UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Master's thesis

Revisiting the Gender Representation Disparities in Physics

From 1970 to 2023

Signe Boldt Bendsen

Advisor: Mathias Spliid Heltberg and Troels Christian Petersen

Submitted: July 1, 2024

This thesis has been submitted to the The Faculty of Science, University of Copenhagen

Acknowledgements

I wish to thank my supervisors; Postdoc Mathias Spliid Heltberg, University of Copenhagen, The Niels Bohr Institute, Department of Biocomplexity, and Associate Professor Troels Christian Petersen, University of Copenhagen, The Niels Bohr Institute, Department of Experimental Subatomic Physics, for embarking with me on this journey and for sharing their expertise and support. A special thanks goes to Associate Professor Mathias Wullum Nielsen, University of Copenhagen, Department of Sociology, for assisting me throughout this project and especially providing excellent guidance on the Sociological and Science of Science aspects; this study would not have happened without his participation. I also wish to thank Professor Martina Erlemann, Freie Universität Berlin, Department of Physics, Gender and Science Studies in Physics, for including me in her Research Seminar - as well as everyone who participated - a forum where I did not only receive valuable feedback on this project but also gathered a lot of inspiration and knowledge. Lastly, I must acknowledge my friends and family for their love and encouragement. Thank you to everyone who have been a part of this process.

Abstract

Dette studie undersøger forskelle i mænd og kvinders akademiske karrierer indenfor fysik fra 1970 frem til 2023. Udover at kortlægge de eksisterende bias har studiet til formål at karakterisere de underlæggende faktorer, der bidrager til vedvarende ulighed mellem køn i akademia. Projektets data udgør 1.432.907 forfatterskaber hvilket har analyserts ved brug af statistiske metoder og Machine Learning teknikker, for at undersøge forskelle mellem køn, specifikt i relation til prestige, aktivitetsniveau og karrierelængder samt repræsentation indenfor fysikkens grene. Studiet viser statistiske signifikante forskelle mellem mandlige og kvindelige forfattere, særligt i forhold til institutionelle tilhørsforhold, publikationsrater, prestigemarkører og karrierelængder. Dog var det ikke muligt, ved brug af Machine Learning, at separere de to køn på baggrund af disse faktorer fuldstændigt. Tilgengæld, viser studiet indikationer på,at de kønsmæssige forskelle indenfor fysik stammer fra homofili, vedvarende tranditionelle kønsroller og forskelle i kulturer. På den anden side, viser studiet, at antallet af kvinder i det fysiske felt er steget over tid. For videre dybdegående undersøgelser af studiets resultater anbefales for eksempel Cox Proportional Hazards Model. Denne model tager højde for tidslige variationer og relevante covaritioner til at undersøge resultaternes aktualitet. Ydermere kunne kvalitative studier bruges som supplement til de kvantitative resultater for at dybdegøre og nuancere deres kompleksitet. Alt i alt, belyser studiet ikke kun systemiske udfordringer for kvinder i fysik, men også muligheder og understreger vigtigheden af at fostre diversitet og inklusion i disciplinen.

keywords:

Gender Disparities, Academic Careers, Physics, Statistics, Machine Learning, Prestige Markers, Survival Rates, Publication Patterns, Diversity in STEM, Gender Equity

Contents

Introduction

As I embarked on my journey as an aspiring physicist at the Niels Bohr Institute, my quest was to unravel the laws of nature and the mysteries of our existence.

Yet, besides my enthusiasm and eagerness to explore the world of physics, a theme of questions started to present themselves. As I was walking down the historical halls of the Niels Bohr Institute I noticed the traditional photo wall capturing the current staff year by year since its establishment back in 1921 [1]. Alongside the iconic images of renowned male physicists, such as Niels Bohr himself, hung the less familiar faces of female scientists as depicted in Figure 1. Despite their presence in the historical record, these women remained largely unrecognised throughout my education. This dissonance between representation and recognition prompted me to delve deeper into the overlooked contributions of women in physics.

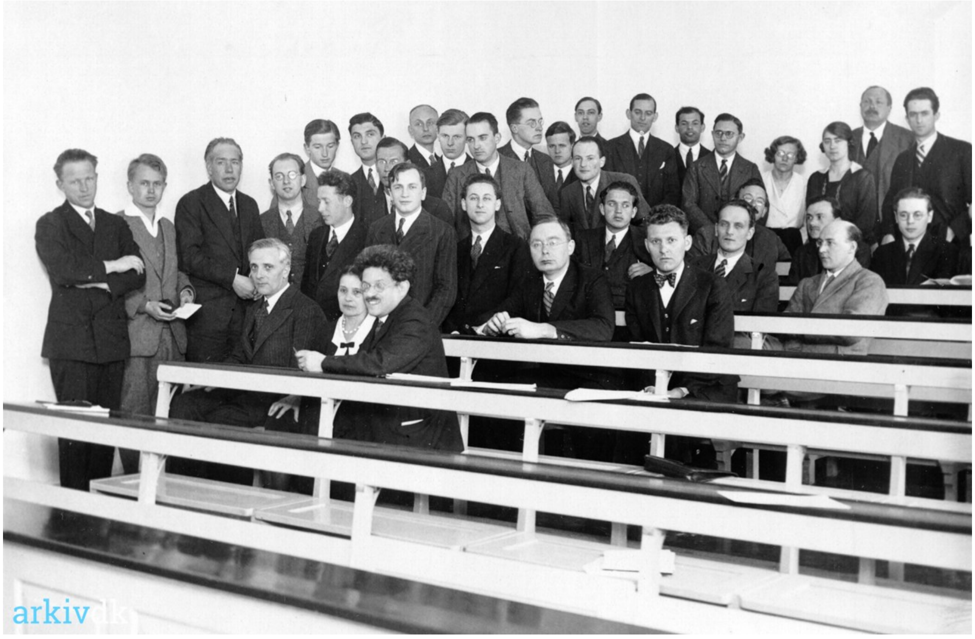


Figure 1: Copenhagen Conference at the University of Copenhagen Institute for Theoretical Physics 1932. Among famous scientists - Werner Heisenberg, Piet Hein, Niels Bohr, and many more - we see Lise Meitner sitting on first row. She is 1 out of 3 women on the photo, the other two being Ingeborg Lam and Eva Rindal. Source: [2].

An example of this was when one of my lecturers dedicated a session to recounting the story of Lise Meitner. In the photo (Figure 1), she is seen sitting in the front row. My professor wanted to share with us the unjust story about Otto Hahn receiving a Nobel Prize in Chemistry in 1944 over his collaborator List Meitner who made a large contribution on the theory behind nuclear fission [3]. Still to this day, she is not granted recognition for her work, as the movie ”Oppenheimer” (2023) about the development of the atomic bomb (”The Manhattan Project”) illustrated. While, in the movie, Otto Hahn and Robert Frisch are mentioned and acknowledged as scientists behind the fission theory, which was crucial in the atomic bomb development, Meitner is completely left out. This systematic bias against acknowledging female scientists' contributions is commonly referred to as the ”Matilda effect” [4].

”This is not part of the curriculum, I just wanted to tell you the story” my professor said. Even though I appreciate the effort and thought behind sharing Meitners story, which is more than the $100 million-budget Blockbuster hit did [5], the exclusion of such stories from the curriculum undermines *the Matilda effect*. It is indicative of the differential treatment of female physicists in academia compared to their male counterparts. In the words of the French existentialist philosopher Simone de Beauvoir, they are ”the second sex” [6].

Lise Meitner represent just one example among many women whose contributions went unrecognised - *the Matilda effect* - despite their significant impact, even being snubbed for the Nobel Prize [7]. In fact, a study has shown that the rate with which female scientists have been granted the Nobel Prize is statistically unlikely, with a probability of gender bias exceeding 96% [8]. Moreover, research within the science of science has shed light on persistent gender biases in physics, including perceptions of scientific talent and brilliance [9, 10], disparities in authorship recognition [11, 12], inequities in patent inventors and innovation benefits [13, 12], and biases in grant submissions [10]. Despite the widely held belief that ”gender diversity leads to better science” [14], these biases do persist.

The latter, presumably shared by the Niels Bohr Institute itself as they are on a mission to increase diversity and make it an attractive place to work for both genders to work. Their mission includes a target of 35% of associate professors and 30% of professor at Niels Bohr Institute to be female by year 2030. Remarkably, these are long-term goals and reveal just how limited presentation is as of today (in March 2021: 20,5% assoc. prof. and 13,8% prof.) However, it is not specified exactly how they are planning on implementing and achieving this [15].

The Niels Bohr Institute is not alone in its growing emphasis on gender diversity in research. Literature underscores that accounting for sex and gender enhances scientific outcomes [16, 17]. Studies have demonstrated a strong positive correlation between author gender, particularly women’s authorship, and the inclusion of gender and sex analysis in research [18]. Neglecting to consider sex and gender in research design can have dire consequences, as highlighted by Londa Schiebinger, a science historian at Stanford University. For instance, between 1997 and 2000, ten drugs were withdrawn from the U.S. market due to life-threatening effects, with eight posing greater risks to women than men. Examples abound, from crash test dummies designed without accounting for shorter individuals (including many women) to facial recognition systems trained on biased datasets. Medical studies often overlook gender phenotypical diseases, while data gaps in city planning lead to inefficient transportation systems [19]. Caroline Criado Perez’s award-winning book extensively documents how women are affected by gender bias in big data collection [20].

In response, the European Commission mandated in 2020 that grant recipients integrate sex and gender analyses into research designs. However, with exceptions for researchers working on topics for which the commission thinks gender and sex would be irrelevant, such as in pure mathematics [16]. This means that, in many cases, physics will be one of the exceptions as well, so why even care about gender representation within physics?

From a philosophical point of view, we could argue the need for a more diverse representation within physics and natural sciences in order to even out the power balance. According to the French philosopher Michel Focault there exists an intertwined relationship between power and knowledge: ”Power creates and recreates its own fields of exercise through knowledge” [21]. Consequently, if knowledge is monopolised by a select group in society, so is power. This link between knowledge and power is evident in studies within the Science of Science discipline. For instance, research has shown a correlation between authors’ backgrounds, genders, and their areas of study or target populations for innovation [13, 10, 17, 22].

Similarly, gender bias and inequality within STEM, computer science, cybersecurity, and the digital sector have significant consequences. Research has documented that artificial intelligence (AI) and machine learning (ML) carry an underlying gender bias, largely because they have predominantly been designed by (white) males. This bias originates from datasets created by homophilic scientist groups, which underrepresent or misrepresent groups, such as women [**zacharia**], as well as black and queer people [23]. These are just a few examples. Since physics is a natural science, innovations, medical research, and computer science are all integral parts of physics and the career paths that stem from a physics degree. While many more examples undoubtedly exist, this selection suffices to illustrate the power imbalance inherent in knowledge. With knowledge comes the power to innovate, discover, and thus shape and influence society.

Hence, I argue that diversity within gender does indeed matter. However, as previously mentioned, our academic world, particularly within physics (and beyond), is far from unbiased and equal [24, 11, 8, 9, 25, 10, 13, 26]. Therefore, my research aims to go beyond these well documented biases. Rather than rediscovering these facts, I wanted to investigate what happens to the women that actually *do* enter physics. As a physicist, I am intrigued by the movement, nature, and patterns of behaviour and development among female physicists. Who are the women depicted in those few photographs? What does their career trajectory entail? How does it differ from that of their male colleagues? If our goal is truly to encourage more women to pursue careers in physics and to bolster the female workforce in the digital and technological sectors [27], understanding these patterns and dynamics could prove crucial, shedding light on the aspects of physics that women excel in or are drawn to compared to other fields.

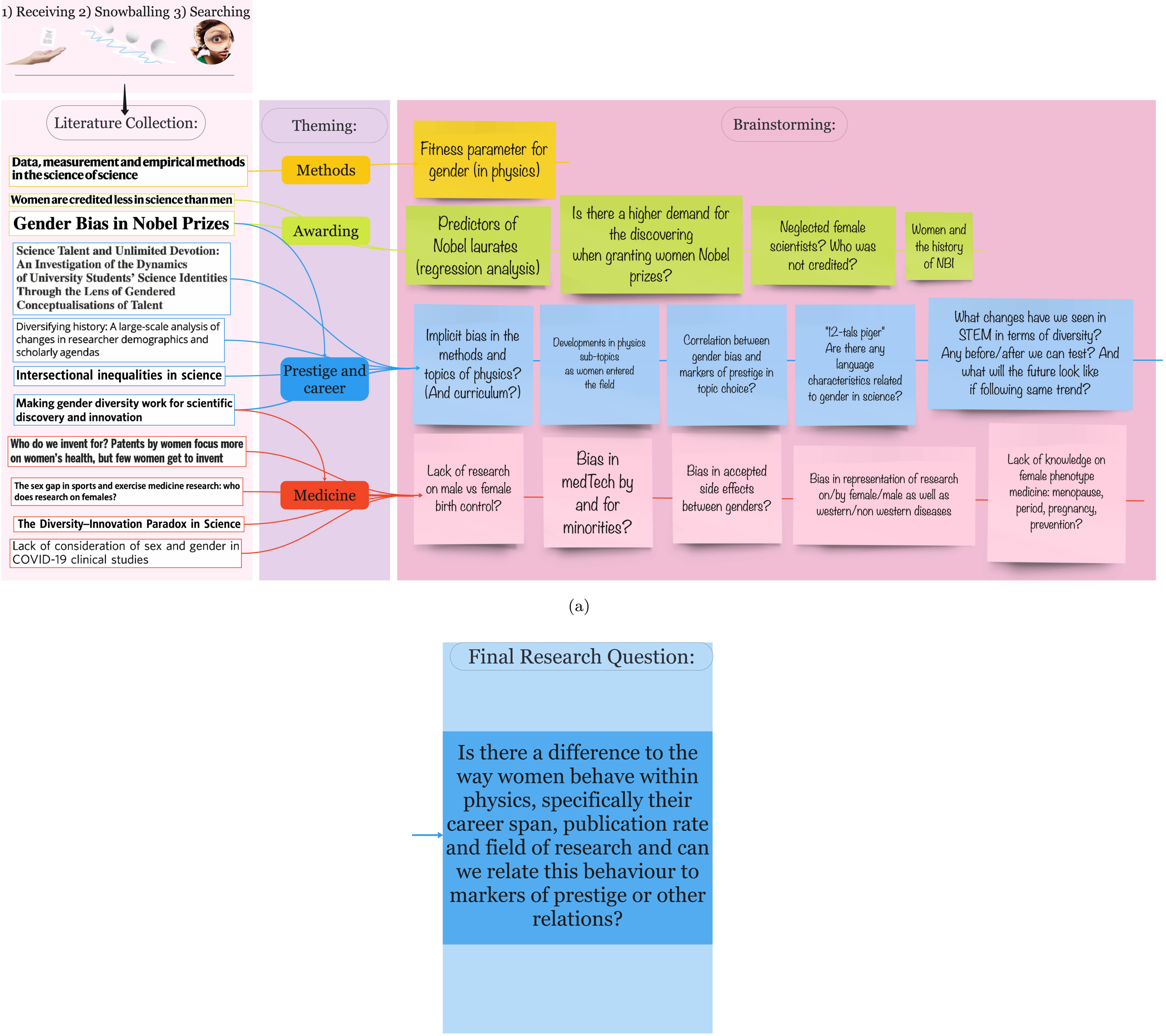
This study builds upon existing research indicating that women are generally more represented in broader, interdisciplinary fields [10]. The study by Kozlowski et al. recommends delving deeper by examining the relationship between these fields of research and biases and markers of prestige. While prior research has focused on academia as a whole, this study specifically concentrates on physics and its sub-fields. Additionally, another article by Erlemann [28] discusses that even the performance of physics may be gendered. For instance, contradictions between the traditionally masculinised intimacy of men with machines and the traditionally feminised concept of care could manifest similarly in the distribution of female and male authors within sub-fields of physics. Therefore, with this study we aim to discover whether this theory will be reflected in the behavioural patterns of women within physics and if there are any relation to this behaviour in terms of prestige markers or other factors.

All of these reflections ultimately led to the final research question: *Is there a difference to the way women behave within physics, specifically their career span, publication rate and field of research and can we relate this behaviour to markers of prestige or other relations?*

Methodology

Literature

Figure 2 provides a visual depiction of the literature search and subsequent brainstorming sessions that shaped our research question.


1) Receiving 2) Snowballing 3) Searching
Literature Collection:
Data, measurement and empirical methods in the science of science
Women are credited less in science than men
Gender Bias in Nobel Prizes
Science Talent and Unlimited Devotion: An Investigation of the Dynamics of University Students' Science Identities Through the Lens of Gendered Conceptualisations of Talent
Diversifying history: A large-scale analysis of changes in researcher demographics and scholarly agendas
Intersectional inequalities in science
Making gender diversity work for scientific discovery and innovation
Who do we invent for? Patents by women focus more on women's health, but few women get to invent
The sex gap in sports and exercise medicine research: who does research on females?
The Diversity–Innovation Paradox in Science
Lack of consideration of sex and gender in COVID-19 clinical studies
Theming:
Methods
Awarding
Prestige and career
Medicine
Brainstorming:
Fitness parameter for gender (in physics)
Predictors of Nobel laurates (regression analysis)
Is there a higher demand for the discovering when granting women Nobel prizes?
Neglected female scientists? Who was not credited?
Women and the history of NBI
Implicit bias in the methods and topics of physics? (And curriculum?)
Developments in physics sub-topics as women entered the field
Correlation between gender bias and markers of prestige in topic choice?
"12-tals piger" Are there any language characteristics related to gender in science?
What changes have we seen in STEM in terms of diversity? Any before/after we can test? And what will the future look like if following same trend?
Lack of research on male vs female birth control?
Bias in medTech by and for minorities?
Bias in accepted side effects between genders?
Bias in representation of research on/by female/male as well as western/non western diseases
Lack of knowledge on female phenotype medicine: menopause, period, pregnancy, prevention?

(a)

Final Research Question:
Is there a difference to the way women behave within physics, specifically their career span, publication rate and field of research and can we relate this behaviour to markers of prestige or other relations?

(b)

Figure 2: An illustration of **a** the iterative process of reviewing literature and generating research questions, leading to **b** our final research question. Each colour in the figure corresponds to a thematic category into which we grouped the literature. The brainstorming process led to several research questions based on their themes and their associated literature. The furthest right side column shows the preliminary question we ended up going forward with based on the brainstorming process and the overall theme *Prestige and Career*

Prior to initiating the study, we conducted a comprehensive review of literature concerning gender bias in STEM fields presented in Figure 2(a). This process aimed not only to inform our research but also to ensure that our study builds upon existing knowledge and potentially contributes novel insights. The literature collection was conducted under the guidance of Mathias Wullum Nielsen, an authority in gender and diversity within scientific domains. Additionally, we employed a technique known as ”backward snowballing” [29], wherein we leveraged the reference lists of relevant papers to identify additional pertinent literature. Moreover, we conducted literature searches using appropriate keywords. Subsequently, we carefully evaluated these questions, ultimately selecting the thematic category represented by the blue colour. Based on the literature presented in this category, and their recommendations and findings, our final research question, Figure 2(b), emerged.

Software and Code

The data analysis of this project is conducted in software Python, Jupyter Notebook. The code repositry for this project can be found in https://github.com/SINEBB/Thesis/tree/main. Specifically, ”Analysis 4.0” is the notebook wherein the final results have been produced.

Database

To address our research question, we analysed academic publication data within the field of physics. Several databases were considered during the initial stages of this project, including Scopus, SciSciNet, and OpenAlex.

- **Scopus** had the advantage of it's python-based API-wrapper, *pybliometrics*, which facilitated easy access to the database, enabling seamless retrieval and extraction of data directly into Jupyter Notebook [30]. However, Scopus had limitations on the number of data retrieval requests [31], why we explored alternative databases.

- **SciSciNet** emerged as a potential alternative, tailored for research within the ”Science of Science” domain [32]. However, the data proved challenging to work with, requiring manual downloading and importing, which was memory-intensive and time-consuming. Additionally, SciSciNet lacked full abstract texts, a crucial metric for our research objectives.

- **OpenAlex** turned out to be the best fit for our purpose. It provides the *pyalex* software as a Python interface to its API [33]. OpenAlex serves as an index of interconnected scholarly papers, authors, and institutions. Similar to Scopus, OpenAlex facilitated direct importation into Jupyter Notebook, offered unlimited access, and contained the requisite data. Consequently, we proceeded with data from OpenAlex for our study.

Data Processing

Table 1 provides an overview of the measures obtained from the OpenAlex Database. The complete dataset consisted of N=533,794 unique papers, comprising N=2,000,305 rows (accounting for multiple authors per paper) after filtering for English articles within the field of physics.

To extract papers only within the physics domain, we utilised the pyalex search function. We retrieved the desired data from OpenAlex by querying for abstracts related to ”physics” published between 1970 and 2023, iterating through each year individually. Given the time-consuming nature of data retrieval, we simultaneously downloaded and organised sample data to facilitate analysis and familiarisation with the dataset structure.

The sample data encompassed 31,168 publications from the year 2023. Organisational tasks involved unpacking JSON files (stored within a single column) and rather saving variables into distinct columns. The initial code structure for organisation on the sample data were used for later reorganising the entire dataset.

Data from each year was stored in a memory-efficient manner to facilitate future access. We opted for the pickle format due to its advantages in terms of speed and storage efficiency [34, 35]. Subsequently, the downloaded and reorganised data from each year spanning 1970 to 2023 were concatenated into a unified dataset.

The choice of 1970 as the starting year was somewhat arbitrary. But, based on an assessment that the period from 1970 to 2023 provided a sufficiently comprehensive time frame for analysis and that the dataset size diminished approaching the year 1970 (approximately 1000 entities), it seemed like a good choice.

Table 1: The OpenAlex database provides detailed information on scientific publications. After importing data on publication within physics from year 1970-2023 and organising JSON files into separate columns we ended up with 22 distinct variables. The variable name and it's comprehensive description is provided here.

Initial Variable List		
	Name	**Description**
1	Article ID	Each article in the dataset is assigned with a unique OpenAlex ID. This ID can be used to distinct the data on article level.
2	DOI	The DOI corresponds to the DOI as it is given in the publication.
3	Publication Year	The year of publication in format 'YYYY'.
4	Publication Date	The date of publication in format 'YYYY-MM-DD'.
5	Title	The title corresponds to the title as it appears in the publication.
6	Cited By Count	The number of times the publication has been cited by others.
7	Grants	If the research has received any grants it will appear in this variable otherwise it is empty.
8	Abstract	The abstract corresponds to the abstract as it appears in the publication.
9	Journal ID	Each journal is assigned with a unique OpenAlex ID. This ID can be used to distinct the data on journal level.
10	Journal Name	The name of the journal as corresponding to it's ID.
11	Total Author Counts	The number of total authors listed on the publication author list.
12	Author Position	The position of the given author. This variable can take one of three values: 'first', 'middle', and 'last'.
13	Author Countries	The country associated with the given author.
14	Is Corresponding	Whether the given author is listed as the corresponding author or not. This is a binary variable in format of True/False.
15	Raw Affiliation String	The affiliation for the given author of the given publication.
16	Raw Author Name	The name of the given author of the given publication.
17	Author Display Name	Similar to "Raw Author Name" but with the name as it is listed in the given publication.
18	Author ID	Each author is assigned with a unique OpenAlex ID. This ID can be used to distinct the data on author level.
19	Institution ID	Each institution is assigned with a unique OpenAlex ID. This ID can be used to distinct the data on institutional level.
20	Institution Country Code	The country code of the institution as corresponding to it's ID.
21	Institution Name	The name of the given institution as corresponding to it's ID.
22	Institution Type	The type of the given institution as corresponding to it's ID. This variable can take one of 9 values: 'education', 'facility', 'company', 'government', 'healthcare', 'nonprofit', 'other', 'unknown', and 'archive'.

Gender Assignment

Gender and Sex Terminology

Where the term "sex" refers to a person's biological classification at birth, "gender" refers to a person's internally recognised sense of their gender identity, which may or may not align with their biological sex [36]. In this study, we use authors' first names as proxies for their gender/sex identification. However, it's important to note that we have no way of knowing or estimating whether individuals identify with the gender typically associated with their name or whether their biological sex aligns with societal expectations linked to that name.

In this study we use the term "gender" referring to the likely gender/sex of an author based on their listed publication name. This terminology choice is primarily dictated by the use of a "gender API" software for author gender/sex determination, aiming to maintain consistency in language. It's crucial to recognise that this terminology diverges from the nuanced concept of gender identity, as we do not have any proxies the individuals' internal gender identity solely from their names. Therefore, this project focuses exclusively on binary gender categorisations, despite the existence of non-binary genders and gender identities.

Gender API

Table 2 presents a Gender API Market Study.

Table 2: Gender API Market Study as presented at [37]. The Market Study presents an overview of several available so-called Gender APIs which return a gender based on a given name (and country if given). It depicts different characteristics of each type of API. For this study we prioritised an option with unlimited free requests and free license as highlighted in bright pink. Therefore, we ended up using *gender-guesser* as highlighted in light pink.

	Gender API	**gender-guesser**	**genderize.io**	**NameAPI**	**NamSor**	**damegender**
Database size	431322102	45376	114541298	1428345	4407502834	57282
Regular data updates	yes	no	no	yes	yes	yes
Handles unstructured full name strings	yes	no	no	yes	no	yes
Handles surnames	yes	no	no	yes	yes	yes
Handles non-Latin alphabets	partially	no	partially	yes	yes	no
Implicit geo-localization	yes	no	no	yes	yes	no
Exists locale	yes	yes	yes	yes	yes	yes
Assignment type	probabilistic	binary	probabilistic	probabilistic	probabilistic	probabilistic
Free parameters	$\text{total}_{\text{names}}$, probability	gender	probability, count	confidence	scale	$\text{total}_{\text{names}}$, count
Prediction	no	no	no	no	no	yes
Free license	no	yes	no	no	no	yes
API	yes	no	yes	yes	yes	future
Free requests	limited	unlimited	limited	limited	limited	unlimited

A Gender API is a tool designed to predict the most likely gender based on input parameters name and, optionally, country. As several Gender APIs exists part of the initial phases of the project was to examine the better option for our specific research and purpose. We proceeded with *gender guesser* mainly due to its provision of unlimited free requests as highlighted in the Table 2. While the highlighted parameters offers clear advantages, particularly considering the absence of external funding, it also comes with limitations in terms of database size, update frequency, and assignment methodology. Nevertheless, *gender guesser* were the most suitable option for our research study.

The process of assigning author gender using *gender-guesser* involved the following three steps:

1. Extracting the Author's First Name.
2. Extracting the Author's Country and Converting the Format.
3. Assigning gender based on the extracted information.

Extracting Author First Name

Table 3 shows a few examples of extracted author first names. Within our dataset, author names were represented in two columns: *Raw Author Name* and *Author Display Name*. Both columns contained the author's full name, including both the first name and surname. However, it was inconsistent whether the *Raw Author Name* column or the *Author Display Name* column provided the most detailed and/or accurate author name, as illustrated in Table 3. To ensure accurate gender assignment based on first names, we aimed to extract the name providing as much information as possible.

To achieve this, we developed a function capable of determining which of the columns (*Raw Author Name* or *Author Display Name*) contained the most detailed name per row based on the number of characters in the name. The function creates a new column, *Selected Full Name*, which replicates the name of the column containing the most detailed name. Then, we extracted the author's first name by splitting the selected full name using space as a separator and retrieving the first element, which was then stored in a new column labelled "Extracted First Name."

Table 3: Examples of raw author names and author display names from the dataset, alongside the resulting selected full name determined by a function replicating the name containing the most information between the two columns. Additionally, the table displays the extracted first name obtained from the selected full name by splitting it and copying the first part. This illustrates the process of selecting the most detailed name per row and extracting first names for gender assignment.

Examples of Selected Full Names and Extracted First Names			
Raw Author Name	**Author Display Name**	**Selected Full Name**	**Extracted First Name**
Harvey Scher	H. Scher	Harvey Scher	Harvey
Richard Zallen	Richard Zallen	Richard Zallen	Richard
R Balian	Roger Balian	Roger Balian	Roger
C Bloch	Claude Bloch	Claude Bloch	Claude

Extracting Author Country

Within our dataset, author countries are provided as ISO 3166-1 alpha-2 country codes, which consist of two-letter country codes [38]. While most authors are associated with a single country, a few authors have an array of countries assigned to them. In such cases, we opt to extract the first country listed, saved as *Author First Country*, for compatibility with the gender API, which only accepts one country per author.

Furthermore, the gender API requires the country name rather than the country code and so we create a function to convert the alpha-2 country code into the corresponding full country name using the pycountry.countries module. The resulting country names are stored in a new column labeled *Author Country Name*.

Furthermore, the gender API is only trained, and therefore only accepts, a selection of countries and so if the given country name weren't in that list it was overwritten as *other_countries* [39] and the result stored in column *Author Country Name API*.

Gender Assignment

Gender assignment is performed using the gender API "gender guesser," which utilises an underlying program called "gender". This program determines the gender associated with a given name, optionally considering the country of origin for that name. The output of the gender guesser API is one of the following categories:

- female
- male
- mostly_female
- mostly_male
- andy
- unknown

In this context, "female" or "male" indicates that the name is primarily associated with the respective sex in the database, while "mostly_female" or "mostly_male" suggests a strong association but with some exceptions. The "andy" category (short for androgynous) denotes names with equal probability of being male or female, while "unknown" indicates that the name was not found in the database [40].

We created a function that takes the*Extracted First Name* and *Author Country Name API* for each row and returns the predicted gender in a new column *Predicted Gender*, assigned by the gender guesser's detector. This process enabled gender assignment based on the given name and country, facilitating gender analysis within the dataset.

Gender Assignment Updating

To improve the initial labelling obtained by utilising *gender-guesser* we applied several updates of the gender assignment following 9 steps:

1. **Update According to Duplicates in Author ID with an Assigned Gender**
2. **Update Without Country**
3. **Drop Rows With Invalid Names**
4. **Update Names Including Special Characters**
5. **Update None Names**
6. **Remove '.', ',', and '-' From Names and Update Gender**
7. **Update With Gender of Identical Name**
8. **Update Using *Gender API***
9. **Final Drop Rows With Invalid Names**

Evaluate Ratio Between Genders

Furthermore, to evaluate the ratio between female and male labels obtained before and after updating updating the labels assigned by the gender API we we conducted a χ^2 contingency test, specifically using *scipy.stats.chi2_contingency.* This function computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table observed. The null hypothesis follows that the observed frequencies are the same in both distributions and that if $p < 0.05$ we can reject the null hypothesis meaning that the frequencies differ significantly between the two distributions. Mathematically, the test is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i - E_i}{E_i} \tag{1}$$

where O is the observed frequency and E is the expected [41].

Hypothesis Testing

We hypothesis that there is a difference between the female and male distribution between several key points in our dataset. To test this hypothesis we use Welsh's T-test and Mann Whitney U-test. Both are a test for the null hypothesis that the underlying distribution for two two independent samples are the same. However, where the T-test has some underlying assumptions the U-test has no distributional assumptions and so both statistics are provided [42, 43].

Topic Model

Another essential part of this study was to assign topics to each publication within the dataset. While the overall topic is physics, we aimed to assign topics based on sub-fields of physics. To achieve this, a Topic Model (TM) was employed. TMs reveal the latent structure (such as topics) of large document collections that cannot be manually annotated or labelled [44]. An online variational Bayes algorithm for Latent Dirichlet Allocation (LDA) implemented in Python, as provided by the Gensim library, was used [45].

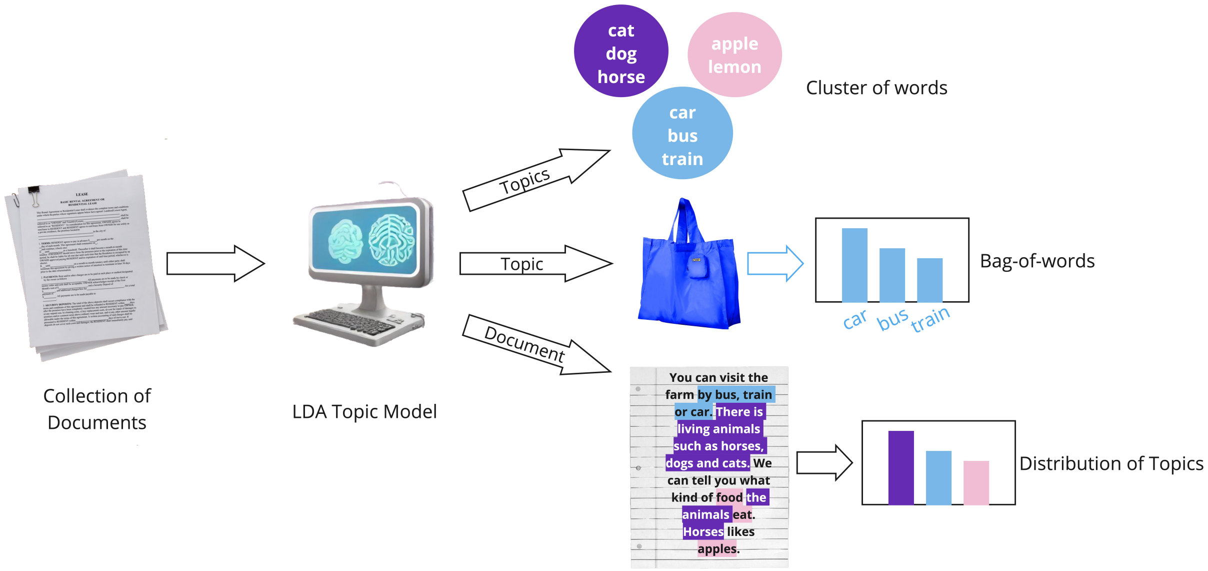
LDA is an unsupervised Bayesian probabilistic model for text documents thus uses probabilities to understand the relationship between words and latent topics in the documents. It encodes assumptions about observed data within a Bayesian framework using Bayesian statistics to update its understanding of the topics words as it sees more and more documents. It revolves around examining the posterior distribution of the topics and words given the documents, model parameters and latent variables (here topics) conditioned on observed data, making it suitable for probabilistic topic modelling; figuring out the different topics in a set of documents and how those topics are related to each other. [44].

LDA operates under the following key assumptions:

1. Documents sharing the same topic will have a lot of words in common.

2. It is a bag-of-words model, meaning it disregards word order or relationships within sentences.

3. It assumes a predefined number of K topics, similarly to the clustering algorithm K-means, aiming to group words and documents into distinct clusters representing topics.



You can visit the farm by bus, train or car. There is living animals such as horses, dogs and cats. We can tell you what kind of food the animals eat. Horses likes apples.

Figure 3: Depiction of the LDA Topic Model. The Model takes a collection of document and by using an unsupervised Bayesian probabilistic model to draw a distribution over topics from a Dirichlet prior distribution. It groups words that are likely to be related into distinct clusters making up each distinct topic. Then, each topic consists of a so-called bag-of-words where each word within the bag has a certain weight. Finally, each document is assigned with a distribution of topics based on the word distribution within the document.

Figure 3 displays an illustration of the LDA Topic Model. In LDA, each document is represented as a mixture of topics, and each topic is represented as a distribution over the weighted bag-of-words. The generative process for LDA involves the following steps for each document d:

1. Draw a distribution over topics θ_d from a Dirichlet prior with hyperparameter α controlling the concentration of topics within documents

2. For each word i in the document, draw a topic z_{di} from the distribution θ_d

3. Draw the observed word w_{di} from the topic z_{di}'s word distribution $\beta_{z_{di}}$ a matrix of word distributions for each topic [46].

This process, known as the ”multinomial PCA” interpretation of LDA, again a similarity to the clustering algorithm K-means, corresponding to the dimensionality reduction step, can be expressed as:

$$\theta_d \sim Dirichlet(\alpha) z_{di} \sim Multinomial(\theta_d) w_{di} \sim Multinomial(\beta_{zdi}) \tag{2}$$

By summing over the topic assignments z, we can obtain the likelihood of observing word w_{di} given the topic proportions θ_d and the topic-word distributions β:

$$p(w_{di}|\theta_d, \beta) = \sum_k \theta_{dk} \beta_{kw} \tag{3}$$

Using LDA to reveal the underlying topics of our set of publications, one should aim to estimate the posterior distribution of the topics given the observed documents. However, this posterior cannot be computed directly. Instead, variational inference provides an approximate solution by introducing variational parameters γ and ϕ to approximate the true posterior distribution.

Online variational inference updates the variational parameters incrementally as new documents are observed. The goal is to maximize the Evidence Lower Bound (ELBO) with respect to these parameters. The ELBO is crucial in variational inference, serving as a lower bound on the log marginal likelihood of the data. By maximising the ELBO we're essentially making the model as good as possible at explaining the observed data, ensuring that it's at least as likely as the value given by the ELBO. This process involves iteratively adjusting the variational parameters for each document while keeping the topic-word distributions β fixed.

In LDA, the ELBO is expressed as:

$$ELBO =_q [log\, p(w, z, \theta, \beta)] -_q [log\, q(\theta, z, \beta)] \tag{4}$$

Maximising the ELBO with respect to the variational parameters γ and ϕ involves optimising the first term of the ELBO, which corresponds to the expected log joint probability under the variational distribution. That is, optimising the average (expected) of the log probabilities of both the observed data and the latent variables (topics) according to the variational distribution. This optimisation step aims to ensure that the variational distribution captures as much of the true posterior structure as possible.

Now, with the optimal variational parameters, γ and ϕ, obtained the next step is to update the actual topics λ to further maximise the ELBO. In this phase, we're aiming to set the topics λ in a way that maximises the ELBO while keeping the variational parameters fixed. This corresponds to maximising the following quantity:

$$L(n, \lambda) =_q [log\, p(n, \gamma, \phi, \lambda)] \tag{5}$$

Equation 5 represents the expected log probability of the data, documents, and topics under the current variational distribution.

After fitting the per-document variational parameters γ and ϕ then the topics λ are updated based on a weighted average of the current topics and the topics that would be optimal given the current document considering the entire corpus. This update process aims to maximise the ELBO, ensuring that the topics capture the underlying structure of the document collection [44].

Finally, a distribution of words per topic per document is obtained and we can start to identify the latent topic in our data.

Text Preprocessing

Before applying the LDA algorithm to our corpus of documents our data undergoes some preprocessing to ensure that our data is consistent and analysable. The text preprocessing includes the following steps:

1. **Tokenization:** Splitting the text into sentences and words, converting words to lowercase, and removing punctuation.

2. **Stopword Removal:** Removing all stopwords, which are commonly used words that do not contribute to the specific meaning of the text.

3. **Lemmatization:** Changing words in the third person to first person and converting verbs in past and future tenses to present tense.

4. **Stemming:** Reducing words to their root form. [47]

Additionally, several general issues in the text data (abstracts) that would add noise to the clustering results were addressed by stripping the abstracts of the following entities:

- **Links:** Abstracts often end with references to publication links. Any text containing URLs (identified by the presence of "https") is removed.

- **Copyright Notices:** Abstracts may include copyright symbols ©associated with universities. These are stripped from the text.

- **Latex Commands:** Abstracts may contain Latex commands, such as , , and , which are removed.

- **Chinese Characters:** Abstracts containing Chinese characters, which are not interpretable for our model, are stripped.

The *Abstract Clean* was saved in a new column and and used for the Topic Model Analysis.

Stopwords specific to the context and purpose of this task, namely identifying sub-fields of physics, were added to the NLTK predefined English stopwords list. These additional stopwords include:

1. **Institution or Affiliation** To avoid clustering based on institutional affiliations and reduce noise in the topic analysis.

2. **Geographic Entities** Similar to the above, to prevent clustering based on geographic locations mentioned in the text.

3. **Months** As some abstracts include publication months, removing them to avoid adding noise to the topic analysis.

4. **Author names** Removing author names to focus solely on the content of the abstracts.

5. **General Words** Words such as *paper, article, abstract*, and *intrudction* etc., as they do not contribute to the topic analysis. Additionally, the word "physics" is added to the list of stopwords, as it is likely to appear in most abstracts but does not provide further information about specific physics subfields.

Following preprocessing, the gensim *filter_extremes* function is applied to filter out tokens that appear in:

- Less than 15 documents (absolute number) or

- More than 0.5 documents (fraction of total corpus size)

- After the above steps, keeping only the first 100,000 most frequent tokens to create the final dictionary.

The *doc2bow* function is then applied to the dictionary to report the frequency of words for each document, resulting in a bag-of-words (BOW) corpus.

Finally, a tf-idf model object is created using *models.TfidfModel* on the BOW corpus. This model is then applied to the entire corpus, resulting in the TF-IDF weighted corpus, which is then ready for training the LDA model in order to predict the latent topics within our data [47].

LDA Model Optimisation

To optimise our LDA model, we use relevant metrics to test different hyperparameter settings, and determine the optimal number of topics.

Ideally, we would loop through a range of topics, α, and β but due to computational constrains, we chose a slightly different approach:

1. **Initial Topic Modeling with Default Hyperparameters:** Train multiple LDA models with varying numbers of topics using default hyperparameters. Evaluate each model's performance using relevant metrics such as coherence score and perplexity.

2. **Select Optimal Number of Topics:** Based on these results, identify the number of topics that provides the best performance according to the chosen evaluation metrics.

3. **Hyperparameter Optimization:** Once the optimal number of topics is determined, focus on optimising the hyperparameters by evaluating the results of several parameter settings and combinations.

4. **Final Topic Modeling with Optimized Hyperparameters:** Train the final LDA model using the optimal number of topics and the best hyperparameters obtained from the above steps.

The performance metrics used specifically for LDA Model Optimisation were Model Perplexity and Coherence Score [48].

Model Perplexity

Perplexity is an intrinsic evaluation metric widely used for language model evaluation [48]. It measures how well a probability model predicts a sample, particularly in Natural Language Processing tasks [49]. In essence, perplexity quantifies the probability that a model represents or reproduces the statistics of unseen data. Mathematically, it is given by:

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}} \tag{6}$$

Where w represents a word and N is the total number of words in the unseen data. Hence, minimising perplexity corresponds to maximising the probability of rightfully understanding and predicting the structure of the unseen data.

However, recent studies indicate that the perplexity score and human judgement may not always agree and that they may even be slightly anti-correlated. Hence, it serves as a guiding metric rather than an absolute truth in model evaluation [48].

Coherence Score

The Topic Coherence score measures the semantic similarity between high-weighted words within a topic, distinguishing interpretable topics from statistical artifacts, thus distinguishing between topics that make sense and those that might be random or unclear. [48].

The Gensim library provides a Coherence Score class that implements the c_v coherence model [50], which comprises four components:

1. **Segmentation:** Utilises the S-one-set method, comparing the words within each topic to the total word set in the entire corpus using word context vectors. Mathematically, it is given by:

   $$S_{set}^{one} = \{(W', W^*)|W' = w_i \in W; W^* = W\} \tag{7}$$

   and corresponds to computing how the words within a topic relate relate to all the words in the data.

2. **Probability Calculation:** Estimates the probability $P(w)$ of a word occurring and the occurrence probability $P(w_1, w_2)$ of pairs of words occuring togehter. It does so by using the Psw(110) method, which employs a sliding window approach.

3. **Confirmation Measure:** Employs an indirect confirmation measure m by comparing words within each pair W' and W^* against all other words W in the dataset. It computes a measure vector for each pair given by:

$$\hat{v}_m = \{ \sum_{w_i \in W'} m(w_i, w_j)\}_{j=1,2,\ldots|W|} \tag{8}$$

   Then, it calculates the cosine similarity between these vectors given by:

$$m_{nlr}(S_i) = \frac{m_{lr}(S_i)}{-log(P(W', W^*) + \epsilon} \tag{9}$$

   which helps determine how similar the words within a pair of words are compared to all other words.

4. **Aggregation:** Finally, it aggregates all confirmations of subset word pairs to derive a single coherence score, computed as the arithmetic mean σ_a of the confirmation measures [51].

The coherence score provides an overall indication of how coherent and meaningful the topics extracted by the LDA model are [50].

Prestige Markers

To explore the relationship between gender and prestige markers, it's essential to define these markers. Initially, we considered collecting Journal Impact Factors (JIF) for each journal represented in the dataset. However, we soon realised that this task would be exceedingly time-consuming, given our dataset spanning more than 50 years and encompassing numerous journals.

As an alternative, we used the the number of citations as a proxy of prestige. We defined two types of rankings: University Ranking and Journal Ranking. These rankings are calculated as follows:

$$\frac{\#\ citations}{\#\ articles}/university\ or\ journal\ or\ author \tag{10}$$

These markers provide an estimation of prestige and were used to investigate relations between prestige and gender, as well as other variables.

Activity and Survival Rate

In addition to investigating gender-related patterns, we aimed to analyse author activity within the field of physics and their survival rates as researchers. Key variables for this part of the analysis include:

- **Author Publication Count:** The total number of publications attributed to each author.
- **Career Span Years:** The duration of an author's publishing career, calculated as the difference between the year of their first publication and the year of their last publication.
- **Author Publication Rate:** The rate at which an author publishes articles within their career duration, calculated as the ratio of the author's publication count to their career span years.
- **Event Observed:** A binary variable indicating whether an author is currently active as a researcher within physics.

To define whether an author is active or not, we established the following assumptions:

1. An author should have their first publication in 1975 or later. If their initial publication occurs more than five years into the dataset, we consider it their first publication within their overall career.

2. An author should have their last publication in 2018 or later. If the author's most recent publication falls within the five most recent years, we assume they are still active; otherwise, we consider them inactive from the year of their last publication.

With these variables defined, we conducted a survival analysis using the Kaplan-Meier Estimator. This non-parametric statistic estimates the survival function from lifetime data, measuring the probability that an author will remain active past a certain time. The Kaplan-Meier Estimate is defined as:

$$\hat{S}(t) = \Pi_{t_i<t} \frac{n_i - d_i}{n_i} \tag{11}$$

where $\hat{S}(t)$ is the estimated survival probability at the time t, n_i is the number of authors at risk of death just prior to the time t, and d_i is the number of death events observed at time t. [52].

Additionally, we use the Log-rank test to compare the survival probability between each gender class. This test assesses whether the hazard functions of the two groups are identical, indicating equal likelihoods of survival. The log-rank test statistic is defined as:

$$Z_i = \frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \tag{12}$$

where Z_i is the log-rank statistic, $O_{i,j}$ is is the observed number of events in group i at time j, $E_{i,j}$ is the expected number of events in group i at time j, $V_{i,j}$ is the variance of the observed events in group i at time j, and J represents the number of observed event times [53].

The survival probability analysis provide insights into author activity patterns and potential gender-related differences in career drop-outs within the physics field.

Classification

In order to investigate the level of separability between female and male authors given our key variables we use a classification model. Classification is a supervised machine learning technique where the model aims to predict the correct label or category for a given input data point based on the observed and labelled data [54].

The classification process involves the following steps:

1. **Training Phase:** In this phase, the model is trained using a labelled dataset, also known as the training data. During training, the model learns patterns and relationships between the input features and the corresponding class labels.

2. **Evaluation Phase:** After training, the model is evaluated using a separate dataset called the test data. This allows us to assess the model's performance on unseen data and determine its accuracy and generalisation ability.

3. **Prediction Phase:** Once the model has been trained and evaluated, it can be deployed to make predictions on new, unseen data. The model takes the input features of the new data points and predicts the corresponding class labels [55].

In our specific case, we aim to use classification to predict author gender based on other variables in our dataset. The extend to which the model succeed at classifying the test data reveals the separability of the genders and how well they can be classified given other variables except the gender. To accomplish this task, we use the *LGBMClassifier()* from the LightGBM library [56]. This classifier was chosen for its efficiency and effectiveness in handling large datasets with high-dimensional features [57].

Before going into classification, it is essential to understand decision trees, which form the basis of many classification algorithms.

Decision Trees

Decision Trees serve as a non-parametric supervised learning method used for classification and regression tasks. The core aim of a decision tree is to construct a model that predicts the value of a target variable by deriving simple decision rules from the input features of the data [58].

Figure 4 depicts an example of a decision tree including explanatory labels. Within a decision tree, each internal node represents a "test" on an attribute, with each branch indicating the outcome of the test, and each leaf node denoting a class label or numeric value. By splitting the data into subsets based on feature values, the decision tree aims to create subsets that are more uniform with respect to the target variable, essentially characterising the features and their values leading to the target variable [59].

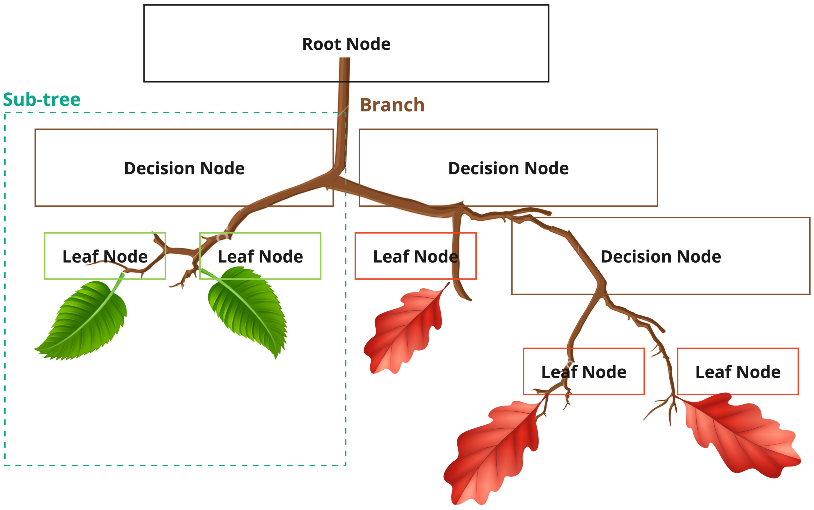


Figure 4: Depiction of a Decision Tree. The Root Node serves as the initial test attribute where to each branch corresponds to the outcome of that test with each node providing an ”intermediate” response and each leaf node the ”final answer” to the question tested from the root node. The decision tree can be though of as a map of how the root node is characterised in terms of each of the other feature nodes.

An elementary concept in decision tree learning for classification tasks is the Gini index [60]. This index serves as a measure of impurity or inequality within a sample, representing the probability of incorrectly classifying random data if it were labelled based on the class distribution of the dataset [59]. The Gini index ranges between 0 and 1, where 0 indicates complete homogeneity (all points in the data belong to one class), and 1 indicates maximal inequality (elements are evenly distributed across all classes) [61].

Mathematically, it is given by

$$Gini\ Index = 1 - \sum_{i=1}^{n} p_i^2 \tag{13}$$

where n is the number of classes, and p_i is the probability of a data point belonging to the i-th class .

In decision tree classification, the algorithm aims to minimise the Gini index at each node by selecting the attribute that results in the best split, leading to the most homogeneous subsets creating the terminal leaf node, meaning that the test attribute leads to a homogeneous result between terminal nodes. Think of it as if you pose a question, the terminal leaf nodes will present distinct (homogeneous) answers such as ”yes” or ”no” which will be true for each data point belonging to a given class. [61].

Standardisation and Centering

Before applying any machine learning technique, it’s essential to standardise the data to ensure that each variable contributes equally to the analysis no matter scale. That is, say we want to compare the weight of a bikes number of gears (count) to its weight (kg) in determining the price, then weight would always be evaluated higher than number of gears as its unit exists on a complete different scale. Standardisation involves centering the mean and scaling the data to have a unit variance. This process ensures that the features are transformed to the same scale, preventing variables with larger magnitudes from dominating the analysis.

The standardisation process is carried out using:

$$z = \frac{x - \mu}{\sigma} \tag{14}$$

where x is the value of the given data point, μ its mean and σ its standard deviation [62].

To standardise the features, we adjust the data to have a mean of 0 and a standard deviation of 1 [63]. This standardisation is achieved using the *StandardScaler* module available in the scikit-learn library for Python [62].

LightGBM

LightGBM is a powerful ensemble learning framework, specifically designed as a gradient boosting method. It constructs a robust learner by iteratively improving upon the mistakes of previous learners. At its core, LightGBM employs Gradient Boosting Decision Trees (GBDT), a widely-used machine learning algorithm. This approach involves combining many weak learners (decision trees) in a sequential manner. Each subsequent model optimizes and corrects the errors of its predecessors by fitting to the residuals of the preceding models. This iterative process aims to minimize a predefined loss function and ultimately produce highly accurate predictions. [64].

In LightGBM, the decision tree nodes are split at the most informative features, leading to maximum gain in evidence. This gain is quantified by measuring the variance improvement after partitioning the data. By sequentially adding subsequent decision trees, each of which contributes to the final prediction based on an appointed learning rate. The final prediction formula can be interpreted as the contributions of each tree, scaled by the learning rate:

$$Y = Base\,tree(X) - lr \cdot Tree_1(X) - lr \cdot Tree_2(X) - lr \cdot Tree_3(X) \tag{15}$$

where Y is the prediction, $Base\,Tree(X)$ is the base decision tree prediction, and $Tree_1(X)$, $Tree_2(X)$, $Tree_3(X)$ are subsequent decision trees. The learning rate lr controls the contribution of each tree to the final prediction.

Determining optimal splits within decision trees, LightGBM calculates variance gain. This metric measures the improvement in variance after dividing the data based on a particular feature and value. By identifying the most informative splits, LightGBM enhances its predictive power and overall performance.

The variance gain of dividing measure j at a point d for a node is calculated as follows:

$$v_{j|0}(d) = \frac{1}{n_O}\left(\frac{(\sum_{\{} x_i \in 0 : x_{ij \leq d} g_i)^2}{n^j_{l|0}(d)} + \frac{(\sum_{\{} x_i \in 0 : x_{ij > d} g_i)^2}{n^j_{r|0}(d)}\right) \tag{16}$$

where n_O is the total number of data points in the node, $n^j_{l|0}(d)$ is the number of data points with feature value less than or equal to d, $n^j_{r|0}(d)$ is the number of data points with feature value greater than d and g_i is the gradient of the loss function with respect to the prediction at point x_i [57].

By employing LightGBM we can test the models ability to classify each gender class in our data using sophisticated machine learning techniques.

Performance Metrics

For classification model evaluation, we use relevant performance metrics to asses the effectiveness of the model. We use F1 Score, Receiver Operating Characteristic (ROC) Curve, Precision-Recall Curve, and Confusion Matrix. These metrics are derived from fundamental measures including Precision, Recall (also known as Sensitivity or True Positive Rate), and Specificity (also known as True Negative Rate) given by:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

and

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \tag{18}$$

and

$$Specificity = \frac{FP}{FP + TN} \tag{19}$$

where TP = True Positives, FP = False Positives, TN = True Negatives and FN = False Negatives [65, 66, 67, 68, 69]. These metrics provide insights into the model's ability to correctly identify positive and negative instances.

F1 score

The F1 score, which represents the harmonic mean of precision and recall, offers a single score of a combined metric optimising both aspects of model performance. It is computed by summing the reciprocals of precision and recall values and then taking the reciprocal of this sum, as described mathematically:

$$F1\,Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{20}$$

[66].

This metric provides a balanced assessment of the model's ability to correctly identify positive instances while minimising false positives and false negatives and thus provides a trade-off between Precision and Recall [66].

An F1 Score > 0.9 is considered excellent, while scores falling between $0.8 - 0.9$ are deemed good. Scores ranging from $0.5 - 0.8$ are considered average. A score below 0.5 indicates poor model performance [70].

We use the F1 score as metric when testing different hyperparameter settings for the LightGBM classification model as well as for evaluating the final classification of gender classes obtained by the model.

ROC Curve

The Receiver Operating Characteristic (ROC) Curve is a graphical tool used for assessing classification models. It presents the trade-off between Sensitivity and Specificity across varying classification thresholds [67].

The ROC curve is constructed based on a "separator" scale, where results for the classes form two overlapping distributions. Complete separation of these distributions signifies a perfectly discriminating test, while complete overlap suggests no discrimination [71]. The proximity of the ROC curve to the upper left corner of the graph indicates higher test accuracy. In this corner, sensitivity equals 1 and the false positive rate equals 0 (specificity equals 1), representing ideal performance. Hence, the ideal ROC curve has an Area Under the Curve (AUC) of 1.0 [67].

Figure 5 provides an illustration of a ROC Curve with explanatory labels.

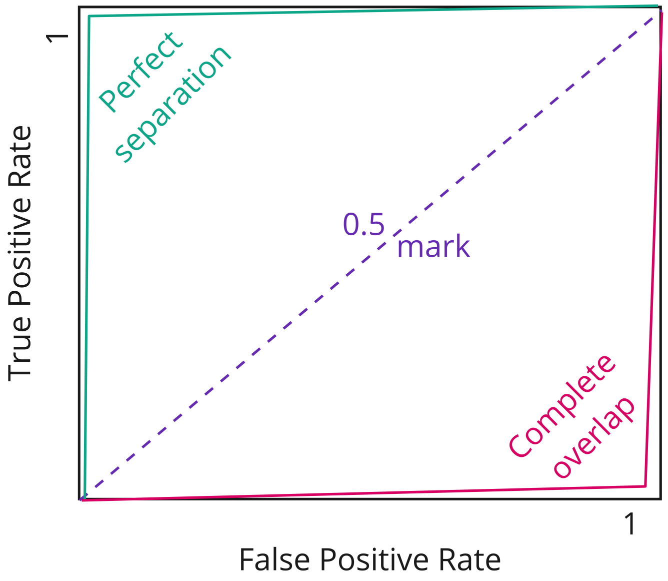


Figure 5: Roc Curve illustration with explanatory labels.

We use the ROC curve for evaluating the final classification of gender classes obtained by the model.

Precision-Recall Curve

The Precision-Recall Curve showcases the balance between Precision and Recall across different classification thresholds, similarly to the ROC curve. A larger area under the curve (AUC) signifies superior performance, reflecting high Precision and Recall. The staircase-like appearance of the plot highlights the sensitivity of Precision and Recall to threshold adjustments [68].

The PR-AUC quantifies how effectively a model can differentiate between classes, considering both its ability to accurately classify negative samples as negative (Precision) and similarly to correctly identify positive samples (Recall). A higher PR-AUC value suggests a better-performing model [72].

Figure 6 provides an illustration of a Precision-Recall Curve with explanatory labels.

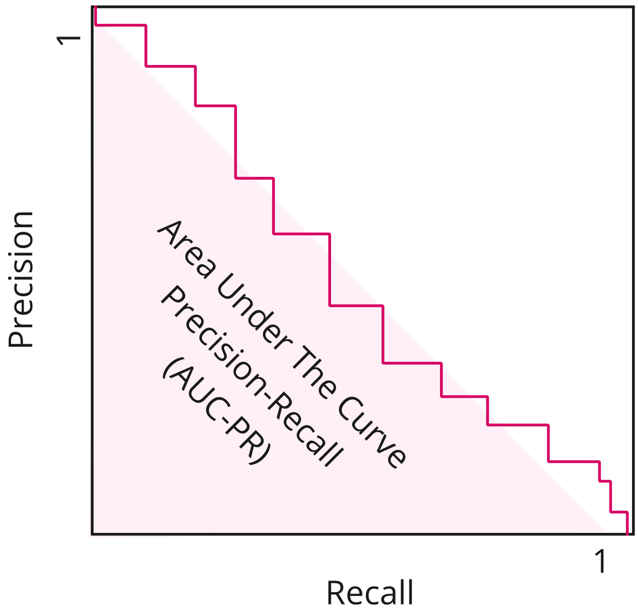


Figure 6: Precision-Recall Curve illustration with explanatory labels.

We use the Precision-Recall Curve for evaluating the final classification of gender classes obtained by the model.

Confusion Matrix

The Confusion Matrix offers a detailed summary of the model's performance by presenting the counts of TPs, FPs, TNs and FNs.

Figure 7 provides an illustration of an Confusion Matrix with explanatory labels.

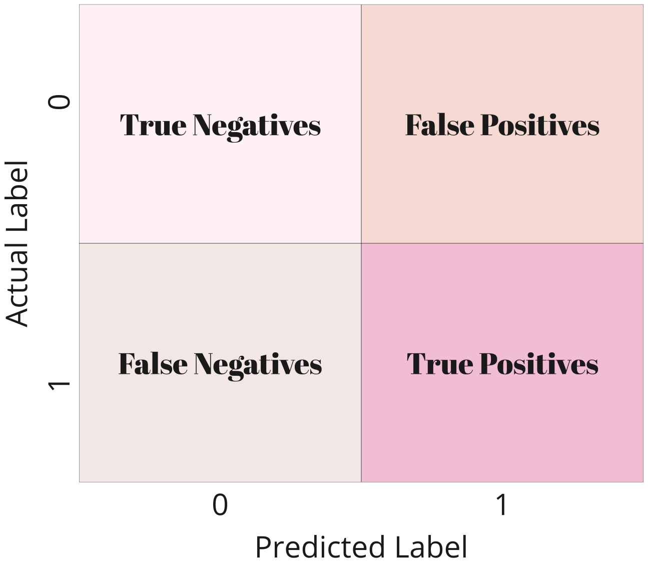


Figure 7: Confusion Matrix illustration with explanatory labels.

We use the Confusion Matrix for evaluating the final classification of gender classes obtained by the model.

Feature Importance

Feature importance quantifies the relevance of each feature in predicting the target variable. Within Scikit-learn, this is computed by considering the reduction in node impurity weighted by the probability of reaching that node.

In each decision tree, the importance of a node j is determined using the Gini Importance method, assuming a binary tree structure:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{21}$$

where ni_j is the importance of node j, w_j is the weighted number of samples reaching node j, C_j is the impurity value of node j, $left_j$ and $right_j$ are the child nodes resulting from splitting node j.

Then, the importance of each feature in a decision tree is computed by aggregating the importance values of nodes that split on that feature:

$$fi_i = \frac{\sum_{j:node\, j\, splits\, on\, feature\, i} ni_j}{\sum_{k \in all\, nodes} ni_k} \tag{22}$$

where fi_i is the importance of feature i, and ni_j is the importance of node j.

These feature importance scores are subsequently normalised to a range between 0 and 1 by dividing each score by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\, features} fi_j} \tag{23}$$

Finally, at the Random Forest level, the overall feature importance is obtained by taking the average of the the normalised feature importance values across all trees:

$$RFfi_i = \frac{\sum_{j \in all\, trees} normfi_{ij}}{T} \tag{24}$$

where $RFfi_i$ is is the importance of feature i calculated from all trees in the Random Forest model, $normfi_{ij}$ is the normalised feature importance for feature i in tree j and T is the total number of trees in the Random Forest model [60].

In our case, the features with the highest feature importance reveal the features which value are most important for classifying and thereby distinguishing between the two genders.

Clustering

Another way to identify the genders and how easily each gender is recognised based on the features describing them is by grouping the genders into clusters. Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) [73].

Similarly to the classification preprocessing, we also standardised our data before undergoing cluster analysis.

UMAP

Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction algorithm aimed at finding a lower-dimensional representation of a dataset while preserving its inherent structure. The goal is to map high-dimensional data points onto a lower-dimensional space while maintaining both local and global relationships among the data points.

Mathematically, UMAP treats data points as samples drawn from a Riemannian manifold, which is a space locally resembling Euclidean space but capable of representing more complex geometries. The algorithm constructs an embedding function ϕ to map these data points from the manifold into a lower-dimensional Euclidean space $\mathbb{R}^m$.

UMAP offers several advantages over other dimensionality reduction techniques. It does not assume linearity in the data, making it suitable for nonlinear datasets. It is computationally efficient and faster compared to alternatives. Moreover, UMAP preserves both local and global structures of the data and can handle large datasets efficiently, including high-dimensional and sparse datasets.

We perform the UMAP analysis and visualisation by emplying the python *UMAP()* function from the umap-learn package. While UMAP is commonly used for standard unsupervised dimensionality reduction, it can also be extended for supervised dimensionality reduction tasks, such as clustering, and metric learning [74, 75].

In our project, we employ UMAP for dimensionality reduction and supervised clustering of female and male authors. We use UMAP results to explore the distributions between gender classes and assess their separability.

Fishers Linear Discriminant Analysis

To assess the degree of separation achieved by the UMAP embeddings and LightGBM classification, respectively, we use Fisher's Linear Discriminant Analysis (FLDA). FLDA aims to reduce the dimensionality of the data to one dimension in a manner that maximises the separation between classes.

The fundamental concept behind FLDA is to find a linear transformation that maximises the separation between classes while minimising the variance within each class. Given a d-dimensional vector $\vec{x} \in \Re^d$, FLDA searches for a one-dimensional projection $\vec{w}^T\vec{x}$, where $\vec{w}$ represents a vector of weights. Mathematically, this projection is defined as:

$$\vec{z} = \vec{w}^T\vec{x} = \sum_{j=1}^{d} \vec{w}j\vec{x}j \tag{25}$$

[76]

By applying Fisher's LDA, we can reduce the initial dimensionality to one dimension and obtain the Fisher weight. This enables us to visualise and evaluate the level of separation between female and male authors.

Label Verification

To verify that our labelling of gender and topic is somewhat reliable we do some quality checks.

Gender Assignment

To sanity check the gender label assigned by the gender API we select 20 danish authors from the Niels Bohr Institute. These are authors that I am familiar with and therefore know their gender. We select a random sample of 20 authors and check how many of their genders that were assigned correctly.

Topic Model

To ensure that the label given by the Topic Model is (somewhat) reliable we look at the best matching abstract within each topic and verify whether the assigned label seems reasonable or not. While this does not guarantee a perfect match between topic label and actual topic it provides us with an idea of how well the topic and abstract text match.

Results

Presenting the results we differ between grouping by distinct authors by referring to "Author Level" (meaning unique authors but publication duplicates for publications with multiple authors) and including author duplicates by referring to "Authorships Level" (meaning duplicate of authors for every publication the author appears on as well as duplicate of articles for every publication with multiple authors). Lastly when grouping by distinct publications we refer to "Publication Level". It is indicated for each section at which level the results were obtained.

Gender Assignment Updating

The original dataset consisted of 2,000,281 rows, whereof 518,351 were unique publications by 810,843 unique authors. The author's gender was initially determined using the gender API, specifically the *gender-guesser* module, as described in Section **Gender API**. The assigned gender is saved in a new column named *Predicted Gender*. The results in this section were obtained at "Authorships Level".

Initial Assignment

The initial gender assignment predicted 44% of the authors to be of "unknown" gender, 35% of "andy" gender, 17% to be "male" and the last 4% were distributed between "female", "mostly male" and "mostly female".

Figure 8(a) illustrates the distribution of the initially predicted genders, while Figures 8(b) and 8(c) provide insights into the characteristics, in terms of author names and countries, for the most frequently occurring cases where the gender assignment is not interpretable as binary (female or male).

Figure 8(a) indicates that more than 50% of the predictions fall into the "andy" or "unknown" categories. Since gender identification is crucial for this study, a series of updates are performed to increase the number of authors with interpretable genders, primarily categorised as (mostly) female or male. These updates include:

1. **Update According to Duplicates in Author ID with an Assigned Gender**
2. **Update Without Country**
3. **Drop Rows With Invalid Names**
4. **Update Names Including Special Characters**
5. **Update None Names**
6. **Remove '.', ',', and '-' From Names and Update Gender**
7. **Update With Gender of Identical Name**
8. **Update Using *Gender API***
9. **Final Drop Rows With Invalid Names**

Following each update, the corresponding figures similar to Figure 8 are generated. However, due to space constraints, the figures for the subsequent 8 updates will be presented in . Please refer to the Appendix to view the figures referenced in the subsequent subsections.

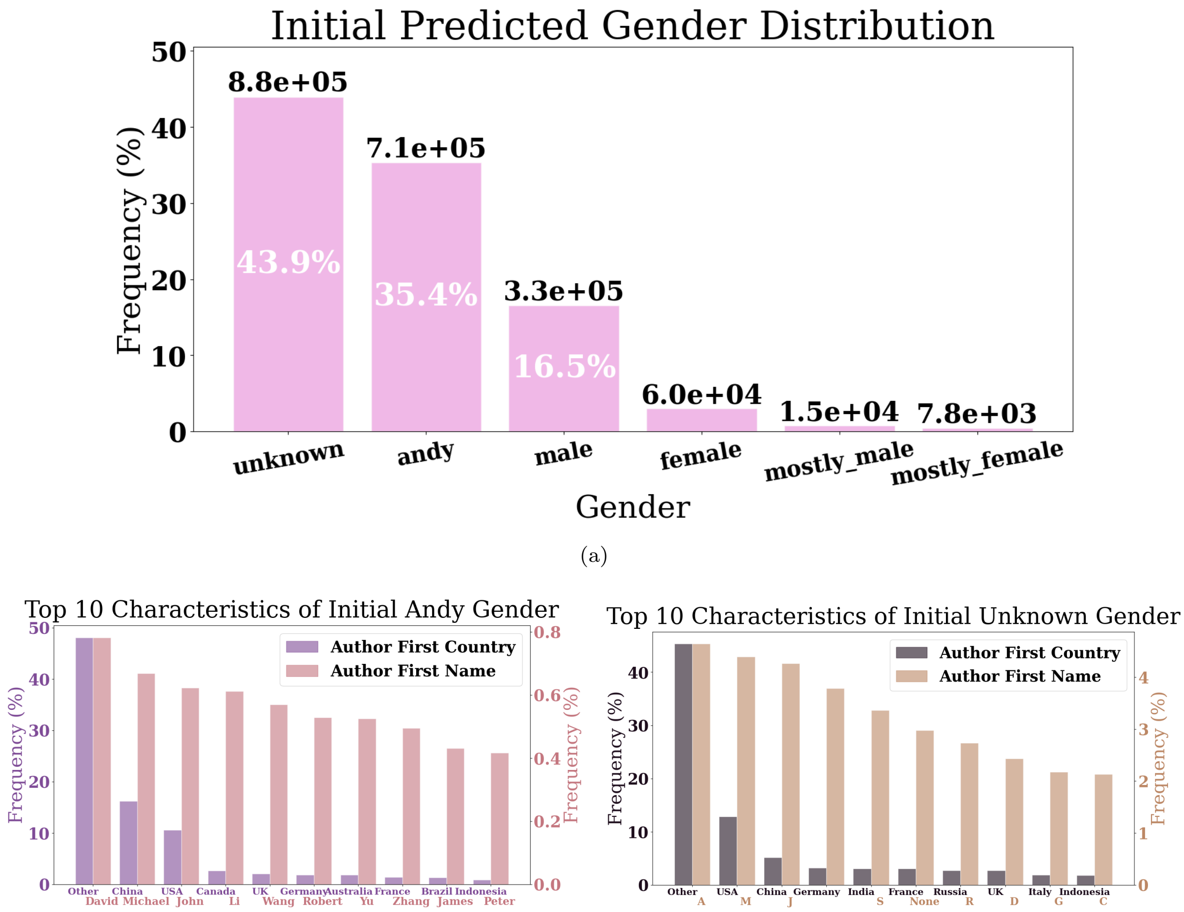
Table 4 shows the difference in the count for the *Predicted Gender* between each of the updates. The detailed descriptions of each update is provided in Section **Update1-Update9**.

Update 1: Update According to Duplicates in Author ID with an Assigned Gender

The distribution of predicted gender after the first update is shown in Figure A1 in Appendix A. In the first update, we developed a function which adjusted the predicted gender based on the most common value of the predicted gender associated with each unique author ID. This function identifies the most common gender value within each unique author ID and replaces all other gender values for that ID with the most occurring one. That is to make sure the assigned gender is aligned between duplicates of a distinct author ID. However, if the most common value is andy or unknown it takes the second most common value in order to get an interprteable label - female or male. Finally, it returns the updated DataFrame.

Table 4: This table highlights the change in the distribution of predicted gender categories after applying each update. The count difference represents the difference in the number of entities assigned to each gender category. The update number is labelled for each column where Δ corresponds to the difference in count between the given update and the count from the previous state. If the count of a given gender has decreased it is marked in red and if it has increased it is marked in green. If the change is zero or close to none (< 100) it is marked in grey.

Difference in Predicted Gender Between Updates	
Gender	**Count Difference**
Δ Update 1	
Unknown	-151,168
Andy	-39,498
Male	+158,860
Female	+22,270
Mostly Male	+6,480
Mostly Female	+3,056
Δ Update 2	
Unknown	-5,475
Andy	-416,085
Male	+318,475
Female	+77,980
Mostly Male	+18,083
Mostly Female	+7,022
Δ Update 3	
Unknown	-390,872
Andy	-40,720
Male	-85,228
Female	-10,983
Mostly Male	-2,899
Mostly Female	-1,127
Δ Update 4	
Unknown	-2,013
Andy	-17,938
Male	+15,869
Female	+2,940
Mostly Male	+1,082
Mostly Female	+60
Δ Update 5	
Unknown	-1,021
Andy	+650
Male	+165
Female	+148
Mostly Male	+27
Mostly Female	+31
Δ Update 6	
Unknown	-13,110
Andy	+13,008
Male	+84
Female	+7
Mostly Male	+3
Mostly Female	+8
Δ Update 7	
Unknown	-18,361
Andy	-66,166
Male	+45,925
Female	+16,338
Mostly Male	+10,101
Mostly Female	+12,163
Δ Update 8	
Unknown	-8,047
Andy	-46,663
Male	+11,216
Female	+22,270
Mostly Male	+1
Mostly Female	+7
Δ Update 9	
Unknown	-35,544
Andy	0
Male	-1
Female	0
Mostly Male	0
Mostly Female	0



(b)

(c)

Figure 8: Distribution of predicted gender across the dataset as initially assigned by the gender API gender-guesser (see Section *Gender API*). **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories.**(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Table 4 reveals a significant decrease in ”andy” and ”unknown” genders, while there’s an increase in ”male” and ”female” categories after the first update. Despite this improvement, more than 50% of the dataset is still flagged as ”andy” or ”unknown”, and so we proceed with further updates.

Update 2: Update Without Country

The distribution of predicted gender after this second update is depicted in Figure A2 in Appendix A.

Upon examining the characteristics of authors assigned the gender label ”andy” (see Figure 1(b)) we observed that prevalent names included ”David” and ”Michael.” These names are globally common names, and therefore we would assume a gender should be recognisable. But looking at the most occuring country label in Figure 1(c) we see that it is ”other”, suggesting that the gender assignment of ”andy” might be due to the gender-guesser API’s lack of training data for certain countries. Similarly, authors labelled with the gender ”unknown”, Figure 1(b), often had single-letter names, making it difficult for the API to assign a gender.

To refine the gender predictions, particularly for authors initially flagged as ”andy”, we implemented a second update. This update focused on re-predicting the gender based solely on the first name while excluding the country information. It also excluded single-letter names, which were prevalent among authors with unknown gender labels. This function iterated through each row, updating gender predictions for rows labeled as ”andy” or ”unknown.”

Table 4 indicates a significant decrease in the count of authors flagged as ”andy”. Following the second update, the prevalence of ”andy” and ”unknown” genders were reduced to counting for less than 50% of the dataset. Despite this improvement, close to half of the dataset remained assigned with a non-binary gender label, prompting further refinement in the third update.

Update 3: Drop Rows With Invalid Names

The distribution of predicted gender after the third update is shown in Figure A3 in Appendix A.

Upon reviewing Figure 2(b) in Appendix A, we observed that a significant portion of authors assigned the ”andy” gender label were Chinese names. This is potentially attributed to the androgynous nature of many Chinese names, where gender determination often depends on the name's perceived gender association or specific character usage [77]. We will get back to dealing with this but for this update, we focused on refining the gender predictions for authors initially assigned the ”unknown” gender label.

The prevalence of ”unknown” gender labels was primarily attributed to single-letter names, Figure 2(c). Hence, for the third update, we opted to remove rows containing single-letter first names. Since assigning gender to single-letter names is unfeasible and these authors would be excluded from the analysis later, dropping such rows resulted in a dataset reduction to 1,468,452 rows.

Table 4 highlights an overall reduction in gender predictions, due to the drop of single-letter names. The reduction is primarily driven by the decrease in the ”unknown” gender group as intended. We proceed to update 4.

Update 4: Update Names Including Special Characters

The distribution of predicted gender after the fourth update is displayed in Figure A4 in Appendix A.

The remaining authors with unknown genders predominantly have names containing special characters, such as Ø in ”Bjørn” or é in ”André”, Figure 3(c). Hence, Update 4 aims to address gender assignment in such cases.

To address this, we first examine the authors' *Raw Author Name* and *Author Display Name* in order to determine if a better method exists for extracting names with special characters. Additionally, we investigate how the gender guesser API handles such characters.

Testing confirmed that the gender guesser API handles special characters fine. However, the issue stemmed from the code used for the extraction of the first name from author data, particularly those inclduing special characters. For the fourth update, we revisited the extraction process for ”unknown” and ”andy” rows, ensuring proper handling of special characters. Subsequently, we reassigned gender based on the extracted first name and country, excluding ”unknown” or ”other” countries as they may cause assignment inconsistencies as described in Section **Update 2**.

The fourth update successfully addressed names with special characters and reduced their presence in the ”unknown” gender category as seen in Table 4. Notably, a significant decrease in the ”andy” and ”unknown” gender groups were achieved. We proceed to update 5.

Update 5: Update None Names

The distribution of predicted gender after the fifth update is shown in Figure A5 in Appendix A.

Figure 4(c) shows that authors flagged as ”unknown” gender had a ”None” name. To address this, we examined the nature of authors with the name ”None” and found instances where it appeared either as their entire name or just as part of their full name.

Therefore, for the fifth update, we refined the process of extracting the first name, excluding instances where the name is listed as ”None.” This was achieved through a function that ensured ”None” was excluded during name extraction. Subsequently, we updated the predicted gender using the revised first name extraction method, following the same criteria as in update 2 and 4, where country information is only considered if valid.

The fifth update effectively reduced the prevalence of ”None” names in the unknown gender category as presented in Table 4, however it hasn't eliminated it completely, why we continue to Update 6.

Update 6: Remove '.', ',', and '-' From Names and Update Gender

The distribution of predicted gender after the sixth update is displayed in Figure A6 in Appendix A.

Following Update 5, it is clear from Figure 5(c) that the ”unknown” gender category still included names with punctuation marks like '.', ',', or '-', which were probably causing the misclassifications. Moreover, the same names but without punctuation appeared in the ”andy” gender category, as depicted in Figure 5(b). To ensure accurate gender assignment and eliminate discrepancies caused by such characters, we proceed to remove these characters from the end of names and update the predicted gender accordingly.

The sixth update involved stripping trailing '.', ',', or '-' characters from names and revising the predicted gender, applying the same condition as in Updates 2, 4 and 5, where country information is excluded if invalid.

In Table 4 we see that Update 6 successfully removed names with trailing punctuation marks from the unknown gender category, and realigned them with the flag andy gender category of the same names seen in Figure 6(b). We proceed to update 7.

Update 7: Update With Gender of Identical Name

The distribution of the predicted gender after the seventh update is shown in Figure A7 in Appendix A.

Following Update 6, the results revealed that the unknown gender category, depicted in Figure 6(c), primarily consisted of invalid names that therefore were challenging to assign with a gender. Furthermore, despite the improvement following Update 1-6, the "andy" gender category still predominantly comprised Chinese names, which, Figuer 2(b), which therefore required further attention.

Therefore, in Update 7, we introduced a function that targeted cases within the "andy" and "unknown" gender categories. This function identified instances of a mislabelled author which shared the same name and country with another classified author. Then, it updated the gender of the mislabelled author so that it matched that of the gender labelled author. In case multiple authors had the same name and country but have different assigned genders, the function selected the most common gender value for that name, irrespective of country. If there was a tie between the most common gender values, no update was made.

From Table 4, we see that Update 7 successfully decreased the amount of "unknown" and "andy" genders. We proceed to update 8.

Update 8: Update Using *Gender API*

The distribution of predicted gender after the eighth update is shown in Figure A8 in Appendix A.

From update 7 we see that most occurring names of authors flagged as "andy" were Chinese names, Figure 7(b). For the "unknown" gender category, Figure 7(c), the names were also of Asian origin or non-sense names such as "Null" or "J.m". To specifically target these cases, we selected the top occurring names flagged as "andy" or "unknown" and predicted their gender using a different gender API, namely *Gender API*. As can be seen from the gender API market study 2, Gender API is trained on a way larger dataset than the API that we used initially. However, Gender API only allowed a limited number of request why we could only use it for a couple hundred of the most appearing names in the andy/unknown group.

As can be seen from Table 4 The predicted gender distribution following Update 8 successfully decreased both the occurrence of "andy" and "unknown". We proceed to update 9.

Update 9: Final Drop Rows With Invalid Names

The distribution of predicted gender after the ninth update is shown in Figure A9 in Appendix A.

Inspecting the results from Update 8, Figure 8(c), we see that the authors of "unknown" were due to names such as "Null" or "J.m". The rest of the names in both the "andy", Figure 8(b), and unknown group seemed valid but *gender-guesser* just doesn't know the gender. Therefore, the final update was to drop the entries including invalid names such as "Null" or names with only a single character before a punctuation. Dropping such rows resulted in a dataset reduction to 1,432,907 rows.

Final Predicted Gender Distribution

Figure 9 displays the initial gender assignment distribution and the distribution obtained proceeding through **Update1** to the final update **Update9**.

As can be seen from Figure 9 and Table 5 we succesfully increased the fraction of authors flagged as "male" (58% vs 17%), "mostly male" (3% vs > 1%), "female" (13% vs 3%), and "mostly female" (2% vs > 1%). Simultaneously, we reduced the fraction of "andy" (35% vs 7%) and "unknown" (44% vs 18%) gender labels.

Table 5 depicts the change in percent point between the original and final gender label count as depicted in Figure 9.

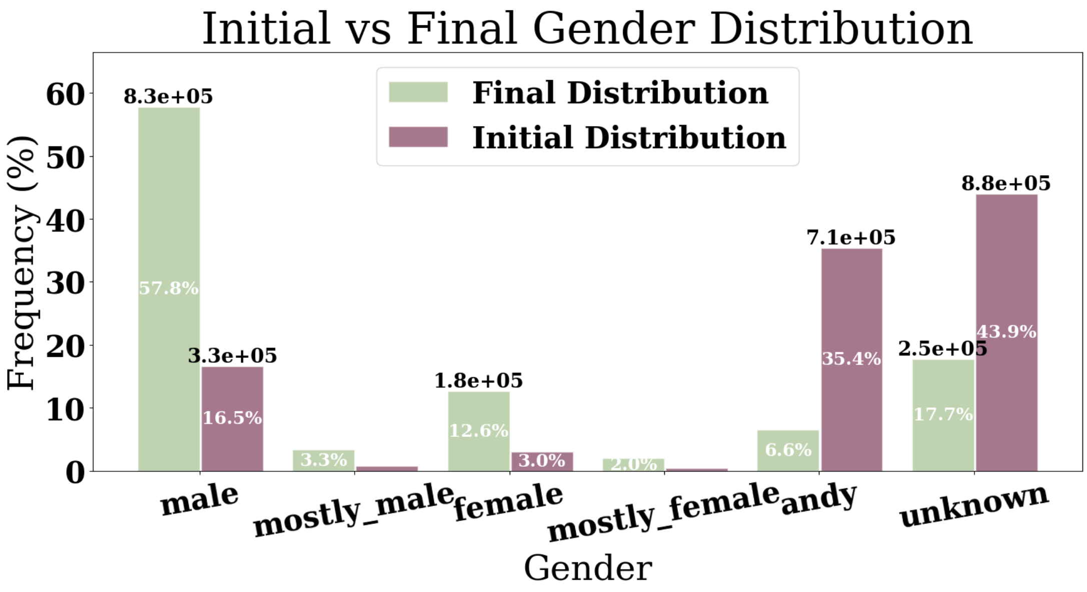


Figure 9: This figure illustrates the change in gender label frequencies between the initial and final gender assignments after 9 updates. The objective was to maximize the number of gendered labels (excluding "andy" or "unknown"). The plot highlights the success of the updates, with the first four bars (mainly male and female after updates, shown in green) significantly increasing compared to the initial distribution (depicted by red bars). Similarly, the last two bars (andy and unknown, shown in green) notably decrease compared to the initial distribution (red bars).

Table 5: Difference in percent point for each gender label between the original and final gender assignment corresponding to before and after completing nine updates targeting an increase in (mostly) male and female labels. A positive change is marked in green and a negative change is marked in red. We see that the gender label updating successfully increased the number of gendered labels, (mostly) female and male, as well as decreased the number of andy and unknown labels. PP stands for Percent Point.

Difference in Predicted Gender Between Original and Final Assignment	
Gender	**PP Difference**
Male	+41
Mostly Male	+3
Female	+10
Mostly Female	+2
Andy	-29
Unknown	-26

Binary Gender Assignment

The final distribution of binary gender in our dataset is depicted in Figure 10. Having obtained our final gender assignment, we create a new column *Binary Gender* in which we save all values of "female" and "mostly female" as 0 (Female) and "male" and "mostly male" as 1 (Male). Gender categories falling outside the binary labels are replaced with "NaN".

Evaluate Ratio Between Genders

Figure 11 show the distribution between binary gender before and after updating. We compare the ratio before and after updating in order to ensure that the updating process did not lead to a sort of "cherry-picking" and created an abundance of male or female labels compared to the initial distribution which we assume is statistically reliable.

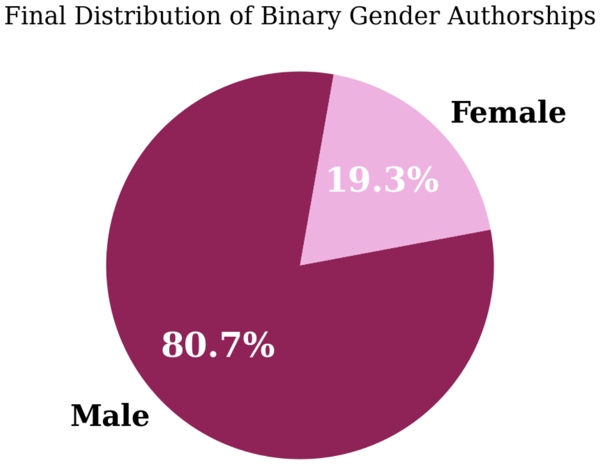


Figure 10: Pie chart depicting the final distribution between female and male authorships in our dataset.

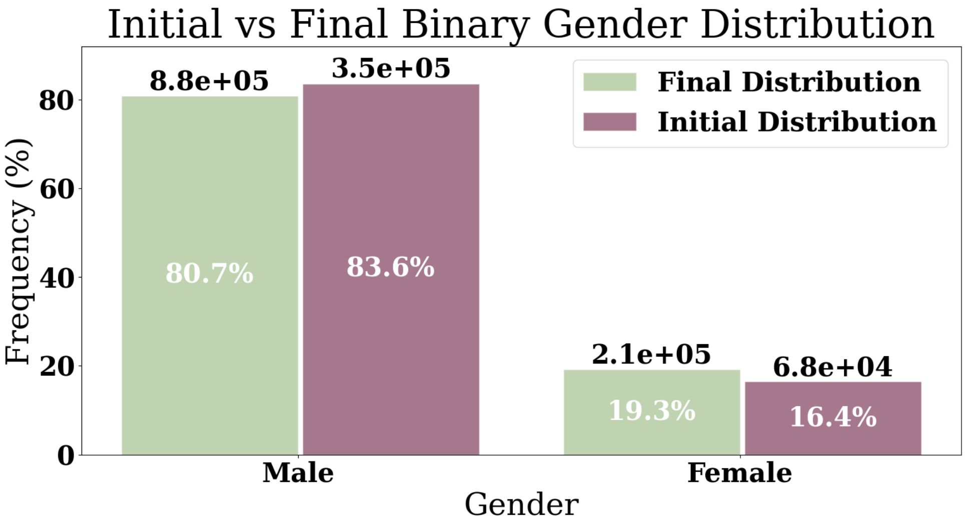


Figure 11: This figure shows the difference in distribution between male and female authorships for the orignal (marked by the red bars) and final (marked by the green) gender assignment as according to before and after updating. The figure shows that the ratio between male and female authors is around $80-20$ both before and after the update which guarantees that we haven't been "cherry picking". However, it is higher for females after the updating so female labels might be overestimated but we accept the final distribution as a reasonable depiction of female and male authorships.

Figure 11 shows us that the distribution between female and male labels were somewhat consistently $80-20$ before and after the update. However, the ratio after the update slightly shifted towards more females which thus might be slightly overestimated in our final dataset. To further test the significance of this difference we conducted a χ^2 contingency test, specifically using *scipy.stats.chi2_contingency*.

The χ^2 contingency test results are presented in Table 6.

The result presented in Table 6 shows that since $p < 0.05$ we can reject the null hypothesis and conclude that there is a significant difference in the ratio of females to males before and after updating. However, if some of the female labels are actually male, it would probably only disturb the analysis so that the difference between genders seem less than it actually is and we can use these results as a "best case" indicator, why we evaluate the results to be fair and feasible for analysis. Compared to distributions of men and women in physics elsewhere in the literature [11] the $80-20$ distribution also seems reasonable.

Table 6: This table show the χ^2 results obtained when conducting a contingency test on our distribution of female and male authorships before and after updating. We get $p < 0.001$ and so the result tell us that the ratio between female and male authors before and after updating are statistically different.

χ^2 Test Results of Male to Female Ratio Before and After Update	
Test	**Result**
Chi-square statistic	1576
p-value	< 0.001
dof	1
Expected male (initial; final)	337,514; 884,612
Expected female (initial; final)	76,494; 200,487
Observed male (initial; final)	345,948; 876,178
Observed female (initial; final)	68,060; 208,921

Gender Label Verification

Table 7 displays the author first name and the gender assigned by gender-guesser hereby as well as the actual author gender. We inspected the names and assigned gender labels of ”The Niels Bohr Institute”-affiliated authors as these were authors that we knew the gender of with (almost) certainty. This worked as a sense of verifying the correctness of the gender labelling.

Table 7: Gender label verification based on ”The Niels Bohr Institute”-affiliated authors. We evaluated the predicted gender as we knew the gender of these authors. A correctly predicted gender is marked in green and a wrongly predicted gender is marked in red.

Gender Label Verification		
Selected Full Name	**Predicted Gender**	**Actual Gender**
Ian Gardner Bearden	Male	Male
Sune Toft	Male	Male
Marina Hesselberg	Female	Female
Charles M. Marcus	Male	Male
Brian M. Andersen	Male	Male
Joachim Mathiesen	Male	Male
Jens Paaske	Male	Male
Kim Sneppen	Male	Male
Helle Astrid Kjær	Female	Female
Thomas Heimburg	Male	Male
Karsten Flensberg	Male	Male
Signe F. Simonsen	Female	Female
Katie Auchettl	Female	Female
Mogens Dam	Male	Male
Yun Jiang	Male	Male
Ilaria Brivio	Female	Female
Ariana Di Cintio	Female	Female
Lene B. Oddershede	Female	Female
Johann B. Severin	Male	Male
Peter Lodahl	Male	Male

From Table 7 we see that 100% of the gender labels were correctly predicted. Even though this is a very small sample we see that the gender API seems reliable - at least for "The Niels Bohr Institute"-affiliated authors.

As we have now obtained the final gender distribution and went through the necessary verification steps we proceed to analysis.

Final Data and Measures

The final dataset consisted of 1,432,907 rows (unique publication·publication total authors) and 43 columns (number of variables). 463,195 were unique publications by 657,355 unique authors. Authorships assigned with a binary gender included 1,085,657 entries leaving 347,250 entries with a "NaN" gender.

Table 8 depicts the variables that were defined from our our original variables (1) for analysis purpose.

Characteristics

Table 9 presents an overview of the characteristics of the publications and authorships.

Initial Variables Distributions Between Genders

Figure 12 display the ratio between female and male authors in our dataset.

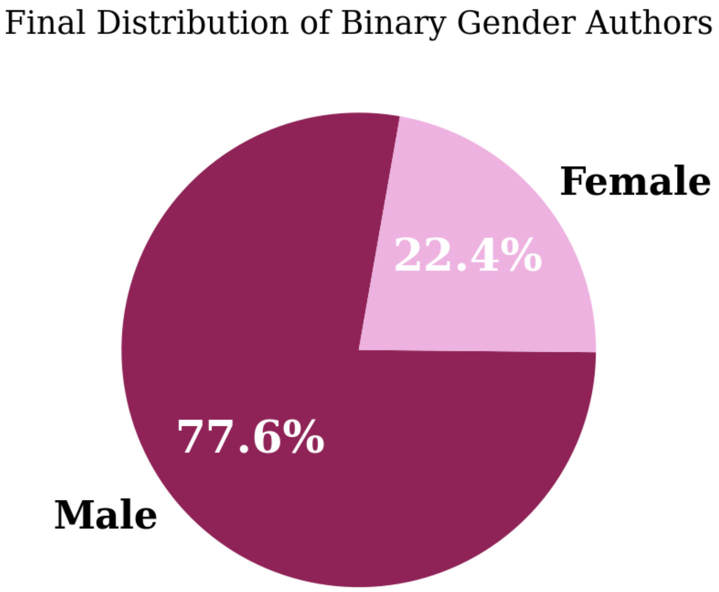


Figure 12: Pie Chart depicting the distribution of male and female authors in the final dataset on author level. That is, the number of unique authors of the given gender in the dataset. Compared to Figure 10 depicting the number of male and female authorships, meaning number of times the given gender appears overall, we see slightly larger female population when looking at distinct authors.

Figure 12 shows that less than 1/4 authors publishing within physics are female.

Table 8: The initial 22 variables were imported from the OpenAlex database (Table 1). These variables were then extended with further 21 variables calculated and extracted from the initial variables and defined as described throughout the Section **Methodology**. The variable name and it's comprehensive description is provided here.

Extended Variable List		
	Name	**Description**
1	Selected Full Name	The selected full name of the given author to be used for the gender assignment as described in section **Extracting Author First Name**
2	Extracted First Name	The first name of the given author extracted from the *Selected Full Name* to be used for the gender assignment as described in Section **Extracting Author First Name**.
3	Author First Country	The first country listed of the given author as described in Section **Extracting Author Country**.
4	Author Country Name	The name of the given country corresponding to the country code in *Author First Country* to be used for the gender assignment as described in Section **Extracting Author Country**.
5	Predicted Gender	The gender of the given author as predicted by the gender API as described in Section **Gender Assignment**. This variable can take one of 6 values: 'female', 'mostly_female', 'male', 'mostly_male', 'andy', and 'unknown'.
6	Binary Gender	The Predicted Gender of the given author categorised into a binary category as either 'female' or 'male'. 'female' and 'mostly_female' were categorised as 'female'. Similarly, 'male' and 'mostly_male' were categorised as 'male'. 'andy' and 'unknown' were replaced as NaN.
7	Abstract Clean	The clean version of the original abstract after undergoing the steps described in Section *Text Preprocessing* as to be used in the Topic Model Analysis.
8	Topic Max Score	Each article were assigned with their topic max score based on the Topic Model result evaluating the score of each topic of each article.
9	Topic Label	The label of the topic of the publication as corresponding to its Topic Max Score.
10	Topic Index	The index of the topic label of the publication as corresponding to its Topic Max Score.
11	Topic Words	The bag-of-words making up each topic as corresponding to its Topic Index and Label.
12	Domain	The domain of the corresponding Topic Label based on the overall theme of the given topic. This variable can take one of 2 values: 'research' or 'education'.
13	Author Ranking	The average number of times the author has been cited by others given by $\frac{\#citations}{\#articles}/author$ as described in Section **Prestige Markers**.
14	Institution Ranking	The average number of times the instituion has been cited by others given by $\frac{\#citations}{\#articles}/institution$ as described in Section **Prestige Markers**.
15	Journal Ranking	The average number of times the instituion has been cited by others given by $\frac{\#citations}{\#articles}/journal$ as described in Section **Prestige Markers**.
	⋮	⋮

	⋮	⋮
16	First Publication Year	The earliest publication year of the given author.
17	Last Publication Year	The latest publication year of the given author.
18	Career Span Years	The number of years between the First Publication Year and Last Publication Year of the given author.
19	Author Publication Count	The total number of publications by the given author.
20	Author Publication Rate	The rate with which the authors has been publishing given by $\frac{Author\ Publication\ Count}{Career\ Span\ Years}$.
21	Event Observed	Whether an author has ceased publishing or not.

Table 9: Authorship and publication characteristics of our main variables in the analysis.

Authorship and Publication Characteristics	
Characteristic	**Entities**
Binary Gender, N (%)	
Male	850,662 (81)
Female	202,356 (19)
Author position distribution, N (%)	
Middle	758,463 (52)
First	434,751 (30)
Last	278,689 (19)
Domain, N (%)	
Research	1,186,707 (82)
Education	256,319 (18)
Author Country Continent, N (%)	
Europe	333,440 (23)
North America	315,979 (21)
Asia	261,494 (18)
South America	16,065 (1)
Africa	8,418 (1)
Oceania	16,749 (1)
Institution Type, N (%)	
Education	607,133 (41)
Unknown	542,913 (37)
Facility	202,917 (14)
Company	46,381 (3)
Government	28,920 (2)
Healthcare	20,149 (1)
Nonprofit	15,039 (1)
Other	7,110 (1)
Archive	1,341 (0)
⋮	⋮

⋮	⋮
Total Author Counts (count), mean (SD)	11.01 (19.56)
Publication Year (years), mean (SD)	2014 (9.307)
Cited By Count (count), mean (SD)	22.66 (149.1)
Author Mean Cites (count/article/author), mean (SD)	22.66 (95.31)
Institution Ranking (count/article/institution), mean (SD)	22.66 (32.88)
Journal Ranking (count/article/journal), mean (SD)	24.01 (55.55)
First Publication Year (year), mean (SD)	2009 (11.05)
Last Publication Year (year), mean (SD)	2017 (8.154)
Career Span Years (years), mean (SD)	7.774 (9.760)
Author Publication Count (count), mean (SD)	8.163 (17.08)
Author Publication Rate (count/year), mean (SD)	0.843 (1.345)
Is corresponding (False), N (%)	1,228,717 (83)
Grants (False), N (%)	1,214,659 (83)

Figure 13 display the distribution of our initial variables. We present the initial variables and their distribution in their raw form, gender aggregated and as a function of time (when relevant). This section provides an overview of the nature of the dataset and the differences between male and female authors for the main variables. It is indicated at each figure whether the distribution is at Authorship, Author or Publication Level.

Figure 13(a) show that there is an increasing number of publications per year and similarly, as depicted in Figure 13(b), the amount of female authors increase as well from circa year 2000 and forward.

Figure 13(c) and 13(d) display the raw number of citation counts and gender aggregated citation count, respectively. We see that both the female and male citation count follow the overall citation count trend. However, there is a disparity in the number of citations in the intermediate range of number of citations, where female authors are not as represented.

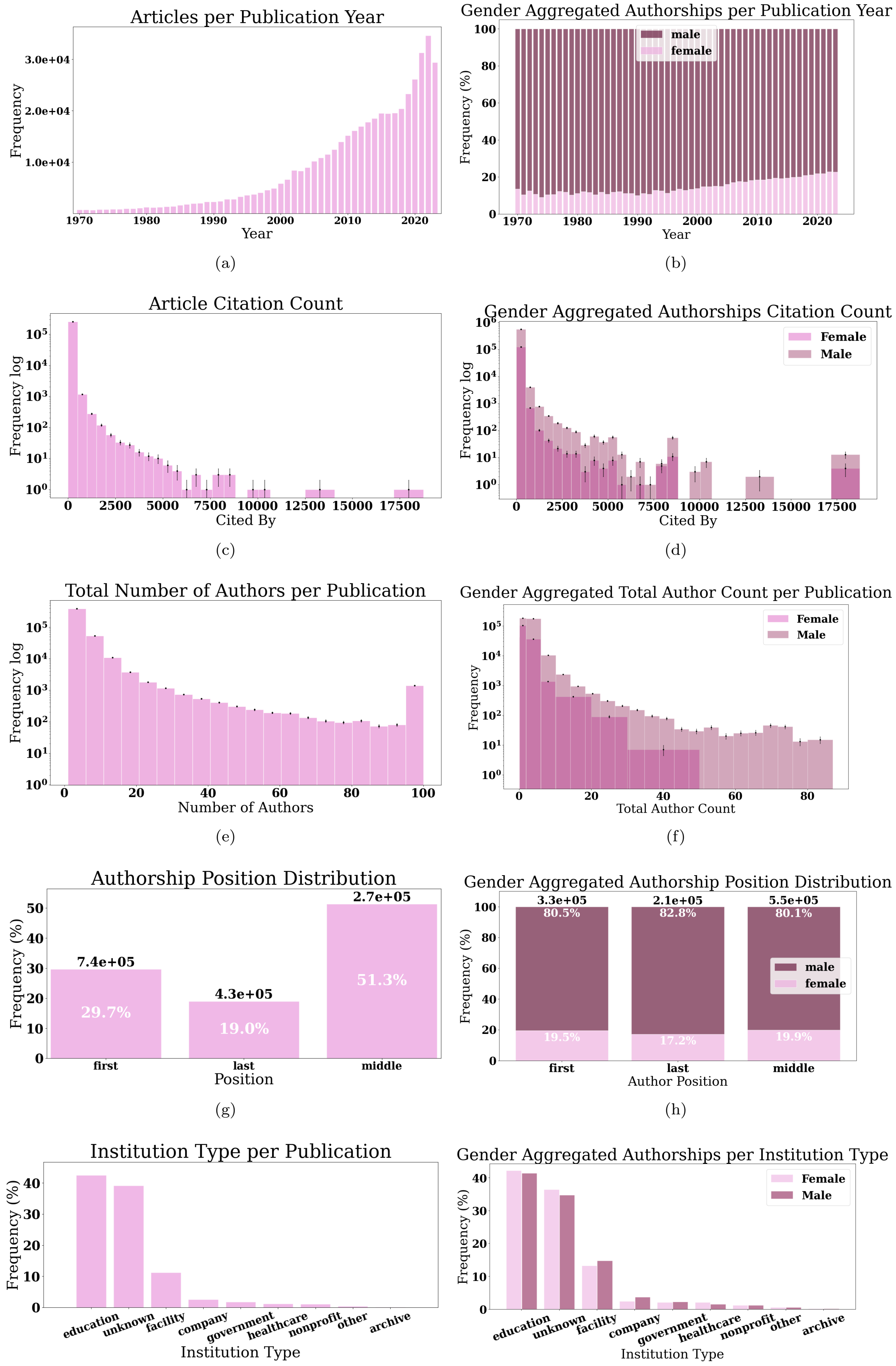
Figure 13(e) and 13(f) show that there is a disparity between the total number of authors between genders. We see that there are no publications with more than 50 female authors whereas male authors are somewhat continuously distributed from zero to around 85 male authors per publication.

Figure 13(g) show that half the authorships are middle positions whereas the first and last position is more rare. The unequal size of first and last author positions show that there are some publications with a single author and therefore only a first author position. The gender aggregated version of authorship positions shown in Figure 13(h) show that while the first author position and middle author position resembles the overall authorship distribution (Figure 10). However, the last author position is more dominated by the male group, compared to the other two position types.

Figure 13(i) show that most of the publications are from an 'education' institution type which makes up around 40% of the publications. That is the same for next biggest proportion which is of an 'unknown' institution type. Then 'facility' makes up around 10% and the rest is mainly distributed between 'company', 'government', 'healthcare', and 'nonprofit'. Looking at the gender aggregated version, shown in Figure 13(j) it follows almost the same pattern for each gender. Where 'facility' and 'company' are more represented by male authors whereas female authors dominate 'healthcare' and are slightly more represented in 'education' and 'unknown'.

Figure 13(k) and 13(l) show that the number of authorships that are corresponding between genders follows the overall distribution i.e. that around 17% is corresponding.

Figure 13(m) and 13(n) show that grants are somewhat equally given between female and male authorships. Furthermore, Figure 13(o) where the received grant count is shown per last author gender, confirms the equal distribution of grants.


Articles per Publication Year
Frequency
3.0e+04
2.0e+04
1.0e+04
1970
1980
1990
2000
2010
2020
Year
Gender Aggregated Authorships per Publication Year
male
female
Frequency (%)
100
80
60
40
20
0
Year
Article Citation Count
Frequency log
Cited By
0
2500
5000
7500
10000
12500
15000
17500
Gender Aggregated Authorships Citation Count
Female
Male
Total Number of Authors per Publication
Number of Authors
Gender Aggregated Total Author Count per Publication
Frequency
Total Author Count
Authorship Position Distribution
7.4e+05
29.7%
4.3e+05
19.0%
2.7e+05
51.3%
first
last
middle
Position
Gender Aggregated Authorship Position Distribution
3.3e+05
80.5%
19.5%
2.1e+05
82.8%
17.2%
5.5e+05
80.1%
19.9%
Author Position
Institution Type per Publication
education
unknown
facility
company
government
healthcare
nonprofit
other
archive
Institution Type
Gender Aggregated Authorships per Institution Type

(a)

(b)

(c)

(d)

(e)

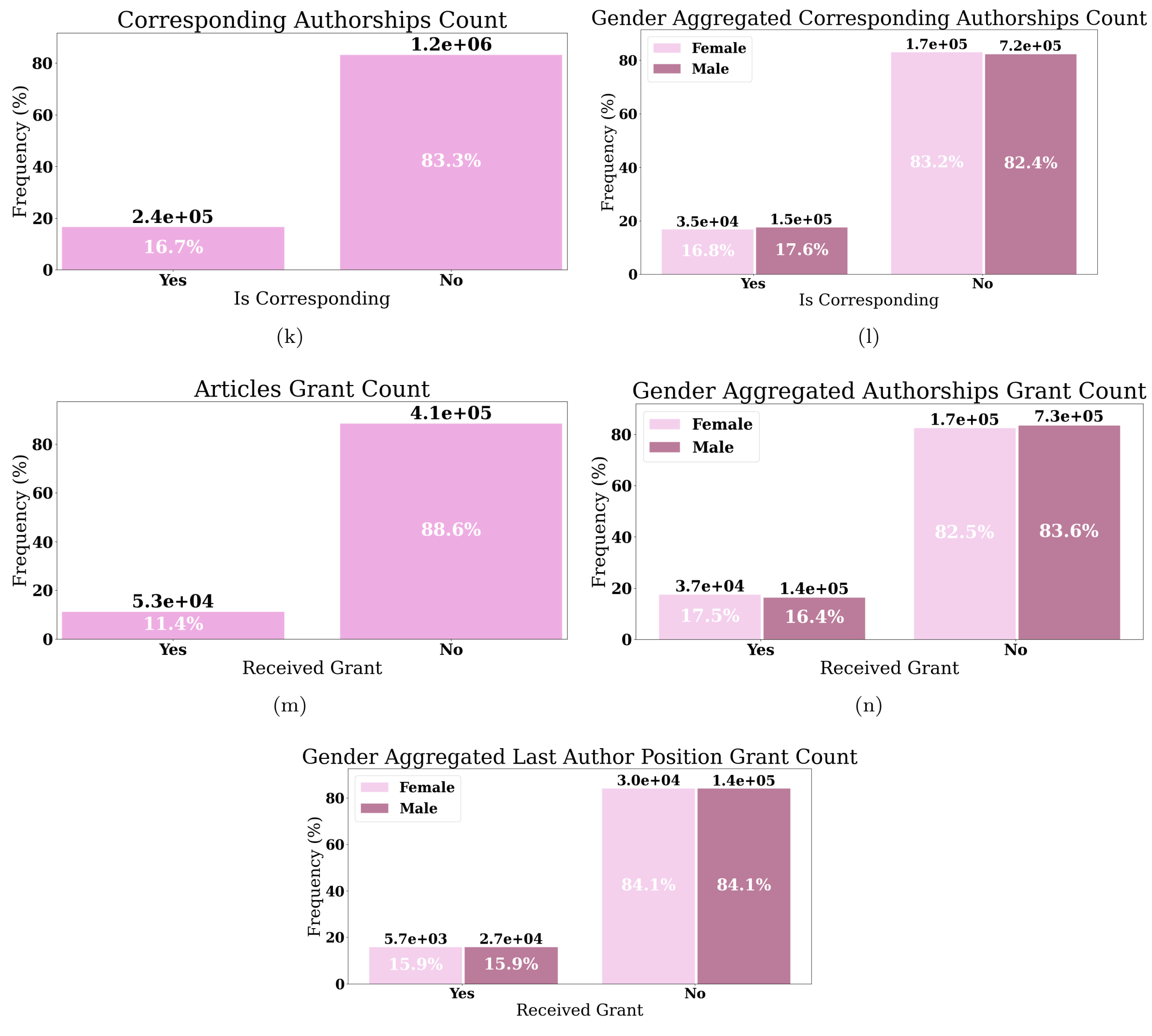
(f)

(g)

(h)

(i)

(j)



(k) (l)

(m) (n)

(o)

Figure 13: Distributions of our initial variables. The raw distribution of each variable is shown in the left column and the gender aggregated distribution is shown in the right column (except for journal and institution distribution where the raw, female and male distribution is in a row each).

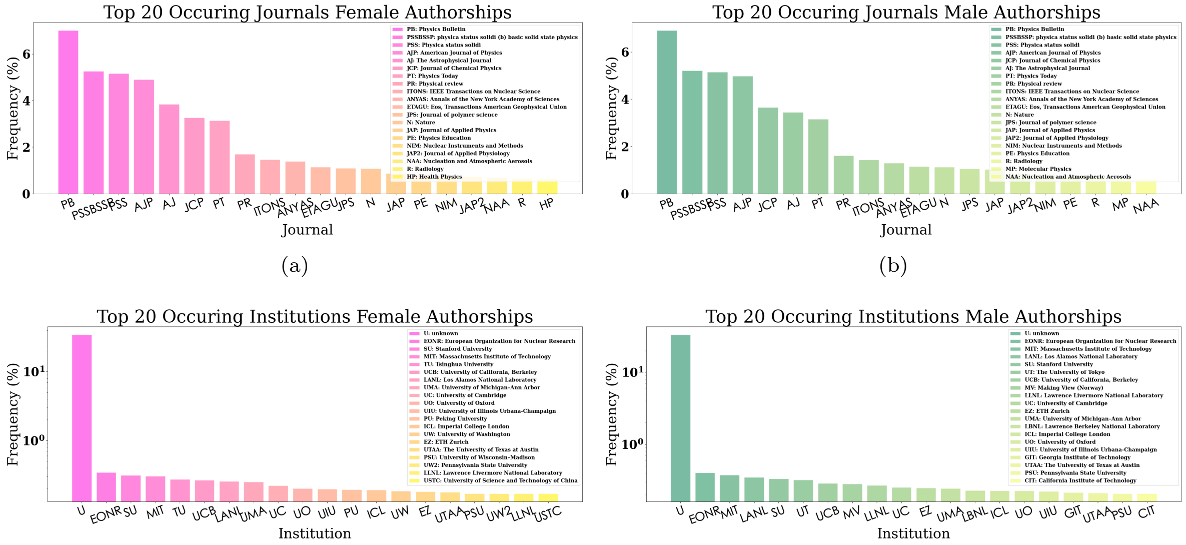
Table 10 presents the statistical test outcomes comparing the means for each variable presented in Figure 13 between female and male authorships. For each variable we compare the mean between female and male authorships to test if the means are statistically different. We used both Welsch's t-test as well as Mann Whitney U-test. We hypothesised that the mean for female authors is less than for male authors and thus used a one-sided test. The significance threshold was set at .05.

From Table 10 we see that the female mean was significantly smaller for each of the distributions tested, except for Author Count per Author Position and Grant.

Figure 14 display the top 20 most occurring distinct journals and institutions grouped by female and male authorships. The top 20 most occuring journals and insititions overall are displayed in Figure A10 in Appendix A.

Table 10: Hypothesis testing results. We tested whether the key variables, presented in Figure 13, were statistically different between the two genders. We hypothesised that the female mean would be less than the male mean. The significance threshold was set at .05. If the test result were significant, meaning the female mean was significantly smaller than the male mean, it is marked in green, and if it was insignificant, meaning the distribution between female and male authorships were equal, it is marked in red. Categorical variables (Institution Type, Is Corresponding, Grant, and Grant per Last Author) were factorised in order to perform the hypothesis test.

Initial Variables Hypothesis Testing Results					
Variable	**Female Mean (SD)**	**Male Mean (SD)**	**T-statistic, p-value**	**U-statistic, p-value**	**Figure**
Authorships per Publication Year	3874 (5575)	16231 (19844)	-4.41, > .001	663, > .001	13(b)
Cited by count	20.34 (142.4)	26.06 (162.6)	-16.04, > .001	8.599e+10, > .001	13(d)
Total Author count per Publication	0.50 (1.02)	2.11 (2.88)	-339, > .001	3.06e+10, > .001	13(f)
Author Count per Author Position	69,732 (37,368)	29,2154 (137,555)	-2.70, .03	0.0, 0.05	13(h)
Institution Type	1.66 (1.25)	1.68 (1.22)	-5.66, > .001	3.71e+10, > .001	13(j)
Is Corresponding	0.17 (0.37)	0.18 (0.38)	-8.51, > .001	9.10e+10, > .001	13(l)
Grant	0.17 (0.38)	0.16 (0.37)	11.7, 1.00	9.27e+10, 1.00	13(n)
Grant per Last Author	0.16 (0.37)	0.16 (0.37)	-0.443, .33	3.06e+09, 0.33	13(o)



(a)

(b)

(c)

(d)

Figure 14: Top 20 most occurring distinct **a** journals and **b** institutions in our dataset as well as the same top 20 grouped by **c, e** female and **d, f** male authorships.

Figures 10(a), 14(a) and 14(b), show that the overall distribution as well as the gender aggregated distribution are very similar for the first journals and then we see some disparities. Where the number one most appearing institution is ”unknown” overall and between genders, the rest of the top occurring institution seem somewhat different between female and male authors as depicted in Figure 10(b), 14(c), and 14(d).

Figure 16 and Figure 15 display the geographical distribution of authorships and gender aggregated authorships.

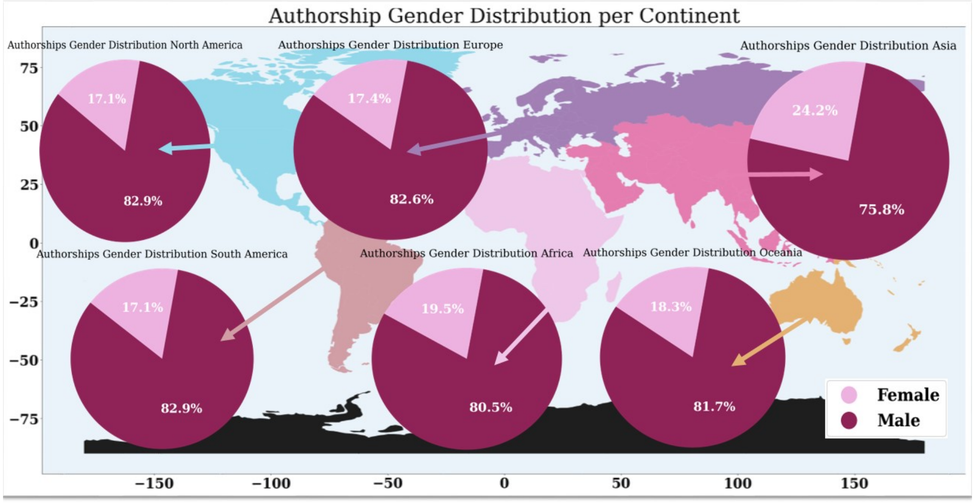

Authorship Gender Distribution per Continent
Authorships Gender Distribution North America
Authorships Gender Distribution Europe
Authorships Gender Distribution Asia
75
50
25
0
−25
−50
−75
17.1%
82.9%
17.4%
82.6%
24.2%
75.8%
Authorships Gender Distribution South America
Authorships Gender Distribution Africa
Authorships Gender Distribution Oceania
17.1%
82.9%
19.5%
80.5%
18.3%
81.7%
Female
Male
−150
−100
−50
0
50
100
150

Figure 15: World map presenting the authorship gender distribution per continent. Each continent is color coded and each pie chart displays the authorship gender distribution of the given continent.

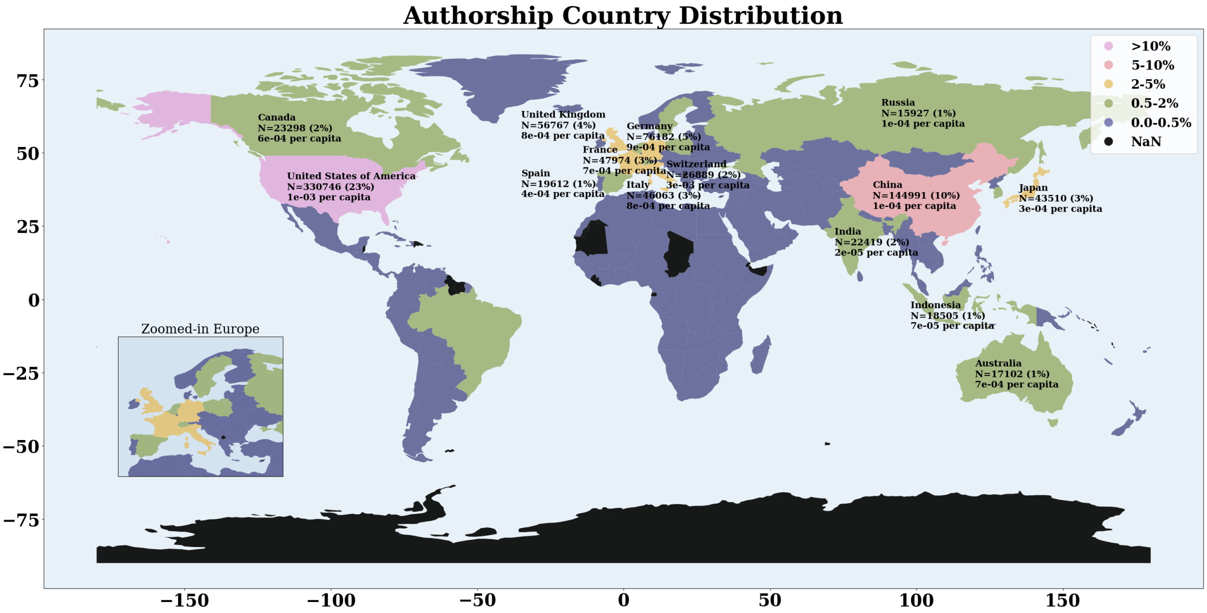


Figure 16: World map presenting the Authorship Country distribution. Each country is color coded according to the proportion of authorship of the given country. Furthermore, a descriptive text is printed on the countries making up at least 1% of the dataset. The text displays the given country name, the authorship count, the authorship count in percentage and lastly the authorship count compared to the capita size of the given country.

The world map, Figure 16 reveal that United States of America dominate the dataset with its 23%. Next is China with its 10%. The following countries make up 5% or less of the authorship countries in the dataset. However, Table 9 shows that grouped per continent Europe is the most presented with its 23% followed by North America, 12% and then Asia, 18%. South America, Africa, and Oceania are responsible for 1%.

Looking at the authorship gender distribution grouped per continent, Figure 15, we see a similar distribution per continent as to the overall distribution between male and female authorships (around 80-20) depicted in Figure 10. However, there are some disparities and the largest female cohort is found in Asia with 24% female authorships. The smallest female cohort is found in North and South America each with 17% female authorships.

Figure 17 displays a 2D histogram of the gender composition per publication.

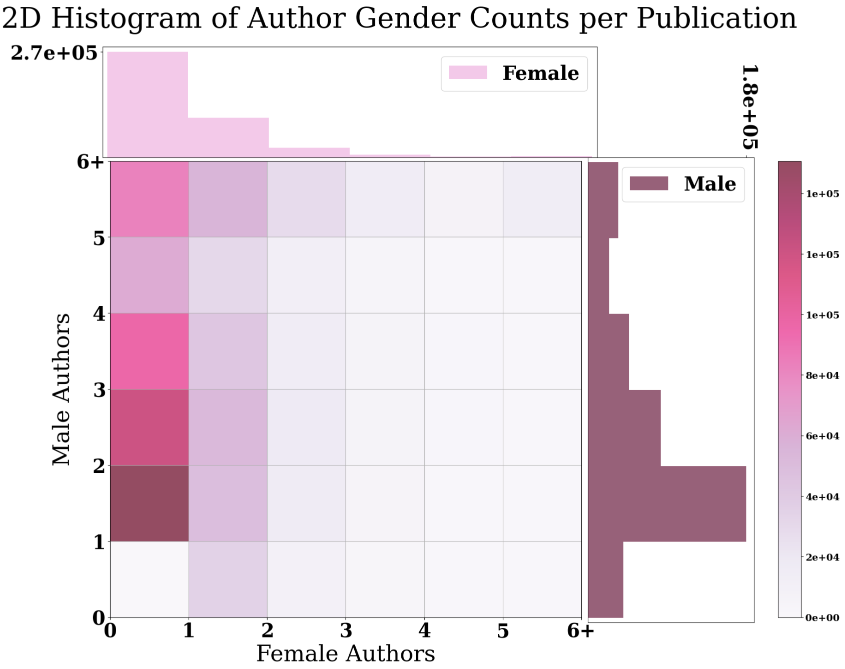


Figure 17: 2D histogram depicting the distribution of collaboration per publication between female and male authors. All publications with more than a total of 6 authors is put in the "6+" bin. The upper x-axis figure displays the 1D histogram of female authors count per publication and the figure on the right displays the 1D histogram of male authors count per publication. The histogram is created such that the values of the returned histogram are equal to the sum of the total number of authors belonging to the samples falling into each bin.

From Figure 17 we see that most publications are by 1 male author. We also see that male author counts are presented along all categories whereas female authors are mostly found in the "0" count category and fewer as the count is increased. This is in line with what we saw from the toal gender count per publication in Figure 13(f).

Figure 18-20 depicts the evolvement of gender aggregated authorship distributions over time.

From Figure 18 we see that there has been an increase of female authors of all three positions. Where in 1970 the female authors occupying the first author position were 15% it was increased to 24% in 2023. The female middle authorships increased from 13% to 23% between 1970 and 2023. Lastly, the ratio of last female authorships increased from 11% in 1970 to 20% in 2023 which again confirm that the last author position is the least preoccupied for female authors. Furthermore, this plot also aligns with Figure 13(b) showing that the increase of female authorships happen especially around year 2000.

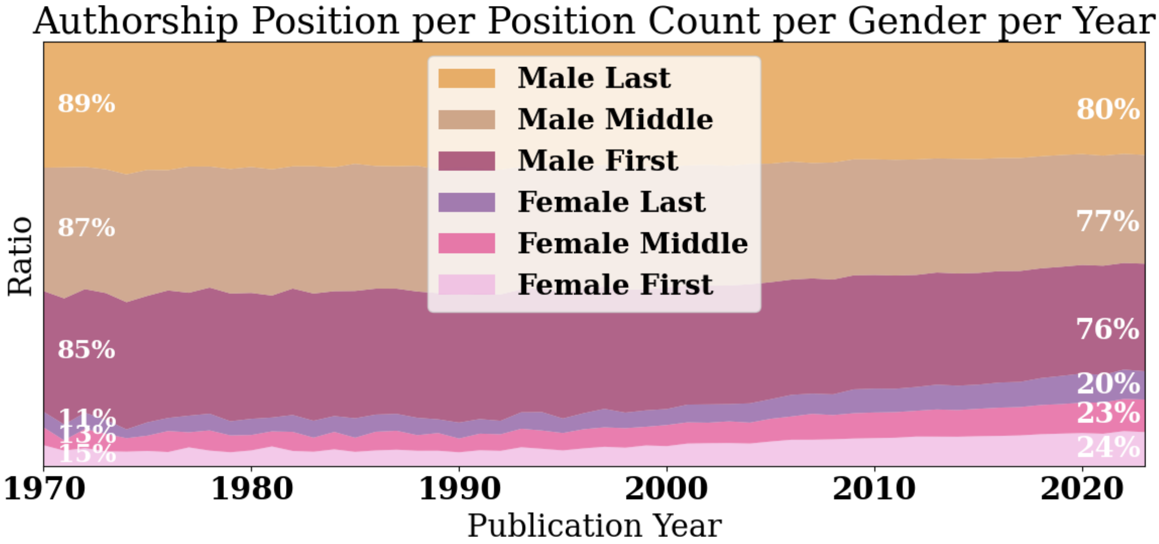


Figure 18: Gender aggregated authorship position distribution per publication year. Each position type per gender is normalised according to the total number of authors in the given position per year to make up for the general increase in number of publications over time. The ratio show the part of respectively female and male authors per position type. The text displayed on the left side shows this ratio between each gender per position in publication year 1970 and the text on the right hand similarly but for year 2023.

Figure 19 displays the total amount of female and male authorships over time normalised according to the total count of authorships per year.

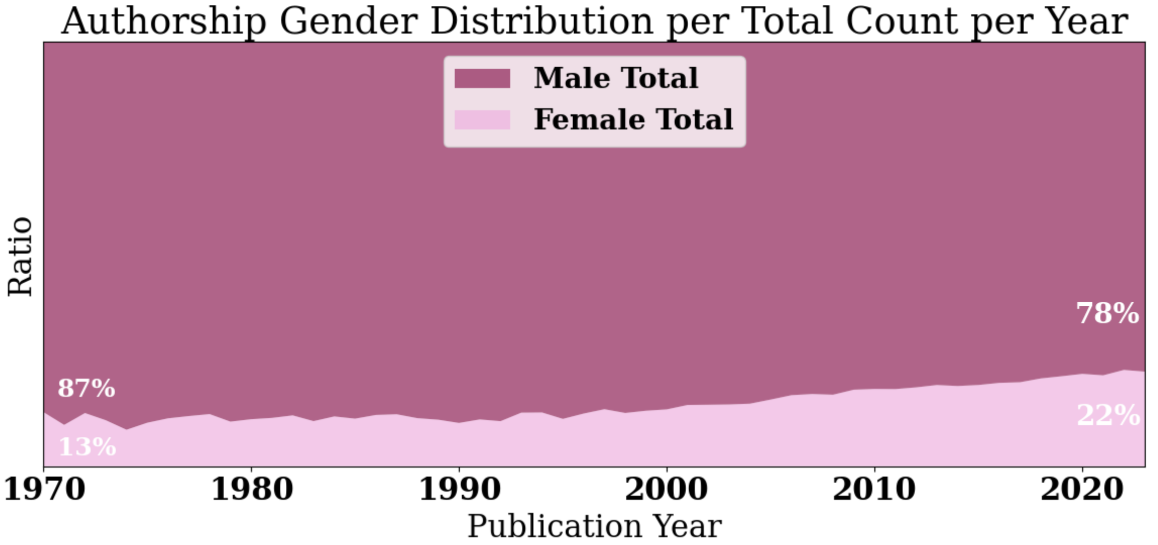


Figure 19: Gender aggregated authorship distribution per publication year. The gender count is normalised according to the total number of authorships per year to make up for the general increase in number of publications over time. The ratio show the part of respectively female and male authors. The text displayed on the left side shows this ratio between each gender per position in publication year 1970 and the text on the right hand similarly but for year 2023.

From Figure 19 we see that there has been a general increase of female authorships over time and it has increased with 9pp.

Figure 20 displays the female authorship distribution over time with a logistic fit predicting the year that the number of female authors will reach 50% of all authorships.

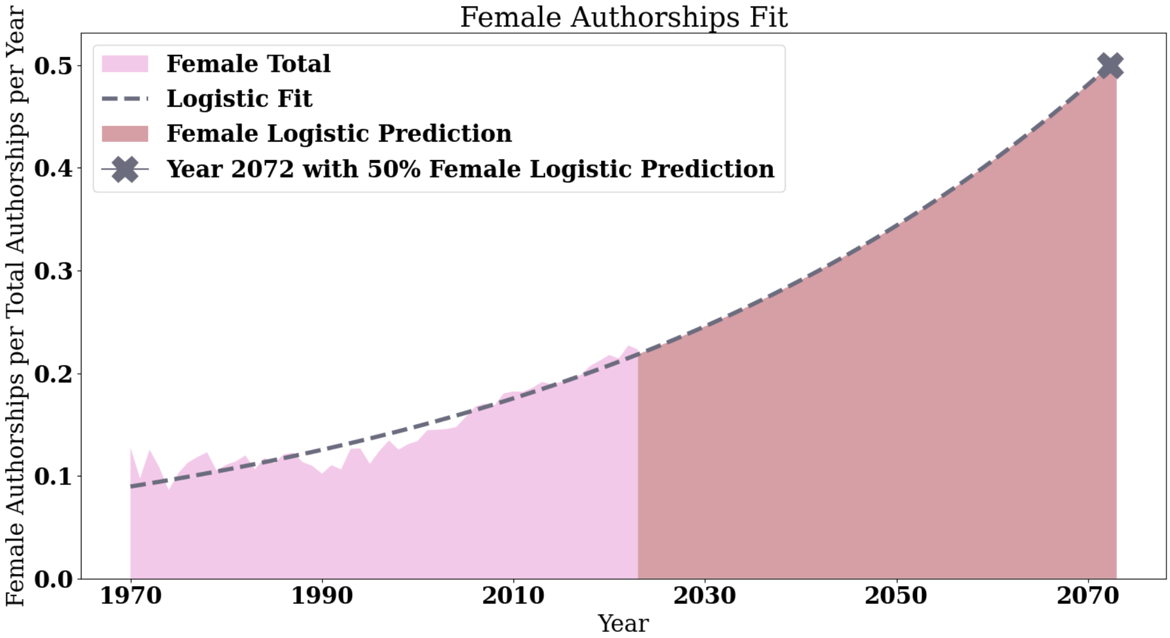


Figure 20: Female authorship distribution per publication year with a logistic fit and prediction. The gender count is normalised according to the total number of authorships per year to make up for the general increase in number of publications over time. The first part denoted in pink is the actually data as presented in Figure 19 whereas the second part denoted in copper is simulated data following the prediction made with the logistic fit assuming that the number of female authorship will increase with the same logistic rate.

From Figure 20 we see that in year 2072 female authorships will make up 50% of all authorships and thereby an equal amount as male authorships (assuming that the number of female authors increase at a logistic rate).

In summation, there is a statistical difference between male and female authors in terms of Publications per Publication Year, Cited by Count, Total Author count per Publication, Institution Type, and Author Is Corresponding. The only variables that are not statistically different are Author Count per Author Position, Grant, and Grant per Last Author. There has been an increasing amount of female authors over time, especially from year 2000 and onward. However, there are some disparities between male and female authors in especially last author position, lack of collaboration (on the male side), citation count, total number of authors per publication, institution, institution type, and (somewhat) journal. As expected, we also see a huge disparity in the overall cohort size of male and female authors distributed around $80-20$ with the largest difference in North and South America and the smallest in Asia. If the number of female authorships continuously increase with a logistic rate we will see an equal amount of female and male authorships 48 years from now in year 2072.

Topics and Domains Between Genders

We move on to labelling each publication with a physics sub-field, as described in Section **Topic Model**, by applying a topic model to the *Abstract Clean* per publication.

LDA Model Optimisation

As described in Section **Topic Model** we use Model Perplexity and Coherence Score to evaluate the most optimal parameter settings for our Topic Model. Remembering that the main parameters of the Topic Model is number of topics, α, and β we look at the optimal parameter settings determined through each step of the evaluation:

Initial Topic Modeling with Default Hyperparameters

Figure 21 show the Perplexity Score and Coherence Score as a function of number of topics. We trained multiple LDA models looping through a range of number of topics from 20-100 at steps of 10. We calculated the coherence and perplexity score for each setting. Remembering that the Perplexity Score should be minimised and coherence score should be maximised we identify the min and max score of each metrics and mark them on the Figure 21.

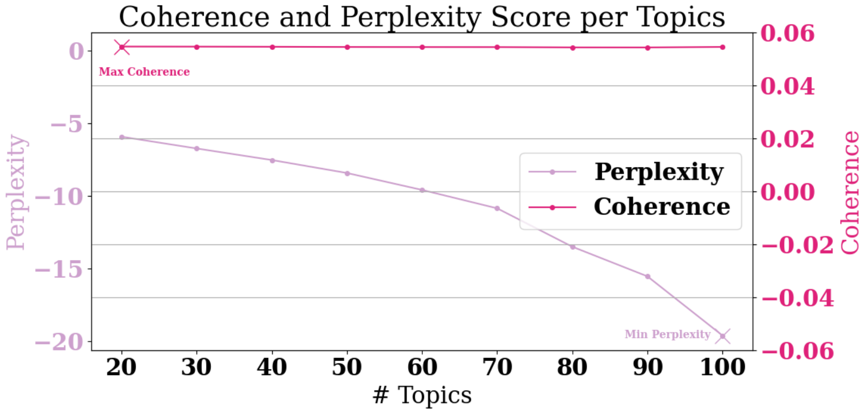


Figure 21: The figure displays the perplexity score and coherence score as a function of number of topics in or topic model. Where the left y-axis represents the perplexity score, the right y-axis shows the coherence score. The best scores, minimum perplexity and maximum coherence, are marked on the plot as well. We see that the max coherence score is at 20 topics and the minimum perplexity at 100 topics why we proceed to find the trade-off between the two scores.

From figure 21 we see that the minimum coherence score is found at 20 topics whereas the maximum perplexity score is found at 100 topics. As these selection criteria do not agree with the optimum number of topics we consider the best trade-off between the two scores as illustrated in Figure 22.

Select Optimal Number of Topics

Looking at Figure 22 we see that the optimal trade-off between Perplexity and Coherence Score is at 70 topics where the intercept between the two is. However, as 70 topics in term of physics sub-fields would be very hard to analyse and evaluate we choose to consider 20 topics for the model. This is also aligned with the max Coherence Score and since Coherence Score reflect interpretablity of the Topics we weight this metric higher than the perplexity score.

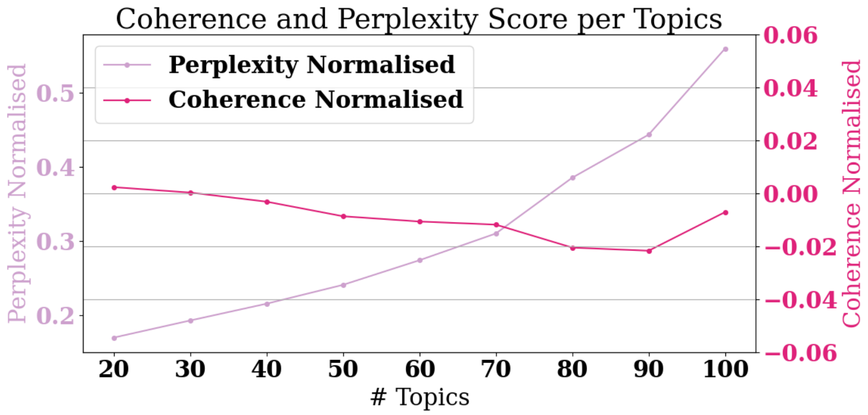


Figure 22: The figure displays the normalised perplexity score, on the left y-axis, and coherence score, on the right y-axis, in order to compare them and find the trade-off between the two scores. We find the intercept at 70 topics which indicates the optimal topics setting based on the coherence and perplexity measure.

Hyperparameter Optimisation

Table 11 show the Coherence Score obtained by each combination of hyperparameter settings tested. Now that we have decided to go forward with 20 topics we find the optimal coherence score between different settings of hyperparameters α and β. We compute the score for each combination of the two scores taking the values 0.01, 0.1, and 1.0.

Table 11: Now that we have decided on 20 topics we test different settings of the hyperparameters α and β to fine tune the topics model. We use coherence score as a measure of the best model. Remembering that the coherence score should be maximised we find that the settings listed in row 9 give the highest coherence score as marked in green. Therefore we run or topic model with settings 20 topics, $\alpha = 0.01$ and $\beta = 1.00$.

Coherence Score Evaluation of Topic Model Hyperparameters				
	# Topics	α	β	**Coherence Score**
1	20	1.00	0.01	0.368601
2	20	1.00	0.10	0.412725
3	20	1.00	1.00	0.481174
4	20	0.10	0.10	0.494421
5	20	0.01	0.01	0.496179
6	20	0.10	0.01	0.500060
7	20	0.10	1.00	0.543652
8	20	0.01	0.10	0.544857
9	20	0.01	1.00	0.563709

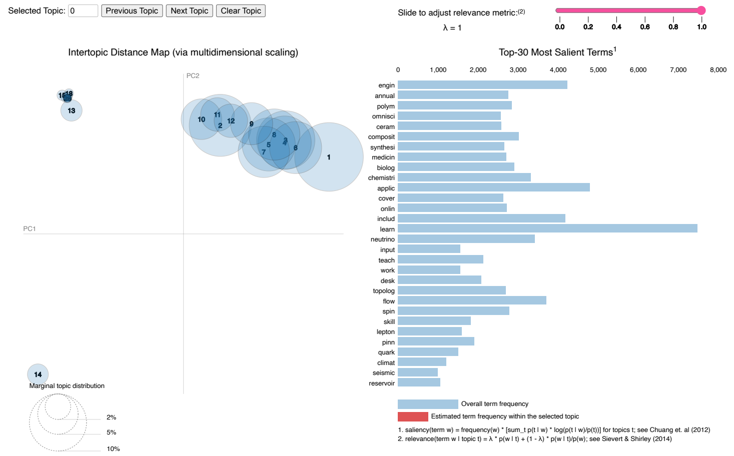
Evaluating the results obtained from Table 11 we see that the maximised and thereby optimum coherence score is found with hyperparameter settings $\alpha = 0.01$ and $\eta = 1.00$ why we ran our final Topic Model with settings:

- **Number of Topics:** 20
- **Hyperparameter α:** 0.01
- **Hyperparameter η:** 1.00

Final Topic Modeling with Optimized Hyperparameters

Topic Labelling

Figure 23 shows a preview of an interactive plot showing the obtained clusters of bag-of-words and their respective weights. You can view the visualisation created with pyLDAvis by clicking here. Running or final LDA model using the optimal settings we obtained 20 distinct clusters of words making up the topics. The Topic Model outputs a bag of words and their corresponding weight in the bag from which we label the topic. Since the Topic Model is an unsupervised clustering method it doesn't provide any labels but rather serves as a guide for easier labelling so the next step was to label the bag-of-words as overall topics.



Selected Topic: 0

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 23: png view of the clusters of topics and their bag of words obtained by the Topic Model. The left hand side of the plot displays the topics as cluster in a 2D space and their size illustrates their prevalence in the dataset. The right hand side displays the words in the bag-of-words that make up their topic. The size of the bar show the weight of each word in the respective topic. The interactive html plot can be viewed here.

Figure A11 in Appendix A show the labelled topics and their underlying bag-of-words with their respective weights in the bag.

Evaluating the theme of the words in the terms of physics sub-fields we decided on the following topics:

1. Particle Physics
2. Medical Physics
3. Computational Physics
4. Engineering
5. Optoelectronics
6. Fluid Dynamics
7. Mathematical Physics and ML
8. Geophysics
9. Undefined Topic 1
10. Undefined Topic 2
11. Undefined Topic 3
12. Quantum Physics
13. Materials Science
14. Undefined Topic 4
15. Undefined Topic 5
16. Undefined Topic 6
17. Meteorology
18. Learning and Teaching
19. Undefined Topic 7
20. Astrophysics

Initial Topic Distribution

The initial topic label distribution and max scores are displayed in Figure A13 and **??**, respectively, in Appendix A. We assigned each publication with it's topic based on the maximum score obtained when matching the given article abstract with each topic (remember that the TM provides a distribution of topic probability for each document and not a hard assignment). Evaluating the initial TM results we argue to discard the "Undefined Topics" from the analysis. Looking at the bag-of-words and their respective weights, presented in Figure A11 in Appendix A, many of them are zero and therefore seems to be noise or at least meaningless. They are also the least represented in the dataset as can be seen from figure A13, except for Topic 13: "Undefined Topic 4". However, even Topic 13 were only assigned to 6% of the dataset.

Figure 24 show the topic score distribution for not only the maximum score but for the distribution of scores assigned by the TM. We see the maximum score and the following 2nd, 3rd, 4th and 5th highest score, respectively.

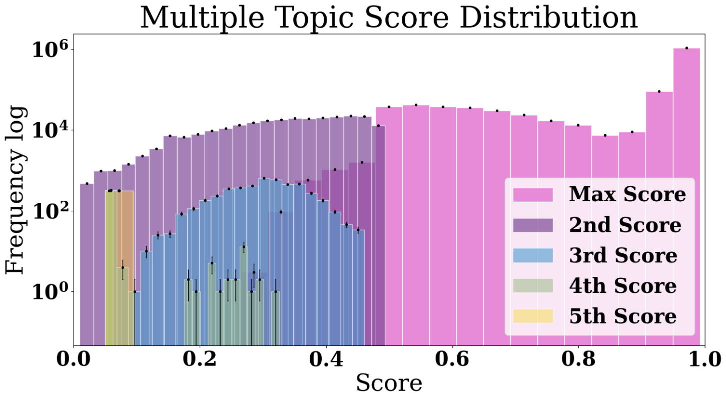
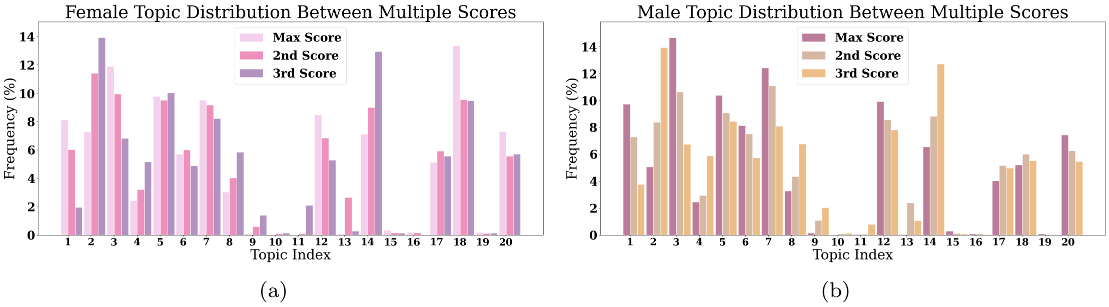


Figure 24: The figure shows the score distribution of each publication assigned by the TM. Where zero is the lowest score, indicating that the topic and the text do not match at all, 1 is the highest score indicating that the topic and the text are very likely to match. We see that some publications have up to 5 potential matches. However, most of the Max Scores are above 0.5 and none of the 2nd Scores and below are above 0.5. That means that in most cases the Max Score is either the only assigned score and if not it is more than 50% likely to be the best matching score.

Figure 25 shows the distributions of the topics between female and male authorships, respectively, when taking into account not only the max topic score but also the second and third highest scoring assigned by the topic model based on the abstract. We see from Figure 24 that while some publications are assigned with a range of probable topics counting up to as much as five topics, most of them only have 1 topic assigned, fewer 2-3 topics and even less above 3 topics assigned. Therefore, we only plot the first three best matching scores when comparing the topic distributions between genders for multiple scores in Figure 25.



(a)

(b)

Figure 25: Topic distribution for **a** female and **b** male authorships comparing the distribution between the best matching scores, Max Scores, as well as the next best matching, 2nd Score and 3rd Score, respectively. The Topic Label of the corresponding Topic Index are listed in Section **Topic Labelling**.

From 25 we see that even when taking into account the 2nd and 3rd best scores we still see the "Undefined" Topic, namely *Topic Index* 9, 10, 11, 14, 15, 16, and 19, are almost zero in their count. However, we see an exception with Topic 14: Undefined Topic 4 which is actually highly represented when taking into account the 2nd and 3rd Score. However, looking at the bag-of-words connected to this topic "desk", "help", "take", "optic", "world", "present", "philosphi", "scientist", "year", and "univers" (Figure 11(n)), we wouldn't know how to interpret it as a physics subfield. Also, the words seem like quite general words that might be in any physics publication which might explain the relatively high matching level. Therefore, we still argue to discard all the Undefined Topics from the analysis for now.

Furthermore, we plot the multiple score topic distributions between genders to ensure that we can "trust" the max score as an efficient marker of the document topic and do not see an entirely different distribution when including the range of scores and possible topics. It might also reveal some insights into the levels of interdisciplinary publications between genders.

Figure 25(a) show that most topics for female authorships somewhat agrees between each score but for Topic 1, 3, 12, 18 and 20 we see fewer matches in the 2nd and 3rd score compared to the Max Score, and for topic 2, and 4 we see less matches by the Max Score compared to the 2nd and 3rd scores.

Figure 25(b) show that most topics for male authorships somewhat agrees between each score but for Topic 1, 3, 5, 7, 12, and 20 we see fewer matches in the 2nd and 3rd compared to the Max Score, and for topic 2, 4, and 8 we see less matches for the Max Score compared to the 2nd and 3rd scores.

For this analysis we evaluated undefined topics as noise and that they were not contributing meaningfully to the analysis why we discarded them from further analysis. We also evaluated that the Max Scores provided a good enough measure to use as topic indicator for the document topic distribution in our dataset.

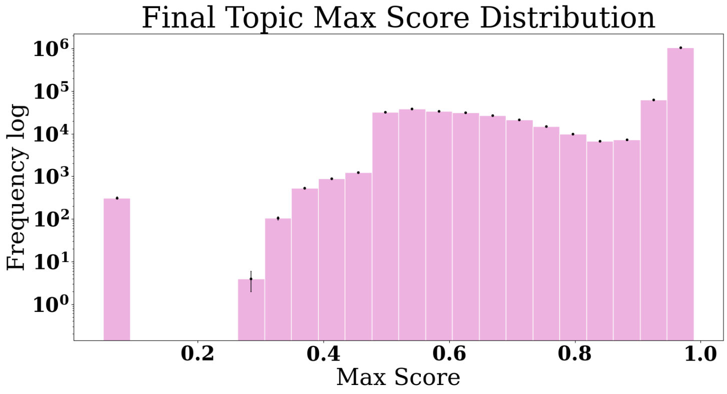
Clean Topic Distribution

We drop the undefined topics from the dataset leading to a dataset reduction to 1,340,627 rows including 408,052 unique publications and 615,377 unique authors.

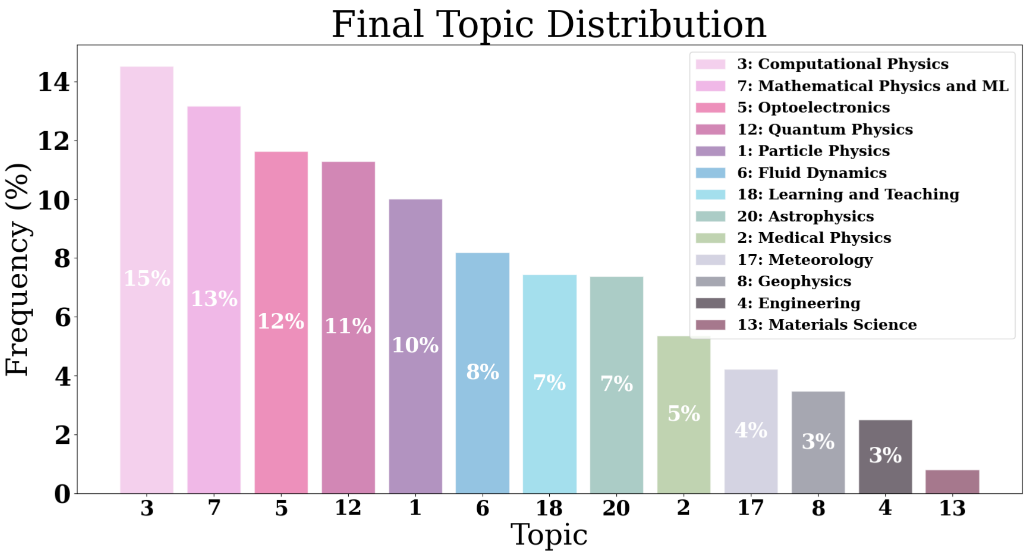
Table 13 provides on overview of each clean topic label, it's bag-of-words and their respective weights as well as an example of the highest scoring abstract for the given topic.

From Table 13 we see that the abstract and its label seems reasonable throughout all examples. All of the matching scores are above 0.9 which means that they probably only match with that given topic.

Figure 26 shows the final *Topic Label* distribution after discarding the undefined topics as well as their corresponding *Topic Max Score* distribution.



(a)



(b)

Figure 26: Distribution of the **a** max scores and **b** topics in our clean topic dataset after discarding the undefined topics from the initial topic distribution obtained. The document topic were assigned by applying the topic model and finding the max probability score when matching a topic and an abstract.

Figure 26(b) shows that the most present topic in our distribution is Topic 3: Computational Physics making up 15% and the least present topic in our distribution is Topic 13: Materials Science making up 1%. From Figure 26(a) we see that most our topics had a matching score < 0.9 and the next most had a score < 0.5. Few had a score > 0.5. This means that for most part, we can expect the abstract to match the topic at the same level as presented in Table 13.

Figure 27 shows the gender aggregated version of the clean topic distribution.

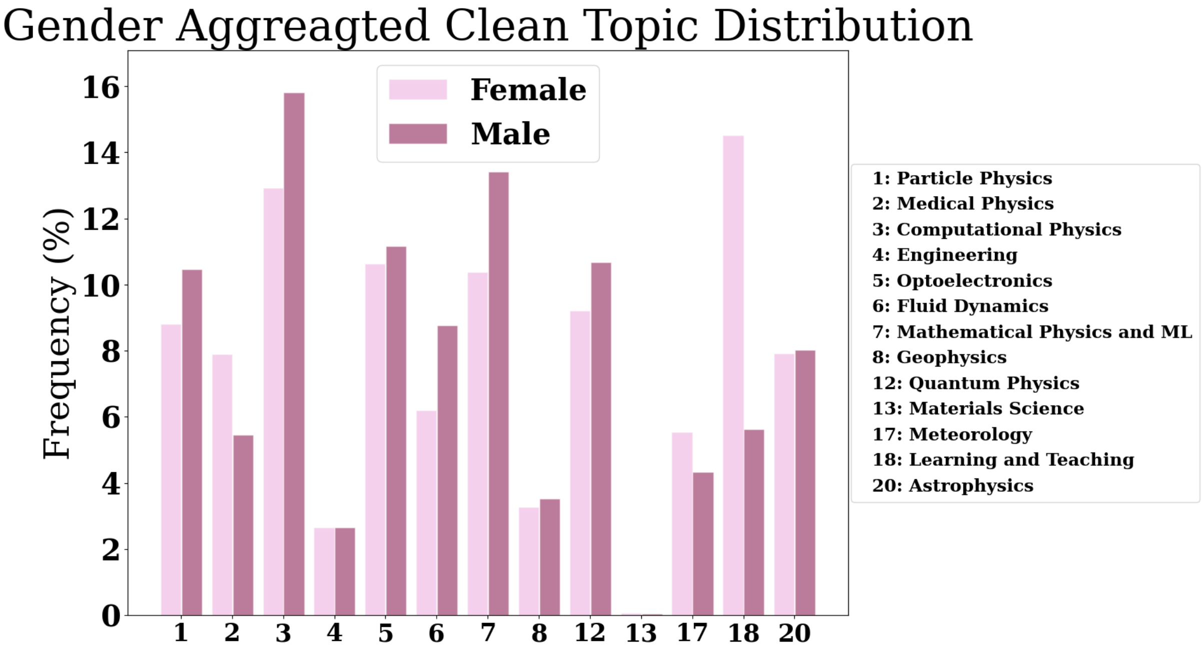


Figure 27: Gender aggregated topic distribution of the clean topics after discarding undefined topics.

Table 12 displays on overview of topic distribution count and percentage point difference between genders. From Figure 27 we see a general overrepresentation of male authorships for most topics with a few exceptions. To explore the exact representation between female and male authors per topic we calculate the count and percentage for each topic per gender and find the difference between them for each topic respectively.

Table 12: Table overview of the topic distribution between genders. The table displays the count of female and male authorships per topic as well as the percentage. Finally, it displays the difference in count and percentage point between female and male authors per topic. The distribution is also depicted in Figure 27. A female dominated topic ($\Delta_{pp} > 0.5$) is marked in pink and a male dominated topic ($\Delta_{pp} < -0.5$) in red whereas an equally dominated topic ($0.5 > \Delta_{pp} > -0.5$) is marked in green.

Topic Distribution Count and Percentage Point Difference Between Genders				
	Topic Label	**Female N (%)**	**Male N (%)**	**Δ N (pp)**
1	Particle Physics	16988 (8.82)	85103 (10.5)	-68115 (-1.65)
2	Medical Physics	15220 (7.90)	44405 (5.46)	-29185 (2.44)
3	Computational Physics	24883 (12.9)	128639 (15.8)	-103756 (-2.90)
4	Engineering	5110 (2.65)	21696 (2.67)	-16586 (-0.02)
5	Optoelectronics	20486 (10.6)	90859 (11.2)	-70373 (-0.54)
6	Fluid Dynamics	11927 (6.19)	71340 (8.77)	-59413 (-2.58)
7	Mathematical Physics and ML	19988 (10.4)	109125 (13.4)	-89137 (-3.04)
8	Geophysics	6292 (3.27)	28642 (3.52)	-22350 (-0.26)
12	Quantum Physics	17754 (9.21)	86872 (10.7)	-69118 (-1.47)
13	Materials Science	143 (0.07)	371 (0.05)	-228 (0.03)
17	Meteorology	10685 (5.55)	35295 (4.34)	-24610 (1.21)
18	Learning and Teaching	27960 (14.5)	45784 (5.63)	-17824 (8.88)
20	Astrophysics	15235 (7.91)	65264 (8.02)	-50029 (-0.12)

From Table 12 we see that while most of the topics are dominated by the male group there are a few exceptions. Topic 2: "Medical Physics", Topic 17: "Metereology", and Topic 18: "Learning and Teaching" stand out with an overrepresentation of the female authorships making up 8%, 6% and 15%, respectively (in terms of the total amount of female authorships). The largest group of male authorships are represented in Topic 3: "Compuational Physics" making up 16%. The fraction of male authors in the female dominant topics, Topic 2, 17, and 18 are 5%, 4% and 6%, respectively. This leaves Topic 17: "Learning and Teaching" as the Topic with largest disparity between the genders with 9pp. Topic 3: "Agricultural Physics", Topic 8: "Geophysics", and Topic 19: "Astrophysics" are the most equally distributed between the male and female authors making up 3%, 4%, and 8%, respectively, for both genders.

Furthermore, we notice Topic 13: "Materials Science" has a very low sample frequency in the overall dataset, presented in Figure 26(b). When grouped by gender it is even lower (as the genders falling into this topic is not meaningfully categorised). This indicates that the Materials Science Topic might not be very well defined and we probably can't conclude much from the Materials Science Topic results in the analysis.

Figure 28 shows the topic distribution over time and Figure 29 depicts the gender aggregated evolution of topics per publication year.

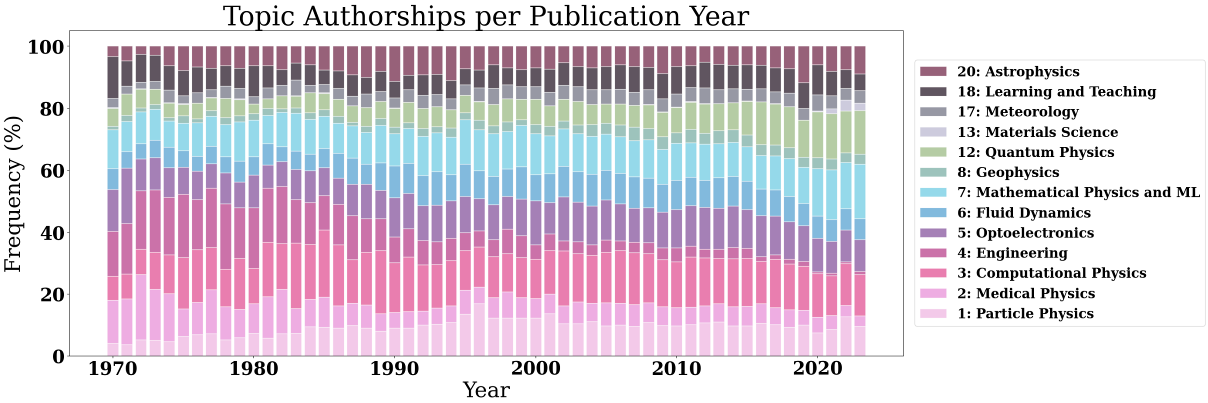


Figure 28: Topic distribution on authorship level over time. Each bar presents the fraction of which each topic made up per publication year.

From Figure 28 we especially notice that the prevalence of Topic 1: Particle Physics increase up till around publication year 1995. Topic 2: Medical Physics decrease over time. Topic 3: Computational Physics increase over time. Topic 4: Engineering decrease a lot over time and is almost eliminated from year 2015 and on. Topic 12: Quantum Physics increases a lot especially from year 2010. Topic 20: Astrophysics increase up till around year 1995.

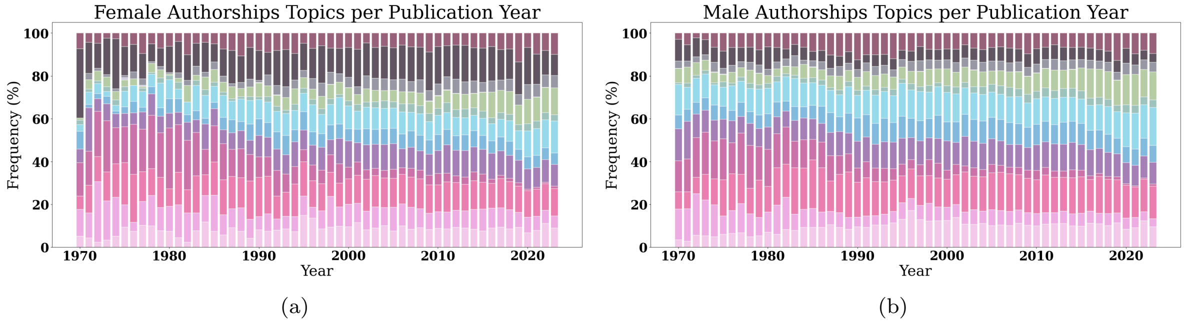


Figure 29: The fraction of topics per publication year per **a** female authorships and **b** male authorships. The color code corresponding legend is displayed in Figure 28.

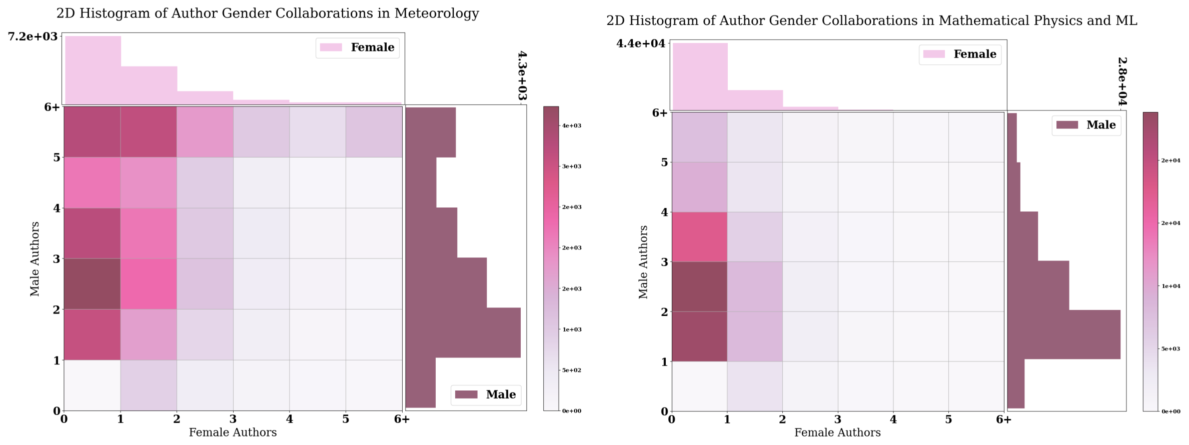
Comparing the overall trend in the topic distribution, Figure 28, over time to the gender aggregated distribution, Figure 29, we see that the trends look very similar. However, investigating the topic distribution over time per female authorships, displayed in Figure 29(a) in we do notice that the frequency of Topic 2: Medical Physics and Topic 18: Learning and Teaching is larger.

Specifically for Topic 2: Medical Physics we see that while the overall distribution decrease from year 1990 and onward the amount of female authors publishing within medical physics seem to stabilise.

Similarly when comparing the change in distribution between topics over time for male authorships, shown in Figure 29(b), we specifically observe that Topic 2: Medical Physics decrease more than it does for female authorships in 1990 and onwards. Topic 1: Particle Physics increase more.

The prevalence of Topic 18: Learning and Teaching is generally lower for male than female authorships throughout time.

Figure 30 displays the two most contrasting topic examples of author collaborations between genders, Meteorology and Mathematical Physics and ML, whereas the distribution between genders for the rest of the topics can be seen in Figure A12 Appendix A.



(a)

(b)

Figure 30: 2D histogram displaying the distribution of female and male authors per topic. We display the topic with the highest rate of including female authors **a** Topic 17: "Metereology" and the topic with lowest collaboration between genders **b** Topic 7: "Mathematical Physics and ML". The figure is a depiction of the level at which female and male authors collaborate. The histogram is created such that the values of the returned histogram are equal to the sum of the total number of authors belonging to the samples falling into each bin.

Overall, from Figure 30 and Figure Figure A12 Appendix A, we see that same pattern of all-male groups dominating in most topics except topic 17/domain "Didactic Research" where the collaborations are more equally divided between the genders but where there is also less authors per publication.

However, we do see some topics with more female representation on different author counts compared to others - this is especially true for the female dominated topics "Metereology", Figure 30(a) and "Medical Physics" Figure 12(b).

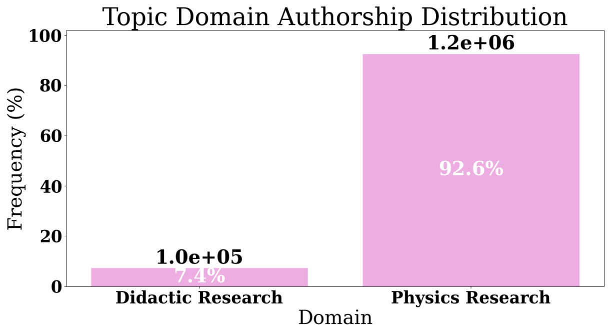
A counter example would be the (lack of) collaboration between genders found within Topic 7: "Mathematical Physics and ML" Figure 30(b).

Topic Domain

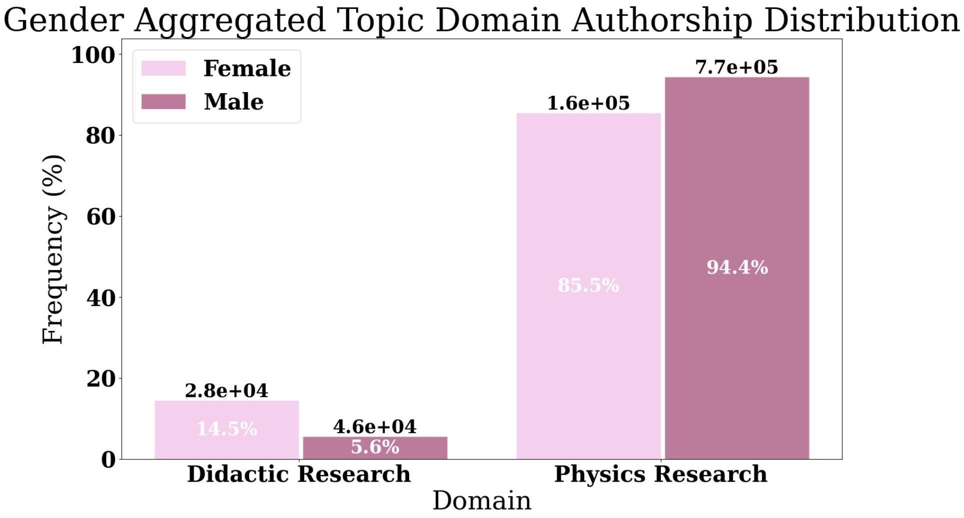
Since 13 distinct topics are a lot to analyse we group them into domains.

Investigating the topics distribution in Figure 26(b) we see that while most of the topics are within (more or less) traditional research fields within physics, Topic 17: "Learning and Teaching" differs therefrom. Therefore, we split the topics into two respective domains where "Learning and Teaching" is labelled as "Didactic Research" and everything else is labelled as "Physics Research".

Figure 31 show the distribution of domains and the gender aggregated domain distribution.



(a)

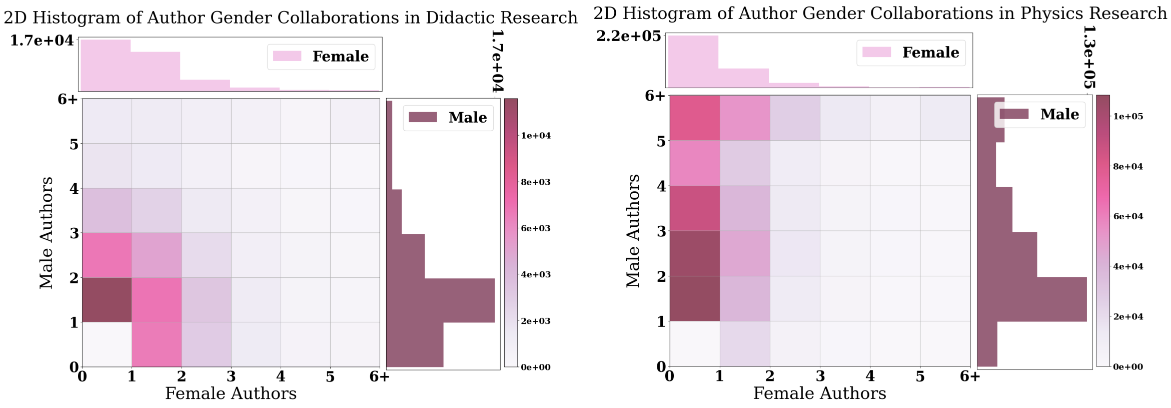


(b)

Figure 31: Distribution of domains in terms of **a** the overall authorships and **b** gender aggregated authorships. The domains are assigned based on the topic where topics labelled "Learning and Teaching" go into the "Didactic Research" domain and all other topics go into the "Physics Research" domain.

From Figures 31(a) and 31(b) we see that "Physics Research" make up most of the dataset with its 93%, but when segregating between female and male authorships the distribution is 95% for males and only 86% for females (in terms of the total count of males and females, respectively). Hence, in terms of domain male authors dominate "Physics Research" while female authors dominate "Didactic Research".

Figure 32 displays a 2D histogram of the gender composition per publication per domain.



(a)

(b)

Figure 32: 2D histogram displaying the distribution of female and male authors per publication. The figure is a depiction of the level at which female and male authors collaborate. The histogram is created such that the values of the returned histogram are equal to the sum of the total number of authors belonging to the samples falling into each bin.

From Figure 32 we see that in both domains the most prevalent case is 1 male author. However, in the Didactic Research domain the group of 1 female author is larger and the collaborations are more equally divided between male and female authors whereas in the Physics Research domain all-male author groups highly dominate.

Figure 33 depicts the evolution of domain authorships per gender over time. The distribution is normalised according to the total number of authorships in the given domain per year. The ratio corresponds to the amount of female and male authorships out of those available per domain per year.

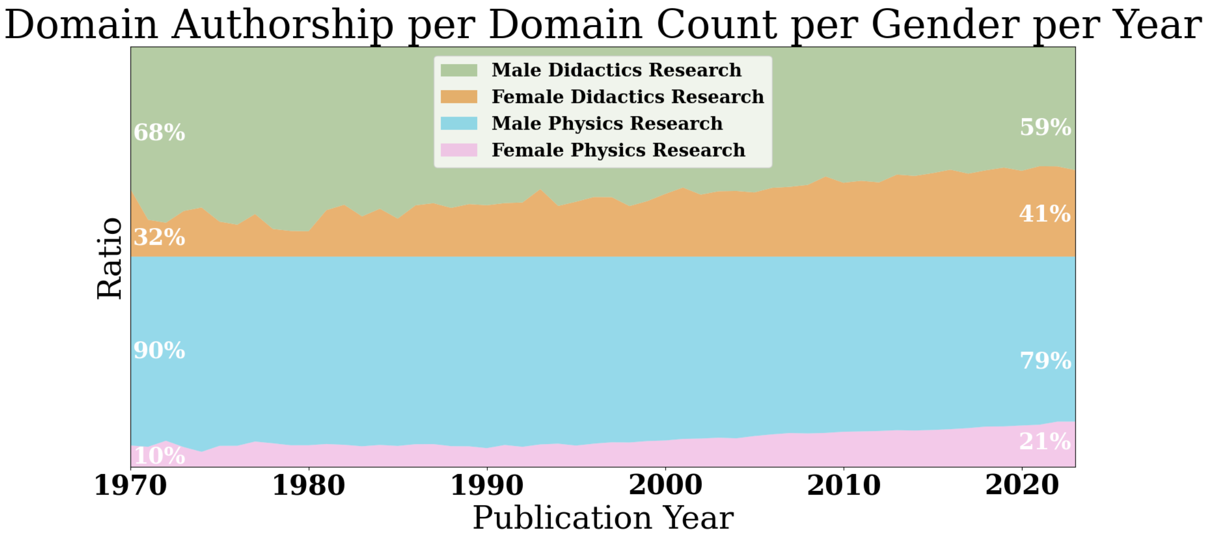


Figure 33: Evolution of distribution of authorships between the two domains per gender. It is normalised according to the number of total authorships within the domain per year. The ratio thus compares the number of authorships of a given gender out of the total count of authorships in a domain, physics or didactics, in the given publication year. That is, the lower half corresponds to the ratio of female and male authorships in Physics Research over time, and the upper half corresponds to the ratio of female and male authorships in Didactic Research over time. The text on the left hand side displays the percentage between genders in publication year 1970 and the percentage displayed on the right hand side corresponds to the ratio in 2023.

From Figure 33 we see that the normalised representation of male authorships are larger than that of female authorships in both domains. We also see that the fraction of female authorships increases over time for both domains. Between year 1970 and 2023 it increased with 9pp in the "Didactics Research" domain and with 11pp in the "Physics Research" domain.

In summation, female authors were mostly represented within topics Medical Physics, Metereology and Learning and Teaching. The topics Engineering, Geophysics, and Astrophysics were equally dominated by male and female authors. The rest of the topics were dominated by male authors.

The evolution of topics over time seemed somewhat stable, except for topics Engineering, Particle Physics, and Medical Physics. This was true when grouped by gender as well.

The level of collaboration between genders were largest within female and equally dominated topics.

When grouped by domain, female authors dominated "Didactic Research" whereas male authors dominated "Physics Research".

The distribution of female authorships have increased over time within both domains.

Table 13: This table presents an overview of the clean topics, their bag-of-words and the respective weight of each word as well as an abstract section of the best scoring abstract and the matching score between that topic and abstract.

Topic Labels and Best Matching Abstract Examples				
	Topic Label	**Bag-of-Words with Weights**	**Best Matching Abstract Section**	**Abstract Score**
1	Particle Physics	1.4”*neutrino*” + .7”*lepton*” + .7”*particl*” + .7”*decay*” + .7”*quark*” + .5”*violat*” + .5”*search*” + .5”*measur*” + .4”*proton*” + .4”*precis*”	...The Lorentz boost of these heavy particles changes the topology of their decay products. The classical method to identify and to reconstruct their hadronic decays is not adequate for these objects. In order to manage this boosted topology, new techniques have been developed that reconstruct the boosted object as a single jet and distinguish it from the background through a substructure analysis. In order to select and to reconstruct top quark pairs at the LHC, two methods for selecting lepton+jets events are discussed: the classical, resolved approach, developed at the Tevatron for top quark production at rest, and a new technique designed specifically to deal with boosted top quark production...	0.984112
2	Medical Physics	.6”*patient*” + .6”*imag*” + .6”*protein*” + .5”*clinic*” + .5”*cancer*” + .4”*treatment*” + .4”*diseas*” + .3”*radiat*” + .3”*therapi*” + .3”*cell*”	...Biomedical applications of systems biology and biological physics’. The aim is to show how systems biology can help us establish a more detailed understanding of the processes underlying the regulation and functioning of living systems and how biological physics can provide us with new effective tools to investigate biological phenomena at the molecular as well as the cellular and the physiological levels. We also demonstrate how these insights and techniques can contribute towards a better treatment of patients with diseases such as cancer and Parkinson’s tremor...	0.988924
3	Computational Physics	.5”*perform*” + .4”*base*” + .4”*control*” + .4”*sensor*” + .3”*develop*” + .3”*optim*” + .3”*neutron*” + .3”*use*” + .3”*algorithm*” + .3”*softwar*”	...Experiments have been conducted with different datasets and on the SuperMUC machine at the Leibniz-Rechenzentrum, the local CoolMAX AMD GPU cluster in Munich, the Phi-accelerated Beacon at University of Tennessee, and the Todi Cray XK7 at the Swiss National Supercomputing Centre. The performance results show that for strong scaling settings, GPUs and coprocessors suffer from lack of parallelism and do not perform as well at large scale. For weak scaling settings however, they always outperform. Several challenges, as stated previously, have been identified in order to create large-scale computing systems that meet current application requirements...	0.988624
4	Engineering	1.7”*work*” + .8”*agricultur*” + .7”*engin*” + .7”*chemistri*” + .6”*microanalysi*” + .6”*supplement*„ + .6”*horticultur*” + .5”*rural*” + .5”*laboratori*” + .4”*correspond*”	...Virtual 17th International Conference on Particle Induced X-ray Emission (PIXE 2021) Toyama, Japan AVS 67 Focus Topic: New Trends in Structural Electronic Characterization of Materials, Interfaces, and Surfaces Using Synchrotron and FEL Based Light Sources (LS) Charlotte, North Carolina, USA ICDD X-ray Fluorescence Clinic ICDD Headquarters, Newtown Square, PA, USA 57th Annual Conference on X-Ray Chemical Analysis...	0.971364
	⋮	⋮	⋮	⋮

	⋮	⋮	⋮	⋮
5	Optoelectronics	.7”*devic*” + .4”*optic*” + .4”*electron*” + .4”*xmln*” + .3”*voltag*” + .3”*applic*” + .3”*puls*” + .3”*base*” + .3”*electr*” + .3”*metal*”	...Herein, the Special Issue focuses on important concepts in artificial photosynthesis for versatile energy- and environmental-related applications such as photo(electro)chemical H2 evolution, photoelectrochemical O2 evolution, photo(electro)catalytic CO2 reduction, pollutant degradation, and so forth. In the 21st century, the widespread utilization of carbon-based fossil fuels (e.g. coal, oil and natural gas) for the generation of energy and electricity driven by rapid economic growth has come at a cost...	0.988034
6	Fluid Dynamics	1.1”*flow*” + .6”*heat*” + .6”*fluid*” + .4”*pressur*” + .4”*simul*” + .4”*numer*” + .3”*thermal*” + .3”*veloc*” + .3”*droplet*” + .3”*predict*”	...The article by Cunha and Andreotti contributes to the characterization of the mechanism of drag reduction with low volume fractions of anisotropic additives in turbulent channel flow.Studies of the flow past a square heated cylinder with its longitudinal axis aligned normal to the direction of the flow are motivated, aside by the fundamental significance, by their importance in applications such as combustion chambers in chemical processes, flow dividers in polymer processing, cooling of electronic components, and compact heat exchangers...	0.987372
7	Mathematical Physics and ML	.8”*equat*” + .7”*neural*” + .6”*learn*” + .5”*pinn*” + .5”*network*” + .5”*problem*” + .5”*propos*” + .4”*function*” + .4”*algorithm*” + .4”*approach*”	... Interpreting the concepts of open and closed systems rigorously means that the SL must be satisfactorily constrained by all field equations. However, in open systems dynamics, in which body forces and energy supplies may have arbitrary values (we assume in this discussion that no further evolution equations occur), only the mass balance equation contains no external source and must be accounted for in the exploitation of the entropy inequality.Apart from this, the source terms of momentum and energy may take any values. In particular, the body force f and the energy supply r may at any material point and any time take values such that they outbalance all other terms in the remainder of these equations...	0.986105
8	Geophysics	1.1”*seismic*” + 1.1”*reservoir*” + .8”*fractur*” + .7”*earthquak*” + .5”*fault*” + .4”*stress*” + .4”*poros*” + .4”*invers*” + .4”*elast*” + .4”*predict*”	...Given that any remediation system has a cost and a carbon/energy footprint, we need every cleanup to have a distinct benefit relative to those costs and co-pollution considerations. This editorial recaps the technical underpinning of Tn, its complexities, its derivations from field tests, and its potential utility in decision making. For consistency in our review, we will use risk-based corrective action (RBCA) concepts in defining various attributes of Tn...	0.984671
12	Quantum Physics	1.0”*topolog*” + 1.0”*spin*” + .7”*magnet*” + .7”*phase*” + .6”*electron*” + .6”*state*” + .6”*transit*” + .4”*lattic*” + .4”*coupl*” + .4”*phonon*”	...Unlike in ordinary metals or semiconductors, the mutual interactions between electrons may not be neglected in materials with partially filled d or f shells. Correlated electron materials display an exceptionally rich variety of quantum-mechanically entangled insulating, metallic, magnetic, or superconducting states. Our Transregio has combined highly advanced instrumentation and theoretical tools for the exploration of correlated quantum matter with new functionalities...	0.988280
	⋮	⋮	⋮	⋮

	⋮	⋮		⋮	⋮
13	Materials Science	7.8”*omnisci*” 7.5”*ceram*” 7.4”*annual*” 6.7”*synthesi*” 6.6”*polym*” 6.2”*medicin*” 6.1”*cover*” 5.8”*composit*” 5.5”*onlin*” 5.3”*biolog*”	+ + + + + + + + +	Advanced Materials Science and Technology is a peer-reviewed open access journal published semi-annual online by Omniscient Pte. Ltd. The journal covers the properties, applications and synthesis of new materials related to energy, environment, physics, chemistry, engineering, biology and medicine, including ceramics, polymers, biological, medical and composite materials and so on.	0.950434
17	Meteorology	1.0”*climat*” .8”*forecast*” .6”*soil*” .5”*atmospher*” .4”*precipit*” .3”*aerosol*” .3”*observ*” .3”*predict*” .3”*rainfal*” .3”*simul*”	+ + + + + + + + +	...A primary aim was to provide data for use in the development of models to describe the interactions between the global climate system, plankton functional biodiversity, and ocean/atmosphere biogeochemistry. The specific objectives were to determine (1) how the structure, functional properties, and trophic status of the major planktonic ecosystems vary in space and time; (2) how physical processes control the rates of nutrient supply, including dissolved organic matter, to the planktonic ecosystem; and (3) how atmosphere–ocean exchange and photo-degradation influence the formation and fate of organic matter...	0.988093
18	Learning and Teaching	2.1”*learn*” 1.0”*teach*” .8”*skill*” .5”*develop*” .5”*use*” .5”*think*” .4”*base*” .4”*subject*” .4”*valid*” .4”*problem*”	+ + + + + + + + +	...In each of these school districts, a middle school and a high school were selected by the school district for participation in the survey. In addition, the sampling frame included students enrolled in grades 5 through 8 in four private elementary schools and students enrolled in grades 9 through 12 in four private high schools. From these 11,200 youth, 4,032 usable surveys were obtained...	0.988921
20	Astrophysics	.7”*observ*” .6”*gravit*” .6”*cosmic*” .6”*star*” .5”*stellar*” .5”*galaxi*” .4”*astrophys*” .4”*emiss*” .4”*binari*” .4”*radio*”	+ + + + + + + + +	...Conceptual Cosmological Process Based on physical and theoretical considerations above, a novel cosmological process is described that links past galaxy formation to unresolved dynamics observed today. In the proposed scenario, matter in the very early universe was subject to highly relativistic conditions which induced flux “confinement,” similar to the experiment above...	0.986057

Prestige Markers Between Genders

We defined rankings in terms of *Institution ID*, *Journal ID*, and *Author ID* as described in Section *Prestige Markers*. This section was conducted on Authorship Level.

Table 14 show the hypothesis testing result testing whether the mean for each ranking type between the genders was statistically different. We hypothesised that the female mean would be less than the male mean. The significance threshold was set at .05.

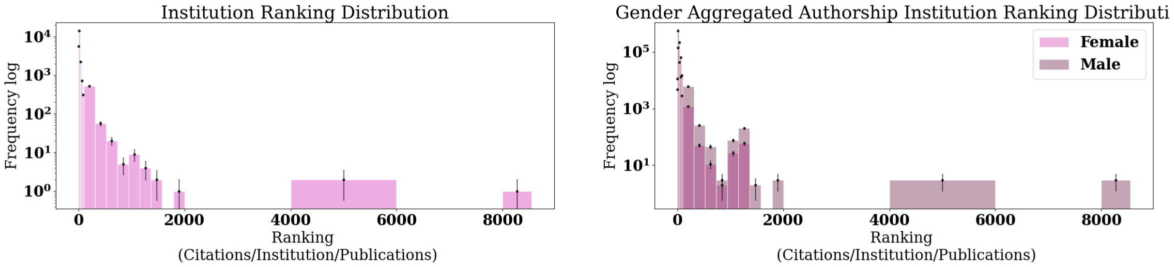
Table 14: Ranking hypothesis testing results. We tested whether the ranking variables were statistically different between the two genders. We hypothesised that the female mean would be less than the male mean. The significance threshold was set at .05. Significant test results, meaning the female mean was significantly smaller than the male mean, are marked in green, and insignificant results, meaning the distribution between female and male authorships were equal, are marked in red.

Ranking Hypothesis Testing Results					
Variable	**Female Mean (SD)**	**Male Mean (SD)**	**T-statistic, p-value**	**U-statistic, p-value**	**Figure**
Institution Ranking	22.14 (31.79)	24.30 (35.50)	-27.23, > .001	8.507e+10, > .001	34(b)
Journal Ranking	22.95 (50.55)	25.25 (57.28)	-17.39, > .001	7.323e+10, > .001	38(b)
Author Ranking	20.20 (97.58)	25.54 (101.3)	-22.32, > .001	8.088e+10, > .001	42(b)

From Table 14 we see that female authors are ranked significantly lower in terms of all three types of ranking: institution ranking, journal ranking, and author ranking.

Institution Ranking

Figure 34 displays the *Institution Ranking* distribution overall as well as gender aggregated.



(a)

(b)

Figure 34: Institution Ranking distribution per **a** *Institution ID* and **b** *Binary Gender*.

From Figure 34(a) we see that most institutions are ranked within 0-2000 citations/institution/publication but with a tail up til ∼ 8000 citations/institution/publication. Figure 34(b) reveals that the tail only includes male authorships and none of the female authorships score an institution ranking above 1400 citations/institution/publication. The hypothesis test results displayed in Table 14 reveal that the difference in institution ranking between genders is statistically significant.

Figure 35 displays the top 20 ranked institutions in our dataset.

From Figure 35 we notice that the highest scoring institution ”Quansight (United States)” ranked at score ∼ 8000 citations/institution/publication is by far the highest ranking. The second place is taken by ”Sirtex (Australia)” with institution ranking ∼ 5000 citations/institution/publication. The lowest ranked institution within top 20 is ”John F. Kennedy Medical Center” with ranking ∼ 1000 citations/institution/publication.

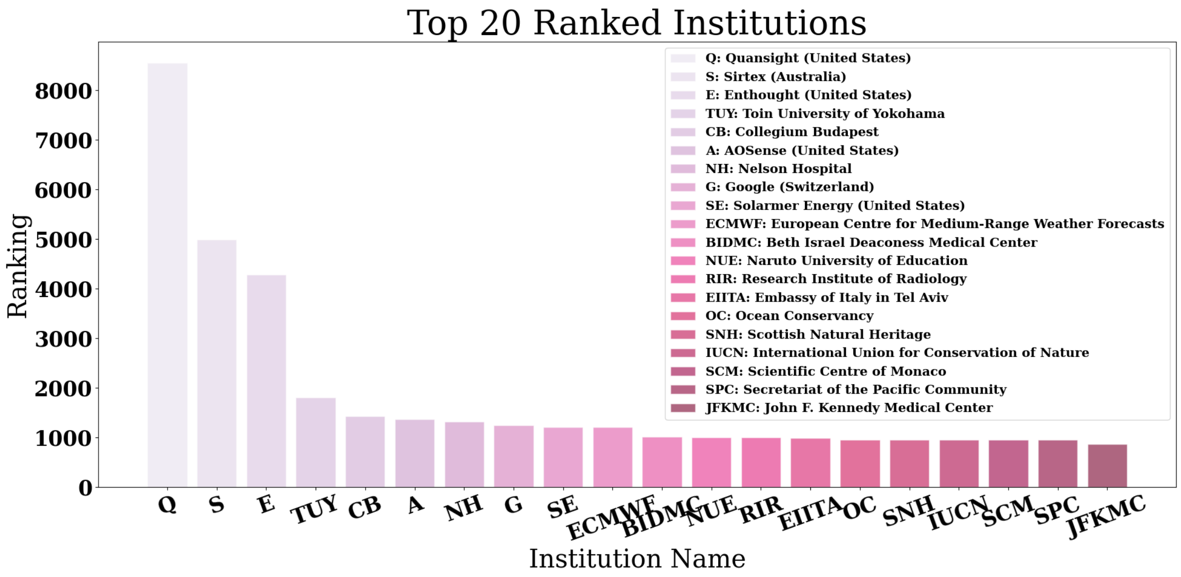


Figure 35: Top 20 highest ranked institutions in our dataset.

Figure 36 displays the percentage wise distribution of female and male authorships for our top 20 ranked institutions.

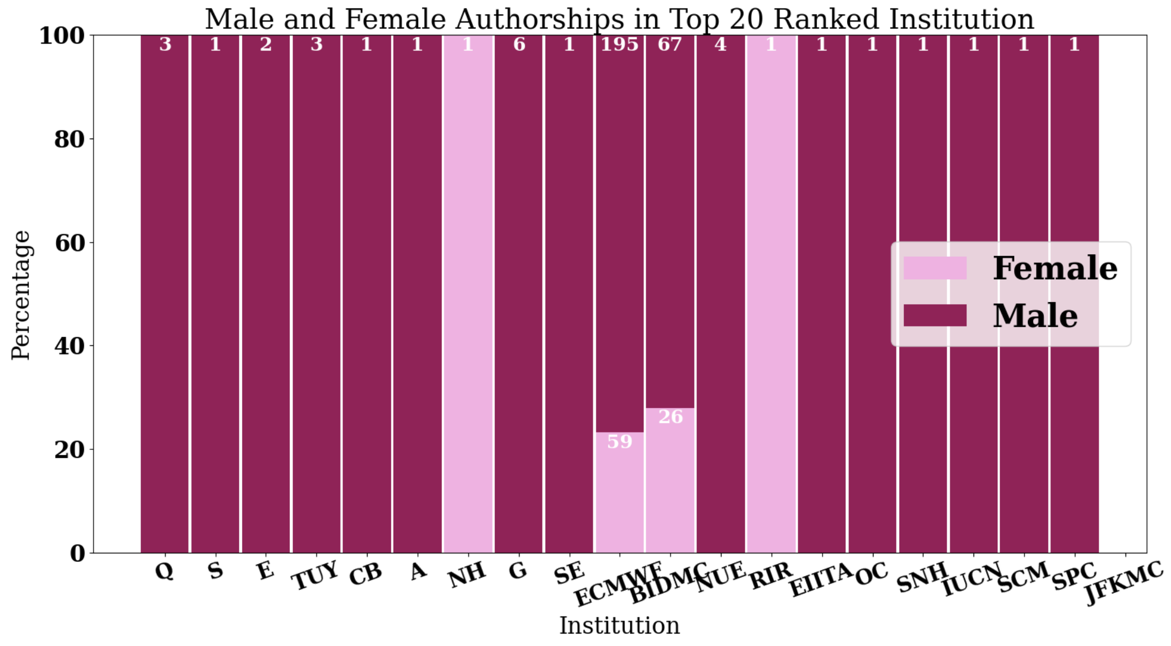


Figure 36: Distribution of female and male authorships in percentage for our top 20 highest ranked institutions in our dataset. The corresponding x-tick legend is displayed in Figure 35. The text displayed on each bar corresponds to the count of authorships of the given for each institution listed.

Figure 36 reveals that only sixteen out of twenty top ranked institution only include male authorships, leaving four to include female authorships the first appearing at top seven. Two of those four are 100% female authors. However, we see that almost all of them are very low statistics including only 1 authorship overall in our dataset. For the two institutions, ”European Centre for Medium-Range Weather Forecasts” and ”Beth Israel Deaconess Medical Center”, in top 20 that include both female and male authorships we see that the number of females are slightly larger (around $25 - 30\%$) compared to our overall gender aggregated authorships distribution in our dataset as presented in Figure 10.

Figure 37 shows the overall distribution of percentage of female and male authorships per institution ID. Since Figure 36 only provides a description of the distribution of female and male authorships for the top 20 ranked institutions We look into the overall distribution of percentage of female and male authorships for all institutions in the dataset displayed in Figure 37.

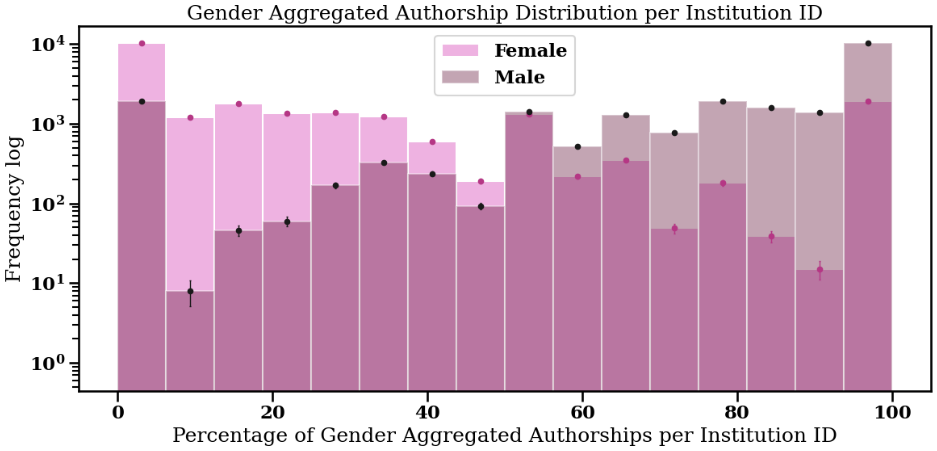
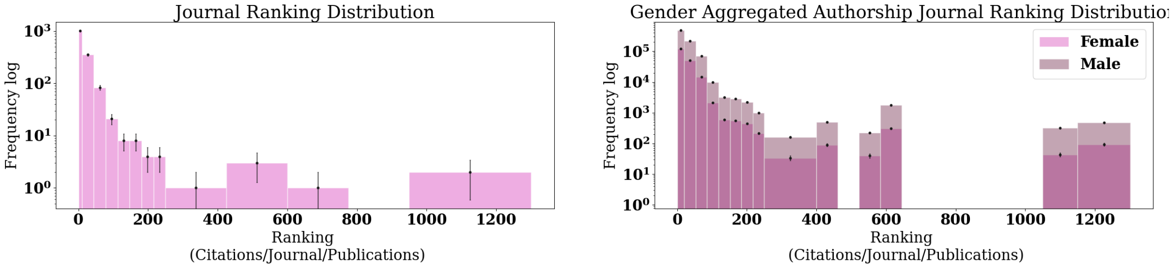


Figure 37: Distribution of percentage of female vs male authors per each distinct institution in the dataset. The x-axis displays the percentage of the given gender and the y-axis the frequency with which that percentage happens per institution ID.

Figure 37 show that the distribution of percentage of female authorships per institution ID is highest at 0% whereas the most occuring percentage of male authorships is 100%.

Journal Ranking

Figure 38 displays the *Journal Ranking* distribution overall as well as gender aggregated.



(a)

(b)

Figure 38: Journal Ranking distribution per **a** *Journal ID* and **b** *Binary Gender*.

From Figure 38(a) we see that most journals are ranked within 0-200 citations/journal/publication but with a tail up till ∼ 1300 citations/journal/publication. Figure 38(b) show that the journal ranking seem equally distributed between male and female authors. The hypothesis test results displayed in Table 14 reveal that the difference in journal ranking between genders is statistically significant.

Figure 39 displays the top 20 ranked journals in our dataset.

From Figure 39 we notice that the three highest scoring journal "Journal of Molecular Structure", "Advances in Physics", and "Journal of Applied Probability" ranked at score ∼ 1300 citations/journal/publication is by far the highest ranked journals ranked double as high as the fourth ranked journal. The fourth ranked journal "The Philosophical Magazine" is ranked at ∼ 600 citations/journal/publication. The lowest ranked journal within top 20 is "Journal of Physics and Chemistry of Solids" with ranking ∼ 200.

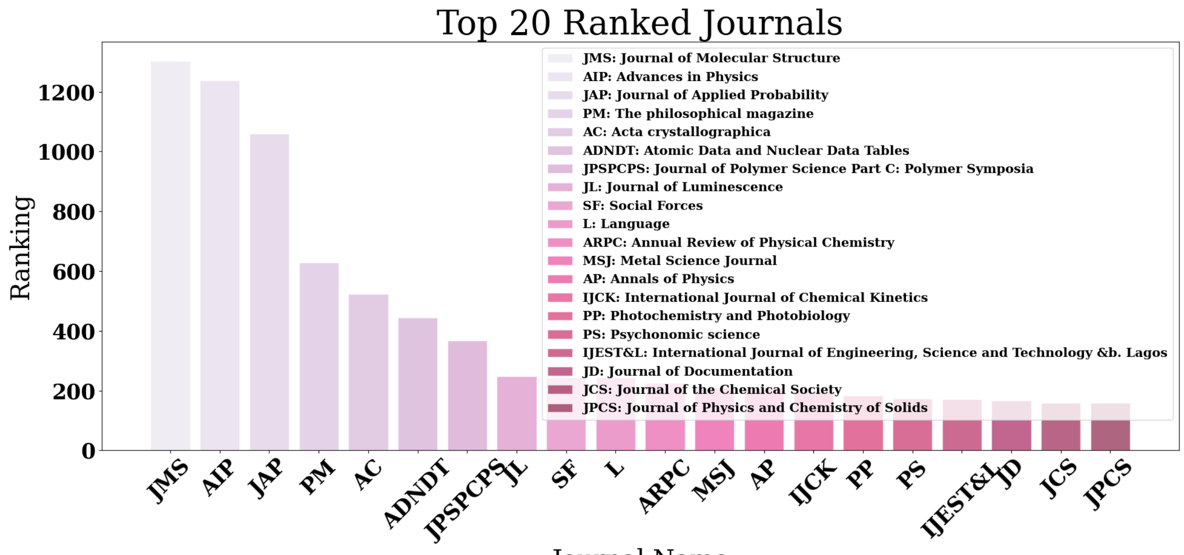


Figure 39: Top 20 highest ranked journals in our dataset.

Figure 40 displays the percentage wise distribution of female and male authorships for our top 20 ranked journals.

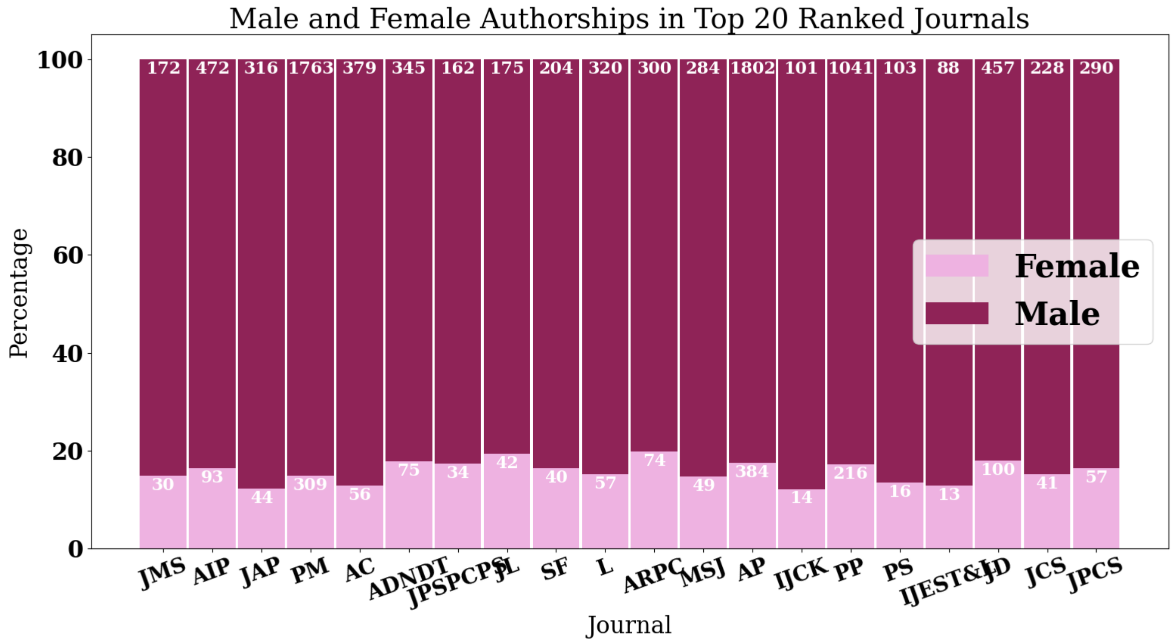


Figure 40: Distribution of female and male authorships in percentage for our top 20 highest ranked journals in our dataset. The corresponding x-tick legend is displayed in Figure 39. The text displayed on each bar corresponds to the count of authorships of the given for each journal listed.

Figure 40 reveals that the distribution of female and male authorships for our top twenty ranked journals follows a distribution somewhat similar to the overall distribution of authorships between each gender as seen in Figure 10 namely around $20 - 80$ even though for our top journals it seemed to be a bit more pushed towards $15 - 85$ rather.

Figure 41 shows the overall distribution of percentage of female and males per journal ID. Since Figure 40 only provides a description of the distribution of female and male authors for the top 20 ranked journals

We look into the overall distribution of percentage of female and male authorships for all journals in the dataset.

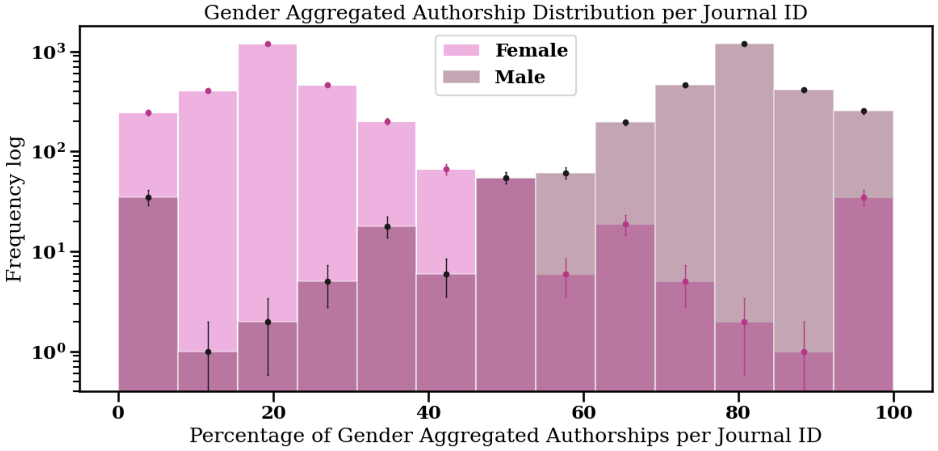


Figure 41: Distribution of percentage of female vs male authors per each distinct journal in the dataset. The x-axis displays the percentage of the given gender and the y-axis the frequency with which that percentage happens per journal ID.

Figure 41 shows that the percentage of female authorships per journal ID has its top point at 20% whereas the distribution of percentage of male authorships has its top point at 80%. This aligns with our general distribution of authorships in the dataset as depicted in Figure 10.

Author Ranking

Figure 42 displays the *Author Mean Cites* distribution overall as well as gender aggregated.

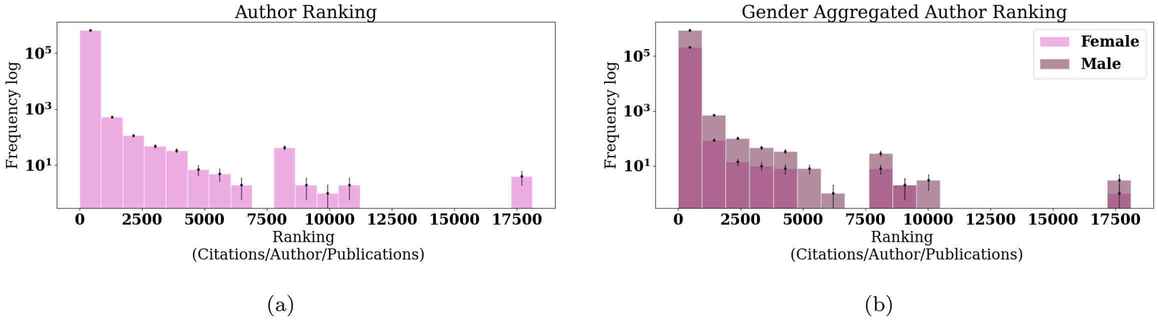


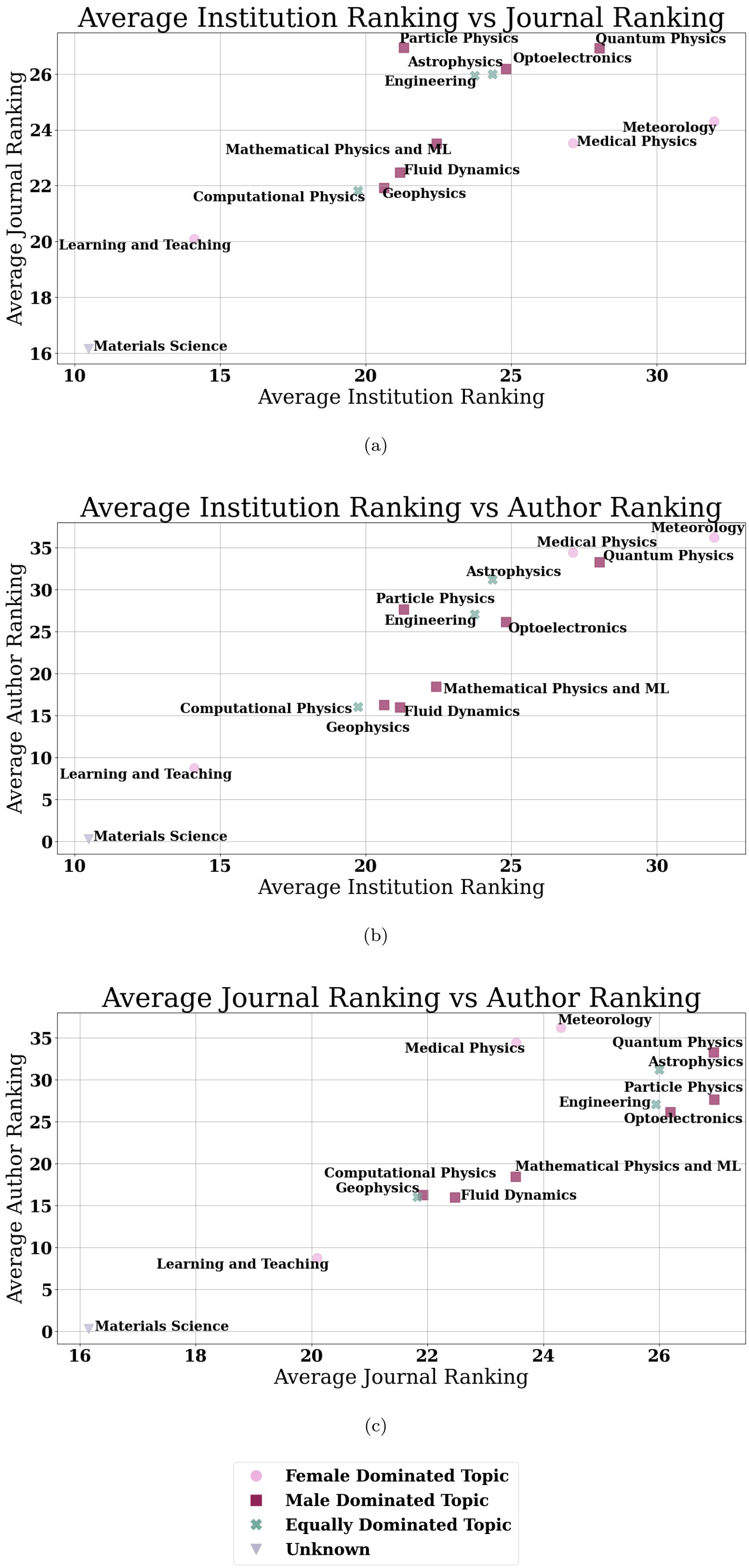
Figure 42: Author Ranking distribution per **a** *Author ID* and **b** *Binary Gender.*

From Figure 42(a) we see that most authors are ranked within range 0-7000 citations/author/publication but with a tail up till ~ 18000 citations/author/publication. From Figure 42(b) we see that the author ranking is somewhat equally distributed between male and female authors, however male authors are more common in the mid-range window between $\sim 5000 - 10000$ citations/author/publication. If we assume that the single bin at ~ 18000 is an outlier, the mid-range window would actually be the upper-range. The hypothesis test results displayed in Table 14 reveal that the difference in author ranking between genders is statistically significant.

Ranking and Topic Correlations

We investigated if and how ranking differentiated between topics.

Figure 43 displays the correlation between mean ranking grouped by topic. It is indicated with distinct markers (Figure 43(d)) whether the given topic is female, male or equally dominated according to what we found from Figure 27.



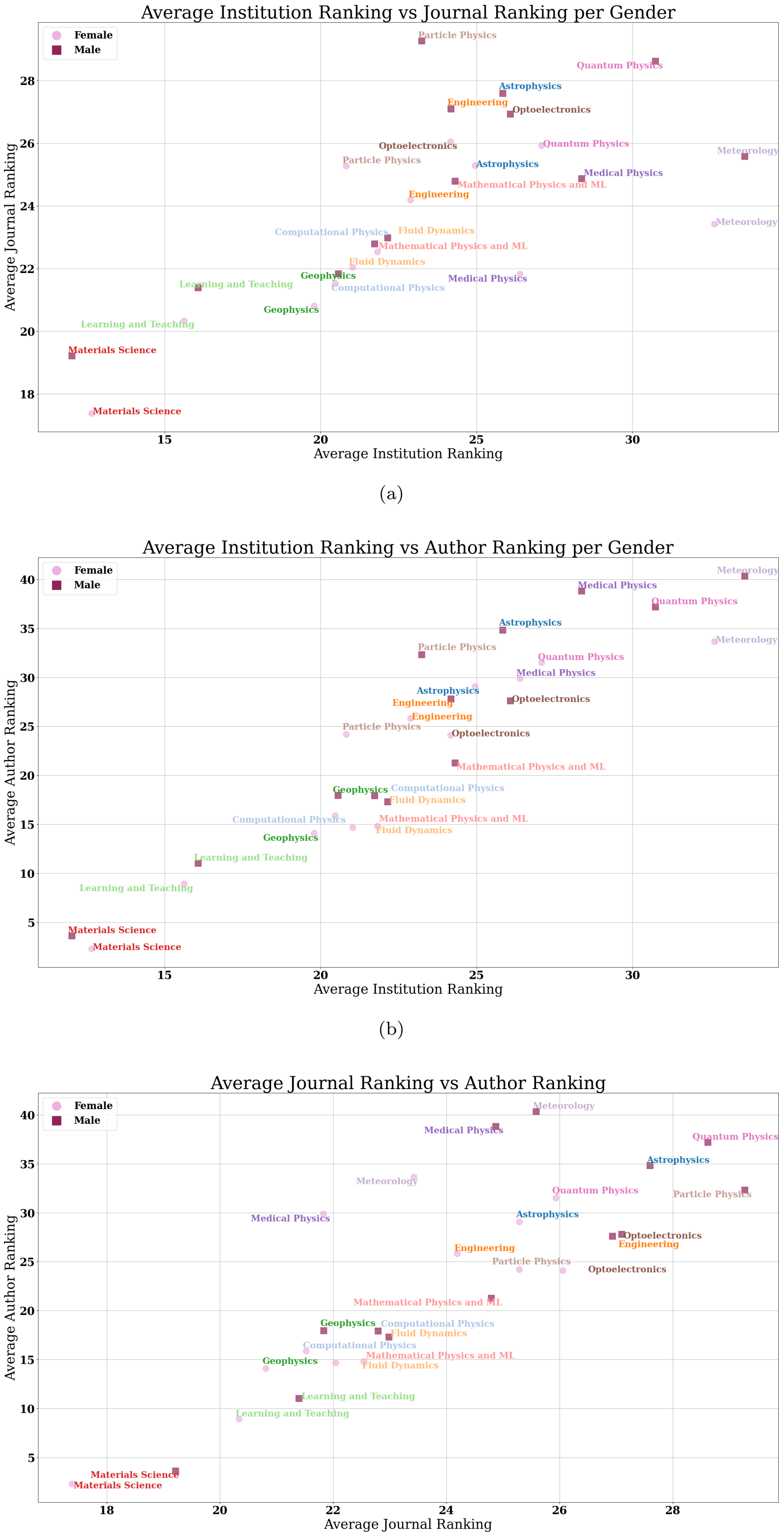
(a)

(b)

(c)

(d)

Figure 43: Mean correlations between **a** institution and journal, **b** institution and author, and **c** journal and author rankings per topic. The upper right corner indicates high correlation, while the lower left corner indicates low correlation. Topics are marked as female, male, or equally dominated (see **d**).



(a)

(b)

(c)

Figure 44: Mean ranking correlations per topic between **a** institution and journal ranking, **b**, institution and author ranking, and **c** journal and author ranking grouped by gender. The average female ranking is marked in pink and the average male ranking in copper. Each topic label is color coded so the same topic appears in the same color for both genders. The further in the upper right hand corner, there is a high correlation between mean ranking for the given topic, and the further placed in the left corner, there is a correlation of low ranking for the given topic

Investigating Figure 43 we notice that for all three rankings Topic 18: "Learning and Teaching" and Topic 13: "Materials Science" are ranked lowest. Ignoring Topic 13, then the most female dominated topic - Learning and Teaching - is the lowest ranked. However, we see that the other two female dominated topics, Topic 2: "Medical Science" and Topic 17: Metereology" are ranked among the highest in all three cases.

The equally dominated topics, Topic 20: "Astrophysics" and Topic 4: "Engineering", are also quite highly ranked. The male dominated "Quantum Physics" are in top three highest ranked in all three cases.

The rest of the male dominated topics are distributed from the middle to top ranking.

Specifically, exploring the Institution and Ranking correlation per topic displayed in Figure 43(a) Quantum Physics (male dominated) has the highest score in terms of journal ranking. However, Meterology (female dominated) has the highest score in terms of institution ranking.

Between Institution and Author Ranking 43(b) Metereology (female dominated) has the highest overall score followed closely by Medical Physics (female dominated) and Quantum Physics (male dominated).

Comparing Journal Ranking and Author Ranking we see that while Quantum Physics has the highest journal ranking, Metereology is ranked highest in terms of author cites. 43(c)

Figure 44 shows the correlation between each type of ranking grouped by gender. To explore whether it is the male or female authors within the female/male dominated topics that drive the ranking that we see in Figure 43 we group the data by gender and plot the same relations.

Figure 44 shows that even though the female dominated topics (except for Learning and Teaching) are highly ranked there is a generally lower ranking for female.

Figure A24 in Appendix displays the descriptive statistics of ranking per topic as a violin plot.

Figure A21-A23 in Appendix A show correlations between each ranking type, institution, journal, and author, respectively.

Activity and Survival Rate Between Genders

We define the necessary variables for conducting an activity and survival analysis as described in Section *Activity and Survival Rate*. Figures 45-47 show the distribution and gender aggregated distributions of these variables. As this part of the analysis include drop out definition the analysis is only conducted on the authors defined as active according to our activity definition. That is, the author should have their first publication in 1975 or later and their last publication in 2018 or later. The dataset containing so-called active authors contain 234,365 rows entailing 70,556 unique authors and 147,599 unique publications. The analysis in this section is conducted on Author Level.

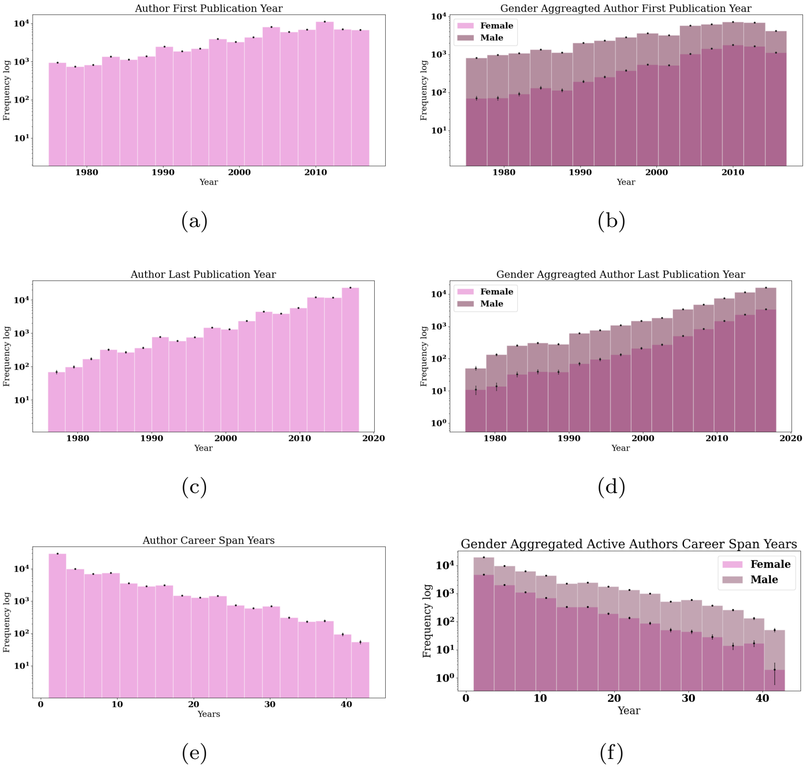
Figure 45 displays the first and last publication year per author as well as the number of years between them.

From Figure 45(a) we see that the frequency of the publication year slightly increase over time. This is in conformity with what we saw in Figure 13(a); that the number of publications increase over time. In the gender aggregated version Figure 45(b) we see the same trend especially for female authors which also aligns with what we saw in Figure 13(b) that the number of female authorships increase over time.

Figure 45(c) depicting the frequency of last publication year we see that this too increase over time, indicating that many authors included for this analysis were still active up till a recent point. The gender aggregated version in Figure 45(c) show the same trend for both genders.

Lastly, Figure 45(e) depicting the number of years between first and last publication year per authors show that as the number of years increase the frequency is lower meaning fewer authors have a longer career. Looking at the gender aggregated version in Figure 45(f) we see that this is especially true for female authors - but remembering Figure 45(b) their first publication year is also skewed towards more recent years.

Figure 46 displays the total number of publications per author as well as their publication rate defined as number of publications per career span years.

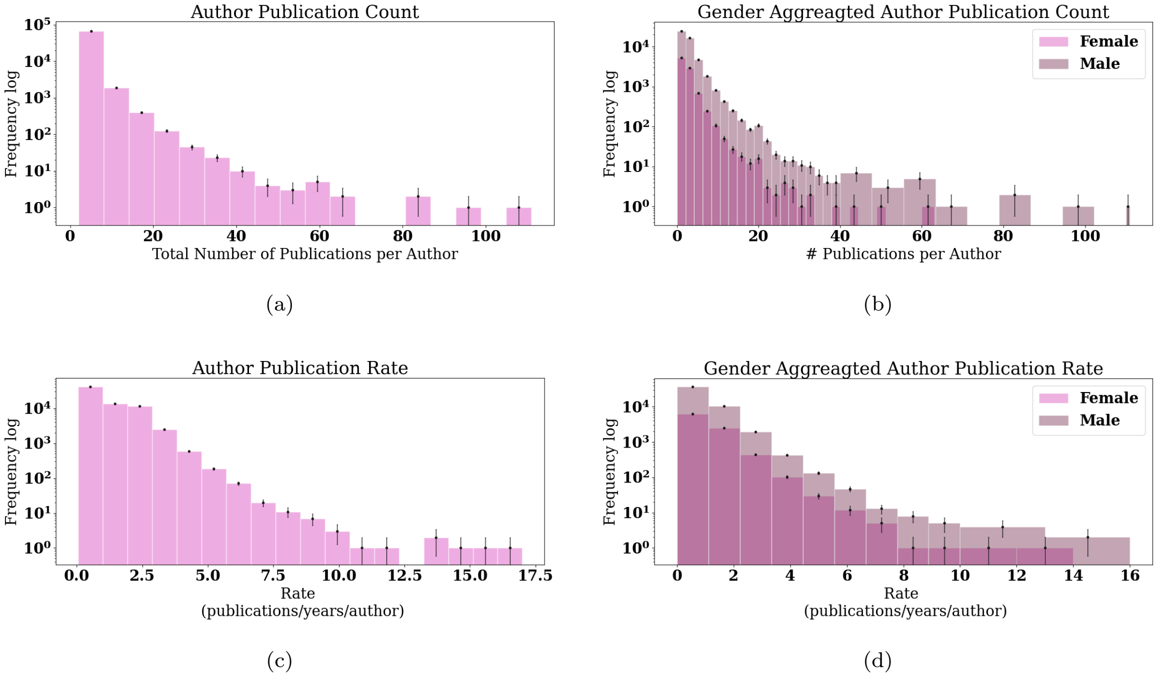


(a) (b)

(c) (d)

(e) (f)

Figure 45: Distribution of **a, b** first publication year, **c, d** last publication year and **e, f** career span years of the authors defined as active within the given activity year window 1975-2018 (as described in Section **Activity and Survival Rate**). Career span years is defined as the number of years between first and last publication year. The figures in the left hand column show the overall distribution and the figures on the right hand side column the gender aggregated distribution.



(a) (b)

(c) (d)

Figure 46: Distribution of author **a, b** publication count and **c,d** publication rate. The publication count represents the number of total publications per author and the publication rate is the number of publications per career span years. The figures in the left hand column show the overall distribution and the figures on the right hand side column the gender aggregated distribution.

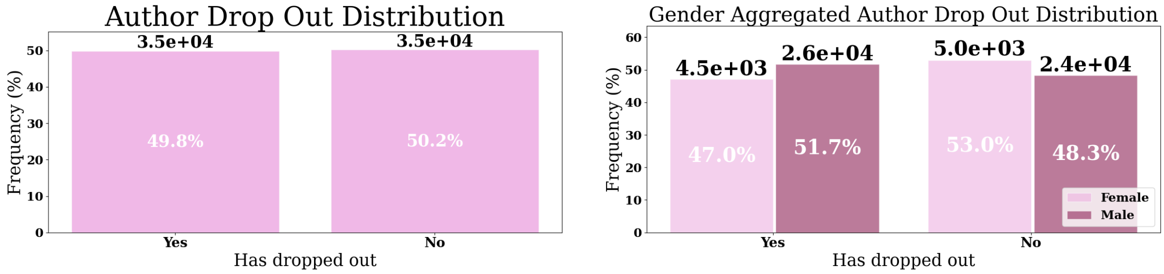
From Figure 46(a) we see that the total number of publications per author within our activity window is ranging from $\sim 1 - 120$ publications with most being distributed in the lower end.

Compared to the gender aggregated distribution in Figure 46(b) we see a similar trend, however no female author published more than 60 publications in total. Looking at the publication rate displayed in Figure 46(c) we see the publication rate ranges from $\sim 1 - 16$ publications per year with most of the authors in the lower end.

Compared to the gender aggregated version in Figure 46(d) we see the same trend, however fewer female authors are distributed above 8 publications/year.

Figure 47 displays the drop out distribution of authors in our activity window.

An author is set to have dropped out if the *Last Publication Year* is in 2013 or earlier as described in Section **Activity and Survival Rate**.



(a)

(b)

Figure 47: Drop out distribution for **a** all authors and **b** grouped by gender. An author is assumed to have dropped out if their *Last Publication Year* was in 2013 or earlier.

The drop out distribution displayed in Figure 47(a) shows that there is an almost 50/50 distribution of the authors who have and haven't dropped out in the activity analysis. The gender aggregated version, Figure 47(b), shows a similar distribution. However, a slightly smaller percentage of the female authors have dropped out - but remembering Figure 45(b) their first publication year is also skewed towards more recent years.

Table 15: Hypothesis testing results. We tested whether the activity rate variables were statistically different between the two genders. For *First Publication Year*, *Last Publication Year* and *Has Dropped Out* we hypothesised that the mean for female authors is greater than for male authors but for the rest of the variables we hypothesised that the mean for female authors is less than for male authors and thus used a one-sided test. The significance threshold was set at .05. Significant test results are marked in green, and insignificant results are marked in red.

Activity Variables Hypothesis Testing Results					
Variable	**Female Mean (SD)**	**Male Mean (SD)**	**T-statistic, p-value**	**U-statistic, p-value**	**Figure**
First Publication Year	2006 (8.38)	2003 (9.99)	31.2, > .001	2.79e+08, > .001	45(b)
Last Publication Year	2012 (6.42)	2011 (7.19)	13.46, > .001	2.54e+08, > .001	45(d)
Career Span Years	5.82 (6.04)	7.86 (7.71)	-28.8, > .001	2.00e+08, > .001	45(f)
Author Publication Count	3.10 (2.38)	3.41 (2.80)	-11.1, > .001	2.18e+08, > .001	46(b)
Author Publication Rate	1.07 (0.90)	0.92 (0.86)	15.5, 1.0	2.67e+08, 1.0	46(d)
Has Dropped Out (no=0, yes=1)	0.47 (0.50)	0.52 (0.50)	-8.34, 1.0	2.26e+08, 1.0	47(b)

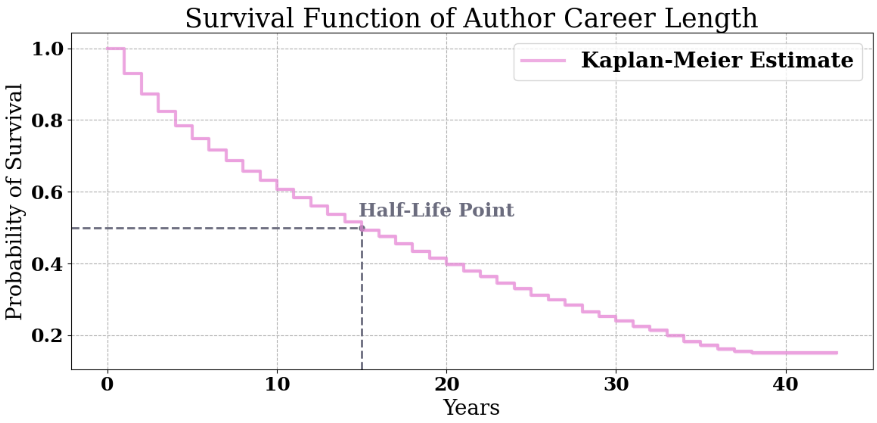
CTable 15 presents the statistical test outcomes comparing the means for each of the activity and survival variables presented in Figure 45-47 between female and male authors.

For each variable we compare the mean between female and male authorships to test if the means are statistically different. We used both Welsch's t-test as well as Mann Whitney U-test.

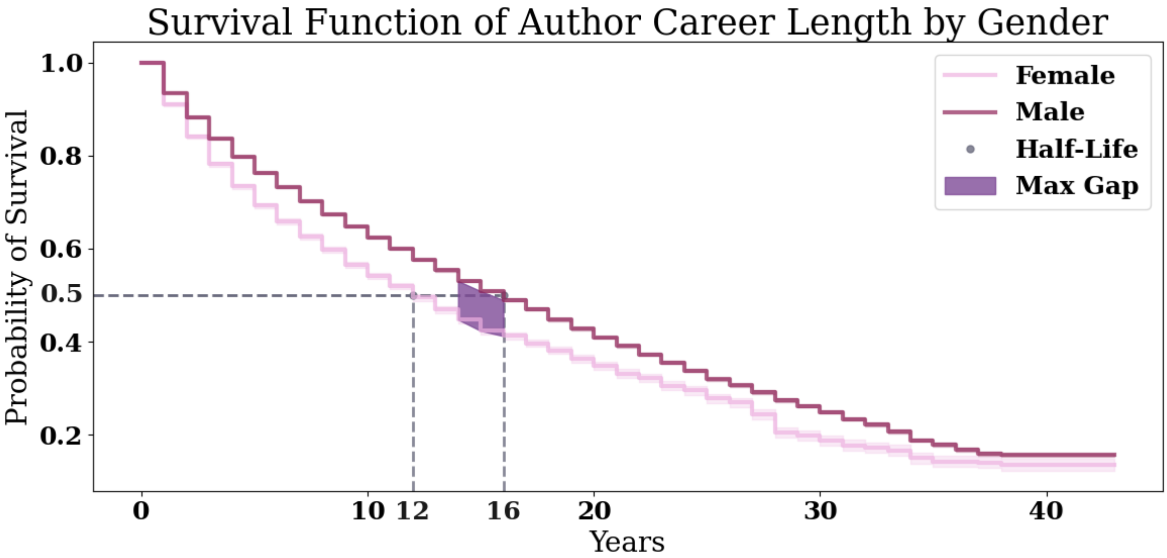
For *First Publication Year* and *Last Publication Year* we hypothesised that the mean for female authors is greater than for male authors but for the rest of the variables we hypothesised that the mean for female authors is less than for male authors and thus used a one-sided test. The significance threshold was set at .05.

From Table 15 we see that all activity and survival variables are statistically different between genders except for *Author Publication Rate*. Where *First Publication Year* and *Last Publication Year* are greater for female authors than for male authors, *Career Span Years*, *Drop Out*, and *Author Publication Count* are smaller. *Author Publication Rate* are equally distributed between female and male authors.

Figure 48 shows the Kaplan-Meier estimated survival rate.



(a)



(b)

Figure 48: The Kaplan-Meier estimated survival rate for **a** all given authors and **b** between genders. The estimator uses the drop out variable as probability estimator variable as a function of career span years when calculating the probability survival.

Looking at the overall survival rate displayed in Figure 48(a) we see that the probability of survival decreases with time as expected. After 15 years half of the authors have stopped publishing.

Similarly, the gender aggregated survival rate displayed in Figure 48(b) shows that female authors are less likely to "survive" across all points in time compared to their male colleges. This is especially true after 15 years where the gap between the two curves is largest. The half life for male authors is 16 years and for female authors 12 years.

Survival Rate per Topic and Domain

We proceeded to investigate survival rates between topics and domain. As this part of the analysis require not only that the auhtor is defined as active but also that the publication is assigned with a valid topic and/or domain the dataset included in this part of the analysis is reduced to 215,642 rows of 131,505 distinct publications and 65,393 distinct authors.

Figure 49-51 show the survival rate between distinct authors grouped by topic and domain. Each author is assigned a single topic or domain based on the mode of which they mostly publish within. The analysis in this section is conducted on Author Level.

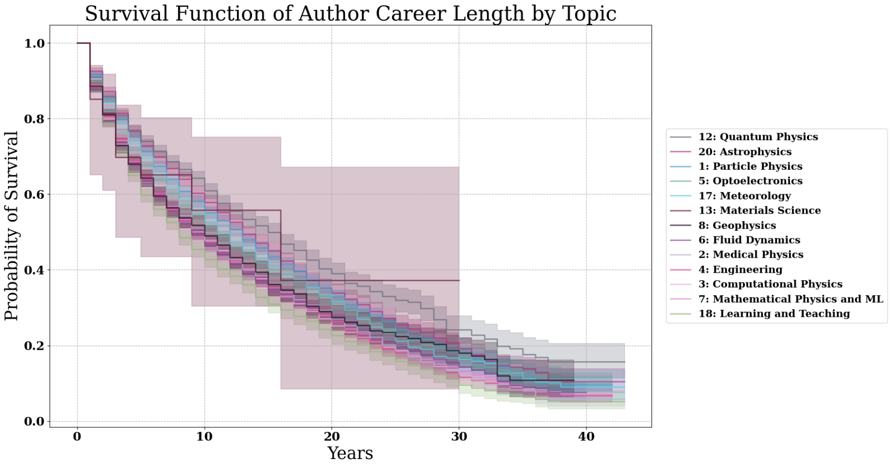


Figure 49: Survival rate for authors publishing within a given topic. Each author is assigned a single topic based on the most occurring topic for that given author (in case they publish within multiple topics). The legend is ordered in the same way as the survival curves (as much as possible) such that the first topic displayed in the legend has the highest probability of survival and the last topic displayed in the legend has the lowest probability of survival.

From figure 49 we see that the survival rate per topic is hard to distinguish. However, we see that the authors with highest probability of survival are those mostly publishing within Topic 12: "Quantum Physics" and the authors publishing within Topic 18: "Learning and Teaching" has the lowest probability of survival.

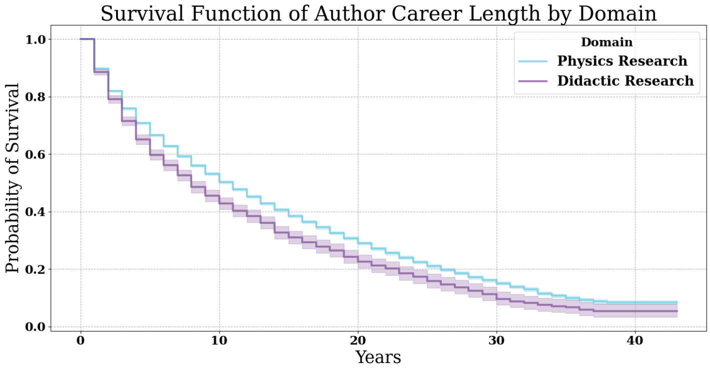


Figure 50: Survival rate for authors publishing within the given domain. Each author is assigned a single domain based on the most occurring topic for that given author (in case they publish within multiple topics).

Figure 50 confirms what we saw from the survival rates between topics, Figure 49, the the authors publishing within Didactic Research has the lowest probability of survival.

Figure 51 displays the survival rate of authors grouped by both domain and gender.

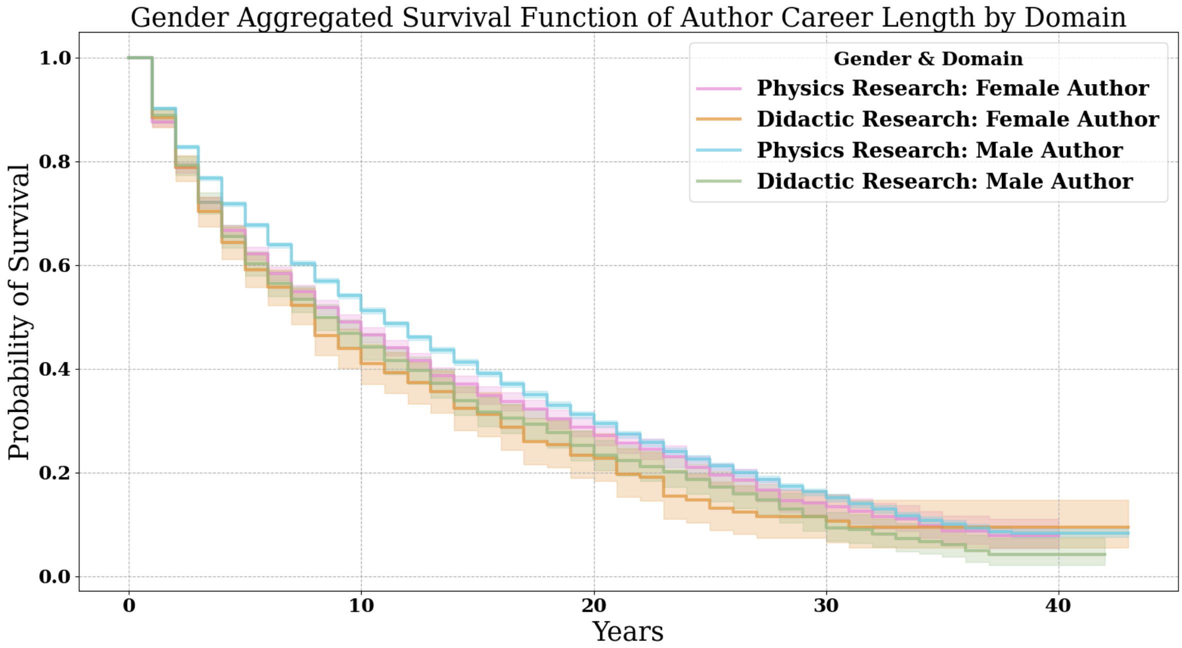


Figure 51: Survival rate for authors publishing within each domain type grouped by gender. Each author is assigned a single domain based on the most occurring topic for that given author (in case they publish within multiple topics).

From figure 51 we see that this trend holds when breaking down survival rate between both domain and gender. That is, female authors publishing within the Didactic Research domain are more likely to drop out. Generally, female authors are more likely to drop out compared between domains. However, female authors publishing within Physics Research are slightly more likely to survive than males publishing withing Didactic Research. While male authors publishing within Physics Research has the overall highest survival rate, we do see that after 15 years female Physics Research authors start approaching male Physics Research authors almost aligning with them around 23 years. The distance between the two groups again increase approaching 30 years and onwards but then decrease again approaching 40 years.

Table 16 presents the log-rank test results. We conduct the log-rank test as described in Section **Activity and Survival Rate** in order to investigate at which statistical level of significance the survival rates are different between genders and domains.

Table 16: Log-rank testing results. We tested whether the difference in probability of surviving was statistically significant between genders and domains as presented in Figures 48-51. If the test result were significant, meaning that variable 1 had a significantly smaller probability of surviving than variable 2, it is marked in green, and if it was insignificant, meaning the two variables had an equal chance at surviving, it is marked in red.

Survival Rate Log-Rank Testing Results			
Variable 1	**Variable 2**	**Test statistic, p-value**	**Figure**
Female Authors	Male Authors	609.23, $> .001$	48(b)
Didactic Research	Physics Research	65.05, $> .001$	50
Female Authors Didactic Research	Male Authors Didactic Research	0.66, .41	51
Female Authors Physics Research	Male Authors Physics Research	63.43, $> .001$	51

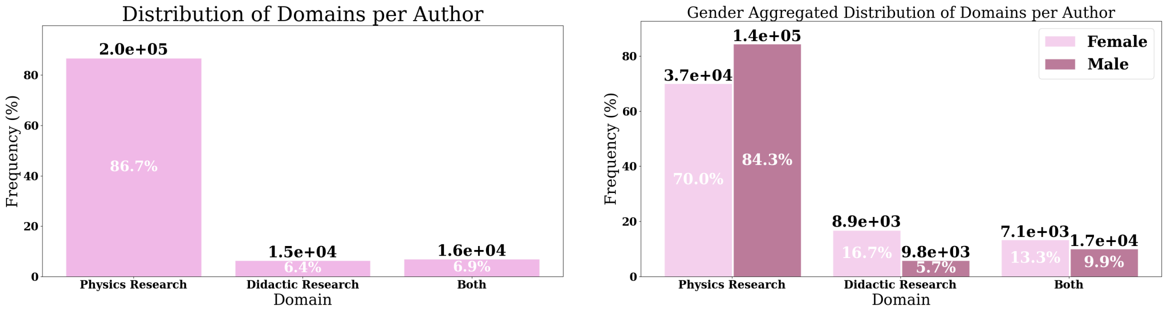
The results in Table 16 show that overall female authors have a statistically significant lower probability of surviving than male authors. Between domains, authors publishing within Didactic Research have a statistically significant lower probability of surviving than authors publishing within Physics Research. However, when grouped by both gender and domain, there is not statistically significant difference in survival probability between female and male authors publishing within didactic research but only for female and male authors publishing within Physics research.

In summation, female authors within the "Didactic Research" domain has the lowest probability of surviving followed by male authors also publishing within "Didactic Research". Male authors within the "Physics Research" has the highest probability of surviving followed by female authors publishing within that same domain. There is no statistically significant difference in survival probability between female and male authors publishing within Didactic Research but only for female and male authors publishing within Physics research.

Domain Transition

We continue using the same subset data only including authors defined as active in the period 1975-2018. We look into whether there is a tendency of transitioning between the domains throughout their career and whether this transitioning is different between female and male authors. The analysis in this section is conducted on Author Level.

Figure 52 shows the distribution of authors publishing solely within one domain or the other or both.



(a)

(b)

Figure 52: Distribution of **a** all authors and **b** gender aggregated authors publishing within one distinct domain or both throughout their career.

From Figure 52(a) we see that most authors publish within only "Phycics Research" (87%) while 6% publish within only "Didactic Research" and 7% publish within both domains through their career. Looking at the gender aggregated domain distribution per author in Figure 52(b) we find that a larger percentage of the male than female authors solely publish within "Physics Research" (84% vs 70%) where as a larger percentage of the female authors publish within "Didactic Research" and both domains compared to the male authors (17% vs 6% and 14% vs 10%, respectively).

Figure 53 displays the transition probability matrix for female and male authors, respectively. We assume that the first domain that an author publish within is the field within which they start researching and that the last publication they do in their career is the final publication that they do.

Figure 53 show that for both female and male authors most of those that start within "Physics Research" also stay within "Physics Research" both for female authors, Figure 53(a), and especially for male authors, Figure 53(b). 96% of female vs 98% of male authors starting publishing within physics end publishing within physics.

Furthermore, we specifically notice in Figure 53(a) that while 80% of female authors starting within "Didactic Research" finish within "Didactic Research" leaving 19% to transition from "Didactic Research" to "Physics Research". The last percentage transition to both. Compared to their male colleagues, Figure 53(b), only 65% of male authors starting publishing within "Didactic Research" finish in "Didactic Research" leaving 34% transitioning to "Physics Research" (and 1% to both).

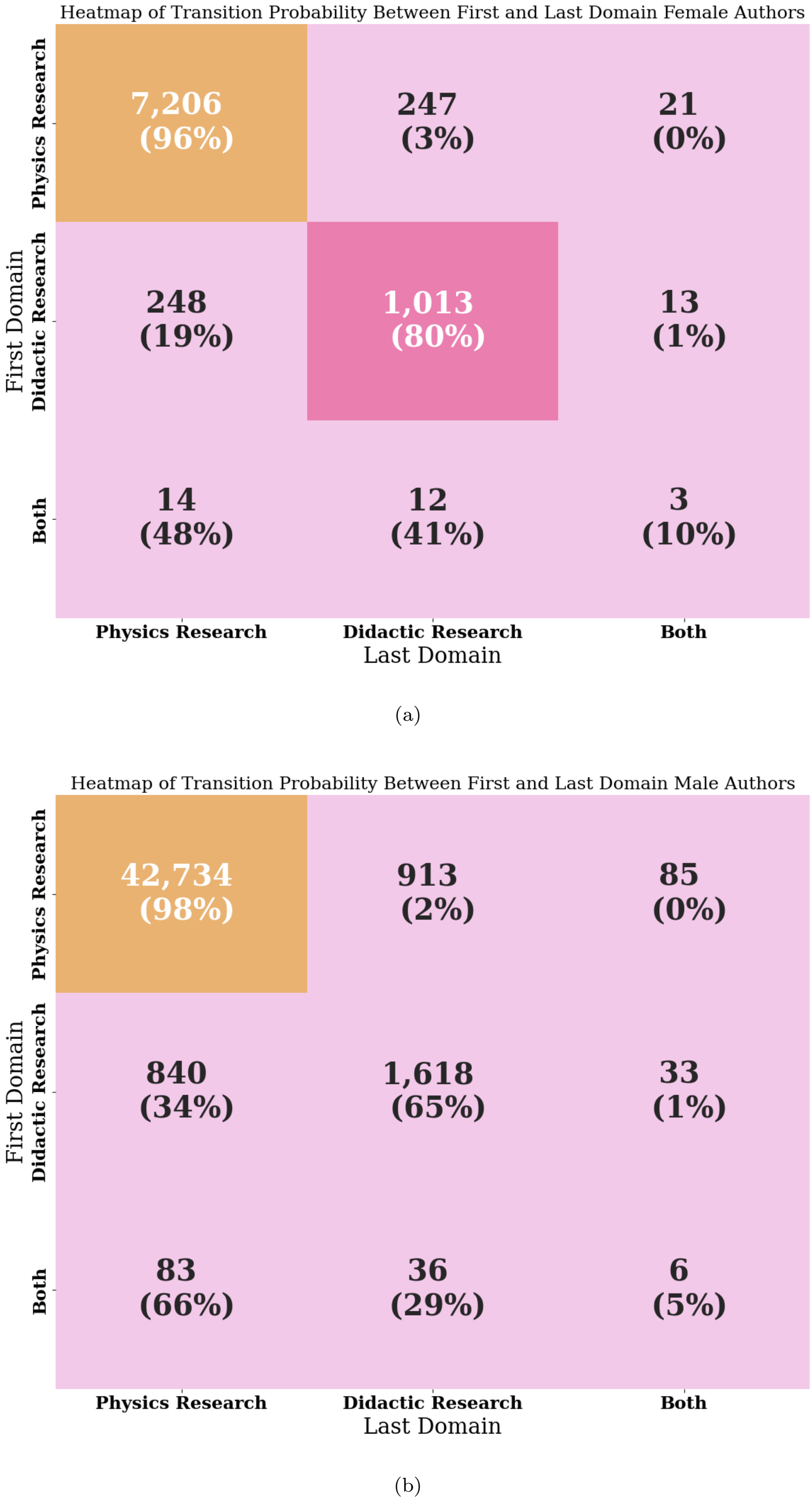


Figure 53: Transition Probability Matrix of **a** female authors and **b** male authors. The y column displays the domain that the author make their first publication within and the x column displays the domain within which the author make their latest publication. The count and percentage corresponds to the number of authors that ended within the given domain out of how many that started within the give domain.

Figure 54 shows a section of an interactive interactive Sankey Diagram displaying the transition between domains (including drop out) per author per year. Due to space constraints the PDF version only displays the first couple years. We look into the domain transitioning over time to investigate any chances between publication years.

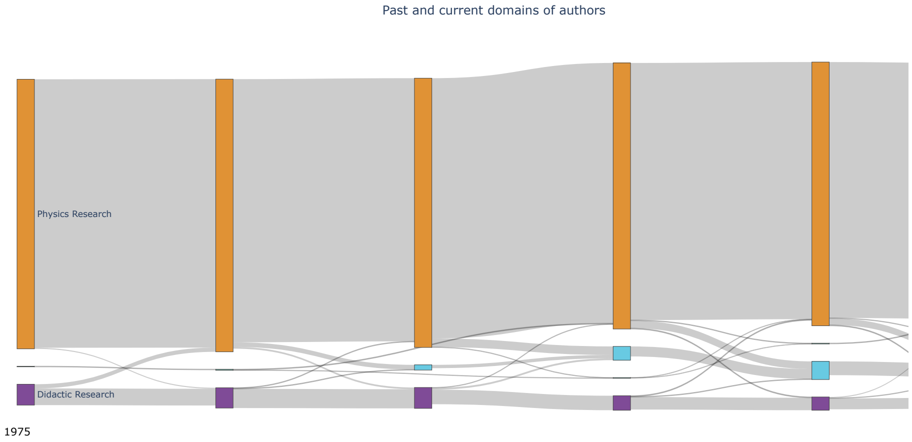
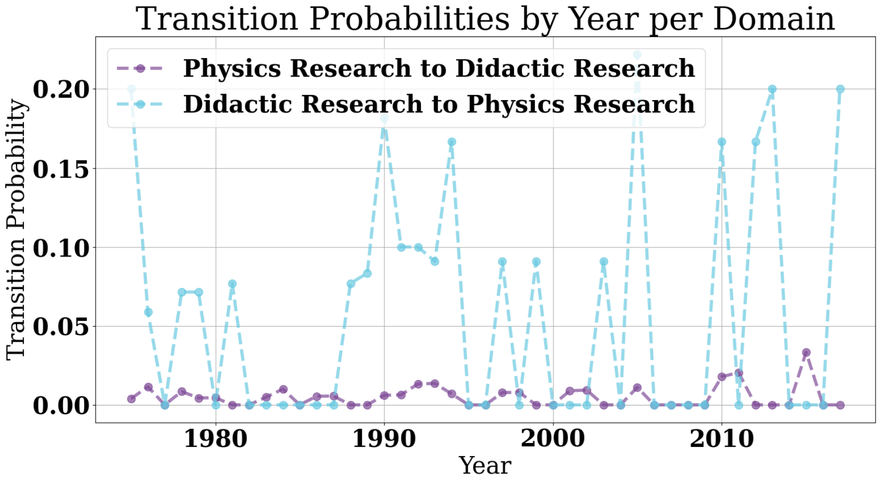


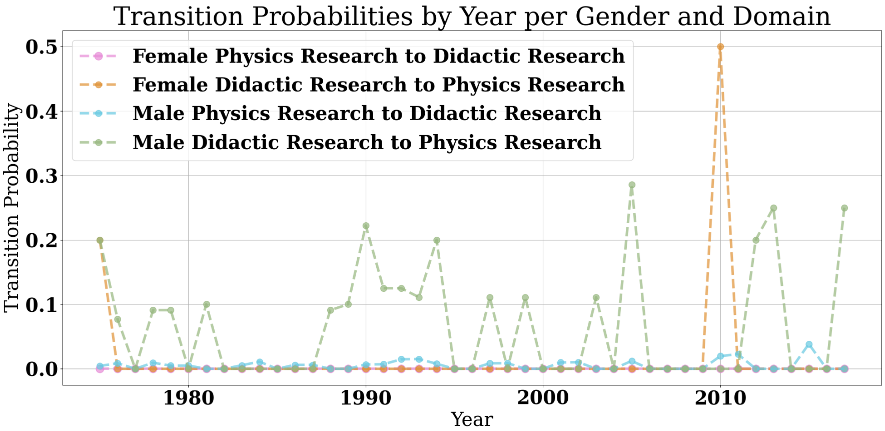
Figure 54: PDF view of the Sankey diagram depicting the transition flow of authors between each domain per year. Due to the large image size only the first couple years are shown. Each node represents a year and the bar size of each category the number of authors publishing within the given domain. The plot also includes the number of authors transitioning to publication stop (drop out). The interactive html plot can be viewed here.

From 54 we see that only a small proportion transition between the two domains per year. However, it is a bit hard to interpret and so we calculate the transition probability per year.

Figure 55 displays the transition probability of transitioning between the two domains.



(a)



(b)

Figure 55: Transition Probability per year for **a** all authors and, **b** per gender. The transition probability is calculated as the number of authors publishing within the given domain per year out of the total of numbers publishing within the given domain the year before.

From Figure 55(a) we see that the probability of transition from "Physics Research" to "Didactic Research" is approximately zero - that is true for all years. The probability of transitioning from "Didactic Research" to "Physics Research" is a little higher with the lowest probability at 0% but the highest at 20% over time. Grouped by gender, Figure 55(b), we see that female authors (almost) do not transition while male authors - if they do transition - only transition from Didactic Research to Phyiscs Research.

In summation, these results show that there is not really a tendency to transition between domains and if it does happen it is only for male authors transitioning from "Didactic Research" to "Physics Research" - that is, if it is not just due to mislabelling and uncertainty from the TM.

Clustering and Classification of Genders

Clustering and Classification are two kinds of Machine Learning techniques that we used to investigate the separability of female and male authors. All variables are converted to numerical features in order for the models to be able to interpret them. We also remove all NaN values. This leaves us with a dataset containing 612,138 rows entailing 318,847 unique authors and 224,191 unique publications for this part of the analysis. The analysis is conducted on Authorship Level.

For clustering and classification we include the following key features:

- *Publication Year*
- *Cited By Count*
- *Grants*
- *Journal ID*
- *Total Author Counts*
- *Author Position*
- *Institution ID*
- *Institution Type*
- *Author First Country*
- *Is Corresponding*
- *Topic Index*
- *Domain*
- *Author Mean Cites*
- *Institution Ranking*
- *Journal Ranking*
- *First Publication Year*
- *Career Span Years*
- *Author Publication Count*
- *Author Publication Rate*
- with *Binary Gender* as target variable

That is, each feature listed is used to predict the target variable *Binary Gender*.

UMAP Clustering

Table 17 shows the parameter settings chosen for the UMAP clustering based on sample data testing. In order to obtain optimal results from our UMAP clustering, we first tested several parameter settings on sample data of size 10,000 before running the final clustering analysis on the entire dataset. We determine the optimal hyperparameter settings one by one based on the visual clustering output. The plots for each hyperparameter settings test are shown in Figure A15 in Appendix A.

Table 17: Table presentation of the hyperparameter settings tested for UMAP clustering. The table displays the parameters tested, the parameter values tested as well as the best value (marked in green) based on the visual interpretation presented in Figure A15 in Appendix A. The final UMAP is conducted with the settings chosen as the best value.

Evaluation of UMAP Clustering Hyperparameter Settings			
	Parameter	**Parameter Values Tested**	**Best Value**
1	n_neighbors	20, 100, 200	200
2	min_dist	0.0, 0.5, 0.99	0.5
3	metric	"hamming", "jaccard", "dice", "russellrao", "kulsinski", "rogerstanimoto", "sokalmichener", "sokalsneath", "yule"	"hamming"

Figure A16 in Appendix A show three metric types as a function of the wall time for UMAP with the given setting. When testing the metric setting we found three equally good looking settings, namely "hamming", "jaccard", "dice, "russelrao", "kulsinski", and "yule" but since the UMAP task was very time consuming we tested which of the settings would optimise the computational time.

As can be seen from Figure A16 the metric setting that optimised the UMAP wall time was "hamming" why we proceeded with that. With our final hyperparamter settings obtained as shown in Table 17 we proceed to run the UMAP clustering technique on the full dataset.

We perform the clustering algorithm for three different cases:

1. **Unsupervised UMAP:** In this case we provide the full list of feature variables for the model. That means the model is given the entire dataset except the target value *Binary Gender*.
2. **Fully supervised UMAP:** In this case we not only provide the full list of feautre variables but also the target variable *Binary Gender*. That means the model is given the entire dataset and "knows" that it should be divided by the gender.
3. **Partially supervised UMAP:** In this case we split the data 50/50 into testing and training. First we provide both the feature variables and the target variable of the training data to create a supervised embedding. Lastly, we predict the feature variables of the testing data by fitting the supervised prediction. We then project the prediction onto the supervised embedding to test the level at which the prediction succeeded at clustering the genders.

Unsupervised UMAP

Figure 56 depicts the clustering obtained through unsupervised UMAP. The unsupervised UMAP performs a dimensionality reduction of our entire dataset without the target data, here *Binary Gender*, as an input. This gives us an idea of the distribution of clusters in it's original state without the model having any training.

The results of the unsupervised UMAP presented in Figure 56 show that the standard UMAP algorithm does not separate the genders according to their type.

Fully Supervised UMAP

Figure 57 shows the clustering result obtained through fully supervised UMAP. That is, in this case the model gets the target data and label of each entry in the dataset. This gives us an idea of the extent at which the model is able to separate the two groups, female and male authors, given that it is provided their label.

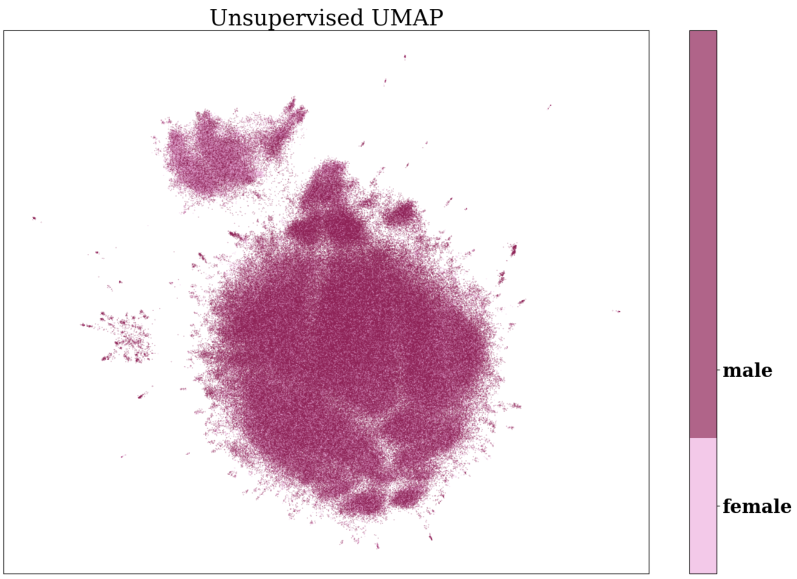


Figure 56: Depiction of the dimensionality reduction performed by UMAP on unsupervised data. That means the algorithm did get the target variable as input and therefor does not seperate the genders. The colorbar shows the color according to each gender label as well as the frequency of each label in the data.

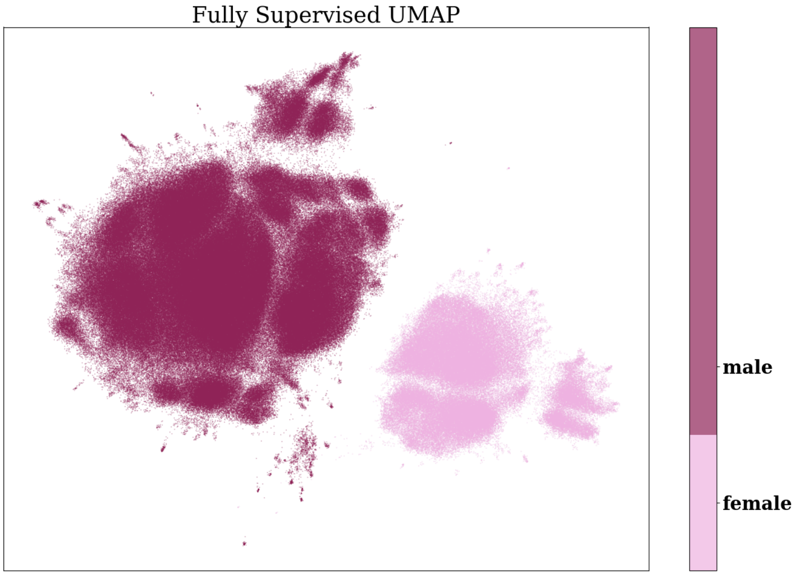


Figure 57: Depiction of the dimensionality reduction performed by UMAP on supervised data. That means the algorithm did get both the target and feature variables as input and therefor does succeeds at separating the genders. The colorbar shows the color according to each gender label as well as the frequency of each label in the data.

The results of the supervised UMAP presented in Figure 57 is a clearly separated set of genders. This is is expected since we gave the algorithm all the target value information. Besides we see some internal structures for each of the gender clusters, respectively. We also note that for each class we retrained a similar structure to that of the unsupervised case shown in Figure 56. Both groups consist of ”one” large blob with a smaller outlier ”blob”. Globally, the larger blobs are placed nearer each other and the small blobs further from each other but closer to their same class large blob. This indicates that the inter-relationships among each gender is also retained.

Partially supervised UMAP

Figure 58-59 displays the clustering results obtained by partially supervised UMAP. The partially supervised UMAP is made in order to test the algorithms ability for label prediction. We split the data into testing and training with a training size 0.5 leaving us with training data of shape (306069, 19), and testing data of identical shape (306069, 19). The training size is relatively large due to a large the imbalance (80-20) between our two groups, female and male authors.

The partially supervised is done in two steps: first creating the supervised embedding with our training feature and target data. Then, the same supervised embedding is fitted to the test feature data in order to predict the labels.

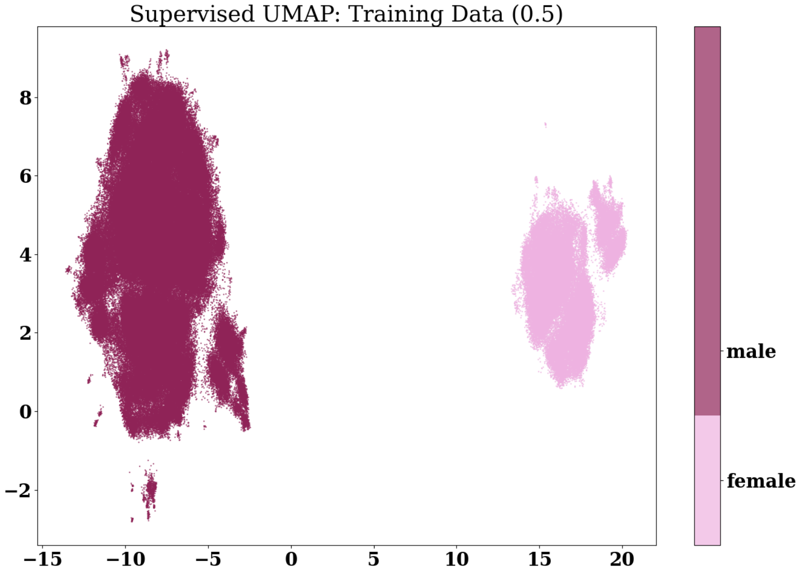


Figure 58: Depiction of the supervised embedding performed by UMAP on the training data. That means the algorithm did get both the target and feature variables as input and therefore does succeeds at seperating the genders. The supervised embedding made on training data will be used for predicting the target variable of the test data. The colorbar shows the color according to each gender label as well as the frequency of each label in the data.

The results of the supervised UMAP training embedding presented in Figure 58 is a clearly (aside a little noise) seperated set of genders similar to what we saw in the fully supervised setup displayed in Figure 57. We again see the same structure of the clusters consisiting of one large and one smaller blob each. The shape is a little different from the shape we saw in the unsupervised and supervised case, as seen Figure 56, and 57.

The partially supervised UMAP, Figure 58, is not as round. Furthermore, the smaller blobs between the two classes are not placed globally in the same place as we saw in the unsupervised and fully supervised case. This might indicate the algorithm didn't retain the inter-relationship. However, the best way to find out is by fitting the supervised UMAP training to the unsupervised testing data and se how well it predicts the gender.

Figure 59 shows the predicted UMAP results.

From Figure 59 we see a very noisy separation. While there is trajectories of the two classes similar to those we saw in the supervised case in Figure 58 the shape is definitely not as sharply outlined. Furthermore, we see that the female and male data points are mixed between both clusters and not clearly seperated as in Figure 58. To better compare the success at separating the two groups in the prediction we plot the supervised result, Figure 58 and the predicted result Figure 59 onto each other.

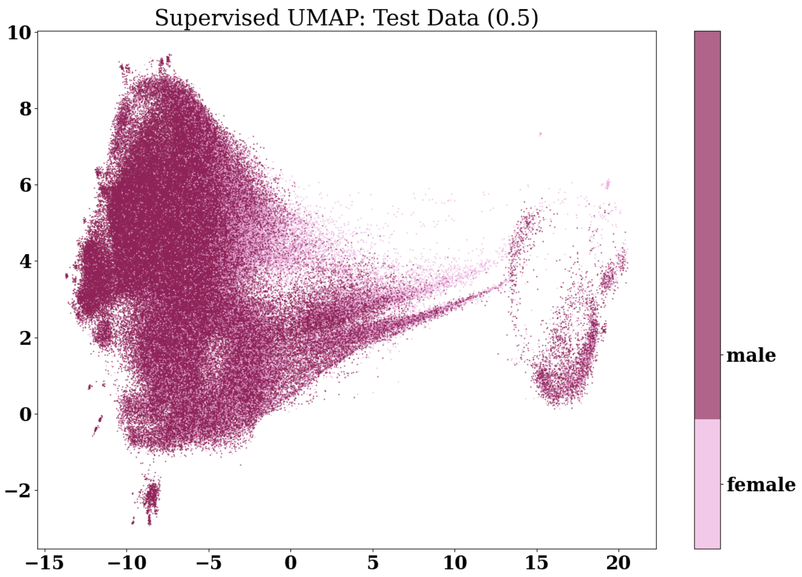


Figure 59: Depiction of the predicted clustering performed by UMAP by fitting the testing data to the supervised training data embedding as seen in Figure 58. That means the algorithm did get both the target and feature variables as input in the first place and then based on that tries to make a similar seperation of the testing data while only getting the feature variables. The colorbar shows the color according to each gender label as well as the frequency of each label in the data.

Figure 60 shows a projection of the predicted testing data onto the supervised training data in a 2D histogram.

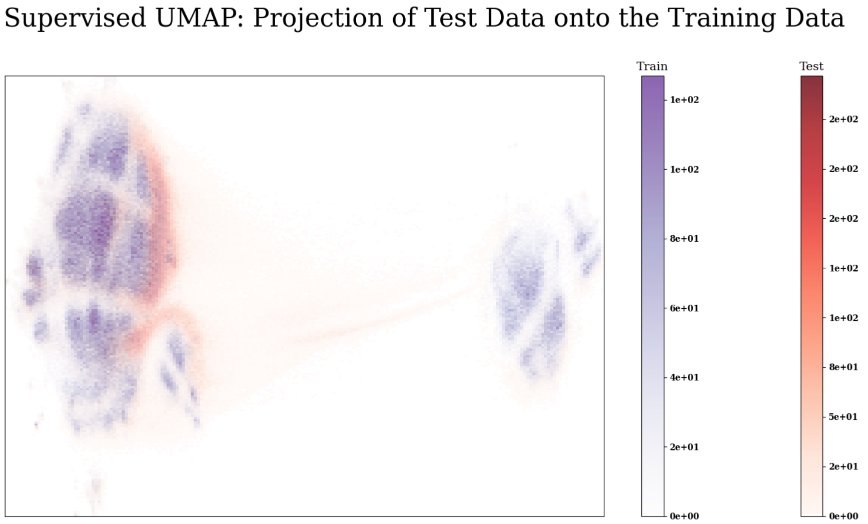
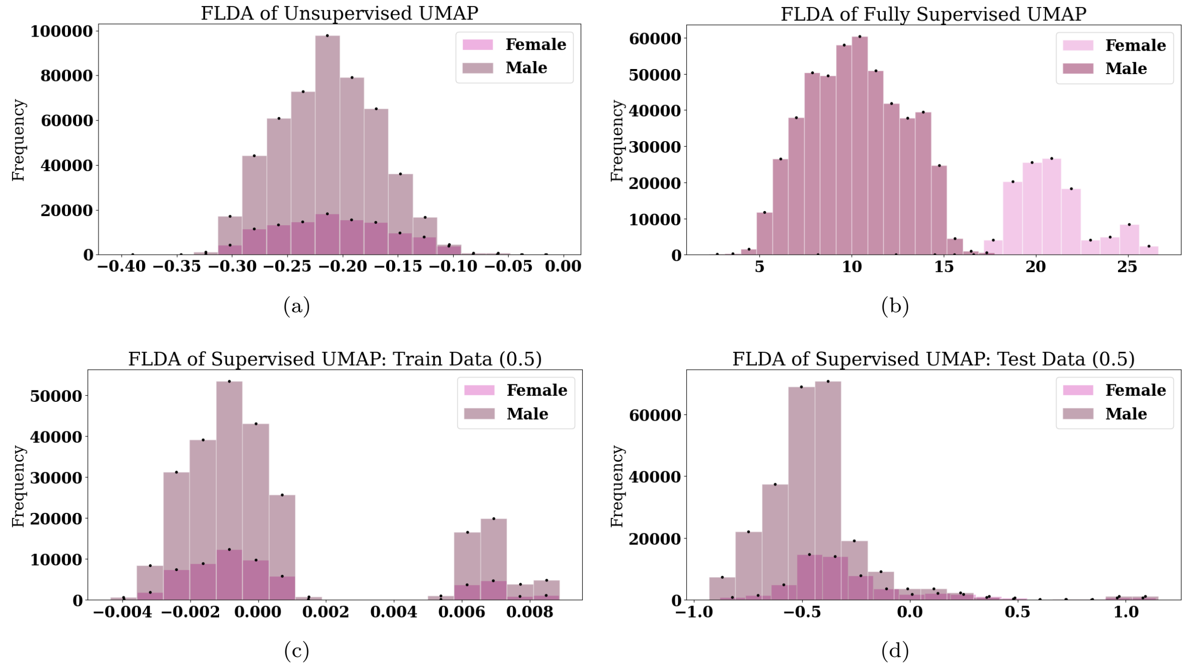


Figure 60: Projection of the predicted testing data clustering from the supervised training data clustering in a 2D histogram. The training data is displayed in blue colors whereas the testing data is shown in reds.

The projection of the testing data onto the training data presented in Figure 60 show that the UMAP prediction of the testing data did not clearly create two distinct classes following the supervised training data. The result marks the rim of the clusters but the separation is very noisy and very few data points of the testing data fall into the cluster on the right representing female authors. In order to further investigate the clusters we use Fisher Linear Discriminant dimensionality reduction to obtain the data in one dimension at the highest separation rate.

Fisher Linear Discriminant Analysis: UMAP

Figure 61 shows results obtained by applying FLDA between the two gender classes on the unsupervised, fully supervised, and partially supervised clustering, respectively. To further analysis our UMAP results we perform a FLDA, as described in Section **Fishers Linear Discriminant Analysis** to view them in a 1 dimensional space.



(a)

(b)

(c)

(d)

Figure 61: The **a** unsupervised UMAP, **b** fully supervised UMAP, **c** supervised training data UMAP, and **d** predicted testing data UMAP, reduced to 1 dimension by employing FLDA between the two binary gender classes.

The separation of the unsupervised UMAP obtained by FLDA as presented in Figure 61(a) confirms that that the standard UMAP algorithm does not separate the genders according to their type as we saw in the unsupervised UMAP depiction Figure 56 as well.

Similarly, Figure 61(b) displaying the supervised FLDA seperation confirms what we saw at the supervised UMAP Figure 56, namely that the genders are well separated. Figure 61(b) also displays the two distributions within each group where we see two peaks in the histogram for each gender.

Lastly, the supervised training data UMAP FLDA dimensionality reduction in Figure 61(c) reveals that while we do see two groups of similar shape to the distributions of male and female authors in the fully supervised case, Figure 61(b), we also see that the male and female authors are mixed within the two groups.

The prediction of the classes based on the testing and training data reduced to one dimension by FLDA is displayed in Figure 61(d). The results confirm what we already saw in the UMAP results in Figure 59, namely that the two classes are not clearly separated and that the shape of the clusters, as seen in Figures 61(a), 61(b) and 61(c), is no longer entailed.

The Fisher weights obtained were $-2.0e-02, 8.6e-05$ for the unsupervised UMAP, $1.2, -0.6$ for the supervsied UMAP, $3.3e-04, 4.2e-04$ for the training data of the partially supervised UMAP, and $6.1e-02, -2.7e-02$ for the testing data of the partially supervised UMAP.

LightGBM Classification

Table 18 shows the F1-Score obtained by testing different hyperparametersettings. Figure A17 shows the F1-Score as a Function of Parameter Value for Each Hyperparameter. The optimal hyperparameter setting was found by testing a range of parameter values for one parameter while keeping the others constant using the F1-score as a performance evaluation metric as described in Section **F1 Score**.

Table 18: Table overview of the highest F1-score obtained from testing each parameter at different settings. We tested one parameter at a time keeping the previous at the optimal found setting and the ones that had not yet been tested on default setting. Each value that was tested for each distinct parameter is displayed in Figure A17 in Appendix A. Remembering that the F1-Score should be maximised we found the settings displayed in this table to be the optimal settings with maximised F1-score 0.833863 as marked in green.

F1-Score Evaluation of LightGBM Classification Model Hyperparameters

	Parameter	**Parameter Value**	**F1-Score**
1	num_leaves	150	0.746204
2	learning_rate	0.5	0.798022
3	subsample_for_bin	100000	0.798471
4	min_child_samples	50	0.798471
5	colsample_bytree	1	0.799658
6	max_depth	50	0.799658
7	n_estimators	200	0.833863

Figure 62 display the correlation matrix between the feature variables used in the model. We checked if any of the variables were highly correlated.

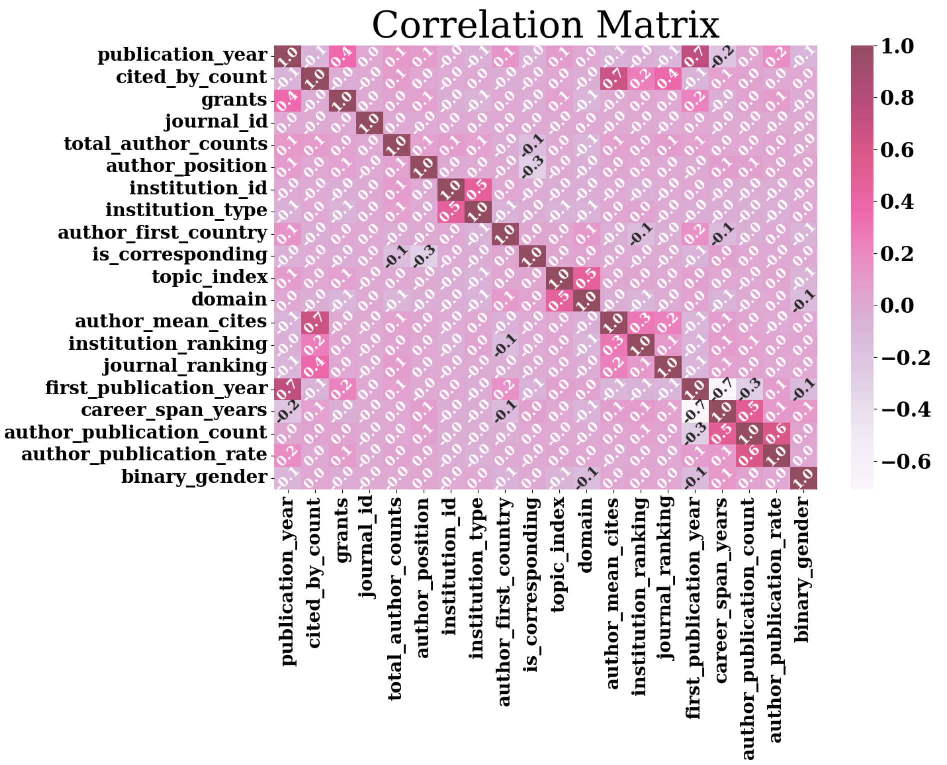


Figure 62: Heatmap depicting the level of correlations between each feature used in the classification model. If the variables are not correlated at all the matrix output 0 whereas if they are completed correlated the output is ±1 depending on the direction of the correlation.

Figure 62 show that only the variables *First Publication Year*, *Career Span Years* and *Publication Year* were highly correlated (which is expected) and with our parameter settings obtained we proceeded to run the LightGBM classification model.

LightGBM I

We split the data into testing and training with a training size 0.3 leaving us with training data of shape (183,641, 19) and testing data of shape (428,497, 19)

Figure 63 shows the resulting evaluation metrics scores obtained for the first run of our LightGBM. We use both to evaluate the performance of the classifier model.

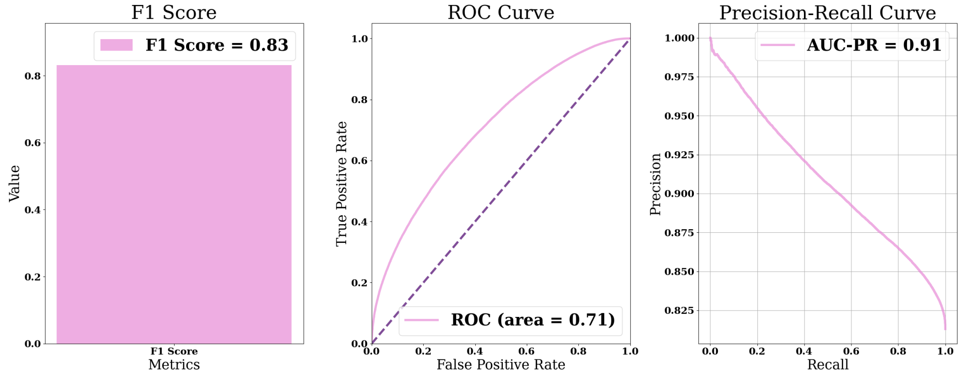


Figure 63: Depiction of the F1 Score, ROC Curve area, and Precision-Recall Curve which are the evaluation metrics we use to evaluate the performance of the classifier model. Each respective metric score is displayed in the legend of each figure.

Figure 64 shows the confusion matrix obtained for the first run of our LightGBM.

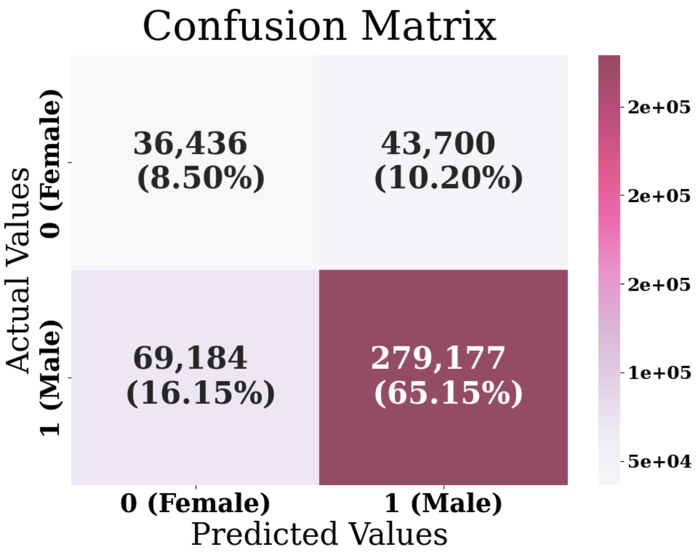


Figure 64: Confusion Matrix. The upper left corner represents the number of TN's, here female authors predicted as females. The lower left corner indicates the number of FN's, here male authors predicted as female. The upper right corner shows the FP's, here the number of female authors predicted as males. The lower right corner shows the TP's, here the male authors predicted as male. The Confusion Matrix displays both the absolute values as well as the percentage (%).

From Figure 63 we see the resulting evaluation metrics score obtained through the first run of our LightGBM model:

- **F1 Score: 0.83** which is considered a good score indicating that the model is good at recognising positive cases while minimising false positives and negatives.
- **ROC Curve Area: 0.71** indicating that the two gender classes are not completely separated but also do not complete overlap.
- **AUC-PR: 0.91** which is a relatively high value indicating that the model somewhat is able to distinguish between the two gender classes.

To further evaluate the model performance we investigate the number and ratio of TP's, TN's, FP's and FN's obtained. From the confusion matrix we obtained the following results:

- **True Positives: 278,246 (65.15%)** predicted male labels that are actually male
- **True Negatives: 36,747 (8.5%)** predicted female labels that are actually female
- **False Positives: 43,489 (10.20%)** predicted male labels that are actually female
- **False Negatives: 70,115 (16.15%)** predicted female labels that are actually male

Figure 65 show the Feature Importance indicating which of the features are most influential for the model to distinguish between the two classes. The feature importance specifically for the split and gain in the decision tree production is shown in Figure A18 in Appendix A.

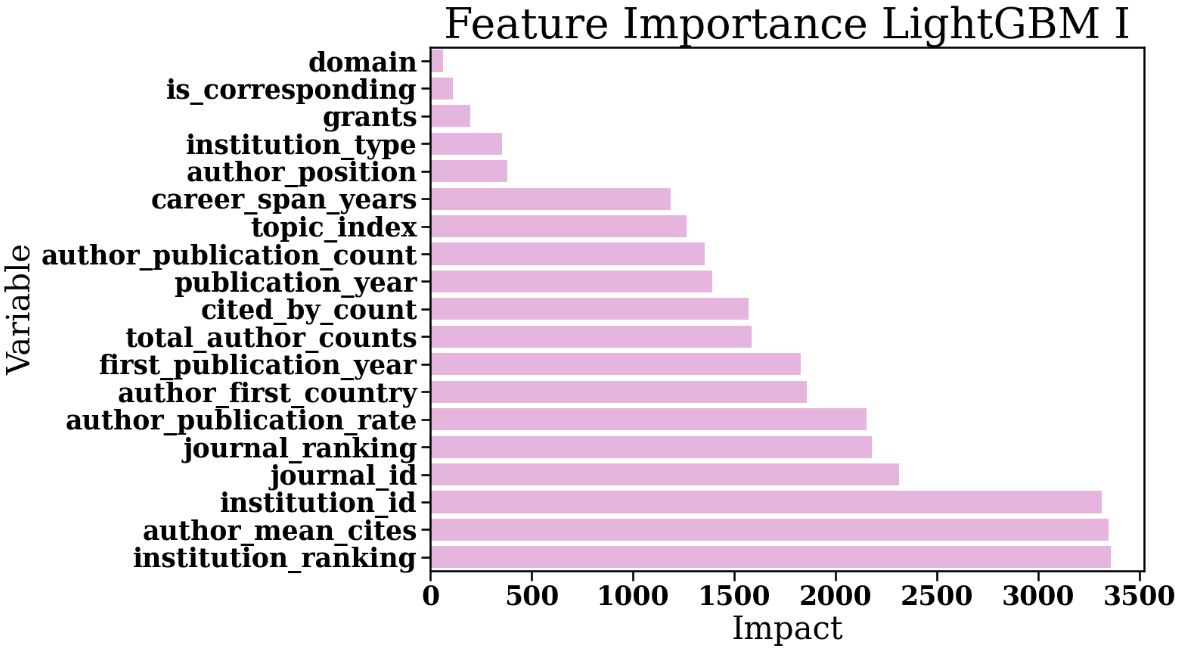


Figure 65: Feature Importance depicting the level of impact each feature has on the models ability to distinguish between the two gender classes. The features appear in an increasing order of importance such that the least influential are listed at the top and the most influential in the bottom of the y-axis. The impact size can be read from the x-axis.

Inspecting Figure 65 we see that the most important features for differentiating and thereby classifying female and male authors are:

1. *Institution ID*
2. *Author Mean Cites*
3. *Institution Ranking*
4. *Journal ID*
5. *Author Publication Rate*
6. *Journal Ranking*

In order to test how important these features really are for classifying male and female authors we try and remove them from the analysis and do a second run of the LightGBM classification model.

LightGBM II

As we drop the 6 most important features obtained from LightGBM I found in Figure 65 we included the following features for LightGBM II:

- *Publication Year*
- *Cited By Count*
- *Grants*
- *Total Author Counts*
- *Author Position*
- *Institution Type*
- *Author First Country*
- *Is Corresponding*
- *Topic Index*
- *Domain*
- *First Publication Year*
- *Career Span Years*
- *Author Publication Count*
- with *Binary Gender* as target variable

Figure 66 show the resulting evaluation metrics obtained for the second run of our LightGBM model in comparison to the first run.

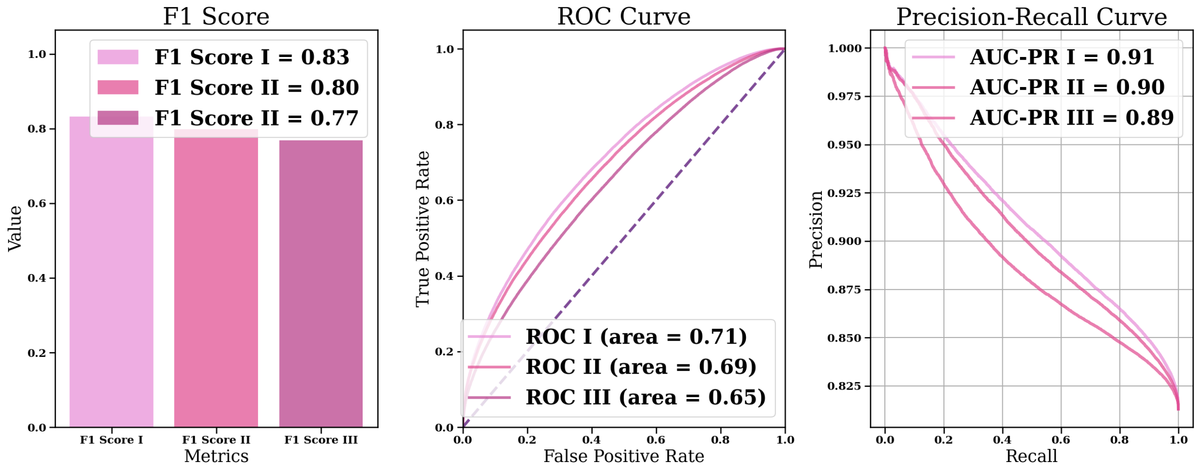


Figure 66: Depiction of the F1 Score, ROC Curve area, and Precision-Recall Curve for LightGBM I and LightGBM II. Where LightGBM includes all the key features in our dataset, LightGBM II does not include the six most important features found by LightGBM I and displayed in Figure 65. Each respective metric score is displayed in the legend of each figure.

From Figure 66 we see the resulting evaluation metrics score obtained through the second run of our LightGBM model in comparison to the first run. Computing the difference in each score between LightGBM I and LightGBM II we get:

- **F1 Score II: 0.80** which is considered an average score with Δ **F1-Score II-I = -0.03** which indicates the model has gotten somewhat worse at recognising positive cases while minimising false positives and negatives.
- **ROC Curve Area: 0.69** with Δ **ROC area II-I = -0.02** indicating that the two gender classes are a little less separated now but still do not completely overlap.
- **AUC-PR: 0.90** with Δ **AUC-PR II-I = -0.01** which is still a relative high value however the model is now less able to distinguish between the two gender classes.

Figure 67 show the resulting confusion matrix obtained at the second run of our LightGBM.

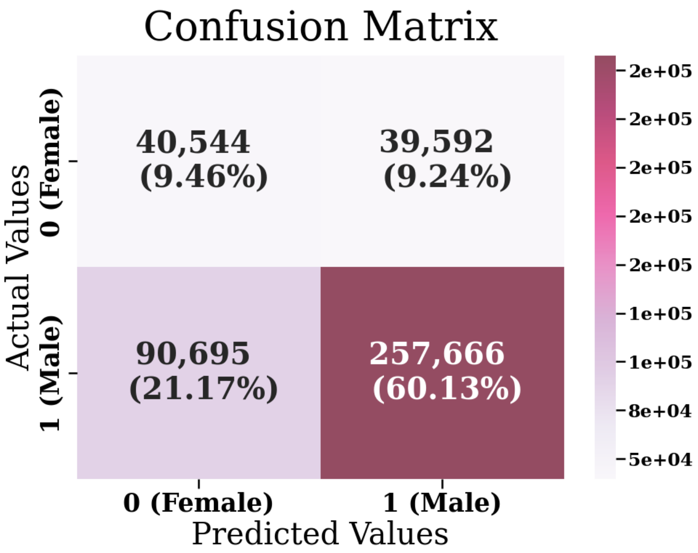


Figure 67: Confusion Matrix for the second run of our LightGBM. See explanation of how to read the figure in Figure 64.

From the confusion matrix presented in Figure 67 we obtain the following results:

- **True Positives: 257,666 (60.13%)** predicted male labels that are actually male
- **True Negatives: 40,544 (9.46%)** predicted female labels that are actually female
- **False Positives: 39,592 (9.24%)** predicted male labels that are actually female
- **False Negatives: 90,695 (21.17%)** predicted female labels that are actually male

Compared to the confusion matrix obtained in the first run of our model and presented in Figure 64 we see a decrease in the number of TP's and FP's but an increase in the number of TN's and FN's.

Figure 68 show the Feature Importance obtained from our second run indicating which of the features are most influential for the model to distinguish between the two classes. The feature importance specifically for the split and gain in the decision tree production is shown in Figure A19 in Appendix A.

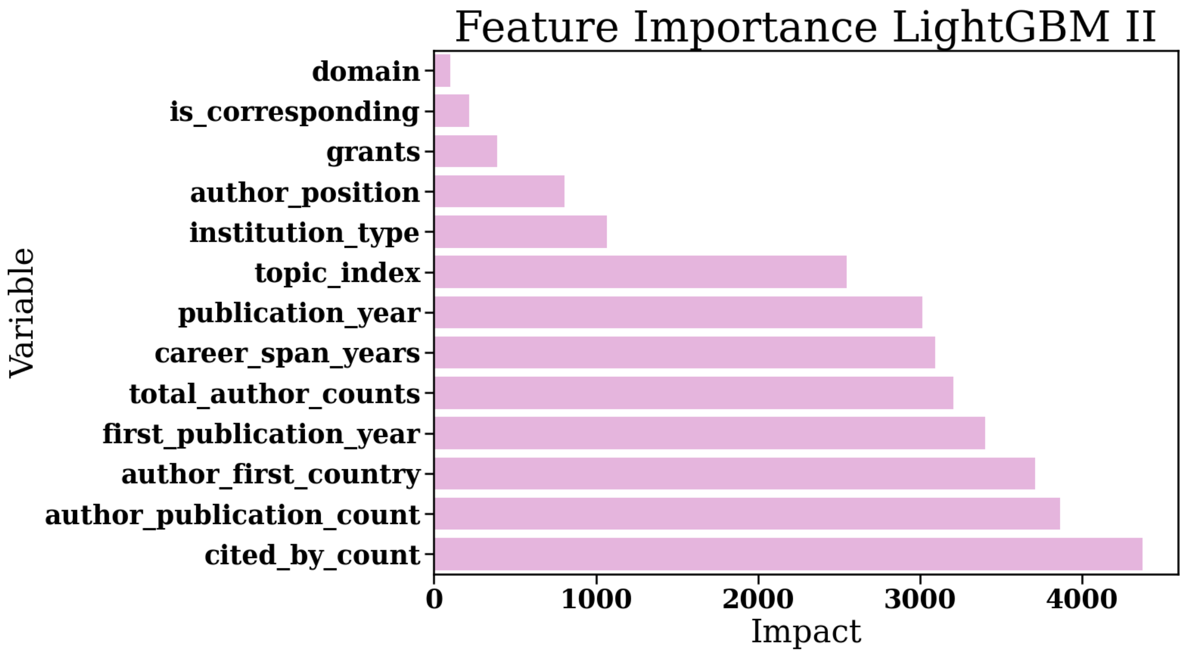


Figure 68: Feature Importance of LightGBM II depicting the level of impact each feature has on the models ability to distinguish between the two gender classes. The features appear in an increasing order of importance such that the least influential are listed at the top and the most influential in the bottom of the y-axis. The impact size can be read from the x-axis.

Inspecting Figure 65 we see that the most important features for differentiating and thereby classifying female and male authors are:

1. *Cited By Count*
2. *Author Publication Count*
3. *Author First Country*

As LightGBM II has shown only a very small decrease in the performance level of the model when dropping the 6 first important features from LightGBM I we proceed to further drop the next 2 most important features displayed in Figure 65. Therefore, we drop *First Publication Year* and *Author First Country*, and run the model for a third time without including these variables.

LightGBM III

As we drop the 8 most important features obtained from "LightGBM I" found in Figure 65 we include the following features for "LightGBM III":

- *Cited By Count*
- *Grants*
- *Total Author Counts*
- *Author Position*
- *Institution Type*
- *Is Corresponding*
- *Topic Index*
- *Domain*
- *First Publication Year*
- *Career Span Years*
- *Author Publication Count*
- with *Binary Gender* as target variable

Figure 69 shows the resulting evaluation metrics obtained for the third run of our LightGBM model in comparison to the first and second run.

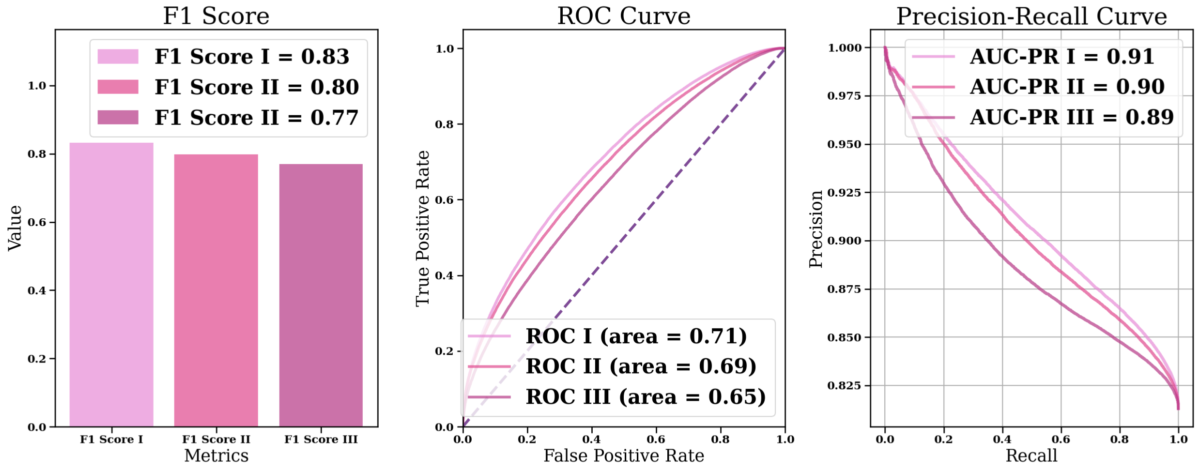


Figure 69: Depiction of the F1 Score, ROC Curve area, and Precision-Recall Curve for LightGBM I, II and LightGBM III. Where "LightGBM I" includes all the key features in our dataset, LightGBM III does not include the eight most important features found from LightGBM I and displayed in Figure 65. Each respective metric score is displayed in the legend of each figure.

From Figure 69 we see the resulting evaluation metrics score obtained through the third run of our LightGBM model in comparison to the first and second run. Computing the difference in each score between LightGBM I and LightGBM III we get:

- **F1 Score III: 0.77** which is considered an average score with Δ **F1-Score III-I = -0.06** which indicates the model has gotten worse at recognising positive cases while minimising false positives and negatives.
- **ROC Curve Area: 0.65** with Δ **ROC area III-I = -0.05** indicating that the two gender classes are way less separated now.
- **AUC-PR: 0.89** which is an average score with Δ **AUC-PR III-I = -0.02** and the model is now even less able to distinguish between the two gender classes.

Figure 70 show the resulting confusion matrix obtained at the third run of our LightGBM.

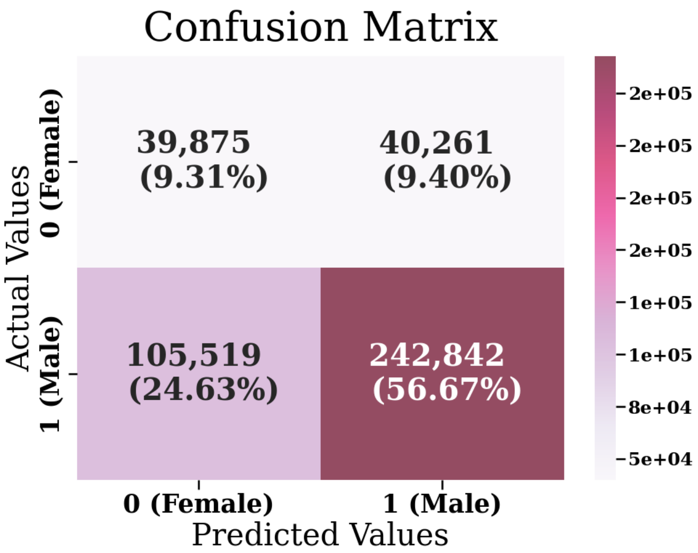


Figure 70: Confusion Matrix for the third run of our LightGBM. See explanation of how to read the figure in Figure 64.

From the confusion matrix presented in Figure 70 we obtain the following results:

- **True Positives: 242,842 (56.67%)** predicted male labels that are actually male
- **True Negatives: 39,875 (9.31%)** predicted female labels that are actually female
- **False Positives: 40,261 (9.40%)** predicted male labels that are actually female
- **False Negatives: 105,519 (24.63%)** predicted female labels that are actually male

Compared to the confusion matrix obtained in the first run of our model and presented in Figure 64 we see a decrease in the number of TP's and FP's but an increase in the number of TN's and FN's.

Figure 71 shows the Feature Importance obtained from our third run of the LightGBM model indicating which of the features are most influential for the model to distinguish between the two classes. The feature importance specifically for the split and gain in the decision tree production is shown in Figure A20 in Appendix A.

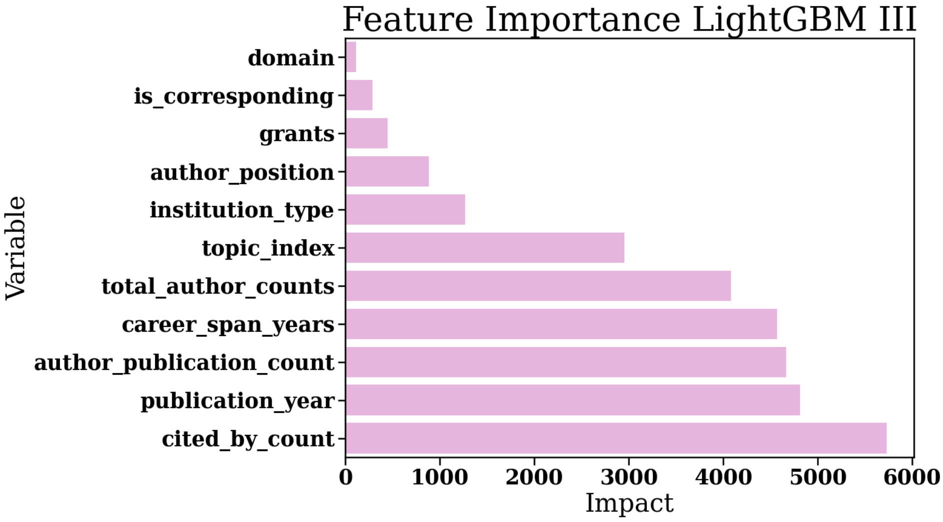


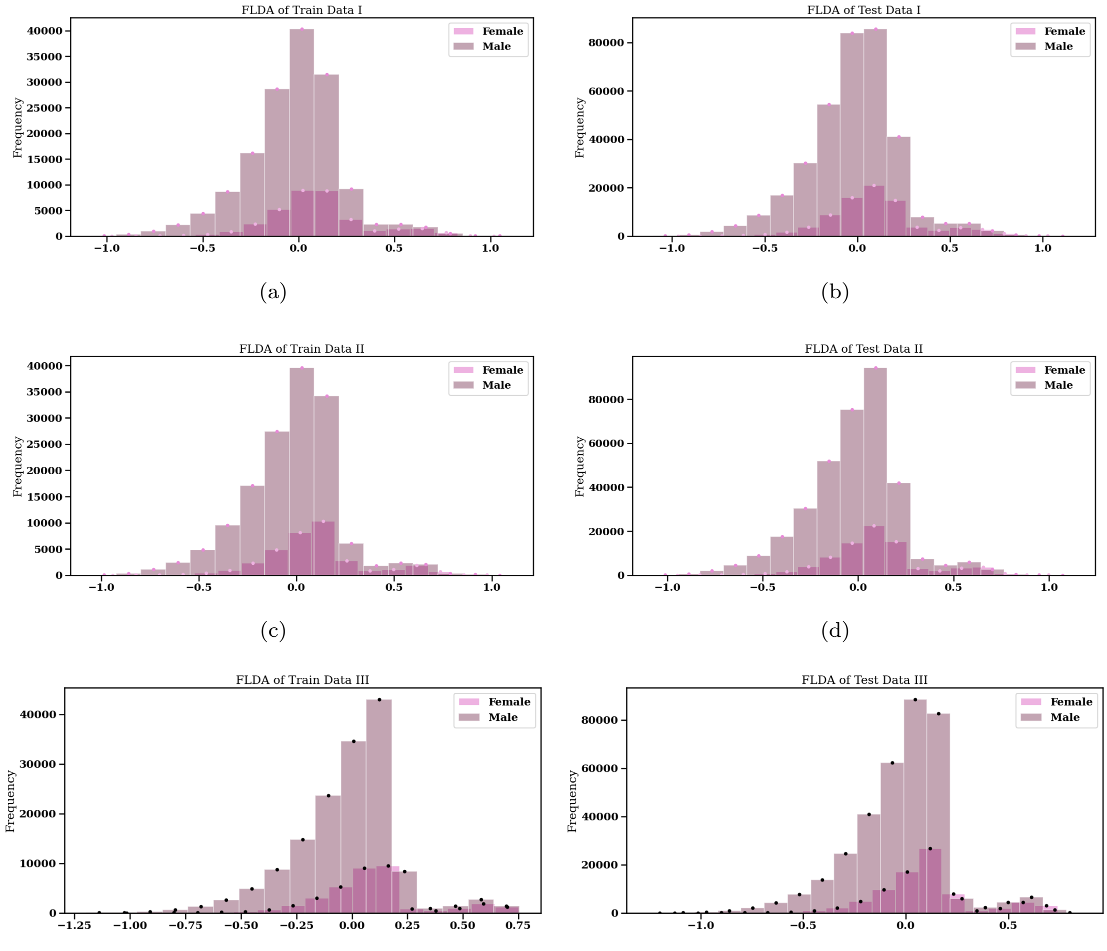
Figure 71: Feature Importance of LightGBM II depicting the level of impact each feature has on the models ability to distinguish between the two gender classes. The features appear in an increasing order of importance such that the least influential are listed at the top and the most influential in the bottom of the y-axis. The impact size can be read from the x-axis.

Inspecting Figure 71 we see that the most important features for differentiating and thereby classifying female and male authors are:

1. *Cited By Count*
2. *Publication Year*
3. *Author Publication Count*
4. *Career Span Years*

Fisher Linear Discriminant Analysis: LightGBM

Figure 72 displays the 1 dimensional distribution of males and females classified by LigthGBM I-III. The dimensionality reduction to one dimension is obtained by applying FLDA.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 72: 1D distribution of females and males in our **a,b** first LightGBM classification, **c,d** second LightGBM classification, and **e,f** third LightGBM classification. The figures in the left column displays the distribution in the training data whereas the figures in the right column displays the predicted distribution of the test data.

Investigating the distribution of male and female authors in our three different LightGBM setup we see a clear visualisation of what we saw from our three obtained confusion matrices displayed in Figure 64, 67 and 70; that the number of females predicted in the test data (Figure 72(b), 72(d), 72(f)) increase. However, when comparing to the distribution of females and males in the training data (Figure 72(a), 72(c), 72(e)) we also see that the number of predicted males decrease. These are especially the males distributed in closer proximity to the females i.e. on the right hand side tail of the male distribution. Again, this is also what we saw from the confusion metrics; namely that the level of TN's increase but at the cost of the level of TP's decreasing.

The Fisher weights obtained for LightGBM I were $4.0e-02, 6.0e-03$ for the training data and $2.3e-02, -6.0e-03$ for the testing data. The Fisher weights obtained for LightGBM II were $4.3e-02, 1.9e-03$ for the training data, and $2.6e-02, -7.5e-03$ for the testing data. The Fisher weights obtained for LightGBM III were $8.6e-02, 1.7e-03$ for the training data, and $8.9e-02, -7.4e-03$ for the testing data.

In summation, we find that we were able to cluster and classify the genders to some extent but far from with certainty.

Discussion

Summary

In this study, we investigated the disparities in academic careers between female and male authors from 1970 to the present day. Our aim was to move beyond simplistic conclusions of bias, which are abundant in existing literature, and instead, to characterise the behaviour and underlying factors contributing to these persistent biases in the field. We did so by analysing a dataset of 1,432,907 authorships and several subsets of our data. We employed different statistical methods as well as machine learning techniques. We found statistically significant disparities between female and male authors specifically in terms of prestige, activity and survival and representation within topics. However, we did not succeed at completely separating the two gender classes based on these disparities by clustering nor classification techniques even though we did see some level of distinguishability between the two genders.

Interpretations

This study includes several sections to interpret; the overall key variable distributions between genders, the gender relation to topic distribution, prestige markers, and activity and survival rate, and lastly clustering and classification of the genders.

Initial Variable Distributions

Interpreting the results found in Section Initial Variables Distributions Between Genders we find an indication of overall disparities between female and male authors(hips). First of, the overall distribution of female and male authors, presented in Figure 12, indicates that the workspace of physics is definitely not gender neutral. We saw that the distribution was somewhat similar worldwide, Figure 15, though North Amercia and South America had the smallest proportion of female authorships (17%) and Asia had the largest (24%). This indicates a global gender related barrier for women in physics, but also that there might be cultural differences influencing the level of these barriers in terms of access and representation within physics. While the unbalanced distribution in itself does not reveal the underlying bias or barriers for female authors to enter the field, but set in the context of our finding that most of the main variables are significantly smaller for female authors than for male authors, indicate that there is a systematic favouring of male authors within the field. This is true when it comes to publications per publication year, number of citations, author count per author position, and institution type with number of citations being the largest disparity (t(3.5e+05)=-16.04, p=< .001). This indicates a systematic ignoring of female authors in physics.

The most similar variables were grants received (t(3.1e+05)=11.68, p=1.0), also when accounting for only the last author (t(2.1e+05)=-0.44, p=0.33), and author count per author position (t(4)=-2.70, p=0.03 and U=0.0, n1=n2=3, p=0.05). While it is positive that there is no indication of females authors being granted less than male authors, there is a lot of uncertainty on this variable, as it is often not listed in the dataset, whether a grant was received or not. The author count per author position is on the borderline between signficant/insignificant but the t-statistics does indicate that if it is significant, the disparity is towards less females per position. However, Figure 18, did show an increasing number of female authorships per position over time. When fitting the total number of female authors with a logistic fit (fit parameters: L=12708, k=0.017, x_0 = 706), we found that with such a trend, we would reach equality by 2072 - that is another 48 years from now. Whether to interpret this as a success or a failure we will leave up to the individual but most importantly the number is not the whole story - if the number of female authors increase but say their number of citations do not - the "equal representation" is without meaning. This is something to remember when working in increasing diversity and inclusion.

Topics and Domain

Beyond disparities in the general publication variables we also found disparities between sub-fields of topics that female and male authors tend to publish within. The topics that the fraction of female authors dominated, Topic 2: "Medical Physics", Topic 17: "Meteorology", and Topic 18: "Learning and Teaching" could indicate that female physicist tend to be more drawn to interdisciplinary fields and fields with a less abstract purpose and application. Specifically, Learning and Teaching, the topic with largest disparity of percentage (9pp) between female and male authorships, respectively, is also the topic mostly related to a traditional female gender role. Similarly, the most male dominated topics (with disparity ¿2pp) Computational Physics, Fluid Dynamics, and Mathematical Physics and ML are very computationally heavy - a field often associated with a traditional male gender role. This indicates, that even as physics is an overall field assigned to traditional male gender role, there seem to be a tendency that even for female authors that break that traditional gender role, the gender roles persist within subfields of physics. However, the more equally distributed topics, especially Astrophysics, Engineering, and Geophysics do not necessarily support this theory. We saw no major changes in the distribution of genders between topics over time indicating that while the global distribution of female physicists do increase (Figure 19) the local changes within topics seem more stagnant. However, Topic 2: Medical Physics has decreased for male authorship proportion over time, potentially "making room" for more female authors - but this would have to be investigated further.

As Learning and Teaching is quite and outlier of traditional physics research the two domains: Didactic Research (Teaching and Learning) and Physics Research (all other topics) are made. While we indicated that the larger proportion of female authors in Didactic Research might be due to a more female tradition gender role one could also assume that this is a field of personal interest for female scientists interested in physics who have experienced the gender bias and barriers as discussed above. One could imagine the interest in Didactic Research could stem from an initial interest in physics and a wish to make physics more attainable for everyone. If this was the case, we wondered if there would be a tendency, specifically for female authors, to transition from Physics Research to Didactic Research. However, we saw no evidence of that. Rather, if anything, there is a tendency for male authors to transition from Didactic Research to Physics Research. However, the evidence that we have is very limited and further analysis would be required. However, the Didactic Research is still within the field of physics, indicating that at a minimum, female authors have some level of interest in physics and specifically how it is taught to new generations of physicist.

While we, as discussed, do see an increase in the amount of female authors over time, when grouped by domain we see that this is the case for both Physics Research (11 pp increase between 1970-2023) as well as Didactic Research (9pp increase between 1970-20023). The larger proportion of females in Didactic Research - 41% females in Didactic Research vs 21% females in Physics Research in 2023 - could be an indication that there might be either a difference in the entry barriers between the two domains or simply indicating that women tend to be more interested in Didactic Research than Physics Research. Obviously the reasons for the larger proportion of female authors in Didactic Research and seemingly particular interest would have to be investigated and discussed further.

One explanation for the larger proportion of female authorships in Didactic Research could lie within the total author and gender collaborations as displayed in Figure 32. The total author count within Didactic Research is generally lower than for that of Physics Research which probably makes the authorship distribution between genders more equal. We see an indication of homophily within Physics Research domain where all-male authorships dominate compared to the Didactic Research domain. Similarly, within the Physics Research domain grouped by topics, we also saw that the most extreme case of homophily was found within Topic 7: "Mathematical Physics and ML", and the most mixed gender publications were found within the female dominated Topic 2: "Medical Physics". This could be another underlying factor in the overall disparity of gender counts within Physics Research, specifically the most male dominated ones - they are simply not "invited" by their male colleagues. However, there is also a lower probability of male authors collaborating with a female author as there are fewer of them.

Prestige Markers

Investigating disparities between genders we found that female authors were significantly lower ranked than male authors for all three rankings: *Institution Ranking* (t(3.5e+05)=-27.22, p=> .001), *Journal Ranking* (t(3.2e+05)=-17.39, p=> .001), and *Author Ranking* (t(3.3e+05)=-22.32, p=> .001). This indicates a systematical prestige barrier for female physicists, especially in terms of institution ranking which could be another explanatory factor of what is withdrawing women from physics. Figure 37 reveals another examples of homophily with an extensive amount of (almost) all male institutions affiliations indicating that physics is closed land for female authors. However, as seen from Figure 36 some of these cases are very low statistics with only a few counts per institution. The gender distributions between journals align more with the overall distribution, indicating that the barriers happen to a further extent between institutions and between authors themselves in terms of citing. However, for the top journals we do see a lower proportion of females than in our overall dataset, again indicating a systematic lower ranking and prestige level for female authors.

Relating prestige to domains, we did see topics within the Physics Research domain defined as female dominated ranked as some of the highest between all three types of ranking. This indicate that female physicist might strategically strive for topics of a higher ranking when pursuing a career within physics.

However, the topic Learning and Teaching within the Didactic Research domain was the lowest ranked, contradicting this theory.

When grouping the ranking correlations between topics per gender we see that female authorships are systematically ranked lower - even within the female dominated topics. This indicates that while female authors might be present within highly ranked sub-fields of physics, it is the male authors driving the ranking.

Activity and Survival Rate

In terms of activity and survival rate we also tested whether there was a statistically significant level of disparity between female and male authors in terms of activity and survival. The first (t(1.5e+04)=31.25, p=> .001) and last publication year (t(1.4e+04)=13.46, p=> .001) of female authors were significantly later for female authors than male authors which was expected following the distribution in Figure 13(b). Naturally following therefrom, the female career span years is significantly shorter (t(1.6e+04)=-28.80, p=> .001). As most the female authors joined later we cannot interpret much from this yet. On the other hand, where the number of publications per female author was significantly smaller than that of their male counterpart (t(1.5e+04)=-11.13, p=> .001) when accounting for their academic age the publication count per year was not statistically smaller for female authors (t(1.3e+04)=15.50, p=1.0). Actually, female authors were more productive than their male colleagues, indicating that the disparities found in citations and level of prestige does not correlate with a lower production rate and that female physicists are not producing research. Lastly, the number of females that have dropped out were significantly less compared to the number of male authors - but this is also expected following their academic age. Instead, to interpret the drop rate we looked into the genders respective survival rate.

The survival rates reveal that female authors are statistically less likely to survive compared to their male colleagues (χ^2 = 609.23, p=> .001). While we did not fully explore the relation of survival to the other variables, one could imagine that with less recognition and prestige received, a lack of motivation follows ultimately leading to a lower probability of surviving. When grouping survival rate by domain and gender, we found no statistically significant differences between female and male authors publishing within "Didactic Research" (χ^2 = 0.66, p=.41), however, it was statistically significant between female and male authors publishing within "Physics Research" (χ^2 = 63.43, p=> .001). This indicates, that the lower survival probability is not necessarily gender related but rather specific to the field. However, the overall survival rate is lower for authors within Didactic Research why the two are not necessarily comparable domains in terms of survival.

Furthermore, while there is a consistent gap between female and male authors within Physics Research it is decreasing and almost closing when passing 20 years indicating that the survival rate is especially low for female author in the early career stage.

The fact that female authors are less probable at survival also indicate that the issue with gender representation within physics is not just a matter of "interest in the field" - fair enough if less women than men find it interesting to pursue a career within physics - then they shouldn't be forced to. However, from the survival rates we see that even for those that are clearly interested (since they are publishing within physics they probably pursued a 5+ year long academic education) but they still drop out a higher rate than their male colleagues. Therefore, the imbalanced distribution of male and female authors can not all be credited to the name of interest, as we are already beyond that at the point of publishing.

Clustering and Classification

Most of the findings discussed so far revealed revealed statistically significant disparities between female and male authors within physics indicating a systematic bias or barrier for female physicists. However, we wanted to explore how much of these disparities were actually related to gender and how much could be assigned to other factors such as academic age, country or topic. To answer this, we employed a clustering and classification model.

The clustering as well as classification results did not reveal a clear separation between genders. However, the clustering result did outline a global structure indicating different behaviours between the genders but also within genders. This indicates that while there might be a pattern between female and male authors, differences within genders exist too making the quantifiable of the genders less precise. Similarly, the classification model were not able to classify the genders with certainty. But, like the clustering, the classification was not completely random and we did obtain some extent of separability. The top features according for the separation decision were *Institution ID*, *Author Mean Cites*, *Institution Ranking*, *Journal ID*, *Author Publication Rate*, *Journal Ranking.* Where *Institution ID*, *Author Mean*

Cites, *Institution Ranking*, *Journal ID*, and *Journal Ranking* were also statistically significant between the genders it makes sense that they were an important measure. Furthermore, it underlines the extent to which female authors are less cited and affiliated with prestigious institutions and/or journals. However, the *Author Publication Rate* was not statistically less for female authors but rather for male authors. Again, an indication that the lack of recognition is not related to a lack of productivity.

However, when removing the top most important variables from the classification model we still saw a good classification (even though not as good). The next most important features were Cited By Count, Author Publication Count, Author First Country, Publication Year and Career Span Years. All of these features, except Author First Country, is related to the authors academic age, which we know is generally lower for the female authors in our dataset. Therefore, part of the disparities and separability between genders might also be explained by the element of time, meaning that the had we conducted this analysis on only recent data or say 10 years from now, we might have seen a smaller gap between the genders. But, at least we can say that from 1970 up till now, the level of prestige is a statistically significant marker of separability between female and male physicists.

Implications

This study brings awareness to potential biases and systemic barriers within the field of physics, particularly affecting women (and other minoritized populations). This awareness is crucial for fostering inclusivity and diversity in the discipline.

The comparison of female and male authors' differences and similarities provides a roadmap for targeted interventions aimed at increasing access and recognition for underrepresented groups in physics. By identifying areas where disparities are most pronounced, such as citation counts and survival rates, efforts can be directed towards addressing these specific challenges.

The distinction between female- and male-dominated subfields of physics offers valuable insights into successful strategies for fostering inclusivity: it indicates which sub-fields that are on to something and which not so much. Concrete examples from these subfields can serve as inspiration for developing more inclusive practices across the discipline.

The analysis of the evolution of female versus male authorship over time provides valuable foresight into future trends. Comparing these trends to established goals, such as those set by The Niels Bohr Institute, which was to have 35% associate professors and 30% professors in 2030, could serve as a proxy for the ambition level of such a target. The prediction result also underscores the importance of continued efforts to promote gender equity in academia.

While the focus of this study was on gender disparities in physics, its methodology and insights can be applied to other fields and social subgroups of interest. By adopting similar approaches, researchers can investigate disparities related to ethnicity, socio-economic status, sexual orientation, and other factors, contributing to a broader understanding of social equity in various work fields.

Comparison With Prior Work

In comparison to prior studies also examining publication patterns of authors, our findings align with their indications and observed trends. For example, Kozlowski et al. (2022) found evidence of ”homophily between identities and topics, suggesting a relationship between diversity in the scientific workforce and the expansion of the knowledge base” [10]. This resonates with our observation that male authors tend to collaborate more with other male authors, while the most diverse topic, ”Learning and Teaching,” exhibits a more equal distribution of collaboration. While our analysis focused solely on gender identities, Kozlowski et al. examined both gender and ethnicity, revealing underrepresentation of women across most ethnicities in Physics, Mathematics, and Engineering, but with more equitable representation among Asians. This parallels our finding of a higher proportion of female physicists from Asia. Furthermore, the sub-field of Physics Research with the most female authors were ”Medical Physics” which corresponds to the their finding of females to be more present within Health. Additionally, their observation of male authors being ”underrepresented in Humanities and Social Sciences and overrepresented in Physics, Engineering, Math, and Chemistry” [10] aligns with our distribution of male authors between the more humanity oriented Didactic Research domain and Physics Research domain.

Kozlowski et al. imply ”a connection between traditional gender roles and topics related to gender-based identity and inequality” [10]. This might back up the theory behind the topic distribution between genders that we found and emphasise the importance of that if we want to increase the diversity within physics we need to make sure that authors of diverse backgrounds and lived experiences can see themselves in the field and see a way of how they can relate and apply it to their own sense of meaning.

Lastly, Kozlowski et al. suggested "a further investigation of the scientific inequalities in relation to markers of prestige such as in terms of institutional affiliations" [10]. While they found women to be in less cited topics they hypothesised that they would overrepresented in fields of lower prestige as well. But in our study, we actually found a larger ratio of female authors in highly prestige topics. However, it was the male authors driving the prestige level and we did, after all, also see a statistically significant lower citation of female authors.

Furthermore, in "Dynamics in the entanglements of gender cultures and disciplinary cultures in science as a key for gender equality: the case of the physical sciences" by Martina Erlemann, she discussed how traditional gender roles may play a crucial role for the representation of female and male authors in physics. Like she points out, there is a traditional masculinised relation between men and machines and a traditional feminised relation with nurture and care. This match our findings from the distributions found between subfields of physics, namely that women are mostly represented within "Learning and Teaching" and men within fields where computation plays an essential role: Computational Physics, Fluid Dynamics, and Mathematical Physics and ML. "Medical Physics" and "Metereology" were also female dominated and where Medical Physics align with this theory, the domination within Metereology may be explained by other underlying factors.

Furthermore, a study by Matthew B. Ross et al. (Year), "Women are credited less in science than men", investigated the gender gap in scientific output and attribution, finding that women are significantly less likely than men to be credited with authorship [11]. This finding complements our observation of disparities in authorship attribution, particularly in the Physics Research domain. A qualitative response analysis by Ross et al. suggests that the gender gap in scientific output is more likely due to differences in attribution rather than scientific contribution, providing insight into the underlying factors behind our finding that the female author publication rate is not lower than that of male authors. This underscores the importance of specific and focused efforts to include and recognise female physicists, addressing issues of attribution and acknowledgement.

Limitations

This study had some limitations. First of all the database OpenAlex had a quite limited information on the authors - we were only provided their name why key variables such as gender, academic age, activity and survival had to be estimated. Similarly, we had to label each publication based on a topic model which also contributes to uncertainty in terms of the analysis done on topics. However, we did work with very large statistics and made our estimations under reasonable assumption, evaluating that the overall statistics presented in this study is somewhat aligned with reality.

The gender API used to estimate the authors gender, namely gender-gueseser, was definetely not the best out of all the gender APIs available. Had we had funding or other economic support we could have probably attained a much more precise description of the gender. Using an API such as Gender Guesser not only provides the name but also with which probability it is determined. This could have been used to underline the certainty of the study findings to a much higher degree. However, we did test a sample of the gender provided by gender-guesser on names that we know with certainty the gender of and found it reasonable. Also, we again rely on the large statistics and assume that even though not all genders are assigned correctly the average is. The distribution of female and male authors is also aligned with what we would expect why it seems reasonable.

For the Machine Learning part we had some computational challenges in terms of for example performing a gridsearch in order to test many more and find optimal hyperparameters. This might mean that the model performance is not its best. Howver, we did do some model tuning even though not as thoroughly and therefore assume that it would be minimally optimised by further tuning and that the results are overall similar and would not reveal something we do not see already from this study.

Lastly, our dataset contains data from 1970 and up till today and the statistics are therefore sort of averaged out over time. Therefore, it can sometimes be challenging to interpret how much of the effect is status quo and how much of it was from 50 years ago. However, part of the analysis was to study the element of time and while it is not always distinguishable it also serves as an indicator of the "historic" influence that we still see in the statistics off of today.

Overall, the study and the data it encompassed were quite broad, which at times made it challenging to draw precise conclusions regarding the underlying factors. However, this broad approach served to present overall trends and offer a preliminary depiction of gender disparities in physics. It also aligns with the aim of our study, which was to provide general insights into the movement, nature, and patterns of behaviour and development among female physicists. While future studies are recommended to delve deeper into specific areas, our study offers a preliminary overview of topics worthy of further investigation to enhance our understanding of women's experiences in physics.

Recommendations

To further understand career disparities between female and male authors future studies could include conducting similar types of analysis on other, or even several, databases. This could be databases such as X (earlier Twitter) and Github which might reveal other underlying patterns also outside the academic publication realm. It would also be interesting to get access to a more detailed dataset, such as Scopus, which include many of the estimated variables as predefined variables to get a more ”True” picture of the distributions. This could also serve as a measure for how well gender-guesser and the topic model worked.

While this study provides insights into the quantitative disparities it would be important to highlight these findings by supplementing with qualitative studies. This could help further investigate the complexity of the experience as a female physicist and how the negative experience could be avoided and the positive experiences enhanced.

Future studies could also focus on really optimising the Machine Learning Models and testing the reliability of gender-guesser, or create a gender API, in order to make sure no bias from the training of the gender API is projected onto the results of this study and also in order to make sure that the models are performing at their highest performance level so not missing out on important results. Specifically for this study, there is also a challenge in the very huge ubiquity of male authors why it can be hard for the model to balance the two classes leading to an overestimate of FPs (female labels predicted to be males) since this is the most probable case.

Furthermore, parts of this study could be further explored such as further investigate disparities between topics and domains and why female/male authors are more/less represented in some fields over others. Furthermore, the clustering results indicated not only a global structure between female and male authors, but also internal clusters between the two classes. The characteristics of these within-gender differences should be investigated. Moreover, it would be interesting to investigate the underlying factors going into the differences found in probability of survival between genders and domains. Many factors could affect this such as maternity/paternity leave (or lack of), career transitioning and academic age.

To further underline these findings, it would be of interest to take into account the timely element. While we did observe a development over time between female and male authors, the analysis conducted in this study mostly did not differentiate between points in time but rather provided a timely average. This approach may overlook important temporal trends and variations within subgroups.

Given that most of our female authors had their first publication year later than the male authors in our dataset, it becomes challenging to directly compare their career span years, activity levels, and survival rates. Addressing this challenge, employing a model such as the Cox proportional hazards model for survival analysis offers a promising approach. By integrating covariates such as academic age or gender into the analysis, we can assess their influence on survival rates while controlling for other factors. This method allows for a nuanced examination of gender disparities in authorship over time, considering the impact of associated variables. Incorporating the Cox model enables us to evaluate the persistence of gender disparities while accounting for temporal variations and relevant covariates. This enhances the depth and accuracy of our analysis, providing valuable insights into the complex dynamics of gender disparities in academia.

Lastly, this is a very broad study including data from the whole world and spanning more than 50 years as well as 13 topics distributed between two distinct domains. Therefore, a subgroup analysis limited to for example a specific country, only first author or sub-field could potentially underline the specific challenges and successes between different cases. This would probably be relevant if the results should be implemented and included in a diversification strategy say like the one presented by The Niels Bohr Institue.

Conclusion

The analysis reveals pervasive gender disparities within the field of physics, indicating systemic challenges for female authors. Despite an increasing amount of female authors the past few years, the distribution of female and male authors remains imbalanced. While the ratio might be increasing, the publication patterns still show statistical significant differences favouring male authors. This is particularly evident in disparities within citation counts and survival rates, suggesting persistent biases and challenges. While numerical parity in authorship may be achievable in the future, substantive equality requires addressing underlying issues such as biases in citation practices and institutional affiliations. Furthermore, the analysis highlights the need for continued research to understand and address the multifaceted nature of gender disparities in physics, emphasising the importance of promoting diversity and inclusion in the field.

Bibliography

[1] *Profile and history.* URL: https://nbi.ku.dk/english/about/profile-and-history/ (visited on 11/19/2023).

[2] NBA. *arkivdk.* URL: https://arkiv.dk/en/vis/5946558 (visited on 05/05/2024).

[3] *Lise Meitner.* URL: https://en.wikipedia.org/wiki/Lise_Meitner (visited on 12/09/2023).

[4] Ann-Louise Bergström. *Lise Meitner, Oppenheimer and the Matilda-effect.* URL: https://movingscience.dk/lise-meitner-oppenheimer-and-the-mathilda-effect/ (visited on 12/09/2023).

[5] *Oppenheimer (film).* URL: https://en.wikipedia.org/wiki/Oppenheimer_(film) (visited on 12/09/2023).

[6] *The Second Sex.* URL: https://en.wikipedia.org/wiki/The_Second_Sex (visited on 05/05/2024).

[7] JANE J. LEE. *6 Women Scientists Who Were Snubbed Due to Sexism.* URL: https://www.nationalgeographic.com/culture/article/130519-women-scientists-overlooked-dna-history-science (visited on 11/19/2023).

[8] Per Lunnemann, Mogens H. Jensen, and Liselotte Jauffred. "Gender Bias in Nobel Prizes". In: *Palgrave Commun* 5.46 (2019), pp. 1–4. DOI: 10.1057/s41599-019-0256-3.

[9] Henriette Tolstrup Holmegaard and Bjørn Friis Johannsen. "Science Talent and Unlimited Devotion: An Investigation of the Dynamics of University Students' Science Identities Through the Lens of Gendered Conceptualisations of Talent". English. In: *Contributions from Science Education Research.* Ed. by Henriette Tolstrup Holmegaard and Louise Archer. Contributions from Science Education Research. Publisher Copyright: © 2023, Springer Nature Switzerland AG. Springer, 2023, pp. 113–140. ISBN: 978-3-031-17642-5. DOI: https://doi.org/10.1007/978-3-031-17642-5_6.

[10] Diego Kozlowski et al. "Intersectional inequalities in science". In: *PNAS* 119.2 (2022), pp. 1–8. DOI: 10.1073/pnas.2113067119.

[11] Matthew B. Ross et al. "Women are credited less in science than men". In: *Nature* 608 (2022), pp. 135–145. DOI: 10.1038/s41586-022-04966-w.

[12] Bas Hofstra et al. "The Diversity-Innovation Paradox in Science". In: *PNAS* 117.17 (2020), pp. 9284–9291. DOI: 10.1073/pnas.1915378117.

[13] Rembrand Koning, Sampa Samila, and John-Paul Ferguson. "Who do we invent for? Patents by women focus more on women's health, but few women get to invent". In: *HScience* 372.6548 (2021), pp. 1345–1348. DOI: 10.1126/science.aba6990.

[14] Mathias Wullum Nielsen et al. "Gender diversity leads to better science". In: *PNAS* 114.8 (2017), pp. 1740–1742. DOI: 10.1073/pnas.1700616114.

[15] *Gender equality and diversity at NBI.* URL: https://nbi.ku.dk/english/about/profile-and-history/gender-equality-diversity/ (visited on 11/22/2023).

[16] "Accounting for sex and gender makes for better science". In: *Nature* 588 (2020), p. 196. DOI: 10.1038/d41586-020-03459-y.

[17] Mathias Wullum Nielsen, Carter Walter Bloch, and Londa Schiebinger. "Making gender diversity work for scientific discovery and innovation". In: *Nat Hum Behav* 2 (2018), pp. 726–734. DOI: 10.1038/s41562-018-0433-1.

[18] Mathias Wullum Nielsen et al. "One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis". In: *Nature Human Behaviour* 1 (2017), pp. 791–796. DOI: 10.1038/s41562-017-0235-x.

[19] Londa Schiebinger. *What is Gendered Innovations?* URL: https://genderedinnovations.stanford.edu/what-is-gendered-innovations.html (visited on 12/09/2023).

[20] Caroline Criado Perez. *Invisible Women: Exposing Data Bias in a World Designed for Men.* New York, NY: Vintage, 2020.

[21] *Power-knowledge.* URL: https://en.wikipedia.org/wiki/Power-knowledge (visited on 12/12/2023).

[22] Emer Brady et al. "Lack of consideration of sex and gender in COVID-19 clinical studies". In: *Nat Commun* 12.4015 (2021), pp. 1987–1994. DOI: 10.1038/s41467-021-24265-8.

[23] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* 838 Broadway, New York, NY: NYU Press, 2018.

[24] Lu Liu et al. "Data, measurement and empirical methods in the science of science". In: *Nature Human Behaviour* 7 (2023), pp. 1046–1058. DOI: 10.1038/s41562-023-01562-4.

[25] Stephan Risi et al. "Diversifying history: A large-scale analysis of changes in researcher demographics and scholarly agendas". In: *PLoS ONE* 17(1).e0262027 (2022), pp. 1–17. DOI: 10.1371/journal.pone.0262027.

[26] Aron Laxdal. "The sex gap in sports and exercise medicine research: who does research on females?" In: *Scientometrics* 128 (2023), pp. 1987–1994. DOI: 10.1007/s11192-023-04641-5.

[27] Thomas Riisgaard Hansen et al. "Mangel på digitale kompetencer kan bremse det danske kvanteeventyr". In: *Børsen* (2023). URL: https://borsen.dk/nyheder/opinion/mangel-paa-digitale-kompetencer-kan-bremse-det-danske-kvanteeventyr (visited on 12/28/2023).

[28] Martina Erlemann. "Dynamics in the entanglements of gender cultures and disciplinary cultures in science as a key for gender equality: the case of the physical sciences". In: (), p. 6.

[29] Amanpreet Kohli. *"Snowballing" in Systematic Literature Review.* URL: https://www.linkedin.com/pulse/snowballing-systematic-literature-review-amanpreet-kohli (visited on 02/08/2024).

[30] Michael E. Rose and John Kitchin. *pybliometrics: Python-based API-Wrapper to access Scopus.* URL: https://pybliometrics.readthedocs.io/en/stable/ (visited on 01/20/2024).

[31] Elsevier B.V. *How much data can I retrieve with my APIKey?* URL: https://dev.elsevier.com/api_key_settings.html (visited on 01/20/2024).

[32] Zihang Lin et al. "SciSciNet: a large-scale open data lake for the science of science research". In: *Scientific Data* 10.315 (2023), pp. 1–22. DOI: 10.1038/s41597-023-02198-9.

[33] Jonathan De Bruin. *PyAlex.* Version 0.8. Jan. 2023. URL: https://github.com/J535D165/pyalex.

[34] Dario Radečić. *Stop Using CSVs for Storage — Pickle is an 80 Times Faster Alternative.* URL: https://towardsdatascience.com/stop-using-csvs-for-storage-pickle-is-an-80-times-faster-alternative-832041bbc199 (visited on 04/08/2024).

[35] Reza Shokrzad. *Pickle, JSON, or Parquet: Unraveling the Best Data Format for Speedy ML Solutions.* URL: https://medium.com/@reza.shokrzad/pickle-json-or-parquet-unraveling-the-best-data-format-for-speedy-ml-solutions-10c3f7bf4d0c (visited on 04/08/2024).

[36] NYU Libraries. *Gender and Sexuality Studies.* URL: https://guides.nyu.edu/genderandsex/terminology (visited on 01/24/2024).

[37] David Arroyo Menendez. *damegender 0.1.45.* URL: https://pypi.org/project/damegender/0.1.45/ (visited on 01/20/2024).

[38] *ISO 3166-1 alpha-2.* URL: https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2 (visited on 01/24/2024).

[39] dukebody. *gender-guesser/gender$_g$uesser/detector.py.* URL: https://github.com/lead-ratings/gender-guesser/blob/master/gender_guesser/detector.py (visited on 04/08/2024).

[40] Israel Saeta Pérez. *gender-guesser 0.4.0.* Version 0.4.0. Jan. 2016. URL: https://pypi.org/project/gender-guesser/.

[41] imajhere. *Discovering Insights with Chi Square Tests: A Hands-on Approach in Python.* URL: https://www.analyticsvidhya.com/blog/2023/03/discovering-insights-with-chi-square-tests-a-hands-on-approach-in-python/ (visited on 04/12/2024).

[42] *scipy.stats.ttest$_i$nd.* URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html (visited on 05/04/2024).

[43] *scipy.stats.mannwhitneyu.* URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html (visited on 05/04/2024).

[44] Matthew Hoffman, David Blei, and Francis Bach. "Online Learning for Latent Dirichlet Allocation". In: vol. 23. Nov. 2010, pp. 856–864.

[45] Matthew D. Hoffman. *onlineldavb.* Version 3. Jan. 2010. URL: https://github.com/blei-lab/onlineldavb/tree/master.

[46] Cory Maklin. *Latent Dirichlet Allocation.* URL: https://medium.com/@corymaklin/latent-dirichlet-allocation-dfcea0b1fddc (visited on 01/25/2024).

[47] Susan Li. *Topic Modeling and Latent Dirichlet Allocation (LDA) in Python.* URL: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 (visited on 01/25/2024).

[48] Shashank Kapadia. *Evaluate Topic Models: Latent Dirichlet Allocation (LDA).* URL: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 (visited on 02/13/2024).

[49] Ms Aerin. *Perplexity Intuition (and its derivation).* URL: https://towardsdatascience.com/perplexity-intuition-and-derivation-105dd481c8f3 (visited on 02/13/2024).

[50] João Pedro. *Understanding Topic Coherence Measures.* URL: https://towardsdatascience.com/understanding-topic-coherence-measures-4aa41339634c (visited on 02/13/2024).

[51] Michael Röder, Andreas Both, and Alexander Hinneburg. "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.* WSDM '15. Shanghai, China: Association for Computing Machinery, 2015, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

[52] Pratik Kumar. *Understanding Kaplan-Meier Estimator (Survival Analysis).* URL: https://towardsdatascience.com/understanding-kaplan-meier-estimator-68258e26a3e4 (visited on 04/06/2024).

[53] *Logrank test.* URL: https://en.wikipedia.org/wiki/Logrank_test (visited on 04/06/2024).

[54] Zoumana Keita. *Classification in Machine Learning: An Introduction.* URL: https://www.datacamp.com/blog/classification-machine-learning (visited on 03/25/2024).

[55] Kiruthika Devaraj. *Top difference between training data and testing data.* URL: https://testsigma.com/blog/difference-between-training-data-and-testing-data// (visited on 05/04/2024).

[56] *Welcome to LightGBM's documentation!* Version 4.3.0. 2024. URL: https://lightgbm.readthedocs.io/en/v4.3.0/.

[57] Arnab Mondal. *Complete guide on how to Use LightGBM in Python.* URL: https://www.analyticsvidhya.com/blog/2021/08/complete-guide-on-how-to-use-lightgbm-in-python/ (visited on 03/25/2024).

[58] *1.10. Decision Trees.* URL: https://scikit-learn.org/stable/modules/tree.html (visited on 03/24/2024).

[59] *What is a decision tree?* URL: https://www.ibm.com/topics/decision-trees (visited on 05/04/2024).

[60] Stacey Ronaghan. *The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark.* URL: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3 (visited on 03/24/2024).

[61] MLMath.io. *Math behind Decision Tree Algorithm.* URL: https://ankitnitjsr13.medium.com/math-behind-decision-tree-algorithm-2aa398561d6d (visited on 03/24/2024).

[62] *sklearn.preprocessing.StandardScaler.* URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (visited on 06/15/2023).

[63] *A Step-By-Step Complete Guide to Principal Component Analysis — PCA for Beginners.* URL: https://www.turing.com/kb/guide-to-principal-component-analysis (visited on 06/15/2023).

[64] shreyanshisingh28. *LightGBM (Light Gradient Boosting Machine)*. URL: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/ (visited on 03/25/2024).

[65] Abhishek Jain. *A Comprehensive Guide to Performance Metrics in Machine Learning*. URL: https://medium.com/@abhishekjainindore24/a-comprehensive-guide-to-performance-metrics-in-machine-learning-4ae5bd8208ce (visited on 04/08/2023).

[66] Ellie Frank. *Understanding the F1 Score*. URL: https://ellielfrank.medium.com/understanding-the-f1-score-55371416fbe1 (visited on 04/06/2024).

[67] Francis Sahngun Nahm. "Receiver operating characteristic curve: overview and practical use for clinicians". In: *Korean J Anesthesiol* 75.1 (2022), pp. 25–36. DOI: 10.4097/kja.21209.

[68] *Precision-Recall*. URL: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html (visited on 04/06/2024).

[69] Sarang Narkhede. *Understanding Confusion Matrix*. URL: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (visited on 04/06/2024).

[70] Nikolaj Buhl. *F1 Score in Machine Learning*. URL: https://encord.com/blog/f1-score-in-machine-learning/ (visited on 04/27/2024).

[71] Karimollah Hajian-Tilaki. "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation". In: *Caspian J Intern Med* 4.2 (2013), pp. 627–635.

[72] Tom Alon. *Ultimate Guide to PR-AUC: Calculations, uses, and limitations*. URL: https://www.aporia.com/learn/ultimate-guide-to-precision-recall-auc-understanding-calculating-using-pr-auc-in-ml/ (visited on 04/27/2024).

[73] *Clustering analysis - Wikipedia*. URL: https://en.wikipedia.org/wiki/Cluster_analysis (visited on 06/15/2023).

[74] Renesh Bedre. *UMAP dimension reduction algorithm in Python (with example)*. URL: https://www.reneshbedre.com/blog/umap-in-python.html (visited on 03/25/2024).

[75] Leland McInnes. *UMAP for Supervised Dimension Reduction and Metric Learning*. URL: https://umap-learn.readthedocs.io/en/latest/supervised.html (visited on 04/06/2024).

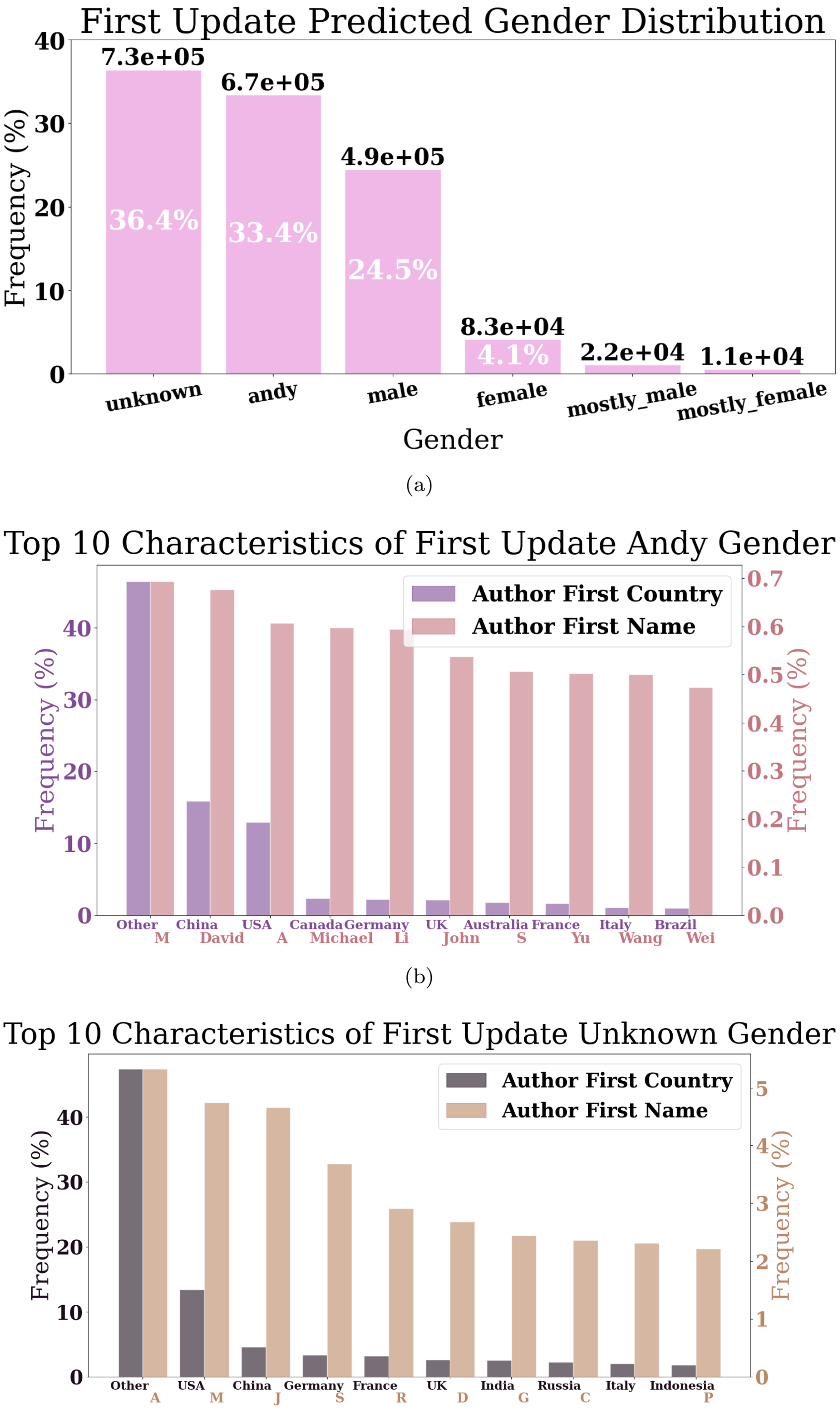
[76] Jonathan Lee. *Statistical Learning- Classification*. URL: https://sas.uwaterloo.ca/~aghodsib/courses/f07stat841/notes/lecture6.pdf (visited on 04/07/2024).

[77] Yana Huang and Tianyu Wang. "MULAN in the name: Causes and consequences of gendered Chinese names". In: *China Economic Review* 75 (2022), p. 101851. ISSN: 1043-951X. DOI: https://doi.org/10.1016/j.chieco.2022.101851. URL: https://www.sciencedirect.com/science/article/pii/S1043951X22001092.

Appendix A

Update Gender Assignment

Update 1: Update According to Duplicates in Author ID with an Assigned Gender



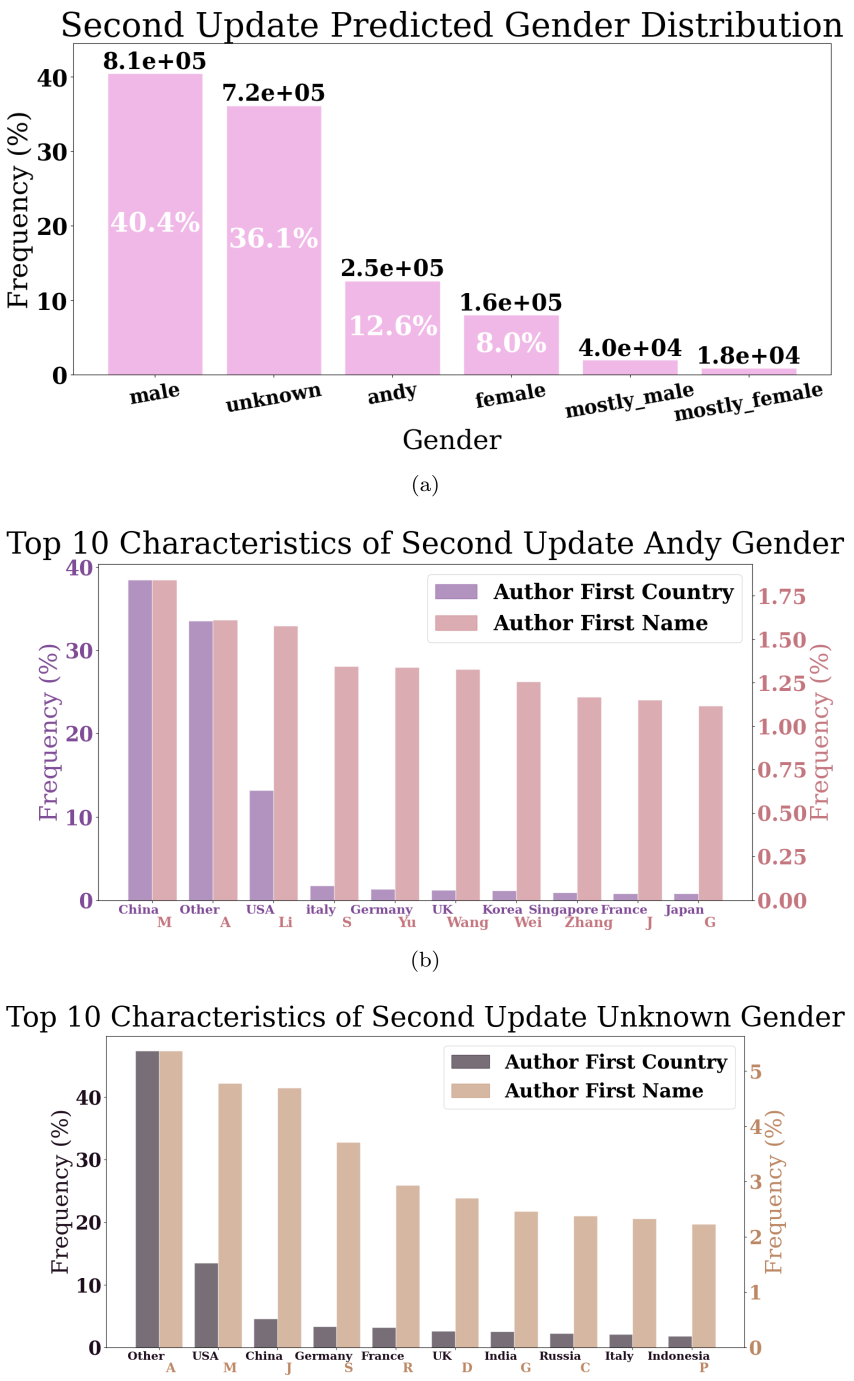
(a)

(b)

(c)

Figure A1: Distribution of predicted gender across the dataset after the first update according to duplicates in author id with an assigned gender. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 2: Update Without Country



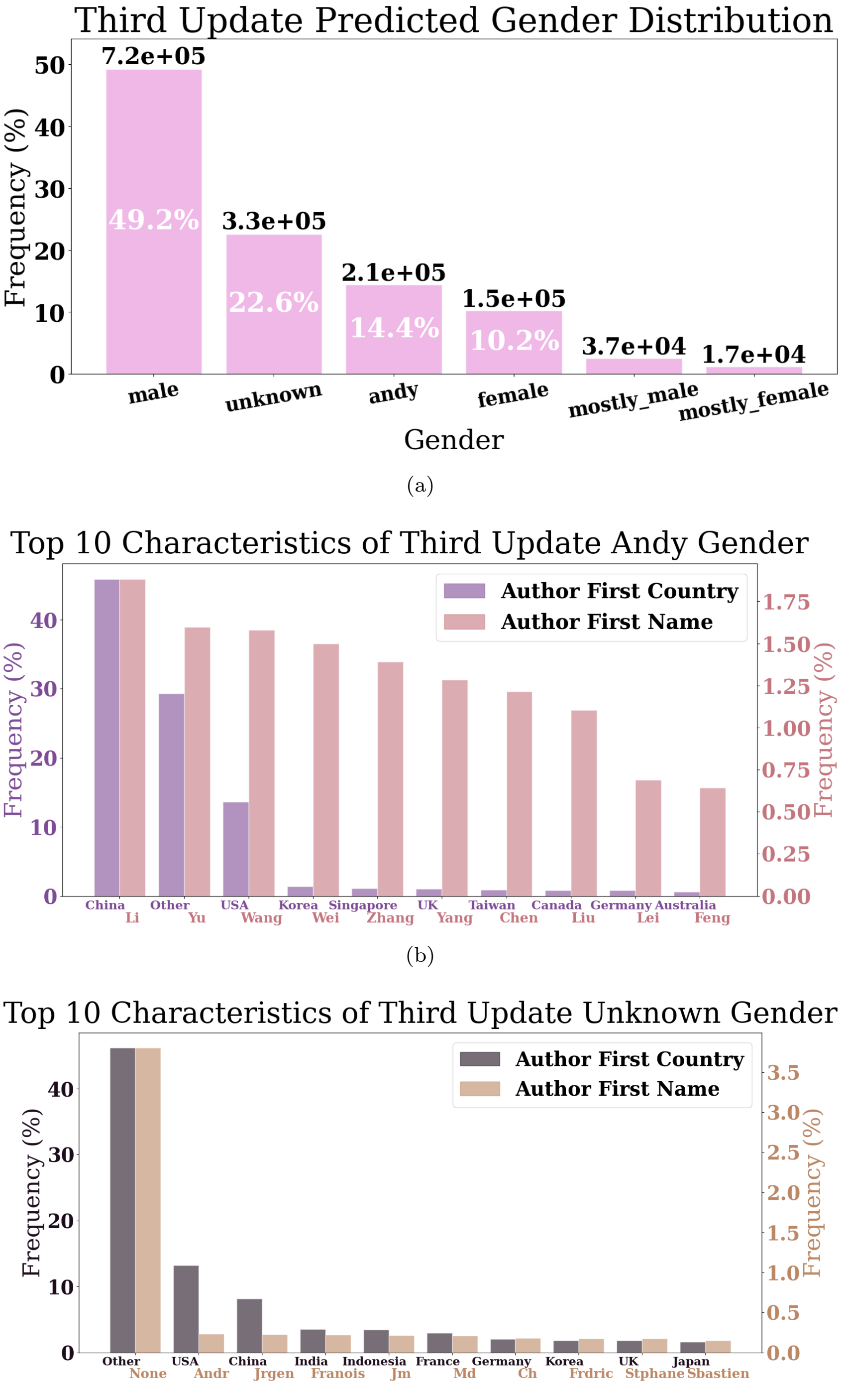
(a)

(b)

(c)

Figure A2: Distribution of predicted gender across the dataset after the second update without giving country as an input. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 3: Drop Rows With Invalid Names



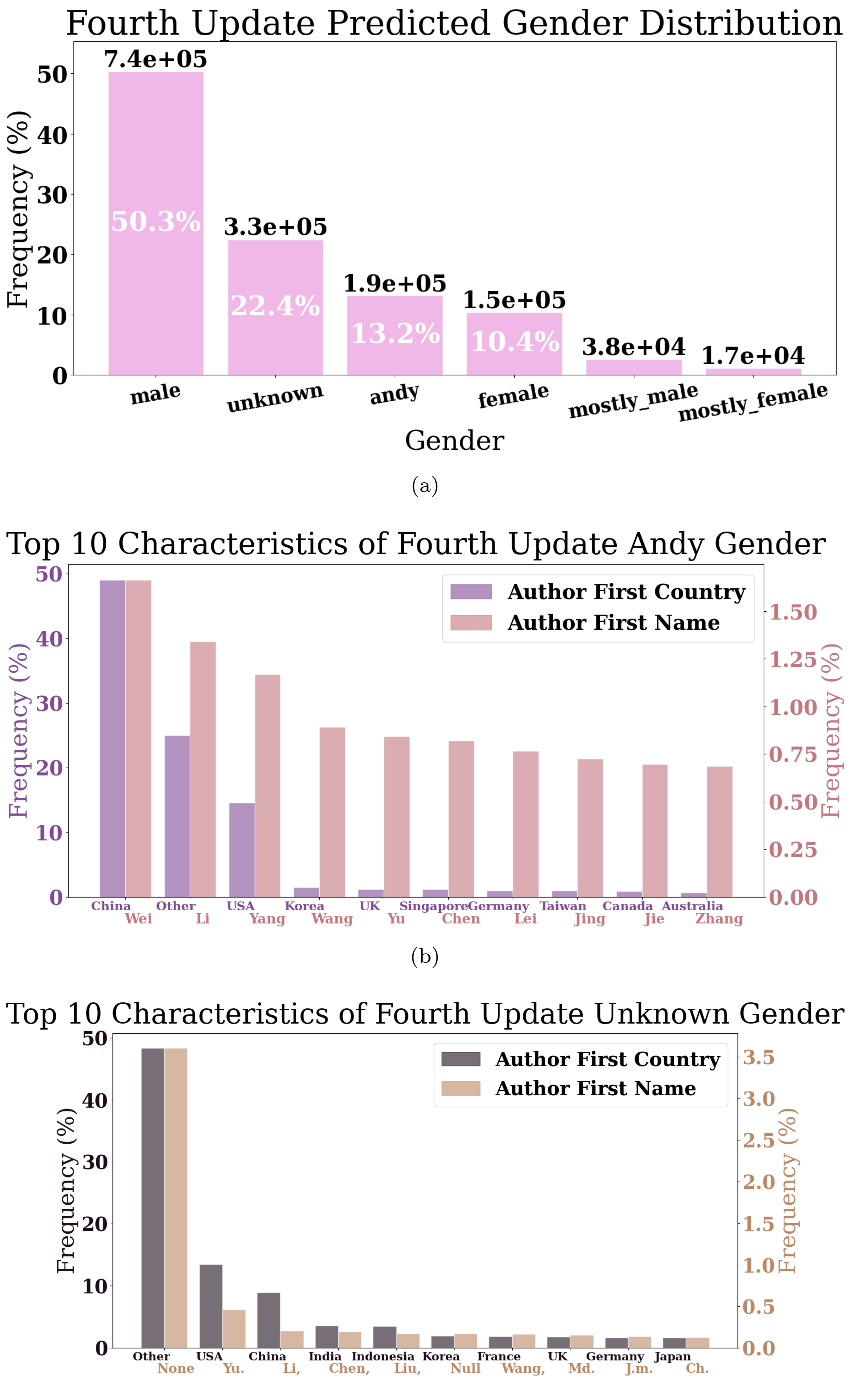
(a)

(b)

(c)

Figure A3: Distribution of predicted gender across the dataset after the third update dropping rows with invalid names. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 4: Update Names Including Special Characters



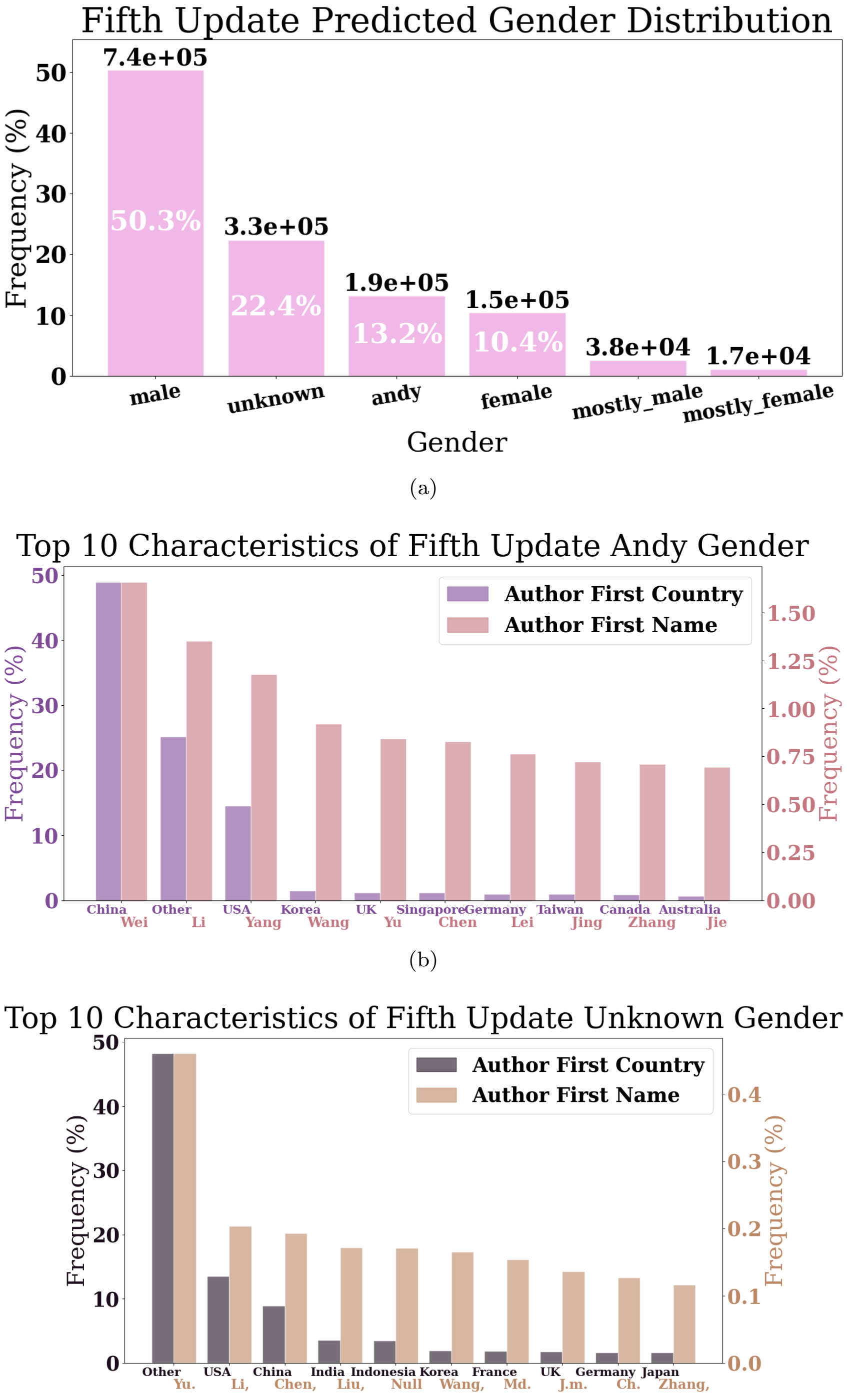
(a)

(b)

(c)

Figure A4: Distribution of predicted gender across the dataset after the fourth update targeting names including special characters. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 5: Update None Names



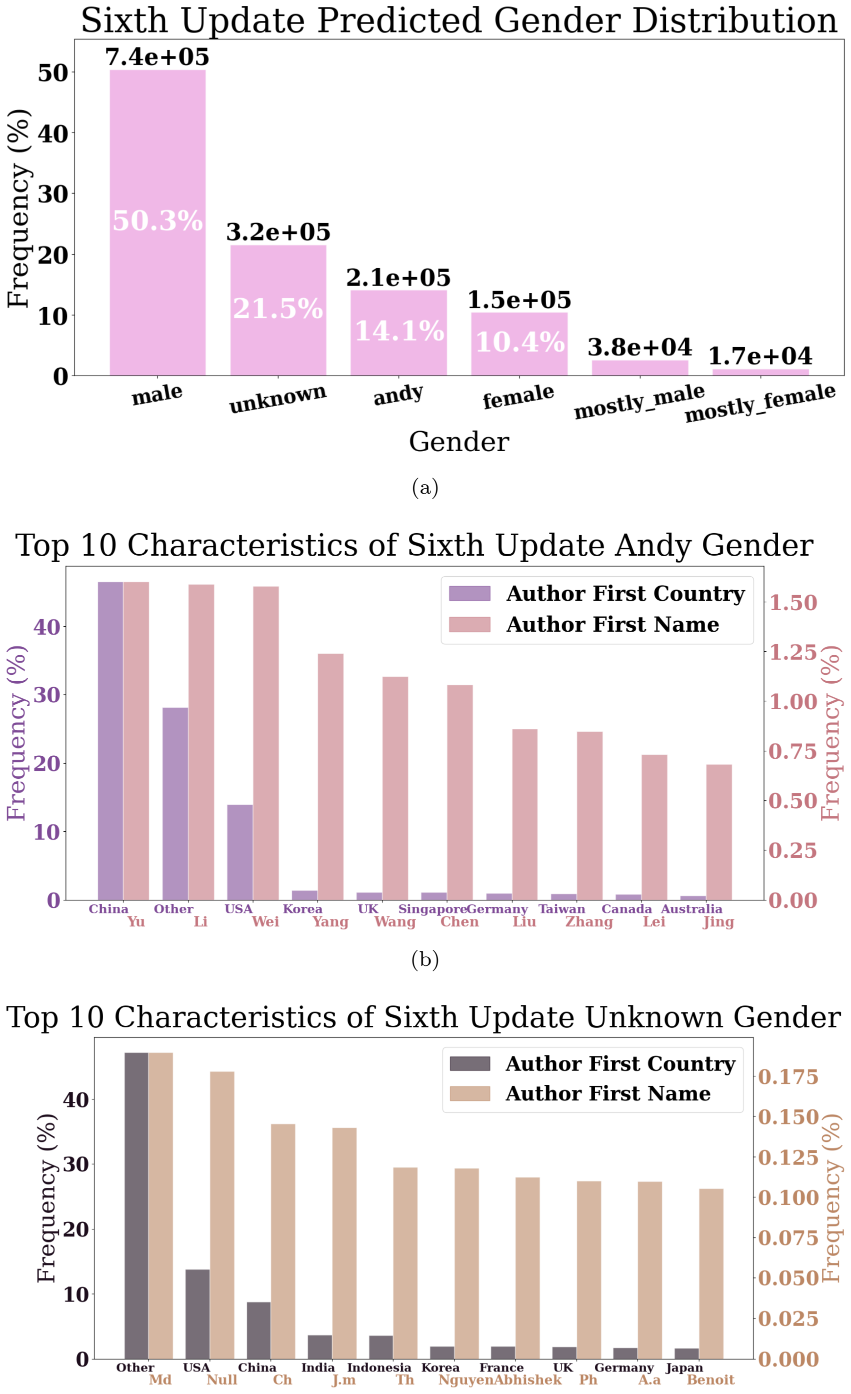
(a)

(b)

(c)

Figure A5: Distribution of predicted gender across the dataset after the fifth update targeting None names. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 6: Remove '.', ',', and '-' From Names and Update Gender



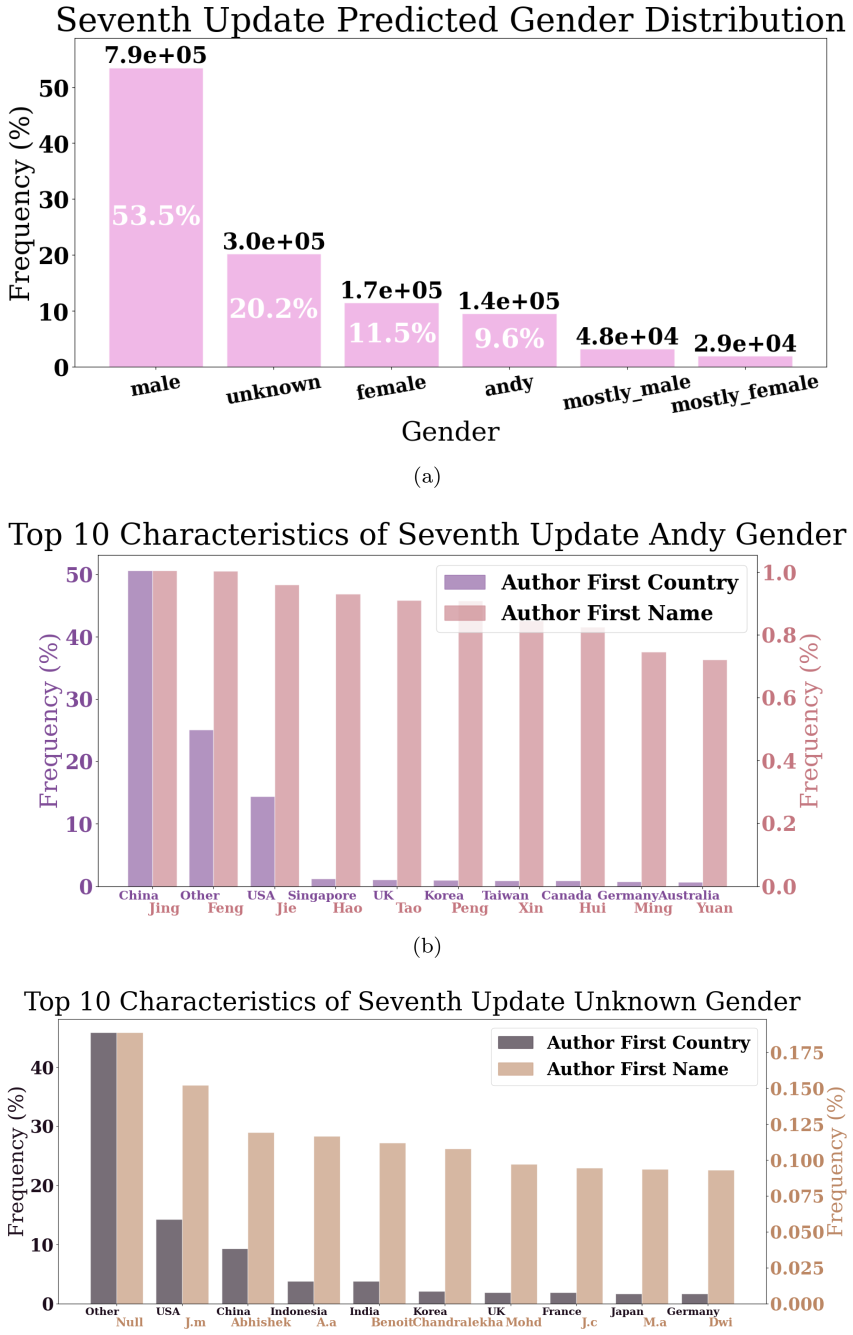
(a)

(b)

(c)

Figure A6: Distribution of predicted gender across the dataset after the sixth update removing punctuation tails. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 7: Update With Gender of Identical Name



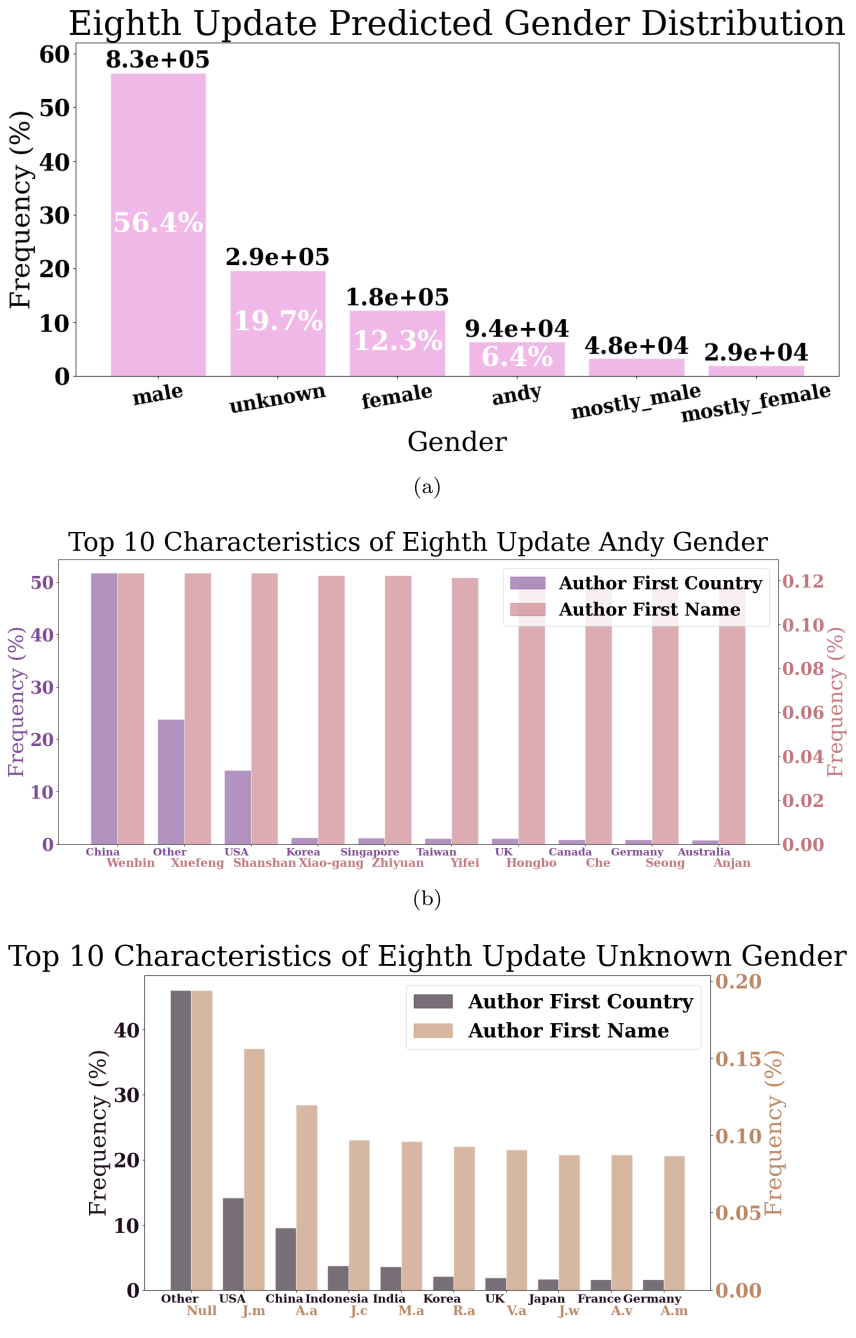
(a)

(b)

(c)

Figure A7: Distribution of predicted gender across the dataset after the seventh update copying genders of identical name (and country). **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 8: Update Using *Gender API*



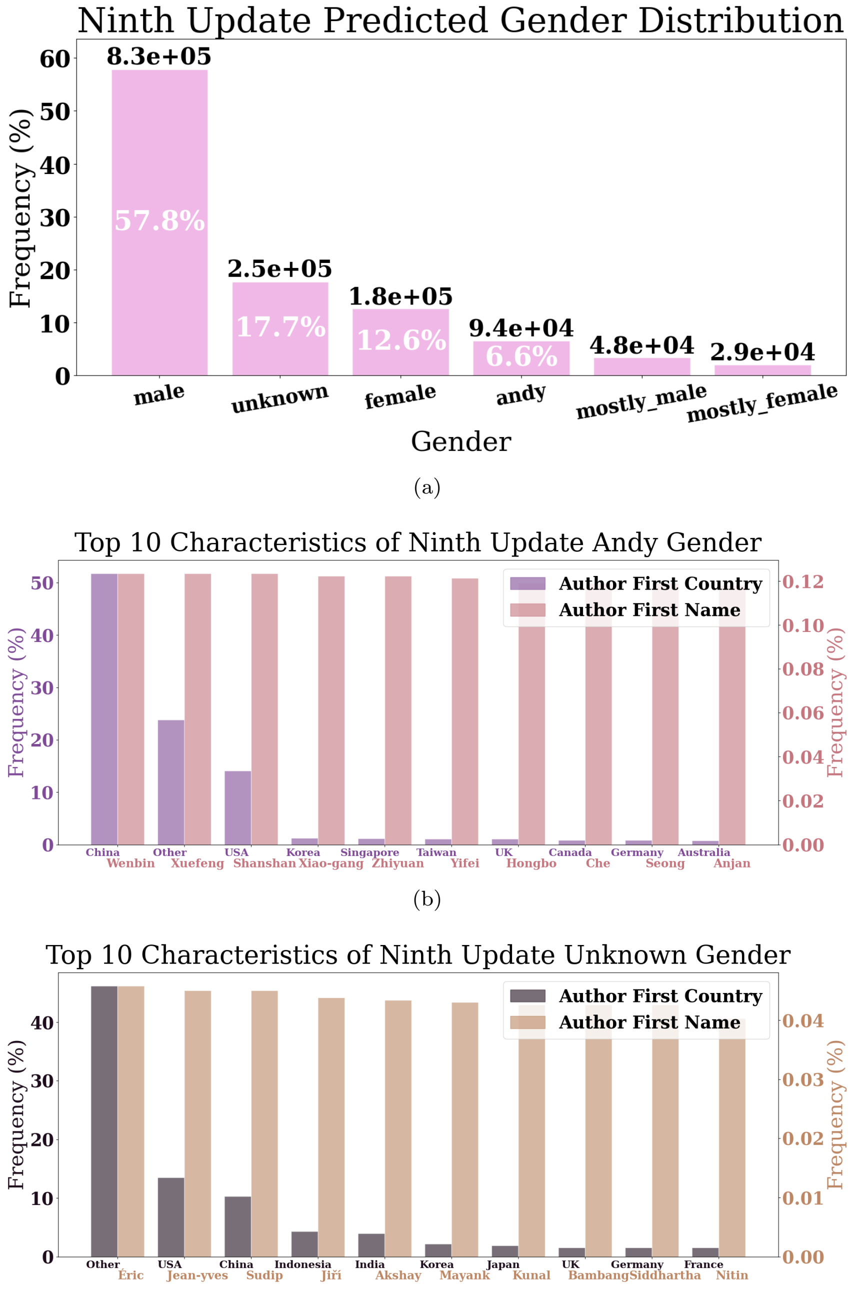
(a)

(b)

(c)

Figure A8: Distribution of predicted gender across the dataset after the eighth update using Gender API. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Update 9: Final Drop Rows With Invalid Names



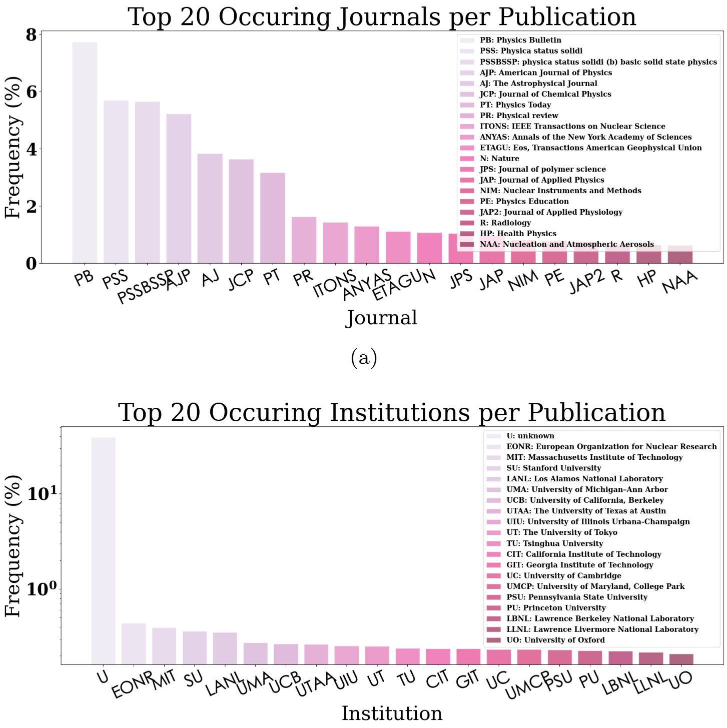
(a)

(b)

(c)

Figure A9: Distribution of predicted gender across the dataset after the ninth and final update dropping rows with invalid names. **(a)** The bar plot shows the frequency distribution of predicted genders, including male, female, andy, and unknown categories. **(b)** The bar plot illustrates the ten most frequent characteristics (author names and countries) associated with andy gender and **(c)** unknown gender.

Initial Distributions

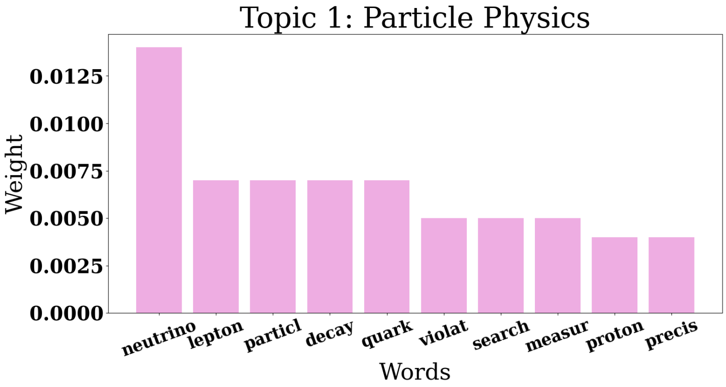


(a)

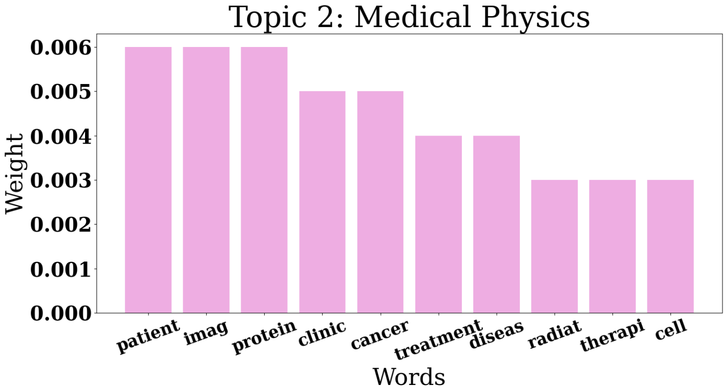
(b)

Figure A10: Top 20 overall occurring journals and institutions in our dataset per publication

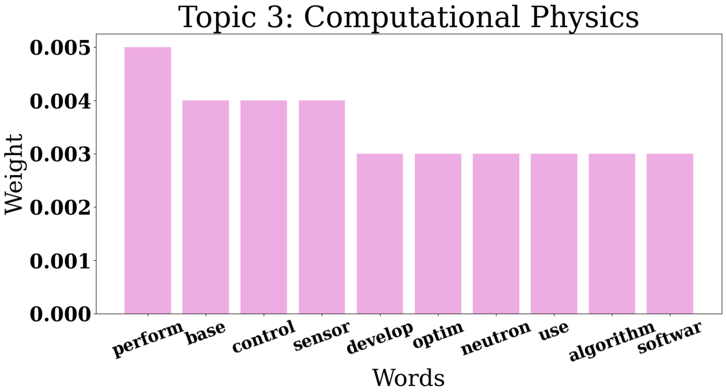
Topic and Domain



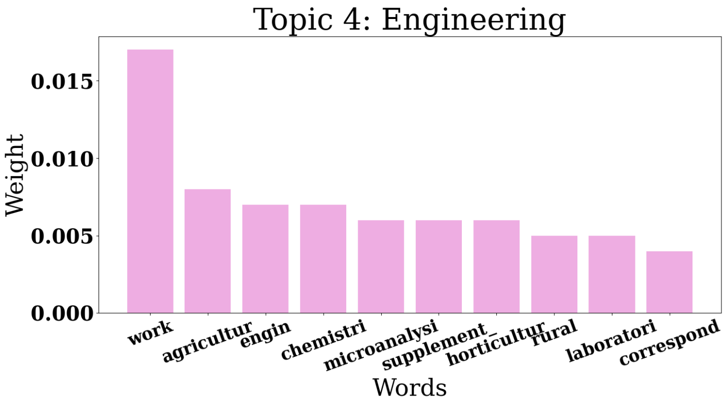
(a)



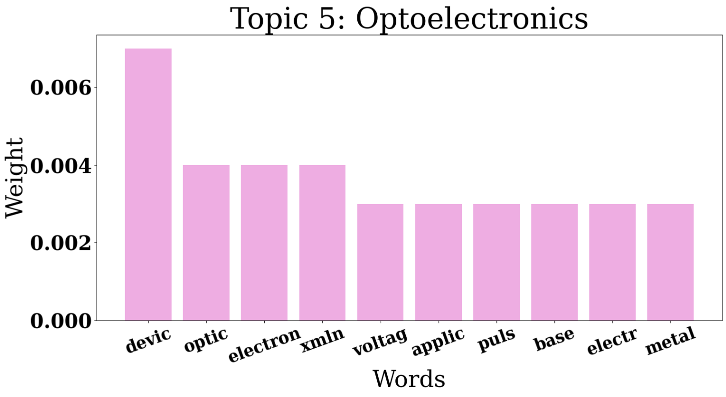
(b)



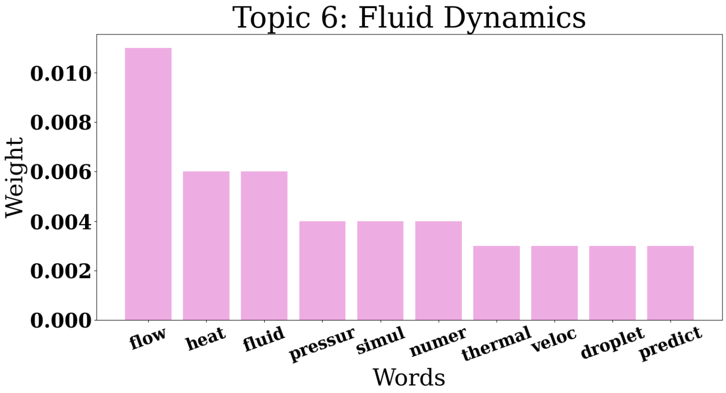
(c)


Topic 4: Engineering
Weight
0.015
0.010
0.005
0.000
work
agricultur
engin
chemistri
microanalysi
supplement_
horticultur
rural
laboratori
correspond
Words

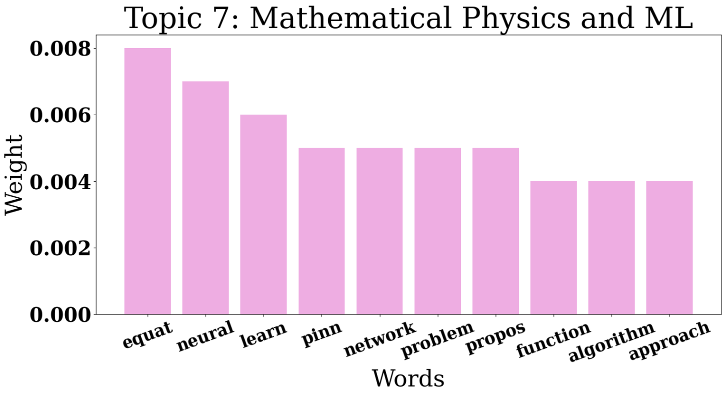
(d)


Topic 5: Optoelectronics
Weight
0.006
0.004
0.002
0.000
devic
optic
electron
xmln
voltag
applic
puls
base
electr
metal
Words

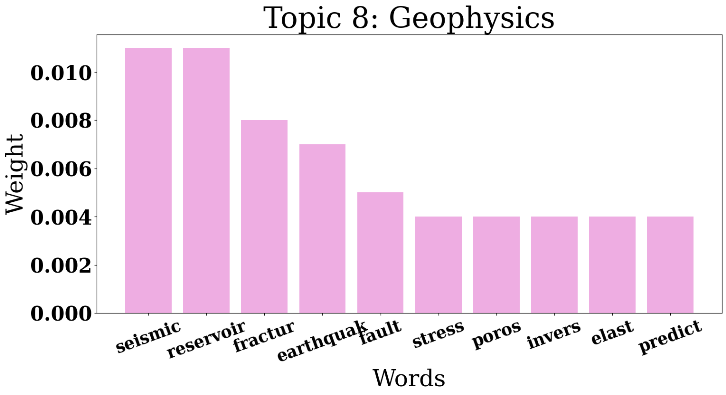
(e)


Topic 6: Fluid Dynamics
Weight
0.010
0.008
0.006
0.004
0.002
0.000
flow
heat
fluid
pressur
simul
numer
thermal
veloc
droplet
predict
Words

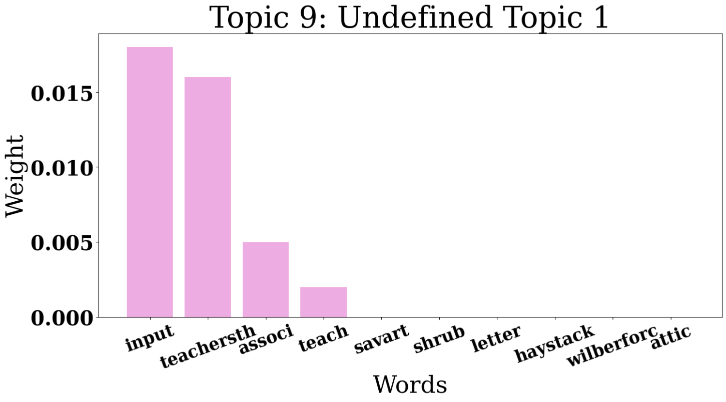
(f)


Topic 7: Mathematical Physics and ML
Weight
0.008
0.006
0.004
0.002
0.000
equat
neural
learn
pinn
network
problem
propos
function
algorithm
approach
Words

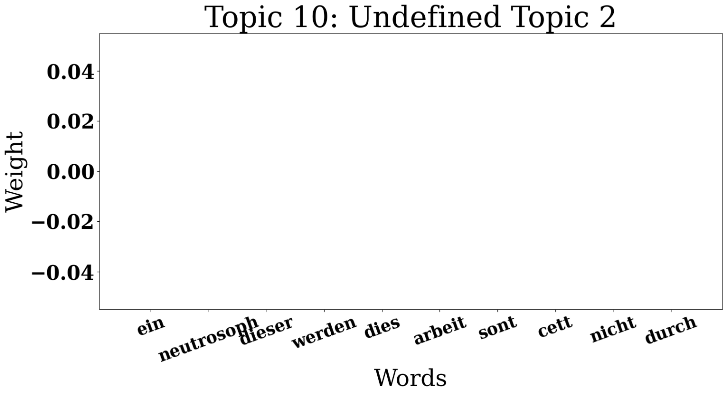
(g)


Topic 8: Geophysics
Weight
0.010
0.008
0.006
0.004
0.002
0.000
seismic
reservoir
fractur
earthquak
fault
stress
poros
invers
elast
predict
Words

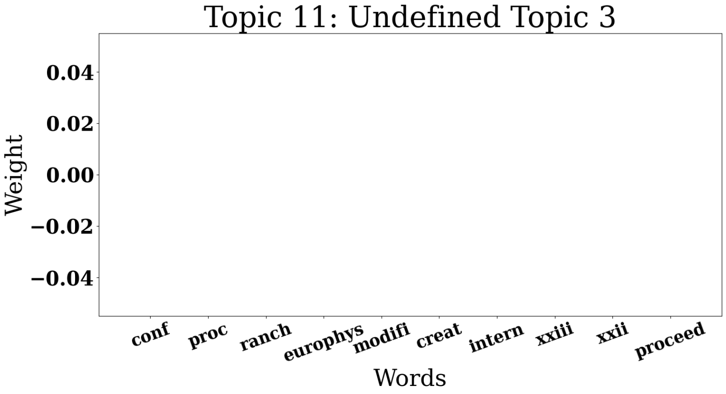
(h)


Topic 9: Undefined Topic 1
Weight
0.015
0.010
0.005
0.000
input
teachersth
associ
teach
savart
shrub
letter
haystack
wilberforc
attic
Words

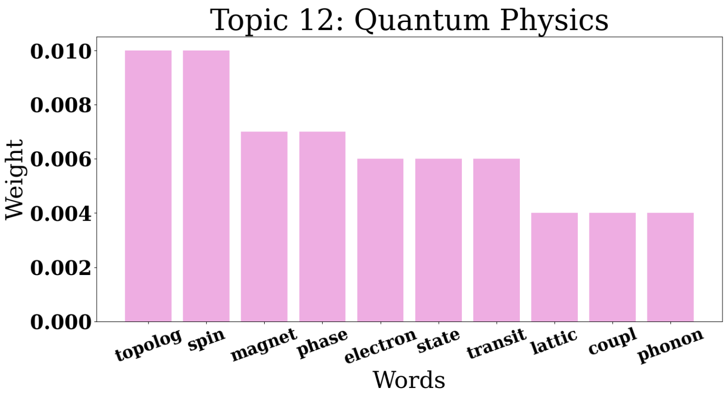
(i)


Topic 10: Undefined Topic 2
Weight
0.04
0.02
0.00
−0.02
−0.04
ein
neutrosoph
dieser
werden
dies
arbeit
sont
cett
nicht
durch
Words

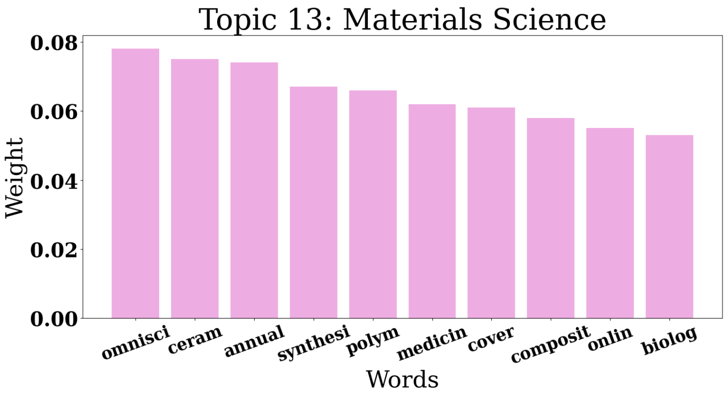
(j)


Topic 11: Undefined Topic 3
Weight
0.04
0.02
0.00
−0.02
−0.04
conf
proc
ranch
europhys
modifi
creat
intern
xxiii
xxii
proceed
Words

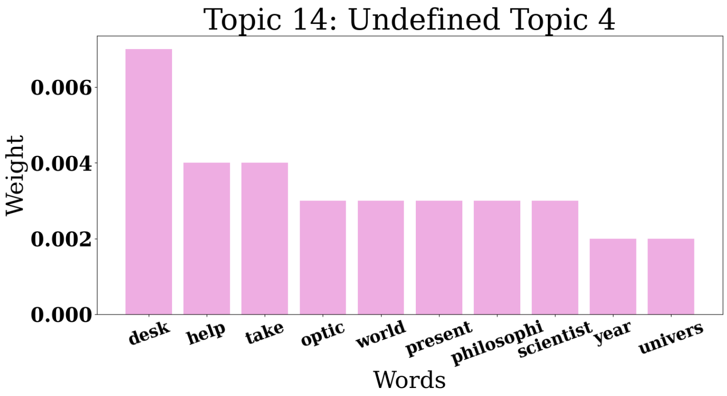
(k)


Topic 12: Quantum Physics
Weight
0.010
0.008
0.006
0.004
0.002
0.000
topolog
spin
magnet
phase
electron
state
transit
lattic
coupl
phonon
Words

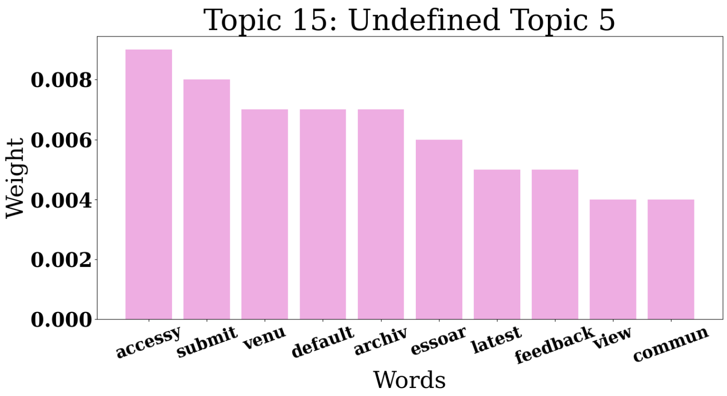
(l)


Topic 13: Materials Science
Weight
0.08
0.06
0.04
0.02
0.00
omnisci
ceram
annual
synthesi
polym
medicin
cover
composit
onlin
biolog
Words

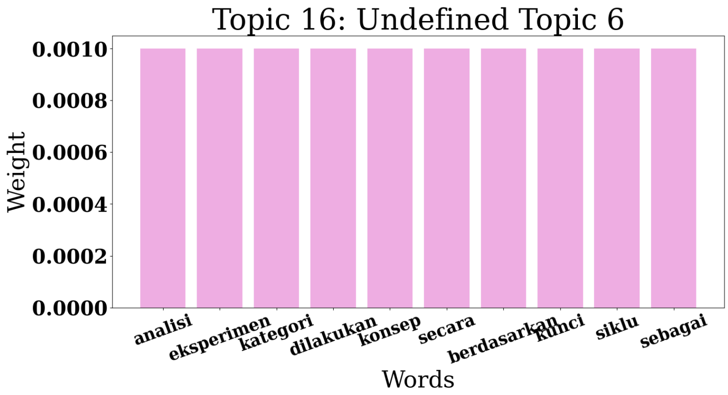
(m)


Topic 14: Undefined Topic 4
Weight
0.006
0.004
0.002
0.000
desk
help
take
optic
world
present
philosophi
scientist
year
univers
Words

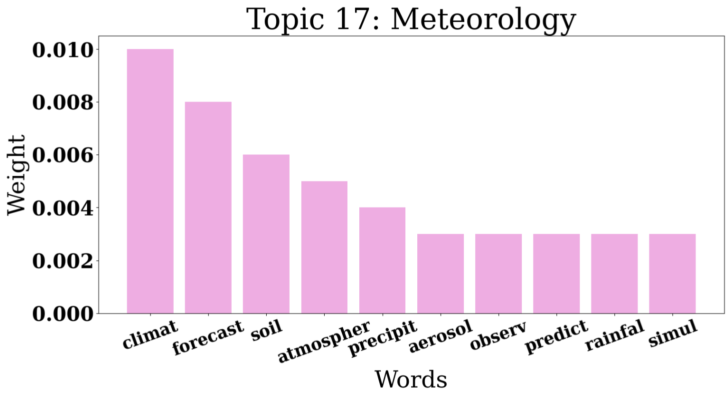
(n)


Topic 15: Undefined Topic 5
Weight
0.008
0.006
0.004
0.002
0.000
accessy
submit
venu
default
archiv
essoar
latest
feedback
view
commun
Words

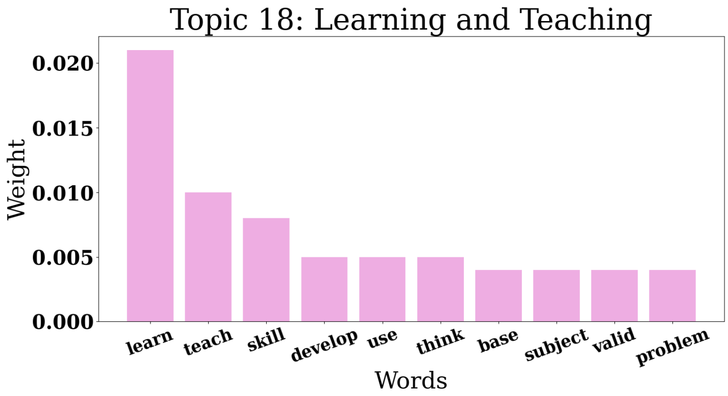
(o)


Topic 16: Undefined Topic 6
Weight
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
analisi
eksperimen
kategori
dilakukan
konsep
secara
berdasarkan
kunci
siklu
sebagai
Words

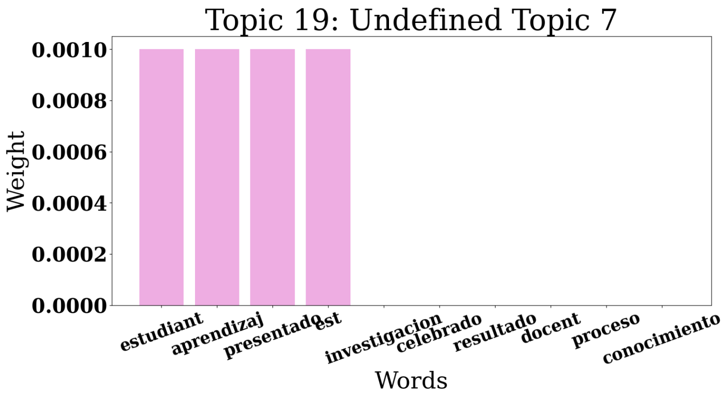
(p)


Topic 17: Meteorology
Weight
0.010
0.008
0.006
0.004
0.002
0.000
climat
forecast
soil
atmospher
precipit
aerosol
observ
predict
rainfal
simul
Words

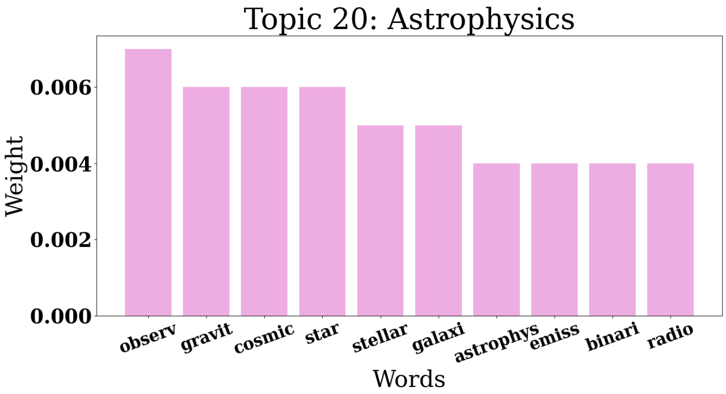
(q)


Topic 18: Learning and Teaching
Weight
0.020
0.015
0.010
0.005
0.000
learn
teach
skill
develop
use
think
base
subject
valid
problem
Words

(r)

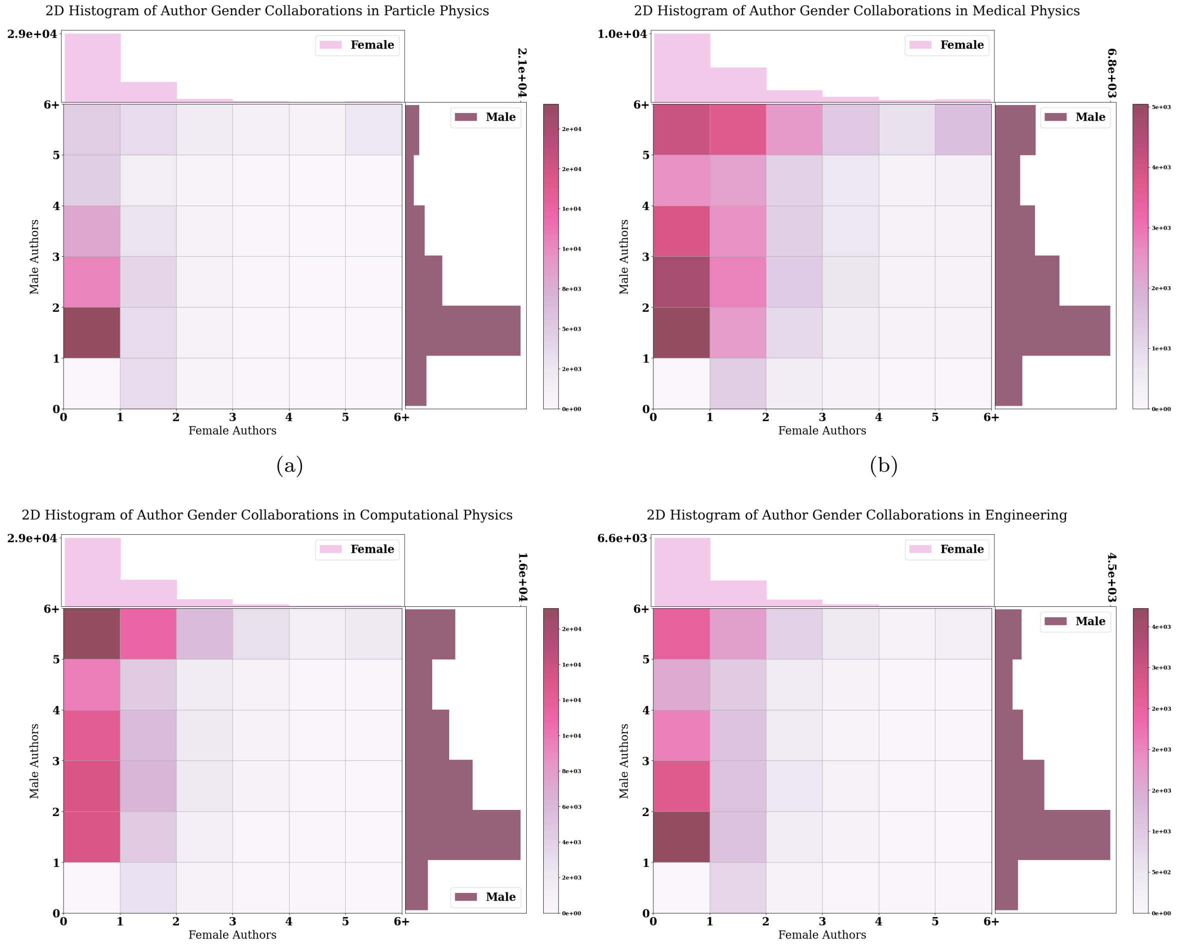

Topic 19: Undefined Topic 7
Weight
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
estudiant
aprendizaj
presentado
est
investigacion
celebrado
resultado
docent
proceso
conocimiento
Words

(s)



(t)

Figure A11: The bag-of-words distribution per Topic Index including the weight of each word within the topic.

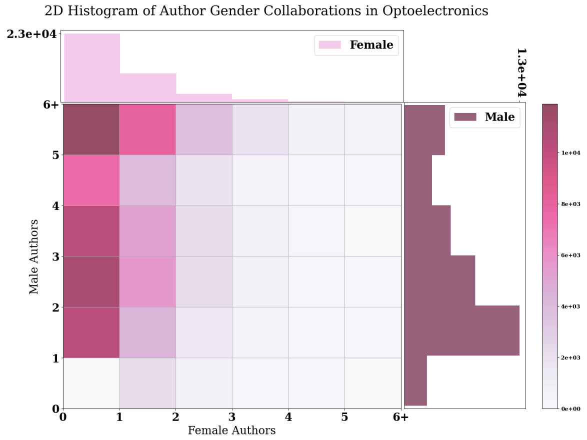


(a)

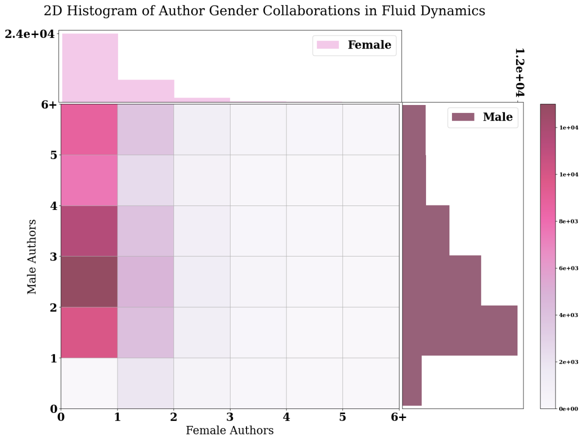
(b)

(c)

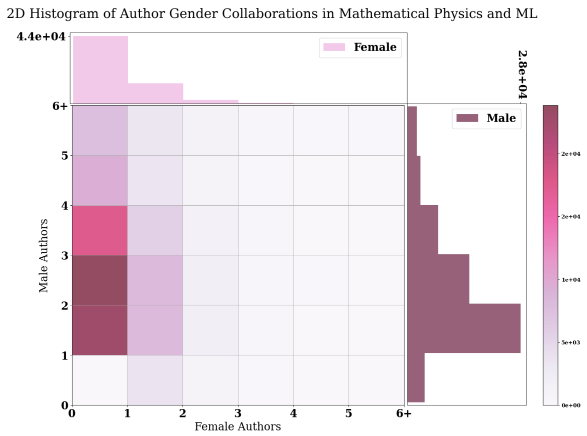
(d)


2D Histogram of Author Gender Collaborations in Optoelectronics
2.3e+04
Female
1.3e+04
Male
Male Authors
Female Authors
0 1 2 3 4 5 6+

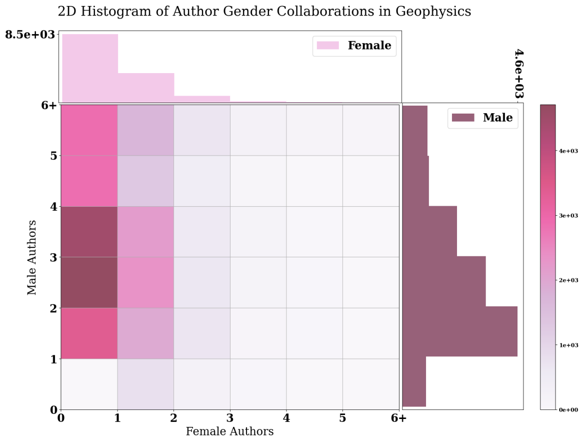
(i)


2D Histogram of Author Gender Collaborations in Fluid Dynamics
2.4e+04
Female
1.2e+04
Male
Male Authors
Female Authors
0 1 2 3 4 5 6+

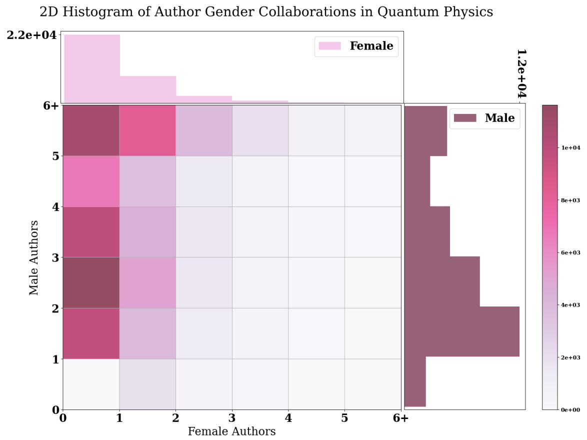
(j)


2D Histogram of Author Gender Collaborations in Mathematical Physics and ML
4.4e+04
Female
2.8e+04
Male
Male Authors
Female Authors
0 1 2 3 4 5 6+

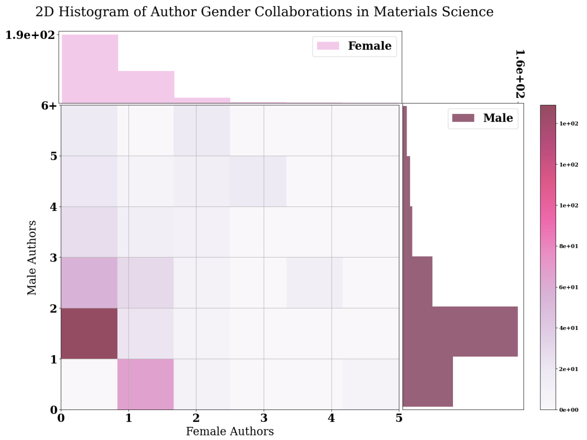
(k)


2D Histogram of Author Gender Collaborations in Geophysics
8.5e+03
Female
4.6e+03
Male
Male Authors
Female Authors
0 1 2 3 4 5 6+

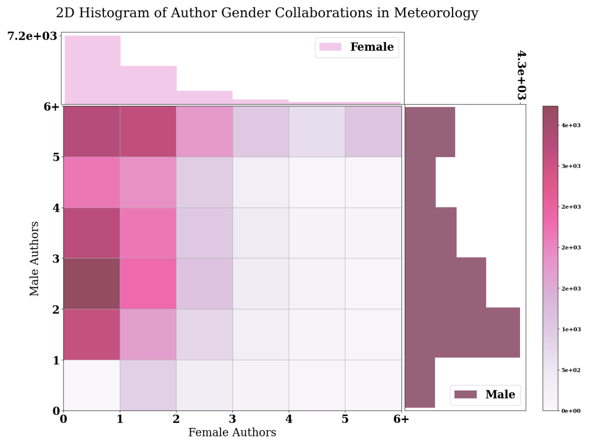
(l)


2D Histogram of Author Gender Collaborations in Quantum Physics
2.2e+04
Female
1.2e+04
Male
Male Authors
Female Authors
0 1 2 3 4 5 6+

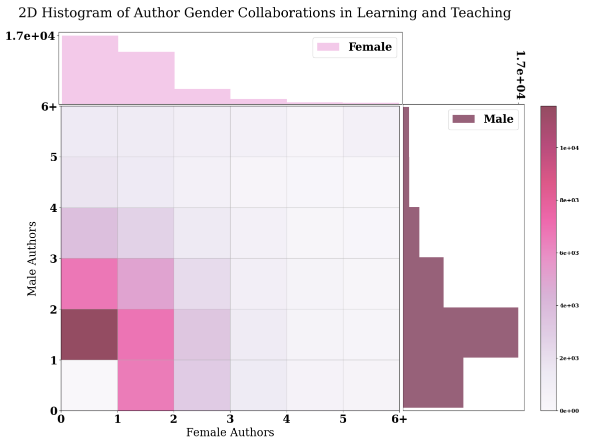
(m)


2D Histogram of Author Gender Collaborations in Materials Science
1.9e+02
Female
1.6e+02
Male
Male Authors
Female Authors
0 1 2 3 4 5

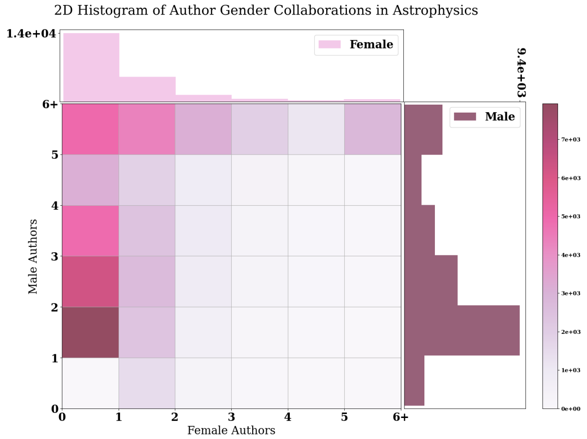
(n)


2D Histogram of Author Gender Collaborations in Meteorology
7.2e+03
Female
4.3e+03
Male
Male Authors
Female Authors
0 1 2 3 4 5 6+

(o)


2D Histogram of Author Gender Collaborations in Learning and Teaching
1.7e+04
Female
1.7e+04
Male
Male Authors
Female Authors
0 1 2 3 4 5 6+

(p)



(q)

Figure A12: 2D histogram displaying author collaborations between genders for each topic.

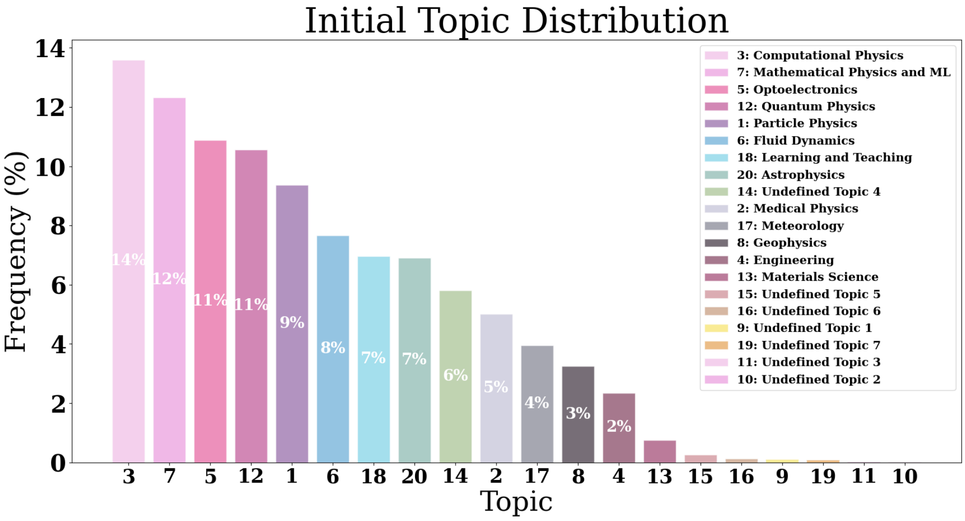


Figure A13: Initial Topic Distribution.

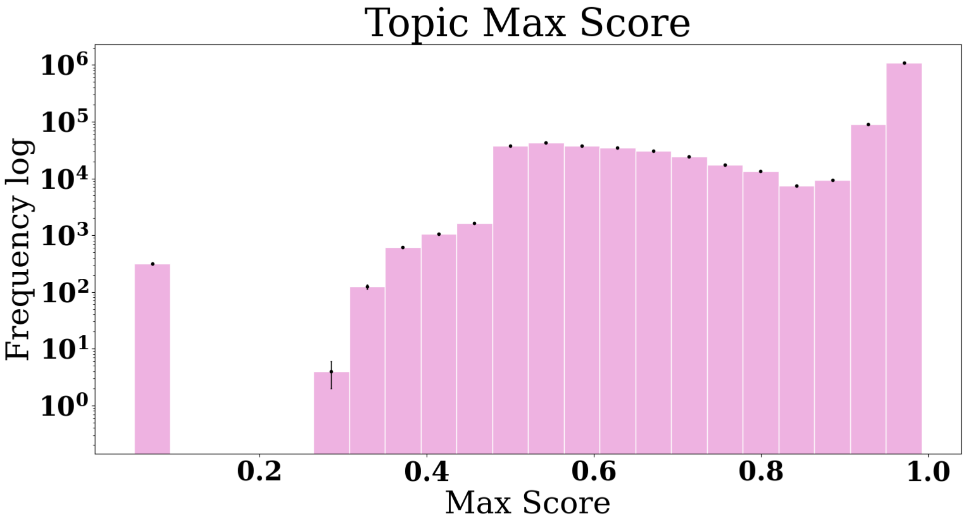
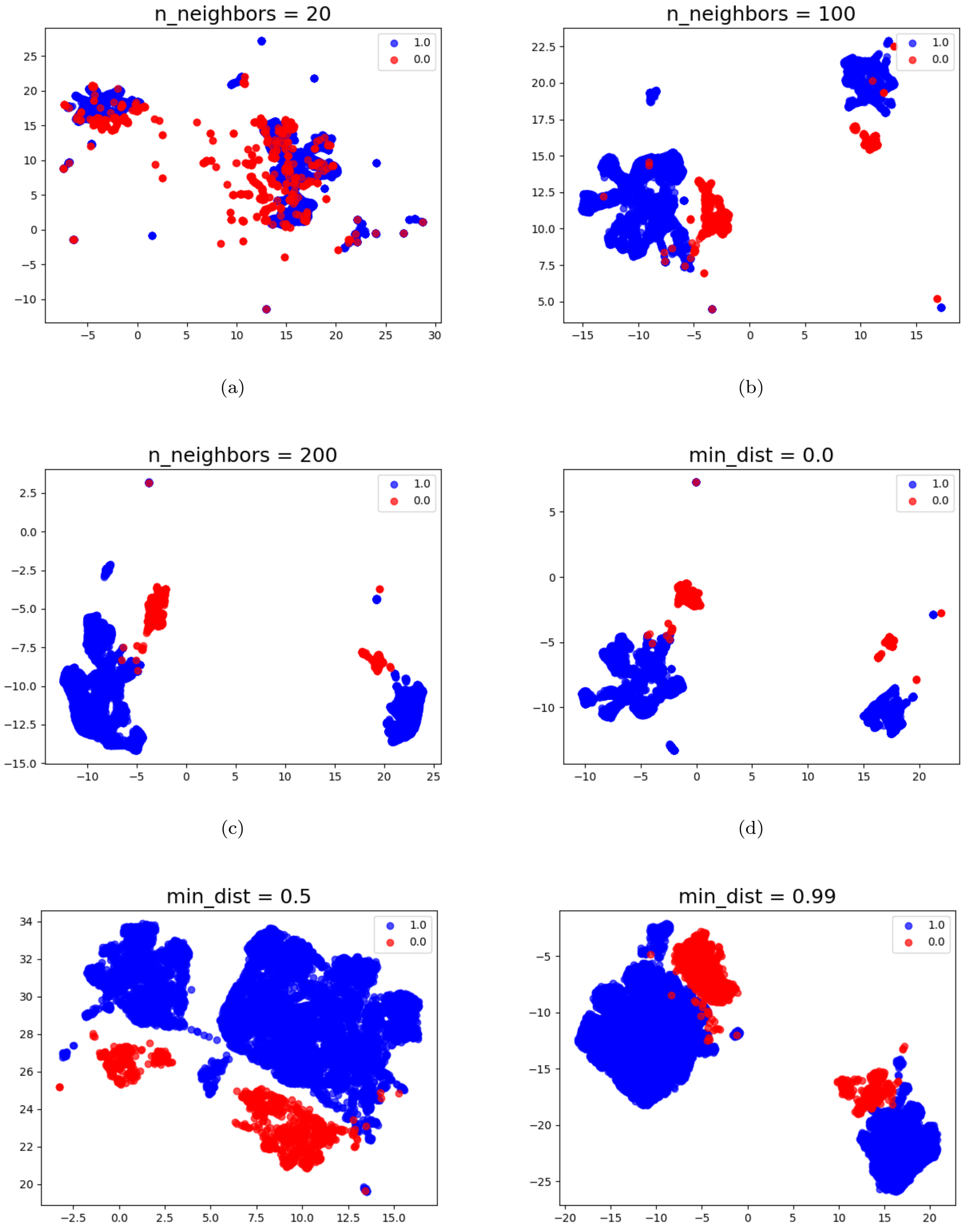


Figure A14: Initial Topic Max Score.

UMAP


n_neighbors = 20
1.0
0.0
25
20
15
10
5
0
−5
−10
−5
0
5
10
15
20
25
30
n_neighbors = 100
1.0
0.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
−15
−10
−5
0
5
10
15
n_neighbors = 200
1.0
0.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
−15.0
−10
−5
0
5
10
15
20
25
min_dist = 0.0
1.0
0.0
5
0
−5
−10
−10
−5
0
5
10
15
20
min_dist = 0.5
1.0
0.0
34
32
30
28
26
24
22
20
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
min_dist = 0.99
1.0
0.0
−5
−10
−15
−20
−25
−20
−15
−10
−5
0
5
10
15
20

(a)

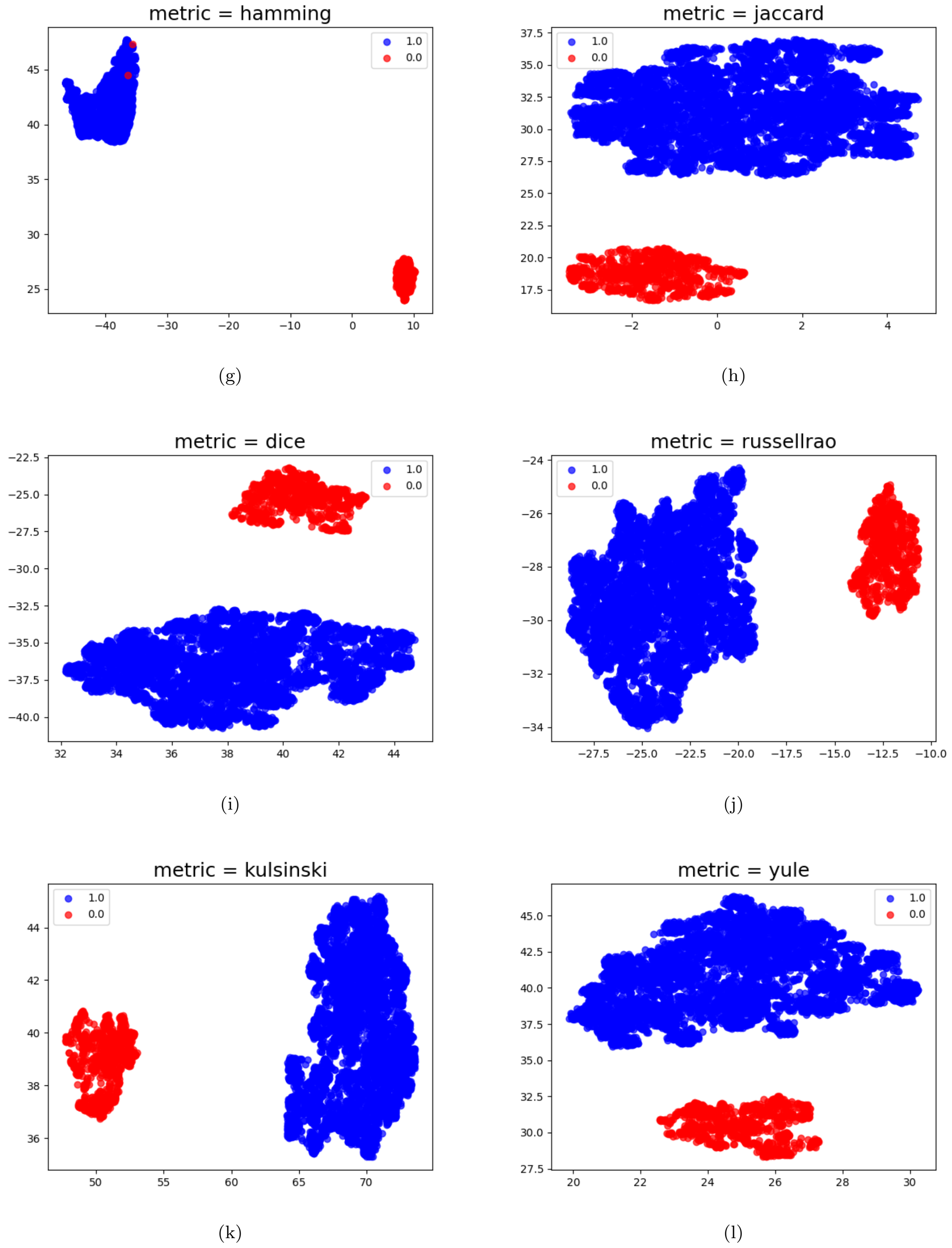
(b)

(c)

(d)

(e)

(f)



(g)

(h)

(i)

(j)

(k)

(l)

Figure A15: UMAP Hyperparameter testing outcomes for each distinct hyperparameter setting tested.

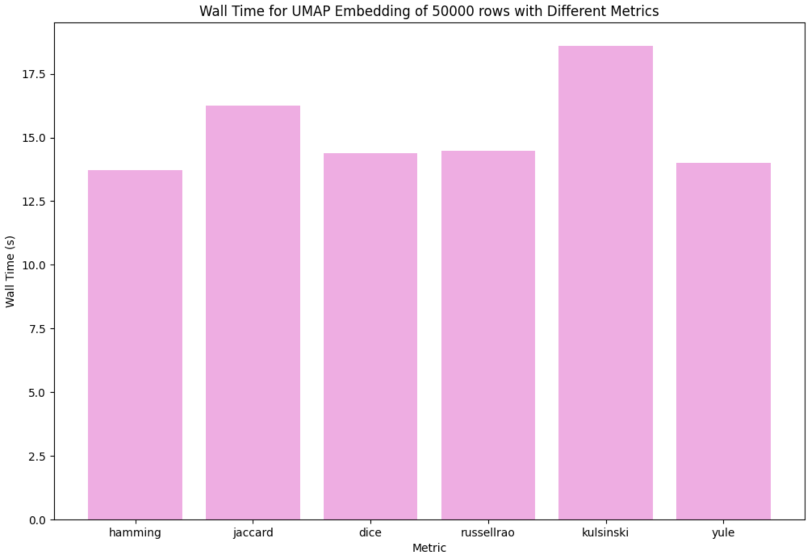


Figure A16: UMAP computational time per metric type.

LightGBM

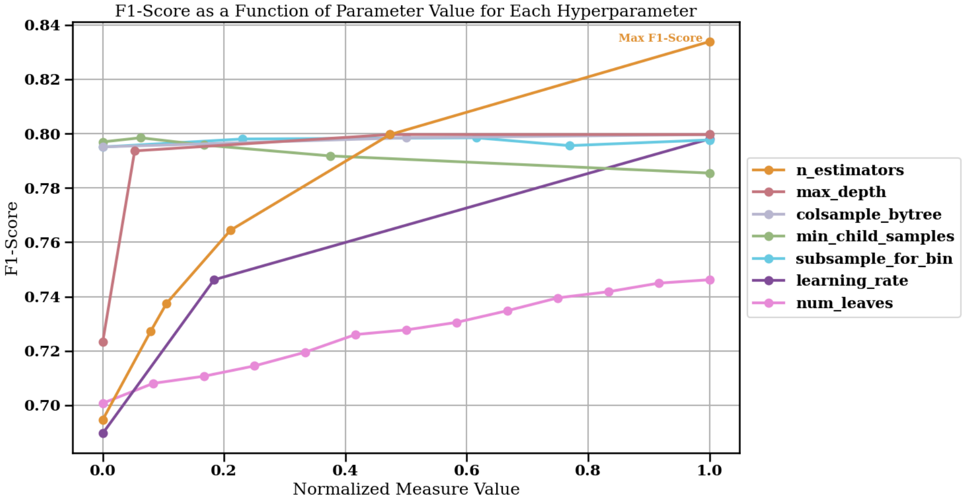


Figure A17: LightGBM Hyperparameter testing outcomes for each distinct hyperparameter setting tested.

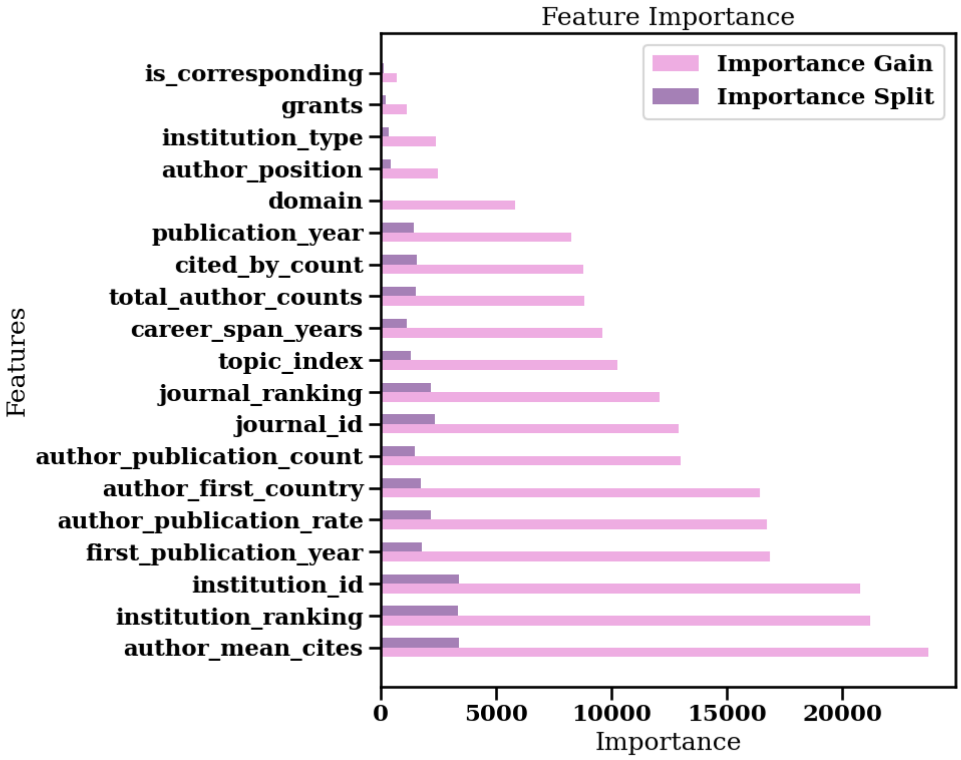


Figure A18: Feature importance between split and gain for the first run of our LightGBM model (LightGBM I)

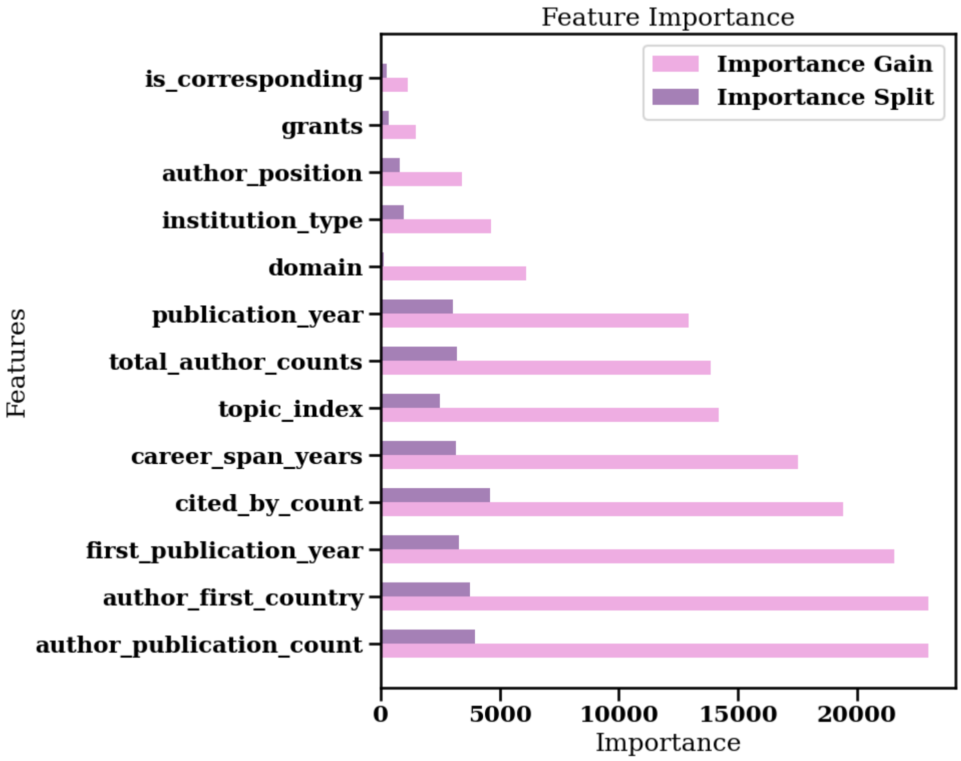


Figure A19: Feature importance between split and gain for the second run of our LightGBM model (LightGBM II)

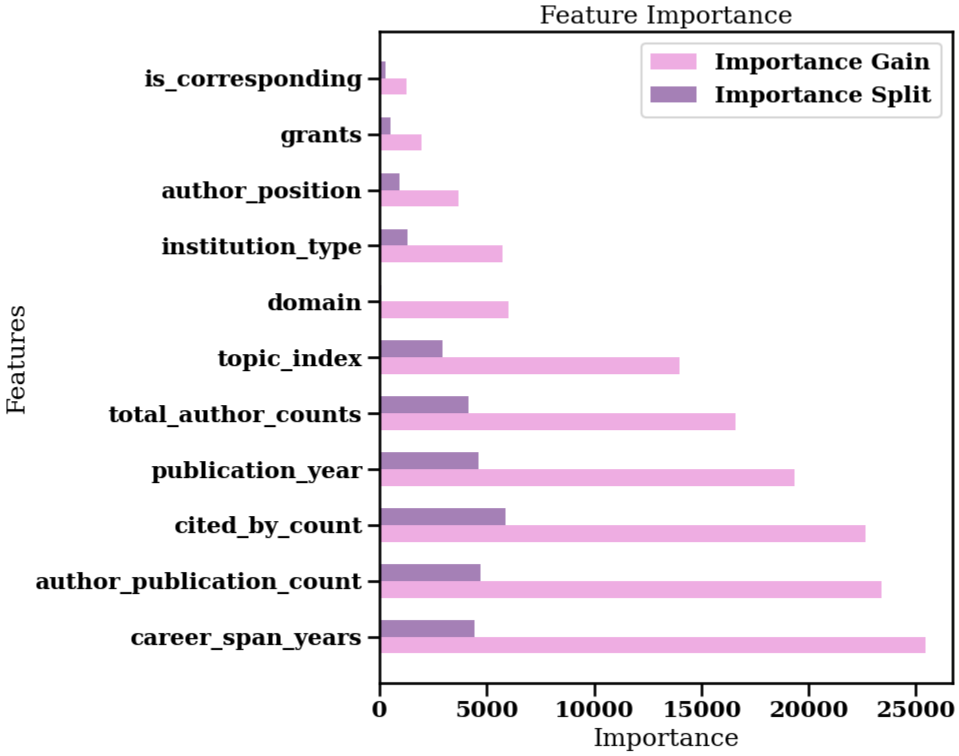


Figure A20: Feature importance between split and gain for the third run of our LightGBM model (LightGBM III)

Ranking Correlations

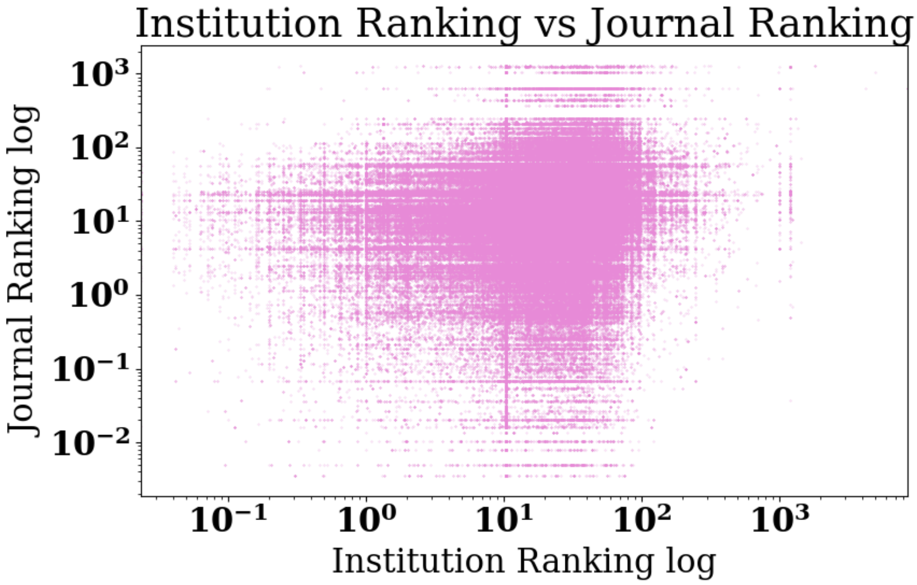


Figure A21: Institution Ranking vs Journal Ranking

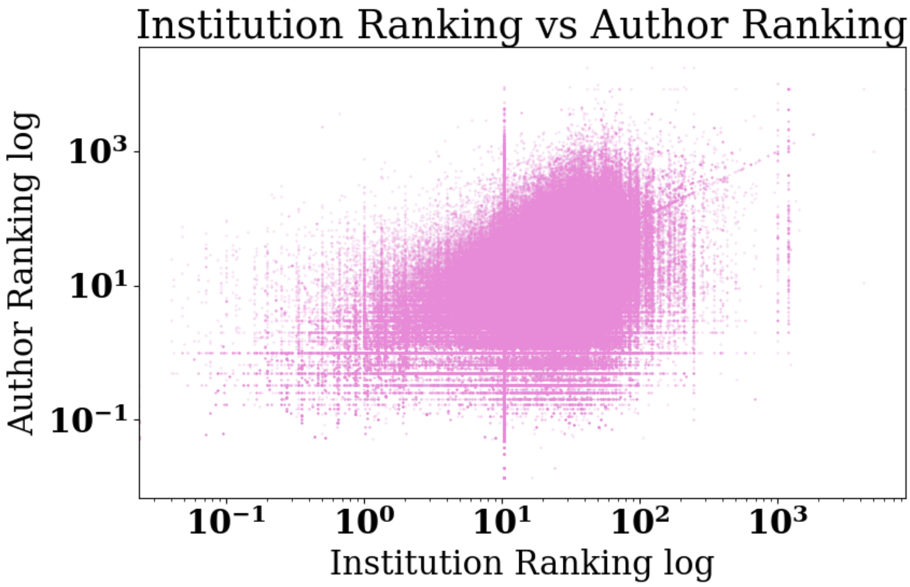


Figure A22: Institution Ranking vs Author Ranking

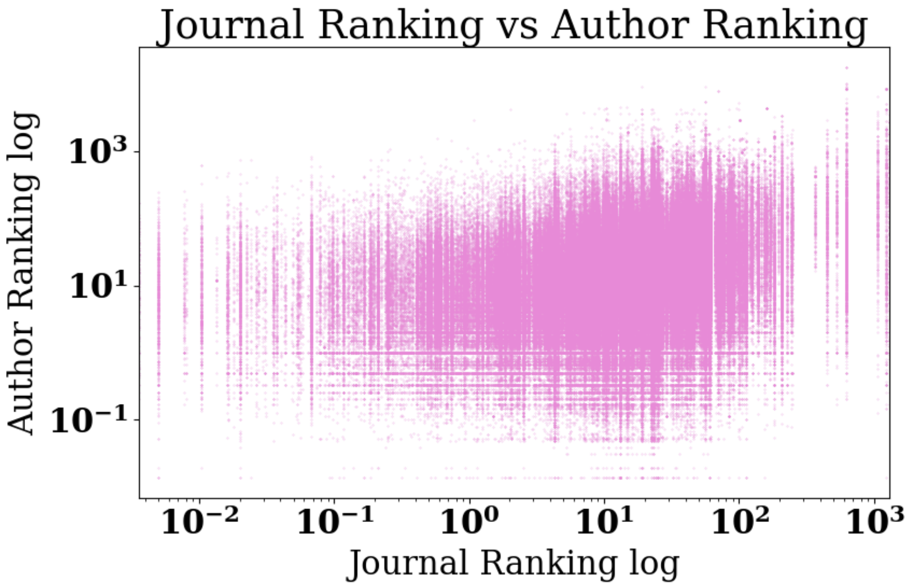
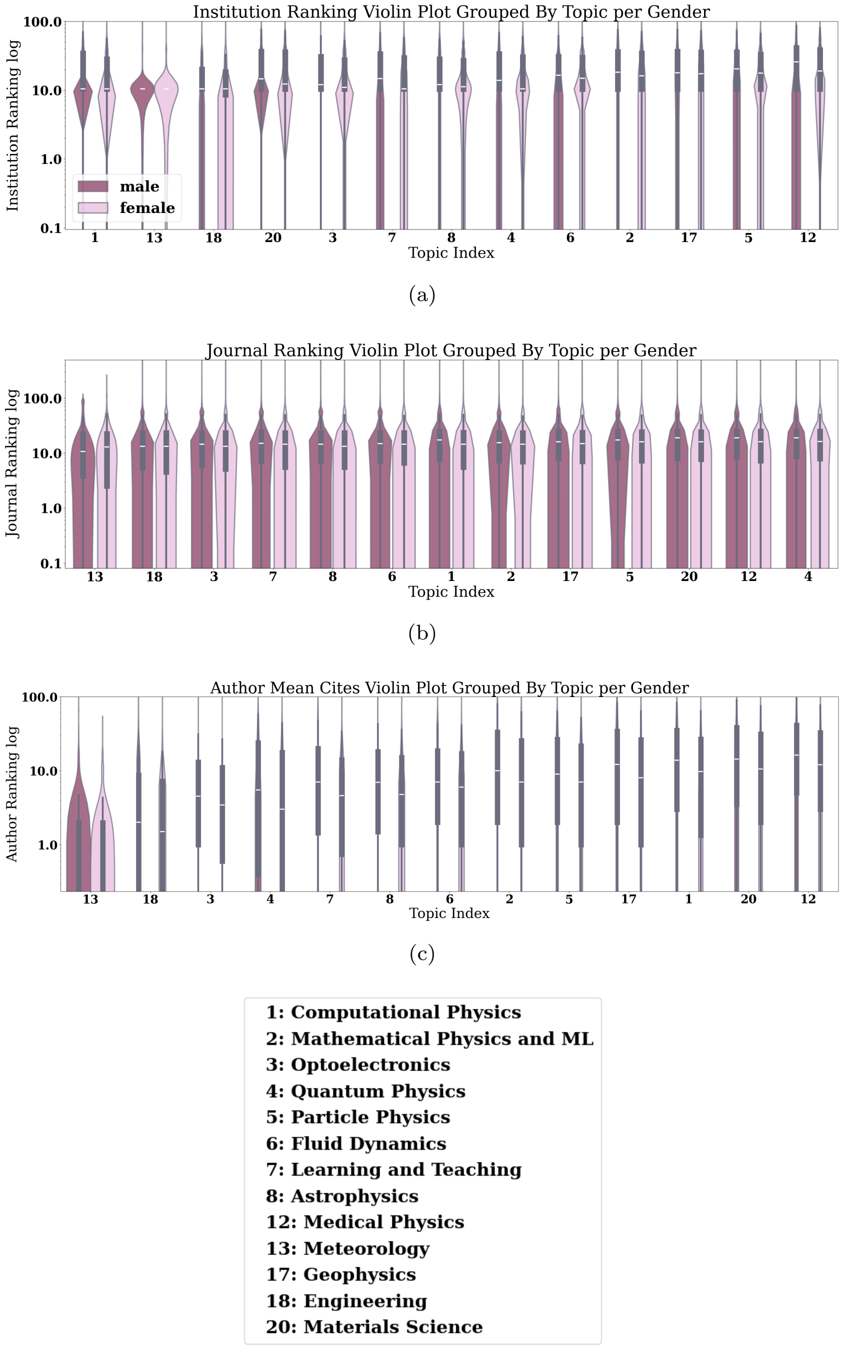


Figure A23: Journal Ranking vs Author Ranking

Topic and Ranking



(a)

(b)

(c)

(d)

Figure A24: Violin plot displaying the descriptive statistics of **a** *Institution Ranking*, **b** *Journal Ranking*, and **c** *Author Ranking* grouped by topic per *Binary Gender*. The order of appearance of topic is set such that the median ranking per topic is increasing along the x-axis. The median is marked as a white dot within each "violin". The y-axis is cut and the figure does thus not display all the outliers. The entire range of the rankings can be found in their respective histograms (Figures 34, 38, and 42). **c** displays the legend showing the corresponding *Topic Label* to the *Topic Index*.