How Much Scientific Data Is Out There, and Are We Using It?

The Data Deluge in Life Science

Chee Ping Ng, Founder
AvatarsBio, The Netherlands

SUMMARY

We surveyed the scale of six major public scientific databases — the Sequence Read Archive (SRA), Gene Expression Omnibus (GEO), Human Cell Atlas (HCA), ChEMBL, PubChem, and PRIDE — compiling growth trajectories from primary sources including raw statistics files, official release notes, and peer-reviewed database update papers. The exercise was prompted by a simple question: how much scientific data actually exists in these archives, and to what extent is it being reused? What we found was consistent across all six domains: exponential accumulation, accelerating submission rates, and a persistent gap between the volume of data deposited and the evidence of meaningful reuse after the originating publication. We close with a brief reflection on what this pattern implies for how science allocates its curatorial effort.

1. WHY WE LOOKED

Scientific databases are frequently cited in aggregate — a petabyte here, a million compounds there — but it is rare to compile verified, primary-sourced figures across multiple domains in a single document. The numbers are often outdated by the time they appear in review articles, sourced from secondary references, or presented without a clear trail back to the underlying statistics.

This survey began with a practical question: across the major archives that life science and chemistry depend on, what does the current scale actually look like when you go directly to the source? We chose six databases spanning genomics, transcriptomics, single-cell biology, drug discovery, cheminformatics, and proteomics — broad enough to be representative, specific enough to be tractable.

As the numbers came together, a second question emerged naturally from the data: given the scale of what already exists, how much of it is being actively reused? The answer, documented in the literature and implicit in the growth curves themselves, prompted the closing reflection in Section 4.

2. HOW WE COMPILED THE DATA

For each database, we identified the primary source for current and historical figures — typically a combination of a live statistics page, a downloadable data file, and the sequence of official database update papers published annually in Nucleic Acids Research. Where a live statistics file was available (SRA, GEO), we downloaded and analysed it directly. Where figures came from published papers, we used the specific release or update paper that reported the number, and noted the version or date. Where live portals were available (PubChem, HCA Data Portal), current figures were drawn directly from those sources; published database update papers provided historical anchor points and methodological context.

We did not rely on secondary citations, review article summaries, or database homepages alone. Every number in Table 1 and Figure 1 is traceable to a named primary source listed in the references.

Historical anchor points for the growth curves were drawn only from verified sources; intermediate years were not interpolated.

This level of verification is more demanding than it might appear. Database statistics pages update continuously, published figures refer to specific release versions, and the same database is often cited with different numbers in different papers depending on the snapshot date. The process took considerable time and revealed that figures for the same database can differ depending on the snapshot date, counting method, or format definition used.

3. WHAT WE FOUND

Table 1. Summary of six major public scientific databases. All figures verified from primary sources as of 2024–2026.

Database	Domain	Latest verified figure	Data source	Established	Ref.
SRA	Genomics	27.2 PB raw data (Jan 2024)	NCBI sra_stat.csv	2008	[1]
GEO	Transcriptomics	8.4M samples; 280k studies (2026 Q1)	NCBI GEO history CSV	2000	[2]
Human Cell Atlas	Single-cell	70.5M cells; 528 projects (2026)	HCA Data Portal	2016	[3],[4]
ChEMBL	Drug discovery	2.5M compounds; 21.1M activities (Ch.35, Dec 2024)	ChEMBL blog release notes	2009	[5],[6]
PubChem	Cheminformatics	123M compounds; 299M bioactivities (Apr 2026)	PubChem statistics page	2004	[7]
PRIDE	Proteomics	29k public datasets; ~3,000 TB; 534 submissions/month (2024)	Perez-Riverol et al. 2025	2005	[8]

Figure 1 shows the growth trajectories for all six databases. The pattern is consistent: exponential accumulation across every domain, with submission rates accelerating rather than plateauing.

GENOMICS · RAW SEQUENCES

Sequence Read Archive

NCBI · est. 2008

27.2 PB

raw sequencing data stored (Jan 2024)

Source: NCBI sra_stat.csv (direct download)

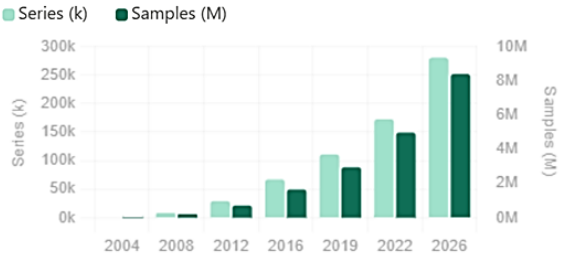
TRANSCRIPTOMICS · EXPRESSION

Gene Expression Omnibus

NCBI · est. 2000

8.4M

biological samples archived (2026 Q1)



Source: NCBI GEO Summary history CSV

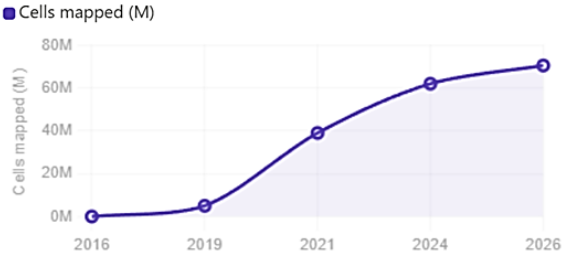
SINGLE-CELL · ATLAS

Human Cell Atlas

Global consortium · est. 2016

70.5M

human cells mapped (2026)



Trends Genetics 2021 · HCA Portal 2026

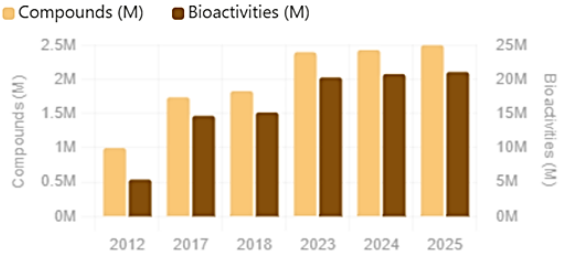
DRUG DISCOVERY · CURATED

ChEMBL

EMBL-EBI · est. 2009

21.1M

bioactivity records (ChEMBL 35, Dec 2024)



ChEMBL blog release notes Ch.14–Ch.35

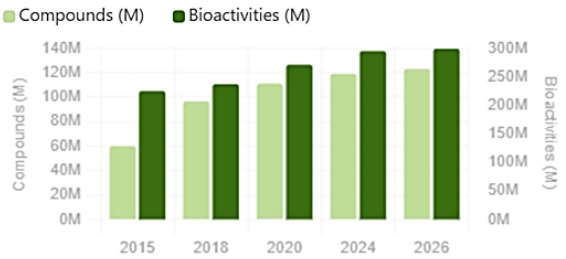
CHEMINFORMATICS · OPEN

PubChem

NCBI · est. 2004

299M

bioactivity datapoints (Apr 2026)



Wikipedia · Kim et al. 2023 · PubChem stats Apr 2026

PROTEOMICS · MASS SPEC

PRIDE Archive

EMBL-EBI · est. 2005

~3,000 TB

cumulative archive size (Aug 2024)

Perez-Riverol et al. NAR 2025 · Fig. 3A & 3B

Six databases, six growth curves — all verified from primary sources. The scientific data deluge in numbers: most of it sits uncurated, waiting for the re-searcher.

Figure 1. Growth trajectories across six major public scientific databases, compiled from verified primary sources. Each panel shows anchor points drawn directly from official statistics files, database release notes, or peer-reviewed update papers. No intermediate values were interpolated.

Sequence Read Archive (SRA)

The SRA has grown from near zero in 2008 to 27.2 petabytes of raw sequencing data by January 2024, based on direct analysis of the NCBI sra_stat.csv file [1]. Early growth was dramatic — roughly doubling every 12–18 months — but has since moderated to approximately 20% per year, implying a doubling time of around 45 months. The archive currently accumulates several petabytes of new data annually.

Gene Expression Omnibus (GEO)

GEO now archives over 8.4 million biological samples across more than 280,000 studies as of 2026 Q1, based on the NCBI GEO quarterly history CSV [2]. Growth has been steady over two decades with a notable acceleration in sample counts from 2022 onwards, likely driven by the proliferation of single-cell RNA sequencing technologies.

Human Cell Atlas (HCA)

The HCA had profiled more than 39 million cells from 15 major organs by 2021 [3], reaching 62 million cells by 2024 and 70.5 million cells across 528 projects and 18 biological networks by early 2026 [4]. The growth reflects both expanding single-cell technology capacity and the coordinated international effort behind the consortium.

ChEMBL

ChEMBL grew from approximately 1.0 million compounds and 5.4 million bioactivity records at its 2012 founding paper [5] to 2.5 million compounds and 21.1 million bioactivity records as of ChEMBL 35 in December 2024 [6]. A notable observation: bioactivity records have grown fourfold while compound count has grown 2.5-fold — suggesting repeated measurement of the same chemical space in different biological contexts, rather than genuine expansion of the explored chemical frontier.

PubChem

PubChem now contains 123 million unique compounds, 344 million substance entries, and 299 million bioactivity datapoints as of April 2026 [7]. The bioactivity count alone has grown from 225 million in 2015 to 299 million today — an increase of 74 million datapoints in approximately a decade.

PRIDE Archive

PRIDE reached 42,036 total datasets by August 2024, of which 29,039 (69%) are publicly available [8]. Average monthly submissions grew from approximately 20 per month in 2013 to 534 per month in 2023, with a single-month record of 636 in July 2024. Cumulative archive size reached approximately 3,000 terabytes by mid-2024.

4. WHAT THIS SUGGESTS

The growth curves in Figure 1 tell a consistent story across six domains: data is accumulating faster than the field's capacity to curate, annotate, and reuse it. Several observations follow from the numbers.

First, deposition and reuse are not the same thing. Journal mandates now require data deposition as a condition of publication, which explains much of the accelerating submission rate. But mandated deposition does not guarantee usable data. PRIDE's own 2025 update paper [8] notes that the vast majority of deposited proteomics datasets show zero reported reanalyses in the scientific literature after their originating publication. GEO faces persistent metadata quality problems that limit cross-study comparisons even where the will to reuse exists.

Second, the data that does get reused tends to be the data that was designed for reuse. ChEMBL is the clearest example among the six: its manual curation model, explicit data standards, and community-facing infrastructure make it the most actively queried database in the drug discovery field. The contrast with PRIDE — where curation is inconsistent and reuse rare — is instructive. Curation is not an optional post-processing step; it is a prerequisite for reuse.

Third, the temporal dimension matters. Vines et al. [9] demonstrated that the availability of raw data from published studies declines by approximately 17% per year after publication — broken links, lost files,

unresponsive authors. Combined with metadata quality problems, this means a significant fraction of the archive is not merely unused but becoming unusable over time.

The word research derives from the Old French *rechercher* — to seek intensely. It was never only about generating the new. To re-search, in its most literal sense, is to go back into what already exists with fresh eyes and better tools. Sikkema et al. integrated Human Lung Cell Atlas [10] illustrates what this looks like in practice: rather than generating new data, the authors harmonised 49 existing datasets into a single reference atlas of 2.4 million cells from 486 individuals — a resource that could not have been produced de novo. The scientific value generated was proportional to the curatorial effort invested, not to the volume of new data produced.

The question the numbers raise is not whether existing archives are large — they clearly are. It is whether the field's allocation of effort between generation and curation is calibrated to extract the maximum scientific value from what already exists.

> ***"The question is no longer whether we can generate more data, but whether we can meaningfully reuse what already exists."***

REFERENCES

1. NCBI. Sequence Read Archive growth statistics. sra_stat.csv. ncbi.nlm.nih.gov/sra/docs/sragrowth · Leinonen R et al. The Sequence Read Archive. Nucleic Acids Research. 2011. https://doi.org/10.1093/nar/gkq1019
2. NCBI. GEO Summary quarterly history CSV. ncbi.nlm.nih.gov/geo/summary/?type=history · Edgar R et al. Gene Expression Omnibus. Nucleic Acids Research. 2002 · Clough E & Barrett T. The Gene Expression Omnibus Database. Methods in Molecular Biology. 2016. https://doi.org/10.1007/978-1-4939-3578-9_4
3. Haniffa M et al. A roadmap for the Human Cell Atlas. Trends in Genetics. 2021.
4. Regev A et al. The Human Cell Atlas. eLife. 2017. https://doi.org/10.7554/eLife.27041 · HCA Data Portal. data.humancellatlas.org
5. Gaulton A et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2012. https://doi.org/10.1093/nar/gkr777
6. Zdrazil B et al. The ChEMBL Database in 2023. Nucleic Acids Research. 2024. https://doi.org/10.1093/nar/gkad1004 · ChEMBL release notes Ch.34–Ch.35. chembl.blogspot.com
7. Kim S et al. PubChem 2023 update. Nucleic Acids Research. 2023. https://doi.org/10.1093/nar/gkac956 · PubChem statistics. pubchem.ncbi.nlm.nih.gov/docs/statistics
8. Perez-Riverol Y et al. The PRIDE database at 20 years: 2025 update. Nucleic Acids Research. 2025. https://doi.org/10.1093/nar/gkae1011
9. Vines TH et al. The availability of research data declines rapidly with article age. Current Biology. 2014. https://doi.org/10.1016/j.cub.2013.11.014
10. Sikkema L et al. An integrated cell atlas of the lung in health and disease. Nature Medicine. 2023. https://doi.org/10.1038/s41591-023-02327-2