Comparative Evaluation of BERT-Based Models for Relation Extraction in a Seaweed Knowledge Graph Visualization Platform

Mursalin[1], Anne Parlina[2], Ahmad Tabrani[3]

ABSTRACT

Seaweed literature contains information related to species, bioactive compounds, biological activities, and geographic distributions. Most of this information is available in unstructured textual form, limiting direct retrieval and analysis. This paper presents a comparative evaluation of BERT-based models for relation extraction in seaweed literature and their integration into a web-based knowledge graph visualization system. A relation extraction dataset was constructed from annotated seaweed literature using four relation categories: Seaweed–Location, Seaweed–Benefit, Seaweed–Compound, and Compound–Benefit. BioBERT, SciBERT, and PubMedBERT were evaluated under the same experimental setting. Model performance was measured using Accuracy, Precision, Recall, and F1-Score. PubMedBERT obtained an Accuracy of 86.92%, Precision of 87.64%, Recall of 86.92%, and F1-Score of 87.02%, while SciBERT and BioBERT obtained F1-Scores of 85.69% and 85.17%, respectively. Predicted relations were converted into subject–relation–object triples and used for knowledge graph construction. The resulting knowledge graph was integrated into a web-based visualization system comprising literature search, entity retrieval, relation retrieval, and graph visualization components. The proposed workflow combines relation extraction and knowledge graph construction to transform unstructured seaweed literature into a graph-based knowledge representation.

Keywords: Relation Extraction; Knowledge Graph; Seaweed Literature; BioBERT; SciBERT; PubMedBERT; Information Visualization.

1. INTRODUCE

Seaweed is utilized in food, pharmaceutical, cosmetic, agricultural, and biotechnology applications. Seaweed species contain bioactive compounds such as polysaccharides, carotenoids, phenolics, and antioxidants associated with anti-inflammatory, antioxidant, antimicrobial, and anticancer activities. The volume of scientific literature on seaweed has expanded, presenting challenges in extracting and mapping relationships among seaweed species, bioactive compounds, biological activities, and geographical locations. While macro-scientometric approaches using tools like CiteSpace are applied to map global seaweed research trends, they analyze co-citation metadata rather than sentence-level entity connections [1]. The increasing number of scientific publications requires computational methods for knowledge discovery and information retrieval.

Manual literature review approaches require time and labor to identify information and establish connections between entities across multiple research articles. Automated knowledge extraction techniques provide a computational mechanism for literature exploration and mapping domain-specific bioactives [2]. Recent developments in Natural Language Processing (NLP), particularly transformer-based language models, have increased F1-scores and accuracy metrics in information extraction tasks compared to traditional rule-based methods. This trend is also evident in the Architecture, Engineering, and Construction (AEC) sector, where transformer-based architectures are increasingly being adopted to overcome the limitations of domain-specific rule-based systems [3]. Relation Extraction (RE) aims to identify semantic relationships between entities within textual data. In biomedical and scientific domains, Pre-trained Language Models (PLMs) such as BioBERT, SciBERT, and PubMedBERT have been applied to extract domain-specific relationships from scientific literature. Benchmarks indicate that variations in pre-training corpora—ranging from multidisciplinary scientific texts to purely biomedical datasets—affect model accuracy in processing specialized classifications [4] and extracting interactions such as protein-protein relations [5].

Relation extraction techniques transform unstructured text into structured representations that serve as the foundation for Knowledge Graph (KG) construction. In a knowledge graph, entities are represented as nodes and relationships as edges, enabling scientific literature to be organized as an interconnected knowledge network. These networks support knowledge discovery, facilitate literature exploration, and map relationships among scientific concepts within pharmaceutical and natural product informatics [6]. While relation extraction and knowledge graph construction are applied in biomedical domains, literature concerning seaweed-specific extraction is less documented. Existing studies concentrate on broader biomedical contexts; a systematic comparison of pre-trained language models for seaweed relation extraction has not been documented. Furthermore as discussed by [7], standard graph databases often require local desktop installations or proficiency in specific query languages (e.g., Cypher, SPARQL) and web-based systems designed

for visualizing knowledge graphs derived from seaweed literature are limited.

Therstudy presents a comparative evaluation of BERT-based models for relation extraction within a web-based seaweed knowledge graph visualization system. Three transformer-based models—BioBERT, SciBERT, and PubMedBERT—are evaluated using precision, recall, and F1-score metrics on a manually annotated seaweed literature dataset. The model with the highest F1-score is subsequently integrated into an automated framework and a web interface [8] to construct and visualize knowledge graphs. This system transforms unstructured data extracted from scientific publications into an interconnected network for mapping relationships among seaweed species, bioactive compounds, health benefits, and geographical locations.

1. Previous studies have evaluated BioBERT, SciBERT, and PubMedBERT for relation extraction tasks in biomedical and scientific domains using domain-specific datasets. However, the application and comparative evaluation of these BERT-based models for relation extraction in seaweed scientific literature have not been specifically investigated. This study evaluates BioBERT, SciBERT, and PubMedBERT using a manually annotated seaweed literature dataset under the same preprocessing procedure, training configuration, and evaluation metrics.
2. A domain-specific relation extraction dataset is developed from seaweed scientific literature to represent semantic relationships among relevant entities. The dataset contains sentence-level annotations describing relations among seaweed species, compounds, benefits, and locations, which are used for training and evaluating transformer-based relation extraction models.
3. The extracted entities and relations are converted into a knowledge graph representation to structure semantic information obtained from seaweed literature. The proposed representation models entities as nodes and relations as edges, forming a graph-based structure that captures relationships identified through the relation extraction process.
4. A web-based knowledge graph visualization system is developed to integrate relation extraction results and graph representation within an interactive application. The system supports literature input, extraction result visualization, statistical summaries, relation filtering, and data export for accessing relationships extracted from seaweed scientific literature.

2. RELATED WORK

The increase in seaweed research publication volume requires computational tools for literature mining and knowledge extraction. While macro-scientometric approaches and bibliometric tools map global research trends using co-citation metadata [1], these methods process document-level metadata rather than sentence-level interactions between biological entities. Extracting specific connections among seaweed species, geographical locations, and bioactive compounds requires semantic analysis algorithms capable of processing unstructured text. Relation Extraction (RE) is a Natural Language Processing (NLP) task designed to identify and classify semantic relationships from unstructured text [9]. The application of Transformer-based Pre-trained Language Models (PLMs), including BioBERT, SciBERT, and PubMedBERT, has altered the methodological framework of RE [4].

The performance of these models, measured by standard classification metrics such as F1-score, is determined by their pre-training corpora. Comparative evaluations of these models are required to identify which architecture yields the highest precision and recall for specific textual domains. Previous research indicates that variations in training corpora influence a model's capacity to process specialized domain terminologies [10]. PLMs are primarily trained on broad biomedical datasets and are less frequently evaluated on marine and agricultural literature, which presents distinct multi-entity classification challenges. The development of corpora such as BioRED [11], indicates a requirement for extraction systems capable of processing interactions across text sequences. To organize the relations extracted by PLMs, Knowledge Graphs (KGs) are utilized to represent data structures as interconnected nodes and edges. Standard graph databases utilized for KG storage often require local installations or specific query languages (e.g., SPARQL, Cypher), which presents operational constraints for users without query programming backgrounds [6].

Additionally, literature indicates that RE model evaluation and KG construction are often conducted as separate tasks within standardized benchmark environments [10], rather than being integrated into specific web applications. Web-based systems specifically designed for seaweed literature informatics are limited. To address these methodological constraints, this research conducts a comparative study of BERT-based models to quantify their relation extraction performance on seaweed texts. The model producing the highest evaluation metrics is subsequently integrated into a web-based visualization system to map the extracted multi-entity relations.

3. METHODOLOGY

3.1 Research Framework

This section outlines the methodological framework for the comparative evaluation of BERT-based models and the development of the web-based seaweed knowledge graph visualization system. The methodology comprises seven phases: data acquisition, relation annotation, model development, model evaluation, knowledge graph construction, system integration, and data visualization. Initially, a corpus of scientific publications is compiled and manually annotated. This approach aligns with literature stating that organizing scientific information into structured knowledge bases requires the extraction of scientific entities and their relations [12]. The resulting annotated dataset serves as the baseline for training and

evaluating the transformer-based relation extraction models. Subsequently, the model yielding the highest F1-score is deployed to generate structured relational data from the literature corpus. In the final phase, the extracted relations are synthesized into a knowledge graph and integrated into a web-based visualization interface for data exploration and semantic analysis.

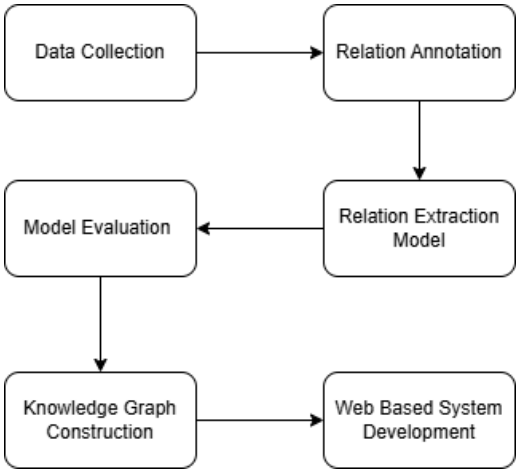

Figure 1. Research methodology framework

3.2 Data Collection

This study uses a seaweed literature dataset provided by the Research Center for Data and Information Sciences, National Research and Innovation Agency (BRIN). The dataset comprises scientific abstracts retrieved from the Scopus database in July 2024 using the query (seaweed OR macroalgae) AND ("functional food*" OR "nutraceutical*"). The collected abstracts serve as the primary corpus for developing the Named Entity Recognition (NER) and Relation Extraction (RE) datasets. The resulting entity and relation annotations are subsequently used for knowledge graph construction

Table 1. Data Characteristics

Characteristic	Description
Data Source	National Research and Innovation Agency (BRIN), Research Center for Data and Information Sciences
Database	Scopus
Domain	Seaweed Literature
Number of Documents	1,122 Documents
Data Format	CSV
Language	English
Data Content	Article Metadata and Abstracts
Purpose	Named Entity Recognition (NER) and Relation Extraction (RE)

3.3 Relation Annotation

The relation annotation process was conducted manually on the collected seaweed literature to construct a supervised dataset for relation extraction, an approach consistent with established text-mining methodologies where relations are annotated from scratch by hand to ensure high data quality [4]. Four entity categories were defined based on the characteristics of seaweed-related scientific information: Seaweed, Location, Benefit, and Compound. Similar to existing data-driven approaches that determine a specific set of key entities and relations most representative of their domain [4], these defined entities capture the essential seaweed species, geographical information, biological activities, and bioactive compounds discussed in the literature.

Table 2. Annotated entity categories

Entity	Description
Seaweed	Name of a seaweed species or type
Location	Research location or geographical origin of seaweed
Benefit	Biological activity or benefit associated with seaweed
Compound	Bioactive compound contained in seaweed

After entity identification, semantic relationships between entity pairs were annotated according to predefined relation types. The annotation focused on four primary relations that were subsequently used for relation extraction model training.

Table 3. Relation types used in annotation

Relation	Description
Seaweed–Location	Indicates the origin, habitat, or research location of a seaweed species
Seaweed–Benefit	Indicates biological activities or benefits associated with a seaweed species
Seaweed–Compound	Indicates compounds contained in a seaweed species
Compound–Benefit	Indicates biological activities or benefits produced by a compound

The complete annotation process produced 10,284 relation instances. The distribution of all annotated relations is presented in **Table 4.**

Table 4. Distribution of annotated relations

Relation	Count
Seaweed–Compound	4,100
Compound–Benefit	3,213
Seaweed–Benefit	2,068
Seaweed–Location	786
Compound–Compound	40
Benefit–Compound	38
Seaweed–Seaweed	18
Benefit–Benefit	12
Compound–Seaweed	8
Benefit–Location	1
Total Relations	10,284

Only the four predefined primary relations were retained for relation extraction model training. The remaining relation types appeared infrequently and were excluded to reduce class imbalance and maintain a consistent relation schema. The final training dataset contained 10,167 relation instances, as summarized in Table 5.

Table 5. Relation types used for RE model training

Relation	Count
Seaweed–Location	786
Seaweed–Benefit	2,068
Seaweed–Compound	4,100
Compound–Benefit	3,213
Total Relations Used	10,167

Although the annotation process produced various relation categories, this study only utilizes relations that are aligned with the predefined information extraction objectives. Therefore, the construction of the RE dataset is focused on four main relation categories, namely the relationships between Seaweed and Location, Seaweed and Benefit, Seaweed and Compound, and Compound and Benefit. These four relation categories were selected because they represent the most relevant information in seaweed literature studies.

3.4 Relation Extraction Model

This study conducts a comparative evaluation of three domain-specific transformer-based architectures: SciBERT, BioBERT, and PubMedBERT, for the Relation Extraction (RE) task. The objective is to compute semantic relationships between entity pairs using contextual representations generated from tokenized inputs, specifically input_ids and attention_mask. This transformation of raw text into token indices follows the standard transformer encoding pipeline, a robust approach recently utilized in clinical text-encoding architecture [13]. To establish a controlled experimental environment for baseline comparison, all candidate models are trained using a supervised learning approach under identical hyperparameter configurations: a batch size of 16, a learning rate of 2×10^{-5}, a weight decay of 0.15, and a maximum limit of 5 training epochs.The dataset is partitioned into 17,195 samples for training and 1,911 samples for validation per epoch.

During training, each batch undergoes forward propagation to output relation class predictions, followed by loss computation against ground-truth labels. Model parameters are updated through backpropagation. To prevent overfitting, an early stopping strategy with a patience parameter of 1 is implemented. Validation loss is calculated at the end of each epoch. If the validation loss does not decrease, training is terminated, and the model weights corresponding to the minimum validation loss are restored to ensure structural evaluation based on empirically observed loss metrics.

3.5 Model Evaluation

The evaluation of the Relation Extraction (RE) models was conducted using standard classification metrics, including Accuracy, Precision, Recall, and F1-score. These metrics were used to measure the model's ability to correctly classify relation categories in a multi-class setting. Accuracy refers to the proportion of correctly predicted samples out of the total dataset. Precision measures the proportion of correctly predicted instances for each class [14], while Recall measures the proportion of relevant instances that are successfully identified by the model. The F1-score is defined as the harmonic mean of Precision and Recall, providing a single measure that balances both metrics [15].

The metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Since the task involves multiple relation classes, both macro-averaged and weighted-averaged scores are reported to provide a more balanced evaluation across classes with different distributions. Macro-averaging treats all classes equally, while weighted-averaging accounts for class frequency in the dataset. All models (SciBERT, BioBERT, and PubMedBERT) are evaluated on the same held-out test set to ensure a fair comparison. The F1-score is used as the main metric for comparing model performance across relation categories.

3.6 Knowledge Graph Construction

The knowledge graph is constructed from the entities and relations produced by the Relation Extraction (RE) models. Within this graph, entities serve as nodes, while their relations are represented as directed edges connecting them [16]. The graph uses four entity types: Seaweed,

Location, Compound, and Benefit. It includes four relation types: Seaweed–Location, Seaweed–Benefit, Seaweed–Compound, and Compound–Benefit. The construction process converts each predicted relation into a triplet format (head entity, relation, tail entity) [17]. These triplets are directly used to form the graph structure. To maintain data quality, duplicate relations are removed and entity names with the same meaning are standardized. The final output is a knowledge graph that connects seaweed-related concepts in a structured form.

3.7 Web-Based System Development

The web-based system is developed to visualize the results of relation extraction and knowledge graph construction[18]. The system integrates processed data into a graph-based interface. The system consists of data processing, backend, and visualization components[19]. Following relation extraction, the backend structures the processed data for graph representation. The visualization component maps the extracted information units to nodes and their relationships to edges [20]. The interface supports interactive graph exploration, enabling users to examine data frequency and connections. Specifically, hovering over an edge reveals the linked nodes and their co-occurrence frequency. The system is deployed as a web application to provide browser-based access.

4. RESULT AND DISCUSSION

4.1 Comparative Evaluation of Relation Extraction Models

The performance of three pretrained language models, namely SciBERT, BioBERT, and PubMedBERT, was evaluated for the relation extraction task under identical experimental settings, including dataset, preprocessing procedure, data split, and training configuration. The evaluation focused on comparing their ability to classify entity relationships into five predefined categories:

No Relation, Seaweed–Location, Seaweed–Benefit, Seaweed–Compound, and Compound–Benefit.

Model performance was assessed using four classification metrics: Accuracy, Precision, Recall, and F1-score. The training behavior of each model was analyzed based on changes in training loss, validation loss, and validation accuracy across epochs. **Table 6.** presents the training and validation performance of the evaluated models, while the corresponding loss and accuracy curves describe the optimization process. A confusion matrix analysis was subsequently conducted for the selected model to analyze the distribution of correct and incorrect relation predictions. All models were evaluated using the same held-out test set containing 1,911 samples. SciBERT, BioBERT, and PubMedBERT employ BERT-based sequence classification architectures with different domain-specific pretraining corpora. SciBERT was pretrained on scientific publications, BioBERT on biomedical literature, and PubMedBERT on PubMed abstracts and full-text articles. Differences in pretraining corpora result in different learned language representations, which may affect relation classification performance on seaweed scientific literature.

Table 6. Training and Validation Performance of SciBERT, BioBERT, and PubMedBERT

Model	Epoch	Training Loss	Validation Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SciBERT	1	0.8703	0.4912	83.36	83.64	83.36	83.43
	2	**0.4121**	**0.4134**	**85.56**	**86.21**	**85.56**	**85.69**
	3	0.3023	0.4486	85.87	87.29	85.87	86.08
BioBERT	1	0.9604	0.5524	78.60	79.53	78.60	78.02
	2	0.4353	0.4532	83.73	84.86	83.73	83.99
	3	**0.3250**	**0.4289**	**84.88**	**86.30**	**84.88**	**85.17**
	4	0.2510	0.4414	85.56	86.29	85.56	85.69
PubmedBERT	1	0.8494	0.4909	82.37	82.38	82.37	81.86
	2	0.3837	0.4349	85.35	86.75	85.35	85.56
	3	**0.2948**	**0.4034**	**86.92**	**87.64**	**86.92**	**87.02**
	4	0.2314	0.4281	88.02	88.46	88.02	88.07

Table 6. Summarizes the training and validation performance of SciBERT, BioBERT, and PubMedBERT across all training epochs. For each model, training progress was monitored using training loss, validation loss, Accuracy, Precision, Recall, and F1-score. The highlighted rows indicate the selected checkpoints, determined by the lowest validation loss for each model. These checkpoints we were used for the subsequent comparative evaluation and analysis.

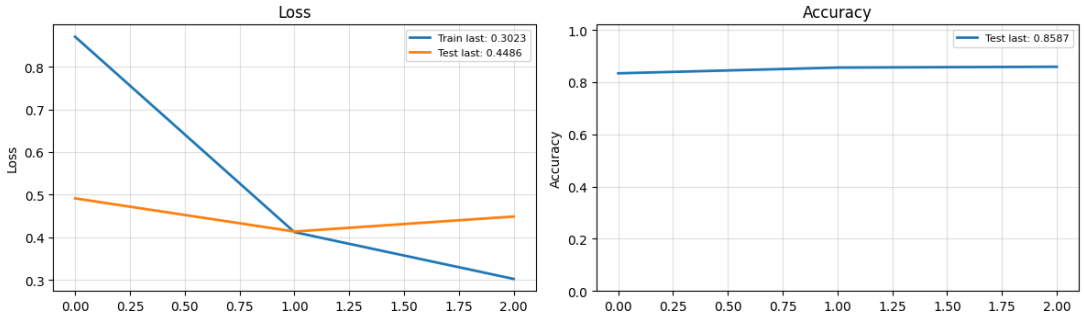

Figure 2. Training Loss and Accuracy of SciBERT

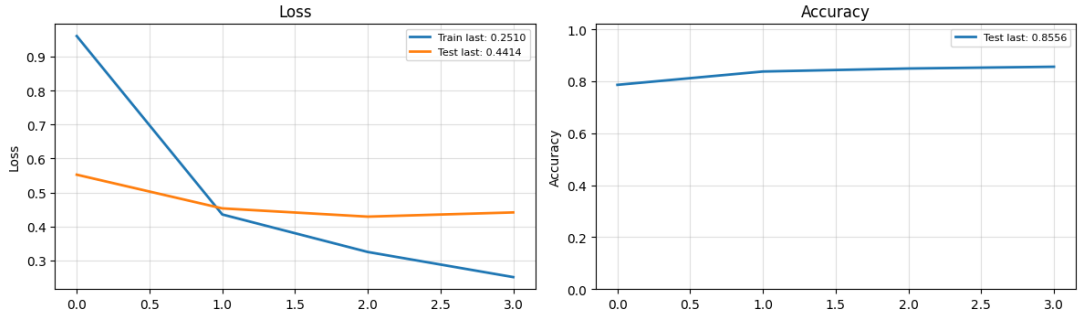

Figure 3. Training Loss and Accuracy of BioBERT

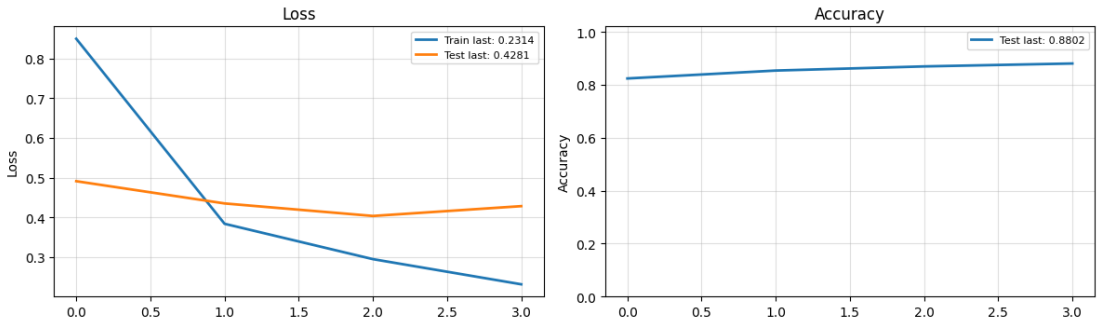

Figure 4. Training Loss and Accuracy of PubmedBERT

The learning dynamics of each model were examined through the training and validation loss curves and validation accuracy trends. **Figure 2.** presents the training behavior of SciBERT across epochs. The training loss decreased from 0.870326 at epoch 1 to 0.302321 at epoch 3, while the validation loss reached its minimum value of 0.413442 at epoch 2 before increasing to 0.448630 at epoch 3. Based on the lowest validation loss, the checkpoint from epoch 2 was retained for SciBERT.

Illustrates the training behavior of BioBERT, where **Figure 3.** the training loss decreased from 0.960425 at epoch 1 to 0.250992 at epoch 4. The validation loss decreased until epoch 3, reaching 0.428929, and increased to 0.441361 at epoch 4. Therefore, the checkpoint from epoch 3 was retained for BioBERT based on the early stopping criterion. **Figure 4**. shows the training behavior of PubMedBERT, with the training loss decreasing from 0.849350 at epoch 1 to 0.231449 at epoch 4. The validation loss reached its lowest value of 0.403423 at epoch 3 and increased to 0.428107 at epoch 4. Accordingly, the checkpoint from epoch 3 was retained for PubMedBERT.

Table 7. Performance Comparison of the Evaluated Models

No	Model	Acc	Prec	Rec	F1
1	BioBERT	84.88	86.30	84.88	85.17
2	SciBERT	85.56	86.21	85.56	85.69
3	**PubMedBERT**	**86.92**	**87.64**	**86.92**	**87.02**

Table 7. summarizes the final performance comparison of SciBERT, BioBERT, and PubMedBERT on the held-out test set. The comparison shows differences in classification performance among the evaluated models

under the same experimental settings. Based on the overall evaluation results, PubMedBERT was selected as the final. relation extraction model for subsequent analysis and knowledge graph construction

Figure 5. Confusion Matrix of PubMedBERT

To further examine the classification performance of the selected model, a confusion matrix analysis was conducted **Figure 5** presents the confusion matrix of PubMedBERT on the held-out test set. The model correctly classified the majority of samples across all relation categories, indicating consistent prediction performance. Misclassifications were primarily observed between relation categories with similar semantic contexts, whereas categories with more distinct contextual characteristics were classified more consistently. These observations complement the quantitative evaluation presented in **Table 7** and provide additional insight into the classification behavior of the selected model.

4.2 Knowledge Graph Construction Results

The relation predictions generated by the PubMedBERT model were utilized to construct the knowledge graph. The graph is structured using semantic triples (subject, relation, object), where each triple denotes a predicted semantic relationship extracted from the literature corpus

Table 8. Extracted Triples for Knowledge Graph Construction

Subject (Entity)	Relation Type	Object (Entity)
C. hypnaeoides	Seaweed–Location	Japan
C. hypnaeoides	Seaweed–Compound	Polysaccharide
C. hypnaeoides	Seaweed–Benefit	ameliorate insulin resistance
Polysaccharide	Compound–Benefit	ameliorate insulin resistance

Figure 6. Visualization of the Constructed Knowledge Graph.

The knowledge graph models the relationships between entities as a network structure. Nodes denote the extracted entities, while edges represent the predicted relations connecting them. This visualization maps the interdependencies among seaweeds, compounds, benefits, and locations.

4.3 Web-Based Visualization System

The extracted entities and relations were integrated into a web-based application for structured information retrieval from the seaweed literature corpus.

4.3.1 System Architecture

The application architecture is divided into two modules: a Home Page for data management and a Visualize Page for data representation. This framework supports the processing, storage, and rendering of the extracted entities and relations within a web interface.

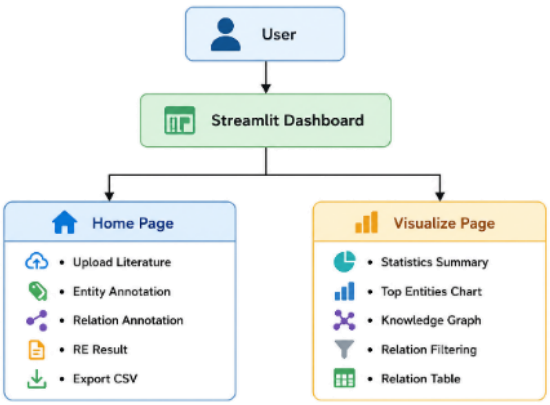

Figure 7. System Architecture

4.3.2 Data Management Interface

The Home Page functions as the data ingestion and processing interface. The implemented modules include document uploading, entity and relation annotation, rendering of the Relation Extraction (RE) outputs, and data exportation to CSV format. This interface centralizes the preprocessing and extraction phases.

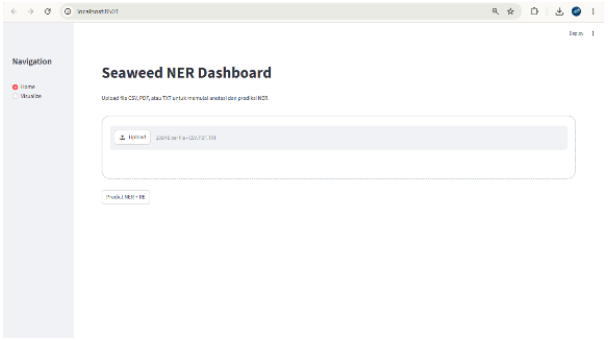

Figure 8. Web Interface for Document Upload and Literature Input

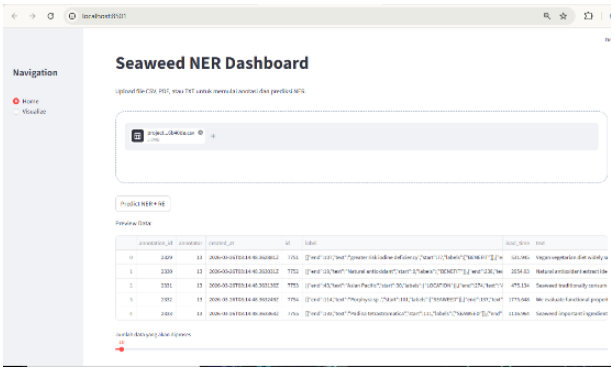

Figure 9. Annotation Interface for Entity and Relation Labeling.

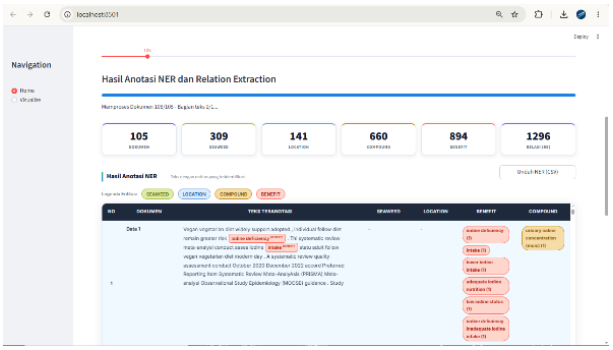

Figure 10. Extracted Entity Results from Named Entity Recognition

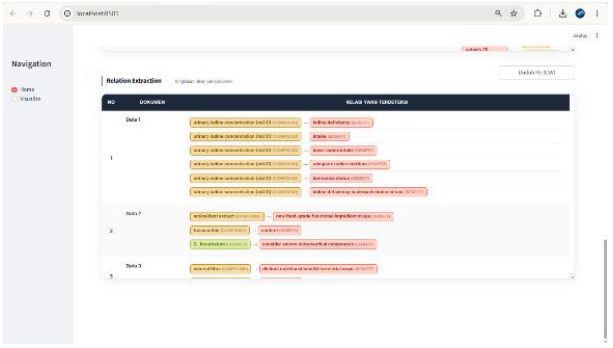

Figure 11. Predicted Relation Results from Relation Extraction.

The web-based system provides an interface for managing literature input and presenting the results of the information extraction pipeline. **Figure 8** shows the document upload interface, which allows users to submit literature documents as the input for subsequent processing stages. The uploaded documents are then processed through the entity recognition and relation extraction modules to obtain structured information from unstructured text. Following the input stage, **Figure 9** presents the annotation interface used for defining entity categories and relation types within the literature dataset.

The annotation process establishes the labeled information required for training and evaluating the relation extraction model, where entities and their corresponding relationships are represented according to the predefined annotation scheme. The extracted entity information generated by the Named Entity Recognition module is presented in **Figure 10.** The system identifies relevant entities from the processed literature, including seaweed, compounds, benefits, and locations. These extracted entities provide the basis for identifying semantic relationships in the subsequent relation extraction stage. **Figure 11** presents the relation extraction results generated by the trained PubMedBERT model. The system predicts relationships between extracted entity pairs and produces structured entity-relation outputs. These outputs are subsequently used as input for knowledge graph construction, where the extracted entities and relations are represented as interconnected graph components.

4.3.4 Analytical and Statistical Visualization

The Visualize Page aggregates the extracted data into quantitative representations. A Statistics Summary module reports numeric metrics, including the total count of processed documents, extracted entities, generated relations, and unique entity types. An Entity Distribution chart illustrates the occurrence frequency of each extracted category within the dataset.

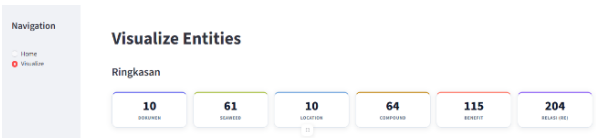

Figure 12. Statistical Summary Dashboard of Extracted Literature Information.

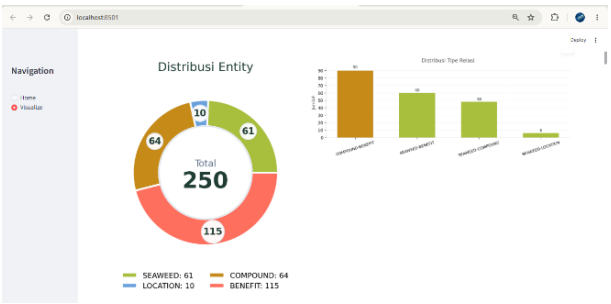

Figure 13. Overall Entity Distribution Visualization.

Figure 14. Seaweed Species Frequency Distribution.

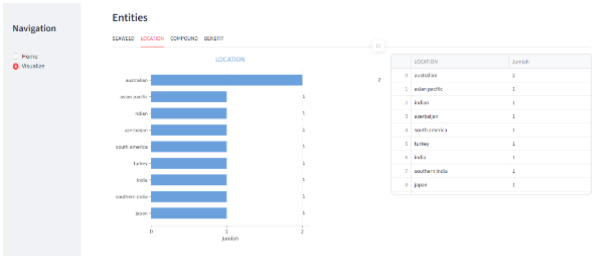

Figure 15. Location Entity Frequency Distribution.

Figure 16. Compound Entity Frequency Distribution.

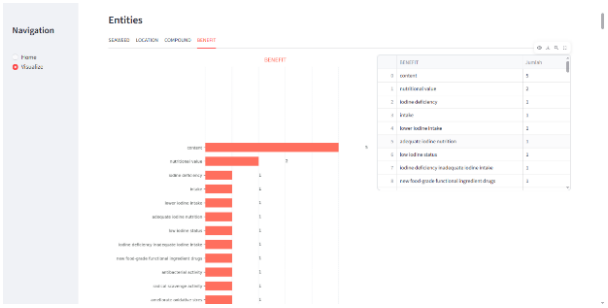

Figure 17. Benefit Entity Frequency Distribution.

The analytical visualization module presents statistical representations of the entities and relations extracted from the seaweed literature corpus. **Figure 12** presents a statistical summary of the processed dataset, including the total number of processed documents, extracted entities, and identified relations generated by the information extraction pipeline. These statistics provide a quantitative description of the structured information obtained from the literature after the NER and relation extraction stages. **Figure 13** presents the overall distribution of entity categories, including seaweed species, locations, compounds, and benefits, thereby illustrating the composition of the extracted entities within the dataset. To further characterize the extracted information, **Figure 14** presents the frequency distribution of seaweed species identified from the literature, reflecting the occurrence of each species across the processed documents**.**

Figure 15 presents the distribution of extracted location entities, describing the geographical locations associated with the reported seaweed studies. **Figure 16** presents the frequency distribution of compound entities identified from the literature, providing an overview of the chemical compounds most frequently associated with the extracted seaweed entities. **Figure 17** presents the distribution of benefit entities, summarizing the biological and functional benefits identified through the information extraction process. Collectively, these statistical visualizations provide complementary perspectives on the extracted dataset by presenting both the overall composition of entity categories and the frequency distribution of individual entity types, thereby supporting subsequent exploration of the constructed knowledge graph.

4.3.5 Relational Filtering and Network Visualization

A Relation Filtering module permits users to isolate data based on relation types or source documents. The filtered data is presented in two formats: a tabular Relation Table for record inspection, and a network visualization (Knowledge Graph). In the graph visualization, nodes denote the extracted entities (seaweeds, compounds, benefits, locations), and edges represent the predicted semantic relations.

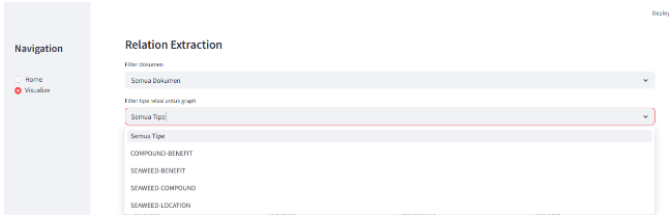

Figure 18. Entity relation filtering interface

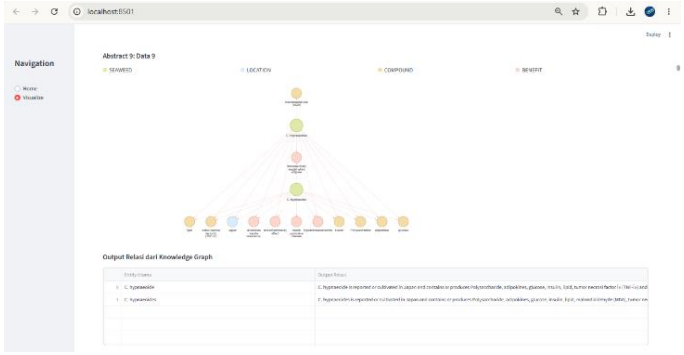

Figure 19. Extracted knowledge graph

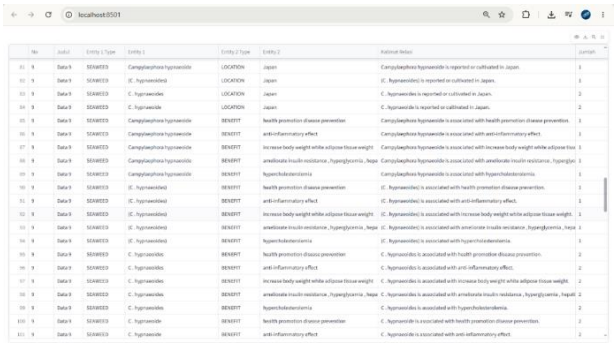

Figure 20. Extracted relation table

The constructed knowledge graph is presented through complementary visualization and tabular representations to facilitate the examination of semantic relationships extracted from the seaweed literature. **Figure 18** presents the entity relationship filtering interface, in which extracted relations are organized according to predefined relation categories. This representation allows each relation type to be examined independently, thereby reducing visual complexity and supporting the interpretation of specific semantic associations within the knowledge graph. **Figure 19** presents the knowledge graph generated from the extracted entity–relation triplets.

In this representation, entities are modeled as nodes and their semantic relationships as connecting edges, forming an interconnected graph that captures the structural organization of knowledge extracted from the literature corpus. The resulting graph illustrates how seaweed species are associated with compounds, biological benefits, and geographic locations through the relations predicted by the selected relation extraction model. Figure 20 presents the same extracted relationships in tabular form, where each record corresponds to an entity pair and its associated relation type. Unlike the graphical representation, the tabular format preserves the extracted relations as structured records, facilitating systematic inspection, verification, and subsequent processing while maintaining consistency with the knowledge graph representation.

4.5 DISCUSSION

4.5.1 Comparison of BERT-Based Models

The evaluation results indicate differences in relation extraction performance among BioBERT, SciBERT, and PubMedBERT. PubMedBERT achieved an Accuracy of 86.92%, Precision of 87.64%, Recall of 86.92%, and F1-Score of 87.02%. SciBERT achieved an Accuracy of 85.56%, Precision of 86.21%, Recall of 85.56%, and F1-Score of 85.69%. BioBERT achieved an Accuracy of 84.88%, Precision of 86.30%, Recall of 84.88%, and F1-Score of 85.17%. The difference between the highest and lowest F1-Score was 1.85 percentage points, while the difference between the highest and lowest Accuracy was 2.04 percentage points.

Across all reported metrics, PubMedBERT ranked first, SciBERT ranked second, and BioBERT ranked third. The three evaluated models differ in their pre-training corpora. PubMedBERT was pre-trained on PubMed literature, SciBERT was pre-trained on scientific publications from multiple disciplines, and BioBERT was further pre-trained on biomedical corpora after initialization from BERT. In this study, no additional experiments were conducted to isolate the effect of pre-training data on model performance. Therefore, the comparison is limited to the observed evaluation results obtained on the same relation extraction dataset and experimental configuration. Based on the reported Accuracy, Precision, Recall, and F1-Score values, PubMedBERT was selected as the relation extraction model for the subsequent knowledge graph construction and web-based visualization stages.

4.5.2 Implications for Knowledge Graph Construction

The outputs of the Relation Extraction (RE) model constitute the structural components of the resulting knowledge graph. The knowledge graph is constructed by converting predicted relations into subject–relation–object triples; consequently, classification outputs during the extraction phase determine the final graph topology. Structural variations in the knowledge graph correspond directly to the false positive and false negative rates in the RE model predictions.

A false positive produces an unverified edge—such as linking a seaweed species to an unannotated geographic location or bioactive compound—creating a structural link not present in the source dataset. A false negative results in a missing edge, where an annotated connection between entities is excluded from the graph. PubMedBERT obtained a Precision of 87.64% and a Recall of 86.92%. In the context of graph construction, the 87.64% Precision score indicates that 12.36% of the generated edges represent false positives. The 86.92% Recall score indicates that 13.08% of the relationships annotated in the source text are excluded from the graph as false negatives. The implementation of PubMedBERT yields a knowledge graph with a structural composition defined by these measured operational parameters and quantified error rates.

4.5.3 Practical Implications of the Visualization System

The integration of the PubMedBERT relation extraction model with the web-based visualization interface constitutes a system for querying extracted seaweed literature. The relation extraction model processes unstructured scientific text to output subject–relation–object triples. The web application subsequently maps these predicted triples into a graphical layout, rendering biological and geographic entities as nodes and their extracted relationships as edges. The system's search functionality accepts input parameters, including specific seaweed species or biological activities, and returns the corresponding relational network. This graphical output displays connected bioactive compounds, biological benefits, and geographic locations as an interactive

structure. Through this pipeline, the integrated extraction and visualization components transform raw scientific text into a structured, queryable data format, bypassing the requirement for manual document parsing.

4.5.4 Limitations and Future Work

The operational scope of the relation extraction and knowledge graph visualization system is constrained by the methodological parameters established during the training phase. The dataset utilized for model evaluation consists exclusively of annotated texts sourced from seaweed literature. Applying the current PubMedBERT architecture to process unstructured texts from divergent marine or botanical domains requires the integration of supplemental annotated corpora.

The classification schema allocates inputs strictly into four predefined categories: Seaweed–Location, Seaweed–Compound, Seaweed–Benefit, and Compound–Benefit. Semantic variables outside this specified schema—such as temporal sequences, extraction methodologies, and dosage measurements—are omitted from the extraction pipeline and the final knowledge graph topology. Subsequent research designs necessitate the expansion of the training corpora to increase the scope of processable domains. The methodological framework will be modified to integrate additional semantic categories into the classification schema. Planned experimental protocols include the application of alternative model architectures to calculate structural variations in false positive and false negative rates during the graph construction phase.

5. CONCLUSION

BioBERT, SciBERT, and PubMedBERT were evaluated for relation extraction in seaweed literature using a dataset containing four relation categories: Seaweed–Location, Seaweed–Benefit, Seaweed–Compound, and Compound–Benefit. PubMedBERT obtained an Accuracy of 86.92%, Precision of 87.64%, Recall of 86.92%, and F1-Score of 87.02%, compared with F1-Scores of 85.69% for SciBERT and 85.17% for BioBERT. Consequently, PubMedBERT was selected for the relation extraction stage. Predicted relations were converted into subject–relation–object triples for knowledge graph construction.

The resulting knowledge graph was integrated into a web-based visualization system comprising literature search, entity retrieval, relation retrieval, and graph visualization components. Entities are represented as nodes and relations as edges, providing a graphical representation of relationships among seaweed species, compounds, biological activities, and geographic locations extracted from the literature.

The presented framework combines relation extraction and knowledge graph visualization within a single workflow for seaweed literature. Future work may involve expanding the annotated corpus, incorporating additional relation categories, integrating automatic named entity recognition modules, and evaluating alternative language model architectures. Further development of the web application may include automated literature ingestion, periodic knowledge graph updates, advanced search and filtering functionalities, graph analytics features, and support for larger-scale knowledge graphs.

6. REFERENCES

[1] T. Chandra, M. Nor, M. I. Mohd, J. Xu, Z. Abdul, and L. Seong, “Heliyon Knowledge mapping analysis of the global seaweed research using CiteSpace,” *Heliyon*, vol. 10, no. 7, p. e28418, 2024, doi: 10.1016/j.heliyon.2024.e28418.

[2] M. Maskur, A. A. Prihanto, M. Firdaus, and R. Nurdiani, “Potential of seaweed bioactives for pharmaceutical applications [Potencial de los bioactivos de las algas marinas para aplicaciones farmacéuticas],” *J. Pharm. Pharmacogn. Res.*, vol. 14, no. 1, pp. 1–2, 2026, doi: 10.56499/jppres_14.1.2477.

[3] H. Hettiarachchi, M. M. Gaber, P. Parsafard, and E. Vakaj, “SNOWTEC: Synthetic Natural language Oversampling With Transformer-based information ExtraCtion for automated compliance checking,” *Mach. Learn. with Appl.*, vol. 24, p. 100911, 2026, doi: https://doi.org/10.1016/j.mlwa.2026.100911.

[4] T. Taslim, S. Handayani, W. Walhidayat, and D. Toresa, “Evaluating Contextual Embedding Models for Multi-Label PICO Classification in Heart Disease: Addressing the Intervention - Comparison Bottleneck,” *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 16, no. 2, pp. 84–95, 2025, doi: 10.31849/digitalzone.v16i2.28375.

[5] H. Rehana *et al.*, “Evaluating GPT and BERT models for protein–protein interaction identification in biomedical text,” *Bioinforma. Adv.*, vol. 4, no. 1, p. vbae133, Jan. 2024, doi: 10.1093/bioadv/vbae133.

[6] J. E. Evangelista *et al.*, “Creating an Interactive Web Interface for Networks Stored in Knowledge Graph Databases,” *Curr. Protoc.*, vol. 5, no. 9, pp. 1–43, 2025, doi: 10.1002/cpz1.70200.

[7] R. Martelo, K. Ahmadiyehyazdi, and R.-Q. Wang, “Towards democratized flood risk management: An advanced AI assistant enabled by GPT-4 for enhanced interpretability and public engagement,” *Environ. Model. Softw.*, vol. 197, p. 106821, 2026, doi: https://doi.org/10.1016/j.envsoft.2025.106821.

[8] R. Rocha and G. Â. de Lima, “Deigmata: A User-Centered Visualization Tool for Knowledge Graph Exploration,” *Knowl. Organ.*, vol. 52, no. 8, 2025, doi: 10.31083/ko46137.

[9] Y. Lee, J. Son, and M. Song, “BertSRC :

transformer - based semantic relation classification," *BMC Med. Inform. Decis. Mak.*, vol. 5, pp. 1–18, 2022, doi: 10.1186/s12911-022-01977-5.

[10] Y. Jia, H. Wang, Z. Yuan, L. Zhu, and Z. Xiang, "Biomedical relation extraction method based on ensemble learning and attention mechanism," *BMC Bioinformatics*, 2024, doi: 10.1186/s12859-024-05951-y.

[11] L. Luo, P. T. Lai, C. H. Wei, C. N. Arighi, and Z. Lu, "BioRED: A rich biomedical relation extraction dataset," *Brief. Bioinform.*, vol. 23, no. 5, 2022, doi: 10.1093/bib/bbac282.

[12] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 3219–3232, 2018, doi: 10.18653/v1/d18-1360.

[13] H. Wang *et al.*, "Predicting ICU in-hospital mortality from text-encoded structured EHR data using adaptive transformer layer fusion," *iScience*, vol. 29, no. 6, p. 116135, 2026, doi: 10.1016/j.isci.2026.116135.

[14] M. Grandini, E. Bagli, and G. Visani, *Metrics for Multi-Class Classification: an Overview*. 2020. doi: 10.48550/arXiv.2008.05756.

[15] J. Doménech, S. Mhiri, O. León, M. S. Siddiqui, and J. Pegueroles, "Robust AI-based intrusion detection for IoMT: A data-efficient approach via optimized feature selection and hybrid data balancing," *Comput. Networks*, vol. 284, p. 112359, 2026, doi: https://doi.org/10.1016/j.comnet.2026.112359.

[16] I. Mouiche and S. Saad, "Entity and relation extractions for threat intelligence knowledge graphs," *Comput. Secur.*, vol. 148, p. 104120, 2025, doi: https://doi.org/10.1016/j.cose.2024.104120.

[17] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022, doi: 10.1109/TNNLS.2021.3070843.

[18] K. Sun, Y. Liu, Z. Guo, and C. Wang, "EduVis: Visualization for education knowledge graph based on web data," *ACM Int. Conf. Proceeding Ser.*, pp. 138–139, 2016, doi: 10.1145/2968220.2968227.

[19] S. Latif *et al.*, "Visually Connecting Historical Figures Through Event Knowledge Graphs," *Proc. - 2021 IEEE Vis. Conf. - Short Pap. VIS 2021*, pp. 156–160, 2021, doi: 10.1109/VIS49827.2021.9623313.

[20] W. Chen, J. Hou, Y. Wang, and M. Yu, "Visualization analysis of concrete crack detection in civil engineering infrastructure based on knowledge graph," *Case Stud. Constr. Mater.*, vol. 21, p. e03711, 2024, doi: 10.1016/j.cscm.2024.e03711.