nature methods

Brief Communication

https://doi.org/10.1038/s41592-025-02687-2

NeuroXiv: AI-powered open databasing and dynamic mining of brain-wide neuron morphometry

Received: 19 September 2024

Accepted: 31 March 2025

Published online: 29 April 2025

Check for updates

Shengdian Jiang[1,2,6], Lijun Wang[1,3,6], Zhixi Yun[1,2,6], Hanbo Chen[4,6], Lijuan Liu[1,3] ✉, Jianhua Yao[4] ✉ & Hanchuan Peng[5] ✉

Neuron morphology has been extensively reconstructed at the whole-brain scale by various projects in recent years. Here, to facilitate interactive exploration in a standardized and scalable manner, we introduce NeuroXiv (neuroxiv.org), a large-scale database containing 175,149 reconstructed neuron morphologies mapped to the Common Coordinate Framework Version 3 (CCFv3). In addition, NeuroXiv incorporates an AI-powered mining engine (AIPOM) for dynamic, user-specific data mining, delivering enhanced performance via a custom client program.

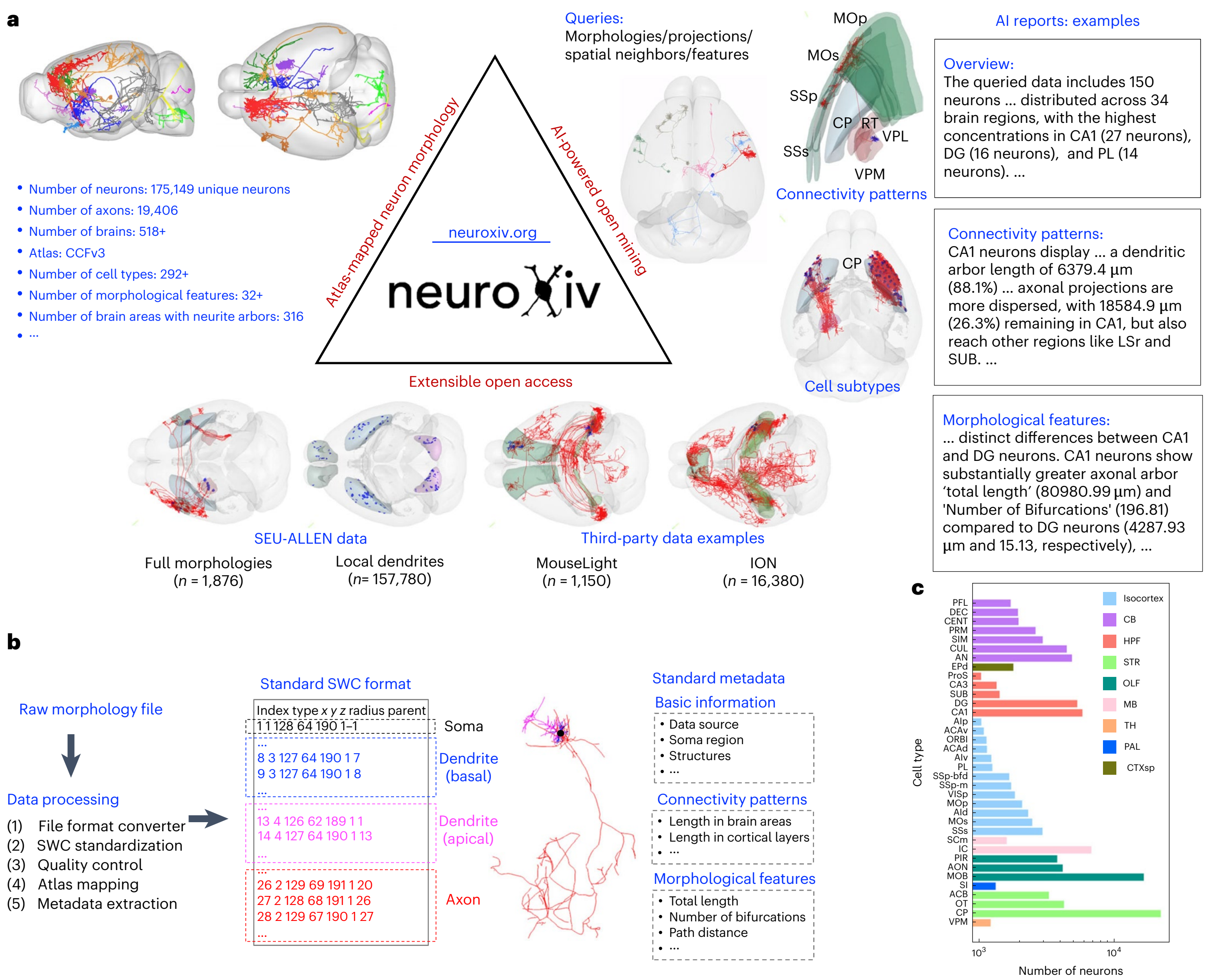
Neuronal morphologies, characterized by their diverse branching patterns and anatomical arborization, provide critical insights into cell types and brain functional networks[1]. Recent advancements, including sparse labeling techniques[2], high-resolution brain imaging[3], terabyte-scale image handling[4] and neuron tracing methods[5], have enhanced our capability to digitize brain-wide neuron morphologies. As a result, there has been a substantial increase in the volume of publicly accessible neuron morphologies, which supports various quantitative analyses, including the morphological characteristics of individual neurons[6], dendritic microenvironments[7], neuron typing[8] and organizational principles of neuron projections[9]. However, a remarkable gap exists: how do we harness these valuable datasets from diverse sources for new discoveries while addressing dynamic needs throughout the development process?

Current data dissemination solutions for neuron morphology[9–13] generally fall into two categories: browser-based platforms, such as Neuron Browser (mouselight.janelia.org) and Digital Brain (mouse.digital-brain.cn), and archiving platforms, including Brain Image Library (brainimagelibrary.org), NeuroMorpho.Org (neuromorpho.org) and Neurons Reunited Portal (neuroinformatics.nl/HBP/neuronsreunited-viewer/). Archiving platforms provide dataset downloads and a broader range of data, while browser platforms offer exploration tools but limited access. The Neurons Reunited Portal stands out by aggregating data from multiple sources and mapping all morphologies onto the Common Coordinate Framework (CCF) for unified visualization. Large-scale offline analyses face challenges such as dataset harmonization, CCF alignment, metadata extraction and transforming neuronal features into insights, requiring both domain expertise and advanced coding skills (Extended Data Table 1 and Supplementary Notes 1 and 2).

We introduce the NeuroXiv platform (neuroxiv.org), currently hosted on Amazon Web Services (AWS), designed to address challenges in databasing and mining brain-wide neuron morphometry. Building upon the foundational work of the Allen Brain Atlas[14] and NeuroMorpho.Org, we have expanded efforts to establish a standardized atlas-oriented database of neuron morphometry (Fig. 1a). To address challenges associated with large-scale analysis of neuron morphology, we have developed the AI-Powered Open Mining (AIPOM) engine. This engine offers functionalities such as searching and visualizing neuron morphometry, enabling analyses including data statistics, cell typing and connectivity studies. Crucially, it incorporates advanced capabilities, such as generating artificial intelligence (AI)-driven mining reports (Figs. 1a and 2a).

We established a server-side data processing pipeline capable of continuously aggregating publicly available datasets into our database (Fig. 1b). Interoperability is maintained through standardization of neuron morphology in the widely used SWC (named after its initial developers Ed Stockley, Howard Wheal and Robert Cannon) format[15].

[1]New Cornerstone Science Laboratory, SEU-ALLEN Joint Center, Institute for Brain and Intelligence, Southeast University, Nanjing, China. [2]School of Computer Science and Engineering, Southeast University, Nanjing, China. [3]School of Biological Science and Medical Engineering, Southeast University, Nanjing, China. [4]Tencent AI, Shenzhen, China. [5]Shanghai Academy of Natural Sciences, Fudan University, Shanghai, China. [6]These authors contributed equally: Shengdian Jiang, Lijun Wang, Zhixi Yun, Hanbo Chen. ✉e-mail: lijuan-liu@seu.edu.cn; jianhuayao@tencent.com; h@braintell.org



Overview:
The queried data includes 150 neurons ... distributed across 34 brain regions, with the highest concentrations in CA1 (27 neurons), DG (16 neurons), and PL (14 neurons). ...

Connectivity patterns:
CA1 neurons display ... a dendritic arbor length of 6379.4 µm (88.1%) ... axonal projections are more dispersed, with 18584.9 µm (26.3%) remaining in CA1, but also reach other regions like LSr and SUB. ...

Morphological features:
... distinct differences between CA1 and DG neurons. CA1 neurons show substantially greater axonal arbor 'total length' (80980.99 µm) and 'Number of Bifurcations' (196.81) compared to DG neurons (4287.93 µm and 15.13, respectively), ...

```
Index type x y z radius parent
1 1 128 64 190 1–1
...
8 3 127 64 190 1 7
9 3 127 64 190 1 8
...
...
13 4 126 62 189 1 1
14 4 127 64 190 1 13
...
...
26 2 129 69 191 1 20
27 2 128 68 191 1 26
28 2 129 67 190 1 27
...
```

Fig. 1 | Overview of NeuroXiv, an open, AI-assisted database for interactive brain-wide neuron analysis. a, NeuroXiv is founded on three pillars: atlas-mapped neuron morphology, AIPOM and extensible open access. The AIPOM engine enables users to efficiently and flexibly retrieve neuronal data and to explore neuron types and connectivity patterns, and offers an intelligent mining tool for generating comprehensive data mining reports. **b**, The dataset standardization process in NeuroXiv is performed server-side, which is crucial for ensuring data reusability and interoperability. This process involves formatting raw morphology files into the standard SWC format and storing them accordingly. In addition, the data are mapped to the same atlas space, and rich metadata are extracted to enhance the dataset's utility. **c**, NeuroXiv has established the largest and most comprehensive dataset of neuron types, covering a wide range of brain regions including the TH, STR, isocortex, HPF and CB. A full list of abbreviations for all brain structures in this study is provided in the Methods.

Reusability is ensured by initially mapping neuron morphologies into the Common Coordinate Framework Version 3 (CCFv3)[14], a widely recognized brain atlas for registering neuroanatomical data. We systematically document comprehensive metadata, encompassing basic information, morphological features and anatomical arborization characteristics of neuron morphology (Supplementary Table 1).

We demonstrate the feasibility and scalability of the databasing method by consolidating data from diverse sources, including our SEU-ALLEN datasets[6,7] and third-party examples[9,10,16,17], culminating in a large database of brain-wide neuron morphologies (Fig. 1a and Extended Data Fig. 1). The database features over 175,149 atlas-oriented reconstructed morphologies of individual neurons derived from more than 500 mouse brains. Importantly, it contains 19,406 fully traced axonal morphologies. Each neuron reconstruction is characterized by its structural components (soma, axon and dendrite), alongside metadata documented using a common standardized description method (Supplementary Table 2). The resource provides access to the most comprehensive atlas of cell types based on soma anatomical locations, encompassing 12 major gray-matter divisions and including data from 292 out of 316 brain structures in the Allen Reference Atlas ontology (Fig. 1c).

Our database offers several advantages for neuron morphology research. By mapping neuron morphologies from diverse sources onto a unified coordinate system, it enables rapid access to neuronal data without the need to switch between different sources. This data aggregation facilitates detailed, data-driven analyses of neuronal morphological characteristics and enhances the study of brain connectivity at the single-cell level. Specifically, we have identified a greater number of incoming neurons that extend their arbors into specific brain regions (Extended Data Fig. 2a) and have uncovered additional projection combinations of target regions formed by individual neurons (Extended Data Fig. 2b). Furthermore, the database's enhanced

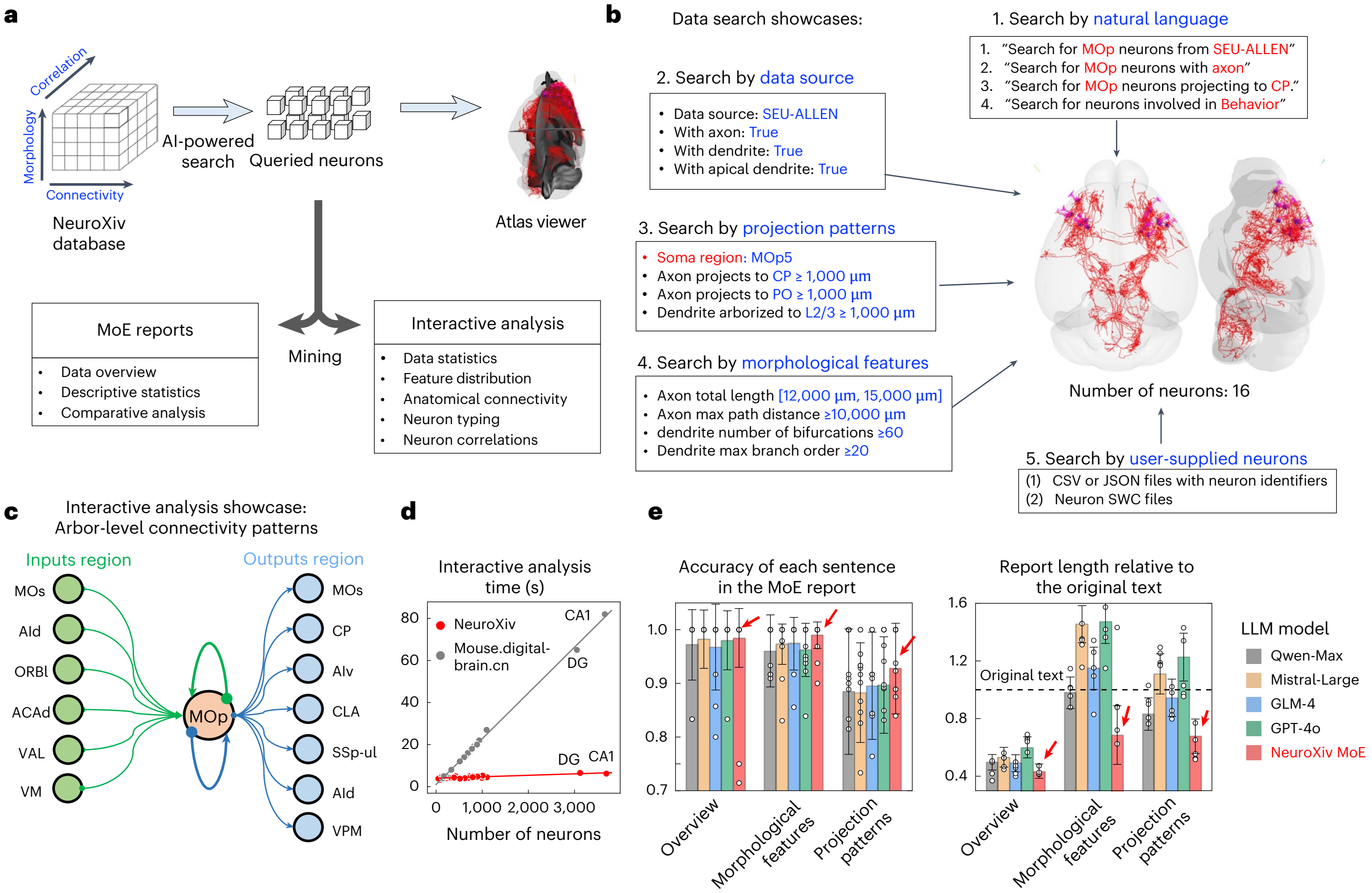


Fig. 2 | AI-powered engine for open analysis and mining of neuron data. **a**, A schematic diagram of data analysis and mining within NeuroXiv. The NeuroXiv database provides extensive morphology data, along with detailed morphological features, connectivity patterns and interdata correlations. The AI-powered search tool assists users in extracting relevant data of interest. Users can then interactively visualize the retrieved data on an atlas viewer and explore data features in depth. In addition, NeuroXiv includes a MoE module that automatically generates reports to describe data characteristics and uncover patterns within the data. **b**, The showcases of data search in NeuroXiv demonstrate five common use cases. Users can search data using natural language queries, followed by searches based on the distribution of specific features provided by the user. In addition, data can be retrieved using a user-supplied list of neurons, further enhancing the flexibility of data searches. **c**, The interactive analysis showcase in NeuroXiv presents an example where users can study arbor-level connectivity patterns. The input and output regions are determined by the spatial proximity of neurons within the database and the retrieved data, allowing detailed exploration of neuronal connectivity. **d**, A comparison of interactive analysis time between the two platforms focuses on a shared cell type analysis scenario. The time measurement includes the entire process from data retrieval to the rendering of analysis charts. **e**, MoE reports were evaluated against four LLMs (Qwen-Max, Mistral-Large, GLM-4 and GPT-4o) for accuracy and text length across 291 random analysis cases. Our MoE showed higher accuracy and conveyed the same information using shorter text. Data are presented as mean values ± s.d.

indexing system allows improved retrieval of neurons based on spatial proximity, morphological similarity or shared arborization patterns (Extended Data Figs. 2c,d and 3). This advanced indexing opens up new research opportunities, such as investigating whether spatially adjacent neurons consistently share similar morphological or projection characteristics (Extended Data Fig. 3a,b).

The AIPOM engine streamlines the knowledge development workflow on the established neuron morphometry database by integrating large language models (LLMs), which have rapidly advanced in recent years and demonstrated effectiveness in various domains[18] due to their robust text comprehension capabilities. AIPOM enables users to define data cohorts using natural language or rule-based queries, with the LLM-based tool generating comprehensive reports on morphology, connectivity and cell type comparisons (Fig. 2a and Supplementary Fig. 1). NeuroXiv also offers interactive tools for visualizing quantitative results and exploring complex neuronal structures (Extended Data Fig. 4 and Supplementary Figs. 2 and 3).

We demonstrate the broad applicability and flexibility of our data search tool through several showcases of querying primary motor area (MOp) neurons (Fig. 2b). We first illustrate that searches can be conducted using an LLM-based method, enabling users to query for specific neuron types, neurons with particular structures (such as axons or dendrites) or those exhibiting specific projection patterns through natural language inputs (Supplementary Fig. 4). Notably, based on the established correspondence between function and neuron location (Supplementary Table 3), users can define neuron retrieval on the basis of functional information. We then show that precise searches can be performed by setting customized criteria based on neuron metadata (Extended Data Fig. 5a–c and Supplementary Figs. 5 and 6), such as searching for neurons with long-range projection patterns or those with specific projection patterns (Supplementary Fig. 7). NeuroXiv further provides a database interface enabling users to index specific neurons via an upload function, facilitating its use as a downstream exploration tool following user-defined neuron classification (Supplementary Fig. 8). Moreover, the search tool supports advanced capabilities such as similarity searches to identify neurons with comparable morphological features and arborization patterns (Extended Data Fig. 3a,b and Supplementary Fig. 9), as well as neighboring neuron queries to explore arbor-level connectivity within the brain (Extended Data Fig. 3c,d and Supplementary Fig. 10).

We highlight the interactive analysis capabilities of AIPOM through two studies. In the first study, users can identify which neuron types in the database provide input to a single neuron and determine the brain regions that receive projections from that neuron (Fig. 2c). In the second study, focusing on projection patterns, we use ventral posteromedial nucleus (VPM) neurons from our database as an example (Extended Data Fig. 6). These VPM neurons, sourced from two datasets, exhibit axonal arbors that extend across the caudoputamen into multiple cortical regions, including secondary motor area (MOs), MOp, primary somatosensory area (SSp) and supplemental somatosensory area (SSs), encompassing various projection subtypes (Extended Data Fig. 6a). In addition, our visualization tools allow users to observe the selectivity of different projection subtypes across cortical regions and layers (Extended Data Fig. 6b,c), as well as compare soma distribution and morphological features among these subtypes (Extended Data Fig. 6d,e). These findings align with prior knowledge of VPM neuron projection patterns[19].

NeuroXiv also demonstrates high efficiency in online analyses (Fig. 2d). We benchmarked NeuroXiv against the Digital Brain platform, finding that NeuroXiv completes most tasks within 4–5 s, regardless of data volume. By contrast, Digital Brain's response time increases linearly with data size, taking over 80 s for the same CA1 neuron results—nearly 20 times slower than NeuroXiv.

To improve the integration of LLMs into AIPOM, we implement two key optimizations. First, to address the challenges of unpredictable outputs and occasional inaccuracies, we developed an advanced Mixture of Experts (MoE) framework for more reliable mining reports (Fig. 2a, Extended Data Fig. 7 and Supplementary Figs. 11 and 12). This framework operates in three stages: first, a program generates standardized reports that capture all relevant data details; second, multiple LLM experts analyze and summarize these reports from a data scientist's perspective; and third, a separate LLM reviews the outputs for accuracy and consistency, producing the final report. This multiexpert approach allows MoE to deliver comprehensive data overviews while effectively identifying morphological and projection differences. Our tests show that the MoE framework yields higher accuracy and more concise reports compared with those generated by a single LLM (Fig. 2e).

Second, to address the computational demands of server-side LLM deployment, we offer a client-side solution using a natural language processing (NLP) model and a supervised decision tree. This approach transforms natural language queries into actionable search operations within 2–3 s, achieving comparable accuracy with a 12.3-fold improvement in response time compared with LLM-based server-side searches (Extended Data Table 2).

In summary, NeuroXiv provides global access to the largest neuron morphometry database, aggregating and standardizing datasets into the SWC format and mapping them to the CCFv3 atlas for enhanced reusability. It integrates search, visualization and analysis tools, leveraging advanced LLMs for intuitive queries and mining reports. AIPOM optimizes performance using the MoE framework and client-side deployment, creating an open, scalable and efficient platform for neuron data reuse in neuroscience.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-025-02687-2.

References

1. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
2. Aransay, A., Rodríguez-López, C., García-Amado, M., Clascá, F. & Prensa, L. Long-range projection neurons of the mouse ventral tegmental area: a single-cell axon tracing analysis. *Front. Neuroanat.* **9**, 59 (2015).
3. Gong, H. et al. High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nat. Commun.* **7**, 12142 (2016).
4. Bria, A., Iannello, G., Onofri, L. & Peng, H. TeraFly: real-time three-dimensional visualization and annotation of terabytes of multidimensional volumetric images. *Nat. Methods* **13**, 192–194 (2016).
5. Manubens-Gil, L. et al. BigNeuron: a resource to benchmark and predict performance of algorithms for automated tracing of neurons in light microscopy datasets. *Nat. Methods* https://doi.org/10.1038/s41592-023-01848-5 (2023).
6. Peng, H. et al. Morphological diversity of single neurons in molecularly defined cell types. *Nature* **598**, 174–181 (2021).
7. Liu, Y. et al. Neuronal diversity and stereotypy at multiple scales through whole brain morphometry. *Nat. Commun.* **15**, 10269 (2024).
8. Xiong, F. et al. DSM: deep sequential model for complete neuronal morphology representation and feature extraction. *Patterns* **5**, 100896 (2024).
9. Qiu, S. et al. Whole-brain spatial organization of hippocampal single-neuron projectomes. *Science* **383**, eadj9198 (2024).
10. Winnubst, J. et al. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. *Cell* **179**, 268–281 (2019).
11. Kenney, M. et al. The Brain Image Library: a community-contributed microscopy resource for neuroscientists. *Sci. Data* **11**, 1212 (2024).
12. Ascoli, G. A., Maraver, P., Nanda, S., Polavaram, S. & Armañanzas, R. Win–win data sharing in neuroscience. *Nat. Methods* **14**, 112–116 (2017).
13. Timonidis, N. et al. Translating single-neuron axonal reconstructions into meso-scale connectivity statistics in the mouse somatosensory thalamus. *Front. Neuroinform.* **17**, 1272243 (2023).
14. Wang, Q. et al. The Allen Mouse Brain Common Coordinate Framework: a 3D reference atlas. *Cell* **181**, 936–953 (2020).
15. Mehta, K. et al. Online conversion of reconstructed neural morphologies into standardized SWC format. *Nat. Commun.* **14**, 7429 (2023).
16. Gao, L. et al. Single-neuron projectome of mouse prefrontal cortex. *Nat. Neurosci.* **25**, 515–529 (2022).
17. Gao, L. et al. Single-neuron analysis of dendrites and axons reveals the network organization in mouse prefrontal cortex. *Nat. Neurosci.* **26**, 1111–1126 (2023).
18. Bzdok, D. et al. Data science opportunities of large language models for neuroscience and biomedicine. *Neuron* **112**, 698–717 (2024).
19. Rubio-Teves, M. *et al.* Beyond barrels: diverse thalamocortical projection motifs in the mouse ventral posterior complex. *J. Neurosci*. **44**, e1096242024, (2024).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Nomenclature and abbreviations of brain regions

The 12 major divisions of gray matter in the Allen Reference Atlas ontology: isocortex, olfactory areas (OLF), hippocampal formation (HPF), cortical subplate (CTXsp), striatum (STR), pallidum (PAL), thalamus (TH), hypothalamus (HY), midbrain (MB), pons (P), medulla (MY) and cerebellum (CB).

Isocortex: primary motor area (MOp), secondary motor area (MOs), primary somatosensory area (SSp), supplemental somatosensory area (SSs), gustatory area (GU), visceral area (VISC), dorsal auditory area (AUDd), primary auditory area (AUDp), posterior auditory area (AUDpo), ventral auditory area (AUDv), primary visual area (VISp), anterior cingulate area, dorsal part (ACAd), anterior cingulate area, ventral part (ACAv), prelimbic area (PL), infralimbic area (ILA), orbital area, lateral part (ORBl), orbital area, medial part (ORBm), orbital area, ventrolateral part (ORBvl), agranular insular area, dorsal part (AId), agranular insular area, posterior part (AIp), agranular insular area, ventral part (AIv), retrosplenial area, ventral part (RSPv) and temporal association area (TEa).

Olfactory areas (OLF): piriform area (PIR).

Hippocampal formation (HPF): hippocampal region (HIP), fields CA1, CA2 and CA3, dentate gyrus (DG), dentate gyrus, molecular layer (DG-mo), entorhinal area, lateral part (ENTl), entorhinal area, medial part (ENTm), parasubiculum (PAR), postsubiculum (POST), presubiculum (PRE), subiculum (SUB) and prosubiculum (ProS).

Cortical subplate (CTXsp): claustrum (CLA).

Cerebral nuclei (CNU): striatum (STR), caudoputamen (CP), nucleus accumbens (ACB), globus pallidus, external segment (GPe), globus pallidus and internal segment (GPi).

Thalamus (TH): ventral anterior–lateral complex (VAL), ventral medial nucleus (VM), ventral posterolateral nucleus (VPL), ventral posterolateral nucleus, parvicellular part (VPLpc), ventral posteromedial nucleus (VPM), ventral posteromedial nucleus, parvicellular part (VPMpc), medial geniculate complex, dorsal part (MGd), lateral geniculate complex, dorsal part (LGd), lateral posterior nucleus (LP), posterior complex (PO), anteromedial nucleus (AM), mediodorsal nucleus (MD), submedial nucleus (SMT), paraventricular nucleus (PVT) and reticular nucleus (RT).

Hypothalamus (HY): subthalamic nucleus (STN) and zona incerta (ZI).

Midbrain (MB): substantia nigra, reticular part (SNr) and midbrain reticular nucleus (MRN).

NeuroXiv platform

The architecture of NeuroXiv (neuroxiv.org) is designed to support large-scale analysis of brain-wide neuron morphometry, utilizing a cohesive and highly integrated technology stack.

Frontend. The frontend of NeuroXiv is developed using Vue.js (v2.6.12), a progressive JavaScript framework known for its efficiency in building dynamic and responsive single-page applications. To create a user-friendly and visually engaging interface, Element Plus (v2.7.3), a Vue 3-based component library, is used. In addition, Three.js (v0.134.0) is integrated into the frontend to handle the rendering of complex three-dimensional (3D) visualizations, including brain regions and neuron reconstructions. This powerful WebGL-based library allows the generation of detailed and interactive 3D models, providing users with an immersive experience in exploring neuroanatomical data.

Backend. The backend is constructed with Python (v3.9.12) and Flask (v3.0.0), a lightweight Web Server Gateway Interface web application framework. Flask serves as the backbone of the server-side architecture, enabling seamless communication between the frontend and the database. SQLite (v3.38.2) is used as the database engine, offering a self-contained, serverless solution for efficient data storage and retrieval. This setup ensures that the platform remains agile and capable of handling the substantial datasets inherent to neuroinformatics research.

To manage web traffic and optimize performance, Nginx (v1.24.0) is deployed as a reverse proxy server. Nginx efficiently distributes incoming requests across backend processes, enhancing the platform's ability to support a high volume of concurrent users while maintaining fast response times and secure connections.

NeuroXiv is hosted on AWS with four central processing units, 32 GB of memory and 12.5 Gbps network bandwidth, leveraging AWS's scalable and resilient cloud infrastructure to provide reliable access for users worldwide. This deployment strategy ensures that the platform remains highly available and capable of scaling in response to increasing user demand, thereby offering a stable and responsive environment for researchers.

Datasets

Currently, the NeuroXiv database reports the integration of several brain-wide neuron morphology datasets shared by the community. Each dataset will be described in detail in the following sections. In the future, we plan to continuously add new mouse brain datasets and encourage users to contribute their own datasets to the NeuroXiv platform. Furthermore, in upcoming updates, we plan to incorporate the BigNeuron Project[5]—a community-contributed resource for benchmarking neuron morphology autotracing algorithms—into NeuroXiv, providing ongoing support for users worldwide.

SEU-ALLEN full dataset. This dataset[6,7] was initially generated using a semi-automatic annotation pipeline with 1,741 neurons and has been expanded to 1,876 neurons with improved quality[20]. Each neuron includes fully traced axonal and dendritic arbors, with 512 apical dendrites additionally annotated. The data mainly cover neurons in the VPM (389 neurons), CP (324 neurons) and many cortical regions.

SEU-ALLEN local dataset. This dataset was produced using an automatic tracing method described in our previous work[7]. Initially, image volumes centered on the cell body (soma) were extracted from whole-brain image data, with a size greater than 200 μm in each dimension, sufficient to capture most of the neuron's dendritic arbor. These image volumes were then processed using image enhancement algorithms[21] to improve image quality. Automatic reconstructions were generated and cross-validated using the APP2[22] and NeuTube[23] algorithms, followed by neurite fiber pruning to remove extraneous signals[24]. In NeuroXiv, we retained only the neurite segments within 100 μm of the soma to ensure consistency. This dataset contains 155,743 neurons, which are extensively distributed across various brain regions such as CP, MOB, OT, AON and PIR.

ION datasets. Currently, NeuroXiv integrates two datasets[9,16,17] from Institute of Neuroscience (ION): a prefrontal cortex dataset comprising 6,357 neurons ('Single-neuron projectome of mouse prefrontal cortex (with dendrite)', Brain Science Data Center, Chinese Academy of Sciences; https://cstr.cn/33145.11.BSDC.1689837400.1681922768243666945 and https://doi.org/10.12412/BSDC.1690164952.20001) and a hippocampus dataset consisting of 10,100 neurons ('Single-neuron datasets for mouse hippocampus', Brain Science Data Center, Chinese Academy of Sciences; https://cstr.cn/33145.11.BSDC.1667284058.1585980235450376194 and https://doi.org/10.12412/BSDC.1667278800.20001). During data integration, 77 neurons with indeterminate soma locations were excluded, resulting in a final dataset comprising 16,380 fully reconstructed axons and 6,106 fully reconstructed dendrites. The neurons are primarily distributed across brain regions such as CA1 (3,657 neurons), DG-sg (2,618 neurons), SUB (934 neurons) and CA3 (887 neurons).

MouseLight dataset. The MouseLight project[10] currently publishes data on 1,200 neurons available at MouseLight NeuronBrowser (http://ml-neuronbrowser.janelia.org). During data integration, 50 neurons with somas located in fiber tracts were excluded, resulting in 1,150 neurons from MouseLight being included in NeuroXiv. This dataset contains 1,150 fully reconstructed axons and 1,138 fully reconstructed dendrites. The neurons are distributed across various brain regions, such as MOs, SUB, PRE, VAL, DG-mo and VPM, with some overlap with data from SEU-ALLEN and ION.

Data aggregation

Data aggregation in NeuroXiv involves collecting datasets from our own datasets (SEU-ALLEN) and third-party sources such as ION and MouseLight, and processing the data to convert it into a consolidated format.

Data format conversion. SEU-ALLEN datasets have already been processed into the standardized SWC format[15] and registered to the CCFv3 atlas. Therefore, our focus here is on processing the datasets from ION and MouseLight. The ION and MouseLight datasets had different format issues. We standardized the ION datasets into the SWC format, aligning the structure domain types, for example, soma (type label = 1), axon (type label = 2), basal dendrite (type label = 3) and apical dendrite (type label = 4). We also converted the neuron reconstruction data from the MouseLight dataset from JavaScript object notation (JSON) files into SWC files.

Quality control. We first performed a quality screening process to ensure data usability, filtering out noncompliant data. The specific steps included:

(1) Single connected tree: ensuring that all nodes have only one parent node, tracing back to a single root node (soma).
(2) Root node (soma) labeling: verifying that there is exactly one node labeled as type = 1 with parent = −1.
(3) Structure domain type correctness: confirming that type attributes 1–4 are valid and that the type attribute remains consistent when tracing from terminal nodes back to the root.
(4) SWC tree structure integrity: checking for the presence of loops and trifurcations in the SWC tree structure.

Atlas mapping. Using mBrainAligner[25,26], we mapped all data points to the CCFv3 atlas. We then resampled the atlas-oriented reconstruction data, ensuring that the distance between parent and child nodes was set to 1 µm.

Data curation. All data points were renamed to follow a standardized format: '<resource_name>_<full/local>_<brainid>_<neuronid>_..._<atlas>' (for example, 'SEU-ALLEN_full_17302_00001_CCFv3').

Metadata extraction. We first extracted basic information such as the soma region for each neuron in the dataset. Then, using the atlas annotation template, we calculated the arborization strength for axons and dendrites across different brain regions based on neurite length. Finally, we extracted morphological features for axons and dendrites (Supplementary Table 1).

We generated lists of morphology- and projection-similar neurons based on feature distances and axonal node point clouds. We also defined axon- and dendrite-neighboring neurons on the basis of arbor overlap. All metadata and similarity tables are stored in an SQL database for easy user access.

Visualization

Brain atlas visualization. We use the visualization toolkit (VTK) to render brain atlases and neuron morphologies. Brain regions are visualized using annotation templates, marching cubes[27] for mesh contours, and Laplacian smoothing[28]. To optimize performance, mesh triangles are reduced with progressive decimation, generating VTK files for 838 brain regions in the CCFv3 atlas.

Neuron morphology visualization. The system includes two components: thumbnail views for quick neuron morphology overviews and 3D atlas visualization for detailed exploration. Thumbnails are generated by resampling morphologies with a 100 µm step size and using principal component analysis for two-dimensional projections. Three-dimensional visualization converts neuron structures into object (OBJ) files for VTK, with soma rendered as a 50-µm sphere using Three.js.

MoE

We have developed a MoE framework that leverages four LLMs, each containing trillions of parameters. This system is specifically designed to collaboratively mitigate errors and hallucinations that are commonly associated with LLM-generated content, thereby producing reliable, accurate and coherent data analysis reports. The MoE framework operates in three distinct stages:

(1) Descriptive report generation: Initially, data retrieved from the database are programmatically organized into a standardized data description format. This ensures consistency and facilitates accurate analysis by the models.
(2) LLM export reports: The organized data are then independently analyzed by three models—Qwen-Max-0428, Mistral-Large-2407 and GLM-4-0520. Each model is tasked with generating an analysis report based on the following prompts (In the following statement, {data} refers to the descriptive reports generated in the previous step):
    2.1 Prompt for overview
        Objective: To provide a concise summary that enhances readability and clarity, with a focus on accurately representing significant numerical values.
        Methodology: The model is instructed to prioritize larger statistics while summarizing key findings in a coherent paragraph without bullet points.
        Data input: 'Original statistical data: {data}'
    2.2 Prompt for morphological features mining
        Objective: To analyze neuronal morphology data, particularly focusing on critical features such as 'total length' and 'number of bifurcations'.
        Methodology: The model generates a comparative summary that emphasizes the importance of these features, ensuring numerical accuracy throughout.
        Data input: 'Original neuronal morphology data: {data}'
    2.3 Prompt for projection pattern mining
        Objective: To analyze neuronal projection data, with a specific focus on axon and dendrite projections, and their implications for neuronal connectivity.
        Methodology: The model produces a summary that highlights the key points related to projection length and strength of connectivity, maintaining numerical precision and coherence.
        Data input: 'Original neuronal projection data: {data}'
(3) Report confirmation:
    The GPT-4o-2024-05-13 model serves as the final synthesis expert. This model evaluates the analysis reports generated by the three previous models against the original data and synthesizes them into a comprehensive, refined analysis report. The process follows a structured evaluation as outlined below (in the following statement, {origin_input} refers to the descriptive reports generated in step 1 and LLM summaries generated in step 2):

    3.1 Prompt for overview
        Objective: To assess the precision of three summaries relative to the original statistical dataset insights.

Methodology: The model ensures that numerical data in the summaries align with the source material. The most accurate summary is then refined into a new summary that enhances readability, brevity and consistency.
Data input: 'Original text: {origin_input}'

3.2 Prompt for morphological features mining
Objective: To meticulously assess the accuracy of three summaries in relation to an original text detailing neuronal morphology data.
Methodology: The model compares numerical values, particularly those related to 'total length', 'number of bifurcations', 'max path distance' and 'center shift', ensuring accuracy and consistency in the summaries.
Data input: 'Original text: {origin_input}'

3.3 Prompt for projection pattern mining
Objective: To evaluate the precision of summaries concerning neuronal projection characteristics, particularly focusing on axon and dendrite projections as indicators of connectivity strength.
Methodology: The model confirms numerical congruity and validates the logical consistency of comparisons in the summaries, generating a final, coherent summary.
Data input: 'Original text: {origin_input}'

MoE evaluation

The evaluation methodology is centered on assessing the accuracy and logical consistency of text summaries by comparing them against a source text. This process is implemented through a custom Python script that systematically evaluates key aspects of the summaries, particularly focusing on numerical data accuracy and logical consistency.

Data accuracy evaluation. The evaluation begins by extracting numerical data from both the source text and the generated summaries. A custom function utilizes NLP tools, such as spaCy, to identify numbers within their contextual surroundings. These extracted numbers are then compared between the source text and the summaries to determine how accurately the numerical data have been represented.

A data accuracy score is calculated by examining the occurrence and contextual integrity of each number in the summaries relative to the source text. This score reflects the proportion of correctly matched numerical values, providing a quantitative measure of how faithfully the summaries represent the original data.

Logic consistency verification. Beyond numerical accuracy, the script also evaluates the logical consistency of the summaries. This involves verifying whether the statements in the summaries logically follow from the information provided in the source text.

The script uses an LLM to perform this verification. It generates a prompt that includes both the source text and the summary statement in question, asking the model to determine whether the summary statement can be logically and numerically inferred from the source. The output from the LLM is then parsed to decide whether the summary is logically consistent. The logic accuracy score is derived by calculating the percentage of summary sentences that were deemed logically valid.

Comprehensive evaluation. The script integrates the results from both the data accuracy and logic consistency assessments to provide a comprehensive evaluation of the summaries. By quantifying the alignment of numerical data and logical coherence, the evaluation method offers a robust approach to determining the quality and reliability of text summaries in capturing the essence of the source material.

Documentation and reporting. The results of the evaluation process, including both data accuracy and logic consistency scores, are meticulously recorded. This documentation includes relevant metadata, such as the models used and the specific instances evaluated, ensuring that the evaluation process is both transparent and reproducible for further analysis and refinement.

AI-powered natural language search

Our framework integrates multiple components to achieve accurate and context-aware natural language understanding and data retrieval.

(1) Entity recognition and intent classification: The core of our NLP framework is built on a combination of machine learning models and rule-based systems. A supervised decision tree classifier, trained on a specialized dataset of neuroscience-related queries, is used to recognize key entities such as neuron types, brain regions and projection relationships. The classifier works alongside rule-based components that handle domain-specific terminology variations, ensuring a robust response to user queries.

(2) Semantic parsing and contextual understanding: The framework uses semantic parsing techniques to accurately extract and interpret user intent from natural language input. It detects complex phrases related to neuroscience, such as neuron classifications and brain region relationships. Using contextual analysis, the system discerns detailed query intents (for example, 'projection from region *X* to region *Y*'), allowing precise and relevant data to be retrieved.

(3) Dynamic mapping and knowledge integration: The framework integrates domain-specific structured schemas to map both full terminologies and their corresponding abbreviations into a standardized format compatible with database queries. This dynamic mapping ensures consistency and accuracy by aligning user input with the system's structured knowledge base. This capability enhances the system's flexibility and robustness in providing comprehensive and relevant responses.

(4) Multistage query processing pipeline: The NLP module operates through a multistage query processing pipeline, encompassing tokenization, entity extraction, context recognition and result formulation. Each stage is designed to maximize the understanding of user input and generate accurate database queries, providing users with precise and comprehensive results.

Frontend deployment and benefits. The NLP framework is deployed on the frontend using Vue.js, which brings two notable advantages:

(1) Protecting user privacy: By processing queries directly on the client side, the framework ensures that user inputs remain private and are not exposed to external servers. This approach is particularly beneficial in sensitive research settings where data privacy is paramount.

(2) Improved query speed and responsiveness: Client-side processing substantially reduces latency by eliminating unnecessary server round trips. This results in faster response times and a more interactive user experience, enabling researchers to explore neuroscience data efficiently.

Implementation and model training. The implementation leverages JavaScript-based libraries combined with tailored AI algorithms optimized for the neuroscience domain. The decision tree model is trained on a diverse set of domain-specific queries to ensure robust performance and generalization.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Atlas-mapped neuronal morphology data and discovery results—including metadata, figures and mining text—are available for direct download via the web portal (https://neuroxiv.org). In addition, we have set up an AWS server (https://download.neuroxiv.org) to facilitate easy access to standardized datasets. VTK files for various brain regions in CCFv3 are also available via GitHub at https://github.com/SEU-ALLEN-codebase/NeuroXiv/VTK. All the materials available from NeuroXiv should only be used exclusively for academic purposes and must adhere to the CC-BY NC license (https://creativecommons.org/licenses/by-nc/4.0/legalcode). Source data are provided with this paper.

Code availability

The source code of the NeuroXiv project is available via GitHub at https://github.com/SEU-ALLEN-codebase/NeuroXiv, where detailed deployment instructions are provided. User manuals and video tutorials are available at our website (https://neuroxiv.org). Data processing utilizes the plugin system provided by the Vaa3D platform (version 4.001, available via GitHub at https://github.com/Vaa3D). The source code for quality control procedures is available via GitHub at https://github.com/Vaa3D/vaa3d_tools/tree/master/hackathon/shengdian/NeuroMorphoLib. The source code for mBrainAlinger is available via GitHub at https://github.com/Vaa3D/vaa3d_tools/tree/master/hackathon/mBrainAligner. Atlas mapping can also be conducted via the mBrainAlinger web portal (http://mbrainaligner.ahu.edu.cn). All code is available under an open-source MIT license.

References

20. Li, Y., Jiang, S., Ding, L. & Liu, L. NRRS: a re-tracing strategy to refine neuron reconstruction. *Bioinform. Adv.* **3**, vbad054 (2023).
21. Guo, S., Zhao, X., Jiang, S., Ding, L. & Peng, H. Image enhancement to leverage the 3D morphological reconstruction of single-cell neurons. *Bioinformatics* **38**, 503–512 (2022).
22. Xiao, H. & Peng, H. APP2: automatic tracing of 3D neuron morphology based on hierarchical pruning of a gray-weighted image distance-tree. *Bioinformatics* **29**, 1448–1454 (2013).
23. Feng, L., Zhao, T. & Kim, J. neuTube 1.0: a new design for efficient neuron reconstruction software based on the SWC format. *eNeuro* https://doi.org/10.1523/eneuro.0049-14.2014 (2015).
24. Zhao, Z.-H., Liu, L. & Liu, Y. NIEND: neuronal image enhancement through noise disentanglement. *Bioinformatics* **40**, btae158 (2024).
25. Qu, L. et al. Cross-modal coherent registration of whole mouse brains. *Nat. Methods* **19**, 111–118 (2022).
26. Li, Y. et al. mBrainAligner-Web: a web server for cross-modal coherent registration of whole mouse brains. *Bioinformatics* **38**, 4654–4655 (2022).
27. Lorensen, W. E. & Cline, H. E. Marching cubes: a high resolution 3D surface construction algorithm. In *Seminal Graphics: Pioneering Efforts That Shaped the Field* https://doi.org/10.1145/280811.281026 (1998).
28. Vollmer, J., Mencl, R. & Müller, H. Improved Laplacian smoothing of noisy surface meshes. *Comput. Graph. Forum* **18**, 131–138 (1999).

Acknowledgements

This project was mainly supported by several grants awarded to H.P. H.P. is a New Cornerstone Investigator and a SANS Senior Investigator. The Southeast University team was also supported by a MOST (China) Brain Research Project, 'Mammalian Whole Brain Mesoscopic Stereotaxic 3D Atlas' (project numbers 2022ZD0205200 and 2022ZD0205204) awarded to L.L. We thank Brain Science Data Center, Chinese Academy of Sciences (https://braindatacenter.cn/) for the open sharing of ION datasets. We thank MouseLight project (http://ml-neuronbrowser.janelia.org) for the open sharing of its dataset. We extend our gratitude to X. Zhao for early involvement in the project's development. We also thank S. Zhao for assisting with the use of the auto-arbor tool to generate arbor data for all axon-containing neurons, and X. Jiang for testing NeuroXiv's functionalities and drafting the NeuroXiv user manual.

Author contributions

H.P. conceptualized and managed this study, and invented AIPOM. Z.Y., L.L., H.C. and J.Y. designed the initial version of the NeuroXiv (neuroxiv.net) platform, hosted originally on Tencent Cloud server. Z.Y. was responsible for backend development of the first-generation platform, while H.C. handled frontend coding. S.J. and L.W. codeveloped the new version of the NeuroXiv (neuroxiv.org) platform and migrated the servers to the AWS cloud platform. L.W. undertook the majority of website development tasks, including project deployment as well as backend and frontend development. S.J. managed the data aggregation tasks, including data standardization processing, metadata generation and the production of required atlases and reconstruction files (OBJ files) for the website. H.P. and S.J. wrote the manuscript with assistance from all authors, who reviewed and revised the manuscript.

Competing interests

H.C. and J.Y. were employed by Tencent AI Lab when this work was done. The company did not influence the research. The other authors declare no competing interests.

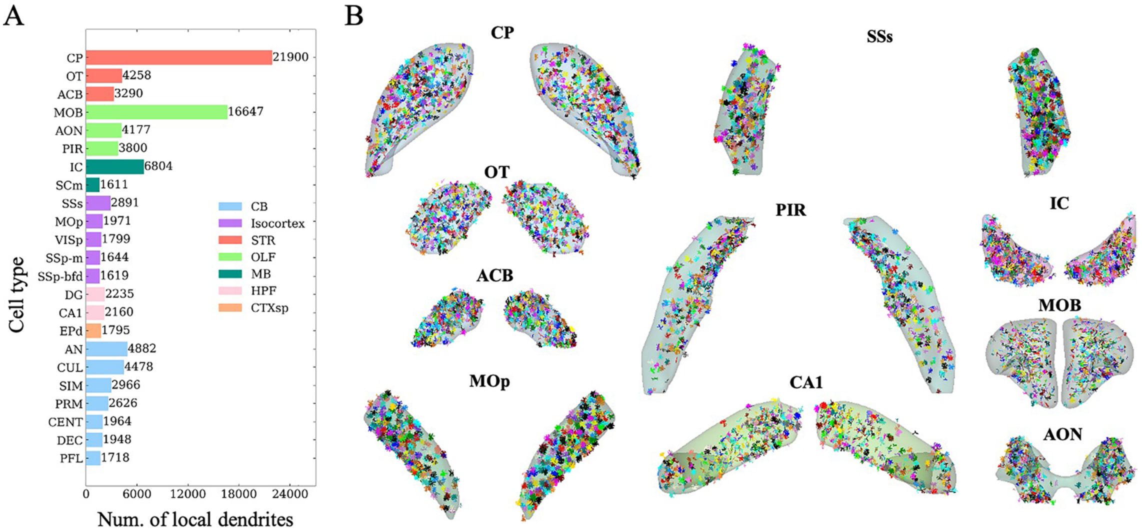
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-025-02687-2.

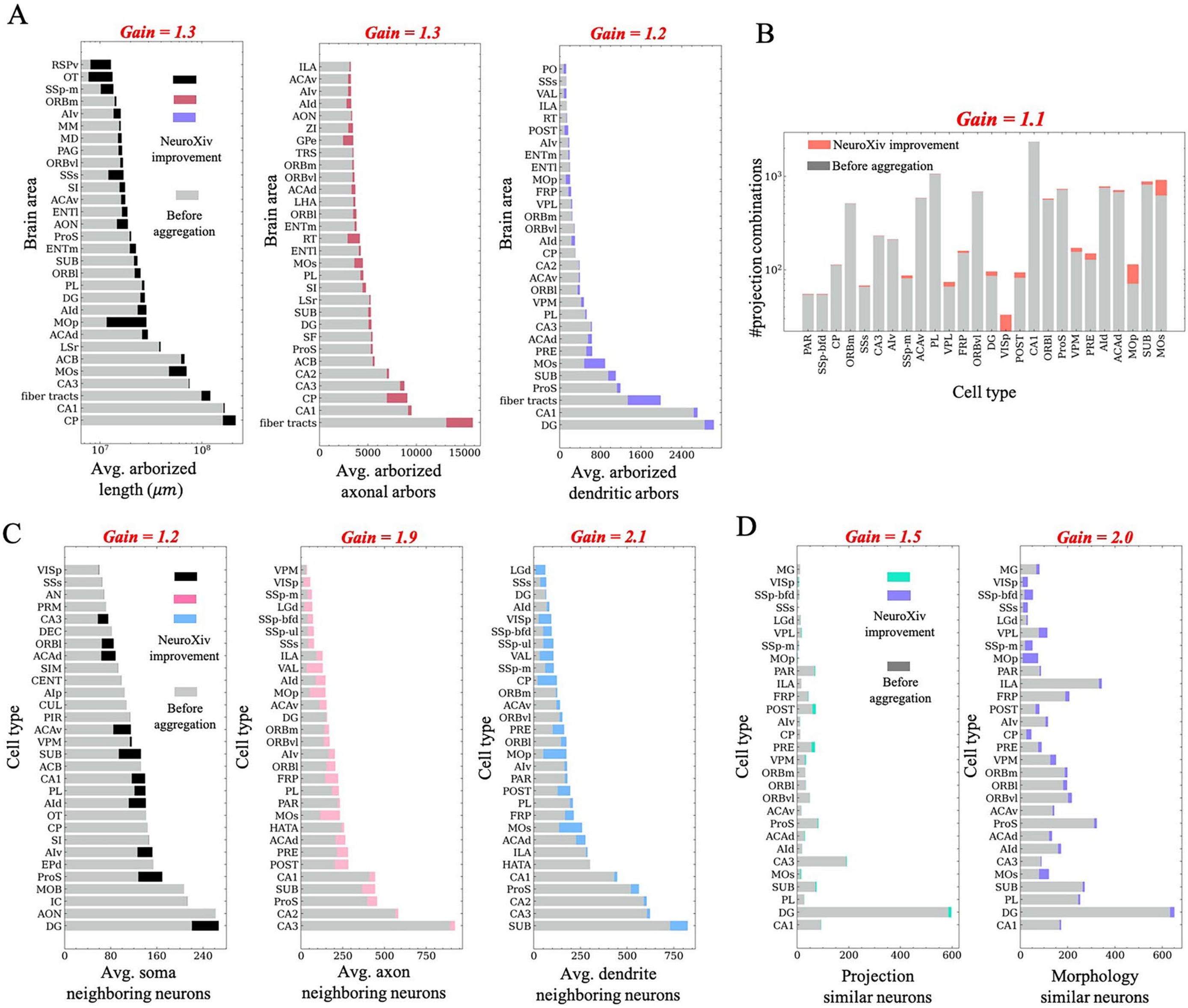
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02687-2.

Correspondence and requests for materials should be addressed to Lijuan Liu, Jianhua Yao or Hanchuan Peng.

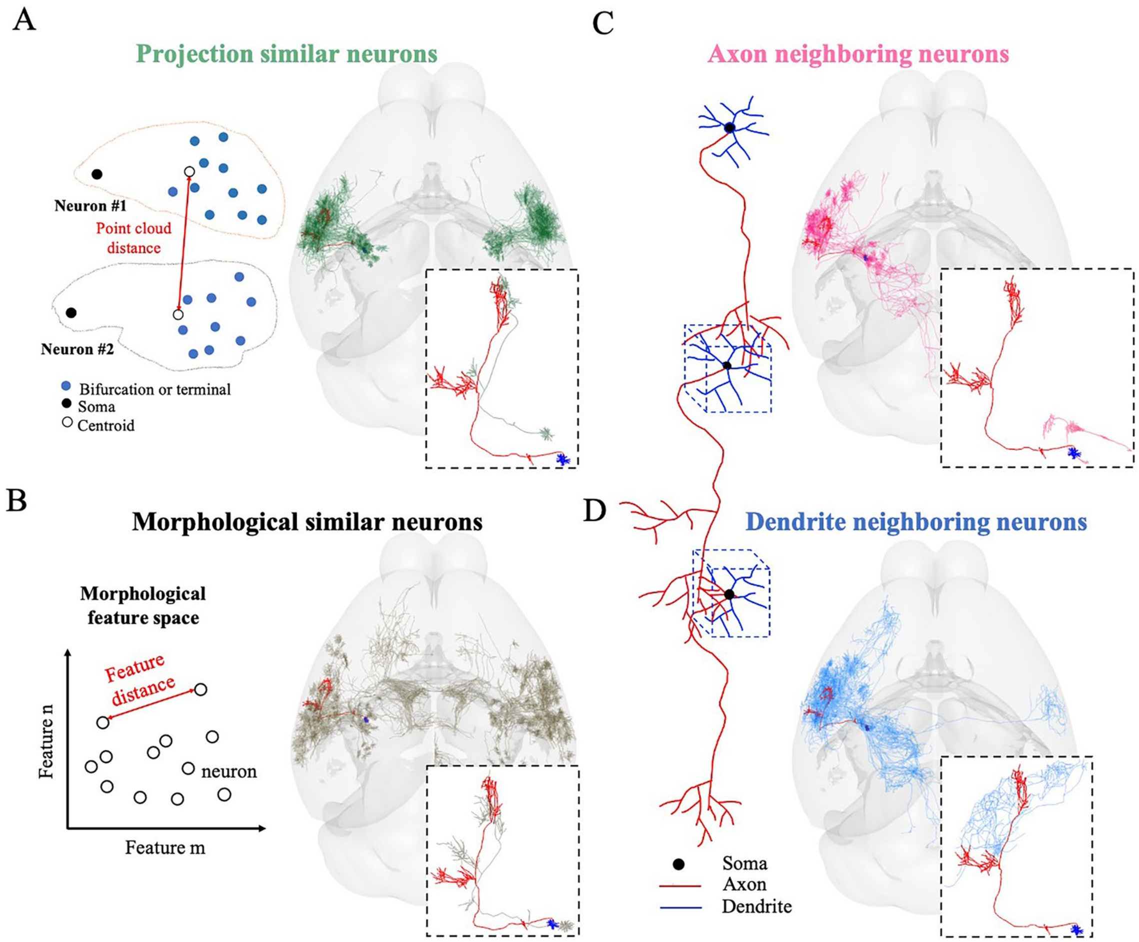
Peer review information *Nature Methods* thanks Yongsoo Kim and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Nina Vogt, in collaboration with the *Nature Methods* team.



Extended Data Fig. 1 | Local dendrites in the NeuroXiv database. a, the statistics of the number of local dendrites in different brain regions within the CCFv3 atlas. **b**, Visualization of local dendritic data across various brain regions. Local dendrites are rendered in distinct colors to enhance differentiation.

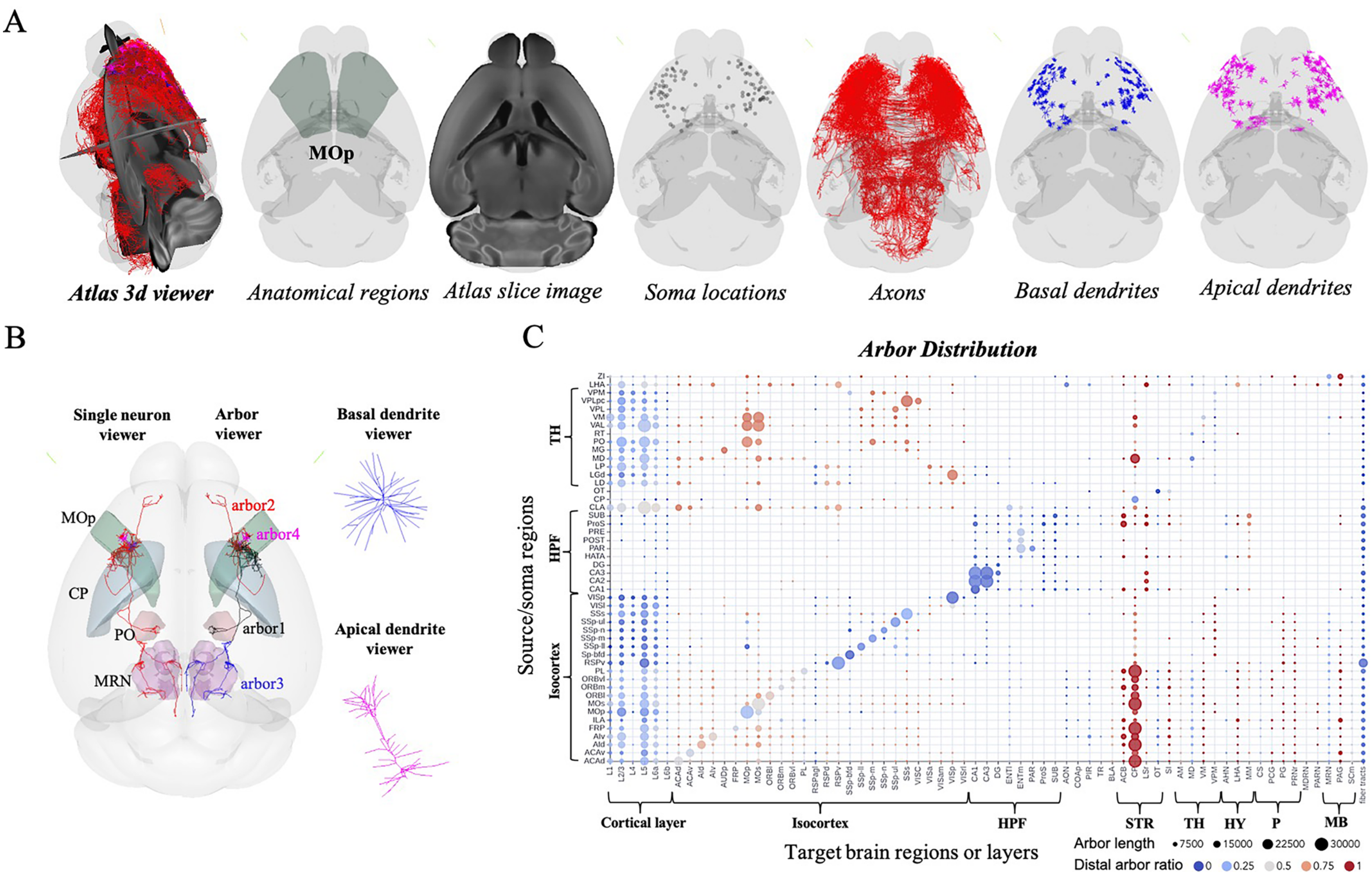


Extended Data Fig. 2 | Gains in improvements resulting from data aggregation. The NeuroXiv database was compared with three other data sources across multiple parameters. In all comparative analyses, gray bars indicate the maximum values obtainable from other individual data sources. Colored bars demonstrate improvements achieved using the NeuroXiv database. **a**, neurite arborization. Total arborized length across brain regions, with separate measurements for axonal and dendritic processes. **b**, projection combinations. Number of distinct projection patterns between brain regions, considering only neurons with processes extending >1000 µm into target regions. **c**, neuronal proximity. three categories of neighboring neurons: soma-neighboring, axon-neighboring and dendrite-neighboring. **d**, neuronal similarity. two classification types: projection-similar and morphology-similar. (See Extended Data Fig. 3 for precise definitions of neighboring and similar neuron classifications.).

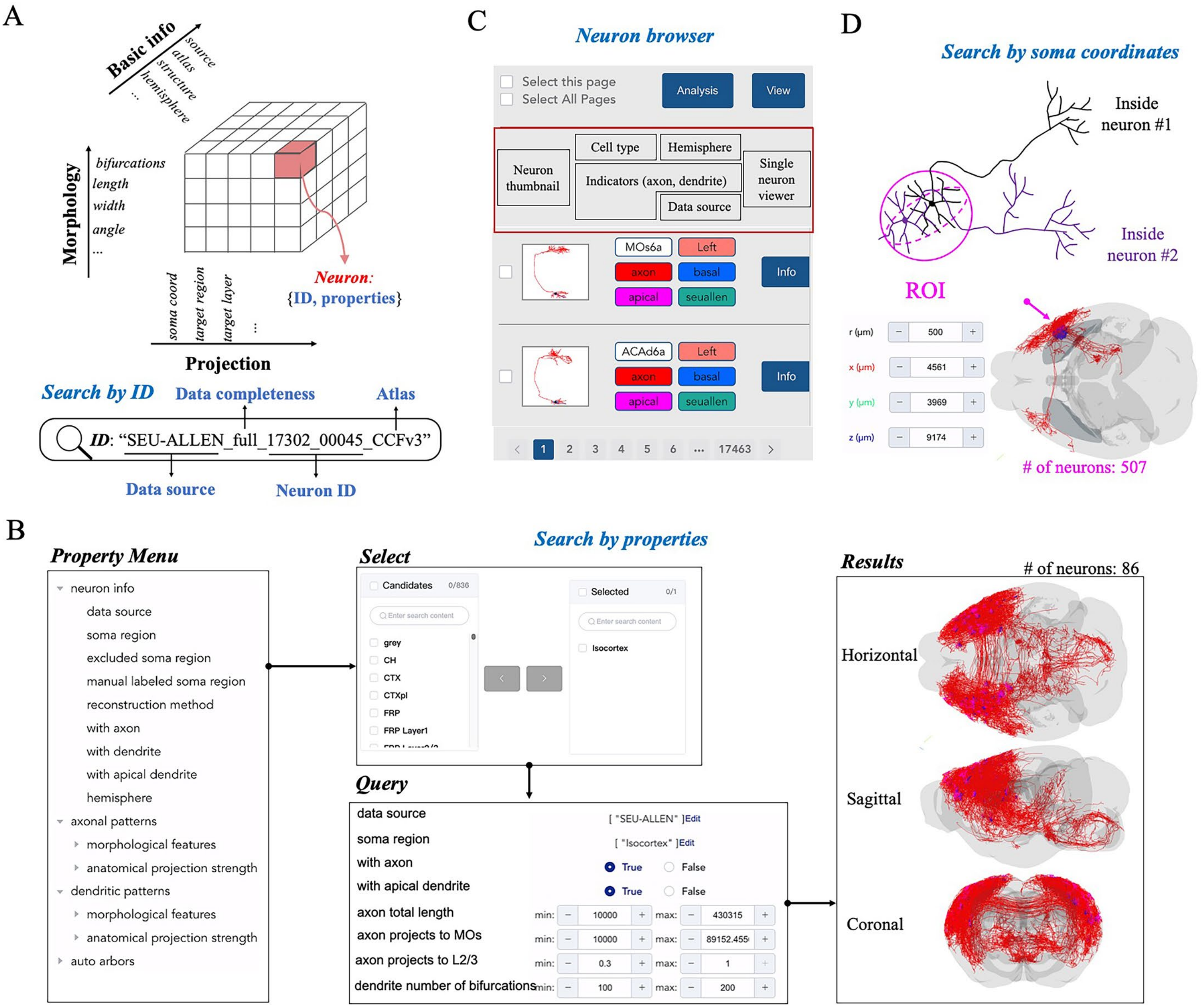


Extended Data Fig. 3 | Illustrative diagrams depicting the definition of correlated neuron data on the NeuroXiv platform. a, projection similarity neurons are defined by the distance measured between key axonal nodes of the neurons, including the soma, bifurcation points, and terminal points. **b**, morphological similarity neurons are defined by calculating the distance between neuron pairs in morphological feature space. In the database, we rank similarity based on the distances, with closer distances indicating greater similarity. **c** and **d**, Axon neighboring neurons are those where the axonal arbor of neurons in the database spatially overlaps with the dendritic arbor of the subject neuron. Conversely, dendrite neighboring neurons are those where the dendritic arbor of neurons in the database spatially overlaps with the axonal arbor of the subject neuron. In the database, we rank neighboring neurons based on the length of the overlapping arbor regions, with greater lengths indicating higher proximity.

https://doi.org/10.1038/s41592-025-02687-2

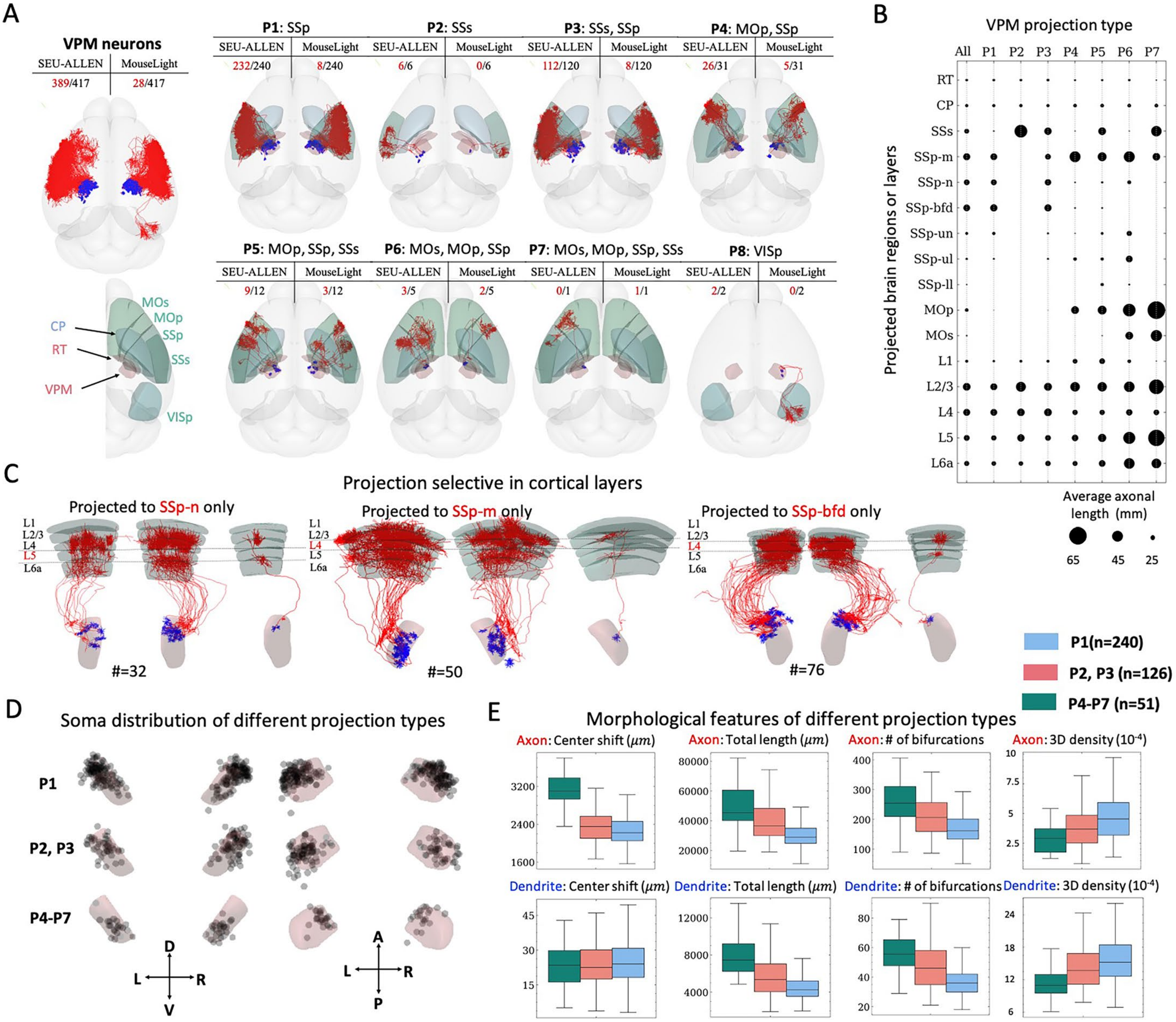


Extended Data Fig. 4 | Interactive visualization diagram of NeuroXiv. **a**, we have implemented a 3D viewer on the web platform, allowing users to visualize the atlas and neurons or neuronal structures of interest. **b**, visualization of individual neuron data, featuring an embedded arbor viewer with zoom-in views of basal and apical dendrites. **c**, arbor distribution of major cell types across various brain regions. We also visualize arbor distribution across different cortical layers.



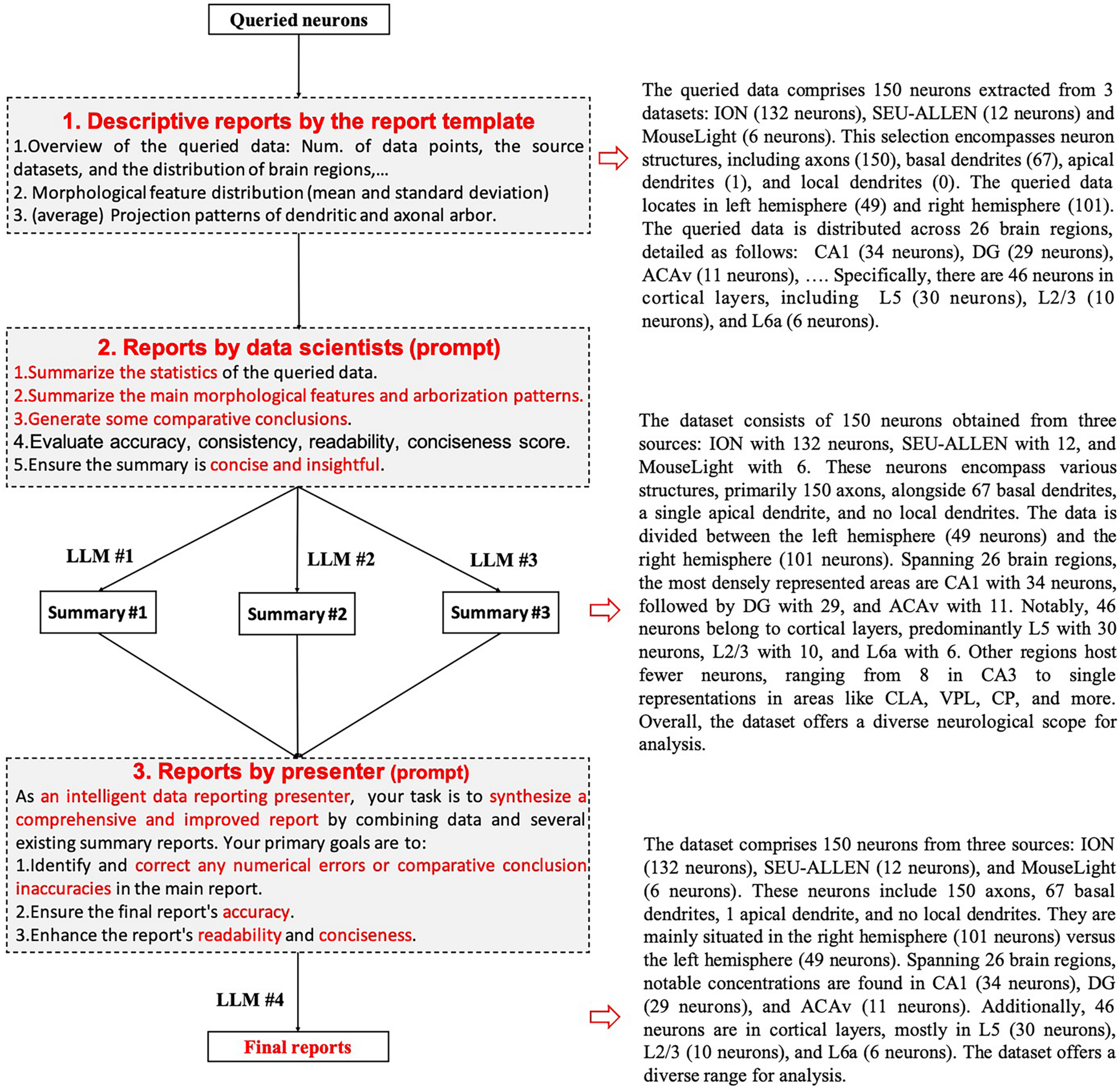
data source	["SEU-ALLEN"]Edit
soma region	["Isocortex"]Edit
with axon	● True ○ False
with apical dendrite	● True ○ False
axon total length	min: 10000 max: 430315
axon projects to MOs	min: 10000 max: 89152.455
axon projects to L2/3	min: 0.3 max: 1
dendrite number of bifurcations	min: 100 max: 200

Extended Data Fig. 5 | An illustrative diagram of data retrieval and filtering on the NeuroXiv platform. a, each neuron reconstruction in the NeuroXiv database is assigned a unique ID and includes three types of metadata: basic information, morphological features, and projection characteristics. **b**, users can customize their search strategies using either the ID or the metadata. **c**, to facilitate efficient data filtering, we have implemented a neuron browser on the web portal. This tool displays the morphology of each neuron (neuron thumbnail) and key information, and includes an entry point for navigating to detailed data pages. **d**, users can define regions of interest (ROI) and then retrieve neurons with soma located within these ROI (Supplementary Fig. 6).



Extended Data Fig. 6 | A case study of neuron projection research using NeuroXiv. **a**, this study includes 417 VPM neurons from two sources. Overall, VPM neurons project through the CP brain region, bifurcating at the boundary of the CP region to target cortical brain areas. Based on combinations of target brain regions, eight subtypes of projections are identified. Subtype P8 (characterized by neurons projecting to VISp) was excluded from subsequent analyses. **b**, differences in projection strength among various projection types in target brain regions and cortical layers are visualized, with projection strength determined by axonal length. **c**, VPM neurons exhibit projection selectivity across cortical layers. For instance, neurons projecting solely to the SSp-n region form clusters while skipping L5, whereas neurons projecting solely to the SSp-m or SSp-bfd regions form clusters while skipping L4. **d** and **e**, differences in soma distribution and morphological characteristics among various VPM projection subtypes are analyzed. In this figure, box edges in box plots show 25th and 75th percentiles, the centre line shows the 50th percentile, and bars show 1.5× the interquartile range (75th percentile − 25th percentile).

https://doi.org/10.1038/s41592-025-02687-2



1. Descriptive reports by the report template

1.Overview of the queried data: Num. of data points, the source datasets, and the distribution of brain regions,...

2. Morphological feature distribution (mean and standard deviation)

3. (average) Projection patterns of dendritic and axonal arbor.

2. Reports by data scientists (prompt)

1.Summarize the statistics of the queried data.

2.Summarize the main morphological features and arborization patterns.

3.Generate some comparative conclusions.

4.Evaluate accuracy, consistency, readability, conciseness score.

5.Ensure the summary is concise and insightful.

3. Reports by presenter (prompt)

As an intelligent data reporting presenter, your task is to synthesize a comprehensive and improved report by combining data and several existing summary reports. Your primary goals are to:

1.Identify and correct any numerical errors or comparative conclusion inaccuracies in the main report.

2.Ensure the final report's accuracy.

3.Enhance the report's readability and conciseness.

The queried data comprises 150 neurons extracted from 3 datasets: ION (132 neurons), SEU-ALLEN (12 neurons) and MouseLight (6 neurons). This selection encompasses neuron structures, including axons (150), basal dendrites (67), apical dendrites (1), and local dendrites (0). The queried data locates in left hemisphere (49) and right hemisphere (101). The queried data is distributed across 26 brain regions, detailed as follows: CA1 (34 neurons), DG (29 neurons), ACAv (11 neurons), …. Specifically, there are 46 neurons in cortical layers, including L5 (30 neurons), L2/3 (10 neurons), and L6a (6 neurons).

The dataset consists of 150 neurons obtained from three sources: ION with 132 neurons, SEU-ALLEN with 12, and MouseLight with 6. These neurons encompass various structures, primarily 150 axons, alongside 67 basal dendrites, a single apical dendrite, and no local dendrites. The data is divided between the left hemisphere (49 neurons) and the right hemisphere (101 neurons). Spanning 26 brain regions, the most densely represented areas are CA1 with 34 neurons, followed by DG with 29, and ACAv with 11. Notably, 46 neurons belong to cortical layers, predominantly L5 with 30 neurons, L2/3 with 10, and L6a with 6. Other regions host fewer neurons, ranging from 8 in CA3 to single representations in areas like CLA, VPL, CP, and more. Overall, the dataset offers a diverse neurological scope for analysis.

The dataset comprises 150 neurons from three sources: ION (132 neurons), SEU-ALLEN (12 neurons), and MouseLight (6 neurons). These neurons include 150 axons, 67 basal dendrites, 1 apical dendrite, and no local dendrites. They are mainly situated in the right hemisphere (101 neurons) versus the left hemisphere (49 neurons). Spanning 26 brain regions, notable concentrations are found in CA1 (34 neurons), DG (29 neurons), and ACAv (11 neurons). Additionally, 46 neurons are in cortical layers, mostly in L5 (30 neurons), L2/3 (10 neurons), and L6a (6 neurons). The dataset offers a diverse range for analysis.

Extended Data Fig. 7 | A schematic illustration of the mixture of experts (MoE) system. The MoE-based report generation process can be divided into three distinct stages. **1) Descriptive reports generation**: A program generates reports in a fixed format, capturing all relevant details of the retrieved data. **2) LLM Expert reports**: Multiple LLM Experts analyze and summarize the descriptive reports from a data scientist's perspective. Although three experts are shown, the process can involve one or more. **3) Report confirmation**: A different LLM Expert evaluates the previous reports for accuracy, readability, and coherence, and refines the final report accordingly. An actual case is shown on the far right, with red arrows pointing to the reports generated at each stage.

Extended Data Table 1 | Comparison of alternative neuron morphology web platforms

Name		NeuroXiv	Neurons Reunited Portal	MouseLight neuron browser	Digital Brain (ION)	NeuroMorpho
Website		neuroxiv.org	neuroinformatics.nl/HBP/neuronsreunited-viewer/	mouselight.janelia.org	mouse.digital-brain.cn	neuromorpho.org
Database	number of neurons (mouse)	177,186	9,928	1,227	16,788	137,457 (unknown completeness)
	number of cell types (anatomy)	294		126	25	72
	structural domains	soma, axon, (basal and apical) dendrite, arbor	soma, axon, dendrite	soma, axon, dendrite	soma, axon, dendrite	soma, axon, (basal and apical) dendrite
	number of data sources	3	6	1	1	13
	data format and standard	standard SWC format	standard SWC format (no quality control)	JSON file containing medata and neuron structures	SWC format file with non-standard and inconsistent structural domain identifiers	standard SWC format (no quality control)
	reference atlases	CCFv3	CCFv3	CCFv2.5, CCFv3	CCFv3	Not exist one for all neurons
Data query	neuron browser	Yes (neuron thumbnail and metadata)	Yes (id based)	Yes (id based)	Yes (id based)	Yes (id based)
	by soma region	Yes	Yes (offline)	Yes	Yes	Yes
	by spatial coordinates	Yes	Yes	Yes		
	by projected regions	Yes	Yes (offline)	Yes	Yes	
	by morphological features	Yes				Yes
	by nature language	Yes				
	by user-supplied neuron id (s)	Yes				
Data visualization	interactive visualization	Yes	Yes	Yes	Yes	2D snapshot only
Data analysis	statistics	Yes			Yes	Yes
	projection patterns	Yes			Yes	
	morphological feature distribution	Yes				measures of one neuron
Data mining	AI report	Yes				
	morphology similar neurons	Yes	Yes (offline)			
	projection similar neurons	Yes				
	spatial neighboring neurons	Yes				
Open access	no authentication	Yes	Yes	Yes		Yes
	(meta) data	Yes		Yes	morphology data only	Yes
	queried (meta) data	Yes		Yes (limited number : 20)	Yes (limited number : 500)	Yes
	discovery data	Yes			only metadata of part of figures	
	processing tools / pipelines	Yes				Yes
	processing required before data reuse		file format conversion and standardization, quality control, atlas mapping, metadata extraction			

Extended Data Table 2 | Comparison of the performance of two types of natural language query methods in AIPOM

query items		server-side method (ms)	client-side method (ms)	Gain
query neuron types	search CA1 neurons	16,599.4	1,687.2	9.8
	search DG neurons	14,286.5	1,569.9	9.1
	search PL neurons	19,531.8	1,975.4	9.9
	search MOs neurons	12,846.5	1,048.0	12.3
	search CA3 neurons	21,297.8	1,482.3	14.4
	search Ald neurons	24,380.3	3,622.0	6.7
	search SUB neurons	18,534.9	3,340.7	5.5
	search ACAd neurons	19,674.8	1,163.7	16.9
	search VPM neurons	13,016.1	1,435.2	9.1
	search CP neurons	12,664.5	2,642.9	4.8
query neurons with particular structure	search CA1 neurons with axon	24,787.2	1,630.7	15.2
	search DG neurons with axon	22,434.6	2,105.8	10.7
	search PL neurons with axon	24,884.9	1,138.6	21.9
	search MOs neurons with axon	27,150.4	1,178.4	23.0
	search CA3 neurons with axon	21,155.0	1,277.4	16.6
	search Ald neurons with axon	18,351.7	2,532.8	7.2
	search SUB neurons with axon	25,905.8	2,792.3	9.3
	search ACAd neurons with axon	17,380.2	3,053.6	5.7
	search VPM neurons with axon	37,543.3	2,326.7	16.1
	search CP neurons with axon	12,683.2	2,251.7	5.6
query neurons with specific projection patterns	search CA1 neurons projecting to ACB	15,865.6	2,568.2	6.2
	search DG neurons projecting to CA3	33,939.7	4,821.4	7.0
	search PL neurons projecting to CP	31,327.7	2,796.8	11.2
	search MOs neurons projecting to SSs	59,691.9	2,008.7	29.7
	search CA3 neurons projecting to LSr	35,424.7	2,596.1	13.6
	search Ald neurons projecting to MOs	31,140.7	2,357.7	13.2
	search SUB neurons projecting to MM	24,500.8	2,155.4	11.4
	search ACAd neurons projecting to CP	49,261.0	2,418.1	20.4
	search VPM neurons projecting to MOp	26,355.9	2,323.7	11.3
	search CP neurons projecting to SNr	31,316.8	2,203.5	14.2
Average		24,797.8	2,216.8	12.3

To achieve this, we defined three common retrieval scenarios: querying neuron types, querying neurons with particular structures, and querying neurons with specific projection patterns, each comprising 10 test cases. To ensure fair testing, both methods used the same computer configuration, eliminating disparities due to varying computational power. The server-side method involved setting up a local NeuroXiv server on the test computer, while the client-side method accessed the NeuroXiv server hosted on AWS directly. As a result, compared to the server-side approach, the client-side method demonstrated an average response time improvement of 12.3.

nature portfolio

Corresponding author(s): Hanchuan Peng

Last updated by author(s): Jan 31, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data collection primarily sourced from publicly available datasets shared across multiple websites, facilitated through downloads provided on these platforms. We described the data collection process in our manuscript. No software was used for data collection.
Data analysis	We utilized the platform (https://github.com/SEU-ALLEN-codebase/NeuroXiv) described in the manuscript for conducting data analysis, and we employed Python 3.9 to carry out several plotting tasks.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Atlas-mapped neuronal morphology data and discovery results—including metadata, figures, and mining text—are available for direct download via the web porta l(https://neuroxiv.org). Additionally, we have set up an AWS server (https://download.neuroxiv.org) to facilitate easy access to standardized datasets. VTK files for various brain regions in CCFv3 are also accessible through our GitHub repository (https://github.com/SEU-ALLEN-codebase/NeuroXiv/VTK). All the materials available from NeuroXiv should only be used exclusively for academic purposes and must adhere to the CC-BY NC license (https://creativecommons.org/licenses/by-nc/4.0/legalcode).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We established a database comprising datasets from three sources, totaling 175,149 neuron morphologies from over 500 mouse brains. These mouse brains were shared publicly by multiple laboratories. Our database integrates all publicly available data; however, due to missing metadata for some neuron data, we are unable to accurately determine the exact number of mouse brains from which all these neurons originate.
Data exclusions	No data was excluded from the analysis
Replication	All the experiments in this study utilized data that is directly accessible on the platform described in our manuscript, ensuring reproducibility. Users can access the data and metadata for each step of the experiment at any time.
Randomization	To showcase the performance of our platform, we randomly selected several groups of neurons without specifically considering their characteristics.
Blinding	The investigators were blinded to the neuron distribution from different sources during both data collection and analysis. In building the database, we collected data from various sources without excluding any specific datasets.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging