Published: 23 June 2025

SSDBM 2025: 37th International Conference on Scalable Scientific Data Management
June 23 - 25, 2025
Columbus, USA

PANEL

Vector Databases for Modelling, Managing and Querying Big Scientific Data: Models, Issues, Paradigms

ALFREDO CUZZOCREA, University of Calabria, Rende, CS, Italy

SSDBM '25: Proceedings of the 37th International Conference on Scalable Scientific Data Management (June 2025)
https://doi.org/10.1145/3733723.3742469
ISBN: 9798400714627

Vector Databases for Modelling, Managing and Querying Big Scientific Data: Models, Issues, Paradigms

Alfredo Cuzzocrea*
iDEA Lab
University of Calabria
Rende, Italy
Department of Computer Science
University of Paris City
Paris, France
alfredo.cuzzocrea@unical.it

Abstract

Inspired by the emergence of scientific data in fields like as *astronomy*, *climate research*, and *genomics*, which presents significant challenges for conventional database systems. This vision paper investigates the adaptation of *vector databases* in order to *describe*, *handle*, and *query* large-scale scientific data. Moreover, we propose a road-map for *embedding-centric data infrastructures*, along with an analysis of current advancements and identification of important future research directions, including *explainability* and *scalability*. This research provides a foundational basis for *next-generation interdisciplinary scientific discovery*.

CCS Concepts

• **Information systems** → **Data model extensions**; **Retrieval models and ranking**.

Keywords

Vector Databases, Advanced Scientific Data Representation, Advanced Scientific Data Management, Big Scientific Data

ACM Reference Format:
Alfredo Cuzzocrea. 2025. Vector Databases for Modelling, Managing and Querying Big Scientific Data: Models, Issues, Paradigms. In *The International Conference on Scalable Scientific Data Management 2025 (SSDBM 2025), June 23–25, 2025, Columbus, OH, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3733723.3742469

1 Introduction

Nowadays, there has been a significant increase in the *scale*, *complexity*, and *diversity* of scientific data. The capacity to handle and extract knowledge from large and complex datasets is essential to current scientific discoveries, from *particle physics experiments* and *climate modeling* to *high-throughput genome sequencing* and *astronomical sky surveys*. The functional and conceptual boundaries of conventional data management systems, i.e., relational and document-oriented databases, have been reached as a result of this development. Although these systems are strong in *transactional* and *schema-based settings*, they frequently fail to capture the *high dimensionality*, *multimodality*, and *semantic richness* of modern scientific data (e.g., [2, 10, 23]).

Currently, developments in *Natural Language Processing* (NLP) and *Machine Learning* (ML) have produced *vector embeddings*, which are *dense*, *high-dimensional data representations* that encode semantic similarity in a feature space. Recently, *Vector Databases* (VDBs) have become popular for managing these embeddings, particularly in *Artificial Intelligence* (AI) applications such as *Large Language Model* (LLM), image retrieval and recommendation systems. They still have a lot of unrealized promise for managing scientific data.

The usefulness and revolutionary potential of *vector databases for handling, modeling, and querying large scientific data* are examined in this vision paper (e.g., [22]). We contend that access to scientific knowledge may be made more flexible, semantic, and scalable by vector-based representations, which can facilitate AI-driven insights, similarity searches, and approximation searches that are not possible with traditional methods. Critical research problems are brought up by the adaptation of VDBs to the scientific domain, such as: *How can scientific data be successfully integrated in vector spaces?* (*ii*) *Which new paradigms for queries and data structures are required?* (*iii*) *How can explainability, reproducibility, and scientific rigor be guaranteed?*

Scientific data, in contrast to many commercial or consumer-focused datasets, frequently has intricate provenance linkages, highly organized metadata, and changing interpretations connected to theoretical or experimental settings. *Units of measurement*, *hierarchical ontologies*, *domain semantics*, and *uncertainty quantification* must all be carefully taken into account when integrating such data into vector spaces. Additionally, scientific workflows are iterative in nature, requiring data models that can track changes over time while preserving vector representation coherence. This is because data undergoes numerous phases of processing, analysis, and reinterpretation.

Another critical issue consists of integrating VDBs with current scientific computing platforms. From *Hierarchical Data Format version 5* (HDF5) and *Network Common Data Form* (NetCDF) to

[*] This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France

SSDBM 2025, Columbus, OH, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1462-7/25/06
https://doi.org/10.1145/3733723.3742469

domain-specific pipelines in astrophysics or bioinformatics, scientific data is usually handled using a patchwork of formats and tools. *Embedding-based systems* must adhere to reproducibility criteria, validation standards, and disciplinary conventions in addition to working with these tools. These limitations necessitate hybrid architectures that integrate *symbolic*, *statistical*, and *vector-based* representations, which might fundamentally alter our understanding of scientific databases, with a special, prominent role played by *multidimensional data* (e.g., [4–6]).

The remaining part of this paper is organized as follows. Section 2 reviews relevant related work in the context of vector databases, scientific data management, and AI-driven retrieval systems. Section 3 provides future research directions in vector-based scientific data systems. Finally, Section 4 concludes the paper along with a discussion on the broader implications of this paradigm shift.

2 Vector, Semantics, and Scalable Scientific Data Management: State-of-the-Art Analysis

Scientific data management has evolved to address several challenges, which have led to emerging research that explores how *embedding-centric approaches* and *AI-enhanced retrieval mechanisms* can be adapted to scientific data context. In this Section, we review relevant developments in vector databases, scientific data management, and AI-driven retrieval systems.

[7] suggests integrating *relational embeddings* with *probabilistic databases* to allow symbolic query of vector spaces. Authors propose the use of relational embeddings for querying high-dimensional vector spaces. Their method enables symbolic reasoning over vector representations to integrate probabilistic models with vector-based data. This paper is significant because it tackles the problem of searching vector databases while preserving *interpretability* and *flexibility*, two crucial components for scientific data applications.

Authors in [20] concentrate on *Nearest Neighbor Search* (NNS) using *Graph Neural Network* (GNN) embeddings in *Tree-Structured Data* stored in vector databases. As explainability is essential in scientific applications, they offer techniques for understanding NNS problems. The methods they suggest may be used to scientific fields that deal with complicated, hierarchical data, such as *chemical structures* and *biological networks*. Their method makes it possible to extract valuable information from *complex graph-structured data*.

With an emphasis on *highway interchange data*, [26] explores the application of *Deep Learning* (DL) methods to enhance vector spatial databases. The approach consists to extract significant patterns from unprocessed spatial data and turning it into *enriched vector representations*. Moreover, authors show how sophisticated deep learning algorithms can improve *spatial queries*. Finally, this research can be pertinent, especially for *complex* and *spatially distributed scientific data*, such as *geospatial data in environmental sciences*.

[24] aims at presenting a novel technique for unsupervised compression of *short-read databases* by transforming *FASTA files*, a popular file format for *DNA sequence data*, into *feature vectors*. Authors demonstrate the creation of feature vectors for biological data, which could significantly lower storage needs while facilitating quick similarity searches. This method shows how vector databases can be crucial in handling massive biological data, a fundamental field for future scientific data management.

The synthesis of *quantum vector databases* using *Grover's search algorithm*, a quantum algorithm used for effectively searching unsorted databases, has been examined in [21]. Authors present how to apply quantum computing to vector databases, which could have important ramifications for accelerating data retrieval procedures. This research highlights how quantum computing and vector databases can be integrated in the future, which could transform scientific data modeling and querying, particularly for large-scale datasets.

According to [14], vector databases can be used for improving resource usage in the emerging context of the *Internet of Things* (IoT). Authors propose methods for effectively managing data using vector databases to boost IoT devices' energy efficiency. Although the approach is focused on the Internet of Things, it can assist in optimizing scientific data management systems for large-scale, resource-constrained environments, such as *sensor networks* used in *environmental* or *industrial monitoring*.

An Innovative technique for dimensionality reduction in vector databases is presented by [3], which uses the *Fast Fourier Transform* (FFT). The method is leveraged for sentence embeddings produced by transformer models, which are often employed in NLP applications. By decreasing the processing cost related to high-dimensional data, this approach helps in increasing the effectiveness of vector databases. Handling complicated scientific datasets is a great benefit provided by their proposed methods, particularly when dealing with high-dimensional representations in *Genomics*.

[9] introduces the use of text embedding models in vector databases to develop *text classifiers*, with a specific example centered on biological and epidemiological data. Specifically, authors demonstrate how vector databases can effectively store and query medical text embeddings by enabling the use of machine learning models for classification tasks. This is directly applicable in sectors where vector databases store and evaluate massive amounts of text data (e.g., *Electronic Health Records*).

[12] presents *Curator*, an *indexing framework meant to effectively manage multi-tenant vector databases*. The proposed approach addresses the issue of scaling vector databases in multi-tenant environments where several users or apps use the same underlying infrastructure. Furthermore, *Curator* can be applied to scientific data management applications where several research teams may need access to large-scale datasets stored in a shared environment.

The idea of personalized *similarity search* in vector databases, where search results are customized for specific users or situations, is the subject of [18]. For scientific applications that need *domain-specific*, *context-aware data retrieval* (e.g., personalized recommendations in drug discovery or tailored climate models), the authors propose new algorithms for implementing personalized search in large-scale vector databases.

In [25], authors investigate how vector databases can speed contextualization in *AI LLM*. They showcase how vector databases can store and retrieve contextual embeddings more efficiently, which is a critical aspect for improving AI model performance. In fields like *medical imaging* or *environmental forecasting*, where big, *context-dependent models* are used on complicated datasets, this method has a crucial impact on scientific data management.

In [19], a novel technique for enhancing image analysis is presented, which leverages *adaptive caching mechanisms* in vector databases in order to extract features efficiently. Authors demonstrate that this method can be an emerging solution for decreasing time and computational cost of processing large-scale image dataset, with applications in disciplines such as *medical imaging*, *remote sensing*, and *biological image analysis*.

[29] explores the constraints of relational databases in managing vector data, using PostgreSQL as a case study. Authors emphasize issues associated with *scalability*, *query performance*, and *feature extraction* in relational databases when dealing with vector data. The obtained results are critical for understanding the limitations of standard database systems and underlining the importance of specialized vector databases in scientific applications.

In [8], authors discuss how to incorporate and reduce overall uncertainty in *geometric length measurements* that are imprecise and inaccurate inside vector databases. The research has significant implications for geographic scientific data, where accurate geometric data measurement and representation are essential.

[30] presents a method for calculating the *approximate Hausdorff distance* in multi-vector databases, which is essential for comparing patterns and structures in high-dimensional areas. For scientific fields like *molecular structures* or *geometric models* that compare complex datasets, this methodology is crucial, by enabling faster, more effective searching and comparison of multidimensional scientific data.

For advanced *Retrieval-Augmented Generation* (RAG) systems, in [16] authors present *TigerVector*, a *system designed to support vector search in graph databases*. Specifically, authors explain how *graph-based structures* and vector search can be leveraged in order to improve retrieval and reasoning for challenging scientific tasks.

3 Paving the Way Forward: Research Direction in Vector-Based Scientific Data Systems

Building on current advancements in vector databases and their emerging applications in scientific data management, several open challenges and promising directions remain to be explored. In this Section, we outline key relevant areas where innovation is needed to fully realize the potential of *next-generation scientific discovery*.

Integration of Domain-Specific Ontologies with Vector Databases. When using vector databases for scientific data, one of the biggest obstacles is how to express *domain-specific knowledge*. Scientific data must be retained and incorporated into vector-based representations, as it is frequently deeply buried in *rich ontologies* or *taxonomies* (such as in biology or chemistry). Future research should concentrate on creating techniques that preserve the semantic structure of *traditional ontologies* while mapping them onto *high-dimensional vector spaces*. By guaranteeing that scientific reasoning is based on both *vector space semantics* and *expert-defined domain knowledge*, such integration might make it easier to search for and retrieve complicated scientific questions.

Scalable and Efficient Vector Database Indexing for Large-Scale Scientific Data. With the exponential growth of scientific datasets, efficient indexing is critical to speedy retrieval. In scientific environments where datasets may approach terabytes or more, current vector database systems such as Milvus and *Facebook AI Similarity Search* (FAISS) sometimes fail to scale, even when they perform well in some applications. Researchers should investigate cutting-edge indexing techniques that can effectively handle *multi-dimensional embeddings*, *high-throughput queries*, and *large-scale scientific datasets*. The hybrid indexing strategies that combine cutting-edge machine learning algorithms with conventional database procedures have to be their main focus. Better scalability and performance in scientific domains like climate modeling, genomics, and high-energy physics will be made possible by this.

Query Optimization for Approximate Search in Scientific Data. The *Approximate Nearest Neighbor* (ANN) search is a crucial function in vector databases; nevertheless, the accuracy and performance of this approach often depend on the dataset's characteristics and the underlying algorithm. Because scientific data often contains intrinsic *noise* and *uncertainty*, optimizing ANN search becomes much more challenging. Future research should concentrate on developing *new query optimization strategies* for scientific applications, where finding a compromise between speed and accuracy may have a significant influence on the quality of the answers. In a scientific context, methods such as *precision-enhanced ANN methods*, *adaptive query processing*, and *hybrid search strategies* may strengthen vector database queries.

Ensuring Explainability and Interpretability of Vector-Based Models. In scientific domains where *transparency* and *interpretability* are crucial, vector databases lack *explainability* presents a problem despite their strong semantic search and data retrieval capabilities. It is important for researchers to come up with ways to make vector-based models easier to understand. Tools that shed light on the connections between data points and visuals that demonstrate how particular embeddings translate into scientific phenomena may be examples of this. Vector embeddings must produce observable, intelligible outcomes when applied in fields like genetics or healthcare, where choices may have immediate repercussions.

Real-Time Data Processing and Embedding Updates. *sensors*, *computer models* and *ongoing experimental* methods offer continuous data across various scientific disciplines. To ensure that vector database searches display the latest information, real-time modifications to vector embeddings are essential, presenting considerable technological challenges. The primary focus of future research should be on developing effective techniques for incrementally updating embeddings without the need to recalculate the complete dataset. These advancements hold significant relevance for applications such as *self-driving vehicles*, *environmental oversight*, and *immediate medical diagnostics* that require rapid and reliable access to data.

Cross-Domain and Multi-Modal Data Integration in Vector Databases. Combining data from various domains and formats—like *text*, *images*, and *sensor measurements* is becoming more crucial for scientific research. Vector databases offer a reliable system for managing diversity by representing data as high-dimensional vectors. Nonetheless, maintaining important semantic connections across domains while integrating data from various modalities into a unified vector space continues to present a considerable challenge. Innovative strategies for *cross-domain representation* and *multi-modal learning* will be necessary to advance this area, enabling the smooth amalgamation of varied scientific data for more profound and comprehensive analysis.

Quantum Computing and Vector Databases. *Quantum computing* has the potential to change and enhance query execution and data processing, particularly in high-dimensional vector fields. By delving into the active literature, there is a rise in investigations of possible integration of quantum algorithms, such as *Grover's search*, with vector databases to improve information retrieval speed and effectiveness. Future directions can use quantum hardware and algorithms progress to create *hybrid quantum-classical systems* that may leverage quantum advantages in practical search tasks. The processing of vast amounts of scientific data might be greatly accelerated by these advancements, which would have an effect on areas such as *genetics*, *materials research*, and *drug development.*

Handling Uncertainty and Incomplete Data in Vector Representations. Scientific data can be *noisy*, *incomplete*, or *uncertain*; as a result, vector databases must have techniques to manage and address uncertainty while ensuring data integrity. Future studies might focus on developing *probabilistic embeddings*, which explicitly include uncertainty into vector representations. These embeddings might then be used to guide querying operations, with the results indicating confidence levels depending on data uncertainty. This technique would be particularly beneficial in domains like *environmental modeling*, where data from sensors might have various degrees of reliability, or in *genomics*, where incomplete data is prevalent.

Automated Data Annotation and Semantic Search Using Vector Databases. The growing amount of scientific data makes *manual data annotation* less and less feasible and less efficient. By linking new data points to preexisting embeddings that have semantic significance, vector databases may be extremely useful in automating the data annotation process. Vector search and machine learning techniques can be an emerging solution to create systems that *cluster* scientific data, *annotate*, and *automatically classify.* In order to develop *self-organizing systems* that can help researchers effectively tag, categorize, and analyze fresh data, research should leverage a combination of vector databases with machine learning and natural language processing models.

Balancing Performance and Accuracy over Large-Scale Big Scientific Datasets. As mentioned, one of the most relevant tasks of the vector database model is represented by *search operation*, which plays a fundamental role for the whole target big data system that is based on that model. On the other hand, one of the most common *search algorithms* is represented by *Approximate k-Nearest Neighbor* (ANN) (e.g., [15]), which leads the search task when *approximation* is tolerated within search results, such as the cases of a wide spectrum of application scenarios, including information retrieval, image retrieval, spatial data retrieval, etc. Here, *indexing data structures*, like *K-D trees* (e.g., [1]), are employed to drive the (approximate) search task. However, this can, of course, introduce approximation, which can become even more problematic on *large-scale big scientific datasets* (e.g., [11]). Therefore, a relevant research problem of the future consists in *balancing performance and accuracy of approximate search tasks*, perhaps based on *probabilistic methodologies.*

Security and Privacy in Vector Databases for Scientific Data. In scientific research, vector databases pose *privacy and data security challenges*, especially in the fields of genetics and medicine. While *efficiency* and *scalability* are given main focus in current vector database systems, robust techniques for *secure access* and *privacy-preserving queries* are fully investigated. Future studies should focus on building secure vector databases that provide *strong querying capabilities* while *preserving data privacy. Differential Privacy*, *Homomorphic Encryption*, and *Federated Learning* (FL) are emerging techniques that can to satisfy the privacy concerns of scientific communities who work with sensitive information.

Interdisciplinary Collaboration for Advancing Vector Databases. In scientific data management, vector databases introduce a multidisciplinary strategy that combines domain-specific research expertise, database systems, and ML. Future investigation should place a high priority on interdisciplinary cooperation between *domain experts*, scientists from different disciplines, including *physics*, *chemistry*, *biology*, and *medicine.* This cooperative approach can be a creative way to employ vector databases in scientific data management that is both technically complex and intimately related to the requirements of scientific researchers.

Integration with Contemporary Big Data Processing Platforms. One of the key success of the vector database modelling approach for big scientific data is, without doubts, played by its full integration with *contemporary big data processing platforms*, such as *Hadoop*, *Spark*, *Azure*, and so forth. This not only will determine the definition and implementation of a plethora of advanced real-life applications and systems based on complex models and algorithms (e.g., [13, 28]), which usually require relevant computational overheads to run, but also promote and stir-up industrial research initiatives in the area, as well as further investigation in the academic setting. On the other hand, developing and running real-life applications and systems will convey in the discovery of new research lines and new research perspectives unexplored till now, with significant advancements to occur within the whole scientific sector (e.g, [17]).

4 Conclusions

In this paper, we delve into *vector databases*, which appear to be a convincing alternative for supporting *semantic-rich*, *high-dimensional*, and *approximate querying capabilities.* This vision paper presents vector representations with transformative potential in modeling and querying complex scientific data, while also identifying a number of unresolved research directions, including *ontology integration*, *indexing scalability*, *explainability*, *privacy*, and *quantum acceleration.*

To fully achieve this potential, there must be an effort to promote multidisciplinary creativity by incorporating domain semantics into vector spaces, developing hybrid architectures that strike a compromise between *symbolic* and *statistical paradigms*, and ensuring *reproducibility* and *interpretability* in data-driven scientific operations. Furthermore, developments in *real-time embedding updates*, *cross-modal integration*, and *privacy-preserving techniques* will be important for developing trustworthy and influential scientific infrastructures. Finally, vector databases are poised to become foundational components of next-generation scientific ecosystems by enhancing the scalability and intelligence of data retrieval processes (e.g., [27]).

Acknowledgments

This work was funded by the Next Generation EU Italian NRRP, Mission 4, Component 2, Investment 1.5 (Directorial Decree n. 2021/3277) - project Tech4You n. ECS0000009.

References

[1] Jon Louis Bentley. 1975. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* 18, 9 (1975), 509–517.

[2] Gustavo Caetano Borges, Júlio Cesar dos Reis, and Claudia Bauzer Medeiros. 2022. SSM: A Semantic Metasearch Platform for Scientific Data retrieval. *RITA* 29, 1 (2022), 91–101.

[3] Vitaly Bulgakov and Alec Segal. 2024. Dimensionality Reduction in Sentence Transformer Vector Databases with Fast Fourier Transform. *CoRR* abs/2404.06278 (2024).

[4] Alfredo Cuzzocrea. 2005. Overcoming Limitations of Approximate Query Answering in OLAP. In *Ninth International Database Engineering and Applications Symposium (IDEAS 2005), 25-27 July 2005, Montreal, Canada*. IEEE Computer Society, 200–209.

[5] Alfredo Cuzzocrea. 2006. Accuracy Control in Compressed Multidimensional Data Cubes for Quality of Answer-based OLAP Tools. In *18th International Conference on Scientific and Statistical Database Management, SSDBM 2006, 3-5 July 2006, Vienna, Austria, Proceedings*. IEEE Computer Society, 301–310.

[6] Alfredo Cuzzocrea and Wei Wang. 2007. Approximate range-sum query answering on data cubes with probabilistic guarantees. *J. Intell. Inf. Syst.* 28, 2 (2007), 161–197.

[7] Tal Friedman and Guy Van den Broeck. 2020. Symbolic Querying of Vector Spaces: Probabilistic Databases Meets Relational Embeddings. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*. 1268–1277.

[8] Jean-François Girres. 2024. Combining Error Models to Reduce the Imprecision of Geometric Length Measurement in Vector Databases. *Transactions in GIS* 28, 2 (2024), 200–222.

[9] Rishabh Goel. 2024. Using Text Embedding Models and Vector Databases as Text Classifiers with the Example of Medical Data. *CoRR* abs/2402.16886 (2024).

[10] Chengyu Gong, Gefei Shen, Luanzheng Guo, Nathan R. Tallent, and Dongfang Zhao. 2024. OPDR: Order-Preserving Dimension Reduction for Semantic Embedding of Multimodal Scientific Data. *CoRR* abs/2408.10264 (2024).

[11] Fabian Groh, Lukas Ruppert, Patrick Wieschollek, and Hendrik P. A. Lensch. 2023. GGNN: Graph-Based GPU Nearest Neighbor Search. *IEEE Trans. Big Data* 9, 1 (2023), 267–279.

[12] Yicheng Jin, Yongji Wu, Wenjun Hu, Bruce M. Maggs, Xiao Zhang, and Danyang Zhuo. 2024. Curator: Efficient Indexing for Multi-Tenant Vector Databases. *CoRR* abs/2401.07119 (2024).

[13] Fujiao Ju, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin. 2019. Probabilistic Linear Discriminant Analysis With Vectorial Representation for Tensor Data. *IEEE Trans. Neural Networks Learn. Syst.* 30, 10 (2019), 2938–2950.

[14] Rani Kumari, Dinesh Kumar Sah, Korhan Cengiz, Ali Nauman, Nikola Ivkovic, and Ivan Mihaljevic. 2023. Optimizing Resource Utilization Using Vector Databases in Green Internet of Things. In *IEEE Globecom Workshops 2023, Kuala Lumpur, Malaysia, December 4-8, 2023*. 1988–1993.

[15] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Mingjie Li, Wenjie Zhang, and Xuemin Lin. 2020. Approximate Nearest Neighbor Search on High Dimensional Data - Experiments, Analyses, and Improvement. *IEEE Trans. Knowl. Data Eng.* 32, 8 (2020), 1475–1488.

[16] Shige Liu, Zhifang Zeng, Li Chen, Adil Ainihaer, Arun Ramasami, Songting Chen, Yu Xu, Mingxi Wu, and Jianguo Wang. 2025. TigerVector: Supporting Vector Search in Graph Databases for Advanced RAGs. *CoRR* abs/2501.11216 (2025).

[17] Carlos Linares López. 2010. Vectorial Pattern Databases. In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings (Frontiers in Artificial Intelligence and Applications, Vol. 215)*, Helder Coelho, Rudi Studer, and Michael J. Wooldridge (Eds.). IOS Press, 1059–1060.

[18] Marek Mahrík, Matús Sikyna, Vladimir Mic, and Pavel Zezula. 2024. Towards Personalized Similarity Search for Vector Databases. In *Similarity Search and Applications - 17th International Conference, SISAP 2024, Providence, RI, USA, November 4-6, 2024, Proceedings*. Springer, 126–139.

[19] Solmaz Seyed Monir and Dongfang Zhao. 2024. Efficient Feature Extraction for Image Analysis through Adaptive Caching in Vector Databases. In *7th International Conference on Information and Computer Technologies, ICICT 2024, Honolulu, HI, USA, March 15-17, 2024*. IEEE, 193–198.

[20] Lukas Pfahler and Jan Richter. 2020. Interpretable Nearest Neighbor Queries for Tree-Structured Data in Vector Databases of Graph-Neural Network Embeddings. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020*.

[21] Cesar Borisovich Pronin and Andrey Vladimirovich Ostroukh. 2023. Synthesis of Quantum Vector Databases Based on Grover's Algorithm. *CoRR* abs/2306.15295 (2023).

[22] Toni Taipalus. 2024. Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cogn. Syst. Res.* 85 (2024), 101216.

[23] Meng Tang, Jaime Cernuda, Jie Ye, Luanzheng Guo, Nathan R. Tallent, Anthony Kougkas, and Xian-He Sun. 2024. DaYu: Optimizing Distributed Scientific Workflows by Decoding Dataflow Semantics and Dynamics. In *IEEE International Conference on Cluster Computing, CLUSTER 2024, Kobe, Japan, September 24-27, 2024*. IEEE, 357–369.

[24] Tao Tang and Jinyan Li. 2021. Transformation of FASTA Files into Feature Vectors for Unsupervised Compression of Short Reads Databases. *Journal of Bioinformatics and Computational Biology* 19, 1 (2021), 2050048:1–2050048:15.

[25] Raad Bin Tareaf, Mohammed AbuJarour, Tom Engelman, Philipp Liermann, and Jesse Klotz. 2024. Accelerating Contextualization in AI Large Language Models Using Vector Databases. In *International Conference on Information Networking, ICOIN 2024, Ho Chi Minh City, Vietnam, January 17-19, 2024*. IEEE, 316–321.

[26] Guillaume Touya and Imran Lokhat. 2020. Deep Learning for Enrichment of Vector Spatial Databases: Application to Highway Interchange. *ACM Transactions on Spatial Algorithms and Systems* 6, 3 (2020), 21:1–21:21.

[27] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei, and Charles Xie. 2021. Milvus: A Purpose-Built Vector Data Management System. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*. ACM, 2614–2627.

[28] Haijun Zhang, Shuang Wang, Xiaofei Xu, Tommy W. S. Chow, and Q. M. Jonathan Wu. 2018. Tree2Vector: Learning a Vectorial Representation for Tree-Structured Data. *IEEE Trans. Neural Networks Learn. Syst.* 29, 11 (2018), 5304–5318.

[29] Yunan Zhang, Shige Liu, and Jianguo Wang. 2024. Are There Fundamental Limitations in Supporting Vector Data Management in Relational Databases? A Case Study of PostgreSQL. In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*. IEEE, 3640–3653.

[30] Dongfang Zhao. 2025. Approximate Hausdorff Distance for Multi-Vector Databases. *CoRR* abs/2503.06833 (2025).