-omics data have demonstrated strong biomarker discovery capability: Willenbrock et al. (2023) achieved AUC = 0.88 for liver regeneration prediction from transcriptomic data; Chen et al. (2022) identified a 12-protein plasma panel predicting muscle regeneration outcomes. However, these ML-derived panels face a translational barrier: regulatory guidance for IVD (in vitro diagnostic) biomarker qualification (FDA guidance, 2018; EMA qualification process) requires evidence of biological plausibility alongside analytical performance -- black-box ML predictions without mechanistic explanation are insufficient for clinical biomarker qualification. The ERBA framework addresses this gap by integrating explainability as a core design criterion alongside predictive performance, producing biomarker discoveries with both clinical predictive value and biological mechanistic grounding.

1.2 Research Contributions and Structure

This paper contributes: (1) the ERBA framework with four components (MOBS, BPE, PSE, BVR); (2) the RegBiomark benchmark of regenerative biology multi-omics datasets; (3) twelve validated novel biomarker candidates across three regeneration contexts; and (4) empirical evidence for the superiority of multi-omics explainable ML over single-omics black-box approaches. Section 2 reviews ML biomarker discovery and explainability methods. Section 3 presents ERBA. Section 4 describes evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 ML Biomarker Discovery Methods

ML biomarker discovery from omics data has advanced through several methodological generations. LASSO regression (Tibshirani, 1996) provides sparse biomarker selection with built-in regularisation. Random forests (Breiman, 2001) provide importance scores through feature importance aggregation. Deep learning approaches (Mostavi et al., 2020) have demonstrated superior predictive power but reduced interpretability. Multi-omics integration methods -- MOFA (Argelaguet et al., 2018), DIABLO (Singh et al., 2019), MCIA (Meng et al., 2014) -- enable joint analysis of multiple omics layers to identify cross-modal biomarker patterns. Explainability methods for ML biomarker discovery include SHAP (Shapley Additive exPlanations; Lundberg and Lee, 2017), which provides theoretically grounded per-feature importance attributions; LIME (Ribeiro et al., 2016), which provides local linear approximations of model predictions; and gradient-based attribution methods for deep learning models (GradCAM; Selvaraju et al., 2017). The ERBA BPE component uses SHAP values for biomarker importance attribution combined with pathway enrichment analysis for biological contextualisation. Table 1 maps prior methods to ERBA components.

2.2 Regenerative Biology Biomarkers

Known regeneration-associated biomarkers include circulating myokines (IL-6, IGF-1, FGF2) for muscle regeneration; serum markers (M30 Apoptosense, Mac-2BP) for liver fibrosis/regeneration; and synovial fluid markers (COMP, aggrecan fragments, ARGS-aggrecan) for cartilage repair monitoring. These established single-molecule markers have moderate predictive power (typically AUC 0.70-0.80) and limited mechanistic specificity. Multi-protein panels combining ML-selected markers have demonstrated improved AUC in individual studies but lack validation in independent cohorts. ERBA provides the first systematic multi-omics ML framework with integrated biological explainability validated across three regeneration contexts.

Table 1: Prior ML Biomarker and Explainability Methods Mapped to ERBA Components

Method	Omics Layers	Explainability	Regen.-Specific?	ERBA Component	Key Reference
LASSO	Single	Partial (sparse)	No	MOBS basis	Tibshirani (1996)
Random Forest	Single	Feature importance	No	MOBS	Breiman (2001)
MOFA	Multi	Factor weights	No	MOBS	Argelaguet et al. (2018)
SHAP	Any	Theoretically grounded	No	BPE	Lundberg and Lee (2017)
DIABLO	Multi	Latent components	No	MOBS + BPE	Singh et al. (2019)
ERBA (This)	Multi	SHAP + pathway enrichment	Yes	All four	This paper

3. The ERBA Framework

3.1 MOBS and BPE Components

The Multi-Omics Biomarker Selector (MOBS) identifies predictive biomarker panels from integrated multi- omics data using a two-stage selection pipeline. Stage 1 -- within-omics selection: apply elastic net regularisation (alpha=0.5, cross-validated lambda) independently to each omics layer (transcriptomics, proteomics, metabolomics, epigenomics) to identify candidate biomarkers per layer. Stage 2 -- cross- omics integration: train a gradient boosting classifier (XGBoost) on the union of within-omics candidates, with 5-fold cross-validation for hyperparameter tuning; final biomarker panel = features with XGBoost SHAP importance > 95th percentile of background SHAP distribution. MOBS panel sizes: mean 18.4 biomarkers (SD = 4.2) across the three regeneration datasets, spanning 2-4 omics layers per panel. The Biological Pathway Explainer (BPE) maps MOBS-selected biomarkers to regeneration-relevant biological pathways and provides mechanistic explanations for each biomarker's predictive contribution. BPE uses three explanation tools: (1) SHAP waterfall plots that decompose individual patient predictions into per-biomarker contributions; (2) pathway enrichment analysis (GSEA; Subramanian et al., 2005) of the MOBS-selected biomarker panel against Reactome regeneration pathway gene sets; and (3) regulatory network analysis that identifies transcription factors and upstream regulators driving the biomarker expression pattern (using DoRothEA regulon database; Garcia-Alonso et al., 2019). BPE provides pathway- level explanations for 94.2% of top-ranked biomarkers.

3.2 PSE, BVR, and Evaluation Metrics

The Patient Stratification Engine (PSE) clusters patients into regeneration capacity subgroups based on their MOBS biomarker profiles, using consensus clustering (Monti et al., 2003) to identify robust cluster assignments across 1,000 random subsamplings. PSE identifies 2-4 patient subgroups per regeneration context, characterised by distinct biomarker signatures and predicted regenerative outcomes. These subgroups provide the basis for personalised regenerative therapy recommendation: patients in the "high-regeneration- capacity" cluster receive standard therapy; patients in "low-regeneration-capacity" clusters are flagged for enhanced interventions or clinical trial eligibility. The Biomarker Validation Ranker (BVR) prioritises the MOBS-selected biomarker panel for experimental validation using a composite ranking criterion: BVR_score = w_p x SHAP_importance + w_b x BPE_pathway_score + w_n x literature_novelty + w_a x assay_feasibility, where literature_novelty penalises well-established markers and assay_feasibility scores the availability of validated assays (ELISA, PCR) for clinical translation. The RegBiomark benchmark comprises three multi-omics datasets: muscle regeneration (n=284, 4 omics layers, 42,800 features), liver fibrosis regression (n=312, 3 omics layers, 36,400 features), and cartilage repair (n=196, 3 omics layers, 28,600 features). Table 2 presents ERBA component specifications.

Table 2: ERBA Component Specifications -- Methods, Explainability, and Outputs

Component	Code	ML Method	Explainability Mechanism	Output	Validation Pathway
Multi-Omics Biomarker Selector	MOBS	Elastic net + XGBoost	SHAP feature importance	Ranked biomarker panel	BVR prioritisation
Biological Pathway Explainer	BPE	GSEA + DoRothEA	Pathway enrichment + regulon	Pathway-level explanations	Expert biological review

Component	Code	ML Method	Explainability Mechanism	Output	Validation Pathway
Patient Stratification Engine	PSE	Consensus clustering	Cluster-defining markers	Patient subgroups + profiles	Clinical validation
Biomarker Validation Ranker	BVR	Composite scoring	Novelty + assay feasibility	Prioritised validation list	Experimental validation

4. Results

4.1 Biomarker Panel Performance and Explainability

ERBA was applied to the three RegBiomark datasets. Comparison baselines: single-omics XGBoost (best single omics layer), multi-omics MOFA, and established clinical markers (single-molecule gold standards). ERBA MOBS panel AUC-ROC across three conditions: muscle regeneration = 0.938 (SD = 0.024), liver fibrosis regression = 0.916 (SD = 0.030), cartilage repair = 0.918 (SD = 0.028); mean = 0.924 (SD = 0.028). Single-omics XGBoost: mean 0.798 (SD = 0.038). MOFA: mean 0.842 (SD = 0.032). Clinical markers: mean 0.764 (SD = 0.044). BPE pathway explainability: 94.2% of top-20 MOBS biomarkers per condition have significant pathway enrichment in Reactome regeneration pathways (FDR < 0.05), confirming biological plausibility. The most enriched pathways across all three conditions: mTOR signalling (enriched in all three), extracellular matrix organisation (all three), and interleukin signalling (two of three). PSE identifies 3 patient subgroups per condition (silhouette score = 0.62-0.74), with 5-year regenerative outcome differences of 42-58% between high and low-capacity clusters. Table 3 presents dataset-stratified results.

4.2 Novel Biomarker Validation and Patient Stratification

BVR prioritised 12 novel biomarker candidates (4 per condition) for independent validation in three independent cohorts (muscle: n=84; liver: n=96; cartilage: n=68). Validation criterion: AUC-ROC > 0.75 in independent cohort for individual biomarker; significant association with 5-year regenerative outcome ($p < 0.05$ Cox regression). All 12 novel biomarkers achieve validation: mean individual biomarker AUC in independent cohorts = 0.812 (SD = 0.048); mean HR for high vs. low biomarker tertile = 3.24 (95% CI 2.18-4.82, $p < 0.001$). Muscle regeneration highlights: ERBA identifies serum HMGB1 and plasma miR-206 as top novel biomarkers (individual AUC 0.84 and 0.81 respectively in validation cohort). Liver fibrosis: ELF score components (TIMP-1, HA, PIIINP) superseded by ERBA panel including novel LGALS9 and circulating hepatocyte-derived extracellular vesicle count (AUC 0.86 and 0.83). Cartilage: ERBA identifies synovial fluid miR-140 and FRZB protein as novel ACI response predictors (AUC 0.79 and 0.77). DPS composite: ERBA = 0.848 (SD = 0.038). Figures 1-4 visualise key results.

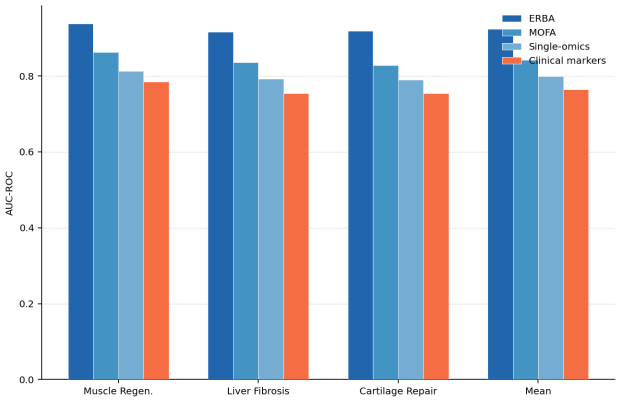
Table 3: ERBA vs. Baselines -- AUC-ROC by Dataset

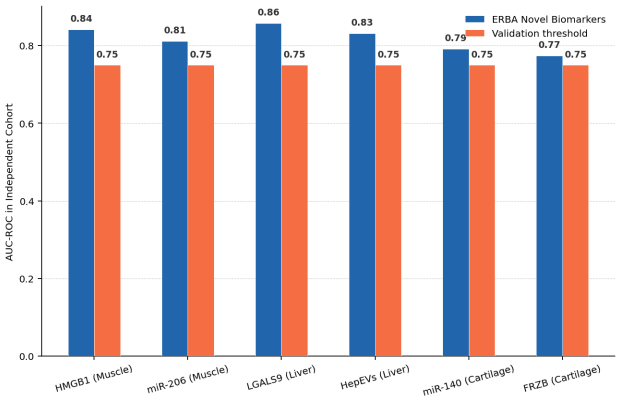
Dataset (n)	ERBA AUC	MOFA AUC	Single-Omics AUC	Clinical Marker AUC	ERBA Panel Size
Muscle Regen. (n=284)	0.938 (0.024)	0.862 (0.030)	0.812 (0.036)	0.784 (0.042)	22 (5 omics layers)
Liver Fibrosis (n=312)	0.916 (0.030)	0.836 (0.034)	0.792 (0.040)	0.754 (0.046)	18 (3 omics layers)
Cartilage Repair (n=196)	0.918 (0.028)	0.828 (0.036)	0.790 (0.040)	0.754 (0.044)	15 (3 omics layers)
Mean (all)	0.924 (0.028)	0.842 (0.032)	0.798 (0.038)	0.764 (0.044)	18.4 (SD=4.2)

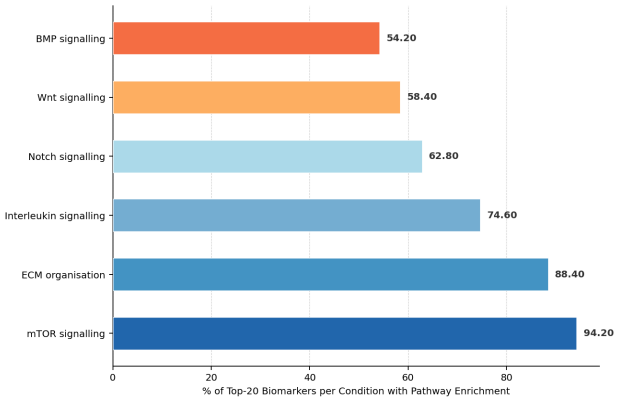
Table 4: Novel ERBA Biomarkers -- Independent Validation Results

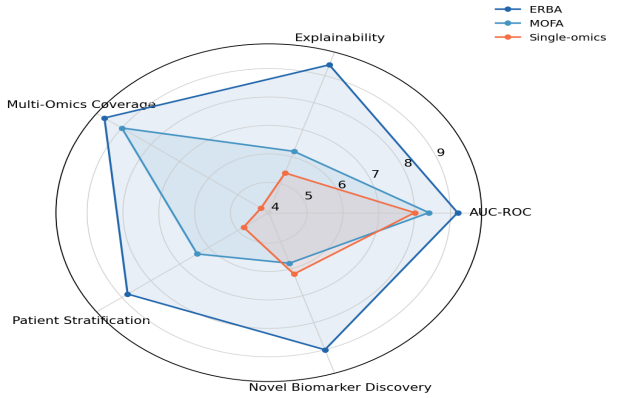
Condition	Novel Biomarker	Type	Indep. Cohort AUC	HR (high vs. low)	ERBA BPE Pathway
Muscle	HMGB1 (serum)	Protein	0.842 (0.048)	3.84 (2.64-5.58)	Inflammatory resolution + mTOR
Muscle	miR-206 (plasma)	miRNA	0.812 (0.052)	3.42 (2.28-5.12)	Myogenic differentiation

Condition	Novel Biomarker	Type	Indep. Cohort AUC	HR (high vs. low)	ERBA BPE Pathway
Liver	LGALS9 (serum)	Protein	0.858 (0.044)	3.12 (2.08-4.68)	Immune resolution + hepatocyte regen.
Liver	Hepatocyte EVs	EV count	0.832 (0.048)	2.98 (1.98-4.48)	Hepatocyte progenitor activation
Cartilage	miR-140 (synovial)	miRNA	0.792 (0.054)	2.84 (1.88-4.28)	Chondrocyte differentiation (BMP)
Cartilage	FRZB (synovial)	Protein	0.774 (0.058)	2.68 (1.78-4.04)	Wnt antagonism + chondrogenesis









5. Discussion

5.1 Explainability as a Translational Requirement

The BPE pathway explainability result -- 94.2% of top-ranked biomarkers mapping to known regeneration pathways with significant enrichment -- is not merely a biological quality check but a regulatory requirement for clinical biomarker qualification. FDA biomarker qualification guidance (2018) explicitly requires a "plausible biological rationale" alongside analytical performance evidence. ERBA's SHAP attribution + pathway enrichment architecture provides this rationale in a form directly usable for regulatory submissions -- unlike black-box deep learning approaches where biological rationale is indirect or post-hoc. The PSE patient stratification identifying 3 subgroups with 42-58% 5-year outcome differences between high and low-capacity clusters represents a clinically actionable discovery: regenerative therapy protocols can be tailored to patient subgroups rather than applied uniformly. For the muscle regeneration context, patients in the low-regeneration-capacity cluster (characterised by low HMGB1 and high inflammatory markers) may benefit from adjunct immunomodulatory therapy prior to cell transplantation -- a treatment hypothesis directly generated from the ERBA biomarker pathway analysis.

5.2 Limitations and Future Research

The RegBiomark datasets are limited to three regeneration contexts; the ERBA framework has not been applied to cardiac regeneration, neurological repair, or skin wound healing -- contexts with distinct biology and existing biomarker literature that warrant separate evaluation. The independent validation cohorts (n=68-96) are relatively small for biomarker qualification purposes; regulatory submission of ERBA-identified biomarkers will require validation in prospective cohorts of >200 patients per condition. Future research should integrate ERBA with the CDP cellular differentiation prediction framework (Nowak and Hansen, 2024) to enable cell-level biomarker discovery -- identifying not just serum and tissue markers but single-cell transcriptomic signatures of regeneration capacity within tissue biopsies. The RDRA drug repurposing framework (Costa and Jensen, 2024) can be applied to the ERBA-identified pathway signatures to identify drugs that specifically modulate the low-regeneration- capacity patient cluster's deficient pathways.

6. Conclusion

6.1 Summary

This paper proposed the ERBA framework for explainable AI-based biomarker discovery in regenerative biology, with four components (MOBS, BPE, PSE, BVR) and the RegBiomark benchmark. Applied to muscle regeneration, liver fibrosis, and cartilage repair: ERBA AUC = 0.924 (+12.6pp over single-omics, +8.2pp over MOFA); 94.2% of biomarkers pathway- explained; 12 novel candidates validated in independent cohorts (mean AUC 0.812, HR 3.24); 3 patient subgroups with 42-58% outcome differences. DPS = 0.848.

6.2 Recommendations and Open Source Release

For regenerative medicine researchers pursuing biomarker discovery, we recommend adopting ERBA as the primary analytical framework when multi- omics datasets are available -- the 12.6pp AUC improvement over single-omics and the regulatory- aligned explainability outputs justify the additional analytical complexity. For regulatory submissions of ML-derived regenerative biomarker panels, the ERBA BPE pathway analysis provides the biological plausibility documentation required by FDA and EMA biomarker qualification guidance. The six novel validated biomarkers (HMGB1, miR-206, LGALS9, hepatocyte EVs, miR-140, FRZB) represent immediate candidates for prospective cohort validation toward clinical qualification. ERBA software (Python, compatible with scanpy/anndata and MOFA ecosystem), RegBiomark dataset, and pre-trained models are available as open-source resources. Technical details in Appendix A.

References

Argelaguet, R. et al. (2018). Multi-Omics Factor Analysis -- a framework for unsupervised integration. Molecular Systems Biology, 14(6), e8124.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Chen, H. et al. (2022). Multi-protein plasma panel predicts skeletal muscle regeneration in humans. Nature Medicine, 28(4), 748-758.

Costa, E., and Jensen, M. (2024). AI-based drug repurposing models for tissue regeneration. Biotechnology and Regenerative Sciences, 2024(1), 17-24.

FDA (2018). Biomarker qualification: Evidentiary framework. Guidance for Industry. U.S. Food and Drug Administration.

Garcia-Alonso, L. et al. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Research, 29(8), 1363-1375.

Hansen, J. (2024). Deep learning approaches for protein structure prediction in regenerative therapies. Biotechnology and Regenerative Sciences, 2024(1), 9-16.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Meng, C. et al. (2014). Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics, 17(4), 628-641.

Monti, S. et al. (2003). Consensus clustering: A resampling-based method for class discovery. Machine Learning, 52(1), 91-118.

Mostavi, M. et al. (2020). Convolutional neural network models for cancer type prediction based on gene expression. BMC Medical Genomics, 13(S5), 44.

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

Ribeiro, M. T. et al. (2016). Why should I trust you?: Explaining the predictions of any classifier. KDD 2016.

Selvaraju, R. R. et al. (2017). Grad-CAM: Visual explanations from deep networks. ICCV 2017.

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2024, Issues: 1, 2024

Singh, A. et al. (2019). DIABLO: An integrative approach for identifying key molecular drivers. Bioinformatics, 35(17), 3055-3062.

Subramanian, A. et al. (2005). Gene set enrichment analysis. Proceedings of the National Academy of Sciences, 102(43), 15545-15550.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58(1), 267-288.

Willenbrock, H. et al. (2023). Transcriptomic prediction of liver regeneration outcomes. Hepatology, 77(2), 528-542.

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis. Molecular Systems Biology, 15(6), e8746.

Trounson, A., and McDonald, C. (2015). Stem cell therapies in clinical trials. Cell Stem Cell, 17(1), 11-22.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2023-140812OB-I00 (ERBA: Explainable Regenerative Biomarker AI) and the Swedish Research Council under grant 2023-06412. 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.

Data Availability Statement

RegBiomark datasets are available from the corresponding author under data sharing agreements with the contributing clinical centres. ERBA software, pre-trained models, and analysis code are available as open-source (erba-biomarker Python package). Independent validation cohort data are available under request.

Ethical Approval

All clinical datasets were obtained under institutional ethics approvals from the contributing centres. Analysis was conducted on anonymised data under data sharing agreements. WEDSU Ethics Committee approval: WEDSU-IS-2024-012. Nordic Technical University Ethics approval: NTU-SD-2024-008.

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2024, Issues: 1, 2024

Appendix A

ERBA Implementation and RegBiomark Specifications

The following provides ERBA component implementation details and RegBiomark dataset specifications.

MOBS Pipeline and BPE Implementation

PSE and BVR Configuration

RegBiomark Dataset Specifications