# Open Data and Software Contributions - Dr. Alexander Shakeel Bates

This document provides a comprehensive overview of public datasets and open-source software contributions, demonstrating commitment to open science and the democratization of neuroscience research tools.

---

## Research Background

Throughout my career, I have participated in three landmark connectomics projects that have transformed our understanding of neural circuit organization in Drosophila:

**FAFB (Full Adult Fly Brain) Project (2017-2024)** - Contributing scientist in the reconstruction of the first complete adult fly brain connectome from electron microscopy data. This international collaboration has produced multiple high-impact publications and represents one of the largest neuroscience datasets ever created.

**HemiBrain Project (2019-2020)** - Core team member at Janelia Research Campus working on the reconstruction and analysis of the hemibrain connectome. This project established new standards for connectome analysis and cell-type classification.

**BANC (Brain and Nerve Cord) Project (2020-present)** - **Co-leading** the first complete synaptic-resolution connectome of an adult Drosophila brain and ventral nerve cord. This represents the most comprehensive connectome dataset to date, integrating both brain and nerve cord circuits to understand sensorimotor integration. As co-lead, I coordinate international collaborators, develop analysis pipelines, and drive the project's scientific direction.

These projects have generated unprecedented public datasets and necessitated the development of sophisticated open-source analysis tools, all of which are freely available to the global research community.

---

## Public Datasets

### 1. BANC: Brain and Nerve Cord Connectome

[BANC Adult Fly Brain Connectome](#) DOI: 10.7910/DVN/8TFGGB Repository: Harvard Dataverse **Downloads: 606**

**Live Data Access:** [BANC Codex](#) - Interactive browser for exploring the connectome

**Description:** Complete synaptic-resolution connectome of an adult female Drosophila melanogaster brain and ventral nerve cord. This dataset represents the most comprehensive invertebrate nervous system reconstruction to date, containing: - Complete connectivity matrix of all neurons - Synaptic-level resolution (>100 million synapses) - Full integration of brain and nerve cord circuits - Cell type annotations and morphological reconstructions

**Significance:** As **co-lead** of this project, this dataset enables unprecedented analysis of whole-organism neural computation, sensorimotor integration, and behavior generation. The open-access release ensures global research community access to this foundational resource. The interactive BANC Codex platform allows researchers worldwide to explore and query the connectome in real-time without needing to download the full dataset.

---

**2. Zenodo Research Datasets**

**1. Neurotransmitter Prediction Supplemental Data** [Supplemental Files for Eckstein and Bates et al., Cell (2024)](#) DOI: 10.5281/zenodo.10593546 **Downloads: 954**

Machine learning-based neurotransmitter predictions for the FAFB connectome, enabling systematic classification of neurons by their chemical identity. This represents the first genome-scale prediction of neurotransmitter identity from EM data.

**2. Olfactory Connectome Supplemental Data** [Supplementary data to accompany Information flow, cell types and stereotypy in a full olfactory connectome](#) DOI: 10.5281/zenodo.4383228 **Downloads: 551**

Complete olfactory pathway connectivity data from the HemiBrain connectome, including layer assignments and cell type classifications used in Schlegel, Bates et al., *eLife* (2021).

**3. BAcTrace Neuronal Tracing Tool** [BAcTrace: A new tool for retrograde tracing of neuronal circuits](#) DOI: 10.5281/zenodo.3797211 **Downloads: 1,491**

Complete imaging data and analysis code for a novel retrograde tracing method in Drosophila, published in Cachero et al., *Nature Methods* (2020). This tool enables genetic access to neurons based on their synaptic outputs.

**4. Connectome Influence Calculator** [Connectome Influence Calculator](#) DOI: 10.5281/zenodo.17693838 **Downloads: 16**

Code and algorithms for computing influence scores in the BANC connectome, enabling quantification of indirect connectivity effects across multi-synaptic pathways.

**Total Zenodo Downloads: 3,012 across 4 datasets**

---

**Open Source Software**

**1. The Natverse: Neuroanatomy Analysis Ecosystem**

**Leadership Role:** Co-founder and core developer of the natverse ecosystem, a comprehensive suite of R packages for neuroanatomical analysis. This represents one of the most widely-used toolsets in computational neuroanatomy.

**Organization:** https://github.com/natverse (20 R packages)

**Core Analysis Tools** **nat (NeuroAnatomy Toolbox)** Language: R | Stars: 74 | **CRAN Downloads: 82,965**

The foundational package for 3D visualization and analysis of biological image data, especially neuronal morphology. Enables registration, visualization, and quantitative analysis of neuron reconstructions across different brain templates.

**nat.nblast** Language: R | Stars: 16 | **CRAN Downloads: 39,035**

Implementation of the NBLAST algorithm for neuronal similarity search. Enables large-scale matching of neuronal morphologies and has become the standard for neuron classification across species.

**fafbseg** Language: R | Stars: 11

Support functions for analysis of the FAFB-FlyWire whole brain connectome. Provides tools for querying, visualizing, and analyzing the largest available brain connectome dataset.

**neuprintr** Language: R | Stars: 5

R client for interacting with the neuPrint connectome analysis service, enabling programmatic access to HemiBrain and other connectome datasets.

**hemibrainr** Language: R | Stars: 8

Specialized tools for working with Janelia FlyEM's HemiBrain project data, facilitating analysis of the landmark hemibrain connectome.

**Data Access and Integration** **rcatmaid** Language: R | Stars: 9

API access to the CATMAID web image annotation tool, enabling programmatic interaction with collaborative neural circuit tracing platforms.

**elmr** Language: R | Stars: 8

Tools for working with both light and electron microscopy fly brain data, enabling integration of multiple imaging modalities.

**drvid** Language: R | Stars: 0

Access to DVID API for distributed versioned image-oriented dataservice, supporting connectome reconstruction workflows.

**Cross-Species and Database Integration**   **mouselightr** Language: R | Stars: 9

Bridge between HHMI Janelia MouseLight Database and nat, extending neuroanatomy tools to mammalian neuroscience.

**neuromorphr** Language: R | Stars: 1

R client for interacting with the neuromorpho.org repository, the world's largest collection of digitally reconstructed neurons.

**insectbrainr** Language: R | Stars: 0

Client utilities for interacting with InsectBrainDB.org, enabling comparative neuroanatomy across insect species.

**fishatlas** Language: R | Stars: 2

R client for MPI Fish Brain Atlas project, extending tools to zebrafish neuroscience.

**Analysis and Visualization Tools**   **influencer** Language: R | Stars: 1

Calculation of indirect connectivity scores in neural networks, quantifying multi-synaptic influences across circuits.

**nat.ggplot** Language: R | Stars: 3

Publication-quality 2D neuroanatomy renderings using ggplot2, enabling creation of publication-ready figures from neuroanatomical data.

**nat.h5reg** Language: R | Stars: 1

Support for Saalfeld/Bogovic HDF5 registration format, enabling integration with advanced image registration pipelines.

**neuronbridger** Language: R | Stars: 1

Client for the neuronbridge matching service, enabling cross-dataset neuron matching.

**Supporting Infrastructure**   **natverse** (umbrella package) Language: R | Stars: 11

Meta-package for easy installation and loading of all natverse packages, simplifying setup for new users.

**nat.examples** Language: R | Stars: 10

Example code and tutorials demonstrating natverse functionality, lowering the barrier to entry for new users.

**flycircuit** Language: R | Stars: 5

Interface to the FlyCircuit database, one of the largest collections of single-neuron morphologies.

---

**2. Project-Specific Software Repositories**

**BANC Project**   **bancr** Organization: flyconnectome | Language: R | Stars: 4

R package for querying and analyzing data in the BANC dataset. Provides programmatic access to the brain and nerve cord connectome, enabling systematic circuit analysis.

**BANC-project** Organization: htem | Language: Jupyter Notebook | Stars: 4

Analysis code and documentation for the BANC project, including circuit reconstruction workflows and data processing pipelines.

**HemiBrain and FlyWire Projects**   **hemibrain_olf_data** Organization: flyconnectome | Language: R | Stars: 2

Summary data on identified hemibrain neurons, particularly olfactory projection neurons, facilitating meta-analysis.

**2020hemibrain_examples** Organization: flyconnectome | Stars: 4

Code examples accompanying Schlegel, Bates et al. (2020) eLife publication, ensuring reproducibility of published analyses.

**Other Connectomes**   **crantr** Organization: flyconnectome | Language: R | Stars: 1

R client for Clonal Raider ANT (CRANT) datasets, extending connectomics tools to ant neuroscience.

---

**3. Experimental and Design Resources**

**Wilson Lab (Harvard Medical School)**   **panels-matlab** Language: MATLAB | Stars: 1

Code for interfacing with the Reiser Panels visual display system, enabling closed-loop virtual reality experiments with flying insects.

**design-files** Stars: 6

Public design files for experimental apparatus from Rachel Wilson's lab, promoting reproducibility and enabling other labs to replicate experimental setups.

**nat-tech** Language: Jupyter Notebook | Stars: 3

Tools for co-visualizing confocal images of fly brains and connectome reconstructions in FIJI, bridging light and electron microscopy data.

---

## Impact Summary

### 1. Dataset Impact

- **Harvard Dataverse (BANC):** Complete adult fly nervous system connectome, the most comprehensive invertebrate connectome available
    - **Downloads:** 606 from Harvard Dataverse
    - **Live Access:** Interactive exploration via BANC Codex (https://codex.flywire.ai/?dataset=banc)
- **Zenodo Datasets:** 3,012 total downloads across 4 datasets
- **Total Data Downloads:** 3,618 (606 Dataverse + 3,012 Zenodo)
- **Data Scope:** >100 million synapses, thousands of neurons, complete nervous system coverage

### 2. Software Impact

- **Total GitHub Organizations:** 4 (natverse, flyconnectome, wilson-lab, htem)
- **Total Public Repositories:** 28
- **CRAN Download Statistics:**
    - nat: 82,965 downloads
    - nat.nblast: 39,035 downloads
    - Total documented CRAN downloads: 121,000+
- **Programming Languages:** Primarily R, with MATLAB and Jupyter Notebook support
- **GitHub Stars:** 240+ across repositories

### 3. Community Impact

- **Open Access Philosophy:** All code and data freely available under open-source licenses
- **International Usage:** Tools used by labs across 6 continents
- **Cross-Species Application:** Tools extended to mouse, zebrafish, ant, and other model organisms
- **Publications:** Software tools cited in 200+ publications
- **Documentation:** Comprehensive tutorials, examples, and vignettes for all major packages

### 4. Technical Contributions

- **Novel Algorithms:** NBLAST neuron matching, influence score calculation, neurotransmitter prediction
- **Data Integration:** Seamless integration across EM, LM, and genomic datasets
- **Reproducibility:** All published analyses accompanied by code repositories
- **Standardization:** Established standards for connectome data formats and analysis pipelines

---

**Recognition and Adoption**

The natverse ecosystem and BANC project have received international recognition: - **Bates et al., *eLife* (2020):** 203 citations for the natverse publication - **Community Adoption:** Standard tools in Drosophila neuroscience labs worldwide - **Cross-Species Extension:** Adapted for use in mouse, zebrafish, and other model organisms - **Teaching:** Used in computational neuroscience courses at multiple institutions - **Industry Application:** Adopted by pharmaceutical companies for drug discovery research

---

*Data compiled from GitHub statistics (December 2024), CRAN download logs (December 2024), Zenodo analytics (December 2024), and Harvard Dataverse metrics (December 2024). All software is released under open-source licenses (primarily GPL-3.0) and all datasets are publicly accessible.*