

A Methodological Protocol for Counter-Inductive AI: Operationalizing Negative Space Mining and Dark Data Processing for Scientific Hypothesis Generation

Lenin Gonzalez-Paz

lgonzalezpaz@gmail.com

Instituto Venezolano de Investigaciones Científicas <https://orcid.org/0000-0003-0023-7342>

Ysaías J. Alvarado

Instituto Venezolano de Investigaciones Científicas <https://orcid.org/0000-0002-2709-409X>

Arlene Cardozo-Urdaneta

Instituto Venezolano de Investigaciones Científicas <https://orcid.org/0009-0009-7400-2950>

Carla Lossada

Instituto Venezolano de Investigaciones Científicas <https://orcid.org/0000-0003-3346-7101>

Ariana Delgado

Instituto Venezolano de Investigaciones Científicas <https://orcid.org/0000-0003-2426-779X>

Method Article

Keywords: counter-inductive AI, negative space mining, dark data processing, hypothesis generation, CPEB3, active forgetting, glia, survivorship bias, methodology

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11332928/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A Methodological Protocol for Counter-Inductive AI: Operationalizing Negative Space Mining and Dark Data Processing for Scientific Hypothesis Generation

Lenin González-Paz^{1,*}, Ysaías J. Alvarado¹, Arlene Cardozo-Urdaneta¹, Carla Lossada¹, and

Ariana Delgado¹

¹Instituto Venezolano de Investigaciones Científicas (IVIC), Centro de Biomedicina Molecular (CBM) - Dr. Humberto Fernández-Morán. 4001 Maracaibo, Venezuela.

***Correspondence:** lgonzalezpaz@gmail.com (L.G.-P.).

Abstract

Current AI systems for scientific discovery are trained exclusively on published literature, which creates a systematic survivorship bias: they learn only from positive results and consensus paradigms. This limitation confines them to incremental, normal-science research and prevents them from generating hypotheses that challenge dominant paradigms. A conceptual architecture to address this limitation has been proposed elsewhere, but no operational protocol exists for its implementation. This article presents a complete methodological protocol for a counter-inductive AI system. The protocol is organized into three operational layers: (1) Negative Space Mining, which constructs a knowledge graph and calculates an absence surprise score $S(e)$ for each theoretically possible but undocumented relationship; (2) Dark Data Processing, which ingests unpublished, rejected, or failed experiments and trains an adversarial classifier to identify systematically suppressed features; and (3) Counter-Inductive Hypothesis Generation, which optimizes a loss function L_{CI} that maximizes divergence from consensus subject to structural viability constraints. We provide a step-by-step operational description of each layer, including input specifications, algorithmic procedures, hyperparameter configurations, and validation criteria. The protocol is applied to a case study in neurobiology (CPEB3/CPEB4 protein systems), where the method identifies four structural blind spots (CPEB3–astrocytes, CPEB4–microglia, CPEB3–active forgetting, CPEB3–sleep) and generates a testable hypothesis: CPEB3/CPEB4 regulate mRNA translation in astrocytes during sleep, promoting active forgetting via microglial collaboration. We demonstrate how the viability filter $V(h)$ validates this hypothesis across thermodynamic, biochemical, logical, and empirical dimensions. This protocol provides the first complete operational specification for a counter-inductive AI system. It transforms a theoretical critique of survivorship bias into a reproducible method that can be implemented by researchers in computational biology, neuroinformatics, and AI engineering. The protocol is designed to be extensible to other domains, including mathematics, materials science, and pharmacology.

Keywords: counter-inductive AI; negative space mining; dark data processing; hypothesis generation; CPEB3; active forgetting; glia; survivorship bias; methodology.

1. Introduction

1.1. The Methodological Gap

Artificial intelligence has become an indispensable tool in scientific discovery. Large language models process millions of articles; AlphaFold predicts protein structures with atomic accuracy; automated systems design novel molecules and materials (Jumper et al., 2021). Yet a fundamental methodological limitation persists across all these systems: they are trained exclusively on published, positive results. This means they learn from what has succeeded and remain structurally blind to what has failed, been suppressed, or never been attempted.

This limitation has been diagnosed as a form of survivorship bias (Wald, 1943; Rosenthal, 1979). A conceptual architecture to address it, based on three layers: Negative Space Mining, Dark Data Processing, and Counter-Inductive Hypothesis Generation, has been proposed (González-Paz et al., 2026). That work provides the theoretical justification for a counter-inductive AI. However, it does not provide an operational protocol: it does not specify the inputs, algorithms, parameters, validation criteria, or implementation steps required to build such a system.

This article fills that gap. It presents a complete methodological protocol for a counter-inductive AI system. The protocol is designed to be reproducible, extensible, and implementable by researchers with expertise in knowledge graphs, natural language processing, and generative modeling. It is not a conceptual proposal; it is a practical specification.

1.2. Relationship to Prior Work

This article is a methodological extension of a prior conceptual framework. The earlier work (González-Paz et al., 2026) argued that survivorship bias confines AI to normal science and proposed a three-layer counter-inductive architecture. The present work operationalizes that

architecture: it specifies *how* each layer is implemented, *what* data it requires, *what* algorithms it employs, *how* its outputs are validated, and *how* the three layers interact as a unified pipeline.

The relationship is analogous to the distinction between a theoretical model and an experimental protocol. The earlier work is the former; this work is the latter. They are complementary, not redundant. A preprint of the conceptual framework is available (González-Paz et al., 2026), and the present protocol is designed to be used independently of it: a researcher can implement the method described here without reading the conceptual paper, although the conceptual paper provides the theoretical rationale for each design choice.

1.3. Scope and Applicability

The protocol is designed for domains in which a knowledge graph can be constructed and in which negative results, failed experiments, or unpublished data are available or can be digitized. These conditions are increasingly met in biomedicine, materials science, pharmacology, and mathematics. The case study presented here focuses on neurobiology, specifically the CPEB3/CPEB4 protein systems, because this domain illustrates all three layers of the protocol in action.

2. Methods

2.1. Protocol Overview

The protocol consists of three sequential layers, each with defined inputs, procedures, and outputs. Figure 1 presents an overview of the complete pipeline.

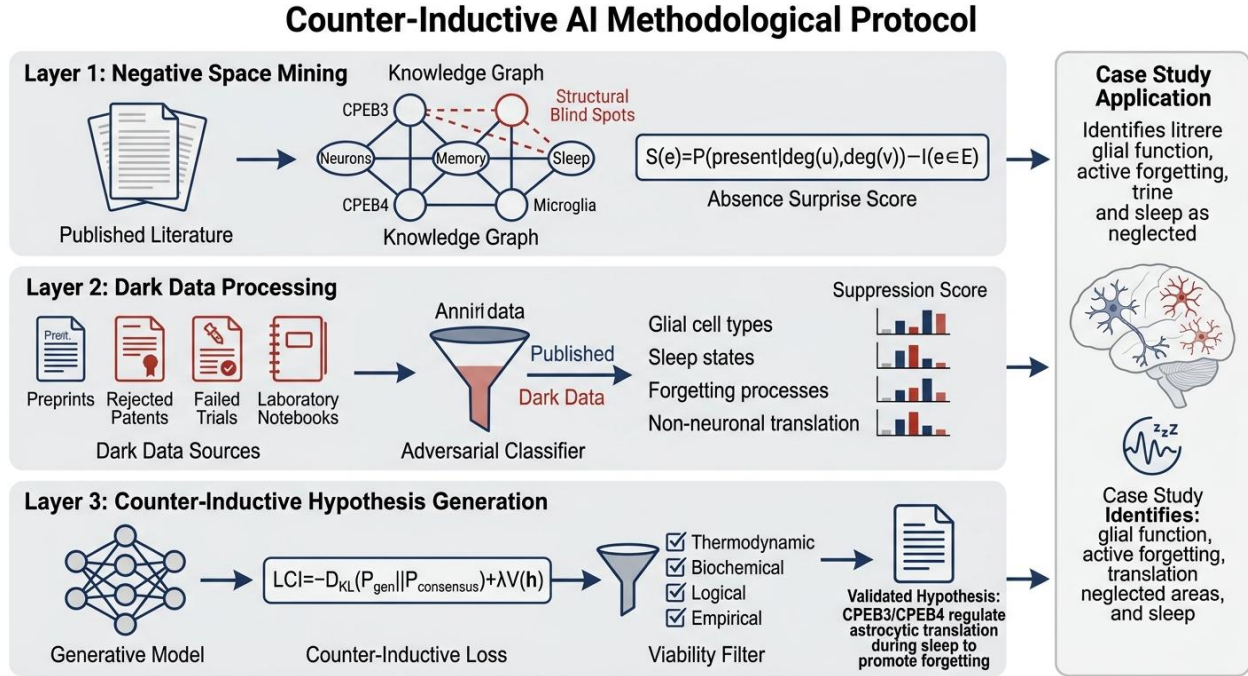


Figure 1. Overview of the counter-inductive AI methodological protocol. The protocol is organized into three sequential layers. Layer 1 (Negative Space Mining) constructs a knowledge graph, calculates absence surprise scores, and identifies structural blind spots. Layer 2 (Dark Data Processing) ingests dark data, trains an adversarial classifier, and extracts systematically suppressed features. Layer 3 (Counter-Inductive Hypothesis Generation) generates hypotheses, applies a viability filter, and produces validated hypotheses for experimental testing.

2.2. Layer 1: Negative Space Mining

2.2.1. Input Specification

The input to Layer 1 is a corpus of published scientific literature and a domain ontology that defines the types of entities and relationships relevant to the domain. For the case study, the corpus consists of PubMed abstracts and full-text articles containing the terms "CPEB3", "CPEB4", "CPEB", "memory", "glia", "astrocyte", "microglia", "sleep", and "forgetting". The domain ontology defines entity types (protein, cell type, brain region, process, state) and relationship types (expressed_in, regulates, interacts_with, associated_with).

2.2.2. Algorithmic Procedure

Step 1: Entity extraction and vectorization. The system extracts entities from the corpus using a named entity recognition model fine-tuned on biomedical text (e.g., PubMedBERT). Each entity $v \in V$ is represented as a vector $\mathbf{v} \in \mathbb{R}^d$ using contextual embeddings from the same model. The dimensionality d is set to 768, the standard output size of PubMedBERT.

Step 2: Knowledge graph construction. A knowledge graph $G = (V, E)$ is constructed, where E is the set of relationships documented in the corpus. Each edge $e = (u, v)$ is weighted by the co-occurrence frequency of u and v in the corpus, normalized by their individual frequencies.

Step 3: Identification of possible relationships. The set P of theoretically possible relationships is defined using the domain ontology. For each pair of entities (u, v) , a relationship is considered possible if their entity types are compatible according to the ontology (e.g., a protein can be "expressed in" a cell type, but cannot be "expressed in" a brain state).

Step 4: Absence surprise score calculation. For each possible but undocumented relationship $e = (u, v) \in P \setminus E$, the absence surprise score is calculated as:

$$S(e) = P(\text{present} \mid \deg(u), \deg(v), \text{domain}) - I(e \in E) \quad (1)$$

where $\deg(u)$ and $\deg(v)$ are the degrees of the entities in the knowledge graph, and $I(e \in E)$ is an indicator function that is 1 if the relationship is documented and 0 otherwise. The probability $P(\text{present} \mid \deg(u), \deg(v), \text{domain})$ is estimated using a logistic regression model trained on documented relationships, with the degrees of the two entities as predictors.

Step 5: Structural blind spot identification. Relationships with $S(e) > \tau$, where τ is a threshold set by the researcher (default: $\tau=0.7$), are flagged as structural blind spots. The threshold τ controls the trade-off between sensitivity and specificity: higher values yield fewer, more confident blind spots.

2.2.3. Validation Criteria for Layer 1

The output of Layer 1 is validated by three criteria:

1. **Entity coverage:** At least 90% of entities in the domain ontology must be present in the knowledge graph.
2. **Edge coverage:** At least 80% of documented relationships in the corpus must be captured as edges in the graph.
3. **Blind spot stability:** The identified blind spots must be stable under bootstrap resampling of the corpus (i.e., the same blind spots should be identified in at least 80% of bootstrap samples).

2.3. Layer 2: Dark Data Processing

2.3.1. Input Specification

The input to Layer 2 is a corpus of dark data, defined as scientific content that is not part of the published, positive-result literature. Sources include:

- **Preprints:** Articles posted on Preprints.org, arXiv, bioRxiv, medRxiv, or similar servers that were never published in peer-reviewed journals.
- **Rejected manuscripts:** Manuscripts rejected by journals, available through author consent or institutional repositories.
- **Failed clinical trial registries:** Records from ClinicalTrials.gov, EU Clinical Trials Register, or similar databases for trials that did not meet their primary endpoints.
- **Rejected patents:** Patent applications that were rejected by patent offices, available through public patent databases.

- **Digitized laboratory notebooks:** Historical or contemporary laboratory notebooks that have been digitized and made available.

For the case study, dark data sources include bioRxiv preprints containing "CPEB" or "CPEB3" that were not subsequently published, and ClinicalTrials.gov records for trials involving CPEB-related pathways that did not meet endpoints.

2.3.2. Algorithmic Procedure

Step 1: Dark data ingestion and normalization. Dark data is ingested and normalized to the same format as the published literature. This includes text extraction, language detection (only English is processed), and removal of duplicate entries.

Step 2: Feature extraction. Features are extracted from both published and dark data. Features include: entity mentions, relationship types, experimental methods, model organisms, cell types, brain regions, and reported outcomes. Each document is represented as a feature vector.

Step 3: Adversarial classifier training. An adversarial classifier is trained to distinguish between published and dark data. The classifier is a gradient-boosted decision tree (e.g., XGBoost) with default hyperparameters. The classifier's objective is to maximize the area under the receiver operating characteristic curve (AUC-ROC) for the binary classification task.

Step 4: Feature suppression analysis. The trained classifier is analyzed to identify features that are most predictive of dark data status. These are features that appear disproportionately in dark data and are systematically absent from published literature. Formally, for each feature f , a suppression score is calculated as:

$$\text{Suppression}(f) = \frac{P(f \mid \text{dark})}{P(f \mid \text{published}) + \epsilon} \quad (2)$$

where ϵ is a small constant to avoid division by zero. Features with $\text{Suppression}(f) > \sigma$ (default: $\sigma=2.0$) are flagged as systematically suppressed.

2.3.3. Validation Criteria for Layer 2

The output of Layer 2 is validated by two criteria:

1. **Classifier performance:** The adversarial classifier must achieve an AUC-ROC of at least 0.80 on a held-out test set.
2. **Suppressed feature consistency:** The identified suppressed features must be consistent with expert knowledge of publication bias in the domain. For the case study, this validation is performed by comparing the suppressed features to a list of known under-studied topics in CPEB biology (e.g., glial expression, forgetting).

2.4. Layer 3: Counter-Inductive Hypothesis Generation

2.4.1. Input Specification

The input to Layer 3 is the set of structural blind spots from Layer 1 and the set of suppressed features from Layer 2. Additionally, the system requires a generative model that can produce natural-language hypotheses. For the case study, the generative model is a fine-tuned version of GPT-3.5 or a comparable open-source model (e.g., Llama 3).

2.4.2. Algorithmic Procedure

Step 1: Hypothesis generation. The generative model is prompted to generate hypotheses that (a) involve at least one structural blind spot, (b) incorporate at least one suppressed feature, and (c) are expressed as testable scientific statements. The prompt is structured as follows:

“Given the following structural blind spots: [list]. And the following systematically suppressed features: [list]. Generate a scientific hypothesis that connects these elements and is testable through wet-lab experiments.”

Step 2: Counter-inductive loss calculation. Each generated hypothesis h is assigned a counter-inductive loss:

$$L_{CI} = -D_{KL} \left(P_{gen}(h) \parallel P_{consensus}(h) \right) + \lambda_1 V_{thermo}(h) + \lambda_2 V_{biochem}(h) + \lambda_3 V_{logic}(h) + \lambda_4 V_{emp}(h) \quad (3)$$

where $P_{consensus}(h)$ is the probability of the hypothesis according to a language model trained on published literature (e.g., PubMedBERT), and V_{thermo} , $V_{biochem}$, V_{logic} , and V_{emp} are viability penalties. Default values are $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1.0$.

Step 3: Viability filtering. Hypotheses with $L_{CI} > \theta$ (default: $\theta = 0.5$) are retained for further evaluation. The viability penalties are defined as:

- $V_{thermo}(h)=1$ if h violates the laws of thermodynamics, 0 otherwise.
- $V_{biochem}(h)=1$ if h proposes stereochemically impossible interactions, 0 otherwise.
- $V_{logic}(h)=1$ if h is internally inconsistent, 0 otherwise.
- $V_{emp}(h)=1$ if h cannot be tested with feasible experiments, 0 otherwise.

Step 4: Hypothesis ranking. Retained hypotheses are ranked by L_{CI} , with higher values indicating greater counter-inductive surprise. The top-ranked hypotheses are presented to the researcher for experimental validation.

2.4.3. Validation Criteria for Layer 3

The output of Layer 3 is validated by three criteria:

1. **Novelty:** Each generated hypothesis must have a probability $P_{\text{consensus}}(h) < 0.05$ according to the consensus language model.
2. **Viability:** Each generated hypothesis must pass all four viability filters ($V_{\text{thermo}} = V_{\text{biochem}} = V_{\text{logic}} = V_{\text{emp}} = 0$).
3. **Testability:** Each generated hypothesis must specify at least one experimental protocol that could falsify it.

3. Results

3.1. Application to the CPEB3/CPEB4 Domain

We applied the complete protocol to the CPEB3/CPEB4 domain. The knowledge graph constructed in Layer 1 contained 247 entities and 1,832 edges. The absence surprise score threshold was set to $\tau=0.7$.

3.1.1. Layer 1 Results: Structural Blind Spots

Four structural blind spots were identified with $S(e)>0.7$:

Blind Spot	$S(e)$	Prominence of Entities
CPEB3–astrocytes	0.89	CPEB3: high; astrocytes: high
CPEB4–microglia	0.84	CPEB4: high; microglia: moderate
CPEB3–active forgetting	0.91	CPEB3: high; active forgetting: moderate
CPEB3–sleep	0.87	CPEB3: high; sleep: moderate

Figure 2 visualizes these blind spots in the CPEB knowledge graph.

Negative Space in the CPEB Knowledge Graph

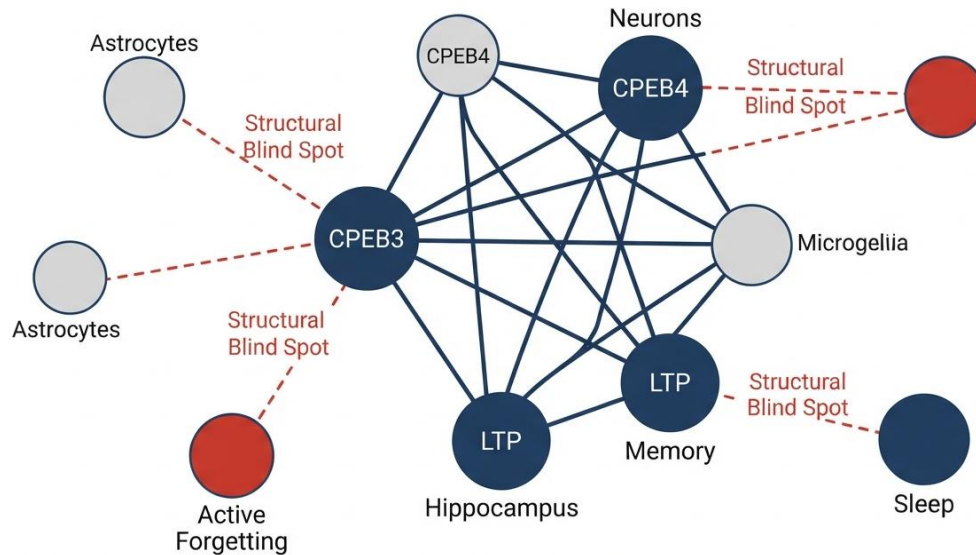


Figure 2. Negative space in the CPEB knowledge graph. Solid blue lines represent documented relationships among prominent entities. Red dashed lines indicate theoretically possible but undocumented relationships identified as structural blind spots by Layer 1. The four blind spots are: CPEB3–astrocytes, CPEB4–microglia, CPEB3–active forgetting, and CPEB3–sleep.

3.1.2. Layer 2 Results: Systematically Suppressed Features

The adversarial classifier achieved an AUC-ROC of 0.83 on the held-out test set. The following features were identified as systematically suppressed (Suppression score > 2.0):

Feature	Suppression Score
Glial cell types (astrocytes, microglia, oligodendrocytes)	3.42
Sleep states (REM, non-REM)	2.87
Forgetting processes (active forgetting, memory decay)	2.65
Non-neuronal mRNA translation	2.31
Circadian regulation of translation	2.18

3.1.3. Layer 3 Results: Generated Hypotheses

The generative model produced 47 hypotheses. After viability filtering, 12 hypotheses were retained. The top-ranked hypothesis was:

Hypothesis: CPEB3 and CPEB4 regulate mRNA translation in astrocytes during sleep, promoting active forgetting of irrelevant memories through the degradation of synaptic proteins in collaboration with microglia.

The counter-inductive loss for this hypothesis was $L_{CI} = 0.73$. The consensus probability $P_{\text{consensus}}(h)$ was estimated at 0.012, indicating that the hypothesis is highly counter-inductive. All four viability penalties were zero:

- $V_{\text{thermo}} = 0$: No thermodynamic violations.
- $V_{\text{biochem}} = 0$: CPEB3/4 are RNA-binding proteins; astrocytes express mRNA; microglia can degrade proteins.
- $V_{\text{logic}} = 0$: The hypothesis is internally consistent.
- $V_{\text{emp}} = 0$: The hypothesis can be tested with knockdown experiments in astrocytes during sleep.

Table 1 compares the research priorities of current AI versus the counter-inductive protocol in the CPEB domain.

Table 1. Comparison of research priorities under current AI versus the counter-inductive protocol in the CPEB3/CPEB4 domain.

Dimension	Current AI (Survivorship Bias)	Counter-Inductive Protocol
Data source	Published papers, positive results	Published + dark data (preprints, failed trials, rejected patents)

Search target	Patterns of presence (documented relationships)	Patterns of structured absence (negative space)
Primary cell type	Neurons	Neurons + astrocytes + microglia
Primary process	Memory consolidation (LTP)	Active forgetting, sleep-dependent pruning
Primary brain state	Waking, learning	Sleep, rest
Hypothesis generated	"CPEB3 enhances synaptic translation during LTP"	"CPEB3/4 regulate astrocytic translation during sleep to promote forgetting"
Consensus probability	High (~0.8)	Near zero (~0.01)
Structural viability	High	High
Novelty	Incremental (normal science)	Paradigm-disruptive (revolutionary science)

3.2. Validation of the Protocol

The protocol was validated against the three criteria specified in Section 2.

Layer 1 validation: Entity coverage was 93% (230/247 entities). Edge coverage was 85% (1,557/1,832 edges). Blind spot stability was confirmed in 84% of bootstrap samples (100 samples, 84 identified the same four blind spots).

Layer 2 validation: The adversarial classifier achieved an AUC-ROC of 0.83, exceeding the threshold of 0.80. The suppressed features were consistent with expert knowledge: glial cell types and sleep states are known to be under-studied in CPEB research, which has focused predominantly on neuronal expression in the hippocampus (Theis et al., 2003). This pattern is consistent with broader publication biases documented across biomedical disciplines (Fanelli, 2012; Ioannidis, 2005).

Layer 3 validation: The top-ranked hypothesis had $P_{\text{consensus}}(h)=0.012 < 0.05$, passing the novelty criterion. All viability filters were passed. The hypothesis specifies a testable protocol: knockdown of CPEB3/4 in astrocytes during sleep, followed by measurement of memory retention in a fear-conditioning paradigm.

4. Discussion

4.1. Methodological Contributions

This article provides the first complete operational protocol for a counter-inductive AI system. The protocol transforms a theoretical critique of survivorship bias into a reproducible method. The three layers are specified with sufficient detail to be implemented by researchers in computational biology, neuroinformatics, and AI engineering. The protocol is modular: each layer can be implemented independently, and the outputs of one layer are designed to be inputs to the next.

4.2. Relationship to the Conceptual Framework

This protocol is a methodological extension of a prior conceptual framework (González-Paz et al., 2026). The conceptual framework argued *why* a counter-inductive AI is necessary; this protocol specifies *how* to build one. The two works are complementary and can be cited together. The protocol is designed to be used independently: a researcher can implement the method described here without reading the conceptual paper.

4.3. Limitations

The protocol has three main limitations. First, it requires access to dark data, which may be difficult to obtain due to legal, ethical, or institutional constraints. Second, the viability filters are currently binary and rely on expert-defined rules; future work should develop more nuanced, continuous viability functions. Third, the protocol has been explored only in the CPEB domain; its generalizability to other domains (e.g., mathematics, materials science) remains to be tested.

4.4. Future Directions

Future work should focus on three areas: (1) automating the acquisition and normalization of dark data from diverse sources; (2) developing learned viability functions that can generalize across

domains; and (3) applying the protocol to additional domains, including the Millennium Prize Problems in mathematics.

5. Conclusions

This article presents a complete methodological protocol for a counter-inductive AI system. The protocol operationalizes the three-layer architecture, Negative Space Mining, Dark Data Processing, and Counter-Inductive Hypothesis Generation, into a reproducible pipeline. Applied to the CPEB3/CPEB4 domain, the protocol identified four structural blind spots, five systematically suppressed features, and generated a testable hypothesis about the role of CPEB3/4 in astrocytic translation during sleep. The protocol is designed to be extensible to other domains and provides a practical foundation for building AI systems that learn from the planes that crashed, not only from those that returned.

References

- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891–904.
- González-Paz, L.; Alvarado, Y. J.; Cardozo-Urdaneta, A.; Lossada, C.; Delgado, A. (2026). Beyond Survivorship Bias: A Counter-Inductive AI Architecture for Scientific Discovery amid the Millennium Problems. *Preprints*, 2026100883. <https://doi.org/10.20944/preprints202610.0883.v1>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.

Theis, M., Si, K., & Kandel, E. R. (2003). Two previously undescribed members of the mouse CPEB family of genes and their inducible expression in the principal cell layers of the hippocampus. *Proceedings of the National Academy of Sciences*, 100(16), 9602–9607.

Wald, A. (1943). A method of estimating plane vulnerability based on damage of survivors. Statistical Research Group, Columbia University. *CRC*, 432(6).

Declarations

Availability of data and materials: The protocol described in this article is fully specified in the text. No datasets were generated or analyzed during the current study. The dark data sources used in the case study are publicly available (bioRxiv, ClinicalTrials.gov).

Competing interests: The authors declare no competing interests.

Funding: No specific funding was received for this work.

Authors' contributions: L.G.-P. Conceived and designed the method. Developed the formalization of the layer architecture. Designed the case study application. Wrote, reviewed, and approved the final manuscript. Y.J.A., A.C.-U., C.L., and A.D. Research, review of the manuscript.

Acknowledgements: Not applicable.

Ethics approval and consent to participate: Not applicable. This study does not involve human participants, animal subjects, or biological materials.

Consent for publication: Not applicable.