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

Edgard Marx[1], Marilia Valli[2], João da Silva e Silva[2] and Paulo Viviurka do Carmo[1]

[1]*Leipzig University of Applied Science, Germany*
[2]*São Paulo University, Brazil*

Over five decades of research on biodiversity, documented in scientific literature, hold immense potential for advancing various scientific domains. When systematically organized and disseminated through knowledge graphs, this wealth of information becomes significantly more accessible. Such structured representation can support the development of diverse scientific disciplines and the creation of environmentally sustainable, high-value products, while also informing public policy to benefit both scientific progress and the strengthening of the bioeconomy.

Despite this potential, most of the structured biochemical data currently available on the Web is manually curated. Given the continuous and rapid publication of new research, manual curation alone is insufficient to keep up with the growing body of knowledge.

The Second International Biochemical Knowledge Extraction Challenge (BiKE) seeks to address this challenge by fostering advancements in automatic biochemical knowledge extraction methods within the Semantic Web research community. The initiative aims to improve the availability of information on natural products, promote the development of eco-friendly innovations, and raise awareness of the critical value of biodiversity. The following works were accepted for publication and presented at the workshop:

- Large Language Models Ensemble for Biochemical Properties Detection in Scientific Articles
- SPHOTA: Knowledge Graph Structure Prediction with a Hybrid Orientation of Textual Alignment using K-BERT

The following teams withdrew during the peer review process:

- KBE - IBM Research by Anaximandro Souza, Viviane Torres, Marcelo Archanjo (IBM, Brazil)
- Aryabhatta by Dr. Sanju Tiwari and Rishit Agarwal (Shodhguru Labs, India)
- Tsotsa Lab by Furel Teguimene, Fidel Jiomekong, Sanju Tiwari (University of Youndé, Cameroon)
- SPARKS KENYA by Ronald Ojino, Titus Muthomi Kimaani, Faridah Atieno Onyango (University of Kenya, Kenya)

BiKE 2025: Second International Biochemical Knowledge Extraction Challenge, Co-located with Text2SPARQL at the ESWC 2025, 1 June 2025, Portoroz, Slovenia
edgard.marx@htwk-leipzig.de (E. Marx); marilia.valli@ifsc.usp.br (M. Valli); jvictor.ssilva@ifsc.usp.br (J. d. S. e. Silva); paulo.carmo@htwk-leipzig.de (P. V. d. Carmo)


Challenge

The Biochemical Knowledge Extraction (BiKE) Challenge invited researchers to participate in either reusing existing methodologies or developing novel and innovative approaches for the extraction of biochemical information. Participants were tasked with extracting pertinent data from biochemical research articles and constructing a Biochemical Knowledge Graph (BKG) using a predefined ontology. A Biochemical Knowledge Graph is a structured representation that includes biological and chemical information derived from living organisms.

Training and Test Datasets

The training and evaluation dataset was compiled from several hundred peer-reviewed scientific articles, encompassing data on over 2,521 distinct natural product extraction scenarios. This dataset was meticulously curated by experts in chemistry, who manually reviewed the literature to annotate four key properties for each natural product described. For the purposes of this challenge, the focus was on five specific NuBBE properties used for training and prediction: (1) the compound name (rdfs:label), (2) bioactivity (nubbe:biologicalActivity), (3) the species from which natural products were extracted (nubbe:collectionSpecie), (4) the collection site of these species (nubbe:collectionSite), and (5) the isolation type (nubbe:collectionType).

In each training split, all papers were included. However, for the corresponding test splits, all links to manually extracted features were removed. Consequently, these test papers were disconnected from the broader knowledge graph. To support reconstruction, the provided Python code—leveraging the networkx library—utilized topics extracted via BERTopic for re-establishing connections within the knowledge graph. Topics assigned in this process were filtered according to a specific criterion: any topic appearing in more than 80% of the examples was excluded due to insufficient discriminatory power.

An additional component of the challenge was to devise alternative methods for reconnecting the knowledge graph using automatically extracted features, such as citation networks, authorship data, and conference associations.

Participants were given access to the original flat dataset, the initial networkx-based knowledge graph, and ten pre-randomized train/test splits. In each split, connections were preserved in the training set and removed in the test set. For every split, a prepared networkx representation of the knowledge graph was also provided. The full source code and documentation for the benchmark, referred to as NatUKE [1], are publicly accessible at: https://github.com/AKSW/natuke.

Evaluation Metrics

The challenge was focused on ranking the correct document prediction of real links that were hidden in the knowledge graph. Together with MRR (Mean Reciprocal Rank), hits@k is a ranking metric for when there is only one correct document. On the other hand, mAP (mean Average Precision) and nDCG (normalized Discounted Cumulative Gain) are designed for ranking when a list of relevant documents is available. The hits@k was chosen because it allows

the evaluation of each characteristic extraction with reasonable expectations by customizing the k value. Following the rule used in NatUKE, the final k values in this table are from 1 to 50, considering values multiples of 5 and two thresholds: (1) a score equal to or higher than 0.50 is achieved; and (2) a score equal to or higher than 0.20 is achieved. Please refer to the NatUKE benchmark paper for further details.

Best Knowledge Extraction Awards

The Best Extraction Method award was established to honor the top three participants who demonstrated exceptional proficiency, dedication, and a deep understanding of the concepts and methodologies involved in extracting relevant information from complex biochemical datasets. These individuals exhibited outstanding analytical reasoning, problem-solving capabilities, and an advanced command of state-of-the-art computational tools and techniques.

Each award was accompanied by a personalized certificate of recognition, highlighting the recipient's name and specific achievements within the context of the workshop. The recipients of the Best Extraction Method award in the second edition of the BiKE Challenge were:

1st **Large Language Models Ensemble for Biochemical Properties Detection in Scientific Articles**
Marcos Paulo Silva Gôlo, Jose Gilberto Barbosa de Medeiros Junior, Gabriele Souza Vilas Boas, Fábio Manoel França Lobato, Diego Furtado Silva and Ricardo Marcondes Marcacini

2nd **SPHOTA: Knowledge Graph Structure Prediction with a Hybrid Orientation of Textual Alignment using K-BERT**
Bharath Chand and Sanju Tiwari

In this challenge version, Gôlo et al. [2] shifts focus from graph-based methods to direct textual analysis using large-scale language models for biochemical property extraction. This work explores three distinct LLM strategies: zero-shot prompting using proprietary models such as GPT-4.1 and Gemini 2.5, few-shot prompting using open-source Qwen3 models, and fine-tuning using QLoRA on a 32-billion-parameter Qwen model. In the zero-shot setting, prompt engineering enables models to extract compound names, bioactivities, species, collection sites, and isolation types directly from scientific texts without additional training. The few-shot approach incorporates labeled examples into prompts to guide open-source models with limited context. The fine-tuning method creates five single-task models, each specialized in extracting a specific property. To enhance performance further, the study proposes an ensemble strategy that merges the outputs of GPT-4.1 and Gemini 2.5, leveraging their complementary strengths. The ensemble consistently outperforms individual models and graph-based baselines, achieving state-of-the-art results on the NatUKE benchmark, particularly in accurately identifying species and compound names.

Chand and Tiwari [3] proposes a hybrid link prediction method that combines K-BERT-generated textual embeddings with graph-based regularization. Building on the EPHEN model, the approach injects contextual information from scientific texts into a knowledge graph and propagates these embeddings to non-textual nodes. Link prediction is performed via nearest-neighbor retrieval in the embedding space. Evaluated on the NatUKE benchmark, the method

outperforms traditional graph embedding techniques by effectively integrating semantic and structural information.

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 would like to thank the advisory team, authors, program committee, and other organizers for their constant support in making this event successful.

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] M. P. S. Gôlo, J. G. B. de Medeiros Junior, G. S. V. Boas, F. M. F. Lobato, D. F. Silva, R. M. Marcacini, Large language models ensemble for biochemical properties detection in scientific articles, in: Proceedings of the Second Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC 2025, BiKE'25, CEUR, 2025. Published Digitally.

[3] B. Chand, S. Tiwari, Sphota: Knowledge graph structure prediction with a hybrid orientation of textual alignment using K-BERT, in: Proceedings of the Second Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC 2025, BiKE'25, CEUR, 2025. Published Digitally.