The Contribution of Chinese Diaspora Researchers to China's Catching Up in Global Science and High-Tech Industries

Abstract

This study examines the contribution of Chinese diaspora researchers – those born in China but working outside the country – to China's catching up in global science to become a leader in academic research and in research intensive industries. Our evidence shows that diaspora researchers produce a large proportion of global scientific papers, with high scientific impact, and are a critical node in the co-authorship and citation networks linking China and the rest of the world. Diaspora researchers also play a special role in China's catch-up by citing more Chinese-based papers and being cited more by Chinese-based papers than other papers written outside the country. Diaspora researchers also make an exceptional contribution in computer science and artificial intelligence, where they both work for high tech Chinese and US firms. By advancing global scientific knowledge and connecting China-based research with the rest of the world, diaspora researchers mutually benefit their country of origin and the countries in which they work.

JEL: F1, I2, J2, J3, J5, O3
Keywords: China, Scientific Research, Catch-up, Diaspora Author, Citation, International Collaboration, R&D Intensive Industries, AI

Qingnan Xie
Nanjing University of Science and Technology
200 Xiaolingwei Street
Xuanwu District, Nanjing, Jiangsu 210094, China
e-mail: <2362626753@qq.com>

Richard B. Freeman
Harvard University and NBER
1050 Massachusetts Avenue
Cambridge, MA 02138, USA
e-mail: <freeman@nber.org>

Oct 4, 2020

ACKNOWLEDGEMENTS:
"We gratefully thank Dr. Xi Hu from University of Oxford for her support in collecting data from Scopus. Qingnan Xie is supported by the National Social Science of China [16ZDA224]."

Introduction

In the latter part of the 20th century and early decades of the 21st century China advanced from the periphery of the global economy, accounting for barely 2% of world GDP and 1% of world trade in the early 1970s, to second largest economy with about 17% of world GDP in 2019 and 12.4% of world trade in 2018.[1] In science and engineering research China made a similarly impressive catch-up to become the top country in papers published, and to second place (to the US) in citations by the late 2010s (Xie and Freeman, 2019). Combining improved research capability with manufacturing prowess, China further advanced in high research-intensive industries, and made an especially notable progress in artificial intelligence, which many view as *the* technology of the future.

China's economic rise from low income to middle income country was predicated on economic reforms that fit well with precepts of development economics, albeit with Chinese characteristics that gave the government a large role in the direction of change (Lin, 2011). The 1970s agricultural reforms increased rural productivity, opening the door for rural workers to migrate to urban areas. Public investments in infrastructure expanded domestic markets. Marketization allowed new private sector firms to enter low skill manufacturing, and eventually to hire upwards of 200 million rural workers who migrated to cities. Finally, China's 2000 membership in the World Trade Organization brought the country into the center of global trade, where its comparative advantage in low skill labor made it the world's manufacturing hub. Debates about the efficacy of particular reforms notwithstanding, there is widespread agreement that decentralization of economic decisions from the state to market contributed to China's growth.

There is no comparable standard path for a developing country to rise in science, where national science, technology, and innovation policies (STI) (Chen et al., 2020) and spending on higher education and research rather than natural resource endowment determine comparative advantage. Traditionally, countries invest in higher education and R&D late in the development process, and develop STI policies even later. In China the 1966-76 Cultural Revolution had devastated the country's universities and research, placing it at rock bottom in all of these areas when Deng Xiaoping initiated economic reforms. From the late 1970s through the 1990s, China expanded its institutions of higher education and developed new ones to enroll and graduate millions of bachelor's degree holders and large numbers of master's and PhDs, mostly in STEM fields. But it did not have the scientific expertise to play more than a minor role in global scientific research nor to make headway in high tech manufacturing and service sectors. Catch-up in science and high-tech industrial production awaited the new century.

In this study we examine the contribution of researchers born in China while doing research outside the country – the *diaspora researchers* of our title – to China's move to the frontier of science and technology in the first two decades of the 2000s. Since, by definition, diaspora researchers are migrants from one country to another, our paper contributes to research on the effect of high skilled immigrants on source and destination economies.

There are two competing views in this area of work. Traditional brain drain literature views emigration as a loss that weakens the ability of the source country to upgrade its productive capacity

---

[1] Trade figures from https://www.wto.org/english/res_e/statis_e/wts2019_e/wts2019_e.pdf

and to catch up with economically advanced countries (Docquier and Rapoport, 2012). These studies assess the costs of the brain drain to the source country and often seek ways to recompense the low-income source country for their loss of the "best and brightest" or to subsidize home-grown researchers (Docquier, Lohest and Marfouk, 2007; Cao, 2008; Ziguras and Gribble, 2015). By contrast, studies in large immigrant-receiving countries focus on the importance of immigrants in the supply of scientists and engineers and the high quality of their work. The US Science and Engineering Indicators 2020 shows that in 2017 the foreign born made up about one-third of S&E researchers in academia, many of whom came to the country as graduate students; about half of post-docs; and that 30% of science and engineering faculty were foreign born. Stephan and Levin (2001) documented the exceptional contribution of immigrants to US academic research and patents.

The "ethnic network view" offers a different perspective. It treats highly skilled migrants as a positive channel of communication and knowledge that allows the source country to access advances in science and technology more rapidly than would otherwise be possible (Kerr, 2008). Research on ethnic networks finds trade links between the country of emigration and the country of immigration (Saxenian and Hsu, 2001; Felbermayr, Jung and Toubal, 2010; Aleksynska and Peri, 2014; Behncke, 2014), and greater diffusion of technology both from origin to destination and from destination to origin countries (Lissoni, 2018), with effects that differ between the most innovative and average innovations (Agrawal, Kapur, McHale and Oettl, 2011). In contrast to the view that both source and destination countries benefit from migrant scientists and engineers, the Trump administration in the US has worried that Chinese diaspora researchers might "steal intellectual property and technology from the United States",[2] particularly in artificial intelligence, which some see as the key to economic and military supremacy in the 21st century. These developments place the role of Chinese diaspora researchers at the forefront of policy discussion on the globalization of science and mobility of scientists.

Section one uses data from the Scopus data base of scientific articles to measure the contribution of diaspora researchers to China's catch-up in scientific research. Section two examines the position of diaspora research in the co-authorship and citation networks that connect China-based research to research in the rest of the world. Section three uses data from the proceedings of major AI conferences to measure the role of diaspora researchers in AI. Section four concludes.

1. Quantity and Quality of Scientific Papers of Diaspora Research Scientists

Do diaspora researchers contribute a sufficiently large number of scientific papers to merit the detailed attention that we give them in this paper?

We answer this question by analyzing the *names* and *addresses* of authors in physical and natural sciences, including engineering and mathematics journal articles published in 2018 contained in the Scopus database.[3] Our analysis goes beyond bibliometric studies that assess the contribution of

---

[2] https://www.whitehouse.gov/briefings-statements/president-donald-j-trump-protecting-america-chinas-efforts-steal-technology-intellectual-property/.

[3] Scopus is the largest bibliometry of scientific journals with wide coverage of China-published English and Chinese language journals. English is the primary language of science and the language for 88% of Scopus journal articles. The 350

countries or groups based on the addresses of authors. We use the last names of Chinese authors to determine their ethnicity and their first names to determine their likelihood of being born in China.

For many groups, names do not identify country of origin – among authors with a US address, John O'Leary could be an Irish immigrant or a US born Irish-American, Ingrid Swenson could be a Swedish immigrant or US born Swedish-American, and so on. But the distinctive names of Chinese born authors allows us to identify ethnicity and place of birth. Last names identify ethnic Chinese – Yang, Lui, Xie are common Chinese last names. First names allow us to differentiate ethnic Chinese born in China from those born outside the country. Mainland born Chinese almost invariably have Chinese first names as well as family names – Xixi, Wei, and Fang – while those born outside China are likely to have a first name that fits their country –Sharon, David, Pierre, and so on.[4]

Based on names, we define a Chinese *diaspora (D) author* as an author with first and last Chinese names writing an academic paper with an address outside China. We define a Chinese *diaspora paper* as a paper with one or more such authors. Thus, a paper by Qing Yang at US address would be a diaspora paper and Qing Yang would be a diaspora author. By contrast, author David Yang at a US-address would not qualify as diaspora and his paper would count as a US paper. We label papers with all non-China addresses but at least one diaspora author as a *non-China diaspora* (NCD) paper.

Since papers have many authors, the number of diaspora authors can differ. Following the convention that gives fractional credit of a paper to a country based on the proportion of authors with that country's address, we measure a papers' "diaspora-ness" by the diaspora proportion of authors. A three authored paper with addresses outside China would be 100% diaspora if all three authors have Chinese first and second names while a paper with all addresses outside China would be 1/3rd diaspora if just one author had Chinese first and second names.

An increasing share of global papers are international collaborations that combine the work of researchers in different countries (including in some cases the same author who reports addresses from more than one country). We label papers written by researchers in China and researchers outside China as China joint papers (CJ). The papers with one or more Chinese named author at a non-China address are China *joint diaspora* (CJD) papers. The diaspora-ness of a CJD paper is the ratio of the number of diaspora authors to all authors, including those with China addresses.

How many?

To estimate the number of Chinese diaspora papers, we gathered Scopus data on 1.6 million English language articles in natural and physical sciences, including engineering and mathematics. Of those articles, 16.8% had all China addresses, and are thus not diaspora. Diaspora authors are found in the 83.2% of papers with all non-China addresses or with both China and non-China addresses. We

---

active Chinese language journals in Scopus make Chinese the 2nd largest language, accounting for 4.8% of 2018 articles.

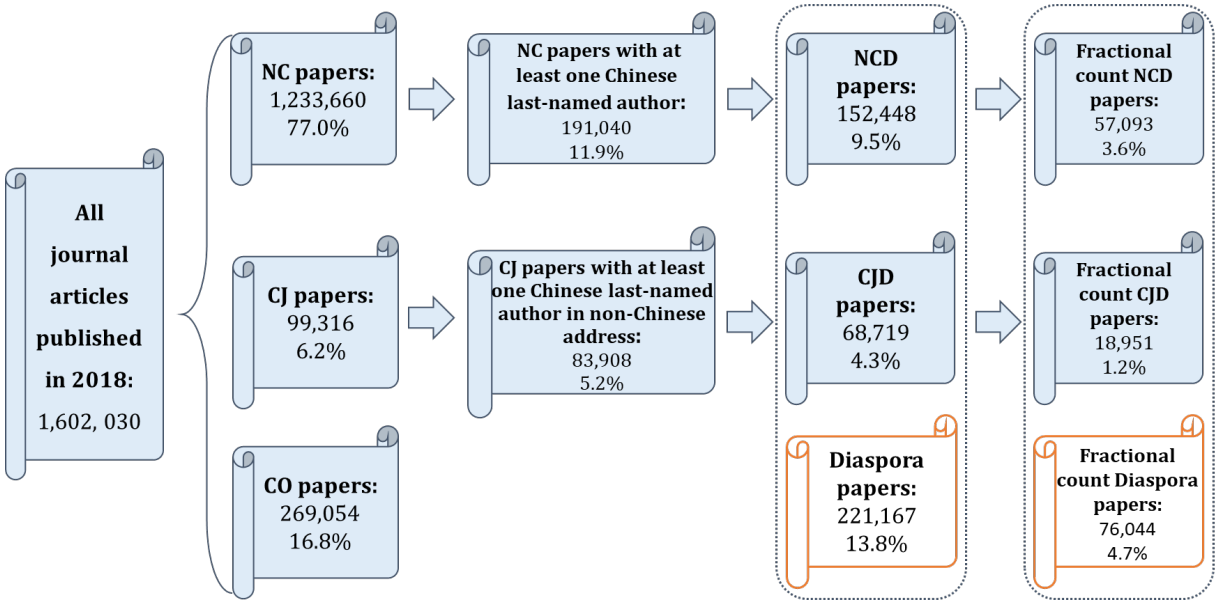
[4] We identify Chinese last-named authors using common Chinese last names from the household registrations of the Chinese Ministry of Public Security: http://www.mps.gov.cn/, accessed June 26, 2017. This list has the Chinese last names of 84.8% of the population, leaving out uncommon last names, including some non-Han minority names. We label observed authors as Chinese by matching their last name with the list of the most common Chinese last names and used differences in Pinyin spelling between Mainland China and other Chinese language areas to distinguish Mainland names. We then manually check the first names of the authors with Mainland Chinese last-names to determine if they are Chinese born authors using the grammar of Mainland Pinyin system. Our codes and name list are available on request from the authors.

estimated the number of diaspora papers in these papers by *counting* the number of papers with at least one Chinese last-named author at a non-China address, and then *estimating* the proportion of authors who had Chinese first names on a sample of papers, as described in detail in Appendix A.

Figure 1 shows the results. The figure assigns papers by address to three exclusive groups: those with China only (CO) addresses, those with non-China only (NC) addresses: and those with joint China and non-China (CJ) addresses. It then gives our counts of papers with at least one Chinese last-named author in the NC and CJ groups followed by our estimated number of papers with Chinese first and last-named authors. The note to the figure gives the acronyms used in this paper for different address-name groups.

The estimates show that the largest number of diaspora papers come from NC addresses (9.5% of all papers), which is over double the 4.3% coming from China joint papers. The 13.5% sum is our estimate of all papers with at least one China diaspora author. Comparing the CJD share of the global total to the CJ share shows that 69% (= 4.3%/6.2%) of collaborations between China-based researchers and non-China based researchers involve a diaspora author, which suggests that diaspora authors have a special role connecting China-based researchers and researchers outside China in collaborative work.

Figure 1. Numbers of Journal Articles by Address and Names of Authors and Numbers Relative to World papers, 2018.



Note: Acronyms for the address-name papers

CO: Papers with China Only addresses;
CJ: Papers with joint China and non-China addresses;
NC: Papers with non-China only addresses;
CJD: CJ papers with at least one Diaspora author;
CJN: CJ papers with no Diaspora author; N
CD: Papers with no Chinese address and with at least one Diaspora author;
NCN: Papers with no Chinese addresses nor Diaspora author.
D author: Diaspora author; Author with Chinese first and last names and a non-Chinese address;

NCN author: non-Chinese addressed and non-Chinese named author

Source: Scopus English language journal articles in physical and natural sciences, including mathematics and engineering. This excludes papers in social sciences; arts and humanities; psychology; business, management and accounting; economics, econometrics and finance; decision sciences, and undefined. Appendix A describes the statistics and the sample of papers used to estimate the proportion of authors with Chinese first as well as last names.

Crediting diaspora research for an entire paper when diaspora researchers make up only part of the authorship arguably exaggerates the diaspora contribution. The natural way to address this problem is to fractionalize the credit for a paper by its diaspora proportion of authors. The statistics on the far right of Figure 1 show that 4.7% of fractionalized 2018 papers are attributable to diaspora researchers.[5] If researchers were a country, this proportion would place them fourth in the world behind China, the US, and India as a producer of papers measured by fractionalized addresses.[6]

Diaspora share of China's scientific publications

Figure 2 organizes the data on diaspora papers to measure their quantitative importance in *China's* publications using different measures of Chinese contribution to publications.

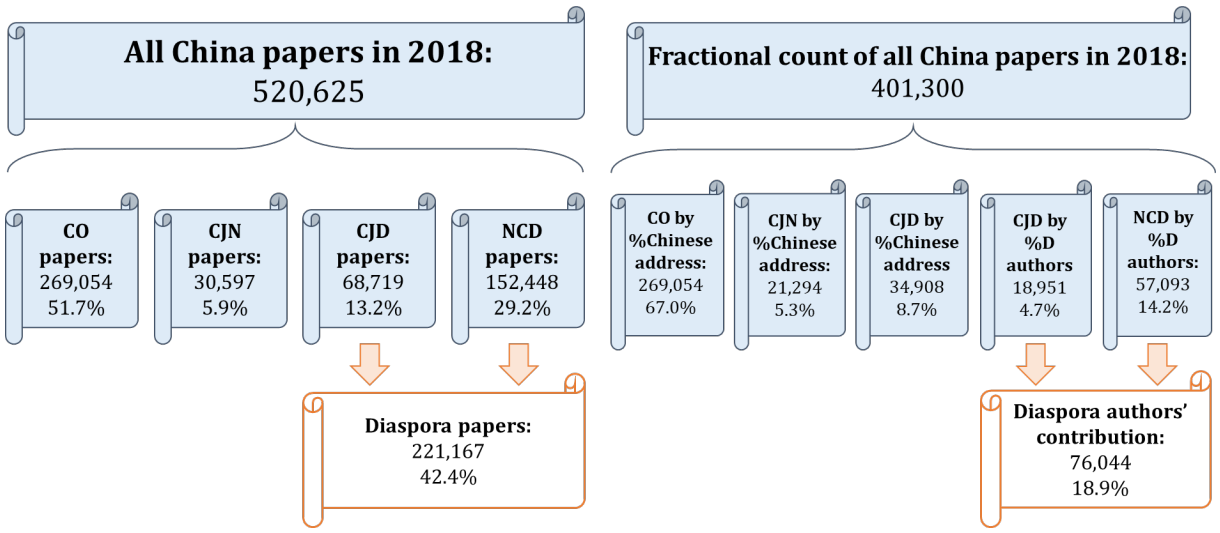
The first measure is China's *presence in scientific literature,* which we count as the sum of papers with at least one Chinese name or address, and which thus includes all CO, CJ, and NCD papers. Counting every paper with a Chinese presence irrespective of the proportion of addresses or authors from China is a maximal measure of China's scientific activity. In 2018 China had a presence on 520,625 scientific papers, 42.4% of which are diaspora papers.

The second measure fractionates China's contribution by the China address or diaspora share of authors. The number of CO papers is the same in the two measures, but the number of CJ and NCD papers is smaller as each paper is discounted to the Chinese proportion of addresses and names on the paper. This measure gives China 401,300 scientific papers in 2018, of which fractionated diaspora papers make up 18.9% of the total.

Figure 2. Journal Articles with Chinese Addressed Authors or Chinese Named Authors by whole count and fractional count, 2018

[5] We prorated the address share of credit for an author with addresses in China and another country on the extent of a paper being diaspora by giving ½ to each of the two country addresses. For example, we credited a paper with one Chinese named author with both a Chinese and a US address 3/4th to China and 1/4ths to non-China. In an n-authored paper, this gives 3/4n to China and 1/4n to non-China. Because non-China named researchers with China addresses are a negligible part of China addressed papers, we ignored them but their contribution could be divided similarly by names and addresses.

[6] National Science Board (2020) Table 5A-1 shows India with 5.3%, Germany with 4.1% and Japan with 3.9% of papers in 2018. We obtain the same ranking in our compilation of English journal articles in Scopus.



Note: All China papers include papers with at least one Chinese addressed author or Chinese named author, which are the union of CO papers, CJN papers, CJD papers, and NCD papers.

Source: Scopus database

How much of China's rise to number one producer of papers is attributable to increased numbers of diaspora papers? Appendix Figure B, which divides global papers in 2000 and 2018 among our address and name categories, shows that the primary force driving the 2000-2018 rise to number one in global papers is the growing number of China only papers, which expanded seven-fold from 2.4% of the world's total papers in 2000 to 16.8% in 2018. Diaspora papers increased rapidly as well – from 8% of China's contribution to world papers in 2000 to 13.8% in 2018 – to account for 27% of China's increased share but could not match the seven-fold expansion of CO papers. The only diaspora papers that kept pace with the seven-fold growth were CJD collaborations with China based researchers, whose share increased from 0.6% of world papers (2000) to 4.3% (2018).

If diaspora papers had similar quality/impact as papers written by non-diaspora researchers, the Figure 1 and Figure 2 and Appendix Figure B evidence would be our bottom line on the diaspora contribution. But Scopus data on citations and the cite scores of journals that published diaspora papers shows that, far from being of average quality/impact, diaspora papers lead the world in citations and cite scores, magnifying their contribution to China's rise in scientific prominence nearly twofold.

Impact/quality of diaspora research[7]

Our first measure of the impact/quality of diaspora research compared to non-diaspora research are the forward citations given to the papers. Forward citations measure the attention an article receives from researchers *after* its publication. To the extent that researchers build on the findings or

[7] Because citations and cite scores depend on social factors such as the size of an author's research network, their fame or the prestige of their university or research center, and their connection to potential reviewers and editors, etc., as well as the paper's "innate quality" we refer to them as indicators of impact/quality.

methods in a paper, they will cite the paper.[8] Expansion of the scientific literature over time increases the number of potential citations to existing papers. With Scopus data that ends in 2018 we focus our citation analysis on 3 year forward citations on papers published in 2015. The 3-year period provides a reasonable indicator of the likely position of papers in citation distributions over time.[9]

Our second measure of impact/quality is the cite score of the journal of publication. Cite scores measure the attention to the journal of publication *before* the given article appears. Scopus computes cite scores as the ratio of total citations to the journal divided by the number of articles over the past three years. Because high cite score journals attract many submissions from which they accept relatively few articles, articles published in a high cite score journal are likely to be of high quality as they have to overcome a high acceptance hurdle.

Citations and cite scores are highly correlated[10] but reflect different evaluation processes by different groups of scientists that justifies our analyzing both. The authors of future papers decide whether or not to cite a published article based on the influence the article had on their thinking or work. The reviewers and editors who choose whether to accept or reject an article in a given journal base their decisions on expectations of the article's validity and importance in the future. Analyzing results with both measures provides an independent replication or robustness test of findings.

Citations

Table 1 records three-year citations for 2015 papers differing in diaspora status. Given the heavy power-law tail of citations, in which many papers receive a few citations and a few receive many, the table gives the median of citations and the mean of the upper decile of the citation distribution as well as the more commonly reported mean. The statistics show that diaspora papers gained roughly **twice** the citations of NC papers without diaspora authors and of CO papers. The *diaspora advantage* is larger in mean than in median citations and is largest in the mean of the upper 10% of papers, indicative of the skew of the citation distribution. Measured by means, NCD papers lead all others but measured by medians, CJD papers top all others. Collaborative papers without diaspora authors (CJN) obtain more citations than CO papers and NCN papers while falling short of diaspora paper citations.

Table 1. Average 3 year forward citations of papers published in 2015

Papers by address-name group	Mean	Median	Mean for top decile of group
1 NCD – NC (Non-China Only) papers with one or more China named authors	**18.3**	**8.0**	**103.9**

[8] Analysis of citations by scite.ai show that about 90% of citing articles mention a paper without much comment while the number of positive comments outweighs critical comments by about 4 to 1.

[9] Three year forward citation in a sample of 5989 papers published in 2000 had correlations of 0.97, 0.89, and 0.68 to their citations 5, 7, and 10 years in the future. An extensive literature examines ways to predict later citations from early citations and other attributes of papers (Bornmann, Leydesdorff, Wang, 2014; Abrishami and Aliakbary, 2019).

[10] We obtained a 0.5 correlation between three year forward citations and cite scores in a sample of 5,540 papers published in 2015 with valid cite scores. The correlation fits with Larivière et al. (2016)'s data on within-journal variation in citations

2 **CJD – CJ papers with diaspora author (CJD)**	**17.5**	**10.0**	**85.5**
3 CJN – CJ papers without diaspora author	12.4	7.0	51.2
4 CO – China Only addressed papers	9.1	5.0	37.4
5 NCN – NC papers with no China named author	8.5	5.0	34.3

Note: The standard errors for the means in citations are 2.1, 1.0, 0.9, 0.3, and 0.3.

Source: All measures are based on 2,000 yearly CO, CJ and NC samples, see Appendix A for details.

Another way to assess the quality/impact of diaspora researchers is to examine the position of these scientists on rankings of scientists by number of citations. In 2011 Clarivate Analytics published the "Top 100 Materials Scientists" based on 2000-2010 citations in its Web of Science data. Table 2 shows that five of the top 10 had Chinese first and last names and worked outside of China – diaspora authors. The five were employed by leading US universities. They all graduated from the University of Science and Technology of China, which suggests their China education contributed to their success. In the entire list, 12 of the top 100 material scientists were diaspora scientists.

Table 2. Top Ten Material Scientists, 2000-10, Ranked by Total Citations

Rank	Name	Current Employer	Bachelor's degree if had China education.	Citations	Papers
1	**Peidong Yang**	Univ Calif Berkeley	University of Science and Technology of China	13,900	36
2	**Younan Xia**	Washington Univ, St. Louis	University of Science and Technology of China	11,936	83
3	**Yiying Xu**	Ohio State	University of Science and Technology of China	9,590	74
4	N. Serdar Sarificitci	Johnnes Kepler Univ, Linz		6,444	74
5	**Yadong Yin**	Univ Calif Riverside	University of Science and Technology of China	6,387	32
6	Alan Heeger	Univ Calif Santa Barbara		5,788	49
7	Frank Caruso	Melbourne		5,589	
8	Michael Huang	National Tsing Hua University, Taiwan		5439	34
9	**Yugang Sun**	Argonne Nat'l Lab	University of Science and Technology of China	**5,231**	**37**
10	Galen Stuckey	Univ Calif Santa Barbara		5,095	72

Note: Our ranking is based on total citations, whereas the Clarivate ranking is based on the ratio of citations to papers, which causes some differences between their statistics and ours. Diaspora researchers are in bold.

Source: Tabulated from *Clarivate Science Watch*, 'Top 100 Materials Scientists'.
http://archive.sciencewatch.com/dr/sci/misc/Top100MatSci2000-10/

Cite scores

Table 3 records the cite scores of papers in different address-name groups.[11] Consistent with the citation data, the cite scores show diaspora papers leading the list. The magnitude of the differences are however, smaller than in citations due to the distribution of cite scores more concentrated around its mean as a result of averaging citations from many articles. Even so, the diaspora advantage is high, with NCD papers have 1.5 times the mean cite score of NCN papers and 1.6 times the mean cite score of CO papers.

Table 3. Average Cite Scores of papers published in 2015

Papers by address-name group	Mean	Median	Mean for top decile of group
1 NCD – NC Papers with one or more China named authors	5.0	4.1	14.4
2 **CJD – CJ papers with diaspora author (CJD)**	4.9	4.1	13.6
3 CJN – CJ papers without diaspora author	4.2	3.4	11.0
4 CO – China Only addressed papers	3.1	2.7	8.3
5 NCN – NC papers with no China named author	3.4	2.7	9.3

Note: The standard errors for means of Cite Scores are 0.2, 0.1, 0.2, 0.1, and 0.2. The Cite Score values are assigned to papers based on the Cite Score of the journals in which they appeared. Scopus does not assign a Cite Score to new or inactive journals so observations on those journals are excluded at the Cite Score calculation. We use the 2017 version Cite Score list issued by Scopus, Downloaded at 25 May 2018.

Source: All measures are based on 2000 yearly CO, CJ and NC samples, see Appendix A for details

Exemplifying the success of diaspora papers in getting into top scientific journals, we examined the diaspora share of papers in *Nature* and *Science* in 2000 and in 2018. Table 4 shows that in 2000 *Nature* and *Science* published virtually no papers with only China addresses and relatively few joint China-other country collaborative papers. The only Chinese born researchers with noticeable representation were authors of diaspora papers, with 16.4% of *Nature* papers and 18.1% of *Science* papers in our NCD group. Between 2000 and 2018, despite the seven-fold increased CO share of all published articles, the CO share of *Nature* and *Science* articles remained low. The big increase in China's presence in *Nature* and *Science* was in diaspora article. In 2018 30.3% of papers in *Nature* and 35.0% in *Science* had a diaspora author.

Since Chinese authors and addresses are only part of the authors and addresses on diaspora papers, fraction counting them by their share of authors and addresses reduces the credit given to China. Even so in 2018 NCD papers had a larger share of *Nature* (3.4%) and *Science* (3.9%) articles than did the far more numerous China only papers (0.9% and 2.6%, respectively).

[11] As cite scores are highly correlated over time, the results should be similar with modestly different year coverage. The correlation for the cite score of Scopus journals is 0.93 between 2017 and 2015, and is 0.87 between 2017 and 2011.

Table 4. Chinese Diaspora Papers in *Nature* and *Science*, 2000 and 2018

	2000	2018	2000	2018
	Nature		*Science*	
Proportion of papers				
1. Papers without Chinese address but with at least one China named authors (NCD)	16.4%	24.6%	18.1%	27.0%
2. China Joint papers with diaspora authors (CJD)	0.2%	5.7%	0.2%	8.0%
3. China Joint papers without diaspora authors (CJN)	0.2%	3.4%	0.5%	2.1%
4. Only China addressed papers (CO)	0.3%	0.9%	0.2%	2.6%
5. Non-China Addressed Papers with no China name author (NCN)	82.8%	65.3%	80.9%	60.3%
Proportion of papers, fractional counts by addresses and names				
1. Papers without Chinese address but with at least one China named authors (NCD)	2.5%	3.4%	3.1%	3.9%
2. China Joint papers with diaspora authors (CJD)	0.1%	1.7%	0.1%	3.2%
3. China Joint papers without diaspora authors (CJN)	0.1%	1.5%	0.2%	0.8%
4. Only China addressed papers (CO)	0.3%	0.9%	0.2%	2.6%
5. Non-China Addressed Papers with no China name author (NCN)	97.0%	92.5%	96.4%	89.4%

Note: Tabulated from every edition of *Nature* and *Science* in the specified year.

Source: Scopus database

The diaspora advantages in citations and cite scores could be due to differences in the attributes of papers and authors beyond addresses and names – for instance their field of study, number of authors, or other factors associated with citations or with publication in more prestigious journals (Börner et al., 2010; Abramo and D'Angelo, 2015). To see if our estimated diaspora advantages hold up in the face of other determinants of citations and cite scores we estimated a linear regression model that linked the number of citations and cite scores to dummy variables for the different address-name groups of papers by themselves and then added a set of other determinants of citations or cite scores – field of study measured by dummy variables for 21 fields and the numbers of authors on a paper. The regression results in Appendix Table C show that while inclusion of field dummies and numbers of authors greatly improves the fit of the equations, their inclusion in the regression reduces the coefficients on the NCD and CJD only modestly.[12]

[12] We explored four non-linear specifications as well: (1) a log regression with one citation added to each observation to keep 0 citation papers in the regression; (2) a log regression limited to positive citation observations with a separate equation that estimates the impact of factors on the probability of positive citations; and (3) a regression with citations and

Bottom line, analysis of citations and cite scores shows that diaspora research had a bigger impact on China's catch-up in science than indicated by numbers of papers. Our regression analysis indicates that a diaspora paper gains 1.9 times as many citations as a non-diaspora paper. Adjusting numbers of diaspora papers upwards for this quality advantage increases the diaspora contribution to global science from 13.8% of papers to 24.9% of "quality adjusted" papers.[13] The effect of cite scores on quality adjusted numbers is smaller but still substantial.

By producing many papers of high impact/quality outside China, diaspora researchers contribute to the global stock of knowledge but because these papers are a common good for all to use, they do not necessarily advance research in China more than anywhere else. To boost China's catch-up in science and engineering research, diaspora work must have a larger impact on Chinese research than on research in the rest of the world. Viewing diaspora research from the perspective of the co-authorship and citation networks of science, we show next that diaspora research is a critical node connecting Chinese research to research outside the country. Diaspora research plays a special role in China's catch-up in science by citing more Chinese-based papers and being cited more by Chinese-based papers than other papers written outside the country.

2. Diaspora Research as Node Connecting China and ROW

The ethnic network view of mobility treats highly skilled immigrants as a conduit of knowledge between source and destination locations that turns "the old dynamic of 'brain drain' … to ... 'brain circulation'" (Saxenian, 2002). In this section we examine the extent to which diaspora research is a quantitatively important conduit of knowledge between China and the rest of the world through the co-authorship and citation networks among research publications.

Co-authorship network

At least since Newman (2004) networks of co-authors in scientific publications have been viewed as "small-worlds" in which most researchers work with a few others near them in geographic space or with similar personal attributes while a few researchers work directly with people far away per the Watts and Strogatz (1998) small world model. The few create long distance connections that reduce the number of links for information to pass through the network.

Researchers coauthor extensively with people near or like them along many dimensions (Yan and Ding, 2012), ranging from country (Schubert and Glänzel, 2006) to ethnicity within the same country (Freeman and Huang, 2016) to gender (Wang, et al 2019). The reason for working with scientists geographically near is the lower cost of connecting with them than with far away co-

---

cite scores scaled into a 0-1 interval by dividing each observation of a variable by its maximum value; and (4) a power-law regression of the Ln of citations on the Ln rank of citations. These results are available as supplementary material from the authors.

[13] The number of diaspora papers was 220,974 (see Figure 1) and the number of all English journal articles was 1,602,030. Using the Appendix Table C column 2 regression coefficients that estimated the relation between citations and types of papers conditional that papers are in the same fields and have the same number of author, we estimated that diaspora share of citations adjusted for relative number of citations as (NCD+CJD) = [(9.44+7.92)*152,255 + (8.55+7.92)*68,719] / [(9.44+7.92)*152,255 + (8.55+7.92)*68,719 + (3.88+7.92)*30,597 + (1.24+7.92)*269,054 + 7.92*1,081,405] = 24.9%. 7.92 is the mean of NCN deleted group in the regression.

researchers. The reason for homophily by gender or ethnicity is presumably the greater ease of communicating with people like oneself.

Table 5 shows a huge division in co-authorship in 2018 papers between persons with a Chinese address and those with a non-China address compared to the distribution that would arise if the formation of co-authorship teams was independent of location and ethnicity. Column (1) gives the distribution of papers in the entire data set, per figure 1. Column (2) gives the distribution of papers having exactly five authors, which is the average number of authors on a paper rounded to an integer. The distributions of all papers and five-authored papers by address-name group differ only modestly.

Table 5. Observed and Expected share of CO, CJN, CJD, NCD, and NCN papers, 2018

	Observed share of all papers	Observed share of all 5-author papers	Expected share of all 5-author papers
CO	16.8%	19.8%	0.1%
CJN	1.9%	2.1%	55.6%
CJD	4.3%	3.7%	17.6%
NCD	9.5%	8.6%	8.9%
NCN	67.5%	65.8%	17.8%

Note: Expected shares are calculated by proportions of Chinese addressed authors published in 2018: 23.2%, proportion of diaspora authors published in 2018: 6.0%, and proportion of non-Chinese addressed and non-Chinese named authors published in 2018: 70.8%.

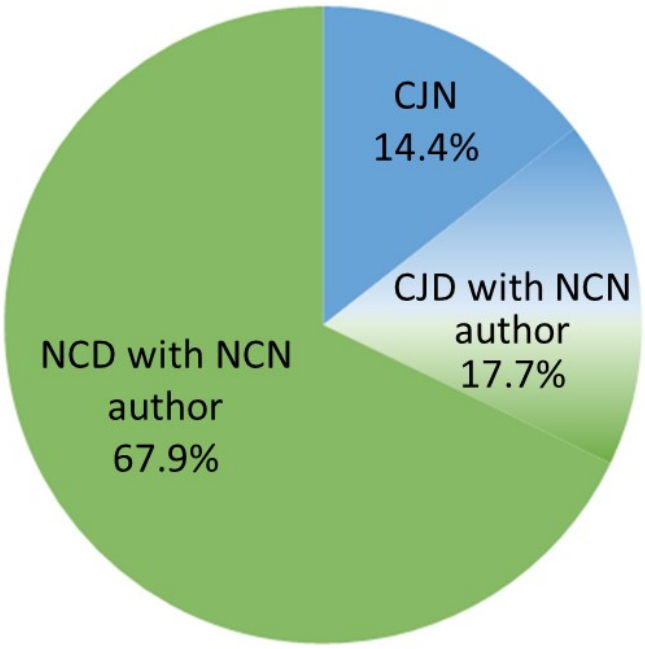
Source: Scopus database

Column (3) gives the expected distribution of papers for five authored papers absent location or ethnic homophily. We calculated this by selecting five authors randomly to be on the same paper and estimating the ethnic and location distribution of those papers. The likelihood of drawing five people from a group with α% of the distribution is α^5 so that there is essentially zero chance of getting papers with all of one address group save for the large NCN category. The actual distributions deviate massively from this. The majority of coauthors work on NCN or CO papers while the proportion doing joint collaborative papers is far smaller than in the hypothetical distribution. The NCD group is the only one in which the actual and hypothetical distributions are similar, because the diaspora ethnicity link to China balances the absence of a Chinese address.

The bifurcation of the distributions between all China and all non-China addressed papers arguably creates a need for some researchers/papers to provide the long connection that speeds the diffusion of knowledge in an efficient network. In our data, there are two long connection: CJ collaborations between China and non-China addressed researchers and NCD diaspora papers, where a Chinese author works in a non-China address. In 2018, the two collaborations accounted for 212,067 papers, which is 13.2% of global papers. The pie diagram in Figure 3 shows that the biggest source of collaborations are through NCD papers. China born researchers and non-China born researchers are far more likely to work together through diaspora papers than through international collaborations.

Figure 3. Division of Papers in Which Chinese born and non-Chinese Born Researchers Worked

Together by Type of Paper, 2018.



Note: Papers in Which Chinese born and non-Chinese born researchers worked together include all papers with Chinese ethnic authors (Chinese named or Chinese addressed authors) and non-Chinese ethnic authors (non-Chinese addressed and non-Chinese named authors). Collaboration papers of Chinese authors and non-Chinese authors is the union of CJN papers, CJD with NCN author, and NCD with NCN author.

Transmitting knowledge through citations

When China addressed papers cite non-China addressed papers, those cites provide a measure of China's "imports" of knowledge from the rest of the world (and commensurately of the rest of the worlds' "exports" of knowledge to China). Similarly, when non-China addressed papers cite China-addressed papers, those cites measure imports of knowledge from China/exports of Chinese knowledge to the rest of the world.

What role, if any, does the presence of Chinese-born researchers on diaspora papers have in these two transmissions in the citation network of science?

To see whether diaspora research has a special role in transmitting non-China based research to China addressed papers, we compare three year forward citations to diaspora papers published in 2015 from papers with only China addresses (CO papers) to three year forward citations from papers with only non-China addresses with no diaspora authors (NCN papers). Column 1 in Panel A of Table 6 shows that a 2015 NCD paper averaged 2.3 forward citations from CO papers whereas a 2015 NCN paper averaged 0.9 forward citations from CO papers– a 2.56 to 1 advantage for diaspora papers. Finding that CO papers cite NCD papers more than NCN papers does not, however, establish that CO papers rely more on those papers because diaspora and Chinese-addressed researchers are closely connected due to ethnicity. The section 2 evidence that diaspora papers are highly cited overall suggests that the diaspora differential could instead reflect the high quality of diaspora papers.

We control for the effect of quality through a difference in difference strategy. We compare the

CO diaspora differential to the analogous differential from the citations that NCN papers give to diaspora and non-diaspora papers. The NCN papers' diaspora differential is 1.64 to 1, a substantial but smaller preference toward citing diaspora papers. Assuming that CO and NCN researchers value the quality of NCD papers similarly, the ratio of the differentials 1.56 (= 2.56/1.64) is a valid "difference in difference" estimate of the tendency of CO papers to rely more on diaspora work via connections with ethnic Chinese authors. Given that NCN papers are far more numerous than NCD papers, however, CO papers invariably cite more of them in the aggregate.

To see if diaspora papers reciprocate and pay greater attention to CO research than do other non-China based papers, we computed the number of three year forward citations that a 2015 CO paper received from NCD and NCN papers. Panel B of Table 6 shows that 2015 CO papers averaged 0.6 citations from NCD papers compared to 2.1 citations from NCN papers, for a ratio of 0.29. In comparison, NCN papers averaged 0.7 citations from NCD papers and 6.4 citations from NCN papers, for a ratio of 0.11. Dividing the ratios, diaspora papers showed a 2.61 (=0/29/0.11) preference for CO papers compared to NCN papers.[14] By the ratio metric, diaspora papers are more attuned to CO papers than CO papers are to diaspora papers.

Table 6. Three Year Forward Citations received by Diaspora Papers and by CO papers compared to NCN papers published in 2015, by Specified Citing Group

Panel A. Citations of NCD papers			
2015 Papers	Three year forward Citations		Col.1/Col.2
	From CO papers	From NCN papers	
NCD papers	2.3	10.5	*0.22*
NCN Papers	0.9	6.4	*0.14*
Row 1/ Row 2	*2.56*	*1.64*	***1.56***
Preference of CO for citing NCD Papers is 1.56			
Panel B. Citations of CO papers			
2015 Papers	Three year forward Citations		Col.1/Col.2
	From NCD papers	From NCN papers	
CO papers	0.6	2.1	*0.29*
NCN Papers	0.7	6.4	*0.11*
Row 1/ Row 2	0.86	0.33	***2.61***

[14] The univariate analysis leaves open the possibility that the mean differences reflect factors associated with the papers beyond name and address. We examined citations to and from China-only addressed papers in a regression format that includes dummy variables for field of study and the number of authors. The results summarized in Appendix Table C show that while field and number of authors impact citations their inclusion in the regression does not substantially change the differentials among address-name groups

Preference of NCD for citing CO papers is 2.61

Note: Citations counts are 3-year forward citation counts. Citations to CO papers are estimated from sample of 2,000 CO papers. Citations to CJD and CJ_N papers are estimated from sample of 2,000 CJ papers. Citations to NCD and NCN papers estimated from 2,000 NC papers, described in Appendix A.

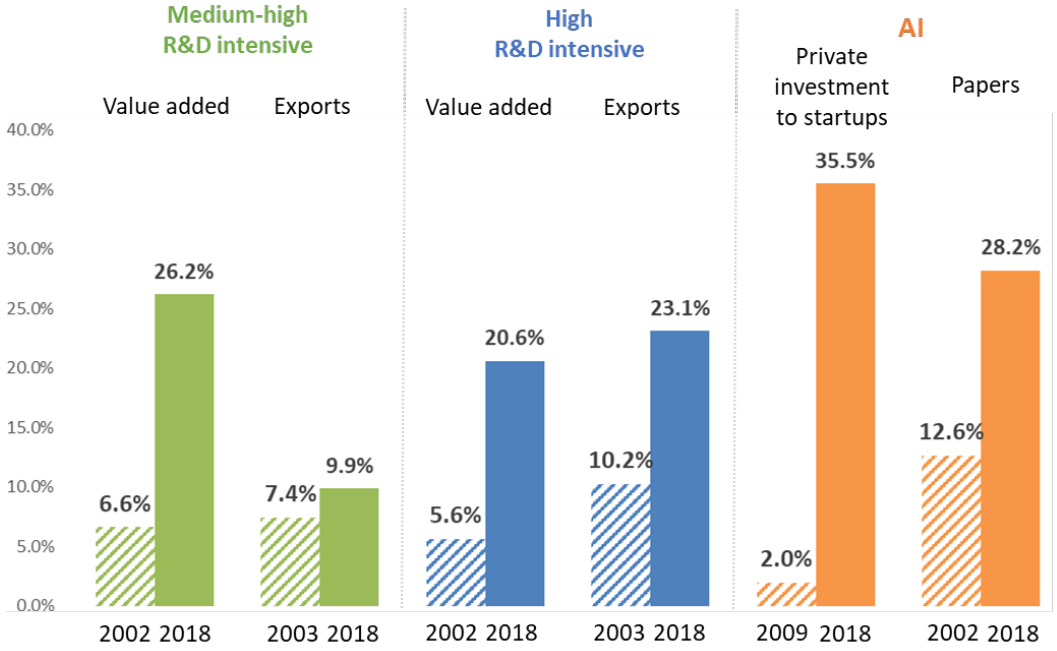
In sum, the citation network data shows that diaspora and China only papers are strongly connected in both the citing and cited directions, which makes diaspora research a potentially critical node in China's catch-up in global science.

3. From Research to High Tech Production to AI

During the same two decades in which China rose in scientific prominence, China upgraded its industrial structure toward medium-high and high research intensity industries, as classified by the OECD. The bar charts under the heading medium-high and high research intensity in Figure 4 document this change in terms of China's share of global value added and share of global exports between 2002 and 2018. The share of value added in medium-high and high research intensity increased fourfold to exceed China's 17% share of world GDP[15]. In contrast, the share of world exports increased at a much lower rate in medium-high research intensity industries than in high research intensity industries. The reason seems to be that much of the output in medium high industries went to meeting domestic consumers' demand for products such as automobiles, while output of high research intensity products was highly demanded by consumers and firms in high income countries. The archetype here is Huawei and ZTE's cutting edge advances in 5G wireless networks that raised US fears that the technology will give the Chinese government access to data for military purposes.

Figure 4. China's Shares of World Value Added and Exports of High / Medium-high R&D Intensive Industries, and China's Shares of World Private Investment in AI Startups and AI Papers (2002, 2009 and 2018).

[15] China' share of world GDP in 2018 is 16.8%. (World Bank: https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD?locations=CN&name_desc=false)



Note: AI papers include all peer reviewed papers indexed in Scopus (article, conference, and reviews). The AI field is defined mostly by keyword; for the detailed method please check method descriptions in *The 2019 AI Index*.

Source: China's shares of world value added and exports of high / medium-high R&D intensive Industries are based on statistics from National Science Board, 2020; Appendix tables, Chapter 6.
China's shares of world private investment in AI startups and AI papers are based on statistics from *The 2019 AI Index* Report by Stanford Institute for Human-Centered Artificial Intelligence; Appendix tables, Chapter 1&4. https://hai.stanford.edu/research/ai-index-2019.

The AI bar graphs in Figure 4 focus on the technology that many view as potentially the most important in the next several decades and as one highly dependent on government policy support (Chen, et al, 2020). The data shows that China was making great strides in raising private startup capital for AI and in AI research before the Chinese government made public its 2017 Plan to turn the country into a $150 billion world leader in AI by 2030 by creating "intelligent manufacturing, intelligent medical care, intelligent cities, intelligent agriculture, national defense construction and other fields".[16] The Plan produced widespread discussion of an AI arms race.[17]

What contribution did diaspora researchers make in these advances?

We answer this question by analyzing journal articles and conference papers in the Scopus database for *Computer Science,* on which many advances in medium-high and high research intensity industries rest and on conference papers from leading *AI conferences*: the Neural Information Processing Systems (NeurIPS) and International Conference on Machine Learning (ICML) conferences.[18]

---

16 https://flia.org/wp-content/uploads/2017/07/A-New-Generation-of-Artificial-Intelligence-Development-Plan-1.pdf

17 https://www.forbes.com/sites/cognitiveworld/2020/01/14/china-artificial-intelligence-superpower/#103f00412f05

18 Analyses of AI research often use NeurIPS and ICML conference papers as the best source of data on the frontier

Table 7 presents the distribution of computer science papers from Scopus among the five address-name groups that differentiate diaspora from other papers and the corresponding distribution of AI papers at the NeurIPS and ICML conferences. Columns 1 and 3 give the number of computer science papers in the Scopus and AI conference data sets while columns 2 and 4 give the distribution of papers.

The distribution of computer science papers by address-name groups resembles the distributions for all fields, with a bifurcation between non-China addressed papers and China-addressed papers and with NCD diaspora papers accounting for about 9% of papers. The distribution of AI conference papers is very different, with a small CO share of papers and a huge NCD share that makes the NCD group second in number of papers to those written at non-China addresses without any China presence.

Table 7. Journal and Conference Papers in Computer Science and Conference Papers in Top AI Conferences 2018.

	Computer Science Papers in Scopus		Conference papers of NeurIPS & ICML	
	Number	Share of World	Number	Share of World
CO	63,060	17.6%	59	3.5%
CJN	5,817	1.6%	9	0.5%
CJD	16,163	4.5%	112	6.7%
NCD	33,270	9.3%	485	29.0%
NCN	239,775	67.0%	1,006	60.2%

Note: Papers in computer science field include all journal articles and conference papers in the computer science field. Proportions of CJD and CJN papers are estimated based on 4,000 sampled CJ papers in 2018, and the number of CJD papers = the proportion of CJD in CJ samples times the number of CJ papers in 2018, and the CJN papers are the rest part. Proportion of NCD papers are estimated based on 4,000 NC papers with Chinese last-named author, and the number of NCD papers = the proportion of NCD in NC papers with Chinese last name times the number of NC papers with Chinese last name in 2018.

NeurIPS and ICML calculations are based on all 1,671 Conference papers in NeurIPS and ICML in 2018.

Source: Scopus database.

Using company affiliations listed by authors,[19] we examine next the address-name distribution of the papers associated with US and Chinese high-tech firm at the AI conferences. The first line of Table 8 shows that US firms were associated with about five times as many AI papers at the 2018 conferences as Chinese firms – 377 versus 73. Both US and Chinese companies have researchers on papers in all address-name groups, with US companies having fewer China only addressed papers and

---

of knowledge (Prates, Avelar and Lamb, 2018; Chuvpilo, 2019; Freire, Porcaro and Gómez, 2020; Banerjee and Sheehan, 2020)

[19] We determine company affiliations from Scopus's name list of the top 160 affiliations on the 1,671 conference papers. We select company affiliations from the list by hand, and label papers with at least one company affiliation address as company papers. For companies outside the top 160, we check for terms "Ltd.", "Inc.", "LLC", and "Co." We find that, in 2018, 28.8% of NeruIPS conference papers and 29.4% of ICML conference papers have at least one company affiliation.

more papers with non-China addresses and all non-Chinese named papers. The proportion of NCD papers is higher for Chinese companies than for US companies – 38.4% vs 31.0%, in part because the leading Chinese firms Alibaba, Huawei, Baidu, and Tencent have major research centers in North America which attract many top China-born researchers. Data on the authors of papers shows that 72% of authors affiliated with Chinese companies are diaspora authors[20], with Alibaba being the most extreme, with 100% of the authors who list Alibaba as the affiliated company being diaspora authors. But given the much greater number of papers associated with US than Chinese firms, US firms produces more diaspora AI papers and hire more diaspora researchers producing papers for the NeurIPS and ICML conferences than Chinese firms.

Table 8. Percentage of NeurIPS and ICML Conference Papers Associated with Major US and Chinese Companies by Address and Name of Authors, 2018

	US companies		Chinese companies	
	Number	Share	Number	Share
ALL	377	100%	73	100%
CO	8	2.1%	8	11.0%
CJN	1	0.3%	1	1.4%
CJD	17	4.5%	32	43.8%
NCD	117	31.0%	28	38.4%
NCN	234	62.1%	4	5.5%

Note: More than 96% of the top 30 companies with the most papers in NeurIPS and ICM are either US or Chinese companies. We include companies based on US or China with more than 5 papers.

US companies included in this analysis are *Google, Microsoft, Facebook, IBM, Amazon, Petuum, Intel*, and *Uber*. Google includes DeepMind Technologies Limited in UK; Microsoft includes Microsoft Research Asia in China; Intel includes Intel Labs China.

Chinese companies included in this analysis are *Tencent, Alibaba, Huawei*, and *Baidu*. Tencent includes Tencent AI Lab addressed in US; Alibaba includes Alibaba Group in US; Huawei includes Huawei Noah's Ark Lab in Hong Kong and Huawei Montréal Research Center in Canada; Baidu includes Baidu Research in US.

Source: Scopus database.

In sum, the evidence suggests that diaspora research plays as important a role in computer science as in science more broadly but is a bigger contributor to AI research, where diaspora researchers work for both leading US and Chinese firms.

4. Conclusion

Standard assessments of country contributions to scientific publications credit a paper to a country based on the authors' addresses listed on the paper. Since addresses do not distinguish Chinese born researchers working outside China from other non-China based researchers, the contribution of

[20] A small part of diaspora authors listed both Chinese address non-Chinese address (joint address authors). The average proportion of the joint address authors affiliated with US companies is 0.3%, and the average proportion of the joint address authors affiliated with Chinese companies is 8.6%.

these *diaspora researchers* has been largely ignored. By developing an address-name analysis of papers and authors, we fill in this gap in knowledge and provide detailed statistics on the role of Chinese diaspora researchers in the physical and natural sciences and in computer science and AI. The evidence shows that despite being relatively few in number, diaspora researchers contributed hugely to global science in terms of numbers of articles and the impact/quality of articles. Diaspora researchers have a presence on 13.8% of all articles published in 2018 and can be credited with 4.7% of 2018 (fractional counted) journal articles. Diaspora articles gained far above average citations and publications in top journals, roughly doubling their share of global science in citation weighted articles.

The analysis further finds that diaspora researchers are critical in the co-authorship and citation networks linking China and the rest of the world. Diaspora researchers co-author with researchers in China, cite Chinese-addressed papers and are cited by China-addressed papers far more than other researchers outside of China. By making these connections, Chinese diaspora researchers contribute to China's catch-up in science in ways beyond numbers and citations to diaspora work. Finally, working at the AI frontier for both US and Chinese high-tech firms, diaspora researchers contribute a larger proportion of papers to the leading AI conferences than to science broadly.

Since countries govern borders, the development of a large Chinese diaspora research community required supportive education, research, and migration policies by China and the US and other advanced Western countries where most diaspora scientists work. In contrast to the former Soviet Union which discouraged its scientists from engaging with scientists in other countries and viewed overseas education as a sign of disloyalty, China has sponsored and supported overseas education and scholarly visits once it emerged from the Cultural Revolution. [21] It continued to support international students and research trips even after the 1989 Tiananmen Square Protests, which led many Chinese students and scholars to seek permanent residence and citizenship overseas. On the other side, Western countries also had open doors to Chinese students and researchers. Stephan, Franzoni and Scellato, (2016) suggest that this is due in part to the exceptional performance of the Chinese international students and their contributions to science and productivity of the host country.

The benefits from the exchange of ideas and researchers across geographic areas to the advance of global knowledge documented in this paper deserve attention in assessing policies toward diaspora research even in a period of tensions over trade and other economic developments, Covid-19 pandemic fears and worries by the Trump administration some that some foreign-born researchers come to the US to steal industrial or military secrets. Most come and succeed in creating new knowledge.

---

[21] In 1978, the Minister of Education made a goal to send 3000 more Chinese students to study overseas and succeeded in sending 4252 government-sponsored Chinese students, including 3006 visiting scholars, 537 graduate students, and 649 undergraduate students. In 1981, the state allowed self-supporting oversea education. After China's accession to the WTO in 2001, the number of Chinese students/researchers going abroad boomed. (Chen, 2009; Miao, Wei, Bai, Long and Chen, 2009).

References

Abramo, G., and D'Angelo, C. A. (2015), 'The relationship between the number of authors of a publication, its citations and the impact factor of the publishing journal: Evidence from Italy', *Journal of Informetrics*, *9*(4), 746-761.

Abrishami, A., and Aliakbary, S. (2019), 'Predicting citation counts based on deep neural network learning techniques', *Journal of Informetrics*, *13*(2), 485-499.

Agrawal, A., Kapur, D., McHale, J., and Oettl, A. (2011), 'Brain drain or brain bank? The impact of skilled emigration on poor-country innovation', *Journal of Urban Economics*, *69*(1), 43-55.

Aleksynska, M., and Peri, G. (2014), 'Isolating the network effect of immigrants on trade', *The World Economy*, *37*(3), 434-455.

Behncke, N. (2014), 'The Structure of Ethnic Networks and Exports: Evidence from Germany', *CEGE Discussion Paper*, No. 198. https://ssrn.com/abstract=2412046or http://dx.doi.org/10.2139/ssrn.2412046

Banerjee, I and Sheehan, M. (2020), 'America's Got AI Talent: US' Big Lead in AI Research Is Built on Importing Researchers', Online. https://macropolo.org/americas-got-ai-talent-us-big-lead-in-ai-research-is-built-on-importing-researchers/?rp=e.

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., ... and Uzzi, B. (2010), 'A multi-level systems perspective for the science of team science', *Science Translational Medicine*, *2*(49), 49cm24-49cm24.

Bornmann, L., Leydesdorff, L., and Wang, J. (2014), 'How to improve the prediction based on citation impact percentiles for years shortly after the publication date?', *Journal of Informetrics*, *8*(1), 175-180.

Cao, C. (2008), 'China's brain drain at the high end: why government policies have failed to attract first-rate academics to return', *Asian population studies*, *4*(3), 331-345.

Chen, J., Yin, X., Fu, X., and McKern B. (2020), Beyond Catch-up: Could China Become the Global Innovation Powerhouse? -- China's Innovation Progress, Challenges and Path towards Global Innovation Leadership', *Industrial and Corporate Change*, forthcoming.

Chen, X. (2009), 'Review on China's Policies for Chinese Students Going Abroad over Last Three Decade', *Journal of Xuzhou Normal University*, *35*(4), 1–8 (In Chinese language).

Chuvpilo, G. (2019), 'AI Research Rankings 2019: Insights from NeurIPS and ICML, Leading AI Conferences', Online.

Docquier, F., Lohest, O., and Marfouk, A. (2007), 'Brain drain in developing countries', *The World Bank Economic Review*, *21*(2), 193-218.

Docquier, F., and Rapoport, H. (2012), 'Globalization, brain drain, and development', *Journal of Economic Literature*, *50*(3), 681-730.

Felbermayr, G. J., Jung, B., and Toubal, F. (2010), 'Ethnic networks, information, and international trade: Revisiting the evidence', *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, 41-70.

Felbermayr, G. J., and Toubal, F. (2010), 'Cultural proximity and trade', *European Economic Review*, *54*(2), 279-293.

Freeman, R. B., and Huang, W. (2015), 'Collaborating with people like me: Ethnic coauthorship within the United States', *Journal of Labor Economics*, *33*(S1), S289-S318.

Freire, A., Porcaro, L., and Gómez, E. (2020), 'Measuring diversity of artificial intelligence conferences', *arXiv preprint arXiv:2001.07038*. https://medium.com/@chuvpilo/ai-research-rankings-2019-insights-from-neurips-and-icml-leading-ai-conferences-ee6953152c1a.

Kerr, W. R. (2008), 'Ethnic scientific communities and international technology diffusion', *The Review of Economics and Statistics*, *90*(3), 518-537.
Larivière, V., Kiermer, V., MacCallum, C. J., McNutt, M., Patterson, M., Pulverer, B., Swaminathan, S., Taylor, S. and Curry, S. (2016), 'A simple proposal for the publication of journal citation distributions', *BioRxiv*, 062109.
Lin, J. Y. (2011), 'China and the global economy', *China Economic Journal*, *4*(1), 1-14.
Lissoni, F. (2018), 'International migration and innovation diffusion: an eclectic survey', *Regional Studies*, *52*(5), 702-714.
Miao, D., Wei, Z., Bai, Y., Long, M., and Chen, X. (2009), 'Memorabilia in the 60 Years of China's Oversea Education', *World Education Information, 10*, 35-40. (In Chinese language)
National Science Board. (2020), *Science & Engineering Indicators 2020*, National Science Foundation. https://ncses.nsf.gov/indicators.
Newman, M. E. (2004), 'Coauthorship networks and patterns of scientific collaboration', *Proceedings of the national academy of sciences*, *101* (suppl 1), 5200-5205.
Prates, M., Avelar, P., and Lamb, L. C. (2018), 'On quantifying and understanding the role of ethics in AI research: A historical account of flagship conferences and journals', *arXiv preprint arXiv:1809.08328*.
Saxenian, A., and Hsu, J. (2001), 'The Silicon Valley–Hsinchu Connection: Technical Communities and Industrial Upgrading', *Industrial and Corporate Change*, *10(4)*, 893-920. https://doi.org/10.1093/icc/10.4.893
Saxenian, A. (2002), 'Brain circulation. How high-skill immigration makes everyone better off', *Brookings Review*, *20*(1), 28-31.
Schubert, A., and Glänzel, W. (2006), 'Cross-national preference in co-authorship, references and citations', *Scientometrics*, *69*(2), 409-428.
Stephan, P. E., and Levin, S. G. (2001), 'Exceptional contributions to US science by the foreign-born and foreign-educated', *Population research and Policy review*, *20*(1-2), 59-79.
Stephan, P., Franzoni, C., and Scellato, G. (2016), 'Global competition for scientific talent: evidence from location decisions of PhDs and postdocs in 16 countries', *Industrial and Corporate Change*, *25*(3), 457-485.
Watts, D. J., and Strogatz, S. H. (1998), 'Collective dynamics of 'small-world' networks', *Nature*, *393*(6684), 440-442.
Wang, Y. S., Lee, C. J., West, J. D., Bergstrom, C. T., and Erosheva, E. A. (2019), 'Gender-based homophily in collaborations across a heterogeneous scholarly landscape', *arXiv preprint*, arXiv:1909.01284.
Xie, Q., and R. B. Freeman (2019), 'Bigger Than You Thought: China's Contribution to Scientific Publications and Its Impact on the Global Economy', *China & World Economy*, 27(1), 1–27.
Yan, E., and Ding, Y. (2012), 'Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other', *Journal of the American Society for Information Science and Technology*, *63*(7), 1313-1326.
Ziguras, C., and Gribble, C. (2015), 'Policy responses to address student "brain drain" an assessment of measures intended to reduce the emigration of Singaporean international students', *Journal of studies in International Education*, *19*(3), 246-264.

Appendix A. The data set of sampled papers and the calculations of diaspora papers

There are two ways to use data from Scopus in analysis. The first method is to download a file that contains bibliographic data on of papers from the Scopus online website https://www.scopus.com using the Scopus query string (https://service.elsevier.com/app/answers/detail/a_id/11365/c/10545/supporthub/scopus/). The second is to make requests to the server of Elsevier and get the response content through its API (Application programming interface). Downloading files from the first channel does not provide the first names of researchers that we need to differentiate mainland-born persons from citizens or permanent residences born in other countries that meets our definition of diaspora researchers. It also does not give sufficiently detailed data to determine the position of diaspora researchers in the citation network of papers. It records the number of citations a paper receives but little about the citing papers. It also does not report the address or name of authors of the papers in the reference part of a paper.

To extract evidence on those aspects of papers, we undertook a two-part analysis.

First, we randomly selected samples of 2000 articles from the Scopus English journal articles with valid address or name information that are the focus of our study. The query string in Scopus allows 2,000 papers to be downloaded in any query. It reports up to 100 pages of data for each query, with each page containing from 20 to 200 items. We specify the result page to show 100 items per page. To draw the random samples, we generated 20 random numbers between 1~100 from the random function in Excel and used the numbers to select 20 pages with papers for our sample. The 100 papers in each of the 20 pages gives us a sample of 2,000 papers out of the 10,000 items in the query. The downloaded files contain the author name and address information and other bibliographic information – the title of paper, the publication year, and the ISSN number of the journal etc. But they don't report the first names of authors nor which publication in Scopus cites the selected papers.

Second, using the paper identifier in the downloaded files, we added the desired information to the samples through Elsevier API. We find information on the first names of authors and the papers that cited the paper using the unique identifier assigned to papers in Scopus – EID (see: https://dev.elsevier.com/guides/ScopusSearchViews.htm) and added the first names and the author and address information of the citers of the selected samples via the API portal provided by Elsevier (see: https://dev.elsevier.com/api_docs.html). To get the address and name information of the references in papers in our sample, we accessed the metadata of papers to get the EID code of the references indexed in Scopus through the Elsevier API. We then obtained the detailed address and name information of those cited papers using their EID also through Elsevier API.

The 2000 paper maximum sample that Scopus allowed for an inquiry gives us an adequate number of observations for generalizing to the larger population of all papers. As most of our statistics are counts that we use to compute proportions of papers in different groups, we calculate the sampling error for estimating a proportion in a random sample of 2,000. It is quite small, with a maximum value on the order of 0.006 for a true proportion of 0.50. This allows us to distinguish modest differences in shares of the magnitudes we observe with a high level of significance. As noted in the text, in the case where we had a substantially smaller sample with just 324 persons with Chinese last names in the 2018 NC sample from which to calculate the proportion with Chinese first names, we drew a much larger sample of 2,000 NC papers with at least one Chinese last-named author and obtained virtually identical estimates of the proportion with Chinese first names as in the smaller sample.

Table A-1 lists the data samples that we created. Our focus on diaspora authors meant that we sampled papers with diaspora authors more intensely than papers with all China addresses. The number of 2,000 samples for CJ papers is particularly large because we wanted to track the change over time carefully for a related project. The 2018 sample of NC papers with China last named authors was

our check on the estimated proportion of China named authors who also had Chinese first names.

Table A-1

Data Sample	Purpose	Years Covered	Total number sampled
Papers with only non-China addresses	Obtain data on largest group of papers; find those with China first and last names.	2000, 2015, 2016-2018	2,000 in each year for total of 10,000
Papers with only non-China addresses and China last named authors	Get larger sample to estimate the proportion of NC papers with Chinese last and first named author in NC papers with Chinese last-named author	2018	2,000 in year for total of 2,000
China Joint papers with China and other country addresses	Obtain large time series sample on international collaborations	2000-2018	2,000 in each year for total of 38,000
China Only papers	Obtain data on largest group of CO addressed papers	2000, 2015, and 2018	2000 papers in each year for total of 6,000 papers

Table A-2 records the number of cited and referenced papers we developed from our samples for 2015.

Table A-2

Data Sample	Number of papers	Number of papers which cite the sampled papers published in 2015	Number of referenced papers of sampled papers published in 2018
Papers with only Non-China addresses	2,000	19,415	70,561
China Joint papers with China and other country addresses	2,000	32,324	80,433
China Only papers	2,000	18,160	76,556

Table A-3 describes how we estimate the number of diaspora papers in 2018 and the fractional count of diaspora papers.

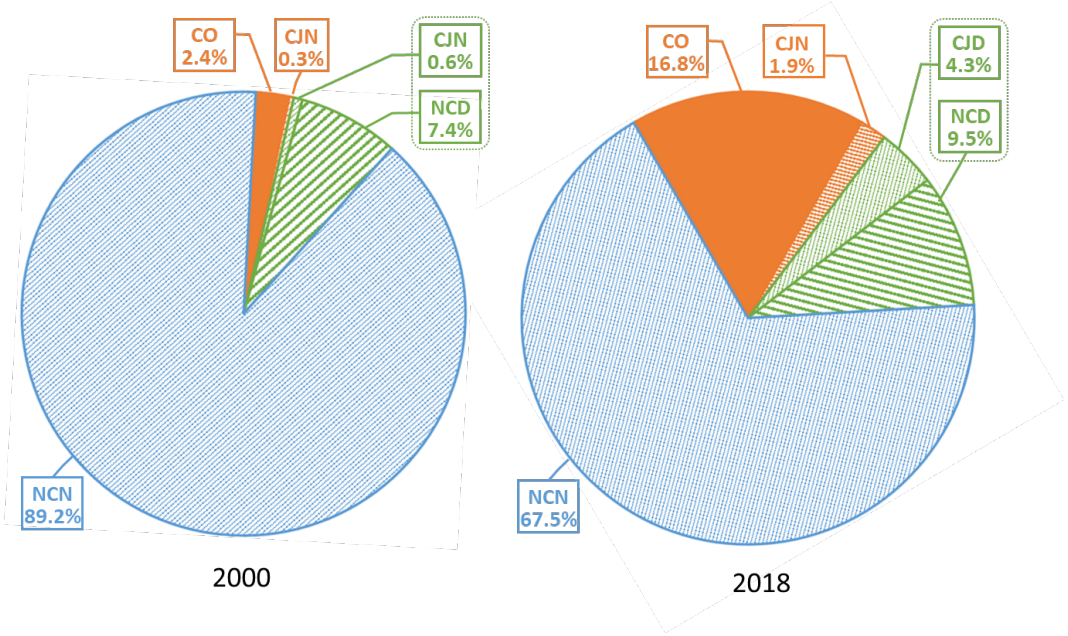
Table A-3.

Definition and Source	Number	Relative to World
All Journal Articles published in 2018	1,602, 030	100%
1. Papers with China Only address (CO)	269,054	16.8%
2. Papers with Only Non-China address (NC)	1,233,660	77.0%
a) NC Papers with at least one Chinese *last-named* author	191,040	11.9%
b) NC Diaspora Papers, estimated from 2,000 NC papers and 2,000 NC papers with at least one Chinese last-named author (NCD)	152,448	9.5%
3. Papers with at least one C and one NC address (CJ)	99,316	6.2%
a) CJ papers with at least one Chinese last name at NC address, estimated from 2,000 CJ papers	83,908	5.2%
b) CJ Diaspora papers (CJD), based on % papers with at least one Chinese first & last-named authors at NC address in 2,000 CJ sample	68,719	4.3%
4. Papers with Chinese names and Non-China Addresses		
a) NC Papers with at least one Chinese *last*-named author, 2a+ 3a	274,948	17.2%
b) NC papers with at least Chinese *first and last*-named author, 2b +3b	220,974	13.8%
5. Diaspora Papers Fractional Counts by Chinese Diaspora Proportion of Authors		
a) Fractional Count NC Diaspora Papers, based on 37.5% share of China names on papers from 2,000 NC sample x line 2b	57,093	3.6%
b) Fractional Count CJD papers based on 27.6% estimated Chinese names on NC address from 2,000 CJ sample x line 3b	18,951	1.2%
c) Fractional Count of all Diaspora Papers (5a + 5b)	76,044	4.7%

Note: China number of papers fractionated by giving China a proportion of each CJ paper dependent on % of authors with China address, with China credited for authors with a C and one or more NC addresses, proportion to China's share of addresses.

All of the codes and the computer prints for the analysis on request from the authors.

Appendix Figure B: China's Presence in Global Scientific Publications, 2000 and 2018



Notes: The 2000 Figure, the measure of CO is accurate number from Scopus data base, the measure of CJD and CJN are estimated based on a sample of 2,000 CJ papers published in 2000, and the measure of NCD and NCN are estimated based on a sample of 2,000 NC papers published in 2000.

Source: English journal articles in Scopus which are published in 2000 and 2018, as described in Appendix Table A-1.

Appendix Table C: Regression Estimates and Standard Errors Relating 3 Year Forward Citations and Cite Scores of 2015 Papers to Groups of Paper Authors, with Field Variables and Number of Authors

Dependent Variable/Group	Citations	Citations	Cite Score	Cite Score
NCD (Diaspora Papers in NC addressed group	10.72 (0.000)	9.44 (0.000)	1.92 (0.000)	1.42 (0.000)
CJD (Diaspora Papers in CJ group)	10.19 (0.000)	8.55 (0.000)	1.84 (0.000)	1.58 (0.000)
CJN (Papers without Diaspora authors in CJ)	4.15 (0.001)	3.88 (0.004)	0.85 (0.000)	0.94 (0.000)
CO (China Only papers)	1.16 (0.137)	1.24 (0.144)	-0.08 (0.414)	-0.15 (0.131)
NCN (Papers with no China address and no diaspora authors)	-	-	-	-
Other Factors				
21 Field	no	yes	no	yes
#Authors	-	0.27 (0.000)	-	0.03 (0.000)
Adj R-squared	0.0333	0.0634	0.0787	0.2293
NOB	5318	5318	5318	5318

Note: NCD is the dummy variable of NCD papers; CJD is the dummy variable of CJD papers; CJN is the dummy variable of CJN papers; CO is the dummy variable of CO papers; NCN is the dummy variable of NC_N papers and also is our benchmark. Cite Score value is assigned to a paper based on the 2017 cite score value of the journal it published on. The 21 fields are: Multidisciplinary; Agricultural and Biological Sciences; Biochemistry, Genetics and Molecular Biology; Chemical Engineering; Chemistry; Computer Science; Earth and Planetary Sciences; Energy; Engineering; Environmental Science; Immunology and Microbiology; Materials Science; Mathematics; Medicine; Neuroscience; Nursing; Pharmacology, Toxicology and Pharmaceutics; Physics and Astronomy; Veterinary; Dentistry; Health Professions.

Source: Tabulated from a sample of 2,000 CO papers, a sample of 2,000 CJ papers, and a sample of 2,000 NC papers published in 2015. Observations without valid address or name information are omitted, papers are also omitted if the journals they published on haven't been assigned a 2017 version of cite scores by Scopus, mainly because those journals are newly established. The number of observations for each group are NCD: 364; CJD: 1269; CJN: 401; CO: 1838; NCN: 1446.