THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Artificial Intelligence-Powered Construction of a Microbial Optimal Growth Temperature Database and Its Impact on Enzyme Optimal Temperature Prediction

Published as part of The Journal of Physical Chemistry B virtual special issue "Women Scientists in China".

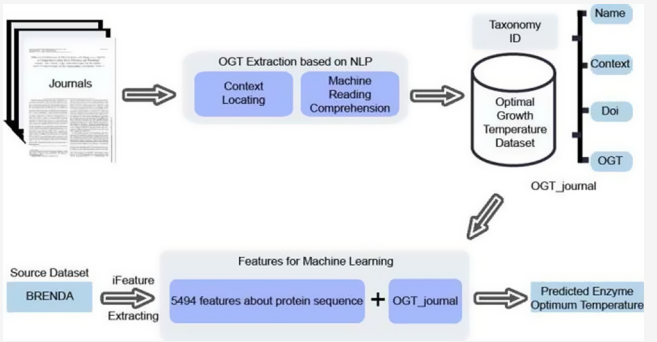
Xiaotao Wang, Yuwei Zong,‖ Xuanjie Zhou,‖ Li Xu, Wei He,* and Shu Quan*

Cite This: *J. Phys. Chem. B* 2024, 128, 2281−2292

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Accurate prediction of enzyme optimal temperature (Topt) is crucial for identifying enzymes suitable for catalytic functions under extreme bioprocessing conditions. The optimal growth temperature (OGT) of microorganisms serves as a key indicator for estimating enzyme Topt, reflecting an evolutionary temperature balance between enzyme-catalyzed reactions and the organism's growth environments. Existing OGT databases, collected from culture collection centers, often fall short as culture temperature does not precisely represent the OGT. Models trained on such databases yield inadequate accuracy in enzyme Topt prediction, underscoring the need for a high-quality OGT database. Herein, we developed AI-based models to extract the OGT information from the scientific literature, constructing a comprehensive OGT database with 1155 unique organisms and 2142 OGT values. The top-performing model, BioLinkBERT, demonstrated exceptional information extraction ability with an EM score of 91.00 and an F1 score of 91.91 for OGT. Notably, applying this OGT database in enzyme Topt prediction achieved an R^2 value of 0.698, outperforming the R^2 value of 0.686 obtained using culture temperature. This emphasizes the superiority of the OGT database in predicting the enzyme Topt and underscores its pivotal role in identifying enzymes with optimal catalytic temperatures.

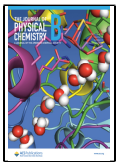


INTRODUCTION

Environmental microorganisms serve as a valuable reservoir of industrially significant enzymes. The preferred enzyme must exhibit stability and activity under bioprocessing conditions with its temperature optimum (Topt) ideally aligned with the operating temperature to ensure optimal bioprocessing outcomes. The majority of industrial enzymes are derived from mesophilic organisms that thrive at moderate temperatures ranging from 20 to 45 °C, with optimum growth temperatures (OGT) typically falling between 30 and 39 °C.[1] In contrast, a smaller subset of industrial enzymes is sourced from extremophilic microbes, including thermophiles inhabiting high-temperature environments (55−121 °C)[2] and psychrophiles adapted to frigid temperatures (−2 to 20 °C).[3] These less-explored thermophiles and psychrophiles present highly promising resources for the discovery of new enzymes, particularly those with Topt values suitable for extreme bioprocessing conditions.

Out of the 34,082 identified enzymes, only 11,250 have reported Topt values in public databases (https://www.brenda-enzymes.org/). Consequently, researchers have sought to develop tools for the in silico prediction of the enzyme Topt. Li et al. discovered a close association between the OGT of an organism and the Topt of enzymes derived from that organism.[4] Leveraging this insight, they developed machine learning models to predict enzyme Topt by training the model with an OGT dataset and enzyme features.[5] Subsequently, Gado et al. enhanced the accuracy of Topt prediction by employing resampling and ensemble learning strategies to address data imbalance issues within the OGT dataset, resulting in a 60% decrease in errors for high Topt values (>85 °C).[6] However, despite these efforts, the accuracy of the enzyme Topt prediction remains below the desired level. One key factor influencing model prediction accuracy is the OGT dataset used. The OGT dataset utilized in the aforementioned studies was compiled by Engquist, who aggregated culture temperatures from over 18,000 organisms across various

Received: September 30, 2023
Revised: February 6, 2024
Accepted: February 15, 2024
Published: March 4, 2024



https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

microbial culture databases.[7] This OGT dataset, in fact, represents the culture temperature rather than the actual OGT, which reflects the long-term environmental adaptation of microbes through an evolutionary process. Therefore, constructing a high-quality actual OGT database and utilizing it for model training could potentially lead to an increase in the prediction accuracy of the enzyme Topt.

The scientific literature stands out as a crucial source for constructing databases. An exemplary case is the Reaxys chemical reaction database by Elsevier, which compiles chemical reaction information from over 16,000 scientific journals. Consequently, there is potential for building a high-quality OGT database by mining relevant scientific papers. However, information retrieval through manual methods is labor-intensive, and the recent exponential growth of scientific publications has exacerbated the inadequacy of such expertise-intensive extraction tasks. Consequently, addressing this challenge has spurred the development of automated methods for extracting structured data from the unstructured literature. With the rise of artificial intelligence (AI) technologies, particularly deep learning, automatic extraction of data from text has become feasible through the application of Natural Language Processing (NLP) techniques. For instance, Guo et al. utilized a large-scale unlabeled corpus extracted from chemical literature for language model pretraining, achieving commendable accuracy in chemical product extraction and reaction role extraction.[8] In the realm of third-generation semiconductor materials (TGSMs), Jiang et al. employed deep learning for the automatic extraction of TGSM-related entities and their relationships, substantially easing the information retrieval burden for researchers.[9] Additionally, with the advent of domain-specific language models in biomedicine, AI-assisted approaches have demonstrated significant enhancements across various NLP tasks on standard biomedical benchmarks, such as bioasq by Tsatsaronis et al.,[10] BLURB by Kanakarajan et al.,[11] MedQA-USMLE by Yao et al.,[12] as well as Pubmedqa by Jin et al.[13] These advancements underscore the feasibility of constructing an actual OGT database through an AI-based approach.

In this study, we successfully constructed an OGT database comprising 1155 organisms and 2142 OGT values through AI-based information extraction from over 100,000 relevant scientific studies. Of these, 1002 organisms are shared between our OGT database and the culture temperature database. Remarkably, these shared organisms tend to exhibit overall higher growth temperature values in our OGT database compared to their corresponding values in the culture temperature database. Additionally, the model trained with the OGT database demonstrates improved performance in predicting enzyme Topt compared with the model trained with the culture temperature database. These findings underscore the distinction between an organism's optimal growth temperature and its culture temperature, emphasizing a stronger correlation between an enzyme's Topt and the OGT value of the organism it originates from, rather than its culture temperature.

METHODS

Software. The machine learning analyses were implemented using scikit-learn. Python version 3.9 was utilized for both deep learning and machine learning workflows. The Python source code, along with all datasets, is publicly available on the GitHub repository (https://github.com/seantaud/OGT-Extraction-and-Application). Relevant information and acquisition addresses for the six Pretrained Language Models (PLMs) in this study are provided in Supplementary Table 1. Deep learning and fine-tuning tasks were conducted within the PyTorch and Hugging Face framework. The deep learning experiments utilized two different GPUs for distinct model architectures, with BERT models trained on a NVIDIA V100 GPU and GPT models executed on an A100 GPU. Machine learning tasks were carried out on clusters of 12 Intel(R) Xeon(R) CPU E5-2696 v3.

Literature Acquisition. The acquisition of the literature dataset involved utilizing the "elsapy" module to interact with the Elsevier API, which is grounded in the literature index provided by Elsevier. Initially, a keyword-based search strategy was employed to retrieve a substantial number of matching articles. In the keyword retrieval strategy, a curated list of biological organism names was generated through web crawling on the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) platform. This list included taxonomy, scientific names, and alternative designations for a total of 18,854 microbial species. The outcome of literature retrieval includes detailed information about the articles, with DOI numbers functioning as key identifiers for linking to the text of each article. The final dataset comprises 518,674 articles, each systematically categorized and stored using biological identifiers.

Context Locating. For data preprocessing, we employed a context-locating approach to extract sentences of constrained length from the literature dataset, which served as input text for the deep learning model. Text segmentation algorithms from the Natural Language Toolkit (NLTK) were employed to divide the extensive text data into discrete sentences. These algorithms consider common abbreviations in organism names, punctuation marks, and omissions such as "*M. fulvus*", "Rhizobium sp.", and "et al.". Once the sentences were segmented, processing was conducted on the full content of each article to locate paragraphs containing organism names and related temperature information. A set of regular expressions was applied to match organism names, followed by another set of regular expressions to identify temperature symbols such as "number" + "°C", "number" + "degrees C", and "degrees Celsius". This process recorded the positions of these names and temperatures within the corresponding sentence fragments.

Pretrained Language Models (PLMs). Three BERT-based models (BioBERT-base,[14] BioBERT-large,[14] and BioLinkBERT[15]), alongside three GPT-based models (BioGPT,[16] BioGPT-large,[16] and ChatGPT[17]), were employed in this work (Supplementary Table 2). These language models, which have undergone pretraining using a self-supervised approach on large-scale unlabeled datasets, could make task-adaptive fine-tuning and achieve remarkable results on multiple NLP tasks. In this work, these models, excluding ChatGPT, were fine-tuned for OGT extraction using our annotated datasets known as OGT-QA. These steps were instrumental in optimizing the models for improved performance in biomedical domain tasks.

Machine Reading Comprehension. We approached the task of extracting OGT information from research articles within the framework of machine reading comprehension (MRC), employing two distinct formats: extractive question answering (EQA) and generative question answering (GQA).

Extractive Question Answering. EQA aimed to identify OGT in biomedical articles by predicting the start position

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

$P_{start}(i)$ and end position $P_{end}(j)$ of the OGT-related information within the context (Supplementary Figure 1). We adopted the methodology introduced by Devlin and colleagues in the biomedical task EQA to predict answer spans related to OGT.[18] The prediction involved applying softmax operations to the dot products of contextual embeddings h_i and h_j and learnable weight matrices W_s and W_e, as illustrated in the following formulas:

$$P_{start}(i) = \text{softmax}(W_s \cdot h_i) \quad (1)$$

$$P_{end}(j) = \text{softmax}(W_e \cdot h_j) \quad (2)$$

Generative Question Answering. Contrastingly, GQA considered the question and context holistically, generating tokens $P(x_t \mid x_{<t}; \theta)$ based on a transformed hidden state (Supplementary Figure 2). In this process, GPT dynamically generated tokens one at a time by predicting their probabilities from a hidden state h_t of GPT and a learnable weight matrix W_o. The token generation process in GQA for biomedical applications is mathematically represented as

$$P(x_t|x_{<t}; \theta) = \text{softmax}(W_o \cdot h_t) \quad (3)$$

Before the EQA or GQA processes, input texts were tokenized using Wordpiece (as applied in BERT) and Byte-Pair Encoding (utilized in GPT). Subsequently, these tokenized inputs were subjected to both EQA and GQA, each with distinctive input formats. This approach leveraged the capabilities of biomedical PLMs to accomplish the task of OGT extraction from the biomedical literature.

Fine-Tuning Method. Three methods, full fine-tuning,[19] Prefix tuning,[20] and P-tuning-v2,[21] were employed for fine-tuning the biomedical PLMs. Full fine-tuning updated all parameters of Transformer blocks and required storage of an entire model replica for each distinct task. Prefix tuning and P-tuning-v2 were lightweight fine-tuning approaches, entailing freezing most of the pretrained parameters and fine-tuning the model with small trainable modules.

The steps for using fine-tuning methods for PLMs were as follows: (1) Import pretrained models based on the Pytorch and Hugging Face framework. (2) Design EQA and GQA classes for language models. If Prefix Tuning or P-Tuning v2 is used, a prefix prompt module is added to models. (3) Set hyperparameters for training, including learning rate and batch size, and add parameters such as num_virtual_tokens (for Prefix Tuning) or prefix sequence length (for P-Tuning v2). (4) Train the models and save the weights. Update the parameters of Transformer blocks only when using full fine-tuning. (5) Load the fine-tuned model for inference.

A total of 5 hyperparameters are tuned for language model training: per_device_train_batch_size, learning_rate, epochs, patience, and the prefix length (this parameter was named as pre_seq_len for P-tuning-v2 or num_virtual_tokens for Prefix tuning). The values for these hyperparameters are provided in Supplementary Table 3. Additionally, for the GPT model, the dataset was augmented by a factor of 16, and the maximum input length was set to 768. For the BERT model, a maximum input length of 384 and doc_stride of 128 for text truncation was applied.

Dataset Construction and Manual Evaluation. *Datasets for Deep Learning.* The OGT-QA dataset was constructed by an iterative, multistage process starting from manual curation of OGT descriptions in the scientific literature and represented in a format similar to Stanford Question Answering Datasets (SQuAD).[22] This dataset was partitioned into training, validation, and test sets in a 6:2:2 ratio, ensuring a balanced distribution of answerable and unanswerable samples.

Datasets for Machine Learning. Three datasets were used: OGT_journal dataset, OGT_cultivation dataset (downloaded from 10.5281/zenodo.1175608), and Topt dataset (collected from BRENDA).

The **OGT_journal dataset** was generated by performing large-scale inference on the literature contexts using the fine-tuned BioLinkBERT language model, followed by manual verification and deduplication of similar data entries, which were carried out as follows: (1) remove data entries without corresponding OGT values; (2) replace incorrectly predicted OGT values with manually reviewed true values; (3) add new entries if a verified context contained OGT description for other organisms or multiple OGT descriptions for the same organism. After manual verification, we observed many data entries sharing identical organisms, identical values of the OGT, and highly similar contextual content. To minimize data redundancy, we categorized the data into distinct groups based on organisms and the values of the OGT and performed deduplication within each group. The deduplication process involved comparing each entry with subsequent ones. Entries overlapped with each other for more than 50% of the original text were marked and the entry with the shortest text was kept (The specific deduplication process was exampled in Supplementary Table 4). This process continued until all of the redundant elements were removed.

The **Topt dataset** was downloaded from BRENDA by web crawling focusing on the Enzyme EC number, experimental comments, and Topt value, followed by a dataset refinement procedure similar to that described by Li et.al.[5] Briefly, several steps were involved: (1) enzymes containing noncanonical amino acids or having fewer than 30 residues were systematically excluded. (2) enzymes for which the Topt values were referenced in the BRENDA "comments" field with the phrase "assay at" were removed. (3) enzymes were considered identical if they have the same Uniprot identifiers, the same Topt values, and the same first three digits of the EC numbers. (4) If an enzyme corresponds to more than one Topt value, the mean and standard deviation are first calculated. Topt values that deviate by more than 5 °C from the standard deviation are removed, and then the mean value is used as the representative Topt value for that enzyme. (5) For Topt values expressed as intervals, the midpoint is used as the representative Topt value. Following these refinement steps, a final Topt dataset comprised of 4728 enzymes was constructed.

To facilitate a fair comparison of the impact of the OGT data and culture temperature in Topt prediction experiments, we preserved the temperature data for organisms that appeared in both the OGT_journal dataset and the OGT_cultivation dataset. Subsequently, we correlated the Topt data with the aforementioned organisms. Approximately 40% of the entries in the Topt_dataset could be associated with organisms that had temperature values in both the OGT_journal and the OGT_cultivation datasets.

Metrics. *Evaluate Deep Learning Models.* In the evaluation phase of our deep learning models, we specifically employed the EM score and the F1 score to assess the extraction capabilities of language models, including BioBERT and other language models. These metrics were chosen to

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

measure the models' proficiency in extracting OGT from diverse contexts in the final "OGT-QA" dataset.

The **exact match (EM) score** measures the percentage of model predictions that exactly match the ground truth values. The EM score is calculated as follows:

$$\mathrm{EM} = \frac{N_{\mathrm{exact-match}}}{N_{\mathrm{total}}} \times 100\% \tag{4}$$

where $N_{\mathrm{exact-match}}$ is the number of predictions that exactly match the ground truth answers, and N_{total} is the total number of samples in the dataset.

The **F1 score**, considering both precision and recall, is calculated as

$$\mathrm{F1} = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{5}$$

where $\mathrm{precision} = \frac{N_{\mathrm{true-positives}}}{N_{\mathrm{predicted-positives}}}$ and $\mathrm{recall} = \frac{N_{\mathrm{true-positives}}}{N_{\mathrm{actual-positives}}}$. $N_{\mathrm{true-positives}}$ is the number of correctly predicted tokens, $N_{\mathrm{predicted-positives}}$ is the total number of predicted tokens, and $N_{\mathrm{actual-positives}}$ is the total number of tokens in the ground truth. The final F1 score is calculated as the average F1 score over all samples in the dataset.

Correlation Coefficient Comparison. To compare the correlation coefficient between the OGT_journal database and Topt and that between the OGT_cultivation database and Topt, their values were transformed into Z scores using Fisher's Z-transformation (eq 6). Then, the test statistic z and probability p are computed (eq 7−9) to assess statistical significance (Diedenhofen et al.[23]).

$$Z = \frac{1}{2}(\ln(1 + r) - \ln(1 - r)) \tag{6}$$

Herein, Z represents Z scores and r represents correlation coefficient.

$$z = \frac{(Z_{12} - Z_{13})\sqrt{n - 3}}{\sqrt{2 - 2c}} \tag{7}$$

where

$$c = \frac{r_{23}(1 - 2\overline{r}^2) - \frac{1}{2}\overline{r}^2(1 - 2\overline{r}^2 - {r_{23}}^2)}{(1 - \overline{r}^2)^2} \tag{8}$$

and

$$\overline{r} = \frac{r_{12} + r_{13}}{2} \tag{9}$$

Specifically in these equations, r_{12} and r_{13} represent the correlation coefficients between the OGT_journal database and Topt, and between the OGT_cultivation database and Topt, respectively; Z_{12} and Z_{13} denote their corresponding Z transformation scores. r_{23} is the correlation coefficient between the OGT_journal database and the OGT_cultivation database; n specifies the size of the dataset from which the correlation coefficients are derived. By convention, if the absolute value of the test statistic z is greater than 1.96, the test is considered statistically significant.

Evaluate Machine Learning Models. We employed four pivotal metrics including the determination coefficient (R^2 score), which gauges the model's ability to elucidate observed variability; the mean absolute error (MAE), a metric assessing the average absolute prediction differences; the mean squared error (MSE), which calculates the overall squared differences between predicted and observed values; and the root mean square error (RMSE), capturing the accuracy of predictions by evaluating residual errors.

The **determination coefficient** (R^2*score*) was calculated by using the following formula:

$$R^2 = 1 - \frac{\mathrm{SS}_{res}}{\mathrm{SS}_{tot}} \tag{10}$$

where SS_{res} stands for the cumulative sum of squared differences between the model's predictions and the ground truth. SS_{tot} is the cumulative sum of squared differences between the observed ground truth and their mean.

The **root mean square error** (RMSE) is calculated by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{11}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the total number of observations.

The **mean absolute error** (MAE) is calculated by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

The **mean squared error** (MSE) is calculated by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

RESULTS AND DISCUSSION

Acquisition and Preprocessing of Scientific Studies Relevant to Microbial Growth. To extract descriptions related to the OGT, we initiated the process by retrieving the full text of relevant scientific papers from the Elsevier publication database. For this purpose, we employed an automatic procedure to search the Elsevier publication database using query statements such as "name + optimal growth temperature", "name + optimal temperature", or "name + optimal". To ensure precise organism identification, we developed a comprehensive look-up table that integrated taxonomy, scientific names, and alternative names of 18,854 microorganisms. This information was gathered from the NCBI website through a web crawler. In total, we obtained 518,674 articles containing both organism names and temperature-related descriptions.

Next, to enhance computational efficiency and reduce the complexity of the OGT extraction, we implemented a data preprocessing step to extract temperature-related sentences from the textual content of these articles. This operation, termed "Context Locating", involved several steps. We started by performing sentence segmentation and then identifying sentences containing organism names. Subsequently, we searched for temperature-related descriptions, such as "number + °C", "number + degree C", or other variations denoting temperature, within the proceeding texts. To ensure efficiency, we limited the search length to 40 sentences, as we observed that most temperature descriptions were located within 20 sentences of the organism name. To construct the "context", we extracted the sentence containing the organism name as the first sentence and the one containing the temperature description as the last sentence. Finally, we compiled a total

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

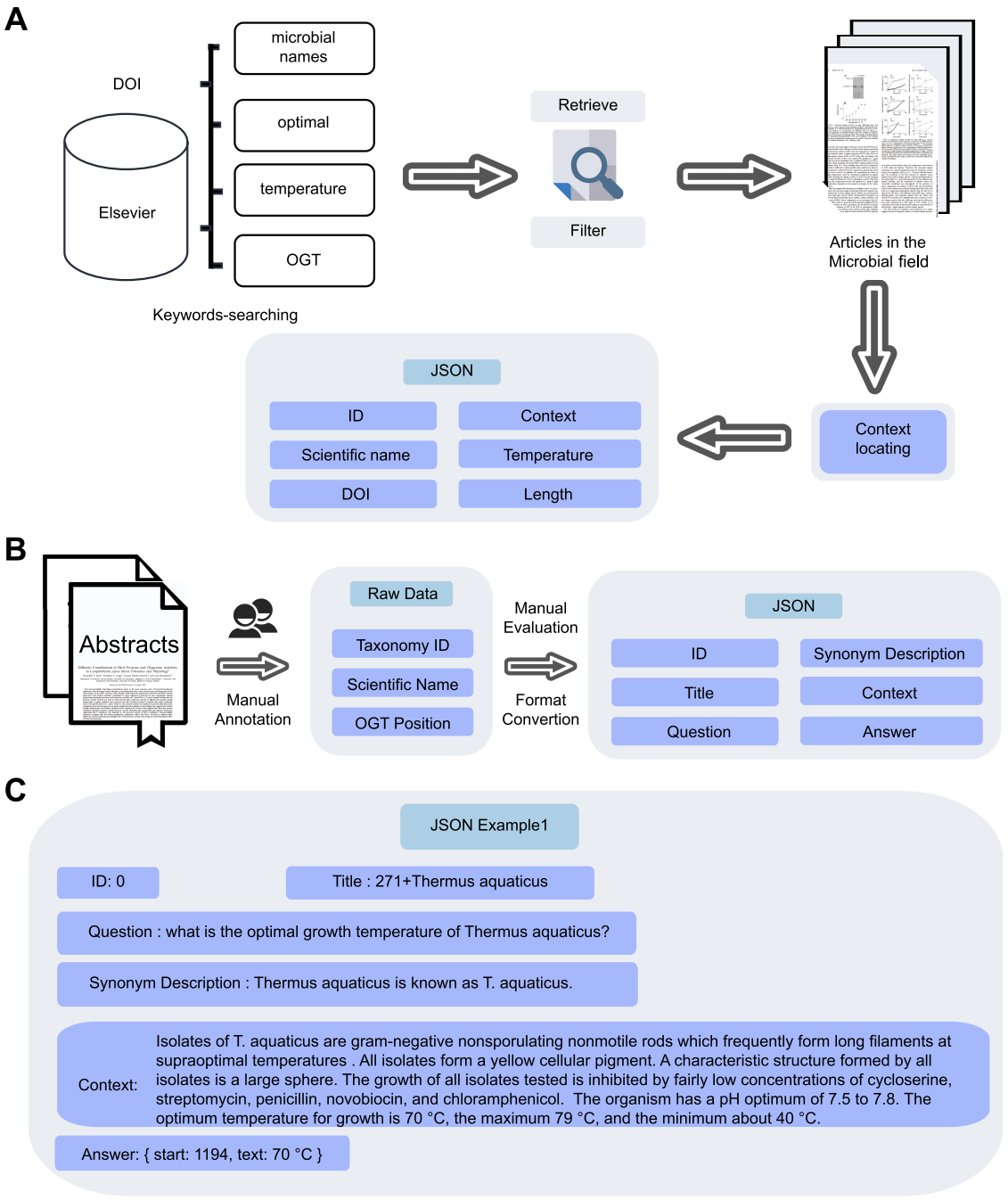


Figure 1. Workflow of literature acquisition, text preprocessing, and data annotation. (A) Keyword searching was performed using the elsapy Python module integrated with the Elsevier publication database. Full-text content retrieval utilized articles' DOI numbers, followed by sentence segmentation. For context locating, sentences containing organism names and temperature-related information were identified using regular expressions. Extracted context was organized into structured JSON files for subsequent analysis. (B) Construction of an annotated OGT-QA dataset involved a manual review of abstract texts. Information in standardized format was stored in JSON files. (C) An example of a positive instance includes an identification number, a title (comprising the taxonomy ID and scientific name), a question about the OGT of the organism, a synonym description of alternative organism names, and the answer containing temperature information with its starting position (precisely indicating the OGT location) along with the surrounding context.

of 983,680 contexts sourced from 108,316 articles and stored them in JSON files, which consumed approximately 8 GB of disk space (Figure 1A).

Creation of an Annotated OGT-QA Dataset. It is crucial to acknowledge that the temperature information within the contexts may contain additional irrelevant temperature data alongside the corresponding OGT information. To isolate authentic OGT information, we proposed leveraging the framework of Machine Reading Comprehension (MRC) and addressing this issue through Extractive/Generative Question Answering (QA) with Natural Language Processing (NLP) models. The reason we chose the MRC framework over typical Named Entity Recognition (NER) models, as implemented by Guo et al.[8] and Jiang et al.,[9] is guided by the specific requirements of our data extraction task.

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

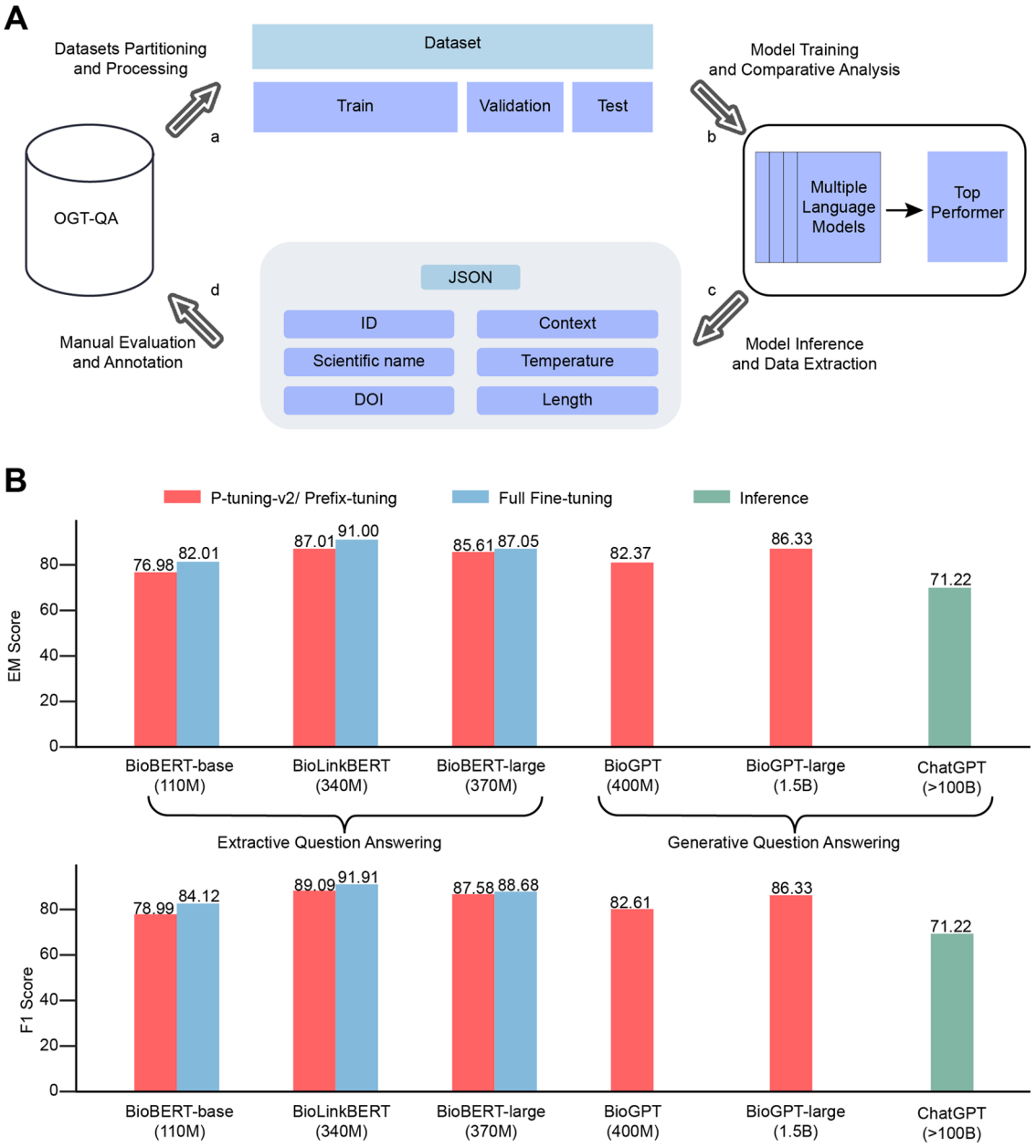


Figure 2. Biomedical language model training process and performance comparison. (A) Illustration of the iterative processes employed in the training of biomedical language models. (B) Comparative analysis of model performance using the final OGT-QA dataset. The BERT-based models underwent both p-tuning-v2 and full fine-tuning, while GPT-based models (BioGPT and BioGPT-large) underwent only prefix tuning, given their substantial model dimensions and the complexity of Generative Question Answering (GQA). Notably, ChatGPT was exclusively employed for answer generation during testing due to the unavailability of a publicly available fine-tuning service by OpenAI. EM score, a metric for text matching; F1 score, the harmonic mean of precision and recall with both the predictions and ground truth treated as bags of tokens in the text (see Methods). The numbers in parentheses indicate the parameter size for each language model.

Our primary objective is the extraction of OGT data for various microorganisms. MRC, with its focus on understanding the content of a passage and answering questions related to it, aligns well with the nature of our task. By utilizing queries encoding prior knowledge, such as organism names and synonym descriptions, MRC enables language models to focus directly on understanding the concept of OGT within the context of literature, as demonstrated in the work by Li et al.[24] Directly querying the models with relevant information allows us to achieve a more precise and focused extraction of the OGT data. In contrast, NER models necessitate token-by-token classification across predetermined categories. This approach can introduce potential redundancy, increase computational costs, and impose a higher annotation burden.[24] Given our specific focus on the extraction of the OGT, MRC proves to be a more efficient and targeted solution for our research objectives.

To train the language model, we first examined the descriptions of OGT in journals and curated an annotated dataset named the OGT-QA through manual editing. We specifically chose over 300 papers published in Microbiology Society (MS) and American Society for Microbiology (ASM) journals, focusing on abstracts that explicitly described the OGT data (Figure 1B). Within each abstract paragraph, we

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

performed annotation, incorporating crucial details such as the scientific name and taxonomy ID of the organism, along with the OGT. Subsequently, we converted these data into a standardized format, encompassing a question that inquired about the OGT of the organism, an answer comprising the temperature text and its starting position, and the context containing the answer (Figure 1C).

During the manual annotation process, we encountered a diverse array of OGT descriptions. While a majority of articles articulated the OGT as "the optimal growth temperature is...," there were various alternative descriptions such as "the fastest growth rate was observed at 25°C" and "optimal growth for *F. equiseti* strains cultivated in wheat and barley media was observed at 20−30°C."[25] To accommodate this diversity in the interpretation of the OGT descriptions, we intentionally selected texts with different narrative styles, resulting in 320 samples. Within these samples, 11 were designated as negative instances, lacking relevant information about the OGT. The remaining 308 samples constituted positive instances, each showcasing the authentic OGT information. This compilation of samples constituted the initial dataset utilized to train the NLP model for extracting OGT, which we denoted as the initial OGT-QA dataset.

Biomedical Language Model Training for OGT Information Extraction. In the development of biomedical language models adept at extracting OGT data from diverse contexts, we established an iterative, multistage model construction process (Figure 2A). This approach fostered a dynamic interplay between human expertise and machine learning, comprising four distinct phases that underwent three successive iterations: dataset partitioning and processing, model training and comparative analysis, model inference and data extraction, as well as manual evaluation and annotation.

To outline the process in greater detail, we first split the initial OGT-QA dataset into training, validation, and test subsets, followed by formatting and tokenization (see Methods). Subsequently, we conducted model training and comparative analyses employing three BERT-based models (BioBERT-base, BioBERT-large, and BioLinkBERT), in conjunction with three GPT-based models (BioGPT, BioGPT-large, and ChatGPT). All models, except ChatGPT, were subjected to training and validating using the designated datasets with either full fine-tuning or prefix tuning/p-tuning-v2 strategies, as outlined in the Methods section. Following this refinement process, the top-performing model on the test dataset was selected to predict OGT-related information within hundreds of contexts, which were randomly sampled from the JSON files containing the entire unlabeled contexts. The generated predictions underwent rigorous manual review, enabling the refinement of the initial OGT-QA dataset: instances deemed irrelevant to OGT descriptions were included as negative samples, while those genuinely pertinent were integrated as positive samples.

After three iterative cycles of the aforementioned procedures, the final "OGT-QA" dataset was established, comprising 1398 samples, with 622 instances classified as negative and 776 as positive (Table 1). Remarkably, the precision in predictions during the evaluation stage experienced a substantial improvement from 0.200 to 0.521, indicating a noteworthy enhancement in the capabilities of the biomedical language models. Among these models, BioLinkBERT, following training with the final "OGT-QA" dataset, emerged as the top performer on the test dataset, achieving the highest EM score of 91.00 and F1 score of 91.91 (Figure 2B). Notably, the BioGPT-large model, featuring a considerably smaller parameter size compared to ChatGPT (1.5 billion vs over 100 billion parameters), substantially outperformed ChatGPT, which was not subjected to our model training protocol. These findings underscore the effectiveness of fine-tuning biomedical models with a meticulously curated dataset, showcasing their remarkable potential in extracting pivotal biomedical insights.

Table 1. OGT-QA Dataset Refinement during the Iterative Model Generation Process[a]

iterations	count of positive predictions	count of accurate OGT descriptions	precision	size of OGT-QA datasets
1st	200	40	0.20	520
2nd	126	36	0.29	646
3rd	752	392	0.52	1398

[a]Note: The second column indicates the count of plausible OGT predictions, while the third column represents the number of predictions that underwent manual verification and were confirmed as accurate OGT descriptions.

OGT Database Construction Using the Fine-Tuned BioLinkBERT Model. We next employed the fine-tuned BioLinkBERT model to extract OGT information from the unlabeled contexts, resulting in 5921 positive predictions. We then eliminated apparent duplicate data entries and converted these OGT descriptions to FLOAT values. In cases where an OGT was provided as a range (e.g., "20−30 °C"), we converted it to a single value by using the midpoint (25 °C). Additionally, we rectified anomalies arising from text conversion errors (e.g., validating '30,035 °C' as "30−35 °C"). Following these steps, a dataset containing 4673 entries was subjected to further analyses.

To assess the accuracy of the extracted OGT information, we compared our dataset with a microbial culture temperature database, which previously served as an OGT database[4] (referred to here as the OGT_cultivation database). This database encompasses 21,498 temperature records for 18,854 different organisms. We calculated the difference in the OGT for the same organism between the two databases. While the OGT for the same organism in the two databases could differ by up to 78 °C, we observed that 43% of the entries in our dataset had an OGT deviation of less than 3 °C compared to the OGT_cultivation database.

For entries with an OGT deviation equal to or greater than 3 °C (57%), we conducted manual validation, revealing an accuracy of 80%. Specifically, we found only 25% (21/84) accuracy for entries with OGT derivations between 20 and 78 °C (20 °C < OGT derivations ≤78 °C), 69% (292/424) for derivations between 10 and 20 °C (10 °C ≤ OGT derivations ≤ 20 °C), 74% (279/377) for derivations between 7 and 10 °C (7 °C ≤ OGT derivations < 10 °C), and an increase to 86% for deviations between 3 and 7 °C (1531/1774) (3 °C ≤ OGT derivations < 7 °C). These results suggest that predicted OGT values are more likely to be correct if they closely align with the culture temperature of the organism. For the remaining OGT entries (0 °C ≤ OGT derivations < 3 °C), randomly sampled entries exhibited an accuracy of 97%. Consequently, the overall estimated accuracy of our extracted OGT was 87% (Table 2). It is essential to note that incorrect OGT values were manually corrected during this validation process, resulting in a high accuracy of 98.7% for the remaining dataset of 4238 entries.

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

Table 2. Manual Verification of the Extracted OGT Information

data statistics	verified (deviation ≥ 3)	remaining (deviation < 3)	total
quantity	2659	2014	4673
proportion	0.569	0.431	1.000
accuracy rate	0.798	0.972	0.873
comments	fully verified	sampled in #2632−2672 (Deviation: 2.7−2.96) and #3289−3319 (Deviation equals 2)	weighted by quantity

We merged these validated records with the positive samples in the OGT_QA dataset to obtain a total of 5014 OGT descriptions related to 1155 microbial organisms. Within this dataset, we observed instances where the same OGT value for a specified organism originated from multiple context descriptions. To mitigate redundancy, entries with identical organism names and OGT values were merged. Following this consolidation, the remaining 2142 OGT descriptions constituted the final OGT database that we referred to as "OGT_journal". A significant portion (71%) of the OGT values in this dataset falls within the range of 20−40 °C (20 °C ≤ OGT values < 40 °C). There is a notable decrease in the number of OGT values at temperatures equal to or above 40 °C (Figure 3A). Most organisms in the OGT_journal database could be matched with entries in the OGT_cultivation database. A total of 1002 distinct organisms were found in both databases, associated with 1972 OGT records in the OGT_journal database and 1407 cultivation temperature records in the OGT_cultivation database, respectively (Figure 3B).

For organisms common to both databases, we compared their temperature data (if there were multiple temperature values available for the same organism, then the average value was analyzed). We found noteworthy differences between these databases. The temperature region containing the highest number of OGT entries was higher in the OGT_journal database (30−40 °C in the OGT_journal database vs 20−30 °C in the OGT_cultivation database) (Figure 3C).

Moreover, we observed that only 15.7% of organisms had identical OGT and cultivation temperatures, while 29.3% of the organisms showed temperature differences of greater than or equal to 5 °C (Figure 3D). In particular, for several organisms we found substantial discrepancies between the literature-described OGTs and their cultivation temperatures (Figure 3E), including *Pseudomonas stutzeri* (70 °C vs 30 °C), *Bacillus tequilensis* (60 °C vs 30 °C), *Agaricus bisporus* (50−55 °C vs 24.5 °C), *Brevibacillus agri* (60 °C vs 33 °C), and *Brevibacillus borstelensis* (60 °C vs 34.5 °C) (Supplementary Table 5). These results indicate that the OGT_journal database could offer a more accurate representation of an organism's OGT compared to the OGT_cultivation database.

Enhanced Performance of Optimal Growth Temperature over Culture Temperature in Topt Prediction. Enzymes must effectively perform catalytic functions at the specific growth temperatures of the organisms to support their survival. Consequently, enzymes often undergo evolutionary adaptations to align their optimal working temperatures with the growth temperatures of their host organisms. Through a comparative analysis of the OGT_journal database and the OGT_cultivation database with a Topt dataset obtained from BRENDA (Figure 4A, containing Topt values for 4728 enzymes from 1623 organisms, see Methods for details), we identified correlations between enzymes' Topt and the corresponding growth temperatures of their sourced organisms. To make the comparison, we concentrated on 1878 enzymes, for which information about the organisms is available in both the OGT_journal and the OGT_cultivation databases. Notably, approximately 80% of the entries display OGT values exceeding their corresponding culture temperatures (Figure 4B). Moreover, we observed significantly different distribution patterns of these temperature values (Figure 4C).

We found that the correlation coefficient for the OGT_journal database and Topt (0.76) surpassed that of the OGT_cultivation database (0.74) (Figure 4D). Additionally, we conducted a Fisher's *Z*-Transformation and an associated two-tailed *z*-test to compare the correlation coefficient between the OGT_journal and Topt against that between the OGT_cultivation and Topt. With a sample size of $n = 1878$, the *z*-test yielded a test statistic z of 5.400 and a *p*-value of less than 0.001 (Table 3). This emphasizes a significantly stronger association between the optimal growth temperatures of organisms and the Topt of enzymes, in contrast to the relationship between organisms' culture temperatures and Topt.

The above observation prompted us to investigate whether the accuracy of Topt predictions could be further enhanced by utilizing authentic optimal growth temperatures instead of culture temperatures of the source organisms. In a previous work, a machine learning approach relying on the OGT_cultivation database and basic sequence information on individual enzymes achieved over 50% determination coefficient (R^2 score) in Topt prediction.[4] Subsequently, the inclusion of an additional 5494 sequence features (Supplementary Table 6) from the iFeature package resulted in an improved R^2 score of 61%.[5] To initiate our investigation, we first compiled a dataset of 1024 enzymes with known Topt values from 136 organisms. Each entry in this dataset included manually curated OGT information (sourced from the OGT-QA positive samples) and culture temperature information from the OGT_cultivation database. This dataset was employed to train and select the best-performing model from five classical machine learning models (Bayesian Ridge, Elastic Net, Decision Tree, Random Forest, and Support Vector Machine), with either OGT values or culture temperature values in conjunction with the 5494 sequence-derived features from iFeature as input. In this process, we employed the grid search method to determine the model hyperparameters (Supplementary Table 7). Furthermore, the datasets utilized in the machine learning experiments were partitioned into a training set and a test set, maintaining a ratio of 9:1, and a random seed value of 212 was employed.

Our analysis revealed that the Random Forest model outperformed all other models, achieving the highest R^2 score when the OGT values were used as input (Supplementary Table 8). Subsequently, we conducted further training and testing of the Random Forest model on a larger dataset (1878 entries, Figure 4B) using a similar 9:1 data partitioning strategy. The endeavor resulted in an enhanced R^2 and smaller MAE, MSE, and RMSE when OGT values were employed as input, in contrast to those obtained using culture temperature as input (Figure 4E, Supplementary Table 9). However, despite the improved performance of the model trained with our OGT data, the analysis of the Topt values predicted from

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

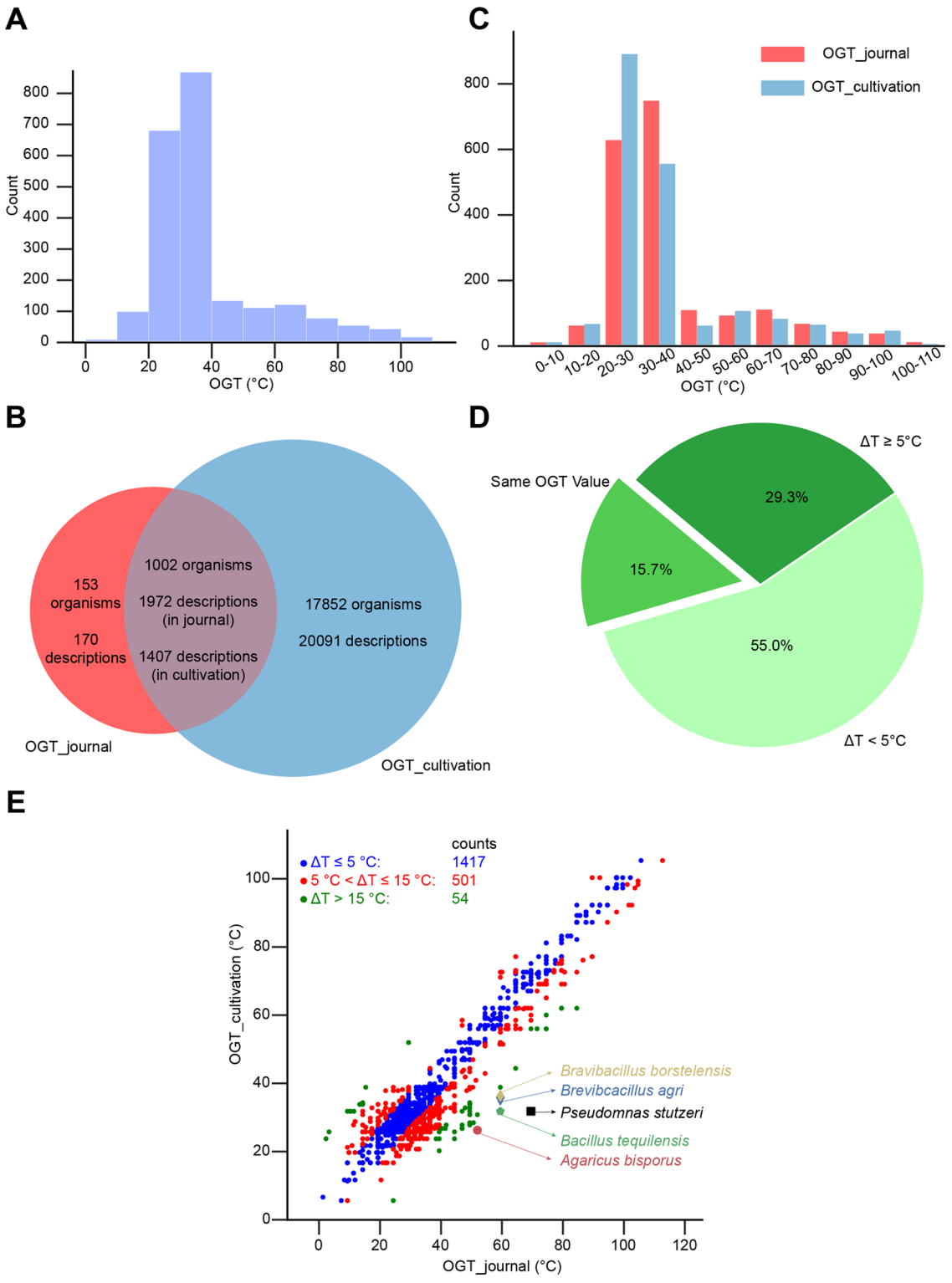


Figure 3. Comparative analysis of the OGT_journal and OGT_cultivation databases. (A) Temperature distributions of the OGT_journal database. (B) Intersection analysis between the OGT_journal database and the OGT_cultivation database. (C) Temperature distributions of the intersection data points from the two databases. (D) Statistical analysis of the temperature deviations of the intersection data points shared in the OGT_journal database and the OGT_cultivation database. (E) Scatter plot analysis of the intersection data points shared in the OGT_journal database and the OGT_cultivation database. Data points with temperature deviation greater than 5 and 15 °C were labeled in red and green, respectively. Notably, five data points with the largest temperature deviation were additionally marked, as indicated in the figure.

the two databases revealed similar distribution patterns (Figure 4F). This result is not entirely surprising, considering the interplay of various factors contributing to the formation of OGT and Topt. First, the relationship between temperature and enzyme activity may be nonlinear, and the model might effectively capture this nonlinearity, leading to improved performance without significantly altering the overall distribution patterns. Second, common features shared by the

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

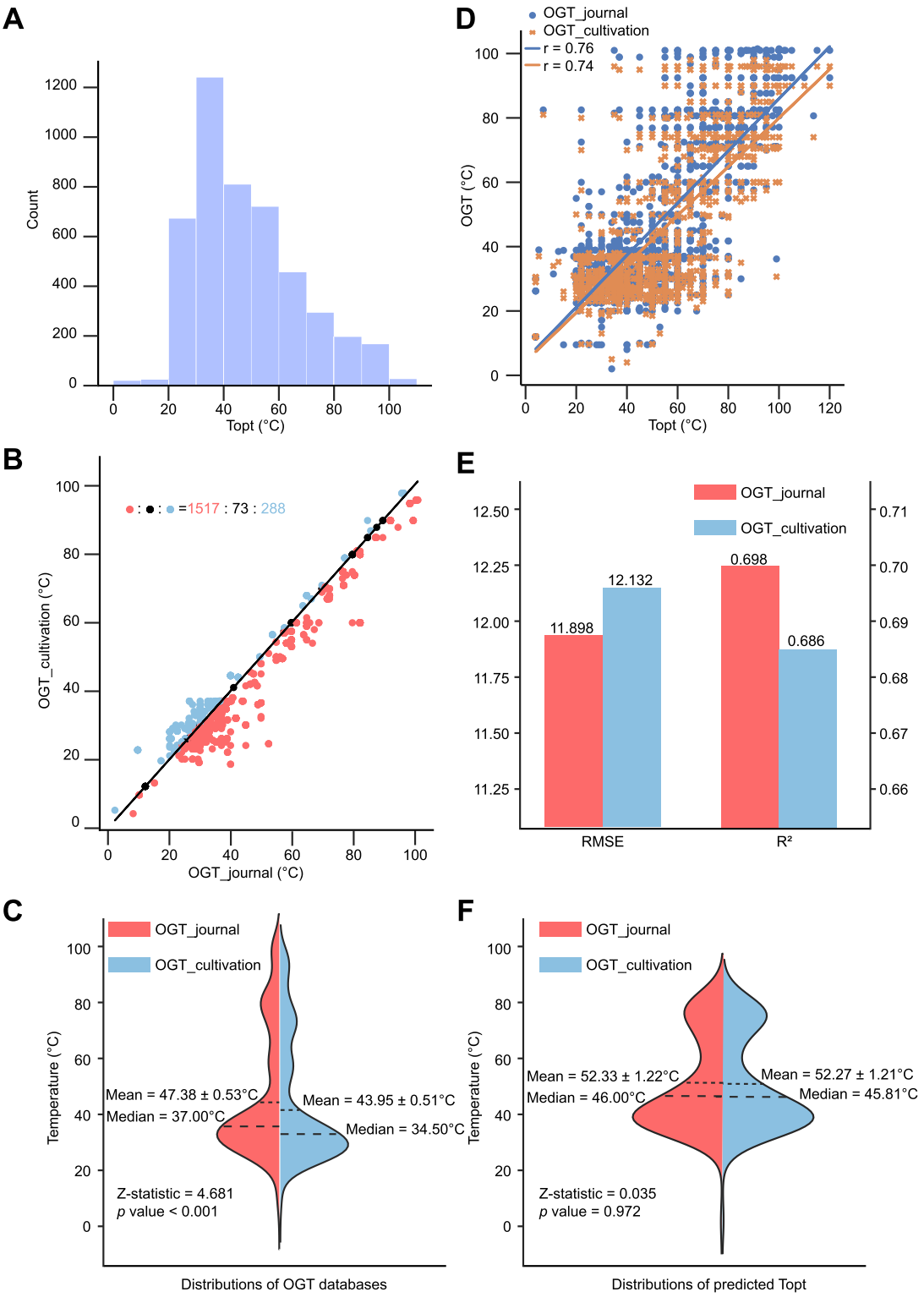


Figure 4. Comparative analysis of the OGT database and the culture temperature database in Topt prediction. (A) Temperature distribution of the Topt dataset (4728 entries). (B) Comparison of the Culture Temperature and the OGT for the same organism in Topt prediction. Red points represent cases where the OGT is higher than the culture temperature, the blue ones indicate lower OGT values and the black dots signify equivalent values. The exact numbers of each category of dots were labeled. (C) Distribution analysis of the OGT_journal and the OGT_cultivation databases. (D) Correlation analysis between the Topt dataset and the OGT_journal database or the OGT_cultivation database with a common sample size of $n = 1878$. r denotes Pearson's correlation coefficient. (E) Topt prediction performance of the Random Forest model using the OGT values or the culture temperature values as input. The bar plot illustrates a higher prediction performance (RMSE: 11.898 vs 12.132, R^2: 0.698 vs 0.686) for Topt when utilizing OGT as the input. (F) Distribution analysis of Topt data predicted by using the OGT_journal database and the OGT_cultivation database as input. For C and F, mean values and median values of each distribution, alongside the corresponding Z-statistics and p-values that statistically indicate distribution differences, were provided.

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292

Table 3. Correlation Coefficients Comparison[a]

n (size of samples)	r_{12}	r_{13}	r_{23}
1878	0.76	0.74	0.97
test statistic z		5.400	
probability p		<0.001	

[a]Note: r_{12} represents the correlation coefficient between the OGT_journal database and Topt; r_{13} represents the correlation coefficient between the OGT_cultivation database and Topt; r_{23} represents the correlation coefficient between the OGT_journal database and OGT_cultivation database.

OGT_journal and the OGT_cultivation datasets might align the distribution patterns, even if the specific values differ. Collectively, these results underscore the effectiveness of our OGT_journal database in predicting the optimal enzyme reaction temperatures.

CONCLUSIONS

In this work, we presented a novel approach to identify and extract OGT information from the scientific literature, employing a Machine Reading Comprehension framework. The proposed method involved a two-stage process, encompassing Context Locating and Extractive/Generative Question Answering. Specifically, six pretrained models were utilized, including three BERT-based models (BioBERT-base, BioBERT-large, and BioLinkBERT) and three GPT-based models (BioGPT, BioGPT-large, and ChatGPT). The top-performing model, BioLinkBERT, demonstrated an impressive performance with an EM score of 91.00 and an F1 score of 91.91 for the OGT extraction. These results underscore the effectiveness of our approach in automatically extracting crucial information from scientific papers.

Utilizing the BioLinkBERT model for OGT information extraction from relevant scientific papers enabled the construction of a robust OGT database, comprising 2142 OGT descriptions associated with 1155 organisms. Remarkably, this database outperformed the culture temperature database in predicting enzyme Topt, underscoring a stronger correlation between the level of OGT and Topt compared to the relationship between cultivation temperature and Topt.

In summary, our work not only delivered a high-quality OGT database, with implications for enzyme discovery and precise Topt prediction, but also established an effective machine learning-based pipeline for information extraction from scientific literature. This pipeline is not limited to the OGT extraction presented in our work but could be broadly applied for data collection and prediction across various biological processes, such as microbial fermentation, cell cultivation, metabolic pathways, as well as other domains like materials science and chemistry. Our work underscores the influential role of AI in advancing scientific research and inspires heightened machine learning-based exploration and innovation in literature analysis, data interpretation, and automated scientific discovery.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.3c06526.

Additional figures and tables related to NLP techniques used for OGT extraction; examples of the database construction process and other details regarding deep learning and machine learning (PDF)

AUTHOR INFORMATION

Corresponding Authors

Wei He − *State Key Laboratory of Bioreactor Engineering, Shanghai Collaborative Innovation Center for Biomanufacturing (SCICB), East China University of Science and Technology, Shanghai 200237, China*; Email: whe@mail.ecust.edu.cn

Shu Quan − *State Key Laboratory of Bioreactor Engineering, Shanghai Collaborative Innovation Center for Biomanufacturing (SCICB), East China University of Science and Technology, Shanghai 200237, China; Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai 201203, China*; orcid.org/0000-0002-6672-4947; Email: shuquan@ecust.edu.cn

Authors

Xiaotao Wang − *State Key Laboratory of Bioreactor Engineering, Shanghai Collaborative Innovation Center for Biomanufacturing (SCICB) and School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China*

Yuwei Zong − *School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China*

Xuanjie Zhou − *School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China*

Li Xu − *School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China*; orcid.org/0000-0001-9234-094X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.3c06526

Author Contributions

‖Equal contribution.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Elsevier for providing us with remote access to their API and text resources. We also express our gratitude to Chi Zhang and Yingchao Zhang for critical suggestions and their assistance on manuscript preparation, Mencius Meng and Xin Huang for helpful discussions. This work was supported by the Undergraduate Training Program on Innovation and Entrepreneurship grant 202210251039 (to X.W., Y.Z., X.Z., and L.X.) and the National Natural Science Foundation of China (NSFC) grants 32222049 (to S.Q.), 32201043 (to W.H.).

REFERENCES

(1) Zuberer, D. A.; Zibilske, L. M. Composting: The Microbiological Processing of Organic Wastes. In *Principles and Applications of Soil Microbiology*, 3rd ed.; Elsevier, 2021; pp 655−679.

(2) Vivek, K.; Sandhia, G.; Subramaniyan, S. Extremophilic lipases for industrial applications: A general review. *Biotechnol. Adv.* **2022**, *60*, No. 108002.

(3) Mesbah, N. M. Industrial biotechnology based on enzymes from extreme environments. *Front. Bioeng. Biotechnol.* **2022**, *10*, No. 870083.

(4) Li, G.; Rabe, K. S.; Nielsen, J.; Engqvist, M. K. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth. Biol.* **2019**, *8* (6), 1411−1420.
(5) Li, G.; Zrimec, J.; Ji, B.; Geng, J.; Larsbrink, J.; Zelezniak, A.; Nielsen, J.; Engqvist, M. K. Performance of regression models as a function of experiment noise. *Bioinform. Biol. Insights* **2021**, *15*, 11779322211020315.
(6) Gado, J. E.; Beckham, G. T.; Payne, C. M. Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *J. Chem. Inf. Model.* **2020**, *60* (8), 4098−4107.
(7) Engqvist, M. K. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. *BMC Microbiol.* **2018**, *18*, 177.
(8) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. Correction to automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model.* **2021**, *61* (8), 4124−4124.
(9) Jiang, X.; He, K.; Yang, B. Automatic information extraction in the third-generation semiconductor materials domain based on DKNet and MANet. *IEEE Access* **2022**, *10*, 29367−29376.
(10) Tsatsaronis, G.; Schroeder, M.; Paliouras, G.; Almirantis, Y.; Androutsopoulos, I.; Gaussier, E.; Gallinari, P.; Artieres, T.; Alvers, M. R.; Zschunke, M. Bioasq: A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering. In *2012 AAAI Fall Symposium Series*; 2012; pp 92−98.
(11) Raj Kanakarajan, K.; Kundumani, B.; Sankarasubbu, M. BioELECTRA: Pretrained Biomedical Text Encoder Using Discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, Association for Computational Linguistics, 2021, pp 143−154.
(12) Yao, Y.; Huang, S.; Dong, L.; Wei, F.; Chen, H.; Zhang, N. Kformer: Knowledge Injection in Transformer Feed-Forward Layers. In *CCF International Conference on Natural Language Processing and Chinese Computing*; Springer, 2022; pp 131−143.
(13) Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; Lu, X. *Pubmedqa: A Dataset for Biomedical Research Question Answering*. 2019, arXiv preprint arXiv:1909.06146 (acccessed Sep 28, 2023).
(14) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36* (4), 1234−1240.
(15) Yasunaga, M.; Leskovec, J.; Liang, P. *Linkbert: Pretraining Language Models with Document Links*. 2022, arXiv preprint arXiv:2203.15827 (acccessed Sep 28, 2023).
(16) Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.-Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **2022**, *23* (6), bbac409.
(17) OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. 2022, https://www.openai.com/research/chatgpt (accessed Sep 28, 2023).
(18) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018, arXiv preprint arXiv:1810.04805 (acccessed Sep 28, 2023).
(19) Howard, J.; Ruder, S. *Universal Language Model Fine-tuning for Text Classification*. 2018, arXiv preprint arXiv:1801.06146 (acccessed Sep 28, 2023).
(20) Li, X. L.; Liang, P. *Prefix-tuning: Optimizing Continuous Prompts for Generation*, 2021, arXiv preprint arXiv:2101.00190 (acccessed Sep 28, 2023).
(21) Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; Tang, J. *P-tuning v2: Prompt Tuning can be Comparable to Fine-tuning Universally Across Scales and Tasks*. 2021, arXiv preprint arXiv:2110.07602 (acccessed Sep 28, 2023).
(22) Rajpurkar, P.; Jia, R.; Liang, P. *Know What You Don't Know: Unanswerable Questions for SQuAD*. 2018, arXiv preprint arXiv:1806.03822 (acccessed Sep 28, 2023).
(23) Diedenhofen, B.; Musch, J. cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one* **2015**, *10* (4), No. e0121945.
(24) Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; Li, J. *A Unified MRC Framework for Named Entity Recognition*. 2019, arXiv preprint arXiv:1910.11476 (acccessed Sep 28, 2023).
(25) Lahouar, A.; Marin, S.; Crespo-Sempere, A.; Saïd, S.; Sanchis, V. Influence of temperature, water activity and incubation time on fungal growth and production of ochratoxin A and zearalenone by toxigenic Aspergillus tubingensis and Fusarium incarnatum isolates in sorghum seeds. *Int. J. Food Microbiol.* **2017**, *242*, 53−60.

https://doi.org/10.1021/acs.jpcb.3c06526
J. Phys. Chem. B 2024, 128, 2281−2292