# Supplementary Materials

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

**Tommy Lepesteur** Independent researcher, France

## Table of Contents



## 1. Field Schema & Units

### Core Identity Fields

| Column | Type | Description | Example | |--------|------|-------------|---------| | `SystemID` | String | Unique identifier | FP_0012 | | `protein_name` | String | Protein name | GCaMP6f | | `canonical_name` | String | Standardized name (lowercase, no spaces) | gcamp6f | | `family` | String | Functional family | Calcium | | `is_biosensor` | Boolean | 1=sensor, 0=standard FP | 1 |

### Contrast Metrics

| Column | Type | Unit | Description | |--------|------|------|-------------| | `contrast_value` | Float | Original unit | Raw extracted value | | `contrast_unit` | String | — | "fold", "deltaF/F0", "percent" | | `contrast_normalized` | Float | fold-change | Normalized to fold (ΔF/F■ → 1+ΔF/F■) | | `quality_tier` | String | — | "A" (CI/n), "B" (measured), "C" (derived) |

### Context Metadata

| Column | Type | Unit | Description | |--------|------|------|-------------| | `context` | String | — | "in_cellulo(HEK293)", "in_vivo(neurons)" | | `temperature_K` | Float | Kelvin | Measurement temperature | | `pH` | Float | — | Buffer pH | | `method` | String | — | "fluorescence", "imaging", "FRET" | | `assay` | String | — | Assay type (e.g., "calcium_imaging") |

### Provenance

| Column | Type | Description | |--------|------|-------------| | `doi` | String | Publication DOI (required for measured) | | `pmcid` | String | PubMed Central ID (if OA) | | `source_note` | String | "Author YYYY Journal, protein_name" | | `license` | String | "CC BY", "varies (see DOI)" | | `curator` | String | Curation stage (v1.3_conservative, etc.) |

**Total columns**: 33 (see TRAINING.METADATA.v1.3.json for complete schema)

# 2. Quality Tiers & Decision Rules

## Tier A — Measured with Confidence Interval

**Criteria:** - Direct experimental measurement - Confidence interval (CI) or standard error (SE) reported - Sample size (n) specified - Traceable to figure/table with error bars or statistical test

**Example** (target for v1.3.1):

```
GCaMP6f: ΔF/F■ = 14.5 ± 2.3 (mean ± SEM, n=12 cells)
Source: Chen et al. 2013 Nature, Fig. 2c
```

**Count in v1.3.0-beta**: 0 (future expansion)

## Tier B — Measured (Point Estimate)

**Criteria:** - Direct experimental measurement - Point estimate only (no CI/SE/n) - Traceable to publication DOI + figure/table

**Example:**

```
dLight1.3b: 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
```

**Count in v1.3.0-beta**: 65

## Tier C — Computed/Derived

**Criteria:** - Computed from other measured quantities - Examples: brightness = QY × ε, relative contrast = sensor_A / sensor_B - NOT used for functional contrast values in current Atlas

**Count in v1.3.0-beta**: 0 (reserved for future brightness proxies)

## Decision Tree

```
Is the value directly measured in an experiment?
■
■■ YES → Does it have CI/SE + sample size?
■ ■
■ ■■ YES → Tier A
■ ■■ NO → Tier B
■
■■ NO → Is it computed from measured quantities?
■
```

■■ YES → Tier C

# 3. Example Evidence Notes

## High-Quality Entry (Tier B)

**GCaMP6s** (Calcium sensor)

| Field | Value | |-------|-------| | SystemID | FP_0014 | | protein_name | GCaMP6s | | family | Calcium | | contrast_value | 26.0 | | contrast_unit | fold | | contrast_normalized | 26.0 | | quality_tier | B | | context | in_cellulo(HEK293) | | temperature_K | 298.0 | | pH | 7.4 | | doi | 10.1038/nature12354 | | pmcid | PMC3777791 | | source_note | Chen et al. 2013 Nature - GCaMP6 suite | | license | CC BY (Nature OA) | | method | fluorescence | | assay | calcium_imaging |

**Provenance Trail:** 1. Original publication: Chen et al. *Nature* 2013, Figure 1d 2. Value extracted: 26-fold change upon saturating Ca²■ 3. Context: HEK293 cells, room temperature (295 K ≈ 298 K) 4. License confirmed: Nature OA article, CC BY

## Moderate Entry (Tier B, in vivo context)

**SF-iGluSnFR** (Glutamate sensor)

| Field | Value | |-------|-------| | SystemID | FP_0036 | | protein_name | SF-iGluSnFR | | family | Glutamate | | contrast_value | 5.8 | | contrast_unit | deltaF/F0 | | contrast_normalized | 6.8 | | quality_tier | B | | context | in_vivo(hippocampus) | | temperature_K | 310.0 | | pH | 7.4 | | doi | 10.1016/j.neuron.2013.06.043 | | pmcid | PMC3650424 | | source_note | Marvin et al. 2013 Neuron, SF-iGluSnFR | | license | CC BY |

**Provenance Trail:** 1. Original publication: Marvin et al. *Neuron* 2013, Figure 3 2. Value extracted: ΔF/F■ = 5.8 in hippocampal slices 3. Context: Mouse hippocampus, physiological temperature (37°C = 310 K) 4. Normalized: 1 + 5.8 = 6.8-fold

## Standard FP (Non-biosensor)

**EGFP** (Enhanced GFP)

| Field | Value | |-------|-------| | SystemID | FP_0009 | | protein_name | EGFP | | family | GFP-like | | contrast_value | 1.2 | | contrast_unit | fold | | contrast_normalized | 1.2 | | quality_tier | B | | context | in_cellulo | | temperature_K | 298.0 | | pH | 7.4 | | doi | 10.1016/j.gene.2005.06.018 | | source_note | Tsien 1998 - reference | | license | CC BY (Gene OA) |

**Notes:** - Standard FPs have low "contrast" (≈1-fold, no ligand-dependent change) - Included for spectral completeness and as ML training negatives - Contrast here refers to brightness vs background (not functional response)

# 4. Build Artifacts List

## Data Files

| Filename | Format | Size | Description | |----------|--------|------|-------------| | `atlas_fp_optical_v1_3.csv` | CSV | ~45 KB | Main dataset, 80 rows, 33 columns | | `atlas_fp_optical_v1_3.parquet` | Parquet | ~28 KB | Binary format (pandas/Arrow) | | `TRAINING.METADATA.v1.3.json` | JSON | ~8 KB | Schema, provenance, license summary | | `SHA256SUMS_v1.3.txt` | Text | ~1 KB | Checksums for integrity verification |

## Reports

| Filename | Description | |----------|-------------| | `reports/AUDIT_v1.3_fp_optical.md` | QA audit: pass/fail per check, blocking issues | | `reports/EVIDENCE_SAMPLES_v1.3.md` | Table of 30+ measured contrasts with sources | | `reports/METRICS_v1.3.json` | Machine-readable counts, statistics, QA results | | `reports/SOURCES_AND_LICENSES.md` | License breakdown per source |

## Scripts & Config

| Path | Description | |------|-------------| | `scripts/etl/build_atlas_v1_3.py` | Main build script | | `scripts/etl/fetch_fpbase_candidates.py` | FPbase GraphQL harvest | | `scripts/etl/extract_pmc_contrast_real.py` | PMC full-text mining | | `scripts/qa/compute_metrics_v1_3.py` | Metrics & QA checks | | `schema/aliases.yaml` | Canonical name mappings | | `config/providers.yml` | API endpoints, rate limits |

# 5. Source Breakdown

## Contribution by Source (v1.3.0-beta)

| Source | Count | Description | |--------|-------|-------------| | `neurotransmitter_preseed` | 11 | Manually curated dopamine, glutamate, ACh sensors | | `metabolic_preseed` | 10 | ATP, cAMP, pH, H■O■ sensors | | `geci_db_preseed` | 9 | Calcium indicator database | | `pmc_fulltext` | 8 | Conservative PMC XML extraction | | `voltage_preseed` | 6 | Voltage indicator database | | `v1.2.1_migration` | 36 | Legacy FP entries from previous build |

**Total**: 80 unique systems (after deduplication)

## FPbase API Status

**v1.3.0-beta**: FPbase GraphQL API was **down** during build window (Oct 2024). Fallback strategy:

1. Use specialist preseeded databases (higher quality, sensor-focused) 2. Conservative PMC mining (8 entries, manual validation) 3. v1.2.1 migration for continuity (36 entries)

**Impact**: Lost ~150 standard FP entries (mCherry, mKate, etc.) that would have come from FPbase. These will be restored in v1.3.1 when API recovers.

**Mitigation**: Current 80 systems prioritize **biosensors** (33 entries) over standard FPs (47 entries), aligning with Atlas focus on functional quantum-enabled sensors.

# 6. License Tracking Details

## Per-Source License Status

| Source | License | Reusability | |--------|---------|-------------| | **FPbase API** | varies (see original publication DOIs) | ■ Open (per FPbase policy) | | **Specialist databases** | varies (see DOI) | ■ Curated for OA publications | | **PMC full-text** | CC BY / CC0 | ■ Open Access only | | **v1.2.1 migration** | Mixed | ■■ Pending audit |

## License Breakdown (v1.3.0-beta)

| License | Count | Percentage | |---------|-------|------------| | `varies (see DOI)` | 36 | 45% | | `CC BY/CC0 (PMC OA)` | 8 | 10% | | `CC BY (Nat Commun OA)` | 4 | 5% | | `CC BY (Nature Methods OA)` | 12 | 15% | | `CC BY (PNAS OA)` | 6 | 7.5% | | `CC BY (Neuron OA)` | 4 | 5% | | *(Other CC BY)* | 10 | 12.5% |

**Notes:** - "varies (see DOI)": Entries from specialist databases where license must be checked per original publication. **Action item (v1.3.1)**: Scrape licenses via Unpaywall API. - All PMC entries confirmed CC BY/CC0 (Open Access filter applied during extraction).

## Reusability Guarantee

**Commitment**: By v1.3.1 (stable release), 100% of entries will have explicit license attribution: - Either CC BY/CC0/CC BY-SA (permissive) - OR explicit publisher OA policy documented

**Current compliance**: ~55% explicit CC BY, 45% pending granular check.

# 7. Normalization Examples

## ∆F/F■ → Fold-Change

**Original**: ∆F/F■ = 15.5 **Normalized**: 1 + 15.5 = **16.5-fold**

**Rationale**: ∆F/F■ represents fractional change from baseline (F■). Adding 1 converts to absolute fold-change (F_max / F■).

## Percent → Fold-Change

**Original**: 340% increase **Normalized**: 1 + (340/100) = **4.4-fold**

**Rationale**: Percent increase is relative to baseline. Dividing by 100 and adding 1 yields fold-change.

### Fold-Change (as-is)

**Original**: 26-fold **Normalized**: **26.0-fold** (no transformation)

## 8. QA Threshold Rationale

### Blocking Thresholds (Production Releases)

| Metric | Threshold | Rationale | |--------|-----------|-----------| | `N_total` ≥ 200 | Comprehensive coverage of major sensor families | | `N_measured` ≥ 120 | Sufficient for robust ML training (10-fold CV with n=12 per fold) | | `families_with_≥5` ≥ 10 | Diversity across functional classes (calcium, voltage, metabolic, neurotransmitters) | | `unique_doi_rate` ≥ 0.85 | Minimize redundancy; each sensor ideally from distinct publication | | `license_ok_rate` = 1.0 | Legal compliance for dataset redistribution |

**v1.3.0-beta Exceptions:**

Beta release serves as **community testing snapshot**. Thresholds relaxed to enable early feedback: - N_total: 80 / 200 (40%) — Acceptable for beta - N_measured: 65 / 120 (54%) — Usable for initial ML prototypes - license_ok_rate: 0.1 — **Not acceptable for stable**; requires v1.3.1 fix

## 9. Future Schema Extensions (Roadmap)

### Planned Additions (v2.0)

| Column | Type | Description | |--------|------|-------------| | `contrast_ci_low` | Float | Lower bound of 95% CI | | `contrast_ci_high` | Float | Upper bound of 95% CI | | `sample_size_n` | Integer | Number of replicates | | `ex_max_nm` | Float | Excitation maximum (nm) | | `em_max_nm` | Float | Emission maximum (nm) | | `quantum_yield` | Float | Fluorescence quantum yield | | `brightness_proxy` | Float | QY × ε (computed) | | `photostability` | String | "high", "moderate", "low" | | `in_vivo_validated` | Boolean | Demonstrated in animal models |

**END OF SUPPLEMENTARY MATERIALS**