ACS SyntheticBiology

pubs.acs.org/synthbio Research Article

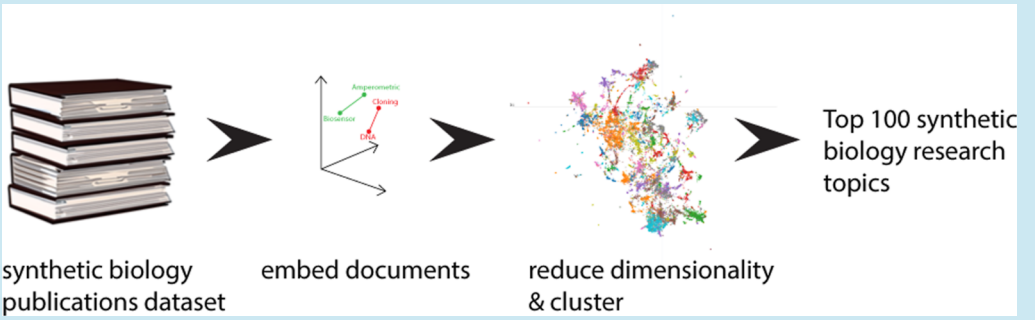
Navigating the Frontier of Synthetic Biology: An AI-Driven Analytics Platform for Exploring Research Trends and Relationships

Published as part of the ACS Synthetic Biology virtual special issue "AI for Synthetic Biology".

Felix Meier, Thom Dixon, Tom Williams, and Ian Paulsen*

Cite This: *ACS Synth. Biol.* 2023, 12, 3229–3241 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The field of synthetic biology has experienced rapid growth in recent years, leading to an overwhelming amount of literature that can make it difficult to comprehend the scope and trends of the discipline. In this study, we employ topic modeling to comprehensively map research topics within synthetic biology, revealing subtopics and their relationships, as well as trends over time. We utilize metadata to identify the most significant journals and countries in the field and discuss potential policy impact on the research output. In addition, we investigate co-authorship networks to analyze collaborations among authors, institutions, and countries. We believe that our findings could serve as a valuable resource for gaining a deeper understanding of synthetic biology and provide a foundation for analyzing other disciplines.

KEYWORDS: *synthetic biology, metabolic engineering, natural language processing, topic modelling, network analysis*

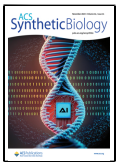
INTRODUCTION

Synthetic biology is an interdisciplinary field that combines the principles of engineering and biology to design and construct new biological parts, devices, and systems that do not exist in nature. The field has grown rapidly in recent years (Figure 1) with advances in genetic engineering, computational biology, and biotechnology providing new tools and strategies for the design and construction of synthetic biological systems.[1,2] The science of synthetic biology has evolved from tuning genetic components and engineering simple feedback circuits in the 2000s to synthesizing entire genomes and engineering complex metabolic pathways.[3−5] As capability grows from engineering genes to genomes and the tools and infrastructure improve, it becomes conceivable to achieve what was once deemed impossible such as realizing an orthogonal central dogma, simulating whole cells, de novo genome and protein design, controllable evolution, living materials, and artificial cells.[6−9] This explosion of knowledge has changed the applications and focus during the second decade of synthetic biology but has also made it more difficult to agree on one fundamental question: What actually is synthetic biology and what does it encompass?[10−12]

To attempt to answer this question, one could conduct a meta-analysis of publications, which usually involves gathering publications on a topic, summarizing the current state of knowledge and open questions, highlighting, critically evaluating differing perspectives, and synthesizing probable future directions. Traditionally, this is quite a manual process, can introduce bias due to the author's perspective on the field, and due to the abundance of scientific literature and rapidly changing nature of a novel field can make it challenging to systematically discern trends and uncover underlying structural changes. Therefore, scalable methods to systematically extract and aggregate information from published texts are required.[13,14]

Natural language processing (NLP) lies at the intersection of linguistics and artificial intelligence and uses statistics to

Received: March 30, 2023
Published: August 30, 2023



https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

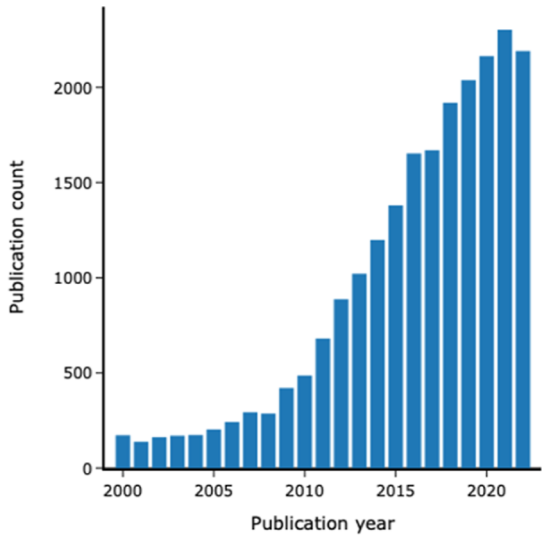


Figure 1. Web of Science publications on synthetic biology.

analyze and interpret large amounts of text, retrieve information, and identify key concepts. It has its origins in the 1940s in machine translation and has undergone several phases of research on the syntactic structure, semantics, and parsing algorithms. This has laid the groundwork for advanced current applications in speech recognition, sentiment analysis, topic modeling, text classification, and text generation.[6,7] The advent of the internet and the growing necessity for extracting information from enormous unstructured data sets has fueled a surge in interest in NLP. Progress has further been driven by significant advancements in deep learning algorithms and exponentially more computer processing power. Of particular interest to this study is the subdiscipline of topic modeling, which aims to identify the themes present within a corpus of documents. Three major advances in topic modeling are relevant to our approach: term frequency—inverse document frequency, latent Dirichlet allocation (LDA), and bidirectional encoder representations from transformers.[15]

The first significant development in 1983 involved comparing the term frequencies of a document to their occurrences in the entire corpus, resulting in the computation of a term frequency-inverse document frequency score (TF-IDF).[16] This method permits the estimation of a word's representativeness in a specific document. Unfortunately, TF-IDF only relies on frequencies and does not take semantic structure into account and the context in which these words often co-occur. A second popular technique published in 2003 is LDA, which is a probabilistic method that can discover hidden (latent) topics within a corpus by assigning topic probability distributions to each word and documenting and then iteratively updating these assessments of topics to words. The LDA model does not consider syntactic structure and lacks the ability to dynamically capture changes in topic distribution over time, a crucial aspect for the scope of this study.[8] More recently, language models based on a transformer architecture have emerged as state-of-the-art models in NLP. The transformer architecture typically consists of an encoder and a decoder, which are composed of multiple layers of self-attention and feedforward neural networks. The encoder takes in the input sequence and produces a set of hidden representations that capture the relevant information in the sequence. The decoder then takes in these hidden representations and generates the output sequence by attending to the relevant parts of the input sequence.[17] Bidirectional encoder representations from transformers (BERTs), published by Google researchers in 2018, is a remarkable example of a transformer-based language model that considers both semantic and syntactic structures to assign topics to a given text.[18] BERT is a transformer-based language model that uses a multi-layer bidirectional encoder to capture the contextual information of a given word. Unlike unidirectional language models like OpenAIs GPT-3, BERT takes into account both preceding and subsequent context of a word, making it highly relevant for topic modeling tasks.[19] Optimal performance is achieved by pretraining BERT on a large corpus of text using masked language modeling and next-sentence prediction tasks. The masked language modeling task involves randomly masking some tokens in a sentence and training the model to predict the missing tokens. The next sentence prediction task involves training the model to predict if two given sentences are consecutive in the text corpus or not. After pretraining, BERT can be fine-tuned for various downstream tasks, including topic modeling. Fine-tuning BERT involves updating its parameters using a smaller corpus of text that is specific to the target task. In topic modeling, the fine-tuning process involves training BERT to assign a topic label to a given text. Once fine-tuned, BERT can accurately predict the topic of a given text.

We seek to apply these advancements in topic modeling to synthetic biology scientific publications to gain a better understanding of how the field is shaped. Popular databases, such as Web of Science (WOS), PubMed, or Scopus usually only offer broad categories (Biochemistry, Molecular Biology, and Immunology) or keywords (Metabolism, DNA, *Escherichia coli*) to understand their publication data sets. Notable comprehensive efforts to map the synthetic biology landscape published in 2012 and 2016 have used keywords and network analysis for mapping the first decade of synthetic biology.[20,21] We instead use topic modeling, metadata analysis, and network analysis to build and improve upon these mapping approaches and primarily focus on mapping the second decade of synthetic biology.[10] This study serves as a future resource to explore the field in more detail, investigate possible latent relationships between sub-topics, and discover emerging trends. In addition to mapping the research areas that scientists are prioritizing, we endeavored to gain insights into the most pertinent journals in the field as a resource for scientists aiming to publish in the field.

RESULTS AND DISCUSSION

Aims and Scope. This paper uses recent advances in topic modeling to more clearly define synthetic biology subdomains, extract relational information between sub-topics, and examine their prominence between 2000 and 2021. We use metadata of published articles to ascertain which journals and countries are most relevant to the field. We use network analysis to uncover relationships between the most prominent individual researchers, countries, and organizations. We also aimed to establish an interactive data set including an open-source code that can be interrogated by others to identify trends and relationships of interest as well as provide a framework that could be applied to other disciplines.

Topic Modeling. In this study, we employed the BERTopic model to identify and analyze the underlying topics

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

Table 1. Present Study Utilized BERTopic to Identify Topics from Abstracts of a WOS Synthetic Biology Publication Data Set[a]

count	topic	count	topic	count	topic
2143	metabolic production pathway	181	biofilms biofilm materials	48	algorithms algorithm evolutionary
1501	circuits biological synthetic	178	electron oneidensis transfer	48	communication mc receiver
1193	noise networks network	175	carotenoids carotenoid astaxanthin	47	tuberculosis mycobacterium mtb
1090	expression promoter promoters	169	enzyme enzymes immobilization	46	wine ionone yeast
987	biology synthetic research	156	plants transgenic cry	46	beta amyloid ad
923	biosynthetic natural products	149	phage phages bacteriophage	45	images segmentation image
893	RNA translation trna	141	microbial consortia communities	39	conservation extinction biodiversity
743	plant biosynthesis production	130	mice liver hepatic	39	space mars earth
667	CRISPR cas editing	123	nitrogen plant nitrogenase	39	circadian clock clocks
664	cyanobacteria photosynthetic pcc	123	fungi aspergillus fungal	38	silk spider spidroins
587	plant plants genes	122	metal heavy cd	36	protease proteases protein
574	DNA assembly cloning	122	cancer bladder patients	35	magnetic magnetosome magnetotactic
551	biosensors biosensor sensing	121	life origin evolution	35	miRNA miRNAs micrornas
465	membrane vesicles lipid	118	bone abm tissue	32	HIV samples specimens
383	virus viral COV	105	receptors receptor gpcrs	30	bont botulinum lc
376	protein proteins peptide	101	fl glycosylation glycan	30	ceramide alpha apoptosis
373	light optogenetic control	99	data network networks	30	atps recovery phase
372	populations genetic population	87	falciparum parasite plasmodium	29	artemisinin artemisinic annua
363	cell cells self	85	drive drives gene	28	hydrogen waste mol
361	sex genetic selection	85	heme cyp cytochrome	27	myoglobin heme hemoglobin
356	was alpha recombinant	83	fluorescence emission agncs	27	ppr editing RNA
340	yeast cerevisiae genome	78	mitochondrial mitochondria mtdna	27	archaea archaeal asgard
336	delivery gene transfection	76	zebrafish vertebrate pax	25	gsh cys ptpase
336	proteins protein secretion	73	editing cas crop	25	ubiquitin cdc ubiquitination
331	wheat chromosomes chromosome	71	actin motors filaments	23	dendrimers dendrimersomes dendrimer
322	genome genes essential	66	muscle smooth contractile	23	degrees thermostable ttc
315	gut bacteria microbiome	65	detection sars COV	23	graphene nanoparticles silver
275	pha phb degradation	64	shell bmcs microcompartments	22	gsoil strain omega
254	lignin biomass cellulose	56	learning deep machine	22	wounding broccoli phenolic
220	car cells cell	54	effects traits qtls	22	meat phytase food
198	microfluidic droplet droplets	52	ginsenoside ginsenosides ginseng	22	methylation imprinted imprinting
183	dna strand nanostructures	51	hia rs association	20	ige allergen allergic
182	DNA rad chromatin	48	nanowires Raman nanowire	19	research publications citations

[a]The top 99 topics along with their corresponding frequency of occurrence are displayed.

within a corpus of ∼23k scientific articles related to synthetic biology queried from the Web of Science database.

As summarized in Table 1, the resulting topics reflect a diverse range of subject matter, with metabolic engineering emerging, circuits, networks, promoters and expression, and synthetic biology research emerging as the largest topic clusters containing over 25 % of the total amount of publications. Overall, we observe a strong presence of well-established topics, such as biosynthetic natural product/plant biosynthesis/plant genes (counts: 923, 743, and 587), CRISPR-Cas genome editing (count 667), and DNA assembly (count 574), which is expected given their significance within the field. Interestingly, we also uncovered surprisingly prevalent topic clusters. For instance, membrane lipid vesicles (count 465) and optogenetics (count 373) featured more prominently than expected by the authors. As we move down the list of topics, we note the emergence of more specialized themes, such as space (count 39) and spider silk (count 38). This, we believe, demonstrates the suitability of BERTopic for modeling a scientific field in sufficient detail to not only emphasize overarching categories but also emerging niche research focal points. While most topic descriptions are sufficiently informative and specific, we acknowledge that there is some overlap. For instance, virus viral COV (count 383) and detection SARS-COV (count 65). Topics such as cell cells self (count 363) could benefit from a more descriptive topic title and this highlights where the model falls short. To explore these relationships further, we can visualize hierarchical relationships between topics (Figure 2).

Hierarchy. Our approach can highlight similarity between topic clusters, such as lignin valorization, plant biosynthesis, natural product biosynthesis, metabolic engineering, bioproduction via cyanobacteria, and transgenic plants. This works quite well in a lot of cases such as grouping evolutionary algorithms, data and networks, machine learning, and image segmentation close together and something like circadian clocks in relative proximity. All of these could broadly be classified as algorithm based. However, the clustering also grouped the seemingly loosely related topics of ATP recovery phase and spider silk together. Conversely, it should be noted that solely because topic-relatedness does not immediately make sense to a human observer, there are underlying language trends in the model that support the inference. The resulting chart, therefore, provides a resource for researchers to explore and identify relevant adjacent research topics that align with their respective fields as well as investigate overarching topic clusters. There remains the need for some judgement in interpreting the identified adjacencies of topics. We have

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

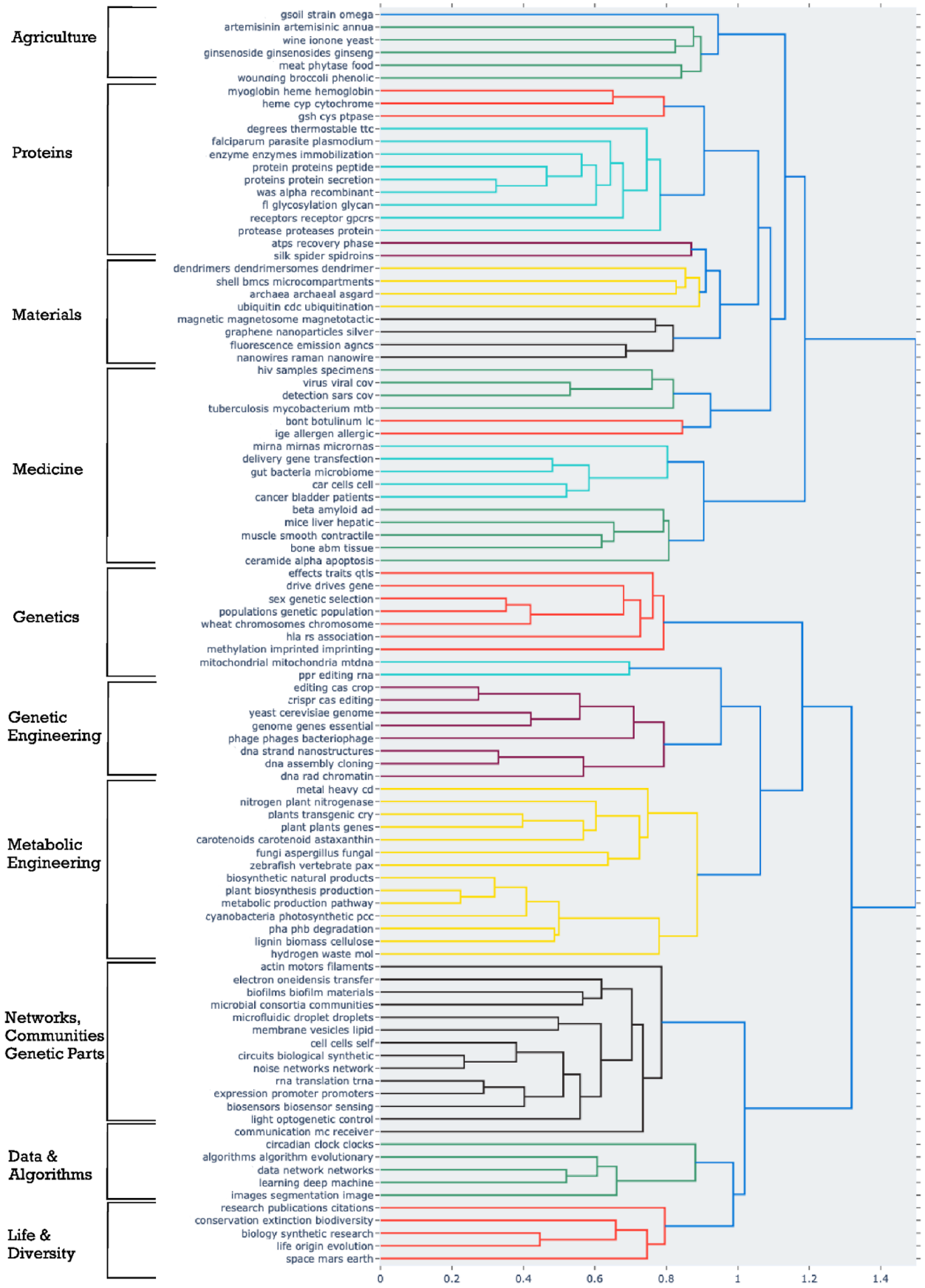


Figure 2. Hierarchical relationship between topic clusters based on cosine similarity between c-TF-IDF scores of each topic. Closeness between two c-TF-IDF representations indicates similar synthetic and semantic structure between the two respective topics. Clusters are denoted by colors. Overarching topic dimensions are shown on the left designated by the authors.

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

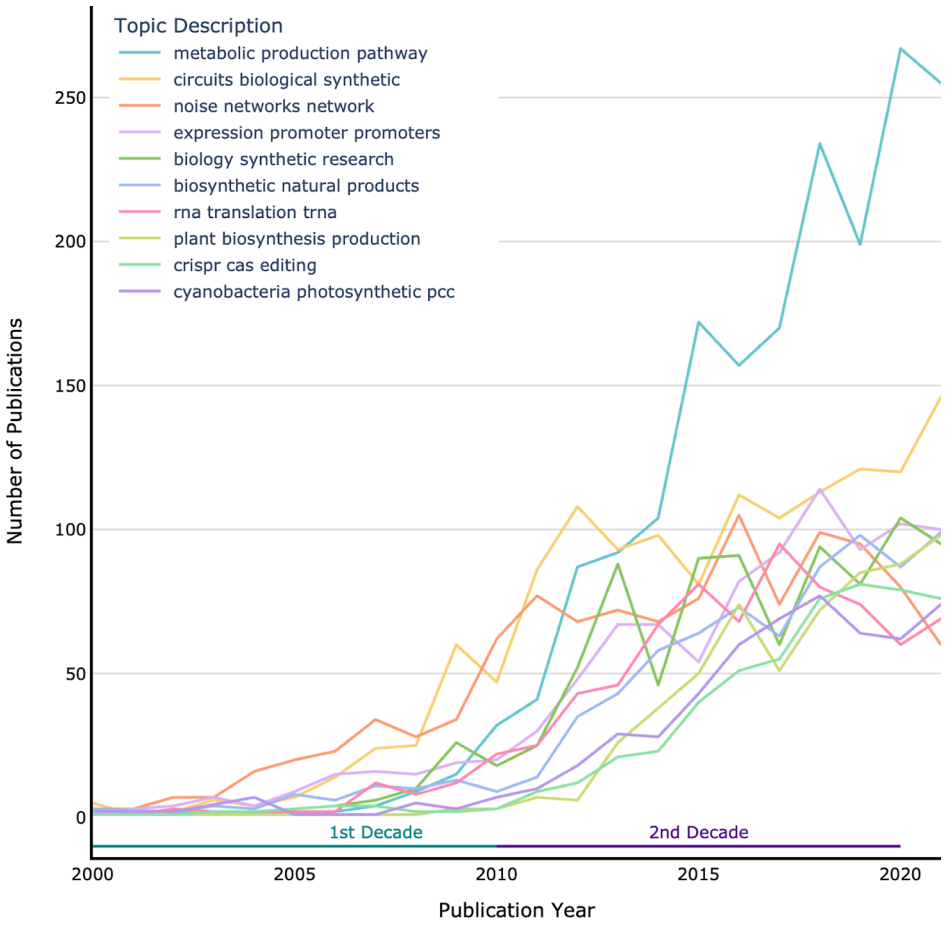


Figure 3. Dynamical topic modeling of 10 largest topic clusters between 2000 and 2021. The *y*-axis indicates the number of publications in the respective year classified into that topic cluster. The first and second decade of synthetic biology are highlighted on the *x*-axis.

grouped the topic clusters into 10 overarching categories that reflect the increasing diversification of the field when compared to the previous analysis in 2012 and 2016, which broadly suggests that the field can be divided into categories, such as metabolic engineering, DNA assembly, minimal genome, gene expression optimization, artificial cells, mammalian cells, gene circuits, genetic networks, and systems biology.[20,21] Particularly, the incorporation of medicinal, agricultural, and material science topics reflects this trend. Another hypothesis could be that topics that fall under the umbrella of traditional genetic engineering are increasingly being associated with synthetic biology.

To provide greater insights into the interrelationships between subtopics, Supplementary Figure 2 highlights similarity scores between different topics and allows for the exploration of similarity scores between individual topic clusters.

Dynamic Topic Modeling. While a static picture of the field and the relationships within can be of interest, most of the time it is of more significance to explore topics from a temporal perspective. Therefore, we present an analysis of the publication frequency of the 10 most significant topic clusters from 2000 to 2021 (Figure 3). To enhance the visual clarity of the figure, we have excluded infrequent topic clusters.

Our findings reveal a sharp upward trend in metabolic engineering and biological synthetic circuits. CRISPR research output has grown consistently but seems to have plateaued after 2018. Notably, the topic representation algorithm has displayed CRISPR-Cas genome editing before the publication of the seminal paper that described the system in 2012.[22] CRISPR-Cas genome editing is statistically the best topic representation for that cluster but other genome editing technologies prior to CRISPR are also grouped into that the cluster due to their similar linguistic structure, resulting in values prior to 2012. Research into CRISPR biology may also have contributed to that cluster prior to 2012. Broadly speaking, growth in the top 10 largest clusters has slowed down and has for most of them stabilized between 50 and 100 annual publications.

Comparing our approach to mapping the synthetic biology field to previous approaches by Oldham et al., two observations stand out: first, synthetic biology has changed substantially through the emergence of the above-mentioned topics and while some core topics remain constant, the field has diversified significantly to cover wider applications and a greater proportion of biological research.[20] This can also possibly help to explain the slowing in growth in the largest topic clusters as a research focus has shifted to incorporate a wider array of topics.

Journals. According to Ulrich's Web serials analysis system, in February 2023 there are over 35 000 English language peer-reviewed scientific journals. The sheer abundance of possible publishing venues can make it difficult to determine which journals to target for maximum impact and alignment with career trajectory. Due to its novelty and broad applicability, synthetic biology papers could see a reasonable alignment with

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

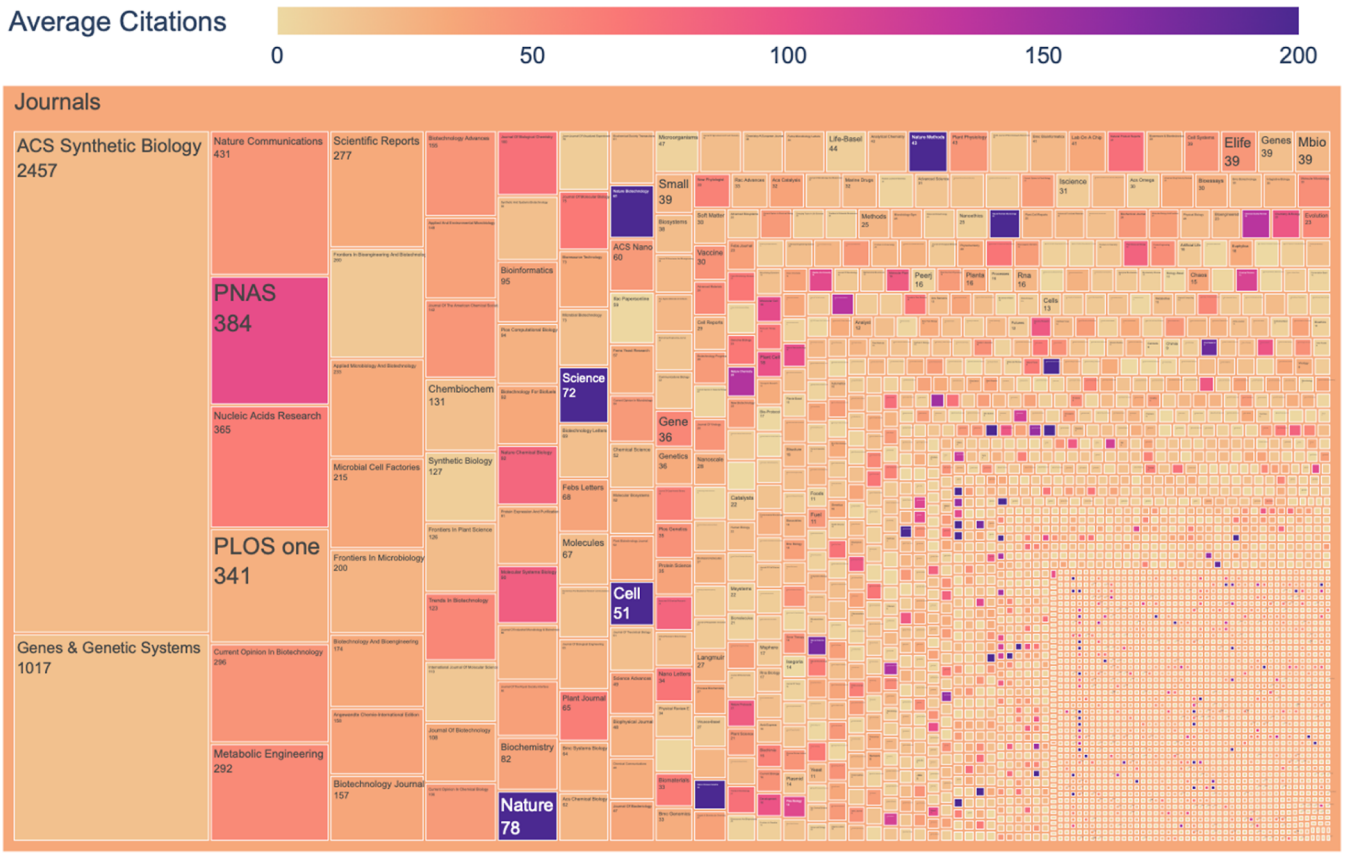


Figure 4. Treemap depicting journal distribution, where the size of each box and the displayed number corresponds to the proportion of articles published in a particular journal. The hue of each box is indicative of the average citation count received by papers published in the respective journal.

anything from a metabolic engineering journal to an ethics or material science journal. The expansive reach necessitates scoping a plethora of journals for fit, impact factor, and potential citation yield. This can be a particularly daunting task, especially for early career researchers, and is rarely done systematically or comprehensively. Figure 4 attempts to provide a systematic overview of journals for the field of synthetic biology. While average citation count per paper is around 36, the highest cited journals such as Nature, Science, and Cell only occupy a small fraction of the data set, but their papers have average citation counts of an order of magnitude larger. We hypothesize that these prominent journals tend to accept papers of interest to a diverse disciplinary audience and articles with potential high impact, both likely contributing to a high number of citations. Out of the top 10 papers in terms of citation count in the data, 3 were published in Cell, 3 in Science, 1 in Nature Methods, 1 in Nature Reviews Drug Discovery, 1 in Scientific Reports, and 1 in Annual Review of Genetics, all of which further skew the average citation count toward these prestigious journals. Publications in the data set are distributed over 2476 journals, 1193 of which only contain a single publication. In terms of popularity, ACS synthetic biology stands out containing ∼10.6% of the total publications of the data set. This most likely can be attributed not only to their relevance to the scope of synthetic biology but also due to the way the query was structured as ACS synthetic biology publications prominently feature the journal name on each page.

Countries. Resources, economic milieu, and education systems supporting research vary across countries, and research focus in synthetic biology research is not uniformly distributed geographically. For this part of the study, we are to categorizing papers by their respective country of origin to elucidate these geographic trends. During our analysis, we have considered all collaborating countries as distinct entities. For example, a paper with eight authors associated with three institutions, one in the US, one in China, and one in Germany will count once toward each country. We decided on this approach due to the collaborative nature of research, the difficulty in disentangling the primary research institution from collaborative ones and the futility of weighing individual contributions appropriately. However, this approach does skew the data toward countries that produce papers with many collaborators. This in turn also means that these publications can count toward the median citation values of multiple countries.

The top 5 countries in aggregate contribute about two-thirds of research output (Figure 5a). The dominance of the United States could possibly be explained by the first mover advantage as the field largely emerged from laboratories in Massachusetts and California.[1] However, this dominance is by no means set in stone and the period 2014−2021 demonstrates the enormous rise in synthetic biology publications originating from China, which has almost reached parity with the United States in terms of paper output by 2022 (Figure 5b). Another noteworthy trend is the increase in synthetic biology publications in the rest of the world (other).

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

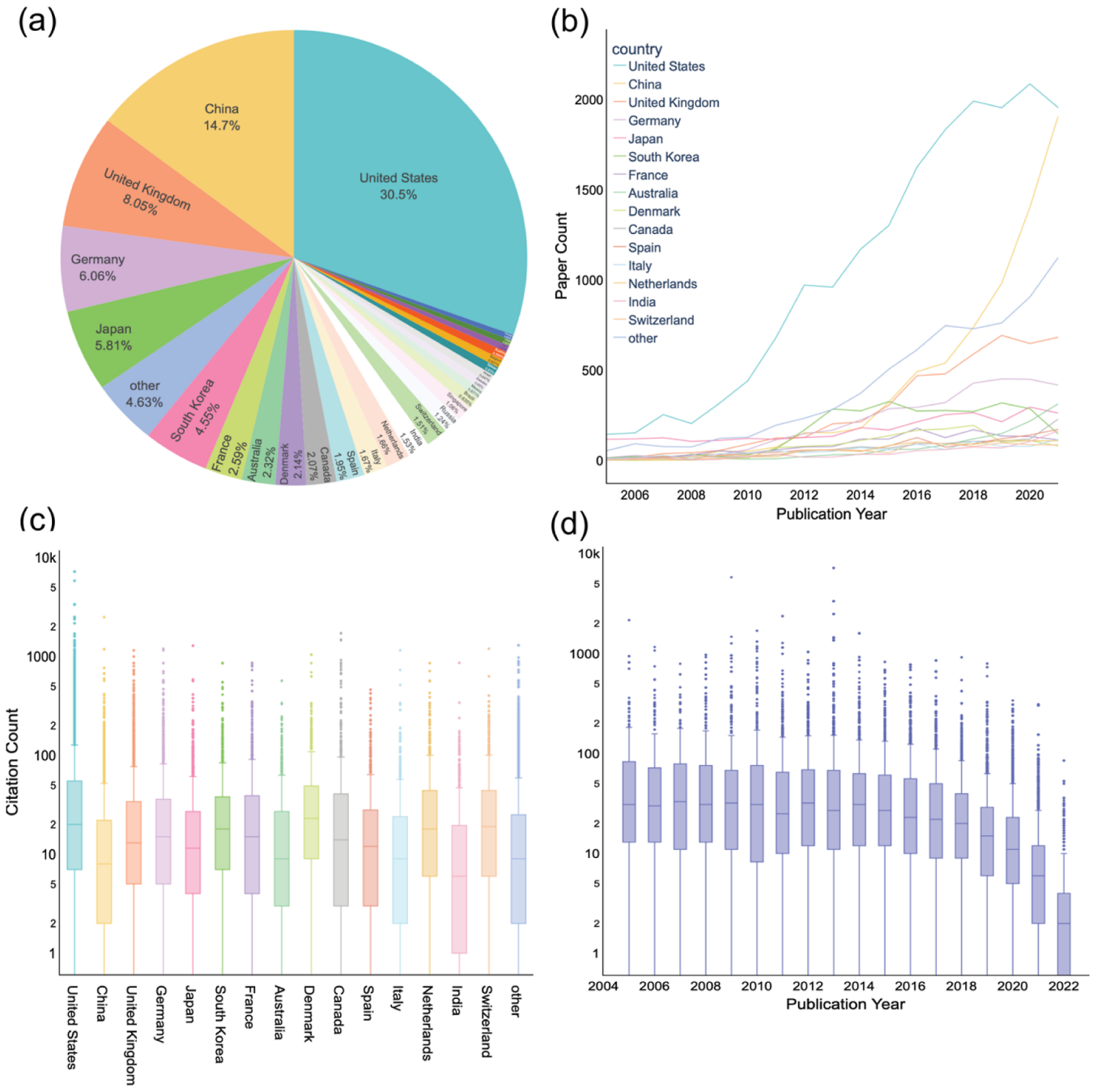


Figure 5. (a) Relative distribution of synthetic biology publication count by country, (b) absolute distribution of publications by top 15 synthetic biology paper producing countries over time, (c) cumulative citation count for top 15 countries in terms of paper count displayed as a boxplot where the vertical line in the box represents the mean, the box the interquartile range between Q1 and Q3, the dot outliers, and the top and bottom lines the minimum and maximum (d) cumulative citation count of all countries combined per year displayed as a boxplot where the vertical line in the box represents the mean, the box the interquartile range between Q1 and Q3, the dot outliers, and the top and bottom lines, the minimum and maximum.

The UK has seen substantial increases whereas South Korea plateaued after 2013 with 250−300 yearly papers. Australia has seen its output increase by a factor of 5 between 2015 and 2021.

However, output does not equal quality, and even though citations are not a perfect measure of quality in a world with increasing numbers of co-authors, longer reference lists, self-citing, and increased publication frequency, they remain a good proxy for interest.[23] When comparing median citations per paper for the top 15 paper producing countries (Figure 5c), the US (median 20) still receives more citations compared to China (median 8). Overall, only China, India, Australia, and Italy have produced papers with median citations at or below the rest of the world ("other", median 9). A plausible conjecture is that studies originating from highly cited countries such as the United States (20) or Denmark (23) are garnering higher levels of attention. Another hypothesis is that it takes time after publication for a paper to reach the community, be cited in other papers, and then for these papers to be published, as shown in Figure 5d. Considering the

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

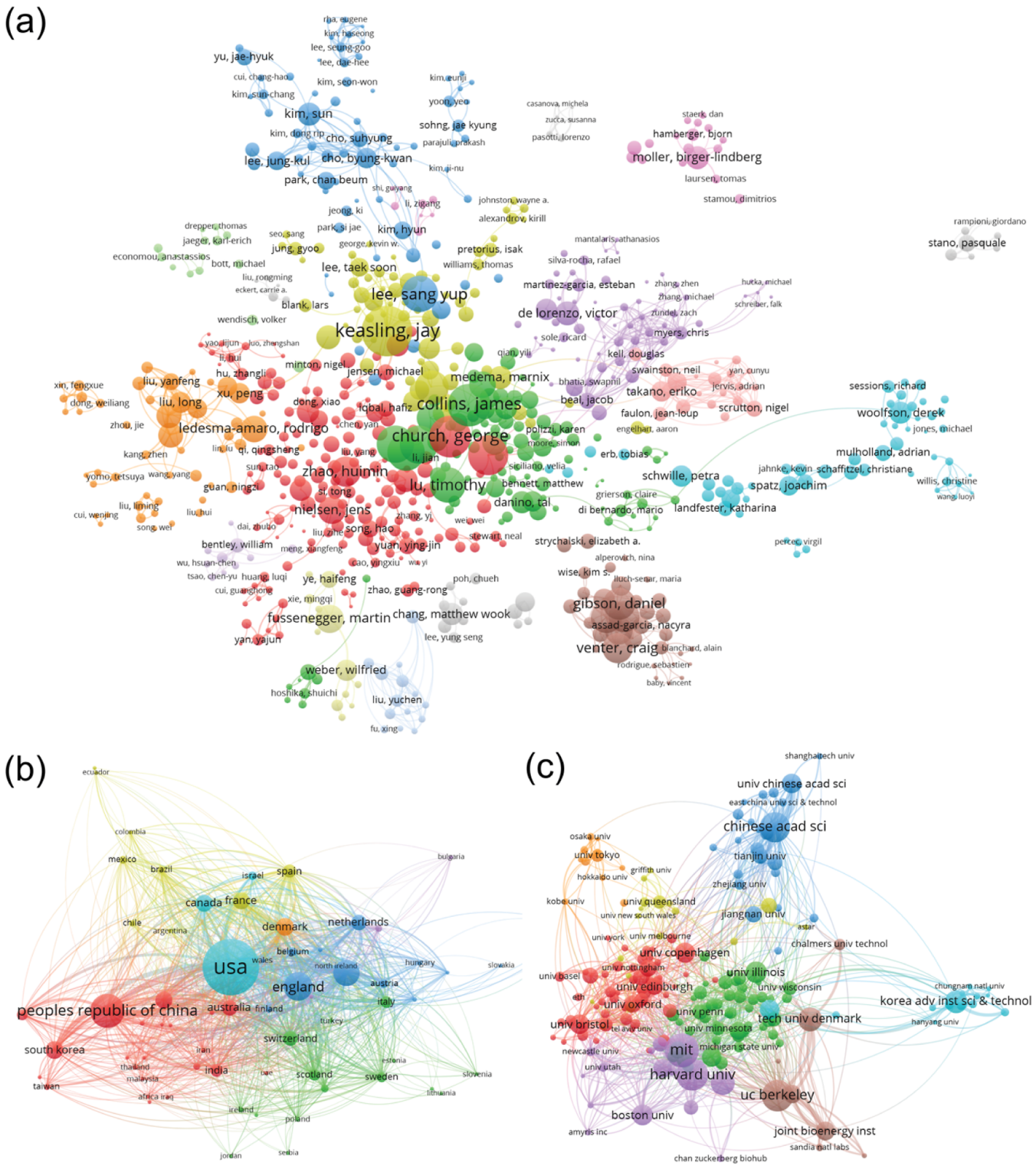


Figure 6. (a) Co-authorship network of top 750 most connected authors with at least 5 publications where size represents normalized citation count and color indicates the cluster (b) network of top 75 most connected countries based on co-authorship with at least 10 publications (c) network of top 200 most connected organizations based on co-authorship with at least 35 publications.

fraction of papers originating from China, Australia, and India post 2015 and therefore, for those not having experienced their highest citation potential, this most likely contributes to the difference between the US and China in terms of difference in median citations per paper. Denmark's output in contrast has peaked in 2018 at 195 publications and has since halved, possibly contributing to its high median citation count.

Policy. There are a diverse set of public policy drivers that may be contributing to the trends we have identified in the data. While the public policy base for synthetic biology in the United States is much more mature, a key question emerges from the data—why have publication numbers in the US stabilized since 2016? This stabilization has occurred at the same time as the Engineering Biology Research Consortium

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

was established and captured the acute phase of the COVID-19 pandemic. Our hypothesis is that the data demonstrates the shift of US synthetic biology research into the commercial sector. While raw publication numbers may have stabilized, this may in fact evidence the embedded maturity of the field in the US. Recent assessments of synthetic biology's contribution to the US bioeconomy,[24] corporate funding trends (Stephanie Wisner, Q1 Shatters Previous synthetic biology Investment Record—Signals Projected 2021 Investment of up to $36 Billion),[25] and public policy announcements in the US have included the Executive Order on Biomanufacturing and the Chips and Science Act all support this hypothesis.[26,27] The Bold Goals For U.S. Biotechnology and Biomanufacturing paper released by the White House in March 2023 highlights ambitious biotechnology targets across climate, food, agriculture, supply chain, health, and data. All three political interventions seek to promote the commercial applications of synthetic biology in the US rather than focus purely on scaling the backbone of basic and applied research being published in support of the emerging bioeconomy.

In the case of China's rapid acceleration in publication numbers over recent years, public policy support for synthetic biology has appeared across policy documents since at least 2010 when the Chinese Academy of Sciences published a science and technology roadmap to 2050.[28] In this document, synthetic biology was identified as one of the four basic science initiatives likely to make transformative breakthroughs. The data shown by Figure 3b appears to evidence a rapid acceleration in basic science capability in China indicating that significant funding has been dedicated to the basic science aspects of scaling synthetic biology. This scaling of basic research may be beginning to change with the 2021 Five Year plan being the first to identify the need to secure protein supply through cellular agriculture and synthetic dairy.[29,30] Meanwhile, the 2023 meeting of The National People's Congress and the Chinese People's Political Consultative Conference has boosted the primacy of science and technology in China and promised an increase in funding levels by 2%, according to a draft budget.[31] This additional funding is expected to go to areas like artificial intelligence and biotechnology, which are likely to further amplify China's work on the AI-driven analysis of sequenced data and the translation of this into automated in silico bio-design protocols. A salient feature of the discourse that emanated from the 2023 "two sessions" event in China underscores the escalating securitization narrative. The country is now publicly emphasizing the pivotal role that scientific and technological advancements can play in fortifying a security agenda, encompassing critical areas such as food and energy security.[32] This is likely to have an ongoing impact on the scale of synthetic biology investment, research, and publication output in China across the coming five years.

Australia is an acknowledged latecomer to the discipline of synthetic biology but has begun to invest significantly in building national capacity in basic and applied research, this is now beginning to right shift into the development of a nascent commercial sector. Key points in this timeline include Macquarie University joining the International Research Consortium Yeast 2.0 in 2013−14, the creation of the Commonwealth Scientific and Industrial Research Organisation's (CSIRO) Future Science Platform in synthetic biology beginning in 2016 and the funding of the Centre of Excellence in 2019.[33,34] Across this same timeframe the Australian Council of Learned Academies published an Horizon Scan on synthetic biology and the CSIRO published a national roadmap for the discipline.[35] Each of these events correlates with an increase in the scale of Australian research as borne out in the data.

Meanwhile, the United Kingdom has had mature public policy support for its synthetic biology research, both in the applied and commercial domains, for more than a decade. The UK government published a national synthetic biology roadmap in 2012,[36] created a national synthetic biology leadership council that same year, released a national strategic plan for the discipline in 2016,[37] and throughout that time SynBiCite, launched in 2013, has provided national commercialization support in collaboration with the London Biofoundry.[36,38,39] Indeed, the Centre for SynBio Research and Innovation (CSynBI) at Imperial College was founded in late 2008 by the Engineering and Physical Sciences Research Council. This goes some way to show the difference in public policy maturity levels between the United Kingdom and Australia, with comparative public policy funding decisions and interventions occurring in Australia up to a decade later than when they first happened in the UK.

Networks. Synthetic biology is an interdisciplinary field with collaborations across multiple disciplines, including biology, engineering, chemistry, robotics, ethics, and computer science. In recent years, the importance of networks to support successful scientific research in synthetic biology has become increasingly apparent.[40] These networks can take many forms and draw from many sources, including academic and public/private partnerships, such as AlphaFold, the Global Biofoundry Alliance, the student international genetically engineered machine competition (iGEM), community standards such as the synthetic biology open language (SBOL), or BioBricks, research consortia such as the first synthetic eukaryotic genome project (yeast 2.0) or the attempts to build a fully synthetic cell (BaSyC https://www.basyc.nl, Build-A-Cell https://www.buildacell.org), and online communities.[41−45]

Collaborations can span different organizational levels but always start and end with individual members of the research community. Interconnectedness varies with some researchers co-authoring a lot of papers and being connected to a large network not only consisting of local researchers. Figure 6a displays a co-authorship network of the top 750 most connected authors among the data set with at least five publications where size corresponds to normalized citation count, color is determined according to the cluster the researcher can be grouped into, and proximity between two nodes in the graph indicates similar co-authorship patterns.

Researchers with high normalized citation counts are typically situated near the center of the network and have comparable co-authorship patterns (closeness in the graph), indicating a strong degree of connectivity. Conversely, citation counts decrease toward the periphery of the network. Anecdotal observations suggest that researchers collaborate most frequently with colleagues who are in their immediate geographic vicinity, as evidenced by highly connected clusters of scholars, such as Kim Sun, Cho Byung-Kwan, and Kim Haseong, who are primarily affiliated with Korean institutions. Similar patterns can be observed around researchers such as Derek Woolfson (UK), Eriko Takano (UK), Craig Venter (USA), B. Lindberg-Møller (Denmark), and Matthew Chang Wook (Singapore).

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

An examination of the top 10 researchers in terms of citations reveals that most of these scholars are in the US and are all are male. Jay Keasling (Berkeley, 16.2k), George Church (Harvard, 12.2k), Jonathan Weissman (MIT, 11.1k), Lei Qi (Stanford, 11.1k), James Collins (MIT, 10.9k), Wendell Lim (UCSF, 10.5k), Daniel Gibson (10.3k, JCVI), Christopher Voigt (MIT, 9.9k), Hamilton Smith (JCVI, 9.8k), and Hutchison Clyde (9.7k, JCVI). In terms of connectedness in the co-authorship network the top 10 authors in descending order are as follows: Jay Keasling (Berkeley), Martin Fussenegger (ETH), Huimin Zhao (UIUC), Michael Jewett (Stanford), Christopher Voigt (MIT), Sun Chang Kim (KAIST), Victor De Lorenzo (CSIC), Guocheng Du (Jiangnan University), Sang Yup Lee (KAIST), and Timothy Lu (MIT). This is consistent with Oldham et al. 2012 and Raimbault et al. in 2016, both of which are highlighting similar authors and their centrality in the field.[20,21] Interestingly, S. Benner and M. Fussenegger occupied the front position in terms of publication amount in 2012, whereas our data shows that the field has diversified and shifted toward a larger number of prominent and high output researchers, possibly as a function of its increasing popularity and topic heterogeneity. A co-authorship analysis can provide a valuable lens into the organizational structure of scientific research. In Korea and Europe, junior faculty members often integrate into broader research collectives steered by senior professors. In the U.S. and other countries, senior professors run large laboratories with a lot of staff who then go on to become new professors at other institutions. These organizational schemes inherently amplify the connectivity and citation count associated with senior professors, reinforcing their prominent standing within the academic community and our analysis. Additionally, researchers who have been active in the field longer as they have had more time to produce papers and for those to accrue potential citations, as exemplified by Supporting Information, Figure S4. Therefore, this type of analysis can provide a perspective on the past structure of the field but cannot necessarily be projected into the future. An alternative network analysis approach relying on citations instead of co-authorship using the same data set excluding reviews, yields broadly similar results (Supporting Information, Figure S5).

When zooming out and observing the community from the national level, six clusters emerge. One dominated by the United States (in cyan) containing Canada and Israel. An Asian cluster with, among others, the Peoples Republic of China, South Korea, Singapore, India, and Australia. European countries are organized in the blue, green, and orange clusters. The yellow cluster contains mostly Spanish speaking countries as well as France. The US features more than twice the link strength and is connected to more countries as illustrated by its central position in the network. Prominent countries in terms of research output tend to be located more toward the center of the network and are better connected (Figure 6b).

The previously alluded-to local focus of collaboration and premier position of elite US institutions is also corroborated by observing co-authorship networks on the institutional level (Figure 6c). Korean institutes have high local connectedness mirroring, as shown in Figure 6a, with its central node around the Korea Advanced Institute of Science and Technology (in cyan). Chinese institutions tend to group around the Chinese Academy of Sciences (colored blue). European universities are mostly represented by red nodes, while orange nodes predominantly represent Japanese institutions, and yellow nodes denote Australian institutions. Of particular interest is the brown cluster, which includes several industry-focused universities and institutions, such as UC Berkeley, Technical University of Denmark, Joint Bioenergy Institute, Chalmers University, and Shenzhen Institute of Advanced Technology. European institutions are generally grouped into the red cluster and American institutions in either the green or purple cluster. Broadly speaking, connections among local institutions are the strongest but institutions that produce a lot of synthetic biology publications such as MIT (594 documents) tend to have many connections not only locally but also internationally. This all highlights the importance of connectivity among authors, countries, and institutions for high impact publications and progress of the field overall. Particularly in times of increasing geopolitical tensions, we argue for maintaining an open and collaborative environment for the free exchange of ideas.

SUMMARY AND FUTURE OUTLOOK

Due to the absence of an official definition for synthetic biology and the fact that many definitions come from self-identified synthetic biologists, there is a notable variety in how it is defined. For example, the final opinion on synthetic biology from the European Commission lists 35 definitions of synthetic biology.[46] Our findings, especially contrasted with findings from Oldham et al. and Raimbault et al., suggest that the definition of synthetic biology is anything but final. While synthetic biology broadly centers around topics such as metabolic engineering, genetic engineering, and (synthetic) genomics, it has grown to also encompass a more heterogenic range of topics. We believe this can partly be attributed to growing capabilities enabling several novel fields of application and research. Nevertheless, we also consider the hype associated with this set of technologies leads to researchers from adjacent fields utilizing the label synthetic biology to attract funding, talent, and publicity. We think the trend to diversification of research focus will continue and will result in synthetic biology having a significant influence on the chemical, medical, and materials industry in addition to its core competencies in biotechnology. Considering the broad implications for our lives and, therefore, also geopolitical relevance, it comes as no surprise that this technology is also the subject of fierce competition between nation states. We argue that an open and collaborative environment is crucial for progress in synthetic biology but also to ensure equitable access for those nations currently not pursuing active industrial biotechnology policy. Considering the trend to ever higher throughput experimental automation and data gathering, access to data will be crucial. As Kai-Fu Lee put in his book AI Superpowers: China, Silicon Valley, and the New World Order: “In deep learning, there’s no data like more data.” Something similar will become the norm for synthetic biology and the champions of tomorrow will be those who can gather, analyze, and deploy insights to market. We believe the shift we observed toward commercialization in the US will be a natural evolution of the field, which other countries will follow. This not only will bring the wonders of synthetic biology to market but also possibly result in a shift of knowledge and innovation from universities to private labs analogous to the shift of AI research from academy to big tech companies. This shift to commercialization inevitably results in knowledge and innovations not necessarily being available to the broader scientific community through publications and could lead to a

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

less open research environment. Finally, we hope that our work and open-source code can serve as a resource for researchers to explore the field in more detail, choose suitable journals for their publications relying not only on anecdotal evidence and impact factors, and possibly as a base to expand connectivity between the nodes of synthetic biology across national, institutional, and individual dimensions. Not only is this study relevant to researchers but we also hope that it could help inform industrial policy for governmental institutions.

METHODS

Data Query and Preprocessing. 29 331 research articles were gathered from the Clarivate Web of Science (WOS) database using the query: *((((((((((((((((((((((((((((ALL=("engineered genetic parts")) OR ALL=("genetic systems")) OR ALL=("engineered metabolic pathways")) OR ALL=("engineered biosensors")) OR ALL=("genetic circuits")) OR ALL=("synthetic biology")) OR ALL=("synthetic gene network")) OR ALL=("synthetic gene")) OR ALL=("standard biological parts")) OR ALL=("synthetic genome")) OR ALL=("synthetic genomics")) OR ALL=("minimal cells")) OR ALL=("minimal genome")) OR ALL=("synthetic circuits")) OR ALL=("artificial gene networks")) OR ALL=("synthetic genes")) OR ALL=("synthetic life")) OR ALL=("genetic circuit design"))) OR ALL=("molecular biology engineering")) OR ALL=("synthetic evolution")) OR ALL=("biosynthetic engineering"))) OR ALL=("engineered biological pathways")) OR ALL=("engineered metabolic pathway")) OR ALL=("synthetic cell")) OR ALL=("synthetic cells")*. Therefore, articles containing an exact match in any field to any of the search terms will have been included in the query. For example, mention of synthetic biology will be included but not an article just including synthetic in one sentence and biology in another.

Out of 29 331 WOS data points, 18 704 were articles, 5053 were reviews, 3516 meeting abstracts, 936 editorial material, and 768 proceedings paper. The remaining were books, corrections, news items, notes, and letters. We focused on articles, reviews, editorial material, and proceedings papers for our study. Out of 24 525 downloaded WOS articles, reviews, editorial material, or proceeding papers, 1602 were missing either abstract, publication year, or address information, all essential for further downstream analysis, and therefore excluded. As a result, 23 110 journal publications were subjected to topic and network analysis.

Topic Modeling. In this study, we utilized the modular BERTopic neural topic modeling procedure to classify the abstracts of articles into different topic clusters. First, the all-mpnet-base-v2 sentence transformer model was applied to embed words and sentences into a 768-dimensional vector space. Subsequently, we employed the UMAP algorithm[47] to reduce the dimensionality of the data, followed by HDBSCAN for clustering.[48] For each cluster, we generated a normalized bag-of-words representation to determine which words occur within the cluster. We then employed a class-based version of the TF-IDF algorithm[16] to identify the words that best represent the topic of each cluster. To improve the topic representations, we deployed OpenAIs GPT-3.5-turbo, the same model behind ChatGPT, to generate topic descriptions accurately representing the topic cluster.[19] Following static modeling, a distinct dynamic instance of BERTopic was utilized to investigate topics related to synthetic biology over a temporal dimension and investigate trends. More explicit details on the topic modeling parameters used in each of the steps of this study can be found at: https://colab.research.google.com/drive/17Vdm004B26G9-M-WBVuDIkpG1F0kdWi5?usp=sharing or https://github.com/SynBioExplorer/SynBio_NLP_analysis/.

Metadata. To examine the distribution of synthetic biology research globally, we conducted a comprehensive analysis of publication metadata. Specifically, we grouped metadata by journal to extract relevant information on the publications in which synthetic biology content appeared. From the address section of each publication, we parsed country and zip codes and geolocated them by matching them to the simplemaps.com world cities database. During this analysis, each collaborating location was displayed separately, resulting in 68 868 locations in total, with an average of approximately 3 locations per paper. By mapping the geographic location of authors and the countries with which they are affiliated, we were able to identify patterns of collaboration and gain insights into the global distribution of synthetic biology research.

Networks. To investigate co-authorship networks and the relationship between organizations and countries, we utilized VOSviewer.[49] This software enabled us to construct a network of co-authorship relationships and to visualize the strength and frequency of these connections. The strength of the links between two entities was determined by the number of co-authorships they shared. Entities with more co-authorships were linked by stronger lines, while entities with fewer connections were linked by weaker lines. The size of each entity was scaled to represent its normalized citation counts. To enhance the clarity of the visualization, we applied filters to the data. For the authorship network, we only included authors with at least five publications. We restricted the visualization to the top 750 most connected authors to focus on the most influential contributors. For institutions, we selected only those with at least 35 documents and reduced the visualization to the top 200 most connected institutions to highlight the most prominent research organizations. Similarly, for countries, we limited the visualization of the countries to the top 75 to emphasize the most significant players in the research landscape.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.3c00192.

URLs to interactive graphs, word cloud keyword visualization, heatmap visualizing individual similarity between BERTopics, bar chart displaying BERTopic frequency within the data set, co-authorship network of most connected authors and their respective average publication year, citation network of most connected authors, and citation network of most cited papers (PDF)

AUTHOR INFORMATION

Corresponding Author

Ian Paulsen − *School of Natural Sciences, Macquarie University, Sydney, New South Wales 2109, Australia; ARC Centre of Excellence in Synthetic Biology, Sydney, New South Wales 2109, Australia;* orcid.org/0000-0001-9015-9418; Email: ian.paulsen@mq.edu.au

https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241

Authors

Felix Meier − *School of Natural Sciences, Macquarie University, Sydney, New South Wales 2109, Australia; ARC Centre of Excellence in Synthetic Biology, Sydney, New South Wales 2109, Australia;* orcid.org/0009-0000-2858-6259

Thom Dixon − *ARC Centre of Excellence in Synthetic Biology, Sydney, New South Wales 2109, Australia; School of Social Sciences, Macquarie University, Sydney, New South Wales 2109, Australia*

Tom Williams − *School of Natural Sciences, Macquarie University, Sydney, New South Wales 2109, Australia; ARC Centre of Excellence in Synthetic Biology, Sydney, New South Wales 2109, Australia;* orcid.org/0000-0002-0594-3441

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.3c00192

Author Contributions

Felix Meier conceptualized the study, conducted the computational analysis, and authored the majority of the initial manuscript. Thom Dixon authored the policy section and provided valuable feedback throughout the writing process. Tom Williams and Ian Paulsen provided guidance on the overall direction of the paper. All authors participated in revising the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

ChatGPT was used to rephrase some sentences but not to write entire paragraphs de novo.

■ ABBREVIATIONS

NLP, natural language processing; SynBio, synthetic biology; WOS, Clarivate Web of Science

■ REFERENCES

(1) Cameron, D. E.; Bashor, C. J.; Collins, J. J. A Brief History of Synthetic Biology. *Nat. Rev. Microbiol.* **2014**, *12*, 381−390.

(2) Stephanopoulos, G. Synthetic Biology and Metabolic Engineering. *ACS Synth. Biol.* **2012**, *1*, 514−525.

(3) Paddon, C. J.; Westfall, P. J.; Pitera, D. J.; Benjamin, K.; Fisher, K.; McPhee, D.; Leavell, M. D.; Tai, A.; Main, A.; Eng, D.; Polichuk, D. R.; Teoh, K. H.; Reed, D. W.; Treynor, T.; Lenihan, J.; Jiang, H.; Fleck, M.; Bajad, S.; Dang, G.; Dengrove, D.; Diola, D.; Dorin, G.; Ellens, K. W.; Fickes, S.; Galazzo, J.; Gaucher, S. P.; Geistlinger, T.; Henry, R.; Hepp, M.; Horning, T.; Iqbal, T.; Kizer, L.; Lieu, B.; Melis, D.; Moss, N.; Regentin, R.; Secrest, S.; Tsuruta, H.; Vazquez, R.; Westblade, L. F.; Xu, L.; Yu, M.; Zhang, Y.; Zhao, L.; Lievense, J.; Covello, P. S.; Keasling, J. D.; Reiling, K. K.; Renninger, N. S.; Newman, J. D. High-Level Semi-Synthetic Production of the Potent Antimalarial Artemisinin. *Nature* **2013**, *496*, 528−532.

(4) Gibson, D. G.; Glass, J. I.; Lartigue, C.; Noskov, V. N.; Chuang, R.-Y.; Algire, M. A.; Benders, G. A.; Montague, M. G.; Ma, L.; Moodie, M. M.; Merryman, C.; Vashee, S.; Krishnakumar, R.; Assad-Garcia, N.; Andrews-Pfannkoch, C.; Denisova, E. A.; Young, L.; Qi, Z.-Q.; Segall-Shapiro, T. H.; Calvey, C. H.; Parmar, P. P.; Hutchison, C. A.; Smith, H. O.; Venter, J. C. Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **2010**, *329*, 52−56.

(5) Elowitz, M. B.; Leibler, S. A. synthetic oscillatory network of transcriptional regulators. *Nature* **2000**, *403*, 335−338.

(6) Liu, C. C.; Jewett, M. C.; Chin, J. W.; Voigt, C. A. Toward an Orthogonal Central Dogma. *Nat. Chem. Biol.* **2018**, *14*, 103−106.

(7) Gallup, O.; Ming, H.; Ellis, T. Ten Future Challenges for Synthetic Biology. *Eng. Biol.* **2021**, *5*, 51−59.

(8) Khalil, A. S.; Collins, J. J. Synthetic Biology: Applications Come of Age. *Nat. Rev. Genet.* **2010**, *11*, 367−379.

(9) Rodrigo-Navarro, A.; Sankaran, S.; Dalby, M. J.; del Campo, A.; Salmeron-Sanchez, M. Engineered Living Biomaterials. *Nat. Rev. Mater.* **2021**, *6*, 1175−1190.

(10) Meng, F.; Ellis, T. The Second Decade of Synthetic Biology: 2010-2020. *Nat. Commun.* **2020**, *11*, 5174.

(11) Serrano, L. Synthetic Biology: Promises and Challenges. *Mol. Syst. Biol.* **2007**, *3*, 158.

(12) Ausländer, S.; Ausländer, D.; Fussenegger, M. Synthetic Biology-The Synthesis of Biology. *Angew. Chem., Int. Ed.* **2017**, *56*, 6396−6419.

(13) Asmussen, C. B.; Møller, C. Smart Literature Review: A Practical Topic Modelling Approach to Exploratory Literature Review. *J. Big Data* **2019**, *6*, 93.

(14) Cohen, K. B.; Hunter, L. Natural Language Processing and Systems Biology. *Artificial Intelligence Methods And Tools For Systems Biology*; Springer Netherlands, 2004; pp 147−173.

(15) Cambria, E.; White, B. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Comput. Intell. Mag.* **2014**, *9*, 48−57.

(16) Salton, G.; Fox, E. A.; Wu, H. Extended Boolean Information Retrieval. *Commun. ACM* **1983**, *26*, 1022−1036.

(17) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. 2023. http://arxiv.org/abs/1706.03762 (accessed Jan 24, 2023).

(18) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 2019, arXiv:10.48550/arXiv.1810.04805. https://doi.org/10.48550/arXiv.1810.04805 (accessed May 24, 2023).

(19) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners. **2020**, arXiv:10.48550/arxiv.2005.14165, accessed May 26 2023

(20) Oldham, P.; Hall, S.; Burton, G. Synthetic Biology: Mapping the Scientific Landscape. *PLoS One* **2012**, *7*, No. e34368.

(21) Raimbault, B.; Cointet, J.-P.; Joly, P.-B. Mapping the Emergence of Synthetic Biology. *PLoS One* **2016**, *11*, No. e0161522.

(22) Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; Charpentier, E. A Programmable Dual-RNA−Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **2012**, *337*, 816−821.

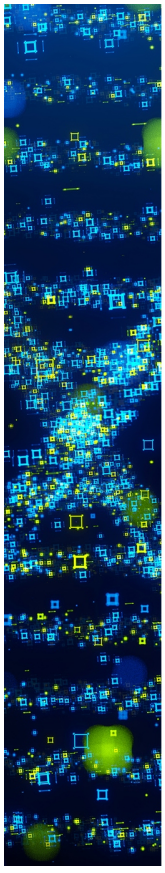
(23) Fire, M.; Guestrin, C. Over-Optimization of Academic Publishing Metrics: Observing Goodhart's Law in Action. *GigaScience* **2019**, *8*, giz053.

(24) Committee on Safeguarding the Bioeconomy. *Finding Strategies for Understanding, Evaluating, and Protecting the Bioeconomy while Sustaining Innovation and Growth; Board on Life Sciences; Board on Agriculture and Natural Resources; Board on Science, Technology, and Economic Policy; Board on Health Sciences Policy; Forum on Cyber Resilience; Division on Earth and Life Studies; Policy and Global Affairs; Health and Medicine Division; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine. Safeguarding the Bioeconomy*; National Academies Press: Washington, D.C., 2020; p 25525.

(25) Q1 Shatters Previous Synthetic Biology Investment Record - Signals Projected 2021 Investment of up to $36 Billion—SynBioBeta. https://www.synbiobeta.com/read/q1-shatters-previous-synthetic-biology-investment-record-signals-projected-2021-investment-of-up-to-36-billion (accessed March 20, 2023).

(26) House, T. W. Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-

sustainable-safe-and-secure-american-bioeconomy/ (accessed March 20, 2023).
(27) H.R.4346—117th Congress (2021-2022): Chips and Science Act | Congress.gov | Library of Congress. https://www.congress.gov/bill/117th-congress/house-bill/4346 (accessed March 20, 2023).
(28) Lu, Y. Science & Technology in China: A Roadmap to 2050: Strategic General Report of the Chinese Academy of Sciences. *Choice Reviews Online* **2010**, *47*, 5393.
(29) Mallapaty, S. China's Five-Year Plan Focuses on Scientific Self-Reliance. *Nature* **2021**, *591*, 353−354.
(30) China's 5-Year Plan is a Blueprint for the Future of Meat. Time. https://time.com/6143109/china-future-of-cultivated-meat/ (accessed March 20, 2023).
(31) Mallapaty, S. China Is Mobilizing Science to Spur Development — and Self-Reliance. *Nature* **2023**, *615*, 570−571.
(32) National People's Congress of the People's Republic of China. *Report on the Implementation of the National Economic and Social Development Plan for 2022 and the Draft National Economic and Social Development Plan for 2023*, 2023.
(33) Williams, G.; Banfield, D.; Towns, A.; Wynn, K.; Liu, M.; Cohen, J. *A National Synthetic Biology Roadmap; CSIRO Futures*; CSIRO, 2021.
(34) 2020 ARC Centre of Excellence in Synthetic BiologylAustralian Research Council. https://www.arc.gov.au/funding-research/discovery-linkage/linkage-program/arc-centres-excellence/2020-arc-centre-excellence-synthetic-biology (accessed March 20, 2023).
(35) Gray, P.; Meek, S.; Griffith, P.; Trapani, T.; Small, i.; Vickers, C. Synthetic Biology in Australia: An Outlook to 2030. *Australian Council of Learned Academies Expert Working Group Horizon Scanning Project* 2018.
(36) Groupcoordination, U. S. B. R.; Clarke, L.; Adams, J.; Sutton, P.; Bainbridge, J.; Birney, E.; Calvert, J.; Collis, A.; Kitney, R.; Freemont, P.; Mason, P.; Pandya, K.; Ghaffar, T.; Rose, N. S.; Woolfson, D.; Boyce, A. *A Synthetic Biology Roadmap for the UK*, 2012.
(37) Clarke, L.; Kitney, R. Developing Synthetic Biology for Industrial Biotechnology Applications. *Biochem. Soc. Trans.* **2020**, *48*, 113−122.
(38) Clarke, L. J.; Kitney, R. I. Synthetic Biology in the UK − An Outline of Plans and Progress. *Synth. Syst. Biotechnol.* **2016**, *1*, 243−257.
(39) Chambers, S.; Kitney, R.; Freemont, P. The Foundry: The DNA Synthesis and Construction Foundry at Imperial College. *Biochem. Soc. Trans.* **2016**, *44*, 687−688.
(40) Molyneux-Hodgson, S.; Meyer, M. Tales of Emergence—Synthetic Biology as a Scientific Community in the Making. *BioSocieties* **2009**, *4*, 129−145.
(41) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(42) Hillson, N.; Caddick, M.; Cai, Y.; Carrasco, J. A.; Chang, M. W.; Curach, N. C.; Bell, D. J.; Le Feuvre, R.; Friedman, D. C.; Fu, X.; Gold, N. D.; Herrgård, M. J.; Holowko, M. B.; Johnson, J. R.; Johnson, R. A.; Keasling, J. D.; Kitney, R. I.; Kondo, A.; Liu, C.; Martin, V. J. J.; Menolascina, F.; Ogino, C.; Patron, N. J.; Pavan, M.; Poh, C. L.; Pretorius, I. S.; Rosser, S. J.; Scrutton, N. S.; Storch, M.; Tekotte, H.; Travnik, E.; Vickers, C. E.; Yew, W. S.; Yuan, Y.; Zhao, H.; Freemont, P. S. Building a Global Alliance of Biofoundries. *Nat. Commun.* **2019**, *10*, 2040.
(43) Smolke, C. D. Building Outside of the Box: IGEM and the BioBricks Foundation. *Nat. Biotechnol.* **2009**, *27*, 1099−1102.
(44) Galdzicki, M.; Clancy, K. P.; Oberortner, E.; Pocock, M.; Quinn, J. Y.; Rodriguez, C. A.; Roehner, N.; Wilson, M. L.; Adam, L.; Anderson, J. C.; Bartley, B. A.; Beal, J.; Chandran, D.; Chen, J.; Densmore, D.; Endy, D.; Grünberg, R.; Hallinan, J.; Hillson, N. J.; Johnson, J. D.; Kuchinsky, A.; Lux, M.; Misirli, G.; Peccoud, J.; Plahar, H. A.; Sirin, E.; Stan, G.-B.; Villalobos, A.; Wipat, A.; Gennari, J. H.; Myers, C. J.; Sauro, H. M. The Synthetic Biology Open Language (SBOL) Provides a Community Standard for Communicating Designs in Synthetic Biology. *Nat. Biotechnol.* **2014**, *32*, 545−550.
(45) Pretorius, I. S.; Boeke, J. D. Yeast 2.0-Connecting the Dots in the Construction of the World's First Functional Synthetic Eukaryotic Genome. *FEMS Yeast Res.* **2018**, *18*, foy032.
(46) European Commission. Directorate General for Health and Consumers. *Opinion on Synthetic Biology I: Definition*; Publications Office: LU, 2014.
(47) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020. https://doi.org/10.48550/arXiv.1802.03426, accessed March 28, 2023.
(48) Campello, R. J. G. B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V. S., Cao, L., Motoda, H., Xu, G., Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Series, Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; Vol. *7819*, pp 160−172.
(49) van Eck, N. J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523−538.



https://doi.org/10.1021/acssynbio.3c00192
ACS Synth. Biol. 2023, 12, 3229−3241