European Commission

Joint Research Centre

Leveraging Artificial Intelligence tools to collect and structure information on human biology-based models used in biomedical research

The Biomedical models Hub (BimmoH) dataset

Deceuninck, P., Barroso, J., Bridio, S., Chinchio, E., Gastaldello, A., Mennecozzi, M., Selfa Aspiroz, L., Straccia, M., Whelan, M.

2026

Contact information
Name: Deceuninck Pierre
Address: Via E. Fermi, 2749. TP126 I-21027 Ispra (VA), Italy
Email: pierre.deceuninck@ec.europa.eu
Tel.: +39 0332 785861

Joint Research Centre
https://joint-research-centre.ec.europa.eu

JRC144237
EUR 40668

PDF ISBN 978-92-68-38821-1 ISSN 1831-9424 doi:10.2760/0117069 KJ-01-26-136-EN-N

Luxembourg: Publications Office of the European Union, 2026

How to cite this report: Deceuninck, P., Barroso, J., Bridio, S., Chinchio, E., Gastaldello, A. et al., *Leveraging Artificial Intelligence tools to collect and structure information on human biology-based models used in biomedical research - The Biomedical models Hub (BimmoH) dataset,* Publications Office of the European Union, Luxembourg, 2026, https://data.europa.eu/doi/10.2760/0117069, JRC144237.

Contents

Abstract

The Biomedical models Hub (BimmoH) is the world largest dataset containing scientific article references making use of human biology-based models. Powered by Artificial Intelligence (AI), it consolidates information on a wide range of models based on human biology, including organ-on-a-chip technologies, 3D cell cultures, and computational models. By bringing together these resources, it enables researchers to design studies that are both relevant and translatable to human health.

This report provides the technical details of the work performed to create the BimmoH dataset, allowing full transparency on the article selection process. It covers the methodological approach, going from data preparation to the identification of the best machine learning approaches for detecting the use of human biology-based models in biomedical research articles and indexing them, as well as the validation of the dataset.

The result is a comprehensive, curated collection of hundreds of thousands of resources that can significantly reduce research time and accelerate innovation, and that will be kept updated over time. Beyond academia, BimmoH offers value to regulators, policymakers, funders, and industry stakeholders. It can inform evidence-based decisions across areas such as drug development, safety assessment, and early-stage biomedical exploration.

Developed under a European Parliament Pilot Project, the initiative also supports the EU's Three Rs principle, reinforcing European leadership in animal welfare and sustainable innovation.

Acknowledgements

BimmoH was funded through a Pilot project of the European Parliament and implemented by the Joint Research Centre of the European Commission.

The work was contracted out to a private consortium composed of:

- FRESCI by SCIENCE&STRATEGY: project management, vocabularies development and scientific validation
- TenWise/Ecomole: data preparation and AI model development
- Ready2Use: dataset preparation and web application development
- Altertox: training packages and documentation

The JRC Systems Toxicology unit would like to thank all students participating to the JRC Summer School 2025 on Non-Animal Approaches in Science for their participation to the testing of the alpha version of the BimmoH dataset in May 2025.

Finally, the authors and contractors would also like to underline the contribution of all stakeholders that participated to the SWOT analysis in March 2024, the workshop on search functionalities organised in October 2024, and the Beta testing of the web application in July 2025.

Data sources used to create the BimmoH dataset comes from PubMed[1], "Courtesy of National Library of Medicine" (Sayers *et al*, 2025) and OpenAlex[2], CC0 (Priem *et al*, 2022).

[1] https://pubmed.ncbi.nlm.nih.gov

[2] https://openalex.org

1. Executive summary

This technical report presents an innovative initiative by the European Commission's Joint Research Centre (JRC). The Biomedical models Hub (BimmoH) aims to promote and foster the use of human biology-based models in biomedical research by creating the largest dataset of its kind, integrating Artificial Intelligence technology to collect routinely and structure information.

The introduction presents the existing biomedical reviews, already published by EU Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM). As they were initially limited to specific disease areas over fixed periods, BimmoH seeks to identify and incorporate a broader range of human biology-based models, including organ-on-a-chip technologies, 3D cell cultures, and computational models, published both before and after these review periods. This extended scope is designed to support researchers, regulators, and policymakers in making evidence-based decisions that advance human-centric research and innovation.

In the objectives section, the report outlines the primary goals of BimmoH to leverage AI techniques to automatically identify and classify scientific articles from PubMed mentioning the use of human biology-based models. This involves not only expanding the pool of included articles but also providing advanced search tools to enable bespoke reviews tailored to specific research needs.

The methodology section details the AI-driven approach, describing the optimisation machine learning classifiers by using various training sets, embeddings, and algorithms. The result is a two-tier classification system that efficiently filters relevant articles from a pre-selected pool of 4.3 million PubMed entries.

The report provides a comprehensive overview of the experimental design and validation phases, explaining how a supervised Machine Learning (ML) classifier was refined through multiple iterations, focusing on maximising precision and specificity while balancing sensitivity to obtain a final dataset comprising approximately 800,000 references.

To facilitate user access and navigation, the dataset is indexed with specific vocabularies developed for anatomy, disease, and modelling approaches, supporting multidimensional search criteria. Validation phases included extensive testing by JRC experts, PhD students during the JRC Summer School 2025, and a diverse group of stakeholders in a beta testing phase. Feedback from these sessions was largely positive, confirming the dataset's relevance and utility for various user communities.

In the discussion section, the report highlights the added value of BimmoH, emphasising its role in promoting the use of human biology-based models and aligning with the EU's Three Rs principle for animal welfare and sustainable innovation making this dataset a crucial resource for accelerating innovation in drug development, safety assessment, and early-stage biomedical exploration.

While the report acknowledges limitations, such as partial coverage due to reliance on titles and abstracts, it underscores the success of the project in vastly expanding the database from an initial 3,000 articles to nearly 800,000. Future improvements are planned to enhance sensitivity and explore access to full-text articles for deeper analysis.

In conclusion, the BimmoH dataset represents a significant European advancement in the promotion and utilisation of human biology-based models, offering a scalable solution to support the evolving

needs of biomedical research and policy development. The dataset is designed for regular updates[3], ensuring it remains current and relevant, thereby continuing to serve as a valuable tool for the research community in the pursuit of human-centric science and innovation.

[3] Initially planned to take place on a quarterly basis

2. Introduction

The Joint Research Centre, Systems Toxicology Unit, runs the EU Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM) that operates according to its mandate set out in Article 48/Annex VII of Directive 2010/63/EU on the protection of animals used for scientific purposes. An important aspect of EURL ECVAM's work is to facilitate the sharing of knowledge on non-animal models across different communities and promoting their uptake for scientific use. In 2022, the number of animals employed in basic, translational, and applied biomedical research was about 4.1 million, representing more than half of animal uses for scientific purposes in the EU (European Commission, 2024).

The significant advances made in recent years in the development of non-animal procedures and techniques (e.g., in vitro, in silico) represent a huge resource for enhancing the modelling and comprehension of human-specific biological processes and pathologies without using animals. However, an important prerequisite for their widest use and further development and application across disciplines, is the ready access to state-of-the-art models and methods adequately described and presented.

In 2022, EURL ECVAM completed the publication of biomedical reviews analysing available and emerging non-animal models used for research in the following seven disease areas:

1. Respiratory tract diseases (Hynes *et al*, 2020)
2. Breast cancer (Folgiero *et al*, 2020)
3. Neurodegenerative diseases (Witters *et al*, 2021)
4. Immuno-oncology (Romania *et al*, 2021)
5. Immunogenicity testing for Advanced Therapy Medicinal Products (ATMP) (Canals *et al*, 2022)
6. Cardiovascular diseases (Celi *et al*, 2022)
7. Autoimmune diseases (Otero *et al*, 2022)

The aim was to identify and describe specific research contexts where animal models have been put aside in favour of non-animal techniques that use, for example, in vitro assays based on human cells and engineered tissues or in silico approaches employing computer modelling and simulation.

However, despite the significant value of the reviews carried out by EURL ECVAM, it is desirable to expand such work to:

1. Identify promising non-animal models that have been published since the cut-off date of the EURL ECVAM reviews.
2. Analyse the availability of non-animal models in other disease areas.
3. Develop search tools that the scientific community can use to carry out bespoke reviews and build customised non-animal model databases that are fit for their own purposes.

This report presents the outcome of the "Biomedical models Hub" (BimmoH) project designed to develop an automated database that collects and structures information on human biology-based models in use for biomedical research, published in scientific journals. Making use of AI to mine the vast body of published literature, BimmoH enables the creation of an up-to-date, state of the art

dataset, allowing the extension of already collected models both in time and in scope without restriction of specific diseases categories.

With a user-friendly search interface, BimmoH supports interested end-users (e.g., scientists working in biomedical research, research projects, evaluation committees, educational institutions, etc.) to easily find information on available human biology-based models in specific biomedical research categories.

This report complements the BimmoH dataset published in the JRC data catalogue (Deceuninck *et al*, 2025), providing complete information on how the dataset was collected and structured, and reviewing its strengths and limitations.

3. Objectives

The main objective of BimmoH is to leverage AI approaches, and in particular supervised ML techniques, to identify human biology-based models published in the scientific literature. With a baseline of approximately 3,000 references of research articles, identified through the past EURL ECVAM biomedical reviews introduced in the previous section, our aim was to create an automated pipeline able to identify scientific articles referenced in PubMed matching our criteria to identify human biology-based models.

3.1. Biomedical reviews extension

The starting point of BimmoH was the collection of models published between 2014 and 2019 coming from the EURL ECVAM biomedical review exercise described in the introduction. Such work identified more than 3,000 relevant scientific publications with certain domain-based specificities (e.g., more frequent use of *in silico* models for cardiovascular diseases). The biomedical reviews highlighted that non-animal models can be identified in all analysed sectors and that the models identified represents only a small fraction of the total number of models in use that keeps growing over time (see section 2).

Making use of ML techniques, we wanted to automatically extend this collection over time and identify new models published before and after the timeframe of our reviews. We also wanted to extend to other areas of biomedical research, using an actionable strategy to avoid the repetition of manual work.

3.2. Stakeholders support

BimmoH's final goal was to go beyond the biomedical reviews (identifying non-animal approaches used in biomedical research) by promoting human biology-based models with a high potential for increasing the translatability of basic and applied research, while reducing the use of animal models. For this purpose, we selected and interviewed a panel of representative stakeholders that would make use of BimmoH's data. This panel was composed of academic researchers, biomedical data scientists, regulators, policy advisors, Three Rs organisations, industry and innovators.

Box 1. Human biology-based model definition

In BimmoH, a model is defined as any system capable of receiving an input and that allows recording an output. This includes human and/or animal models based on data or biological or biochemical materials, and may be:

in vitro: Living biological components outside a living organism.

in silico: Computational-based models.

in chemico: Chemical systems used for testing and analysis.

A human biology-based models is a model used to replicate or simulate human biological processes, diseases, or drug responses for research purposes, based on non-animal models.

Studies using biological materials or data solely for analysis (i.e., biomarker detection in histological studies, omics studies, etc.) are not considered as models. Humans are not considered models for ethical reasons. They are test subjects under the governance of clinical trial regulations.

3.3. Human biology-based model identification

With the objective of increasing the adoption of methods that could improve the translation of biomedical research, we developed a clear definition for the “human biology-based” models to be selected by our pipeline (Box 1). This proved to be an important task as we realised during the consolidation of data coming from the reviews that biomedical experts have sometime used slightly different definitions for identifying non-animal models (Box 2). The overall decision process for selecting articles relevant to the BimmoH dataset is available in section 4.3.1 (Figure 3).

4. Methodology

To fulfil the above-mentioned objectives, we looked at the different options offered by AI solutions. Although chatbots based on Large Language Models (LLMs) had already gained wide popularity, at the start of our project in 2023, we opted for the well-established supervised Machine Learning approach.

The main reasons for this choice were driven by different elements: (i) the supervised aspect of the process, basing the computer learning from human expertise; (ii) the production of predictable and reproducible results, allowing knowledge transfer; (iii) the possibility to analyse the classifier behaviour, brining transparency and interpretability; and (iv) the low computational power requirements allowing the quick running of experiments to improve classifiers performance. Moreover, this type of classifiers requires a reasonable number of pre-classified articles, estimated around a few thousands (Page *et al*, 2021).

Regarding the research articles data source, we opted for PubMed since it contains the largest aggregated set of information on biomedical research and medicine publications free to use for research purposes, with more than 39 million entries (Sayers *et al*, 2024). Using PubMed data[4], a supervised ML classifier can be trained based on information from publications' title, abstract and, when available, MeSH[5] (Medical Subject Headings) terms.

Considering the systematic review process, our classifier can thus be considered as implementing the automation of "Title and abstract screening" step which is the identification of relevant articles based on title and abstract against defined inclusion/exclusion criteria (Page *et al*, 2021).

The main outcome of the BimmoH project is a data processing pipeline that allows to identify human biology-based models published in the scientific literature and present in the PubMed database. This pipeline was used to create the BimmoH dataset published in the JRC data catalogue and will be launched regularly to update its content over time, starting on a quarterly basis.

Figure 1 describes the BimmoH data processing pipeline, a combination of Python scripts relying on existing APIs and services (PubMed, OpenAlex) to retrieve the data. The pipeline begins with the data preparation that collects the articles to be classified. It starts by retrieving the unique identifiers (PMIDs) of articles from PubMed database using the command-line tool Entrez Direct (step 1). Once the PMIDs are collected, the full texts of the articles are downloaded in Medline text format (step 2) by using Entrez Direct to obtain the plain text versions of the articles metadata, which are then saved as .txt files. Then text files are structured and converted into JSON format (step 3). During this step, data cleaning operations are performed[6].

The second phase of the pipeline deals with the machine learning classification. The cleaned articles are transformed into numerical representations, or embedded, as described in section 4.3.1.2 (step 4). Each embedding is processed by a combination of different ML classifiers (see section 4.3), generating a prediction classification score for each article (step 5). The generated scores are normalised and aggregated using arithmetic means to create an ensemble classifier. Average

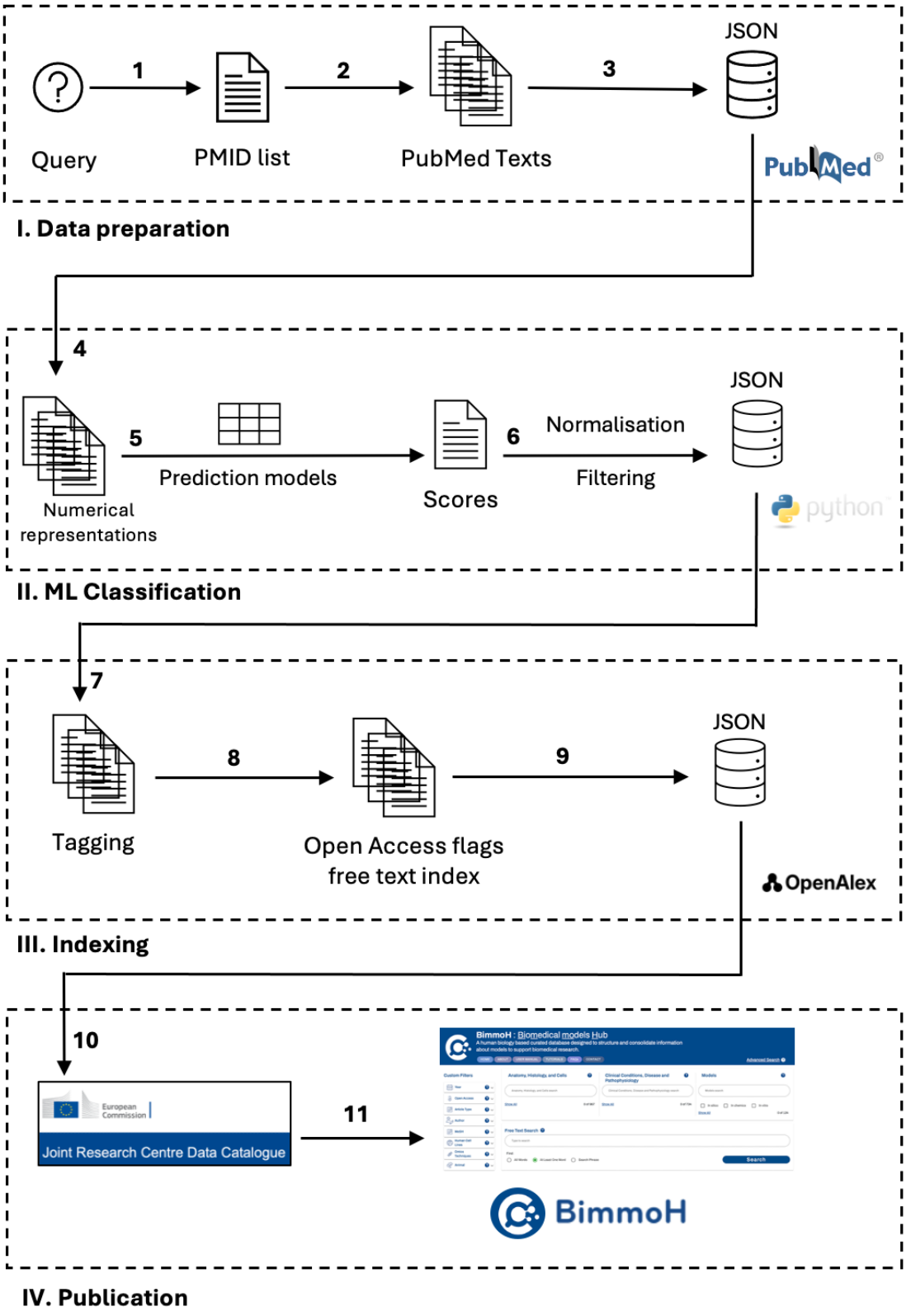
[4] Although the full body of a scientific article contains all the information necessary to decide whether human biology-based models are used, such content is copyrighted and often not allowed to be used to train AI models.

[5] https://www.ncbi.nlm.nih.gov/mesh/

[6] the "year" field is cleaned by removing day and month; the Doi string is removed from the DOI field; articles without a title or abstract are discarded; only complete and properly formatted articles are retained.

scores are computed and only the articles with a score above a defined threshold are kept in the dataset (step 6).

Figure 1. BimmoH dataset creation and update pipeline



Source: EU Commission – Joint Research Centre

The third phase of the pipeline deals with the indexing of selected articles. The tagging script assigns semantic tags (such as Anatomy, Disease, Models, etc.) to each article using regular expressions applied to the title and abstract (step 7). It calculates tags for each file in the source folder in parallel by using controlled vocabularies described in section 4.4.1. Lastly, additional information is retrieved from the OpenAlex dataset to identify if the articles contained in our dataset are published in open access, and to add the index of abstract keywords allowing the search of information using free text (step 8). Finally, the data is aggregated and consolidated in a single JSON file (step 9).

The last phase of the pipeline concerns the BimmoH dataset publication. It is stored in the JRC Data catalogue BimmoH page (step 10) and then imported in the BimmoH web application (step 11).

4.1. Requirements collection

We collected the requirements for the creation of the BimmoH dataset in two phases, interacting with relevant end-user communities (see Annex 2).

In March 2024, we performed a SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis[7], interviewing 16 stakeholders coming from the regulatory domain (4), research or academia (4), industry (4), Three Rs organisations (2) and biomedical research-related consultancy (2). We followed an agile product development interview process, guiding the sessions with a standardised questionnaire focused on user needs, pain points, goals, and feedback on existing solutions. Priority order was subsequently established by ranking these items according to the frequency with which similar requirements were cited during the interviews. To ease the discussions, we developed an early BimmoH prototype to showcase the main concepts.

In October 2024, we organised a workshop at JRC premises to present this early prototype to a panel of 14 experts representing researchers (3), project evaluators (4), industry/pharmaceutical companies (3) and Three Rs promotion organisations (4). The workshop helped us to collect concrete information on how the BimmoH dataset could be supporting the activities of the four identified groups, helping us to finalise our list of requirements.

4.1.1. Must-have key characteristics and functionalities

Based on these two consultations (SWOT analysis and workshop) described above, we agreed in focusing on the implementation of the requirements that we identified as the most appropriate for supporting the BimmoH dataset end-users in their activities.

The ML classifier should detect articles making use of human biology-based models (even if they also incorporate animal models) and exclude those using only animal-based models. To detect potential use of both models in an article, a flag could be added if an animal species name is mentioned in the title or the abstract.

Data consistency and accuracy should be ensured through regular updates thanks to the implementation of an automated pipeline allowing the identification of new articles. It should be comprehensive and cover all diseases and conditions: the ML classifier should look for the use of human biology-based models in articles independently of the biomedical research domain.

[7] Strategic planning tool used to identify and evaluate an organization's Strengths, Weaknesses, Opportunities, and Threats to inform decision-making and strategy development.

To support advanced search capabilities, including filters and keywords, the BimmoH dataset should rely on specific vocabularies developed to ease targeted research of human biology-based models. It should provide detailed metadata for each model type (*in vitro*, *in silico*, *in chemico*), by including hundreds of known models. It should also provide direct link access to original publications using the DOI (Digital Object Identifier) provided by PubMed.

It should allow to export data and to search results in various formats through the provision of a responsive and intuitive web interface for easy navigation, which ensures high performance and responsiveness of the database search functionality. The complete dataset should also be made available for download to be used by data scientists.

Training materials and resources for new users should be prepared, for example: user manual and tutorial videos, including visual aids and infographics to enhance data presentation.

4.1.2. Deprioritised requests

Some of the requirements we received during the interviews and workshop could not be implemented for different reasons, linked to their technical or legal feasibility. Most of these requirements are legitimate requests, and it is important to acknowledge them to avoid misunderstanding on the scope of the BimmoH dataset.

The BimmoH dataset does not provide bibliographic references and citations to avoid cognitive bias. This is a decision we took to avoid the impression that one model would be more relevant than another based on this information.

Distribution of the full text of articles is not always allowed for both open access[8] and pay-walled articles. As scientific articles are copyrighted, their content cannot be shared by third parties. Nevertheless, the BimmoH dataset should contain the DOI link to the article, allowing the access to the full text in case of the article is open access. This is why BimmoH mentions if the article is in open access or not.

Similarly, though it would increase its performance, making use of the Materials and Methods sections by the ML classifier is not allowed for copyright reasons**.** Most of the information relevant to understand the details of a model are presented in this section, but as this part of the articles is copyrighted[9], it cannot be used for the BimmoH dataset.

The BimmoH dataset leverages the power of supervised ML classification to treat a large amount of data that could not be performed manually. For this reason, the dataset is not intended be manually corrected with the addition or removal of specific articles. Instead, the datasets created to train the ML algorithm are manually curated and peer reviewed.

BimmoH's pipeline selects relevant articles making use of human biology-based models. It does not have neither the objective nor the technical capability to judge or interpret the content of the identified models. Therefore, we do not assess model performance or validation. It is not possible either to identify if the models adhere to specific guidelines.

[8] Distribution of open-access articles is only possible when they are open-licensed

[9] Only articles published in open access under CC-BY 4.0 licence can be used freely, which represents less than 10% of the total number of articles available in PubMed: https://ncbiinsights.ncbi.nlm.nih.gov/2021/09/01/pubmed-central-article-datasets-cloud

Box 2. Main differences between the scope of the BimmoH dataset and the seven EURL ECVAM biomedical reviews

The feedback received by stakeholders led us to important conceptual extensions compared to the dataset we had created in the context of our biomedical reviews (Section 2).

Not limited to fixed biomedical research areas. Instead of focusing on dedicated areas, BimmoH identifies relevant models in all fields of biomedical research.

Not limited to a specific period. BimmoH algorithm can be launched regularly to capture new articles that are published and it also allows the identification of models published before 2014, the starting date of our biomedical reviews.

Including articles containing both human biology-based and animal-based models. As many publications report the use of a combination of human biology-based and animal-based models, BimmoH includes all articles mentioning the use of at least one human biology-based model.

Enriching data with specific vocabularies. To index information, BimmoH uses specific vocabularies that were created with the aim of being more precise than the metadata used in PubMed for retrieving relevant human biology-based models.

Including both reviews and original research articles. While the biomedical reviews were focusing only on peer-reviewed articles, BimmoH includes reviews as well, as they can provide useful information in specific areas.

Limiting the analysis to title and abstract for selection. While the previously developed biomedical reviews followed a tiered approach for selecting articles in which the full content of the articles was reviewed by human experts in the last step, BimmoH algorithms only review title and abstract of articles (for performance and intellectual properties reasons).

Identifying animal-derived material is not in the scope of BimmoH but the dataset will contain a flag indicating whether the title or the abstract mention animals in the text.

4.2. Data preparation

Data preparation is an essential step for building a ML classifier. This section explains which strategy we put in place to identify relevant articles making use of human biology-based models in PubMed.

4.2.1. Biomedical reviews data consolidation

Our primary data source comprises a collection of seven Excel files available in the JRC data catalogue, each containing a set of carefully curated articles that focus on human biology-based models across seven distinct disease areas (breast cancer, immunogenicity testing for ATMPs, autoimmunity, immuno-oncology, respiratory, neurodegenerative, and cardiovascular diseases).

Table 1. Consolidated biomedical reviews metadata structure

Name	**Definition**
Disease	Names the disease or pathogenetic process studied
Organ	Physiological organ targeted in the research
Tissue	Identifies the specific tissue studied
Method Domain	Cell-free; e*x vivo; in silico; in vitro*
Method Type	Algorithms; Biochemical; Biopsies; Cells; Computational; Finite-element model (FEM); Mathematical; n/a; Simulation; Tissue Explant; Whole Organ
Cell Type	Immortalised cells; n/a; Primary cells; Stem cells; Stem cell-like
Cell Culture Type	Co-Culture; Culture; n/a; MPS (microphysiological systems)
3D Model Type	n/a; Organ slice; Organoids; Scaffolds; Spheroids; Whole organ
Use of Omics	Yes; No; n/a
Title	Study title
Abstract	Study abstract
Identifiers	DOI or PMID or PMCID or ISSN: identifiers to retrieve or link to the publication, facilitating access to the source
First author name	Name of the first author of the peer-reviewed article
Corresponding author name	Name of the corresponding author of the peer-reviewed article
MeSH terms	Medical Subject Headings terms used by PubMed

Source: EU Commission – Joint Research Centre

To satisfy the project objective of extending this collection to additional sectors of the biomedical research (see section 3.1), the contractor's biomedical experts created a selection of additional articles related to gastro-intestinal, metabolic diseases, and infectious diseases. They grouped all

articles' references in a single Excel file[10] containing the metadata listed in Table 1 (Deceuninck *et al*, 2025).

During the consolidation, some of the articles listed as entries in the file could not be retrieved from PubMed. In this case, they were removed from the consolidated list.

4.2.2. PubMed data

To access a large set of biomedical research articles, we decided to use the PubMed database provided by the US National Library of Medicine. PubMed is a free search engine primarily accessing the MEDLINE database of references and abstracts on life sciences and biomedical topics. It contains a wide range of data as well as Application Programming Interfaces (APIs) allowing automated content retrieval in machine readable formats.

PubMed contains research articles, reviews, clinical studies, case reports, editorials and commentaries, correspondences, and systematic reviews. Focusing on the identification of human biology-based models and based on the feedback received by our stakeholders, we analysed exclusively original research articles and reviews. For each publication, PubMed provides a standard format covering most of the information we needed for performing an ML-based classification to identify the ones using human-biology based models.

Table 2. PubMed metadata structure

Name	**Definition**
Title	Title of the article as published in the journal.
Abstract	Summarised version of the article that provides an overview of the study's objectives, methods, results, and conclusions.
Full-Text Link	While PubMed itself does not host full-text articles, it includes links to the publisher's site or repositories where the full text can be accessed. Some articles are available for free, while others may require a subscription or purchase.
Citation Information	Detailed citation data including authors, journal name, publication date, volume, issue, and page numbers.
MeSH Terms (Medical Subject Headings)	Standardised keywords and phrases used to index articles, helping users find related content effectively.
DOI and PMID	Digital Object Identifier (DOI) and PubMed Identifier (PMID) are unique identifiers for articles, facilitating easy retrieval and citation.

Source: National Library of Medicine

Overall, PubMed serves as a comprehensive repository of biomedical literature, offering a broad spectrum of resources for research and clinical practice, matching our data sources with respect to the format of the metadata (Table 2).

[10] https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/EURL-ECVAM/datasets/BimmoH/LATEST/BiomedReviews_Consolidated.xlsx

4.2.3. PubMed candidates pre-filtering

PubMed contains more than 39 million citations and abstracts of biomedical literature that can be retrieved using key filters. Since our study focused on models, we decided to reduce the number of entries to be analysed by designing a PubMed query that contains:

1. **Human biology-based models related terms:** all relevant models that we have identified, according to inclusion criteria (e.g., Organ-on-chip, cell culture, organoid)
2. **Exclusion criteria:** for scientific articles unlikely to contain models as they do not focus on biomedical research experiments relying on model usages (e.g., clinical trials, case study, prognosis)
3. **Specific article types:** limiting the selection to research articles and reviews.

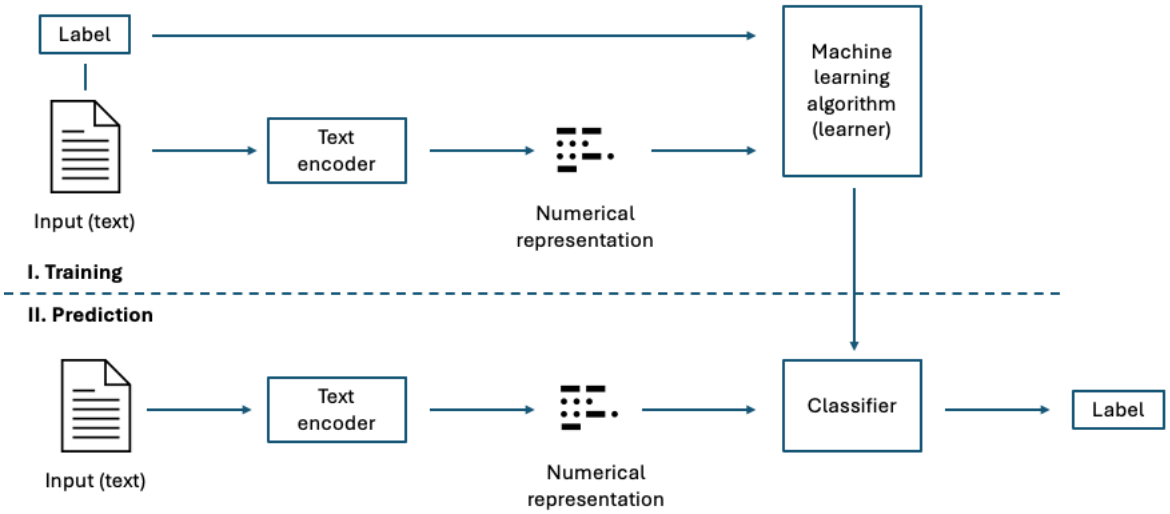
The designed query retrieved a pool of approximately 4.6 million PubMed references more likely to make use of human biology-based models, ensuring a wide and representative sampling of the domain. The complete PubMed query is available in Annex 3.

Some results (less than 10%) contained titles or abstracts that were missing or truncated due to embedded HTML tags or encoding errors. These incomplete or missing records were excluded by adding a filtering process limiting the selection of candidate articles to 4.3 million.

4.3. Supervised Machine Learning Classification

Supervised machine learning is composed of two steps: training and prediction. During the training phase, the computer learns form labelled inputs how to make the classification. During the prediction phase, the classifier uses what it learned to make a prediction and associate a label to unknown inputs (Figure 2).

Figure 2. Main steps of supervised machine learning classification



Source: EU Commission – Joint Research Centre

Training a machine learning classifier requires an initial set of texts (in our case scientific articles) associated to a label providing information on the category they belong to (see section 4.3.1.1). These texts are transformed into numerical representations by a text encoder (see section 4.3.1.2)

that can be read by the computer and used as a source of information by a machine learning algorithm (see section 4.3.1.3).

The result is a classifier that can be used for predicting the category an unknown article belongs to. The prediction of the classifier is a score ranging from 1 (articles making use of human biology-based model(s)) to 0 (articles not containing any human-biology based model) used to assign a new label (only articles scoring above this threshold are relevant).

ML is efficient with a relatively low amount of trained data for the classification of large data sets. Knowing that the above candidate data set contained more than four million scientific articles, this could be achieved with a training set constituted of few thousand articles (Koshute *et al*, 2021; Riekert and Klein, 2021).

4.3.1. Methodology

To build the ML classifier used by the BimmoH pipeline, we carried out various experiments to test different training sets, numerical representations, algorithms, and validation strategies. This helped us to understand how to improve the classification step by step but also to learn about the key factors to maximise the performance of our classifier.

4.3.1.1. Training set

Training sets are the foundation for supervised ML classifiers as they need to be trained on labelled data that will allow them to decide to which category the data belong to. In our case, positive instances are articles that contain human biology-based models and negative data all the others.

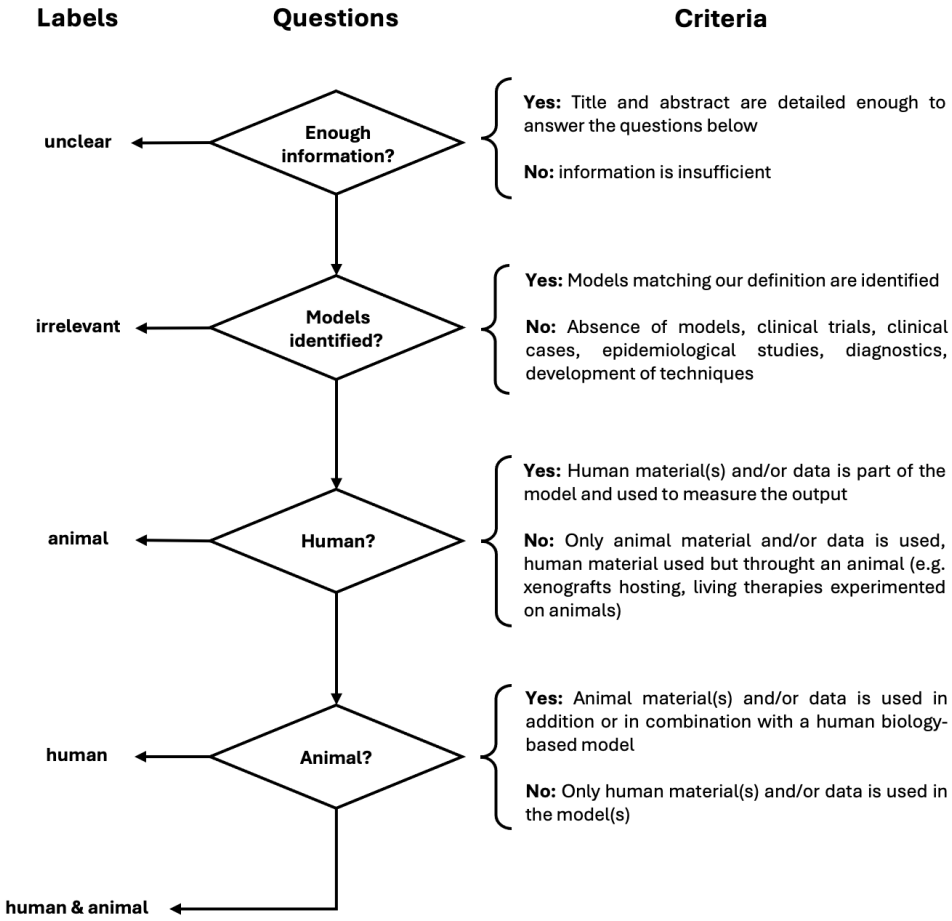
Building the training set is a challenging task. To perform well on target data (in our case the data coming from the PubMed candidates query described in section 4.2.3), the training set needs to be representative. This means that the composition of the training set needs to contain all types of articles we can find in the data we are going to classify.

If the training set is not representative, the ML classifier will be biased. It will be able to classify properly the training data, but not the target data as it will encounter data that were never seen during its training. Retrieving representative data in a sufficient amount to be able to train the supervised ML classifier is one of the keys. Proceeding in an iterative way, we decided to build our training set by identifying useful labels and reflecting the diversity of the target data, step by step, relying on complementary approaches for the data collection.

Firstly, two curated lists of scientific articles were used for the construction of our training set. The consolidated selection of biomedical reviews articles was our starting point (see section 4.2.1). In addition to it, a relevant reference of articles making exclusive use of animal-based models was essential as our primary exclusion criteria. This one is composed of a curated set of articles retrieved from an animal-based gene bank[11] (Ali Khan, 2023) and filtered to ensure exclusive focus on animal experimentation and biology.

[11] https://www.infrafrontier.eu/about-us/bibliography/infrafrontier-bibliography/

Figure 3. Labelling process for the manual validation of articles



Source: EU Commission – Joint Research Centre

Secondly, we assessed manually the best predictions of this baseline classifier recording the expert decision by assigning labels to the articles (see process described in Figure 3 above). This allowed us to extend significantly the training set with another 3,000 articles manually labelled by biomedical research experts. Analysing in detail the composition of this dataset, we convened to label our training set using the following categories:

- **Positives**: articles making use of human biology-based models
    - **Exclusive use of human biology-based models**
    - **Combined use of both human biology and animal-based models**
- **Negatives**: articles not making use of human biology-based models
    - **Irrelevant** (not making use of any model), as dealing with:
        - **Clinical trials** (biomedical research study involving human participants, conducted to evaluate the safety, efficacy, or optimal use of medical interventions such as drugs, devices, or treatment strategies.)
        - **Diagnosis** (process of identifying a disease, condition, or injury in a patient through the evaluation of clinical signs, symptoms, medical history, and diagnostic tests.)

- **Profiling** (analysis of a set of biological characteristics to comprehensively characterise a sample, condition, or patient group.)
- **Sequencing** (determination of the precise order of nucleotides in a DNA or RNA molecule to analyse genetic information at the molecular level.)
- **Other** irrelevant articles

- **Exclusive use of animal-based models**

This part of the training set came from the different review sessions assessing the quality of intermediate ML classifiers. To avoid biases each article was peer-reviewed by domain experts, blinded to the model prediction.

Thirdly, to balance the training sets and strengthen the negative training set with a representative repartition, articles from PubMed, focused on topics such as profiling, sequencing, and diagnosis, were incorporated and labelled as excluded. The new references were retrieved using a series of targeted PubMed queries, each designed to capture specific thematic areas by applying boolean expressions.

Finally, expert-designed queries were developed using advanced combinations of keywords to retrieve articles specifically matching the criteria for the four subgroups identified (Human only, Human and Animal, Animal only, Irrelevant), indexing a subset of 200,000 articles randomly sampled from the query described in Annex 3. This resulted in a selection of more than 18,000 articles for which a curated classification was created by applying a peer-review process on random samples.

By analysing the type of articles present in the negative sets we were also able to estimate the proportion of each category of articles needed to represent properly target data. To balance the number of negative data in the training set, we selected from these PubMed queries around 8,000 irrelevant articles dealing with clinical trials, sequencing, profiling, and diagnosis.

The result was a set of text files associating the PubMed Identifiers (PMIDs) to the correct label for almost 40,000 articles. Table 3 describes the repartition of labels present in the training set and the origin of the data (steps of the selection process).

Table 3. Training set labels repartition

Selection process steps	Positive set		Negative set	
	Human	**Human and Animal**	**Irrelevant**	**Animal only**
1. Curated lists	3046	0	0	2888
2. Expert reviews	2214	1621	2134	520
3. PubMed queries	0	0	8000[1]	0
4. Expert queries	4629	2187	6523	4958
Total	9889	3808	16657	8366

Source: EU Commission – Joint Research Centre

[1] With the following breakdown: Clinical trials: 2,000, Diagnosis: 3,000, Profiling: 1,500, Sequencing: 1,500

4.3.1.2. Numerical representations

Numerical representation is the process of converting unstructured text data (words, sentences, documents) into structured numerical vectors, in a continuous space, where the geometry of that space captures meaningful relationships between the items. Similar inputs (e.g., "cell" and "neuron") are mapped to vectors that are close together, while unrelated ones (e.g., "cell" and "hospital") are farther apart. They allow ML algorithms to work with semantic meaning instead of raw text, enabling tasks like classification, clustering, search, and recommendation. We investigated the use of the following three types of numerical representations: TF-IDF, sentence transformers, and Doc2Vec.

TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word in a document relative to a corpus (Spärk Jones, 1972). TF-IDF combines two metrics: Term Frequency (TF), which measures how often a word appears in a document, and Inverse Document Frequency (IDF), which measures how important a word is across the corpus. This method is effective in reducing the weight of common words and highlighting unique terms. In text classification, TF-IDF transforms text data into numerical vectors, facilitating the use of ML classifiers.

We tested two different TF-IDF configurations, using the python library "NumPy" v1.26.4 (Harris, 2020):

- TF-IDF 1: 200,000 terms (groups of one to three words, excluding function words like "a", "the", etc.) coming from 10,000 articles returned by PubMed for the query "disease", and keeping only terms appearing in more than 1% and less than 90% of these articles[12].
- TF-IDF 2: 40,000 terms (groups of one to three words, excluding function words like "a", "the", etc.) coming from 200,000 articles randomly selected from target data (section 3.2.3) and keeping only terms appearing in more than 0.05% and less than 90% of these articles[13].

Doc2Vec

Doc2Vec is a neural network-based technique that converts text into continuous vector representations, capturing semantic relationships between words and the overall context of the text (Le and Milikov, 2014). Unlike Word2Vec, which generates embeddings for individual words, Doc2Vec produces a unique vector for each document, enabling the representation of variable-length texts as fixed-size vectors. This document-based vector is then used for generating the embedding for the given text.

We tested two different Doc2Vec configurations, using the python library "Gensim" v4.3.2:

- Doc2Vec 1: 50 vector sized representation of articles created from a set of 100,000 articles returned by PubMed for the query "disease" with 40 iterations (epochs) keeping words appearing more than 3 times[14].

[12] TfidfVectorizer(stop_words='english', ngram_range=(1, 3), min_df=0.01, max_df=0.90)

[13] TfidfVectorizer(stop_words='english', ngram_range=(1, 3), min_df=0.0005, max_df=0.90)

[14] Doc2Vec(vector_size=50, min_count=3, epochs=40)

- Doc2Vec 2: 768 vector sized representation of articles created from a set of 200,000 articles randomly selected form target data (section 4.2.3) with 30 iterations (epochs), keeping only the most frequently appearing (more than 15 times in the collection)[15].

Sentence transformers

Sentence transformers is a framework for generating sentence and text embeddings[16] using transformer-based models, such as BERT, RoBERTa, and others (Reimers and Gurevych, 2019). Sentence transformers provide embeddings for entire sentences or texts (instead of single words), capturing their semantic meaning. These embeddings are highly effective in tasks requiring understanding of sentence or document-level semantics, such as text classification. In addition, these vectors can be used for document clustering and semantic searches. Many approaches are available for creating transformer-based embeddings.

We used two of them, downloaded from https://huggingface.co :

- Transformer 1 (all-MiniLM-L6-v2): lightweight Sentence Transformer that generates high-quality 384-dimensional sentence embeddings, optimised for semantic similarity and clustering tasks, offering an excellent balance between speed and accuracy[17].
- Tranformer 2 (pubmedbert-base-embeddings): PubMedBERT-base fined-tuned using sentence-transformers. It maps sentences & paragraphs to a 768-dimensional dense vector space and can be used for tasks like clustering or semantic search. The training dataset was generated using a random sample of PubMed title-abstract pairs along with similar title pairs[18].

4.3.1.3. Machine Learning algorithms

The algorithm's goal is to generalise from the training data so it can correctly predict labels for new, unseen data. The algorithm applies a learning process, adjusting its internal parameters to minimise error between predicted labels and true labels. Using python library "NumPy" v1.26.4 (Harris, 2020), we investigated the following four algorithms for the BimmoH project: Random Forest, Logistic regression, Adaptive boosting and Gradient boosting.

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their outputs to improve classification accuracy (Breiman, 2001). Each tree is trained on a random subset of the data and features, which helps in reducing overfitting and increasing robustness. In text classification, Random Forest can handle large feature spaces and complex decision boundaries. It is less sensitive to noisy data and provides feature importance scores, which can be useful for understanding the significance of individual features. Combined with a human-interpretable feature matrix, such as generated by TF-IDF, this allows for interpretation and understanding of the models in terms of which words are used for classification.

---

[15] Doc2Vec(vector_size=768, min_count=15, epochs=30)

[16] In machine learning, embedding is a representation learning technique that maps complex, high-dimensional data into a lower-dimensional vector space of numerical vectors: https://en.wikipedia.org/wiki/Embedding_(machine_learning)

[17] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[18] https://huggingface.co/NeuML/pubmedbert-base-embeddings

Logistic Regression

Logistic Regression is a linear model used for binary and multiclass classification tasks (Bewick *et al*, 2005). It models the probability of a text belonging to a specific class based on a linear combination of input features. In text classification, it is often employed with feature extraction techniques such as TF-IDF or word embeddings. Logistic Regression is interpretable, efficient, and performs well with large datasets.

Adaptive Boosting

AdaBoost (Adaptive Boosting) is an ensemble method that combines the predictions of multiple weak classifiers to form a strong classifier (Freund and Schapire, 1997). It works by iteratively training weak learners on the dataset and assigning more weight to misclassified instances in subsequent rounds. This method helps in focusing on difficult-to-classify instances. In text classification, AdaBoost can effectively handle high-dimensional data and improve model accuracy by reducing overfitting and enhancing generalisation.

Gradient Boosting

Gradient Boosting constructs an ensemble of weak learners, where each new model is trained to minimise the loss function (e.g., mean squared error or cross-entropy) of the previous model using gradient descent, so that the combined ensemble forms a strong classifier (Friedman, 2001). In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimise this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met. In contrast to AdaBoost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of the predecessor as labels.

4.3.2. ML classifiers construction

4.3.2.1. Experimental design

For our experiments, we relied on 12 combinations of numerical representations and algorithms (Table 4) to create classifiers that could be trained on different types of training sets. By evaluating many learners, we wanted to identify the combination(s) that would construct the best classifier given our data set.

Table 4. Combinations of learners evaluated during our experiments

	Random Forest	**Logistic Regression**	**Adaptive Boosting**	**Gradient Boosting**
TF-IDF	Model 1 (TF/RF)	Model 4 (TF/LR)	Model 7 (TF/AB)	Model 10 (TF/GB)
Doc2Vec	Model 2 (D2V/RF)	Model 5 (D2V/LR)	Model 8 (D2V/AB)	Model 11 (D2V/GB)
Sentence transformer	Model 3 (Trans/RF)	Model 6 (Trans/LR)	Model 9 (Trans/AB)	Model 12 (Trans/GB)

Source: EU Commission – Joint Research Centre

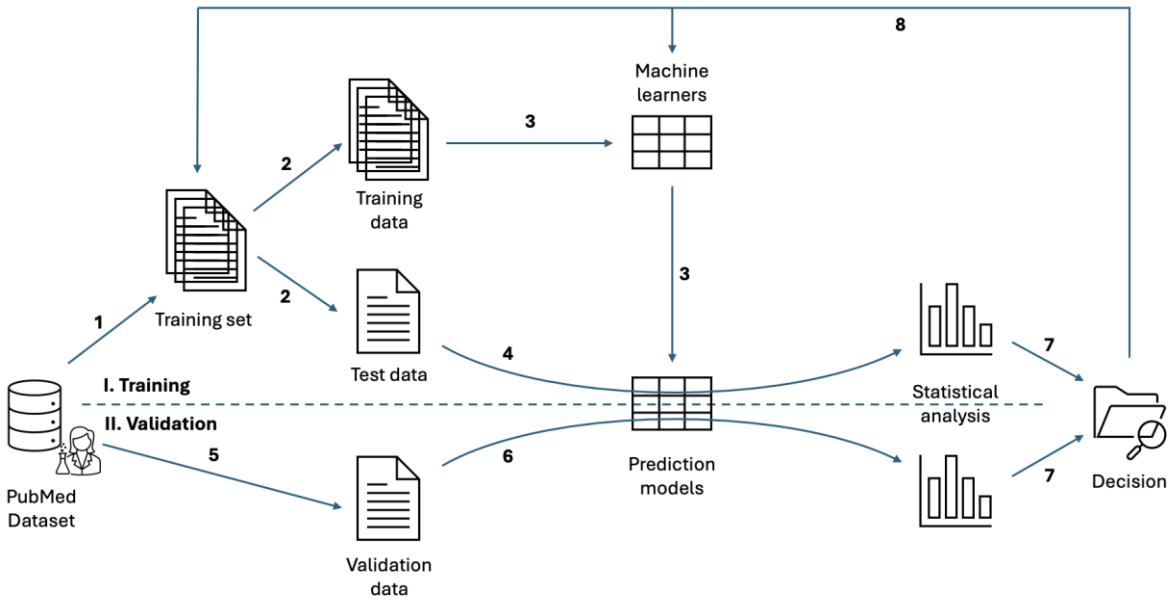
To achieve the construction of our ML classifier, we decided to perform various experiments, based on a defined experimental approach (Figure 4), playing on its different components.

In the first part (I. Training), we choose a training set composed of PubMed data selected within our section 4.3.1.1 training set (step 1), the training set is split in two (step 2). 80% is used as training data to train the machine learners and create the prediction models (step 3) scoring articles between 0 and 1. The remaining 20% is used for testing the prediction models performance by scoring each article and performing a statistical analysis of the results (step 4).

In the second part (II. Validation), we select validation data (step 5), coming from the section 4.2.3. PubMed filtered data, manually classified by experts, constituted of a balanced set of positives and negatives, sampled based on the iteration needs (e.g., focusing on assessing high positives or negatives or on the contrary, more complex cases, harder to classify). We then classify the validation data using the prediction models created during the training part (step 6) and perform a statistical analysis of the results.

We then compare the results expected by the test data issued from the training set with the results obtained on validation dataset (step 7). This allows the assessment of the performance of the models which is essential to make sure that the training set is representative of the whole PubMed candidate dataset. Depending on outcome, we adjust them and repeated the process in a further iteration (step 8).

Figure 4. Experimental approach



Source: EU Commission – Joint Research Centre

4.3.2.2. Evaluation Metrics

To assess the performance of the prediction models, we used a confusion matrix (Table 5) for comparing the model predictions with the known data labelled by experts (Test data or Validation data).

Table 5. Confusion matrix

		Known data	
		Positive	**Negative**
Predicted data	**Positive**	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

Source: EU Commission – Joint Research Centre

This matrix provides a detailed view of the types of errors a prediction model makes and forms the basis for following metrics:

Sensitivity (or Recall): percentage of positives correctly identified by the classifier. It tells us how performant the classifier is to identify the positives. The higher this percentage is, the better the classifier identifies articles making use of human biology-based models.

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN}$$

Specificity: percentage of negatives correctly identified by the classifier. It tells us how performant the classifier is to identify the negatives. The higher this percentage is, the better the classifier can exclude articles that are either irrelevant or making use of animal-based models.

$$Specificity = \frac{TN}{TN + FP}$$

Precision: percentage of real positives amongst the predicted positives by the classifier. It tells us how good the data selected by the classifier is. The higher this percentage is, the better the dataset of articles making use of human biology-based models will be (the lower the number of incorrect articles it will contain).

$$Precision = \frac{TP}{TP + FP}$$

Accuracy: percentage of correct predictions made by the classifier. It tells us how often the classifier correctly predicts if an article is positive or negative. The higher this percentage is, the better the classifier is in making predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

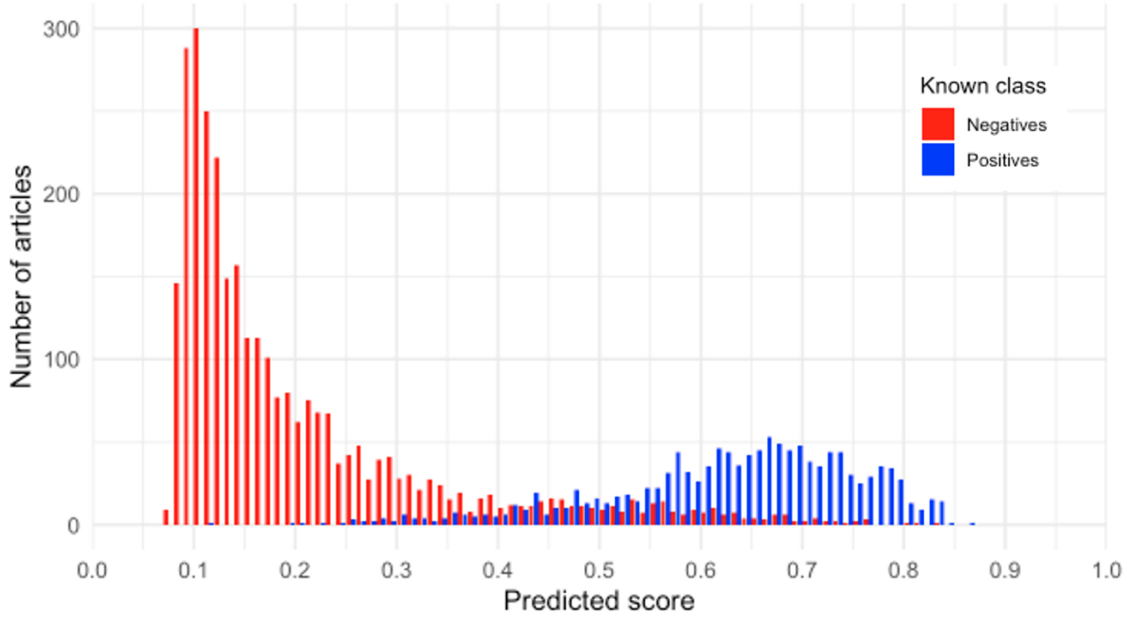
F1 score: single number that measures how well a model balances precision (how many predicted positives are correct) and recall (how many real positives are found). It tells us how the classifier performs on both recall and precision at the same time. This number is high only if both precision and recall are high.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Knowing that our objective is to create a dataset containing the lowest number of incorrect articles and considering that our pre-filtered PubMed dataset was likely to contain more negatives than positives, our priority was to obtain the highest possible **Specificity (above 95%)** targeting a high **Precision (above 90%)**. To retrieve enough articles making use of human biology-based models but not knowing exactly how many of them our PubMed pre-filtered dataset would contain, we targeted a **Sensitivity of 50%**, ensuring that more than half of them would be retrieved.

Figure 5 shows an example of the repartition of scores for test data. In red, the negative articles that should be excluded while the ones in blue should be included in the final data set. To be able to define articles that are predicted as positives or negatives, we need to define a threshold for which the articles with a lower score will be predicted as negatives while the articles with a higher score will be predicted as positives. Between 0.2 and 0.8 there is an overlap of positive and negative articles: depending on the chosen threshold, we can play on the specificity (and therefore on the sensitivity as well). The higher the threshold is, the higher will be the specificity, but the lower the sensitivity will be.

Figure 5. Repartition of predicted scores for test data

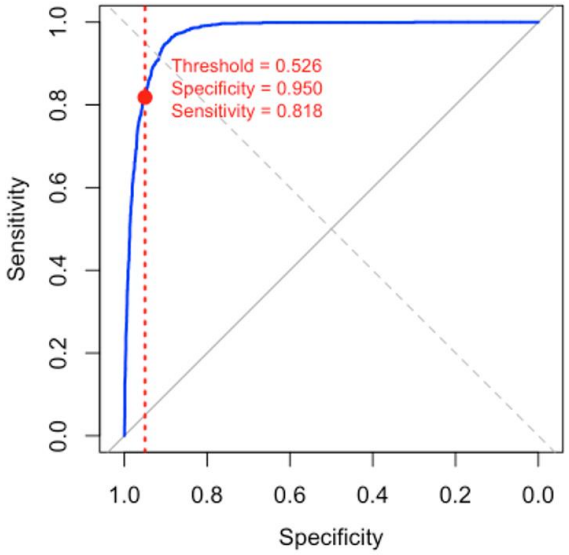


Source: EU Commission – Joint Research Centre

This can be achieved using the **ROC curve (Receiver Operating Characteristic)**: a graph that shows how well a classification model can distinguish between positive and negative samples (Fawcett, 2006). It illustrates the trade-off between a model's Sensitivity represented on the Y-axis and Specificity represented on the X-Axis. The ROC curve shows how a classifier's ability to detect true positives compares to the rate of false positives, across all possible decision thresholds.

Drawing the ROC curve can be achieved by calculating both Sensitivity and Specificity for all decision thresholds going from 0 to 1. The result is the curve presented in Figure 6. It starts with a Specificity of 1 (100%) and a Sensitivity of 0 as, for the threshold fixed at 0, all articles are considered as negatives, meaning that it predicts successfully all negatives. On the contrary, all positives are predicted as negatives meaning that none of them are predicted positives.

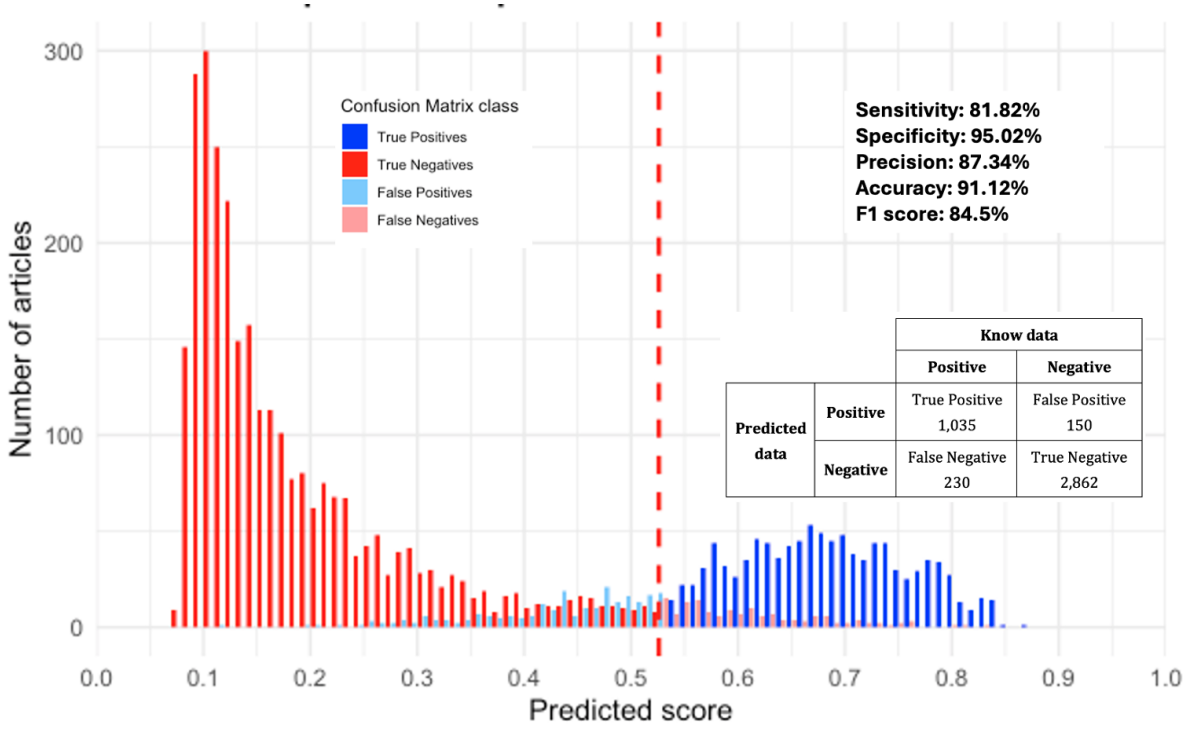
Figure 6. ROC curve for Test data presented in Figure 3



Source: EU Commission – Joint Research Centre

With the ROC curve, we can define both Sensitivity and Specificity for each threshold. We can also fix a defined Specificity or Sensitivity and determine the threshold to obtain it. In our case, we were particularly interested by having a very high Specificity while we could afford a lower Sensitivity. In line with the objectives described above, we decided to fix the Sensitivity at 95%. Therefore, we calculated the threshold at this level. Figure 7 shows that for the Test data represented in Figure 5, the threshold should be fixed at 0.526 to obtain a Specificity of 95% and a Sensitivity of 81.8%. Once the threshold is fixed, other metrics can be calculated.

Figure 7. Metrics calculations based on the threshold determined by Figure 4



		Know data	
		Positive	Negative
Predicted data	Positive	True Positive 1,035	False Positive 150
	Negative	False Negative 230	True Negative 2,862

Source: EU Commission – Joint Research Centre

4.3.2.3. Data analysis

As described in Figure 4, we relied on different datasets to estimate the performance of the classifiers we created. We named them: Test dataset, Validation dataset and Target dataset.

Test dataset

The test dataset is composed of a fixed set corresponding to 20% of the articles in the training set and is unknown to the classifiers (typically 1,250 positive and 3,000 negative articles).

To assess the overall performance of our classifiers on the test data, we computed the mean of each performance metric across independently trained classifier configurations. As we are interested in highly specific classifiers, we set the threshold for each classifier to achieve a specificity of 95%.

The variability across models was quantified using the empirical variance of the metric values. Ninety-five percent confidence intervals were constructed using the normal approximation based on the standard error of the mean across models.

Validation dataset

The validation dataset consists of 3,000 articles manually classified by experts (600 clear positives, 900 positives close to the threshold, 900 negatives close to the threshold and 600 clear negatives).

To estimate generalisation performance on the validation data, we built an ensemble classifier by averaging all individual scores and calculated the threshold to obtain a specificity of 95%.

Ninety-five percent confidence intervals were constructed using binomial confidence intervals for metrics that can be expressed as proportions (sensitivity, specificity and accuracy). For derived metrics, asymptotic or approximate methods were applied.

Target dataset

The target dataset consisted of 2,150 articles randomly sampled from the PubMed candidates described in Section 4.2.3, representing the intended articles to be classified.

To estimate performance on this target dataset, we built an ensemble classifier by averaging the individual scores of the three best-performing classifiers and fixed the decision threshold to achieve a target specificity of 95%.

Ninety-five percent confidence intervals were constructed using binomial methods for metrics expressed as proportions (sensitivity, specificity, and accuracy). For derived metrics, asymptotic or approximate methods were applied.

4.3.3. Experiments

To prove the validity of the supervised machine learning classifier approach to create the BimmoH dataset, we started by building a basic proof of concept, using the consolidated biomedical reviews as positives and a random set of PubMed articles containing the word “Disease”. We built the 12 classifiers described in section in table 4, using the protocol described in Annex 4 (5-fold cross validation).

This basic approach gave positive results for all 12 models on the test data. By manually assessing validation data coming from the 3,000 highest prediction scores, we obtained a precision of 85%,

confirming the promises expected by the chosen approach. We then investigated how to optimise it playing on the different components of the classifiers.

4.3.3.1. Training set optimisation

In our first experiment, the training set was composed of about 6,250 positive articles (3,000 coming from the consolidated list of biomedical reviews and 3,250 of expert reviews performed on the proof of concept described above) and 15,000 negatives (3,000 articles dealing with animal-based models only, 10,000 articles related to clinical trials and 2,000 irrelevant articles not containing any model).

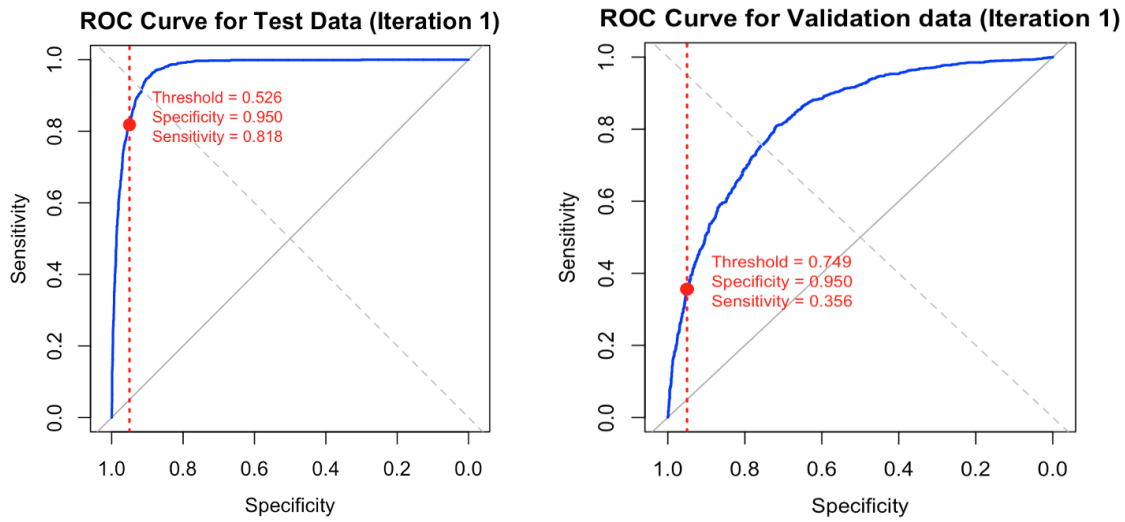
Table 6. Experiment 1 results (see also Annex 5)

	Sensitivity	Specificity	Precision	Accuracy	F1 Score
Test Data	68.84 ±7.46%	95.03 ±0.02%	84.94 ±1.46%	87.28 ±2.21%	75.59 ±5.20
Validation Data	35.60 ±2.61%	95.01 ±1.08%	82.88 ±3.40%	71.00 ±1.61%	49.81 ±2.62

Source: EU Commission – Joint Research Centre

Although test data results met the objectives established in section 4.3.2.2, this was not the case for our Validation data (Table 6). Sensitivity was significantly lower (35.6% against 68.84%), as well as the accuracy (69.40% against 87.28%). This indicated a generalisation issue that could indicate a lack of alignment between our training set and the PubMed candidate data as shown by the ROC curves (Figure 8).

Figure 8. Experiment 1 ROC curves



Source: EU Commission – Joint Research Centre

For the second experiment, we refined the composition of the negative training set, introducing more different types of irrelevant articles. This time it was composed of around 13,000 articles (3,000 articles dealing with animal-based models only, 2,000 with clinical trials, 3,000 with diagnosis related topics, 1,500 with profiling, 1,500 with sequencing and 2,000 other irrelevant kind of articles). We kept the positive training set untouched and used the same validation data.

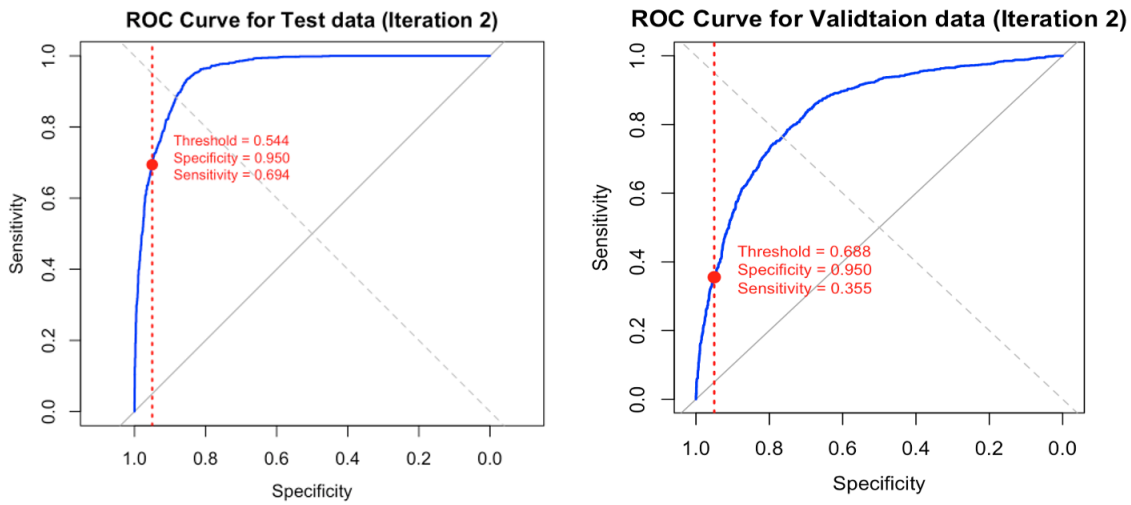
Table 7. Experiment 2 results (see also Annex 5)

	Sensitivity	Specificity	Precision	Accuracy	F1 Score
Test Data	54.96 ±7.65%	95.03 ±0.01%	83.97 ±1.82%	81.75 ±2.54%	65.82 ±6.08
Validation Data	35.53 ±2.61%	95.01 ±1.08%	82.88 ±3.40%	70.97 ±1.61%	47.73 ±2.62

Source: EU Commission – Joint Research Centre

Test data results were slightly lower but still close to our objectives which can be explained by the greater variety of negative articles making the classification task more difficult. Regarding Validation data we could not observe any meaningful change in the results. Comparing the ROC curves, we observed that for this second experiment, the ROC curve of the test data was slightly closer to the validation data (Figure 9). We decided to keep this training set for the following experiments related to embeddings comparison.

Figure 9. Experiment 2 ROC curves



Source: EU Commission – Joint Research Centre

4.3.3.2. Numerical representations comparison

In our third experiment, we changed the methods used to create the numerical representation of texts, moving from generic types (used in experiment 2) to more representative types (see section 4.3.1.3).

We assessed the performance of six methods used for creating the numerical representations of texts described in section 4.3.1.2. Although we noticed some improvement with transformer 2 method (pubmedbert-base-embeddings, trained on PubMed title and abstracts), TF-IDF was the best performing approach.

We kept the TF-IDF-2 numerical representation for the following experiments.

Table 8. Experiment 3 results (see also Annex 5)

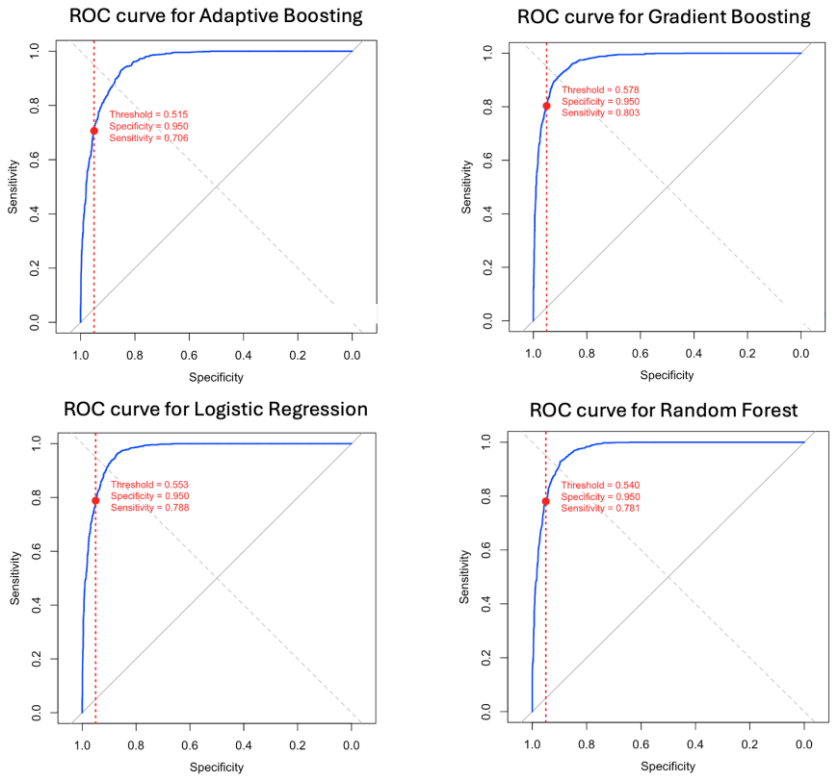
	Sensitivity	Specificity	Precision	Accuracy	F1 Score
TF-IDF 1	71.52 ±4.47%	95.03 ±0.02%	87.68 ±0.66%	87.24 ±1.47%	78.74 ±3.03
TF-IDF 2	71.36 ±4.31%	95.02 ±0.00%	87.64 ±0.68%	87.18 ±1.43%	78.63 ±2.94
Doc2Vec 1	43.16 ±3.38%	95.03 ±0.02%	81.10 ±1.27%	77.84 ±1.13%	53.08 ±3.22
Doc2Vec 2	42.77 ±6.74%	95.04 ±0.04%	80.82 ±2.34%	77.71 ±2.23%	55.78 ±6.22
Transformer 1	50.20 ±7.51%	95.02 ±0.00%	83.12 ±2.22%	80.16 ±2.44%	62.42 ±6.48
Transformer 2	64.80 ±7.33%	95.03 ±0.02%	86.52 ±1.25%	85.01 ±2.38%	73.97 ±5.08

Source: EU Commission – Joint Research Centre

4.3.3.3. ML algorithm selection

In our fourth experiment 4, we curated positive training set and identified several incorrectly labelled articles which we corrected. This led to an improvement of the classifier performance (all results available in Annex 5).

Figure 10. Experiment 4 ROC curves for each type of algorithm using TF-IDF-2 as numerical representation (Test Data)



Source: EU Commission – Joint Research Centre

Then, we assessed the impact of the ML algorithms to select the best preforming one(s), using the TF-IDF 2 numerical representation. As shown in Figure 11, the best performing models were Logistic regression, Random Forest, and Gradient boosting algorithms with a sensitivity on test data ranging from 78 to 80% while Adaptive boosting performed sensibly worse, achieving a sensitivity of 70%.

As all these three classifiers were performing equally, we built an ensemble classifier by averaging their prediction scores. To validate further the performance of this algorithmic choice, we also ran the classifier on the target data (see section 4.3.2.3).

On the target data, the performance remained comparable to the validation data, with a higher mean but wider standard deviation almost meeting our objectives (Table 9).

Table 9. Experiment 4 results (keeping the three best classifiers)

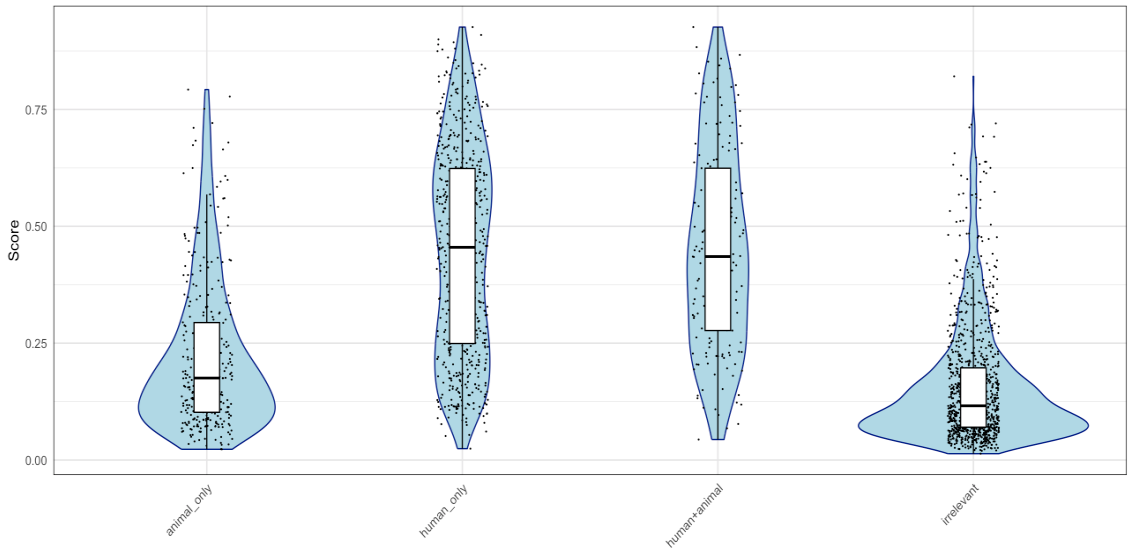
	Sensitivity	Specificity	Precision	Accuracy	F1 Score
Test Data (3 best)	79.08 ±1.30%	95.02 ±0.00%	88.15 ±0.17%	89.94 ±0.42%	83.37 ±0.80
Validation Data	44.74 ±2.74%	95.01 ±1.08%	85.88 ±2.86%	74.70 ±1.54%	58.83 ±2.42
Target Data	46.27 ±3.58%	95.01 ±1.30%	84.09 ±3.85%	77.34 ±1.85%	59.70 ±3.07

Source: EU Commission – Joint Research Centre

4.3.3.4. Tiered approach

By further analysing the distribution of scores according to our four subgroups (listed in section 4.3.1.1), we realised that for articles making exclusive use of animal models (labelled "animal only") scores were higher and more spread than for the articles not containing models (labelled "irrelevant"). This could mean that the classifier was discriminating better on "irrelevant" (Figure 11) and reducing the classification performance of the "animal only" category.

Figure 11. Distribution of score per subgroups



Source: EU Commission – Joint Research Centre

To assess whether these differences were statistically significant, we performed pairwise comparisons using the Wilcoxon rank-sum test. The results showed statistically significant differences between most subclasses ($p < 10^{-13}$), including comparisons involving *"animal only"* and *"irrelevant"*. In contrast, no statistically significant difference was observed between the *"human only"* and *"human+animal"* subclasses ($p = 0.73$), indicating that the classifier assigns similar score distributions to these two categories.

This prompted us to try another approach, inspired by method used by our biomedical expert while doing the labelling of Validation data: identifying irrelevant articles in a first step, and looking at the type of model used in the articles in a second. We decided to build two independent classifiers:

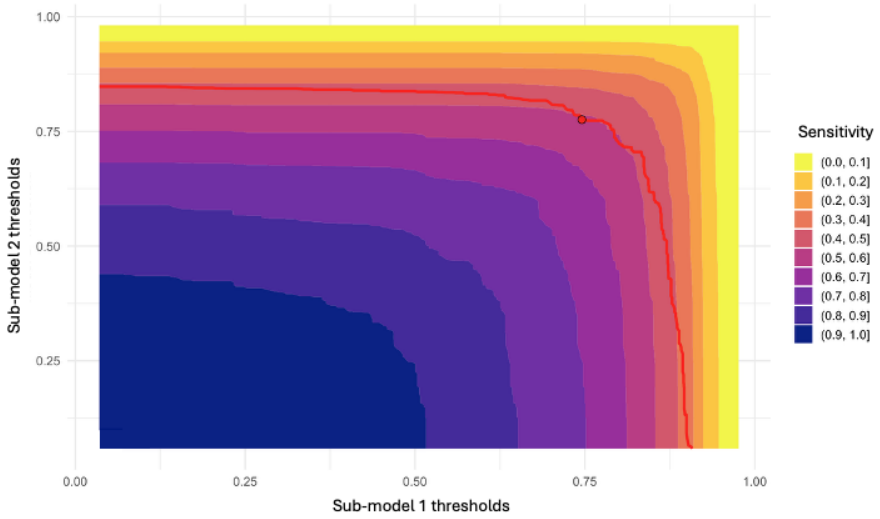
1. **Model detection (sub-classifier 1)**: identifying irrelevant articles in the whole corpus (and keep only articles making use of models)
2. **Animal model presence detection (sub-classifier 2)**: Identifying articles making exclusive use of animal-based models amongst relevant articles.

For this, we used the same models as in experiment 4 (TF-IDF-2 embedding and average of normalised scores for Logistic Regression, Random Forest and Gradient Boosting algorithms) with two new training sets (expended from the one used in experiment 4), as they needed to be expended to contain a sufficient number of articles to train two classifiers (instead of one).

For the sub-classifier 1, we used 7,500 positive articles coming from a representative set of Positive articles used for Iteration 4, completed with the Validation data and 7,400 Negative articles labelled "irrelevant".

For the Animal model detection classifier, we built the training set with 7,000 animal-only articles for the negatives and 8,000 human only / human-animal for the positives. Determining the optimised threshold was more complex for these two-tier models, as a single ROC curve cannot be used for the optimisation (the animal detector classifier should work on the outcome of the model detector classifier). For this purpose, we calculated the sensitivity for all combinations of thresholds for the two sub-models on the target data (Figure 12). By fixing the specificity at 95% (red line), we looked for the best sensitivity that we could obtain (red dot).

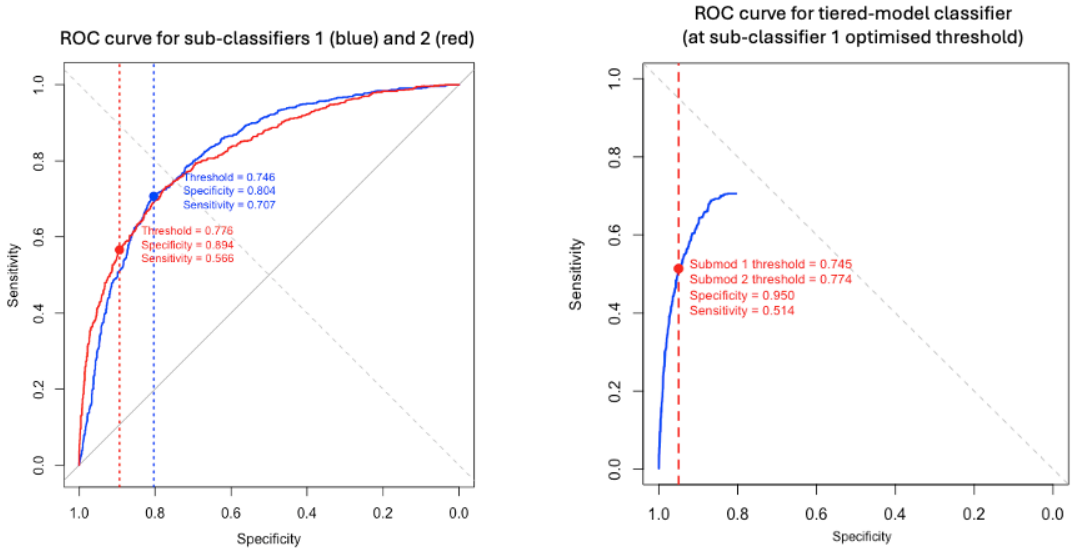
Figure 12. Tiered-model sensitivity for all combinations of thresholds for the two sub-classifiers



Source: EU Commission – Joint Research Centre

Looking at the ROC curves for these established thresholds (Figure 13), we can see the benefits of this two-tiered approach. Individually, both models perform far worse that the outcome of experiment 4. For a specificity of 95%, the model detector classifier would have a sensitivity around 20% (as it cannot predict whether a model is human biology-based or not) and the animal model detector would be around 40% (still good, but it would miss some irrelevant articles). At the optimised thresholds, the outcome would be low in precision as the specificity falls respectively at 80% and 90% (instead of the targeted 95%).

Figure 13. ROC curves of individual models and tiered-model classifier



Source: EU Commission – Joint Research Centre

Figure 13 (right part) shows the ROC curve for sub-classifier 2 thresholds for a sub-classifier 1 threshold fixed at 0.776 (optimised value determined in Figure 12)[19]. Analysing it, we can see an important improvement in the steepness of the curve, leading to a sensitivity of 51.37% for a specificity of 95%.

For the identified threshold, we obtained a 5% gain in sensitivity at a specificity of 95%, leading to a precision of 85.43% and a F1 score of 64.16 demonstrating a fair classifier meeting our objectives on target data. We decided to stop the experiments at this stage to assess the number of articles we could retrieve with this model with the intention to improve it further, in case the number of retrieved articles would be too limited.

Table 10. Experiment 5 results (final tiered-model classifier)

	Sensitivity	**Specificity**	**Precision**	**Accuracy**	**F1 Score**
Target Data	51.37 ±3.61%	95.01 ±1.30%	85.43 ±3.56%	77.39 ±1.80%	64.16 ±2.91

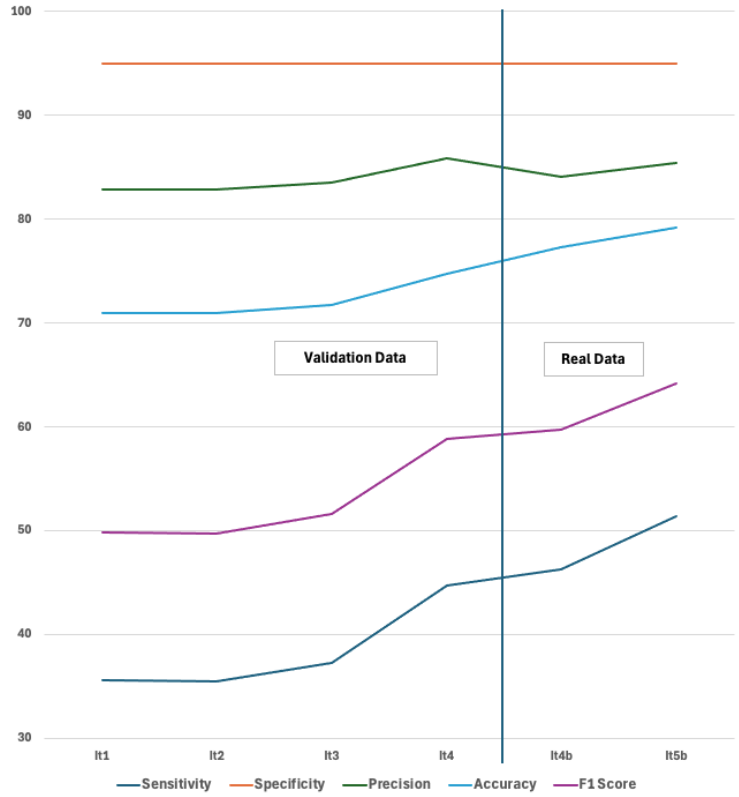
Source: EU Commission – Joint Research Centre

[19] As in the tiered-model classifiers, articles are classified as positives if their two scores are above both sub-classifier 1 and 2 thresholds, the sub-classifier ROC curve is truncated: when sub-classifier 2 threshold reaches 0, the sensitivity and specificity becomes those of sub-classifier 1 at its fixed threshold.

4.3.4. Results

Our final BimmoH classifier was constructed iteratively, by levelling the different parameters involved in the building of a machine learning classifier (the training sets, the embeddings for the text representation and the machine learning algorithms) and by creating specialised sub classifiers.

Figure 14. Evolution of the performance metrics across experimental iterations



Source: EU Commission – Joint Research Centre

As shown in Figure 14, starting from a baseline model with a sensitivity of 36% we managed to improve the performance and get a final sensitivity above 51% (at a fixed specificity of 95%), leading to a 15% increase.

Central to this model is the application of Term Frequency-Inverse Document Frequency (TF-IDF) embedding, which involves the transformation of text data into numerical vectors. This process leverages a comprehensive set of 40,000 features derived from 200,000 titles and abstracts sourced from the PubMed database. The TF-IDF approach is instrumental in highlighting significant terms within the corpus, thus enabling the model to capture the nature of models described in the abstracts.

The classifier's architecture incorporates two main sub-classifiers, each serving a specific purpose in the classification process. These classifiers operate by averaging the outputs of three robust algorithms: Random Forest, Gradient Boosting, and Logistic Regression (see section 4.3.3.3, Figure 10). The first classifier is tasked with the identification of irrelevant articles within the entire dataset collected through the PubMed query. This step is crucial as it filters out non-pertinent entries, allowing the model to focus on potentially valuable articles. The second classifier is dedicated to discerning articles that exclusively use animal models from those that include human

biology-based models. This differentiation is vital for refining the dataset to include only those articles that align with the project's objectives.

Additionally, the model incorporates a sophisticated threshold optimisation process. This involves adjusting the classification thresholds to maximise performance metrics such as precision and sensitivity for a fixed specificity, ensuring that the model not only accurately identifies relevant articles but also minimises the inclusion of false positives. This iterative optimisation process was critical in fine-tuning the model's decision boundaries, thereby enhancing its reliability and effectiveness.

4.4. Indexing

To improve information retrieval on the articles that we selected with the ML classifier, we indexed the title and abstract content with vocabularies especially created for the needs of stakeholders identified during the requirement phase (section 3.1).

4.4.1. Vocabularies

The importance of specific vocabularies designed to retrieve relevant articles became essential when we decided not to limit the retrieval of human biology-based models to specific areas of the biomedical research but to cover the entirety of the biomedical research[20].

4.4.1.1. Main categories

Consulting stakeholders, we realised that they needed a specific but flexible way the retrieve information. Instead of complex hierarchical ontologies, we decided to establish controlled vocabularies of keywords spanning across three main complementary categories to enable efficient search, while allowing a sufficient degree of flexibility to accommodate a vast coverage of research topics, as described below.

Anatomy, Histology, and Cells

This lists the human body from whole organ systems down to cellular and developmental structures. It includes regional anatomy terms (e.g., abdomen, thorax, limb regions), organ names and their substructures (e.g., heart, aorta, ventricles; brain regions like the cerebrum, hippocampus, cerebellum; gastrointestinal organs such as the stomach, duodenum, colon), and tissue types (e.g., epidermis, cartilage, adipose, connective tissue).

It also encompasses cell types such as epithelial cells, neurons, glia, immune cells (lymphocytes, neutrophils, macrophages), germ cells, endocrine cells, and specialised secretory or structural cells such as osteoblasts, hepatocytes, pneumocytes, and melanocytes.

Many terms describe microscopic anatomy, including organelles (mitochondria, Golgi, lysosomes), extracellular components (matrix, collagen, basement membrane), and specialised structures (cilia, axoneme, sarcomere). Developmental biology is represented through embryonic and foetal structures (blastocyst, mesoderm, amnion, placenta) and cell potency states (totipotent, pluripotent, multipotent).

[20] Complete vocabularies used to index the BimmoH dataset are available at: https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/EURL-ECVAM/datasets/BimmoH/LATEST/Vocabulary_5g_2025_08_19.xlsx

This vocabulary also includes vascular and lymphatic terminology, neural pathways, glandular structures, barriers and membranes (blood–brain barrier, peritoneum, meninges), and physiological systems ranging from endocrine and respiratory to reproductive and immune systems.

Clinical Conditions, Disease, and Pathophysiology

This list spans the full spectrum of clinical conditions, infectious diseases, genetic syndromes, and pathophysiological processes encountered in medicine. It includes infectious diseases caused by bacteria (e.g., Staphylococcus, Salmonella, Mycobacterium), viruses (influenza, HIV, SARS-CoV-2, rabies), fungi (Candida, Aspergillus), parasites (Plasmodium, Giardia, helminths), and prions (Creutzfeldt–Jakob disease). Many entries refer to inflammatory and autoimmune disorders such as asthma, dermatitis, lupus, rheumatoid arthritis, multiple sclerosis, and inflammatory bowel diseases.

It also covers cancers and neoplastic processes, including carcinomas, sarcomas, leukemias, lymphomas, and various tumors—along with terms related to tumor biology, such as carcinogenesis, and metastasis.

The list encompasses genetic and congenital conditions (Down syndrome, Marfan syndrome, Wilson disease, Fanconi anemia), metabolic and endocrine disorders (diabetes, hyperlipidemia, thyroid disease), and neurodegenerative diseases (Alzheimer disease, Parkinson disease, Huntington disease) as well as neurological events like stroke, seizures, and demyelination.

There are numerous terms describing cardiovascular pathologies (arrhythmias, aneurysms, atherosclerosis, cardiomyopathy), respiratory illnesses (COPD, pneumonia, bronchiectasis), and gastrointestinal and hepatic diseases (gastritis, cirrhosis, NAFLD/NASH, cholangitis).

Additionally, the list contains descriptors of pathophysiological mechanisms (e.g., inflammation, necrosis, apoptosis, fibrosis, ischemia, immunodeficiency, hypersensitivity, toxicity, mutagenicity) and clinical manifestations such as pain, fever, edema, anemia, or fatigue.

It gives a comprehensive overview of how the body can be affected by infections, genetic abnormalities, immune dysfunctions, metabolic derangements, toxic exposures, and malignant transformations.

Modelling approaches

This list brings together cellular, tissue-engineered, microphysiological, and computational modelling approaches used across modern biomedical research. They range from basic *in vitro* culture systems to highly advanced *in silico* simulations and digital reconstructions of human physiology. Together, they represent the ecosystem of tools used to understand biology, predict drug responses, model disease, and build human-relevant test systems.

In vitro

It includes cell and tissue-based experimental systems such as:

- 2D-systems like simple monolayer cultures (e.g., 2D culture, 2D cell culture, cell line, adherent cell) useful for controlled mechanistic work, imaging, and high-throughput screening;
- 3D systems like 3D culture, spheroids, organoids, assembloids, gastruloids or blastoids that provide architecture, cell–cell interactions, and microenvironments closer to *in vivo* tissues;

- specialised culture formats like air-liquid interface, hanging drop culture, gel-based culture, electrospinning culture, transwell culture, microgravity culture, airborne cell culture used for differentiation, epithelial modelling, exposure studies, and unique biomechanical conditions;
- Human-relevant cellular sources like primary cells, stem cells, progenitor cells, iPSC/hiPSC/hPSC, immortalised cells that allow modelling of development, disease, precision medicine, and patient-specific biology;
- tissue engineering systems like engineered tissue, bioprinted tissue, scaffold-based model, scaffold-free tissue, whole organ, employed for regenerative medicine, implantation research, and drug testing;
- bioreactor-based culture or dynamic culture used to apply flow, shear stress, and long-term physiological conditions.
- microphysiological and advanced *in vitro* systems such as organ-on-chip also called human-on-a-chip, organotypic culture or microphysiological system that are microengineered devices that recreate tissue–tissue interfaces, perfusion, and biomechanics;
- microfluidic technologies like microfluidic chip or 3D microfluidic culture that enable controlled gradients, flow conditions, and precise microenvironment engineering;
- high-complexity constructs like aggregoids, spheroids, assembloids or embryoid bodies that recapitulate multi-cellular development and organ formation processes.
- *ex vivo*, explant culture, tissue, whole organ, bridging *in vitro* work with physiological realism by maintaining intact tissue structure.

In chemico

This covers chemical, toxicological and experimental assays used for mechanistic toxicology, sensitisation testing, and regulatory non-animal assessment.

In Silico

This relates to computational, mathematical, and simulation models covering the full spectrum of *in silico* biomedical modelling such as:

- mechanistic and physics-based models like finite element model, fluid-structure interaction, computational fluid dynamics, electromechanical model, rigid-body model, smoothed-particle hydrodynamics that represent biomechanics, blood flow, cardiac mechanics, tissue deformation, or physical forces;
- biological mechanistic models like Physiologically-Based Pharmacokinetics (PBPK), pharmacokinetic/pharmacodynamic (PK/PD), quantitative systems pharmacology (QSP) to predict drug absorption, distribution, metabolism, and complex biological responses;
- stochastic and statistical models like gaussian process model, stochastic model, regression model, reduced-order model, statistical shape model, used for uncertainty quantification and shape/parameter variability;
- AI / machine learning making use of neural network models, convolutional neural networks, classification models, deep learning models, machine learning for image analysis, biomarker discovery, pattern recognition, and predictive modelling;

- agent-based and boolean models representing interacting cells, molecules, or organisms in rule-based frameworks;
- molecular modelling like molecular dynamics, protein–ligand interaction prediction, pharmacophore model, Quantitative Structure-Activity Relationship (QSAR) model, Swiss-model used in structural biology, drug design, and chemoinformatics;
- emerging systems-scale digital models like digital patient, virtual patient, digital twin, virtual twin aiming at whole-body computational surrogates for precision medicine and treatment optimisation;
- other approaches: data-driven model, mechanistic model, phenomenological model, k-omega model, Markov state model, lumped-parameter model that are common in cardiovascular research, system simulation, and pharmacology;
- more generic labels (often used in articles' title or abstracts) like mathematical model, numerical model, surrogate model, multiscale model, multibody model, constraint-based model.

4.4.1.2. Other categories

For refining the search further, we created the following vocabularies for indexing the articles:

Omics Techniques

A wide array of omics technologies that collectively measure the molecular landscape of cells and tissues at the DNA, RNA, protein, and metabolite levels:

- Genomics and epigenomics, with sequencing-based techniques such as ATAC-seq, DNase-seq, MNase-seq, ChIP-seq, MeDIP-seq, RRBS, and other methylation-sensitive assays that profile chromatin accessibility, histone modifications, DNA–protein interactions, and epigenetic marks, alongside broader platforms like whole-genome sequencing, exome sequencing, and DNA microarrays.
- Transcriptomic approaches (e.g-, RNA-seq, Ribo-seq, ribosome profiling, gene expression microarrays) that captures gene expression, RNA abundance, and translation dynamics.
- Proteomic methods, relying on mass spectrometry technologies such as LC-MS, GC-MS, FT-ICR-MS, MALDI-ToF, tandem MS, and complementary platforms like Reverse Phase Protein Arrays to quantify proteins, interactions, and post-translational modifications.
- Metabolomic analyses, using MS-based and NMR-based techniques to profile small molecules and metabolic signatures that reflect cellular state.

Together, these omics tools create a multilayered view of biological systems, enabling researchers to connect genomic variation, epigenetic regulation, transcriptional output, protein networks, and metabolic activity into integrated, system-level insights.

Human Cell Lines

This list contains the 2,500 most frequently used cell lines present in the BimmoH dataset. They are *in vitro* biological systems derived from human tissues that retain many genetic and phenotypic features of their origin and are used for controlled experiments in biology and medicine.

A cell line is a permanently established cell culture that will proliferate indefinitely given appropriate fresh medium and space. Cell lines differ from cell strains in that they become immortalised.

Animal species and cell lines

List of animal species and animal-derived cell line models to study physiology, genetics, toxicology, pharmacology, development, and disease mechanisms. The species portion of the list spans vertebrate and invertebrate models.

Complementing these animal models is an extensive set of animal cell lines and many cancer, fibroblast, neuronal, endocrine, and immune-derived lines from numerous species. Together, the species and cell lines in this list form a comprehensive cross-species overview of animal models used in biomedical research.

We created this index to give the user the choice to include or exclude from the search articles that use animal models (referring to animal species in title or abstract) in addition to human biology-based models.

4.4.2. Tagging approach

The tagging step is assigning keywords detected in articles' title and abstract to the different categories mentioned in the above paragraph. The tagging implementation covered two aspects: the technical implementation, allowing detection of keywords and associated variants, on one side, and the fine tuning of the vocabularies to optimise them for the search, on the other, to enhance their retrieval performance during the search.

4.4.2.1. Technical implementation

To support information retrieval, each term in the above-mentioned vocabularies is translated into a regular expression designed to robustly match its appearance within titles and abstracts of the PubMed biomedical research articles.

These expressions are constructed to capture plural forms, upper and lower-case variations, UK/US spelling differences, and the variety of ways authors may write biological names, including hyphenation, spacing differences, abbreviations, and taxonomic synonyms.

This flexible regex-based indexing strategy ensures comprehensive retrieval of articles involving the anatomical terms, diseases, and models listed above, enabling accurate classification, trend analysis, knowledge extraction, and integration of heterogeneous datasets across the biomedical research landscape.

For the tagging evaluation, random samples of articles were selected and assessed by experts to identify discrepancies. Experts checked that indexed terms were present in the article's title or abstract, and that all relevant terms were correctly detected.

4.4.2.2. Vocabularies optimisation

One of the challenges of the vocabularies was to avoid redundancies between the three main categories, while maintaining a flexibility and allowing sufficient complementarity of the terms to retrieve relevant articles and their models.

Maximising the performance of vocabularies also meant reducing the number of keywords they contained to the most relevant ones. The number of keywords occurrences inside a representative

dataset of 200,000 articles was counted to keep only the most frequent terms (especially removing those non appearing terms).

Cleaning the vocabularies was another important task, for example to remove duplicates that could be due to spelling variations, in which case only one term was kept and variations added to the regular expression matching the term; similarly, acronyms were removed if they were already available in plain text.

We also focused on the detection of false positive keywords retrieval, when a term was tagged correctly from a technical point of view, but had another meaning within the abstract (for example, “on the other hand” does not refer to the anatomical part “hand”).

5. End users validation

Validation is a crucial step when dealing with data products based on AI techniques. As the BimmoH data is automatically curated, we need to ensure that the dataset meets the expectations of the end users from both selection and indexing perspectives.

Thanks to the various analysis performed during the experimental part of the project described in the previous section, we managed to acquire a good understanding of the performance of the ML classifier and indexing script. Nevertheless, it was important to verify that the outcome of this work, at the heart of the automated construction of the BimmoH dataset, is relevant to end users and will allow significant progress for the easy retrieval of human biology-based models used in biomedical research.

The external validation phases of the BimmoH platform were designed to complement internal development efforts with structured feedback from representative user groups. These evaluations ensured that both technical performance and scientific coverage were assessed under realistic conditions by audiences ranging from PhD students to domain experts and regulatory professionals.

Testing proceeded in successive stages:

- an EC-JRC Full Validation Test, where JRC scientific experts systematically evaluated tagging accuracy, retrieval performance, and interface refinements;
- an Alpha Test, conducted in a live educational setting with early-career researchers to identify usability issues and content gaps;
- a Beta Test, engaging a smaller but more specialised set of stakeholders to assess readiness for broader release.

Together, these phases provided critical insights into functionality, reliability, and relevance, guiding iterative improvements and preparing the platform for production deployment.

5.1. Internal validation (JRC experts)

We set a team of experts in specific technologies (organ-a-chip, *in vitro* techniques, stem cells, *in silico* approaches) to conduct an internal qualitative evaluation focusing on the relevance of content within a newly enhanced database platform.

During the month of June 2025, we assessed a dataset comprising over half a million entries: 522.820 entries filtered and scored by the classifier developed during experiment 4, setting the prediction score to a conservative threshold of 0.6 ensuring very high specificity (higher than the one used in table 11 that was 0.47).

Table 11. Statistical analysis of classification results used for the internal validation and alpha testing

	Sensitivity	**Specificity**	**Precision**	**Accuracy**	**F1 Score**
Target Data	28.37 ±3.17%	97.84 ±0.93%	88.21 ±4.70%	72.64 ±1.96%	42.93 ±3.69

Source: EU Commission – Joint Research Centre

The average satisfaction score and perceived relevance of results was around 70%, suggesting that while testers often found the platform useful, there is significant room for improvement. Tags were

marked correct in only 61% of cases, a critical gap for a tag-driven search tool. Testers successfully retrieved at least one key reference 80% of the time, but missing known references were reported for 68% of queries. Irrelevant articles or articles relying only on animal models were relatively rare (7%). The platform often found what experts looked for, but only with the right query gymnastics, and sensitivity/recall was a bigger challenge than precision (Table 11).

Experts quickly learned that relying solely on tags was risky: many relevant references were invisible without adding free-text terms. For example, searching for computational models like finite element or agent-based often failed unless these exact phrases were typed in, and even then, the platform didn't merge free-text into the main Boolean query, making results unpredictable.

The testing of this BimmoH dataset version demonstrated that the core concept was viable and, when the tags were there and correct, researchers can focus on highly relevant literature in ways that standard search engines could not match.

However, the combination of indexing issues, low sensitivity, and user interface limitations meant the platform wasn't yet ready for high-stakes systematic searching.

5.1.1. Alpha testing (JRC summer school 2025)

The alpha testing phase for BimmoH, was conducted during the JRC Summer School (May 19-23, 2025) with 120 students and approximately 10 external experts. The primary objectives were to evaluate the IT functionality, system performance, and the relevancy of the database content. Over the course of three days (May 20-22, 2025) we ran four sessions, each with 30 PhD students, who interacted with the database interface to perform a series of free-form searches and tasks. The dataset we used was the one also tested by the JRC experts in the Internal validation.

Students were instructed to assess the relevance and completeness of the data they retrieved as well as technical issues they may encounter using the web interface. The initial feedback from the participants was overwhelmingly positive, indicating a strong alignment between the database's capabilities and user expectations. Detailed analytics and findings were compiled to refine the database further and to prepare for the next phase of testing.

Most of the JRC summer school testers that completed the satisfaction survey (65) identified themselves (81,5%) as PhD students (34) or Master students (19) that employed traditional two-dimensional (2D) cell culture methods, such as monolayer cell lines, adherent, or suspension cultures, when conducting their own research (35). A sizable minority (24) also worked with three-dimensional (3D) cell culture systems such as spheroids, organoids, or bioprinted tissues, while smaller subsets used advanced organ-on-a-chip systems (13), *ex vivo*/primary tissue models (3), *in silico* simulations (11), or *in chemico* assays (2). Almost every participant (62) reported familiarity with PubMed, and many also used Google Scholar, Scopus, or Web of Science in their research workflows. Only a handful had experience with EuropePMC or OpenAlex prior to this test.

Because these testers were already familiar with biomedical and life-science literature repositories, they brought clear expectations regarding search relevance, filtering options, and result ranking.

When asked whether the search returned key references they expected, 46 participants (71%) answered "yes," while 19 (29%) felt that certain key items were not retrieved. Similarly, 45 respondents (69%) observed that at least one essential reference was missing from the database, indicating gaps in our indexed corpus. At the same time, 23 participants (35%) reported encountering irrelevant results, studies that did not match their query intentions, while the remaining 42 (65%) did not see such unwanted items.

When asked to identify the single most important improvement, the most common responses revolved around relevance ranking and filter granularity. Many participants wanted search results sorted by a composite measure of relevance, combining citation count, publication date, and textual match, rather than simply listing items alphabetically or chronologically. Others requested more detailed subcategories under "Pathophysiology" or "Human-based models," enabling users to compare specific organ systems (e.g., liver versus central nervous system) or *in vitro* technologies (e.g., organoids, organ-on-a-chip, microfluidics) in one unified view.

Although most testers felt the platform returned important references, nearly 70% reported missing essential content in this specific Alpha version. Many of these missing items were high-impact articles in 2D and 3D cell-culture research, suggesting that our indexing did not fully capture key articles from journals or conference proceedings relevant to these fields. Because most testers rely heavily on PubMed and Google Scholar, participants noted that BimmoH's content coverage must expand to include top-tier journals and specialised repositories that are commonly cited in their own work. At the same time, 35% encountering irrelevant results implied the need of a ranking mechanism, so that the most pertinent references could appear at the top of the results list.

5.1.2. Beta testing (Stakeholder community)

After addressing the shortcomings identified by the alpha testing (especially by improving the sensitivity as shown in Table 12), the beta-test assessed whether the system returned relevant, complete and well-tagged literature references; how easily users could refine and export results; and whether the user interface (UI), filters, and data structures supported expert workflows.

17 experts participated to the beta testing phase, representing a complete panel of candidate users of the BimmoH dataset: Academic researchers (7), NGO/non-profits (6), Governmental / national authorities (5), Three Rs platform developers (4), Consultants (2), Ethics board members (2), Science communicators (2), Animal-welfare bodies (2), Funding agencies (1), Pharmaceutical companies (1), Regulatory scientists (1), Regulatory authorities (1) with a research background in 2D cell culture (7), 3D/organoids (5), *ex vivo*/primary tissue (4), *in silico* modelling (2), advanced MPS (1), *in chemico* (1)[21].

The evaluation combined (a) a structured survey with Likert ratings, multiple-choice items, and facet-usage telemetry, and (b) open-ended comments on what worked well, bugs, and improvement ideas.

The test took place in July 2025 on a dataset comprising over 750,000 articles (filtered and scored by the machine learning model trained in Iteration 5, with a prediction score thresholds of 0.8 or higher for both models, higher than the ones used in Experiment 5, see Table 12) ensuring that only the most relevant and high-confidence results are included in the review.

Table 12. Statistical analysis of classification results used for the beta testing

	Sensitivity	Specificity	Precision	Accuracy	F1 Score
Target data	44.71 ±3.56%	96.65 ±1.11%	88.37 ±3.62%	77.81 ±1.84%	59.38 ±3.19

Source: EU Commission – Joint Research Centre

[21] Experts could belong to more than one group

Although the number of testers answering the satisfaction survey was lower than for the alpha-test, (17 vs. 65) This time, the relevance of records reported by testers reached more than 76%, but reaching 100% when considering if key articles were present in the results of the query. Ease to use of the web interface scored 84%, and 85% compared to similar types of databases for searching human biology-based models.

Identified positives aspects were:

- identification of relevant hits quicker than with PubMed or Google Scholar;
- faceted search works well, especially by models, clinical conditions, and Anatomy/Cells;
- charts give immediate landscape view (yearly trends) for scoping a field;
- customisable columns & saved results support curation;
- fast search and intuitive navigation after brief onboarding.

Requests for improvement related to the export function (allowing all results to be exported easily), improving the advanced search usability, and allowing for saving and sharing queries. These requests where considered in the final version of the user interface.

6. Results

6.1. BimmoH dataset AI performance

The first published dataset (published on 01/12/2025) contains 791,797 references of articles potentially mentioning the use of a human biology-based model (according to the definition provided in box 1) and is based on the classifier described in section 4.3.3, at the threshold of 0.8 used for the beta-testing (Table 13). As the dataset was gathered using an AI and not manually curated, this list of articles is not exhaustive, and misclassification may happen.

Table 13. BimmoH dataset performance statistics (95% confidence interval)

	Sensitivity	Specificity	Precision	Accuracy	F1 Score
BimmoH data	44.71 ±3.56%	96.65 ±1.11%	88.37 ±3.62%	77.81 ±1.84%	59.38 ±3.19

Source: EU Commission – Joint Research Centre

BimmoH database is therefore highly precise and a search may return, on average, one article out of ten that is out of the scope. This is because up to 97.4% of articles that are either irrelevant for BimmoH or make exclusive use of animal models are excluded properly from our PubMed candidate query. On the downside, slightly less than half of relevant articles is retrieved, a situation that could be improved in the future (see section 6.4). However, knowing that the use of AI allowed us to increase the number of included articles in the biomedical reviews from 3,000 to almost 800,000, it can be considered that the project's objectives were successfully achieved.

6.2. Data access

BimmoH data can be accessed in two ways: accessing raw data by directly downloading the dataset from the JRC data catalogue[22] or using the web application[23] providing a user interface to search and navigate before exporting a reduced subset of relevant data.

6.2.1. Full dataset

The dataset is provided free for use and can be downloaded in a JSON file. The JSON file is composed of a list of PubMed PMIDs that each contains the JSON structure described in Table 14.

Table 14. BimmoH dataset JSON file format

Field	Type	Origin	Description
pmid	Integer	PubMed	Unique number assigned to each PubMed citation
doi	String (DoI format)	PubMed	Digital object identifier
author	String	PubMed	All authors with full name

[22] https://data.jrc.ec.europa.eu/dataset/ba511666-1c31-4ac0-a4a4-97567e480aba

[23] https://bimmoh.eu

first_author	String	PubMed	First author full name
title	String	PubMed	The title of the article
content	Enumeration (Strings)	OpenAlex	Index of the abstract keywords in alphabetical order
year	Date	PubMed	The date the article was published
journal	String	PubMed	Full journal title from PubMed cataloguing data
volume	Integer	PubMed	Volume number of the journal in which the article was published
issue	Integer	PubMed	The number of the issue, part, or supplement of the journal in which the article was published
article_type	Enumeration (Strings)	PubMed	The article type, from the PT field in the MEDLINE records
pages	String	PubMed	The full pagination of the article
issn	String	PubMed	International Standard Serial Number of the journal
mesh	String	PubMed	Concatenation of Medical Subject Headings (MeSH) terms in a single string
mesh_terms	Enumeration (Strings)	PubMed	MeSH terms used by PubMed
human_anatomy	Enumeration (Strings)	BimmoH	Specifies the physiological organ, tissue or cell targeted in the research study.
model	Enumeration (Strings)	BimmoH	Type of research models used in the article/reported in the review paper.
disease	Enumeration (Strings)	BimmoH	Names the disease or pathogenic process studied, aligning with clinical relevance.
omics	Enumeration (Strings)	BimmoH	Omics technologies (e.g., genomics, proteomics, metabolomics) used in the research study.
cell_line_type	Enumeration (Strings)	BimmoH	Further details on the specific cell-based models/methods used, referencing the glossary for definitions (e.g., primary cell cultures).
animal	Enumeration (Strings)	BimmoH	List of animal species mentioned in the abstract.
has_animals	Boolean	BimmoH	TRUE if any animal species of the vocabulary is mentioned in the title or in the abstract.
open_access	Boolean	OpenAlex	TRUE if the article is open access.
model_cats	Enumeration (Strings)	BimmoH	Classifies the research based on the type of model used (*in vitro, in chemico, in silico*).
omics_cats	Enumeration (Strings)	BimmoH	Indicates if any 'omics' approach has been undertaken in the study (e.g., genomics, epigenomics, transcriptomics, metabolomics).

Source: EU Commission – Joint Research Centre

6.2.2. User interface

BimmoH also allows end users to search for data and navigate through the results to easily identify models that are relevant to their research through an intuitive user interface. We designed the web application viewer to answer the needs of end-users by offering an intuitive interface that can be easily accessed by users with limited expertise in information extraction. This design ensures that users, regardless of their technical background, can easily navigate data to locate and access human biology-based models pertinent to their research interests.

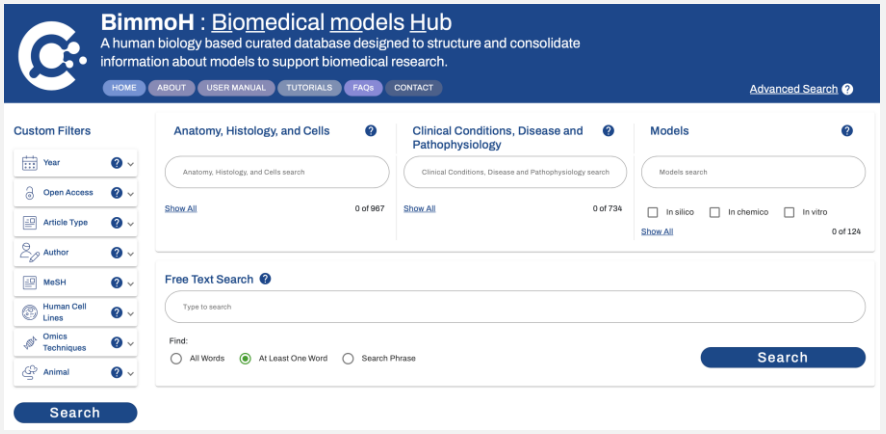
A distinct feature of the application is its support for multidimensional search criteria, which is particularly valuable given BimmoH's comprehensive scope across various biomedical research domains. Users are provided with three primary categories to refine their search and tailor it to their specific needs, using Boolean operators to combine them (AND/OR/NOT). These categories include:

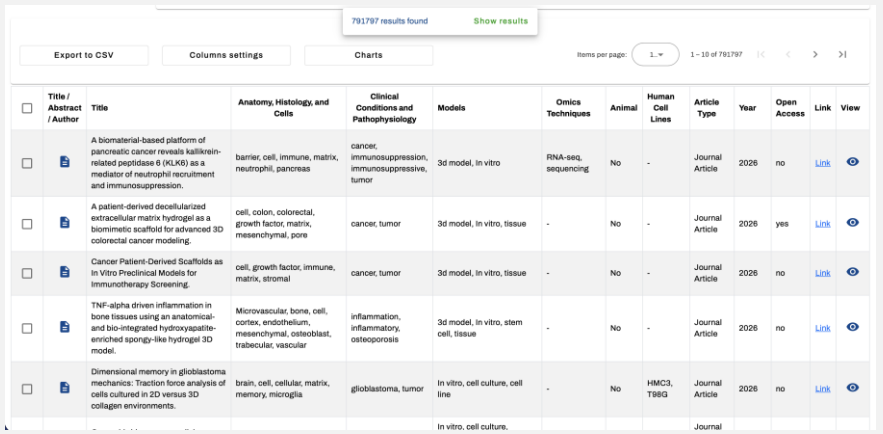
- Anatomy, Histology, and Cells;
- Clinical Conditions, Disease, and Pathophysiology;
- Models, which encompass *in vitro*, *in chemico*, and *in silico* methodologies.

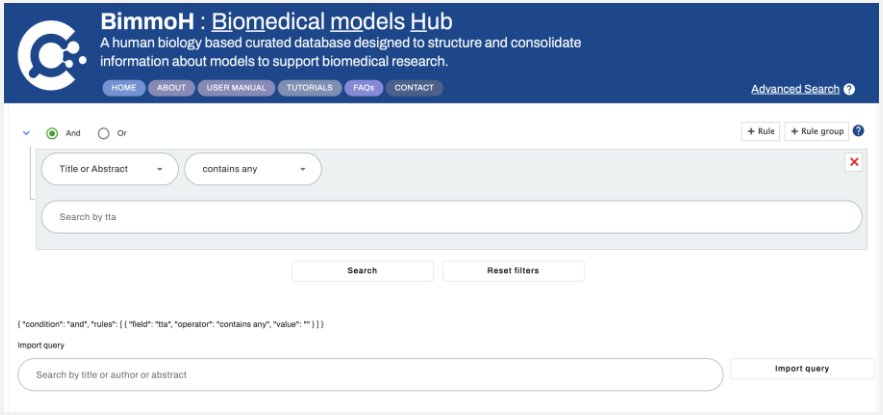
To accommodate more complex needs, the user interface also incorporates an advanced query mode. This feature empowers users with the flexibility to create detailed and customised queries, including the use of free text inputs. Such a capability is crucial for researchers who require precision and specificity in their search parameters, ensuring that they can effectively pinpoint the exact data relevant to their work. When users retrieve substantial datasets, the user interface provides interactive charts (bar charts, pie charts and heatmaps), a powerful tool for data visualisation. These charts offer a macro-level view of the data, facilitating a deeper understanding of trends and patterns that may not be immediately evident through raw data alone. This visual representation aids in the analysis and interpretation of complex datasets, enhancing the user's ability to draw meaningful conclusions from their searches.

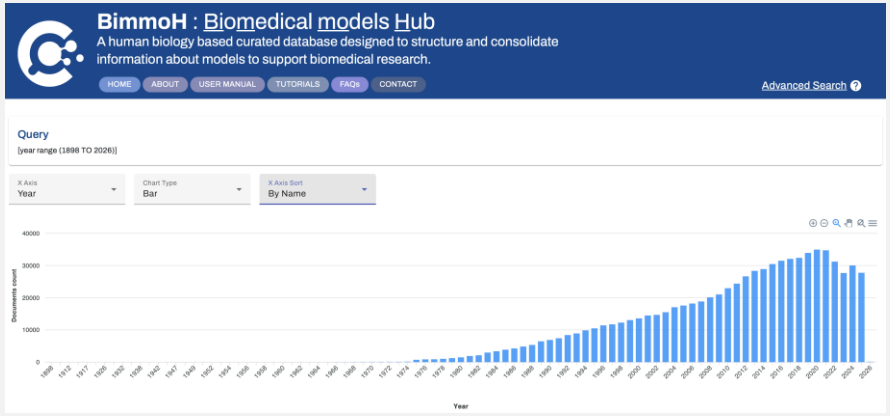
Moreover, the user interface (Box 3) supports the functionality to save and share queries, allowing collaboration and efficiency. Users can easily revisit their searches, share their findings with colleagues, or use them as a foundation for further exploration. Additionally, the platform's export capabilities (tabular CSV format) ensure that users can integrate the data into their existing workflows, enabling them to utilise the information in presentations, publications, or further analysis.

Box 3. The BimmoH user interface









https://bimmoh.eu

7. Discussion and next steps

7.1. ML classifier behaviour interpretation

The interpretation of AI algorithmic behaviour within the BimmoH project reveals the effectiveness of the various ML strategies employed.

A key observation is the performance of TF-IDF embeddings, which are particularly well-suited for handling the short text formats typically found in scientific articles' abstracts and titles. This method effectively captures the importance of terms within documents relative to a larger corpus, thus enhancing the ML model's ability to classify and index the texts accurately (for example, when analysing the terms with heavier weight in the algorithmic decision, we can see that human cell lines are ranked very high in the ML classifier decision making).

Among the tested ML algorithms, the AdaBoost algorithm, while useful, is considered less advanced in this specific application. Consequently, its utility is more limited within the context of this project. By excluding the AdaBoost algorithm and instead averaging the prediction scores of three other combinations of embeddings and ML algorithms (Table 4), the final model achieves a more robust output when determining classification thresholds. This methodological refinement ensures greater stability and reliability in the model's performance across diverse datasets.

One of the pivotal strategies identified for improving classification performance is the alignment and tuning of training datasets to target data coming from the PubMed candidate query, (reducing the number of articles that needs to be classified, especially reducing the number of negatives). By carefully adjusting these sets, the model's efficiency in classification is significantly enhanced. However, it is important to note that such tuning may lead to a decrease in performance if there are substantial changes in the candidate query, necessitating adjustments and refinements of the ML classifier such as changing the threshold or retraining the classifiers with more representative training sets.

An additional strategy that proved beneficial is the splitting of the negative training dataset into two distinct classes and processing them in a stepwise manner with two sub-classifiers. This approach enhances the classification performance by providing a clearer distinction between relevant and irrelevant articles, thereby refining its ability to discern and categorise the data accurately. The implementation of a two-step classification process has significantly improved the recall rates, allowing the model to capture a broader range of relevant articles.

Our analysis of the ML classifiers functioning in the BimmoH project suggests a similar approach to that used by human experts. It often relies more on keywords analysis rather than on semantic interpretation of the articles title and abstracts, which can be easily understood by the limited length of the analysed text. It also focuses on excluding negative articles in two steps, firstly checking their relevance and secondly looking if they are not making exclusive use of animal models.

7.2. Added Value of the BimmoH dataset

The BimmoH dataset provides the world largest database of scientific articles' references making use of human biology-based models in biomedical research. With this first release of almost 800,000 references, published between 1900 and 2025 and covering all areas of the biomedical research, BimmoH supersedes greatly our pioneering work achieved by the biomedical reviews that

we started to publish in 2021. This dataset is a very solid base for anyone that would like to benefit from a very large set of articles making use of human biology-based models.

Characterised by a very high precision and specificity, the BimmoH dataset ensures that users can access relevant data for their research needs, offering comprehensive coverage across all disease types, making it an essential tool for biomedical researchers.

The ML models employed in BimmoH are fully interpretable, offering clarity and transparency in their operation. This characteristic not only aids in understanding the model's decision-making processes but also presents strong potential for future improvements. Feedback from end users during Alpha, EC-JRC, and Beta testing phases has been, in general, positive, underscoring the project's success in meeting user needs and expectations.

The user interface and tailored vocabularies structured to support the search of human biology-based models guarantees efficient information retrieval through harmonised indexes, a major limitation in existing large databases of medical citations.

Finally, the BimmoH database is designed for regular updates and retraining, which ensures not only that it will remain current and relevant, but that it will also continue to evolve in line with the latest scientific advancements. This adaptability is crucial in maintaining the database's relevance and utility over time.

7.3. Limitations

Despite its strengths, the BimmoH project does encounter certain limitations. The database is non-exhaustive yet, capturing approximately half of the estimated relevant articles. While this represents a significant achievement in expanding the database's scope, there remains room for enhancement to achieve more comprehensive coverage.

Additionally, the reliance on information extracted solely from titles and abstracts poses a challenge, as the most detailed and pertinent information is often located in the Materials and Methods section of scientific articles. This legal constraint limits the depth of analysis the database can provide.

One potential other limitation is a bias towards articles making use of *in vitro* models, which may affect the comprehensiveness of the database in selecting articles referring to other types of models equally.

7.4. Potential improvements

To overcome the limitations listed above, the BimmoH dataset will require a careful analysis of its content to acquire a deeper understanding of its composition. If we want to improve the Sensitivity and correct eventual underperformance in some specific areas (for example diseases or models), we need first to characterise these limitations and put in place strategies to understand where they come from.

Some of the current ideas to improve the sensitivity are to improve further the performance of the ML classifier, using larger portion of articles for the classification (particularly the material and methods sections) and/or using other AI techniques such as large language models or cluster-based approach (for retrieving positive articles similar to those already selected, but using other criteria such as the authors, journals, etc.)

Access to the material and methods section of the articles would very likely bring the ML classification to very high levels of performance as it contains detailed information of the models. As these sections are longer than the title and abstract and contain specific language related to models, using sentence transformers could be a good candidate for creating the embeddings. Provided that we can get access to enough articles using those published in open access under the CC-BY NC 4.0 licence or by acquiring licences allowing full text processing, this could be a way to improve the weaker areas of the current BimmoH dataset.

8. Conclusions

Going from an initial set of 3,000 identified biomedical models to almost 800,000 references of articles making use of human biology-based models, we successfully reached the objectives of the project to leveraging the use of AI to automate the systematic review process for identifying relevant articles based on their title and abstract.

By choosing well-established machine learning techniques, we created a robust classifier, trained to identify relevant articles from a pre-selection of 4.3 million candidates extracted from PubMed using a tailor-made query. This classifier favours the exclusion of articles not containing any models or making exclusive use of animals, leading to a very high precision in the final BimmoH dataset, close to 90%.

Making it publicly accessible to the worldwide research community in its entirety in the JRC data catalogue and providing a user interface for navigating easily through its content, the BimmoH dataset is expected to be adopted by a large community of users, such as scientists working in biomedical research, research projects, evaluation committees or educational institutions.

With its innovative approach, BimmoH is an important milestone in the promotion and usage of human biology-based models, pushing for more human-centric science and aligning with the Strategy for European Life Sciences[24], which emphasises innovation in biotechnology while reducing animal testing.

[24] https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/jobs-and-economy/strategy-european-life-sciences_en

References

Ali Khan A., Valera Vazquez G., Gustems M., Matteoni R., Song F., Gormanns P., Fessele S., Raess M., Hrabě de Angelis M., INFRAFRONTIER Consortium, *INFRAFRONTIER: mouse model resources for modelling human diseases*, Mamm Genome, 2023 Sep;34(3):408-417, http://doi.org/10.1007/s00335-023-10010-7

Bewick, V., Cheek, L., Ball, J., *Statistics review 14: Logistic regression*, Critical Care, 2005, 9(1), 112–118. https://doi.org/10.1186/cc3045

Breiman, L, *Random Forests*, Machine Learning, 2001, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Canals, J. M., Romania, P., Belio-Mairal P., Nic, M., Dibusz, K., Novotny, T., Busquet, F., Rossi, F., Straccia, M., Daskalopoulos, E. P., and Gribaldo, L., *Advanced Non-animal Models in Biomedical Research – Immunogenicity testing for advanced therapy medicinal products*, EUR 30334/4 EN, Publications Office of the European Union, Luxembourg, 2022, http://doi.org/10.2760/7190

Celi, S., Cioffi, M., Capellini, K., Fanni, B. M., Gasparotti, E., Vignali, E., Positano, V., Haxhiademi, D., Costa, E., Landini, L., Daskalopoulos, E. P., Piergiovanni, M. and Gribaldo, L., *Advanced non animal models in biomedical research – Cardiovascular diseases*, Publications Office of the European Union, Luxembourg, 2022, http://doi.org/10.2760/94608

Deceuninck, P., Straccia, M., Whelan, M., 'BimmoH - Biomedical Models Hub', European Commission, Joint Research Centre (JRC), 1 December 2025, Dataset PID: http://data.europa.eu/89h/ba511666-1c31-4ac0-a4a4-97567e480aba

European Commission: *2022 Statistical Report: Summary Report on the Statistics on the Use of Animals for Scientific Purposes in the Member States of the European Union and Norway*, Commission Staff Working Document SWD (2024) 185 final. European Commission, 2024, https://environment.ec.europa.eu/topics/chemicals/animals-science_en

Fawcett, T., An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Folgiero, V., Romania, P., Rossi, F., Caforio, M., Nic, M., Dibusz, K., Novotny, T., Busquet, F., Straccia, M. and Gribaldo, L., *Advanced Non-animal Models in Biomedical Research: Breast Cancer*, EUR 30334/1 EN, Publications Office of the European Union, Luxembourg, 2020, http://doi.org/10.2760/618741

Freund, Y., Schapire, R. E., *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences, 1997, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H., *Greedy function approximation: A gradient boosting machine*, Annals of Statistics, 2001, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Harris, C. R., Millman, K. J., van der Walt, S. J., et al., *Array programming with NumPy*. Nature, 2020, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hynes, J., Marshall, L., Adcock, I., Novotny, T., Nic, M., Dibusz, K. and Gribaldo, L., *Advanced Non-animal Models in Biomedical Research: Respiratory Tract Diseases*, EUR 30334 EN, Publications Office of the European Union, Luxembourg, 2020, http://doi.org/10.2760/725821

Koshute, P., Zook, J., McCulloh, I., *Recommending Training Set Sizes for Classification*, 2021, https://arxiv.org/pdf/2102.09382

Le, Q., Mikolov, T., *Distributed Representations of Sentences and Documents*, Proceedings of the 31st International Conference on Machine Learning, in Proceedings of Machine Learning Research, 2014, 32(2):1188-1196, https://proceedings.mlr.press/v32/le14.html

Otero, M.J., Canals, J.M., Belio-Mairal, P., Nic, M., Dibusz, K., Novotny, T., Busquet, F., Rossi, R., Gastaldello, A., Gribaldo, L. and Straccia, M., *Advanced Non-animal Models in Biomedical Research – Autoimmune Diseases*, Publications Office of the European Union, Luxembourg, 2022, http://doi.org/10.2760/617688

Page M.J., McKenzie J.E., Bossuyt P.M., Boutron I., Hoffmann T.C., Mulrow C.D., et al. *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*, Systematic Reviews, 2021;10:89. http://doi.org/10.1186/s13643-021-01626-4

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts.* ArXiv. https://arxiv.org/abs/2205.01833

Reimers, N., Gurevych, I., *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3982–3992), 2019, Association for Computational Linguistics, https://arxiv.org/abs/1908.10084

Riekert, M., Klein, A., *Simple Baseline Machine Learning Text Classifiers for Small Datasets*, SN COMPUT. SCI., 2021, vol. 2, 178, https://doi.org/10.1007/s42979-021-00480-4

Romania, P., Folgiero, V., Nic, M., Dibusz, K., Novotny, T., Busquet, F., Rossi, F., Straccia, M., Daskalopoulos, E. P., and Gribaldo, L., *Advanced Non-animal Models in Biomedical Research: Immuno-oncology*, EUR 30334/3 EN, Publications Office of the European Union, Luxembourg, 2021, http://doi.org/10.2760/393670

Sayers E.W., Beck J., Bolton E.E., Brister J.R., Chan J., Connor R., Feldgarden M., Fine A.M., Funk K., Hoffman J., Kannan S., Kelly C., Klimke W., Kim S., Lathrop S., Marchler-Bauer A., Murphy T.D., O'Sullivan C., Schmieder E., Skripchenko Y., Stine A., Thibaud-Nissen F., Wang J., Ye J., Zellers E., Schneider V.A., Pruitt K.D., *Database resources of the National Center for Biotechnology Information in 2025*. Nucleic Acids Research, 2025 Jan 6;53(D1):D20-D29. https://doi.org/10.1093/nar/gkae979

Spärck Jones, K. (1972), *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, Journal of Documentation, 1972, 28 (1): 11–21, http://doi.org/10.1108/eb026526

Witters, H., Verstraelen, S., Aerts, L., Miccoli, B., Delahanty, A., Gribaldo L., *Advanced Non-animal Models in Biomedical Research – Neurodegenerative Diseases*, EUR 30334/2 EN, Publications Office of the European Union, Luxembourg, 2021, http://doi.org/10.2760/386

List of abbreviations and definitions

Abbreviations	Definitions
AI	Artificial Intelligence
BimmoH	Biomedical Model Hub
EU	European Union
EURL-ECVAM	EU Reference Laboratory for alternatives to animal testing
JRC	Joint Research Centre
ML	Machine Learning
NAM	Non-Animal Model

List of boxes

List of figures

List of tables

Annexes

Annex 1 – Glossary (key terms definitions used in this report)

Artificial Intelligence: Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

BimmoH Dataset: Database of scientific articles references making use of human biology-based models selected by BimmoH's ML Classifier and indexed with BimmoH's vocabularies.

BimmoH Vocabularies: list of key terms along with their definitions, used to index articles present in the BimmoH dataset.

EURL-ECVAM Biomedical Reviews: series of studies to review available and emerging non-animal models being used for research in seven disease areas published by EURL-ECVAM.

Human biology-based models: Systems or methods used to replicate or simulate human biological processes, diseases, or drug responses for research purposes, based on non-animal models.

in chemico model: Experimental approaches performed using chemical systems or reactions, often used to assess toxicity without animal testing.

in silico model: Computer-based modelling and simulation of biological systems, processes, or experiments.

in vitro model: Experiments conducted outside a living organism, typically in controlled laboratory environments such as petri dishes or test tubes.

Machine Learning classifier: Subset of Artificial Intelligence algorithms which build a model based on training data, to make predictions or decisions without being explicitly programmed to do so. In our case, the detection of research articles mentioning the use of human biology-based models.

Non-Animal model: research method that studies biological processes without using live animals or animal derived material (e.g., cells, organs).

Omics techniques: Comprehensive analytical methods such as genomics, proteomics, and metabolomics used to study biological molecules and systems.

PudMed: free database of biomedical and life sciences research articles, maintained by the U.S. National Library of Medicine used by BimmoH as primary source of data.

Annex 2. Outcome of the stakeholders' consultations

SWOT Analysis:

Features	Requests and Solutions	Priority	STRENGHTS	WEAKENESSES	OPPORTUNITIES	THREATS
Include studies that present both Human-based approaches and Animal experiments	Include reviews and peer-reviewed articles using both, animal and human, approaches and human-only approaches, tagging them accordingly.	High	Increases the inclusion of human-biology based Facilitates transition from animal to non-animal Provides study background information	Mix human and animal biology data	Larger information on human Potential engage of a larger end-	Automatic and/or direct extraction of It can be seen as not relevant by
Include Paywalled articles	Abstract and Material and Methods' sections can be sufficient.	High	Comprehensive data retrieval on human biology- Avoids major loss of studies using human It boosts relevancy for all researchers.	Facing legal issues.	Abstract and Material and Methods' Extreme relevance for all	Licensing and copyrights. Hampers open access.
Incorporate Detailed Methodological Information	Incorporate Materials and Methods section and Statistical Extract endpoints, molecules (testing compounds), cell Tag articles reporting the use of Omics (Genomics, Human physiological system, Organ, Tissue, Species categorisation can be better	High High High High	Essential for credibility and reproducibility Enhances quick evaluation of study quality Use of Omics enhance study credibility. Categorisation by physiological system can It provides better filtering for pre-clinical and	 It is a proxy assumption. Categorisation by disease can induce loss None	It eliminates the need for name Not necessary, IF Material and Not necessary, IF Material and Major engagement of extended	 Disease focus can be reduced.
Incorporate Detailed Bibliographic Information	Extract and display of: Authors' list, Corresponding Report or Link "Cited by" from PubMed	High High	It represents a proxy for credibility for certain Help end users.	None None	It can incentivise scientists in using Help finding similar studies.	It can induce a bias based on None
A Hierarchical, tiered approach to searching, indicating a preference for	Multi-step guided hierarchical search through filters Search box suggesting terms present inside the BD	High High	Need for accurate search results Facilitate data usage across research contexts	None None	It can be also guided based on Engage by easing the process.	Entry level end users can ask for None
Regular Updates and Information Retrieval Features as often as possible,	Implement regular updates, such as weekly, and real-time alerts	High	Ensures database remains relevant and reliable. Keeps users informed about new research.	Continue supervision of new entries and	Ensures database remains relevant	Maintanance costs
User profiles for saving and share:	Enable a user profile section for sharing of search outputs	High	Builds collaborative knowledge sharing	It requires dedicated and continuous IT	It boosts the end users use of the	
Systems for reporting miscategorised entries, online form with real time	Implement a transparent issue tracking system for reported miscategorized entries or issues. I.e. Including a	High	Enables continuous improvement. Encourages user engagement.	Back-end use of resources can be intense. It requires dedicated and continuous IT		If not properly followed,
Accessibility Across Multiple Platforms	Create an adaptive web application responsive to different One-page web format, no multiple windows.	High High	Necessary for access in various settings Broadens the database's user base			
Graphically engaging	Ensure a very visually appealing interface, that aligns in	High	Makes navigation intuitive and engaging	None	Attracts extended communities of	
Data Visualization and Interactive	Tabular view of search results and entries is welcome.	High	Makes navigation intuitive and engaging	None		
Video tutorials	Video must be short or cut into chapters, and cover	High	Better understanding of the navigation inside the	None		
PDF instructions with screenshots	Comprehensive PDF guides with annotated screenshots	High	For offline reference.	None		
FAQ	Incorporation of regularly update FAQ section based on	High	Online relevant Q/A to consult.	None		
Include Human Epidemiological studies	Include reviews and peer-reviewed articles on human	Medium/High	Highly relevant for Human Endpoints and		Ease the access and identification to	Overcomplicate the DB
Data Visualization as Interactive	Integrate advanced visualization tools, like Tableau, for	Medium-High	Aids comprehension of complex datasets.		Engage entry level users and allows	It can require IT supervision.
Externals Identificators or Ontologies	Link materials to external IDs or ontologies	Medium	Supports transparency and open science	There are not universally accepted	Very interesting future upgrade to	
Provide information on model	None	Medium	Facilitates standardised comparison among			
Regulatory sciences Information	Interconnect with other regulatory databases for up-to-date compliance information	Medium	Provides a comprehensive view of the Ensures content meets current standards		For the future upgrade.	
Validation/qualification and commercialization potential.	As proxy for TRL, we can identify if companies are among authors, by affiliation.	Medium	To increase study's credibility and quality	We have no tools and/or parameters to In the biomedical literature,	Future upgrade to target a specific	Developing parameters to be
IT and scientific supervised/curated	Continue supervision by IT and Scientific EC-JRC officers	Medium	Builds collaborative knowledge sharing	Continue supervision by IT and Scientific		
The preference for graphical abstracts	Not feasible.	Medium	To identify models and methods used.		Future upgrade, by using GenAI	
Focus on Clinical Research & Development	Categorize content following ICD-10 and MeSH terms Extract and provide clinical relevance information	Low Low	Aligns with practical clinical researcher needs. Supports translational clinical research.	It requires an extra IT development to	Very interesting future upgrade to	
Tag for use of animal-derived components.	Tagging of articles as suggested in Project Proposal	Low	Useful for a smaller community	Oversaturates information for the majority Achievable only in full-text analysis, not	Useful for future implementation of	
Retrieve of information on the use of	Create a dedicated section for guideline adherence if the	Low	To increase study's credibility		Not needed if M&M are shown	
E-mail alerts for user's defined	E-mail alerts integration	Low	Keep the end users up to date and engaged.	Only if user profile is available.	Upgrade phase to boost	
Establish a system for user content submission and model contribution	Living Document with User-Generated Content	Low	Allows organic growth with user contributions Includes diverse models and approaches			
Summary of result	The authors' abstract should be sufficient and relevant.	Low	To quickly identify relevance of the study.			
Lay summary based on the full text.	Incorporate AI-powered summaries and interpretations of study results, perhaps through external open-sources free services.	Low	To easily identify relevance of the study.			

Workshop feedback on prototype:

Researchers

1) Potential Applications and/or Interests:
- Use case of searching for models related to diseases (e.g., Zika virus) and organs.
- Benefits of avoiding epidemiological and clinical studies while searching for models.
- Interest in knowing the availability of cell lines (in-house vs. commercially available).
- Potential for including categories related to therapies, vaccines, and compounds tested on specific models.

2) Strengths:
- Noise reduction: The database helps filter out irrelevant studies (e.g., epidemiological and clinical studies) to focus on models.
- User-friendly format: The table format is considered more navigable and functional compared to systems like PubMed.
- Focus on model information: The system allows researchers to focus on models of interest, even if they have not been applied to specific organs yet.

3) Weaknesses:
- Limited scope: The database relies solely on open-access resources, which could limit the relevance and up-to-dateness of the content.
- Broad classification: The current classification of the biomedical area is seen as too broad, which may limit its usefulness for specific research needs.
- Missing categories: Important categories, such as pathogen types and specific cell type classifications (primary vs. immortalised cells), are absent or underdeveloped.
- Unnecessary data: The inclusion of specific years for models is questioned, as visual representations like graphs might suffice.
- Need for clearer definitions of acronyms, such as MPS (Microphysiological Systems).

4) Suggestions:
- Category enhancement: Add more detailed classifications for pathogen types and primary cell types (e.g., distinguishing between primary, immortalised, and other cell types).
- Classification of cell types and defining the tagging process based on histology and omics.
- Definition clarity: Provide clearer definitions for acronyms and technical terms, such as MPS, to help unfamiliar users.
- Model availability: Include information on whether cell lines are in-house or commercially available.
- Therapies and treatments: Add categories to track therapies, vaccines, or compounds tested on specific models to identify suitable models for testing various treatments (e.g., antivirals or antibiotics).

Project evaluators

1) Potential Applications and/or Interests

- The database can be used during early consultations with applicants, collaborators, or stakeholders for guidance on replacement alternatives in experimental designs.
- It is particularly useful for grant reviews, helping to justify the use of new animal models over in vitro alternatives.
- Evaluators can identify gaps in applications based on the models included in the database which is often a challenging process for competent authorities.

2) Strengths

- The database provides a valuable resource for justifying alternative models during grant reviews and evaluations.
- It helps users identify gaps and opportunities in experimental designs based on available models.
- The system has potential to provide regulatory reassurance by tagging in vitro models that have regulatory acceptance.

3) Weaknesses

- The current focus on disease models might alienate researchers conducting basic research.
- Users with limited backgrounds in specific disease terminologies may struggle with the interface.
- There is a lack of clarity in the cell type filters and information about commercially available versus in-house models.
- The database's focus on nine [A/N: ten] biomedical areas may confuse users from other research fields.
- The quality of papers included in the database could be inconsistent, particularly those not adhering to guidelines for reporting in vitro experiments.

4) Suggestions

- The interface should be made more intuitive to accommodate users with varying levels of expertise.
- Broader search options (e.g., organ types) should be included early in the search process to allow for more flexible exploration.
- Clear communication about the focus on nine [A/N: ten] biomedical areas should be provided to avoid confusion.
- Tooltips or question marks should be added to guide less experienced users through the database.
- Tagging papers that comply with guidelines for reporting in vitro experiments could improve content quality.
- The tool should minimise the need for extensive tutorials, ensuring it remains quick and easy to use.
- The ontology and terminology should be universally applicable and clearly defined.
- A feature that tracks a model's progress and acceptance over time, by linking papers, models, or research groups, would be valuable.

Industry/Pharma representatives

1) Potential Applications and/or Interests

- Method developers are interested in increasing the visibility of their publications while ensuring clients can find relevant studies.
- Identifying competitors for specific techniques is seen as important to better advise clients, particularly for regulatory submissions.
- The potential to discover alternative models derived from animal data is of interest for both research and industry use.
- The speaker suggests leveraging tThe database could be leveraged to find commonly combined methods and view related studies or techniques used by the same research groups.

2) Strengths

- The database has the potential to enhance the visibility of publications, benefitting service providers in terms of outreach and recognition.
- It could be a useful tool for advising on regulatory submissions by identifying competitors and alternative models.
- The ability to find related essays or studies, similar to Amazon's recommendation feature, would provide added value to users.

3) Weaknesses

- The database currently lacks certain bacterial disease models, which limits its comprehensiveness.
- The limited focus on specific disease areas may make it difficult for new researchers to find relevant models.
- A key limitation is that the database only relies on open-source information , which restricts access to a broader range of data, particularly proprietary or in-house information.
- A potential bug was reported when searching for publications related to breast cancer in vitro, with expected results not appearing.
- Negative results are often underreported, which poses a challenge when evaluating competitors and research data.

4) Suggestions

- Expand the database to include missing bacterial disease models and more disease areas to make it more useful for a wider range of researchers.
- Consider contacts in the chemical industry for insights into toxicology labs and their studies, which could enrich the database's toxicology-related content.
- Implement a feature that shows which methods are often combined in studies and how they are related to research conducted by the same groups.
- Introduce a recommendation system similar to Amazon's, where users can see related essays or studies based on their search.
- Address the bug related to breast cancer in vitro searches to ensure users receive accurate results.
- Explore ways to integrate proprietary or in-house data, where possible, to make the database more valuable for industry users who may otherwise guard such information.

Three Rs promotion

1) Potential Applications and/or Interests

- The database could be used to screen for additional data, particularly for regulatory and toxicological contexts.
- The inclusion of rare diseases in the database is seen as potentially valuable for researchers trying to locate specific models.
- A filtering option by country could facilitate connections with local experts and resources, adding to the practicality of the database for international research collaborations.

2) Strengths

- The database could be a powerful tool for identifying gaps in research fields, particularly in areas like reproductive toxicity.
- Including techniques used to analyse or retrieve biological endpoints could enhance its utility, especially for those in regulatory or toxicological roles.
- Connecting the database with existing resources could help avoid redundancy, optimise funding, and ensure more comprehensive access to information.

3) Weaknesses

- Certain metadata and columns from reports are missing from the current prototype, limiting usability.
- The lack of biological endpoints (as distinguished from biomarkers) reduces the database's relevance for regulatory or toxicological users.
- The limited disease areas currently available make it difficult for new researchers to find relevant information, which could impact the database's overall utility.
- There is uncertainty about how to accurately define the country of origin for research, as authorship may not always reflect this clearly.
- There are concerns about the quality control processes for data included or excluded from the database, which may affect trust in the database's content.

4) Suggestions

- Enhance the prototype by incorporating more comprehensive metadata and additional report columns to improve usability.
- Include biological endpoints in the database to cater to regulatory and toxicological users, and ensure clear distinctions between biomedical and regulatory terminology (biomarkers vs. endpoints).
- Expand the database's categories to cover more research fields, particularly in underserved areas like reproductive toxicity, and address gaps in available models.
- Introduce the ability to filter by country to help researchers connect with local experts and resources, although care should be taken in defining the country of origin for research.
- Consider integrating the database with existing resources to avoid redundancy, making it more efficient and maximising the use of available funding.
- Include rare diseases and ensure they are easily searchable through specific search terms.
- Standardise the categoriszation of cell lines and other research materials to improve consistency and effectiveness across the database.
- Provide greater transparency regarding quality control measures for data included in or excluded from the database, ensuring users can trust the information.

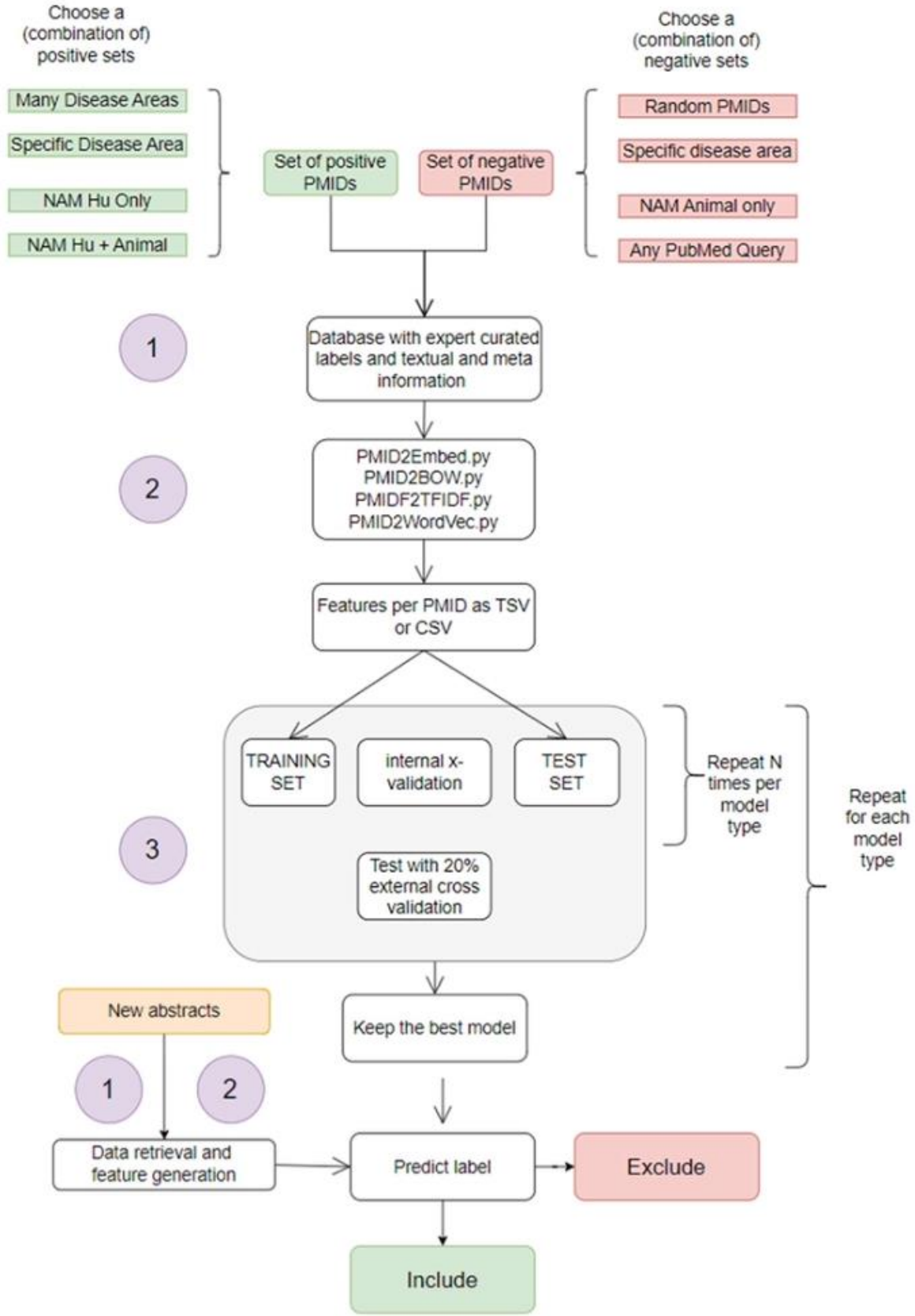
Annex 3. PubMed Pre-filtering Query

Article categories	PubMed query	Hits on 18/10/2024
Included	(Human* OR Homo Sapiens OR H. Sapiens OR "human-based model" OR "human-based models" OR human model OR patient) AND ("agent-based model" OR "agent-based models" OR "2d model" OR "2d models" OR "3d model" OR "3d models" OR "3d bioprint" OR "3d bioprints" OR "abm" OR "adherent cell" OR "adherent cells" OR "adherent suspension" OR "adherent suspensions" OR "adult stem cell" OR "adult stem cells" OR "aggregoid" OR "aggregoids" OR "air-liquid interface cell culture" OR "air-liquid interface cell cultures" OR "air-liquid interface culture" OR "air-liquid interface cultures" OR "airborne cell culture" OR "airborne cell cultures" OR "algorithm" OR "algorithms" OR "ali" OR "ann" OR "artificial intelligence" OR "artificial intelligences" OR "artificial neural network" OR "artificial neural networks" OR "assembloid" OR "assembloids" OR "bayesian inference" OR "bioprinted cell" OR "bioprinted cells" OR "bioprinted tissue" OR "bioprinted tissues" OR "biopsy" OR "biopsies" OR "bioreactor culture" OR "bioreactor cultures" OR "bioreactor cell culture" OR "bioreactor cell cultures" OR "blastoid" OR "blastoids" OR "boolean model" OR "boolean models" OR "cancer stem cell" OR "cancer stem cells" OR cell OR "cell aggregate" OR "cell aggregates" OR "cell co-culture" OR "cell co-cultures" OR "cell coculture" OR "cell cocultures" OR "cell line" OR "cell lines" OR "cell sheet" OR "cell sheets" OR "cell suspension" OR "cell suspensions" OR "cell-free" OR "cfd" OR "chemical gradient culture" OR "chemical gradient cultures" OR "chemical reactivity tests for oxidative stress induction" OR "civm" OR "classification model" OR "classification models" OR "cnn" OR "complex in vitro model" OR "complex in vitro models" OR "computational fluid dynamic" OR "computational fluid dynamics" OR "computational mechanic" OR "computational mechanics" OR "computational model" OR "computational models" OR "constraint-based model" OR "constraint-based models" OR "convolutional neural network" OR "convolutional neural networks" OR "data-driven model" OR "data-driven models" OR "deep learning model" OR "deep learning models" OR "digital patient" OR "digital patients" OR "digital twin" OR "digital twins" OR "dpra" OR "dynamic culture" OR "dynamic cultures" OR "electromagnetic model" OR "electromagnetic models" OR "electromechanical model" OR "electromechanical models" OR "electrophysiological model" OR "electrophysiological models" OR "electrophysiology model" OR "electrophysiology models" OR "electrospinning culture" OR "electrospinning cultures" OR "embryoid body" OR "embryoid bodies" OR "embryonic stem cell" OR "embryonic stem cells" OR "engineered tissue" OR "engineered tissues" OR "ex vivo" OR "explant culture" OR "explant cultures" OR "fea" OR "fem" OR "fetal cell" OR "fetal cells" OR "finite element analysis" OR "finite element analyses" OR "finite element model" OR "finite element models" OR "finite element simulation" OR "finite element simulations" OR "fluid-structure interaction model" OR "fluid-structure interaction models" OR "fsi model" OR "fsi models" OR "gastruloid" OR "gastruloids" OR "gaussian process model" OR "gaussian process models" OR "gel-based culture" OR "gel-based cultures" OR "genome alignment" OR "genome alignments" OR "genome-scale metabolic model" OR "genome-scale metabolic models" OR "gwas" OR "hanging drop culture" OR "hanging drop cultures" OR "heat transfer model" OR "heat transfer models" OR "hips" OR "hipsc" OR "homology model" OR "homology models" OR "immortalised cell" OR "immortalised cells" OR "immortalized cell" OR "immortalized cells" OR "in silico" OR "induced pluripotent stem cell" OR "induced pluripotent stem cells" OR "induced stem cell" OR "induced stem cells" OR "ips" OR	8,641,621

	"ipsc" OR "krigin model" OR "krigin models" OR "machine learning model" OR "machine learning models" OR "ml model" OR "ml models" OR model OR "mathematical model" OR "mathematical models" OR "mechanistic model" OR "mechanistic models" OR "microfluidic" OR "microfluidic chip" OR "microfluidic chips" OR "microgravity culture" OR "microgravity cultures" OR "microphysiological system" OR "microphysiological systems" OR "microplate culture" OR "microplate cultures" OR "mps" OR "multibody model" OR "multibody models" OR "multipotent stem cell" OR "multipotent stem cells" OR "multiscale model" OR "multiscale models" OR "nanobit" OR "neural network model" OR "neural network models" OR "nn model" OR "nn models" OR "numerical model" OR "numerical models" OR "ooc" OR organ OR "organ-on-a-chip" OR "organ-on-chips" OR "organ-on-chip" OR "organ-on-chips" OR "organoid" OR "organoids" OR "organoids on chip" OR "organoids on chips" OR "organoids-on-chip" OR "organoids-on-chips" OR "organoids-on-a-chip" OR "organoid-on-a-chip" OR "organoid-on-chip" OR "organoid on chip" OR "organotypic culture" OR "organotypic cultures" OR "patient-specific model" OR "patient-specific models" OR "pbpk model" OR "pbpk models" OR "pdo" OR "pharmacokinetic / pharmacodynamic model" OR "pharmacokinetic / pharmacodynamic models" OR "pharmacokinetic model" OR "pharmacokinetic models" OR "pharmacophore model" OR "pharmacophore models" OR "phenomenological model" OR "phenomenological models" OR "physiologically based pharmaco-kinetics model" OR "physiologically based pharmaco-kinetics models" OR "pinn" OR "pk" OR "pk/pd" OR "pluripotent stem cell" OR "pluripotent stem cells" OR "precursor cell" OR "precursor cells" OR "primary cell" OR "primary cells" OR "primary culture" OR "primary cultures" OR "primary tissue" OR "primary tissues" OR "primary tumour tissue" OR "primary tumour tissues" OR "progenitor cell" OR "progenitor cells" OR "protein-ligand interaction prediction" OR "qsar model" OR "qsar models" OR "quantitative structure-activity relationship" OR "reduced order model" OR "reduced order models" OR "regression model" OR "regression models" OR "rigid-body model" OR "rigid-body models" OR "scaffold 3d model" OR "scaffold 3d models" OR "scaffold-based model" OR "scaffold-based models" OR "scaffold-free tissue" OR "scaffold-free tissues" OR "scaffold model" OR "scaffold models" OR "sequence alignment" OR "sequence alignments" OR "sequencing" OR "simulation model" OR "simulation models" OR "smoothed-particle hydrodynamic" OR "sph" OR "spheroid" OR "spheroids" OR "ssm" OR "static culture" OR "static cultures" OR "statistical shape model" OR "statistical shape models" OR "stem cell" OR "stem cells" OR "stochastic model" OR "stochastic models" OR "surrogate model" OR "surrogate models" OR "swiss-model" OR "three-dimensional model" OR "three-dimensional models" OR tissue OR "tissue construct" OR "tissue constructs" OR "tissue explant" OR "tissue explants" OR "totipotent stem cell" OR "totipotent stem cells" OR "transwell culture" OR "transwell cultures" OR "two-dimensional model" OR "two-dimensional models" OR "virtual patient" OR "virtual patients" OR "virtual twin" OR "virtual twins" OR "whole organ" OR "whole organs") AND (review[pt] OR journal article[pt] OR systematic review[pt]) AND (english[Filter])	
Excluded	"Controlled Study" OR (Priority AND Journal) OR "Major Clinical Study" OR "Prognosis" OR "Follow Up" OR "Follow-Up" OR "Retrospective Stud*" OR "Prospective Study" OR "Case Control Study" OR "case stud*" OR "case-stud*" OR "Psychology" OR "Case Report" OR questionnaire* OR "Diagnostic Imaging" OR "Mammography" OR cross-sectional OR survey* OR "Meta-Analysis" OR "meta-analysis" OR "clinical trial*" OR gvhd OR "qualitative study" OR workshop OR sympos* OR "conference* proceeding*" OR cohort OR descent OR ancestr* OR participant* OR population OR "in silico trial" OR "in silico trials" OR (letter[pt] OR editorial[pt]	16,115,953

	OR case reports[pt] OR news[pt] OR comment[pt] OR interview[pt] OR biography[pt] OR bibliography[pt] OR congress[pt] OR directory[pt] OR festschrift[pt] OR introductory journal article[pt] OR lecture[pt] OR legislation[pt] OR news[pt] OR newspaper article[pt] OR overall[pt] OR portrait[pt])	
Total	#Included NOT #Excluded	4,601,409

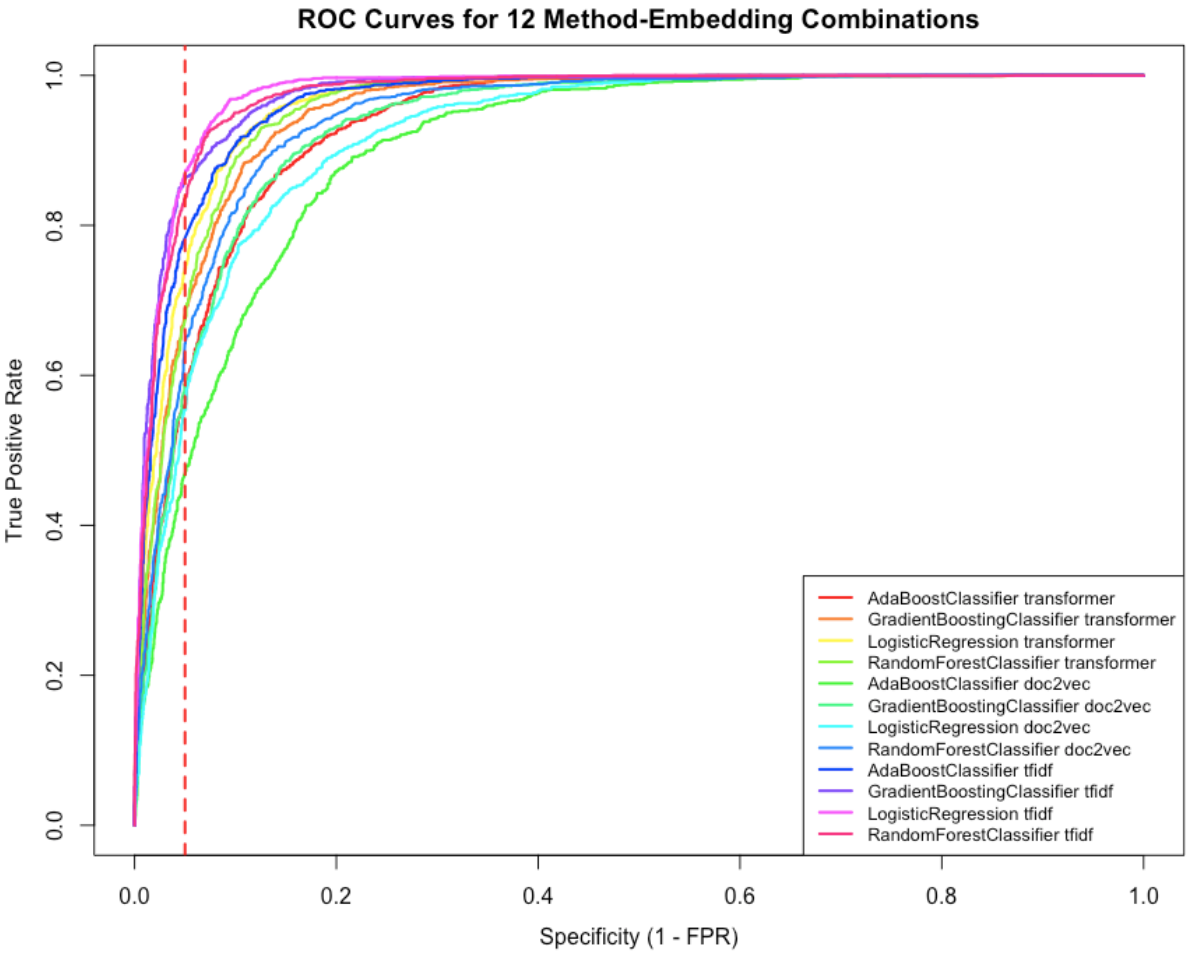
Annex 4. Classifier building workflow



Annex 5. Additional experimental results

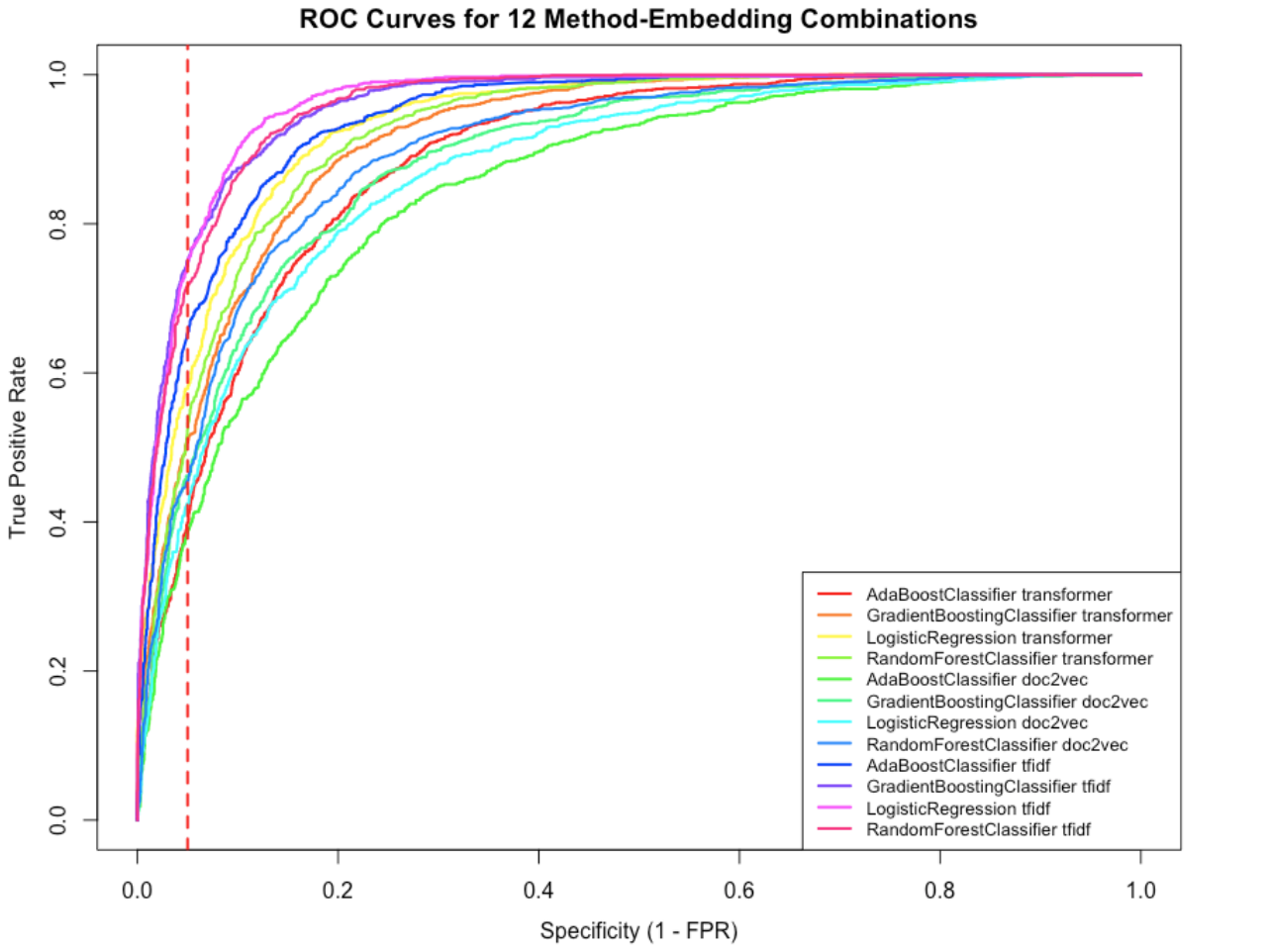
Experiment 1:

Method	Embedding	Threshold	Sensitivity	Specificity	Precision	Accuracy	F1_Score
AdaBoostClassifier	transformer	0.519387	0.560474	0.950199	0.825378	0.834931	0.667608
GradientBoostingClassifier	transformer	0.593359	0.688538	0.950199	0.853085	0.872808	0.76203
LogisticRegression	transformer	0.613065	0.742292	0.950199	0.862259	0.888707	0.797791
RandomForestClassifier	transformer	0.5315	0.672727	0.950199	0.85015	0.868132	0.751103
AdaBoostClassifier	doc2vec	0.528036	0.467194	0.950199	0.797571	0.807342	0.589232
GradientBoostingClassifier	doc2vec	0.592162	0.588142	0.950199	0.832215	0.843114	0.689208
LogisticRegression	doc2vec	0.678587	0.55415	0.950199	0.823737	0.833061	0.662571
RandomForestClassifier	doc2vec	0.5165	0.641107	0.950199	0.843913	0.85878	0.728661
AdaBoostClassifier	tfidf	0.504985	0.779447	0.951195	0.870256	0.900397	0.822352
GradientBoostingClassifier	tfidf	0.514217	0.86166	0.950199	0.879032	0.924012	0.870259
LogisticRegression	tfidf	0.499527	0.868775	0.950199	0.879904	0.926116	0.874304
RandomForestClassifier	tfidf	0.5435	0.836364	0.950199	0.875828	0.91653	0.855641



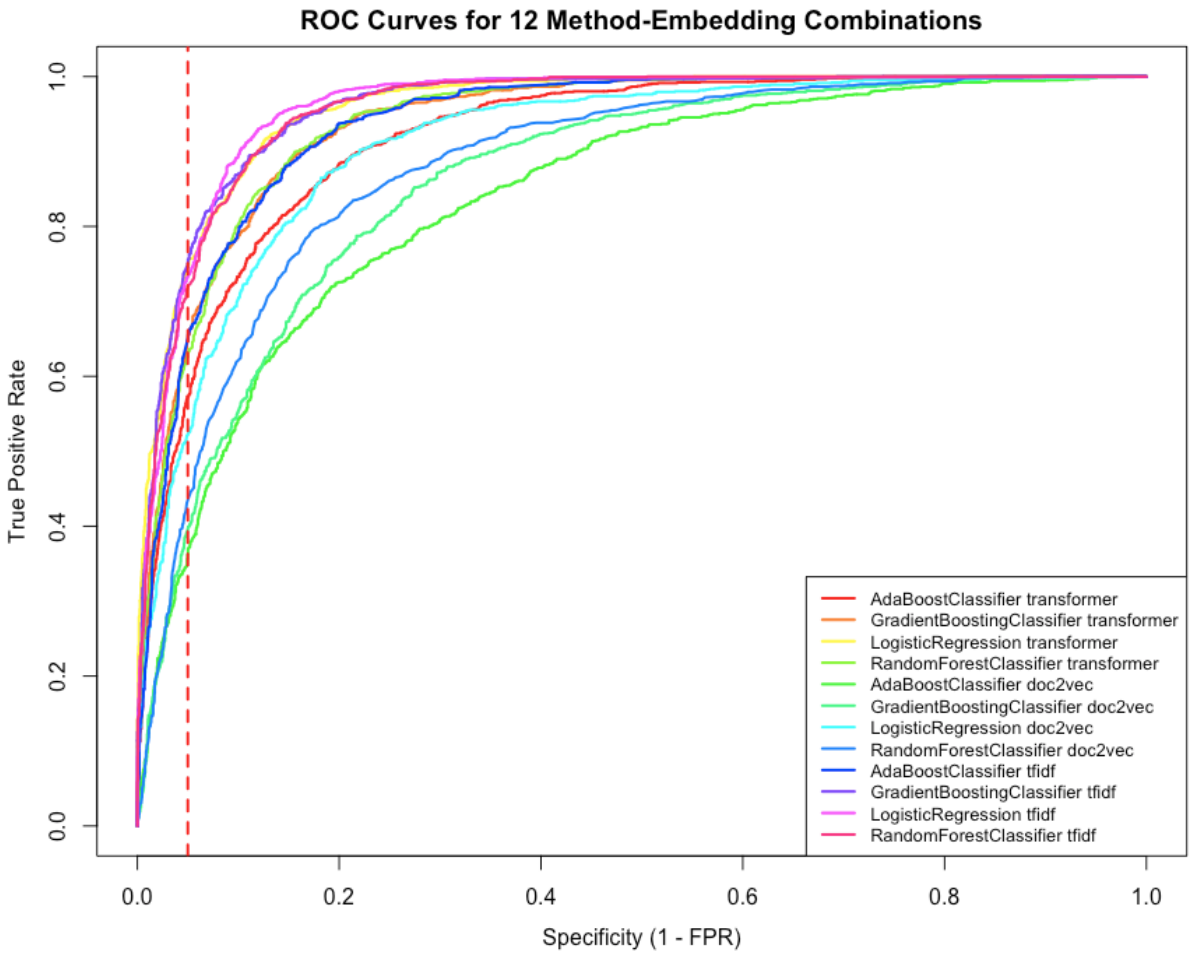
Experiment 2

Method	Embedding	Threshold	Sensitivity	Specificity	Precision	Accuracy	F1_Score
AdaBoostClassifier	transformer	0.534237	0.399209	0.950216	0.799051	0.767558	0.53242
GradientBoostingClassifier	transformer	0.646464	0.506719	0.950216	0.834635	0.803197	0.630595
LogisticRegression	transformer	0.685066	0.578656	0.950216	0.852154	0.827044	0.689266
RandomForestClassifier	transformer	0.5395	0.52332	0.950216	0.839037	0.8087	0.644596
AdaBoostClassifier	doc2vec	0.533658	0.385771	0.950216	0.793496	0.763103	0.519149
GradientBoostingClassifier	doc2vec	0.628395	0.46166	0.950216	0.821378	0.78826	0.591093
LogisticRegression	doc2vec	0.691149	0.424506	0.950216	0.808735	0.775943	0.556765
RandomForestClassifier	doc2vec	0.5485	0.454545	0.950608	0.820257	0.786164	0.584944
AdaBoostClassifier	tfidf	0.516868	0.650593	0.950608	0.867229	0.851153	0.743451
GradientBoostingClassifier	tfidf	0.580193	0.750988	0.950216	0.88208	0.884172	0.811272
LogisticRegression	tfidf	0.572973	0.743083	0.950216	0.880975	0.881551	0.806175
RandomForestClassifier	tfidf	0.5595	0.716206	0.950216	0.877057	0.872642	0.788512



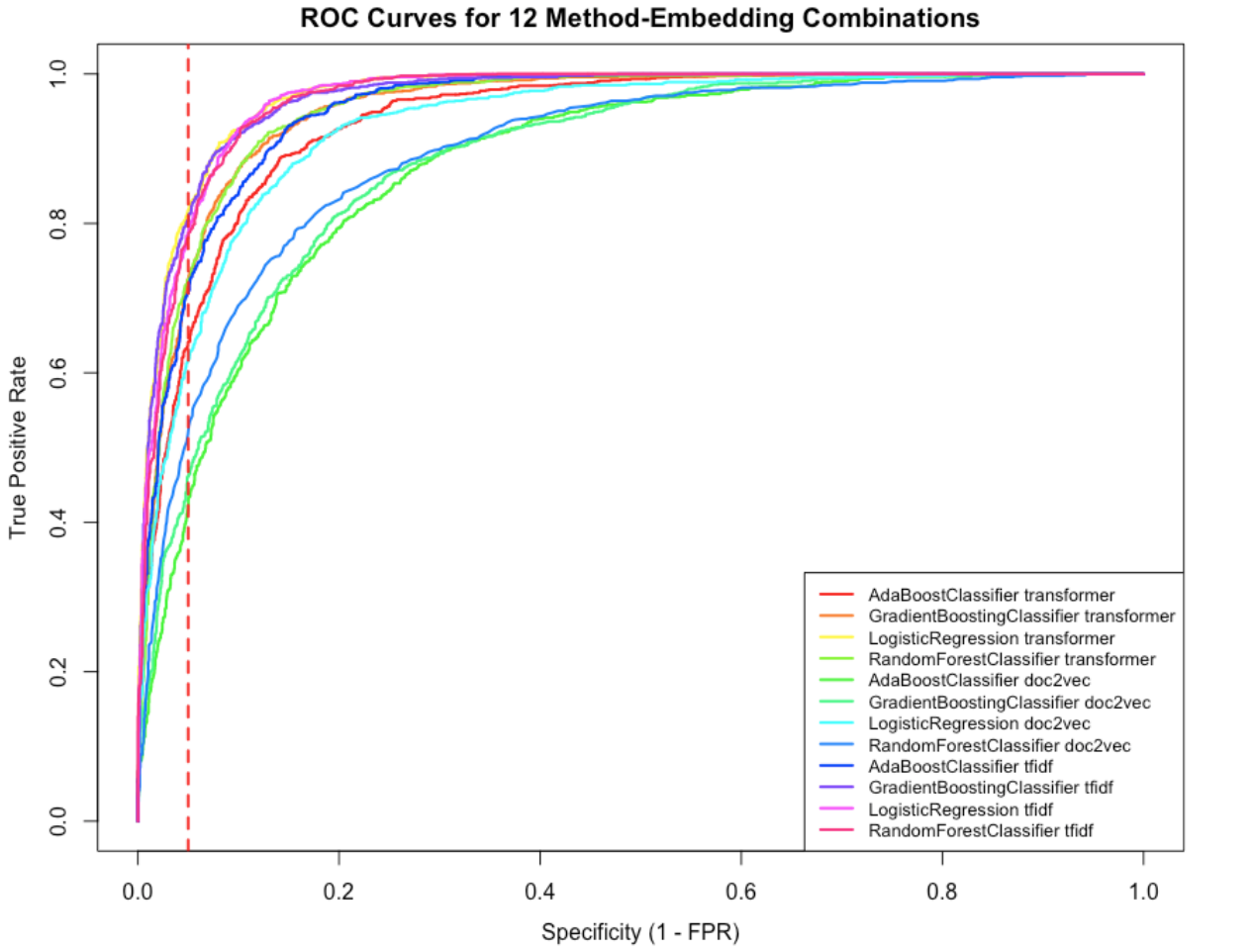
Experiment 3

Method	Embedding	Threshold	Sensitivity	Specificity	Precision	Accuracy	F1_Score
AdaBoostClassifier	transformer	0.523301	0.573123	0.950216	0.850939	0.82521	0.684932
GradientBoostingClassifier	transformer	0.621082	0.64664	0.950216	0.865608	0.849581	0.740271
LogisticRegression	transformer	0.703637	0.747826	0.950216	0.88164	0.883124	0.809239
RandomForestClassifier	transformer	0.5245	0.624506	0.950608	0.862445	0.842505	0.724438
AdaBoostClassifier	doc2vec	0.531041	0.362055	0.950216	0.782906	0.755241	0.495135
GradientBoostingClassifier	doc2vec	0.552166	0.396047	0.950216	0.797771	0.766509	0.529319
LogisticRegression	doc2vec	0.789856	0.521739	0.950216	0.838628	0.808176	0.643275
RandomForestClassifier	doc2vec	0.4355	0.43083	0.951	0.813433	0.778564	0.563307
AdaBoostClassifier	tfidf	0.514856	0.650593	0.950216	0.866316	0.850891	0.743115
GradientBoostingClassifier	tfidf	0.586268	0.750988	0.950216	0.88208	0.884172	0.811272
LogisticRegression	tfidf	0.579243	0.733597	0.950216	0.879621	0.878407	0.8
RandomForestClassifier	tfidf	0.5685	0.719368	0.950216	0.877531	0.87369	0.790617



Experiment 4

Method	Embedding	Threshold	Sensitivity	Specificity	Precision	Accuracy	F1_Score
AdaBoostClassifier	transformer	0.520153	0.641004	0.950216	0.857783	0.851575	0.733716
GradientBoostingClassifier	transformer	0.584086	0.725523	0.950216	0.872233	0.878537	0.792143
LogisticRegression	transformer	0.665804	0.817573	0.950216	0.884964	0.907902	0.849935
RandomForestClassifier	transformer	0.4935	0.723849	0.950608	0.872856	0.87827	0.7914
AdaBoostClassifier	doc2vec	0.530721	0.423431	0.950216	0.799368	0.782168	0.553611
GradientBoostingClassifier	doc2vec	0.536347	0.461925	0.950216	0.81296	0.794447	0.589114
LogisticRegression	doc2vec	0.749311	0.6159	0.950216	0.852839	0.843566	0.715258
RandomForestClassifier	doc2vec	0.4225	0.517992	0.950216	0.829759	0.812333	0.637816
AdaBoostClassifier	tfidf	0.515169	0.706276	0.950216	0.869207	0.872397	0.779317
GradientBoostingClassifier	tfidf	0.57829	0.803347	0.950216	0.883165	0.903364	0.841367
LogisticRegression	tfidf	0.552941	0.788285	0.950216	0.881197	0.898558	0.832155
RandomForestClassifier	tfidf	0.5395	0.780753	0.950216	0.880189	0.896156	0.827494



Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/write-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

Scan the QR code to visit:

Joint Research Centre
https://joint-research-centre.ec.europa.eu

Publications Office
of the European Union