Scientometrics
https://doi.org/10.1007/s11192-026-05868-8

Marebito: citation matching interface for the digital library zbMATH open

Madhurima Deb[1] · **Isabel Beckenbach**[1] · **Norman Meuschke**[2] · **Olaf Teschke**[1] · **Moritz Schubotz**[1]

Received: 28 May 2025 / Accepted: 6 October 2026
© The Author(s) 2026

Abstract
zbMATH Open is a leading curated bibliographic database covering approximately five million publications in mathematics. This paper presents Marebito, a production citation-matching system developed for zbMATH Open, and documents its architecture, domain-adapted feature design, and operational performance at scale. Marebito combines bibliographic parsing with GROBID, candidate retrieval with Elasticsearch, and a supervised Random Forest trained on domain-specific string-similarity features. Evaluation on internal zbMATH Open datasets yields precision and recall exceeding 98% on modern DOI-indexed records. Evaluation across two bibliographically distinct subsets reveals that classifier performance is strongly conditioned on metadata completeness, particularly title and author string quality, rather than DOI availability alone, documenting a concrete challenge for citation matching in mathematical literature. A preliminary comparison with Crossref on a manually verified subset of 2,037 entries illustrates the limitations of general-purpose matchers on specialist mathematical bibliography. A neural reranking baseline confirms that Elasticsearch-derived hard negatives are essential for abstention. We also discuss downstream bibliometric applications within zbMATH Open.

Keywords ZbMATH open · Citation matching · Bibliometrics · Digital library

✉ Madhurima Deb
Madhurima.Deb@fiz-karlsruhe.de

Isabel Beckenbach
Isabel.Beckenbach@fiz-karlsruhe.de

Norman Meuschke
Meuschke@uni-goettingen.de

Olaf Teschke
Olaf.Teschke@fiz-karlsruhe.de

Moritz Schubotz
Moritz.Schubotz@fiz-karlsruhe.de

[1] Department of Mathematics, FIZ Karlsruhe Leibniz Institute for Information Infrastructure, Franklinstrasse 11, 10587 Berlin, Germany

[2] Faculty of Mathematics and Computer Science, University of Göttingen, Papendiek 14, 37073 Göttingen, Germany

Published online: 04 November 2026

Introduction

zbMATH Open[1] is one of the most meticulously curated collections of bibliographic metadata in mathematics. Formerly known as Zentralblatt MATH, it provides bibliographic data and content details for publications across all branches of mathematics and its applications. As of 2024, zbMATH indexes approximately 5 million publications dating back to 1826, with roughly 120,000 new records added annually (Hulek & Teschke, 2023). The database became open-access in 2021 and encompasses mathematical software, formulae, reviews, serials, and classification data. It serves a diverse user base, including bibliographic clients (Wikimedia, arXiv, Zotero), aggregators (OpenAIRE/ERC, NFDI/DFG), archives (Software Heritage, Internet Archive), community platforms (MathOverflow), and individual researchers (Schubotz & Teschke, 2021).

Citations are a key source of information in scientometrics and a crucial metric for evaluating scientific output (Turki et al., 2021). A publication's credibility is often assessed by its citation count (Wilkinson et al., 2014), and citation analysis has advanced across classification, clustering, database analysis, and ranking (Parthasarathy & Tomar, 2015). Although citation counts are not an appropriate proxy for scientific excellence in mathematics (Adler et al., 2009; Arnold & Fowler, 2011; Bannister et al., 2017; Hulek & Teschke, 2024), their distribution may signal problematic behaviour such as the gaming of indicators (Hulek & Teschke, 2023). Citation matching, connecting reference strings to correct bibliographic entries in a database (Councill et al., 2006), plays a vital role in such analyses. Key challenges include extracting content from different versions of articles without noise, building global identifiers, and annotating citations with those identifiers (Färber et al., 2018).

Efficient citation matching enhances user experience, supports deduplication, and enables downstream analysis at zbMATH Open. It strengthens tasks such as author name disambiguation, self-citation mapping, and annotated dataset construction. Accurate matching also improves discoverability, helps researchers track the impact of their work, and facilitates navigation of research topics. A study on author disambiguation for DBLP (Kim, 2018) motivated our investigation into the citation-matching interface at zbMATH Open.

This paper documents the design, implementation, and evaluation of Marebito, a citation-matching system developed within zbMATH Open. Its primary contribution is a domain-adapted pipeline for mathematical literature, a domain whose bibliographic characteristics (metadata and citation norms distinct from those of other fields) present challenges that are underrepresented in the existing literature. The contributions are as follows.

First, we present Marebito: a supervised binary classification pipeline combining bibliographic parsing via GROBID,[2] candidate retrieval via Elasticsearch,[3] and a supervised classifier trained on domain-specific string similarity features. While the deployed classifier is a Random Forest, we also compare it with other supervised classifiers (Section "Classifier selection and comparison"). Evaluated on internal zbMATH Open datasets, Marebito achieves precision and recall exceeding 98% on modern DOI-indexed records, demonstrating that a carefully feature-engineered classical machine learning approach remains competitive for field-specific bibliographic matching.

[1] https://zbmath.org/.
[2] https://github.com/kermitt2/GROBID.
[3] https://www.elastic.co/.

Second, we provide a training and evaluation methodology grounded in two bibliographically distinct test sets, a modern DOI-indexed dataset (DOI) and an older dataset with sparse metadata (DLMF), and adopt domain-adapted *Informedness* as the primary evaluation metric to handle class imbalance inherent in operational citation-matching data. *Informedness* is selected as the most appropriate existing measure for this evaluation setting, given the skewed distribution of positives and negatives in both datasets; it is not introduced here as a novel metric.

Third, we present a focused comparison against the Crossref citation matcher on a manually verified subset of 2,037 DLMF entries, offering an externally grounded reference point for Marebito's performance on older, less structured mathematical literature and illustrating the limitations of general-purpose matchers in this domain.

The remainder of the paper is organized as follows. Section "Related work" reviews related work. Section "Methods" describes Marebito's architecture, training data generation, and evaluation procedure, including a comparison with Crossref and a neural baseline study. The *h*-index analysis in Section "Application of the citation matching interface: *h*-index as a downstream bibliometric use case" illustrates a downstream bibliometric application that accurate citation matching enables in zbMATH Open, thereby motivating the system's operational value beyond matching accuracy alone. Section "Discussion and conclusion" provides discussion and conclusions.

Related work

The introduction of citation indexes in science can be traced back to Eugene Garfield's 1955 publication (Garfield, 1955), which marked the inception of the commercial database Web of Science. Influential participants in bibliographic metadata curation include Web of Science (Stahlschmidt & Stephen, 2022), Scopus, Google Scholar, and CiteSeerX. Major contributors in the open citation domain are Crossref, OpenCitations, OpenAlex, and WikiCite.

CiteSeer, introduced in 1998, was among the first systems to combine citation parsing with similarity-based matching, enabling automated document interlinking in computer science and serving as an architectural precedent for domain-specific citation-matching systems such as Marebito (Giles et al., 1998). Text normalization steps in CiteSeer, including case folding, hyphen removal, and abbreviation expansion, illustrate how preprocessing choices directly condition the quality of downstream string similarity features, a consideration that also shapes Marebito's pipeline design. The string comparison algorithm, LikeIt, based on a technique similar to the Levenshtein distance (Giles et al., 1998), was also employed. Additionally, they implemented a document similarity detection method based on word vector space and TF-IDF (Term Frequency-Inverse Document Frequency). The parameter 'relatedness' of citations was determined using the CCIDF (Common Citation X Inverse Document Frequency) method, which outperformed the LikeIt-based automatic similar document extraction methods. In CiteSeer, a combination of these methods for similarity detection was used, resulting in a similarity measure equal to the weighted sum of the normalized similarity measures (Giles et al., 1998).

The ability to provide citation context has enabled interlinking and annotating citations across documents on CiteSeerX. Färber et al. developed a large-scale gold standard of 15.5 million sentences from arXiv computer science papers, linked to DBLP, to support

citation-based tasks, illustrating the importance and difficulty of constructing reliable annotated datasets for citation matching (Färber et al., 2018).

In the Crossref architecture, reference string parsing serves as the foundation of their citation-matching algorithm. They tested their citation-matching tool using an automatically generated dataset of 7,374 metadata records. Their search-based matching attained an average F1 score of 84.5%, and average precision of 99.2%. After parsing, they assessed the similarity between the input reference and candidate target documents to identify the most similar target document. The similarity calculation was performed field-wise (title/author/journal).[4] Additionally, the Crossref REST API is a useful option for citation matching[5,6] along with their new matching program.[7]

In the case of DBLP, a major digital library for computer science, existing citation metadata is made available through three platforms: CrossRef (the open metadata API), OpenCitations, and I4OC (Initiative for Open Citations). Upon successful matching of the target reference via its DOI, DBLP provides users with the record metadata, along with external links and export options for the articles.[8]

The LOC-DB project explored distributed infrastructure for open citation indexing, reflecting a broader effort in the field to build scalable, open-access citation management systems that motivates Marebito's design within zbMATH Open.[9] Another citation-matching study demonstrated DOI matching for the bibliometric databases WoS, Scopus, and Dimensions, yielding highly accurate matching results despite missing DOIs in 7%, 10%, and 1% of articles and reviews indexed in WoS, Scopus, and Dimensions, respectively, during the years 2016-2019. The DOI-based citation-matching method was validated by computing the Jaro-Winkler distance between the title strings of the versions across the three databases. They also included the difference in normalized citations between databases as a parameter (Stahlschmidt & Stephen, 2022). A tool called "Cited Reference Search" in Web of Science (WoS) can provide citations to a source and assess its influence in academic publications (Tomaszewski, 2021).

Very few traditional resources for mathematical research provide high-level quality control in the inclusion and attribution of data. MathSciNet and zbMATH Open are the leading candidates in this domain. MathSciNet, created by the American Mathematical Society, is a prominent database for abstracting and indexing mathematics literature (Dommermuth, 2017; Tomaszewski, 2021). The open access status of zbMATH Open has made it a favorable option for mathematical information retrieval. The correct interpretation of bibliometric data is an essential process in any digital library, and zbMATH Open has incorporated that by developing a citation matching tool and author disambiguation methods (Hulek & Teschke, 2023).

Figure 1 presents a snapshot of the Marebito API frontend, the user interface, and a schematic that illustrates the core functioning of Marebito's architecture at the inset.

[4] https://www.crossref.org/blog/matchmaker-matchmaker-make-me-a-match/.
[5] https://www.crossref.org/documentation/retrieve-metadata/rest-api/.
[6] https://www.crossref.org/labs/resolving-citations-we-dont-need-no-stinkin-parser/.
[7] https://www.crossref.org/community/special-programs/metadata-matching/.
[8] https://blog.dblp.org/2019/11/11/open-citation-data-and-dblp/.
[9] https://www.dagstuhl.de/en/dblp/news/2019/open-citation-data-and-dblp.

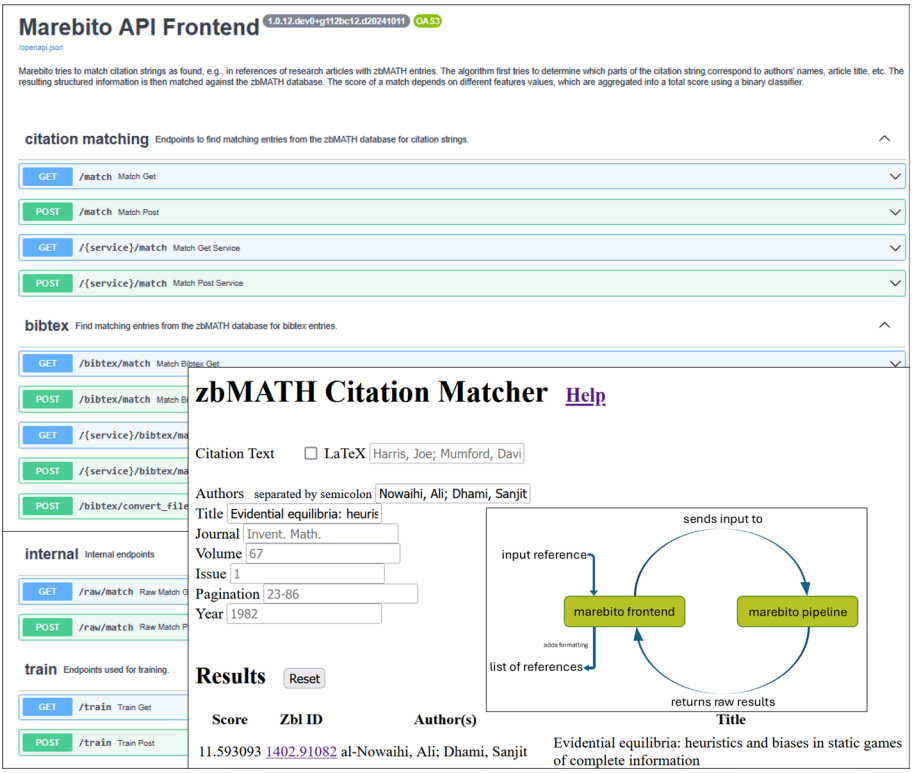


Fig. 1 Schematic showing the core functioning between the Marebito frontend and the pipeline

Methods

This section presents the citation-matching tool Marebito, describes its architecture, compares its performance with the Crossref citation matcher, and presents a neural re-ranker baseline study.

Architecture of Marebito

Marebito is a tool for solving the citation-matching problem at zbMATH Open. Formally, it accepts a reference string as input, along with optional structured information, such as the title, author, journal name, or publication year of the reference. The output is a list of corresponding zbMATH Open entries, with scores indicating how well they match the input query.

The main idea behind Marebito is to generate a set of candidates for each input reference, compute several similarity features for each candidate, aggregate those features into a score, and return the candidates that exceed a certain score threshold, see Fig. 2 for an overview. The architecture can be split into three parts:

- Parsing of the reference string
- Candidate Generation

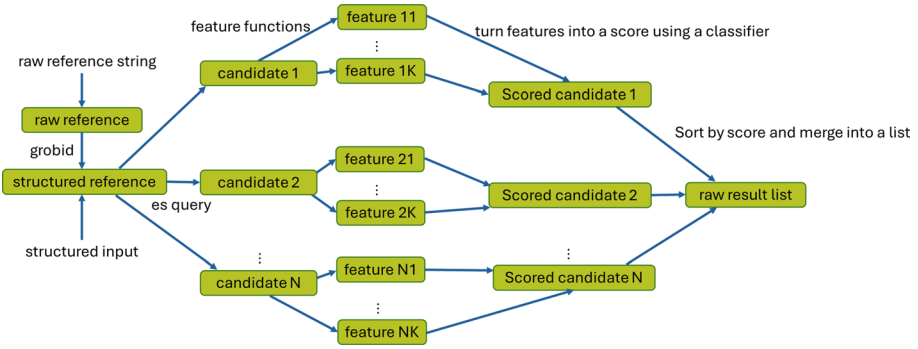


Fig. 2 A detailed view of the citation matching pipeline

- Scoring

For the parsing, we employ GROBID, a powerful machine-learning library specifically designed for parsing and extracting bibliographic data,[10] to extract structured details from the raw reference string.

Candidate generation is performed using Elasticsearch. Given an input reference string, a set of candidate items is created by building a search query using all fields available from the input reference string. The flexibility of Marebito enables the implementation of other candidate generation approaches; however, Elasticsearch was a fast and efficient solution in our setting.

For the scoring phase, candidates are first rated based on different features. Then, the features of each candidate are aggregated into a single score using a binary classifier.

The features Marebito uses to rate candidates can be adjusted via a configuration file. The flexible architecture makes it possible to easily add new features by implementing a new feature class that inherits from a given abstract base class. Currently, the following features are used:

- authors_lev: Measures the similarity of author names using the Levenshtein distance (Petrera et al., 2021).
- doi: The similarity of the input and candidate DOI is measured by the fraction of the overlap between the two DOIs.
- issue: A feature which takes the values 1 or -1, where 1 indicates that the input and candidate issue are the same, and -1 indicates that they differ.
- journal_prefix: A feature that computes the similarity of the input and candidate journal. It splits journal names into parts and checks whether parts of one journal name occur as a prefix of those of another.
- numbers_dice: Measures the similarity of all numbers occurring in the input and candidate item using the Dice-Sørensen coefficient (Sørensen, 1948) of the set of numbers of the input item and that of the candidate item.
- title_trigrams: The trigram similarity of the input and candidate title, where we use character-based trigrams.

[10] https://github.com/kermitt2/GROBID.

- raw_ref: A feature computing the character-based trigram similarity of the raw input reference string and the candidate reference string.
- volume: Similar to the issue feature, this feature is 1 if the volumes of the input and candidate features are the same, and -1 otherwise.
- pages: The similarity of the input pages and candidate pages is computed using the length of the intersection of the pages as a set.

The feature set above reflects deliberate adaptation to the bibliographic characteristics of mathematical literature. Journal names in mathematics are frequently cited in abbreviated or inconsistent forms across publications and time periods, motivating the journal_prefix feature, which checks prefix-based partial matches rather than requiring exact string equality. The numbers_dice feature captures the high diagnostic value of numerical strings, volume, issue, and page numbers, in mathematical references, where titles are often formula-rich and therefore, less discriminative as text strings than in other disciplines. The title_trigrams and raw_ref features together provide robustness to typographic variation and formula-rendering inconsistencies common in mathematical bibliographies. Taken together, these design choices constitute the primary locus of domain adaptation in Marebito: the pipeline architecture is general-purpose, but the feature engineering is grounded in the specific structural properties of mathematical reference strings.

Classifier selection and comparison

The binary classifier used to aggregate all features into one score is a Random Forest classifier. Other implemented classifiers include Decision Tree, Logistic Regression, and Stochastic Gradient Descent (SGD). They all inherit from an abstract base class, making it easy to add other classifier types to Marebito. All implemented models are supervised and require training, which we explain in the next subsection. *Informedness* serves as the basis for comparing classifiers: Marebito's architecture is classifier-agnostic by design, accommodating any supervised classifier without modification to the parsing, retrieval, or feature layers. Table 1 reports *Informedness*, precision, recall, accuracy, and confusion-matrix counts for four classifiers: Logistic Regression, Decision Tree, Random Forest, and SGD, trained on the same training set and evaluated on the same held-out set under identical conditions. Among the four classifiers, Logistic Regression achieves the highest *Informedness* (0.9844) and Random Forest the lowest (0.9636). This paper documents the Random Forest deployment of Marebito; re-evaluating the full pipeline with Logistic Regression is a concrete and prioritized direction for future work. While the Decision Tree produces fewer absolute false matches (FM: 77) and false positives (FP: 7), the weighted *Informedness* criterion penalizes false matches more heavily via the α parameter (explained in detail in Section "Informedness") and is therefore the appropriate metric for this asymmetric-cost

Table 1 Cross-classifier comparison on a fixed 50,000-pair evaluation subset

Classifier	Inf	Precision	Recall	Acc (%)	FM	FP
Logistic Regression	**0.9844**	0.9954	**0.9901**	**98.86**	166	16
Decision Tree	0.9741	**0.9978**	0.9767	97.99	**77**	**7**
SGD	0.9642	0.9907	0.9869	98.24	174	192
Random Forest	0.9636	0.9943	0.9728	97.42	165	54

task, rather than raw error counts alone. This comparison was conducted on a fixed subset of 50,000 evaluation pairs and 10,000 training pairs, drawn from the splits described in Section "Training procedure", to enable computationally feasible inclusion of the MiniLM-L6-V2 neural baseline (Section "Comparison with neural baseline") under CPU-only conditions.

Training procedure

Although the architecture of Marebito is classifier-agnostic, all implemented models are supervised and require training. The training procedure is part of Marebito, and works as follows.

- Generate a set of correct gold data, consisting of pairs of references together with the identifier of the correct zbMATH entry.
- Use this gold data to generate a set of feature vectors together with positive/negative labels.
- Train the classifier on the set of feature vectors.

The gold data consists of input references along with the corresponding zbMATH document. Each input reference in the gold data is a structured citation string derived from the raw bibliographic text as it appears in the reference list of a published article, collected from external sources, including publishers and arXiv. It is generated using the existing reference and DOI information at zbMATH Open. Currently, zbMATH Open contains more than 56 million references from various external sources. For some references, we have the corresponding DOI. References with a DOI are matched against zbMATH entries; successful matches yield labeled training pairs. In this way, we get approximately 3,971,839 labeled pairs. These are partitioned into a training set (2,780,315 pairs, 70%), an evaluation set used for model selection and classifier comparison (993,065 pairs, 25%), and a held-out test set used for final performance reporting (198,459 pairs, 5%). The test set is only used for the final evaluation and is not exposed during training. Random Forest was the classifier deployed for the evaluation reported here. The reduced-subset comparison in Table 1 shows a different ordering, with Logistic Regression achieving the highest *Informedness*; re-evaluating the full pipeline with Logistic Regression is a prioritized direction for future work Discussion and conclusion.

The construction of the gold data consists of pairs of input references and zbMATH documents that yield positive samples for the classifier to be trained, i.e., real positives (RP). Due to incomplete DOI coverage at zbMATH Open, we cannot construct real negatives (RN) with this approach. However, for each input reference in the gold data, we can use the Elasticsearch candidate generation procedure to construct N negative samples, namely, all candidates that do not correspond to the correct zbMATH document given in the gold data.

For each input reference and correct zbMATH pair of the gold data, we rate them using the chosen features. If we have k features, then we get a k-dimensional vector in this way. The target value of this vector is 1 if the candidate equals the correct zbMATH document in the gold data, and 0 otherwise.

The resulting training and target vectors are then used to train a scikit-learn model (Pedregosa et al., 2011), in our case, one of Random Forest, Decision Tree, SGD, and most recently, Logistic Regression. All classifiers were trained using scikit-learn default

Table 2 Random Forest feature importance scores (MDI)

Feature	Importance
numbers_dice	0.282
title_trigrams	0.207
raw_ref	0.181
pages	0.162
volume	0.078
authors_lev	0.066
journal_prefix	0.014
doi	0.003
issue	0.002

hyperparameters; no formal grid search or cross-validated hyperparameter optimization was performed, and this remains a direction for future work. For the classifier comparison reported in Table 1, a fixed random subset of 50,000 pairs was drawn from the evaluation set. This subset was used uniformly across all classifiers: Logistic Regression, Decision Tree, SGD, Random Forest, and the MiniLM-L6-V2 neural baseline described in Section "Comparison with neural baseline", to ensure a fair and computationally feasible comparison. The neural baseline evaluation is conducted on CPU only, which constrains the evaluation subset size.

Feature importance scores from the trained Random Forest, computed via scikit-learn's mean decrease in impurity (MDI) criterion,[11] are reported in Table 2. The three most important features are numbers_dice (0.282), title_trigrams (0.207), and raw_ref (0.181), together accounting for approximately 57% of the total importance. The doi feature carries near-zero importance (0.003), indicating that classifier performance is not primarily conditioned on DOI availability but rather on string-similarity signals present even in older records with sparse metadata. In Table 2, feature importance is computed via mean decrease in impurity (MDI) averaged across trees in the trained Random Forest classifier. Scores sum to 1.0.

The feature importances in Table 2 are impurity-based (MDI) importances, computed from the fitted Random Forest, and therefore, characterize the trained classifier globally rather than any individual test set. They should not be interpreted as test-set-specific importances for the DOI or DLMF subsets, because MDI is a property of the trained model; computing it separately for each evaluation set would return the same values. We note that training separate classifiers on the DOI and DLMF subsets could, in principle, reveal domain-specific differences in feature salience; we identify this as a direction for future work. MDI can also overestimate the importance of high-cardinality features.

Evaluation

We evaluate the performance of Marebito on a set of input pairs, each with an input reference and the corresponding correct zbMATH document from the gold data. For an input reference, there are three possibilities for the Marebito output:

[11] https://scikit-learn.org/0.23/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.feature_importances_.

- Marebito outputs the correct zbMATH document with score > 0.5 as the best match, which we call a true positive.
- Marebito returns no document with score > 0.5, which means it fails to find the correct zbMATH document. This is called a false negative.
- Marebito returns a document that differs from the correct zbMATH document as the best result with a score > 0.5, which we call a false match.

Usually, in a binary classification problem, one only considers true positives and false negatives, not false matches. In our setting, it makes sense to distinguish between false negatives and false matches, as false matches are worse. A false match means that at zbmath.org, a wrong link is displayed to the user, whereas a false negative means that a link is missing. Users perceive incorrect links as worse, on average, than missing ones.

The gold data construction yields no confirmed real negatives, since an unmatched DOI cannot be unambiguously interpreted as evidence that the referenced work is absent from zbMATH Open. Due to incomplete DOI coverage, we cannot construct such references using the DOI information. If a reference has a DOI and this DOI does not correspond to a zbMATH document, it could be that the article with this DOI is not indexed at zbMATH Open, or it could be that the DOI is just missing at the corresponding zbMATH document. Therefore, another approach is needed to create 'artificial real negatives'. We fix a positive integer K and assume that all database entries with IDs divisible by K are removed. Thus, each reference in the evaluation set, whose internal identifier is divisible by K, is considered as a negative sample.

We acknowledge that training and evaluation negatives are constructed differently. Training negatives are hard negatives produced by Elasticsearch retrieval: candidates that are bibliographically similar to the query but incorrect and therefore represent the most confusable non-matches the classifier encounters in practice. Evaluation negatives, by contrast, are constructed via the K removal method, which is not conditioned on bibliographic similarity. This asymmetry is an inherent consequence of the absence of confirmed real negatives in zbMATH Open: because a missing DOI may reflect missing metadata rather than a genuine absence from the database, we cannot construct verified non-matches directly from the index. Despite this asymmetry, the K construction constitutes a valid evaluation proxy because it simulates the operationally relevant failure mode, a query whose correct answer is absent from the index, and therefore directly tests the classifier's ability to abstain from returning a false positive when no retrievable correct match exists, which is the error of greatest consequence in production use.

For an input reference in Real Negatives (**RN**), there are the following possibilities:

- Marebito returns at most one zbMATH document with score > 0.5, and this is the correct zbMATH document from the evaluation data (which we treat as "missing"). This case is called a true negative (**TN**).
- The best match is a different document with score > 0.5, which is treated as a false positive (**FP**).

In summary, if A is the total number of test pairs, they are divided into RP real positive pairs and RN real negative pairs. After running Marebito on the pairs, we get TP true positives and TN true negatives, which are pairs matched correctly, and FN false negatives, FM false matches, as well as FP false positives, which are pairs matched incorrectly. In total, we get the following equations:

$$\begin{aligned} A &= RP + RN \\ RP &= TP + FM + FN \\ RN &= TN + FP \end{aligned}$$

As a final evaluation metric we use the so-called *informedness*[12] given by:

$$\text{inf} = \frac{TP}{RP} - \frac{FP}{RN}$$

Performance comparison on DOI and DLMF data

We have evaluated the Marebito classifier using techniques such as Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves to gain a complete understanding of the evaluation process, on two test datasets: 'DOI' and 'DLMF'. These two data sets are quite different: whereas the first is characterized by perfect DOI coverage and quite good data quality, the second has many missing DOIs and contains many older articles than the average at zbMATH Open.

The DOI dataset

The dataset, containing 198,355 items, consists of citation-matching results on zbMATH Open articles and is denoted as 'DOI'.

- key: This refers to the BibTeX key of the input item.
- correct_de: This refers to the DE number of the corresponding item in zbMATH Open.
- match_de: The DE number given as an output of the citation matching algorithm.
- score: This refers to the score given by the citation matching algorithm. The higher the score, the higher the probability that the matching result is correct. The algorithm gives a score of -1 when there is no match.
- rank: This refers to the rank of the correct result in the result list of the given algorithm. If the correct result is not among the top 10, the rank is set to -1. If the best result is the correct result, then the rank is 1.

The DLMF dataset

DLMF, or the digital library of mathematical functions, is an online resource containing reliable information on special mathematical functions, mathematical formulae, and their relevant information. DLMF is well-integrated with zbMATH Open, and the 36 chapters and 75% of its 2800 bibliographic references are directly linked to zbMATH Open (Ehsani et al., 2023). DLMF provides its citation information in BibTeX format. The BibTeX endpoint of Marebito produces the citation-matching results of DLMF articles indexed in zbMATH Open. These citation-matching results form our dataset named 'DLMF'. This dataset contains 2963 items. The excerpt from the DLMF dataset is shown in Fig. 3, along with the excerpt from the DOI dataset. The definitions of the columns are given below:

[12] https://rvprasad.medium.com/informedness-and-markedness-20e3f54d63bc.



Fig. 3 Excerpt from the DOI and DLMF dataset showcasing TP, TN, FP, FN, and FM

- key: This refers to the BibTeX key of the input item.
- dlmf: This refers to the Zbl number of the corresponding item in zbMATH Open. In this regard, we define a real positive as a BibTeX entry with a Zbl number and a real negative as a BibTeX entry with no Zbl number given. The real negative is indicated by a '-1' in the 'dlmf' column of the dataset.
- zbMATCH: The Zbl number given as an output of the citation matching algorithm.
- score: This refers to the score given by the citation matching algorithm. The higher the score, the higher the probability that the matching result is correct. The algorithm gives a score of -1 when there is no match.
- rank: This refers to the rank of the correct result in the result list of the given algorithm. If the correct result is not among the top 10, the rank is set to -1. If the best result is the correct result, then the rank is 1.

We studied the false positive rate (FPR) and false match rate (FMR) over classifier scores from 0.5 to 1 (equal to `predict_proba`), for both the DOI and DLMF datasets. Higher scores correspond to higher probabilities of a positive sample: a score ≤ 0.5 yields a predicted negative, and a score > 0.5 yields a predicted positive.

The number of false positives and false matches of the DOI dataset over this range are shown in Fig. 4a and b. The FPR corresponding to score 0.5 (number of false positives ≈ 400) is around 1.2%, and the FMR corresponding to the score 0.5 is 1.4% (number of false matches ≈ 500). Both the FPR and FMR decrease monotonically as `predict_proba` increases.

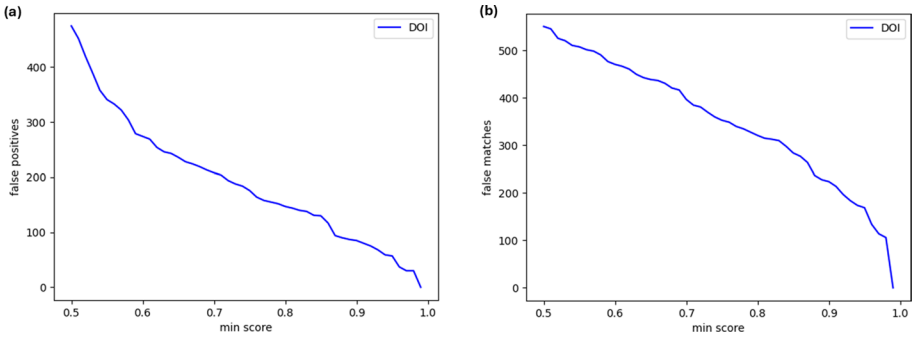


Fig. 4 False positives and false matches for the DOI dataset for the minimum score range between 0.5 and 1

Similarly, the false positives and false matches of the DLMF dataset are shown in Fig. 5a and b, respectively. For DLMF, the FPR exceeds 3% (number of false positives $\approx$ 16), and the FMR exceeds 8% (number of false matches $\approx$ 50), substantially higher than the DOI dataset. This performance degradation is attributable to feature-level differences between the two datasets. The three highest-importance features (numbers_dice: 0.282, title_trigrams: 0.207, raw_ref: 0.181; Table 2) collectively account for $\sim$ 57% of classifier decisions. DLMF records, which predominantly contain older publications with sparse and inconsistent metadata, offer diminished signals on these features: formula-rich titles yield weaker trigram matches, incomplete pagination removes discriminative page-number information, and inconsistent journal abbreviations reduce journal_prefix matching accuracy. Conversely, the doi feature (importance: 0.003), which could partially compensate for sparse metadata, provides a negligible signal for the $\sim$ 26% of DLMF entries lacking DOI coverage. The DOI dataset, by contrast, provides complete metadata across all high-importance features, resulting in stronger candidate discrimination. Step-like drops in FMR and FPR are observable as the minimum score increases; this behaviour is expected for a dataset as small as DLMF (2,963 entries). For example, the number of false matches was approximately 40 for minimum scores in [0.7, 0.8), dropping to approximately 20 for [0.8, 0.9).

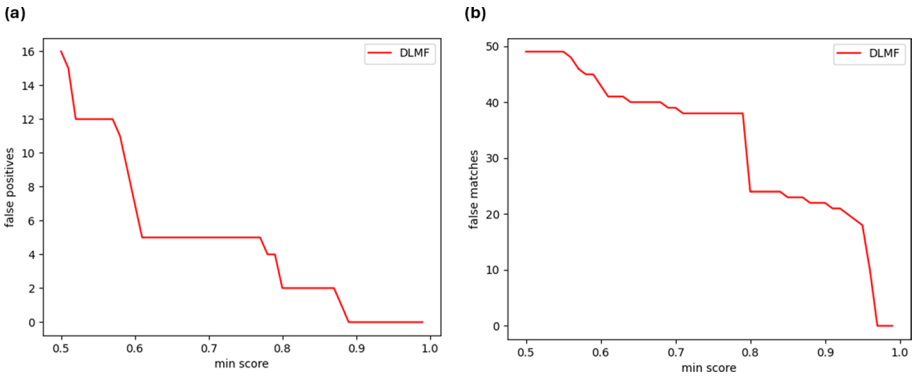


Fig. 5 False positives and false matches for the DLMF dataset for the minimum score range between 0.5 and 1

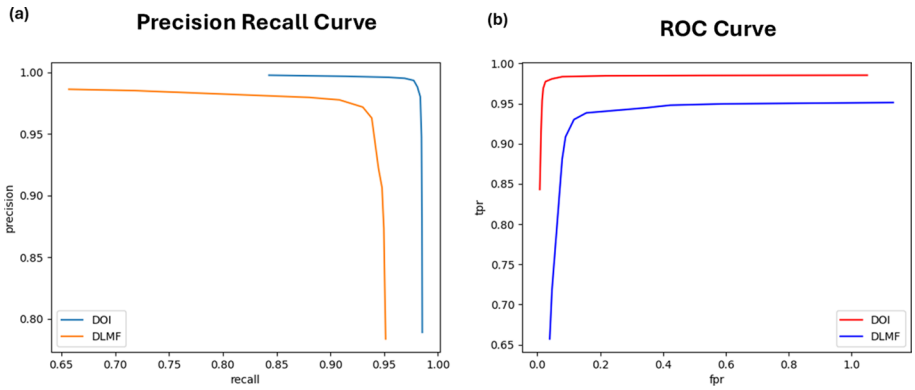


Fig. 6 Precision Recall Curve and ROC Curve for the citation matcher on DLMF and DOI datasets

There is a significant size disparity between the DOI (198,355 entries) and DLMF (2,963 entries) datasets, which affects all evaluation parameters discussed below.

The DLMF dataset also contains articles indexed multiple times in zbMATH Open due to multiple published versions. For example, the entry `Rowe:1931:PTE` has input identifier 57.0422.01,[13] but an alternative version exists with Zbl number 0001.21301.[14] Although Marebito matches the correct article, it returns a different identifier from the one in the DLMF dataset, producing a false negative (FN) rather than a true positive (TP).

False matches arise for analogous reasons. The entry `Courant:1953:MMP` has Zbl number 0051.28802 (English version),[15] but Marebito returns 0053.02805 (German version),[16] constituting a false match (FM). These cases together explain the inferior performance of Marebito on the DLMF dataset relative to the DOI dataset.

Evaluation methods

In this section, we introduce further methods to evaluate the performance of Marebito and apply it to the DOI and DLMF datasets described in the previous section. Both datasets are skewed datasets with positive cases as the majority. We chose to evaluate the performance of Marebito using three different methods which are suitable for such datasets:

- Precision-Recall analysis
- ROC analysis
- Informedness

[13] https://zbmath.org/?q=57.0422.01.
[14] https://zbmath.org/0001.21301.
[15] https://zbmath.org/0051.28802.
[16] https://zbmath.org/0053.02805.

Precision-recall and ROC

The precision-recall curve or PR curve plots the precision against recall for different values of 'threshold', in our case termed as 'minimum score' or 'min score'. The precision-recall curve is plotted based on the performance of the citation matching tool on the two datasets, DLMF and DOI. The precision value for the DLMF dataset reached nearly 1 at a recall value of 0.65, implying that the precision was quite high but the recall was low. The highest recall value attained for the DLMF dataset was 0.95, as shown in Fig. 6a. For the DOI dataset, the precision value is 1 at a recall value of 0.82. The higher recall value for the DOI dataset proves the classifier's ability to identify larger portions of true positives as compared to the DLMF dataset. The DOI PR curve is closer to the point (1,1) in the graph as compared to the DLMF PR curve, and it also substantiates the better performance of the classifier towards the DOI dataset. The performance difference between the two datasets is more accurately attributed to dataset composition and feature availability than to dataset size. The DOI test set consists entirely of modern, DOI-indexed mathematical publications for which all or most of Marebito's features: author name similarity, DOI overlap, journal prefix, title trigrams, volume, issue, and page information are populated. The DLMF dataset, by contrast, consists largely of older publications spanning the full historical bibliography of special functions literature, for which structured metadata is frequently sparse. This is consistent with the feature importance analysis in Section "Training procedure", which shows that the doi feature contributes negligibly to classifier decisions. As documented in Section "Performance comparison on DOI and DLMF data", the elevated error rates for DLMF dataset are characteristic of cases where candidate disambiguation must rely entirely on title, author, and journal string similarity, precisely the conditions under which hard negatives (references that are genuinely similar but distinct) are most likely to confuse the classifier. Inspection of DLMF false matches reveals some recurring structural patterns. First, edition and translation pairs, as discussed in Section "Performance comparison on DOI and DLMF data", can produce false matches when edition-discriminating metadata is absent. Second, entries for which no structured title or author information can be extracted from the raw reference string by GROBID, common in older citation formats that do not conform to modern bibliographic conventions, result in false negatives, as the candidate set generated by Elasticsearch is insufficiently constrained as further documented in Section "Hard-negative case analysis".

We analyzed the Receiver Operator Characteristic (ROC) curve, which plots the TPR, i.e., true positive rate, vs. the FPR, i.e., false positive rate. This helps us visualize and organize our classifier to analyze the relative trade-off between benefits (true positive) and costs (false positive) (Fawcett, 2006). This turned out to be beneficial for our skewed dataset. In some studies, ROC for unbalanced datasets has shown misleading model evaluation when the negative class is small, because, with a small negative class, even many false positives will not significantly inflate the false positive rate (FPR). The dominance of true positives can significantly impact the curve's shape, effectively masking errors related to the minority class (Movahedi et al., 2023). PR curve focuses on the performance of the model in identifying positive cases. As precision is sensitive to false positives, the PR curve will showcase the model's ability to avoid misclassifying negative cases as positive. As the focus is on the positive predictions, the large amount of true positives will not mask the effects of the false positives as much as they do in the ROC curve. For this reason, we evaluated the Marebito classifier using both the PR curve and the ROC curve. We

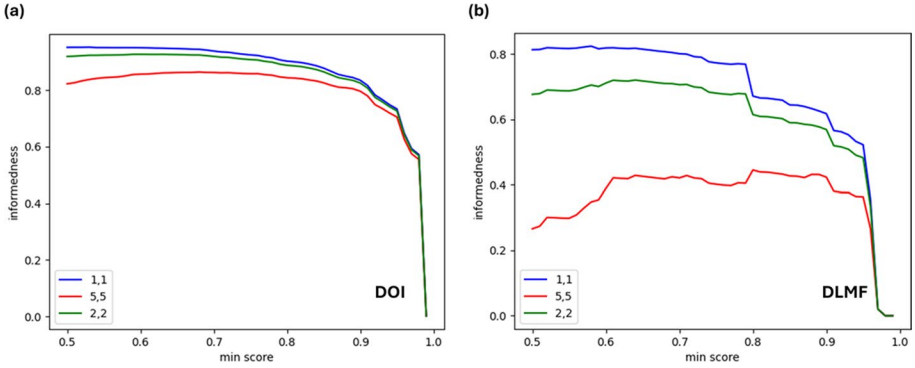


Fig. 7 Informedness for Marebito on DOI and DLMF datasets (Deb et al., 2024)

incorporate the third metric, *informedness*, in the next section to take the unequal cost of classification error into account.

When analyzing the ROC curve, it is well known that a perfect classifier would have a score in the top left corner, having FPR = 0% and TPR = 100% (Powers, 2020), whereas the worst case classifier has a score in the bottom right corner, having FPR = 100% and TPR = 0%.

It was observed that the DOI ROC lies closer to the point (0,1) compared to the DLMF ROC, for the entire range of FPR, as shown in Fig. 6b. The highest TPR value reached for the DOI dataset was around 0.98 and the TPR values were in the range of 0.84 and 0.98. We conclude by analyzing the PR curve and the ROC curve that the overall performance of the classifier is excellent for both datasets and slightly better for the DOI dataset, which is quite expected.

Informedness

Informedness[17] quantifies classification effectiveness by balancing the true positive rate (TPR) against the false positive rate (FPR). It is selected as the primary evaluation metric for Marebito over precision, recall, and ROC, as it is well-suited for the skewed class distributions, with varying rates of real positives (RP) and real negatives (RN), present in both evaluation datasets (Müller & Teschke, 2016). In Marebito, the formula for *informedness* is slightly modified from Müller and Teschke (2016) to account for the relative penalties assigned to false negatives, false matches, and false positives.

The formula defining *informedness*, denoted as $\inf_{\alpha,\beta}$ is given below.

$$\inf_{\alpha,\beta} = \frac{\mathrm{TP}}{\mathrm{RP}} - (\alpha - 1)\frac{\mathrm{FM}}{\mathrm{RP}} - \beta\frac{\mathrm{FP}}{\mathrm{RN}} \tag{1}$$

Equation 1 can be written as,

$$\inf_{\alpha,\beta} = 1 - \frac{\mathrm{FN}}{\mathrm{RP}} - \alpha\frac{\mathrm{FM}}{\mathrm{RP}} - \beta\frac{\mathrm{FP}}{\mathrm{RN}} \tag{2}$$

[17] https://rvprasad.medium.com/informedness-and-markedness-20e3f54d63bc

α represents the “cost" or “penalty" of a false match (FM), quantified as a weight assigned to it. Similarly, β denotes the “cost" or “penalty" of a false positive (FP), while 1.0 is the cost of a false negative (FN). For zbMATH Open, false matches (FM) or false positives (FP) are considered worse than false negatives (i.e., missing matches). Users typically recognize a false match more readily than a missing match, which is why a greater penalty is assigned to FM compared to FN. The closer the value of *informedness* is to 1, the better is the performance.

Figure 7 plots *informedness* against the minimum score (0.5-1) for both datasets under $\text{inf}_{1,1}$, $\text{inf}_{2,2}$, and $\text{inf}_{5,5}$. The DOI vs. DLMF contrast in Fig. 7 documents the performance of the deployed Random Forest classifier across these two bibliographically distinct datasets. As expected, increasing α and β lowers *informedness* by amplifying the penalty for false matches and false positives. The DOI dataset attains *informedness* above 0.8 throughout, while the DLMF dataset yields lower values, consistent with Deb et al. (2024).

As documented in Section "Classifier selection and comparison", this paper reports the Random Forest deployment; Logistic Regression’s higher Informedness (Table 1) marks it as the prioritized classifier for future re-evaluation. The elevated error rates on the DLMF dataset are attributable to degraded title and author metadata in older records rather than absent DOI signals, a finding consistent with the near-zero importance of the `doi` feature (0.003) in Table 2, and motivates the domain-adaptation argument of this paper.

Hard-negative case analysis

To characterize Marebito’s failure modes beyond the DLMF aggregate statistics, we tested twelve hard-negative cases across four structural categories: edition/translation pairs, JFM dual-identifier entries, title ambiguity under varying metadata completeness, and book-form citations from the same author cluster, as presented in Table 3. Of twelve case-input combinations, eight yielded true positives, two false matches, and two false negatives. The most diagnostically fragile category was JFM dual-identifier entries, where score margins between competing records fell below 0.001, rendering classifier decisions effectively arbitrary. False negatives arose exclusively from retrieval-stage failures, incomplete pagination and absent journal metadata in book-series citations, rather than from classifier misranking, as illustrated in Section "Performance comparison on DOI and DLMF data". Input string formatting further conditioned outcomes: edition descriptors and accent normalization each shifted results between TP and FM within the same underlying case.

Preliminary comparison with the crossref citation matcher

This section presents a focused, preliminary comparison between Marebito and Crossref’s citation matcher on a manually verified subset of the DLMF dataset. The comparison is intentionally scoped to the DLMF subset for two principled reasons. First, Crossref’s matcher accepts reference strings or structured bibliographic metadata and returns DOIs as output; evaluating it, therefore, requires ground-truth DOIs that have been independently verified. The manually checked DLMF subset provides exactly this, whereas the DOI dataset, despite containing DOI-indexed records, was not subject to the same entry-level manual verification and does not support a controlled head-to-head comparison on equal terms. Second, and more importantly, the scientifically relevant question this work addresses, how well a general-purpose matcher handles older mathematical literature with incomplete or absent DOI coverage, is present in the DLMF dataset but absent from the DOI dataset,

Table 3 Representative failure cases from the retrieval and classification pipeline

Category	Representative Case	Out.	Failure mode description
Edition/translation pair	Courant & Hilbert 1931 (edition descriptor present)	FM	Semantically accurate input string misleads edition disambiguation
JFM dual-identifier	Bouligand 1939 (raw string)	FM	JFM and Zbl records indistinguishable at classifier confidence threshold
Title ambiguation / incomplete metadata	Weiss 2026	FN	Pre-publication pagination absent; Elasticsearch retrieval fails entirely
Book-form / same-author cluster	Atiyah 1989	FN	Series descriptor yields no journal or page signals for candidate generation

(a)

key	correct_doi	correct_zbl
Abernathy:1993:APC	10.1145/168173.168405	0894.65080
Ablowitz:1991:SNE	10.1017/cbo9780511623998	0762.35001
Ablowitz:1981:IST	10.1137/1.9781611970883	0472.35002
Abramov:1996:UPU		-1
Salem:2013:SPE	10.2989/16073606.2013.779946	1274.41061
Abramov:1975:RCS	10.1016/0041-5553(75)90181-0	0326.65069
Adams:1997:SMD		-1
Abramov:1989:RSL	10.1016/s0041-5553(89)80002-3	0719.65063

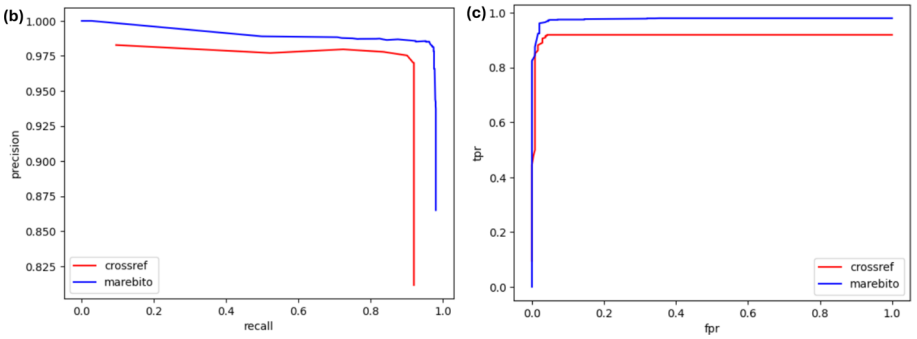


Fig. 8 **a** Excerpt from the modified DLMF dataset for the comparison study, **b** comparison of the Precision-Recall curve between Crossref and Marebito, and (c) comparison of the ROC curve between Crossref and Marebito

which by construction contains only records with confirmed DOIs. Evaluating Crossref on the DOI dataset would reduce the comparison to a setting where both matchers have access to reliable DOI signals, making it uninformative for the domain-adaptation question that motivates this work.

While working on the DLMF dataset, we stumbled upon some phenomena that we would like to present elaborately. There were missing DOIs for many entries in the DLMF dataset. There were around 700 entries with missing DOIs, but when we ran them through the Crossref API, it returned the correct DOIs. If an element in our input data has no DOI, we cannot be sure whether it was just missing or the entry has no DOI. So, we manually checked the entries in the DLMF dataset and added the missing DOIs.

For the Crossref matcher, the input is a JSON file generated from the BibTeX file. The result obtained was a list of BibTeX entries. The algorithm assigns a score to the result, 0 being the lowest score when no match is found in the Crossref database, and the corresponding 'DOI' field shows 'None'. When the algorithm can find a match in the Crossref database, it assigns a score close to 1.

To conduct a fair comparison study between the citation-matching interfaces of Crossref and Marebito, we have selected DLMF entries that are either present in both databases or absent from both: Crossref and zbMATH Open. So, we have considered all entries with the following characteristics:

– The entry has a correct DOI **AND** a Zbl number
– The entry has no correct DOI **AND** it has no Zbl number

2,037 items were identified in this category, consisting of 1,797 RP items and 240 RN items. An excerpt from this modified DLMF dataset is displayed in Fig. 8a. We would like to note that we encountered some unusual cases during this study.

- We have come across some entries where one entry leads to multiple articles. As this did not return the exact article, we considered these entries to be real negative for our analysis.
- Some DOIs obtained from the Crossref matching tool did not return the articles themselves but rather reviews of the articles. We considered them to be a real negative for our analysis.
- Some DOIs pointed to the proceedings and not the correct article that falls under the proceedings. We considered these entries to be real negatives as well.

On this manually verified subset of 2,037 DLMF entries, comprising older mathematical literature with incomplete DOI coverage, precisely the domain condition where Marebito's domain-specific feature engineering is most directly relevant, Marebito performs comparably to and marginally better than the Crossref matcher,[18] as shown in Fig. 8b and c. This result should not be interpreted as a claim of general superiority over Crossref; the subset is narrow and the comparison is deliberately scoped. Rather, it illustrates that a domain-adapted feature set, designed for the structural properties of mathematical reference strings, including inconsistent journal abbreviations and formula-rich titles, provides a meaningful advantage precisely in the conditions where general-purpose matchers are expected to struggle: older literature with sparse and inconsistent metadata. We acknowledge this as a preliminary finding and note that a broader comparative evaluation remains a direction for future work.

Comparison with neural baseline

As a diagnostic reference experiment, rather than a competitive neural benchmark, we implemented a lightweight neural reranking baseline using a frozen cross-encoder (MiniLM-L-6-v2, 22 M parameters)[19] as an additional scoring feature alongside four auxiliary signals: exact DOI match, exact year match, and token-level Jaccard similarity on titles and author surnames. Candidates were retrieved via the zbMATH Open REST API[20] rather than Elasticsearch, yielding an average of approximately 1.5 candidates per reference, far fewer hard negatives per reference than Marebito's Elasticsearch-based retrieval.

[18] https://gitlab.com/crossref/search_based_reference_matcher/-/tree/main
[19] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2.
[20] https://api.zbmath.org/v1/.

The zbMATH Elasticsearch index is an internal production system at FIZ Karlsruhe and is not externally accessible; the neural baseline, therefore, relies on the publicly available zbMATH Open REST API, which returns substantially fewer and less diverse candidates per reference, resulting in a sparse hard-negative training regime. A Random Forest classifier was trained on 105,899 labeled pairs derived from 100,000 references; no fine-tuning of the cross-encoder was performed. On both evaluation and test splits (K=5, threshold=0.5), the neural baseline achieves near-perfect ranking: Precision=0.791, Recall=0.999, with zero false matches (FM=0) and Hits@1 $\approx$ 0.9996. However, *informedness*, $\text{inf}_{1,1}$ collapses to approximately -0.0005 across both splits, driven entirely by FP/RN = 1.0: the classifier matches nearly every artificial-negative reference to some candidate, failing to learn abstention. The absence of false matches alongside a near-total false positive rate reflects the sparse hard-negative training regime: with, on average, fewer than two candidates per reference, the model learned to rank confidently when candidates compete, but never learned to reject when no correct match exists. This result supports two observations. First, $\text{inf}_{1,1}$ correctly identifies a systemically broken classifier that Precision and Recall alone, computed without accounting for false positives, would misclassify as near-perfect. Second, under this constrained retrieval regime, Elasticsearch-derived hard negatives are essential for abstention: without them, a classifier learns to rank but cannot learn to reject. Because the cross-encoder is frozen and the candidate pool is generated by the weaker REST-API retrieval, these observations concern hard-negative construction and abstention rather than the merits of neural representation learning (see Table 4).

Application of the citation matching interface: *h*-index as a downstream bibliometric use case

This section presents a downstream bibliometric application of Marebito's matched citation data to illustrate that its output is of sufficient operational quality to support a substantive bibliometric task: computing and analyzing *h*-index distributions for zbMATH Open authors. This application is motivated by the observation, documented in the zbMATH Open citation analysis literature (Krattenthaler, 2021), that citation-count distributions in mathematics have properties distinct from those of other scientific disciplines, making the quality of the underlying citation-matching infrastructure directly consequential for any derived bibliometric indicator.

Table 4 The neural reranking baseline (frozen MiniLM cross-encoder combined with a Random Forest classifier) on evaluation and test splits. The neural baseline is included as a diagnostic reference point illustrating that, under a constrained REST-API retrieval regime with sparse hard negatives, a frozen cross-encoder fails to learn abstention. It is not a competitive neural benchmark and does not support conclusions comparing neural and classical representation learning

System	Split	Precision	Recall	Accuracy	$\text{inf}_{1,1}$
Neural Baseline (MiniLM + RF)	Eval	0.7908	0.9992	79.03%	− 0.0004
Neural Baseline (MiniLM + RF)	Test	0.7924	0.9992	79.18%	− 0.0006

Neural baseline test results are from the held-out test split of the zbMATH dataset used throughout Section "Methods". These splits are not directly comparable, and no cross-system ranking is implied

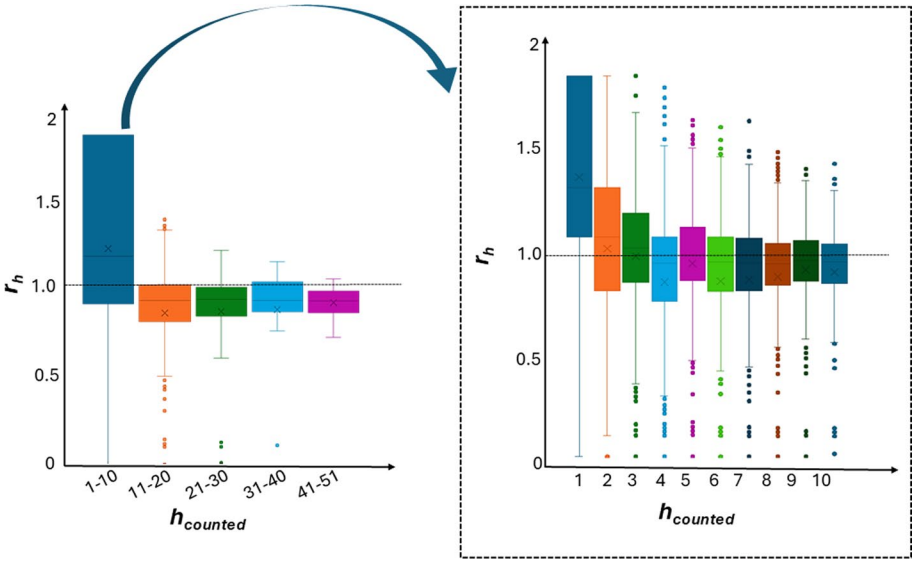


Fig. 9 Distribution of the ratio r_h for zbMATH Open authors, illustrating the relationship between observed and predicted research impact for zbMATH Open authors. The dashed horizontal line marks $r_h = 1$, i.e., exact agreement between $h_{counted}$ and $h_{predicted}$; values above indicate $h_{\text{counted}} > h_{\text{predicted}}$ and values below the reverse

Table 5 r_h vs. number of authors for the most closely matched authors at zbMATH Open FAST TRACK journals

r_h	**Number of Authors**
1.0...	38 970
1.00...	761
1.000...	34
1.0000...	3

The h-index was introduced by Hirsch in 2005 (Hirsch, 2005) as a metric for evaluating a researcher's impact: a researcher has h-index h if they have h papers each cited at least h times. Hirsch also proposed the relation

$$h = c\sqrt{N_{Zit}}$$

where N_{Zit} is the author's total citation count and $c \in [0.44, 0.57]$ varies by Hirsch (2005).

A subsequent study (Krattenthaler, 2021) argued that the h-index is largely redundant with total citation count, and showed that for large citation counts c converges to the constant 0.54 independently of the individual citation distribution. We adopt this asymptotic value as an approximation for the zbMATH Open author population. As the analysis below shows, the approximation holds well for authors with higher h_{counted} values but degrades for $h_{\text{counted}} \leq 10$, a finding of bibliometric interest for a discipline with characteristically lower citation volumes than the natural sciences.

Building on Krattenthaler (2021), we compute the h-index for zbMATH Open authors in two ways. The value derived from total citations via Hirsch's formula is termed $h_{\text{predicted}}$; the directly counted value is termed h_{counted}. A comparative analysis of these two quantities tests whether the concerns raised by Krattenthaler (2021) can be substantiated in the zbMATH Open context, with citation matching as the ground tool for identifying their divergence.

To quantify agreement between the two measures, we define the ratio

$$r_h = h_{\text{counted}}/h_{\text{predicted}}$$

and plot its distribution against h_{counted} in Fig. 9.

Our sample consists of 294,909 authors indexed in zbMATH Open's FAST TRACK journals, the top-tier journals selected by subject editors and periodically updated to reflect changes in the publication landscape. This subset was chosen as a high-quality, manipulation-resistant data source.

Figure 9 shows that prediction accuracy increases with h_{counted}: r_h approaches 1 as h_{counted} grows. For authors with $h_{\text{counted}} \in [1, 10]$, the upper whisker is absent (maximum equals the third quartile), and the high number of outliers in the ranges $h_{\text{counted}} \in [1, 10]$ and $[11, 20]$ confirms that prediction accuracy is substantially lower at low h-index values. A value $r_h > 1$ indicates $h_{\text{counted}} > h_{\text{predicted}}$; $r_h < 1$ indicates the reverse.

Among the 294,909 authors, 39,768 show the closest agreement between h_{counted} and $h_{\text{predicted}}$. Table 5 breaks down the r_h values for this subset.

Of the 39,768 closely matched authors, 3 have $r_h = 1$ to four decimal places, 34 to three, 761 to two, and 38,970 to one decimal place. Across the full sample, 88,927 authors have $r_h < 1$ (predicted exceeds counted); 116,900 have $r_h \in [1.0, 1.49)$; and 89,082 have $r_h \in [1.5, 1.89)$. Authors with $r_h > 1$ are dominated by those with $h_{\text{counted}} = 1$.

Discussion and conclusion

This paper documents the Random Forest deployment of Marebito; re-evaluating the full pipeline with Logistic Regression, which achieves higher *Informedness* on the reduced-subset comparison in Table 1, is a concrete and prioritized direction for future work. The comparison of reduced subsets in Table 1 was conducted for computational feasibility to allow the inclusion of the neural baseline (Section "Comparison with neural baseline") under CPU-only constraints. A plausible contributor to the ordering is the dataset scale: which of the two classifiers performs better depends on the characteristics of the dataset rather than being fixed (Couronné et al., 2018), and logistic regression, in particular, tends to achieve stable performance at smaller sample sizes than more complex algorithms (Silvey & Liu, 2024). All performance analyses in this paper - precision-recall curves, ROC curves, *Informedness* plots, and the Crossref comparison, use the Random Forest classifier as deployed.

One of the present drawbacks of our methodology is that we cannot test whether Marebito also correctly identifies references that are not indexed in zbMATH Open. Currently, it only matches references for items indexed in zbMATH Open. Another noteworthy aspect for researchers is that training Marebito takes time, and labeled data is required, as it is a supervised learning process. Furthermore, there remains a risk of overfitting, and the resulting classifier might depend on the actual training data. One has to take into consideration the possibility of biases in the training data or of the data not being representative

of the whole dataset. The feature importance analysis relies on mean decrease in impurity, which may favour high-cardinality features; a permutation-based study and a formal feature ablation experiment on the full dataset remain open directions. The Crossref comparison is scoped to 2,037 manually verified DLMF entries, and no external *h*-index benchmark from independent citation-matching data was available, both of which limit the reach of the comparative claims made in this work.

One of the most important components of zbMATH Open is its author disambiguation framework, and it is closely related to citation matching. The citation matching interface at zbMATH Open is integrated with other zbMATH Open APIs to further develop the author disambiguation framework, an ongoing project at zbMATH Open. As mentioned in Hulek and Teschke (2023), the outcome of good citation matching is good citation analysis. The exhaustive analysis of the citation data presented in the report (Hulek & Teschke, 2023) is a valuable addition to the field of citation analysis in the mathematics domain, and the data utilized in this analysis (Hulek & Teschke, 2023) was extracted using a previous version of the citation matching tool at zbMATH Open. Marebito has improved our citation-matching data as compared to the previous version of the citation-matching tool used at zbMATH Open. Systematic hyperparameter optimization for the Random Forest classifier, re-evaluation of the full pipeline with Logistic Regression as the operational classifier, and broader comparisons with large language model-based matchers are prioritized directions for future development.

While Marebito was developed and evaluated specifically for mathematical literature, its core architectural design principles merit discussion for broader applicability across scientific domains. The two-stage pipeline: candidate retrieval via indexed field-based search followed by learned ranking via string similarity features, reflects a general strategy applicable beyond mathematics. Recent work on multi-domain citation matching demonstrates that string similarity comparison functions (Jaro-Winkler, Jaccard, Levenshtein) form the foundation of citation matchers across diverse fields (Ghavimi et al., 2019), suggesting that Marebito's feature-engineering approach is generalizable. However, successful adaptation to other domains (e.g., chemistry, biology, physics) would require (1) domain-specific preprocessing and feature engineering tailored to field-characteristic citation formats. For instance, chemical literature requires different string similarity thresholds and weights than general literature (Yamaguchi et al., 2011), (2) retraining on representative labeled data from the target domain, and (3) field-specific evaluation on manually curated gold standards comparable to our DLMF subset. The modularity of Marebito's design: distinct parsing, retrieval, and classification stages support such future domain adaptations, though the relative importance of individual design choices (e.g., hard negative sampling via Elasticsearch, specific string similarity metrics) would need re-evaluation for each field.

Moreover, the citation matching the interface is integrated with the Links API. Links to external platforms such as DLMF and OEIS are already available via the Links API at zbMATH Open. One of the main objectives of the Links API is to integrate external partners, such as arXiv and MathOverflow, into the new zbMATH Open Links API. With the advancement of the Links API, Marebito will incorporate citation matching for articles on external platforms, and this is one of the major research goals to be fulfilled in the future. Additionally, author collaboration identification is a crucial future research area to venture into, another application area of the citation-matching tool. The construction of collaboration graphs can help us assess the effectiveness of researcher collaborations. This comprehensive study of Marebito's architecture and methodology can lead to advancements in the field of bibliometrics and help researchers gain in-depth knowledge of citation-matching interfaces in digital libraries.

Acknowledgements We sincerely appreciate the valuable insights provided by Fabian Müller and Christine Rimmert on this topic. We also extend our gratitude to Liteshwar Rao for his assistance with the Crossref algorithm and to Michael Jost for his support in conducting the *h*-index study.

Author contributions Conceptualization: Isabel Beckenbach, Madhurima Deb, Moritz Schubotz; Methodology: Isabel Beckenbach; Formal analysis and investigation: Madhurima Deb; Writing - original draft preparation: Madhurima Deb; Writing - review and editing: Isabel Beckenbach, Norman Meuschke, Moritz Schubotz, Olaf Teschke; Supervision: Moritz Schubotz.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was supported by Deutsche Forschungsgemeinschaft (DFG) under Grant Agreement No.460135501, NFDI, 29/1 "MaRDI - Mathematische Forschungsdateninitiative".

Data availability The datasets generated and analyzed in the current study are not publicly available because they contain internal reference data subject to institutional restrictions. The authors' own derived datasets are available from the corresponding author upon reasonable request, subject to permission from the relevant data holders.

Code availability The analysis relies in part on a proprietary internal system whose source code cannot be made publicly available due to institutional restrictions. However, the system's REST API is publicly accessible and can be used to replicate the core functionality described in this study. The code used for training, evaluation, and testing of the MiniLM-L6-V2 neural baseline is available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declared no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adler, R., Ewing, J., & Taylor, P. (2009). Citation statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). *Statistical Science, 24*(1), 1–14. https://doi.org/10.1214/09-STS285

Arnold, D. N., & Fowler, K. K. (2011). Nefarious numbers. *Notices of the American Mathematical Society, 58*(3), 434–437.

Bannister, A., Hulek, K., & Teschke, O. (2017). Citation practice in mathematical papers. Some remarks. *The Deutsche Mathematiker-Vereinigung (DMV), 25*(4), 208–214. https://doi.org/10.1515/dmvm-2017-0064

Councill, I. G., Li, H., Zhuang, Z., Debnath, S., Bolelli, L., Lee, W. C. & Giles, C. L. (2006). Learning metadata from the evidence in an on-line citation matching scheme. *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (pp. 276–285). https://doi.org/10.1145/1141753.1141817

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

Deb, M., Beckenbach, I., Petrera, M., Ehsani, D., Fuhrmann, M., Hao, Y. & Schubotz, M. (2024). An overview of zbmath open digital library. *Proceedings of the 24th ACM/IEEE joint conference on digital libraries* (pp. 1–5). https://doi.org/10.1145/3677389.3702597

Dommermuth, E. (2017). Mathscinet. *The Charleston Advisor, 18*(3), 29–33. https://doi.org/10.5260/chara.18.3.29

Ehsani, D., Petrera, M., & Teschke, O. (2023). The integration of OEIS links in zbmath open. *European Mathematical Society Magazine, 128*, 60–64. https://doi.org/10.4171/mag/132

Färber, M., Thiemann, A., & Jatowt, A. (2018). A high-quality gold standard for citation-based tasks. *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*. Retrieved from https://aclanthology.org/L18-1296/

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

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Ghavimi, B., Otto, W., & Mayr, P. (2019). Exmatcher: Combining features based on reference strings and segments to enhance citation matching. https://doi.org/10.48550/arXiv.1906.04484, *arXiv preprint* arXiv:1906.04484

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). Citeseer: An automatic citation indexing system. *Proceedings of the third ACM conference on digital libraries* (pp. 89–98). https://doi.org/10.1145/276675.276685

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hulek, K., & Teschke, O. (2023). How do mathematicians publish?-some trends. *European Mathematical Society Magazine, 129*, 36–41. https://doi.org/10.4171/mag/160

Hulek, K., & Teschke, O. (2024). zbMATH Open as a tool for bibliographical studies. *90 years of zbmath* (pp. 75–91). Berlin: European Mathematical Society (EMS). https://doi.org/10.4171/ZBL90/17

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: A case of dblp. *Scientometrics, 116*(3), 1867–1886. https://doi.org/10.1007/s11192-018-2824-5

Krattenthaler, C. (2021). Was der h-index wirklich aussagt. *Mitteilungen der Deutschen Mathematiker-Vereinigung, 29*(3), 124–128. https://doi.org/10.1515/dmvm-2021-0050

Movahedi, F., Padman, R., & Antaki, J. F. (2023). Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores. *The Journal of Thoracic and Cardiovascular Surgery, 165*(4), 1433–1442. https://doi.org/10.1016/j.jtcvs.2021.07.041

Müller, F., & Teschke, O. (2016). Progress of self-archiving within the dml corpus, with a view toward community dynamics. *Intelligent computer mathematics: 9th international conference, cicm 2016, bialystok, poland, july 25-29, 2016, proceedings 9* (pp. 63–74). https://doi.org/10.1007/978-3-319-42547-4_5

Parthasarathy, G., & Tomar, D. (2015). Trends in citation analysis. *Intelligent Computing, Communication and Devices: Proceedings of ICCD 2014*, Vol. 1, pp. 813–821. https://doi.org/10.1007/978-81-322-2012-1_88

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830. https://doi.org/10.5555/1953048.2078195

Petrera, M., Trautwein, D., Beckenbach, I., Ehsani, D., Müller, F., Teschke, O. & Schubotz, M. (2021). zbmath open: Api solutions and research challenges. https://doi.org/10.4171/zbl90/16, *arXiv preprint* arXiv:2106.04664.

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. https://doi.org/10.48550/arXiv.2010.16061, *arXiv preprint* arXiv:2010.16061

Schubotz, M., & Teschke, O. (2021). zbmath open: Towards standardized machine interfaces to expose bibliographic metadata. *European Mathematical Society Magazine, 119*, 50–53. https://doi.org/10.4171/mag/12

Silvey, S., & Liu, J. (2024). Sample size requirements for popular classification algorithms in tabular clinical data: Empirical study. *Journal of Medical Internet Research, 26*, Article e60231. https://doi.org/10.2196/60231

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter (Kongelige Danske Videnskabernes Selskab), 5*, 1–34.

Stahlschmidt, S., & Stephen, D. (2022). From indexation policies through citation networks to normalized citation impacts: Web of science, scopus, and dimensions as varying resonance chambers. *Scientometrics, 127*(5), 2413–2431. https://doi.org/10.1007/s11192-022-04309-6

Tomaszewski, R. (2021). A study of citations to stem databases: Acm digital library, engineering village, IEEE Xplore, and MathSciNet. *Scientometrics, 126*(2), 1797–1811. https://doi.org/10.1007/s11192-020-03795-w

Turki, H., Hadj Taieb, M. A., Ben Aouicha, M., Fraumann, G., Hauschke, C., & Heller, L. (2021). Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata. *Frontiers in Research Metrics and Analytics, 6*, Article 694307. https://doi.org/10.3389/frma.2021.694307

Wilkinson, D., Sud, P., & Thelwall, M. (2014). Substance without citation: Evaluating the online impact of grey literature. *Scientometrics, 98*, 797–806. https://doi.org/10.1007/s11192-013-1068-7

Yamaguchi, A., Yamamoto, Y., Kim, J-D., Takagi, T., & Yonezawa, A. (2011). Discriminative application of string similarity methods to chemical and non-chemical names for biomedical abbreviation clustering. *2011 IEEE international conference on bioinformatics and biomedicine* (pp. 544–549). https://doi.org/10.1109/BIBM.2011.98

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.