BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100216

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Full Length Article

Tensor databases empower AI for science: A case study on retrosynthetic analysis

Xueya Zhang [a], Guoxin Kang [b,*], Boyang Xiao [c], Jianfeng Zhan [b]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[c] The University of Manchester, Manchester, United Kingdom

ARTICLE INFO

Keywords:
Tensor database
Approximate retrieval
Retrosynthetic

ABSTRACT

Retrosynthetic analysis is highly significant in chemistry, biology, and materials science, providing essential support for the rational design, synthesis, and optimization of compounds across diverse Artificial Intelligence for Science (AI4S) applications. Retrosynthetic analysis focuses on exploring pathways from products to reactants, and this is typically conducted using deep learning-based generative models. However, existing retrosynthetic analysis often overlooks how reaction conditions significantly impact chemical reactions. This causes existing work to lack unified models that can provide full-cycle services for retrosynthetic analysis, and also greatly limits the overall prediction accuracy of retrosynthetic analysis. These two issues cause users to depend on various independent models and tools, leading to high labor time and cost overhead.

To solve these issues, we define the boundary conditions of chemical reactions based on the Evaluatology theory and propose BigTensorDB, the first tensor database which integrates storage, prediction generation, search, and analysis functions. BigTensorDB designs the tensor schema for efficiently storing all the key information related to chemical reactions, including reaction conditions. BigTensorDB supports a full-cycle retrosynthetic analysis pipeline. It begins with predicting generation reaction paths, searching for approximate real reactions based on the tensor schema, and concludes with feasibility analysis, which enhances the interpretability of prediction results. BigTensorDB can effectively reduce usage costs and improve efficiency for users during the full-cycle retrosynthetic analysis process. Meanwhile, it provides a potential solution to the low accuracy issue, encouraging researchers to focus on improving full-cycle accuracy.

1. Introduction

Retrosynthetic analysis is an important method for exploring efficient synthetic pathways for target molecules. It holds significant importance in fields such as chemistry, materials science, and pharmaceuticals [1]. The main goal of retrosynthetic analysis is to identify the appropriate reactants and reaction conditions for the efficient synthesis of the target molecule. For example, when a target product is inputted, the work aims to obtain the correct reactants, reaction conditions, and multiple pathways for its synthesis.

People are always committed to developing efficient and user-friendly tools to help scientists conduct retrosynthetic analysis more quickly and conveniently. Since the 1960s, computer-aided synthesis planning (CASP) has been a key tool in this area [2]. Particularly, in today's era of rapid development of artificial intelligence (AI), AI for Science (AI4S) brings about significant changes in many fields. It also injects new vitality into retrosynthetic analysis technology.

More and more machine learning-based methods, especially deep learning models, are being used in the field of retrosynthetic analysis. These AI technologies significantly improve the efficiency and accuracy of it [3–10]. In 2022, Liu et al. [11] first highlighted three major contradictions facing machine learning in materials science: data characteristics, model interpretability, and result authenticity [12–15]. These also apply to retrosynthetic analysis. On further analysis, existing prediction models are found to treat reactants and reaction conditions as separate factors, which gives rise to two key issues as follows:

a. **Lacking a unified model that can provide full-cycle service for retrosynthetic analysis.** The full-cycle service for retrosynthetic analysis involves a step-by-step prediction of reactants and reaction conditions, ultimately yielding complete candidate chemical reaction equations that are ready for direct experimental validation. Current research on AI-based retrosynthetic

* Corresponding author.
E-mail addresses: zhangxueya21@mails.ucas.ac.cn (X. Zhang), kangguoxin@ict.ac.cn (G. Kang), boyang.xiao@postgrad.manchester.ac.uk (B. Xiao), zhanjianfeng@ict.ac.cn (J. Zhan).

https://doi.org/10.1016/j.tbench.2025.100216
Received 28 January 2025; Received in revised form 20 April 2025; Accepted 28 May 2025
Available online 16 June 2025

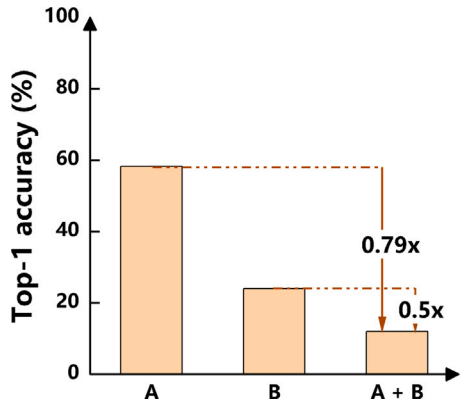

Fig. 1. The entire process of retrosynthetic analysis involves a step-by-step prediction of reactants and reaction conditions. The prediction of reactants requires a target molecule as input, while the prediction of reaction conditions requires a reaction equations without reaction conditions. We choose model RetroTRAE [16] as model A for reactants prediction and model Parrot [17] as model B for reaction conditions prediction. For detailed experimental descriptions and workload information, please refer to Sections 4.3.2 and 3.2. The figure indicates that the individual accuracies of Model A and Model B are quite low, at 58.3% and 24%, respectively. When Model A's output is used as input for Model B, the combined accuracy of the prediction results (Model A + B) drops even further to a mere 12%. This represents a significant decrease of 79% for Model A and 50% for Model B compared to their original accuracies. Given these substantial declines in performance, there is an urgent need for a unified model to improve the overall accuracy.

analysis models mainly focuses on one direction of single-step or multi-step prediction of synthetic routes. These directions can be categorized into two types based on their prediction targets, which usually include reactants and reaction conditions. However, when scientists use retrosynthetic analysis tools, they hope to directly obtain a set of complete reaction equation candidates. Therefore, it is necessary to design a full-cycle service that can provide a set of complete reaction equation candidates, including both reactants and reaction conditions.

b. **Significant bottleneck in the prediction accuracy of full-cycle retrosynthetic analysis.** We have noticed that the accuracy of individual prediction models is often low, and when used together, their combined accuracy tends to decrease even more. For each prediction models, the prediction accuracy is defined as the proportion of correct results among the Top-k generated predictions. However, as shown in Fig. 1, when the Top-1 accuracy of the reactant prediction model RetroTRAE [16] (A) is 58.3% and that of the reaction condition prediction model Parrot [17] (B) is 24%, the overall Top-1 accuracy of the final predicted result (A+B) is only 12%. Therefore, optimizing the accuracy of a specific type of model is insufficient for full-cycle retrosynthetic analysis. And there is an urgent need for new work focusing on overall prediction accuracy.

To address above issues, this paper proposes BigTensorDB. BigTensorDB is the first tensor database designed to provide full-cycle service for retrosynthetic analysis. It integrates storage, prediction generation, search and analysis functions. Its main contributions are summarized as follows:

(1) We design a tensor format for storing chemical reactions. This format efficiently stores all key information related to chemical reactions, including reactants, products, and reaction conditions (such as solvents and reagents). We provide users with services for storing and retrieving chemical reactions.

(2) We integrate multiple retrosynthetic analysis prediction models including reactants prediction and reaction conditions prediction. We also integrate SMILES embedding models. These integrations offer full-cycle retrosynthetic analysis services to users. Our work reduces usage costs and improves the pipeline's efficiency.

(3) We provide search and analysis services. Through these services, we re-rank and analyze the final prediction results. At the same time, we provide real reaction equations similar to the predicted results for user reference. These works can enhance the accuracy and interpretability of the final outcomes.

2. Background and related work

2.1. Background

Retrosynthetic analysis was established by E.J. Corey [2]. It involves identifying target molecules, deconstructing them in a reverse manner, and devising synthetic routes to achieve the desired synthesis. The purpose of retrosynthetic analysis is to assist scientists in finding more efficient synthetic routes to synthesis more useful molecules. The target molecules often originate from diverse application scenarios. They typically cannot be synthesized via known reactions or have very inefficient existing synthetic pathways. Thus, there is a need to explore new and more efficient synthetic routes by retrosynthetic analysis.

To solve this problem, scientists usually need to search through a vast space of possible transformations of the target molecules. This can be done either by hypothetically disconnecting bonds or by converting one functional group into another which goal is to match existing reaction templates. However, this process demands that scientists have a rich knowledge base and extensive synthetic experience. Additionally, it requires significant time and material costs for experiments to verify the correctness of hypotheses.

The earliest computer-aided tools work by first enumerating possible reaction types for the target molecule. Then, they use search algorithms to recursively enumerate and search for potential reaction pathways. This process continues until viable starting materials are identified. However, these methods essentially do not create new reactions but rather rearranged existing knowledge. Today, with the rapid development of AI, an increasing number of machine learning-based models are being applied to retrosynthetic analysis. This helps scientists become more creative in their retrosynthetic analysis works.

These machine learning models, particularly deep learning models, mainly fall into two categories: prediction reactants and prediction reaction conditions. In real-world scenarios, scientists must select from a wide range of models first. Then they use the chosen reactant prediction model to generate a set of reaction equations without reaction conditions. It is essential to emphasize that the reaction conditions play an important role in chemical reactions. The same reactants can undergo different reactions under different conditions, leading to different products. Therefore, scientists must also select a reaction condition prediction model for another round of predictions. Afterward, they need to conduct theoretical analysis and experimental verification manually. Each experimental verification of a reaction requires substantial time and material costs. Thus, the accuracy of prediction models is vital for real experiments. It also determines the efficiency of retrosynthetic analysis.

2.2. CASP's related work

2.2.1. Prediction models

Researchers have developed the Simplified Molecular Input Line Entry System [18] (SMILES) notation, a text-based method that encodes molecular graphs into simple, human-readable character sequences. Some prediction models extract functional groups from SMILES expressions to analyze a target molecule's reaction-related structural, spatial, and functional group features, achieving prediction. Others ignore reaction structures, spatial configurations, and functional groups, instead directly using SMILES for sequence-to-sequence [19–25] (Seq2Seq)

prediction. Generative AI, including the Seq2Seq method, is crucial for discovering new substances or materials. It enables predictive models to transcend existing knowledge and create new knowledge. However, these models show lower prediction accuracy on United States Patent and Trademark Office (USPTO) datasets. Notably, some models achieve high accuracy and better user-friendliness by calling large language model Application Programming Interfaces (APIs) or fine-tuning these models [26] to generate products and recommend reaction conditions. As discussed earlier, deep learning-based machine learning models in this field can be categorized into reactants prediction models and reaction condition prediction models. We will detail these models based on this classification.

From the perspective of the methods used, reactant prediction models can be classified as template-based, template-free, and semi-template-based models [27,28].

Template-based methods play an important role in retrosynthesis prediction. These methods employ reaction templates extracted from chemical databases to guide the retrosynthesis process through template-target molecule matching. The templates, which can be manually curated or automatically generated, enable models to identify optimal chemical transformations [29]. Multiple approaches [30–33] have been developed for template prioritization [29]: RetroSim [30] ranks candidate templates using molecular fingerprint comparisons, NeuralSym [31] employs a deep neural network classifier, and GLN [32] evaluates template-reactant compatibility with a conditional graph logic network. While template-based models provide interpretability and ensure molecule validity, their practical applications [34] are constrained by limited generalization capability and scalability [29].

Template-free methods aim to eliminate dependency on predefined templates. It achieves retrosynthesis prediction through data-driven or innovative architectural design, opening up a new direction of exploration in this field. Most existing methods turn the task into a Seq2Seq problem [19–25], using the SMILES [18] format to represent molecules. This is first to use by Liu et al. [19] who proposed a long short-term memory [35] (LSTM)-based Seq2Seq model to change the SMILES of a product into the SMILES of reactants. Meanwhile, there are some studies treat this task as a graph-to-sequence problem, using molecular graphs as input [36]. For example, Graph2SMILES [36] combined a graph encoder with a Transformer decoder to keep SMILES order the same. However, in recent studies, such as MEGAN [37], MARS [38], and Graph2Edits [39], end-to-end molecular graph editing model is widely used. These models represent chemical reactions as a series of changes to molecular graphs. Fang et al. [40] created a way to decode at the substructure level by finding parts of product molecules that stay the same. Although template-free methods are entirely data-driven, they face challenges related to the interpretability, chemical validity, and diversity of the molecules they generate [29,34].

Semi-template-based methods involve a two-stage strategy. The first stage decomposes target molecules into synthons via reactive site identification, while the second stage converts synthons to reactants through techniques like leaving group selection [27], graph generation [41], or SMILES generation [1,42]. RetroXpert [42] first identified the reaction center of the target molecule using an edge-enhanced graph attention network to obtain synthons. Then it generated the corresponding reactants based on these synthons. RetroPrime [1] incorporated the chemist's retrosynthesis strategy, which finely split the retrosynthesis process into decomposing the synthetic moiety and adding an appropriate leaving group to finally generate the reactant. These methods better match the intuitive problem-solving approach of scientists. However, the two stages in the framework are independent. This increases computational complexity. Moreover, it is challenging to transfer the knowledge and insights gained from predicting reactive sites to the completion of reactants [29].

Reaction condition prediction typically involves inputting a complete SMILES-formatted equation and outputting suitable reaction conditions. However, recommending conditions from scratch is a challenging and under-explored problem that heavily depends on the knowledge and experience of chemists. Neural network models can predict the chemical environment, including catalysts, solvents, and reagents, as well as the most suitable temperature for any given organic reaction [17,43]. There are also some works [26] based on large language models that can also achieve reaction condition recommendation and prediction. MM-RCR [26] model is a text-enhanced multimodal large language model that learns a unified reaction representation by multi-source information from SMILES, reaction graphs, and text corpora. It also demonstrated strong generalization capabilities on out-of-domain and high-throughput experimental datasets, providing new momentum for high-throughput reaction condition screening.

2.2.2. Automated feature engineering

Data quality critically impacts machine learning model performance. High-quality data can greatly boost a model's predictive accuracy and reliability. In contrast, low-quality data may degrade performance and lead to results conflicting with domain-expert understanding [44–49]. Thus, ensuring the effectiveness of data augmentation and feature extraction for retrosynthetic analysis tools is extremely important [50–52]. So, here we introduce the Automated feature engineering models related to our works.

SMILES [18] (simplified molecular input line entry system) are text-based representations that encode a molecular graph in a simple, human-readable sequence of characters [53,54]. Transformer models are effective in cheminformatics for processing SMILES strings and extracting molecular representations, benefiting from bidirectional context for enhanced understanding of chemical environments, transfer learning, and fine-tuning.

BERT [55] is a pioneering model that captures bidirectional context from SMILES strings, making it suitable for tasks like property prediction and drug discovery, though it has high computational and memory demands. MOLBERT [56] is a chemistry-specific adaptation of BERT that excels in predicting physicochemical properties and molecular interactions, but it requires large labeled datasets for optimal performance. SMILES-BERT [57] is designed to learn molecular representations directly from SMILES strings without extensive feature engineering, making it effective for predicting molecular properties, though it also demands significant computational resources. ChemBERTa [58] and ChemBERTa-2 [59] enhance BERT [60] with domain-specific training for a variety of property predictions, improving accuracy while maintaining high complexity and resource demands. The RoBERTa-based Model [61] refines BERT by using more data and longer sequences for better property prediction and molecular classification, although it increases computational requirements for training and inference. Mol-BERT [62] and MolRoPE-BERT [63] are BERT-based models for predicting molecular properties from SMILES, differing mainly in their position embedding approaches, with MolRoPE-BERT using rotary PE to address limitations of absolute PE in Mol-BERT.

2.3. Tensor retrieval related work

In the fields of data science and artificial intelligence, the demand for managing high-dimensional vector data is growing rapidly, driven mainly by the rapid development of unstructured data and machine learning technologies [64].

For vector similarity search, some systems have been put into application, such as Alibaba's AnalyticDB-V [65] and PASE (PostgreSQL) [66]. However, they do not support multi-vector queries. Vearch [67,68] is another system designed specifically for vector search, but it is not efficient in handling large-scale data and also does not support multi-vector queries. However, as a data management system specifically built for the needs for more efficient and flexible vector data management, Milvus focuses on the storage and search of large-scale vector data, supporting various query types, including vector similarity search and multi-vector query processing. In particular, the Milvus 2.5 version introduced the Sparse-BM25 algorithm, achieving hybrid search of sparse and dense vectors, further improving search efficiency.

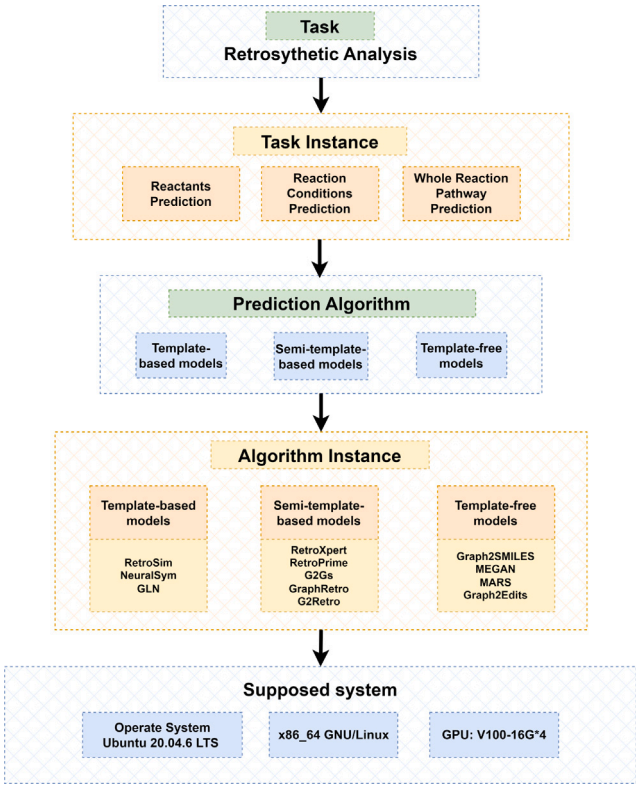

Fig. 2. Valid Evaluation Condition (EC) configurations [69].

3. System architecture of BigTensorDB

3.1. Methodology

The development of BigTensorDB is predicated on a meticulous evaluation of the existing work on retrosynthetic analysis. Our work is the first tensor database designed to provide full-cycle service for retrosynthetic analysis. To demonstrate BigTensorDB's performance, we evaluate its effectiveness in the retrosynthetic analysis process. Inspired by the Evaluatology mentioned in the [69], we conduct a comprehensive survey of the current predictive efforts in retrosynthetic analysis, covering the entire process, including reactants prediction and reaction conditions prediction. Through developing the valid and well-defined Evaluation Condition (EC) configurations, we establish a solid foundation for the motivation of our works.

As shown in Fig. 2, we construct an evaluation condition configurations for the retrosynthetic analysis task inspired by the methodology in [69]. We first clarify that the problem task of our work is retrosynthetic analysis (E'). The specific task instance (E) is the concrete problem that needs to be solved in the process of retrosynthetic analysis, such as reactants prediction (E_1) based on the target product, reaction conditions prediction (E_2) based on the reaction equation and whole reaction pathway prediction (E_3). Under the task instances, we find different algorithms (A') to solve them, which mainly include template-based, semi-template-based, and template-free models method. The specific algorithm instance (A) is each concrete predictive model itself. Each algorithm (A') has its corresponding algorithm instance (A). For example, the template-based models method (A'_1) includes instances such as RetroSim (A_{11}), NeuralSym (A_{12}) and GLN (A_{13}).

3.2. Dataset sources and workload choices

There are many open-source datasets of chemical reaction expressions in the field of retrosynthetic analysis, such as USPTO-50K and REACTION INDEX SYSTEM (REAXYS). However, none of these datasets completely include the reaction conditions in the chemical equations. Obtaining a dataset containing reaction conditions requires a lot of work.

Parrot [17] has organized two large datasets, USPTO-Condition and Reaxys-TotalSyn-Condition, which record reaction equations and reaction conditions, including solvents, reagents, catalysts, etc. They used the reaction classifier to subdivide the dataset categories, and designed an external verification experiment. Therefore, we select the USPTO-Condition dataset. We remove the data with more than two reactants and merge the reaction conditions according to our boundary settings. We finally establish a workload consisting of a training set with 490,398 data entries and a test set with 100 data entries.

3.3. BigTensorDB design and implementation

The design of BigTensorDB is divided into four layers, as shown in Fig. 3. We aim to provide a one-stop, full-cycle service for retrosynthetic analysis, in order to improve user efficiency, reduce costs, and explore ways to overcome the performance bottlenecks in overall prediction accuracy. Users only need to input a target molecule and select the desired models for reactants prediction and reaction conditions prediction. They then can receive a re-ranked prediction candidate set containing complete reaction equations and real reactions for reference.

We are committed to reordering and referencing prediction candidates by retrieving similar templates from a large-scale real chemical reaction database. To achieve this, we have designed the following four layers. In the storage layer, we carefully select feature extraction tools to extract features from chemical reaction datasets and store them in tensor format. In the prediction generation layer, we integrate multiple prediction models and provide them with a unified interface. In the search and analysis layer, we provide similarity retrieval and analysis processing services. The four layers mentioned earlier will be detailed in Sections 3.3.1 3.3.2 3.3.3 3.3.4.

3.3.1. Storage layer

In the storage layer, we determine a tensor format boundary for chemical reaction equations. A complete chemical equation involves reactants, products, and many reaction conditions, including temperature, reagents, solvents, and so on. If we use vector format to store this information, we need to embed all the above information into one vector, which will cause a huge loss in the dimension of chemical information. Therefore, we hope to use a tensor composed of multiple vectors to preserve all the information in the chemical reaction equation.

Based on Evaluatology described in [69], we define reactants, products, solvents and reagents as the four-dimensional parameters of each tensor. We also specify that the dimension size within each dimension of the tensor is 384. Using the feature extraction tool ChemBERTa-77M-MLM model, we convert the chemical equations from SMILES format to 384 dimensional vectors. They then are stored in the tensor format, thereby establishing a tensor-based knowledge base of known chemical equations.

3.3.2. Prediction generation layer

In the prediction generation layer, the user-input target molecule serves as the input for the models. Predictions and generations are carried out step-by-step according to the user-selected reactant prediction model and reaction condition prediction model. We assume the input target molecule is T, the selected reactant prediction model is A, and the selected reaction condition prediction model is B. By inputting the target molecule T into model A, we obtain a set of reactant prediction candidates S_1 without reaction conditions. We then input each candidate from S_1 into model B to obtain a set of complete chemical equation prediction candidates S_2. Assuming that models A and B generate n and m prediction candidates for each input, respectively, the size of the complete reaction prediction set S_2 is $n * m$.

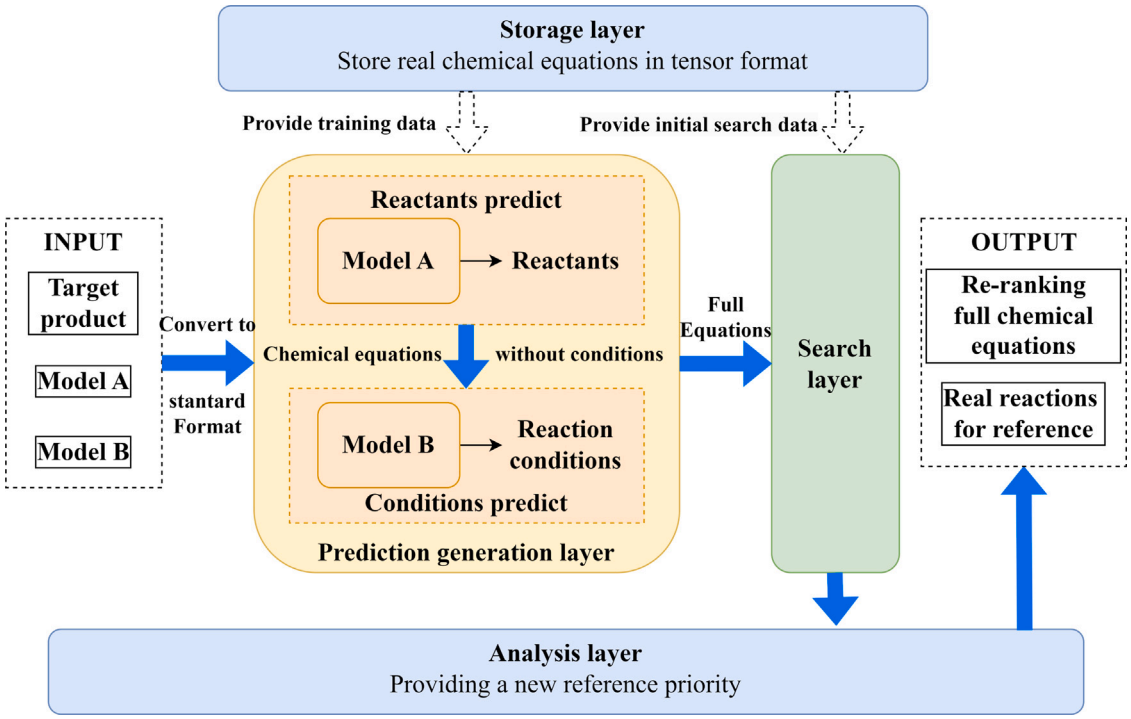

Fig. 3. Overview of the BigTensorDB workflow.

3.3.3. Search layer

In the search layer, we utilize the multi-vector search and apply a weighted ranker to set weights for multi-vector searches. We perform similarity searches for each prediction candidates in the set S_2 against the database of real reactions, obtaining search scores and the similar real reactions. Here, we have designed a preliminary experiment to test the effectiveness of different weight allocations for tensor dimensions during retrieval.

3.3.4. Analysis layer

In the analysis layer, we re-rank the reaction prediction candidate set S_2 based on the search scores obtained in the search layer. The search results return the top 5 real reactions with the highest similarity scores and their corresponding search scores. Our re-ranking strategy prioritizes candidates that achieve higher similarity scores during the search. Through preliminary experiments, we have selected the optimal weight sequence for re-ranking: $a_1 = 0.3$, $a_2 = 0.9$, and $a_3 = a_4 = 0.1$. According to this ranking information, we output to users a re-ranked prediction candidate set with improved accuracy, along with corresponding similar real reactions as reference suggestions.

4. Performance evaluation

4.1. Experiment setup

The server is equipped with 2 Intel Xeon 5218R CPUs running at 2.10 GHz, 512 GB of memory, and an NVIDIA V100-PCIE-16 GB GPU connected via PCIe 3.0. Each CPU has 20 physical cores with hyper-threading enabled, resulting in a total of 80 hardware threads, all of which were utilized. The operating system is Ubuntu 20.04 with the Linux kernel version 5.15.0. The GPU driver version is 535, and CUDA 12.2 is used for GPU computing. All experiments were conducted using Python 3.10 and Docker 26.1.

4.2. Experiment design

Our experiments include validating the research motivation and assessing our work's performance. The research motivation validation consists of theoretical analysis and experimental verification, with the latter providing the baseline model's performance metrics. The performance assessment of our work involves two main comparisons: one for retrosynthetic analysis and another for database performance. For retrosynthetic analysis, we focus on two key metrics: predictive accuracy and time cost. In the tensor database field, we evaluate tensor retrieval recall and throughput.

4.3. Motivation verification

4.3.1. Theoretical analysis

As shown in Fig. 4, we assume the input target molecule is T, the selected reactant prediction model is A, the selected reaction condition prediction model is B, the Top-k accuracy of model A is $E_A(k)$ and the Top-k accuracy of model B is $E_B(k)$.

By inputting the target molecule T into Model A, we obtain a reactant prediction candidate set S_1 without reaction conditions, with a size of n. Taking one candidate reaction equation i from S_1 as input into Model B, we obtain a complete reaction equation prediction candidate set S_2^i with a size of m. By inputting each candidate reaction equation from S_1 into Model B, we obtain the final reaction equation prediction candidate set S_2 with a size of $n * m$. The Top-k accuracy of the candidate reaction equations in S_2 can be obtained as:

$$E(k) = E_A(i) * E_B(j), \quad k = m * (i-1) + j$$

It is easy to see that $E_A(i) < 1$ and $E_B(j) < 1$. Given $k = m * (i-1)+j$, we can deduce that $i \leq k$ and $j \leq k$. Moreover, since E_A and E_B are non-decreasing functions, it follows that:

$$E(k) = E_A(i) * E_B(j) < E_A(i) \leq E_A(k)$$

$$E(k) = E_A(i) * E_B(j) < E_B(j) \leq E_B(k)$$

Thus, there is a significant bottleneck in the prediction accuracy of full-cycle retrosynthetic analysis.

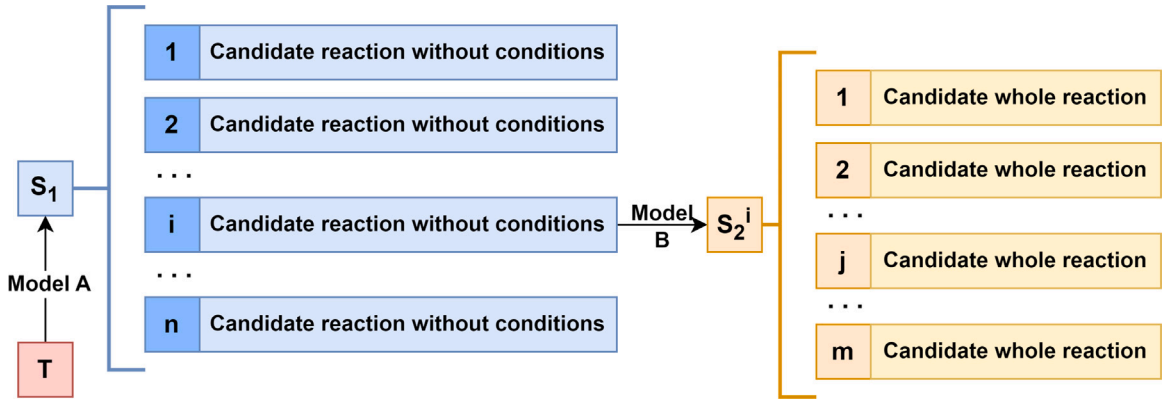

Fig. 4. Model A for reactants prediction, Top-k accuracy is $E_A(k)$, candidate set S_1's size is n. Model B for reaction conditions prediction, Top-k accuracy is $E_B(k)$, candidate set S_2^i's size is m.

Table 1
Full cycle prediction accuracy of inverse synthesis analysis.

k	1	3	5	10
Accuracy (%)	0.12	0.16	0.19	0.2

Table 2
Comparison of prediction accuracy between BigTensorDB and the baseline.

Model	Top-1	Top-5	Top-10	Top-50	Top-100
Model(A+B)'s accuracy	0.12	0.19	0.20	0.26	0.26
BigTensorDB's accuracy	0.12	0.18	0.18	0.24	0.26

4.3.2. Experiment verification

We choose model RetroTRAE [16] as model A for reactants prediction and model Parrot [17] as model B for reaction conditions prediction. Then we use the workload built in Section 3.2.

We first generate $n = 10$ candidate reaction equations without reaction conditions per test. Then we input each candidate to model B to generate $m = 15$ candidate reaction equations including reaction conditions per input. At the end, we get totally 12096 candidate complete chemical equations for 100 tests. We then calculate the Top-k accuracy is shown in Table 1. Also as shown in Fig. 1, compared with the original accuracies, each Top-k accuracy decreases sharply.

4.4. Performance results

4.4.1. Retrosynthetic analysis performance comparison

According to Section 3.1, our work is the first tensor database designed to provide full-cycle service for retrosynthetic analysis. So we evaluate BigTensorDB's effectiveness in the retrosynthetic analysis process to demonstrate BigTensorDB's performance without conducting comparative experiments with other database system.

Based on Evaluatology, we select reactants, products, solvents and reagents as the storage format for chemical equations. We choose the ChemBERTa-77M-MLM model to embed SMILES. This model converts SMILES into 384-dimensional vectors through a neural network. After generating predictions for the test set, we embed the result set and conduct similarity searches. We select the multi-vector search algorithm from the Milvus vector database, using the IVF-Flat index and Euclidean distance metric. We perform retrieval experiments under different weights. Based on the search results, we re-rank the candidate set and recalculate the Top-k accuracy. As shown in Fig. 5, we define four variables as the parameter weights for retrieval ranking. Among them, a_1 corresponds to the reactants, a_2 corresponds to the products, and a_3 and a_4 correspond to the solvents and reagents, respectively.

In Figs. 5(a) and 5(b), we conduct controlled variable experiments for parameters a_1 and a_2, respectively. In Fig. 5(a), we set $a_2 = 0.2, a_3 = a_4 = 0.2$ and a_1 to different values. Then we observe the re-ranked Top-k accuracy. The results shows that the performance is better when a_1 is in the range of 0.2 to 0.4. In Fig. 5(b), we set $a_1 = 0.2, a_3 = a_4 = 0.2$ and a_2 to different values. Then we observe the re-ranked Top-k accuracy. The results shows that the performance is better when a_2 is 0.9.

Then, we set $a_1 = 0.2, 0.3, 0.4$, $a_2 = 0.9, 0.95$ to conduct controlled variable experiments for parameters a_3 and a_4. The results are shown in Fig. 6. The Figs. 6(a) 6(b) and 6(c) shows the results of $a_1 = 0.2, 0.3, 0.4$ respectively, where $a_2 = 0.9$. The Figs. 6(d) 6(e) and 6(f) shows $a_2 = 0.95$'s results. We can find out the performance is better when a_3 and a_4 is 0.1.

Therefore, We currently find that the re-ranking performance is better when a_1 is 0.3, a_2 is 0.9, a_3 and a_4 is 0.1. However, more fine-grained experiments are still needed.

After exploring the parameter space, we conducted comparison experiments to compare our work with the baseline model.

We first compared the top-k accuracy of BigTensorDB's re-ranking strategy with that of the original A + B baseline model, and the results are shown in the Table 2. The baseline model's results were reported in Table 1 in Section 4.3.2. BigTensorDB's prediction accuracy does not surpass the baseline models. This stable accuracy shows our work does not degrade the original prediction models, providing a solid base for further development.

In BigTensorDB, time consumption involves several parts: data cleaning for Model A, (Model A training), running Model A, organizing the results of Model A into the data format required for Model B, (Model B training), running Model B, combining the results of Model A and Model B, retrieving and re-ranking the results, and finally outputting the results. Under the workload from Section 3.2, we recorded BigTensorDB's total experimental time consumption, with the following results. The parts in bold represent the extra overhead from the BigTensorDB system.

As shown in the Table 3, BigTensorDB's extra time cost accounts for less than 5% of the entire process. Thus, when providing full-cycle services, BigTensorDB does not bring extra time consumption that cannot be tolerated.

4.4.2. Tensor database performance comparison

The recall and throughput of Milvus-IVFFlat based multi-vector search in this task scenario are 72% and 45 vectors per second, respectively. This performance is sub-par for vector search and indicates significant bottlenecks. The reason is that Milvus multi-vector search relies on weighted sorting after single-vector search rather than a genuine tensor index.

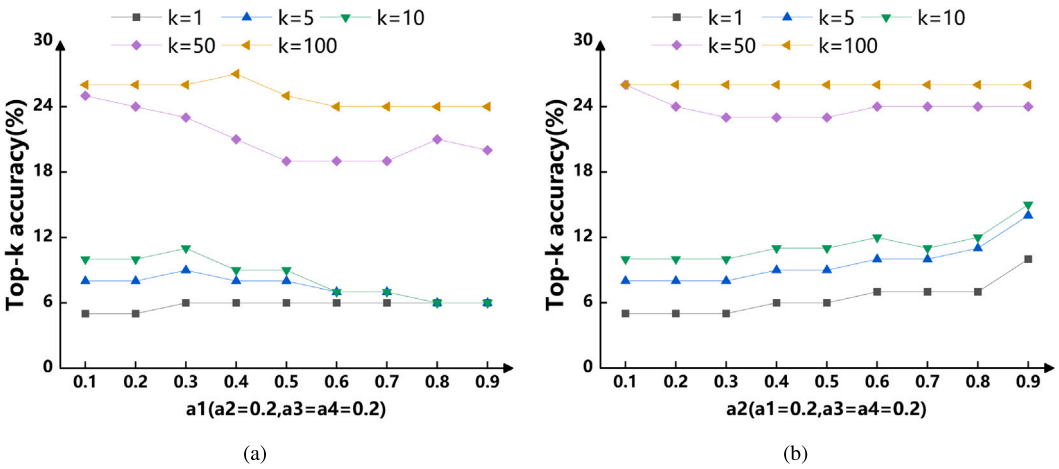

Fig. 5. Controlled variable experiments results of a_1 and a_2.

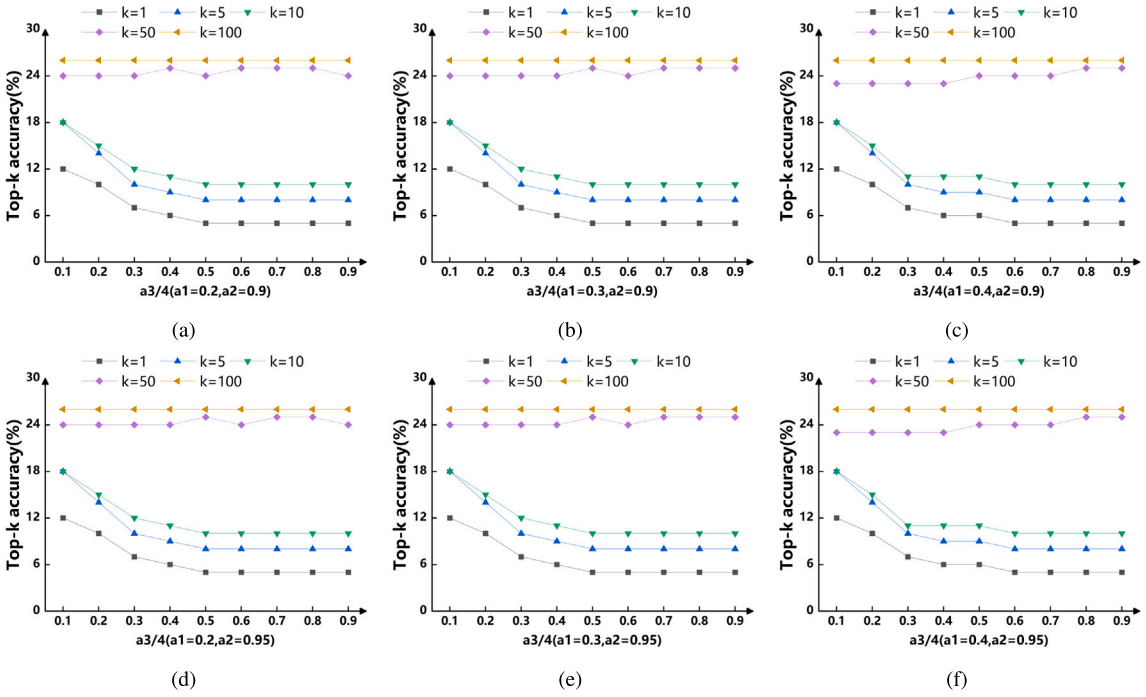

Fig. 6. Controlled variable experiments results of a_3 and a_4.

Table 3
The time cost of each link in BigTensorDB.

Operation	Data cleaning (A)	Model A	Data cleaning (B)	Model B	**Search**	**Re-rank and output**
Time (h)	1	10	1	10	**1**	**0.01**

5. Lessons and future directions

5.1. Limitations of BigTensorDB

In our work, despite conducting meticulous experiments and careful theoretical analysis, we still identify certain shortcomings:

(1) The selection of prediction models is not diverse enough. There are only 2 models for reactant prediction and 4 models for reaction conditions prediction in our study.

(2) The accuracy of the top-k results after re-ranking still has significant space for improvement.

5.2. Future work of BigTensorDB

In our future work, we will primarily focus on the following four directions:

Firstly, we will expand and explore more prediction models. Our goal is to integrate more state-of-the-art and advanced prediction models. These models will include those for predicting reactants and predicting reaction conditions. By integrating different prediction models, we are committed to provide users with a unified and standardized environment. We will offer a broader one-stop selection space.

Secondly, we need to further explore the boundary conditions of the storage structure. In the current work of this study, the boundary conditions of the reaction equation are stored as four parts: reactants,

products, solvents and reagents. Given the vast variety and diverse types of chemical reaction conditions, the boundary definition of these conditions still requires more rigorous experimental analysis. Moving forward, Our goal is to determine whether using these conditions as the key conditions for the reaction equation is accurate. We plan to continue applying methods from evaluation science. We will also deploy more comparative experiments.

Thirdly, we need to conduct more diversified explorations of the embedding models that convert reaction equations from SMILES format to vectors. Currently, we have chosen ChemBERTa-77M-MLM as the embedding model. However, related work shows that there are many other embedding options available. We plan to deploy richer comparative experiments to explore the embedding models that can best preserve chemical information. We aim to ensure that the embedding models we use accurately reflect all the chemical information contained in the reaction equations. This will enable our vector retrieval work to be accurate and efficient.

Lastly, within the retrieval layer, we still need to explore the more fine-grained weight parameters corresponding to different conditions to maximize the accuracy after re-ranking. This is crucial because the corresponding weights can significantly enhance the accuracy and reliability of the final outcomes.

6. Conclusion

This paper proposes the first tensor database system, BigTensorDB, to help scientists in retrosynthetic analysis field. Our work effectively addresses the critical issues of the absence of a unified model capable of providing full-cycle service for retrosynthetic analysis and the significant bottleneck in the prediction accuracy of full-cycle retrosynthetic analysis.

Specifically, BigTensorDB designs an innovative tensor format that efficiently stores all key information related to chemical reactions, including reactants, products, and reaction conditions such as solvents and reagents. This format not only provides users with robust services for storing and retrieving chemical reactions but also lays the foundation for more accurate and comprehensive analysis. Additionally, the integration of multiple retrosynthetic analysis prediction models, including those for reactants and reaction conditions, along with SMILES embedding models, offers a seamless full-cycle retrosynthetic analysis service to users. This integration significantly reduces usage costs and enhances the efficiency of the entire pipeline. Moreover, by providing advanced search and analysis services, this work re-ranks and analyzes the final prediction results, offering real reaction equations similar to the predicted results for user reference. These efforts collectively enhance the accuracy and interpretability of the final outcomes, thereby advancing the state of the art in AI-based retrosynthetic analysis.

CRediT authorship contribution statement

Xueya Zhang: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Formal analysis, Conceptualization. **Guoxin Kang:** Writing – review & editing, Methodology, Conceptualization. **Boyang Xiao:** Writing – original draft, Validation, Resources. **Jianfeng Zhan:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The author Jianfeng Zhan is an Editor-in-Chief for BenchCouncil Transactions on Benchmarks, Standards and Evaluations and was not involved in the editorial review or the decision to publish this article.

The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Innovation Funding of Institute of Computing Technology Chinese Academy of Sciences, China under Grant No. E461070 and Beijing Municipal Natural Science Foundation of Beijing Municipal Science and Technology Commission and Zhongguancun Science Park Administrative Committee, China under Grant No. QY24378.

References

[1] X. Wang, Y. Li, J. Qiu, G. Chen, H. Liu, B. Liao, C.-Y. Hsieh, X. Yao, RetroPrime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions, Chem. Eng. J. (2021) 129845.
[2] E.J. Corey, A.K. Long, S.D. Rubenstein, Computer-assisted analysis in organic synthesis, Science 228 (4698) (1985) 408–418.
[3] A. Heifets, I. Jurisica, Construction of New Medicines Via Game Proof Search, AAAI Press, 2012.
[4] M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature 555 (2017) 604–610.
[5] A. Kishimoto, B. Buesser, B. Chen, A. Botea, Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning, Curran Associates Inc., Red Hook, NY, USA, 2019.
[6] J.S. Schreck, C.W. Coley, K.J.M. Bishop, Learning retrosynthetic planning through simulated experience, ACS Central Sci. 5 (6) (2019) 970–981.
[7] K. Lin, Y. Xu, J. Pei, L. Lai, Automatic retrosynthetic route planning using template-free models, Chem. Sci. 11 (2020) 3355–3364.
[8] B. Chen, C. Li, H. Dai, L. Song, Retro*: Learning retrosynthetic planning with neural guided a* search, in: I.. Hal Daumé, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, vol. 119, PMLR, 2020, pp. 1608–1616.
[9] J. Kim, S. Ahn, H. Lee, J. Shin, Self-improved retrosynthetic planning, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, vol. 139, PMLR, 2021, pp. 5486–5495.
[10] S. Ishida, K. Terayama, R. Kojima, K. Takasu, Y. Okuno, AI-driven synthetic route design incorporated with retrosynthesis knowledge, J. Chem. Inf. Model. 62 (2022) 1357–1367.
[11] L. Yue, Z. Xinxin, Y. Zhengwei, S. Siqi, Machine learning embedded with materials domain knowledge, J. Chinese Ceramic Soc. 50 (3) (2022) 863–876.
[12] Y. Liu, L. Ding, Z. Yang, et al., Domain knowledge discovery from abstracts of scientific literature on nickel-based single crystal superalloys, Sci. China Technol. Sci. 66 (2023) 1815–1830.
[13] Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, S. Shi, Generative artificial intelligence and its applications in materials science: Current situation and future perspectives, J. Mater. 9 (4) (2023) 798–816.
[14] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. Mater. 3 (3) (2017) 159–177, High-throughput Experimental and Modeling Research toward Advanced Batteries.
[15] S. Siqi, T. Zhangwei, Z. Xinxin, S. Shiyu, Y. Zhengwei, L. Yue, Applying data-driven machine learning to studying electrochemical energy storage materials, Energy Storage Sci. Technol. 11 (3) (2022) 739–759.
[16] U.V. Ucak, I. Ashyrmamatov, J. Ko, J. Lee, Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments, Nat. Commun. 13 (2022).
[17] X. Wang, C.-Y. Hsieh, X. Yin, J. Wang, Y. Li, Y. Deng, D. Jiang, Z. Wu, H. Du, H. Chen, Y. Li, H. Liu, Y. Wang, P. Luo, T. Hou, X. Yao, Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center, Research 6 (2023) 0231.
[18] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[19] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q.L. Nguyen, S. Ho, J.L. Sloane, P.A. Wender, V.S. Pande, Retrosynthetic reaction prediction using neural sequence-to-sequence models, ACS Central Sci. 3 (2017) 1103–1113.
[20] S. Zheng, J. Rao, Z. Zhang, J. Xu, Y. Yang, Predicting retrosynthetic reactions using self-corrected transformer neural networks, J. Chem. Inf. Model. (2019).
[21] I.V. Tetko, P. Karpov, R.V. Deursen, G. Godin, State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis, Nat. Commun. 11 (2020).
[22] E. Kim, D. Lee, Y. Kwon, M.S. Park, Y.-S. Choi, Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables, J. Chem. Inf. Model. 61 (1) (2021) 123–133.
[23] S. Seo, Y.Y. Song, J.Y. Yang, S. Bae, H. Lee, J. Shin, S.J. Hwang, E. Yang, GTA: Graph truncated attention for retrosynthesis, in: AAAI Conference on Artificial Intelligence, 2021.
[24] Y. Jiang, Y. Wei, F. Wu, Z. Huang, K. Kuang, Z. Wang, Learning chemical rules of retrosynthesis with pre-training, in: AAAI Conference on Artificial Intelligence, 2023.

[25] Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M.-Y. Wu, T. Hou, M. Song, Root-aligned SMILES: A tight representation for chemical reaction prediction, Chem. Sci. 13 (2022) 9023–9034.
[26] Y. Zhang, R. Yu, K. Zeng, D. Li, F. Zhu, X. Yang, Y. Jin, Y. Xu, Text-augmented multimodal LLMs for chemical reaction condition recommendation, 2024, ArXiv abs/2407.15141.
[27] V.R. Somnath, C. Bunne, C.W. Coley, A. Krause, R. Barzilay, Learning graph models for retrosynthesis prediction, NIPS '21, Curran Associates Inc., Red Hook, NY, USA, 2021.
[28] Y. Wan, C.-Y. Hsieh, B. Liao, S. Zhang, Retroformer: Pushing the limits of end-to-end retrosynthesis transformer, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, vol. 162, PMLR, 2022, pp. 22475–22490.
[29] Y. Han, et al., Retrosynthesis prediction with an iterative string editing model, Nat. Commun. (2024).
[30] C.W. Coley, L. Rogers, W.H. Green, K.F. Jensen, Computer-assisted retrosynthesis based on molecular similarity, ACS Central Sci. 3 (2017) 1237–1245.
[31] M.H.S. Segler, M.P. Waller, Neural-symbolic machine learning for retrosynthesis and reaction prediction, Chemistry 23 25 (2017) 5966–5971.
[32] H. Dai, C. Li, C.W. Coley, B. Dai, L. Song, Retrosynthesis Prediction with Conditional Graph Logic Network, Curran Associates Inc., Red Hook, NY, USA, 2019.
[33] S. Chen, Y. Jung, Deep retrosynthetic reaction prediction using local reactivity and global attention, JACS Au 1 (10) (2021) 1612—1620.
[34] J. Dong, M. Zhao, Y. Liu, Y. Su, X. Zeng, Deep learning in retrosynthesis planning: datasets, models and tools, Brief. Bioinform. (2021).
[35] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[36] Z. Tu, C.W. Coley, Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction, J. Chem. Inf. Model. (2021).
[37] M. Sacha, M. Blaz, P. Byrski, P. Wlodarczyk-Pruszynski, S. Jastrzebski, Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits, J. Chem. Inf. Model. (2020).
[38] J. Liu, C. chao Yan, Y. Yu, C. Lu, J. Huang, L. Ou-Yang, P. Zhao, MARS: A motif-based autoregressive model for retrosynthesis prediction, Bioinformatics 40 (2022).
[39] W. Zhong, Z. Yang, C.Y.-C. Chen, Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing, Nat. Commun. 14 (2023).
[40] L. Fang, J. Li, M. Zhao, L. Tan, J.-G. Lou, Single-step retrosynthesis prediction by leveraging commonly preserved substructures, Nat. Commun. 14 (2023).
[41] C. Shi, M. Xu, H. Guo, M. Zhang, J. Tang, A graph to graphs framework for retrosynthesis prediction, in: International Conference on Machine Learning, 2020.
[42] C. chao Yan, Q. Ding, P. Zhao, S. Zheng, J. Yang, Y. Yu, J. Huang, RetroXpert: Decompose retrosynthesis prediction like a chemist, 2020, ArXiv arXiv:2011.02893.
[43] H. Gao, T.J. Struble, C.W. Coley, Y. Wang, W.H. Green, K.F. Jensen, Using machine learning to predict suitable conditions for organic reactions, ACS Central Sci. 4 (2018) 1465–1476.
[44] Y. Liu, Z. Yang, X. Zou, S. Ma, D. Liu, M. Avdeev, S. Shi, Data quantity governance for machine learning in materials science, Natl. Sci. Rev. 10 (7) (2023) nwad125–.
[45] Y. Liu, X. Ge, Z. Yang, S. Sun, D. Liu, M. Avdeev, S. Shi, An automatic descriptors recognizer customized for materials science literature, J. Power Sources 545 (2022) 231946.
[46] S. Siqi, SUN, M. Shuchang, Z. Xinxin, Q. Quan, L. Yue, Detection method on data accuracy incorporating materials domain knowledge, J. Inorg. Mater. 37 (12) (2022) 1311–1320.
[47] L. Yue, Y. Wenxuan, L. Dahui, D. Lin, Y. Zhengwei, L. Wei, Y. Tao, S. Siqi, Named entity recognition driven by high-quality text data accelerates the knowledge discovery of nickel-based single crystal superalloys, Acta Metall. Sin. 60 (10) (2024) 1429–1438.
[48] L. Yue, L. Da-Hui, G. Xian-Yuan, Y. Zheng-We, M. Shu-Chang, Z.Z. Yi, S.S.-Q. 2, A high-quality dataset construction method for text mining in materials science, Acta Phys. Sin. 72 (7) (2023) 41–54.
[49] L. Yue, M. Shuchang, Y. Zhengwei, Z. Xinxin, S. Siqi, A data quality and quantity governance for machine learning in materials science, J. Chinese Ceramic Soc. 51 (2) (2023) 427–437.
[50] Y. Liu, J.M. Wu, M. Avdeev, S.Q. Shi, Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties, Adv. Theory Simul. 3 (2) (2020).
[51] Q. Zhao, L. Zhang, B. He, A. Ye, M. Avdeev, L. Chen, S. Shi, Identifying descriptors for li+ conduction in cubic li-argyrodites via hierarchically encoding crystal structure and inferring causality, Energy Storage Mater. 40 (2021) 386–393.
[52] Y. Liu, X. Zou, S. Ma, M. Avdeev, S. Shi, Feature selection method reducing correlations among features by embedding domain knowledge, Acta Mater. 238 (2022) 118195.
[53] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, E. Ozkirimli, Exploring chemical space using natural language processing methodologies for drug discovery, Drug Discov. Today (2020).
[54] X. Li, D. Fourches, SMILES pair encoding: A data-driven substructure tokenization algorithm for deep learning, J. Chem. Inf. Model. 61 (4) (2021) 1560–1569.
[55] I. Lee, H. Nam, Infusing linguistic knowledge of SMILES into chemical language models, 2022, ArXiv abs/2205.00084.
[56] B. Fabian, T. Edlich, H. Gaspar, M.H.S. Segler, J. Meyers, M. Fiscato, M. Ahmed, Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020, ArXiv abs/2011.13230.
[57] S. Wang, Y. Guo, Y. Wang, H. Sun, J. Huang, SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019.
[58] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020, ArXiv abs/2010.09885.
[59] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa-2: Towards chemical foundation models, 2022, ArXiv abs/2209.01712.
[60] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019.
[61] T. Tran, C. Ekenna, Molecular descriptors property prediction using transformer-based approach, Int. J. Mol. Sci. 24 (2023).
[62] Y. Liu, R. Zhang, T. Li, J. Jiang, J. Ma, P. Wang, MolRoPE-BERT: An enhanced molecular representation with rotary position embedding for molecular property prediction, J. Mol. Graph. 118 (2022) 108344.
[63] Y. Liu, R. Zhang, T. Li, J. Jiang, J. Ma, P. Wang, MolRoPE-BERT: An enhanced molecular representation with rotary position embedding for molecular property prediction, J. Mol. Graph. 118 (2022) 108344.
[64] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, K. Yu, Y. Yuan, Y. Zou, J. Long, Y. Cai, Z. Li, Z. Zhang, Y. Mo, J. Gu, R. Jiang, Y. Wei, C. Xie, Milvus: A purpose-built vector data management system, in: Proceedings of the 2021 International Conference on Management of Data, 2021.
[65] C. Wei, B. Wu, S. Wang, R. Lou, C. Zhan, F. Li, Y. Cai, AnalyticDB-V: A hybrid analytical engine towards query fusion for structured and unstructured data, Proc. VLDB Endow. 13 (12) (2020) 3152–3165.
[66] W. Yang, T. Li, G. Fang, H. Wei, PASE: Postgresql ultra-high-dimensional approximate nearest neighbor search extension, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020.
[67] Vearch: A distributed system for embedding-based retrieval, 2020, URL: https://github.com/vearch/vearch.
[68] J. Li, H.-F. Liu, C. Gui, J. Chen, Z. Ni, N. Wang, Y. Chen, The design and implementation of a real time visual search system on jd E-commerce platform, in: Proceedings of the 19th International Middleware Conference Industry, 2018.
[69] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (1) (2024) 100162.