Evaluating Medical Entity Recognition in Healthcare: A Comprehensive Analysis of BERT-Based Models

Shengyu Liu, Anran Wang, Xiaolei Xiu, Ming Zhong, Sizhu Wu

Submitted to: JMIR Medical Informatics
on: April 23, 2024

Table of Contents

Evaluating Medical Entity Recognition in Healthcare: A Comprehensive Analysis of BERT-Based Models

Shengyu Liu[1]; Anran Wang[1]; Xiaolei Xiu[1]; Ming Zhong[1]; Sizhu Wu[1]

[1]Department of Medical Data Sharing Institute of Medical Information & Library Chinese Academy of Medical Sciences & Peking Union Medical College Beijing CN

Corresponding Author:
Sizhu Wu
Department of Medical Data Sharing
Institute of Medical Information & Library
Chinese Academy of Medical Sciences & Peking Union Medical College
3 Yabao Road
Chaoyang District
Beijing
CN

Abstract

Background: Named Entity Recognition (NER) models play a pivotal role in deciphering unstructured medical texts by identifying diseases, treatments, and conditions, thereby advancing clinical decision-making and research. Machine learning innovations, especially in deep learning, have notably enhanced NER capabilities. Yet, their performance is inconsistent across medical datasets due to the complexity of medical terminology and linguistic variety. Prior studies have predominantly analyzed general NER performance, overlooking specific applications in medical scenarios and the challenges therein. Moreover, an in-depth analysis of how leading models and macro-factors, such as linguistic composition, affect NER accuracy is needed. This deficiency impedes the refinement of NER models for medical applications, which is vital for improving patient outcomes and the efficiency of healthcare services.

Objective: This study aims to meticulously evaluate the performance of BioBERT, RoBERTa, BigBird, and DeBERTa NER models within medical text analysis, concentrating on varied medical datasets to determine how complex medical terminology and linguistic diversity affect entity recognition accuracy. It also examines the role of macro-factors, including the lexical composition of entity phrases, in influencing the efficacy of specific models. The goal is to bridge the current research gap by offering insights that facilitate refining NER models for medical use, ultimately advancing patient care and healthcare service efficiency.

Methods: This study conducts a thorough evaluation of four prominent NER models: BioBERT, RoBERTa, BigBird, and DeBERTa. The focus is assessing prediction accuracy, training efficiency, computational resource use (CPU and GPU), etc. We utilized three diverse medical datasets-Revised JNLPBA, BC5CDR, and AnatEM-selected for their relevance to the medical field. Furthermore, the study explores the impact of significant macro-factors, like the number of words in an entity phrase, on the models' performance. A systematic analysis of these factors' influence on prediction accuracy across the datasets was performed, aiming to gain an in-depth understanding of the impact of different macro-factors on the prediction accuracy of the medical NER model.

Results: The analysis shows that the BioBERT model exceeded the performance of other models in prediction accuracy across the Revised JNLPBA, BC5CDR, and AnatEM medical datasets, highlighting its superior proficiency in identifying medical entities. Nevertheless, its accuracy was not consistently superior across all entity types. Additionally, the research confirmed that macro-factors, such as the number of words in an entity phrase, markedly affect the prediction accuracy of the models.

Conclusions: This study highlights the essential role of NER models in medical informatics, emphasizing the imperative for model optimization via precise data targeting and fine-tuning. The insights from this study will notably improve clinical decision-making and facilitate the creation of more sophisticated and effective medical NER models.

(JMIR Preprints 23/04/2024:59782)
DOI: https://doi.org/10.2196/preprints.59782

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Evaluating Medical Entity Recognition in Healthcare: A Comprehensive Analysis of BERT-Based Models

Shengyu Liu[1], Anran Wang[1], Xiaolei Xiu[1], Ming Zhong[1], Sizhu Wu[1]

[1] Department of Medical Data Sharing, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College

Abstract

Background: Named Entity Recognition (NER) models play a pivotal role in deciphering unstructured medical texts by identifying diseases, treatments, and conditions, thereby advancing clinical decision-making and research. Machine learning innovations, especially in deep learning, have notably enhanced NER capabilities. Yet, their performance is inconsistent across medical datasets due to the complexity of medical terminology and linguistic variety. Prior studies have predominantly analyzed general NER performance, overlooking specific applications in medical scenarios and the challenges therein. Moreover, an in-depth analysis of how leading models and macro-factors, such as linguistic composition, affect NER accuracy is needed. This deficiency impedes the refinement of NER models for medical applications, which is vital for improving patient outcomes and the efficiency of healthcare services.

Objective: This study aims to meticulously evaluate the performance of BioBERT, RoBERTa, BigBird, and DeBERTa NER models within medical text analysis, concentrating on varied medical datasets to determine how complex medical terminology and linguistic diversity affect entity recognition accuracy. It also examines the role of macro-factors, including the lexical composition of entity phrases, in influencing the efficacy of specific models. The goal is to bridge the current research gap by offering insights that facilitate refining NER models for

Corresponding author:

Sizhu Wu (PhD)

Department of Medical Data Sharing, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, 100020, Beijing, China.

Phone: +86-10-5232-8760

Email: wu.sizhu@imicams.ac.cn

medical use, ultimately advancing patient care and healthcare service efficiency.

Methods: This study conducts a thorough evaluation of four prominent NER models: BioBERT, RoBERTa, BigBird, and DeBERTa. The focus is assessing prediction accuracy, training efficiency, computational resource use (CPU and GPU), etc. We utilized three diverse medical datasets-Revised JNLPBA, BC5CDR, and AnatEM-selected for their relevance to the medical field. Furthermore, the study explores the impact of significant macro-factors, like the number of words in an entity phrase, on the models' performance. A systematic analysis of these factors' influence on prediction accuracy across the datasets was performed, aiming to gain an in-depth understanding of the impact of different macro-factors on the prediction accuracy of the medical NER model.

Results: The analysis shows that the BioBERT model exceeded the performance of other models in prediction accuracy across the Revised JNLPBA, BC5CDR, and AnatEM medical datasets, highlighting its superior proficiency in identifying medical entities. Nevertheless, its accuracy was not consistently superior across all entity types. Additionally, the research confirmed that macro-factors, such as the number of words in an entity phrase, markedly affect the prediction accuracy of the models.

Conclusions: This study highlights the essential role of NER models in medical informatics, emphasizing the imperative for model optimization via precise data targeting and fine-tuning. The insights from this study will notably improve clinical decision-making and facilitate the creation of more sophisticated and effective medical NER models.

KEYWORDS

Artificial intelligence; AI; Model evaluation; macro-factors; medical named entity recognition models

Introduction

Medical Named Entity Recognition (NER) is a crucial component of Natural Language Processing (NLP) in medical informatics, vital for analyzing unstructured textual data such as medical records. Misclassifications in NER can lead to severe consequences, such as misdiagnosis or inadequate treatment plans, emphasizing the critical need for precise handling of medical terminology [1,2,3]. Given the significant implications, rigorous evaluation of NER models is essential to ensure their accuracy, reliability, and appropriateness for medical contexts, ultimately supporting clinical decision-making [4].

In the rapidly evolving field of medical informatics, the introduction of BERT (Bidirectional Encoder Representations from Transformers)-based NER model variants [5], including RoBERTa, BigBird, DeBERTa, and BioBERT models, represents a significant breakthrough, markedly improving medical data analysis. These models utilize advanced feature extraction technologies from mask language models, ELMo, and Transformer architecture, establishing new benchmarks for precisely analyzing diverse medical datasets [6]. Specifically, RoBERTa introduced enriched training data and improved masking patterns [7]. BigBird's ability to efficiently process extended sequences [8] and DeBERTa's innovative attention mechanisms have effectively overcome previous models' scalability and comprehension issues [9]. BioBERT, through its specialized pre-training on comprehensive medical texts, highlights the effectiveness of domain-specific adaptations, achieving unparalleled precision in recognizing medical entities [10]. This strategic focus on the subtleties of context and specialized terminology significantly enhances the quality of patient care decisions. Furthermore, the adoption of advanced methodologies such as reinforcement learning, exemplified by Deng et al.

through their adaptive resource allocation strategy, represents the ongoing efforts to refine NER models for greater efficiency and accuracy [11].

Given the proficiency of the RoBERTa, BigBird, DeBERTa, and BioBERT models, alongside their exceptional ability to discern nuanced contextual differences and medical terminology, we have chosen them for forthcoming evaluations of medical NER models. These models not only excel in prediction accuracy but also in operational efficiency. However, despite significant progress in NLP, the Large Language Model (LLM) still encounters considerable challenges in medical NER tasks, which include limited prediction efficiency, extended runtimes, and substantial hardware requirements, as documented by Tian et al. [12], Zhao et al. [13], and Hu et al. [14]. Specifically, Tian et al. observed that the complexity and specificity of medical terminology pose significant challenges for the GPT model under LLM. They noted suboptimal prediction accuracy when using medical gold standard datasets such as BC5CDR-disease, JNLPBA, and NCBI-disease [12]; Zhao et al. highlighted that despite advancements, the performance of the LLM in NER tasks significantly lagged behind that of supervised baselines. This discrepancy was primarily attributed to the inherent differences in large and NER models' optimization objectives; they also noted the substantial computational demands and extended runtimes associated with using LLM for NER tasks [13]; Hu et al. underscored the necessity for substantial hardware resources to run these models effectively. These findings indicate that while LLM, such as GPT, has demonstrated promise in general language processing tasks, their application in the medical domain, particularly for medical NER tasks, necessitates significant adaptations to align with the elevated standards of medical data interpretation and utilization [14]. Moreover, drawing from these RoBERTa, BigBird, DeBERTa, and BioBERT models, numerous Transformer models have undergone rigorous training, refinement, and calibration while maintaining a coherent underlying architecture, as evidenced by Kalyan et al. [15].

Evaluating the RoBERTa, BigBird, DeBERTa, and BioBERT models is fundamental in establishing standards for enhancing the precision and longevity of the models. Recent studies by Freund et al. [16], Ahmad et al. [17], and others highlight a shift towards more comprehensive evaluation metrics tailored to the specific needs of medical informatics. This shift facilitates the development of advanced NER applications, significantly improving patient care by enhancing clinical data management. This marks a substantial advancement in the integration of technology and healthcare. However, the evaluations' primary focus has been assessing the predictive capabilities of NER models using metrics such as precision, recall, and F1-Score. Research by Yoon et al. [18], Yu et al. [19], and Yadav et al. [20]. have employed these metrics to evaluate these capabilities. Erdmann's group has also compared various NER tools for literary text corpora with human annotators using the same metrics, highlighting the importance of such evaluations in the digital humanities [21]. Further, the work of Usha et al. [22] and Nagaraj et al. [23], which includes advanced techniques such as confusion matrices, ROC, and PR curves, offers a more nuanced understanding of classification accuracy and errors in NER models. As Ozcelik's team and Akhtyamova explored, error analysis is critical for understanding performance nuances, especially in identifying and categorizing short- and long-term entities. This comprehensive evaluation approach underscores the complexities and challenges in developing accurate and efficient NER models for medical informatics [24,25].

The above study predominantly utilizes standard metrics such as precision, recall, ROC curve, and F1-score to evaluate model performance. However, these metrics fall short in capturing performance variations across different medical NER datasets and in conducting a detailed analysis of how dataset characteristics affect model performance and their interdependencies. Moreover, although these metrics are easy to compute, their interpretation proves challenging. This difficulty mainly arises from the metrics' failure to explain the broader reasons behind model outcomes, as they often depend on processing micro-vector features that are not intuitively understandable. Consequently, researchers focused on enhancing medical NER datasets through NER models encounter significant

hurdles in devising effective optimization strategies. In response, some researchers have shifted to customizing macro-factors for evaluating explanatory NER models. For example, Fu's team has developed an evaluation framework that outlines eight distinct factor types to analyze their correlation with the models' F1-Score rankings [26]. Then, Zhou et al. proposed an ant colony optimization (ACO) algorithm based on parameter adaptation. They designed a new dynamic parameter adjustment mechanism to adaptively adjust the pheromone importance factor. This algorithm is also suitable for selection of macro-factors. By adaptively changing the macro-factors, the algorithm can determine which macro-factors affect the prediction accuracy of the NER model [27]. And Yao's team has also enhanced this domain with their groundbreaking AMMSE algorithm, which adjusts the scale of structural elements to measure macro-factors' impact on model prediction accuracy [28]. These advancements have led to increasingly sophisticated NER model evaluations, resulting in more precise and resilient models.

Based on the abovementioned evaluation methods, this study adopts a comprehensive evaluation approach that merges traditional and innovative techniques to enhance the accuracy and reliability of NER applications in medical informatics. This method integrates performance analysis of hardware indicators, such as training duration and CPU/GPU usage, with precision assessment across various medical entity types in RoBERTa, BigBird, DeBERTa, and BioBERT models. Additionally, it introduces the Multi-layer Factor Elimination (MFE) algorithm that combines linear and machine learning strategies. By employing multi-layer factor filtering, this algorithm examines the influence of macroscopic factors on prediction accuracy. This comprehensive evaluation method significantly contributes to the overlapping fields of artificial intelligence and healthcare, advancing patient care quality through more accurate medical data analysis.

Methods

This section outlines two methods: "Training and evaluating medical NER models" and "Further assisted evaluation". The first method evaluates the prediction accuracy of the RoBERTa, BigBird, DeBERTa, and BioBERT models across different medical entity types and overall model effectiveness. The second method further assesses the prediction accuracy of merged entity types within these models' post-classification and examines macro-factors' influence on model performance.

Training and evaluating medical NER models

This method trains, validates, and tests NER models using sophisticated hyperparameter optimization techniques. We employed systematic strategies, such as grid search, for hyperparameter tuning across models like RoBERTa and BioBERT, utilizing datasets like Revised JNLPBA and AnatEM. Our evaluation aimed at enhancing model accuracy by making precise adjustments to learning rates, batch sizes, and other parameters.

Dataset selection

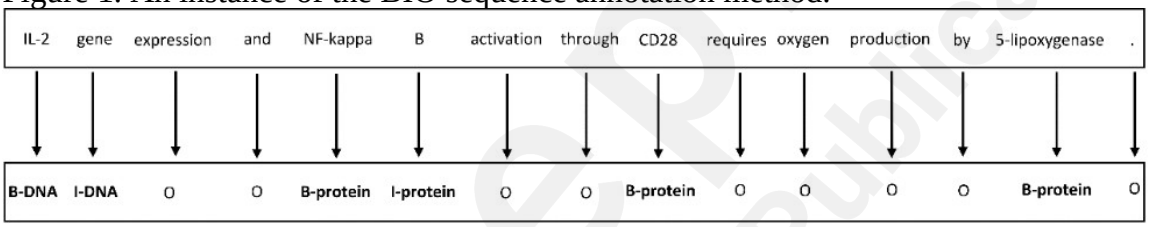
Three medical NER datasets were utilized to evaluate the prediction accuracy of models across different medical contexts (refer to Table 1). These datasets use the BIO sequence annotation method (refer to Figure 1); among them, the Revised JNLPBA dataset, provided by Huang et al. [29], was selected because it retains the original semantic annotation type while addressing known vulnerabilities from the original JNLPBA dataset. It features five entity types (DNA, RNA, protein, cell line, and cell type), offering a focused scope on biological entities. The BC5CDR dataset,

officially released by Li et al. [30], was chosen for its two distinct entity types - Disease and Chemical, which present unique challenges due to their complex and overlapping terminology. Lastly, the AnatEM dataset [31], focusing on anatomical entities in medical fields, was used due to its broad range of twelve different entity types, providing a wide spectrum of medical terms. This diversity in entity numbers and types - most of which are distinct across the datasets - strengthens the evaluation by exposing the models to varied linguistic challenges, thereby reducing the randomness and contingency of the experimental results and enhancing the overall credibility of the experiment.

Table 1. Descriptive Statistics of the medical NER datasets.

Dataset	Medical Entity types	Number of entity types	Number of annotations
Revised JNLPBA	DNA, RNA, protein, cell line and cell type	5	52785
BC5CDR	Disease and Chemical	2	38596
AnatEM	Organism Subdivision, Anatomical System, Organ, Multi-Tissue Structure, Tissue, Cell, Developing Anatomical Structure, Cellular Component, Organism Substance, Immaterial Anatomical Entity, Pathological Formation and Cancer.	12	11562

Figure 1. An instance of the BIO sequence annotation method.



Model training, cross-validation and test

In this section, we achieved optimal prediction accuracy through meticulous hyperparameter tuning during the model training. Specifically, we trained models such as RoBERTa, BigBird, DeBERTa, and BioBERT using datasets carefully divided into training, validation, and test sets. These datasets were sourced from the Revised JNLPBA, BC5CDR, and AnatEM collections. For hyperparameter tuning, a grid search strategy was employed to systematically investigate various values for key hyperparameters, including learning rate, batch size, epochs, and dropout rate. The objective was to identify the optimal combination that maximized performance on the validation set. Subsequently, each model's prediction accuracy was evaluated on the test set using F1-Scores for specific entity types and overall performance metrics, namely micro (MICRO_F) and macro (MACRO_F) averages [32,33].

The selection and tuning of hyperparameters were guided by cross-validation, a robust method for model validation that ensures the model performs well on unseen data. In this context, hyperparameter optimization aims to identify a global or satisfactory local optimum that maximizes medical NER model performance, which involves adjusting several key hyperparameters:

Learning Rate Adaptation Strategy: We implemented a dynamic learning rate adjustment strategy that adapts to the behavior of gradient descent during training, thereby enhancing the model's efficiency in converging to an optimal local minimum. We explored learning rates of 0.001, 0.0001, and 0.00001 to establish the best balance for our models. This method draws on research by

Nakamura et al., which underscores the advantages of adaptive learning rate techniques in deep learning applications for NLP [34].

Batch Size Considerations: Informed by the research on neural network training dynamics [35], computational constraints and the impact of batch size on model generalization guided our batch size selection. We tested batch sizes of 10, 12, 20, and 50 to enable more frequent model updates, thereby facilitating a more granular approach to convergence.

Epoch Configuration: The number of training epochs was carefully adjusted according to the dataset's complexity and initial performance metrics [36]. This approach facilitated an adaptive training period, minimizing the overfitting risk. Specifically, epoch values were assigned as 1, 5, and 10 to ensure practical model training without succumbing to overfitting.

Dropout for Regularization: We utilized dropout as a regularization technique to mitigate model overfitting on training data [37]. The dropout rates were adjusted to 0.1, 0.2, 0.3, 0.4, and 0.5 based on interim validation performance, which ensured robust generalization across unseen medical texts and promoted model reliability.

Evaluative Metrics and Model Assessment

After the training, cross-validation, and test phases, we evaluated the models using an extensive array of F1-Score metrics, which assessed each medical entity type and micro and macro averages (MICRO_F and MACRO_F). This evaluation strategy effectively addresses disparities in entity frequency within the dataset [33]. Furthermore, we introduced entity-specific precision-recall curves to examine the effects of varying classification thresholds, thereby deepening our understanding of the models' sensitivity in medical contexts. This comprehensive method thoroughly assesses the models' performance across various conditions, establishing a new benchmark for NLP model evaluation in medical applications.

To enhance our evaluation framework, we assessed the training time efficiency and the use of computational resources, including CPU and GPU utilization. These metrics offer insights into the operational demands of each model, enabling us to evaluate their prediction accuracy and practicality for real-world applications.

Further assisted evaluation

This method further evaluates the relationship between medical data and NER models, focusing on categorical relaxation to reduce entity classification complexity and improve prediction accuracy. We detail the process of merging similar entity types into broader categories and evaluating the impact on model performance through a comprehensive analysis of macro-factors such as sentence length and entity density. These methods are applied across several datasets, such as Revised JNLPBA, BC5CDR, and AnatEM, to assess and refine NER models systematically.

Academic Revision on Categorical Relaxation

Categorical relaxation in NER classification, particularly within the medical domain, entails merging similar medical entity types to diminish ambiguity and enhance the classification performance of NER models. This technique simplifies the classification landscape and bolsters the models' capacity to generalize from training data to unseen clinical examples. In this study, we implemented categorical relaxation by consolidating specific medical entity types, such as merging different DNA or RNA names. This method was guided by empirical evidence suggesting that merging reduces misclassifications and boosts prediction accuracy, particularly in diagnosing conditions and recommending treatments.

In the Revised JNLPBA dataset, we adopted a merging strategy based on the principles described by Tsai et al. [38]. Biologically related categories such as DNA, RNA, and proteins were consolidated into a single “Macromolecule” category. Similarly, entities categorized as cell lines and cell types were combined into a single “Cell” category.

In the AnatEM dataset, we reclassified 12 entity types into four broader categories relevant to human health, following the classification guidelines from Pyysalo et al. [31]. This reorganization is predicated on the premise that broader categories more effectively capture essential information and reduce the noise caused by particular and infrequently occurring entities.

Following categorical relaxation, we comprehensively evaluated the RoBERTa, BigBird, DeBERTa, and BioBERT models. This assessment compared the prediction accuracy of these models on the newly consolidated entity types. The objective was to determine the effects of entity type consolidation on model performance, particularly whether simplifying categories enhanced prediction accuracy across different model architectures. This method simplifies entity classification and capitalizes on the inherent similarities among entity types to improve model training and evaluation. By reducing the granularity of entity types, we posited that the models would achieve higher accuracy and more effectively address the complexities of medical texts. The specific outcomes of this categorical relaxation are detailed in Table 2.

Table 2. Merged Entity Types.

Dataset	Medical entity types	Merged entity types
Revised JNLPBA	DNA, RNA, protein	Macromolecule
	cell line, cell type	Cell
BC5CDR	Disease	Disease
	Chemical	Chemical
AnatEM	Organism Subdivision, Anatomical System, Organ, Multi-Tissue Structure, Tissue, Cell, Developing Anatomical Structure, Cellular Component	Anatomical Structure
	Organism Substance	Organism Substance
	Immaterial Anatomical Entity	Immaterial Anatomical Entity
	Pathological Formation, Cancer	Pathological Formation

Constructing and Evaluating NER Macro-Factors’ Datasets

1) Macro-factor metrics definition

Entity’s macro-factor metrics were defined in an entity (entity phrase fragment) or on the entire dataset, and its different attributes were described. Based on the eight-factor types defined by Fu’s team [26], we categorized into six macro-factor metrics, which were sentence length (*sLen*), entity phrase length (*eLen*), entity density (*eDen*), the number of entity words in each entity phrase (*eNum*), total entity word counts in each entity type (*tEWC*), and entity label consistency (*elCon*).

Sentence length (*sLen*): This metric refers to the string length of the sentence containing the entity phrase. It quantifies the contextual space in which entities appear. Notably, extreme values are excluded to prevent distortion in average calculations.

Entity phrase length (*eLen*): This metric quantifies the string length of each entity phrase, which can comprise one or more entity words. Like *sLen*, extreme values are excluded to yield a more precise average entity phrase length for each entity type.

Entity density (*eDen*): This metric is calculated as the ratio of *eLen* to sLen; this metric quantifies how dense entities are populated within the text.

Number of entity words in each entity phrase (*eNum*): For datasets like Revised JNLPBA, labeled with the BIO sequence, this metric counts the entity words in a phrase, adjusting for labeling specifics such as combining “B-Entity label” and “I-Entity label” into a single count to avoid underrepresentation in the data.

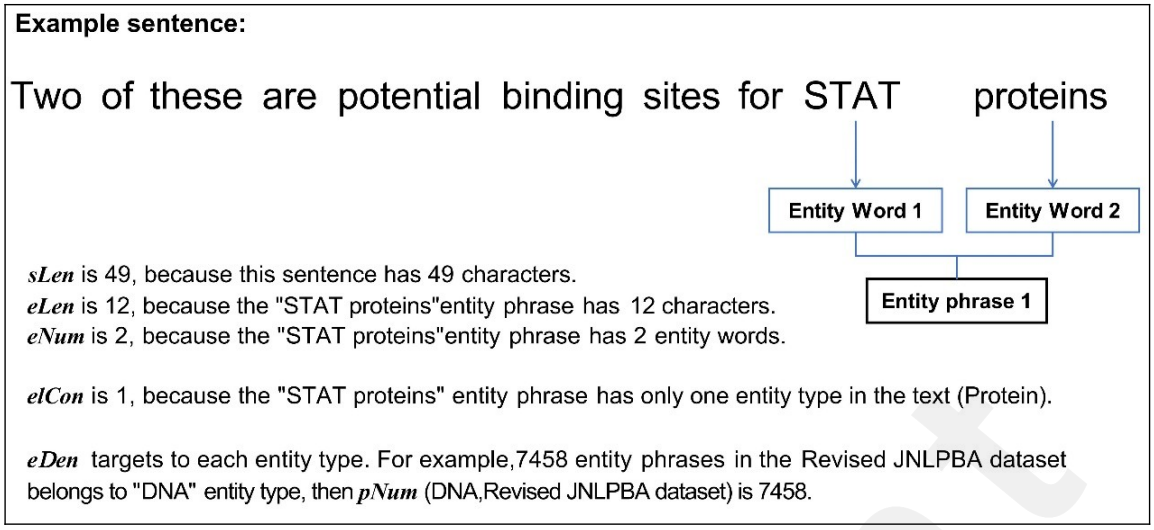
Total entity word counts in each entity type (*tEWC*): This metric quantifies the cumulative number of entity words within each specific entity type across datasets such as Revised JNLPBA, BC5CDR, and AnatEM.

Entity label consistency (*elCon*): This metric evaluates the consistency of entity-type assignments to medical terms across various contexts, which is essential in datasets like the Revised JNLPBA dataset, where terms can possess multiple semantic interpretations. For instance, “lymphocyte” may be categorized as “B-cell_type” in discussions about receptor counts and as “I-cell_type” in contexts involving blood samples. To calculate *elCon*, each term from the dataset is cataloged along with its associated entity types. For example, “lymphocyte” would be documented with both “B-cell_type” and “I-cell_type”. A weight ω assigned to each term based on the inverse number of its entity types (ω=1/Number of entity types), indicating that terms linked to fewer entity types tend to demonstrate higher prediction accuracy. Terms associated with a single entity type are assigned a weight of 1, while those with two types are given a weight of 0.5. We calculate the average weight by the formula: average weight value ($\acute{\omega}$) = $\sum_{i}^{n} \omega_i - \sigma\left(\omega_{\sigma_1} + \omega_{\sigma_2}\right)$, excluding extreme values, to avoid data skew.

2) The new macro-factors’ datasets creation

Then, we derived three macro-factors’ datasets from the Revised JNLPBA, BC5CDR, and AnatEM NER datasets. Each dataset includes columns for metrics such as *sLen*, *eLen*, *eNum*, *eDen*, *elCon*, *tEWC*, and labels. Among these, *sLen*, *eLen*, *eNum*, *eDen*, and *elCon* were calculated for each entity word within each entity type, and their respective average values were computed. Additionally, we included *tEWC* as a separate column to represent the total count of entity words for every kind across the datasets, quantifying the overall entity word volume. Figure 2 displays the systematic computation of these metrics, except *tEWC*, to augment the macro-factors’ datasets.

Figure 2. *sLen*, *eLen*, *eNum*, *eDen* and *elCon* values of an entity word.



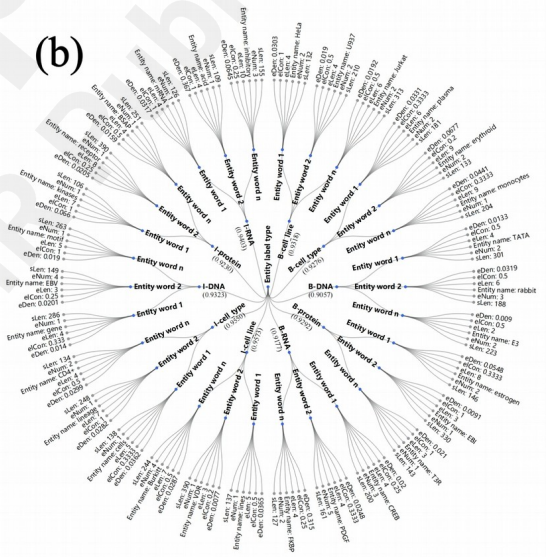
After training, cross-validating, and testing the RoBERTa, BigBird, DeBERTa, and BioBERT models, we selected the model with the highest F1-Score as the labels for the different macro-factors' datasets. The labels reflect the F1-scores for each entity type, as Figure 3(a) depicts. Figure 3(b) provides an example, illustrating the dataset's structure revised based on these macro-factors.

Figure 3. Consists of two parts illustrating different components of each macro-factors' dataset. Panel (a) displays the labels. Panel (b) presents a visualization of *sLen*, *eLen*, *eNum*, *eDen*, and *elCon* macro-factor metrics using the Radial Tree layout algorithm, offering a structured view of these metrics' interrelationships.

(a)

	Level	F1-Score
0	B-cell_type	0.9276
1	I-cell_type	0.955
2	B-protein	0.9292
3	I-protein	0.923
4	B-RNA	0.9177
5	I-DNA	0.9323
6	B-cell_line	0.9318
7	I-cell_line	0.9573
8	B-DNA	0.9057
9	I-RNA	0.9403
	AVG_MICRO	**0.932**
	AVG_MACRO	**0.9298**

(b)



3) Preliminary macro-factors evaluation

Our preliminary analysis adopted a comprehensive method to assess the interactions between prediction accuracy for various entity types (e.g., diseases, DHA, RNA) and six macro-factors. Initially, we extracted prediction accuracy data for each entity type from four NER models (BigBird, RoBERTa, DeBERTa, and BioBERT), and we calculated the average values for macro-factor metrics such as *sLen*, *eLen*, *eNum*, *eDen*, *elCon*, and *tEWC*. We then explored the impact of each metric on the prediction accuracy for each entity type. We considered using trend analysis to provide deeper insights into the relationships between these metrics and accuracy levels. Additionally, the

visualization of multi-model predictive trends offered a comprehensive view of model robustness across different entity types.

4) In-depth macro-factors selection

Our subsequent research focused on identifying which macro-factor significantly impacted the model's prediction accuracy. We developed a Multi-layer Factor Elimination (MFE) algorithm and conducted macro-factor selection. MFE is an improved algorithm based on Recursive Feature Elimination [39], divided into three layers.

a. First Layer - Factor Ranking and Selection:

Process: The MFE algorithm inputs the calculated values for *sLen*, *eLen*, *eNum*, *eDen*, *tEWC*, and *elCon* for each entity word.

Analysis: Perform linear correlation analysis for each factor using the following function. The correlation coefficient (r) is a statistical measure used to quantify the strength and direction of the relationship between two variables. x_i represents individual macro-factor measurements, $\acute{x}$ is the mean of all macro-factor measurements, y_i represents individual model performance measurements, and $\acute{y}$ is the mean of all model performance measurements. This calculation quantifies the relationship between each macro-factor and the model's prediction accuracy.

$$r=\frac{\sum (x_i-\acute{x})(y_i-\acute{y})}{\sqrt{\sum (x_i-\acute{x})^2 \sum (y_i-\acute{y})^2}}$$

Selection: Utilizing a method similar to the Pearson Correlation method [40], factors are ranked by their correlation scores, and the top four are retained for further analysis. This step ensures that only the factors with the highest potential impact are advanced, thereby efficiently streamlining the feature space.

b. Second layer - Random Forest Evaluation:

Process: A random forest model performs multiple training iterations with the selected macro-factors from the first layer.

Evaluation: After each training session, cross-validation is utilized to evaluate model performance. The macro-factor with the lowest feature importance score, indicating minimal contribution to prediction accuracy, is systematically excluded from subsequent analyses. This iterative refinement process prioritizes factors that consistently enhance model accuracy [41].

c. Third Layer - Regression Model Optimization:

Process: Integrate the refined set of macro-factors into a regression model.

Optimization: Sequentially eliminate the least impactful macro-factors based on the influence of their coefficients on the model, which is calculated as follows:

$$Coefficient\ Influence = \beta$$

β is the coefficient related to the macro-factors; the model is retrained after each removal, continuing until no further improvement in performance is detected. This final step ensures that only the most impactful factors are retained, optimizing the model's efficiency and effectiveness.

Results

Model prediction results

Following extensive training, testing, and hyperparameter optimization, the NER model's predicted outcomes are summarized in Table 3. Notably, BioBERT achieved the highest AVG_MICRO and AVG_MACRO scores among all models. An analysis of Tables 3, 4, and 5 shows that despite BioBERT's overall superior performance, there is significant variability in the F1-Scores for different entity types across all three datasets. For instance, the F1-Scores for "B-Developing_anatomical_structure" (0.9275) and "I-Immaterial_anatomical_entity" (0.1538) demonstrate a pronounced disparity of 0.7737. This significant difference highlights substantial inconsistencies in the model's ability to accurately predict various entity types.

Table 3. Models' prediction accuracy comparison.

Model	Revised JNLPBA dataset		BC5CDR dataset		AnatEM dataset	
	AVG_MICRO	AVG_MACRO	AVG_MICRO	AVG_MACRO	AVG_MICRO	AVG_MACRO
BioBERT	0.932	0.9298	0.8726	0.858	0.8494	0.6975
RoBERTa	0.9133	0.9133	0.8313	0.8152	0.8201	0.6501
BigBird	0.9277	0.9218	0.8461	0.8321	0.8147	0.6451
DeBERTa	0.9256	0.921	0.8471	0.8335	0.806	0.6131

Table 4. The best prediction accuracy of BioBERT on the Revised JNLPBA and BC5CDR datasets.

Revised JNLPBA dataset		BC5CDR dataset	
Entity type	F1-Score	Entity type	F1-Score
B-cell_type	0.9276	B-Disease	0.861501
I-cell_type	0.9550	I-Disease	0.799018
B-protein	0.9292	B-Chemical	0.931006
I-protein	0.9230	I-Chemical	0.840546
B-RNA	0.9177	AVG_MICRO	0.872602
I-DNA	0.9323	AVG_MACRO	0.858018
B-cell_line	0.9318		
I-cell_line	0.9573		
B-DNA	0.9057		
I-RNA	0.9403		
AVG_MICRO	0.9320		
AVG_MACRO	0.9298		

Table 5. The best prediction accuracy of BioBERT on the AnatEM dataset.

AnatEM dataset	
Entity type	F1-Score

B-Tissue	0.749064
I-Tissue	0.617834
B-Immaterial_anatomical_entity	0.559585
I-Immaterial_anatomical_entity	0.153846
B-Developing_anatomical_structure	0.927536
I-Developing_anatomical_structure	0.4
B-Pathological_formation	0.711409
I-Pathological_formation	0.626374
B-Anatomical_system	0.646154
I-Anatomical_system	0.388889
B-Organism_subdivision	0.660633
I-Organism_subdivision	0.4
B-Cancer	0.915955
I-Cancer	0.894004
B-Organ	0.836975
I-Organ	0.677966
B-Cell	0.905293
I-Cell	0.913877
B-Multi-tissue_structure	0.761991
I-Multi-tissue_structure	0.740496
B-Organism_substance	0.851153
I-Organism_substance	0.740741
B-Cellular_component	0.850694
I-Cellular_component	0.809717
AVG_MICRO	0.849372
AVG_MACRO	0.697508

Model resource utilization results

In evaluating prediction accuracy among the BioBERT, BigBird, DeBERTa, and RoBERTa models, we documented their training time, CPU usage, and GPU memory consumption performance, as shown in Figure 4.

Training Time: DeBERTa consistently required the longest training times across all datasets, reaching a peak of 35.32 minutes in the Revised JNLPBA dataset. In stark contrast, BioBERT was notably more efficient, completing training in only 11.74 minutes in the same dataset and a mere 3.43 minutes in the AnatEM dataset.

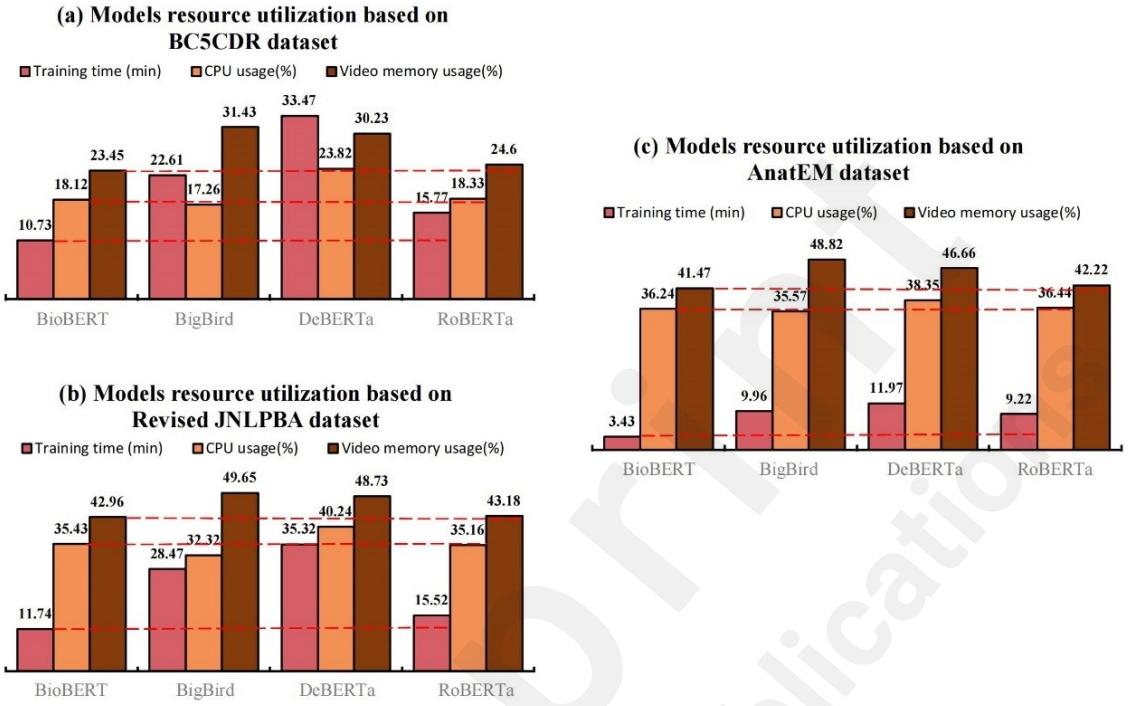
CPU Usage: DeBERTa recorded the highest CPU utilization at 40.24% in the Revised JNLPBA dataset. In contrast, BigBird demonstrated minimal CPU requirements, utilizing only 32.32% in the same dataset and significantly less, at 17.26%, in the BC5CDR dataset.

GPU Memory Consumption: BigBird exhibited the highest GPU usage, consuming 49.65% in the Revised JNLPBA dataset. Conversely, BioBERT showed the lowest GPU usage, consuming 42.96% in the same dataset and achieving a minimal usage rate of 23.45% in the BC5CDR dataset.

Overall, the higher resource requirements of DeBERTa (the highest CPU requirements) and BigBird (the highest GPU requirements) may reflect their greater computational intensity, which could translate to higher accuracy but at the expense of increased operational resources. Conversely,

BioBERT's low resource consumption suggests it is an efficient model that is well-suited for environments with strict resource constraints. RoBERTa displayed moderate resource utilization, positioning them between the extremes.

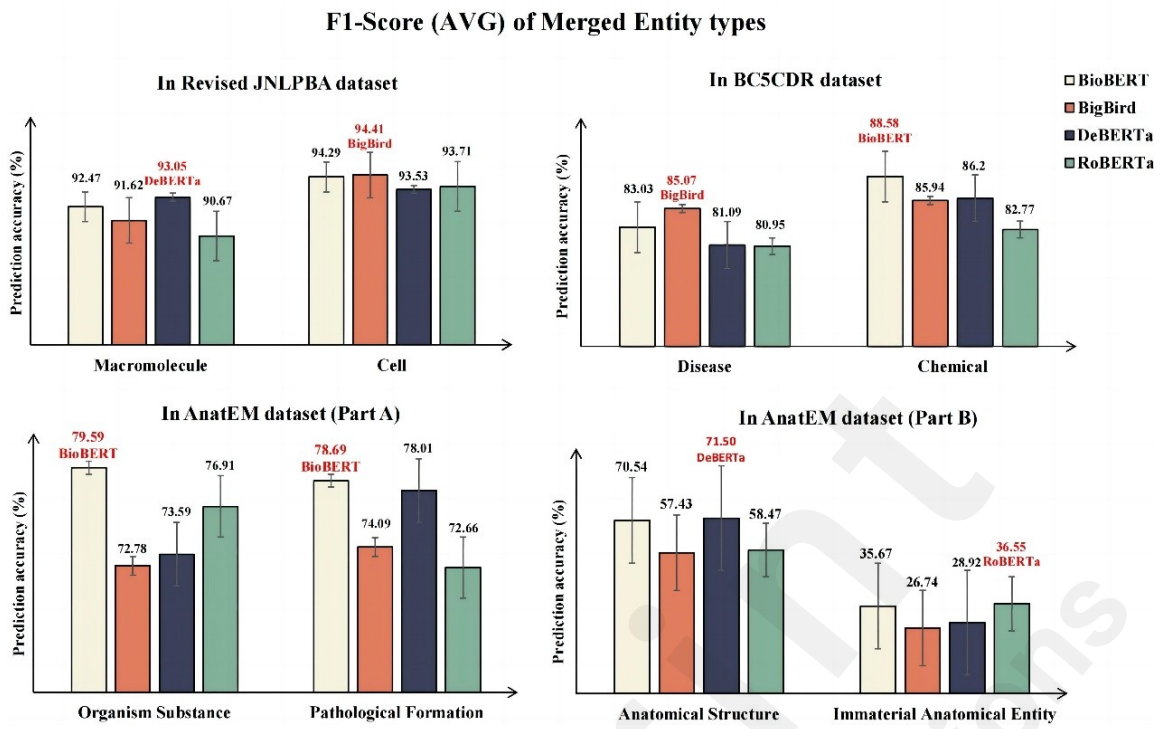
Figure 4. Training time, CPU and GPU usages of models



Entity type prediction accuracy results

Despite BioBERT achieving the highest overall prediction accuracy, it does not consistently excel across all entity types. By employing the categorical relaxation method to consolidate entity types, we observed distinct performance variations among models. As illustrated in Figure 5, BioBERT demonstrated superior prediction accuracy for the Chemical, Organism Substance, and Pathological Formation categories. Conversely, BigBird outperformed others in the Cell and Disease categories, DeBERTa led in the Macromolecule and Anatomical Structure categories, and RoBERTa excelled in predicting Immaterial Anatomical Entity types.

Figure 5. Relationship between models' prediction accuracy and merged entity types.



Macro-factor trends impacting model prediction accuracy

The results of this section primarily involved the trend relationships between the prediction accuracy of four NER models and six macro-factor metrics, focusing on the average values of these metrics across various datasets.

In the AnatEM dataset, as depicted in Figure 6, there is a distinct correlation: entity types with higher prediction accuracies correspond to increased values in *sLen*, *eLen*, *eNum*, *eDen,* and *tEWC*. This pattern indicates that the model's ability to accurately predict entities is enhanced with the rising complexity and volume of data associated with those entity types. Conversely, an inverse relationship is observed with *elCon*, which decreases as the other metrics increase. For instance, BioBERT recorded a high prediction accuracy of 90.96% in the "Cell" entity type. This outstanding performance correlated with the highest metrics observed in the dataset: *sLen* peaked at 216.78, *eLen* at 8.32, *eNum* at 1.94, and *eDen* at 0.03838. Then, the "Cell" entity type's total entity word count (*tEWC*) was notably high at 2436, demonstrating the model's ability to process extensive textual data effectively. However, *elCon* was significantly lower at 0.4642, the minimum average value compared to other types. This indicates that although BioBERT is adept at managing complex and voluminous data, it does not consistently ensure accurate entity labeling, suggesting a trend where higher accuracy and complexity metrics do not correspond with label consistency.

The Revised JNLPBA dataset substantiated these observations, as shown in Figure 7, the "cell_line" entity type consistently exhibited the highest accuracy across the four models, displaying significantly elevated values for *sLen*, *eLen*, *eNum*, and *eDen*, with average values recorded at 187.12, 7.76, 1.91, and 0.0372, respectively. This indicated that the model achieved high accuracy and effectively handled denser entity distributions. Moreover, there seemed to be a negative correlation between higher values of *elCon* and *tEWC* and these accuracies. In addition, a comparable trend emerged in the BC5CDR dataset where the "Chemical" entity type recorded the highest accuracy across RoBERTa, DeBERTa, and BioBERT models, aligning with the highest measurements of *sLen*, *eLen*, *eNum*, *eDen*, and *tEWC*, coupled with the lowest value of *elCon*.

Figure 6. Relationship between macro-factors and model's prediction accuracy in AnatEM dataset.

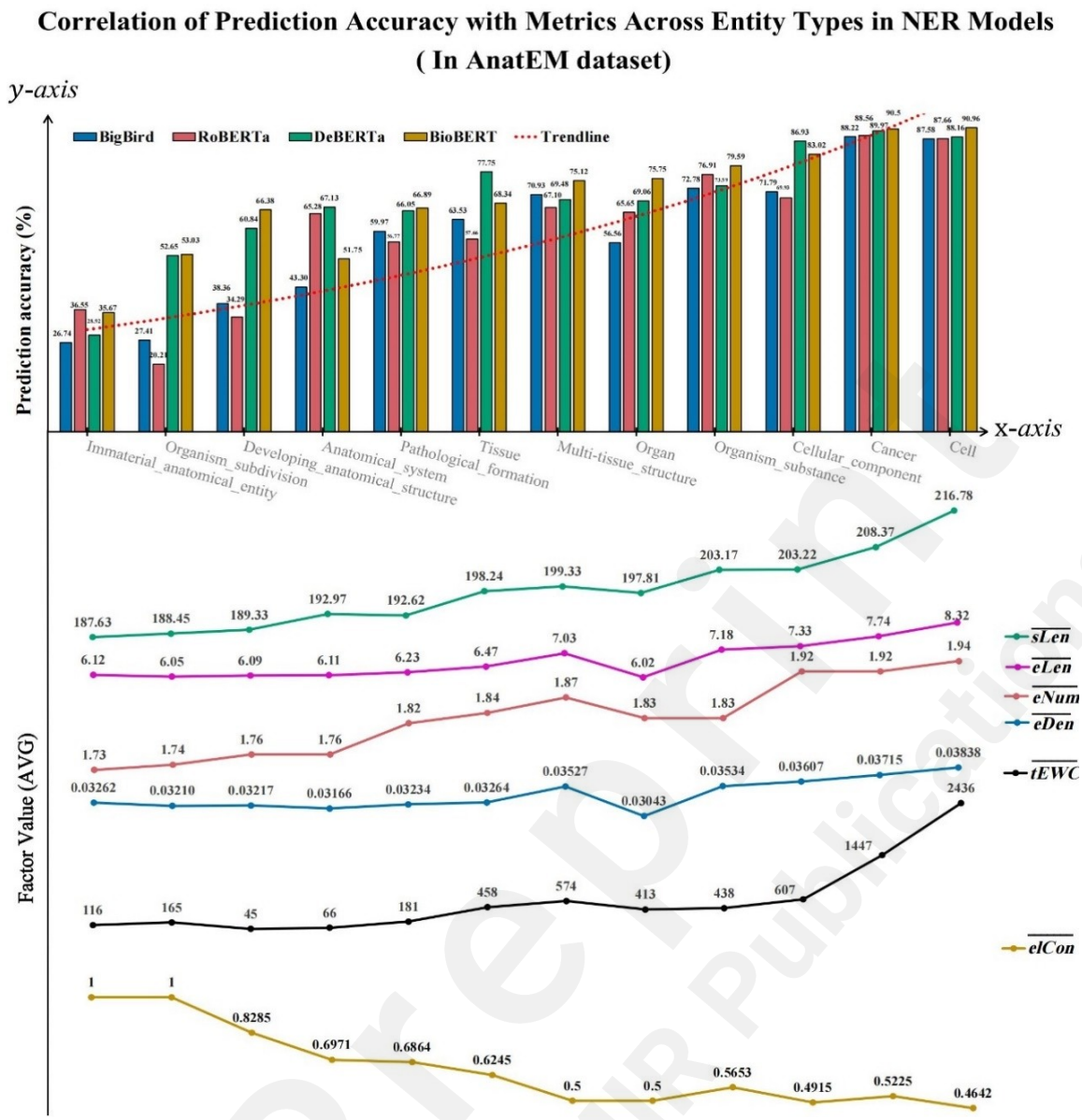
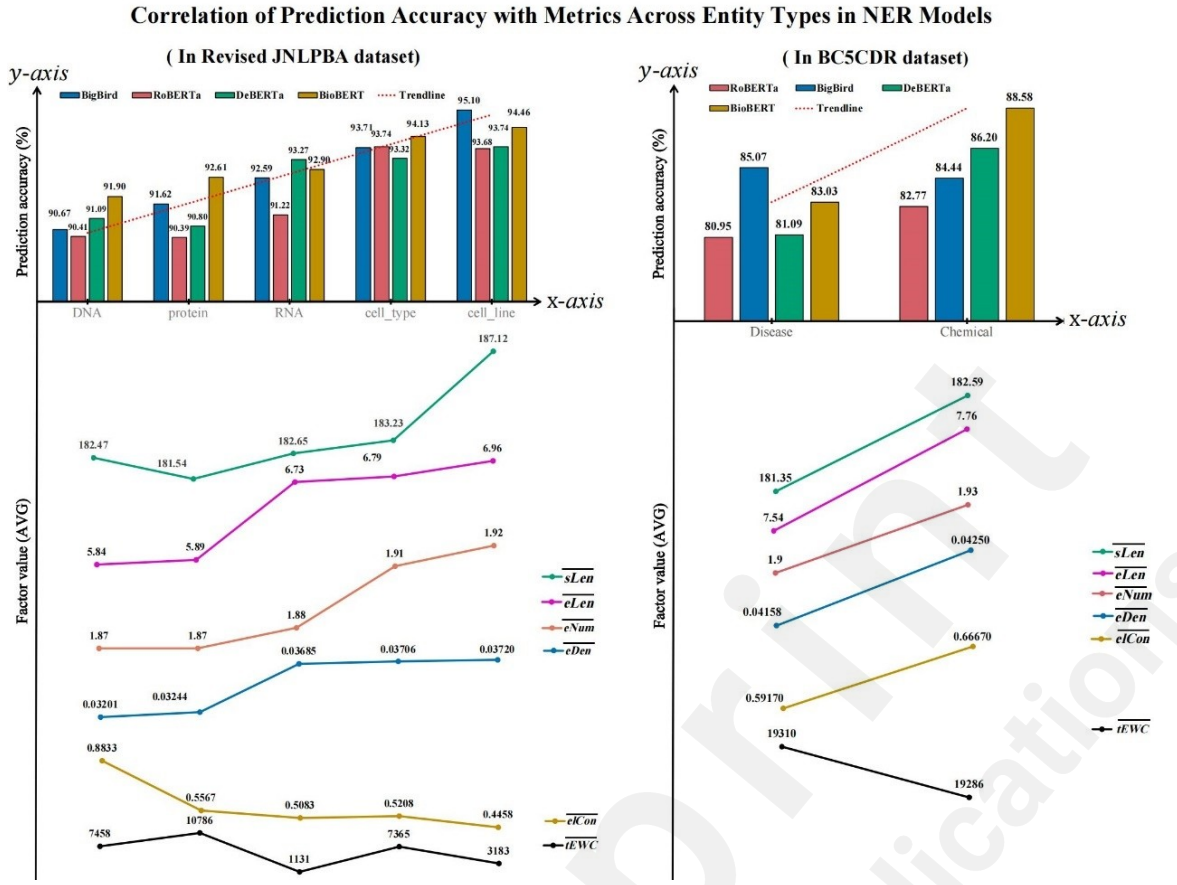


Figure 7. Relationship between macro-factors and model's prediction accuracy in Revised JNLPBA and BC5CDR datasets.



Macro-factors selection results

Through the hierarchical factor screening of the MFE algorithm, number of entity words in each entity phrase (*eNum*) was ultimately selected (see Table 6). Despite varying macro-factor combinations in the initial two layers, *eNum* was consistently chosen for the final layer. This indicates that *eNum* significantly influences the model's prediction accuracy among the six macro-factors.

Table 6. Selected macro-factors by MFE algorithm.

MFE algorithm	Based on Revised JNLPBA dataset	Based on BC5CDR dataset	Based on AnatEM dataset
Input	*sLen*, *eLen*, *eNum*, *eDen*, *elCon* and *tEWC*	*sLen*, *eLen*, *eNum*, *eDen*, *elCon* and *tEWC*	*sLen*, *eLen*, *eNum*, *eDen*, *elCon* and *tEWC*
Layer 1	*eLen*, *eNum*, *elCon*, *tEWC*	*eLen*, *eNum*, *eDen*, *tEWC*	*sLen*, *eLen*, *eNum*, *eDen*
Layer 2	*eLen*, *eNum*	*eNum*, *tEWC*	*eLen*, *eNum*
Layer 3	*eNum*	*eNum*	*eNum*

Discussion

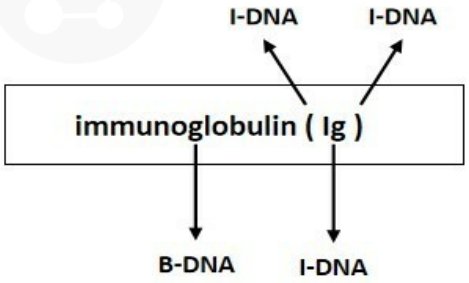
Impact of BioBERT training on medical NER performance

BioBERT's distinction in attaining the highest AVG_MICRO and AVG_MACRO scores among various models stems from key factors inherent in its design and training methodologies. As a derivative of BERT, BioBERT significantly benefits from extensive pre-training on broad general-domain corpora, followed by fine-tuning specialized medical texts, including PubMed abstracts and PMC full-text articles. This training strategy not only equips BioBERT with a nuanced understanding of complex medical terminology through layered contextual embeddings that capture bidirectional context but also optimizes its neural network specifically for medical applications. This domain-specific training substantially enhances its ability to identify and accurately predict medical entities more effectively than models trained on general datasets.

Despite its overall efficacy, BioBERT exhibits considerable variability in its prediction accuracy across different entity types, as evidenced by the marked disparities in F1-scores for entities such as "B-Developing_anatomical_structure" and "I-Immaterial_anatomical_entity". This variability indicates that while BioBERT is proficient at processing well-represented entity types, it encounters difficulties with those less represented or inherently more complex, often due to uneven training example distribution or the specific challenges of certain medical terminologies.

Further analysis of Tables 3, 4, and 5 confirms that despite achieving the best overall prediction results, BioBERT's performance is inconsistent across various entity types in all datasets. Two primary factors contribute to this inconsistency: firstly, uneven data distribution and noise in datasets like the Revised JNLPBA, BC5CDR, and AnatEM impede the model's ability to fully understand and accurately predict underrepresented entity types, affecting its overall performance. For example, the "protein" and "cell_type" entities in the Revised JNLPBA dataset are significantly more abundant (5256 and 2070 entity words, respectively) compared to "cell_line" and "RNA" (which have only 404 and 161 entity words, respectively). Secondly, multiple entity types for a single entity word complicate prediction. For instance, in the Revised JNLPBA dataset, certain symbols function as entity words within entity phrases. As depicted in Figure 8, the parentheses in the entity phrase "immunoglobulin (Ig)" from the training set are categorized as the "I-DNA" entity type. However, when predicting the entity types of parentheses in the test set, the model occasionally misclassifies them as "I-DNA" even though they do not belong to this entity type in some contexts. Specifically, the phrase "(PCBA)" in the test set is not recognized as an entity phrase. Yet, the model may erroneously assign the parentheses to the "I-DNA" entity type, leading to prediction errors.

Figure 8. Entity type division.



Evaluation of hardware performance across NER models

In evaluating hardware performance across the BioBERT, BigBird, DeBERTa, and RoBERTa

models, significant differences in training efficiency, CPU usage, and GPU memory consumption were observed, reflecting diverse computational demands and optimizations. DeBERTa demonstrated the highest computational intensity, requiring the longest training times and consuming the most CPU resources, peaking at 35.32 minutes and 40.24% CPU usage in the Revised JNLPBA dataset. This suggests that DeBERTa's architecture, possibly featuring more complex attention mechanisms, necessitates substantial computational power, impacting its deployment efficiency negatively in resource-sensitive environments. In contrast, BioBERT showcased remarkable efficiency, completing training significantly faster and with considerably lower GPU usage. It highlights its optimization for medical texts where streamlined processing and architecture adaptations reduce training time and resource consumption.

Furthermore, BigBird's high GPU usage, reaching 49.65% in the Revised JNLPBA dataset, indicates a model designed to leverage modern GPU's deep parallel processing capabilities, likely facilitating the handling of larger or more complex datasets efficiently. This contrasts with BioBERT, which maintains minimal GPU and CPU usage, underscoring its suitability for applications with strict resource constraints, such as mobile or embedded devices in clinical settings. The varied resource utilization across these models influences their operational applicability and suggests a trade-off between computational resource demands and potential accuracy. Models like DeBERTa and BigBird may offer higher accuracy due to their capability to process extensive data through more robust computational resources, albeit at greater operational costs.

The analysis of these NER models underscores the crucial balance between achieving high accuracy and maintaining resource efficiency. For researchers deploying NER tasks, selecting the appropriate model requires careful consideration of the specific operational context, including assessing available computational resources and weighing the importance of rapid deployment against high accuracy. This nuanced understanding aids in developing NER strategies that align with technical capabilities and strategic goals, ensuring that the selected model effectively supports the organization's broader objectives.

Variability in prediction accuracy across entity types in NER models

BioBERT's high overall prediction accuracy in medical NER tasks masks its variable performance across different entity types, highlighting the intricate relationship between model architecture, training data, and the distinct characteristics of entity categories. BioBERT excels at identifying entities such as Chemicals, Organism Substances, and Pathological Formations. This proficiency likely derives from the rich and clearly defined terminology in medical literature, significantly contributing to its training data. These categories, characterized by precise and comprehensive representations, bolster the model's ability to create robust contextual embeddings. However, BioBERT's suboptimal performance in other categories underscores the impact of entity complexity and data representation on outcomes.

In contrast, models like BigBird and DeBERTa demonstrate strengths in areas where BioBERT falters, suggesting that different architectural features might be more appropriate for certain entity types. For example, BigBird performs better in identifying Cell and Disease entities, a capability possibly linked to its ability to manage longer sequences effectively—crucial for capturing the complex interactions typical of these entities. DeBERTa excels in recognizing macromolecules and anatomical structures, attributes likely due to its sophisticated attention mechanisms that enhance interpreting complex syntactic structures in scientific texts. Similarly, RoBERTa's proficiency in predicting Immaterial Anatomical Entities can be attributed to its training with dynamically masked language models, facilitating a deeper understanding of abstract terminology common in such categories.

This analysis reveals the necessity for a refined approach to improving medical NER models, emphasizing that while extensive dataset training can improve overall accuracy, architectural adjustments and targeted training are essential to mitigate disparities in model performance across diverse entity types. Therefore, enhancing the accuracy and robustness of NER models involves increasing the variety and volume of training data and optimizing model architectures to address the specific challenges posed by less-represented or more intricate entity types. Strategic enhancements are crucial for advancing NER technology, especially in specialized domains like medicine, where accurate entity recognition is paramount.

Influence of macro-factors on prediction accuracy in medical NER

In analyzing how macro-factor metrics influence prediction accuracy across various medical entity types within four NER models, we observe significant nuances that reflect the complexity of modeling in medical NER tasks. Notably, entities characterized by higher values in *sLen*, *eLen*, *eNum*, and *eDen* generally exhibit better prediction accuracy. This correlation suggests that entities with more extensive and detailed textual representations tend to be predicted more accurately, highlighting the models' capacity to handle intricate data structures effectively. However, a notable exception arises with *elCon*, which inversely correlates with these metrics, indicating a potential trade-off between detailed data processing and consistent label accuracy.

This phenomenon is further complicated by the varying impact of the *tEWC* on prediction accuracy across different datasets. For example, in the Revised JNLPBA dataset, lower *tEWC* values are associated with higher accuracies, suggesting that models perform better with more concise entity representations. In contrast, other datasets show that higher *tEWC* values, indicative of richer contextual data, enhance model performance. This inconsistency underscores the complex influence of data characteristics on model effectiveness and suggests that the optimal balance of data quantity and quality varies significantly across datasets. Therefore, tailored model training and data preparation strategies are essential to optimize prediction accuracy according to the unique characteristics of each dataset.

These observations necessitate a strategic approach to model training and data preparation that considers the unique demands of each dataset. While handling more extensive data sets can lead to better entity recognition in some contexts, balancing this with the need for precision and consistency in entity labeling is crucial. As NER technologies evolve, it becomes imperative to refine model architectures and training methodologies to ensure that models can manage the dual challenges of complexity and volume without sacrificing accuracy. This tailored approach will be essential for advancing NER applications, particularly in specialized fields like medicine, where accurate and efficient data processing is critical for effective decision-making and research advancements.

Optimizing *eNum* handling to enhance NER model accuracy

Using the MFE algorithm for hierarchical marco-factor screening, *eNum* in each entity phrase has the greatest impact on the prediction accuracy of the NER models. Unlike broader textual metrics such as *sLen*, *eLen*, or *tEWC*, which provide general context, *eNum* directly measures the entities' complexity. This factor significantly influences how models process and interpret dense information, with a higher *eNum* generally indicating semantically rich entities are potentially more challenging to analyze. The consistent *eNum* selection in the final layer of the screening process across various datasets underscores its pivotal role in enhancing the precision of entity recognition.

The impact of *eNum* on model accuracy suggests that entities with a higher density of words require sophisticated model capabilities to discern and categorize the detailed information they contain accurately. This necessitates models that can manage the sheer volume of data and effectively interpret each entity's intricate relationships and contextual nuances.

Therefore, refining the models' abilities to analyze entities with higher *eNum* values could thus be a strategic approach for advancing NER technologies, particularly in domains like medicine, where precise and reliable entity recognition is crucial. Enhancing how models manage and utilize the detailed information encapsulated in *eNum* will improve the robustness and effectiveness of medical NER models, ensuring they meet the complex demands of varied and extensive datasets.

Conclusion

Medical NER is a crucial component of medical informatics, essential for identifying and categorizing named entities within unstructured medical text data. A proficient NER model significantly enhances various downstream applications, such as medical text classification, question answering, and information retrieval. Developing a high-performance NER model requires a meticulous approach that includes selecting relevant macro-factors, designing the model's architecture, and curating specialized training data tailored to medical contexts.

Our evaluation method for NER models extends beyond general metrics like accuracy, recall, and F1-score by incorporating an extensive analysis of macro-factors relevant to medical entities. This comprehensive approach enables a multi-dimensional evaluation of the models, providing insights into how different entity types, attributes, and contextual factors influence performance. For example, our findings indicate that while "Disease" frequently occurs in medical texts and requires high accuracy, entities like "Immaterial anatomical entity" may not require the same precision. This discrepancy highlights the need for targeted optimization strategies for different entity types, crucial for advancing medical NER models.

Additionally, our study explores the characteristics of entities to improve the creation of high-quality medical NER datasets and documents. This focus enhances the NER models' ability to identify entities accurately and addresses the specific needs of medical texts. While our analysis extensively covers macro-factors and their impact, it does not delve into misclassifications of entity labels or the fine-grained interactions between entity words. These areas could further refine our understanding of model accuracy. Moreover, examining hardware performance illuminates these models' internal efficiency and resource utilization, which is crucial for their deployment in real-world scenarios.

In conclusion, evaluating medical NER models in this study is essential for developing effective and precise NLP applications in healthcare. It gives medical researchers the insights to select and refine NER models suited to various medical scenarios, ultimately improving these systems' accuracy, robustness, and reliability. This foundational work sets the stage for future research that could explore the intricate relationships within NER systems, further enhancing the capabilities of medical informatics.

Conflicts of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval

Ethical approval was obtained from the Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College (Ethical Approval Code: IMICAMS/01/22/HREC). After getting ethical approval, the Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College wrote an official Ethics Approval Statement.

All medical NER datasets used in this article are public datasets; no personal or sensitive information has been collected, and they complied with local institutional guidelines and legislation. It is unnecessary to provide written or verbal informed consent from the participants. The experimental protocols and datasets in this study were approved by the Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College. All methods were performed according to the relevant guidelines and regulations.

Funding

This work was funded by Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences Program (Grant No. 2021-I2M-1-057).

References

1. Li, J., Wei, Q., Ghiasvand, O., Chen, M., Lobanov, V., Weng, C., & Xu, H. (2022). A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Medical Informatics and Decision Making*, *22*(3), 1-10.

2. Gridach, M. (2017). Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics*, *70*, 85-91.

3. Tianjiao Yang, Ying He, Ning Yang, "Named Entity Recognition of Medical Text Based on the Deep Neural Network", Journal of Healthcare Engineering, vol. 2022, Article ID 3990563, 10 pages, 2022. *https://doi.org/10.1155/2022/3990563.*

4. Kundeti, S. R., Vijayananda, J., Mujjiga, S., & Kalyan, M. (2016, December). Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 1937-1945). IEEE.

5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

6. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692.

8. Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020). Big bird: Transformers for longer sequences. *arXiv preprint* arXiv:2007.14062.

9. He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint* arXiv:2006.03654.

10. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.

11. Deng, W., Ni, H., Liu, Y., Chen, H., & Zhao, H. (2022). An adaptive differential evolution algorithm based on belief space and generalized opposition-based learning for resource allocation. *Applied Soft Computing*, *127*, 109419.
12. Tian, S., Jin, Q., Yeganova, L., Lai, P. T., Zhu, Q., Chen, X., ... & Lu, Z. (2024). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1), bbad493.
13. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint* arXiv:2303.18223.
14. Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., ... & Xu, H. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, ocad259.

15. Kalyan, K. S. (2023). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 100048.
16. Ahmad, P. N., Shah, A. M., & Lee, K. (2023, April). A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. In *Healthcare* (Vol. 11, No. 9, p. 1268). MDPI.
17. Freund, F., Tamla, P., & Hemmje, M. (2023, December). Towards Improving Clinical Practice Guidelines Through Named Entity Recognition: Model Development and Evaluation. *In 2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)* (pp. 1-8). IEEE.
18. Yoon, W., So, C. H., Lee, J., & Kang, J. (2019). Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics*, 20(10), 55-65.
19. Yu, H., Mao, X. L., Chi, Z., Wei, W., & Huang, H. (2020, August). A Robust and Domain-Adaptive Approach for Low-Resource Named Entity Recognition. In *2020 IEEE International Conference on Knowledge Graph (ICKG)* (pp. 297-304). IEEE.
20. Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.
21. Erdmann, A., Wrisley, D. J., Brown, C., Cohen-Bodénès, S., Elsner, M., Feng, Y., ... & de Marneffe, M. C. (2019). Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. Association for Computational Linguistics.
22. Usha, M. S., Smrity, A. M., & Das, S. (2022, December). Named Entity Recognition Using Transfer Learning with the Fusion of Pre-trained SciBERT Language Model and Bi-directional Long Short Term Memory. In *2022 25th International Conference on Computer and Information Technology (ICCIT)* (pp. 460-465). IEEE.
23. Nagaraj, P., Dass, M. V., Mahender, E., & Kumar, K. R. (2022, December). Breast Cancer Risk Detection using XGB Classification Machine Learning Technique. In *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)* (pp. 1-5). IEEE.
24. Ozcelik, O., & Toraman, C. (2022). Named entity recognition in Turkish: A comparative study with detailed error analysis. *Information Processing & Management*, *59*(6), 103065.
25. Akhtyamova, L. (2020, April). Named entity recognition in Spanish biomedical literature: Short review and BERT model. In *2020 26th Conference of Open Innovations Association (FRUCT)* (pp. 1-7). IEEE.
26. Fu, Jinlan, Pengfei Liu, and Graham Neubig. “Interpretable Multi-dataset Evaluation for Named Entity Recognition.” *arXiv preprint*. arXiv:2011.06854 (2020).
27. Zhou, X., Ma, H., Gu, J., Chen, H., & Deng, W. (2022). Parameter adaptation-based ant colony optimization with dynamic hybrid mechanism. *Engineering Applications of Artificial Intelligence*, *114*, 105139
28. Yao, R., Guo, C., Deng, W., & Zhao, H. (2022). A novel mathematical morphology spectrum entropy based on scale-adaptive techniques. *ISA transactions*, *126*, 691-702.
29. Huang, M. S., Lai, P. T., Lin, P. Y., You, Y. T., Tsai, R. T. H., & Hsu, W. L. (2020). Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*, 21(6), 2219-2238.
30. Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., ... & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.
31. Pyysalo, S., & Ananiadou, S. (2014). Anatomical entity mention recognition at literature scale. Bioinformatics, 30(6), 868-875.

32. Zhou, Y., Cahya, S., Combs, S. A., Nicolaou, C. A., Wang, J., Desai, P. V., & Shen, J. (2018). Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets. *Journal of chemical information and modeling*, *59*(3), 1005-1016.

33. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

34. Nakamura, K., Derbel, B., Won, K. J., & Hong, B. W. (2021). Learning-rate annealing methods for deep neural networks. *Electronics*, *10*(16), 2029.

35. Martin, C. H., & Mahoney, M. W. (2021). Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, *22*(165), 1-73.

36. Marchisio, A., Massa, A., Mrazek, V., Bussolino, B., Martina, M., & Shafique, M. (2020, November). NASCaps: A framework for neural architecture search to optimize the accuracy and hardware efficiency of convolutional capsule networks. In *Proceedings of the 39th International Conference on Computer-Aided Design* (pp. 1-9).

37. Jabir, B., & Falih, N. (2021). Dropout, a basic and effective regularization method for a deep learning model: a case study. *Indonesian Journal of Electrical Engineering and Computer Science*, *24*(2), 1009-1016.

38. Tsai, R. T. H., Wu, S. H., Chou, W. C., Lin, Y. C., He, D., Hsiang, J., ... & Hsu, W. L. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC bioinformatics*, *7*, 1-8.

39. Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics*, 19(1), 1-6.

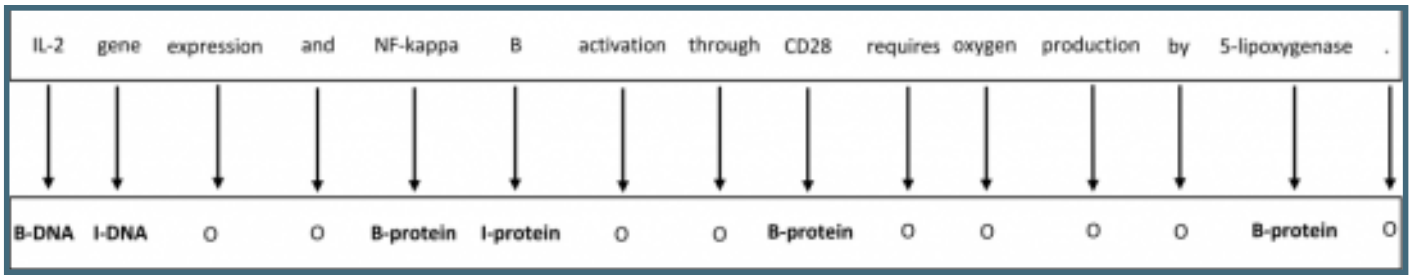
40. Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. *In Noise reduction in speech processing* (pp. 1-4). Springer, Berlin, Heidelberg.

41. Zhou, J. Y., Song, L. W., Yuan, R., Lu, X. P., & Wang, G. Q. (2021). Prediction of hepatic inflammation in chronic hepatitis B patients with a random forest-backward feature elimination algorithm. *World Journal of Gastroenterology*, 27(21), 2910-2920.

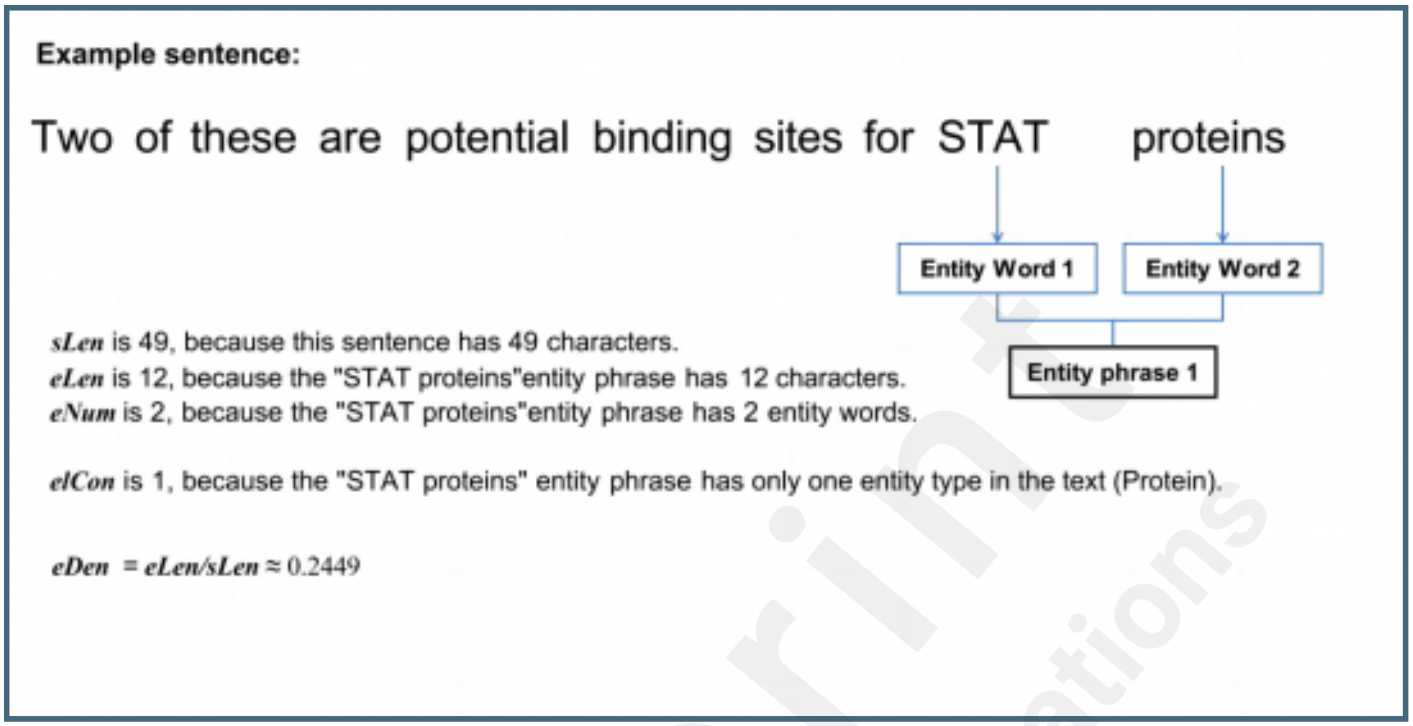
Supplementary Files

Figures

An instance of the BIO sequence annotation method.



sLen, eLen, eNum, eDen and elCon values of an entity word.

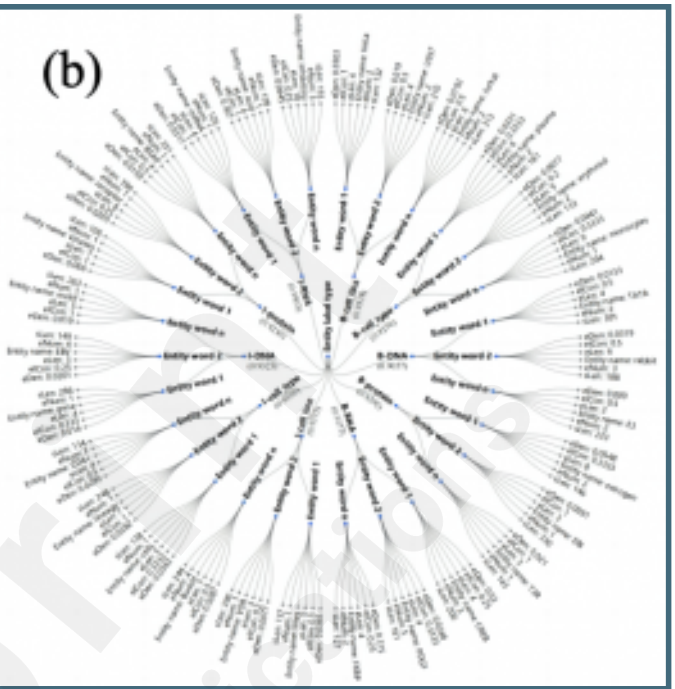


Consists of two parts illustrating different components of each macro-factors' dataset. Panel (a) displays the labels. Panel (b) presents a visualization of sLen, eLen, eNum, eDen, and elCon macro-factor metrics using the Radial Tree layout algorithm, offering a structured view of these metrics' interrelationships.

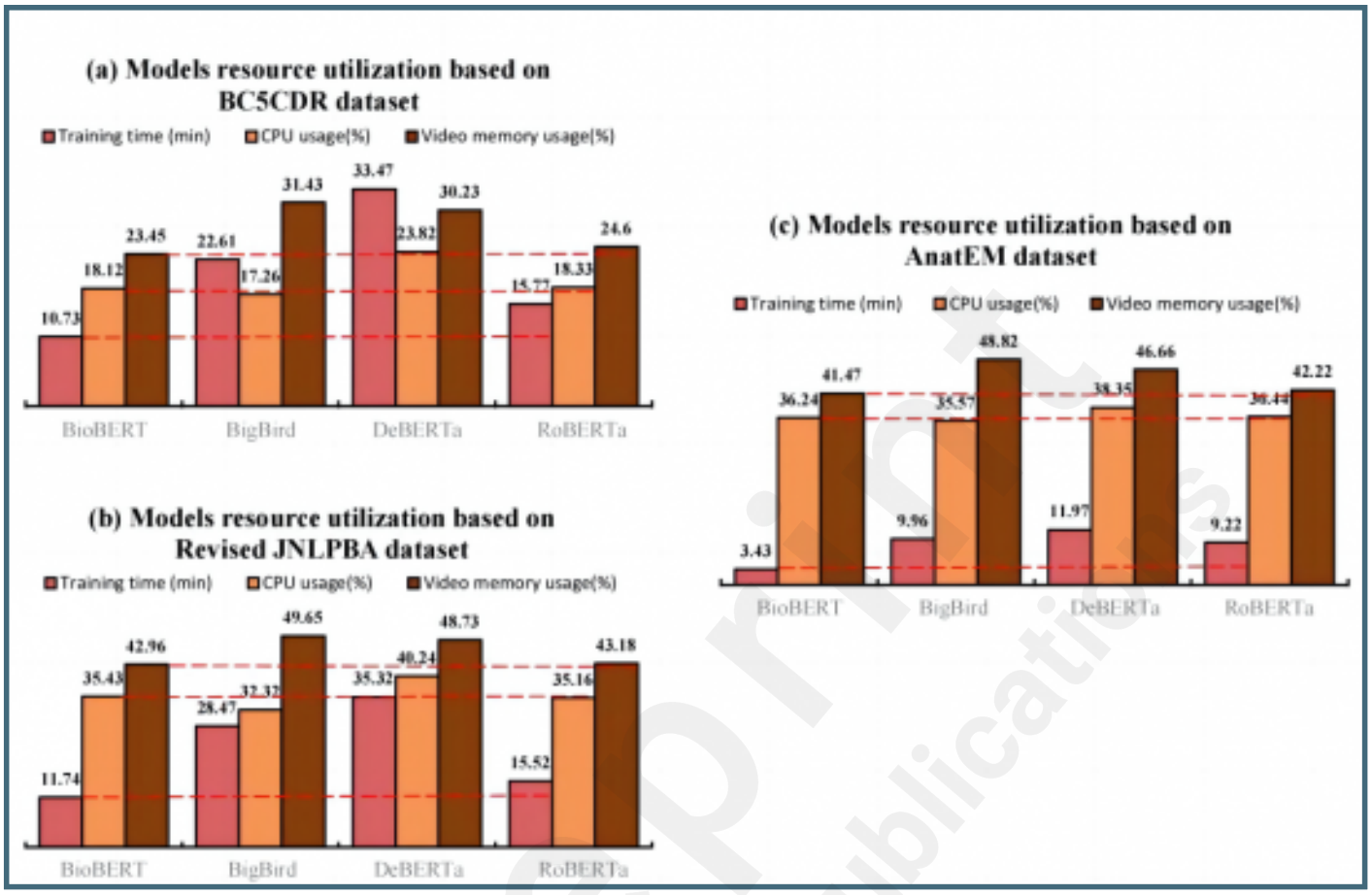
(a)

	Level	F1-Score
0	B-cell_type	0.9276
1	I-cell_type	0.955
2	B-protein	0.9292
3	I-protein	0.923
4	B-RNA	0.9177
5	I-DNA	0.9323
6	B-cell_line	0.9318
7	I-cell_line	0.9573
8	B-DNA	0.9057
9	I-RNA	0.9403
	AVG_MICRO	**0.932**
	AVG_MACRO	**0.9298**

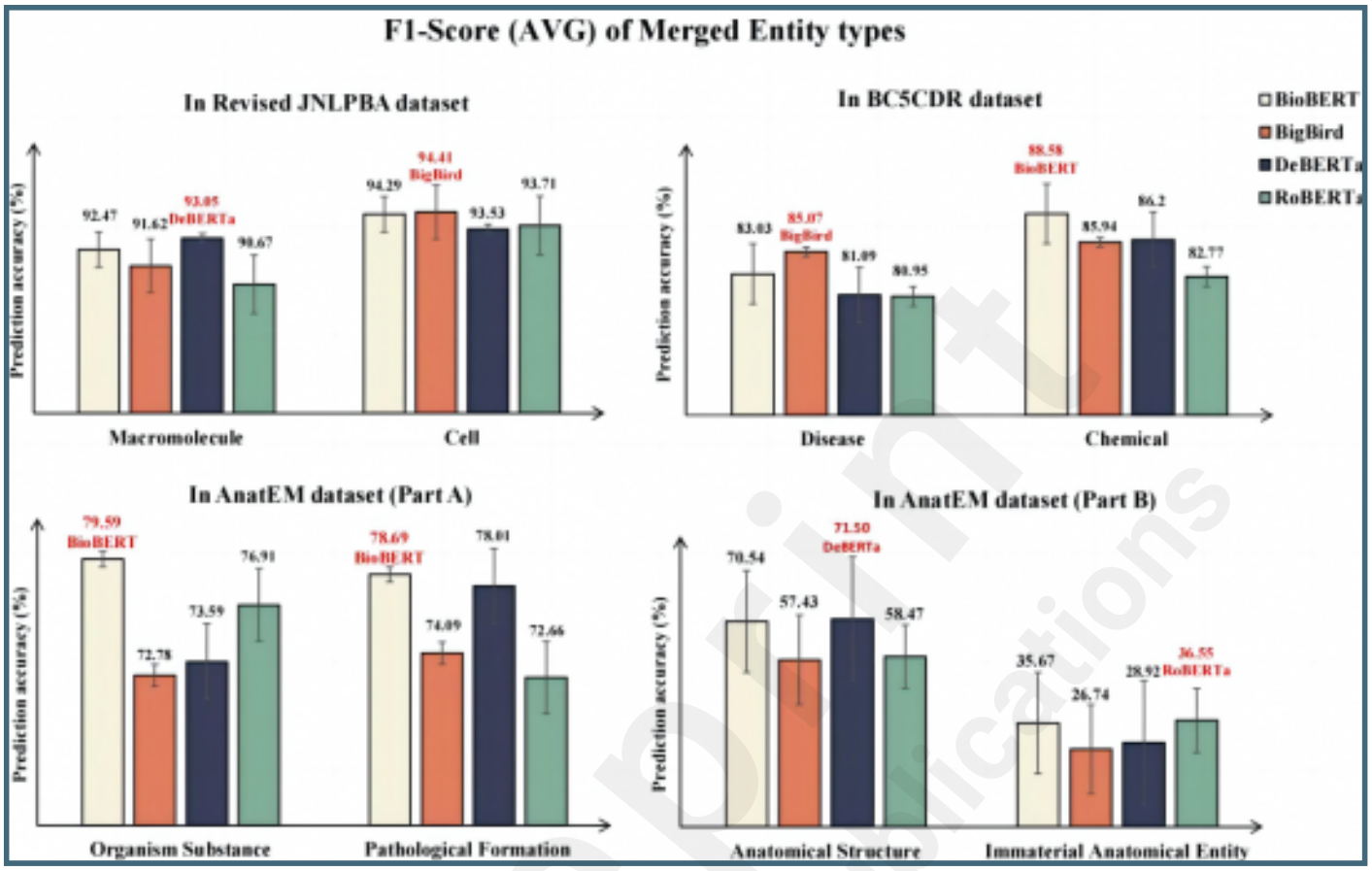
(b)



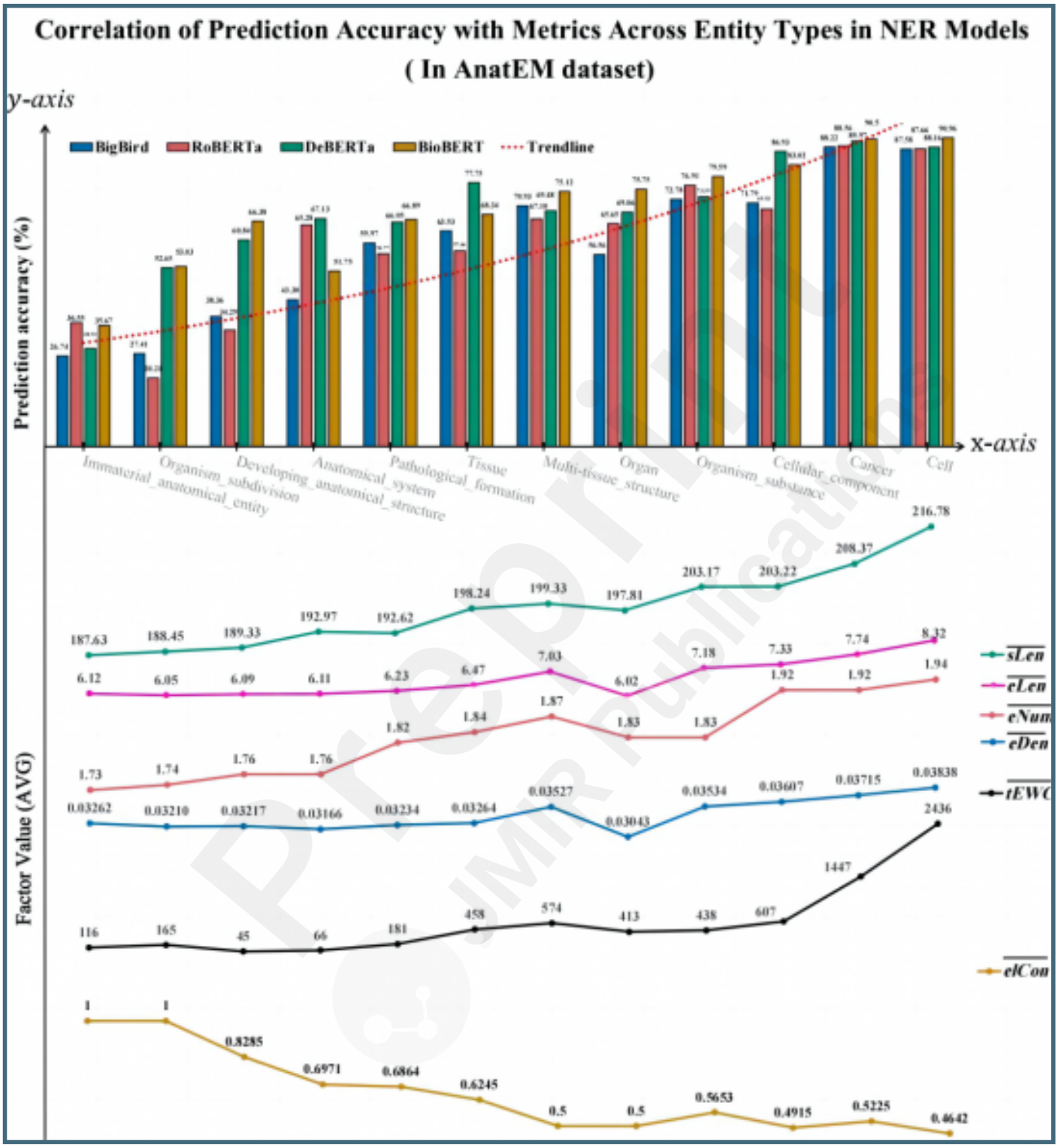
Training time, CPU and GPU usages of models.



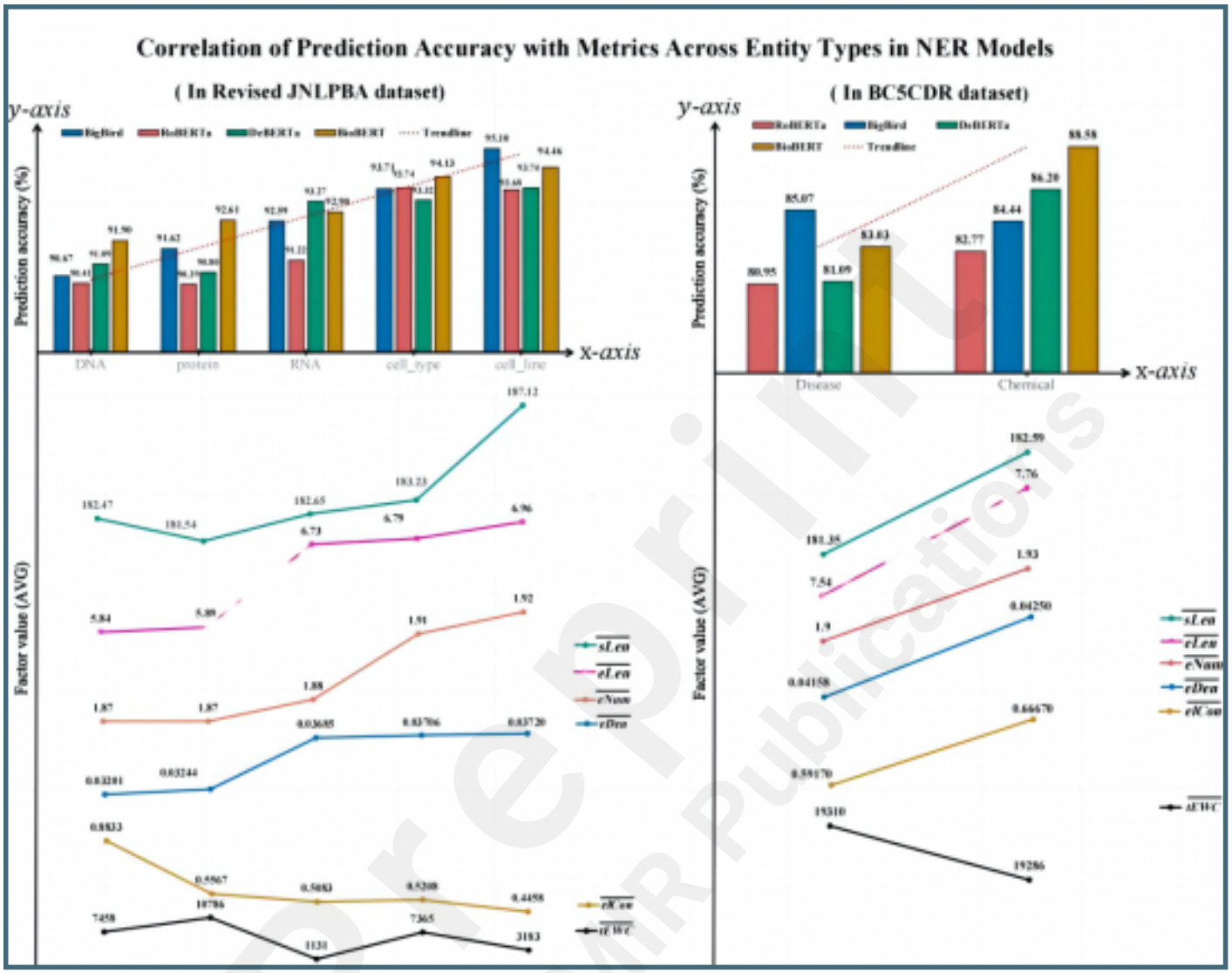
Relationship between models' prediction accuracy and merged entity types.



Relationship between macro-factors and model's prediction accuracy in AnatEM dataset.



Relationship between macro-factors and model's prediction accuracy in Revised JNLPBA and BC5CDR datasets.



Entity type division.