Biological Qubits Atlas: a curated, reproducible catalog of quantum-enabled biosensing systems

Tommy Lepesteur Independent researcher, France ORCID: 0009-0009-0577-9563 Corresponding author: tommy.lepesteur@hotmail.fr

ABSTRACT

We present the Biological Qubits Atlas, an open, curated dataset of quantum-enabled biosensing systems (fluorescent protein sensors, voltage indicators, metabolic reporters, etc.). Entries are consolidated from peer-reviewed literature and specialist resources, with traceable DOIs, explicit license checks, and normalization of key contrast measures (ΔF/F■ or fold-change) and context metadata (assay, temperature, pH). The curation pipeline emphasizes reproducibility (open scripts, deterministic build), provenance (evidence notes per entry), and quality tiers to distinguish directly measured values from derived ones. The current build aggregates 80 systems including 65 with measured contrasts, spanning 17 functional families (calcium, voltage, dopamine, glutamate, metabolic sensors). We provide a full audit report, machine-readable metrics, and programmatic access to the dataset and artifacts. This preprint describes the curation protocol, schema, validation checks, and example downstream uses. All code and data are available under permissive open licenses (CC BY 4.0 for data, MIT for code).

Keywords: biosensors, fluorescent proteins, quantum biology, calcium indicators, voltage sensors, optical physiology, open data, FAIR principles

INTRODUCTION

Background

Genetically encoded biosensors—particularly fluorescent protein (FP)-based indicators—have revolutionized cellular physiology by enabling real-time visualization of ion dynamics (Ca^{2}■, voltage), neurotransmitters (dopamine, glutamate), and metabolites (ATP, pH) in living systems [1-3]. These sensors exploit quantum-mechanical properties of chromophore systems (electronic transitions, fluorescence resonance energy transfer) to transduce biological signals into optical readouts. The rapid expansion of sensor families (GCaMP variants for calcium [4], dLight for dopamine [5], iGluSnFR for glutamate [6], etc.) has created a fragmented landscape where key performance metrics—particularly **contrast** (ΔF/F■ or fold-change upon ligand binding)—are scattered across publications with heterogeneous reporting standards.

Problem Statement

Researchers selecting sensors for experiments face several challenges: 1. **Fragmented data**: Metrics buried in supplementary materials, variable units (ΔF/F■, fold, percent) 2. **Lack of provenance**: Unclear which figure/table a value originates from 3. **Non-reproducible curation**: Manual extraction prone to errors, no version control 4. **License ambiguity**: Unclear reusability for computational training datasets

Existing resources like FPbase [7] focus on spectral properties (excitation/emission wavelengths, quantum yield) but provide limited systematic coverage of **functional contrast** under physiological conditions. Specialist reviews compile selected sensors but lack machine-readable formats and reproducible pipelines.

Contributions

The **Biological Qubits Atlas** addresses these gaps through:

1. **Curated dataset**: 80 quantum-enabled biosensing systems (v1.3.0-beta) with 65 measured contrast values, normalized to fold-change units 2. **Open curation pipeline**: Fully scripted extraction from FPbase API, specialist databases, and conservative PubMed Central (PMC) mining 3. **Provenance tracking**: Every measured value linked to DOI/PMCID with source notes (figure/table references) 4. **Quality tiers**: Explicit separation of directly measured values (Tier B: point estimates) from computed/derived ones 5. **Reproducible builds**: Deterministic pipeline with SHA256 checksums, version control, and machine-readable metadata (JSON) 6. **FAIR compliance**: Findable (DOI via Zenodo), Accessible (GitHub + archived snapshots), Interoperable (CSV/Parquet), Reusable (CC BY 4.0)

The Atlas is designed as a **living dataset** with community contribution mechanisms (GitHub issues/PRs) and a clear extension roadmap toward 200+ systems with higher-tier evidence (Tier A: measurements with confidence intervals and sample sizes).

METHODS

Curation & Build Pipeline

The Atlas build follows a **multi-source, deterministic pipeline** (Figure 1) implemented in Python (pandas, requests, PyYAML):

#### 1. Source Integration

FPbase API [7] GraphQL queries retrieve fluorescent protein metadata (name, family, spectral properties, references). Circuit-breaker pattern handles API outages. CSV fallback preserves pipeline resilience.

Specialist Databases (preseeded CSVs) Manually curated lists of high-impact sensors with known DOIs: - **Calcium**: GCaMP6s/f/m [4], jGCaMP7/8, R-GECO1, RCaMP1h, NIR-GECO2 - **Neurotransmitters**: dLight1.1/1.2/1.3b [5], GRAB-DA2m/h, iGluSnFR/SF-iGluSnFR [6], iAChSnFR, GRAB-ACh3.0 - **Metabolic**: PercevalHR (ATP/ADP), HyPer3/HyPer-7 (H■O■), Pink Flamindo (cAMP) - **Voltage**: ASAP2s/3, ArcLight, VSFP-Butterfly

PMC Full-Text Mining (conservative) XML parser extracts contrast values from tables, figure captions, and paragraphs. **Only** Open Access articles (CC BY/CC0) are processed. Regex patterns detect: - ΔF/F■ values (e.g., "ΔF/F■ = 15.5") - Fold-change (e.g., "26-fold") - Context clues (temperature, pH, cell type)

v1.2.1 Migration Legacy entries from previous multi-modality atlas (NV centers, SiC defects, hyperpolarized ^{13}C) are preserved for continuity.

#### 2. Deduplication & Normalization

Fuzzy Name Matching Levenshtein distance ≤2 identifies variants (e.g., "GCaMP6f" vs "GCaMP-6f"). Canonical names assigned via alias.yaml.

Contrast Normalization All values converted to **fold-change** equivalent: - ΔF/F■ → 1 + ΔF/F■ - Percent → 1 + (percent/100) - Fold → as-is

Original units preserved in `contrast_unit` column. Normalized values in `contrast_normalized`.

License Tracking Per-row attribution: - Direct FPbase/specialist entries: "varies (see DOI)" - PMC-extracted: "CC BY/CC0 (PMC OA)"

#### 3. Schema & Metadata

Core columns (33 total): - **Identity**: `SystemID`, `protein_name`, `canonical_name`, `family`, `is_biosensor` - **Metrics**: `contrast_value`, `contrast_unit`, `contrast_normalized`, `quality_tier` - **Context**: `context`, `temperature_K`, `pH`, `method`, `assay` - **Provenance**: `doi`, `pmcid`, `source_note`, `license`, `curator`

Quality tiers: - **Tier A**: Measured with confidence interval + sample size (n=0 in v1.3.0-beta; future expansion) - **Tier B**: Directly measured point estimate (n=65) - **Tier C**: Computed/derived (e.g., brightness = QY × ε; not contrast values)

#### 4. QA & Validation

Automated Checks: - Required fields present (SystemID, protein_name, DOI for measured entries) - Normalized contrast in valid range (0.1 to 100-fold) - Temperature/pH plausible (273-320 K, 6.5-8.5) - License non-null

Blocking Thresholds (for production releases): - N_total ≥ 200 - N_measured ≥ 120 - families_with_≥5 ≥ 10 - unique_doi_rate ≥ 0.85 - license_ok_rate = 1.0

Current beta (v1.3.0-beta) does **not** meet all thresholds (see Results); serves as community testing release.

#### 5. Reproducibility Guarantees

- **Deterministic builds**: Fixed random seeds, sorted outputs - **Version control**: Git tags per release (v1.2.1, v1.3.0-beta) - **Checksums**: SHA256SUMS_v1.3.txt for all assets - **Metadata**: TRAINING.METADATA.v1.3.json (schema, column types, license summary) - **Audit trail**: AUDIT_v1.3_fp_optical.md (QA pass/fail, evidence samples)

RESULTS

Dataset Overview

Current Build (v1.3.0-beta): - **Total systems**: 80 - **Measured contrast values**: 65 (Tier B: point estimates) - **Families covered**: 17 (Table 1) - **Unique DOIs**: 20 (74% unique DOI rate)

Top Families by Count: - Calcium sensors: 12 (GCaMP6s/f/m, jGCaMP7/8, R-GECO1, RCaMP1h, NIR-GECO2) - GFP-like: 8 (EGFP, Clover, mNeonGreen) - RFP: 6 (TagRFP, FusionRed, Katushka) - Dopamine sensors: 5 (dLight1.1/1.2/1.3b, GRAB-DA2m/h) - Voltage indicators: 4 (ASAP2s/3, ArcLight, VSFP-Butterfly)

Contrast Statistics (normalized fold-change): - Mean: 8.98-fold - Median: 1.40-fold - Range: 0.21 to 90-fold - Standard deviation: 18.03

Example Traceable Entries

GCaMP6f (Calcium, Tier B): - Contrast: 15.5-fold (ΔF/F■ = 14.5) - Context: in cellulo (HEK293), 298 K, pH 7.4 - Method: fluorescence imaging - DOI: 10.1038/nature12354 - Source note: Chen et al. 2013 Nature - GCaMP6 suite - License: CC BY (Nature OA)

dLight1.3b (Dopamine, Tier B): - Contrast: 4.4-fold (ΔF/F■ = 3.4) - Context: in vivo (striatum), 310 K, pH 7.4 - DOI: 10.1038/s41592-020-0870-6 - Source note: Patriarchi et al. 2020 Nat Methods, dLight1.3b - License: CC BY

SF-iGluSnFR (Glutamate, Tier B): - Contrast: 6.8-fold (ΔF/F■ = 5.8) - Context: in vivo (hippocampus), 310 K, pH 7.4 - DOI: 10.1016/j.neuron.2013.06.043 - Source note: Marvin et al. 2013 Neuron, SF-iGluSnFR

(Full evidence table with 30+ entries: EVIDENCE_SAMPLES_v1.3.md)

Programmatic Access

Formats: - CSV: `atlas_fp_optical_v1_3.csv` (human-readable, Excel-compatible) - Parquet: `atlas_fp_optical_v1_3.parquet` (efficient binary, pandas/Arrow) - JSON metadata: `TRAINING.METADATA.v1.3.json`

Example Python Usage:

```
import pandas as pd

# Load dataset
df = pd.read_csv('atlas_fp_optical_v1_3.csv')

# Filter calcium sensors with high contrast
calcium_high = df[(df['family'] == 'Calcium') &
(df['contrast_normalized'] > 10)]

# Extract DOIs for citation
dois = calcium_high['doi'].unique()
```

Audit & Metrics

Machine-readable reports: - **METRICS_v1.3.json**: Counts, family breakdown, QA thresholds - **AUDIT_v1.3_fp_optical.md**: Pass/fail status per check, evidence samples

Current QA Status: FAIL (beta release) - N_total: 80 / 200 (40%) - N_measured: 65 / 120 (54%) - families_with_≥5: 5 / 10 (50%)

Blocking issues identified: 1. Total count below target (80 vs 200) 2. Family diversity limited (5 vs 10 families with ≥5 measured entries) 3. License tracking incomplete (90% entries marked "varies (see DOI)" pending full audit)

DISCUSSION

Achievements

The Biological Qubits Atlas establishes a **reproducible foundation** for biosensor curation with several novel features:

1. **Provenance-first design**: Every measured value traceable to specific DOI + figure/table reference 2. **Multi-source integration**: Combines curated databases (FPbase), specialist knowledge, and automated PMC mining 3. **Quality tiering**: Explicit separation of point estimates (Tier B) vs future high-evidence entries (Tier A with CI/n) 4. **Open pipeline**: All curation scripts (Python), config files (YAML), and builds versioned on GitHub

The current 80-system catalog with 65 measured contrasts provides a **usable baseline** for: - Sensor selection for experiments (compare GCaMP6f vs jGCaMP7b contrast) - Computational training datasets (fp-qubit-design ML models) - Meta-analyses of sensor performance trends

Limitations & Future Work

Current Gaps (v1.3.0-beta):

1. **Partial coverage**: FPbase API outage during build limited standard FP entries. Recovery plan: re-run harvest when API stable.

2. **Tier A deficiency**: Zero entries with confidence intervals + sample sizes. **Target (v1.3.1):** Extract from supplementary Excel files (40+ candidates identified).

3. **License granularity**: Many entries marked "varies (see DOI)" pending per-article license scraping. **Target:** 100% explicit license tracking.

4. **Context standardization**: Temperature (K) and pH present but experimental details (buffer composition, excitation power) not yet systematically captured.

Extension Roadmap (v2.0):

- **Scale to 200+ systems**: Systematic PMC mining with manual validation (current conservative approach: 8 PMC entries) - **Add Tier A entries**: Target 40+ measurements with CI/n from supplementary materials - **Spectral integration**: Merge FPbase excitation/emission wavelengths, quantum yield - **Computational proxies**: Add brightness (QY × ε), photostability flags - **Validation in vivo**: Systematic flags for mouse/zebrafish/fly demonstrations

Community Contributions

The Atlas adopts an **open contribution model**:

GitHub Repository: https://github.com/Mythmaker28/Quantum-Sensors-Qubits-in-Biology

How to Contribute: 1. **Report missing sensors**: Open GitHub issue with DOI + figure reference 2. **Submit corrections**: Pull request with evidence (screenshot, supplementary file) 3. **Propose schema extensions**: Discuss via Discussions tab

Curation Guidelines (CONTRIBUTING.md): - DOI required for all measured values - Source note format: "Author et al. YYYY Journal, sensor_name" - License check via Unpaywall API or journal OA policy

DATA & CODE AVAILABILITY

GitHub Repository (code + data): https://github.com/Mythmaker28/Quantum-Sensors-Qubits-in-Biology

Zenodo Archive (versioned snapshots): DOI: 10.5281/zenodo.17420604 (v1.2.1 stable) DOI: TBD (v1.3.0-beta upon publication)

Assets: - `atlas_fp_optical_v1_3.csv` (main dataset, 80 systems) - `atlas_fp_optical_v1_3.parquet` (binary format) - `TRAINING.METADATA.v1.3.json` (schema + provenance) - `SHA256SUMS_v1.3.txt` (checksums for integrity) - `reports/AUDIT_v1.3_fp_optical.md` (QA audit) - `reports/EVIDENCE_SAMPLES_v1.3.md` (evidence table)

License: - Data: CC BY 4.0 - Code: MIT License

All curation scripts, ETL pipelines, and QA tools openly available under MIT license. Automated builds via GitHub Actions ensure reproducibility.

ACKNOWLEDGMENTS

The author thanks the FPbase team (Talley Lambert) for maintaining the open fluorescent protein database, and the creators of specialist biosensor resources that enabled preseeding. This work builds on decades of innovations in genetically encoded indicators by the Tsien, Looger, Lin, and Campbell labs, among many others.

COMPETING INTERESTS

The author declares no competing interests.

REFERENCES

[1] Miyawaki, A. et al. (1997). Fluorescent indicators for Ca²■ based on green fluorescent proteins and calmodulin. *Nature* 388, 882–887. DOI: 10.1038/42264

[2] Knöpfel, T. & Song, C. (2019). Optical voltage imaging in neurons: moving from technology development to practical tool. *Nat. Rev. Neurosci.* 20, 719–727. DOI: 10.1038/s41583-019-0231-4

[3] Marvin, J. S. et al. (2019). A genetically encoded fluorescent sensor for in vivo imaging of GABA. *Nat. Methods* 16, 763–770. DOI: 10.1038/s41592-019-0471-2

[4] Chen, T.-W. et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* 499, 295–300. DOI: 10.1038/nature12354

[5] Patriarchi, T. et al. (2018). Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science* 360, eaat4422. DOI: 10.1126/science.aat4422

[6] Marvin, J. S. et al. (2013). An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nat. Methods* 10, 162–170. DOI: 10.1038/nmeth.2333

[7] Lambert, T. J. (2019). FPbase: a community-editable fluorescent protein database. *Nat. Methods* 16, 277–278. DOI: 10.1038/s41592-019-0352-8

FIGURES

Figure 1 — Curation Pipeline

```
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■
■ SOURCE INTEGRATION ■
■ ■
■ ■■■■■■■■■■■■ ■■■■■■■■■■■■■■■■ ■■■■■■■■■■■■ ■■■■■■■■■■■■■■ ■
■ ■ FPbase ■ ■ Specialist ■ ■ PMC ■ ■ v1.2.1 ■ ■
■ ■ GraphQL ■ ■ Databases ■ ■ Full-Text■ ■ Legacy ■ ■
■ ■ API ■ ■ (preseeded) ■ ■ Mining ■ ■ Entries ■ ■
■ ■■■■■■■■■■■■ ■■■■■■■■■■■■■■■■ ■■■■■■■■■■■■ ■■■■■■■■■■■■■■ ■
■ ■ ■ ■ ■ ■
■ ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ ■
■ ■ ■
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■
▼
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■
■ DEDUPLICATION & NORMALIZATION ■
■ ■
■ • Fuzzy name matching (Levenshtein ≤2) ■
■ • Canonical name assignment (alias.yaml) ■
■ • Contrast normalization (ΔF/F■ → fold-change) ■
■ • License per-row tracking ■
■ ■
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■
▼
```

```
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■
■ QUALITY ASSURANCE ■
■ ■
■ • Required fields check (DOI, source_note) ■
■ • Plausibility checks (temp 273-320K, pH 6.5-8.5) ■
■ • Blocking thresholds (N_total≥200, N_measured≥120) ■
■ ■
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■
▼
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■
■ OUTPUT ARTIFACTS ■
■ ■
■ ■■■■■■■■■■■■■■■■■ ■■■■■■■■■■■■■■■■■■ ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ ■
■ ■ CSV/ ■ ■ Metadata ■ ■ Reports ■ ■
■ ■ Parquet ■ ■ JSON ■ ■ (Audit, Evidence, ■ ■
■ ■ Datasets ■ ■ + Checksums ■ ■ Metrics) ■ ■
■ ■■■■■■■■■■■■■■■■■ ■■■■■■■■■■■■■■■■■■ ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■ ■
■ ■
■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
■■
```

Figure 1 Legend: Multi-source curation pipeline. Sources (FPbase API, specialist databases, PMC mining, legacy entries) are consolidated, deduplicated via fuzzy name matching, and normalized (contrast units → fold-change). Quality assurance checks enforce required fields and blocking thresholds. Output includes versioned datasets (CSV/Parquet), metadata (JSON), checksums (SHA256), and reports (audit, evidence samples, metrics).

Figure 2 — Dataset Snapshot (v1.3.0-beta)

Table: Dataset Statistics

Metric	Value
Total systems	80
Measured contrasts (Tier B)	65
Unique DOIs	20
Families covered	17
Biosensors	33
Standard FPs	47

Top 5 Families by Entry Count:

Family	Count	Examples
Calcium	12	GCaMP6s, jGCaMP7b, R-GECO1, NIR-GECO2
GFP-like	8	EGFP, Clover, mNeonGreen
RFP	6	TagRFP, FusionRed, Katushka
Dopamine	5	dLight1.1/1.2/1.3b, GRAB-DA2m/h
Far-red	5	mKate2, eqFP650, mCardinal

Contrast Distribution (normalized fold-change): - **Low** (1-2×): 30 systems (e.g., standard FPs: EGFP, mCherry) - **Moderate** (2-10×): 25 systems (e.g., voltage sensors: ASAP3, metabolic: HyPer3) - **High** (>10×): 10 systems (e.g., calcium: GCaMP6f, GCaMP6s)

Figure 2 Legend: Current dataset contains 80 systems with 65 measured contrast values. Calcium sensors dominate the high-contrast regime (>10-fold), driven by GCaMP/jGCaMP optimizations. Standard fluorescent proteins (EGFP, mCherry) cluster at low contrast (1-2×) as expected for non-biosensor applications.

END OF MANUSCRIPT