Department for
Science, Innovation
& Technology

National Vision for Engineering Biology

Technical annex

December 2023

Statistic used (or Figure)	Data sources & method outline	Caveats & key limitations
	Research metrics	
The UK ranks 5th for the number of engineering biology research publications behind China, the US, and India, and very narrowly behind Germany. (2018 to 2022) **Figure 2: Scholarly outputs and research impact for the top 10 countries producing engineering biology publications (2018 to 2022).**	Number of Engineering Biology publications between 2018 to 2022, as defined by a *Government Office for Science* (GOS) developed bibliometric keyword search adapted by DSIT to apply to the SciVal® database. The research performance metrics are derived using bibliometric data from *SciVal*®, which tracks bibliographic information from *Scopus* and other data sources. (an abstract and citation database licensed by *Elsevier*). *Scopus* data has been used for former BEIS performance releases since 2011 and it covers multi-lingual and global peer-reviewed literature, published in journals, book series and conference proceedings among other features of research performance. Scholarly output in the *SciVal*® platform indicates the prolificacy of an entity (here, a country) and is defined as the number of publications an entity has indexed in Scopus. The keyword search we applied in *SciVal*® was originally developed with the help of subject matter experts at *Government Office for Science*, with consultation across government on the suggested keywords and multiple rounds of quality assurance sampling for false positives.	**We have medium confidence in the reliability of this indicator. The most significant source of uncertainty comes from the use of keywords to identify relevant research activity.** The source data has high but not complete coverage of publications worldwide, with higher coverage in Anglophone countries. For more details of database coverage, see Elsevier's Research Metrics Guidebook. The search strategy, given Engineering Biology's wide scope and the scale of the number of publications, may capture some false positives and miss some true positives. An internationally co-authored paper is counted under the tally of two or more nations. Authorship is according to the location of the institution listed by the authors as their affiliation. The nationality of authors is unknown. Different countries may have different propensities to publish their findings, due to culture, or incentives for researchers. This metric does not correct for this. Scopus is frequently updated and so certain indicators, especially those linked to citations, may

	For each subsequent search a *Government Office for Science* analyst undertakes the following quality control: • Compares the outputs to the last data export and against other database benchmarks. • Checks the data in graph visually for anomalous results. We have since adapted this keyword search strategy for *SciVal*®, making minor tweaks to the keyword list (due to the lack of the NEAR Boolean operator, which will mean that there will be potentially more false negatives, but also fewer false positive).	retrospectively change. This analysis was based on a last update date of 01/11/2023, so figures may not be able to be replicated exactly.
Among the top ten nations producing engineering biology scholarly outputs across 2018 to 2022, the UK ranks 4th for the impact of its engineering biology research. **Figure 2: Scholarly outputs and research impact for the top 10 countries producing engineering biology publications (2018 to 2022).**	Field-Weighted Citation Impact (FWCI) of Engineering Biology publications between 2018 to 2022, as defined by a *GOS* developed bibliometric keyword search. This has been adapted by DSIT to apply specifically to the *SciVal*® database as per the description above. FWCI is a measure of the scholarly impact of a set of publications. It compares how a number of citations for a given set of publications compares to the average number of citations received by all world publications in the same field. A value of 1.0 represents the world average FWCI. The FWCI is calculated by dividing the number of citations a paper has received by the average	**We have medium confidence in the reliability of this indicator. The most significant source of uncertainty comes from the use of keywords to identify relevant activity.** Citations might not always be a genuine indicator of quality. For example, a publication could be cited a lot because a paucity of other sources – indicating impact perhaps, but not necessarily quality. Other caveats listed above within this section.

	number received by documents published in the same year and in the same Fields of Research (FoR) category. The 10 countries with the greatest publication output are defined based on the research output figures for the five year period 2018 to 2022, as used in the indicator above. The dataset was restricted to only these top 10 producers of engineering biology research, to avoid skewing the conclusions by the inclusion of a nation with a very small number of publications but a high FWCI. This is to make meaningful comparisons against nations that are producing a larger number of engineering biology publications.	
The UK already excels at [international collaboration]. From 2018 to 2022, 65% of all UK engineering biology publications featured a collaboration with at least one non-UK author **Figure 7: International collaboration share for the top 10 countries producing engineering biology publications (2018 to 2022).**	Percentage of Engineering Biology publications between 2018 to 2022 with at least one foreign collaborator of all research publications, as defined by a *GOS* developed bibliometric keyword search adapted by DSIT to apply to the *SciVal*® database as above. Publications are assigned to one of four mutually exclusive geographical collaboration types: international, national, or institutional co-authorship, and single authorship. An international publication is a publication which was co-authored by at least two researchers affiliated to institutions in different countries.	**We have medium confidence in the reliability of this indicator. The most significant source of uncertainty comes from the use of keywords to identify relevant activity.** Care must be taken when interpreting, as international collaborations are defined when there is just one non-UK-based author (when the extent of international collaboration on any one paper may range from extensive to minimal). Other caveats listed above within this section.

	Since international authorship is according to the location of the institution listed by the authors as their affiliation, some types of international collaboration will be missed – such as researchers from different countries currently working for institutions in the same country; and some included where the collaboration could be between authors of the same nationality, currently working in institutions in different countries. The 10 countries with the greatest publication output are defined based on the research output figures for the period 2018 to 2022, as used in the indicator above.	
Statistic (or Figure)	**Data sources & method outline**	**Caveats & key limitations**
Businesses & Finance metrics		
The UK also has an impressive cohort of EB firms that have fundraised over £5.2 billion from 2017 to 2022	The total funds raised by Engineering Biology firms between 2017 and 2022, as defined by a *Government Office for Science* (GOS) developed "engineering biology" keyword search of the *PitchBook* investment database *(JUL23B)*. *PitchBook* is an online platform used to find details of deals across the public and private equity markets—including information on: funders, funding rounds and post-money valuations. We have adjusted the $6.8 bn figure to £ by adjusting each year's amount fundraised by the spot GBP:USD annual average spot rate from FRED.	**We have medium confidence in the reliability of this indicator. The most significant source of uncertainty comes from the partial reporting of deals.** Data is based on publicly reported investment in privately held companies (such as Venture Capital, Grants and Venture Debt). This means that this dataset excludes: • Publicly listed companies fundraising is excluded (those on the stock market). • Within firm investment by public companies is not identified, such as Google investing into its own UK subsidiary.

	The most significant sources that contribute to the £5.2bn over the period are: - Later stage VC (27%) - Early stage VC (19%) - Mergers and acquisitions (19%) - PIPE transactions (13%) - IPO (7%) - Buyout (5%) - Seed round (4%) <u>Keyword search terms:</u> Technologies have been defined using a combination of keywords with *PitchBook*'s inbuilt "verticals" and "industries. The keyword search was developed with the help of a subject matter expert at *GOS*, involving consultation across government on the suggested keywords and multiple rounds of quality assurance sampling for false positives. Since the initial keyword search strategy was developed, for subsequent searches a *GOS* analyst undertakes the following quality control: • Compares the outputs to the last data export and against other database benchmarks. • Checks the data in graph visually for anomalous results. Keyword searching requires strong associative and semantic match between the keywords and	There could additionally be some important private equity deals that are not reported publicly and therefore are not included in the statistics. There will be some lag in the data and so the most recent period may be updated at a later date once more deals are visible. However we do not expect this to affect the numbers much, and if it does it would most likely affect 2022. Due to the different approach to keyword searching, the companies identified will not overlap precisely with the companies used for other metrics (e.g. research activity).

	language used on the *PitchBook* platform. *GOS* undertook a high level of quality assurance for the most significant companies in each technology but did not check individual deal amounts. *PitchBook* verticals and industries use keyword searching to identify relevant investment in each technology. This search has not been reviewed by *PitchBook* analysts.	
The UK ranks 3rd globally in total private investment in Engineering Biology from 2017 to 2022, behind the US and China.	Uses *PitchBook* data with the same method as above. Excludes the Cayman Islands in 3rd place, as this is not likely to be where much of the activity funded by this investment takes place.	**We have medium confidence in the reliability of this indicator. The most significant source of uncertainty comes from the partial reporting of deals.** As above.
The United Kingdom leads other European countries in the number of new biotech start-ups and funding for those companies between 2017 to 2022.	As above: the total number of privately listed Engineering Biology firms and total funds raised between 2017 to 2022, as defined by a Government Office for Science developed "engineering biology" keyword search of the PitchBook investment database. Firms counted here are privately held firms only. If they grow to larger scale they may perhaps later list on the stock market.	**We have medium confidence in the reliability of this indicator. The most significant source of uncertainty comes from the partial reporting of deals.** As above.

Figure 3 – UK Engineering Biology Companies by Subsector	*The Data City* is a company focused on mapping emerging economic sectors of the UK. Using a combination of supervised machine learning with a database combining the text of company websites matched to their Companies House Records, emerging sectors can be specified. The DSIT Engineering Biology team has worked with *The Data City* to develop a bespoke "Real Time Industrial Classification" (RTIC) for the sector over summer 2023. We used within DSIT PhD Engineering Biology scientific expertise, the platform's machine learning classification system, broad cross-government engagement, and substantial analyst time to develop a taxonomy for Engineering Biology applications and the supply chain. We have visualised the result of this significant research effort. The company numbers do not include every company identified within the RTIC (which includes multiple subsidiaries of any one entity). Instead, where there are multiple companies with the same web address, we have kept just one company. Why we have done this is because in this context, this is the best representation of the number of substantive independently operating companies within the sector (and the relative size of the subsectors).	**We have high confidence in the reliability of this map of company activity. The most significant source of uncertainty comes from the inclusion / exclusion of boundary cases in the map.** *Note, the scope of the firms included in this figure differ from those using PitchBook (due to different identification strategies and scopes of the respective databases)* *Also note, subsectors in Figure 3 will not sum to 1,162 due to 16% of firms being in overlapping subsectors.* The process has undergone extensive expert input and quality assurance. The final list has been scanned for false positives by both DSIT experts, those across-government and The Data City, though: - On false positives: Given the size and scope of this classification it could still potentially contain a few false inclusions. - In relation to false negatives: DSIT has drawn upon as many existing databases and suggestions as possible to minimise the chance of missing out on key companies, which are then in-turn used to identify other companies for consideration via the machine learning.

<table>
<tr>
<td></td>
<td>Each firm can be classified in multiple subsectors, as per the below table of the number of firms (after the above-described removal of duplicates).

<table>
<tr><th></th><th>Total # Firms</th><th>Within supply chain</th><th>Within applications</th><th>Both Supply Chain & Applications</th><th>…Firms with >1 subsector?</th></tr>
<tr><td>#</td><td>1162</td><td>541</td><td>684</td><td>63</td><td>189</td></tr>
<tr><td>%</td><td>100%</td><td>47%</td><td>59%</td><td>5%</td><td>16%</td></tr>
</table>
</td>
<td>The removal of duplicates by URLs was done crudely, keeping one entity. This means the number of substantive companies may be a minor underestimate, as there may be in some cases, multiple substantive independently operating firms with the same URL (when we have included just one entity per URL).

The Cambridge Industrial Innovation policy map of institutions is not necessarily exhaustive or exclusive. However, we believe it captures some relevant examples of infrastructure and capability relevant to Eng Bio around the UK.</td>
</tr>
<tr>
<td>Figure 4: Number of engineering biology international patent families by inventor country (2010 to 2023).</td>
<td>Patent data was retrieved by the Intellectual Property Office querying PatentSight. This work used a bespoke search strategy developed for DSIT and the Government Office for Science in 2022, combining keywords and relevant patent classifications. After being identified through PatentSight, relevant patent families were matched to PATSTAT, a statistical patent database offered by the European Patent Office. This allows access to a wider range of data compared to using only PatentSight.

This search uses International Patent Families (IPFs), which are families with an application filed in at least two authorities. IPFs were used because they are a more reliable measure of inventive activity than using absolute counts of published patent applications. IPFs are a neutral</td>
<td>We have medium confidence in the reliability of this indicator. The main source of uncertainty comes from the tailoring of keywords for use in the search strategy within the platform.

There is an inherent 18-month delay in patent publication, meaning patent data from 2022 to 2023 may be incomplete. Incomplete data from these years have been included in the chart’s total count since this can help provide more up to date information on ongoing trends even if incomplete.

Search strategies predominantly capture English language which may cause some countries to be underrepresented.</td>
</tr>
</table>

	proxy for inventive activity because they provide a degree of control for patent quality and value by only representing inventions deemed important enough by the applicant to seek protection internationally. The quality assurance of this search involves a designated assurer (a patent examiner with familiarity with Engineering Biology IP) who will work through the strategy, methodology and results with the IPO analyst. This includes checking if there were false positives or false negatives being picked up over multiple rounds. The analyst and assurer then modify the keyword list, to improve results to ensure they are an accurate reflection of the patent landscape. When this search has been updated for the latest data, a *GOS* analyst has undertaken the following quality control: • Compares the outputs to the last data export and against other database benchmarks. • Checks the data in graph visually for anomalous results.	
Statistic (or Figure)	**Data sources & method outline**	**Caveats & key limitations**
Maps		
Figure 51: Map of companies that are applying engineering	Developed through a combination of data from *The Data City*, and another map from Cambridge Industrial Innovation Policy.	**We have high confidence in the reliability of this map of company activity. The most significant**

biology, or are part of its supply chain, overlaid with key UK clusters and capabilities (October 2023 snapshot)	The Data City is a company focused on mapping emerging economic sectors of the UK. Using a combination of supervised machine learning with a database combining the text of company websites matched to their Companies House Records, emerging sectors can be specified. The DSIT Engineering Biology team has worked with the Data City to develop a bespoke "Real Time Industrial Classification" (RTIC) for the sector over summer 2023. We used within DSIT PhD Engineering Biology scientific expertise, the platform's machine learning classification system, broad cross-government engagement, and substantial analyst time to develop a taxonomy for Engineering Biology applications and the supply chain. We have mapped the result of this significant research effort. The companies plotted here, in contrast to the above Figure 3, include every company identified within the RTIC including multiple subsidiaries of any one entity (1744 companies). This is to best capture the geographic spread of the sector in the UK. Cambridge Industrial Innovation Policy previously created a map, annotating on key Engineering Biology R&D related institution locations. DSIT have overlaid these onto the company map that we generated with the Data City.	**source of uncertainty comes from the inclusion / exclusion of boundary cases in the map.** *Note, the scope of the firms included in this figure differ from those using PitchBook (due to different identification strategies and scopes of the respective databases)* The process has undergone extensive expert input and quality assurance. The final list has been scanned for false positives by both DSIT experts, those across-government and The Data City, though given its size and scope it could still potentially contain a few false inclusions. In relation to false negatives, we've drawn upon as many existing databases and suggestions as possible to minimise the chance of missing out on key companies, which are then in-turn used to identify other companies for consideration via the machine learning. The Cambridge Industrial Innovation policy map of institutions is not necessarily exhaustive or exclusive. However, we believe it captures some relevant examples of infrastructure and capability relevant to Eng Bio around the UK. To note, some postcodes are plotted on top of each other.

Figure 6: Total investment fundraising by Engineering Biology Firms (2017 to 2022), with notable companies highlighted.	This map draws upon PitchBook data on number of engineering biology companies and total funds raised, as per the description above between 2017 to 2022. Some manually selected key companies have been highlighted on the map. These are important and illustrative key companies in the sector but have not been selected using a formula. They were chosen using expert judgement from PitchBook / other sources considering: company scale, name recognition, ensuring a variety of sectors etc.	**We have medium confidence in the reliability of this indicator. The most significant source of uncertainty comes from the partial reporting of deals.** PitchBook, as above.