Claim Graphs for eLife Neuroscience Papers: Findings from a 10-Paper Prototype

Zachary F. Mainen Champalimaud Research, Lisbon

Prepared for Damian Pattinson and Andy Collings (eLife) — March 2026

1. Introduction

A claim graph is a structured representation of what a paper asserts and on what grounds. The basic unit is the claim: a single declarative sentence asserting what one analysis showed, grounded in a specific figure panel. One analysis, one result, one data source. Each claim names the phenomenon, the direction and magnitude of the result, and the key statistics if the result is quantitative. Each claim carries a pointer to the data and the analysis script that transforms data into the result. And each claim carries explicit dependency edges to the claims it structurally requires — not by citation, but by logical necessity. If this claim would be undermined if that one were wrong, that dependency is made explicit.

The format is not a new theory of science. It is a precise operationalization of something peer reviewers already do informally. When a reviewer reads a paper and traces an argument back through its figures, looking for the weakest link, they are traversing a dependency graph. The claim format makes that graph explicit, names each node, and records its reproducibility status.

Panel-level granularity is the right level for this purpose. A paper-level reproducibility designation is nearly meaningless: a paper can be "reproducible" in the sense that one figure is reproducible while another is not, or that the main result holds while a supporting claim that structures the interpretation does not. Claim-level, conversely, tends to be too granular — individual statistical tests stripped of their analytical context lose the meaning that makes reproduction meaningful. Panel-level is the level at which analysis decisions are made, at which data are transformed into results, and at which peer review is most often exercised.

The dependency structure between claims carries editorial information that the paper's narrative does not directly surface. A paper's argument integrates its dependency structure into a rhetorical sequence designed for human readers: it foregrounds what is striking, backgrounds what is assumed, and presents conclusions before the methods that ground them. This is appropriate for human reading. It is also lossy in specific ways: assumptions that span many claims appear only in methods sections; structural features of the argument

— that this claim is architecturally decoupled from that one, or that the headline finding depends on a parameter choice two dependency levels deep — are invisible in the narrative. The claim graph makes them visible.

The prototype applied this method to all 10 papers eLife provided. The present document reports what was found. It is not a proposal for how eLife should change its processes — it is a report from a field experiment, and the findings are what they are.

A note on the 3-agent extraction method: the method specification calls for three independent reading passes (Agent A reads results prose, Agent B reads figure captions, Agent C reads methods and code) before reconciliation. In this prototype, all three passes were performed by a single analyst, reading in the order specified and withholding prior pass outputs. This is not three independent analysts and should not be treated as such. The practical benefit of the three-pass structure — different reading angles surface different claims and reduce systematic blind spots — was preserved; the statistical independence was not.

---

2. The Prototype Corpus

The 10 papers span four research modalities and a substantial range of data types, computational requirements, and data-sharing practices.

#	Paper (short title)	Type	Claims	Verified	Primary reproduction
1	Rozak — NOVAS3D neurovascular pipeline	Computational / deep learning	17	1 (assessment)	GPU required for infere
2	Gädeke — Insula guilt and responsibility	fMRI / behavioral	18	0	MATLAB + SPM12 dependency throughou
3	Kolb — iGABASnFR2 GABA sensor	Wet-lab tools	15	1 (assessment)	Sensor constructs and stopped-flow equipmer
4	Scheller — Self-prioritization TVA	Psychophysics / Bayesian	14	0	Stan/R not yet executed path)
5	Bouyeure — Fear and reversal RSA	fMRI / RSA	22	1 (assessment)	BrainIAK searchlight co multi-language stack
6	Headley — Inhibitory rhythms	Computational biophysics	17	9 (7 result + 2 assessment)	Not yet executed (clear pre-computed data)
7	Wengert — KCNC1 epilepsy mouse	Wet-lab / genetics	18	1 (assessment)	Knock-in mouse line re
8	Artiushin — Spider brain atlas	Anatomical atlas	17	1 (assessment)	BIL TIFF volumes (~ten GB); atlas inspection re
9	Ejdrup — Dopamine dynamics model	Computational	20	13 (9 result + 4 assessment)	matplotlib 3.8 API breal tissue-scale compute
10	Kämmer — Foveal feedback fMRI	fMRI / MVPA	15	1 (assessment)	OpenNeuro data volum MVPA compute

Total: 173 claims extracted across 10 papers. Approximately 28 verified (by execution or structural code inspection), approximately 125 unverified with clear reproduction paths not yet attempted, approximately 18 unverified:no-data due to wet-lab or atlas barriers, and 2 unverified:no-code.

The corpus was deliberately diverse. Computational papers, where the full analysis chain is software-embodied, sit alongside wet-lab tools papers, where the primary contribution is a physical artifact. fMRI papers with preregistered designs sit alongside a biophysical simulation paper parameterized from the literature. This range was the right choice: it surfaced issues that a computationally homogeneous corpus would not have.

3. Paper-by-Paper Findings

Ejdrup — Dopamine dynamics model (computational, 20 claims) SILVER

This paper builds a reaction-diffusion model of striatal dopamine dynamics to explain the functional difference between dorsal and ventral striatum. The primary finding — that DAT Vmax is the dominant determinant of regional contrast — is well-supported by the parameter sweep in Figures 3B–G, and the claim `vmax-only-parameter-driving-regional-difference` is among the most strongly epistemic claims in the corpus. Thirteen of the 20 claims were verified in this prototype run, including 9 result claims and 4 assessment claims verified by code inspection.

What the claim graph exposed that the paper's narrative did not foreground is the architectural decoupling of the nanoclustering model from the main tissue model. The paper presents both as parts of a unified argument: DAT nanoclustering in VS creates a diffusion-limited clearance bottleneck, which contributes to explaining the lower effective Vmax in VS. The claim graph makes this coupling explicit as an assessment claim — `nanoclustering-model-varicosity-scale` — and records that the two models are architecturally separate, with no formal coupling in the code. The nanoclustering result supports the main model's interpretation but does not participate in its execution. A reader following the paper's prose would not identify where the connection is formal and where it is interpretive.

There is also a numerical discrepancy in `dat-clustering-greater-in-vs`. The paper states the dSTORM nanoclustering result is significant at p=0.012 (n=12/13). A prior reproduction attempt logged in the repository records p=0.029 with sample sizes reversed (n=13/12). This has not been resolved in the current deposit. It does not affect the qualitative conclusion, but it is a specific, checkable claim where the paper and a logged reproduction attempt do not agree.

Three claims are blocked by a matplotlib 3.8 API change (`Axes3D.w_xaxis` was renamed to `xaxis`). This is a one-line fix per script. The simulation code is correct; only the figure rendering is broken.

Headley — Inhibitory rhythms (computational biophysics, 17 claims) SILVER

This paper uses a multicompartmental NEURON model of a layer 5 pyramidal neuron to show that distal-dendritic inhibition at beta frequencies gates Ca^{2+}/NMDA spike occurrence while perisomatic inhibition at gamma frequencies modulates action potential threshold — two distinct computational operations performed by two distinct circuit motifs. The model is a single cell, and the assessment claim `15-model-single-cell-scope` records this explicitly: all 17 result claims are predictions of a single-cell model with no network dynamics or population-level effects.

The paper's synthesis claim `pv-gamma-sst-beta-correspondence` provides mechanistic grounding for the empirical association of PV+ interneurons with gamma and SST+ interneurons with beta. This is the claim the paper is most commonly cited for. In the dependency graph, it sits two levels above any directly verifiable result: it requires both `beta-gates-distal-apical-inputs` and `gamma-gates-proximal-basal-inputs`, each of which in turn requires multiple upstream claims about frequency-dependent gating. The claim is well-supported within the model's assumptions, but the path from verifiable simulation result to the PV/SST interpretation passes through a mechanistic analogy that the model does not directly demonstrate — interneuron types are not modeled; they are invoked as interpretive matches. The claim graph makes this interpretive step visible as a gap in the dependency chain.

Seven of the 15 result claims are verified from pre-computed CSV data in the repository; 7 remain unverified pending download of the Dryad oscillation frequency-sweep arrays; 1 synthesis claim is unverified:no-code. The reproduction path for the remaining claims is clear: pre-computed data arrays are in the Dryad deposit and the GitHub repository provides per-figure notebooks. The barrier is time, not access.

Scheller — Self-prioritization (psychophysics, 14 claims) SILVER

This psychophysics paper uses a temporal order judgment task and hierarchical Bayesian TVA modeling across two experiments (N=69, N=71) to show that arbitrary self-associations enhance attentional selection at the perceptual feature level — and, counterintuitively, that this advantage disappears when social decoding is required.

The paper has two main findings. The headline finding, which the title foregrounds, is self-prioritization: self-associated stimuli are processed faster in the perceptual decision condition (1.5 Hz advantage, HDI: −0.16 to 3.2 Hz). The more theoretically important finding, captured as `self-prioritization-absent-social-decision`, is the inversion: in the social decision condition there is no self-advantage, and other-associated stimuli show a processing rate advantage instead. This is the claim that constrains theories of self-prioritization — it means the effect is not about social salience per se but about the absence of explicit identity decoding. The claim graph captures both as first-class nodes in the dependency tree, but their relative theoretical weight is inverted relative to the paper's narrative emphasis. The headline finding is the less surprising one.

The reproduction path is clear. Data and notebooks are on OSF; the preregistration specifies the analysis plan. The primary dependency is Stan/R for the hierarchical Bayesian model.

Kämmer — Foveal feedback (fMRI/MVPA, 15 claims) NO BADGE / GAP

This paper uses a gaze-contingent fMRI paradigm to show that peripheral saccade targets, which disappear before fixation in 99.27% of trials, are nonetheless decodable from foveal V1 BOLD signal (57.43% decoding accuracy, $t(27)=8.81$, $p<0.001$). The preregistered

design is a structural strength: the analysis plan is fully documented before data collection, making the reproduction path completely unambiguous.

The claim `ips-candidate-driver-foveal-feedback` is classified as `unverified:no-code`. From the available materials, it is unclear whether the IPS driver claim is supported by a functional connectivity analysis reported in the paper or whether it is an inference drawn in the discussion. If the latter, it is a discussion claim masquerading as a result claim — a distinction the claim graph makes explicit by requiring a code pointer. This was not resolvable in the initial prototype pass, which relied on HTML fetching; the JATS XML re-extraction subsequently confirmed from the Results section that the IPS analysis is fully reported with statistics. The claim has accordingly been updated to `unverified:compute-infeasible`.

Five of the seven claims are `unverified:compute-infeasible`. The standard Python neuroimaging stack (nilearn, sklearn) is available; the bottleneck is the OpenNeuro data download volume and MVPA compute time across cross-validation folds.

The re-extraction added the assessment claim `preregistration-submitted-after-manuscript`, verified by OSF timestamps. The preregistration was submitted after the manuscript had been prepared — a detail that does not appear in the paper's description of its preregistered design. This changes the evidential status of the assessment claim `preregistered-design-validates-mvpa`, which was initially classified as `verified:strong` on the grounds that preregistration fully constrains the analysis plan. That classification has been downgraded to `verified:moderate`. The analysis plan is documented, and the preregistration does match the reported analysis — but the protection preregistration affords against post-hoc analysis choices derives from the preregistration preceding data collection and analysis. When the preregistration is submitted after the manuscript, the documentation value is preserved while the chronological protection is not. Reviewers reading that the paper has a preregistered design would normally treat this as strong evidence against HARKing; the OSF timestamp removes that inference. The analysis may be exactly what was planned before data collection — there is no evidence of irregularity — but the claim graph cannot record that on the basis of the preregistration alone, and the epistemic weight is adjusted accordingly.

Wengert — KCNC1 epilepsy mouse (wet-lab, 18 claims) SILVER

This paper generates and characterizes the Kcnc1-A421V/+ knock-in mouse model. The primary claims require this specific mouse line — its generation is the contribution, and the measurements (electrophysiology, survival curves, immunostaining) require both the animal and appropriate equipment.

The claim `excitatory-neurons-unaffected-juvenile` warrants particular attention. In a disease model paper, the specificity claim — that the phenotype is restricted to inhibitory neurons — is as important as the positive findings about those neurons. It is the claim that justifies the conclusion of selective inhibitory dysfunction. It is also the least scrutinized,

because absence-of-effect claims in heterozygous models are particularly sensitive to statistical power, and the paper's primary power calculations were almost certainly designed around the positive findings in PV interneurons. The claim graph assigns this claim `moderate` epistemic weight rather than `strong`, reflecting the asymmetry between detecting an effect and confirming its absence.

G-Node contains the raw electrophysiology traces, which would allow statistical figure reproduction from deposited data. But reproducing figures from deposited data is not the same as re-measuring the phenotype. The claim graph records both the accessibility of deposited data and the nature of what it does and does not enable.

Bouyeure — Fear and reversal RSA (fMRI/RSA, 22 claims) SILVER

This paper uses fMRI and representational similarity analysis across a multi-day fear conditioning paradigm to show that reversal learning employs two simultaneous representational strategies: generalization for newly threatening cues, and item-specificity for cues whose threat value changed.

The most specific predictive claim in the paper — `pfc-context-specificity-predicts-renewal` — is the claim that individual differences in PFC context-specificity during reversal predict subsequent fear renewal at test. This is the paper's most informative finding in the sense of most constraining: it links a neural measurement during learning to behavior after learning, and it makes a directional individual-differences prediction. It is also the claim that sits behind the most computationally inaccessible analysis node in the paper: the LSS beta series estimation that precedes the RSA searchlight is the most expensive step in the pipeline, and it must be completed before any of the RSA results can be computed. NeuroVault pre-computed maps are available, which would allow figure reproduction without re-running the searchlight — but the individual-differences prediction requires individual data, not group maps.

The assessment claim `rsa-roi-derived-from-searchlight` surfaces a structural feature that the narrative does not foreground. The ROIs used in Figures 4B and 5D are derived from the same searchlight analysis conducted on the same dataset — a procedure known as double dipping or circularity of ROI selection. The searchlight identifies regions with significant pattern similarity; subsequent RSA analyses in those same regions using the same data then inherit that significance by construction. FDR correction across voxels corrects for multiple comparisons in the searchlight, but it does not correct for the selection bias introduced by using searchlight-identified ROIs for confirmatory RSA in the same participants. All result claims downstream of Figures 4B and 5D depend on this assessment node. The pattern similarity results in those figures are not independent tests of the RSA hypotheses — they are confirmations in regions pre-selected to show the effect. This is a common analysis pattern in fMRI RSA research and does not invalidate the paper's

findings, but it does constrain what those findings demonstrate: the effect is real in these regions, in this sample; whether it would replicate in an independently selected ROI in a new sample is an open question the current design cannot answer.

Gädeke — Guilt and insula (fMRI/behavioral, 18 claims) GOLD

This fMRI paper studies the neural basis of interpersonal guilt through a social monetary decision task. The analysis chain is documented in detail: the GitHub repository maps each MATLAB script to a figure, behavioral data are in `.mat` files, and pre-computed group NIfTI results enable figure reproduction without re-running first-level GLMs.

All eight claims are unverified because the primary analysis environment is MATLAB/SPM12. There is no Python path. The pre-computed results enable figure reproduction from processed data, but the complete analysis chain — from behavioral `.mat` files through computational guilt model to first-level GLM — requires MATLAB. This is a MATLAB dependency, not a data availability problem.

Rozak — NOVAS3D pipeline (deep learning, 17 claims) NO BADGE / COMPUTE

This paper presents a deep learning pipeline for volumetric vessel segmentation. The contribution is the pipeline itself; the quantitative claims (24 ± 28% vessel radius heterogeneity, 4% efficiency increase, 152 ± 65% assortativity increase) are measurements from the pipeline applied to a specific dataset and mouse model.

The assessment claim `dl-model-scope-single-pipeline` is the most important node in the dependency graph. All result claims are conditional on the pipeline having been trained and applied to a specific preparation (Thy1-ChR2-YFP, optogenetic stimulation, 6–12 month old mice). The claims are therefore not generalizable to other preparations without additional validation. This scope constraint appears in the methods but is not foregrounded in the abstract. The claim graph makes it a first-class node.

The re-extraction added `responder-threshold-2sd-untested` as an assessment claim. The paper classifies vessels as responders using a 2×SD threshold applied to baseline variability. All quantitative spatial claims — the 4% efficiency increase, the 152 ± 65% assortativity increase, the comparisons between dilations and constrictions — and all downstream network biology findings are conditional on this threshold. The paper includes a sensitivity analysis in the appendix, but it is qualitative: it notes that the pattern of results is preserved at alternative thresholds without providing quantitative comparison across thresholds. The assessment claim records that the threshold has not been tested quantitatively. This propagates epistemic uncertainty upward through every result claim that depends on the responder classification, which is the majority of the paper's quantitative spatial findings. The effect is not that the results are wrong — it is that their robustness to the central threshold choice has not been formally demonstrated.

Kolb — iGABASnFR2 sensor (wet-lab tools, 15 claims) NO BADGE / STRUCTURAL

This paper presents iGABASnFR2, an improved GABA sensor resulting from high-throughput mutagenesis screening of 3,947 variants. The primary claims require the sensor constructs and specialized equipment (stopped-flow kinetics, two-photon spectroscopy). Zenodo deposits contain the source data for figure reproduction from pre-measured traces, but not the measurements themselves.

The assessment claim `screening-scope-wet-lab-only` is the most important and most immediately verifiable node: it records that the reproduction barrier is material, not informational. Source data are available. Code is available. What is not available is the capacity to perform the measurements. This distinction — data availability versus measurement availability — is normally invisible in print. The claim graph makes it explicit.

Artiushin — Spider brain atlas (anatomical atlas, 17 claims) NO BADGE / STRUCTURAL

This paper describes a 3D immunofluorescence atlas of the *Uloborus diversus* synganglion. The claims are anatomical observations — neuropil identification, neurotransmitter immunoreactivity patterns, novel structures. Verification means atlas inspection, not analysis re-execution.

The claim `protocerebral-bridge-candidate-central-complex` is assigned `weak` epistemic weight because it is a homology interpretation: the paper proposes that a structure in the spider brain is a candidate homolog of the insect protocerebral bridge. This interpretation requires comparative neuroanatomy expertise that lies outside the usual reproducibility framework. The claim graph records it as a first-class claim with appropriate epistemic weight, rather than folding it into the paper's narrative framing.

---

4. Cross-Corpus Patterns

Three reproducibility tiers

The 10 papers fall into three tiers based on the structural nature of their reproduction barrier, not on the quality of their data-sharing.

Tier 1 (papers 4, 6, 9) — computational papers with clear reproduction paths. All three have code in GitHub, data in standard deposits, and pre-computed arrays that allow plotting without re-running expensive simulations. Paper 9 (Ejdrup) achieved the highest verification count in the corpus: 13 of 20 claims verified. The remaining barriers are a one-line matplotlib fix and compute time. Paper 6 (Headley) has all pre-computed data available on Dryad; the barrier is execution time. Paper 4 (Scheller) has all data on OSF and a preregistered analysis; the barrier is Stan/R environment setup. All three satisfy eLife's data availability requirement — and in all three, that requirement is genuinely met in the sense that the data needed to reproduce the figures is accessible.

Tier 2 (papers 1, 2, 5, 10) — fMRI and deep learning papers where figure reproduction from pre-computed outputs is feasible but full re-analysis is blocked by MATLAB, GPU requirements, or data volume. Paper 2 (Gädeke) requires MATLAB/SPM12 throughout; pre-computed NIfTI results enable figure reproduction but not analysis chain reproduction. Papers 5 and 10 require compute-intensive MVPA pipelines, but NeuroVault and standard Python stacks make figure reproduction tractable. Paper 1 requires GPU inference, but a pre-trained model and Tutorial notebook provide a clear path. In all four, the data availability statement accurately describes what was deposited. What varies is whether the deposited materials enable re-analysis versus figure reproduction, and whether the required compute environment is broadly accessible.

Tier 3 (papers 3, 7, 8) — wet-lab and atlas papers where the primary reproduction barrier is material rather than informational. Paper 3 (Kolb) requires sensor constructs and specialized optical equipment. Paper 7 (Wengert) requires a knock-in mouse line. Paper 8 (Artiushin) requires large-scale confocal volumes. All three deposit data and code — the deposited data allow figure reproduction from pre-measured traces, and in Wengert's case G-Node contains raw electrophysiology enabling statistical figure reproduction. The claim graph is the first thing that makes visible the difference between "data available" and "primary phenomenon re-measurable."

Data availability and executable analysis are not the same

All 10 papers satisfy eLife's data availability requirement at the level of the data availability statement. In 4 of 10 papers, that satisfaction is genuine in the strongest sense: the data are accessible, the code runs, and the figures can be reproduced. In 3 of 10 papers, the data are accessible and the deposited traces allow figure and analysis reproduction, but re-collection of primary measurements would require equipment the statement does not mention. In the remaining 3, reproduction of primary claims requires compute infrastructure (GPU, MATLAB, MVPA pipelines) that the statement also does not mention. These are distinct situations. They currently look identical in print — both appear as "data available on Zenodo" or equivalent.

This is not a criticism of any paper. It is a consequence of the current data availability statement format, which records where data are deposited but not what the data enable. A statement that a dataset is on Zenodo answers the question "was the data deposited?" It does not answer "can I reproduce Figure 4 from this deposit?" The two questions have different answers in most papers in this corpus.

Assessment claims as an invisible claim type

Every paper in the corpus contains claims about its own scope, parameterization, and analytical assumptions that function as claims but do not appear as such. These are what the method calls assessment claims: structural properties of models, analytical frameworks, or scope delimitations that downstream result claims depend on.

Assessment claims share two properties that make them important. First, they are consistently verifiable: they can be confirmed by reading the code, the methods section, or the preregistration. Unlike result claims, which may require data, specialized equipment, or significant compute, assessment claims are almost always accessible. Second, they are almost always present: every paper in this corpus had at least one assessment claim that, once explicit, changed what a reader could know about the downstream results.

Across 10 papers, 14 assessment claims were extracted and 13 were verified (by code inspection, preregistration inspection, or methods reading). This 93% verification rate is the highest of any claim type in the corpus, and it holds regardless of paper type — wet-lab, computational, or fMRI. Assessment claims are a universal and consistently accessible claim type that is currently invisible in the published record.

The narrative hiding pattern

In 9 of 10 papers, the claim graph exposed a structural feature that the prose did not foreground. The cases are worth enumerating because the pattern is consistent and not attributable to any deficiency in the specific papers.

In Ejdrup, the architectural decoupling of the nanoclustering model. In Headley, the interpretive gap between simulation result and PV/SST interneuron correspondence. In Scheller, the inversion of narrative emphasis — the null result in the social decision condition is more theoretically important than the positive result in the perceptual decision condition. In Kämmer, the ambiguity of the IPS driver claim's evidential grounding — and, from the re-extraction, the post-manuscript preregistration timestamp, which changes the protection that "preregistered design" affords. In Wengert, the epistemic asymmetry of the excitatory-sparing claim. In Bouyeure, the position of the individual-differences prediction behind the most compute-intensive analysis node — and, from the re-extraction, the double-dipping structure of the searchlight-derived ROIs used in Figures 4B and 5D. In Rozak, the absence of a quantitative sensitivity analysis for the 2×SD responder threshold that conditions all spatial and network biology claims.

That is 9 of 10 papers with structural features the claim graph surfaces that the narrative did not foreground. The one exception is Kolb, where the primary assessment claim is about material barriers to reproduction rather than about an analytical choice or structural gap in the argument.

These are not failures of those papers. Narrative writing integrates argument structure into a rhetorical sequence — it foregrounds what is striking, sequences results for readability, and presents conclusions before their grounds. This is appropriate for human readers. It is also why the structural features a claim graph reveals are not visible in the paper: they were suppressed in the service of a coherent argument. The claim graph does not improve the papers. It makes a different kind of structure visible.

5. What the Method Learned About Itself

Six methodological issues arose during the corpus run and were resolved during processing. They are documented here not as limitations to disclaim but as findings about papers and how they work.

Issue 1: Assessment claims need their own status vocabulary — and their existence as a claim type is itself a finding. The method initially had no clear way to classify a claim verified by reading code rather than running it. The fix was simple: assessment claims are `verified` when inspection confirms the structural property, with a note that verification was by code reading rather than execution. But the necessity of this fix revealed something more fundamental: papers contain a large class of structural commitments that are checkable by code inspection, consistently present, and currently invisible in publication. Making them first-class claims changes what peer review can address.

Issue 2: "Unverified" is not a status — it is an absence of status. The method initially used bare `unverified` as a catch-all. During processing it became clear that `unverified` conceals more than it reveals: `unverified:no-data`, `unverified:no-code`, `unverified:code-error`, `unverified:compute-infeasible`, and `unverified` (not yet attempted) are editorially different. The full status vocabulary matters because each status implies different remedies: a `no-data` claim is addressed by deposition, a `code-error` claim by a software fix, a `compute-infeasible` claim by pre-computed arrays.

Issue 3: Visualization code breaks differently from analysis code — the distinction is editorially meaningful. In Ejdrup, three claims were blocked by a matplotlib 3.8 API change that affected only the 3D figure rendering code, not the simulation. The underlying simulation outputs are reproducible; the figure generation is not. A code error in the visualization step should not be treated as a failed analysis.

Issue 4: "Verification" is not one act — it depends on claim type. For result claims in computational papers, verification means re-executing the analysis and comparing output to figure. For assessment claims, it means structural code inspection. For anatomical atlas claims, it means atlas inspection. The method initially had a single verification concept. The atlas paper (Artiushin) forced explicit recognition that verification is a family of acts, and the reproduction block should specify which act was performed.

Issue 5: HTML pages truncate under automated fetching; eLife's JATS XML is the right programmatic interface. eLife article HTML pages truncated during automated retrieval in this prototype — results sections, figure captions beyond approximately Figure 2, and methods sections were unavailable via HTML fetch. eLife's JATS XML CDN endpoint returns complete, machine-readable documents and is the correct interface for programmatic access to article content. The method was updated accordingly. The long-term recommendation: use eLife's JATS XML CDN as the primary extraction interface.

Issue 6: The wet-lab/atlas ceiling is material, not informational — the claim graph makes this visible rather than solving it. Papers 3, 7, and 8 have primary measurements that cannot be re-collected without the physical sensor constructs, the knock-in mouse, or the confocal microscope. The deposited data allow figure and analysis reproduction from the deposit, which is the standard applied here. What the deposit cannot provide is the capacity for independent re-measurement. A data availability statement that says "source data available on Zenodo" conceals this distinction. An assessment claim that says "all primary claims require sensor constructs and specialized fluorescence equipment" makes it first-class and visible.

---

6. Implications for eLife's Editorial Process

The findings above suggest four questions for eLife to consider. These are offered as genuine questions, not recommendations. eLife may have editorial reasons — practical, policy, or community-relations — to answer any of them differently than the prototype suggests.

Could submission require a structured claim table in addition to a prose data availability statement? The current data availability statement records where data were deposited. A structured claim table — panel, data deposit DOI, script name, compute requirements, feasibility flag — would record what the deposited data enable. The two answers are not the same, and in 6 of the 10 papers in this corpus they differ in ways that are editorially relevant. A structured table would not require authors to do more work: it would require them to make explicit what they already know about their own data and code.

Should eLife distinguish between figure reproduction from deposited traces and independent re-measurement? In 5 of the 10 papers, deposited data enable figure reproduction from pre-processed or pre-computed outputs without enabling re-analysis from raw data or re-measurement from scratch. In 3 papers (the Tier 3 cases), deposited data enable figure reproduction from pre-measured traces but not the primary measurements. These are different levels of openness, and they currently look identical in print.

For computational papers where reproduction is part of the contribution: should container specifications be required rather than environment files? In this corpus, Papers 6, 4, and 9 — the Tier 1 papers — each had a conda environment file. In Paper 9, the matplotlib breakage was precisely a version incompatibility that a container specification would have prevented. A Docker or Singularity container provides a reproducible environment by definition, rather than a best-effort specification. For papers where the computational environment is the artifact — pipeline papers, simulation papers — the case for containers is stronger.

Should peer review surface claims that rest on discussion reasoning rather than reported analysis? The Kämmer paper's IPS driver claim is classified `unverified:no-code` because it could not be determined whether it is supported by a reported functional connectivity analysis or is an inference from the discussion. This distinction — reported analysis versus discussion inference — is meaningful for what a reader should believe. Asking reviewers to flag claims in this category is a modest change to review guidance that would surface a class of problem currently invisible in peer review.

---

7. What a Claim-Ready Paper Looks Like

Two papers in the corpus most closely approach what a claim-ready submission would require: Kämmer (paper 10) and Headley (paper 6).

Kämmer's preregistered design is the single most useful structural feature in the corpus for claim extraction. The preregistration specifies the analysis plan, the hypothesis, and the analysis decisions, and the correspondence between the preregistration and the reported analyses is exact — there is no question about which analysis was the primary one, whether any parameters were adjusted after seeing the data, or whether the reported figures represent the pre-specified tests. The preregistration allowed the assessment claim `preregistered-design-validates-mvpa` to be classified as `verified:moderate` by inspection. The ambiguity of the IPS claim is the exception that proves the rule: because everything else is preregistered, the one unregistered claim stands out.

Headley's GitHub repository provided a mapping from notebook to figure, a conda environment specification, pre-computed data on Dryad, and documentation in the README sufficient to assign every claim to a panel and a code path without reading the full paper. The reproduction path was clear at code inspection, before any execution. What would get Headley from "clear path" to "executed" is straightforward: the environment specification would need to be hardened, and the Dryad deposit would need a brief README explaining which data file corresponds to which figure.

What a claim-ready author would need to add to either paper is modest. A structured claim table — ten to twenty rows, one per panel, with a data pointer and a script name — would answer the questions that took the most time in this prototype. For most papers, the answers already exist in the GitHub README or in the methods section; the structured table would consolidate them. The incremental authorial effort is small; the information created is substantial.

---

8. Integration with the eLife Reading Experience

The standalone site demonstrates the concept. The practical question for eLife is how the claim graph enters the reading experience — whether it requires a reader to navigate away from the article, or whether it can be present alongside it. Three integration models exist, each requiring a different level of eLife involvement.

Reader-installed extension. A browser extension that detects eLife article URLs and injects a collapsible claim graph panel into the article page requires no eLife involvement at all. This is available today for any reader who installs it. It provides a proof of concept for what integrated reading looks like. It does not scale to readers who do not install extensions, and it is not a partnership in any meaningful sense. Its value is demonstrative, not distributional.

Embedded widget. eLife includes a lightweight script tag or iframe on article pages for papers that have associated claim graphs. The panel renders inline — collapsible, sitting between abstract and introduction, or as a sidebar — without the reader leaving the article. This requires minimal eLife engineering: one script include per article, conditional on a claim graph existing for that paper. The claim graph site serves the data; eLife controls the placement, the visual integration, and the decision about which papers carry the widget. This is the concrete near-term proposal — what eLife could trial with the ten papers in this corpus without committing to any policy or infrastructure change. Authors of those papers would know their claim graphs are being surfaced to real readers; eLife would have a real-traffic test of whether readers engage with the panel; and the entire trial is reversible by removing a script tag.

Native integration. Claim tables become part of the submission and publication record. eLife stores claim data in their own infrastructure, renders claim graphs natively in their article template, and surfaces claim status as part of the editorial record. This requires three changes that the embedded widget does not: a policy change (structured claim table required at submission), an editorial workflow change (reviewers engage with the claim table as part of their assessment), and engineering investment (claim data in eLife's own storage and rendering pipeline). It is also the only integration model that creates the feedback loop between submission, review, and publication that makes claim graphs editorially valuable rather than retrospective annotations.

The embedded widget is what this document asks for. It is also where the conversation is pointing. If the widget generates reader engagement, if authors of the ten papers find the claim graphs recognizable and editorially accurate, and if reviewers find that the claim table format surfaces information they would have wanted during their review, then the case for native integration follows from the evidence rather than from advocacy.

9. The Website

A website at https://zmainen.github.io/elife-claim-trees/ presents browsable claim graphs for all 10 papers. Each paper has an index page with the claim table, the dependency graph, and the reproduction status for each claim. Individual claim pages record the proposition, the panel, the data pointer, the code pointer, and the reproduction status with notes.

We would welcome eLife sharing the site with the 10 corresponding authors and inviting their responses. Authors are the people best positioned to identify errors in the claim extraction — a claim assigned to the wrong panel, a dependency edge that is incorrect, a reproduction status that mischaracterizes what the deposit enables. They are also the people most likely to have views on whether the claim graph adds information they think a reader should have, or whether it misrepresents the paper in ways the narrative format would not. Author feedback on their own claim graphs is the next most valuable data point after the corpus run itself.

The most useful author responses would address three questions: whether the claims as written accurately represent what the paper asserts; whether the dependency edges correctly capture what depends on what; and whether the reproduction status for each claim matches the author's own knowledge of what the deposited materials enable. Disagreement on any of these points is informative. If authors consistently find that the claim graph makes their papers legible in ways they recognize and endorse, that is a signal. If they consistently find that it misrepresents something that the narrative format represents correctly, that is also a signal.

The prototype is one team's reading of ten papers, performed by a single analyst with assistance, under the constraints described above. Its value is not in any individual claim or status — it is in the pattern across ten papers, in the six methodological issues it surfaced, and in the demonstration that the extraction is feasible and produces editorially relevant information. What it cannot do is validate itself. That requires the authors.

Correspondence: z.mainen@neuro.fchampalimaud.org