

Perturbation Catalogue: harmonised, AI-ready genetic perturbation data for research and drug discovery



Challenges

Genetic perturbation studies are generating rapidly growing volumes of CRISPR screening, multiplexed variant-effect and expression-level data. These datasets are distributed across repositories, use heterogeneous formats and scores, and often contain incomplete or inconsistent metadata. Finding comparable experiments and combining evidence across studies therefore requires substantial specialist effort, limiting reuse in target discovery, functional genomics and machine learning.

Technology

The Perturbation Catalogue is an open platform that collects, curates and standardises human genetic perturbation data in a common framework. It brings together gene-level perturbations such as CRISPR screens, variant-effect assays such as MAVE, and expression-level experiments such as Perturb-seq. The public resource includes curated data from MaveDB, DepMap, scPerturb and PerturBase. Researchers can search across studies, inspect harmonised metadata and experimental summaries, and explore perturbation effects through a web interface or programmatic access.

The resource integrates contextual information and cross-references from EMBL-EBI and partner resources. Its open-source, cloud-native architecture supports scalable access and provides a foundation for cross-study analyses, predictive modelling and deployment alongside organisation-specific data.

Core Expertise

- Scientific curation, data cleaning and metadata standardisation across heterogeneous perturbation experiments
- A unified data model and ontology-based representation spanning genes, variants, phenotypes and experimental context
- Normalisation and preparation of comparable, machine-learning-ready datasets
- Scalable data engineering, search, APIs, interactive exploration and visualisation
- Integration with established resources such as Ensembl, UniProt, ChEBI and the Open Targets Platform
- User-centred design and iterative validation with academic and pharmaceutical researchers

Internal EMBLEM Reference

2026-049

Project Lead

Dr. Alexey Sokolov, [Genomics Technology Infrastructure Team Leader](#),
EMBL-EBI, Hinxton, UK

EMBLEM TECHNOLOGY TRANSFER GMBH

Boxbergering 107
D-69126 Heidelberg
Germany

Tel.: +49 (0) 6221 363 22 10

info@embl-em.de

www.embl-em.de

Dr. Birgit Kerber

kerber@embl-em.de

Perturbation Catalogue: harmonised, AI-ready genetic perturbation data for research and drug discovery



Applications

- Compare gene or variant effects across experiments, cell models and disease contexts
- Identify causal links between genetic variation, gene function and cellular phenotypes
- Prioritise therapeutic targets, disease-relevant networks and pathway vulnerabilities
- Support cross-study meta-analysis, benchmarking and AI model development using curated data
- Combine public evidence with internal datasets in organisation-specific analytical workflows

Intellectual Property

The software is publicly available under the Apache License 2.0; public datasets remain subject to source-specific terms. Any tailored deployment, integration, curation, analysis or support should be agreed with EMBLEM and the project team.

Further Reading

^[1] [Perturbation Catalogue](#) ^[2] [EMBL-EBI launch article](#) ^[3] [Open-source repository](#) ^[4] [EMBL-EBI in 2025 \(Nucleic Acids Research\)](#)

Seeking

The team is interested in research and industry collaborations that extend the value of perturbation data. Potential areas include:

- Onboarding and harmonising additional public or partner-provided perturbation datasets
- Developing tailored analyses, visualisations and target-discovery workflows
- Integrating the catalogue with internal data and validating AI-enabled industry use cases
- Supporting customised data curation, integration and analytical solutions based on specific partner needs

The open and extensible infrastructure provides opportunities for tailored collaborations in data integration, functional genomics, target discovery and AI-driven analysis.

Keywords

Genetic Perturbation # CRISPR Screening # Perturb-seq # Variant Effect # Functional Genomics # Data Harmonisation # Target Discovery # Drug Discovery # AI-ready Data # Machine Learning # Bioinformatics # Data Integration # Open Data

Benefits

- One discoverable entry point across previously fragmented perturbation data sources
- Expert curation and consistent metadata that improve cross-study comparability and reproducibility
- Reduced preparation time for data-driven target discovery and functional interpretation
- Interactive and programmatic access for exploratory research and computational pipelines
- Open-source infrastructure that can be extended for new datasets and deployment needs

EMBLEM TECHNOLOGY TRANSFER GMBH

Boxberggring 107
D-69126 Heidelberg
Germany

Tel.: +49 (0) 6221 363 22 10

info@embl-em.de

www.embl-em.de

Dr. Birgit Kerber
kerber@embl-em.de