

# A Living Review Pipeline for AI/ML Applications in Accelerator Physics

A. Ghribi

*CNRS – GANIL, Caen, France*

(Dated: October 14, 2025)

We present an open-source pipeline for generating a *living review* of artificial intelligence (AI) and machine learning (ML) applications in accelerator physics and technologies. Traditional review articles provide static snapshots that are quickly outdated by the rapid pace of research. The presented system automatically harvests publications from multiple bibliographic sources (arXiv, InspireHEP, HAL, OpenAlex, Crossref, and Springer), deduplicates entries, applies semantic filtering to ensure accelerator and ML relevance, and classifies papers into thematic categories. The resulting curated dataset was exported in JSON, HTML, PDF, and BibTEXformats, enabling continuous updates and integration with web frameworks. We describe the methodology, including semantic similarity filtering using sentence-transformer embeddings, threshold calibration, and expert-informed classification. The results demonstrate the robust filtering of  $\sim 12000$  raw papers/month into a focused corpus of  $\sim 2\%$  relevant works. The pipeline provides the basis for an evolving community-driven review of AI/ML in accelerator science.

## I. INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) are reshaping the way scientific research is conceived and conducted. In the field of particle accelerators, these methods are increasingly employed for beam dynamics optimisation, feedback control, anomaly detection, surrogate modelling, computer vision-based diagnostics, and reinforcement learning for autonomous tuning. The pace of progress is accelerating: while early efforts were isolated demonstrations, entire research programs are now devoted to exploring AI and ML as key enablers of next-generation accelerator design and operation [1–3].

Traditional review articles, while invaluable for consolidating knowledge, inevitably provide only a *static snapshot* of a rapidly evolving landscape of research. Given the exponential growth in AI-related publications and the continuous diversification of methods, such static reviews become outdated within a few months. This challenge has already been recognised in high-energy physics (HEP), where the community maintains the *Living Review of Machine Learning for Particle Physics* [4, 5]. It is worth noting that, alongside the Living Review of Machine Learning for Particle Physics [6], a community fork exists that applies a similar InspireHEP-based workflow to accelerator-related studies[7]. Continuously updated resources of this kind, curated by experts, are essential for keeping pace with emerging cross-disciplinary advances. This study introduces a complementary, fully automated, multi-source, and FAIR-compliant pipeline for accelerator science.

Inspired by this model, we present an open-source, fully automated pipeline for generating a living review of artificial intelligence (AI) and machine learning (ML) applications in accelerator physics. Unlike traditional reviews that require extensive manual updates, our approach automates the process end-to-end: it harvests publications from multiple bibliographic databases, deduplicates entries, applies semantic filtering to retain only those works at the intersection of accelerators and

ML, classifies them into thematic categories, and exports the curated dataset in interoperable formats—JSON, HTML, BibTEX, and PDF.

This work aims to provide the accelerator community with the following:

- a continuously updated and openly accessible map of AI/ML research relevant to accelerator science;
- a transparent and reproducible methodology for literature filtering and classification;
- standardized outputs that integrate seamlessly with web platforms, data repositories, and citation workflows.

By establishing a living, FAIR-compliant survey of the field, we aim to catalyse collaboration, reduce duplication of effort, and accelerate the responsible adoption of AI/ML methods in accelerator research and operations. The living review is accessible on <https://aghribi.github.io/acc-ml-living-review/> and contributions remain open through the following repository <https://github.com/aghribi/acc-ml-living-review>.

## II. SYSTEM ARCHITECTURE

The *living review* framework is organized as a modular pipeline (Fig. 1) designed to be robust, extensible, and reproducible. Each module corresponds to a logically independent task, allowing future updates or replacement without modifying the rest of the workflow. The main components are summarised as follows.

### A. Fetchers

Dedicated routines query multiple bibliographic databases relevant to accelerator physics and engineering, including arXiv, InspireHEP, HAL, OpenAlex and

Crossref. Each fetcher implements a uniform interface and returns a list of structured `Paper` objects (title, authors, abstract, venue, date, identifiers, and links). This redundancy ensures broad coverage across physics, engineering, and computer science venues.

## B. Deduplication

Because the same article may appear across several sources (e.g., a preprint on arXiv, a conference entry in Inspire, and a DOI record in Crossref), a deduplication step merges duplicates based on canonical identifiers (DOI, arXiv ID) or fuzzy title matching. The resulting dataset contains only unique entries.

## C. Semantic Filtering and Exclusion

To retain only works at the intersection of AI/ML and accelerator science, and to suppress domain noise, the pipeline combines *positive*, *negative*, and *neutral* semantic queries.

Each paper is represented by a joint embedding of its title and abstract using a lightweight transformer model (*MiniLM*) [8]. Three query embeddings are used: an “accelerator physics” reference query, a “machine learning” query, and a “noise” query capturing unrelated topics (e.g., particle detectors, HEP calorimetry, atomic spectroscopy, and infrastructure computing). The semantic similarity of each paper to these anchors is computed as  $s_{\text{accel}}(p)$ ,  $s_{\text{ml}}(p)$ , and  $s_{\text{noise}}(p)$ . A paper is retained only if it satisfies the following criteria:

$$\begin{aligned} \text{keep}(p) = & [s_{\text{accel}}(p) \geq \theta_{\text{accel}}] \\ & \wedge [s_{\text{ml}}(p) \geq \theta_{\text{ml}}] \\ & \wedge [s_{\text{accel}}(p) > s_{\text{noise}}(p)]. \end{aligned} \quad (1)$$

This ensures that a paper is simultaneously relevant to both accelerators and ML, while being semantically distant from the noise domains.

In addition, a curated list of *negative keywords* (e.g., “Higgs,” “dark matter,” “FPGA accelerator,” “beam welding,” “earthquake”) is applied to explicitly remove papers whose content is outside the scope of accelerator science or pertains to hardware engineering rather than scientific accelerators. Together, the semantic and keyword filters achieve high precision in isolating the relevant literature.

## D. Classification

Papers passing the filtering step are classified into thematic categories such as *Beam Dynamics and Control*, *Diagnostics*, *HPC and Data Management*, *Medical Applications*, and *Novel Tools and Libraries*. The classifier uses semantic similarity against predefined category

descriptions, complemented by keyword heuristics and rule-based overrides (e.g. “surrogate model” or “review” papers).

## E. Statistics

Aggregate statistics provide a global perspective on the field. These include publication counts per year, per venue, and per category, as well as keyword trends, and monthly publication dynamics. All metrics are exported alongside list of papers for downstream visualisation.

## F. Exporters

The curated results were exported in multiple interoperable formats.

- **JSON:** structured data for programmatic access and website integration;
- **HTML:** human-readable, interactive review pages rendered with Jinja2 templating engine;
- **BibTeX:** citation-ready files for reference managers;
- **PDF:** automatically generated static summaries using the ReportLab toolkit.

This modular design enables continuous updates, straightforward integration with web frameworks, and reproducible curation of AI/ML literature on accelerator physics.

## III. METHODOLOGY

The goal of the pipeline is to identify and structure the subset of scientific literature that lies at the intersection of accelerator physics and machine learning. Achieving this requires moving beyond keyword matching toward semantic representations capable of generalizing across disciplines, and publication venues.

### A. Semantic Embeddings

Each paper  $p$  is represented as a dense embedding vector  $\mathbf{e}(p) \in \mathbb{R}^d$  is computed from its title and abstract. We employ the `sentence-transformers/all-MiniLM-L6-v2` model from the *Sentence-Transformers* library [8, 9], a lightweight transformer pretrained on general-purpose natural-language inference and Semantic textual similarity tasks. The model maps natural-language text into a semantic space, where related concepts are close in cosine similarity. Formally,

$$\mathbf{e}(p) = f_\theta(\text{title}(p) \parallel \text{abstract}(p)), \quad (2)$$

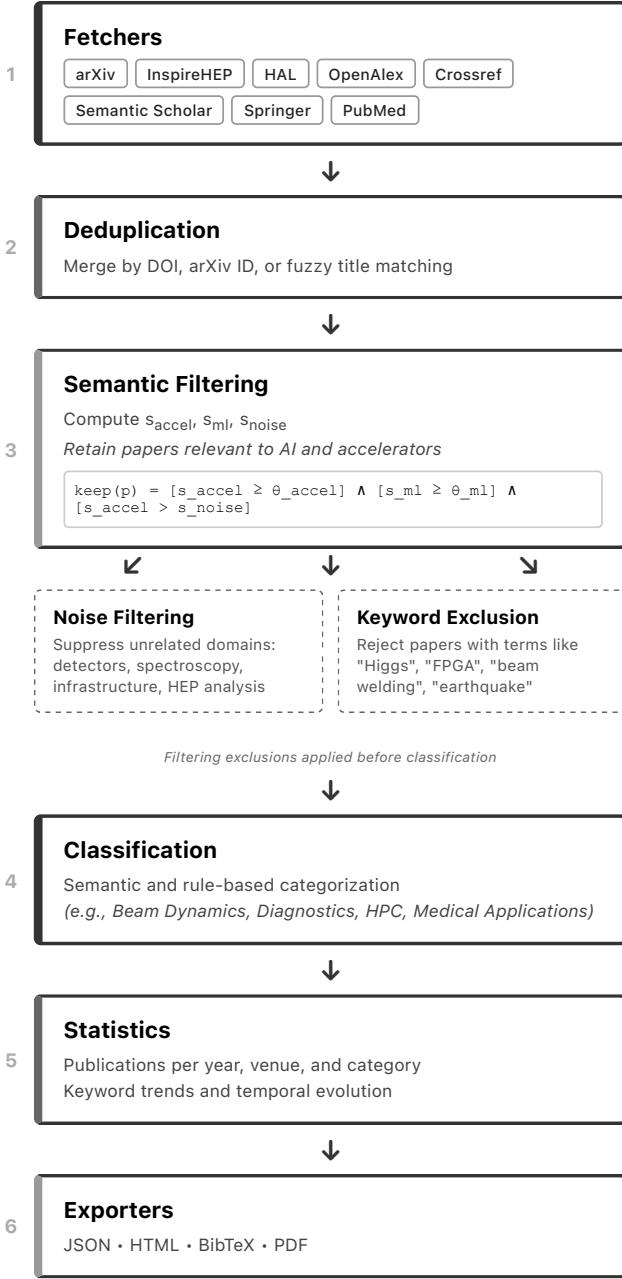


FIG. 1. System architecture of the `living_review` pipeline. Publications are collected from multiple sources, deduplicated, filtered by semantic relevance (including noise suppression and keyword exclusion), classified, and exported in multiple formats for dissemination.

where  $f_\theta$  denotes the transformer encoder and  $\parallel$  represents the concatenation.

## B. Reference Anchors

To quantify relevance, the embedding  $\mathbf{e}(p)$  is compared to three reference vectors representing the accelerator,

machine-learning, and noise domains:

$$s_{\text{accel}}(p) = \cos(\mathbf{e}(p), \mathbf{e}_{\text{accel}}), \quad (3)$$

$$s_{\text{ml}}(p) = \cos(\mathbf{e}(p), \mathbf{e}_{\text{ml}}), \quad (4)$$

$$s_{\text{noise}}(p) = \cos(\mathbf{e}(p), \mathbf{e}_{\text{noise}}). \quad (5)$$

Here,  $\mathbf{e}_{\text{accel}}$  encodes a curated accelerator-physics reference corpus,  $\mathbf{e}_{\text{ml}}$  represents the AI/ML corpus, and  $\mathbf{e}_{\text{noise}}$  captures unrelated domains such as particle detectors, spectroscopy, high-energy physics analyses, or infrastructure computing. The cosine similarity is defined as

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}. \quad (6)$$

## C. Thresholding and Exclusion Rule

A paper is retained only if it is simultaneously relevant to both the accelerator and ML domains and sufficiently distant from the noise domains:

$$\text{keep}(p) = [s_{\text{accel}}(p) \geq \theta_{\text{accel}}] \wedge [s_{\text{ml}}(p) \geq \theta_{\text{ml}}] \wedge [s_{\text{accel}}(p) > s_{\text{noise}}(p)]. \quad (7)$$

Default thresholds  $\theta_{\text{accel}} = 0.13$  and  $\theta_{\text{ml}} = 0.18$  were empirically tuned to balance recall (capturing genuine AI-for-accelerator papers) and precision (excluding tangential or generic ML work).

In addition to the semantic filter, a curated list of *negative keywords*—including, for example, “Higgs,” “calorimeter,” “FPGA,” “chip,” “beam welding,” and “earthquake”—is used to explicitly exclude papers outside the scope of accelerator science, or focused on unrelated hardware engineering. This dual semantic–lexical filtering strategy minimizes false positives and improves dataset quality.

## D. Classification

For papers passing the relevance filter, thematic classification is applied. A hybrid approach was used:

1. **Keyword-assisted rules:** a curated list of accelerator and ML terms (e.g., “beam dynamics”, “reinforcement learning”, “uncertainty quantification”, “proton therapy”) is matched against titles and abstracts.
2. **Semantic clustering (optional):** embeddings may also be grouped using unsupervised clustering to reveal emerging topics not yet captured using predefined categories.

## E. Statistics and Trends

From the curated corpus, descriptive statistics are computed, including:

- publication counts per year and per venue;
- category distributions;
- keyword frequencies over time;
- monthly publication trends.

All metrics are exported to structured JSON for downstream visualization on the Hugo-based website.

#### F. Export and Reproducibility

All outputs (JSON, HTML, BibTeX, and PDF) are automatically generated from the same curated dataset to ensure reproducibility. The pipeline code is version-controlled and designed for continuous re-execution as new publications appear, enabling a fully automated and verifiable living Review of AI/ML applications in accelerator physics.

The resulting dataset forms the basis for the analyses and case studies presented in Sec. IV.

## IV. RESULTS AND CASE STUDIES

### A. Overall Corpus

The pipeline was executed on publications spanning 2000–2025, yielding a curated corpus of  $N = 244$  papers at the intersection of machine learning and accelerator physics. This represents less than 1% of the total records initially retrieved from arXiv, InspireHEP, HAL, OpenAlex, and Crossref, underscoring the selectivity and precision of the semantic filtering process.

### B. Temporal Trends

Figure 2 presents the yearly publication count. The first identified paper dates to 2000, describing early neural-network-based beam diagnostics. A pronounced acceleration begins after 2016 and becomes explosive from 2021 onward, with 70 and 68 papers published in 2024 and 2025, respectively. This corresponds to a three-fold increase relative to the 2016–2020 period, confirming the rapid mainstream adoption of AI in accelerator research.

### C. Category Distribution

Automatic classification revealed 16 thematic groups. The most populated are:

- **Novel Applications** (183 papers,  $\sim 35\%$ ) — cross-domain uses of AI within accelerator environments, including simulation surrogates, anomaly detection, and experimental optimisation.

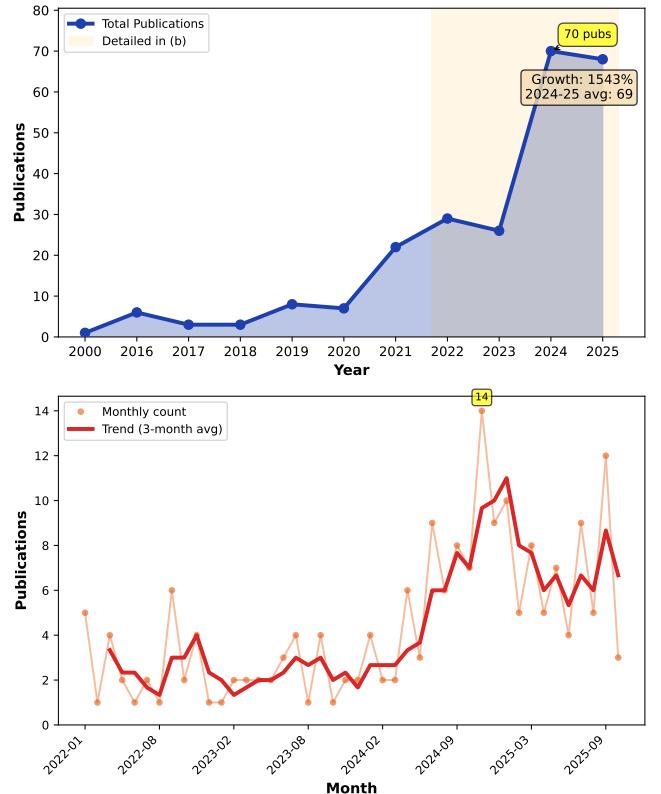


FIG. 2. Publication trends analysis showing temporal patterns at two scales: (a) Annual overview from 2000–2025 displaying long-term growth trajectory with significant acceleration from 2021 onwards and peak activity in 2024–2025 (highlighted region); (b) Detailed monthly activity from 2022–2025 revealing short-term fluctuations and sustained productivity with notable peaks, including 3-month moving average trend line for clarity.

- **Tools and Libraries** (116) — reusable software frameworks, datasets, and benchmark platforms.
- **Surrogate Models** (34) — fast emulators of beamline and RF components;
- **Reinforcement Learning and Autonomous Systems** (28) — adaptive control and tuning strategies
- **Beamline Design and Simulation** (29) — optimization of optical lattices and injection systems.

Smaller but active areas include *Anomaly Detection and Fault Prediction* (22), *Data Management* (22), and *Reviews* (15). Figure 3 shows the relative proportions of

### D. Keyword Analysis

Keyword frequency analysis confirms the prominence of “beam” (109 occurrences), “control” (56), and “optimization” (47) as central themes, reflecting the field’s

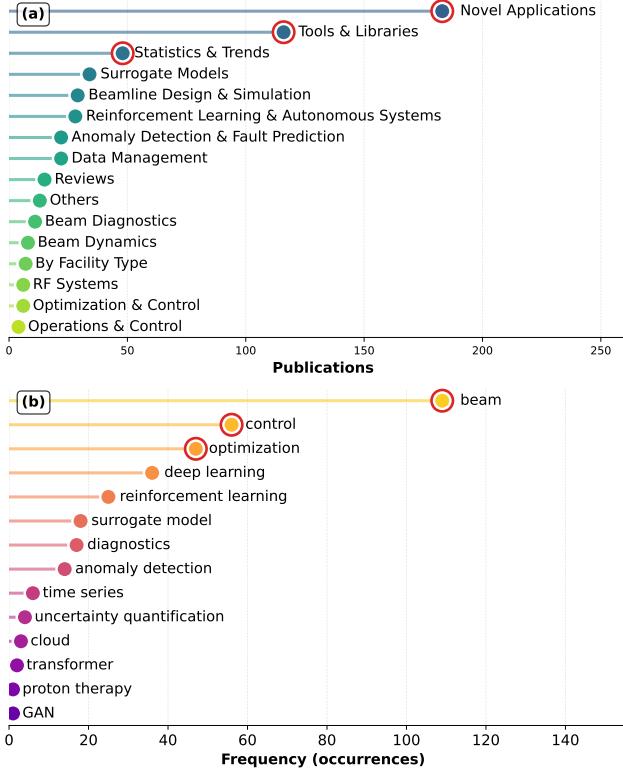


FIG. 3. Thematic distribution: (a) Research categories ranked by publication count ; (b) keyword frequency distribution revealing specific research topics.

strong focus on data-driven tuning and beamline performance. Machine-learning methods such as “deep learning” (36) and “reinforcement learning” (25) appear increasingly in control and diagnostics studies. Emerging topics include “transformers”, “foundation models”, and “federated learning”, indicating the diffusion of advanced AI architectures in accelerator contexts. Mentions of “proton therapy” and “GAN” highlight the growing overlap with medical and imaging application.

### E. Recent Momentum

Monthly publication dynamics reveal sustained high activity throughout 2024–2025, with peaks in November 2024 (14 papers) and September 2025 (12). This consistency suggests the formation of a maturing and expanding research community at the AI–accelerator interface, supported by cross-institutional collaborations and open data-sharing efforts.

### F. Case Studies

Three representative case studies illustrate the diversity of approaches.

**1. Beam tuning with reinforcement learning:** actor–critic and policy-gradient algorithms have been applied to storage-ring and linac tuning, achieving order-of-magnitude reductions in setup time compared with the manual procedures.

**2. Medical quality assurance with generative models:** adversarial and diffusion networks perform MRI-to-CT translation for proton-therapy dose prediction and reducing patient imaging requirements.

**3. Surrogate modeling of RF and beamline components:** deep neural networks trained on high-fidelity electromagnetic simulations yield millisecond-scale predictions of cavity fields and beam optics. This enables interactive design exploration.

### G. Comparison to HEP-ML Living Review

Unlike the HEP-ML Living Review [6], which provides a manually curated overview of machine-learning use across high-energy physics, the present pipeline targets the accelerator domain specifically and automates the entire curation processes. This enables near-real-time updates, quantitative trend analysis, and seamless integration with FAIR data infrastructures.

## V. DISCUSSION

### A. Strengths of the Pipeline

The `living_review` pipeline provides a scalable and reproducible as an alternative to traditional manually curated surveys. Its modular design and integration of multiple bibliographic APIs (arXiv, InspireHEP, HAL, OpenAlex, Crossref) enable the automatic retrieval and filtering of thousands of publications in minutes. By relying on semantic embeddings rather than keyword matching, the system robustly identifies works situated at the accelerator–ML interface, reducing bias and capturing cross-disciplinary research that conventional search strategies might overlook. Because the entire process is automated, the review can be regenerated on a regular basis, maintaining a continuously updated and FAIR-compliant record of research activity that integrates seamlessly with Open science infrastructure.

### B. Limitations

Several limitations highlight the need for improvement. The use of fixed semantic thresholds ( $\theta_{\text{accel}} = 0.13$  and  $\theta_{\text{ml}} = 0.18$ ) introduces sensitivity to boundary cases, occasionally excluding relevant papers or admitting tan-

gential ones. Some degree of source bias persists, as InspireHEP and Crossref remain dominant, while arXiv preprints are frequently merged or overwritten during deduplication, leading to partial underrepresentation of the preprints. In addition, incomplete metadata—such as missing abstracts, inconsistent author identifiers, or absent DOIs—can limit the accuracy of trend analysis and classification.

### C. Comparison with Manual Reviews

Relative to human-curated reviews, the automated approach optimizes for *completeness*, *timeliness*, and *reproducibility*, albeit with a reduced contextual precision. Expert-led reviews remain invaluable for detailed methodological interpretation, whereas automated curation offers a scalable and unbiased quantitative backbone. In practice, both approaches are complementary: automation ensures broad and reproducible coverage, whereas expert judgment provides depth and validation.

### D. Future Directions

Several enhancements are planned to extend both the robustness and interpretability of the framework

- **Model improvements:** Integration of larger transformer-based language models with domain-adapted embeddings to enhance semantic discrimination and reduces misclassification.
- **Calibration and benchmarking:** Validation of classification thresholds using expert-annotated subsets and creation of benchmark datasets to measure reproducibility and precision across releases.
- **Hierarchical taxonomy:** Development of a multi-level category structure (e.g., *Beam Physics → Control, Diagnostics, Simulation*) to reflect thematic hierarchies and subfields.
- **Explainability:** Implementation of SHAP-based and centroid-based visualizations to interpret classification outcomes and provide transparent insights into category assignments.

- **Uncertainty-aware analytics:** Use of probability-weighted statistics to visualize uncertainty and confidence in category trends within the Living Reviews dashboard.

- **Extended coverage:** Inclusion of additional bibliographic and grey-literature sources (e.g., technical design reports, conference proceedings, and internal notes) to broaden the corpus completeness.

## VI. CONCLUSION AND OUTLOOK

We have introduced the **Living Review pipeline**, an open and reproducible framework for the automated collection, filtering, and classification of research at the intersection of accelerator physics and machine learning. The system integrates multiple bibliographic sources, semantic filtering, and lightweight categorization to produce a machine-curated, FAIR-compliant survey of the field that can be continuously updated as new publications are released.

This work complements the community-driven *Living Review of Machine Learning for Particle Physics* [6] and its accelerator-focused derivatives, which rely on InspireHEP-based workflow. The present framework extends these efforts by introducing a fully automated, multi-source, and semantic-driven approach designed specifically for the accelerator science.

As accelerators become increasingly complex and data-rich, and as AI techniques become integral to their operation, diagnostics, and design, such tools are vital to maintaining a comprehensive and transparent view of the evolving research landscape. Beyond cataloguing progress, the `living_review` framework provides a quantitative foundation for evidence-based policy, collaborative research planning, and the responsibly integrate AI methods across the accelerator ecosystem.

## ACKNOWLEDGMENTS

The author thanks the developers of the HEP-ML Living Review for inspiring this work and acknowledges the open-source tools and APIs (InspireHEP, arXiv, HAL, OpenAlex, and Crossref) that made this project possible.

- 
- [1] Ghribi, Adnan, Cassou, Kevin, Dalena, Barbara, Eichler, Annika, Guler, Hayg, Mistry, Andrew K., Oeftiger, Adrian, Shea, Thomas, Valentino, Gianluca, and Welsch, Carsten P., *Europhysics News* **56**, 15 (2025).
- [2] J. Edelen, M. Biedron, B. Chase, P. Stabile, *et al.*, *IEEE Transactions on Nuclear Science* **67**, 1568 (2020).
- [3] M. Reuter, C. Emma, T. Maxwell, *et al.*, *Physical Review Accelerators and Beams* **26**, 104801 (2023).
- [4] D. Guest, K. Cranmer, and D. Whiteson, *Annual Review of Nuclear and Particle Science* **68**, 161 (2022).
- [5] A. Radovic *et al.*, *Nature* **560**, 41 (2018).
- [6] D. Guest, K. Cranmer, *et al.*, Hepml living review: Machine learning for particle physics, <https://iml-wg.github.io/HEPML-LivingReview/> (2025), accessed October 2025.
- [7] <https://github.com/MALAPA-Collab/AccML-LivingReview>.
- [8] N. Reimers and I. Gurevych, Sentence-bert: Sentence

- embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2020).
- [9] N. Reimers and I. Gurevych, all-minilm-l6-v2: A lightweight general-purpose sentence transformer model, Hugging Face model repository (2021), version released as part of the Sentence-Transformers library.