Preface of the Third International Biochemical Knowledge Extraction Challenge (BiKE)

Edgard Marx[1], Marilia Valli[2], João da Silva e Silva[2], Paulo Viviurka do Carmo[1], Thomas Riechert[3] and Guilherme Lehmann Pereira[2]

[1]*Institute for Applied Informatics, Germany*
[2]*University of São Paulo, Brazil*
[3]*Leipzig University of Applied Science, Germany*

More than half a century of biodiversity research, embedded within scientific literature, holds tremendous promise for propelling various scientific fields forward. When this vast expanse of data is methodically organized and distributed via knowledge graphs, it becomes exponentially more usable. This kind of structured format not only accelerates different areas of research and aids in the design of sustainable, high-value eco-products, but it also helps shape public policies that simultaneously boost scientific innovation and fortify the bioeconomy.

Even with this enormous potential, the bulk of the Web's structured biochemical data is still curated by hand. Because new studies are published at a relentless pace, manual extraction simply cannot scale to accommodate the swelling volume of scientific findings.

The Third International Biochemical Knowledge Extraction Challenge (BiKE) was established to tackle this exact bottleneck. By encouraging the Semantic Web community to pioneer automated data extraction techniques, the initiative strives to make information regarding natural products more accessible, spur eco-friendly discoveries, and highlight the critical importance of global biodiversity.

The following papers were successfully peer-reviewed, accepted for publication, and showcased during the workshop:

- A Multi-Stage Agent Framework with Knowledge Graph Alignment for Natural Product Literature Extraction
- CareWeavers BiKE 2026: BERT-Initialised Graph Embeddings for Natural Product Knowledge Extraction
- BiKE-SPHOTA with Ontology Enhancement

During the review phase, the following teams opted to withdraw their submissions:

- GraphMinds - Microcrispr Pvt. Ltd. by Daksh Sammi, Rahul Kharkwal, Khyati Patni
- Intrepid - University of Lagos by Olusola Afuwape and Others

BiKE 2026: Third International Biochemical Knowledge Extraction Challenge, Co-located with Text2KG at the ESWC 2026, 10 May 2026, Dubrovnik, Croatia
marx@infai.org (E. Marx); marilia.valli@ifsc.usp.br (M. Valli); jvictor.ssilva@ifsc.usp.br (J. d. S. e. Silva); carmo@infai.org (P. V. d. Carmo); thomas.riechert@htwk-leipzig.de (T. Riechert); guilhermelp@usp.br (G. L. Pereira)

Challenge Overview

The BiKE Challenge encouraged researchers to dive into the automated extraction of biochemical data by either adapting established methodologies or inventing entirely new ones. The core objective was to pull relevant insights from biochemical research papers and assemble a Biochemical Knowledge Graph (BKG) based on a designated ontology. Ultimately, a BKG serves as an organized framework that maps out chemical and biological traits sourced from living organisms.

Training and Test Datasets

The core dataset for training and testing was assembled from hundreds of peer-reviewed articles, capturing details on more than 2,521 unique natural product extractions. Chemistry domain experts meticulously vetted this collection, manually annotating four primary properties for each extracted compound. For the scope of this challenge, the models targeted five specific NuBBE properties: (1) compound name (rdfs:label), (2) bioactivity (nubbe:biologicalActivity), (3) source species (nubbe:collectionSpecie), (4) location of collection (nubbe:collectionSite), and (5) extraction method (nubbe:collectionType).

Every paper was retained in the training splits. However, in the evaluation splits, the manually annotated links were stripped away, leaving the test documents entirely detached from the overarching knowledge graph. To facilitate the rebuilding of these links, organizers provided Python scripts using the networkx library. These scripts leveraged BERTopic to classify texts and re-establish graph edges based on topical similarity. To ensure these topics remained highly discriminative, any topic appearing in more than 80% of the dataset was filtered out.

Participants were also challenged to invent alternative graph-reconnection techniques utilizing automatically derived metadata, such as authorship networks, citation graphs, and conference venues.

Competitors received the raw flat dataset, the baseline networkx graph, and for this year a single random fold with a pre-shuffled train/test split. For each category a split contained 60% training and 40% testing data. In each provided split, graph edges were kept intact for training data but erased for test data, with a ready-to-use networkx layout supplied for each. The NatUKE benchmark's complete source code and instructions [1] are openly available at: https://github.com/AKSW/natuke.

Evaluation Metrics

The primary goal of the challenge was to accurately rank the predicted connections that had been masked within the knowledge graph. The hits@k metric, alongside Mean Reciprocal Rank (MRR), is ideal for scenarios where a single correct target document exists. Conversely, mean Average Precision (mAP) and normalized Discounted Cumulative Gain (nDCG) cater to situations with multiple relevant targets. The hits@k metric was selected because it offers a flexible way to evaluate different extraction properties by adjusting the k threshold. In alignment with NatUKE's methodology, the final scores report k values from 1 up to 50 (in multiples of

5), utilizing two distinct performance thresholds: (1) achieving a score of at least 0.50, and (2) achieving a score of at least 0.20. Detailed metric definitions can be found in the NatUKE publication.

Best Knowledge Extraction Awards

The Best Extraction Method award was created to celebrate the top participants who showed remarkable skill, persistence, and a profound grasp of the techniques needed to parse intricate biochemical information. These teams showcased superior problem-solving aptitude, sharp analytical thinking, and a mastery of modern computational frameworks.

Winners were presented with personalized certificates acknowledging their specialized contributions to the workshop. For this third iteration of the BiKE Challenge, the Best Extraction Method awards were presented to:

1st **A Multi-Stage Agent Framework with Knowledge Graph Alignment for Natural Product Literature Extraction**
Shixiong Zhao, Yanming He, Lakshan Karunathilake and Hideaki Takeda

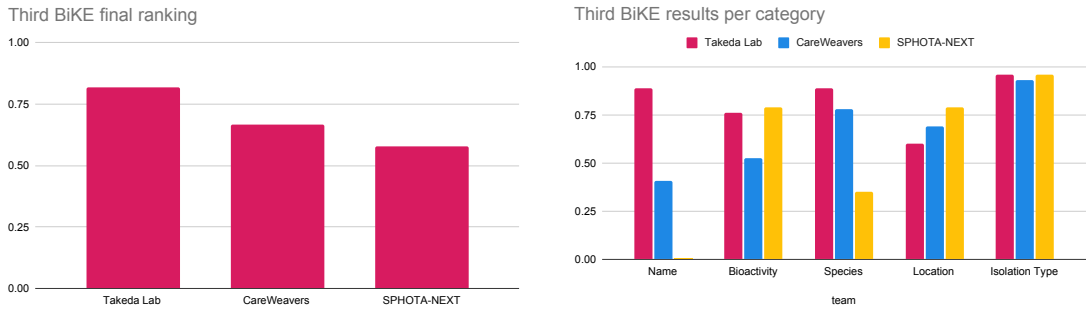
2nd **CareWeavers BiKE 2026: BERT-Initialised Graph Embeddings for Natural Product Knowledge Extraction**
Khyati Patni, Daksh Sammi, Diksha Kashyap and Rahul Kharkwal

3rd **BiKE-SPHOTA with Ontology Enhancement**
Bharath Chand, Thara Prabhakaran and Sanju Tiwari

In Figure 1 we present the overall and extraction task rankings for the best extractions. In 1a we can see a gradual improvement among the average results for all teams. When analyzing by category in 1b we can see that the different strategies yield better results in different. For example, SPHOTA-NEXT achieved the best results for bioactivity and location extractions and achieves 0.00 $hits@50$ for compound name extraction. At the same time the winning team Takeda-Lab achieves the overall good performance across all tasks except for the lowest $hits@20$ for location extraction. This indicates that different types of solutions still perform differently among the biochemical knowledge extraction tasks.

In this challenge, Zhao et al. [2] presents a document-level, multi-stage agent framework designed to extract five core biochemical attributes—compound name, bioactivity, source species, collection site, and isolation type—from natural product research articles. To manage information load and prompt complexity, the proposed pipeline separates tasks into sequential stages: converting raw PDFs to text via Mistral OCR, reading the full text with a large language model to propose candidate values, retrieving localized candidates from property-specific knowledge graphs, and finally making normalized entity decisions using a secondary language model. Experimental results on the BiKE test setting demonstrate the system's strong performance across all target attributes, achieving near state-of-the-art results for compound name extraction and matching the state-of-the-art for isolation type extraction, validating the efficacy of combining multi-stage agent coordination with localized knowledge graph alignment.

Patni et al. [3] introduced a BERT-initialised knowledge graph embedding pipeline for biochemical link prediction on the NatUKE benchmark. The authors' system retrieves paper



(a) Overall ranking.

(b) Ranking by extraction task.

Figure 1: Best knowledge extraction overall and extraction task rankings.

metadata via the CrossRef API, encodes it using SciBERT and PubMedBERT, and constructs a five-relation heterogeneous knowledge graph to connect publications with their corresponding biochemical properties. By systematically evaluating eleven different configurations—including cosine similarity baselines, topology-driven models, and BERT-initialised graph models like RotatE and GraphSAGE—the study reveals that domain-specific text representations significantly enhance structure-only baselines. Ultimately, the research highlights that optimal performance requires per-property model selection rather than a fixed ensemble, as different models excel depending on the specific label-space structure and textual distinctiveness of each target property.

Chand et al. [4] propose the SPHOTA-NEXT framework, an ontology-enhanced extension of their previous work aimed at overcoming the structural sparsity of biochemical knowledge graphs. This novel approach enriches entity representations by aligning compounds with RDKit-derived chemical classifications, generating hierarchical ontology triples that are subsequently injected into K-BERT. By introducing these external domain relationships, the framework creates shared semantic neighborhoods that allow the language model to exploit both textual context and higher-level conceptual categories during embedding generation. Evaluated using the NuBBE DB dataset and NatUKE benchmark, the ontology-driven enrichment significantly improved link prediction accuracy—especially for bioactivity and isolation type properties—proving the value of integrating domain ontologies into language-model-driven knowledge graph completion.

General Chair

- Edgard Marx, Leipzig University of Applied Sciences (HTWK), Germany

Organizing Committee

- Marilia Valli, São Paulo University (USP), Brazil
- João Victor da Silva e Silva, São Paulo University (USP), Brazil

- Paulo Ricardo Viviurka do Carmo, Leipzig University of Applied Sciences (HTWK), Germany

Advisory Committee

- Vanderlan da Silva Bozani, Sao Paulo State University (UNESP), Brazil
- Adriano Defini Andricopulo, Sao Paulo University (USP), Brazil
- Thomas Riechert, Leipzig University of Applied Sciences (HTWK), Germany
- Alan Pilon, Sao Paulo University (USP), Brazil

Acknowledgements

The editors extend their deepest gratitude to the advisory board, the contributing authors, the program committee, and all co-organizers for their tireless dedication to making this workshop a resounding success.

References

[1] P. V. Do Carmo, E. Marx, R. Marcacini, M. Valli, J. V. Silva e Silva, A. Pilon, NatUKE: A Benchmark for Natural Product Knowledge Extraction from Academic Literature, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), 2023, pp. 199–203. doi:10.1109/ICSC56153.2023.00039.

[2] S. Zhao, Y. He, L. Karunathilake, H. Takeda, A multi-stage agent framework with knowledge graph alignment for natural product literature extraction, in: Proceedings of the Third Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC 2026, BiKE'26, CEUR, 2026. Published Digitally.

[3] K. Patni, D. Sammi, R. Kharkwal, D. Kashyap, Careweavers bike 2026: Bert-initialised graph embeddings for natural product knowledge extraction, in: Proceedings of the Third Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC 2026, BiKE'26, CEUR, 2026. Published Digitally.

[4] B. Chand, T. Prabhakaran, S. Tiwari, Sphota-next: Ontology enhanced knowledge injection for biochemical link prediction, in: Proceedings of the Third Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC 2026, BiKE'26, CEUR, 2026. Published Digitally.