Heliyon 10 (2024) e23781

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

What's next? Forecasting scientific research trends

Dan Ofer, Hadasah Kaufman, Michal Linial *

Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

ARTICLE INFO

Keywords:
PubMed
NLP
MeSH
Bibliometrics
Time series
Machine learning
Citation analysis

ABSTRACT

Scientific research trends and interests evolve over time. The ability to identify and forecast these trends is vital for educational institutions, practitioners, investors, and funding organizations. In this study, we predict future trends in scientific publications using heterogeneous sources, including historical publication time series from PubMed, research and review articles, pre-trained language models, and patents. We demonstrate that scientific topic popularity levels and changes (trends) can be predicted five years in advance across 40 years and 125 diverse topics, including life-science concepts, biomedical, anatomy, and other science, technology, and engineering topics. Preceding publications and future patents are leading indicators for emerging scientific topics. We find the ratio of reviews to original research articles informative for identifying increasing or declining topics, with declining topics having an excess of reviews. We find that language models provide improved insights and predictions into temporal dynamics. In temporal validation, our models substantially outperform the historical baseline. Our findings suggest that similar dynamics apply across other scientific and engineering research topics. We present SciTrends, a user-friendly webtool for predicting future publication trends for any topic covered in PubMed.

1. Introduction

The progress of science relies on the discovery and dissemination of new knowledge through scientific publications. In an era marked by the rapid evolution of scientific knowledge, the ability to forecast trends is paramount for strategic decision-making for researchers, policymakers, and industries. Researchers can proactively identify emerging fields, guiding their own studies and collaborations, while policymakers can allocate resources more effectively in response to evolving scientific demands. Furthermore, industries can harness this technology to stay ahead of the curve, adapting their products and services to align with the latest advancements in science. Groundbreaking technologies such as 'quantum computing' were started as niche fields, and the prediction of science trends can direct investigators and investors to such promising fields. On the other hand, researchers, investors, or countries, will be able to avoid investing in fields that are likely to decline in their public relevance. For example, investing in the field of 'fossil fuel consumption' over the years will be less relevant in the shadow of the growing global awareness and efforts to reduce fossil fuel usage. Predicting such trends by machine learning models can be a pioneering step towards enhancing our understanding of the ever-changing scientific landscape.

The number of scientific papers published has been accelerating for at least four decades, and citation and annotation behaviors

* Corresponding author. Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra, Jerusalem, 91904, Israel.

E-mail addresses: dan.ofer@mail.huji.ac.il (D. Ofer), hadasah.kaufman@mail.huji.ac.il (H. Kaufman), michall@cc.huji.ac.il (M. Linial).

https://doi.org/10.1016/j.heliyon.2023.e23781
Received 4 May 2023; Received in revised form 6 December 2023; Accepted 13 December 2023
Available online 15 December 2023
2405-8440/© 2023 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

have changed with it [1]. Changes might be attributed to the continuous increase in research institutes and researchers, the increasing impact of publications for funding agencies, and academic careers [2,3]. In addition, acceleration in publication may reflect better automated data indexing (e.g., Science Citation Index (SCI) [4]), and the establishment of a keyword annotation scheme (e.g., the medical subject headings system, MeSH [5]). Other changes within the last four decades include the expansion of open access policies, increased research originating in industry, and the ongoing increase in the total number of researchers and expected research productivity output in many fields. While presenting the current state can be based on a historical view of analytical methods, predicting future trends is far more challenging [6,7].

Some fields, such as methods in computer science and biotechnology, have fast dynamics due to impactful technological breakthroughs (e.g., machine learning, CRISPR gene editing) [8]. In other domains, such as medicine, topics are often less dynamic over time (e.g., cancer). A topic's popularity is also influenced by social factors bias [9]. Statistical machine learning models are commonly used in attempting to forecast future events and trends [10,11].

Traditionally, most research on scientific publication behavior focused on descriptive analyses of past trends [12], or citation networks [13,14]. Such studies aimed to detect existing trends, such as defining topics with increasing popularity. Alternatively, predictive or prescriptive analytics approaches aim to predict future behaviors in order to answer questions such as which topics will become more (or less) popular or what can make a topic popular [15–17]. In this study, we aim to predict the future behavior of scientific topics. Note that "popular" is not referring to the number of researchers involved or to an absolute measure of the size of the scientific active community. Here, popularity refers to the relative fraction of all publications focused on a specific topic. Furthermore, changes in popularity address the time-dependent relative popularity topic compared to its own past history, as directly comparing popularity between different domains is problematic without context due to widely varying base levels of popularity.

A limited number of studies have investigated time-series data for forecasting future topic citations. Noorden et al. analyzed the absolute number of citations for a specific paper [18]. Other studies focused mainly on narrow domains (e.g., predicting domain-specific conferences [19,20]). Tattershall et al. looked at binary trends within 5 years in computer science terms without continuous fine-grained prediction of the target or exogenous variables [21]. Studies analyzing the relationships between patents and research publications [22,23] suggested that these types of publications carry mutual information. In the field of virology and human health, forecasting virus outbreaks is of utmost importance, as practical measures in a limited time frame are needed (e.g., designed vaccination, animal eradication, population isolation measures) [24]. Based on 16 studies from literature (PubMed and Google Scholar) the potential of forecasting influenza outbreaks at different scales globally was assessed. It showed promise in capturing the outbreak measures, while raising the need for further evaluation in real-time predictions [25].

Our goal is to predict the future popularity of diverse scientific topics, with an emphasis on life science, experimental science, and biomedical domains. We developed a methodology that accounts for the overall increase in the number of publications over time. We discuss novel exogenous factors, such as patents and per-domain publication trends, that can indicate if and how a field's popularity is going to change. We show meaningful results when predicting topic popularity five years into the future and discuss the potential impact of such predictions and predictive insights. We further present a user-friendly webtool for predicting future publication trends for 1–6 years for any topic covered in PubMed.

2. Methods

2.1. Data compilation

We constructed a diverse list of topics focused mostly on life science. We included neuroanatomical regions, experimental methods, and emerging biomedical technology. We also included domain-expert selections for science, technology, and engineering research topics. For example, ultrasound was selected as a technology with emerging use in broad medical diagnostics. Topics such as opioids had clear past historical and funding trends. Other topics were derived from Wikipedia's biomedical "emerging technologies" page. All together, we present 125 topics featured in PubMed publications (Supplementary Table S1). The counts of topical publications are based on PubMed. The PubMed database contains over 35 million published scientific works in biomedicine, life sciences, and related fields.

2.2. Normalized publication measurement – "popularity"

The total quantity of publications is increasing over time. For example, the NIH's Medline database recorded 274K citeable scientific works in 1979 but 774K in 2021 (www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html). This escalation in total publications has the potential to skew the perception of a topic's popularity if measured solely by the raw number of publications, as otherwise most topics will have more publications over time due to the confounding background trend over time. This necessitates a methodology that considers and overcomes naïve quantitative comparisons.

We adapted a measure of "popularity" that normalizes per-topic publication counts relative to the total annual output. Specifically, per year, the popularity of a given topic is calculated by dividing the number of that topic's publications by the total number of citeable, indexed publications in PubMed in the corresponding year.

We multiplied the popularity score by 100,000 in all results and figures for visual ease. Historical PubMed topic popularity was extracted using PubMed by Year (https://esperr.github.io/pubmed-by-year). Search queries are parsed by NCBI's automatic term mapping algorithm, which supports for term and acronym expansion. This normalization helps control for the overall increase in publications over time and is similar to the approach used in other works on trend and popularity analysis, such as Google Trends, where popularity signifies the proportionate presence of a topic within a total volume rather than absolute quantity, thereby offering a more accurate measure of prominence over time.

2.3. Features

The most recent historical ("Lagged") value was extracted for each topic, as well as for different transformations of its past values (including 1st, 2d order differenced and percent change values relative to the preceding year). Additional derived candidate features were explored using the SparkBeyond Automated Machine Learning framework (as used in Refs. [10,26]) and included different lags and lagged interaction features (e.g., the difference and ratio of review to research papers for a topic), aggregated time-window-based features (e.g., the historical average popularity between 5 and 10 years beforehand), as well as differencing and similar transformations of each the remaining exogenous time-series variable (see below).

In addition to the time-series features above, based on each topic popularity, we added exogenous static and dynamic features based on other variables. For each topic, we calculated the relative ratio and difference in popularity due to review and non-review (i.e., original research publication) articles. The absolute quantity of publications (rather than the per-year normalized frequency) per topic was estimated by multiplying by the number of Medline publications that year. The total fraction of US publications out of all Medline-indexed publications was added per year. The number of patents per topic, per year, was acquired by searching the U.S. Patent and Trademark Office's PatentsView database (patentsview.org/download) for each topic, using the USPTO search API (https://github.com/ddofer/Trends/blob/main/patentsView_api_req.ipynb. The topic-relative fraction of all patents that year was also calculated. The first occurrence date of each topic (starting from 1946) was included, as was the first "valid" date (defined as an occurrence with at least 4 previous occurrences in the preceding 5 years) and the time elapsed between those dates, as well as from the prediction date. We experimented with adding the popularity of each topic's associated major MeSH term. We found that this did not improve the models, so it was removed from the final model.

2.4. Statistical modeling

Features were derived from the above inputs with a forecast horizon of 5 years (i.e., all time-dependent features were from at least 5 years before the target time of prediction), and the target is popularity 5 years in the future.

Machine learning regression models were trained using the CatBoost [27] and Scikit-learn libraries [28] with default hyperparameters. The pretrained deep language model used the sentence-transformers package [29]. The linear model is a Scikit-learn RidgeRegressionCV model. Boosting Tree is a CatBoost regression tree model with target encoding of the topic as a categorical feature. Tree and embedding augmented is the CatBoost model with the topic's name embedded as additional features, extracted using the deep learning all-MiniLM-L12-v2 model [30]. Other approaches, including additional scikit-learn models and pure deep learning or statistical forecasting models, were evaluated but had significantly worse results and inferior stability, so they were not used (not shown). Boosting tree models are interpretable, fast, relatively robust, and performant in most predictive tasks and were thus used as the representative model [31]. All models used do not explicitly account for non-stationary targets, which is an issue due to the increase in average popularity over time across targets, however they can learn it from corollary, dynamic features such as the year, or the lagging values.

As the last four decades better represent modern trends in scientific research, we limited the training data to the years 1979–2019. These years cover the historical USPTO PatentsView database. We do use prior historical data when generating features and conducting retrospective analyses. For modeling, we defined for each topic the relevant starting point as the first year in which it had citations in at least 4 of the past 5 years and at least 5 valid occurrences in the 1979–2019 period. This threshold was used to overcome observed errors in PubMed, with some topics appearing just a few times over decades, apparently due to erroneous annotations.

Model performance was evaluated using step-forward temporal cross validation, implemented in scikit-learn, over all topic time series simultaneously. The test set was defined using scikit-learn's temporal cross-validation protocol with 30 splits.

A humanized version summarizing the distribution and summary statistics of the topics is provided as Supplementary Table S2. The table includes each topics popularity and descriptive features over time. An extended table with features per topic, per year, as used in training the models and evaluation is provided as Supplementary Table S3 (total 4176 rows, ~400 columns).

2.5. Application and data implementation

A stand-alone implementation is provided in the code repository (https://github.com/ddofer/Trends). Additionally, we developed SciTrends, a GRADIO web application, for viewing predicted normalized PubMed occurrences of any topic, and it is automatically updated with new data. The application is available at: https://huggingface.co/spaces/hadasak/SciTrends. The website code for acquiring data, extracting features, training models and loading the provided pretrained models, as a webserver is provided in https://github.com/HadasaK1/SciTrends/tree/main/SciTrends.

Data and implementation details are provided in Supplementary Table S2 The table includes 4684 lines related to the 125 selected topics. Each line includes quantified information with rich, dynamic data. The code repository is available at https://github.com/ddofer/Trends. We have used USPTO search API. The code is provided in the repository. (https://github.com/ddofer/Trends/blob/main/patentsView_api_req.ipynb). We have developed a web application called SciTrends that allows the user to view the normalized PubMed occurrences by year for the topic of interest for the years 2023–2028. In addition, we provide an unlimited searching mode for any term in PubMed. For all these instances, we provide future prediction trend for 1–6-years. The application is available at:

https://huggingface.co/spaces/hadasak/SciTrends.

3. Results

3.1. *The dynamics of scientific terms popularity levels*

Fig. 1 presents an overview of 125 diverse topics and their breakdown into broad themes. Detailed information on the topics and their associated scientific domains is available in Supplementary Table S1. Obviously, this list is not exhaustive, although the collection of topics represents a wide range of emerging and established scientific topics. In this study, we analyze these topics as a showcase.

A sample of topics and their trends over time are illustrated in Fig. 2A. We follow topics' popularity (see Methods) over the past 45 years and show that topics gain, lose, and sometimes regain popularity over time. For example, while the popularity of *opioids* is quite stable, the differences over the years for *stem cells* and *neuropeptides* show very different levels of popularity and dynamics (Fig. 2A). There was a clear change in popularity in 1990 and 2010 for *neuropeptides* and *stem cells*, respectively.

Fig. 2B shows the dynamics for 62 years (1960–2022). Some topics display complex dynamics. For example, *RT-PCR*, which was only introduced in the early 90s, exhibits a sharp increase in popularity within 5 years (2005–2010) and a similar sharp decline that only stabilized in recent years. A similar trend was associated with *restriction enzymes*, which were the force behind molecular biology in the early days (1985–1995) and were then replaced by more simple and versatile technologies, including library preparation, CRISPR, and such. On the other hand, topics like *species conservation* and *climate change* monotonically increased in popularity, albeit only after two decades for the latter. The topic *single cell* shows a doubling in popularity that occurred in 1974, increasing popularity for two decades, and then remaining high but stable for the last 25 years.

We observed that the (normalized) mean popularity of our topics increased over time, despite our normalization. The overall change in popularity over time for all discussed topics is shown in Supplementary Fig. S1.

3.2. *Correlations with the popularity of scientific topics*

Since topic popularity includes reviews, the popularity of just review articles on a topic is significantly correlated with its overall popularity. Specifically, the total popularity of all publications (combining review and non-review articles) in that same year shows a high Pearson correlation coefficient (r = 0.87) compared to 5 years before (r = 0.85), as is expected. We list the ratios of review articles to research works for each topic and year (Supplementary Table S2). Note that it varies greatly by topic and the ongoing dynamics of a topic's popularity (Supplementary Table S2). Overall, a popular topic will have a higher fraction of reviews, while a rapidly growing topic may have fewer review articles relative to original research works. We conclude that the relation between the feature and its

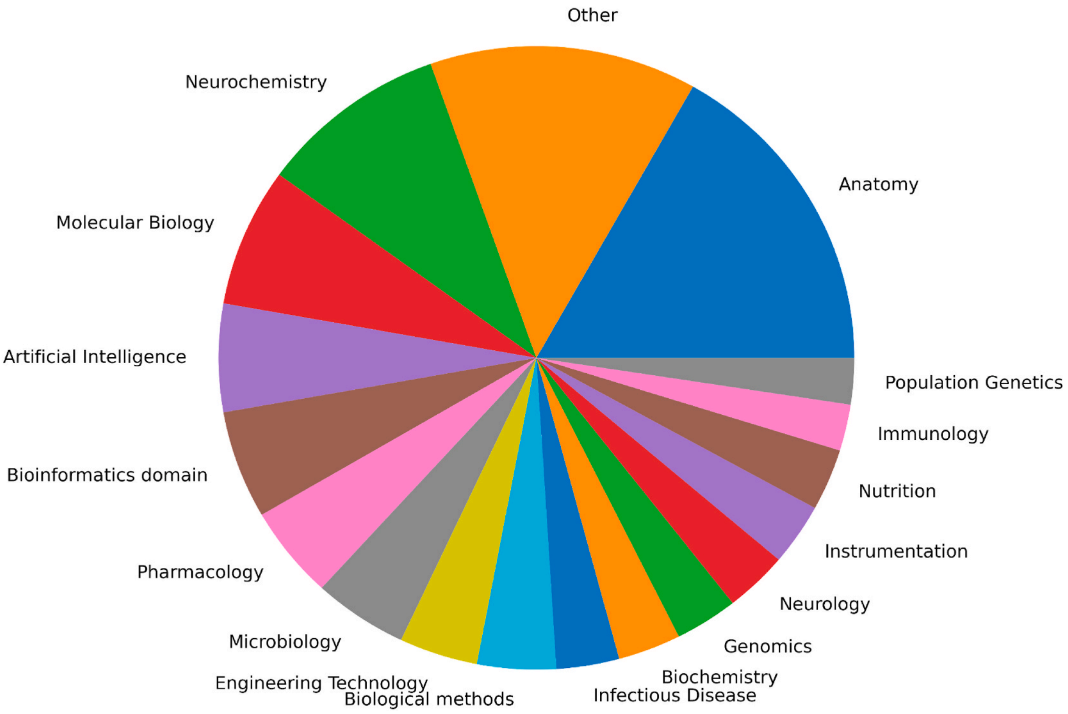

Fig. 1. Topics by domain. Overview of the 125 topics clustered by their association with high-level field domains. A full list is available in Supplementary Table S1.

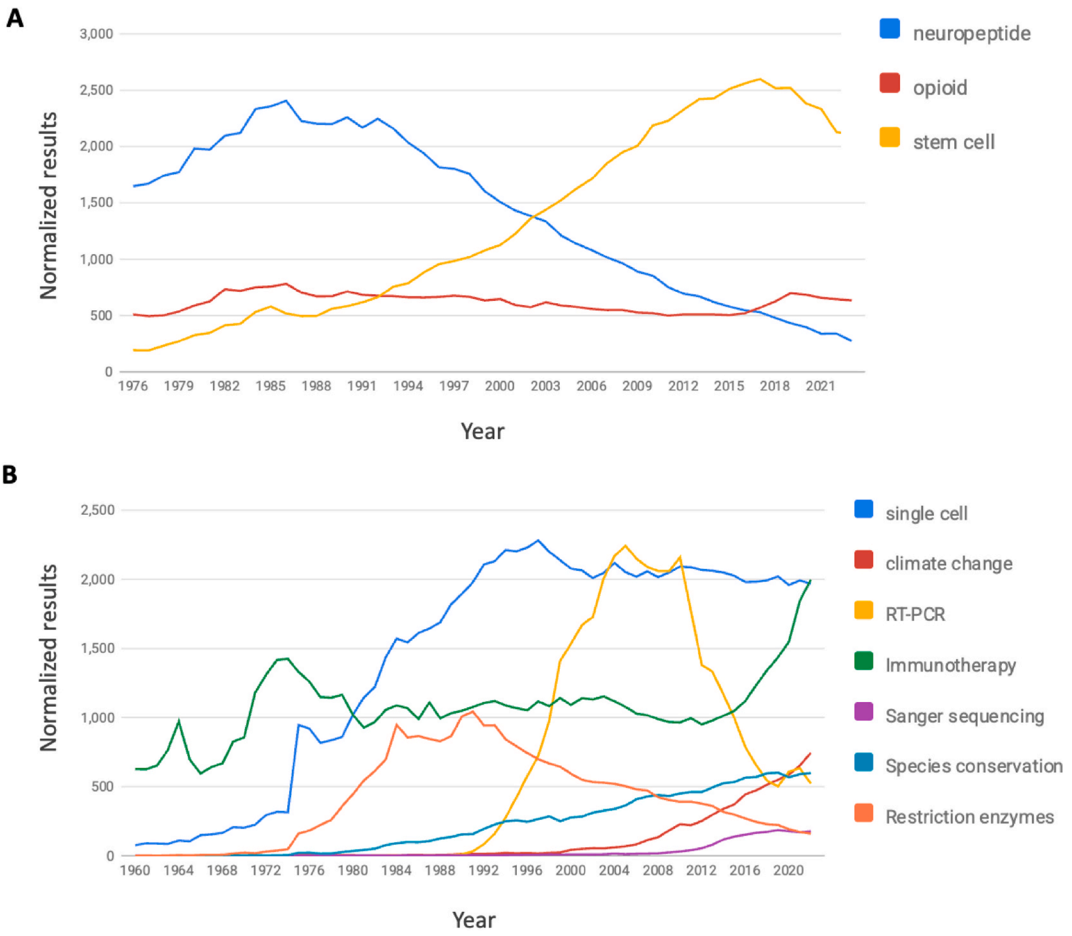

Fig. 2. Different levels of popularity in PubMed for the indicated topics. The popularity of each topic (y-axis) is normalized per 100,000 citations out of all citations that year. **(A)** Sample of topics in the years 1976–2022. **(B)** Changes in the levels of popularity as extracted from PubMed covering the years 1960–2022. Source: PubMed by Year (see Methods).

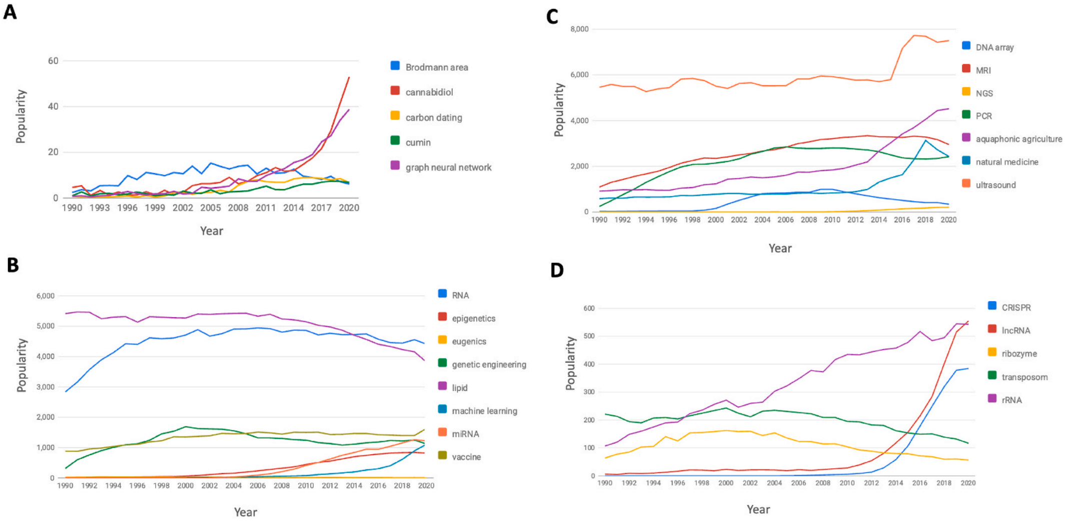

Fig. 3. Examples from the dataset showing different levels of popularity and dynamics over three decades (1990–2020) for selected representative topics. **A-D** cover a wide range in popularity, from 50 to 8000. **(D)** Subtopics in the molecular aspects of RNAs. Note that popularity is scaled by 100,000. Image source: PubMed by year (see Methods).

trend (i.e., changes in its popularity in the future) is different from the static popularity at the same point in time. A popular topic may have many review articles, but a growing topic may have relatively fewer reviews.

We further anticipated that patents would reflect the popularity of the topic. Indeed, overall popularity is correlated with the number of patents in that year with a Pearson rank correlation ($r = 0.37$) and 5 years before ($r = 0.40$). This suggests that patents may be a leading indicator for scientific publication trends, with patents preceding "valuable" research. This observation is in line with the regulation of patent applications, where confidentiality is requested to prevent its public disclosure in scientific publications before its acceptance.

We observe that topic popularity over time is highly correlated with their own past values, e.g., from 5 years beforehand, as we might expect ($r = 0.953$, $R^2 = 0.969$).

Fig. 3 displays a diverse set of topics that we analyzed within a narrower timeframe (1990–2020). We show that popularity can change sharply with 4–5 fold increases within 5 years (e.g., *cannabidiol,* Fig. 3A). Many such cases represent breakouts for topics with relatively low (<50 per 100K normalized citations) previous popularity. Popular topics may also exhibit high growth and overall popularity for decades (e.g., 8000 for *ultrasound*, Fig. 3B). Classical topics like *vaccines* and *lipids* display rather stable dynamics for decades (Fig. 3C). We emphasize that some topics are quite general (e.g., *DNA*, *RNA*) and represent many subtopics that often exhibit distinctive popularity dynamics. We illustrate it by partitioning of the term *RNA* with respect to more specific related subtopics (Fig. 3D). Although RNA in general shows a stable trend (Fig. 3C), *CRISPR* (based on gRNA) and long noncoding RNA *(lncRNA)* have seen a 10-fold increase in popularity within 10 years, while transposons and ribozymes reached their maximum popularity in 2000. An interesting example refers to the 5-fold monotonic increase in rRNA (Fig. 1D). It is likely reflecting the emerging fields of microbiome and metagenomics, in which rRNA is used for sample characterization and microbial species identification. All these informative changes are deemed necessary by considering RNA as a general term. The normtable salized historical trends by years based on the PubMed database for each of the PubMed topics are available in the SciTrends application (see Methods).

3.3. Natural language as a clue to topics' dynamics

As shown in Fig. 3, the topics we discuss cover not just different scientific domains but also vary in how general or specific they are, as well as representing different concepts. Specifically, there are topics that concern methods such as *NGS* (next generation sequencing) or *PCR* (polymerase chain reaction), while others cover biological fields that are precisely defined, such as anatomical regions. We expect different fields and entities to show different behaviors and temporal dynamics. Furthermore, topics greatly vary by the size and nature of their community, e.g., medical clinicians versus ethical researchers.

We would like to incorporate relevant information into our models to enhance the potential for learning latent dynamics. For example, we anticipate that technologies might become obsolete or be replaced by modern ones (e.g., the methodologies for transgenic mice or creating cell lines). Nevertheless, terms for well-defined anatomical entities such as the *hippocampus* are unlikely to be superseded, even if interest in them varies, and the topic will remain despite being subjected to new research technology. For example, *fMRI* (functional magnetic resonance imaging) is a leading technology associated with neuroscience and cognitive psychology that exhibits massive growth in popularity thanks to its ability to allow psychology experiments with live subjects in real time. But a topic such as genetic engineering will most likely not match the dynamics of the term eugenics. The latter has a negative historical context and thus is not expected to correlate well with genetic engineering, despite their shared roots and semantic similarity.

We used a pretrained deep learning language model, all-MiniLM-L12-v2 [30] to extract a quantitative representation (embeddings) for each topic, using the name of the topic as input, and used this embedding as an additional input feature (see Tree & embedding augmented Table 1). The language model was trained in advance on a general text corpus and was not fine-tuned. Information can be efficiently represented using such representations [32,33].

Table 1 shows the model's forecasting performance. The combined tree and embedding models have the best explained variance (R^2) scores and overall results. All results are reported for the test set (see Methods). Lag baseline is the last valid value of the target, or the transformed target, from 5 years beforehand in a regularized linear regression model.

Table 1
Model results for 5-year forecasting.

Target	Model	R^2 coefficient	Mean absolute error	Median absolute error	RMSE (Root mean square error)
$\hat{Y}$ - popularity	Lag baseline	0.973	113.80	44.37	229.12
	Linear model	0.974	101.19	45.11	214.82
	Boosting Tree (CatBoost)	0.981	75.58	26.32	183.67
	Tree & embedding augmented	**0.991**	**58.53**	**23.22**	**132.30**
$\Delta\hat{y}$ - % change in popularity	Lag baseline	0.004	115.42	83.38	289.98
	Linear model	0.03	114.83	66.19	286.22
	Boosting Tree (CatBoost)	0.306	60.56	18.47	241.96
	Tree & embedding augmented	**0.447**	**48.29**	**15.11**	**216.04**

3.4. Predictive model results

The scale of the analyzed data and its structure supported the use of tree-based boosting algorithms such as CatBoost [34]. Several machine learning models including linear regression and boosting tree (CatBoost) were trained on different targets: (i) the popularity of each topic in 5 years ($\hat{y}$), and (ii) the percent change in a topic's popularity in 5 years relative to the present ($\Delta\hat{y}$). The latter target is a more challenging one that also implicitly neutralizes the naïve lagging baseline and reduces the bias due to differing mean levels of popularity between topics. We found that non-linear boosting tree models gave the best results (Table 1). Our models outperform the historical lag baseline, which presents a proxy for human guesses. Similar performance for our models was achieved by other gradient boosting methods including XGBoost and lightGBM [34]. We further studied the features that contributed to the model. We further evaluated at the binary level: predicting if a topic would go up or down in popularity (i.e., binary prediction). At that level, we show 88 % accuracy compared to a 70 % baseline (most topics increased in popularity over time). These results were stable over time.

Supplementary Fig. S2 shows the SHAP analysis (based on Shapley values) for the model trained on the percent change target using the augmented deep learning embedding features. The top features of importance to the tree and text-augmented model are listed along with their values. We found that review articles strongly contribute to the model's performance. Numerous features associated with reviews are among the features that exhibit strong SHAP values.

3.5. Evaluation of unseen topics

Our results implicitly assume predicting changes in known topics. To reflect this, we added an evaluation of predicting completely novel topics over the 40 years covered in the data (Table 2). In this setup, the train-test split is performed at the topic level rather than the time-level, using 30-fold groupwise splits while still predicting 5 years ahead. This is considerably more challenging, as it reflects the problem of "What will the popularity of a completely unknown scientific topic be like over many decades?". This framework is unrealistically challenging, as we would expect a novel topic's behavior over decades to be predictable in advance, but it can be viewed as a proxy for lower-bound performance over completely unseen topics. We observe reduced performance and, surprisingly, no clear benefit from the text embeddings. We view the temporal evaluation setup as the more relevant one.

3.6. Predicting the rise and fall in topics' popularity

We limited the training data up to 2019, partially due to the extreme societal changes and publication biases in the past 2 years during the COVID-19 epidemic [35]. Nevertheless, we examined the model predictions for the present time, i.e., the per-topic model predictions for 2022, using 2017 data.

Table 3 provides a sample of topics predicted to have the greatest relative ($\Delta\hat{y}$) change, selected from the models' top results. Model predictions for 2022 were sorted by the highest absolute predicted change relative to 2017 popularity, then selected. We list topics that had declining popularity before 2017 and are predicted to continue to decline until 2022. An example of an erroneous model prediction is *influenza*, likely due to the 2020 COVID-19 epidemic knock-on effects. We further show several topics with inverse directionality (Table 3). These are other topics that were increasing in popularity in 2017 relative to their popularity level in 2016, which we predicted would decline in 2022, or that were predicted to show the opposite change trend. Specifically, their popularity decreased in 2017 relative to 2016, but nevertheless, we predict their trend to reverse. The list is sorted by the success order of the predictive model in each section.

Manual analysis showed most predictions to be correct, at least at the binary trend level (increasing or decreasing), with examples such as *cumin* and *graph neural networks* (Table 3). We included cases where the trend for a topic (increasing or decreasing popularity) is predicted to reverse, as was indeed the observed case (e.g., *MRI, antibiotics*, Table 3). Fig. 4 shows the actual changes in popularity for 2022 based on PubMed with a resolution of 1-year, 3-years, and 5-years for selected topics from Table 3. Fig. 4A shows representative topics with declined in popularity, while Fig. 4B shows topics that do not agree with the previous year's trend.

To improve the generality of our study, we developed an application for using PubMed's current information (as of 2022), allowing the user to activate the ML model of any topic or term of interest. Fig. 5 shows a screenshot of the results for two sets of terms that were new to the model (Fig. 5, top) and terms that were used for the training (Fig. 5, bottom). Normalized publication trends for viruses and vaccinations (Fig. 5, top) show a non-monotonic and quite complex historical trend for *influenza*, with a maximal popularity in 2010. This peak in popularity is a reflection of the outbreaks in North America (April 2009), where the new H1N1 influenza virus spread rapidly around the world within a few months [36]. The prediction for 2023–2028 supports a large variation for influenza, most likely following the suppression of publications and minimal occurrence of influenza during the COVID-19 pandemic, while the *RNA vaccine*

Table 2
Model predictions for topic-level splits.

Target	Model	R^2 coefficient	Mean absolute error	Median absolute error	RMSE (Root mean square error)
Topic-Level split $\hat{Y}$ - popularity	Linear Model	0.919	268.25	187.34	394.51
	Boosting Tree	0.86	134.45	33.69	507.23
	Tree & embedding augmented	0.748	198.61	58.37	697.18

Table 3
Selected predictions for 2022 with information limited to the year 2017.

Predicted popularity	Topics
a) Predicted to be more popular in 2022 Popularity: (2022 > 2017)	miRNA, drug repurposing, nanopore, carbon nanotubes, synthetic biology, metabolome, mononucleosis, illumina, NGS, connectome, lithium, cannabidiol, natural medicine, graph neural network, biosimilar, cumin, lncRNA, CRISPR, machine learning
b) Predicted to be less popular in 2022 Popularity: (2022 < 2017)	medulla oblongata, serotonin, DNA array, norepinephrine, neuropeptide, histamine, influenza, junk DNA, pituitary gland, ancient DNA, hypothalamus, somatosensory cortex, acetylcholine, cocaine, ribozyme
c) Predicted to reverse direction by 2022, relative to the 2016 to 2017 trend Popularity: (2017 > 2016 & 2022 < 2017) or (2017 < 2016 & 2022 > 2017)	eugenics, cerebellum, mononucleosis, hippocampus, MRI, antibiotic, norepinephrine, ancient viruses, zebra fish, neocortex, carbon nanotubes, carbon dating, HMM, savant

is expected to keep gaining popularity relative to 2022 (by ~3 folds).

Reanalysis of terms that were used for our model training (Fig. 1) indicates a large increase in *IL6*, the major interleukin that reflects the inflammatory response. The trend of other terms such as GABA and APOe seems to decline. Both terms have been extensively studied over the last 20 years, and while GABA is indicative of brain function and mental illness, APOe is a major genetic signature relevant for Alzheimer's disease. The SciTrends application provides a browsing option displaying the actual data from PubMed and the prediction trend by year for 1–6 years ahead.

4. Discussion

While bibliometric publication trend statistics are not a perfect approximation for scientific research, they can provide valuable insights into trends in a field and help researchers and investors make informed predictions about the future direction of research. It is possible to identify emerging trends in a field and predict, to some extent, how trends may evolve over time. With respect to the known quote from Niels Bohr (1970) "It is difficult to make predictions, especially about the future", we aim to predict scientific topics in 6 years in the dynamic life science and technology fields.

Often in experimental science, such as biology, new concepts and methods precede a vast increase in related research. A classic example is the use of a microscope, which established the germ theory and preceded a vast increase in research in human health, microbiology, vaccines, and more. A similar lag can be attributed to the discovery of DNA structure, which led to the evolution of molecular biology and its key technologies (RT-PCR, CRISPR, and NGS). We show that it is possible to outperform the naïve baselines by which what is popular today will be popular in the coming years. Long-term forecasting is non-trivial and has been thoroughly discussed in the electronic media [37].

We present a non-exhaustive list of 125 topics, representing a wide range of emerging and established scientific and biological topics. We then searched PubMed, using the data from PubMed by Year [38] and used it to identify trends in scientific research over decades. We observed that the mean popularity of all 125 topics increased over time, despite our year-based normalization (Supplementary Fig. S1). We hypothesize this might be due to improved automated annotation methods resulting in more keywords being annotated across all studies. This pattern persists even with further attempts at naively detrending the popularity target (not shown) or the percent change target ($\Delta\hat{y}$). Additionally, it might be explained as an outcome of survival bias. Specifically, we chose topics that are valid and active in the present time and most likely ignored the topics that became completely obscure and disappeared over time. Such bias may cause a shift toward more popular topics.

We used the data to train forecasting models to predict topics' popularity six years in advance. We show that our models are capable of predicting future trends in existing topics and outperforming the historical (lag) baseline. Numerous studies also used quantitative approaches while analyzing publications and citations per topic with or without multi-year dynamics [39–42]. Additionally, studies used publications to identify hidden connections between topics [43,44]. A systematic identification of patterns in such large datasets can accelerate innovation and augment creativity [45,46]. In tracing the evolution of topics, abstracts were parsed by rhetorical framing [47]. They found that topics referred to in results sections tend to decline, while the opposite holds for topics appearing in methods sections. Such findings are in accordance with our study, where the abundance of reviews relative to original research articles often precedes stagnating or declining topic popularity.

An important concept in forecasting tasks are leading indicators, which are exogenous variables that provide predictive, potentially causal information about targets of interest. Patents are an intriguing candidate for predicting similar turning points in scientific research [23]. Commercially relevant works are advised to obtain defensive patents prior to publication. In this case, patents should most likely precede publications. We examine the CRISPR-based gene editing technology as a case study [48]. We observe from the data that the number of patents in earlier years is a strong, leading indicator and better predicts research papers than the number of patents in that same year (Supplementary Table S4). Recall that, typically, temporal distance reduces the correlation between variables. The Pearson correlation between the number of CRISPR publications and patents filed at the same year is $r = 0.93$, while the correlation with patents filed 1 year before is $r = 0.98$. Using patent values from 1 year into the future gives an even lower correlation with popularity ($r = 0.88$). This is in line with the hypothesis that patents might be a leading indicator for research publications. This observation is also in line with the regulation of patent applications where confidentiality is requested to prevent its public disclosure in scientific publications before its acceptance.

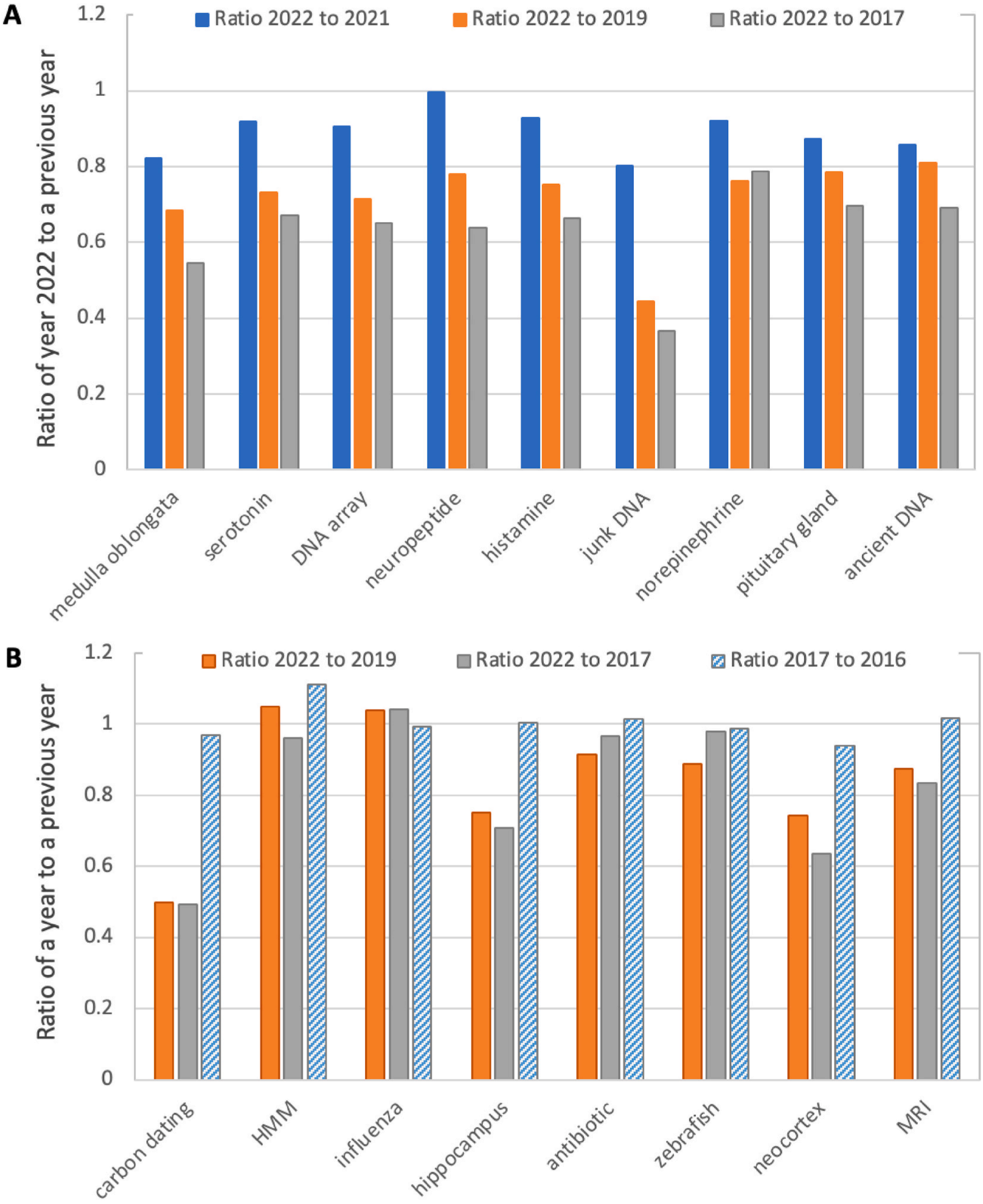

Fig. 4. Actual changes in popularity for 2022 based on PubMed. (A) 2022 predictions for a decline in popularity with the trend of 1-year, 3-years, and 5-years. (B) Predictions for 2022 that do not agree with the previous year's trend. The change in popularity for 3-years, 5-years, and 2017 to 2016 (stripped bar) is shown. The topics are representatives from Table 3.

In this study, we found that a relatively high number of reviews correlate with reduced future topic popularity. This could be because review articles are often published when a field or a technology is mature. When a field is in its early stages, a lot of new research is published, especially in experimental sciences. In other instances, such as the recent COVID-19 global pandemic, the field is so dynamic that the relative number of reviews was suppressed in view of the flood of original publications. The 3-year period of COVID-19 also affected peer-reviewed protocols, the time lag in publications, and the abrupt change in science that was made in the fields that are related to public health, epidemiology, virology, and vaccination. It is likely that once a field matures, both conceptual and literature surveys will be presented, and it may stagnate. This is supported by different works on the relative locations and context of topics in articles. Examples include MRI and fMRI in cognitive and neuroscience research, or deep learning in computer vision [49], language, and biology [7,16,50].

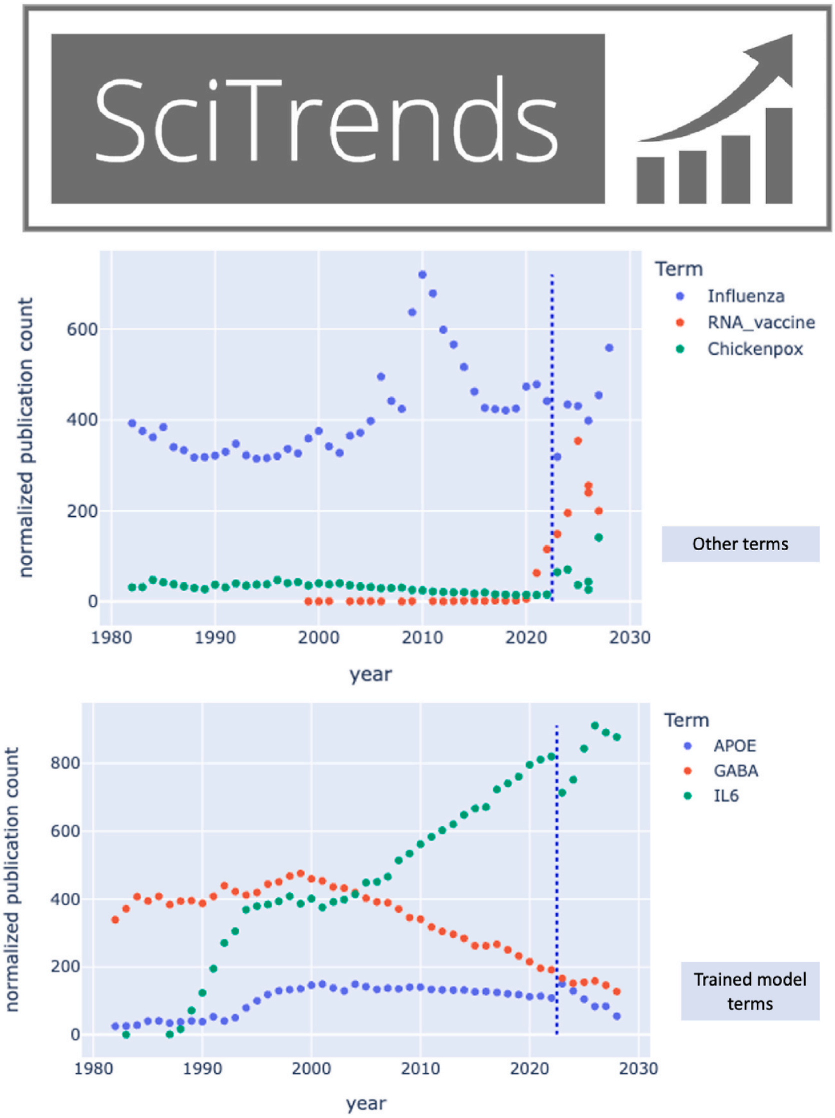

Fig. 5. A screenshot of the SciTrends application (Users can upload any topic or term that appears in PubMed (up to 10 topics can be loaded in a single run). The split between the normalized real data and the prediction trend is marked in 2022 (a vertical dashed line). The predicted trend is shown for six years from 2023 to 2028. User can search and browse in https://huggingface.co/spaces/hadasak/SciTrends, or use the attached API with the associated files.

This research is subject to several limitations. We have used PubMed as a ground-truth source for publication counts. However, PubMed does not include records from new channels for publishing in science, such as conference proceedings, open archives (e.g., BioRxiv), or online collections. Moreover, PubMed focuses mainly on medicine and biomedical sciences, and coverage of other research topics is limited. The generality of our prediction was not yet tested on other publication resources, such as scientific blogs, or domains that are not generally published on PubMed, such as social sciences or linguistics.

A cautionary aspect is cases of semantic transitions in terminology switching, where a topic may be referred to using different nomenclatures over time, e.g., NGS for deep sequencing. PubMed search algorithms help mitigate this by handling many acronyms automatically (e.g., "DNA" and "deoxyribonucleic acid"), but this is an aspect that our framework does not directly handle.

Defining something as popular is fuzzy and contextual. It can be viewed in relation to itself, i.e., a relative increase or decrease in popularity compared to the past (for example, neuropeptides, influenza, and mRNA vaccines), or in relation to other topics. For example, cancer research is an extremely active field, as it is a widely used term in immunology, cell biology, computational biology, medicine, and other scientific fields. In contrast, mentioning the genes that drive cancer (e.g., TP53, BRCA1, ATM) is far less represented in the cancer research literature.

The trends in scientific and technological topics are influenced by confounding effects ranging from human curiosity (e.g., space science), media coverage, and public awareness (e.g., gene therapy) or pressing challenges in public health (e.g., vaccines, climate change). These factors shape public and institutional interest in specific areas of science and technology. Thus, applications of this methodology to predict popularity as the basis for planning must also take human social context into account. For example, to help predict if something is a temporary "flash in the pan" or a meaningful, breakout topic that will remain relevant for years to come.

We found that unstructured text embeddings of topics using just their names provided additional information. The best predictive models out of those tested likely reflect the added capacity to learn latent dynamics between domains [7,26]. However, these results should be viewed with caution, as the underlying language model was trained on data from 2018. While we were not exposed to our tasks, the possibility of future information leakage cannot be discounted, barring a full retraining of the model for every year covered. Moreover, the limitations of the methods used should be acknowledged. Typical deep learning approaches do not support missing values and are not ideal for sparse time points per series or topic. The focus of this study is to share a useful tool with the broad scientific community, with the goal of improving forecasting accuracy in future work.

5. Conclusions

The dynamics and value ranges of the topics vary greatly by topic and time and may span over 4 orders of magnitude, even after normalization. Despite these dynamic and scale-varying targets, our results suggest that scientific publication trends are predictable years in advance using historical data as well as patents and in-domain publication trends, such as the number of reviews relative to research articles. We suggest that such methods can be of great benefit for planning critical decisions regarding career development, technological implantation, training, and education, as well as for early-stage researchers investing in infrastructure and training. To empower the utility of our prediction models, we developed SciTrends as an interactive application that presents the profile of any topic of interest covered by PubMed and its trend prediction for the following 6 years.

Data availability

Data and implementation details are provided with summary statistics in Supplementary Table S2. The variables used for the model are listed in Supplementary Table S3. The SciTrends application is available at: https://huggingface.co/spaces/hadasak/SciTrends with a search and browsing capabilities. Users can also use the attached API with the associated files. The website code for acquiring data, extracting features, training models and loading the provided pretrained models, as a webserver is provided in https://github.com/HadasaK1/SciTrends/tree/main/SciTrends. The code repository is available in https://github.com/ddofer/Trends.

CRediT authorship contribution statement

Dan Ofer: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hadasah Kaufman:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources. **Michal Linid:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Michal Linial reports financial support was provided by Hebrew University of Jerusalem. Michal Linial reports a relationship with Hebrew University of Jerusalem that includes: employment. None.

Acknowledgments

We thank the members from D. Shahaf and M. Linial laboratories for sharing their ideas and valuable discussions. We thank N. Rappoport and R. Zucker for supporting the web application. This work was partially supported by the Center for Interdisciplinary Data Science Research (CIDR, #3035000440) at the Hebrew University, Jerusalem.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e23781.

References

[1] L. Bornmann, R. Mutz, Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references, Journal of the Association for Information Science and Technology 66 (2015) 2215–2222.
[2] R. Fairclough, M. Thelwall, Questionnaires Mentioned in Academic Research 1996–2019: Rapid Increase but Declining Citation Impact, vol. 35, Learned Publishing, 2022, pp. 241–252.
[3] L. Tang, P. Shapira, J. Youtie, Is there a clubbing effect underlying C hinese research citation Increases? Journal of the Association for Information Science and Technology 66 (2015) 1923–1932.

[4] P. Larsen, M. Von Ins, The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index, Scientometrics 84 (2010) 575–603.
[5] Y. Mao, Z. Lu, MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank, J. Biomed. Semant. 8 (2017) 15.
[6] M.J. Salganik, Predicting the future of society, Nat. Human Behav. 7 (2023) 478–479.
[7] D. Ofer, N. Brandes, M. Linial, The language of proteins: NLP, machine learning & protein sequences, Comput. Struct. Biotechnol. J. 19 (2021) 1750–1758.
[8] J.S.G. Chu, J.A. Evans, Slowed canonical progress in large fields of science, Proc. Natl. Acad. Sci. USA 118 (2021), e2021636118.
[9] D.B. Klein, C. Stern, Groupthink in academia: majoritarian departmental politics and the professional pyramid, Indepen. Rev. 13 (2009) 585–600.
[10] S. Cohen, N. Dagan, N. Cohen-Inger, D. Ofer, L. Rokach, ICU survival prediction incorporating test-time augmentation to improve the accuracy of ensemble-based models, IEEE Access 9 (2021) 91584–91592.
[11] E. Spiliotis, V. Assimakopoulos, S. Makridakis, V. Assimakopoulos, The M5 Accuracy competition: results, findings and conclusions, Int. J. Forecast. 38 (2022) 1346–1364.
[12] L. Bornmann, R. Haunschild, Empirical analysis of recent temporal dynamics of research fields: annual publications in chemistry and related areas as an example, Journal of Informetrics 16 (2022), 101253.
[13] S. Effendy, R.H. Yap, Analysing Trends in Computer Science Research: A Preliminary Study Using the Microsoft Academic Graph. Paper Presented at the Proceedings of the 26th International Conference on World Wide Web Companion, 2017.
[14] P.H. Abelson, Trends in Scientific Research: rapid evolution of the frontiers is a hazard for scientists young and old, Science 143 (1964) 218–223.
[15] G. Serrano Najera, D. Narganes Carlon, D.J. Crowther, TrendyGenes, a computational pipeline for the detection of literature trends in academia and drug discovery, Sci. Rep. 11 (2021), 15747.
[16] P. Savov, A. Jatowt, R. Nielek, Identifying Breakthrough Scientific Papers, vol. 57, Information Processing & Management, 2020, 102168.
[17] N. Mazov, V. Gureev, V. Glinskikh, The methodological basis of defining research trends and fronts, Sci. Tech. Inf. Process. 47 (2020) 221–231.
[18] R. Van Noorden, Formula predicts research papers' future citations, News, Nature 3 (2013).
[19] F.G. Nezhad, F. Osareh, M.R. Ghane, Forecasting the Subject Trend of International Library and Information Science Research by 2030 Using the Deep Learning Approach, vol. 20, International Journal of Information Science & Management, 2022.
[20] T.M. Abuhay, Y.G. Nigatie, S.V. Kovalchuk, Towards predicting trend of scientific research topics using topic modeling, Proc. Comput. Sci. 136 (2018) 304–310.
[21] E. Tattershall, G. Nenadic, R.D. Stevens, Detecting bursty terms in computer science research, Scientometrics 122 (2020) 681–699.
[22] M. Park, E. Leahey, R.J. Funk, Papers and patents are becoming less disruptive over time, Nature 613 (2023) 138–144.
[23] T.D. Griffin, S.K. Boyer, I.G. Councill, Annotating patents with Medline MeSH codes via citation mapping, in: Paper Presented at the Advances in Computational Biology, 2010.
[24] C. Viboud, A. Vespignani, The future of influenza forecasts, Proc Natl Acad Sci U S A 116 (2019) 2802–2804.
[25] E.O. Nsoesie, J.S. Brownstein, N. Ramakrishnan, M.V. Marathe, A systematic review of studies on forecasting the dynamics of influenza outbreaks, Influenza Other Respir Viruses 8 (2014) 309–316.
[26] D. Ofer, M. Linial, Inferring microRNA regulation: a proteome perspective, Front. Mol. Biosci. 9 (2022) 989.
[27] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: Gradient Boosting with Categorical Features Support, 2018 *arXiv preprint arXiv:181011363*.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[29] N. Reimers, I. Gurevych, Sentence-bert: Sentence Embeddings Using Siamese Bert-Networks, 2019 *arXiv preprint arXiv:190810084*.
[30] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, Minilmv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers, 2020 *arXiv preprint arXiv:201215828*.
[31] C. Zhang, C. Liu, X. Zhang, G. Almpanidis, An up-to-date comparison of state-of-the-art classification algorithms, Expert Syst. Appl. 82 (2017) 128–150.
[32] D. Ofer, D. Shahaf, Cards against AI: Predicting Humor in a Fill-In-The-Blank Party Game, 2022 *arXiv preprint arXiv:221013016*.
[33] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (2022) 2102–2110.
[34] J.T. Hancock, T.M. Khoshgoftaar, CatBoost for big data: an interdisciplinary review, J Big Data 7 (2020) 94.
[35] T. Ahmad, M.A. Murad, M. Baig, J. Hui, Research trends in COVID-19 vaccine: a bibliometric analysis, Hum Vaccin Immunother 17 (2021) 2367–2372.
[36] D. Gatherer, The 2009 H1N1 influenza outbreak in its historical context, J. Clin. Virol. 45 (2009) 174–178.
[37] The Forecasting Collaborative, Insights into the accuracy of social scientists' forecasts of societal change, Nat. Human Behav. 7 (2023) 484–501.
[38] E.V. Sperr, Libraries and the future of scholarly communication, Mol. Cancer 5 (2006) 1–2.
[39] R.R. Braam, H.F. Moed, A.F. Van Raan, Mapping of science by combined co-citation and word analysis. I. Structural aspects, J. Am. Soc. Inf. Sci. 42 (1991) 233–251.
[40] I. Tahamtan, L. Bornmann, What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018, Scientometrics 121 (2019) 1635–1684.
[41] L. Bornmann, The problem of citation impact assessments for recent publication years in institutional evaluations, Journal of Informetrics 7 (2013) 722–729.
[42] J.B. Voytek, B. Voytek, Automated cognome construction and semi-automated hypothesis generation, J. Neurosci. Methods 208 (2012) 92–100.
[43] W. Marx, L. Bornmann, Change of perspective: bibliometrics from the point of view of cited references—a literature overview on approaches to the evaluation of cited references in bibliometrics, Scientometrics 109 (2016) 1397–1415.
[44] R. Kleminski, P. Kazienko, T. Kajdanowicz, Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification, J. Inf. Sci. 48 (2022) 349–373.
[45] T. Hope, J. Chan, A. Kittur, D. Shahaf, Accelerating innovation through analogy mining, in: Paper Presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.
[46] A. Kittur, L. Yu, T. Hope, J. Chan, H. Lifshitz-Assaf, K. Gilon, F. Ng, R.E. Kraut, D. Shahaf, Scaling up analogical innovation with crowds and AI, Proc. Natl. Acad. Sci. USA 116 (2019) 1870–1877.
[47] V. Prabhakaran, W.L. Hamilton, D. McFarland, D. Jurafsky, Predicting the rise and fall of scientific topics from trends in their rhetorical framing, in: Paper Presented at the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics *(Volume 1: Long Papers)*, 2016.
[48] J.D. Sander, J.K. Joung, CRISPR-Cas systems for editing, regulating and targeting genomes, Nat. Biotechnol. 32 (2014) 347–355.
[49] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.
[50] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018 *arXiv preprint arXiv:181004805*.