Rapport de stage de 4ème année
présenté par

Noah Roussel
Filière MAM
Année 2025 -2026

A Quantitative History of Pharmacology Research

A Case Study of the British Journal of Pharmacology

Tuteur de l'entreprise : Roland MOLONTAY + Csaba KISS
Tuteur de l'école : Sarah DELCOURTE

Janvier 2026

POLYTECH Lyon - UNIVERSITE CLAUDE BERNARD LYON 1
Domaine Scientifique de La Doua – 15, Boulevard Latarjet
69622 VILLEURBANNE CEDEX
Tél. (33) 04.72.43.12.24 - Fax. (33) 04.72.43.12.25
http://polytech.univ-lyon1.fr/

Confidentiality

Acknowledgements

I would like to thank Dr. Roland Molontay and all the HSDSLab for their great welcome and the opportunity, particularly Csaba Kiss for his precious support and advice. I would also like to thank Gabriel Vignon for his work, which served as a source of inspiration throughout this project.

Contents

Abstract **2**

1 Introduction **3**

2 Company Presentation **5**

3 Description of the Work **6**

3.1 Interactions within the Organization 6
3.2 Journal Selection and Data 6
3.3 Descriptive Statistics 8
3.4 Collaboration Networks 10
3.5 Topic Modelling 11
3.6 Statistical Analysis 12
3.7 Modelling 12
3.8 Precisions 14

4 Results **15**

4.1 General Metrics 15
4.2 Authors, Countries, Institutions 21
4.3 Topics 28
4.4 Feature Importance 32

Conclusion **38**

Appendices **39**

Bibliography **66**

Abstract

Abstract:

This study presents a data-driven analysis of the British Journal of Pharmacology (BJP) within Q1 pharmacology journals, using OpenAlex data on nearly 690,000 publications between 1909 and 2024, including over 26,000 BJP articles. Given the limited availability of records before 1950, the study focuses on publications from 1950 onward. The analysis combines bibliometric indicators, network analysis, topic modeling, and machine-learning approaches to examine trends in different metrics.

General impact metrics show that BJP ranks among the leading Q1 journals in terms of publication volume and total citations, with citations fairly well spread across articles. Over time, international collaboration has grown, as well as the number of authors, references, countries, and institutions per paper. Looking at the most productive authors, countries, and institutions also helps highlight who contributes most to the journal.

Topic analyses, using both OpenAlex categories and BERTopic modeling, show that citation patterns vary across research areas, with some less common topics receiving relatively high impact. Finally, statistical tests and machine-learning models indicate that the number of references, past institutional performance, and topic-related factors are the main drivers of citation impact, although the overall predictive power of the models remains limited.

Keywords: Pharmacology, Science of science, Network science, Research collaboration, Topic Modelling

Résumé:

Cette étude présente une analyse basée sur les données du British Journal of Pharmacology (BJP) au sein des revues de pharmacologie classées Q1, en s'appuyant sur les données d'OpenAlex couvrant près de 690 000 publications (1909–2024), dont plus de 26 000 articles du BJP. L'analyse se concentre sur la période postérieure à 1950, en raison de la disponibilité limitée des données pour les années antérieures. Elle combine des indicateurs bibliométriques, des analyses de réseaux, de la modélisation thématique et des méthodes d'apprentissage automatique afin d'examiner les tendances de différentes métriques.

Les indicateurs généraux d'impact montrent que le BJP figure parmi les principales revues Q1 en termes de volume de publications et de nombre total de citations, celles-ci étant relativement bien réparties entre les articles. Au fil du temps, on observe une augmentation de la collaboration internationale, ainsi que du nombre d'auteurs, de références, de pays et d'institutions par article. L'analyse des auteurs, pays et institutions les plus productifs permet également de mettre en évidence les principaux contributeurs à la revue.

Les analyses thématiques, fondées à la fois sur les catégories d'OpenAlex et sur la modélisation BERTopic, soulignent que les dynamiques de citation varient d'un domaine à l'autre, avec des thématiques moins fréquentes qui peuvent néanmoins présenter un impact élevé. Enfin, les tests statistiques et les modèles d'apprentissage automatique indiquent que le nombre de références, les performances institutionnelles passées et les facteurs liés aux thématiques constituent les principaux déterminants de l'impact en termes de citations, bien que le pouvoir prédictif global des modèles demeure limité.

Mots-clés : Pharmacologie, Science de la science, Science des réseaux, Collaboration scientifique, Modélisation thématique

Introduction

This report summarizes the research project conducted at the Human and Social Data Science Lab (HSDSLab)[1] of the Budapest University of Technology and Economics. The objective of this work is to provide a data-driven analysis of the British Journal of Pharmacology (BJP) and to position it within the broader ecosystem of Q1 pharmacology journals, using large-scale bibliographic data and methods from the Science of Science [1].

The project initially focuses on the British Journal of Pharmacology, a long-established and influential journal in the field of pharmacology, in celebration of its 80th anniversary since its founding in 1946. By examining its publication activity, citation patterns, collaboration structures, institutional and geographical diversity, and thematic specialization, the study aims to better understand the mechanisms underlying its scientific impact. To place these findings in context, the analysis systematically compares BJP with other leading Q1 pharmacology journals.

The study relies on publication-level data extracted from OpenAlex, covering nearly 690,000 scientific works published between 1909 and 2024, including over 26,000 articles from the BJP. Due to limited data availability before 1950, the longitudinal analysis focuses on the 1950–2024 period. In this project, a *work* refers to a scientific publication, primarily journal articles, as defined by the data source.

The notion of scientific impact is inherently ambiguous. In this report, *impact* is understood as the recognition received from the scientific community and is approximated using citation-based indicators, including normalized metrics. While such measures are imperfect and subject to well-known biases, they remain central tools in bibliometric research and are extensively discussed in the related literature. Similarly, the term *factor* refers to a measurable characteristic associated with publications, such as references, authorship, national/institutional background, collaboration patterns, or research topics, that may influence citation outcomes. As is common in observational studies, the analysis focuses on statistical associations rather than causal relationships, and potential confounding factors are acknowledged.

Methodologically, this study combines classical bibliometric indicators, network analysis (co-authorship, institutional and international collaborations), topic modelling approaches, and statistical / machine-learning models to explore citation dynamics. These complementary methods allow for both descriptive and explanatory insights into the evolution of BJP and its position relative to Q1 journals.

Several bibliometric studies have applied methods similar to those employed in this project to analyze publication dynamics and scientific impact within pharmacology and related fields. For instance, a study by Darnalis et al. (2025) examined the 100 most cited articles in pharmacokinetic and pharmacodynamic modelling, using Scopus data to map thematic trends, international collaborations, and methodological developments [3]. Similarly, Basol and Seifert (2024) conducted a longitudinal bibliometric analysis of *Naunyn–Schmiedeberg's Archives of Pharmacology*, highlighting thematic specialization through keyword co-occurrence networks and citation distributions [4]. Finally, Li and Chen (2021) analyzed trends in U.S. pharmacology publications, combining bibliometric indicators with institutional and geographical mapping to assess productivity and collaboration patterns [5].

Research problem. How has the scientific impact of the British Journal of Pharmacology evolved since 1950, and how does it compare to other Q1 pharmacology journals?

[1] According to [2], Science of Science is a transdisciplinary field that studies science itself using large-scale data and quantitative methods to understand its structure, dynamics, and evolution.

Contributions

This project provides the following contributions:

- Construction and curation of a large-scale bibliographic dataset, based on OpenAlex, covering nearly 690,000 pharmacology-related publications, including more than 26,000 articles from the British Journal of Pharmacology, with a longitudinal focus on the 1950–2024 period.
- Comprehensive bibliometric analysis of the British Journal of Pharmacology, examining publication trends, references, citation patterns, normalized impact metrics, and citation distribution, comparing these results with other Q1 pharmacology journals.
- Analysis at the level of authors, countries, and institutions, identifying major contributors, and their influence on publication output and scientific impact.
- Analysis of collaboration structures, using network-based approaches to study which entities collaborate with each other, with a focus on the most collaborative entities.
- Topic-level analysis of pharmacological research, combining OpenAlex classifications with data-driven topic modelling (BERTopic) to investigate thematic evolution and heterogeneous citation/publication dynamics across research areas.
- Evaluation of impact determinants through statistical tests and machine-learning models, assessing the relative influence of references, authorship, institutional performance, collaboration patterns, and topics on normalized citation impact.

GitHub Repository of the Project

Figure 1: British Journal of Pharmacology

Company Presentation

The Human and Social Data Science Lab (HSDSLab) at the Budapest University of Technology and Economics (BME), led by Dr. Roland Molontay, is an interdisciplinary group combining mathematics, data science, and societal applications. While based in the Institute of Mathematics, the lab collaborates closely with the Faculty of Economics and Social Sciences and other academic units, reflecting its focus on complex problems at the intersection of data, society, and policy.

The laboratory applies data-driven and network science methods to tackle challenges in social, behavioral, and health-related domains. In particular, its research in network science focuses on structural analysis and modeling of social and co-authorship networks, while in educational data science the lab develops tools to predict student outcomes, evaluate interventions, and design explainable AI solutions for higher education. At the same time, in health data analytics, HSDSLab works on predictive modeling and decision support based on clinical datasets. Furthermore, the lab investigates anomaly detection, addressing high-dimensional and high-frequency data to identify unusual patterns or system failures.

The lab also offers consulting and collaborative services. It supports academic and industry partners in formulating data and network science problems, designing experiments, selecting appropriate methods, and interpreting results. HSDSLab engages in joint research and development projects to create tailored analytic models and innovative solutions, and provides education and training through undergraduate and graduate courses, specialized workshops, and off-site trainings in statistical modeling, machine learning, and network science.

HSDSLab's contributions are recognized internationally through publications in high-impact journals, presentations at conferences, and participation in collaborative projects. Its work combines new methods with practical applications, promoting collaboration across fields and creating impact in both research and real-world problems.

Finally, The lab is organized as a collaborative team, bringing together researchers, PhD students, and postdoctoral fellows with diverse backgrounds in mathematics, data science, and social sciences. It regularly hosts international interns and visiting students, including participants from France and other countries, providing them with hands-on experience in research projects and applied data science.

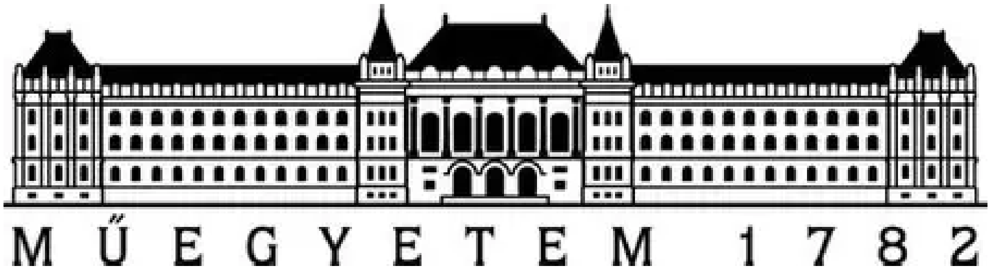


Figure 2: Budapest University of Technology and Economics

Description of the Work

3.1 Interactions within the Organization

I joined the laboratory as an intern for a total of 20 weeks, supervised by dedicated mentors. My primary supervisor for this project, Csaba Kiss, provided regular guidance through weekly meetings during which we reviewed my progress, proposed adjustments, and defined the next steps. I also had some meetings with the head of the laboratory, Roland Molontay, who oversaw the project from a broader scientific perspective.

The working schedule was flexible: I managed my time autonomously, maintaining an average workload of approximately 40 hours per week, combining on-site and remote work. This flexibility allowed me to work efficiently while adjusting my schedule according to computational demands (which were particularly significant) and the different stages of the project.

I worked in a pleasant professional environment that was conducive to concentration. I had a personal office which I shared with another international intern, creating a collaborative and supportive environment. Over the course of the internship, the only noticeable changes mainly concerned the working environment. The office gradually became more occupied than at the beginning, and it was not uncommon for other people to use the space occasionally, particularly professionals from same/other universities or research centers involved in inter-laboratory collaborations. I also had the opportunity to attend several scientific presentations.

3.2 Journal Selection and Data

3.2.1 Dataset Organisation and Building

The first step of the project involved constructing two distinct datasets: the BJP dataset and the Q1 dataset, each corresponding to a different subset of publications. Due to its smaller size, the BJP dataset could be processed with certain manipulation techniques that were impractical for the larger Q1 dataset. Consequently, although the analysis approaches are broadly similar, data preparation and handling differ between the two cases.

For the Q1 dataset in particular, batching [2] was necessary after extraction to prevent memory saturation. Additionally, Parquet files were used instead of CSV files and, more importantly, the Polars library instead of Pandas.

For each selected Q1 journal, publication-level data were collected programmatically using the OpenAlex API. The list of 81 Q1 pharmacology journals was obtained from the SCImago Journal Rank website[6], which is powered by Scopus. Journal names were cleaned and normalized to ensure consistent identification of journals when queried via the OpenAlex API.

3.2.2 Data Extraction

For each journal and both datasets, the data extraction process retrieves all available works associated with the journal identifier, without imposing an upper limit on the number of records. For each publication, the following information was collected:

- Bibliographic metadata (title, publication year, journal)

[2] Splitting the dataset into several chunks

- Citation indicators (total citations, yearly citation counts, normalized citation percentiles)
- Authorship information (author identifiers and positions)
- Institutional and country affiliations
- Reference counts and referenced works
- Topic-related information (primary topic, field, subfield, domain)
- Textual content (abstracts and keywords, when available)

This process resulted in two structured datasets with rich, publication-level granularity.

3.2.3 Data Structuring and Preparation

After collecting bibliographic data from the OpenAlex database, the analysis builds on both the original indicators provided by the database and a set of additional derived metrics. These metrics are designed to enhance the characterization of publication activity, collaboration patterns, and citation impact:

- **H-index[3]:** calculated for journals and aggregated entities, capturing the combined effect of productivity and citation impact.
- **Gini coefficient of citations[4]:** computed at the journal level to quantify inequality in citation distributions.
- **Mean Normalized Citation Score (MNCS) [5]:** computed under two complementary normalization frameworks:
  - relative to all Q1 pharmacology journals.
  - relative only to publications from the BJP.

  This dual normalization enables fair comparisons both across journals and within BJP itself.
- **Publication-type indicators:** review[6] articles and meta-analyses[7] were identified using keyword-based detection in titles and abstracts, supporting the analysis of systematic citation differences across publication types.
- **Publication Age and Age-adjusted citation indicators:** computed at the journal level to characterize publication popularity while accounting for time effects, from two complementary perspectives:
  - Cumulative popularity, based on total citation counts (`cited_by_count`) normalized by publication age.
  - Citations per Year by Publication Age, examining citation patterns by relating the number of citations received since 2012 to the age of publications, independently of their original publication year.

[3] According to [7], H is defined as the number of papers that have received at least H citations each.

[4] The Gini coefficient is defined as the ratio of the area between the Lorenz curve of a distribution and the line of perfect equality to the total area under the line of perfect equality [8]. It ranges from 0 (perfect equality) to 1.

[5] MNCS is defined as $\mathrm{MNCS} = \frac{\mathrm{Ncitations}}{\mathrm{Ncitations}_{\mathrm{year,\ field}}}$, where $\mathrm{Ncitations}_{\mathrm{year,\ field}}$ is the mean number of citations received by works published the same year in the same field.

[6] A review article is an article that synthesizes and critically discusses existing research on a given topic without presenting new original empirical results.

[7] A meta-analysis is a quantitative study that statistically combines results from multiple independent empirical studies addressing the same research question.

3.3 Descriptive Statistics

3.3.1 General Metrics and Journal-Level Ranking

In addition to the variables constructed in the previous section, the following indicators were directly obtained from OpenAlex and analysed as baseline metrics in the general analysis:

- **Publication counts**, aggregated by year, journal, author, institution, or country.
- **Total citation counts**, capturing cumulative scientific impact.
- **Yearly citation trajectories (since 2012)**: time-resolved citation series recording the number of citations received by each journal on a yearly basis from 2012 onward. Unlike cumulative metrics, this indicator reflects current citation performance independently of the publication year of the cited articles.
- **Citation percentile**[8], indicating how a paper performs relative to its contemporaries (e.g., top 10% or top 1% most cited papers)
- **Referenced works count**, corresponding to the number of references cited by each publication and serving as an indicator of article depth.

These raw indicators form the empirical foundation of the study and allow for descriptive rankings and temporal trend analyses without additional transformation.

Journals were then ranked according to multiple criteria, including total publications, cumulative citations, mean citations per work, MNCS, citations received since 2012, and author/country/institution diversity. This multi-dimensional ranking framework enables the identification of journals performing strongly across these dimensions, with a specific focus on evaluating the ranking and comparative standing of the BJP.

Also, particular attention was given to highly cited publications, as a small number of extreme outliers can disproportionately influence aggregate metrics. By explicitly identifying and analyzing top-cited works, the analysis evaluates how journal-level indicators depend on a restricted set of high-impact articles. For visualization purposes, publication titles were shortened by retaining only the first and last words of each title, improving readability while preserving identifiability.

Finally, Time-based analyses incorporate information on journal longevity and historical publication records. For each journal, the year of first appearance in the dataset was identified to distinguish long-established journals from more recent entrants. This information is used to track the temporal evolution of active journals among the most popular ones, and to highlight potential limitations of the database.

3.3.2 Author, Country, and Institution Analysis

To analyze structural patterns in scientific production and collaboration, publication-level information on authors, countries, and institutions was transformed into entity-centric datasets. In each case, publication records were reshaped so that each row corresponds to a single entity–publication relationship. This unified representation enables the computation of productivity, citation impact, and temporal trends at the author, national, and institutional levels.

For authors, publication-level author lists were converted into an author-centric dataset, allowing the analysis of individual productivity, citation performance, and career dynamics over time. For selected subsets of

[8] In OpenAlex[9], the citation percentile of a work is defined as a quantile between 0 and 1 representing the rank of its citation count relative to other works of the same type, publication year, and subfield.

authors, such as the most prolific or most cited, additional metadata were retrieved by querying the OpenAlex Authors endpoint. This step enables the recovery of author display names and the distribution of affiliated countries[9] across publication histories, providing insights into international presence and mobility.

A similar reshaping procedure was applied to countries and institutions. For each publication, all associated countries and institutions were extracted and converted into long-format datasets. Countries were represented using their standard two-letter country codes (e.g., FR for France, US for the United States, HU for Hungary), ensuring consistent aggregation and facilitating international comparisons.

Some institution names were further simplified and harmonized to prevent potential identification issues and to improve visualization clarity, particularly in cases where formatting variations or embedded OpenAlex identifiers were present in the raw metadata.

Across all three entity types, the analysis focuses on entities ranked at the top in terms of publication volume and citation impact. This allows the identification of the most influential and most present authors, countries, and institutions[10], as well as the examination of mismatches between productivity and popularity. For example, entities that publish frequently but receive relatively fewer citations, or conversely, those with high citation impact despite lower publication counts. Temporal analyses further assess whether influential entities are long-established or recently emerging, highlighting shifts in the structure of scientific contributions over time.

This framework also enables the study of diversity patterns across journals, both within the broader set of Q1 pharmacology journals and within the British Journal of Pharmacology specifically. Diversity is examined in terms of the number and distribution of contributing authors, countries, and institutions. These diversity measures are then related to citation impact, using comparative plots to evaluate how article popularity correlates with international and institutional diversity, particularly at the country and institution levels.

Additional analyses were conducted for authors on gender-related patterns. Author gender was determined using a Gender Detector Python library[11], which predicts the likely gender of a first name by matching it against a first-name based reference database. Names are classified as male, female, or unknown. This approach provides a simple and widely used tool for large-scale gender analysis, while remaining inherently imperfect.

Prior to gender inference, author names represented only by initials were excluded from the prediction step, as gender cannot be inferred reliably in such cases. These authors were retained in the dataset but omitted from gender-based statistics. In addition, names of non-Western origin are more likely to be classified as unknown due to limitations in the underlying name database. These sources of uncertainty are explicitly acknowledged and considered when interpreting the results.

The analysis focused on three complementary author subsets: most cited authors, most active[12] authors, and a general overview[13].

3.3.3 Topic Analysis

Research topics were first analyzed using thematic information provided by OpenAlex, with a particular focus on concepts and keywords[14]. Concepts offer a broad representation of research areas, while keywords

[9] Affiliated countries refer to the countries of the institutions associated with the author's published works.

[10] For institutions, affiliations were further analyzed by institutional type (universities, private companies, and national or public research institutes) to assess which organizational categories are most represented and most impactful.

[11] The `gender-guesser` Python library [10] was used.

[12] Activity is defined as having the highest number of publications in a given year.

[13] For technical reasons, the general overview was restricted to the top 1,000 most cited authors per year, as retrieving complete author-level metadata for the full corpus would require an impractically large number of OpenAlex API calls.

[14] In OpenAlex[9], concepts are standardized representations of the main scientific topics of a work, while keywords are textual terms extracted from titles, abstracts, or author-provided metadata.

provide a more precise and fine-grained description of article content, making them especially informative for identifying specific scientific focuses.

Both concepts and keywords were studied using metrics analogous to those applied to authors and institutions. Topics were first ranked by total publications, and citation performance of the top 20 was analyzed using mean MNCS, highlighting that some highly published concepts or keywords are not necessarily the most cited or impactful.

In addition, co-occurrence[15] analyses were conducted for both concepts and keywords, revealing thematic associations and clusters of frequently co-studied research areas.

The Primary Topic variable provided by OpenAlex, which assigns a single dominant topic to each publication, was analyzed separately to enable clearer thematic attribution. Primary topics were examined in terms of most published and most cited topics, allowing direct comparisons between thematic prevalence and citation impact.

Temporal dynamics of the top topics were analyzed both in terms of frequency and cumulatively, and by time periods (years and decades) to capture long-term trends and shifts in research focus. The evolution of dominant topics over time was further illustrated using word cloud visualizations, which provide an intuitive summary of changing thematic landscapes.

3.4 Collaboration Networks

3.4.1 BJP Co-Authorship Network

The analysis of collaboration patterns through co-authorship networks was conducted on both the BJP and the broader set of Q1 pharmacology journals. For each case, co-authorship networks were constructed in which nodes represent authors and edges indicate co-authorship relationships, defined by at least one jointly authored publication.

Due to computational constraints associated with large-scale network visualization, graphical representations were produced only for the BJP co-authorship network, which remains of manageable size. These visualizations were generated using the *ForceAtlas2* layout algorithm in Gephi[16], as it is well suited for medium-sized networks and facilitates the identification of densely connected author groups.

While network visualizations can provide intuitive insights, they are inherently subjective and sensitive to layout choices and parameter settings. Consequently, the analysis primarily relies on standard network metrics to extract objective and reproducible information about collaboration structures, including measures of connectivity, centrality, clustering, and cohesion.

3.4.2 Top Collaboration Networks

Beyond co-authorship, collaboration networks were constructed using a unified methodological framework applicable to authors, institutions, and countries. After reshaping the bibliographic data into long format, all entities associated with a given publication were identified and deduplicated, ensuring that each entity appears only once per work.

For each publication, the resulting set of entities was used to generate pairwise co-occurrence relationships. These co-occurrences were aggregated across the corpus, producing weighted edges and nodes that reflect the intensity of collaboration, measured by the number of shared publications.

[15] Co-occurence : two entities are connected if they jointly appear in at least one publication.

[16] Gephi, an open-source software for network visualization and analysis - `https://gephi.org`

Due to computational and visualization constraints, collaboration networks were not constructed on the full set of entities. Instead, for each level of analysis, entities were ranked according to their total collaboration activity, and networks were built only on the top 25 most prolific entities, showing which entities collaborate most frequently. All collaboration edges involving these entities were retained, allowing meaningful comparison across networks.

Prior to network construction, entity identifiers were standardized to ensure consistent aggregation ; Institution names were simplified and harmonized, while authors and countries were primarily identified using OpenAlex identifiers. When necessary, additional metadata was used to resolve ambiguities.

3.5 Topic Modelling

3.5.1 General Presentation

Topic modelling was conducted using the **BERTopic** framework on the complete set of publications from Q1 pharmacology journals, rather than on BJP alone, in order to ensure greater thematic coherence and more stable topic definitions. Once the global topic structure was established, publications from the British Journal of Pharmacology were extracted by matching common bibliographic fields (notably titles and identifiers).

BERTopic combines modern transformer-based language models with clustering techniques to identify semantically coherent research topics. Textual data (titles and abstracts) are first tokenized and transformed into dense semantic representations using the pretrained *all-MiniLM-L6* transformer model. Document-level embeddings are obtained via sentence-level pooling using the Sentence-Transformers library, allowing contextual similarities between publications to be captured beyond simple word co-occurrence.

Because these embeddings lie in a high-dimensional space, dimensionality reduction is applied using *UMAP*, followed by clustering with *HDBSCAN*, a density-based algorithm particularly well suited for large and heterogeneous textual corpora. An important property of *HDBSCAN* is that, depending on chosen parameters, it does not force all documents into clusters: publications that do not clearly belong to any dense region of the embedding space remain unassigned, which explains the presence of works without an associated topic in the final dataset.

Once topics are identified, the model provides a ranked list of representative terms for each topic. Each term is associated with a score reflecting its relative importance within the topic, as derived from the *c-TF-IDF* weighting scheme. For interpretability and clarity, topic labels were constructed by selecting the top two highest-scoring words for each topic, resulting in concise and descriptive thematic names.

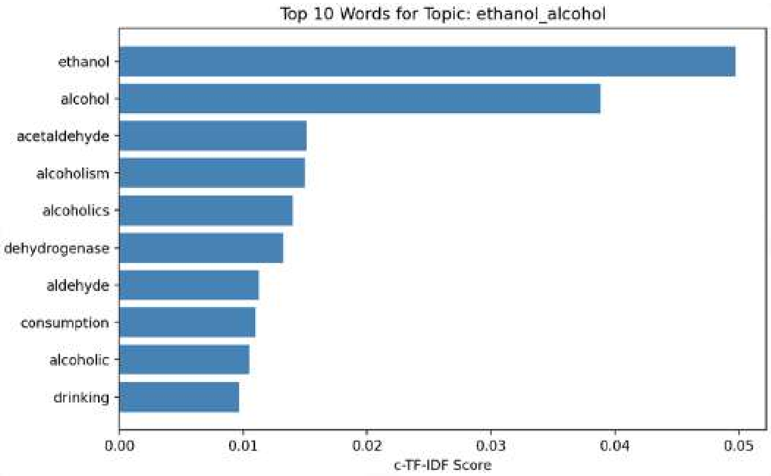


Figure 3: Example of the top 10 words for a modelled topic

Topic similarity was illustrated by visualizing topics in the reduced embedding space (see Appendix A). Topics that appear close to one another correspond to related research areas, reflecting the fact that research themes are not strictly separated. Consequently, due to the conceptual closeness of certain topics, some publications assigned to one topic may also be relevant to a closely related topic. For instance, topics covering morphine and anesthesia are closely related, making it reasonable for a work to be associated with both.

3.5.2 Output Overview

Compared to OpenAlex's topic system, which includes approximately 3,077 manually assigned and highly fragmented topics, BERTopic yields a more compact set of around 580 algorithmically inferred topics, within which publications are grouped in a more semantically meaningful way. This data-driven approach avoids manual labeling biases and enables a more coherent representation of the research landscape.

Due to the extraction and filtering process, a subset of publications could not be retained. Specifically, the number of BJP works included in the topic modelling analysis decreased from 26,115 to 22,308. More generally, not all publications could be assigned to a topic by the model: 348,193 Q1 publications remained without an assigned topic, reducing the Q1 corpus from 689,391 to 341,198 works, and 11,052 BJP publications were similarly excluded, leaving 11,256 BJP works distributed across 311 topics out of the 580 created.

The resulting BERTopic topics were subsequently analyzed in terms of publication volume, citation impact, and temporal evolution, both cumulatively and for recent periods (notably since 2012), offering an additional perspective on the structure of research themes and their evolution over time.

3.6 Statistical Analysis

To assess the relationship between bibliographic depth and citation impact, Spearman and Pearson correlation tests[17] were performed on the number of referenced works. Similar correlation analyses were also conducted on the number of authors, countries, and institutions in order to examine whether diversity in authorship, international collaboration, and institutional involvement is associated with higher citation impact.

3.7 Modelling

3.7.1 Motivation and General Principles

Scientific publications differ along many dimensions, including authorship, institutional affiliations, collaboration patterns, topics, and reference practices. All these characteristics may influence how often a publication is cited and, more generally, its scientific impact. However, the combined effect of these factors is difficult to assess using descriptive analyses alone, as many variables are correlated and interact in non-trivial ways.

Modelling approaches are therefore introduced to disentangle these effects and to identify which factors are most strongly associated with citation impact. In this study, modelling aims to explain variations in impact, measured through the MNCS, by leveraging publication-level characteristics.

To achieve this, machine-learning[18] (ML) models were trained on a large set of publications, using their observable features as inputs and MNCS-related outcomes as targets. Beyond predictive performance, these models allow the estimation of the relative contribution of each feature to the predicted impact, providing insight into which factors matter most[19].

[17] Spearman and Pearson correlation tests are statistical methods that measure the association between two variables, with Spearman evaluating monotonic relationships and Pearson evaluating linear relationships.

[18] A machine-learning model can be understood as a statistical tool that learns patterns from examples and establishes relationships between input variables (e.g., number of authors, references, institutions, topics) and an outcome variable.

[19] Note: several features were computed before applying the modelling step, see Appendix L

3.7.2 Modelling by Decade

To capture long-term dynamics and temporal heterogeneity, a decade-based modelling strategy was implemented. The dataset was divided into consecutive decades, and a separate model was fitted for each period. This approach is particularly well suited to identifying temporal trends and "hype" effects, especially at the topic level, which may vary substantially across historical periods. For each decade, a simple linear regression model[20] was estimated.

This framework allows direct interpretation of coefficients as decade-specific associations between publication characteristics and citation impact. By comparing significant[21] coefficient values across decades, it becomes possible to observe how the influence of certain features, particularly topics, changes over time, highlighting periods of increased or decreased relevance.

Model fit and explanatory power were evaluated using standard regression diagnostics, with particular emphasis on the % coefficient of determination[22] (R²) by decade. Comparing R² values across decades makes it possible to assess whether citation impact becomes more or less predictable over time, and whether the selected features capture an increasing or decreasing share of the mechanisms driving impact.

This decade-based modelling therefore goes beyond descriptive trends by explicitly quantifying how well different factors explain citation outcomes in each historical period, offering a clearer understanding of how the drivers of scientific impact evolve over time.

3.7.3 General Modelling and Machine-Learning Approaches

In addition to decade-specific linear models, a general modelling framework was applied to the full dataset using multiple ML algorithms[23]. The focus of this analysis is placed on variables that go beyond specific topic assignments, including authorship characteristics, institutional affiliations, collaboration intensity, and past performance indicators.

Rather than predicting raw MNCS values, the models were designed to predict MNCS quantile categories, thereby transforming impact prediction into an ordinal classification problem. Each publication was assigned to one of five classes according to its MNCS quantile:

- $[0, 0.25[$ → class 0
- $[0.25, 0.50[$ → class 1
- $[0.50, 0.70[$ → class 2
- $[0.70, 0.90[$ → class 3
- $[0.90, 1.00]$ → class 4

This formulation reduces sensitivity to extreme values and emphasizes the relative positioning of publications in terms of scientific impact.

The models were then trained on more than 550,000 publications and evaluated on a held-out test set of over 138,000 works. Several ML models were compared in terms of both predictive performance and

[20] A simple linear regression model expresses the dependent variable (e.g., MNCS) as a linear combination of explanatory variables, such that $\text{MNCS} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$, where $X_1, \ldots, X_n$ denote the input features, $\beta_0, \beta_1, \ldots, \beta_n$ are the corresponding regression coefficients, and ε represents the error term (which is not considered in the analysis).

[21] Coefficients with <0.05 p-value.

[22] % R² shows how much MNCS variation the model explains per decade.

[23] Those ML models will be detailed and explained in the 4.4 section.

feature importance. In addition, a collaboration matrix was incorporated to capture structured information about collaborative patterns among entities.

Model performance was measured using multiple complementary metrics:

- **Recall**, representing the fraction of true samples correctly identified (derived from the confusion matrix).
- **Accuracy**, measuring the overall proportion of correct predictions.
- **Precision**, computed per class to evaluate prediction quality across impact categories (visualized through scatter plots).
- **Root Mean Squared Error (RMSE)**, capturing the average distance between predicted and true classes.
- **Quadratic Weighted Kappa (QWK)**[24], which measures ordinal agreement between predictions and true labels[25].

3.8 Precisions

3.8.1 Results Presentation

The results are structured into thematic sections covering the main components of the analysis: general metrics, authors, countries and institutions, research topics, and modelling. Moreover, certain results were presented following this plan during a global presentation to Csaba Kiss and Roland Molantay.

3.8.2 Other Contributions

Short videos were produced to illustrate the evolution of top-cited authors over time, both cumulatively and over rolling five-year periods. The same approach was applied to primary topics and to the modelled topics. Moreover, Interactive plots were developed, notably using **BERTopic** for the topic modelling analysis, as well as more generally through the use of Plotly for result visualization.

These visualizations have not been et will not be presented because they are not static images and therefore cannot be directly included in a traditional report or presentation.

3.8.3 Difficulties

The main difficulty at the start was the large size of the dataset, with over 720,000 publications from Q1 journals plus the BJP, each containing extensive metadata. Processing these data was slow and sometimes caused scripts to crash, requiring code optimization and improved computation methods.

Most of these initial problems were then resolved. Remaining challenges were minor and focused on refinements, such as clearer visualizations or better choices, rather than obstacles that could significantly slow the project.

More generally, other difficulties encountered during the project were related to learning and applying new methods and software. This was particularly the case for the collaboration network analysis and, above all, topic modeling, which required using a GPU-compatible environment[26] that was not available on the computer.

[24] 1 indicates perfect agreement, 0 corresponds to random prediction, so values above 0 indicate performance better than chance.

[25] To resume : The higher the accuracy/QWK, the better. The lower the RMSE, the better.

[26] Limited GPU resources on Google Colab were used to run the code successfully.

Results

4.1 General Metrics

4.1.1 Popularity

The starting point of the analysis is the SJR[27] Score :

Table 1: Top 20 Journals in Pharmacology by SJR Score - Date : 10/01/2026

Journal	SJR	Journal	SJR
Nature Reviews Drug Discovery	30.506	Pharmacological Research	2.664
Pharmacological Reviews	6.603	Cell Chemical Biology	2.659
Drug Resistance Updates	5.282	**British Journal of Pharmacology**	**2.344**
Journal for ImmunoTherapy of Cancer	4.220	Cellular and Molecular Life Sciences	2.299
Annual Review of Pharmacology & Toxicology	4.029	Acta Pharmacologica Sinica	2.297
Molecular Therapy	4.008	BioDrugs	2.290
Pharmacology and Therapeutics	3.800	Asian Journal of Pharmaceutical Sciences	2.251
Trends in Pharmacological Sciences	3.771	npj Vaccines	2.236
Medicinal Research Reviews	3.169	Apoptosis	2.233
Neuropsychopharmacology	2.944	Clinical Pharmacology and Therapeutics	2.127

The BJP has a strong SJR score and is ranked 13th among Q1 pharmacology journals, positioning it as one of the most recognized journals in the field. Notably, at the time of data collection[28], BJP was ranked 12th.

For each journal, this score reflects several contributing factors, in particular the cumulative citation counts of their published works.

Table 2: Top 5 Q1 Pharmacology Journals by Cumulative Citations

Rank	Journal	Cumulative Citations
1	Antimicrobial Agents and Chemotherapy	1,821,702
2	Journal of Pharmacology and Experimental Therapeutics	1,433,266
3	European Journal of Pharmacology	1,302,949
4	**British Journal of Pharmacology**	**1,193,827**
5	Pain	1,187,131

As shown in Table 2, the BJP ranks 4th in terms of cumulative citations among leading pharmacology journals, highlighting its substantial overall impact and visibility within the field.

Table 3 shows that while BJP's average citations per publication (45.17) place it at 18th, this indicates that individual articles receive fewer citations on average compared to top journals, suggesting a moderate per-article influence despite the journal's high total impact.

[27] SJR score is the metric used by Scimago to rank and determine Q1 journals - Scimago Journal Rank – Pharmacology Q1

[28] September 2025

Table 3: Top Q1 Pharmacology Journals by Average Citations per Work

Rank	Journal	Average Citations per Work
1	Pharmacological Reviews	185.61
2	Nature Reviews Drug Discovery	110.33
3	Pharmacology & Therapeutics	92.31
4	Drug Resistance Updates	73.71
18	**British Journal of Pharmacology**	**45.17**

Next, the citations per publication across Q1 pharmacology journals are examined to provide an overview of individual article impact.

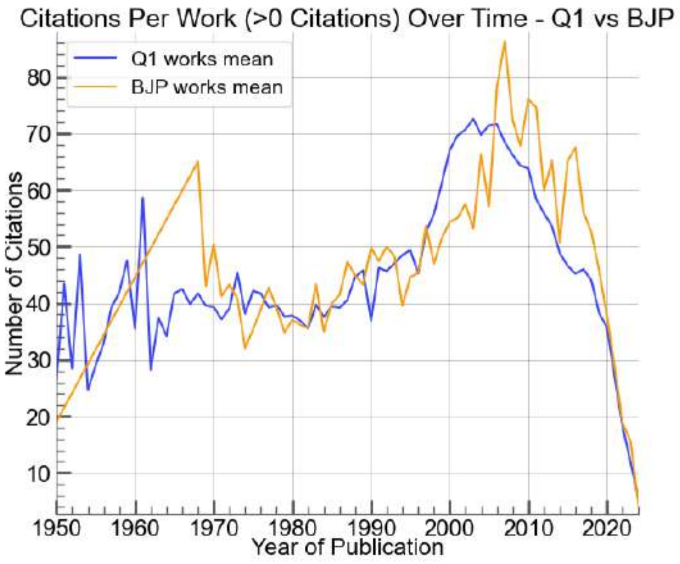


Figure 4: Number of Citations per Work (non-null)

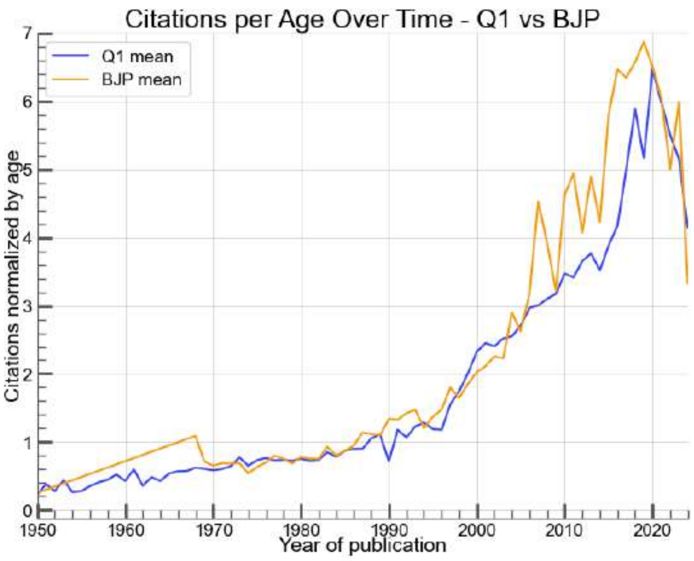


Figure 5: Citations Normalized by Publication Age

The number of citations per publication[29] appears to increase year by year, except since the late 2000s. During this period, the BJP shows a slightly higher average than other Q1 journals, confirming the patterns observed in the tables. The subsequent decline in citations is primarily due to recent publications having had less time to accumulate citations. To account for this, the same plot was normalized by considering citations per publication age[30], which reveals that the apparent decrease occurs much later when controlling for publication age. (see Appendix B for citation-related age analysis.)

Important Note

The following analysis focuses on different metrics which have already been described in the previous sections (*Data Structuring and Preparation*, Section 3.2.3 - *General Metrics and Journal-Level Ranking*, Section 3.3.1).

First of all, As illustrated in Table 4, the British Journal of Pharmacology maintains a high position in

[29] Only non-null citations were considered to limit errors in the dataset, as some publications with no citations may be incomplete or missing data.

[30] Citations per publication age were computed for each work as Number of citations / Publication age.

Table 4: Top Q1 Pharmacology Journals by Citations Received since 2012

Rank	Journal	Number of Citations
1	Antimicrobial Agents and Chemotherapy	451,209
2	Pain	410,767
3	Journal of Ethnopharmacology	369,586
4	Nature Reviews Drug Discovery	309,736
6	**British Journal of Pharmacology**	**220,058**

citations received since 2012, suggesting that its current popularity reflects recent impact rather than long-standing citation history.

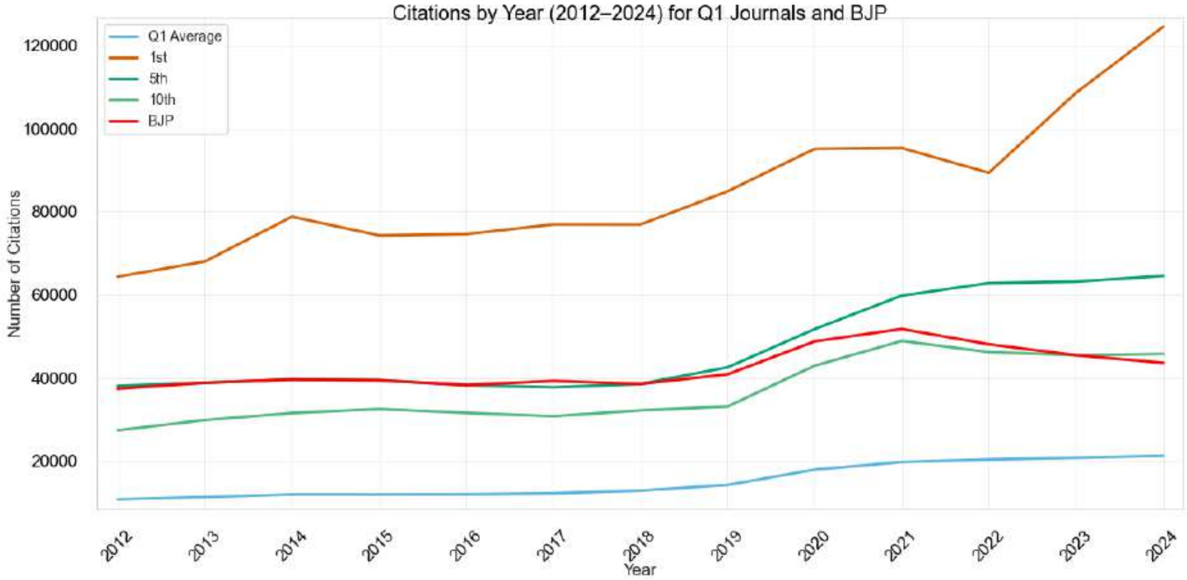
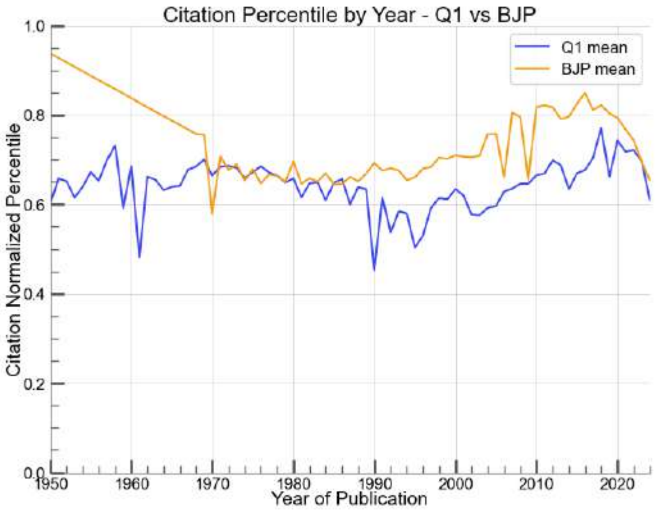


Figure 6: Number of Citations Received Each Year

Nevertheless, as shown in Figure 6, the number of citations received each year has been declining. Indeed, the journal ranked within or near the top 5 between 2012 and 2020, whereas it now falls outside the top 10.



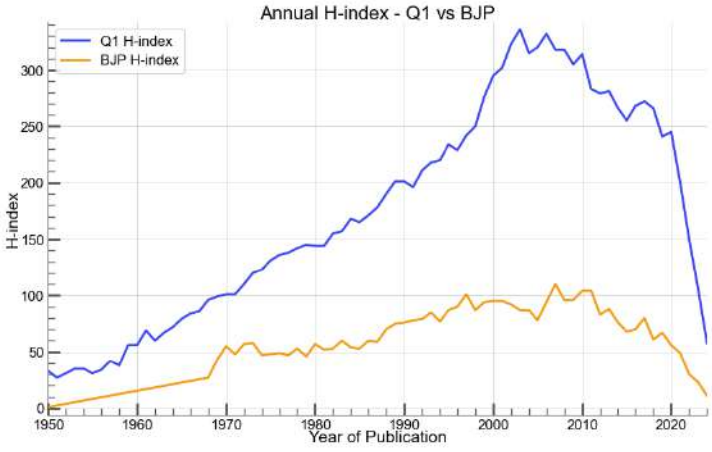


Figure 7: Citation Percentile Per Work

Figure 8: Q1 Journals H-Index

The citation percentile metric confirms that publications from the BJP are highly popular within the field of pharmacology, as well as among Q1 publications (Figure 7).

The H-index (Figure 8) further shows that the BJP consistently produces a significant number of highly cited publications annually, especially compared to other Q1 journals[31]. A peak is observed for the BJP in 2007, with over 110 works receiving more than 110 citations.

Finally, the analysis focuses on the mean MNCS of selected pharmacology journals, with particular attention to the BJP :

Rank	Journal	Mean MNCS
1	Pharmacological Reviews	4.80
2	Nature Reviews Drug Discovery	2.71
3	Pharmacology & Therapeutics	2.34
4	Drug Resistance Updates	2.24
25	**British Journal of Pharmacology**	**1.12**

Table 5: Top Q1 Pharmacology Journals by Mean MNCS

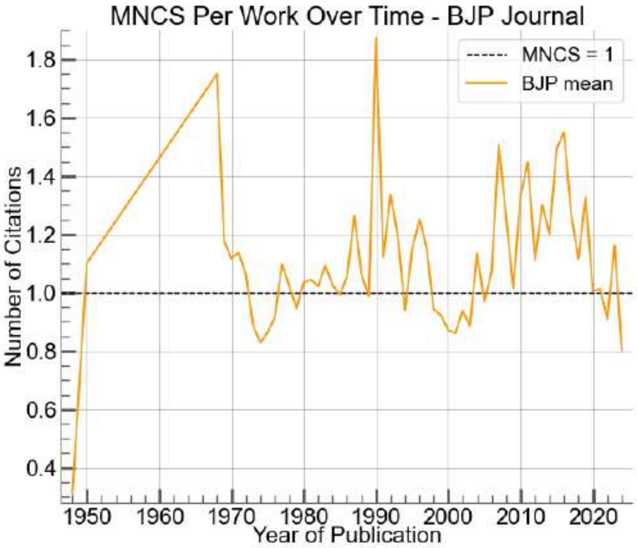


Figure 9: BJP Works Mean MNCS Over Time

Table 5 indicates that while the BJP produces a substantial number of highly cited works (confirmed with Figure 9), the average citation impact per publication is moderate compared to top-ranked journals such as Pharmacological Reviews or Nature Reviews Drug Discovery

4.1.2 Other Metrics

Before analysing the factors influencing a journal's popularity, we first examine some general characteristics of the journals and their publications.

Journal	First year of publication
Journal of Pharmacology and Experimental Therapeutics	1909
Cellular and Molecular Life Sciences	1945
British Journal of Pharmacology	**1948**
Pharmacological Reviews	1949
...	...
Medicine in Drug Discovery	2019
NEJM Evidence	2021

Table 6: Q1 Pharmacology Journals First Year of Publication

[31] The Q1 dataset includes all publications from the considered journals, including BJP; therefore, the H-index could be expected to be higher, but it remains relatively moderate.

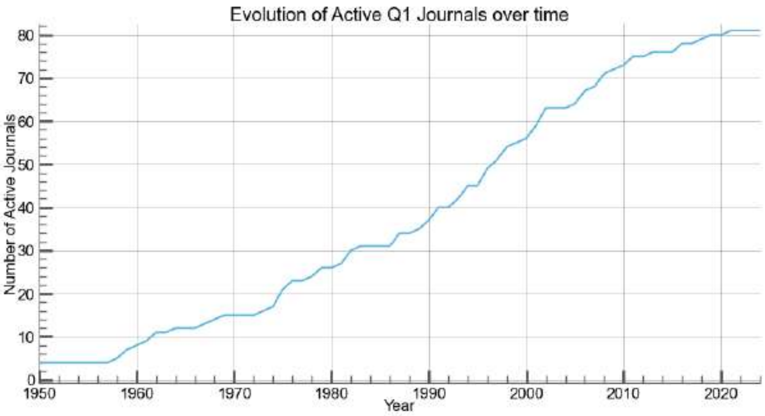


Figure 10: Evolution of Active Journals

The BJP is one of the oldest journals among the Q1 category (see Table 6), even when accounting for certain limitations of the database. Additionally, the number of currently popular journals (Q1-ranked) increases in an approximately linear manner over time (Figure 10), indicating a relatively homogeneous temporal distribution of active[32] journals.

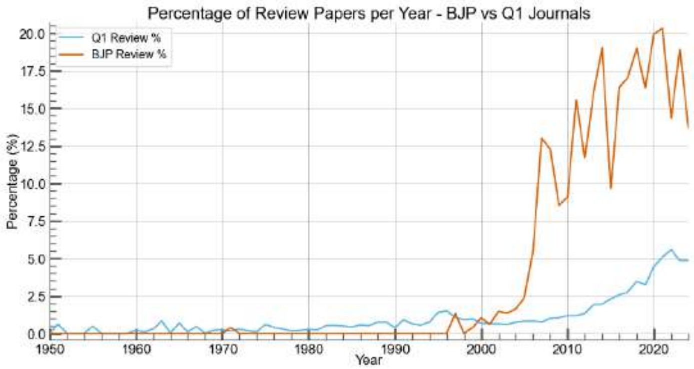


Figure 11: Review Percentage Over Time

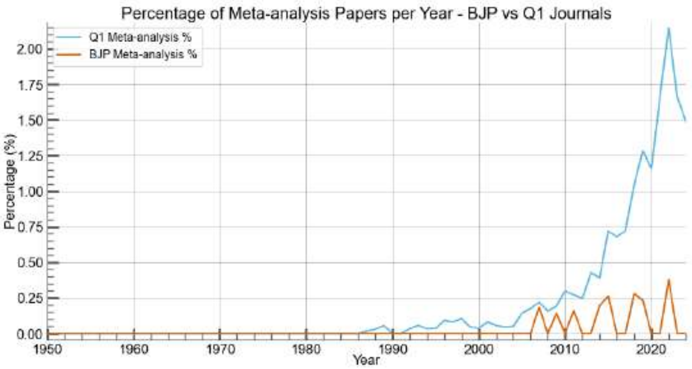


Figure 12: Meta-analysis Percentage Over Time

As shown in Figures 11 and 12, the BJP publishes a markedly higher share of review articles than the Q1 average, particularly from the 2000s onward, whereas the opposite pattern is observed for meta-analyses, which are proportionally more prevalent in Q1 journals on average.

Table 7: Top 5 Cited Q1 Works

Rank	Title	Citations
1	A … activity	25,983
2	Nitric … pharmacology.	15,661
3	Antibodies … manual	12,751
4	Relationship … reaction	12,747
5	A … reactions	11,115

Table 8: Top 5 Cited BJP Works

Rank	Title	Citations
1	Animal … guidelines	3,465
2	Principles … discovery	2,474
3	Measuring … mean?	2,200
4	Guide … edition	2,073
5	Characterization … *in vivo*	1,876

A quick note on the top works (Tables 7, 8) : these publications are expected to have the greatest impact on the popularity analysis presented in later sections. *Note : No highly cited Q1 works are from the BJP.*

[32] A journal is considered active if, from a given year onward, the database contains at least one work published in that journal.

4.1.3 Influential Factors

In this section, we first focus on the number of publications in the BJP and Q1 journals, with particular attention to their evolution over time.

Table 9: Top Q1 Pharmacology Journals by Number of Publications

Rank	Journal	Number of Publications
1	European Journal of Pharmacology	42,326
2	Journal of Pharmacology and Experimental Therapeutics	38,530
3	European Neuropsychopharmacology	34,749
4	Antimicrobial Agents and Chemotherapy	33,052
6	**British Journal of Pharmacology**	**26,115**

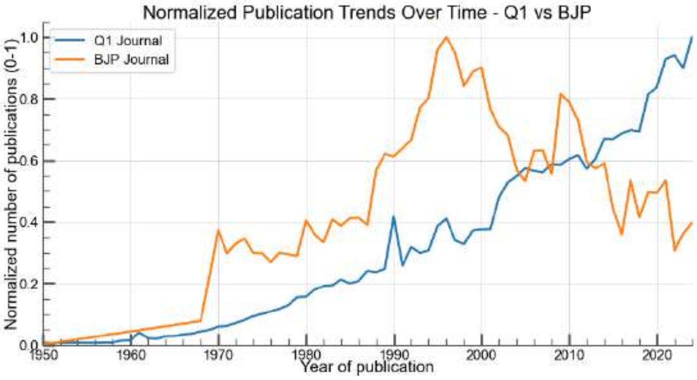


Figure 13: Publication Trends

As shown in Table 9, the British Journal of Pharmacology is among the most active journals in the field of pharmacology. However, its publication volume exhibits a declining trend over time (Figure 13), in contrast to other Q1 journals, which display a continuous increase in publication output[33].

Let's finally take a look at the number of references :

Rank	Journal	Avg. References
1	Pharmacology & Therapeutics	163.73
2	Pharmacological Reviews	125.37
3	Medicinal Research Reviews	121.42
4	Current Neuropharmacology	103.29
48	**British Journal of Pharmacology**	**40.11**

Table 10: Top Q1 Pharmacology Journals by Average Number of References per Publication

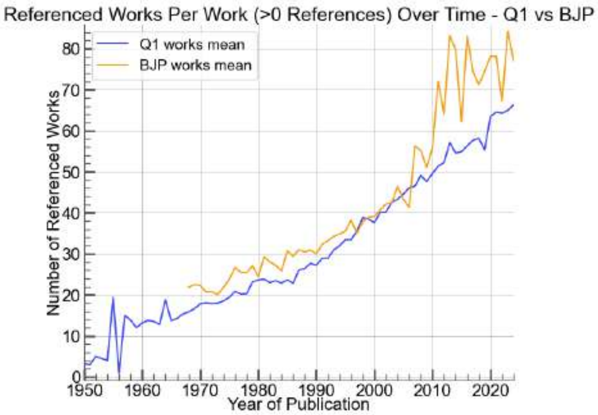


Figure 14: Number of References per Work

[33] Here, *normalized* refers to a normalization procedure in which yearly publication counts are divided by the maximum number of publications observed in a single year, allowing trends across journals to be compared on a common scale.

Within the set of Q1 journals, the British Journal of Pharmacology (BJP) occupies a mid-range position in terms of the average number of references per publication (Table 10), although a clear improvement (Figure 14) in this metric can be observed since the 2000s[34].

4.2 Authors, Countries, Institutions

4.2.1 Authors

The following section focuses on authors, examining publication activity and collaboration patterns.

Rank	Journal	Avg. Authors
1	npj Vaccines	12.42
2	Journal for ImmunoTherapy of Cancer	11.03
3	NEJM Evidence	8.81
4	Cell Chemical Biology	8.18
57	**British Journal of Pharmacology**	**4.29**

Table 11: Top Q1 Pharmacology Journals by Average Number of Authors per Publication

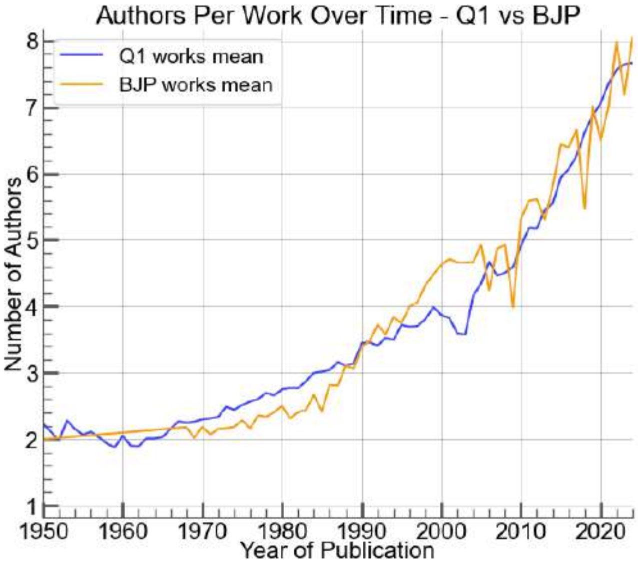


Figure 15: Number of Authors per Work

Although the average number of authors per publication has steadily increased for both Q1 journals and the BJP, with nearly identical growth curves (Figure 15, the journal still ranks relatively low in terms of author counts (Table 11).

The top 20 authors (see Appendix C) by number of citations already provides some insights into the most prominent countries and institutions[35] in the field.

The analysis now moves on to the examination of author gender (detailed in Section 3.3.2 : *Author, Country, and Institution Analysis*).

In the field of pharmacology, male authors are generally more prevalent than female authors, both in terms of overall representation (Figure 17) and among the most highly cited authors (Figure 16). Nevertheless, we can still see with the proportion of female authors gradually increasing over the years. Also, the BJP (Figures 18 and 19) appears to exhibit slightly greater gender diversity, with the proportion of female authors consistenly slightly higher over time. Finally, the notable rise in the percentage of authors with *unknown* gender can be attributed to a factor that will be discussed later: the growing contribution of authors from Asian countries.

[34] Only publications with a non-null number of references were retained, for the same reasons as those applied to citation counts.

[35] Author countries are based on the countries of their institutions. For example, S. P. H. Alexander is British, Erik De Clercq is Belgian.

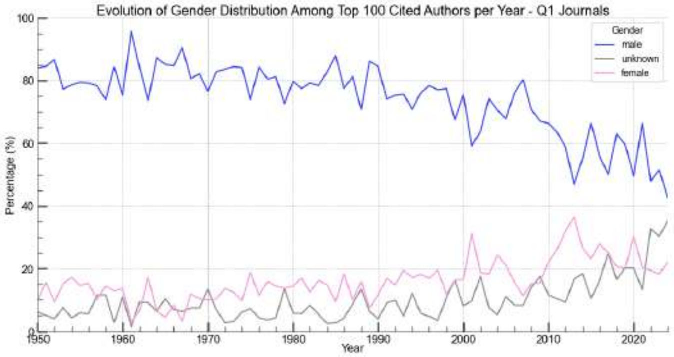


Figure 16: Gender Distribution: Top 100 Authors (Q1 Journals)

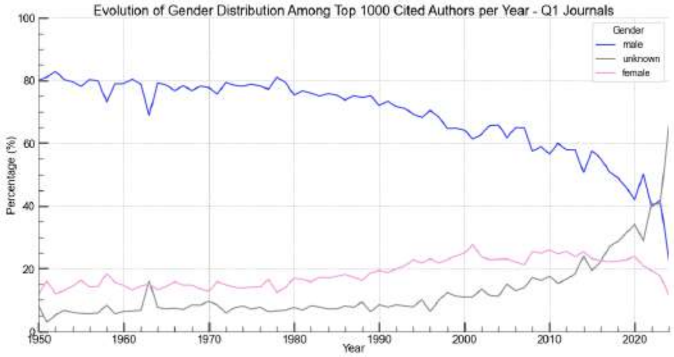


Figure 17: Gender Distribution: Top 1000 Authors (Q1 Journals)

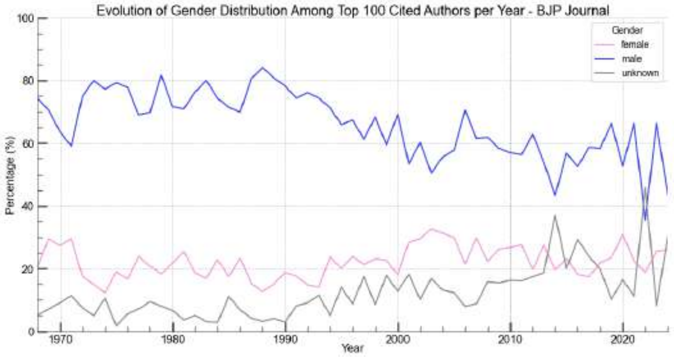


Figure 18: Gender Distribution: Top 100 Authors (BJP)

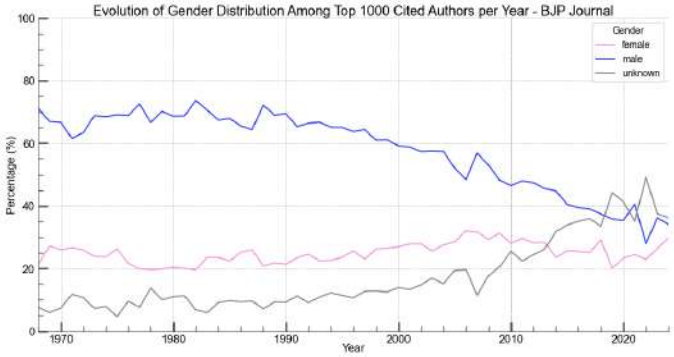


Figure 19: Gender Distribution: Top 1000 Authors (BJP)

Finally, the study turns to collaboration patterns :

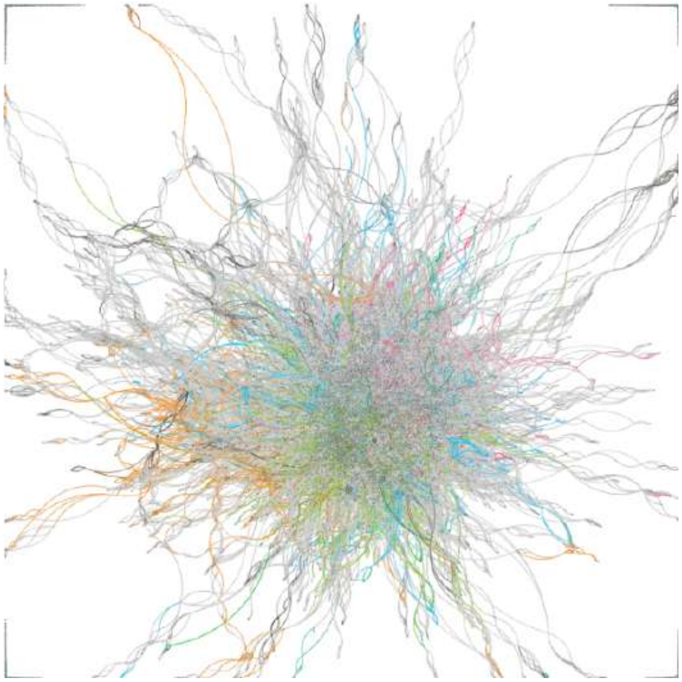
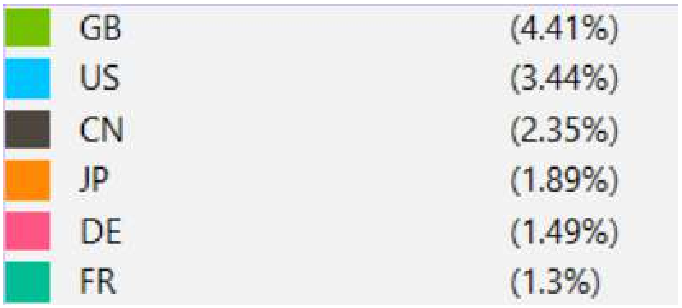


Figure 20: BJP Co-Authorship Network



🟩	GB	(4.41%)
🟦	US	(3.44%)
⬛	CN	(2.35%)
🟧	JP	(1.89%)
🟥	DE	(1.49%)
🟩	FR	(1.3%)

Figure 21: BJP Co-Authorship Network Countries Distribution

According to the BJP co-authorship network (Figures 20 and 21), certain countries stand out both in terms of presence and positioning.[36] For example, there is a high density of authors from Great Britain and the United States near the center of the network, while authors from Japan appear toward the bottom left and those from Germany toward the top right.

Among all available co-authorship network metrics[37], only the analysis of the largest connected component **(LCC)** has been presented[38]:

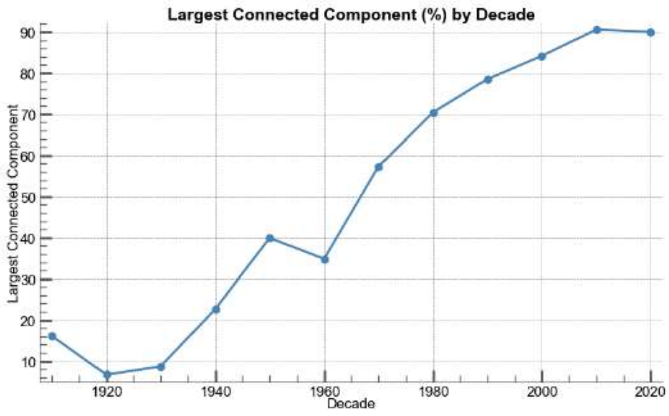


Figure 22: Largest Connected Component Network (Q1 Journals)

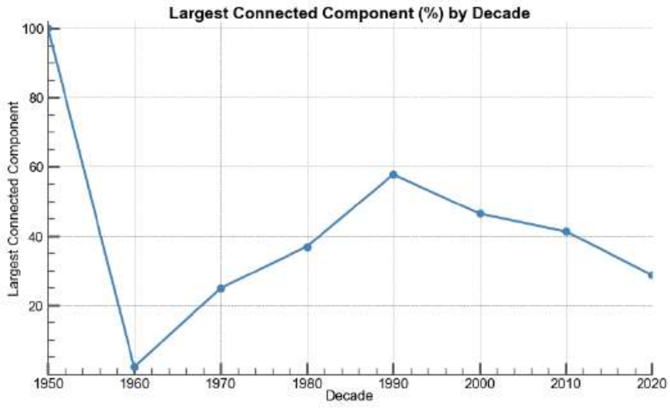


Figure 23: Largest Connected Component Network (BJP)

For Q1 journals (Figure 22), the connectedness of scientists has shown a strong evolution over the decades, reaching a peak of approximately 91% in the 2010s. In contrast, for the BJP (Figure 23), connectedness increased steadily from the 1960s to the 1990s, peaking around 58%, but then experienced a gradual decline toward the 2020s, reaching roughly 30%. One possible explanation for this trend is the decrease in the number of BJP publications, which reduces the number of authors appearing in the journal.

Based on the collaboration graphs among top authors (see Appendix D), two distinct groups can be observed: one composed of Italian authors and the other of French authors. These countries and authors collaborate extensively within Q1 pharmacology research. In contrast, the BJP collaboration graph appears much more diversified, showing a wider distribution of collaborative links across different authors and countries.

4.2.2 Countries

This section focuses on countries, highlighting those with the highest presence and citations, their evolution over time, and patterns of international collaboration.

As with authors, the average number of distinct countries per publication has increased for both Q1 journals and the BJP (Figures 24, 25). Notably, since 2015, the BJP exhibits even greater diversity on average, in contrast to Q1 journals, which show a slight decline in mean country counts[39]. Consequently, the BJP ranks higher than it does for authors, although its position remains relatively moderate (Table 12).

[36] Some nodes appear in gray because the countries of certain authors could not be retrieved.

[37] Other metrics include degree centrality, betweenness centrality, clustering coefficient, and network density.

[38] The **LCC** is defined as the largest subset of nodes in a network in which every node is reachable from any other node via a path of edges.

[39] This will be explained later in this section.

Rank	Journal	Avg. Number of Countries
1	NEJM Evidence	2.11
2	npj Vaccines	1.72
3	Journal for ImmunoTherapy of Cancer	1.68
4	Therapeutic Advances in Neurological Disorders	1.62
32	**British Journal of Pharmacology**	**1.29**

Table 12: Top Q1 Pharmacology Journals by Average Number of Distinct Countries per Publication

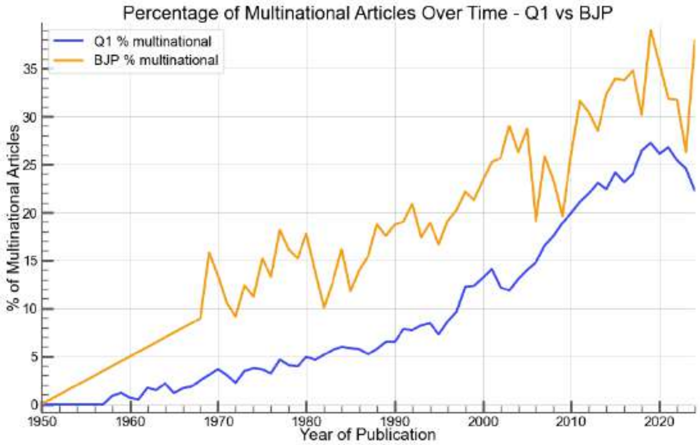


Figure 24: Percentage of Multi-National Articles

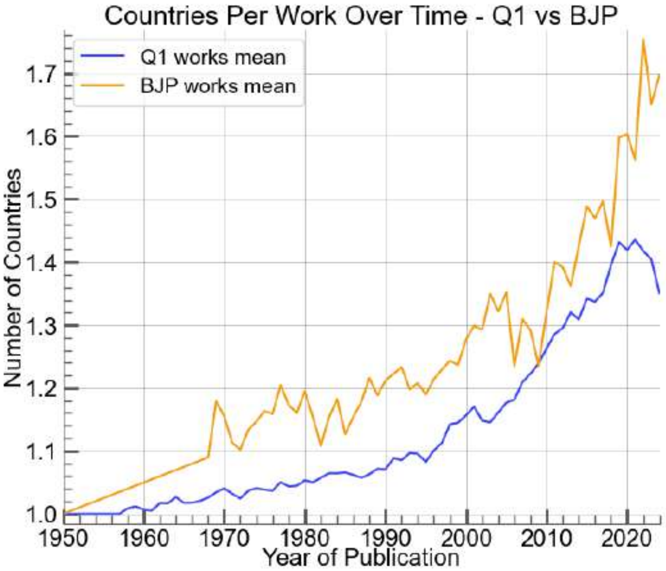


Figure 25: Number of Distinct Countries per Work

Regarding the analysis of top countries, the United Kingdom is highly represented and influential in BJP publications (Figure 26), with the United States also contributing significantly, but to a smaller degree. In contrast, for Q1 journals overall (Figure 27), the United States and China dominate the publication landscape. An important observation is that Asian countries, particularly China and Japan, produce a relatively large proportion of articles compared to the citations they receive; in other words, they publish a lot but their works are cited less frequently. This is also confirmed comparing the countries with their mean MNCS[40]. (see Appendix E)

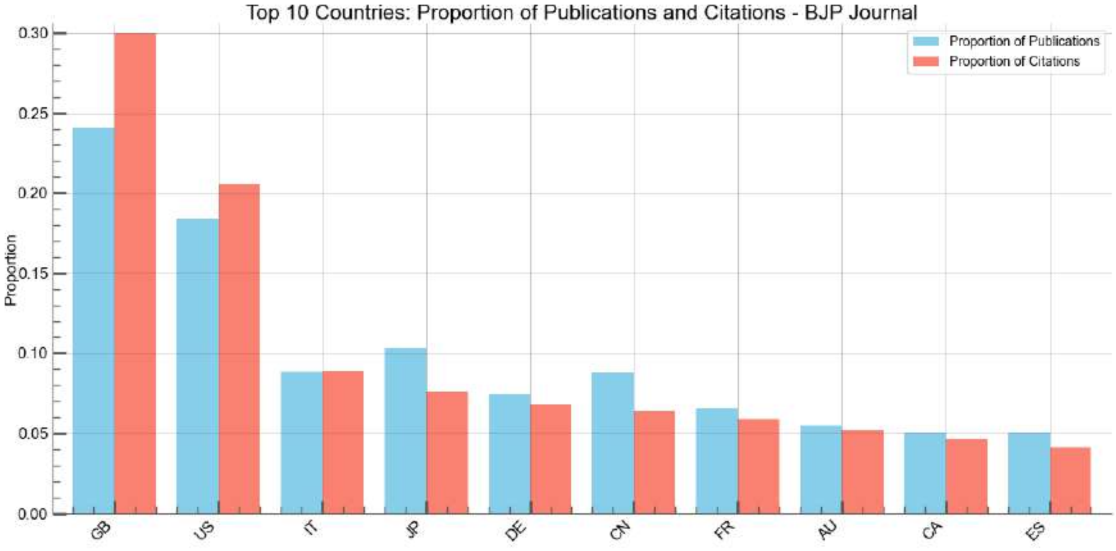


Figure 26: Proportion of Publications and Citations for Top 10 BJP Countries

[40] Only countries with more than 100 publications were retained for visualization and consistency purposes, as countries with very few publications can have disproportionately high mean MNCS values.

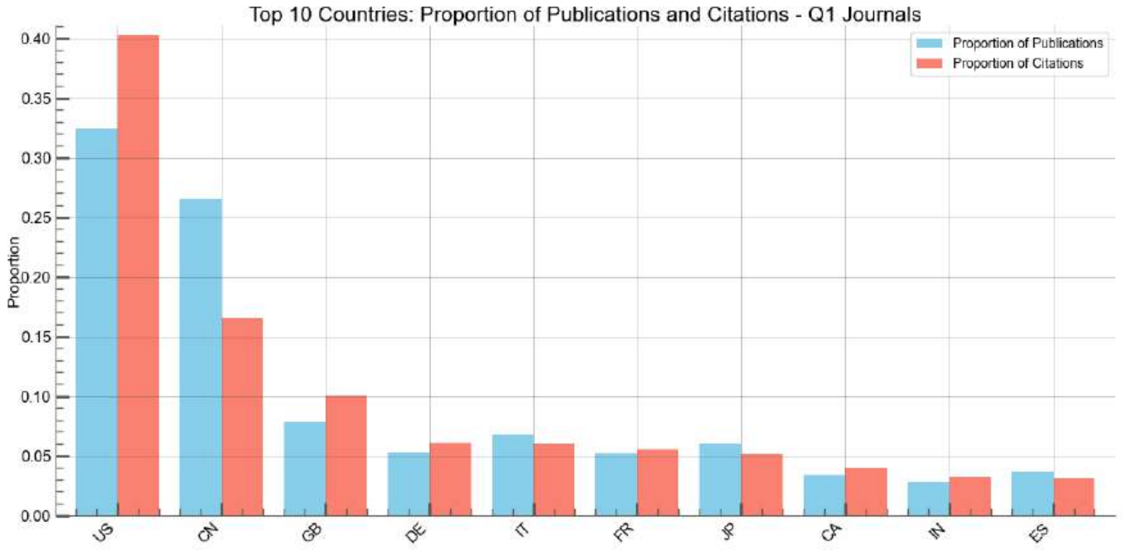


Figure 27: Proportion of Publications and Citations for Top 10 Q1 Countries

As shown in Figures 28 and 29, the United States consistently occupy a dominant position across journals throughout the period. In contrast, China exhibits a strong and rapid emergence starting in the 2000s–2010s. Focusing specifically on the British Journal of Pharmacology, the United Kingdom dominated the journal during the 1970s to 1990s (corresponding to the journal's peak activity) before experiencing a marked decline that persists to the present day[41].

Figure 30 confirms this observation, showing that Chinese publications are, on average, more recent. Additionally, because of its publication rate, publications in the BJP are generally older on average, with Japan constituting a notable exception.

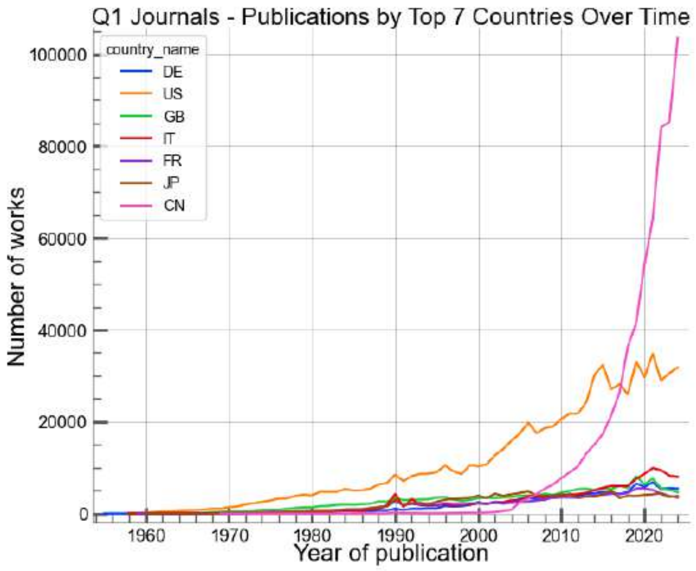


Figure 28: Annual Number of Publications by Top 7 Countries (Q1).

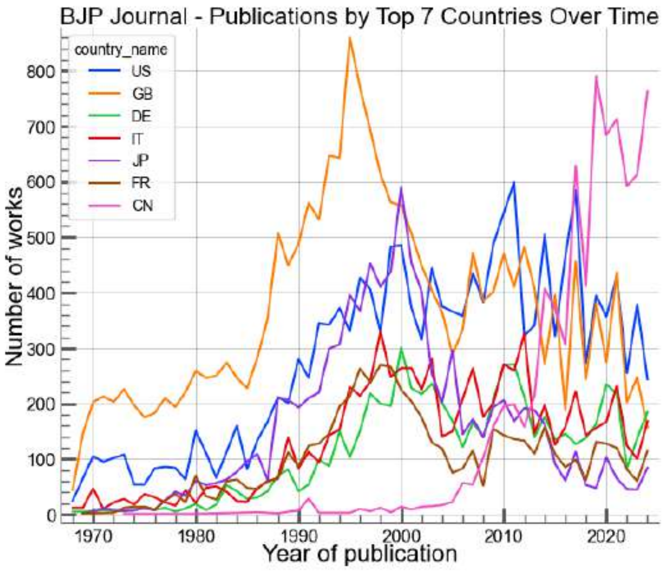


Figure 29: Annual Number of Publications by Top 7 Countries (BJP).

[41] The top seven countries were retained as they are common to both databases.

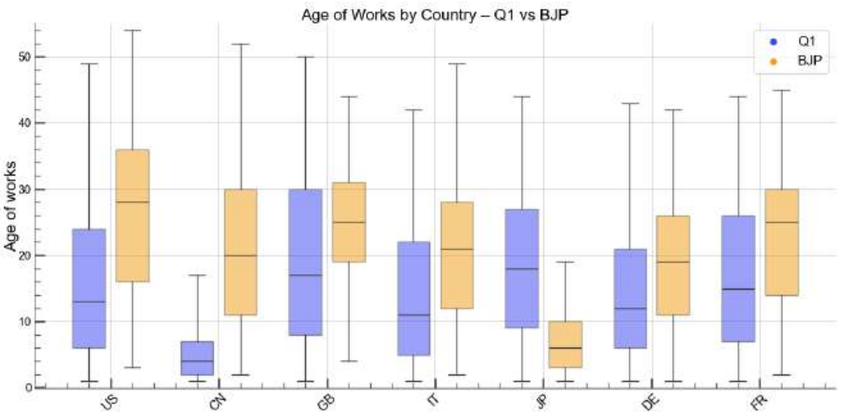


Figure 30: Mean age of publications for the top 7 countries.

Beyond the age of publications, such as the relative recency of Chinese research, differences in international collaboration networks (see Appendix F) may help explain variations in the popularity of Chinese works compared with those from the United States and the United Kingdom. Indeed, the United States and the United Kingdom exhibit dense and diverse collaboration patterns, both between each other and with a wide range of other countries, whereas Asian authors tend to collaborate less internationally and more frequently within regional or domestic networks[42] . This pattern is reflected in the collaboration graphs, where Asian countries display relatively small node sizes despite their high publication output.

4.2.3 Institutions

Rank	Journal	Avg. Number of Institutions
1	NEJM Evidence	6.13
2	Journal for ImmunoTherapy of Cancer	4.84
3	Therapeutic Advances in Neurological Disorders	4.05
4	npj Vaccines	3.77
63	**British Journal of Pharmacology**	**1.88**

Table 13: Top Q1 Pharmacology Journals by Average Number of Distinct Institutions per Publication

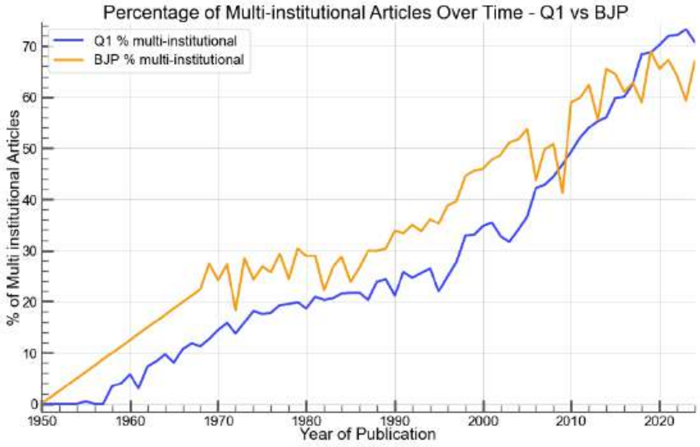


Figure 31: Percentage of Multi-Institutional Articles

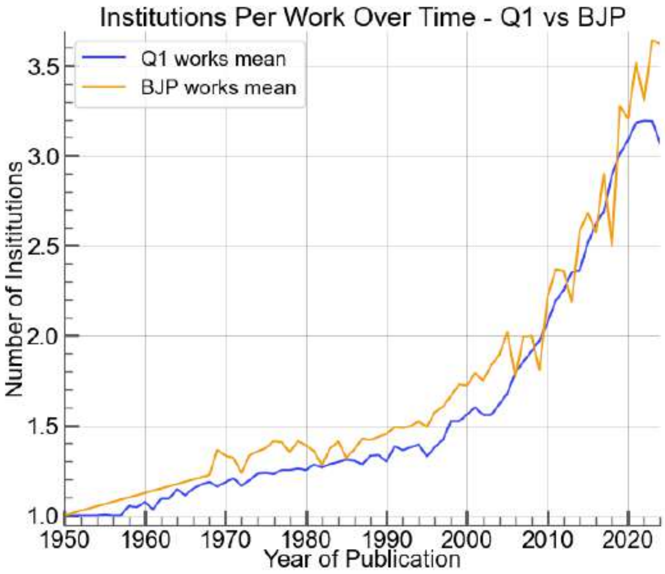


Figure 32: Number of Distinct Institutions per Work

[42] For the matrix : Rows represent each country's collaboration profile, indicating how a country distributes its collaborations across partner countries (each cell corresponds to the percentage of the country's total collaborations). Columns reflect each country's importance as a collaboration partner, measuring how often other countries collaborate with that country.

We will focus less on this *Institutions* section, as institutions are somewhat less informative and largely mirror the patterns observed at the country level, given that country data are derived from institutional affiliations.

For example, the BJP shows nearly comparable performance in terms of institutional diversity as it does for country diversity, even though the journal achieves a higher overall ranking (Table 13). Moreover, the number of distinct institutions evolves in a similar manner, showing a growth pattern comparable to that of countries (Figures 31, 32).

However, the following observations merit attention:

- (see Appendix G)[43] For the BJP, many British universities dominate the ranking of top institutions. Notably, some institutions such as the University of Edinburgh are highly influential despite a relatively lower publication volume. For Q1 journals, a larger number of American and Chinese institutions appear among the top ranks, along with some institutions in common with the BJP, such as French institutions (e.g., Inserm, CNRS[44]). Notably, the NIH[45] is the most popular institution on average in the pharmacology field, even though Inserm has a higher cumulative number of citations.

- (Figures 33, 34) The plots of top citations by institution type[46] show that, overall, national institutions appear more dominant than universities, while universities themselves are more impactful than companies in the pharmacological research landscape. This pattern is somewhat less pronounced for the BJP, which likely relies more heavily on universities.

- (see Appendix H) In terms of collaboration, institutions from the same country or continent are generally clustered. For BJP, only European institutions are present, with a particularly strong influence from British institutions. In Q1 journals, there is notable collaboration among French institutions, and surprisingly, also among Chinese institutions, although to a lesser extent.

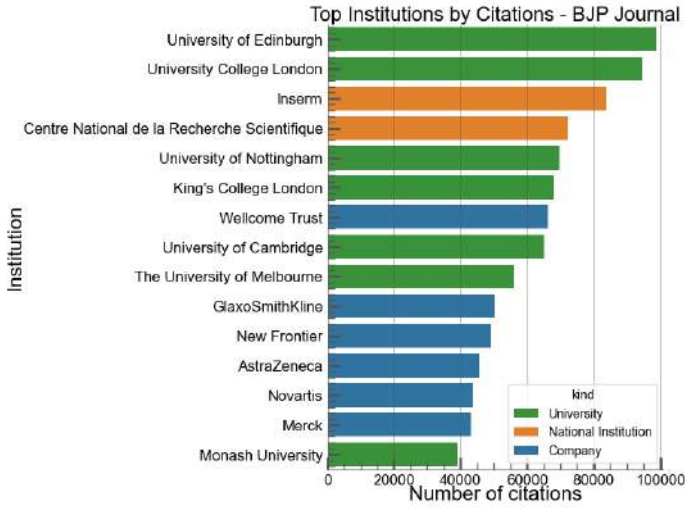


Figure 33: Citations of Top Institutions by Type (BJP)

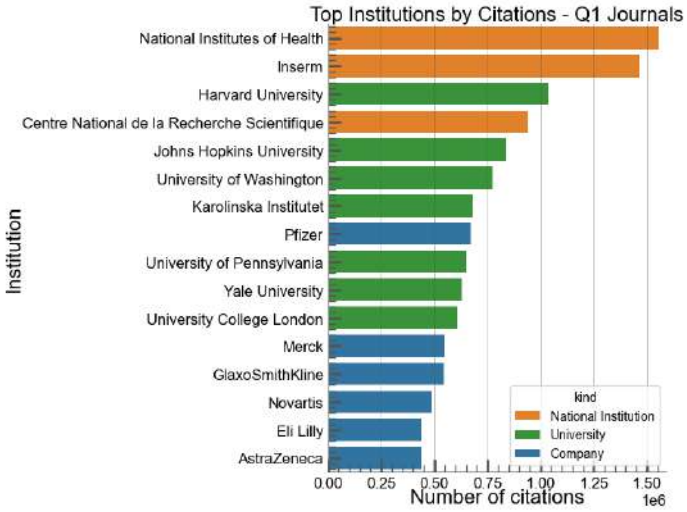


Figure 34: Citations of Top Institutions by Type (Q1)

[43] The mean MNCS and publication counts are normalized relative to the top 20 institutions. All institutions have a good mean MNCS, but a higher number of publications generally results in a lower mean MNCS performance. The horizontal line indicates which institutions over- or under-perform relative to this top 20 benchmark.

[44] Centre National de la Recherche Scientifique

[45] National Institutes of Health

[46] For both plots, only the top 3 national institutions, the top 7 universities, and the top 6 companies were retained.

4.3 Topics

Important Note

The following analysis focuses on different metrics which have already been described in the previous sections (3.3.3, *Topic Analysis* - 3.5, *Topic Modelling*).

4.3.1 OpenAlex Topics

Firstly, metrics related to concepts and keywords will be quickly explored (see Appendix I).

Regarding concepts, the analysis reveals that for the BJP, eight concepts clearly stand out in terms of publication volume, whereas this pattern is less pronounced for Q1 journals, where the distribution is more evenly spread. In terms of popularity, most of these concepts achieve relatively high mean MNCS values, with the notable exception of *Computer Science* (Figures 77, 78).

Regarding keywords, given their much larger number compared to the more general concepts, we focus here on the differences between the two sets of works (Figures 79, 80). For example, the most frequently published keywords for Q1 journals do not even appear in the BJP top 20[47]. In terms of popularity, some keywords are highly cited (e.g., *Drug Development* for Q1 journals, *Capsaicin* for BJP), while others are less so (e.g., *Clinical Pharmacology* for Q1 journals, *Contractility* for BJP). Finally (Figure 81), a group of keywords consistently co-occurs within the general corpus, including *Tolerability*, *Drug Development*, and *Pharmacodynamics*.

The following analysis focuses on the primary topics identified in the corpus. The plots of the top published primary topics (see Appendix J), which operate in the same way as those presented in the institutions section, allow us to identify the most frequently published topics as well as their average popularity. In particular, for BJP, the primary topic *Pain Mechanisms and Treatments* is highly popular on average relative to its number of publications. For Q1 journals, the primary topic *Cannabis and Cannabinoid Research* is not only among the most frequently published topics, but also achieves the highest mean citation count.

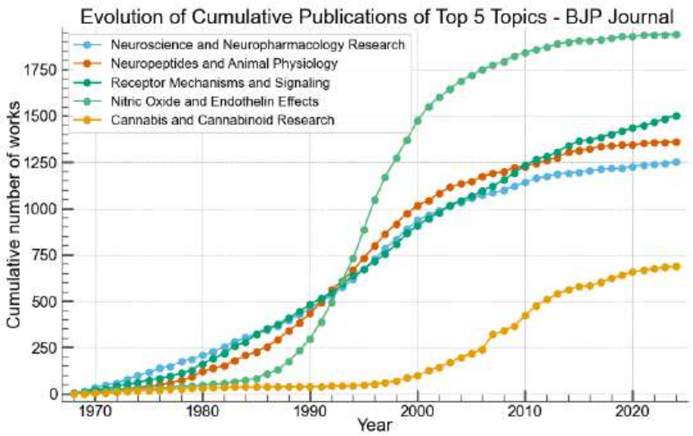


Figure 35: Top 5 BJP Topics Evolution of Publications

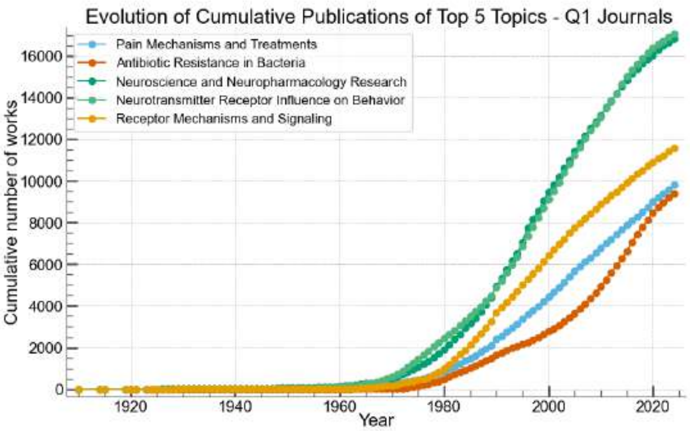


Figure 36: Top 5 Q1 Topics Evolution of Publications

A closer examination was conducted on the five most published primary topics.

Regarding the evolution of their cumulative publication counts (Figures 35, 36), the most published topics in BJP experienced a sharp increase between 1985 and 2000 before stabilizing, with the exception of *Cannabis*

[47] In fact, there are no keywords in common between the two top 20 lists.

and Cannabinoid Research, which shows a more recent surge. This pattern is consistent with the overall publication trends previously observed for the journal. In Q1 journals, the cumulative number of publications shows a much more gradual, steady increase since the 1990s, except for *Antibiotic Resistance in Bacteria*, which emerged more recently. These observations indicate that most of these topics have been consistently popular over time rather than representing emerging research areas.

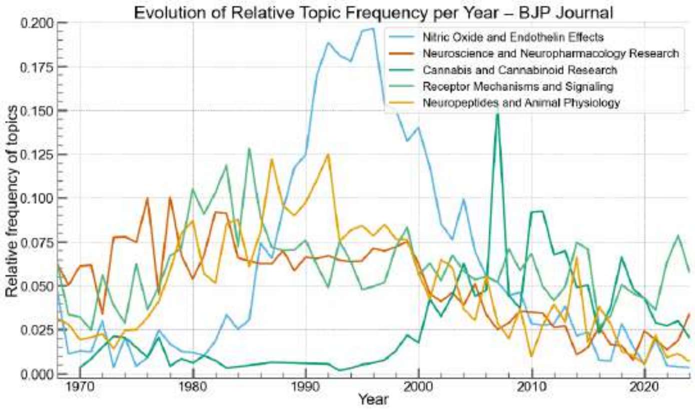


Figure 37: Top 5 BJP Topics Relative Frequency

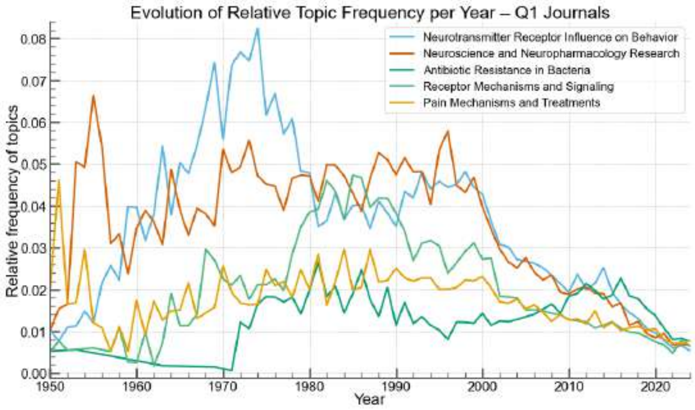


Figure 38: Top 5 Q1 Topics Relative Frequency

Regarding the analysis of *Relative Frequency*[48] (Figures 37, 38). Likely due to the smaller overall number of publications, the top primary topics in BJP appear more prominent and dominant compared to those in Q1 journals. Indeed, the relative frequencies of these topics are considerably higher; for example, *Nitric Oxide and Endothelin Effects* reaches nearly 20% between 1990 and 2000 (coinciding with the peak publication period of the BJP which is unexpected !) In contrast, for Q1 journals, the peak frequency is lower, around 8–9%, as observed for *Neurotransmitter Receptor Influence on Behavior* in the 1970s.

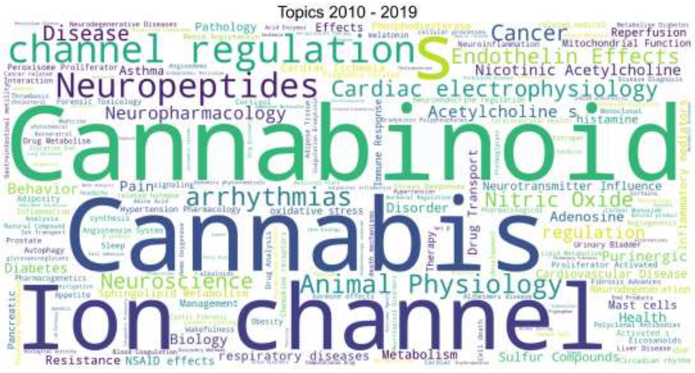


Figure 39: BJP Best Topics from 2010 to 2019

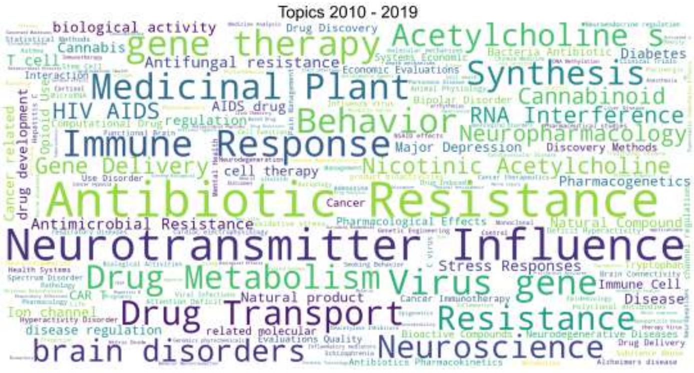


Figure 40: Q1 Best topics from 2010 to 2019

Finally, some word clouds illustrate the differences between BJP and Q1 journals. Notably, in relation to a recent event (Figures 41, 42) BJP has devoted little to no attention to COVID-19, unlike other journals in the pharmacological research field. It can also be noted that the words in the word clouds are generally larger, reinforcing the impression that certain topics dominate BJP publications more prominently.

[48] *Relative frequency* corresponds to the proportion of works in a given year that include the topic. For example, a relative frequency of 0.1 indicates that 10% of the publications in that year addressed this primary topic.

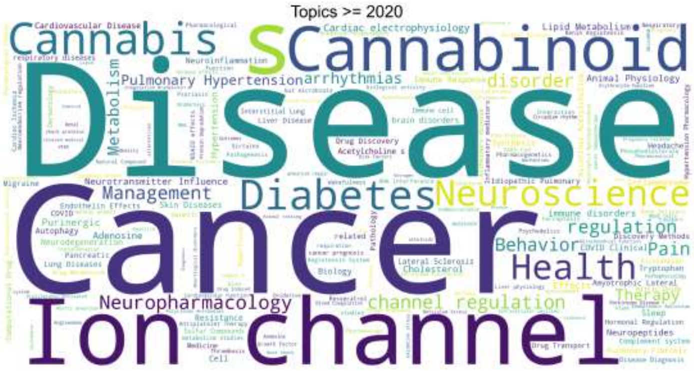


Figure 41: BJP Best topics from 2020 to 2024

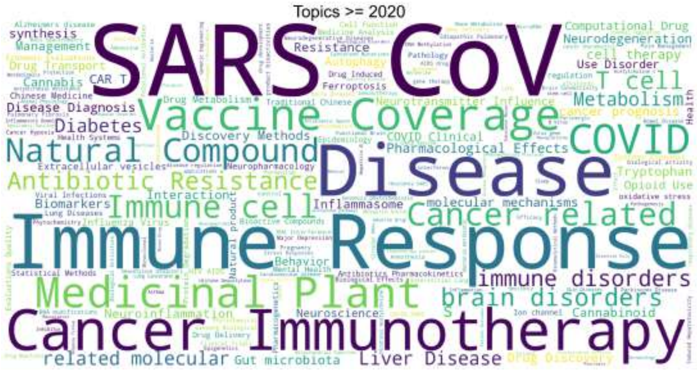


Figure 42: Q1 Best topics from 2020 to 2024

4.3.2 Topic Modelling

The topics section ends with an analysis of topics modeled using **BERTopic**, providing an alternative approach to examining thematic trends in Q1 journals and, more specifically, in BJP.

The exploration initially examines the well-published[49] topics and examine their relationship with MNCS (see Appendix K). For Q1 journals, the "over-published" topics include *schizophrenia*, *ethanol_alcohol*, and others, while the "under-published" topics comprise *design_receptor*, *pain_questionnaire*, etc. For the BJP, the "over-published" topics are *platelets*, *muscarinic_receptors*, and similar, whereas the "under-published" topics include *ht_ht1a*, *hydroxytryptamine*, among others.

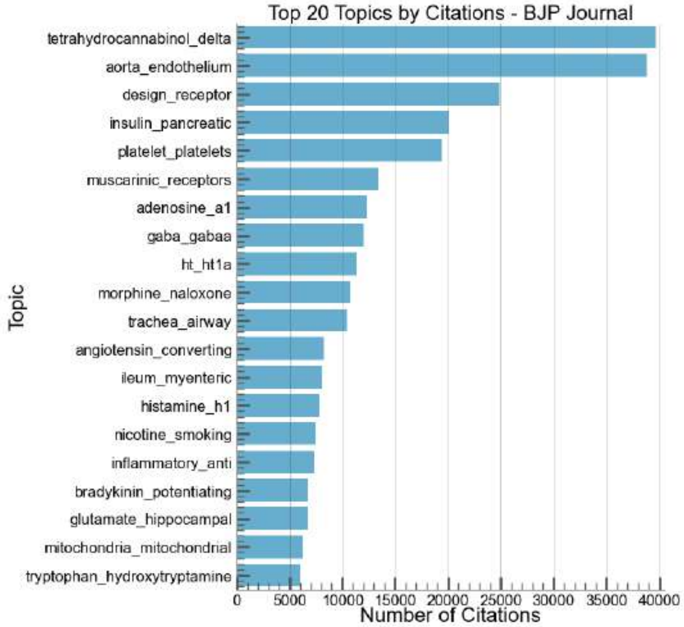


Figure 43: Top 20 Cited BJP Modelled Topics

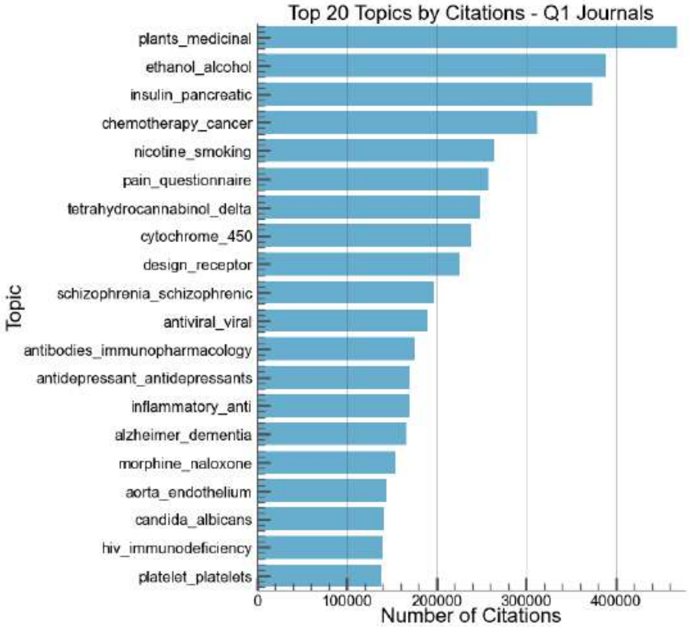


Figure 44: Top 20 Cited Q1 Modelled Topics

Regarding the most cited topics (Figures 43, 44), two topics stand out in BJP: *delta* Topic and *aorta_endothelium*, each with slightly fewer than 40,000 cumulative citations. In contrast, for Q1 journals, the distribution is more evenly spread, with *plants_medicinal* dominating with over 400,000 citations.

[49] For BJP, only topics with more than 100 publications are retained, while for Q1 journals, only those with more than 2,000 publications are considered.

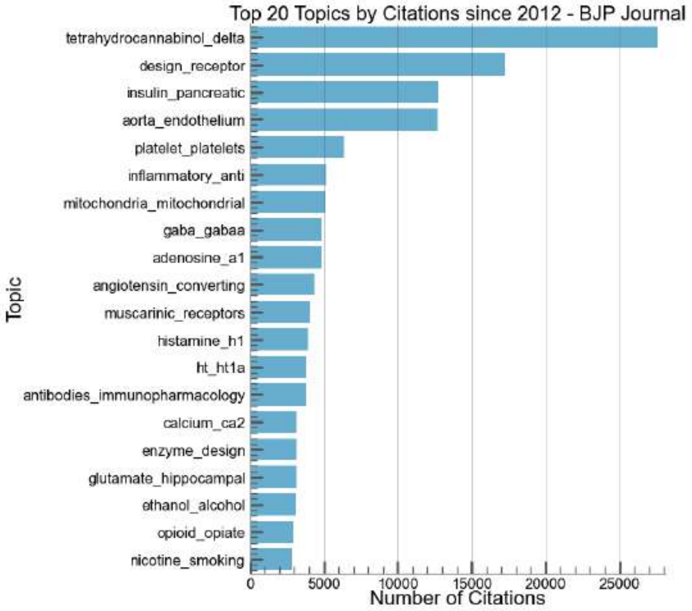


Figure 45: Top 20 Cited (since 2012) BJP Modelled Topics

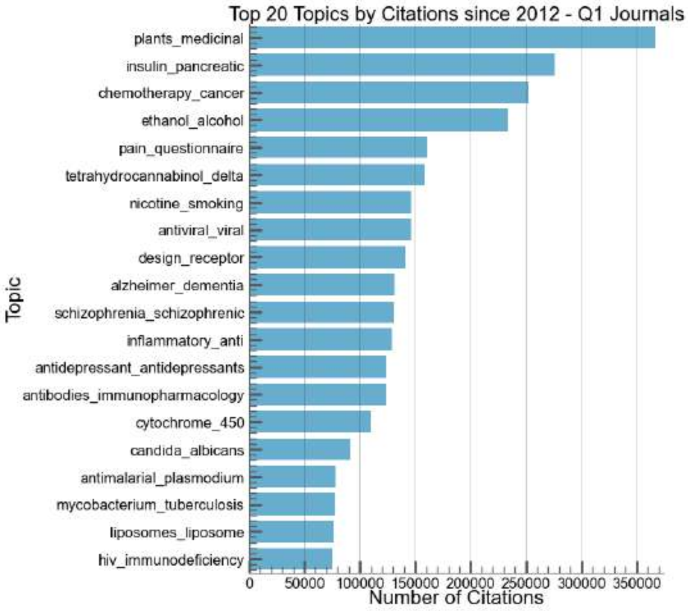


Figure 46: Top 20 Cited (since 2012) Q1 Modelled Topics

However, one aspect of particular interest is the recent popularity of these topics, especially since 2012 (Figures 45, 46). For BJP, it is evident that *delta* Topic has dominated citations received during this period, accounting for approximately 70% of its total citations. In contrast, *aorta_endothelium* obtained the majority of its citations earlier. For Q1 journals, the distribution is similar, with *plants_medicinal* remaining the top topic, accumulating around 75% of its cumulative citations since 2012.

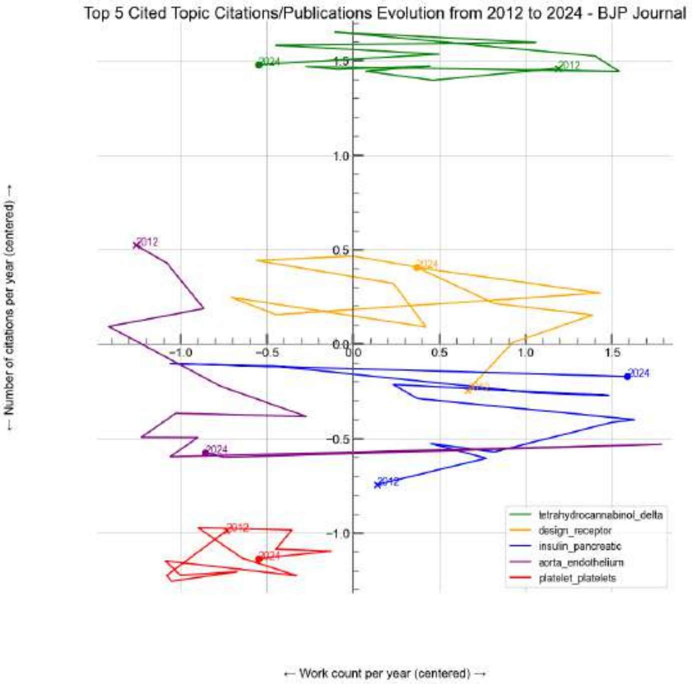


Figure 47: Top 5 Cited BJP Modelled Topics Evolution

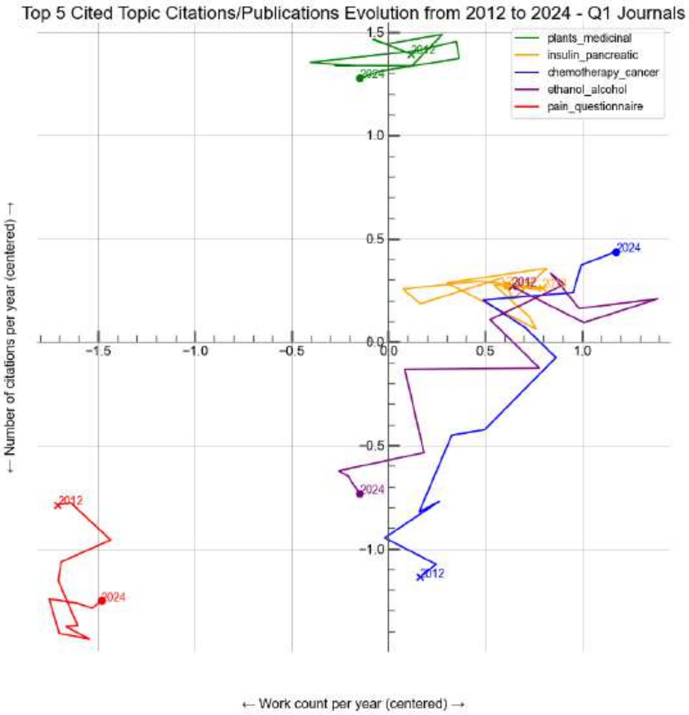


Figure 48: Top 5 Cited Q1 Modelled Topics Evolution

Finally, among the most cited topics since 2012 (Figures 47, 48)[50], their evolution over the years can be analyzed :

- For the BJP, *Tetrahydrocannabinol_Delta* remains at the top, *Insulin_Pancreatic* shows a small increase, and *Aorta_Endothelium* experiences a slight decline in terms of publications[51].
- For Q1 journals, *Plants_Medicinal* remains at the top, *Chemotherapy_Cancer* shows a small increase, and *Ethanol_Alcohol* exhibits a slight decline[52].

4.4 Feature Importance

4.4.1 Descriptive and Statistical Analysis

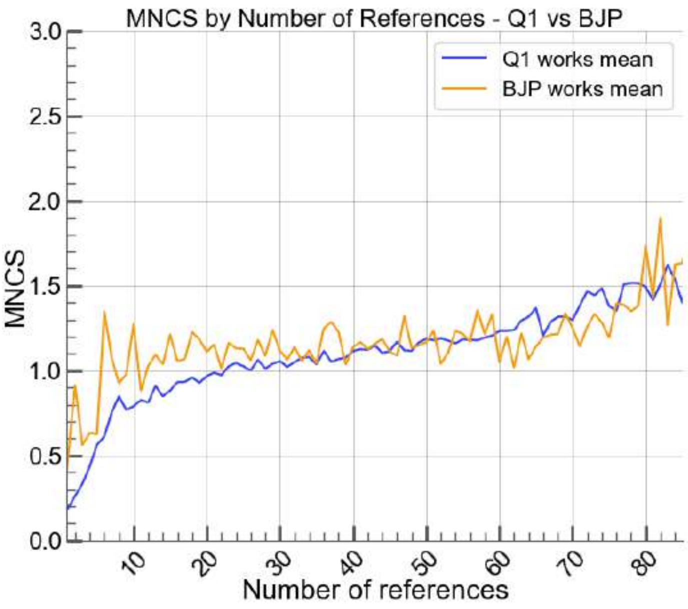


Figure 49: MNCS vs Number of References

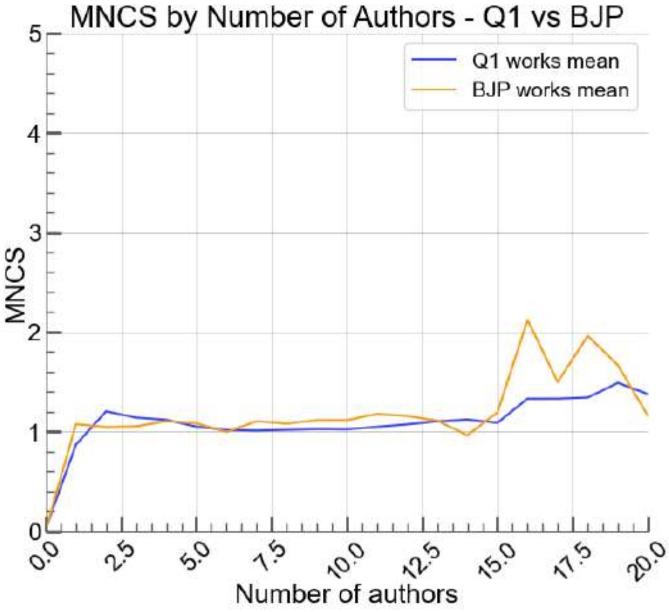


Figure 50: MNCS vs Number of Authors

Before moving on to formal tests and statistical analyses, we first explore the impact of the different metrics on citation counts through visualizations. These plots already suggest that the number of references (Figure 49) and the number of authors (Figure 50) may have a noticeable influence on popularity (MNCS)[53], especially for the number of references.

Similar plots were produced for the number of countries (Figure 51) and institutions (Figure 52), leading to the same observation: as the number of countries and institutions increases, the mean MNCS of the corresponding works also tends to rise. This pattern is particularly pronounced for BJP, although some variability and irregularities begin to emerge at higher values.

Finally, additional plots were made to compare works involving a single institution or country with those involving multiple institutions or countries (Figures 53, 54). These comparisons clearly highlight differences in popularity: in particular, between approximately 1995 and 2015, the gap in average citation counts is far from negligible, during which collaborative works received up to twice as many citations on average in certain years.

[50] x-axis : works count, y-axis: citations count. Centered within the top 5.

[51] In the figures, *Tetrahydrocannabinol_Delta* is represented in green, *Insulin_Pancreatic* in blue, and *Aorta_Endothelium* in purple.

[52] In the figures, *Plants_Medicinal* is represented in green, *Chemotherapy_Cancer* in blue, and *Ethanol_Alcohol* in purple.

[53] For visualization purposes, the plots are limited to a maximum of 85 references and 20 authors, beyond which the variability becomes too large.

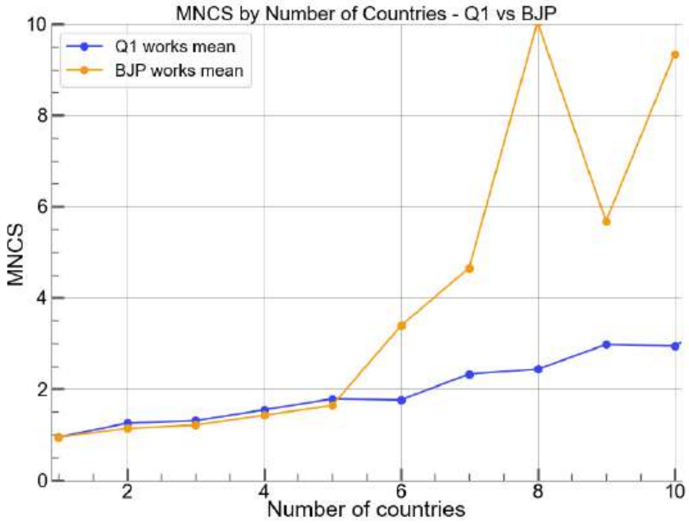


Figure 51: MNCS vs Number of Countries

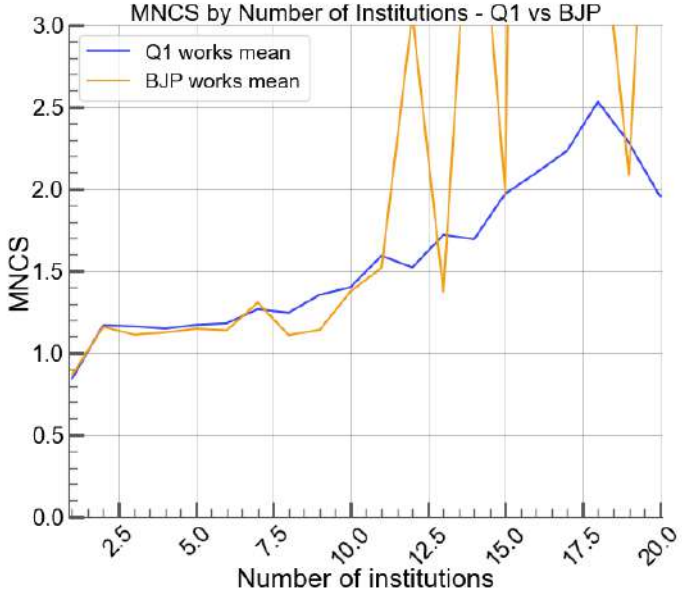


Figure 52: MNCS vs Number of Institutions

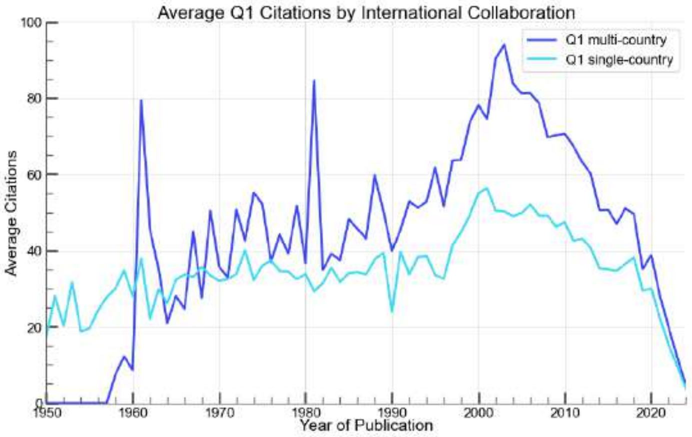


Figure 53: Citations by International Collaboration

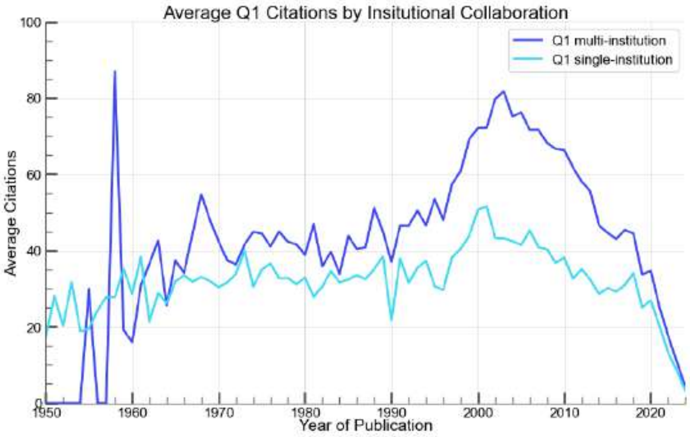


Figure 54: Citations by Institutional Collaboration

The statistical tests (Table 14) show that all considered features are positively associated with citation counts, although with varying strengths. In particular, the number of references exhibits the strongest relationship, especially when measured using Spearman's correlation, suggesting a clear monotonic association. The number of authors, countries, and institutions displays more moderate correlations, indicating that collaborative and structural factors still contribute to citation impact, albeit to a lesser extent. Overall, these results are consistent with the patterns observed in the visual analyses[54].

Table 14: Correlation coefficients between bibliometric features and citation counts

Feature	**Pearson** r	**Spearman** ρ
Number of references	0.238	0.557
Number of authors	0.045	0.279
Number of countries	0.061	0.148
Number of institutions	0.061	0.252

[54] All reported correlations are statistically significant, with p-values extremely close to 0.000.

4.4.2 Modelling by Decade

We now turn to the decade-based modeling section to identify which primary topics, modeled topics, or, more generally, which features have a positive impact on MNCS.

Before focusing on the coefficients, the percentage of R^2 explained by the models for each decade is examined to assess how effectively the models were able to identify the importance of the features.

As illustrated in Figure 55, the R^2 values are high in the early decades due to the small number of publications. They decrease in the 1970s–1990s and 2020s (around 13–15%), while still capturing meaningful patterns. In contrast, R^2 is higher in the 2000s–2010s (approximately 25–30%), reflecting stronger trends and topic ”hypes”.

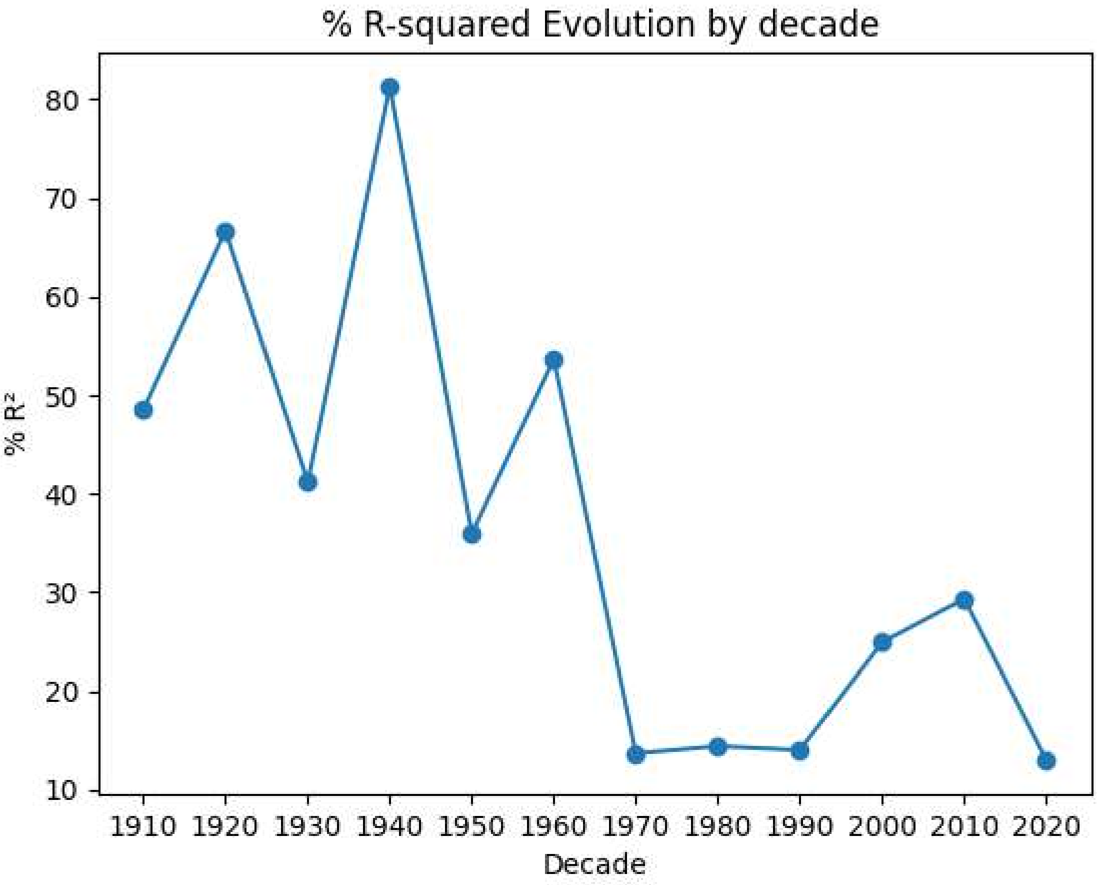


Figure 55: % R^2 by Decade

(See Appendix L) Overall, the estimated coefficients[55] (See Appendix M) reveal consistent patterns across decades. Indeed, the regression results show that both topic-related and non-topic features contribute to variations in MNCS across decades, although their roles differ in nature.

Several topics display large coefficients within specific periods : For instance, in the 1950s and the 1980s, the topic *curves_dose* appears among the highest positive coefficients. In the 2000s, *Advanced Text Analysis Techniques* and *Delphi Technique in Research* are among the top positive coefficients, while in the 2010s, *Reliability and Maintenance Optimization* and *Plant responses to water stress* also appear with high positive coefficients.

[55] Regression tables were obtained for all decades from the 1910s to the 2020s; however, only four representative tables are displayed for readability.

Non-topic features also show recurrent patterns across decades. The variable *mean_past_mncs_institutions* consistently presents a strong positive coefficient, while collaboration-related variables, such as the number of distinct countries and institutions, also display positive coefficients in multiple decades. In contrast, variables related to past volume, such as past contributions at the topic or institutional level, often show small or negative coefficients. However, suprisingly[56], the variable *mean_past_mncs_topic* appears with negative coefficients in several decades[57].

4.4.3 General Modelling

Before applying the machine learning models, the correlation matrix (See Appendix N) already allows observing links between certain features and the MNCS.

The results from the decade-by-decade modelling are broadly consistent, showing the strong influence of *mean_past_mncs_institutions* and the relatively low impact of diversity metrics (number of countries, institutions, and authors). Some changes are nevertheless observable, including a considerably higher importance of *referenced_works_count* and a more neutral role for *mean_past_mncs_topic*, where the effect noted previously appears to balance out.

For the general modelling part, We first compare the performance of the different models[58], as summarized in Table 15. Overall, ensemble methods such as XGBoost and LightGBM achieve the highest Quadratic Weighted Kappa (QWK) and lowest RMSE, while simpler models like Logistic Regression and GaussianNB show lower predictive accuracy. These results suggest that gradient boosting techniques better capture the patterns in the data compared to linear or probabilistic models.

Table 15: ML Models Performance Metrics

Model	**Accuracy**	**QWK**	**RMSE**
Logistic Regression	0.4642	0.5471	1.2207
Random Forest	0.5326	0.6505	1.0687
XGBoost	**0.5412**	**0.6657**	**1.0555**
LightGBM	0.5406	0.6614	1.0627
LDA	0.4180	0.4698	1.2726
kNN	0.4446	0.4698	1.2726
GaussianNB	0.3648	0.5356	1.2256
MLP	0.5314	0.4290	1.2628

We will now examine and analyze in detail the performance of the XGBoost model :

First, we visualize the predictions of the XGBoost model (Figure 56) against the true classes of the publications (Figure 57). Some variations in the colors[59] for classes 2 and 4 are already visible, indicating that the model is less precise in predicting publications of medium and higher popularity.

[56] The negative coefficient reflects a regression-to-the-mean effect. Topics with historically high average MNCS values are already positioned above the field average, which mechanically reduces the remaining variability available for new articles to exceed this benchmark. As a result, once topic-specific effects are controlled for, higher past topic-level MNCS is associated with a lower marginal contribution to article-level impact.

[57] Although *mean_past_mncs_topic* is statistically significant with a p-value close to zero, this significance is mainly driven by the large number of observations. Its standard error is higher (around 0.3) than that of most other features, which present near-zero standard errors.

[58] The models are explained in Appendix O

[59] These variations may be accentuated by light color shades.

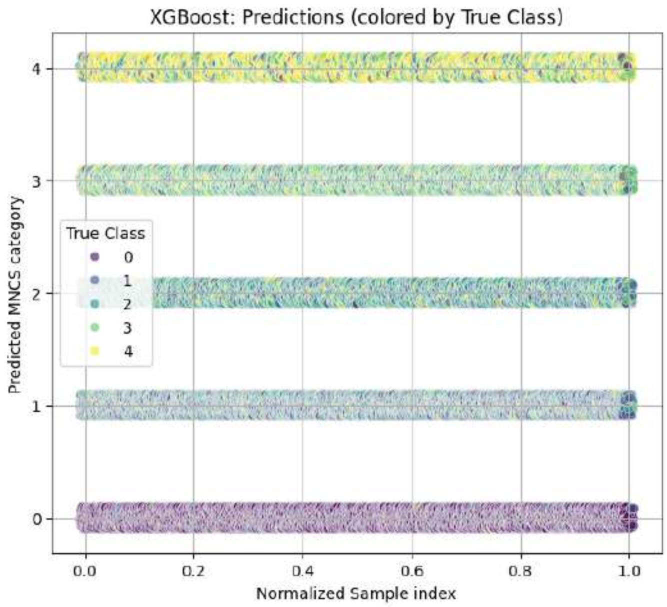


Figure 56: Predicted Classes by XGBoost Model

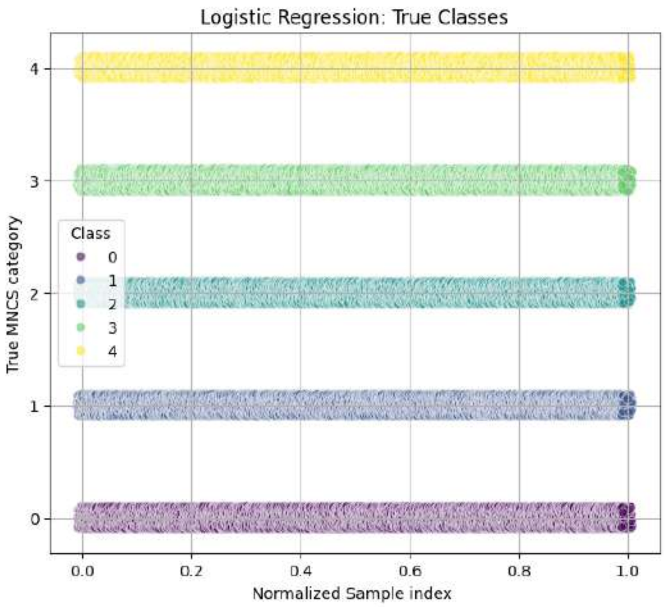


Figure 57: True Classes

These impressions are further supported by the detailed performance metrics (Table 16) and the confusion matrix (Figure 58), which confirm the model's lower accuracy for medium and highly cited publications. Indeed, The model shows high accuracy for the extreme classes, with 0.81 for class 0 and 0.61 for class 1, but performs poorly for intermediate and highly cited publications, achieving only 0.25 for class 2 (despite 0.67 spread across adjacent classes) and 0.37 for class 4 (similarly matched by 0.37 in class 3), confirming that medium and highly popular works are frequently misclassified into neighboring categories.

Class	Precision	Recall	F1-score	Support
0	0.85	0.81	0.83	34,481
1	0.45	0.61	0.52	34,489
2	0.42	0.25	0.32	27,613
3	0.42	0.48	0.45	27,507
4	0.55	0.37	0.45	13,783
Macro Avg	0.54	0.51	0.51	137,873
Weighted Avg	0.55	0.54	0.53	137,873

Accuracy = 0.54
RMSE = 1.0555
QWK = 0.6657

Table 16: XGBoost Detailed Performance Metrics

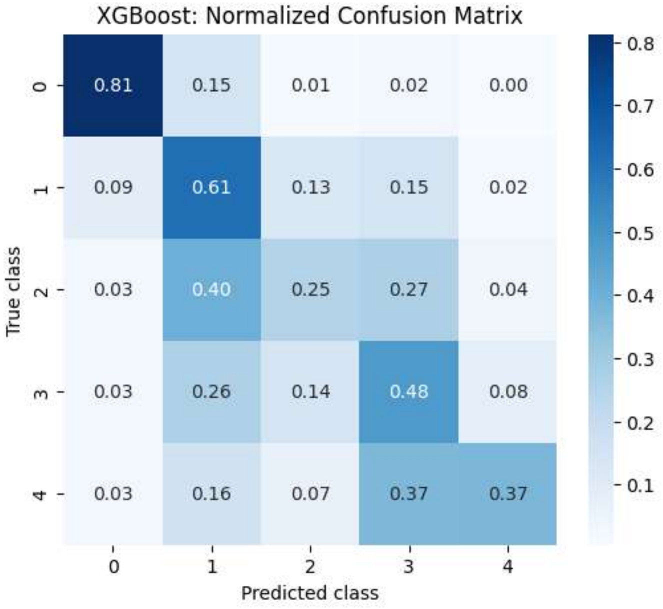


Figure 58: XGBoost Confusion Matrix

Next, the importance of features (Figure 59) in this model will be examined to identify which factors contribute most to its predictions :

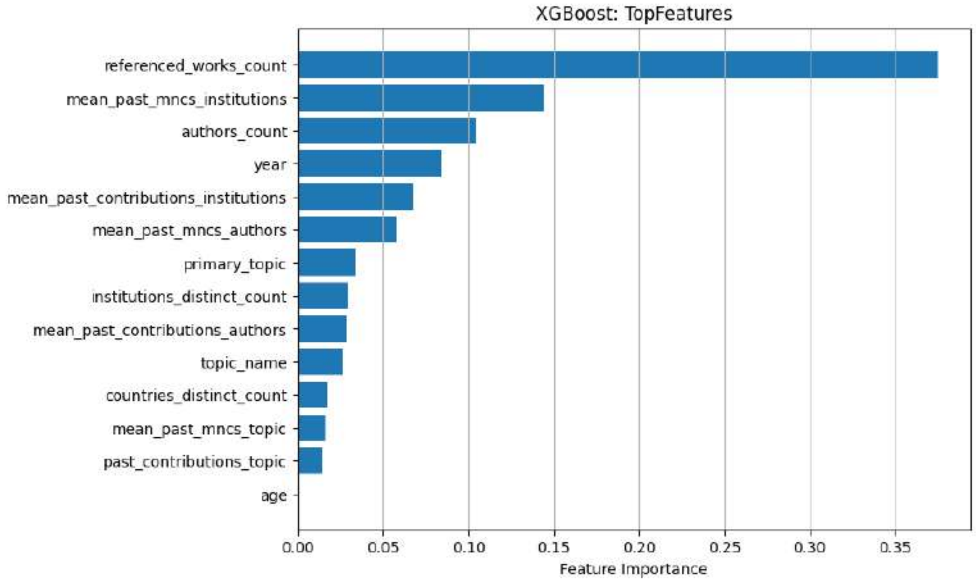


Figure 59: XGBoost Feature Importance

For the XGBoost model, the number of references was the most important feature in determining publication classes[60], while mean past MNCS of institutions played a much smaller role. It should be noted, however, that different models exhibit different feature importance profiles :

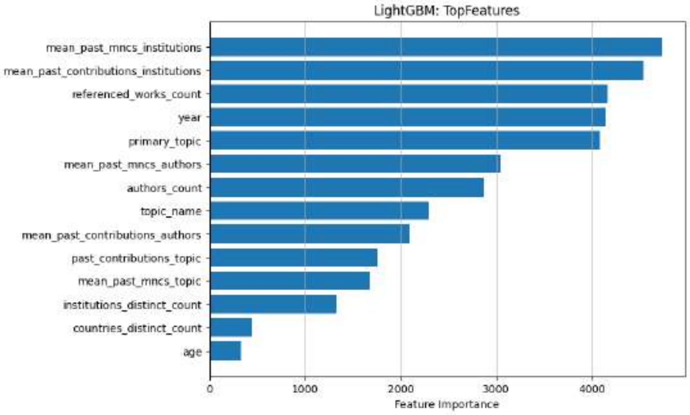


Figure 60: LightGBM Feature Importance

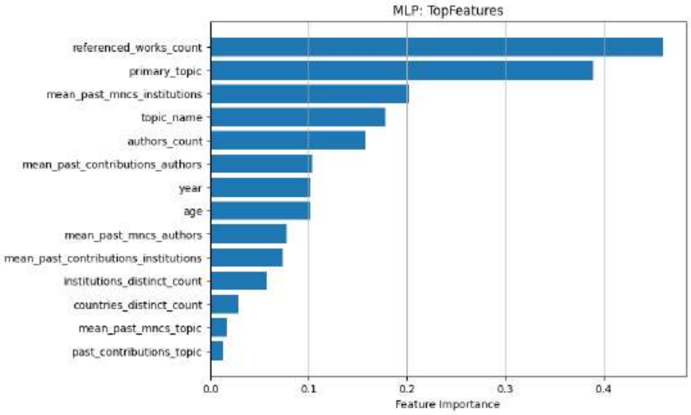


Figure 61: MLP Feature Importance

For instance, the LightGBM model (Figure 60) considers a larger set of features as important while achieving performance comparable to XGBoost, including mean past contributions of institutions, publication year (which is even more influential), and the primary topic. Similarly, the MLP (Figure 61) model not only highlights the number of referenced works but also identifies the primary topic as its top feature.

Overall, the modelling analysis demonstrates that both the performance of the models and the features they prioritize reveal meaningful relationships between publication characteristics and popularity, with the number of references being important, but particularly the mean MNCS of the institutions involved.

[60] This metric should be interpreted with caution, as the presence of meta-analyses and review articles, which are highly cited and contain many references, can inflate its apparent importance. The models were actually initially trained without these publications.

Conclusion

Overall, this work highlights key structural and thematic trends shaping pharmacological research, with particular attention to the British Journal of Pharmacology and other leading journals. BJP remains among the top Q1 journals in the field, with citations generally consistent across publications, although a recent decline in publication rate may affect its visibility and citation impact. Most contributions originate from British and European authors and institutions, and while the journal globally covers popular topics similar to other top journals, it places particular emphasis on certain subfields.

Globally, pharmacology publications have increased in number over time, reaching wider audiences and achieving greater citation and sharing rates. Authorship has become more diverse, both in terms of countries and institutions, and research topics have broadened accordingly. The number of authors, countries, and institutions involved in a publication, as well as the level of international collaboration, all have a non-negligible impact on citation counts. This is illustrated by the example of certain Asian countries, which tend to collaborate less with other countries and consequently receive fewer citations. Citation prediction models confirm the presence of measurable relationships, performing significantly better than random. In particular, indicators such as the mean past MNCS of institutions, as well as the number of references, emerge as highly significant predictors of future citation impact. Topic popularity also shifts over decades, highlighting the evolving interests within the field.

However, This work has several limitations. The dataset, although large, does not capture all publications in pharmacology, and metadata is not always complete for every work. Moreover, the concept of *impact* used in this study, while informative, remains subjective, and alternative measures may provide complementary insights.

Future research could also broaden journal coverage and improve metadata to capture emerging trends more fully. More advanced topic modeling and network analyses could provide deeper insights into the evolution of subfields and the role of collaboration in shaping scientific impact. Developing predictive citation tools and more interactive visualizations would provide actionable insights for researchers and institutions, helping to anticipate and shape the future of pharmacology research.

Appendix A : Topic Modelling Similarity

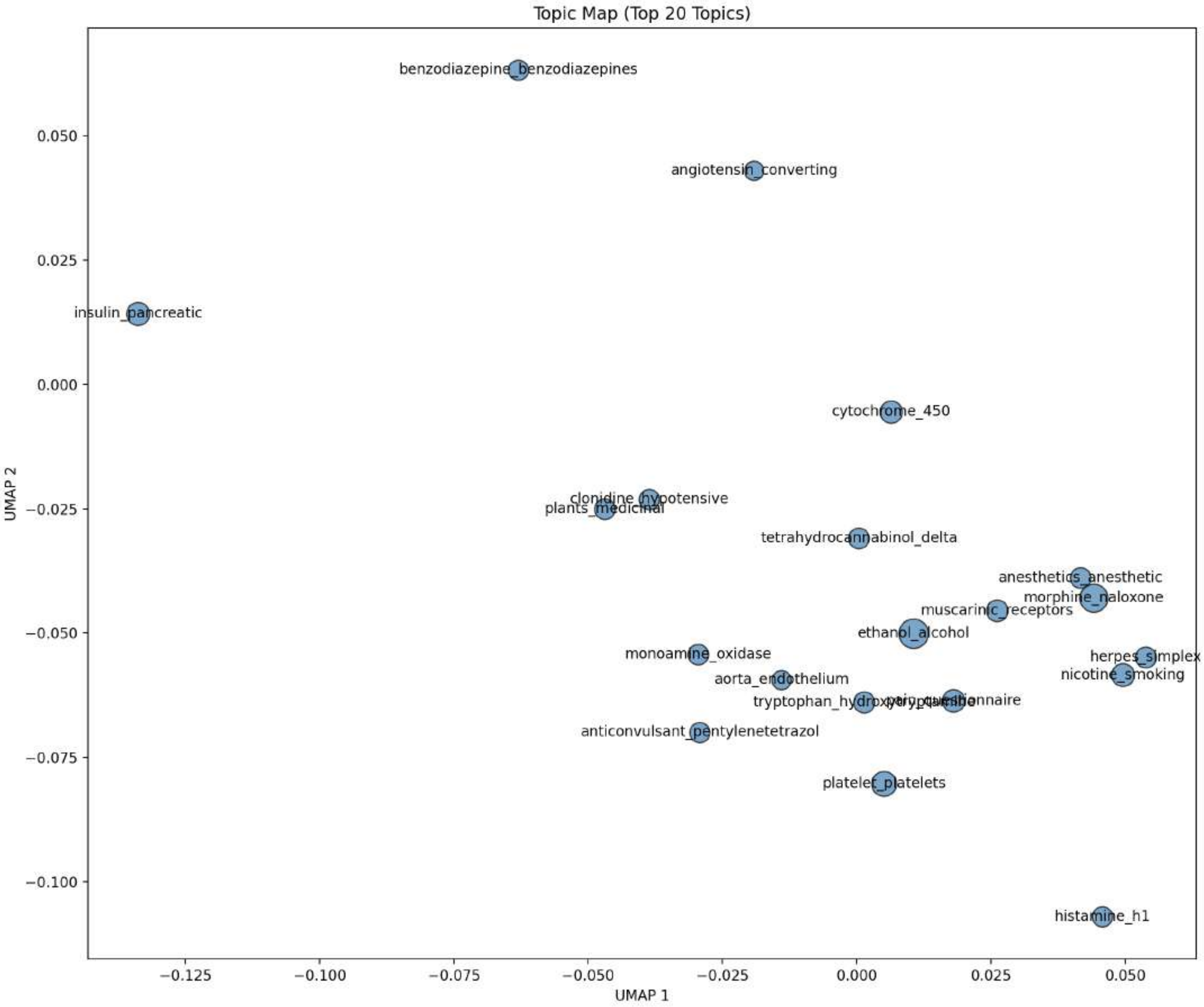


Figure 62: Top 20 Published Modelled Topics Similarity Map

Appendix B : Citation-Related Age Analysis

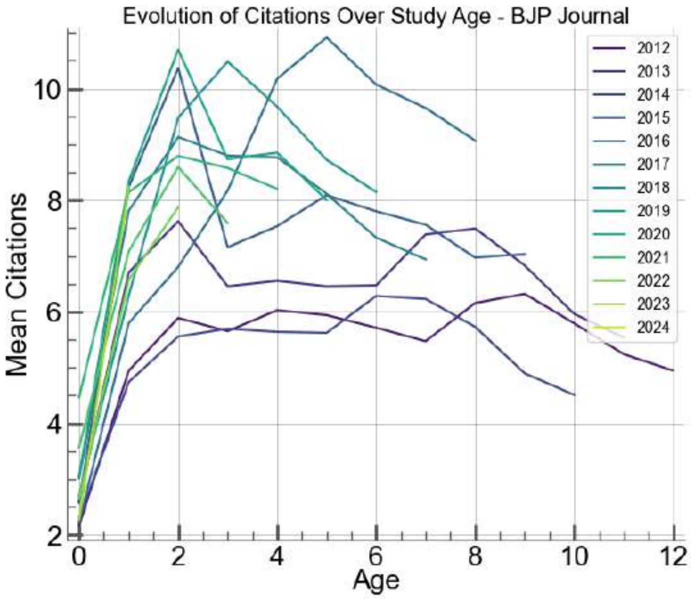


Figure 63: Evolution of Citations over BJP Works Age

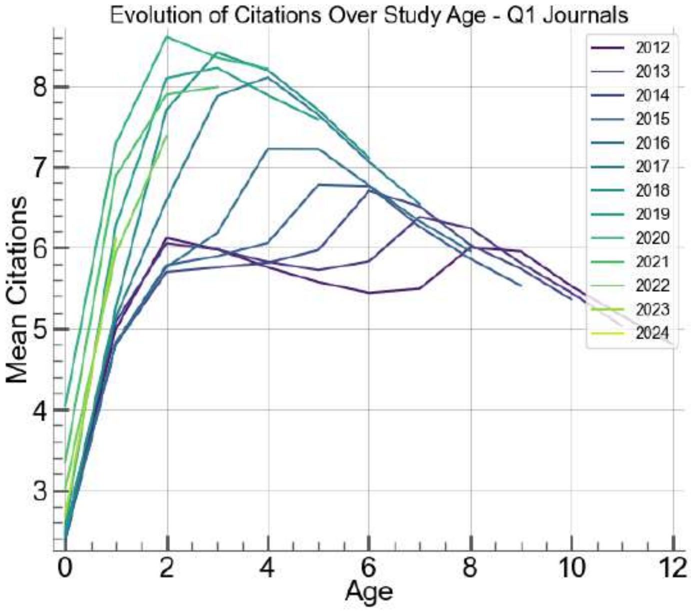


Figure 64: Evolution of Citations over Q1 Works Age

Appendix C : Top 20 Authors

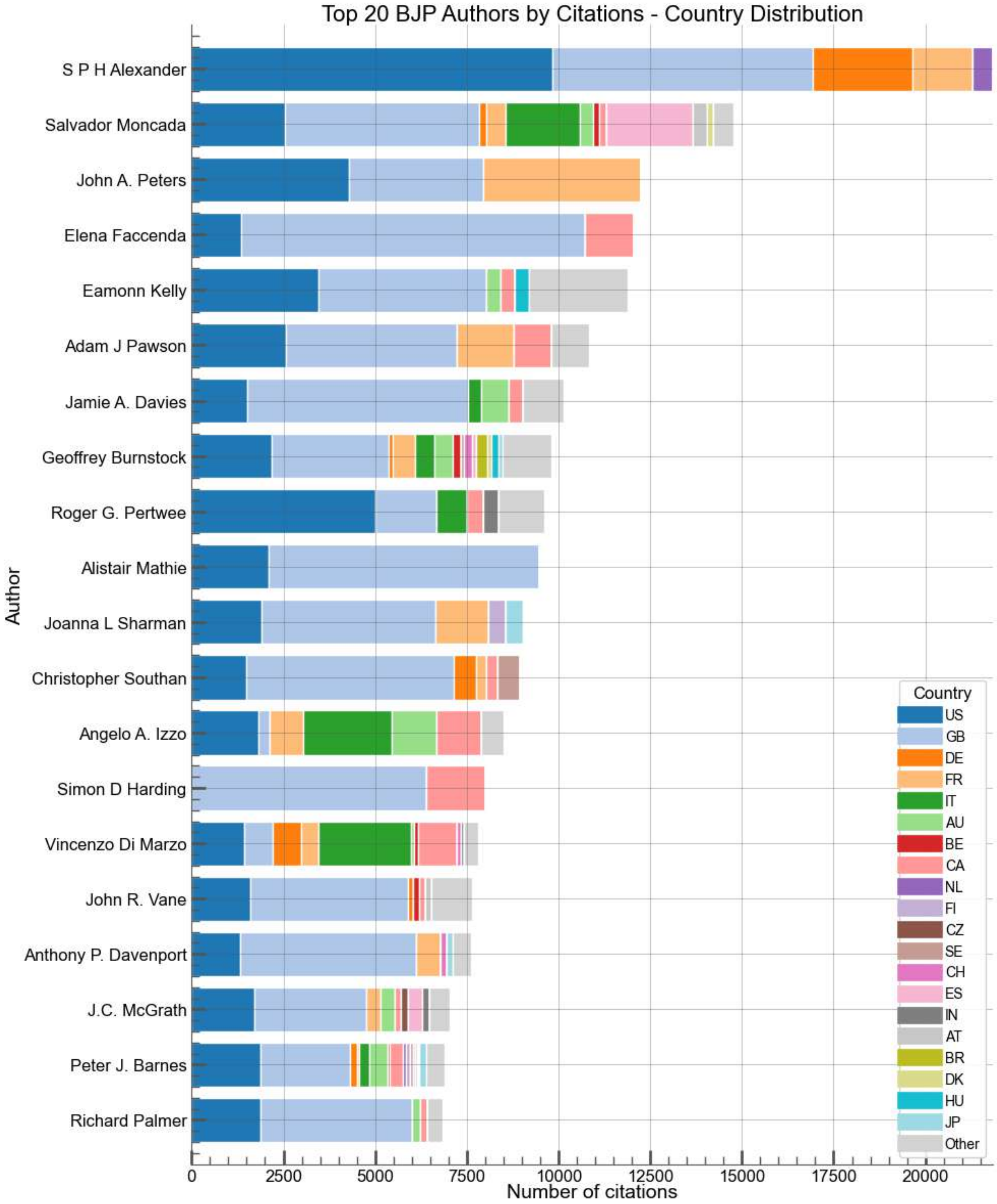


Figure 65: Top 20 BJP Authors

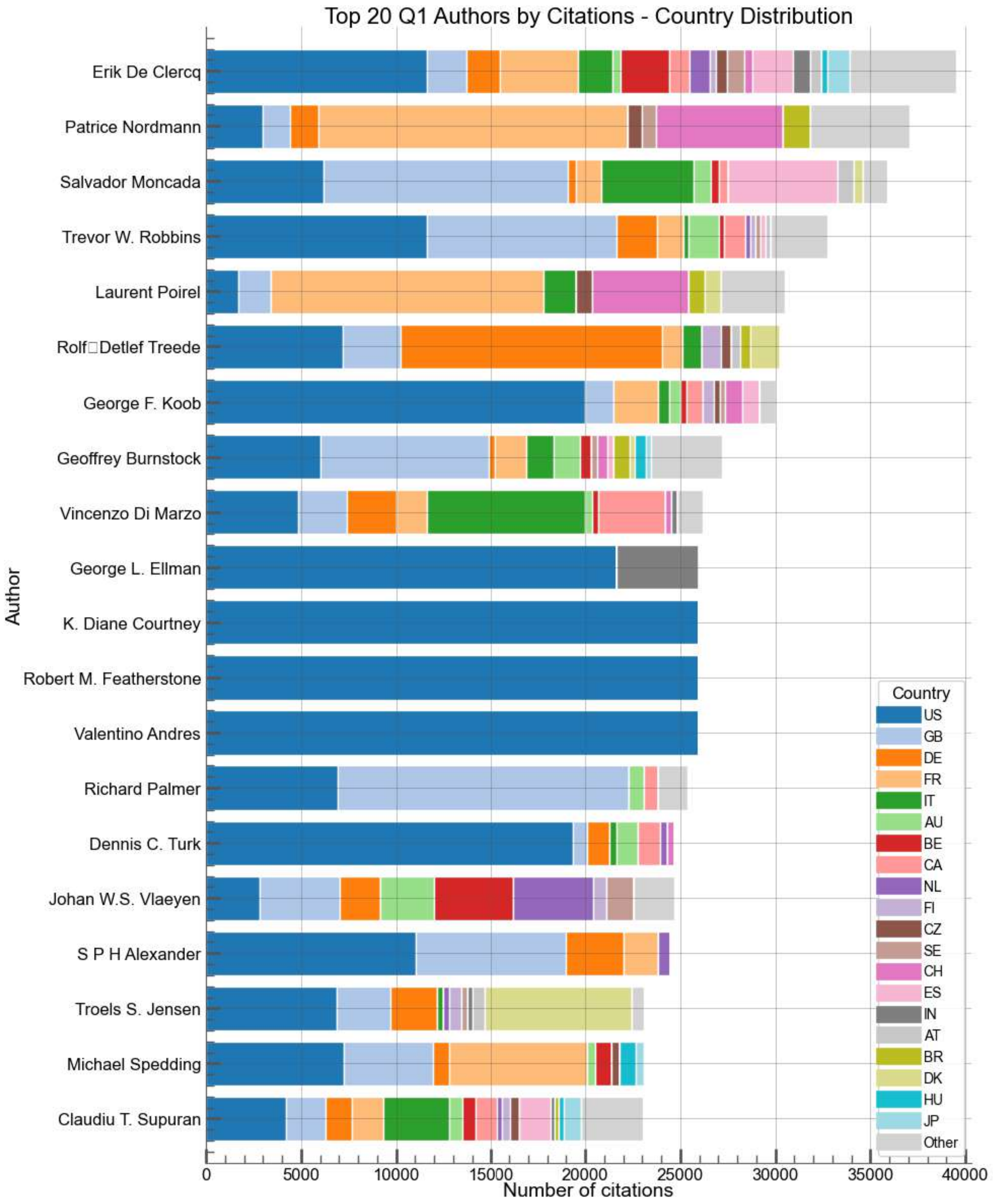


Figure 66: Top 20 Q1 Authors

Appendix D : Top Authors Collaboration

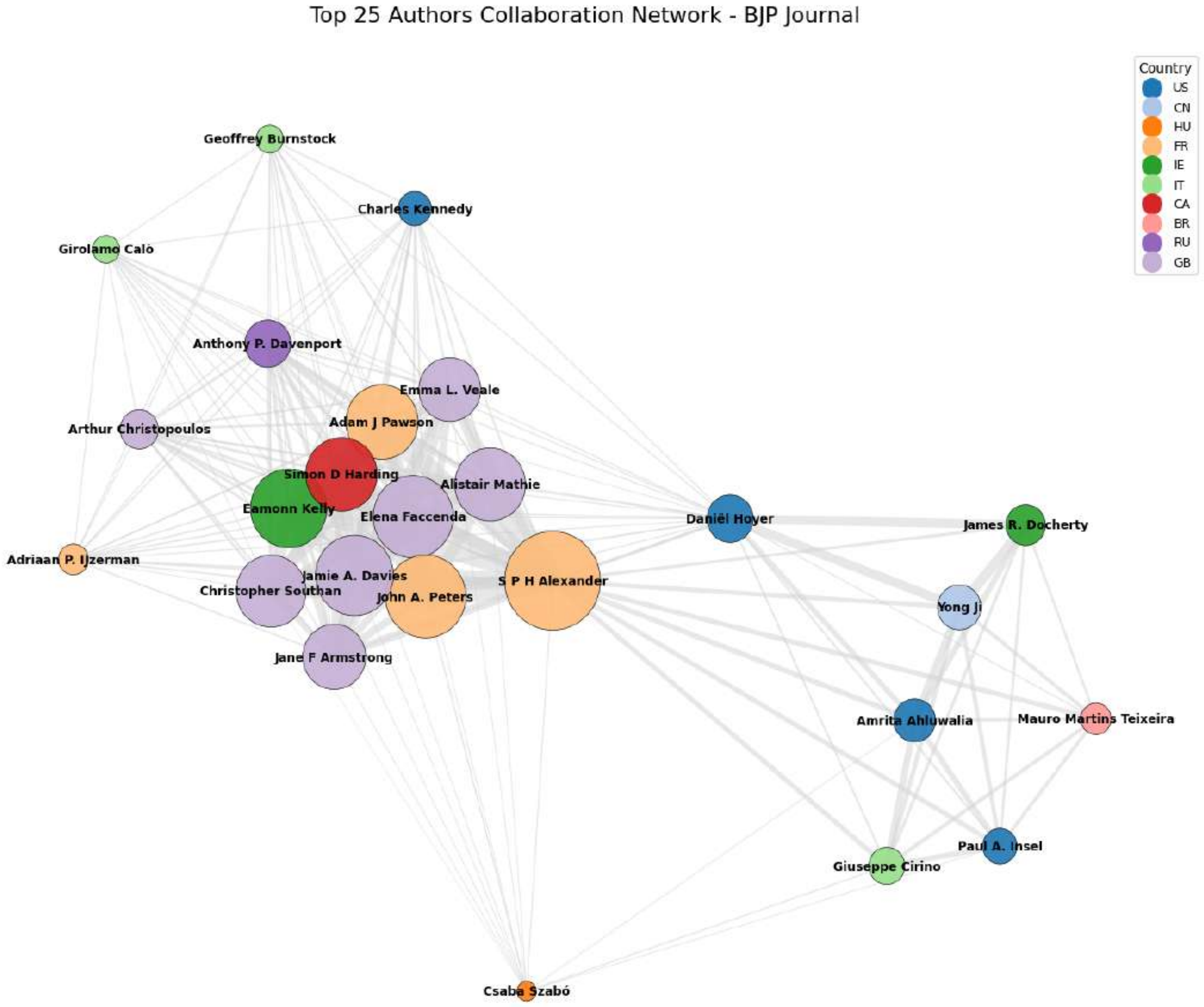


Figure 67: Top 25 BJP Authors Collaboration Graph

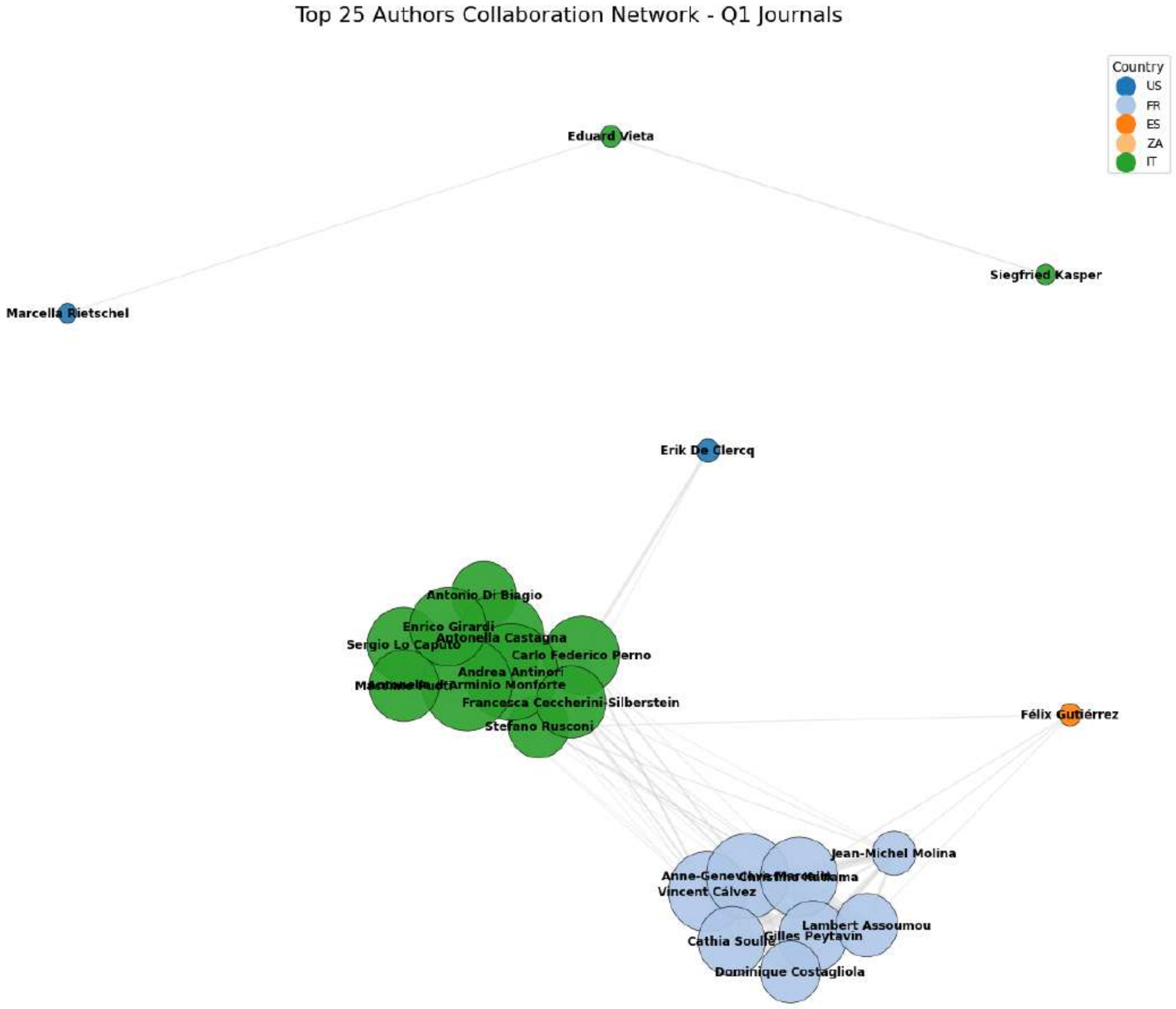


Figure 68: Top 25 Q1 Authors Collaboration Graph

Appendix E : Countries – MNCS vs Publications

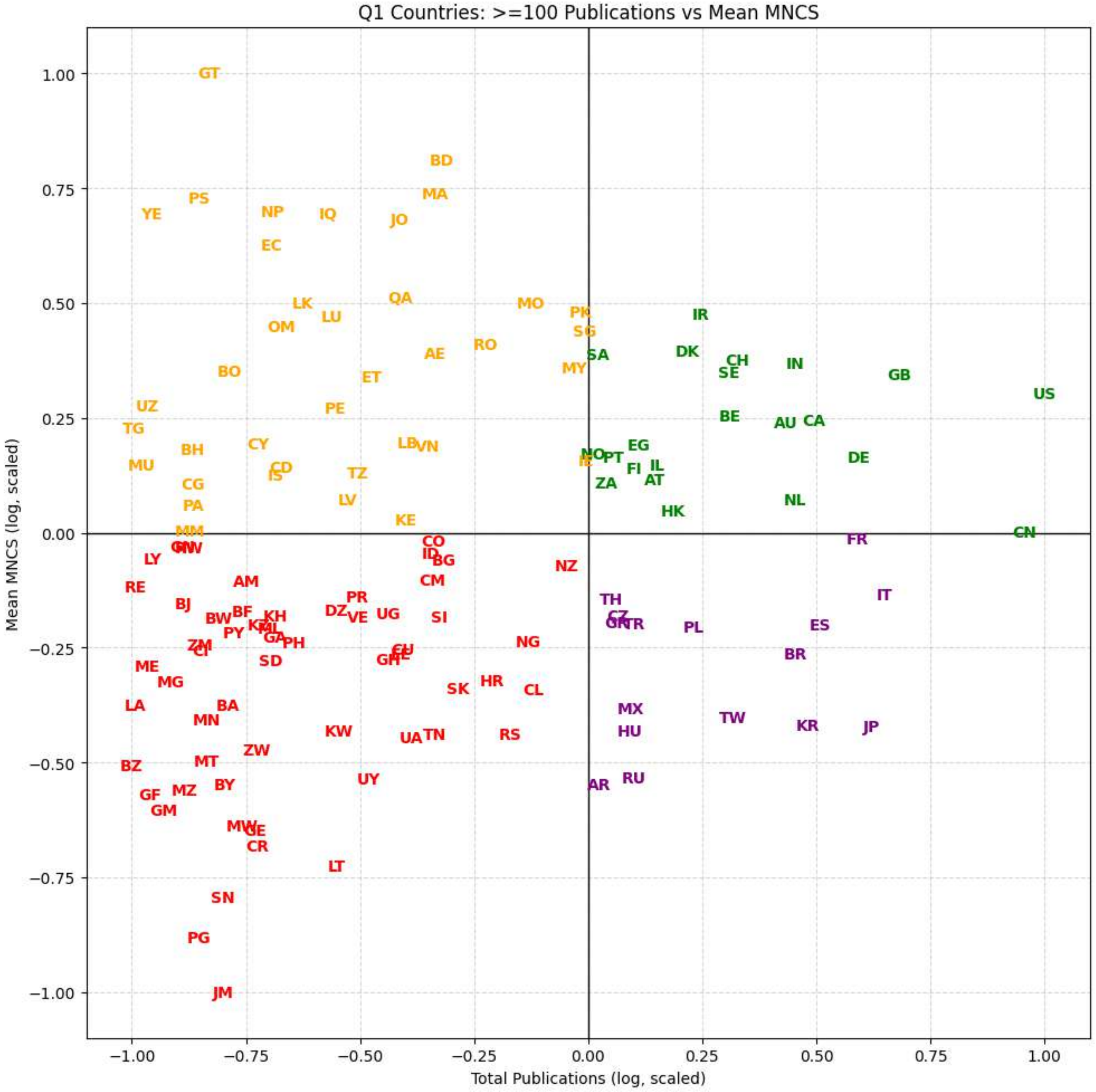


Figure 69: Mean MNCS vs. Publication Count for Q1 Countries

Appendix F : Top Countries Collaboration

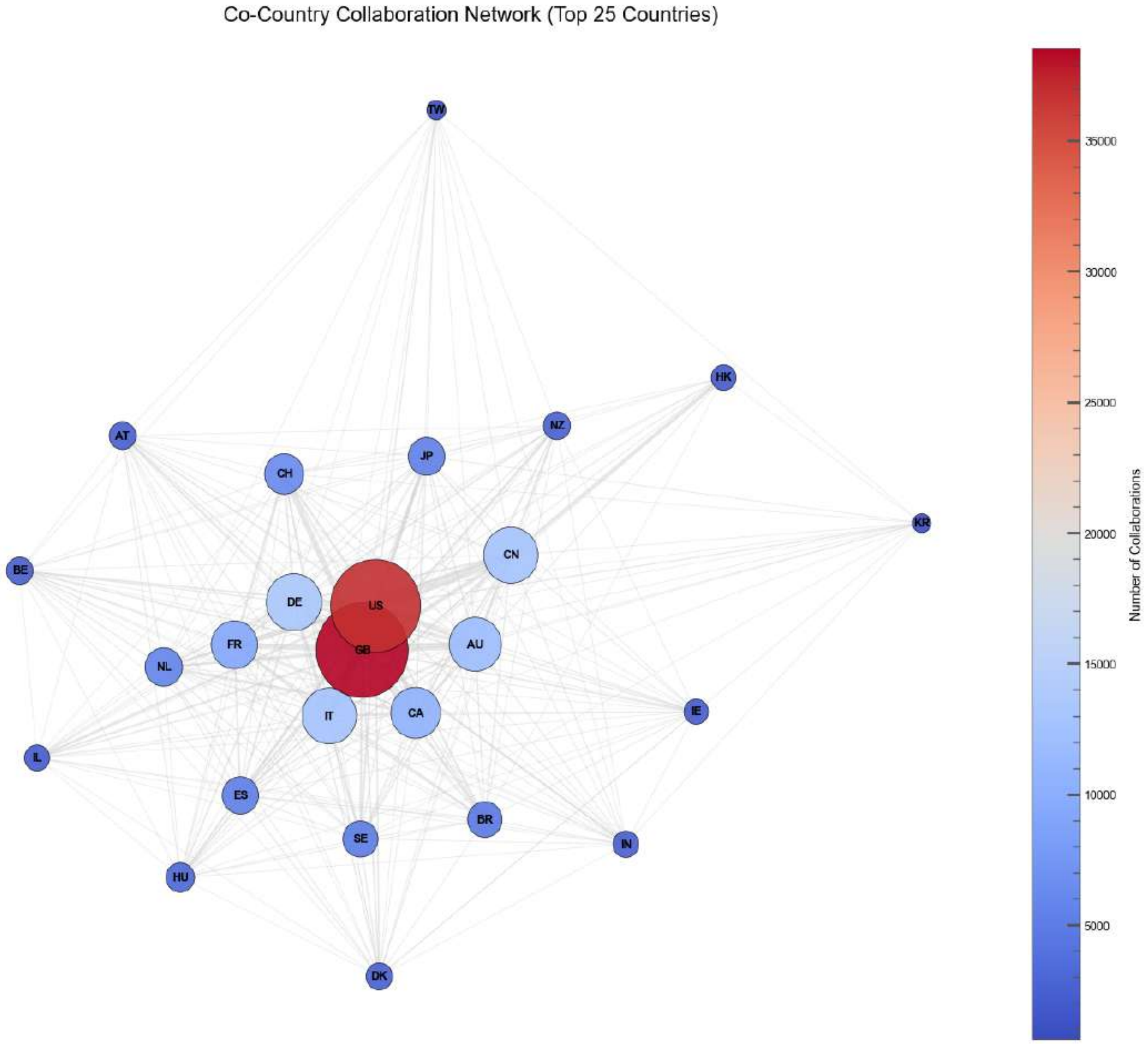


Figure 70: BJP Top 25 Countries Collaboration Graph

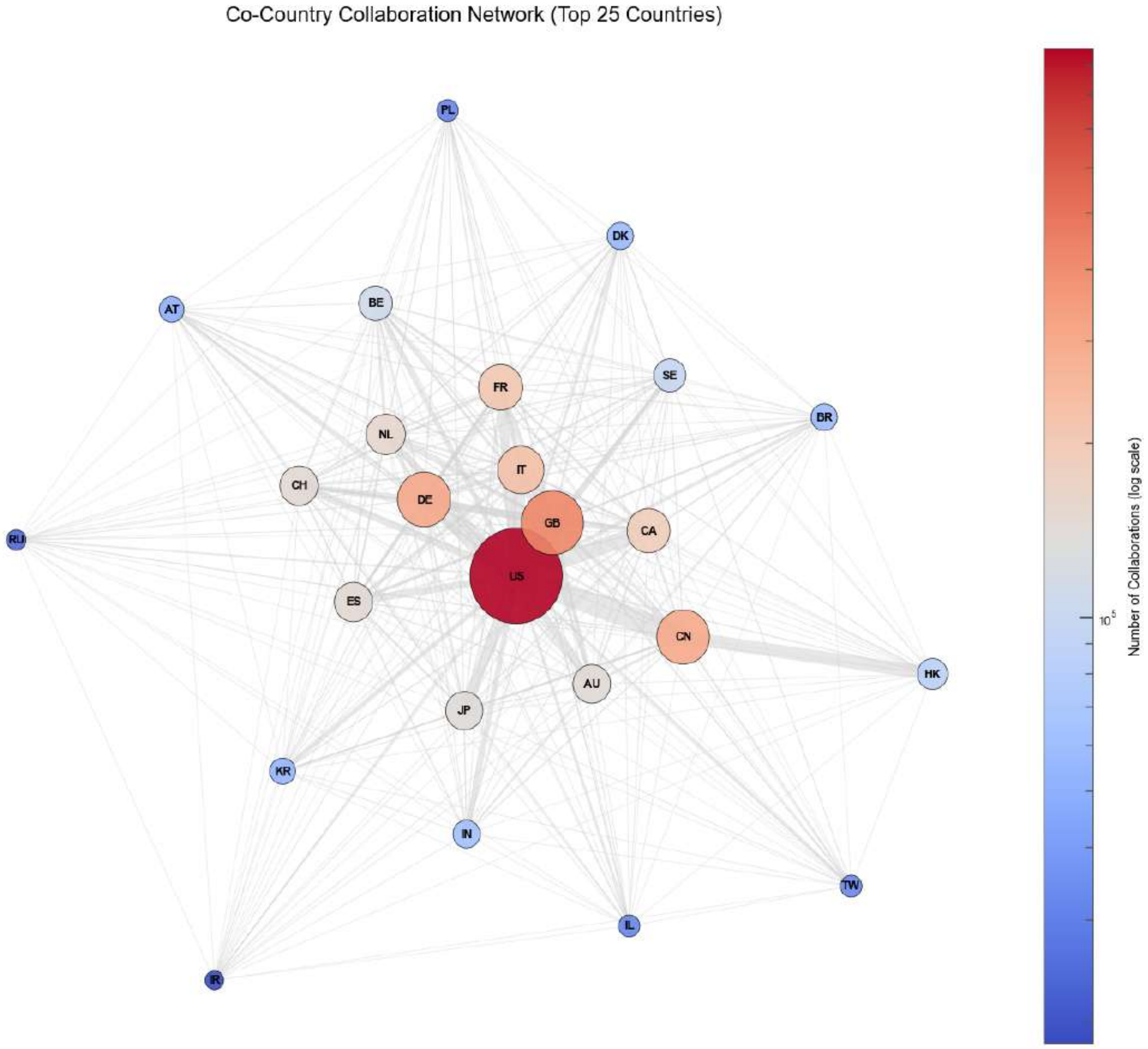


Figure 71: Q1 Top 25 Countries Collaboration Graph

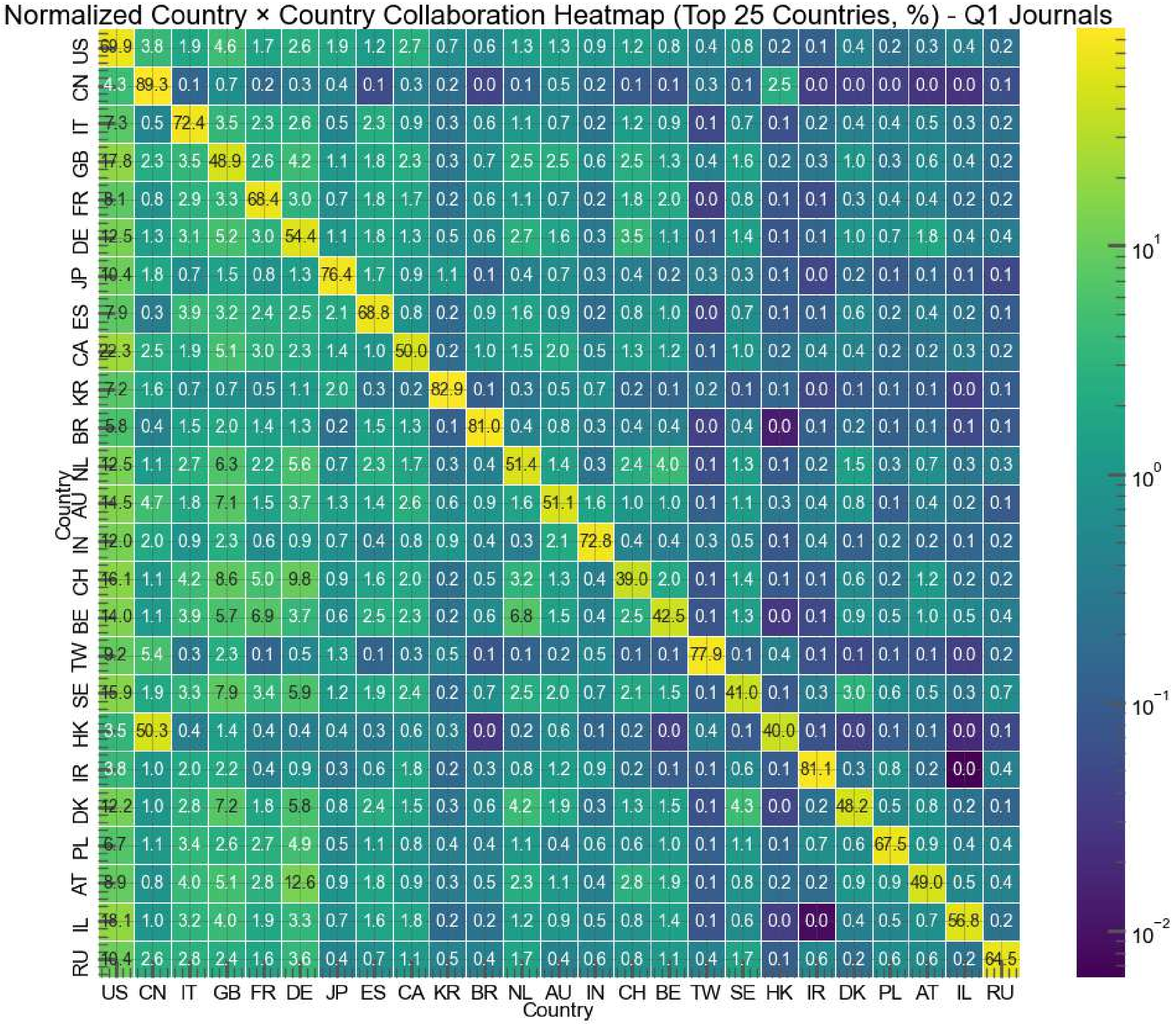


Figure 72: Q1 Top 25 Countries Collaboration Matrix

Appendix G : Institutions – MNCS vs Publications

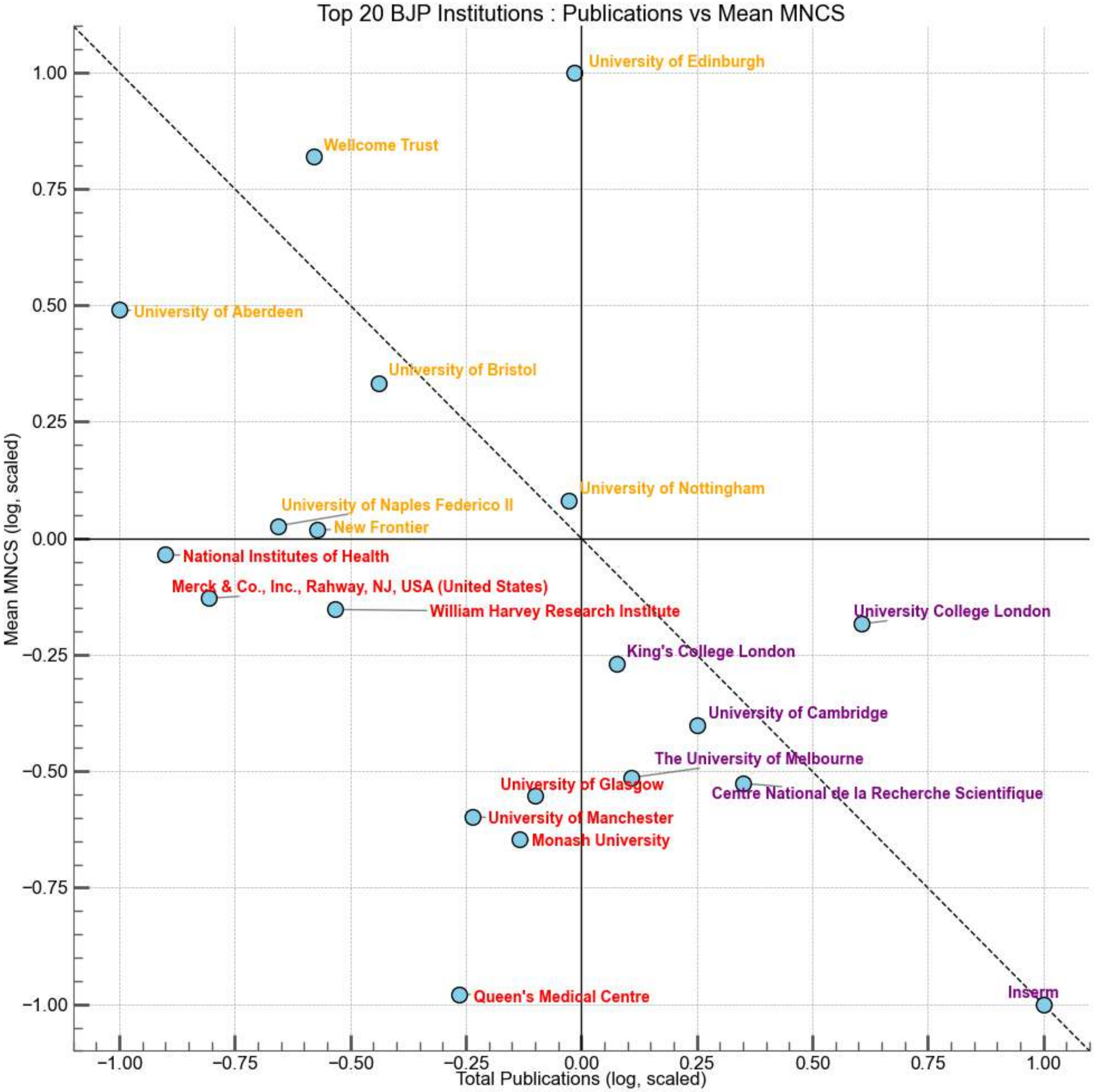


Figure 73: Mean MNCS vs. Publication Count for Top BJP Institutions

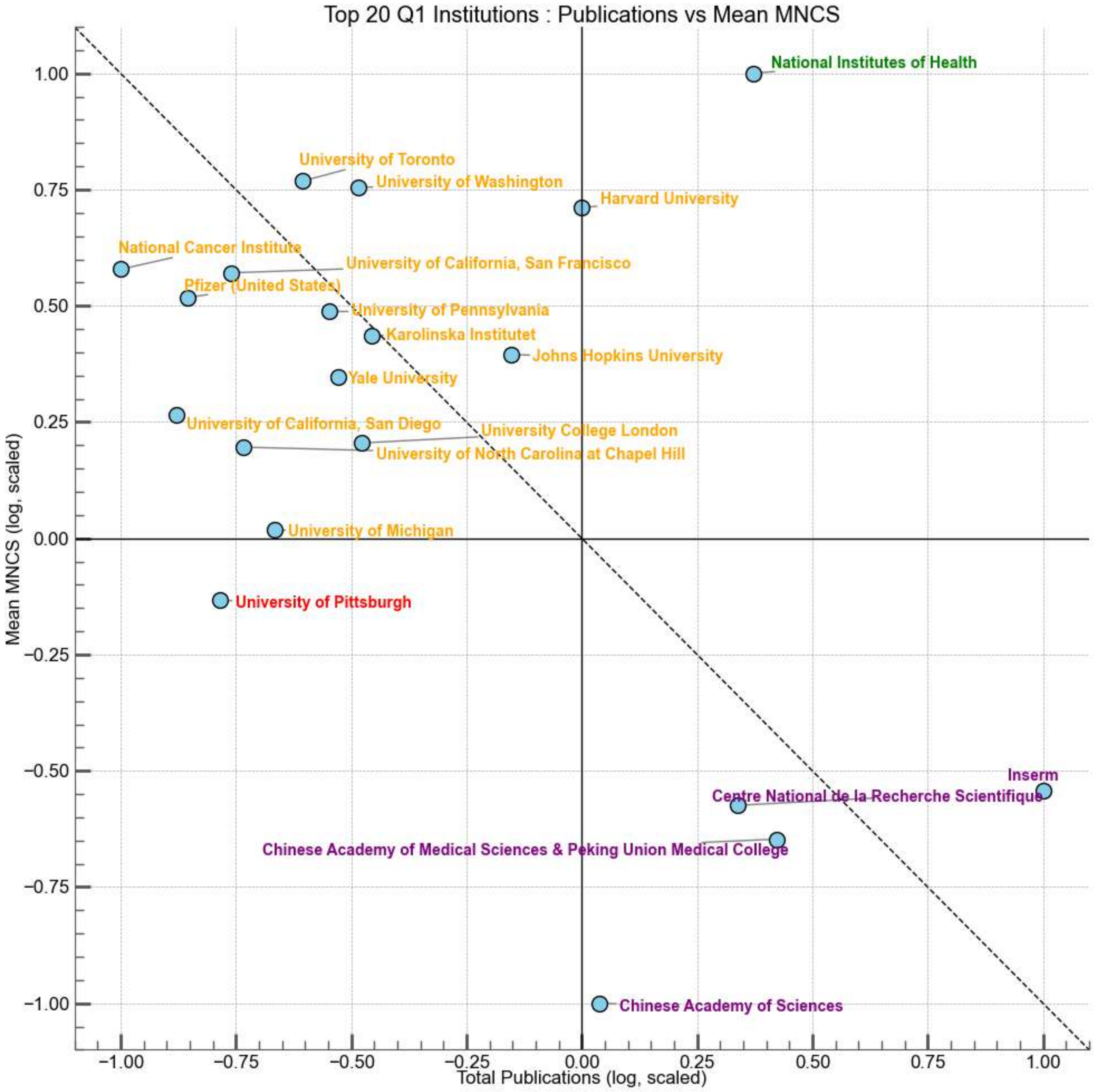


Figure 74: Mean MNCS vs. Publication Count for Top Q1 Institutions

Appendix H : Top Institutions Collaboration

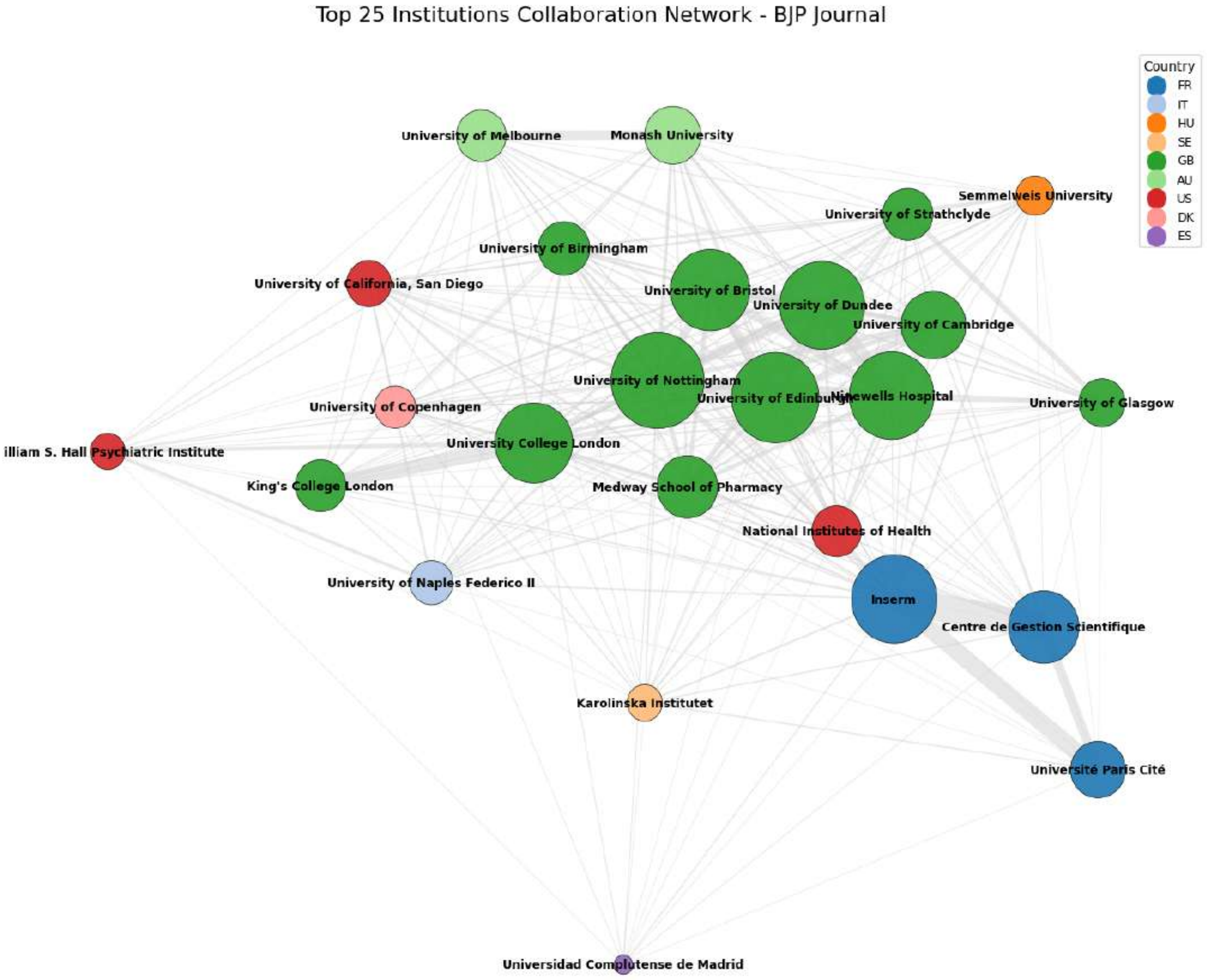


Figure 75: BJP Top 25 Institutions Collaboration Graph

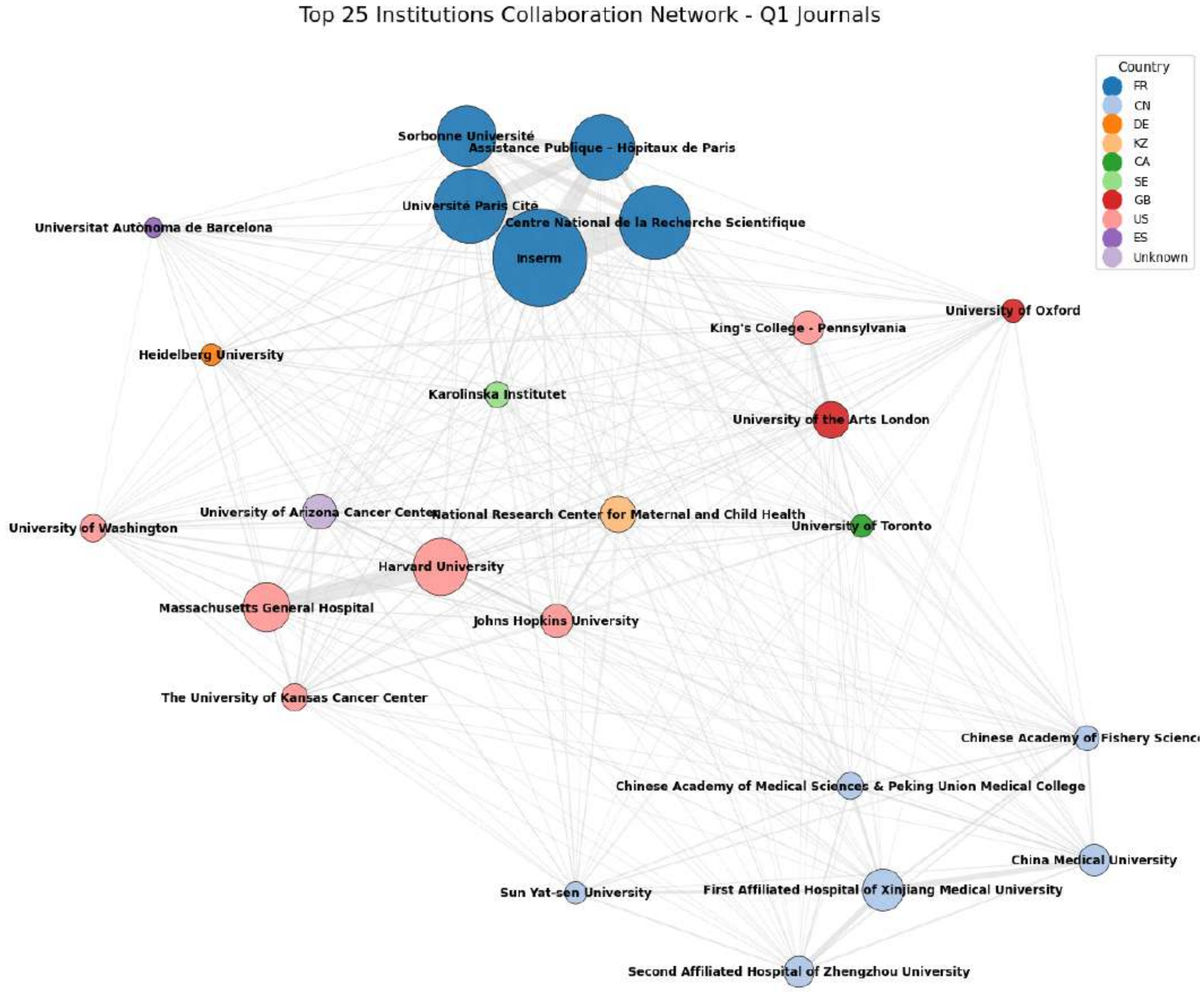


Figure 76: Q1 Top 25 Institutions Collaboration Graph

Appendix I : Concepts and Keywords

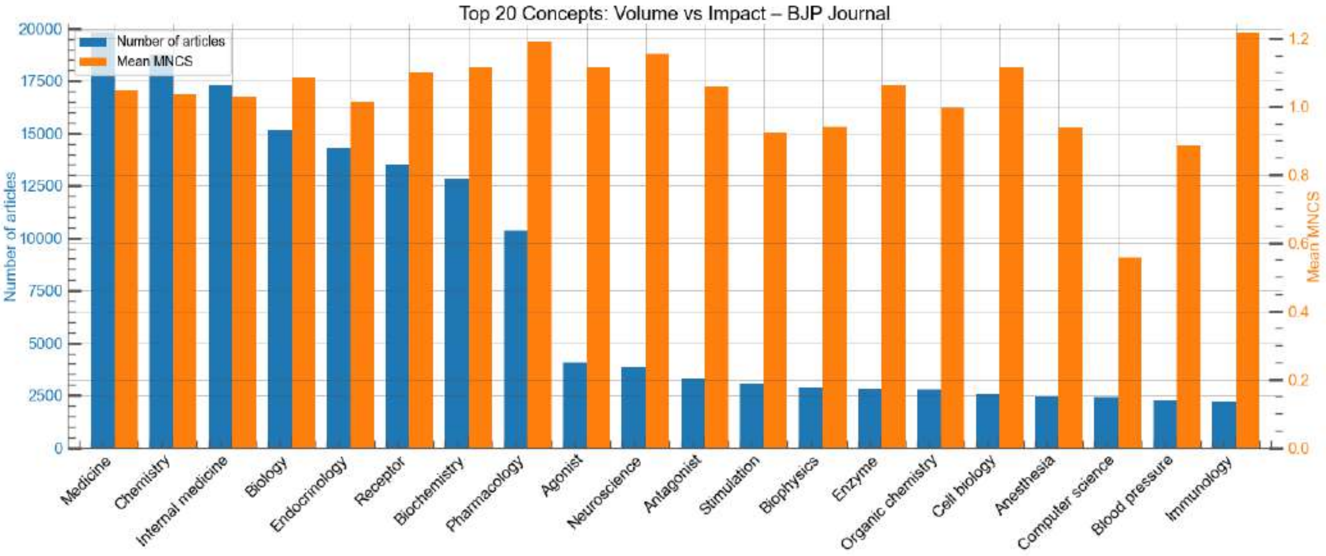


Figure 77: Mean MNCS vs Number of Articles - Top 20 BJP Concepts

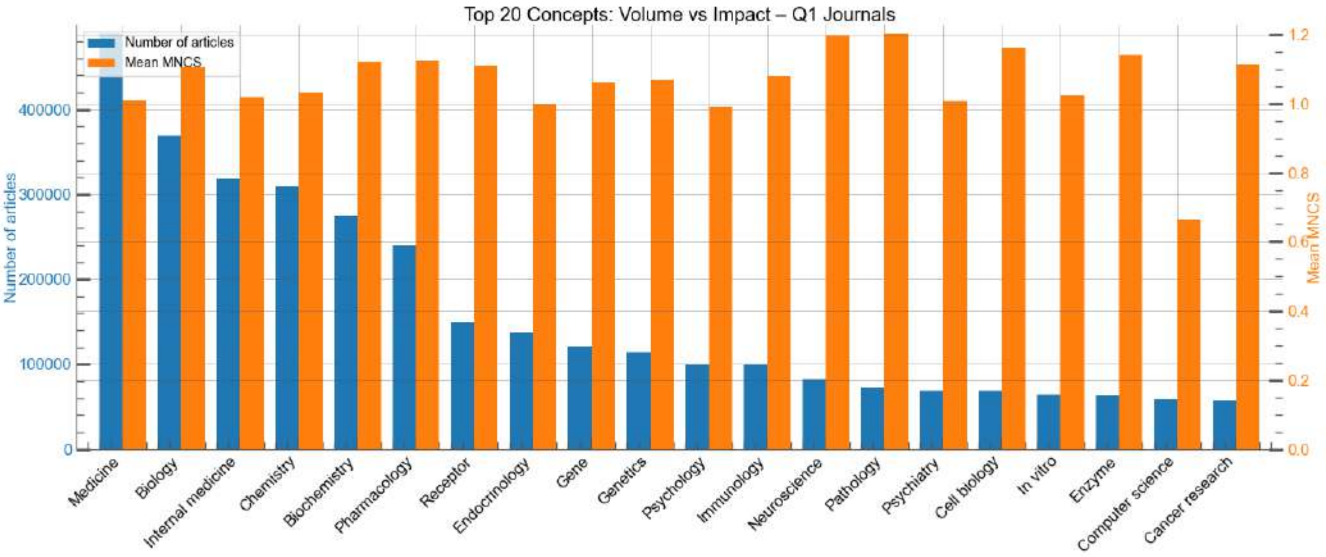


Figure 78: Mean MNCS vs Number of Articles - Top 20 Q1 Concepts

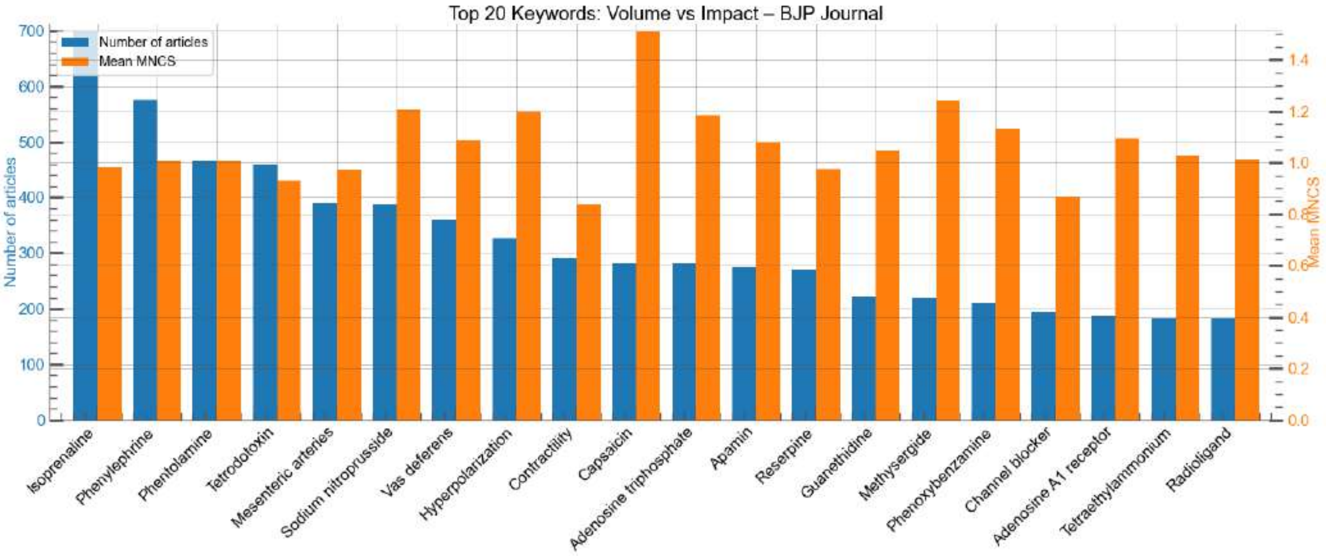


Figure 79: Mean MNCS vs Number of Articles - Top 20 BJP Keywords

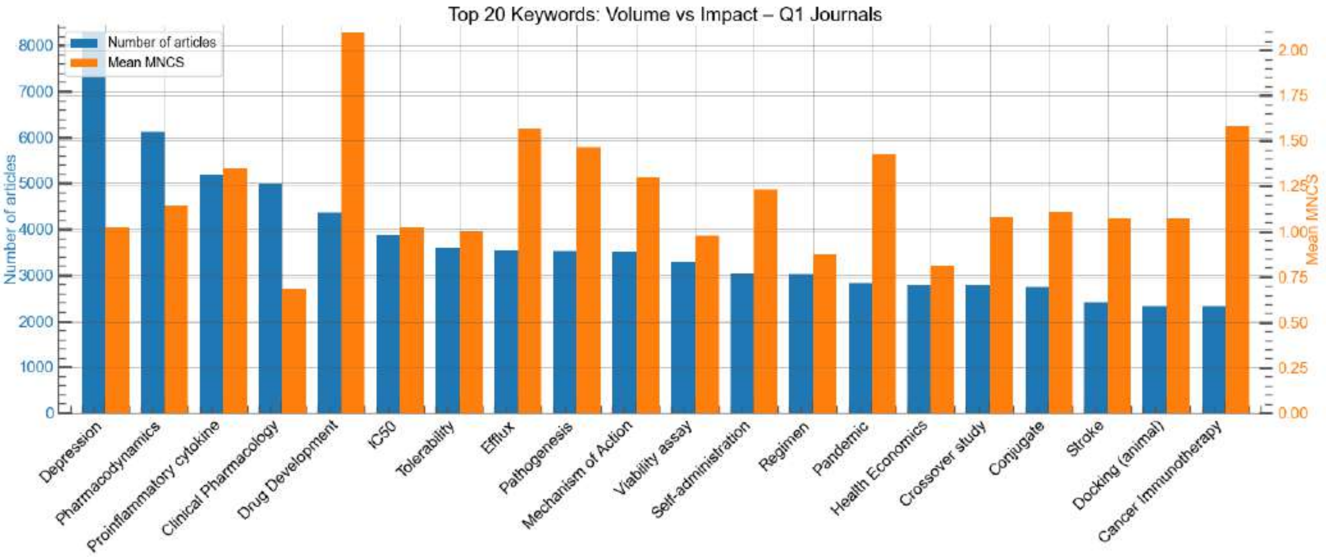


Figure 80: Mean MNCS vs Number of Articles - Top 20 Q1 Keywords

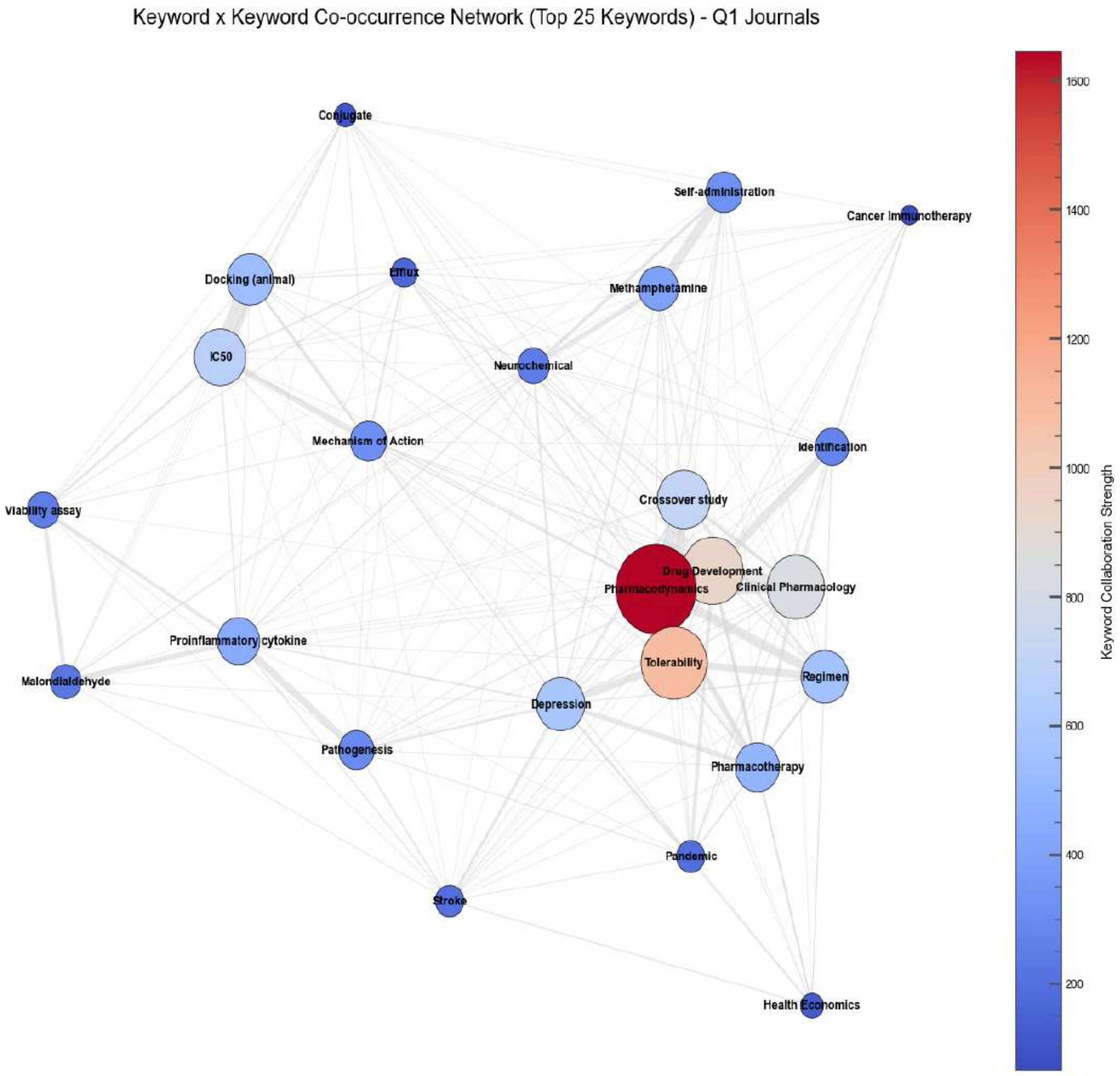


Figure 81: Q1 Top 25 Keywords Co-Occurence Graph

Appendix J : Top Primary Topics

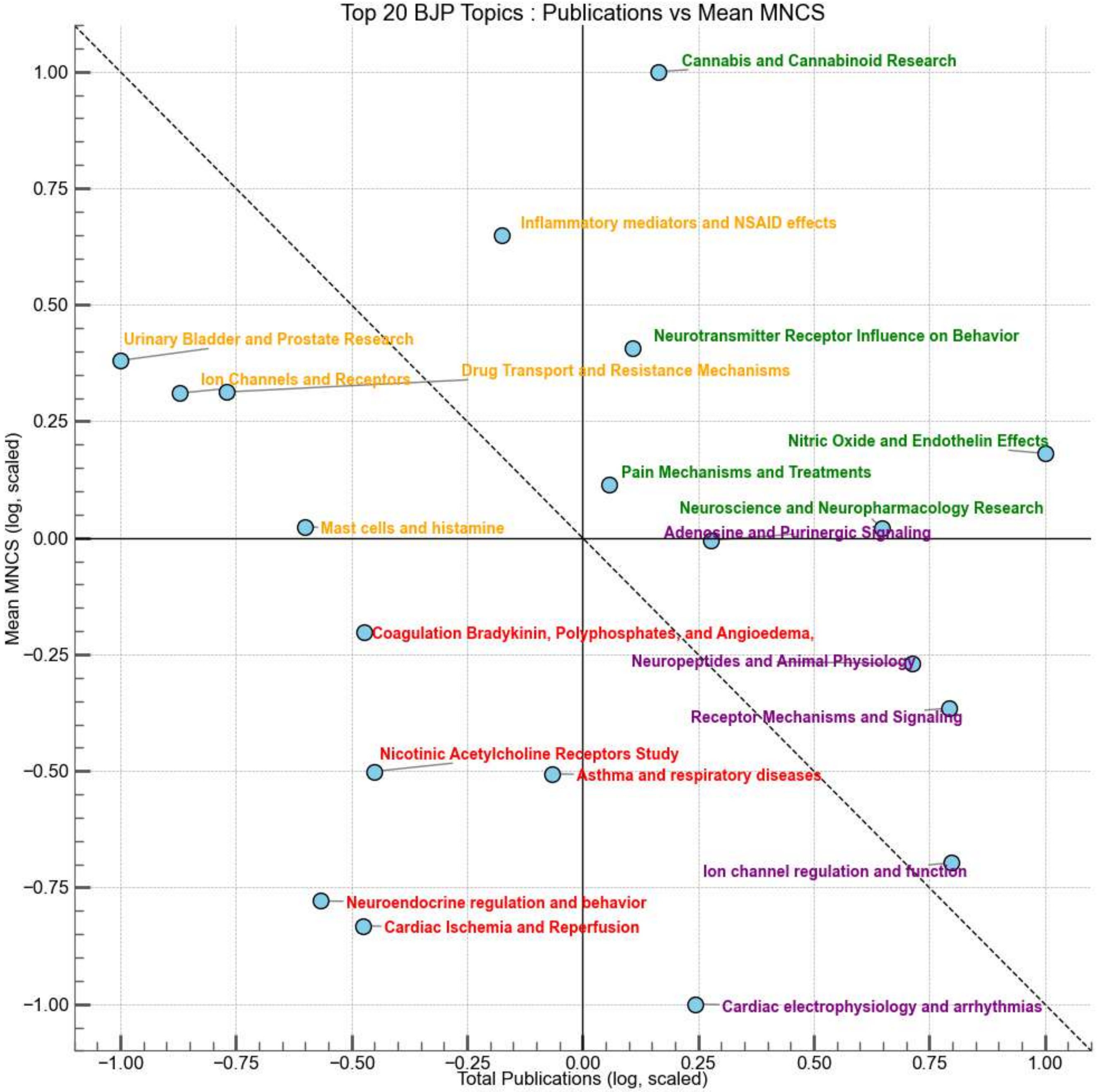


Figure 82: Mean MNCS vs. Publication Count for Top BJP Topics

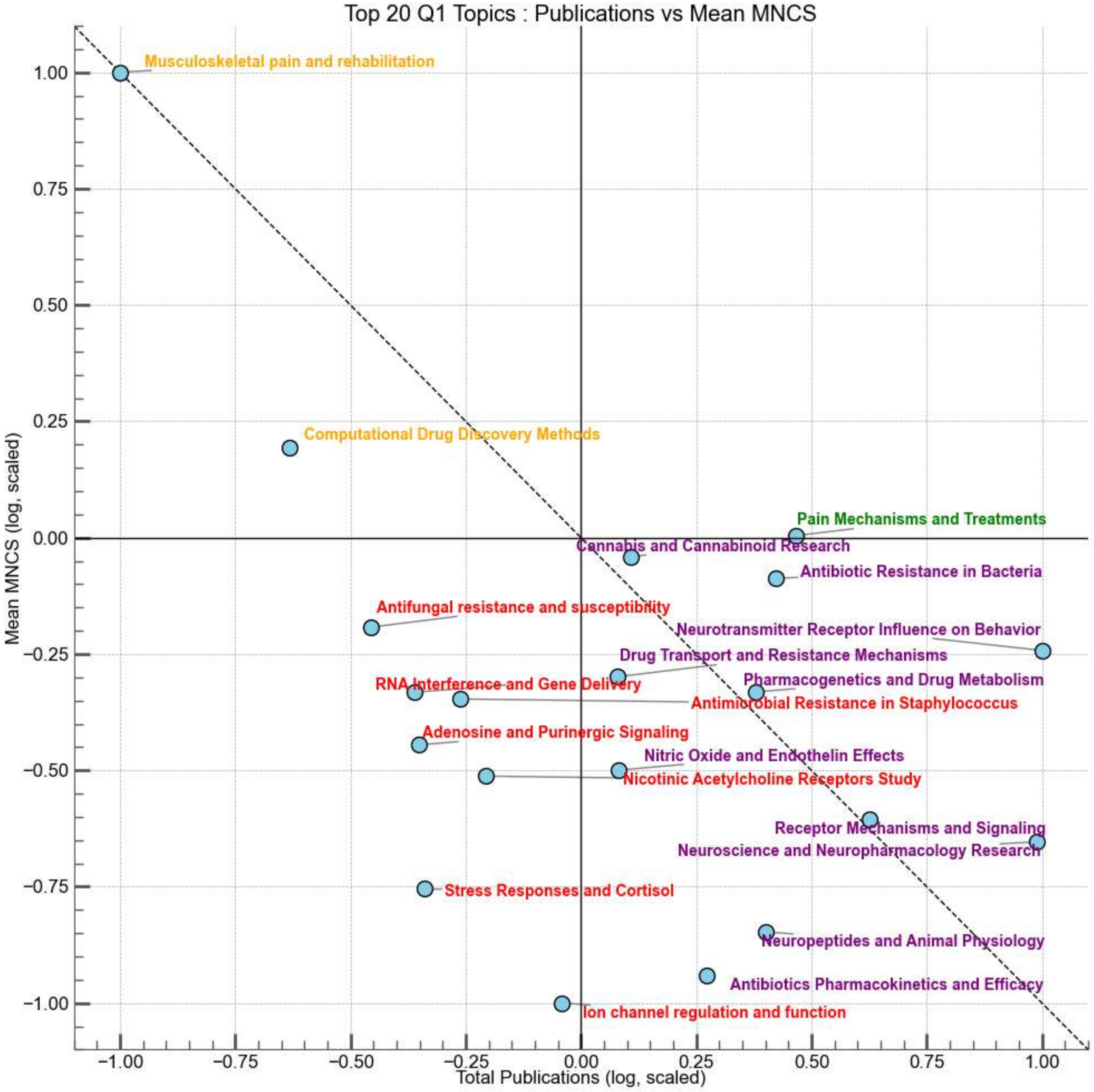


Figure 83: Mean MNCS vs. Publication Count for Top Q1 Institutions

Appendix K : Top BERTopic Topics

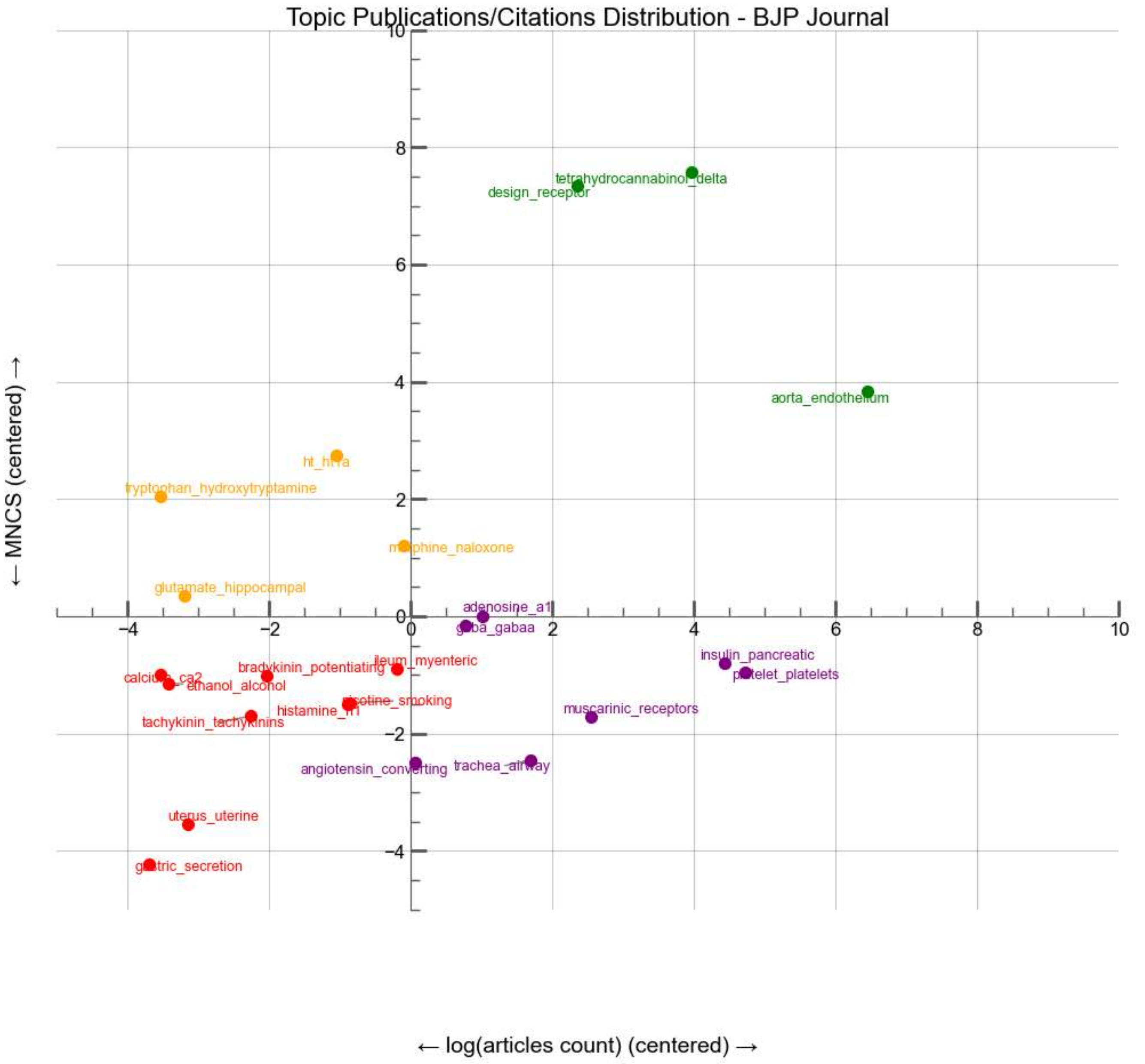


Figure 84: Mean MNCS vs. Publication Count for Top BJP Modelled Topics

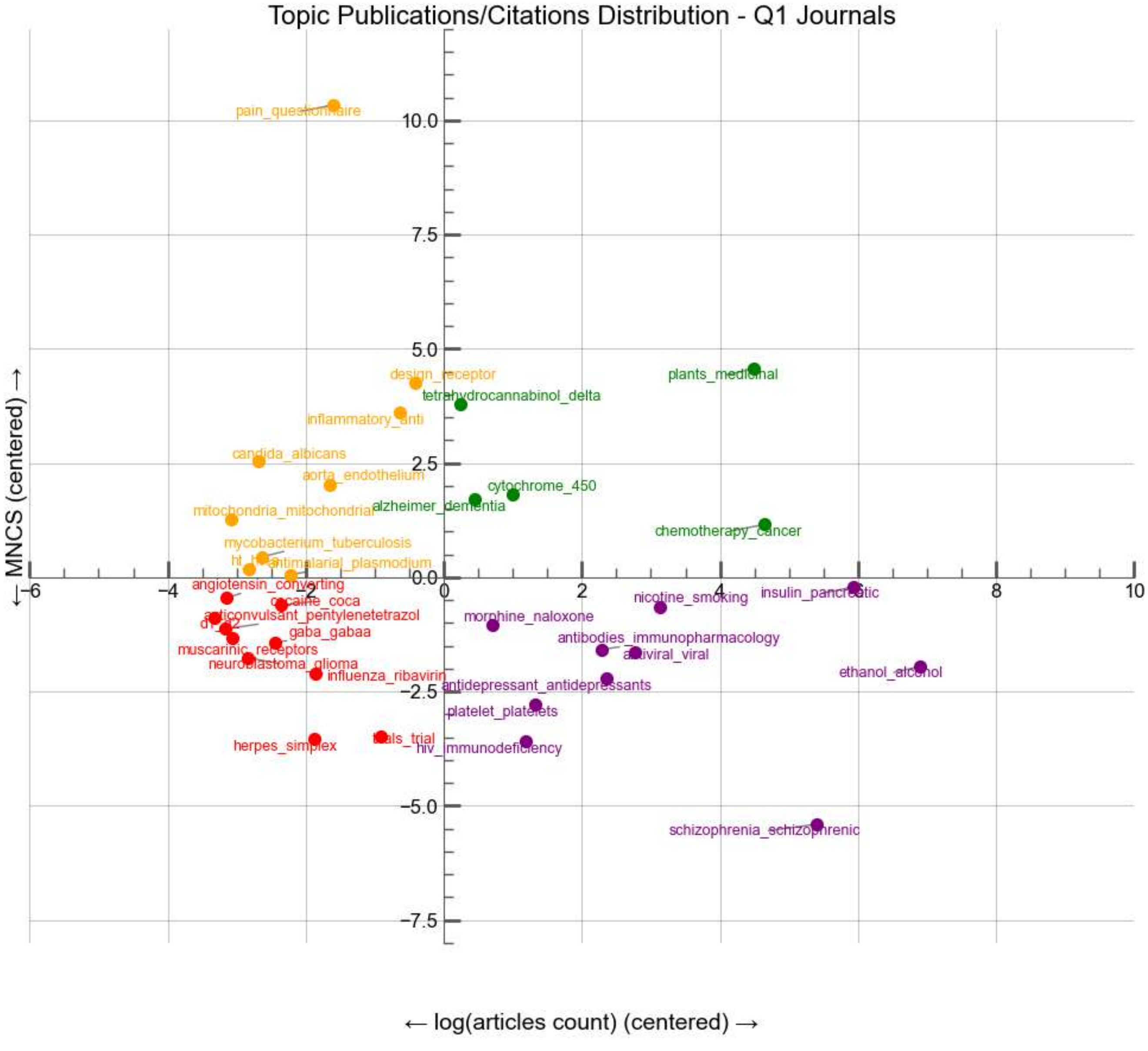


Figure 85: Mean MNCS vs. Publication Count for Top Q1 Modelled Topics

Appendix L : New Modelling Variables

Important Note

Before moving to the modelling step, several new historical features were computed to quantify past contributions and impact of authors, institutions, and topics :

- **mean_past_contributions_authors**: the average number of previous publications of the authors involved in the paper. Formally, for a paper i published in year y with a set of authors A_i:

$$\text{mean\_past\_contributions\_authors}_i = \frac{1}{|A_i|} \sum_{a \in A_i} NbPublications \text{ by author } a \text{ before year } y$$

- **mean_past_mncs_authors**: the average past MNCS of the authors :

$$\text{mean\_past\_mncs\_authors}_i = \frac{1}{|A_i|} \sum_{a \in A_i} meanMNCS \text{ of all previous publications by author } a$$

- **mean_past_contributions_institutions**: the average number of previous publications of the institutions involved in the paper. For a paper i with a set of institutions I_i:

$$\text{mean\_past\_contributions\_institutions}_i = \frac{1}{|I_i|} \sum_{j \in I_i} NbPublications \text{ by institution } j \text{ before year } y$$

- **mean_past_mncs_institutions**: the average past MNCS of the institutions :

$$\text{mean\_past\_mncs\_institutions}_i = \frac{1}{|I_i|} \sum_{j \in I_i} meanMNCS \text{ of all previous publications by institution } j$$

- **past_contributions_topic**: the cumulative number of publications of a topic before the publication year of paper i:

$$\text{past\_contributions\_topic}_i = \sum_{\substack{w \in \text{topic } T \\ \text{year}(w) < y}} 1$$

- **mean_past_mncs_topic**: the cumulative average MNCS of previous publications of a topic :

$$\text{mean\_past\_mncs\_topic}_i = \frac{\sum_{\substack{w \in \text{topic } T \\ \text{year}(w) < y}} \text{MNCS}_w}{\text{past\_contributions\_topic}_i}$$

Appendix M : Decade Modelling Tables

Table 17: Regression coefficients for the 1950s decade (p-value below 0.05)

Feature	Coefficient	Type
Top 5 positive coefficients		
topic_curves_dose	79.9487	BERTopic Topic
primary_topic_Musculoskeletal pain and rehabilitation	50.1623	Primary Topic
topic_pain_questionnaire	50.1623	BERTopic Topic
topic_lithium_carbonate	20.4930	BERTopic Topic
primary_topic_Nausea and vomiting management	19.8368	Primary Topic
Top 3 negative coefficients		
primary_topic_Chalcogenide Semiconductor Thin Film	-12.7395	Primary Topic
primary_topic_History and advancements in chemical analysis	-12.9758	Primary Topic
primary_topic_Material Properties and Applications	-13.7056	Primary Topic
Some non-topic features		
mean_past_mncs_institutions	0.6420	Non-topic
mean_past_mncs_authors	0.2295	Non-topic
mean_past_mncs_topic	-1.6474	Non-topic

Table 18: Regression coefficients for the 1980s decade (p-value below 0.05)

Feature	Coefficient	Type
Top 5 positive coefficients		
topic_adverse_reactions	16.2709	BERTopic Topic
primary_topic_Cardiovascular Health and Risk Factors	12.8763	Primary Topic
topic_curves_dose	10.9885	BERTopic Topic
primary_topic_Asymmetric Hydrogenation and Catalysis	9.8270	Primary Topic
primary_topic_Pharmacovigilance and Adverse Drug Reactions	9.2308	Primary Topic
Top 3 negative coefficients		
topic_british_society	-3.8505	BERTopic Topic
primary_topic_Dermatological diseases and infestations	-8.3501	Primary Topic
primary_topic_Educational Assessment and Pedagogy	-12.1736	Primary Topic
Some non-topic features		
mean_past_mncs_institutions	0.4201	Non-topic
authors_count	0.0621	Non-topic
mean_past_mncs_topic	-1.0507	Non-topic

Table 19: Regression coefficients for the 2000s decade (p-value below 0.05)

Feature	**Coefficient**	**Type**
Top 5 positive coefficients		
primary_topic_Advanced Text Analysis Techniques	20.2940	Primary Topic
primary_topic_Delphi Technique in Research	10.8827	Primary Topic
primary_topic_Silicon Effects in Agriculture	8.7316	Primary Topic
primary_topic_Microbial bioremediation and biosensing	6.0671	Primary Topic
primary_topic_Enzyme-mediated dye degradation	6.0266	Primary Topic
Top 2 negative coefficients		
topic_trigeminal_neuralgia	-1.8097	BERTopic Topic
topic_transit_intestinal	-2.5854	BERTopic Topic
Some non-topic features		
mean_past_mncs_institutions	0.3335	Non-topic
countries_distinct_count	0.0697	Non-topic
institutions_distinct_count	0.0423	Non-topic
age	0.0342	Non-topic

Table 20: Regression coefficients for the 2010s decade (p-value below 0.05)

Feature	**Coefficient**	**Type**
Top 5 positive coefficients		
primary_topic_Reliability and Maintenance Optimization	12.5585	Primary Topic
primary_topic_Plant responses to water stress	11.3972	Primary Topic
topic_methyltyrosine_amphetamine	11.0419	BERTopic Topic
primary_topic_Military Defense Systems Analysis	10.9744	Primary Topic
primary_topic_Speech and Audio Processing	9.8688	Primary Topic
Some non-topic features		
mean_past_mncs_institutions	0.5110	Non-topic
countries_distinct_count	0.0473	Non-topic
referenced_works_count	0.0118	Non-topic
mean_past_contributions_institutions	-0.0002	Non-topic
past_contributions_topic	6.754e-07	Non-topic
mean_past_mncs_authors	-0.0400	Non-topic
mean_past_mncs_topic	-1.7840	Non-topic

Appendix N : Correlation Matrix

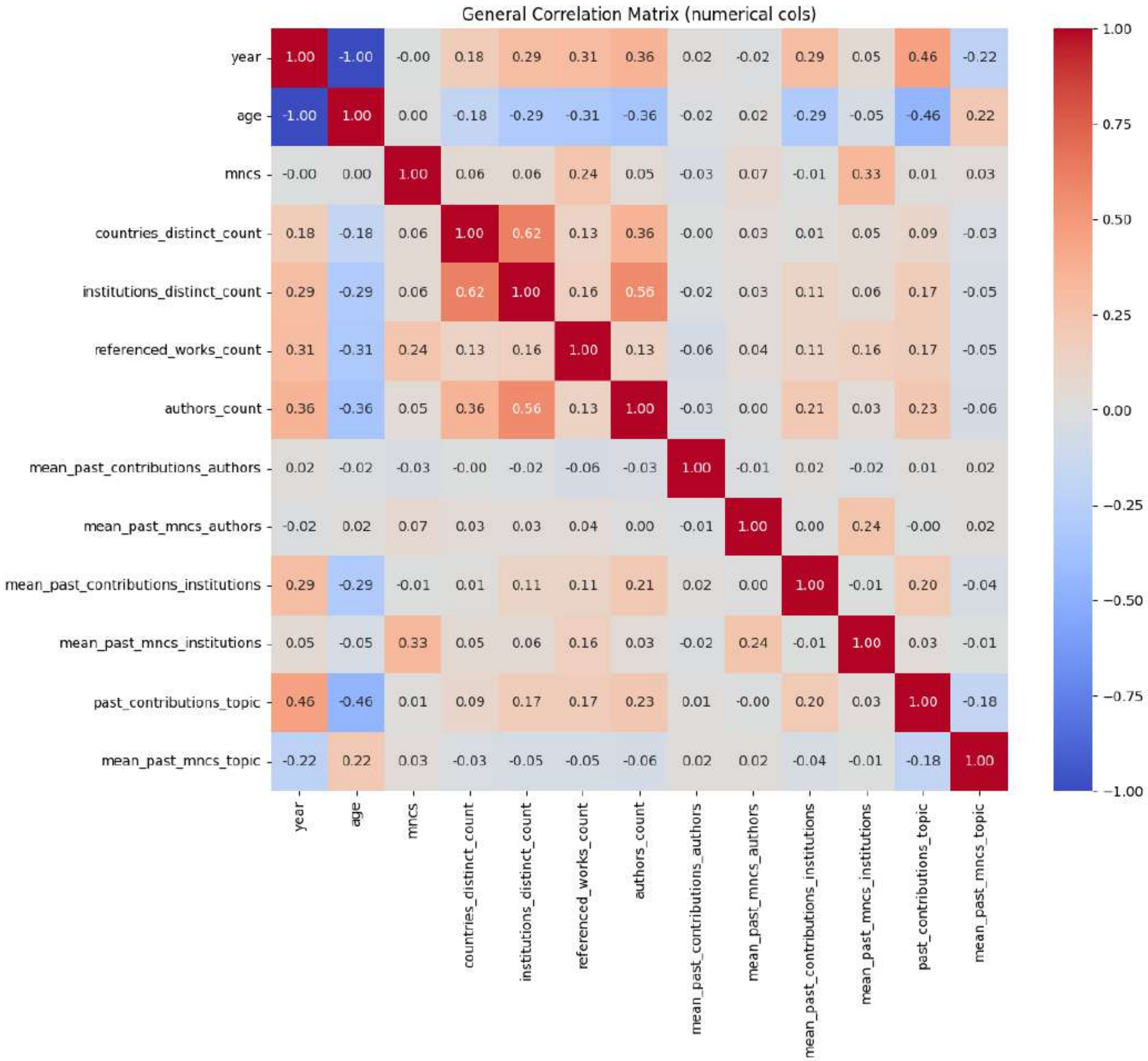


Figure 86: Correlation Matrix

Appendix O : Machine Learing Models

The following machine learning models were applied to the dataset:

- **Logistic Regression** is a linear classification model that estimates the probability of class membership using a logistic function:

  $$P(y=1\mid \mathbf{x}) = \frac{1}{1+e^{-(\beta_0+\beta^\top \mathbf{x})}}.$$

  Model parameters are estimated by maximizing the likelihood, and the decision boundary is linear in the feature space.

- **Random Forest** is an ensemble learning method that builds a collection of decision trees trained on bootstrap samples of the data. Each tree produces a class prediction, and the final prediction is obtained by majority voting:

  $$\hat{y} = \mathrm{mode}\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}.$$

  This approach reduces variance and improves generalization by decorrelating individual trees.

- **XGBoost** is a gradient boosting algorithm that sequentially adds decision trees to minimize a regularized objective function:

  $$\mathscr{L} = \sum_i \ell(y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

  where ℓ is a loss function and Ω penalizes model complexity. Each new tree corrects the residual errors of the previous ensemble.

- **LightGBM** is a gradient boosting framework optimized for efficiency, relying on histogram-based splitting and leaf-wise tree growth. It approximates continuous features into discrete bins, significantly reducing computational cost while maintaining predictive performance.

- **Linear Discriminant Analysis (LDA)** is a probabilistic classifier that assumes class-conditional normal distributions with shared covariance matrices. It projects data onto a lower-dimensional space by maximizing the ratio:

  $$\frac{\text{between-class variance}}{\text{within-class variance}},$$

  leading to linear decision boundaries between classes.

- **k-Nearest Neighbors (kNN)** is a non-parametric method that assigns a class label based on the majority class among the k closest observations according to a distance metric, typically the Euclidean distance:

  $$d(\mathbf{x}, \mathbf{x}_i) = \|\mathbf{x} - \mathbf{x}_i\|_2.$$

  The model makes no explicit assumptions about data distribution.

- **Gaussian Naive Bayes** is a probabilistic classifier based on Bayes' theorem:

  $$P(y \mid \mathbf{x}) \propto P(y) \prod_j P(x_j \mid y),$$

  assuming conditional independence between features. Each feature is modeled using a Gaussian distribution within each class.

- **Multi-Layer Perceptron (MLP)** is a feedforward neural network composed of multiple fully connected layers. Each layer performs a transformation of the form:

  $$\mathbf{h}^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right),$$

  where σ is a non-linear activation function. Model parameters are learned using backpropagation to minimize a loss function.

Bibliography

References

[1] HSDSLab. *Human & Social Data Science Lab*. https://hsdslab.math.bme.hu/. 2026.

[2] Santo Fortunato et al. "Science of science". In: *Science* 359.6379 (2018), eaao0185. DOI: 10.1126/science.aao0185.

[3] N. M. Darnalis, H. Zainal, and E. Germovsek. "Bibliometric Analysis to Explore Trends of the 100 Most Cited Articles in Population Pharmacokinetic and/or Pharmacodynamic Modelling". In: *Journal of Scientometric Research* 14.2 (2025), pp. 706–736. DOI: 10.5530/jscires.20250983.

[4] M. E. Basol and R. Seifert. "Bibliometric analysis of Naunyn–Schmiedeberg's Archives of Pharmacology (1947–1974)". In: *Naunyn–Schmiedeberg's Archives of Pharmacology* 397 (2024), pp. 7141–7168. DOI: 10.1007/s00210-024-03078-8.

[5] D. F. Thompson. "Bibliometric Analysis of Pharmacology Publications in the United States: A State-Level Evaluation". In: *Journal of Scientometric Research* 7.3 (2018), pp. 167–172. DOI: 10.5530/jscires.7.3.27.

[6] SCImago Lab. *SCImago Journal & Country Rank (SJR)*. https://www.scimagojr.com/. 2026.

[7] J. E. Hirsch. "An index to quantify an individual's scientific research output". In: *Proceedings of the National Academy of Sciences* 102.46 (2005), pp. 16569–16572. DOI: 10.1073/pnas.0507655102. URL: https://doi.org/10.1073/pnas.0507655102.

[8] C. Gini. *Variabilità e mutabilità*. Rome: Libreria Eredi Virgilio Veschi, 1912.

[9] OpenAlex. *Work Object — OpenAlex API Documentation*. https://docs.openalex.org/api-entities/works/work-object. 2024.

[10] lead-ratings. *gender-guesser: Python library for gender detection from first names*. https://github.com/lead-ratings/gender-guesser. 2024.