IEP@BU Working Paper Series

The Laboratory of The World?

Territorial vs Headquarter R&D in the European Union, China and the United States

Daniel Gros, Mauro Molteni, Giorgio Presidente

Università Bocconi
IEP@BU
Institute for European Policymaking

January 2026

The *Laboratory of The World*?

Territorial *vs* Headquarter R&D in the European Union, China and the United States

Daniel Gros[1], Mauro Molteni[1], and Giorgio Presidente[*†1]

[1] *Institute for European Policymaking, Bocconi University*

January 13, 2026

Abstract

This paper examines the global allocation of business research and development (R&D) by multinational firms, distinguishing between where R&D is financed and where it is performed. We compare innovation ecosystems in the European Union, the United States and China, and show that the United States is consistently more successful than the EU in attracting and retaining R&D activity within its territory, while China emerges as the largest net recipient of foreign R&D worldwide. China's R&D ratio has fallen rapidly over the last decade and is converging to the EU level as large Chinese firms go increasingly abroad. However, the ratio remains exceptionally high in the manufacturing sector, suggesting that China is the world's laboratory for manufacturing. At the same time, in the ICT sector China lags behind the other two regions, while the EU ratio convergences to that of the United States since the mid-2010s.

[*] Corresponding author: giorgio.presidente@unibocconi.it

[†] The views expressed in this paper are of the authors and do not necessarily reflect those of the affiliating institutions.

1 Introduction

China has long been described as the *factory of the world*—a place where global production networks concentrate their manufacturing capacity. Yet if production has been globalized and spatially unbundled, the same question can be asked of innovation: where is the world's laboratory?

This paper explores the geography of research and development (R&D) in a world where technological creation, much like production, increasingly transcends national borders. In particular, it examines how much R&D activity performed within a country actually belongs to foreign firms headquartered elsewhere, and conversely, how much of domestic firms' R&D takes place abroad.

We focus on a comparison between the European Union (EU), China, and the United States (US)—the latter being often considered at the top of the technological frontier.

The ratio of territorial-to-headquartered R&D expenditures is our primary metric. This ratio, though a proxy rather than a direct measure of cross-border R&D flows, provides a tractable and comparable framework for analyzing how multinational firms distribute their innovation activities across borders. While the measure can be sensitive to R&D reporting practices and data limitation, it offers useful insights into the extent to which firms conduct R&D outside their home territories.

Specifically, a high ratio signals that the domestic economy is a relatively attractive location to conduct R&D—being that due to equipment and skills availability, or a favourable regulatory environment.

On a normative side, a high R&D ratio can be interpreted as desirable because employment gains and the technology and knowledge externalities generated by R&D largely accrue locally. In this sense, a high R&D ratio indicates a net positive return from hosting R&D, as the local benefits from knowledge creation, talent formation, and ecosystem deepening can exceed the value that would arise solely from owning R&D conducted elsewhere.

Based on our metric, we find that on average, the US is a more attractive location for R&D than the EU. However this is not the case for China, which economy-wide has the highest R&D ratio.

However, we show that the primacy of China is entirely driven by the manufacturing sector. When the latter is left out, the US has the highest ratio, at least until 2019, when the US and EU ratio converge due to a steady increase of the EU ratio in Information and Communication.

2 Data and Empirical Approach

Our analysis combines two data sources. For headquarter R&D, we use the EU Industrial R&D Investment Scoreboard, which reports R&D spending by headquartered companies regardless of where the R&D is performed.

2.1 Headquarter R&D

The most recent edition of the Scoreboard (2024) includes data for the top global 2000 corporate R&D spenders, as well as the top 800 EU-based companies.[1] The top 2000 industrial R&D investors represents 85% of global business sector R&D (BERD), while the 800 European ones for about the 17%. The US companies among the top 2000 ones account for about the 36% of global BERD, while the Chinese ones for about the 14.6%.

The distribution of R&D expenditure across firms is highly concentrated. A log-log representation of R&D investment rank, versus the natural logarithm of R&D expenditure confirms a robust Pareto distribution (Appendix Figure A1). Therefore, the Scoreboard provides a reliable basis for international comparisons.

2.2 Territorial Business R&D

For territorial R&D, we use the OECD's Business Enterprise R&D (BERD) data, which reports R&D activity occurring in a jurisdiction, whether by domestic or foreign-owned firms.

The OECD's Business Enterprise R&D (BERD) statistics draw on information submitted to the OECD and Eurostat as part of their joint international data collection on R&D resources.[2] BERD captures intramural R&D expenditures performed by firms within national borders following a territorial principle.[3] These data are compiled by national statistical authorities through dedicated surveys and administrative sources, following the methodological guidelines of the OECD Frascati Manual, which is the internationally recognized standard in this area.

OECD BERD statistics are available by 2-digit NACE economic activity, which is the same classficiation used by the Scoreboard, which allows us to calculate our metric for specific industries. Further details on the sample's coverage by industry are provided in Appendix A, while the details on data construction are presented in Appendix B.

[1] https://iri.jrc.ec.europa.eu/scoreboard/2024-eu-industrial-rd-investment-scoreboard.

[2] https://www.oecd.org/en/data/datasets/analytical-business-enterprise-research-and-development.html.

[3] For instance, if a U.S. multinational funds research at a German subsidiary, the expenditure is recorded as German BERD.

Our analysis focuses on BERD, which in 2020 accounted for 54% percent of all R&D expenditures performed within a country's territory in EU, 76% US and 47% for China.[4]

2.3 Territorial-to-Headquartered R&D Ratio

To examine the geographical allocation of R&D activities, we construct a measure that relates firms' R&D expenditures as reported by their country of headquarters to the R&D expenditures observed within national territories. The territorial-to-headquartered R&D ratio (T-H ratio) provides a tractable indicator of the balance between domestically headquartered and territorially executed R&D activity.[5]

For each country c and year t, the T-H ratio is expressed as:

$$\text{T-H Ratio}_{c,t} = \frac{\text{Territorial R\&D Expenditure}_{c,t}}{\text{Headquartered R\&D Expenditure}_{c,t}}$$

A ratio lower than 1 suggests that firms headquartered in a country perform less R&D abroad than foreign firms perform within that country. Conversely, a ratio higher than 1 indicates that the country is a net recipient of R&D from overseas headquartered companies.

3 Economy-Wide R&D Ratio

Panel (a) of Figure 1 presents the ratio of territorial-to-headquartered R&D (henceforth, R&D ratio) for the EU, United States and China. The EU consistently maintains a lower R&D ratio than the US. This suggests that EU-based firms have a greater propensity to conduct R&D activities outside their home territories, whereas the US innovation ecosystem appears more effective at attracting and retaining these activities domestically.

The unexpected finding is China. The R&D ratio is consistently above the US, although converging over the years of the sample.[6] In 2008—the first year of the sample, the R&D ratio is 1.1 for the EU, 1.4 for the US, and 5.43 for China. The R&D ratio is broadly stable for the first two regions, but it steadily falls for China. In 2021—the last year of the sample, the R&D ratio is 1.03 for the EU, 1.24 for the US, and 1.49 for China.

In Panel (b), we decompose changes in the Chinese R&D ratio between movement in the numerator and denominator. The figure shows that the denominator—Chinese territorial R&D—

[4] In accounting terms, Gross Domestic Expenditure on R&D (GERD) = BERD + GOVERD (by government institutions) + HERD (by universities and other higher education institutions) + PNPRD (by private non-profit organizations).

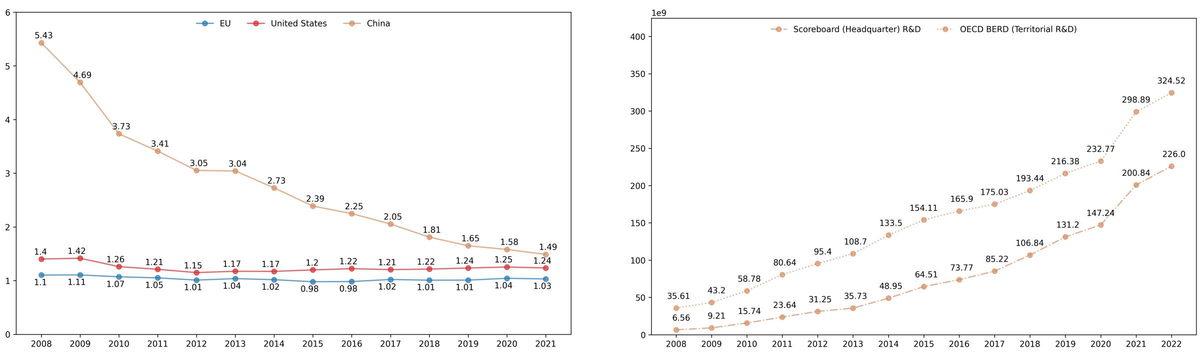
[5] A similar approach based on patents has been proposed by Gkotsis, Hernandez, and Vezzani (Gkotsis et al.).

[6] Previous evidence pointed out that the share of global corporate R&D spent by Chinese firms matches the share of the European Union: https://www.economist.com/business/2024/07/18/china-is-the-wests-corporate-randd-lab-can-it-remain-so.

grows substantially in the last years of the sample. Thus, we conclude that the Chinese R&D ratio in Panel (a) is below 1 not just because headquarter R&D is low, but because there is considerably R&D from abroad made in China, led by companies such as Siemens, Apple and Tesla.[7]

This suggests that China is a net recipients of R&D. To our metric, China is *The Laboratory of The World*.

Figure 1: EU vs USA vs China: Ratio of Territorial to Headquartered R&D Spending



(a) (b)

Source: Authors' calculations based on EU Industrial R&D Scoreboard (2024) and OECD Research and Development Statistics.

4 The Key Role of Manufacturing

A key finding of this policy brief is that the high economy-wide R&D ratio of China is entirely driven by the manufacturing sector.

Figure 2 presents R&D ratios for the manufacturing (Panel (a)) and non-manufacturing (Panel (b)) sectors. Panel (a) shows that the EU ratio is consistently below 1. The US ratio is instead well above 1, and the disparity is even more pronounced in specific manufacturing industries for which sectoral information is available, such as pharmaceuticals and electronics (Appendix Figure A5).

The key finding in Panel (b) of Figure 2 is that for non-manufacturing sectors, China's R&D ratio is well above 1, and the highest across the regions. This suggests that in high-tech sectors such as Information and Communication, companies headquartered in China like Alibaba, Tencent, and ByteDance—now conduct substantial R&D abroad, and that foreign companies in intangible-intensive industries do not find China particularly appealing to conduct

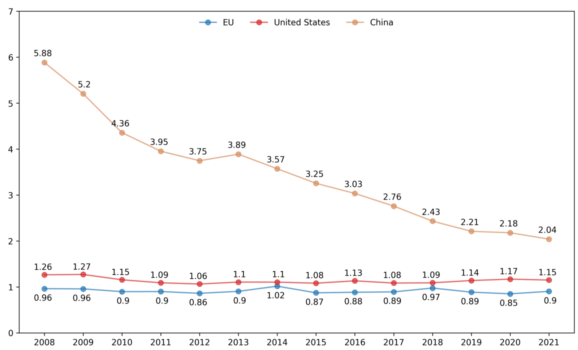
[7] See https://www.uscc.gov/sites/default/files/2025-11/Chapter_6--Interlocking_Innovation_Flywheels_Chinas_Manufacturing_and_Innovation_Engine.pdf.

R&D.

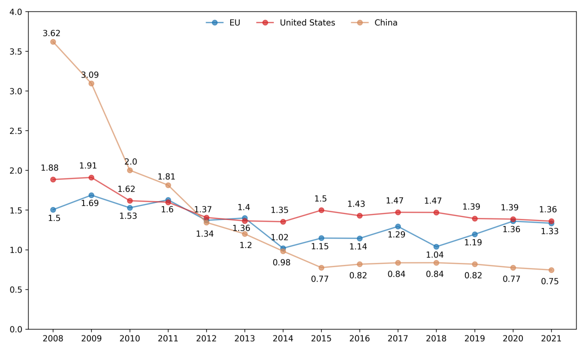
Another finding in Panel (b) is that between 2014 and 2018, the EU had a R&D ratio substantially lower than the US, but from 2019 the EU ratio increases and converges to the US level in 2020.

A key non-manufacturing sectors responsible for the reduction of the gap of R&D ratio between the EU and the US is Information and Communication. This is shown in Panel (c) of Figure 2, where the EU ratio increases above the US level after 2015. The finding is somewhat surprising and deserves further research.

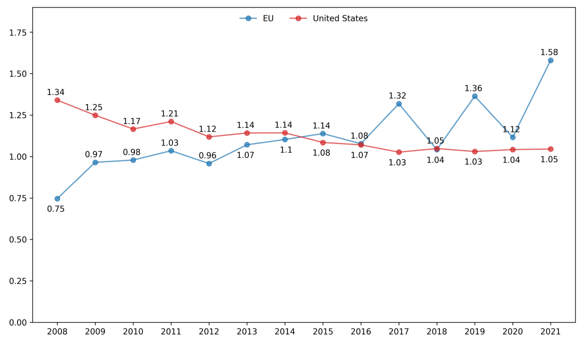
Figure 2: EU vs USA vs China: Ratio of Territorial-to-Headquartered R&D Expenses



(a) Manufacturing



(b) Non-Manufacturing



(c) Information and Communication

Source: Authors' calculations based on EU Industrial R&D Scoreboard (2024) and OECD Research and Development Statistics.

5 Robustness

This section presents a battery of robustness checks.

First, we test the robustness of our main finding that China has an exceptionally high R&D ratio. Specifically, we have two main concerns.

The first concern is that China's high ratio might simply reflect the low representativeness of Chinese firms in the Scoreboard. The Scoreboard featured only 216 firms in 2008, against 768

EU and 1001 US firms. However, Appendix Figure A2 shows that the number of Chinese firms in the Scoreboard increases substantially in the first years of the sample and stabilizes around the mid of the sample and then converges to values of other regions. In 2021, the Scoreboard includes 950 US, 923 EU and 800 Chinese companies. Therefore, while the low early coverage might have explained the low initial values of the Chinese R&D ratio (and some of its initial dynamic), this is not the case in the final part of the sample.

The second concern is that the EU Industrial R&D Scoreboard may systematically under-represent Chinese state-owned enterprises, biasing the headquarter measure downward. To examine the issue, we examine official R&D statistics from the China Statistical Yearbook on Science and Technology.[8] This survey covers mining, manufacturing, and utilities, representing approximately 81% of China's total business R&D in 2021. The advantage of the Chinese official statistics enable direct identification of state-owned enterprise R&D, which amounts to about 9.7 billion euro.[9]

Appendix Figure A3 presents the adjusted China headquarter-to-territorial ratio incorporating the Chinese official state enterprise data. Even adding state R&D to China's Scoreboard denominator yields adjusted ratios of 2.3–1.47 over the years 2012–2021, well above the United States.

Another message from Appendix Figure A3 is that while SoEs have had a substantial weight relative to BERD until 2018, its contribution to total R&D has become negligible thereafter.

Next, we examine whether missing information about Amazon's R&D in the Scoreboard is responsible for an artificially high R&D ratio. Specifically, Amazon follows the accounting practice to merge together traditional R&D activities with operational technology maintenance (store development, design, and maintenance), content operations (curation and display activities), and infrastructure costs (servers, networking equipment, data center depreciation). For this reason, its R&D spending is not included in the Scoreboard. Our strategy is extracting the aggregate figure year by year from SEC filings (Form 10-K) and add it to the denominator of the US R&D ratio. The results are presented in Appendix Figure A4. The figure shows that even considering the adjusted figure, which is clearly inflated (Appendix Table A1), the US ratio is still higher than EU's one.

[8] The Yearbook reports R&D expenditure by enterprise registration type for all industrial firms with annual revenue $\geq$ 20 million RMB in 2011–2017, $\geq$ 15 million RMB thereafter.

[9] Specifically, we aggregate R&D from all state-affiliated categories in Appendix Tables A2 and A3 (Explicitly state-owned enterprises; Collective-owned enterprises (local government ownership); State joint ownership corporations, and State sole-proprietorship limited liability corporations. This yields total state-affiliated R&D of 73.91 billion RMB in 2021, or about 9.7 billion euro.

6 Conclusions

This paper asks Where is the world's laboratory? By comparing territorial and headquarter-based measures of R&D, we document systematic differences in how innovation is geographically organized across the European Union, the United States, and China. Our results show that the EU consistently exhibits a lower territorial-to-headquartered R&D ratio than the United States, indicating a stronger outward orientation of R&D by EU-based firms and a relative difficulty in retaining R&D activity domestically. The US, by contrast, appears structurally more attractive as a location for business R&D.

China stands out as a special case. At the aggregate level, China displays the highest R&D ratio in the sample, making it a net recipient of foreign R&D and, according to our metric, "the laboratory of the world." Importantly, this result is not driven by a lack of Chinese firms in headquarter-based data, nor by the omission of state-owned enterprises. Robustness checks incorporating official Chinese R&D statistics confirm that China's territorial R&D substantially exceeds the R&D financed by firms headquartered there, even after accounting for state-owned enterprise activity

Academic work supports shows that foreign-based firms are committing R&D resources abroad, with China being a prominent destination for these investments due to its dynamic technological ecosystem and expanding pool of researchers and engineers (Allen et al., 2024). Recent reporting and official sources confirm this trend: foreign multinationals are actively expanding research centers in China, driven by market growth and strategic positioning. According to a 2025 overview, many multinational corporations have been "expanding their research and development spending and accelerating the establishment of R&D centers" in China, signifying a clear uptick in foreign R&D activity.[10] In Shanghai alone, there were nearly 575 foreign-funded R&D centers tied to multinational firms as of mid-2024, underscoring the scale of this phenomenon.[11] These facilities often serve dual roles: developing products for local markets and contributing to global innovation pipelines, while cooperating with local partners, universities, and research institutes.[12] High-profile corporate commitments illustrate this trend at the firm level. For example, Porsche announced plans to open a new R&D centre in Shanghai to develop China-specific vehicle technologies, demonstrating how global automakers are aligning technical innovation with local consumer preferences.[13] Likewise, Nvidia has signaled plans for a Shanghai research centre focused on advancing its AI and chip technologies in the Chinese context, despite

[10] https://en.qstheory.cn/2025-10/31/c_1136797.htm.
[11] https://www.stdaily.com/web/English/2024-08/22/content_216463.html.
[12] https://www.chinadailyhk.com/hk/article/605957.
[13] https://www.reuters.com/business/retail-consumer/porsche-set-up-rd-centre-shanghai-local-government-says-2

broader geopolitical challenges.[14] Furthermore, pharmaceutical leader AstraZeneca committed US$2.5 billion to expand R&D and manufacturing operations in Beijing, reinforcing China's role in its global research footprint.[15]

Thus, China has evolved not just as a manufacturing hub but increasingly as a strategic location for international corporate research and development. However, a key contribution of this paper is to show that China's primacy is entirely sector-specific. The exceptionally high R&D ratio is driven almost exclusively by manufacturing. Once manufacturing is excluded, China no longer appears as an attractive destination for foreign R&D. In non-manufacturing and intangible-intensive sectors—most notably information and communication—foreign firms conduct relatively little R&D in China, while Chinese firms increasingly perform R&D abroad. In these sectors, the United States retains its position as the most attractive R&D location, with the EU gradually converging toward U.S. levels after 2015

---

[14] https://www.ft.com/content/c886a4c0-da75-4ea7-8230-6ffd18815fa4.

[15] https://www.theguardian.com/business/2025/mar/21/astrazeneca-to-invest-in-drugs-research-and-manfacturing-

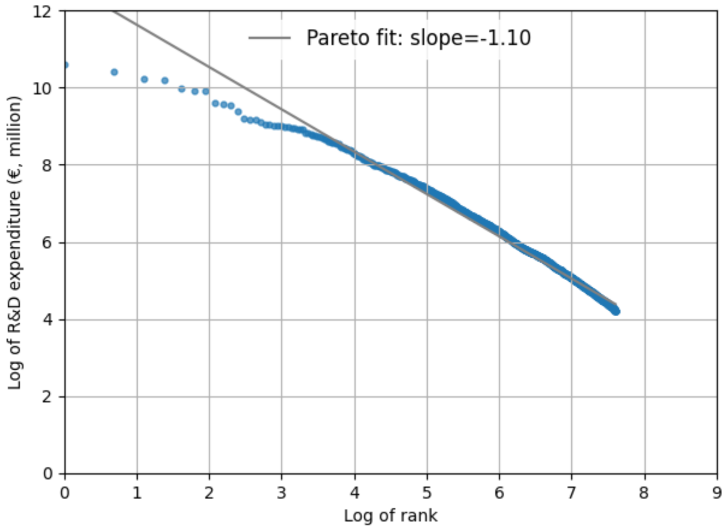
References

Allen, M. M., M. Demirbag, M. L. Allen, S. Bhankaraully, and G. Wood (2024). Multinational enterprises' r&d commitments in chinese provinces: A configurational approach. *Journal of International Management 30*(4), 101158.

Gkotsis, P., H. Hernandez, and A. Vezzani. Estimating territorial business r&d expenditures using corporate r&d and patent data. JRC Research Reports JRC109031, Joint Research Centre.

A Figures and Tables Appendix

Figure A1: Scorebard 2023 Top 2000 World Companies: Pareto Analysis



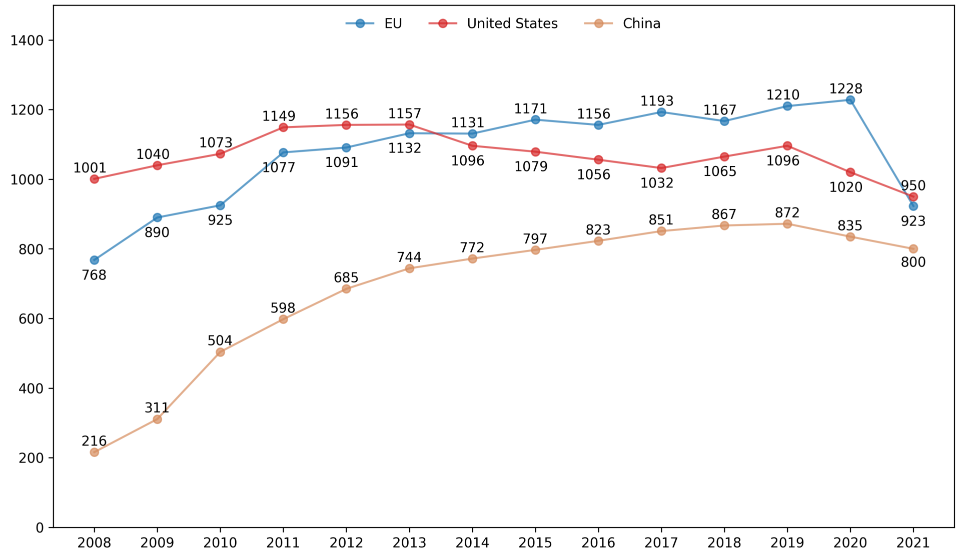
Source: Authors' calculations based on EU Industrial R&D Scoreboard (2024).

Table A1: R&D-to-CapEx Intensity Ratios: Apple, Meta, and Alphabet

Year	Mean Ratio	Median Ratio	Ratio of Amazon on 10-k Data
2009	7.50	3.51	3.32
2010	1.27	0.94	1.77
2011	0.90	0.64	1.61
2012	1.18	1.13	1.21
2013	0.89	1.04	1.91
2014	0.99	0.90	1.90
2015	1.28	1.21	2.73
2016	1.14	1.32	2.39
2017	1.10	1.15	1.89
2018	0.88	0.83	2.15
2019	1.18	1.10	2.13
2020	1.67	1.24	1.06
2021	1.53	1.33	0.92
2022	1.59	1.25	1.15
2023	1.81	1.35	1.62
Average 2009-2023	**1.78**	**1.26**	**1.85**
Average 2010-2023	**1.26**	**1.11**	**1.75**

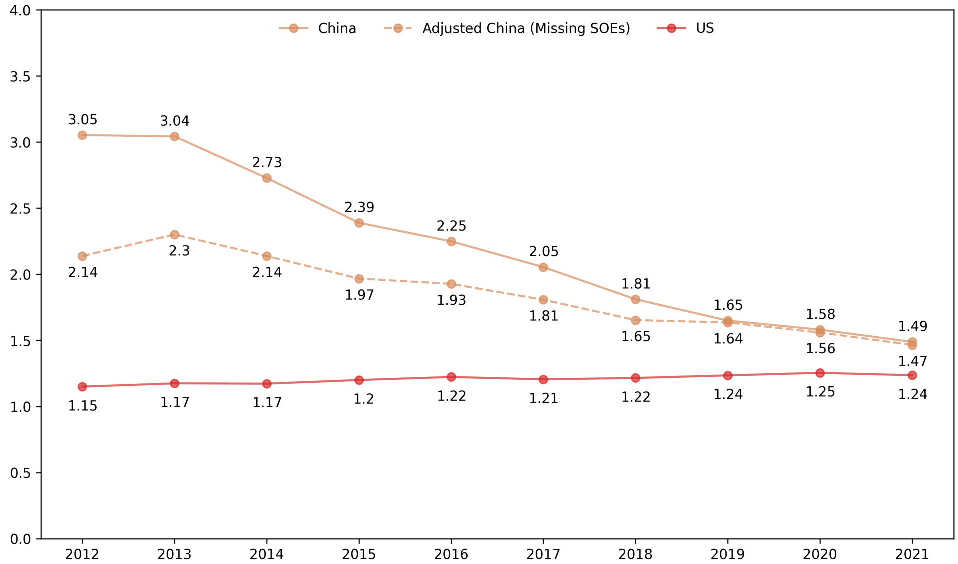
Note: R&D-to-CapEx ratios calculated from EU Industrial R&D Investment Scoreboard data for Apple, Meta (Facebook), and Alphabet (Google). The peer group follows Scott Galloway's "Gang of Four" technology firms, excluding Amazon. Mean and median ratios provide alternative estimates for Amazon's R&D intensity adjustment.

Figure A2: EU vs USA vs China: Number of companies in the EU Industrial R&D Scoreboard



Source: Authors' calculations based on EU Industrial R&D Scoreboard (2024) and OECD Research and Development Statistics.

Figure A3: USA vs China: Territorial-to-Headquarter Ratios with Chinese Official Statistics



Source: OECD Research and Development Statistics; China Statistical Yearbook on Science and Technology. Adjusted China ratio incorporates state-affiliated enterprise R&D from Chinese official statistics (Tables A2–A3).

Table A2: R&D Expenditure by Registration Status (2012-2016)

Status of Registration	2012	2013	2014	2015	2016
Total	**720.06**	**831.84**	**925.43**	**1,001.39**	**1,094.47**
Large & Medium-sized	599.23	674.41	731.97	779.24	828.95
Domestic Invested Enterprises	**543.70**	**630.33**	**710.35**	**771.24**	**852.54**
State-owned	56.21	30.84	32.57	32.24	28.39
Collective-owned	7.52	5.87	6.22	8.14	6.11
Cooperative	2.39	0.88	0.74	0.79	0.63
Joint Ownership	1.14	0.70	0.49	0.07	0.08
State Joint Ownership	0.97	0.68	0.46	0.02	0.04
Limited Liability Corp.	222.47	283.09	315.91	338.89	375.49
State Sole-prop.	45.21	58.83	62.48	57.35	55.51
Share-holding Corp. Ltd.	124.56	138.55	150.46	153.47	161.28
Private	124.65	169.01	202.68	236.36	280.05
Other	4.76	1.38	1.28	1.29	0.50
Enterprises with Investment from HK/Macao/Taiwan	**67.24**	**77.22**	**85.23**	**94.77**	**101.36**
Joint-venture	28.36	33.22	36.37	38.23	38.24
Cooperative	0.77	0.93	1.31	1.26	1.22
Sole-proprietorship	32.94	38.21	42.47	47.52	54.31
Share-holding Corp. Ltd.	4.74	4.83	5.01	7.51	6.52
Foreign Invested Enterprises	**109.13**	**124.29**	**129.85**	**135.39**	**140.57**
Joint-venture	57.65	67.77	73.43	74.68	75.50
Cooperative	1.57	2.29	2.08	1.70	1.62
Sole-proprietorship	43.60	46.64	47.28	51.30	56.31
Share-holding Corp. Ltd.	6.02	7.40	6.82	7.33	6.77

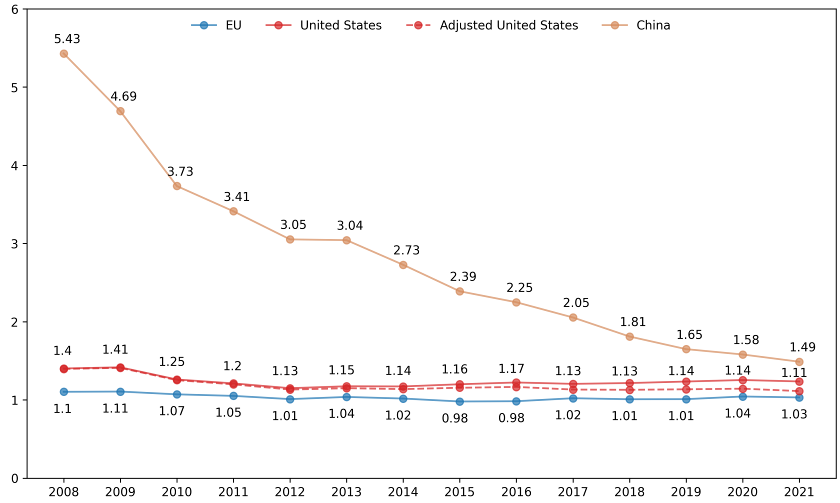
Note: R&D expenditure in billions of yuan (original data in 10,000 yuan units).
Source: China Statistical Yearbook on Science and Technology, Table 20-5 (2012-2016).

Table A3: R&D Expenditure by Registration Status (2017-2021)

Status of Registration	2017	2018	2019	2020	2021
Total	**1,201.30**	**1,295.48**	**1,397.11**	**1,527.13**	**1,751.42**
Large & Medium-sized	897.62	954.27	999.69	1,077.23	1,228.18
Domestic Invested Enterprise	**942.30**	**1,027.20**	**1,121.90**	**1,227.27**	**1,413.68**
State-owned	21.34	8.34	8.32	15.73	21.85
Collective-owned	6.14	0.78	0.57	0.68	0.78
Cooperative	0.60	0.70	0.63	0.84	0.92
Joint Ownership	0.08	0.03	0.04	0.28	0.31
State Joint Ownership	0.04	0.01	0.01	0.04	0.14
Limited Liability Corp.	410.21	427.93	444.98	426.25	485.67
State Sole-prop.	63.42	71.07	62.33	51.59	51.20
Share-holding Corp. Ltd.	184.72	202.52	213.91	216.93	223.81
Private	318.81	385.16	451.67	564.70	678.20
Other	0.40	1.74	1.76	1.87	2.14
Enterprises with Investment from HK/Macao/Taiwan	**111.51**	**113.08**	**113.84**	**125.62**	**144.81**
Joint-venture	44.13	41.18	41.82	44.59	47.05
Cooperative	1.36	1.12	1.39	1.40	1.40
Sole-proprietorship	57.15	60.91	63.47	68.19	69.57
Share-holding Corp. Ltd.	7.67	9.01	6.59	10.84	25.51
Foreign Invested Enterprises	**147.49**	**155.20**	**161.38**	**174.24**	**192.93**
Joint-venture	78.26	82.21	84.47	87.16	89.90
Cooperative	1.63	1.65	1.78	1.46	1.53
Sole-proprietorship	57.09	61.78	66.53	74.74	82.48
Share-holding Corp. Ltd.	9.57	8.42	7.10	9.41	17.96

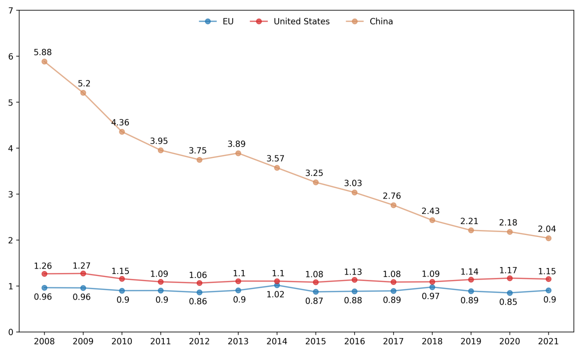
Note: R&D expenditure in billions of yuan (original data in 10,000 yuan units).
Source: China Statistical Yearbook on Science and Technology, Table 20-5 (2017-2021).

Figure A4: Ratio of Territorial-to-Headquarter R&D with USA figures adjusted to include Amazon inflated R&D spending.

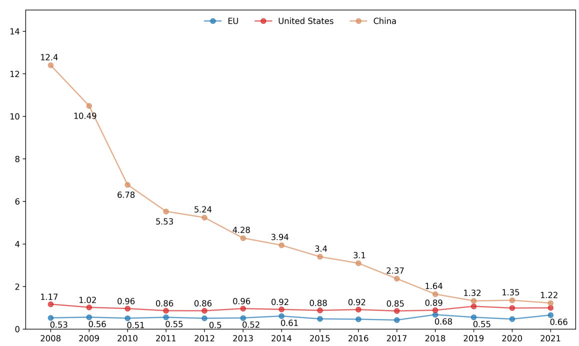


Source: Authors' calculations based on EU Industrial R&D Scoreboard (2024) and Amazon's 10-K filings.

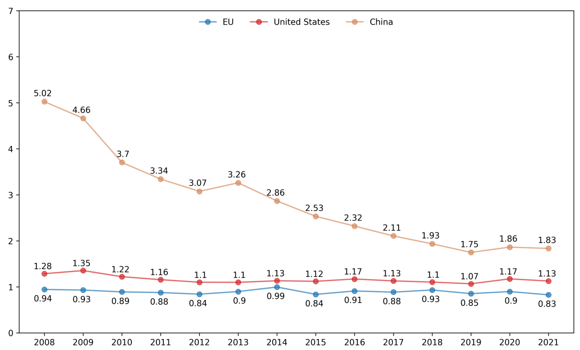
Figure A5: EU vs USA vs China: Ratio of Territorial to Headquartered R&D Expenses in detailed manufacturing sectors



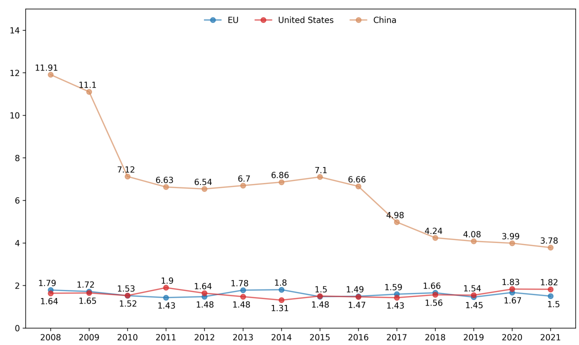
(a) C



(b) C21



(c) C25T30



(d) C28

Source: Authors' calculations based on EU Industrial R&D Scoreboard (2024) and OECD Research and Development Statistics.

B Data Appendix

The initial Industrial Scoreboard dataset comprised 94,214 rows, requiring comprehensive cleaning and standardization to ensure data quality and consistency.

NACE codes were standardized to a consistent 4-digit format by attaching leading zeros where necessary. This ensured uniformity across the dataset and prevented potential matching issues in subsequent analyses. For instance, a code like '123' was transformed to '0123' to maintain a consistent code length.

A crucial step involved handling missing R&D values whose rows were removed, reducing the dataset size. Analysis revealed that the number of companies with missing R&D values varied significantly by year and the dataset exhibited a notable trend of decreasing missing entries over time. In 2003, 1,563 entries lacked R&D values, while by 2022, this number had dramatically reduced to just 54 missing entries. Notably, the 2023 dataset contained no missing R&D values.

Descriptive analysis of missing values revealed:

- Mean of missing entries: 921.45
- Median of missing entries: 1,072
- Standard deviation: 504.67
- Minimum: 54
- Maximum: 1,563

Rows with completely missing R&D values were removed to maintain data integrity.

The sector classification underwent rigorous normalization. Inconsistencies in the Industry Classification Benchmark (ICB) classification were resolved and names were standardized to eliminate variations (e.g., "Automobiles & Parts" vs "
tAutomobiles & Parts" were consolidated to a single, consistent format). Missing sector classifications were inferred by leveraging other observations from the same company across different years.

To address missing sector and country classifications we used a double approach. For companies with partial historical data, missing values were inferred from existing observations of the same company. Additionally, when NACE codes were missing and could not be inferred from other years, The ORBIS database was leveraged to fill these gaps. To validate this approach, Cohen's kappa statistic was computed, indicating a strong agreement:

- NACE Rev2 New vs. ORBIS: 0.756 agreement

- NACE Rev2 vs. ORBIS: 0.676 agreement

These high agreement scores provided confidence in using ORBIS data to fill missing classifications.

Finally, two additional columns were introduced to facilitate future data matching with the OECD R&D dataset:

- NACE sector column, including the sector letter prefix / which appends the sector letter to the four-digit NACE code
- An ISO 3166-1 alpha-3 country code column that complements the existing two-letter country codes and facilitate international comparisons

Since the values are reported in Millions, we multiplied the values to 1M to convert the data.

The cleaned dataset provides a robust foundation for subsequent analysis, with enhanced consistency and comparability across variables.

At the end of the data cleaning stage, the Scoreboard dataset was reduced from 94,214 to 75,780 rows, significantly improving data quality while maintaing the most informative and complete entries.

The merge between the Scoreboard and the OECD datasets is on the information (country, year, economic activity), where country is represented as an ISO 3166-1 alpha-3 code, and the economic activity as a NACE sector code.

In cases when the OECD datasets has both single and grouped NACE codes, e.g. 'C12' and 'C10T12' (meaning C10, C11 and C12) and 'C10_11' (meaning C10 and C11). Before merging the two datasets, we "unpack" the grouped nace to make the match the sectors in the Scoreboard dataset.

The merge then is done in a recursive way: since the NACE sector codes of the Scoreboard dataset are at usually at a finer level than the ones in the OECD dataset, we iteratively scale up the code. For example, the NACE code "A0111" in the Scoreboard doesn't have an exact match in the EOCD dataset but matches against the "A" code or the code "C1395" matches "C13". At the same time, some code matches just grouped codes ("C1091" matches "C10T12") while others match multiples ("C1089" matches "C10T12", "C10_11" and "C")

Once the datasets are merged we end up with a new dataset whose rows represent the R&D expenditure by a company in a specific year and a specific sector together with the total R&D expenditure in that specific country, for that specific year and specific activity.

The business R&D expenditure is represented as euro at current year (year of reporting) while the total expenditure in national currency at current year. To convert the total expenditure to

Euro, we created a map between currency and year to exchange rate using the exchange rate that we get from the Scoreboard dataset (for each year and country the R&D expenses are reported both in Euro and in local currency, so it's easy to get the exchange rate) and we add include data from the IMF for the missing combinations (currencies for Argentina, Iceland and Mexico).