Please cite as: Zhang, Y., Wu, M., Hu, Z., Ward, R., Zhang, X., and Porter A. (2020), Profiling and predicting the problem-solving patterns in China's research systems: A methodology of intelligent bibliometrics and empirical insights, *Quantitative Science Studies*, DOI: 10.1162/qss_a_00100.

Profiling and predicting the problem-solving patterns in China's research systems: A methodology of intelligent bibliometrics and empirical insights

Yi Zhang[1,*], Mengjia Wu[1], Zhengyin Hu[2], Robert Ward[3], Xue Zhang[2], Alan Porter[3,4]

[1]Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia.

[2]Chengdu Library and Information Centre, Chinese Academy of Sciences, China.

[3]Program in Science, Technology & Innovation Policy (STIP), Georgia Institute of Technology, USA.

[4]Search Technology, Inc., USA.

Email: yi.zhang@uts.edu.au (corresponding author); mengjia.wu@student.uts.edu.au; huzy@clas.ac.cn; rward48@gatech.edu; zhangxue@mail.las.ac.cn; alan.porter@isye.gatech.edu.

ORCID: 0000-0002-7731-0301 (Yi Zhang), 0000-0003-3956-7808 (Mengjia Wu), 0000-0002-5699-9891 (Zhengyin Hu), 0000-0001-6876-8620 (Robert Ward), 0000-0003-3329-8977 (Xue Zhang), 0000-0002-4520-6518 (Porter Alan).

Abstract

Uncovering the driving forces, strategic landscapes, and evolutionary mechanisms of China's research systems is attracting rising interest around the globe. One such interest is to understand the problem-solving patterns in China's research systems now and in the future. Targeting a set of high-quality research articles published by Chinese researchers between 2009 and 2018, and indexed in the Essential Science Indicators database, we developed an intelligent bibliometrics-based methodology for identifying the problem-solving patterns from scientific documents. Specifically, science overlay maps incorporating link prediction were used to profile China's disciplinary interactions and predict potential cross-disciplinary innovation at a macro level. We proposed a function incorporating word embedding techniques to represent subjects, actions, and objects (SAO) retrieved from combined titles and abstracts into vectors and constructed a tri-layer SAO network to visualize SAOs and their semantic relationships. Then, at a micro level, we developed network analytics for identifying problems and solutions from the SAO network, and recommending potential solutions for existing problems. Empirical insights derived from this study provide clues to understand China's research strengths and the science policies beneath them, along with the key research problems and solutions Chinese researchers are focusing on now and might pursue in the future.

Keywords: Bibliometrics; Text analytics; Network analytics; Research evaluation; Chinese research policy.

1. Introduction

China's accelerating development in science, technology and innovation over the past decades has sparked interest in the driving forces, strategic landscapes, and evolutionary mechanisms behind it. Profiling China's achievements in science and technology (Zhou & Leydesdorff, 2006; Mu & Qu, 2008) and discussing issues and challenges for refining China's research systems (Cao & Suttmeier, 2017; Tang, 2019) draw research policy analysis attention. Bibliometrics and bibliometric data sources (e.g., scientific publications and patents) are recognized as sturdy tools to identify and answer research questions about the research landscape – for example, the influence of national scientific funding on emerging research (Huang et al., 2016)] or empirical studies on examining the research strengths of China's specific practical sectors (Huang et al., 2014). Further, with the big data boom and rise of artificial intelligence, bibliometrics has benefited greatly from advanced information tools. Zhang et al. (2020a) call these "development and applications of intelligent models for recognizing patterns in bibliometrics" *Intelligent Bibliometrics*, highlighting its tremendous capability and potential to lead a new research thread in bibliometrics.

Returning to the rising interest in China's far-reaching impact on global science, technology and the economy, one specific question is: What are the problem-solving patterns in China's research systems now and in the future? Identifying such patterns will help us understand the mechanisms of China's research systems and China's competitive advantage on the international stage. Further, more knowledge of which problems Chinese researchers solve, and how, could support policy studies. Those could help uncover potential driving forces in China's science policy, and eventually benefit the global technological evolution and economic revolution.

Some studies have touched on intelligent bibliometrics by combining semantic approaches and expert knowledge to identify insights from scientific documents. Some analysts have employed subject-action-object (SAO) analysis as an effective tool for extracting patterns (e.g., independent problems and solutions) (Heffernan & Teufel, 2018; Yang et al., 2018). However, how to develop bibliometric methods to recognize problem-solving patterns in a convincing and semi-automatic way remains elusive. Moreover, how to predict the directions of advance of such science policy patterns is a further challenge.

To address these questions, we assembled a dataset of 27,971 research articles published by Chinese researchers between 2009 and 2018 and indexed as "top papers" in the Essential Science Indicators database on the Web of Science (ESI WoS). We then developed an intelligent bibliometric methodology to profile the problem-solving patterns of China's research systems and predict potential solutions in the future. We analyzed the macro-level landscape and investigated China's disciplinary interactions through science overlay maps (Rafols et al., 2010). The results revealed the evolution of disciplinary emphases in China's research systems. At the micro-level, we followed the assumption raised in one of our pilot studies that problem-solving patterns are reflected in SAO structures (Zhang et al., 2014a), and, hence, we constructed a tri-layer SAO network with subjects, actions, and objects, each in their own layer. Word embedding techniques are incorporated for representing SAOs and network analytics, such as community detection and link prediction, are then used to identify the problem-solving patterns and predict potential connections between existing problems and possible solutions. The empirical results from this study should provide insights into China's research systems. They should be of interest to those studying, devising policies, managing, and/or engaging in China's science, technology, and innovation processes.

The rest of this paper is organized as follows. Section 2 reviews previous studies on advanced bibliometric techniques. Section 3 presents the intelligent bibliometrics-based methodology in detail. An empirical study on profiling the problem-solving patterns of China's research systems and predicting potential solutions is given in Section 4. Section 5 discusses the technical implications and possible applications of the proposed method and concludes this study.

2. Related work: Advanced bibliometric techniques

Dating back to the 1990s, van Raan (1996) highlighted the benefits of advanced bibliometrics with publication and citation data to gain "insights into the international position of actors at the research front in terms of influence and specializations, as well as into patterns of scientific communication and processes of knowledge dissemination", rather than "only numbers". Compared to using the traditional bibliometric indicator -- citation statistics -- text segmentation, with the aid of natural language processing techniques has provided a new angle of conducting topic analysis in terms of semantics (Porter & Detampel, 1995). Interest in analyzing the full text of publications was raised in the early 2000s (Glenisson et al., 2005) and, together with sentiment analysis and behavior analysis, it is an emerging topic in the bibliometric community (Boyack et al., 2018).

Along with the engagement of new data sources and indicators, such as technology opportunity analysis (Ma et al., 2014)[1], technology roadmaps (Li et al., 2015), and triple helix models to describe university-industry-government collaborations (Leydesdorff & Zhou, 2014; Zhang et al., 2014b), the interactions between bibliometrics and information technologies are increasing. In turn, bibliometric solutions are becoming more effective – e.g., large-scale data analytics and mapping (Boyack et al., 2011; Börner et al., 2012), accurate knowledge extraction and representation (Zhang et al., 2018b; Zhang et al., 2019a), full-text analytics (Boyack et al., 2018), and social network analytics (Yan & Guns, 2014; Rost et al., 2017). Driven by diverse practical needs, incorporating computational models, particularly artificial intelligence techniques, with bibliometric indicators and approaches is spearheading new research frontiers – for example, information visualization enhances the ability and adaptability of decision support (Chen, 2006; Waltman et al., 2010). This research route moves forward by either developing more effective algorithms, approaches and tools for visualization (Ping & Chen, 2018; Chen & Song, 2019) or facilitating network analytics to uncover latent relationships by analyzing the topological structures of science maps (Rost et al., 2017; Zhang et al., 2018c). Specifically, SAO analysis with the capability of understanding syntax from sentences receives

[1] Technology opportunity analysis was introduced by Porter and Detampel (1995), highlighting the identification of opportunities related to technological R&D (e.g., key technological components, inventors/owners of a technology, and key players of a technological area), and has been broadly extended to a wide range of studies in technology analysis, assessment, and forecasting.

great attention from the communities of bibliometrics and technology management. For instance, Zhang et al. (2014a) defined problem-solving patterns from scientific documents based on TRIZ theory[2], SAO analysis, and a substantial amount of expert knowledge. Yang et al. (2018) followed a similar approach, constructing an SAO network and using indicators of the network's topological structure to identify independent problems or solutions. Comparably, Heffernan and Teufel (2018) applied a set of classification approaches for distinguishing problems or solutions through supervised learning and a feature space specifically designed for the task.

Topic analysis, as another stream in advanced bibliometrics, has gained from topic models, which exploit latent Dirichlet allocation and its extensions for performing unsupervised clustering tasks (Yau et al., 2014; Suominen & Toivanen, 2016). In parallel, community detection approaches, which are associated with network analytics, group similar nodes as topics based on their topological features (Waltman & Van Eck, 2013; Huang et al., 2018). Incorporating community detection with word embedding techniques has led to novel solutions for knowledge representation and topic extraction (Zhang et al., 2018b). Further, as a key sub-area of topic analysis, topic detection and tracking can be traced back decades (Allan, 2002), but investigating changes in topics over time has long been a challenge to not only the bibliometric community but also a wide range of practical sectors. Machine learning techniques and advanced data analytics are bringing new thoughts and tools for handling these issues – for example, Tang & Popp (2016) studied technological change through a learning process, while Zhang et al. (2017) identified predecessor-descendant relationship over time through streaming data analytics.

Referring to the description of advanced bibliometrics given by van Raan (1996), we define these techniques, approaches, and methodologies of advanced bibliometrics based on computational models (particularly advanced data analytic techniques and artificial intelligence techniques, such as network analytics, streaming data analytics, fuzzy systems, and machine learning) as *intelligent bibliometrics*, highlighting "the development and application of intelligent models for recognizing patterns in bibliometrics" (Zhang et al., 2020a).

3. Methodology

The purpose of this study is to propose an intelligent bibliometrics-based methodology for profiling the problem-solving patterns in China's research systems and predicting possible solutions in future. The framework of this method is given in Figure 1.

The data used in the analysis are the bibliographical information in scientific articles retrieved from WoS, such as titles, abstracts, author keywords, keywords plus (a unique field in the WoS database[3] containing terms that frequently appear in the titles of an article's references), WoS categories, and affiliations. Note that the methodology is adaptable to other bibliometric

[2] TRIZ stands for the theory of inventive problem solving. Specifically, Zhang et al. (2014a) projected problems, solutions, and the type of solutions as objects, subjects, and actions (i.e., verbs) respectively, and thus, profiled problem-solving patterns in a semantic way.

[3] More information on KeyWords Plus could be found on the website: https://support.clarivate.com/ScientificandAcademicResearch/s/article/KeyWords-Plus-generation-creation-and-changes?language=en_US

databases by adapting the specific tags. For example, WoS categories can be, to some extent, replaced by International Patent Classification codes, and keywords retrieved from titles and abstracts can take the place of keywords plus.

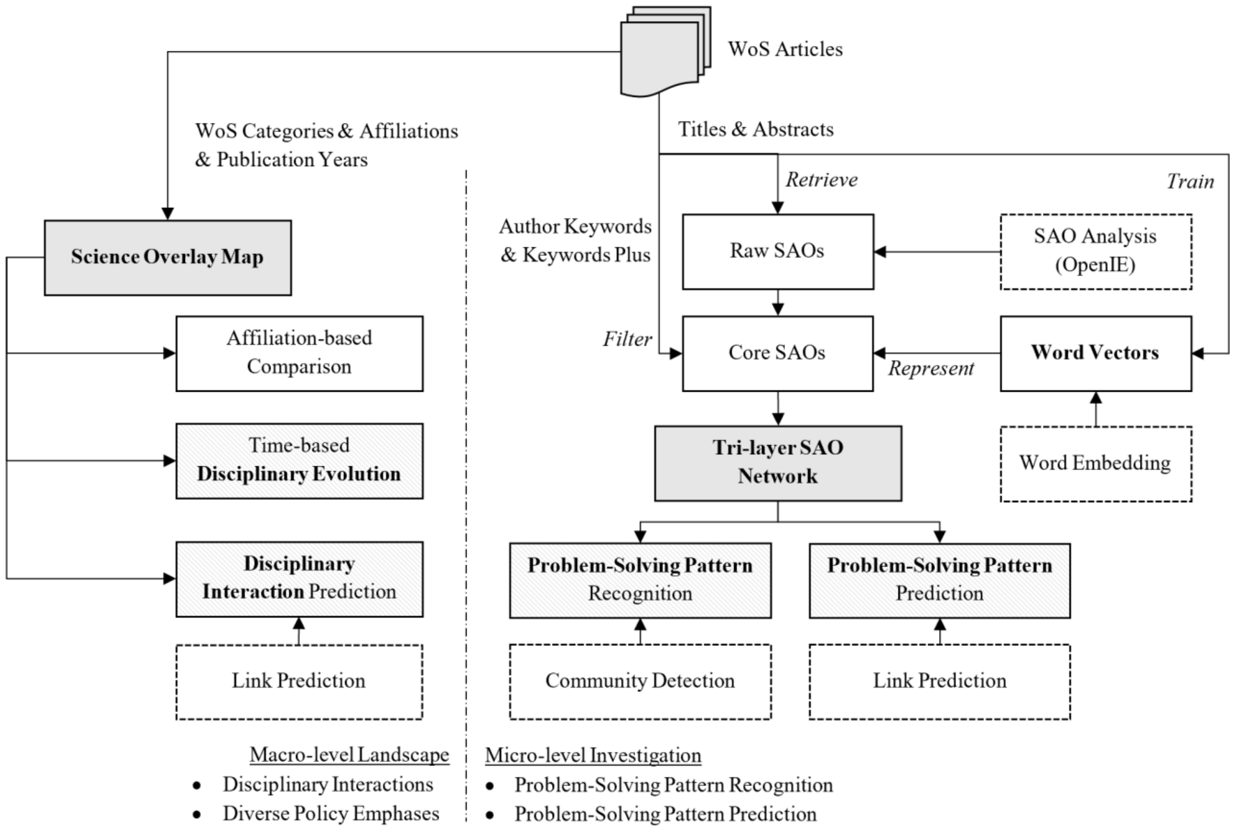


Figure 1. Research framework

3.1. Macro-level investigation: Science overlay maps for profiling disciplinary interactions

Science overlay maps, known as an effective tool for illustrating relationships among scientific disciplines (Rafols et al., 2010), can reveal empirical insights and strategic emphases (Rotolo et al., 2017). Thus, profiling the interactions between scientific disciplines, as well as the diverse policy emphases within China's research systems, and predicting potential cross-disciplinary innovation could benefit individual researchers in reviewing the landscape of related disciplines and extending their knowledge bases for further innovative activities. More importantly, in terms of science policy, these insights could support decisions of governments and their agencies to act pre-emptively (e.g., scoping national R&D and strategic plans by identifying emerging disciplines/directions and allocating research investments). Given these circumstances, we design this macro-level investigation incorporating science overlay maps with an approach of link prediction. Specifically, based on WoS categories and their co-occurrence statistics, we construct a base map that illustrates China's disciplinary interactions from a bird's eye view, and then conduct three tasks:

- *Compare the emphasis of policies between research institutions and university systems*: Targeting the most representative entities at the top levels of research, we generate a science overlay map for each selected institution. We compare these maps to gain empirical insights for understanding the diverse emphases among China's research institutions and university systems in terms of science policy.

- *Track the evolution of scientific disciplines*: We can divide the entire dataset into a set of smaller sequential datasets and generate a science overlay map for each time period. From changes in the nodes on the map, we can track the evolution of various disciplines.
- *Predict disciplinary interactions*: We introduce an algorithm of link prediction with an index of resource allocation (Zhang et al., 2020b) to make predictions by considering the base map as a complex network where each node represents a discipline, and each weighted edge represents the co-occurrence frequency between connected nodes. This process infers missing links between existing nodes, which represent potential interactions between disciplines, and results in a ranked list of discipline pairs based on the weights of their edges. High ranking pairs indicate likely cross-disciplinary directions in the future based on China's current scientific strengths.

3.2. Micro-level investigation: Subject-action-object analysis for recognizing and predicting problem-solving patterns

The subject-action-object (SAO) structure of a sentence is key to translating free text into structured formats (Zhang et al., 2014a). SAO is the most basic grammatical syntax in English, following the form "someone [subject] did [action] something [object]". Our micro-level investigation concentrates on SAO strings retrieved from raw text in the combined title and abstract fields of the articles. However, this SAO analysis includes two novel techniques: 1) enhancing the representation of SAOs by incorporating word embedding techniques; and 2) constructing a tri-layer SAO network and exploiting network analytics to uncover insights from the network's topological structures. The assumptions and steps of the method follow.

Definition 1: $T(s, a, o)$ denotes an SAO structure consisting of three components c: a subject s, an action a, and an object o.

Definition 2: A tri-layer SAO network is represented as $G(G_s, G_a, G_o, E^G)$, in which G_s, G_a and G_o represent the subject, action, and object layers, respectively, and E^G is the set of edges among those layers.

Definition 3: Each layer of the network (G_s as an example) is described as $G_s(N_s, E_s)$, in which N_s is the set of nodes and E_s is the set of edges on that layer.

Step 1: Retrieve raw SAO structures $T(s, a, o)$ from the raw text via OpenIE (Angeli et al., 2015) – a well-recognized and popular tool developed by the Stanford Natural Language Processing Group.

Step 2: Filter the raw SAO structures by matching against a combined list of author keywords and keywords plus. That is, SAOs that do not contain any keywords from the combined list will be removed.

Step 3: Identify core SAOs $T'(s, a, o)$ by matching with core actions (i.e., verbs), and the stepwise results of refining core actions are given in Table 1.

Table 1. Stepwise results of 'action' cleaning and consolidation

	Step
3.1	Actions retrieved from the filtered SAOs
3.2	Rules-based cleaning and consolidation - i.e., consolidating verbs with the same stem (e.g., increase/increases/increased); removing copulas and modal verbs; and removing general verbs (e.g., suggest, introduce, and study)[1]
3.3	Removing actions with a frequency of less than 5
3.4	Dictionary-based refinement - i.e., consolidating actions based on knowledge and rules summarized in the literature[2]
3.5	Screening and selecting core actions with human intervention

Notes: 1) This is based on a thesaurus summarized from our previous experiments and knowledge. 2) Two key sources are applied to help summarize key verbs: a project Semantic Knowledge Representation granted by the US National Library of Medicine summarized 30 key ontological predicate definitions as a semantic predication gold standard for the biomedical literature (Kilicoglu et al., 2011); and a platform provided by AULIVE Inc. based on patent analysis summarized 37 key functional verbs[4].

Step 4: Apply a Word2Vec approach (Mikolov et al., 2013) to the raw text to represent each individual word as an abstract vector $\overrightarrow{\theta_w}$.

Step 5: Assemble the abstract vectors of individual words into SAO components, and then, assemble all components into an SAO. Assembling strategies are required.

$$c \rightarrow \vec{c} = \sum_i \alpha_i \overrightarrow{\theta_i}$$

$$T \rightarrow \vec{T} = \beta_s \vec{c}_s + \beta_a \vec{c}_a + \beta_o \vec{c}_o$$

where $\vec{\theta}$ represents a word vector that comprises the vector of a component $\vec{c}$, α and β represent the weights of the word vector $\vec{\theta}$ and the component vector $\vec{c}$, respectively, and $\sum \alpha = 1$ and $\sum \beta = 1$.

Step 6: Construct the tri-layer SAO network G according to Definition 2, in which non-weighted edges E^G among layers represent their original SAO relationships, and edges E within one layer are the semantic relationships between nodes (e.g., subjects and subjects). The semantic relationships are weighted by Salton's cosine similarity between the vectors of related components, i.e.,

$$e^G(n_x, n_y) = \begin{cases} 1 \text{ if } \mathrm{n_x} \text{ and } \mathrm{n_y} \text{ belong to any same SAOs} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{else} \end{cases}$$

where $e^G(n_x, n_y)$ is the weight of an edge between any two nodes n_x and n_y in different layers of the network G.

$$e_s(n_s^x, n_s^y) = \cos\left(\vec{c}_{n_s^x}, \vec{c}_{n_s^y}\right) = \frac{\vec{c}_{n_s^x} \cdot \vec{c}_{n_s^y}}{\left|\vec{c}_{n_s^x}\right| \left|\vec{c}_{n_s^y}\right|}$$

where $e_s(n_s^x, n_s^y)$ is the weight of an edge between any two nodes n_s^x and n_s^y in the network G_s, and $e_a(n_a^x, n_a^y)$ is calculated in the same way.

Step 7: Follow an approach of fluid community detection (Parés et al., 2017) to group the subject and object layers of the network G into communities (i.e., research topics). More specifically, initialize k communities randomly with a density $d(c)$ in the range

[4] More information could be found on the website: http://www.productioninspiration.com/

(0, 1]. Then, apply the modularity approach (Newman, 2006) to help decide the optimal number k of communities. Here, a larger modularity indicates a better division of the network's communities:

$$d(c) = \frac{1}{n \in c}$$

where $n \in c$ is the number of nodes in community c .

Then maximize the aggregated density of each node to update its community information:

$$\mathcal{L}'_{n_x} = argmax_{c \in \mathcal{L}} \sum_{n_y \in \Gamma(n_x)} d(c) \times \delta(c(n_y), c)$$

$$\delta(c(n_y), c) = \begin{cases} 1, if\ c(n_y) = c \\ 0, if\ c(n_y) \neq c \end{cases}$$

where n_x is the node being updated, $\mathcal{L}'_{n_x}$ is the set of new candidate communities for n_x, $\Gamma(n_x)$ represents the set of nodes neighboring n_x, and $argmax$ is the abbreviation of the arguments of the maxima, representing an operation that seeks an argument to achieve the maximum value from a target function in a learning process.

This process iterates until the community structure of the network converges. Ultimately, each community takes on the label of the node with the highest weighted centrality (Freeman, 1978), i.e.,

$$CT_{n_x} = \frac{\sum_{n_y \in \Gamma(n_x)} e(n_x, n_y)}{N - 1}$$

where $\Gamma(n_x)$ denotes the set of nodes neighboring n_x, $e(n_x, n_y)$ is the weight of the edge between n_x and n_y, and N is the total number of nodes in the layer.

Step 8: Assuming that subjects may relate to solutions and objects may relate to problems, apply the algorithm of link prediction with a weighted index of resource allocation (Zhang et al., 2019b) over the tri-layer SAO network to infer missing edges between the two layers E_{SO}^{G}. We choose this algorithm on the assumption that every node owns a unit of unspecified resource, and the neighbors shared between two nodes are resource transmitters that allocate the resource to each node connected to it. The weight of the edge replaces the number of resource units needed to improve accuracy for the task of link prediction. The formula to predict the weight of the link WRA_{n_s,n_o} between the given subject node n_s and the object node n_o is calculated as:

$$WRA_{n_s,n_o} = \sum_{n_{cn} \in \Gamma(n_s) \cap \Gamma(n_o)} \frac{e(n_{cn}, n_s) + e(n_{cn}, n_o)}{\sum_{n_k \in \Gamma(n_{cn})} e(n_{cn}, n_k)}$$

where $\Gamma(n_s)$ denotes the set of nodes neighboring n_s, and $e(n_{cn}, n_s)$ is the weight of the edge between n_{cn} and n_s. The larger the value of WRA_{n_s,n_o}, the higher the possibility that a link will form between the two nodes in future.

The raw outputs of this procedure include a list of predicted subject-object pairs ranked by the weight of the predicted edges E_{SO}^{G} and a set of similarity matrices between problems, between solutions, and between problems and solutions.

4. Empirical study: What are the problem-solving patterns of China's research systems?

The Essential Science Indicators (ESI) database in the Web of Science (WoS) is designed to reveal emerging science trends and influential entities (e.g., papers, journals, individuals, institutions, and countries), covering a 10-year rolling file[5]. It collects two types of papers: (1) *highly cited papers* are the top 1% papers in a given discipline in a specific year, based on citation counts received from the WoS database, indicating their permanent impact; and (2) *hot papers* are the top 0.1% papers in a given discipline in the most recent two-month period, based on citation counts received from the WoS database, indicating the emerging interests of related communities. *Top papers* in the ESI database contain both types, representing a set of well-recognized and high-quality research articles in a discipline. Thus, the ESI database has been widely used for profiling research disciplines and areas (Zhang et al., 2018a; Liao et al., 2019) and evaluating the research performance of a given entity (Csajbók et al., 2007; Fu et al., 2011).

Aiming to focus on high-quality research conducted by Chinese researchers, this case study exploited the WoS ESI database, and, on November 15, 2019, assembled a dataset of 27,971 highly cited articles published by Chinese researchers between 1 January, 2009, and 31 December, 2018, with the following search criteria:

Countries/Regions = China AND Results = Top Papers

Note that considering the setting of the ESI database, Chinese researchers are affiliated with at least one Chinese institution, and both first authors and co-authors are counted. Moreover, since the strict selection criterion (e.g., top 1% of the citation counts) of those *top papers* in the ESI database, ESI in total contains approximately 155,000 top papers from the globe, and thus, the coverage of the 27,971-article dataset could be convincing for portraying the research landscape of the Chinese research system. Brief information about the dataset is given in Table 2.

Table 2. General information of the collected dataset

Indicator	*Number*	*Indicator*	*Number*
Records	27,971	Author keywords	44,263
Authors	164,445	Keywords Plus	54,801
Affiliations	17,160	WoS categories	211
Countries	175	Journals	2,427

Note: 1) WoS categories were refined by Clarivate Analytics as "research areas" in early 2019, but we used the finer grain version of the WoS Categories, given the period of study ends in 2018; and 2) the numbers provided above are all raw numbers prior to cleaning.

In addition to our main task of profiling the problem-solving patterns in China's research systems, we make overall observations from the data, which could be of interest to stakeholders in science policy and related academic researchers, or be investigated in further studies.

1) On average, each article has 5.9 authors, indicating that Chinese researchers might prefer relatively large research teams.

[5] More information on the WoS ESI could be found at the website: https://clarivate.com/webofsciencegroup/solutions/essential-science-indicators/

2) China's researchers have established collaborations with researchers from 174 countries and regions, demonstrating a high degree of collaborative diversity.
3) Articles published by China's researchers span 211 of the 254 WoS categories, indicating a high coverage of disciplines in China's research systems, which may be supported by the central government's national strategies and science policies.

Note that this statistical information is based on top papers. An analysis of all articles may produce different results.

4.1. Macro-level investigation: Science overlay maps for profiling disciplinary interactions

Science overlay maps specifically focus on the WoS Categories and profile disciplinary interactions to help uncover and understand the science policies behind those patterns. The top 30 WoS Categories in which Chinese researchers publish articles are listed in Table 3.

Table 3. Top 30 WoS categories in which Chinese researchers published articles

No.	*WoS category*	*#A*	*No.*	*WoS category*	*#A*
1	Chemistry, Multidisciplinary	5812	15	Automation & Control Systems	1099
2	Materials Science, Multidisciplinary	4451	17	Plant Sciences	729
3	Chemistry, Physical	3953	18	Biochemistry & Molecular Biology	624
4	Nanoscience & Nanotechnology	3270	19	Telecommunications	609
5	Physics, Applied	2815	20	Oncology	605
6	Engineering, Electrical & Electronic	2281	21	Computer Science, Info. Systems	543
7	Multidisciplinary Sciences	1983	22	Physics, Multidisciplinary	524
8	Physics, Condensed Matter	1978	23	Mathematics	523
9	Energy & Fuels	1969	24	Mechanics	494
10	Engineering, Chemical	1921	25	Food Science & Technology	486
11	Environmental Sciences	1862	26	Geosciences, Multidisciplinary	456
12	Engineering, Environmental	1780	27	Engineering, Mechanical	405
13	Mathematics, Applied	1225	28	Biotech. & Applied Microbiology	393
14	Computer Science, Artificial Intel.	1194	29	Green & Sustainable Science & Tech.	389
15	Automation & Control Systems	1099	30	Cell Biology	388

Note that since one journal may belong to multiple WoS categories, the set of articles assigned to WoS categories may overlap.

As shown, science disciplines, such as chemistry, physics, and biology, lead the list, followed by engineering and computer science. These rankings may reflect two drivers: 1) China has a long history of establishing policies that make the natural sciences a priority. Thus China's research systems provide more funding in those disciplines than to the social sciences (including arts and humanities); and 2) the global publication systems follow Western traditions, which can mean it is difficult for Chinese researchers to publish in high-quality social science journals due to language barriers and differences in culture and values. However, the balance between the two areas is potentially changing. We generated a science overlay map for China's research systems in Figure 2, in which China's specific research interests in chemistry, electrical and electronic engineering, applied mathematics, and multidisciplinary sciences are observed.

Further, in Appendix A, we profiled the interactions between disciplines and uncovered reasons for the observed patterns through two sets of science overlay maps and a forecasting study based on an approach of link prediction.

- Affiliation-based maps (see Appendix A-1): we generated science overlay maps for the Chinese Academy of Sciences (CAS), Tsinghua University, and Peking University, respectively, and particularly tracked the outputs of Chinese researchers in social sciences from top Chinese journals indexed by the Chinese Academy of Social Sciences (CASS). This study revealed the diversity of science policies in China's research systems between universities and government-funded research institutions.
- Time-based maps (see Appendix A-2): we zoomed in on two time periods -- an earlier period between 2009 and 2011, and a later period between 2016 and 2018 to analyze the evolution of disciplinary interactions of China's research systems.
- Link prediction (see Appendix A-3): we applied an approach of link prediction to the base map of China's research systems (i.e., Figure 2) to foresee potential cross-disciplinary interactions.

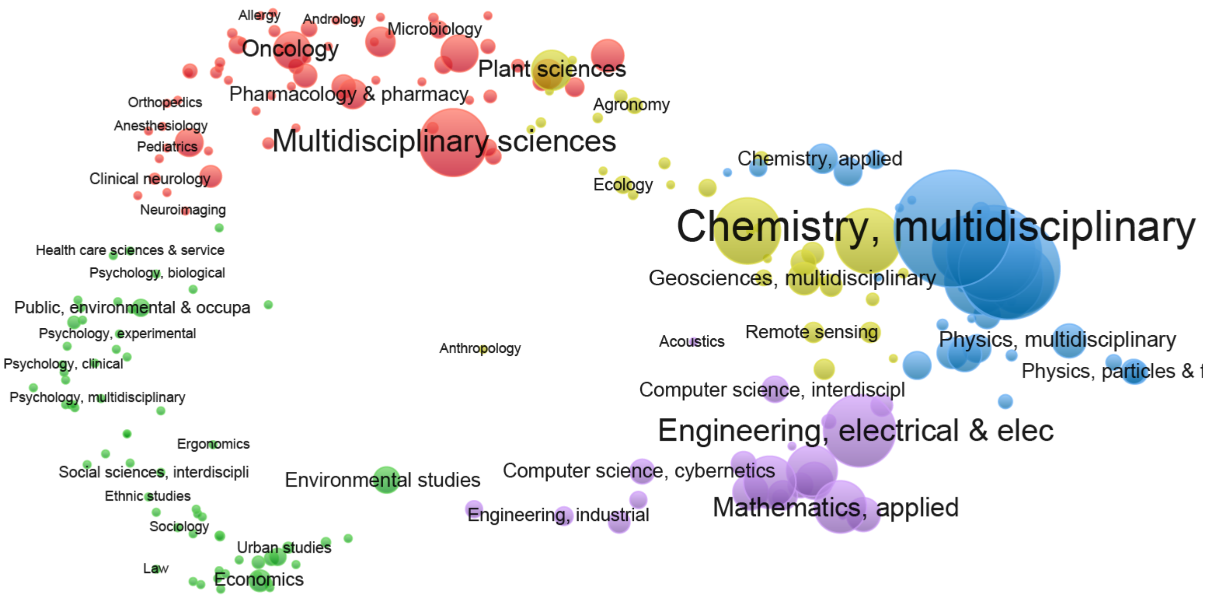


Figure 2. Science overlay map for China's research systems

We concluded our key findings from this macro-level analysis as follows:

- Over the past decade, China has pursued a balanced strategy of encouraging academic research in all scientific disciplines, but China's efforts in social science disciplines are not as advanced as that of natural sciences on the global stage.
- Interactions within natural sciences can be clearly traced for each of the three affiliations, as well as the base map. However, how Chinese researchers will conduct cross-disciplinary studies between the natural and social sciences, where gaps still exist, is elusive so far.
- Computer science and related disciplines are one of China's research strengths, and driven by artificial intelligence techniques and the visionary applications of internet of things, as well as 5G and robotics, interactions between computer science (e.g., artificial intelligence, information systems, and cybersecurity) and its applications in engineering areas, such as

electrical and electronic engineering, telecommunications, and automation are rapidly spearheading a cutting-edge direction.

- With research strengths in chemistry, biology, and material sciences, also, a cutting-edge area that holds strong interest with China's researchers is sustainable technologies -- e.g., 3D printing. New materials and novel manufacturing processes in the areas of chemical engineering and biological engineering are among the most significant innovations these days, and nanotechnologies should further enhance the practical capability of those inventions.

4.2. *Micro-level investigation: Subject-action-object analysis for recognizing and predicting problem-solving patterns*

With the aid of OpenIE (Angeli et al., 2015), we extracted 195,188 raw SAO structures from our corpus and then conducted the cleaning and consolidation process (Steps 1-3 in Section 3.2) to identify 4,528 core SAOs, including 35 core actions (i.e., verbs), 4,308 subjects (i.e., phrases and terms), and 4,409 objects (i.e., phrases and terms). Then, based on the 145,265 word-vectors trained by the Word2Vec model, these core SAOs were represented by SAO vectors (Steps 4-5 in Section 3.2).

The tri-layer network was constructed from the 4,528 SAOs. 1) The subject and object layers consisted of 4,308 nodes and 4,409 nodes, respectively. An edge between two nodes in the same layer was only created if the cosine similarity between the two corresponding SAO vectors was above average and the similarity was then set as the weight of the edge. 2) The action layer with the 35 core verbs was treated as a virtual layer with no edges. 3) The natural connections within each SAO structure were used as non-weighted edges among the three layers. The descriptive statistics of the tri-layer SAO network are given in Table 4.

Table 4. Descriptive statistics of the tri-layer SAO network

	No. of nodes	*No. of edges*	*Distribution of the weights of edges*			
			Max.	*Min.*	*Mean*	*Std. Dev.*
Subject layer	4,308	4,273,606	0.999	9.2e-9	0.293	0.186
Object layer	4,409	4,707,372	0.999	3.0e-7	0.348	0.185

Subsequent to establishing the current network, we conducted community detection to identify the key problems and solutions in China's research systems, followed by the link prediction for predicting the potential problem-solving patterns that Chinese researchers could be contributing to in the near future.

4.2.1. Community detection for identifying key problems and solutions

Given the relatively large number of subjects and objects, cluster analysis provided a way to explore representative subjects and objects and then identify key problems and solutions. Thus, we applied a fluid community detection approach to the subject and object layers of the network. To identify the optimal number *k* of communities over an interval of [2, 150], we plotted the resulting modularity values in a series of experiments and then selected the optimal number of communities based on the two criteria: 1) since the subjects and objects were retrieved from more than 200 WoS categories, a relatively large number of communities may better reflect reality; and 2) a higher modularity value may indicate a better result. Hence, we chose 80 and 90 as the numbers of communities for the subject and object layers, respectively.

The weighted centrality of each subject/object was calculated, and the subject/object with the highest value of the centrality was selected as the label for its community. It may be criticized, but, we employed these labels to represent the entire community and linked those communities with actions of those labels - i.e., when community C^A was labeled with a subject S, the actions connected with subject S were considered to be actions associated with the community C^A. 22 clumps were collected, in which one action acts as the core and is connected to either a set of subjects or a set of objects. Considering that clumps with missing subjects or objects do not adequately reflect a complete problem-solving pattern, 7 clumps were discarded, leaving 15 complete clumps with 68 subjects and 78 objects[6]. These might represent the key problems and solutions achieved by China's research systems over the past decade. It is intriguing that among those identified subjects and objects, the following seven WoS categories cover 71.19% of the combined set. They are: "chemistry, multidisciplinary", "chemistry, physical", "materials science", "multidisciplinary", "physics, applied", "nanoscience & nanotechnology", and "engineering, environmental".

Figure 3 shows a visualization of the tri-layer (using a procedure developed in one of our pilot studies on knowledge discovery in biomedical research – see Hu et al., (2018)), which helped us to analyze the situation more deeply. It is clear that "affect", "provide", "present", "promote", and "construct" are the top 5 clumps. From a semantic perspective and together with concepts in technology management (Li et al., 2018), we classified these 15 clumps into the following three aspects:

- Breakthrough technologies: the clumps of "provide", "inhibit", and "produces" may contain ideas for inventions, such as proposing new manufacturing processes, identifying new materials, and creating new products. These clumps contain 15 subjects and 16 objects, with a coverage of 21.23%.
- Technological refinements: the clumps of "affect", "combine", "estimate", "induce", and "promote" may indicate improvements to existing solutions. These clumps involve 18 subjects and 31 objects, with a coverage of 33.56%.
- Potential innovative solutions: the remaining 7 clumps, covering 45.21% of the network, include "associate with", "construct", "involve", "observe", "process", "present", and "use". Further investigation is needed to understand their contents in detail, but it is reasonable to consider that these solutions may be ancillary to potential innovation, e.g., general observations or the use of existing approaches, as well as possibly containing some significant findings -- e.g., an impactful recombination.

[6] Available at https://github.com/IntelligentBibliometrics/QSS

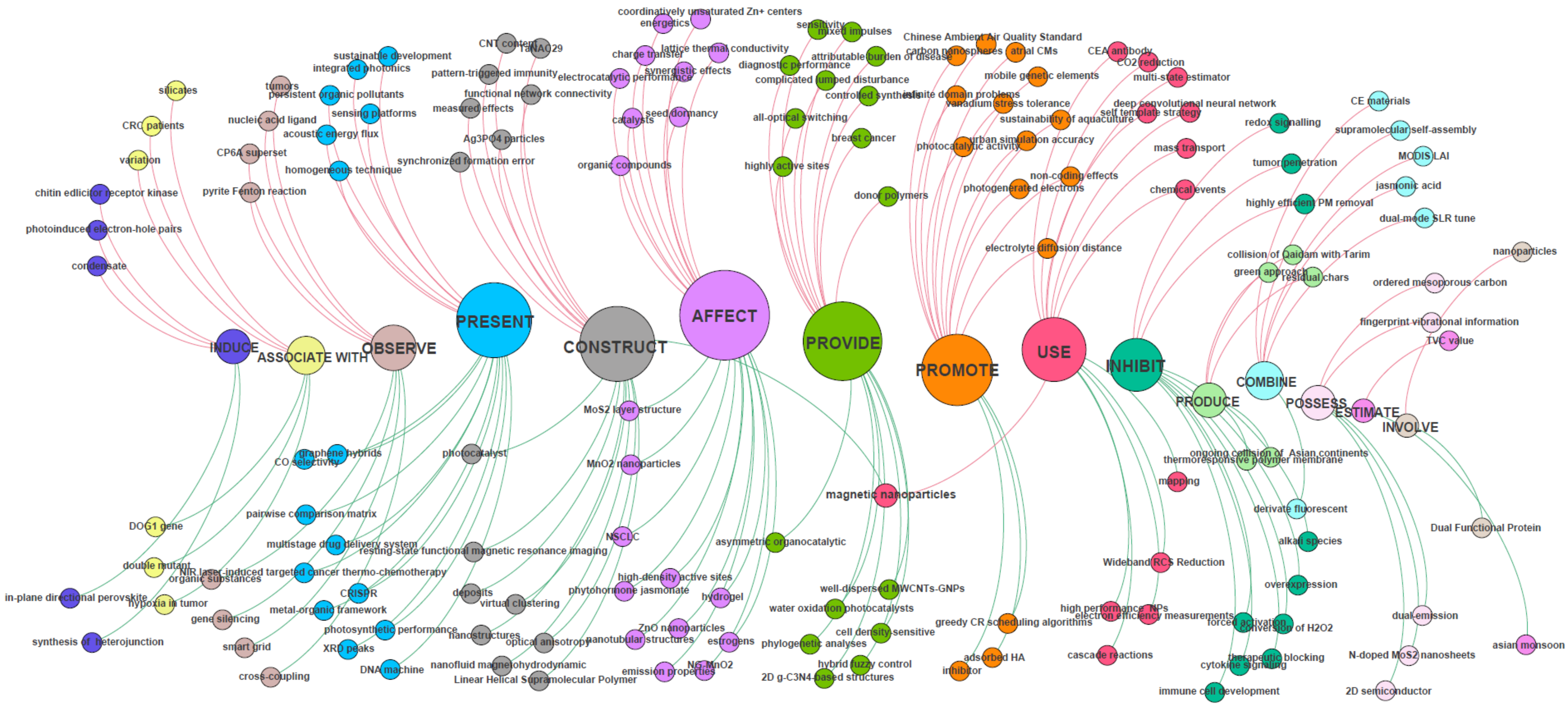


Figure 3. The tri-layer SAO network of key research problems (bottom layer) and solutions (top layer) contributed by Chinese researchers

Note: For readability, we have only included the key terms of subjects and objects. The size of actions (middle layer) indicates the number of connected subjects and objects. The color of the nodes indicates the clumps. A high-definition version could be found: https://github.com/IntelligentBibliometrics/QSS.

4.2.2. Link prediction for predicting potential problem-solving patterns

We applied the refined link prediction approach to predict the links in the tri-layer SAO network. Based on the predicted weight of the connection between a subject and an object, 10,000 problem-solving patterns (i.e., subject-object pairs) were identified[7] that could be considered as potential solutions (i.e., subjects) for certain problems (i.e., objects). Two sets of efforts were conducted to briefly demonstrate these 10,000 patterns: 1) a science map for visualizing the WoS-category co-occurrence among predicted solutions to review potential disciplinary interactions in China's research systems; and 2) 19 highlighted problem-solving patterns to provide examples on how potential solutions are recommended to a specific problem.

Given that we linked the subjects and objects with the WoS categories via their related journals and, hence, one solution (subject) may belong to multiple WoS categories. The breakdown of WoS categories in the predicted network closely follows the proportions given in Table 3. We also generated a co-occurrence matrix between categories, and then a science overlay map, as shown in Figure 4, to reveal the interactions between predicted solutions at a discipline level. Despite an outcome of the micro-level investigation based on network analytics on an SAO network, insights derived from this map could provide clues from a macro-level landscape as to how multiple disciplines might fuse based on China's current research strengths. Such insights might be interesting in terms of science policy and strategic management.

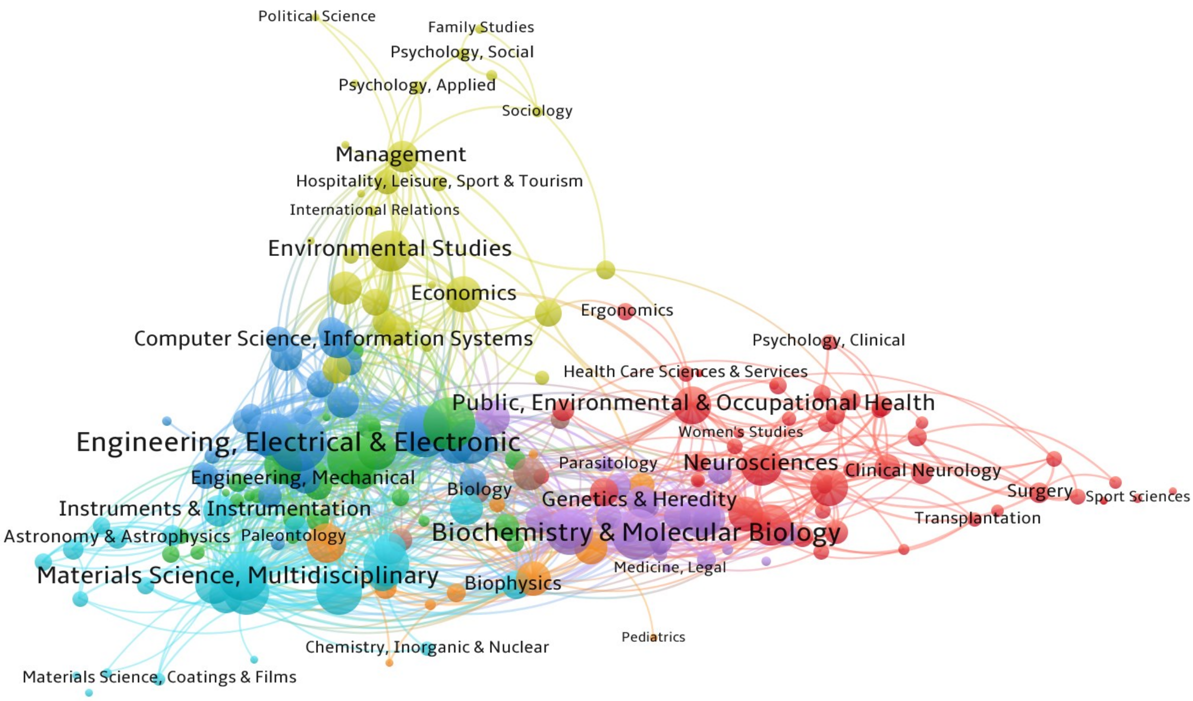


Figure 4. WoS-category co-occurrence map of predicted solutions

Unlike our observations from the science overlay maps above (c.f., Figure 2 and Appendix A), Figure 4 emphasizes the predicted disciplinary interactions that may occur in China's research systems in the near future, and thus, from this analysis, we glean the following insights.

[7] Also available at https://github.com/IntelligentBibliometrics/QSS

- Computer science is bridging engineering disciplines (e.g., engineering, electrical & electronic) and business disciplines (e.g., management and economics). Such a combination is to be expected as a frontier that provides solutions for social sciences, with support of China's research advantages in computer science disciplines.
- Extensive fusions may occur with a broad range of engineering disciplines. Also, the strengths of the connections among disciplines, such as materials science, chemistry, biology, and neurosciences, might be further enhanced.
- Genetics and heredity seem to be a key that could launch a cutting-edge direction in the medical and healthcare sciences. Similarly, the bridging role of public, environmental & occupational health in connecting disciplines, such as biology, neuroscience, clinical neurology and psychology, and economics is observed.

As representative cases, we selectively highlighted 19 problem-solving patterns in Table 5, based on the following steps and criteria:

1) According to the predicted value calculated by the link prediction method, which represents the potential strength of the connection between a subject and an object, each pattern could be ranked.
2) The WoS category of the subject was set as a prior indicator, and for each category, we selected the top 3 high-ranked patterns.
3) Duplicate subjects and objects were removed -- i.e., for each subject/object, only the pattern with the highest rank was retained.

Following a traditional approach in literature-based discovery for seeking supportive evidence from the literature, we randomly picked a few patterns in Table 5:

- Current research on "Plant sciences" (#4) is working toward proof that melatonin treatments protect antioxidant enzyme activities, which could regulate oxidative stress (Emamgholipour et al., 2016).
- In #9, while t-distributed stochastic neighbor embedding is a machine learning technique for reducing dimensions and then visualization (Van der Maaten & Hinton, 2012) and measuring the uptake of carbon dioxide by leaves is an approach of representing gross primary production, the bridge of the two "apparently" irrelevant items is earth system modelling and visualization, a frontier area in plant sciences (Wang et al., 2017).
- In Dermatology, #17, an acne-like skin rash is one of the most common side-effects of treating cancer with a combination of cetuximab and oxaliplatin, which has been reported in diseases such as cholangiocarcinoma (Gruenberger et al., 2008).

Table 5. Selected potential problem-solving patterns that may be achieved by Chinese researchers.

	WoS category of subjects	*Subject (potential solutions)*	*Object (possible problems)*
1	Engineering, environmental	excellent microwave absorption (MA) performance	a nonvolatile rewritable memory effect with the function of flash
2	Thermodynamics	The regional differences in impact factors on CO2 emissions	CRISPR-Cas9
3*	Microbiology	respiratory syncytial virus	as a target of miR-448
4	Plant sciences	Melatonin	significantly Increased activities of catalase
5	Gastroenterology & hepatology	Pin1 inhibition by API-1	for the MnO2/Mn/MnO2 sandwich-like nanotube arrays in solution of 1.0 M Na2SO4
6	Medicine, research & experimental	StarBase V2.0	to detect Pb2+ in practical samples
7	Geochemistry & geophysics	a deep convolutional neural network	high-quality images
8*	Virology	non-Latinized binomial species names	species richness
9	Computer science, artificial intelligence	t-distributed stochastic neighbor	accurate representation of gross primary production
10	Biochemical research methods	machine learning methods	to identify feature-related wavebands for developing models for monitoring the oxidative damage of pork myofibrils during frozen storage
11	Marine & freshwater biology	Tris (1,3-dichloroisopropyl) phosphate exposure	reactive oxygen species
12*	Agriculture, dairy & animal science	Gene expressions of antioxidant enzymes	by real-time polymerase chain reaction
13	Peripheral vascular disease	Human Induced pluripotent stem cells	better durability under harsh hydrogen evolution reaction cycling conditions than commercial Ir/C
14*	Behavioral sciences	enzyme-linked immunosorbent assay	an improved photocatalytic performance under visible light irradiation
15	Nuclear science & technology	The analysis of Fourier transforms infrared spectroscopy	in a thermogravimetry/differential scanning calorimetry coupled with Fourier transform infrared spectroscopy

16*	Genetics & heredity	hepatocellular carcinoma up-regulated long non-coding RNA	hepatocellular carcinoma progression
17	Dermatology	acnes species	the optical properties of the anticancer function of oxaliplatin
18	Materials science, characterization & testing	The results from the 3-D FE model	statistical data concerning changes in the microenvironment of amide moieties in response to different doses of multi-walled carbon nanotubes
19	Orthopedics	Osteoblasts	tyrosine kinase inhibitors resistance

Note that 1) since one subject may be connected with multiple WoS categories, the WoS category of subjects in this table only lists the category with the largest number of SAOs in the predicted 10,000 subject-object pairs; and 2) we do not provide the WoS category of objects since one solution can be easily assigned to a given research area but one problem may be a combination of multiple disciplines.

4.2.3. Empirical validation for the prediction of problem-solving patterns

Aiming to evaluate the performance of the proposed method, an empirical validation was conducted to examine the predicted potential problem-solving patterns through link prediction. This design follows two reasons. 1) The retrieval of SAO structures was conducted by the software OpenIE, which is a popular tool for SAO analysis and has already been examined in the NLP area (Angeli et al., 2015). Thus, we assume the collected SAOs in our study (i.e., core SAOs retrieved in the data pre-processing) are acceptable. 2) Since the empirical dataset is unlabeled and previous SAO approaches are mostly case-driven and semi-automatic, expert knowledge-based empirical validation is the best option for us under these circumstances. Despite the other task of recognizing problem-solving patterns, the prediction of those patterns' recombination is the final outcome of the micro-level investigation. Therefore, targeting to these predicted problem-solving patterns, empirical evaluation was conducted from two aspects: 1) the validation of selected problem-solving patterns in Table 5, and 2) the validation of problem-solving patterns predicted for a specific problem.

(1) Validation of selected predicted problem-solving patterns

These 19 problem-solving patterns cover a broad range of research disciplines, challenging the organization of a relevant expert panel for conducting the empirical evaluation. Thus, we specifically picked up 5 problem-solving patterns (marked with a "*" in Table 5) aligning with biology and life sciences. We formed an expert panel, including five early career researchers (e.g., Research Follows and PhD candidates) in related areas from two CAS's institutes: the Institute of Zoology and the Guangzhou Institutes of Biomedicine and Health. We interviewed these five experts and requested them to mark the relevance of the solutions to problems against five levels, where A means 'exactly relevant' (equal to '1'), and 'E' means 'totally irrelevant' (equal to '0'). The scores for the five problem-solving patterns are given in Appendix B (see Supplementary Table 4).

In general, an average score for the five patterns is 0.58, which could be acceptable, considering the two patterns (i.e., #3 and #14) with the lowest scores received one B score at least. Then, we arranged an online workshop gathering the five experts to delve into the three patterns (i.e., #8, #12, and #16) and empirically discussed their potential. We conclude as follows:

- Pattern #8 refers to the naming issue in virology, raised by the irregularity of naming viruses in early days and the species richness, and then the International Committee on Taxonomy of Viruses (ICTV) introduced the rule of using non-Latinized binomial to name virus species in 2011 (Van Regenmortel et al., 2010). Thus, the connection between the subject and object of pattern #8 is promising, but considering this is not a potential problem-solving pattern, rather than an existing one, it is reasonable to mark it with a score of 0.65.
- Pattern #12 in fact exposes a shortage of SAO analysis, which could not effectively distinguish subjects and objects in a passive tense. In this case, detecting the gene expression of antioxidant enzymes is a problem in animal science, and the approach of real-time polymerase chain reaction (RT-PCR) could be an effective and approvable solution (Yin et al., 2014). That is to say, the SAO analysis failed to clearly identify the roles of the two items, but, intriguingly, the preposition 'by' before RT-PCR could be such an excuse.

However, considering this is an evaluation for link prediction – seeking the relationships between subjects and objects, we made a good hit.

- Pattern #16 uncovers the correlation between long noncoding RNAs (IncRNA) and the hepatocellular carcinoma, and evidence for this potential pattern could be traced in some most recent papers published in related top-level journals (Xiong et al., 2017), considering the dataset only covers publications before January 1, 2019. Thus, we consider this pattern demonstrates good agreement between our prediction and expert knowledge.

According to this workshop, the experts agreed that the proposed method could gain advantages in connecting problems and solutions, and such problem-solving patterns are the recombination of existing knowledge, which might be innovative for related research communities. However, the experts also pointed out that since these patterns were identified from scientific articles, if those predicted patterns were based on relatively old articles, such a recombination might be realized already (e.g., #8). We agree and anticipate that focusing on the most recent publications could increase the practical significance of the proposed method.

(2) Validation of the problem-solving patterns predicted for a specific problem

In this section, we targeted a specific problem and validated whether the set of potential solutions for this problem could be empirically feasible. Considering our own background, we noticed that "Computer science, artificial intelligence (AI)" contains 461 problem-solving patterns. Of these, the problem "to identify feature-related wavebands for developing models for monitoring the oxidative damage of pork myofibrils during frozen storage" in the "food science & technology" category had 15 potential solutions, as listed in Table 6. Evaluating each of these solutions may prove to be an interesting future empirical study we could undertake to demonstrate the feasibility of the link prediction approach for predicting problem-solving patterns.

Table 6. 15 potential AI solutions for one specific problem in food science and technology

	Subject (Potential Solution)	*Level*	*Note*
1	nonlocal hierarchical dictionary learning	A	Methods for feature selection (Zhu et al., 2016a)
2	a supervised inductive manifold hashing framework	C	Manifold hashing could be feasible for feature representation (Song et al., 2017) but may not be suitable for feature selection
3	a manifold embedding algorithm	A	Methods for feature selection (Yao et al., 2017)
4	a reinforcement learning algorithm	E	Inapplicable for this case
5	hyperspectral images	D	Related to the problem but not a solution
6	nonparametric manifold learning	A	Methods for feature selection (Cai et al., 2010)
7	A deterministic learning technique	C	Methods in quantum computing, which may be theoretically applicable
8	a multi-objective discrete particle swarm optimization algorithm	A	Widely applied methods for feature selection (a large number of articles have been published in journals such as *Expert Systems with Applications*)

9	a multimodal deep support vector classification (MDSVC) approach	B	Support vector machine is one classical approach for feature selection, and the use of deep learning for extracting explicit features may be a challenge.
10	a multi-kernel learning strategy	A	Classical methods for feature selection (Zeng & Cheung, 2010)
11	a projection-based TODIM method with MVNSs for personnel selection	A	Classical optimization approaches, similar to #8
12	new hashing techniques	B	Hashing techniques are well known for data storage and transmission, but some work could be traced in the literature (Zhang et al., 2015).
13	existing sparse coding algorithms	A	Methods for image/graph feature selection (Zhu et al., 2016b)
14	spectral embedding	B	Spectral analysis for feature selection could be traced in the literature (Li et al., 2012)
15	a novel low-rank multi-view	E	Not a solution

Note that the 15 solutions were ranked based on the predicted value calculated by the proposed approach of link prediction.

The description of the problem suggests that the solution may be a task of feature selection for pattern recognition (i.e., waveband patterns). We followed the same approach and discussed these 15 solutions with researchers in the Australian Artificial Intelligence Institute at the University of Technology Sydney with particular expertise in machine learning and computer vision. Based on these discussions and scores marked by the experts, the average score over the entire set of 15 solutions was 0.7, which could be considered an “acceptable” result. Additionally, we added references to support the experts’ judgments, noting that all cited references are sourced from top-level journals and conferences in the area of artificial intelligence.

As a conclusion, the empirical validation gained a score of 0.58 for the predicted high-ranked patterns in the area of biology and life science, and a score of 0.7 for the predicted AI solutions for a specific problem in food science and technology. Despite certain limitations raised by the experts, we all agreed that the performance of this prediction is acceptable, and the proposed method could have practical significance for actual uses.

5. Discussion and conclusions

In this paper, we present a methodology based on intelligent bibliometrics to investigate the problem-solving patterns in China’s research systems. The methodology leverages science overlay maps to profile the interactions among research disciplines, plus subject-action-object (SAO) analysis, with network analytics, to identify key problems and solutions, as well as to predict the potential solutions that Chinese researchers might achieve in future. We derive insights from an empirical study focusing on top papers published by Chinese researchers between 2009 and 2018, from which we derive evidence of China’s research strengths and conjectures about the science policies that drive these strengths, multidisciplinary interactions, key research problems and solutions, and potential solutions to existing problems.

5.1. *Key findings*

Research strengths and China's science policies: The proportion of top papers in the WoS categories, as well as the distribution of identified problems and solutions, indicates that China's science policies emphasize research in the natural more than the social sciences (including arts and humanities). However, it is also plausible the imbalance may be due to the difficulties Chinese researchers have with publishing in social science fields as a result of differences in knowledge bases, cultural backgrounds, and, of course, language barriers, which are much higher in social science journals. In general, China's research strengths concentrate in disciplines like chemistry, materials science, applied physics, engineering, and computer science. In particular, "nanoscience and nanotechnology" stands forth as a multidisciplinary strength in China's research systems.

Multidisciplinary interactions: Following the trends in multidisciplinary interactions, China's efforts are competitive internationally. China is spearheading two main cross-disciplinary directions: 1) Computer Science (highlighted by artificial intelligence techniques) and its applications in engineering areas; and 2) nanotechnology and its relevance to chemical and biological engineering. Solid evidence, in the form of affiliation-based science overlay maps, identify key research problems and solutions. Predicted patterns of problems and solutions build from the research outcomes to date.

Problem-solving patterns: Arguably, the actions associated with the key problems and solutions indicate that around 20% of the research relates to breakthrough technologies, 30% to technological refinements, and the remaining 50% to extraneous research activity. While the disciplines of identified problems and solutions coincide with that of China's research strengths, the predicted solutions that might be achievable for Chinese researchers are based on their current accomplishments, which demonstrate China's extensive capabilities in spearheading cross-disciplinary research.

5.2. *Methodological implications and potential applications*

Intelligent bibliometrics could be an effective toolkit for a broad range of empirical studies, both in practical sectors and for specific research questions. The proposed methodology of intelligent bibliometrics emphasizes the use of certain advanced technologies in information retrieval -- i.e., word embedding, subject-action-object (SAO) analysis, and network analytics -- and the empirical results soundly demonstrate its feasibility and reliability. The proposed framework also has the potential for a high level of flexibility since it could easily be applied to many academic databases, such as PubMed and Scopus, or to patents, without major modifications. The main technical implications of the proposed methods are highlighted here.

- Incorporating network analytics (e.g., community detection and link prediction) with bibliometric approaches (e.g., science overlay maps) offers benefits for further knowledge discovery. For example, this combination can help predict future interactions between disciplines and help identify potential solutions for existing problems. However, our experiences with this case study indicate that the complexity of the network analytics algorithms might be sensitive to network structures. Ways to maintain the balance between large networks and efficient analytics should be considered.

- SAO analysis creates additional dimensions for understanding semantics and discovering latent relationships, compared to individual word- and term-based analysis. Involvement of word embedding techniques can further enhance the capability of measuring similarities among SAOs. However, we also acknowledge that the modularity of retrieved SAOs is much more scattered than that of terms. This substantially increases the complexity of further network analytics and may lead to reduced performance when directly applying some traditional approaches, e.g., clustering.

5.3. Limitations and future directions

We note limitations and potential refinements in four directions. 1) The SAO structures were represented as the means of their constituent word vectors built by Word2Vec. However, applying a machine learning technique to train a weighting strategy may better reflect actual situations. For instance, the term "data mining" may give a higher weight to "data" than to "mining". 2) The validation of SAO network analytics, including identifying key problems and solutions based on SAO network analytics, and predicting potential problem-solving patterns might require quantitative approaches to complement expert knowledge to reduce potential biases. 3) Deciding on the optimal number of communities in the community detection step might be better handled with an optimization technique or techniques. 4) When investigating the problem-solving patterns of China's research systems, the use of WoS categories may raise a concern that non-existent interactions may simply mean no journal covers those two disciplines, rather than no research articles covering topics in those disciplines. Thus, engaging multiple data sources and conducting further text analytics may provide a more comprehensive perspective on the landscape under study.

Acknowledgments

We acknowledge support from the Australian Research Council under Discovery Early Career Researcher Award DE190100994 for Yi Zhang and Mengjia Wu, the Library and Information Service Promotion Special Project of Chinese Academy of Sciences (Grant No: Y9290002) for Zhengyin Hu and Xue Zhang, and the US National Science Foundation (Award # 1759960) to Search Technology, Inc., and Georgia Tech for Alan Porter and Robert Ward.

We acknowledge Dr Rongqian Zeng from Southwest Jiaotong University, China for his assistance in visualizing SAO networks. We also appreciate expert knowledge provided by our colleagues at the Australian Artificial Intelligence Institute, University of Technology Sydney, Australia, and by the five researchers from the Institute of Zoology and the Guangzhou Institutes of Biomedicine Health, Chinese Academy of Sciences, China.

Author contributions

Conceptualization (YZ, ZH, and AP); Data curation (ZH and XZ); Formal Analysis and Methodology (YZ, MW, ZH, RW, and XZ); Validation (YZ, MW, ZH, and XZ); Visualization (ZH, RW, and XZ); Writing (YZ, MW, and AP).

Competing interests

Author Alan Porter was employed by the company, Search Technology, Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Allan, J. (2002). *Topic detection and tracking: Event-based information organization*. US: Springer.

Angeli, G., Premkumar, M. J. J., & Manning, C. D. (2015). *Leveraging linguistic structure for open domain information extraction.* Paper presented at the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., . . . Boyack, K. W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS One, 7*(7), e39464.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., . . . Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS One, 6*(3), e18029.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*(1), 59-73.

Cai, D., Zhang, C., & He, X. (2010). *Unsupervised feature selection for multi-cluster data.* Paper presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.

Cao, C., & Suttmeier, R. P. (2017). Challenges of S&T system reform in China. *Science, 355*(6329), 1019-1021.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Chen, C., & Song, M. (2019). Visualizing a field of research: A methodology of systematic scientometric reviews. *PLoS One, 14*(10).

Csajbók, E., Berhidi, A., Vasas, L., & Schubert, A. J. S. (2007). Hirsch-index for countries based on Essential Science Indicators data. *73*(1), 91-117.

Emamgholipour, S., Hossein-nezhad, A., Sahraian, M. A., Askarisadr, F., & Ansari, M. (2016). Evidence for possible role of melatonin in reducing oxidative stress in multiple sclerosis through its effect on SIRT1 and antioxidant enzymes. *Life sciences, 145*, 34-41.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks, 1*(3), 215-239.

Fu, H.-Z., Chuang, K.-Y., Wang, M.-H., & Ho, Y.-S. J. S. (2011). Characteristics of research in China assessed with Essential Science Indicators. *88*(3), 841-862.

Glenisson, P., Glänzel, W., Janssens, F., De Moor, B. J. I. p., & management. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *41*(6), 1548-1572.

Gruenberger, B., Schueller, J., Kaczirek, K., Bergmann, M., Klose, W., Bischof, M., . . . Gruenberger, T. (2008). Efficacy results of cetuximab plus gemcitabine-oxaliplatin (GEMOX) in patients with advanced or metastatic cholangiocarcinoma: A single centre phase II study. *Journal of Clinical Oncology, 26*(15_suppl), 4586-4586.

Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics, 116*(2), 1367-1382.

Huang, L., Jia, X., Zhang, Y., Zhou, X., & Zhu, Y. (2018). *Detecting Hotspots in Interdisciplinary Research Based on Overlapping Community Detection.* Paper presented at the 2018 Portland International Conference on Management of Engineering and Technology (PICMET).

Huang, L., Zhang, Y., Guo, Y., Zhu, D., & Porter, A. L. (2014). Four dimensional science and technology planning: A new approach based on bibliometrics and technology roadmapping. *Technological Forecasting and Social Change, 81*, 39-48.

Huang, Y., Zhang, Y., Youtie, J., Porter, A. L., & Wang, X. (2016). How does national scientific funding support emerging interdisciplinary research: A comparison study of big data research in the US and China. *PLoS One, 11*(5), e0154509.

Kilicoglu, H., Rosemblat, G., Fiszman, M., & Rindflesch, T. C. (2011). Constructing a semantic predication gold standard from the biomedical literature. *BMC bioinformatics, 12*(1), 486.

Leydesdorff, L., & Zhou, P. (2014). Measuring the knowledge-based economy of China in terms of synergy among technological, organizational, and geographic attributes of firms. *Scientometrics, 98*(3), 1703-1719.

Li, M., Porter, A. L., & Suominen, A. (2018). Insights into relationships between disruptive technology/innovation and emerging technology: A bibliometric perspective. *Technological Forecasting and Social Change, 129*, 285-296.

Li, X., Zhou, Y., Xue, L., & Huang, L. (2015). Integrating bibliometrics and roadmapping methods: A case of dye-sensitized solar cell technology-based industry in China. *Technological Forecasting and Social Change, 97*, 205-222.

Li, Z., Yang, Y., Liu, J., Zhou, X., & Lu, H. (2012). *Unsupervised feature selection using nonnegative spectral analysis.* Paper presented at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

Liao, H., Tang, M., Li, Z., & Lev, B. (2019). Bibliometric analysis for highly cited papers in operations research and management science from 2008 to 2017 based on essential science indicators. *Omega, 88*, 223-236.

Ma, T., Porter, A. L., Guo, Y., Ready, J., Xu, C., & Gao, L. (2014). A technology opportunities analysis model: Applied to dye-sensitised solar cells for China. *Technology Analysis & Strategic Management, 26*(1), 87-104.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Mu, R., & Qu, W. (2008). The development of science and technology in China: A comparison with India and the United States. *Technology in Society, 30*(3-4), 319-329.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577-8582.

Parés, F., Gasulla, D. G., Vilalta, A., Moreno, J., Ayguadé, E., Labarta, J., . . . Suzumura, T. (2017). *Fluid communities: a competitive, scalable and diverse community detection algorithm.* Paper presented at the International Conference on Complex Networks and their Applications.

Ping, Q., & Chen, C. (2018). LitStoryTeller+: an interactive system for multi-level scientific paper visual storytelling with a supportive text mining toolbox. *Scientometrics, 116*(3), 1887-1944.
Porter, A. L., & Detampel, M. J. (1995). Technology opportunities analysis. *Technological Forecasting and Social Change, 49*(3), 237-255.
Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871-1887.
Rost, K., Teichert, T., & Pilkington, A. (2017). Social network analytics for advanced bibliometrics: referring to actor roles of management journals instead of journal rankings. *Scientometrics, 112*(3), 1631-1657.
Rotolo, D., Rafols, I., Hopkins, M. M., Leydesdorff, L. J. J. o. t. A. f. I. S., & Technology. (2017). Strategic intelligence on emerging technologies: Scientometric overlay mapping. *68*(1), 214-233.
Song, T., Cai, J., Zhang, T., Gao, C., Meng, F., & Wu, Q. (2017). Semi-supervised manifold-embedded hashing with joint feature representation and classifier learning. *Pattern Recognition, 68*, 99-110.
Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human‐assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(19), 2464‐2476.
Tang, L. (2019). Five ways China must cultivate research integrity. In: Nature Publishing Group.
Van der Maaten, L., & Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. *Machine learning, 87*(1), 33-55.
van Raan, A. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397-420.
Van Regenmortel, M. H., Burke, D. S., Calisher, C. H., Dietzgen, R. G., Fauquet, C. M., Ghabrial, S. A., . . . Horzinek, M. C. J. A. o. v. (2010). A proposal to change existing virus species names to non-Latinized binomials. *155*(11), 1909-1919.
Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B, 86*(11), 471.
Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629-635.
Wang, H., Prentice, I. C., Keenan, T. F., Davis, T. W., Wright, I. J., Cornwell, W. K., . . . Peng, C. (2017). Towards a universal model for carbon dioxide uptake by plants. *Nature Plants, 3*(9), 734-741.
Xiong, H., Ni, Z., He, J., Jiang, S., Li, X., Gong, W., . . . Zhang, N. J. O. (2017). LncRNA HULC triggers autophagy via stabilizing Sirt1 and attenuates the chemosensitivity of HCC cells. *36*(25), 3528-3540.
Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics, 8*(2), 295-309.
Yang, C., Huang, C., & Su, J. (2018). An improved SAO network-based method for technology trend analysis: A case study of graphene. *Journal of Informetrics, 12*(1), 271-286.
Yao, C., Liu, Y.-F., Jiang, B., Han, J., & Han, J. (2017). LLE score: A new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition. *IEEE Transactions on Image Processing, 26*(11), 5257-5269.

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics, 100*(3), 767-786.

Yin, J., Wu, M., Xiao, H., Ren, W., Duan, J., Yang, G., . . . Yin, Y. J. J. o. a. s. (2014). Development of an antioxidant system after early weaning in piglets. *92*(2), 612-619.

Zeng, H., & Cheung, Y.-m. (2010). Feature selection and kernel learning for local learning-based clustering. *IEEE transactions on pattern analysis and machine intelligence, 33*(8), 1532-1547.

Zhang, L., Lu, H., Du, D., & Liu, L. (2015). Sparse hashing tracking. *IEEE Transactions on Image Processing, 25*(2), 840-849.

Zhang, N., Wan, S., Wang, P., Zhang, P., & Wu, Q. J. S. (2018a). A bibliometric analysis of highly cited papers in the field of Economics and Business based on the Essential Science Indicators database. *116*(2), 1039-1053.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018b). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*(4), 1099-1117.

Zhang, Y., Porter, A. L., Cunningham, S. W., Chiavetta, D., & Newman, N. (2020a). Parallel or intersecting lines? Intelligent bibliometrics for investigating the involvement of data science in policy analysis. *IEEE Transactions on Engineering Management, to appear*.

Zhang, Y., Wang, X., Huang, L., Zhang, G., & Lu, J. (2018c). *Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology.* Paper presented at the 2018 Annual Meeting of the Association for Information Science and Technology, Vancouver, Canada.

Zhang, Y., Wu, M., Zhu, Y., Huang, L., & Lu, J. (2020b). Characterizing the potential of being emerging generic technologies: A prediction method incorporating with bi-layer network analytics. *Scientometrics*, to appear.

Zhang, Y., Zhang, C., & Li, J. (2019a). Joint Modeling of Characters, Words, and Conversation Contexts for Microblog Keyphrase Extraction. *Journal of the Association for Information Science and Technology*.

Zhang, Y., Zhou, X., Porter, A. L., & Gomila, J. M. V. (2014a). How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: "Problem & Solution" pattern based semantic TRIZ tool and case study. *Scientometrics, 101*(2), 1375-1389.

Zhang, Y., Zhou, X., Porter, A. L., Gomila, J. M. V., & Yan, A. (2014b). Triple Helix innovation in China's dye-sensitized solar cell industry: Hybrid methods with semantic TRIZ and technology roadmapping. *Scientometrics, 99*(1), 55-75.

Zhang, Y., Zhu, Y., Huang, L., Zhang, G., & Lu, J. (2019b). *Characterizing the potential of being emerging generic technologies: A methodology of bi-layer network analytics.* Paper presented at the International Conference of the International Society for Scientometrics and Informetrics, Rome, Italy.

Zhou, P., & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research Policy, 35*(1), 83-104.

Zhu, P., Hu, Q., Zhang, C., & Zuo, W. (2016a). *Coupled dictionary learning for unsupervised feature selection.* Paper presented at the Thirtieth AAAI Conference on Artificial Intelligence.

Zhu, X., Li, X., Zhang, S., Ju, C., & Wu, X. (2016b). Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE transactions on neural networks and learning systems, 28*(6), 1263-1275.

Appendix A Partial Results of The Macro-Level Analysis

This appendix contains two sets of science overlay maps to profile the interactions between disciplines and further understand the reasons behind them:

- Affiliation-based maps: These emphasize the differences between universities and research institutions, revealing the diversity of science policies in China's research systems.
- Time-based maps: These provide a way to trace the evolution of disciplinary interactions over the past decade.

(1) Affiliation-based comparison

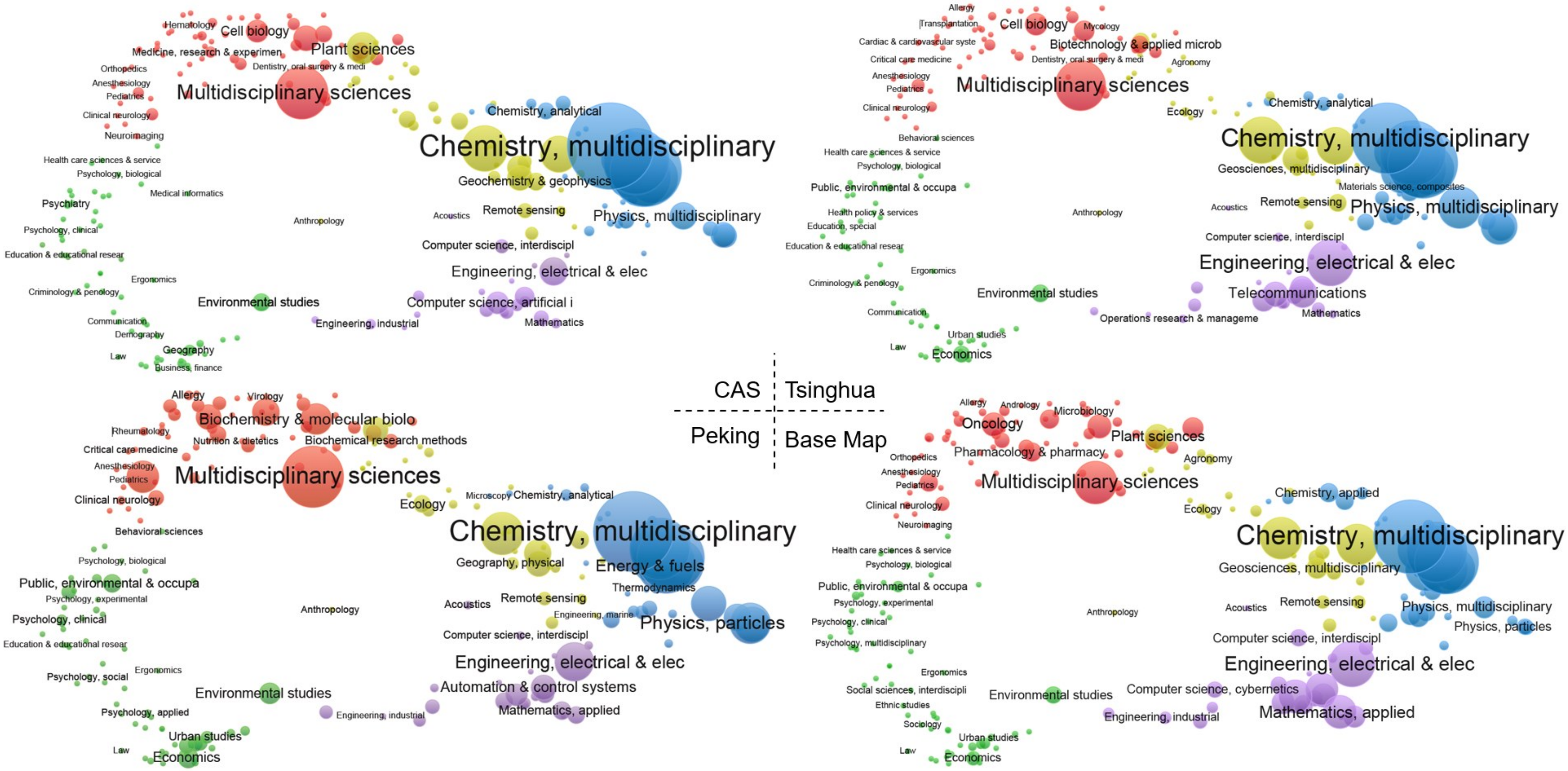
In terms of basic record counts, the Chinese Academy of Sciences (CAS) published the most papers in the dataset with 5,449 articles. Tsinghua University and Peking University follow with 1,631 and 1,309 articles, respectively. Interestingly, these three entities are quite representative of two types of affiliations in China's research systems - i.e., research institutions mainly supported by the government and universities. In terms of the QS World University Rankings[8], as well as several Chinese ranking systems, Tsinghua University performs best in natural sciences while Peking University leads the social sciences in the country.

With the aid of VantagePoint[9], we generated a base science map of the entire dataset along with three further maps for each of these top performers, as illustrated in Supplementary Figure 1. Our observations follow.

- In general, these three affiliations have similar strengths in certain disciplines, such as "chemistry, multidisciplinary", "multidisciplinary sciences", and "engineering, electrical & electronic", and considering the balance between natural sciences (e.g., the red, yellow, blue, and purple nodes), plus the social sciences (green). But strength in the former areas overwhelmingly exceeds the latter. This trend accords with the base map.
- Despite having common interests, the three affiliations each have their own strengths in the natural sciences – i.e., CAS in "computer science, artificial intelligence" and "plant sciences", Tsinghua University in "telecommunications" and "biotechnology & applied microbiology", and Peking University in "automation & control systems", "mathematics, applied", "ecology" and "biochemistry & molecular biology."
- There is no solid evidence to judge the different preferences of Tsinghua University and Peking University in natural sciences and social sciences. However, compared to the other two, Peking University appears to have more strength in business disciplines, such as "economics" and "urban studies", as well as disciplines such as "public, environmental & occupational health" and "psychology, clinical".

[8] More information can be found on the website: https://www.topuniversities.com/university-rankings/world-university-rankings/2020

[9] More information can be found on the website: www.theVantagePoint.com



Supplementary Figure 1. Science overlay maps for the Chinese Academy of Sciences (top-left), Tsinghua University (top-right), and Peking University (bottom-left), and the base map for China's research systems (bottom-right).

As we discussed above, due to language barriers and differences in culture and values, as well as certain internal criteria for performance evaluation (e.g., China's central government starts to encourage researchers to publish papers in Chinese[10]), it is reasonable to track the outputs of Chinese researchers in social sciences from top Chinese journals and, specifically, take the Chinese Academy of Social Sciences (CASS), a government-funded research institution that specializes in the social science disciplines, into consideration. Given the circumstances, we conducted a search in the Chinese Social Sciences Citation Index (CSSCI) database[11], the most reputable and recognized database for Chinese social science studies, with the following search strings (but in Chinese), and the results are given in Supplementary Table 1.

Affiliation= ("Tsinghua University" OR "Peking University" OR "Chinese Academy of Sciences" OR "Chinese Academy of Social Sciences") AND Publication Year= (2009-2018) AND Article Type= (Research Articles)

Supplementary Table 1. Number of Chinese social science publications in the CSSCI

No.	*Affiliation*	*Num of Articles*
1	Tsinghua University	11,980
2	Peking University	19,301
3	Chinese Academy of Sciences	9,448
4	Chinese Academy of Social Sciences	20,616

Supplementary Table 1 reveals that CASS leads China's social sciences, and Peking University is a competitive counterpart, both considerably exceeding Tsinghua University and CAS. This coincides with our observation from QS and other ranking systems. Even though CSSCI and ESI are not necessarily at the same stage, this 'baby' search provides an external indicator to understand the situation of China's social science disciplines and could be complementary to the main body of this affiliation-based comparison. In general, at a macro level we observe:

- Over the past decade, China has pursued a balanced strategy of encouraging academic research in all scientific disciplines, but China's efforts in social science disciplines are not as advanced as that of natural sciences on the global stage.
- Interactions within natural sciences can be clearly traced for each of the three affiliations as well as the base map, but how Chinese researchers will conduct cross-disciplinary studies between the natural and social sciences, where gaps still exist, is elusive so far.

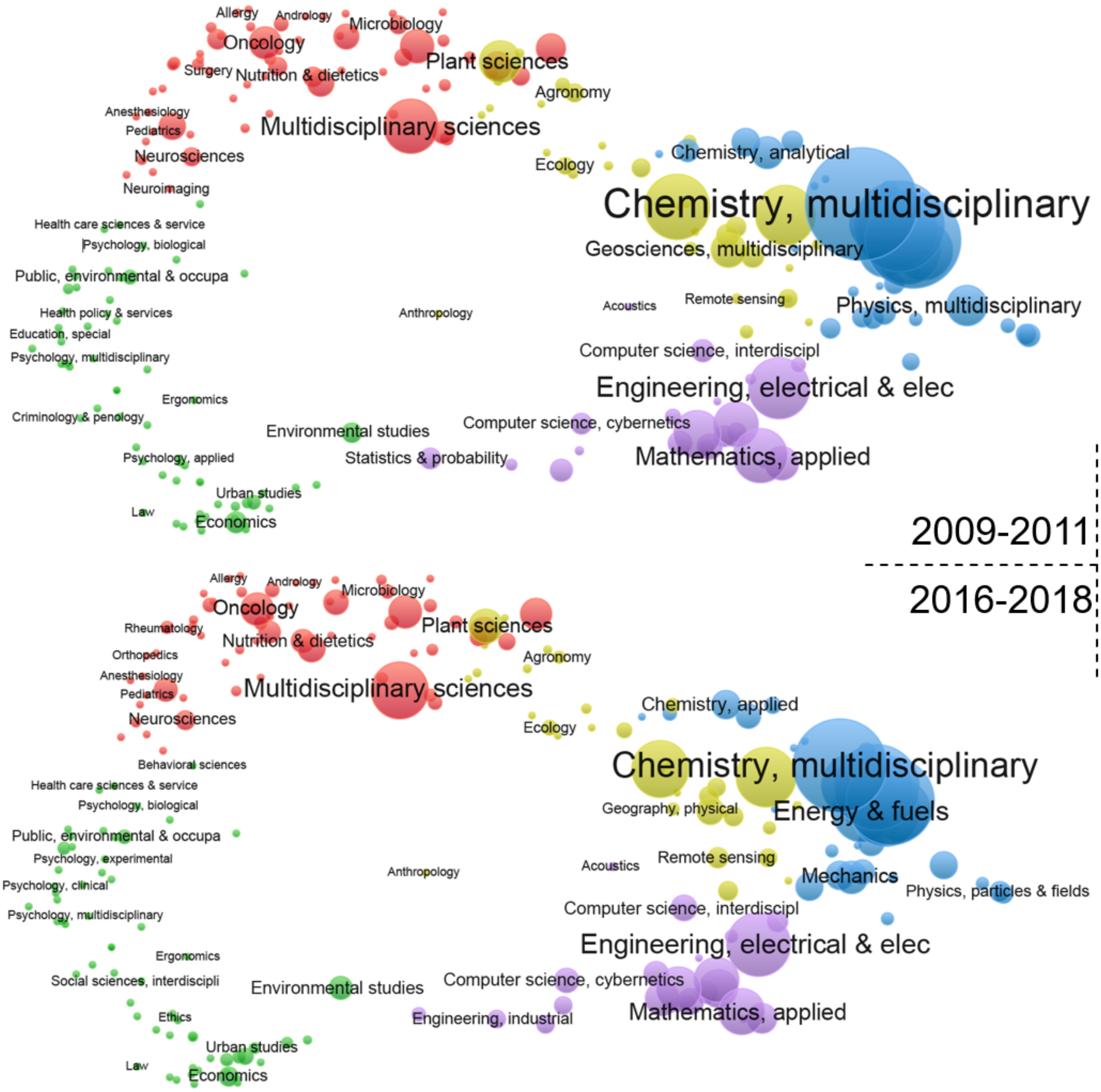
(2) The evolution of disciplinary interactions

To analyze the evolution of disciplinary interactions of China's research systems, we specifically zoom in on two time periods -- an earlier period between 2009 and 2011, and a later period between 2016 and 2018. We generate separate science overlay maps, as shown in Supplementary Figure 2. From the number of articles in each WoS category for each period, we find:

[10] More information could be found at the website: https://scholarlykitchen.sspnet.org/2020/03/03/guest-post-how-chinas-new-policy-may-change-researchers-publishing-behavior/

[11] See its introduction on the WoS's website: https://clarivate.com/webofsciencegroup/solutions/webofscience-chinese-science-citation-index/

- The disciplinary layout of China's research systems has enriched from 2009 to 2018, with an increase from 174 categories to 203.
- Most established disciplines have been maintained and strengthened: 1) the top five categories remain the same, at 4,354 articles (30.5%) of the 14,279 articles in the later period. These are "chemistry, multidisciplinary", "materials science, multidisciplinary", "chemistry, physical", "nanoscience & nanotechnology", and "physics, applied"; and 2) publishing output in 172 categories (84.7%) of the 203 categories has increased at a rate higher than 1.



Supplementary Figure 2. Science overlay maps for the periods 2009-2011 (top) and 2016-2018 (bottom).

To further explore the evolution in disciplines covered, we identified 16 emerging categories that met the following criteria: 1) a rate of increase in the number of articles is higher than 5; and 2) showing more than 100 articles in the later period. The categories are given in

Supplementary Table 2. The differences across periods in these disciplines not only reflect the changing interests of the research community, but also changes in the driving forces behind, such as national strategies and science polices. Results show some interesting patterns.

- Computer science and related disciplines are definitely one of China's research strengths, and, motivated by the rise of artificial intelligence techniques, such strengths were further supported by China's national strategy in 2017[12]. Driven by information technology, telecommunications (advancements like 5G) and robotics were boosted as well, which illustrates active interactions between computer sciences and engineering disciplines.
- With research strengths in chemistry, biology, and material sciences also, a cutting-edge area that holds strong interest with China's researchers is sustainable technologies -- e.g., 3D printing. Investigating such sustainability is not only associated with multidisciplinary studies in natural sciences but also with social sciences.

Supplementary Table 2. Sixteen emerging WoS categories in China (2009-2018)

No.	*WoS category*	*#L*	*#E*	*Inc.*	*Rate*
1	Telecommunications	415	22	393	17.9
2	Remote Sensing	130	8	122	15.3
3	Green & Sustainable Science & Technology	280	26	254	9.8
4	Mathematics, Interdisciplinary Applications	181	17	164	9.6
5	Mechanics	313	30	283	9.4
6	Imaging Science & Photographic Technology	113	11	102	9.3
7	Computer Science, Information Systems	356	38	318	8.4
8	Engineering, Mechanical	253	28	225	8
9	Engineering, Multidisciplinary	128	15	113	7.5
10	Instruments & Instrumentation	214	28	186	6.6
11	Thermodynamics	195	28	167	6
12	Engineering, Chemical	1142	167	975	5.8
13	Medicine, Research & Experimental	146	22	124	5.6
14	Energy & Fuels	1111	172	939	5.5
15	Computer Science, Cybernetics	183	29	154	5.3
16	Environmental Studies	159	26	133	5.1

Note that #L and #E mean the number of articles in the later period and the earlier period, respectively, and Inc. stands for an increased number of articles between the two periods.

Science overlay maps give a vivid solution for exploring the relationships among research disciplines and providing strategic insights for decision support. However, considering science maps are also a type of network, network analytics can be useful for further delving into their topological structures to deepen and expand our knowledge of those interactions (Zhang et al., 2018c). Our following discussion is based on the following assumptions:

- Science overlay maps describe disciplinary interactions through a network, in which a node represents a research discipline and an edge connecting two nodes represents the existing interactions between the two disciplines (e.g., the frequency of co-occurrence).

[12] More information can be found on the website: https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy

- It is reasonable to believe that if two disciplines do not interact with each other, yet are indirectly connected by a number of the same disciplines (i.e., common neighbors), the two disciplines will have a high probability of interacting with each other in the near future.

Given these two assumptions, we applied a link prediction approach based on a weighted index of resource allocation to the base map in Supplementary Figure 1. The results, given in Supplementary Table 3, fall into two categories: cross-disciplinary interactions that are likely to be maintained and new cross-disciplinary interactions that are likely to emerge.

Supplementary Table 3. Link prediction for cross-disciplinary interactions

Type	*Interactive pairs between WoS categories*	
Existing interactions	Materials Science, Multidisciplinary	Chemistry, Physical
	Engineering, Electrical & Electronic	Computer Science, Artificial Intelligence
	Engineering, Electrical & Electronic	Computer Science, Interdisciplinary Applications
	Chemistry, Physical	Nanoscience & Nanotechnology
	Engineering, Electrical & Electronic	Telecommunications
	Computer Science, Artificial Intelligence	Automation & Control Systems
	Biochemistry & Molecular Biology	Biotechnology & Applied Microbiology
	Materials Science, Multidisciplinary	Nanoscience & Nanotechnology
	Engineering, Electrical & Electronic	Automation & Control Systems
	Engineering, Electrical & Electronic	Computer Science, Information Systems
Potential interactions	Oncology	Neurosciences
	Engineering, Electrical & Electronic	Statistics & Probability
	Economics	Public, Environmental & Occupational Health
	Environmental Sciences	Agronomy
	Cell Biology	Pharmacology & Pharmacy
	Chemistry, Multidisciplinary	Engineering, Electrical & Electronic
	Nanoscience & Nanotechnology	Biochemistry & Molecular Biology
	Engineering, Electrical & Electronic	Mechanics
	Chemistry, Multidisciplinary	Engineering, Environmental
	Chemistry, Physical	Optics

Supplementary Table 3 highlights two sets of existing cross-disciplinary interactions that coincide with China's research strengths and, pursuant to this, that are supported by national strategies:

- Driven by artificial intelligence techniques and the visionary applications of internet of things, as well as 5G and robotics, interactions between computer science (e.g., artificial intelligence, information systems, and cybersecurity) and its applications in engineering areas, such as electrical and electronic engineering, telecommunications, and automation are rapidly spearheading a cutting-edge direction.
- New materials and novel manufacturing processes in the areas of chemical engineering and biological engineering are among the most significant innovations these days, and nanotechnologies should further enhance the practical capability of those inventions.

In terms of potential interactions, the predicted pairs of disciplines are all promising and explainable, such as oncology and neuroscience. However, an interesting observation is that some disciplines already appear to have established interactions. For example, statistical models and probability models are widely used in the area of electrical and electronic engineering, and economics has long been connected with public health and related disciplines. One note here is that WoS categories are based on the subject areas of journals. Hence, non-existent interactions may simply mean no journal covers those two disciplines at the moment. In other words, there could be a number of existing studies and published papers covering topics in those disciplines, e.g., a study concerning economics and public health may have been published in a journal dedicated to economics. Therefore, the predicted interactions could be considered as either: 1) an interaction between two indirectly linked disciplines; or 2) a cutting-edge research direction based on a relatively innovative idea.

Appendix B Supplementary Tables

Supplementary Table 4. Evaluation for the five problem-solving patterns in biology and life sciences.

No. in Table 5	*E1*	*E2*	*E3*	*E4*	*E5*	*Avg.*
#3	E	D	D	B	E	0.25
#8	B	D	A	B	C	0.65
#12	B	B	B	A	B	0.8
#14	B	E	C	D	D	0.35
#16	C	B	A	A	A	0.85
Total						**0.58**

Note that we calculated the average score by transferring the scale of A-E as 1, 0.75, 0.5, 0.25, and 0 respectively.