analytical chemistry

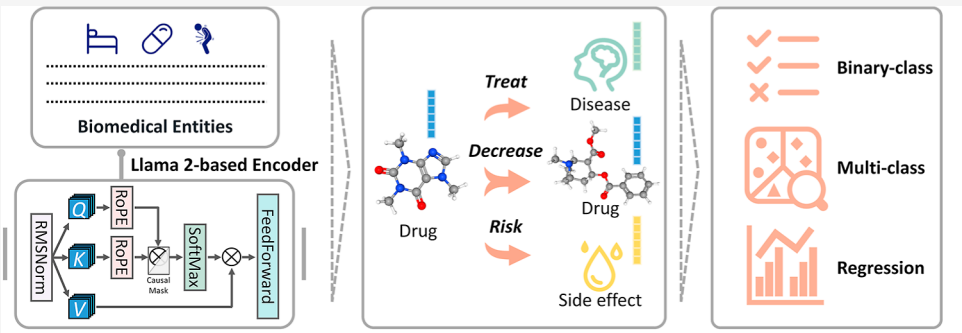
pubs.acs.org/ac Article

Large Language Model-Based Natural Language Encoding Could Be All You Need for Drug Biomedical Association Prediction

Hanyu Zhang, Yuan Zhou, Zhichao Zhang, Huaicheng Sun, Ziqi Pan, Minjie Mou, Wei Zhang, Qing Ye, Tingjun Hou, Honglin Li,* Chang-Yu Hsieh,* and Feng Zhu*

Cite This: *Anal. Chem.* 2024, 96, 12395−12403 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Analyzing drug-related interactions in the field of biomedicine has been a critical aspect of drug discovery and development. While various artificial intelligence (AI)-based tools have been proposed to analyze drug biomedical associations (DBAs), their feature encoding did not adequately account for crucial biomedical functions and semantic concepts, thereby still hindering their progress. Since the advent of ChatGPT by OpenAI in 2022, large language models (LLMs) have demonstrated rapid growth and significant success across various applications. Herein, LEDAP was introduced, which uniquely leveraged LLM-based biotext feature encoding for predicting drug-disease associations, drug−drug interactions, and drug-side effect associations. Benefiting from the large-scale knowledgebase pre-training, LLMs had great potential in drug development analysis owing to their holistic understanding of natural language and human topics. LEDAP illustrated its notable competitiveness in comparison with other popular DBA analysis tools. Specifically, even in simple conjunction with classical machine learning methods, LLM-based feature representations consistently enabled satisfactory performance across diverse DBA tasks like binary classification, multiclass classification, and regression. Our findings underpinned the considerable potential of LLMs in drug development research, indicating a catalyst for further progress in related fields.

INTRODUCTION

With the advent of ChatGPT by OpenAI in 2022, large language models (LLMs) exemplified by generative pre-trained transformer have witnessed explosive growth in multiple fields, including medical research.[1] These powerful models have demonstrated remarkable versatility, with numerous investigations highlighting their widespread applicability across biomedical domains.[2] For example, LLMs have shown promise in developing effective treatments through drug virtual screening and de novo molecular generation,[3] aiding protein design and engineering by generating functional protein sequences across diverse families,[4] as well as enhancing biomedical question answering by distilling knowledge from vast biomedical corpora.[5] In addition to traditional natural language models such as BioBERT,[6] a burgeoning field of LLMs tailored for biomedical language processing has emerged, encompassing the simplified molecular input line entry system (SMILES) language model for molecules,[7] the protein language model for amino acid sequences,[8] and the gene language model for single-cell transcriptomic data.[9] All these LLM-based approaches have held immense potential for revolutionizing the modeling and representation of intricate bioentities.[10]

Analyzing biomedical interactions has been a crucial part of drug discovery and development.[11] Concurrently, the understanding of drug-related associations has been continually refined through the accumulation of bioassays and screening

Received: April 6, 2024
Revised: June 1, 2024
Accepted: July 2, 2024
Published: July 16, 2024

results.[12] By leveraging the latest insights, researchers have endeavored to unravel the multifaceted roles that drugs play in their associations with diverse biomedical entities.[13] These efforts have sparked human knowledge accumulation in a range of fields,[14−17] such as drug design,[18] disease treatment,[19] and biochemical mechanism study.[20] In this pursuit, numerous artificial intelligence (AI)-based methods have been proposed to analyze drug biomedical associations (DBAs) such as drug repurposing, drug interactions, and adverse drug reactions. Notably, knowledge graph (KG)-based machine learning (ML) methods have been widely applied and have shown promising results in DBA analysis.[12,21,22] Additionally, plenty of intrinsic characteristic-based deep learning (DL) methods have also achieved great success by utilizing well-designed feature engineering to generate representations for model input, which are typically related to molecular characteristics, structural information, physicochemical properties, etc.[23,24]

However, the existing DBA analysis tools were still hampered by their representation methods for biomedical entities.[25] For pharmacological terminologies such as diseases, side effects, etc., their representations were predominantly premised on expert-curated structurized data.[26] This was typically done by constructing KGs for the high-dimensional graph embedding[21,27] and feature encoding via semantic similarity profiles computed from ontologies.[12,24] However, due to the complexity of the graph embedding algorithms, these methods were often limited in terms of information extraction effectiveness.[28,29] Meanwhile, feature representation strategies based on similarity spectra might result in a loss of information.[30,31] As for biomedical molecules such as drugs, their feature representations largely relied on intrinsic molecular constitution and characteristics.[32,33] Classic molecular fingerprints,[34] physicochemical descriptors,[35] biochemical language-based embeddings,[7,36] etc. were commonly applied to represent such substances. However, these features focused solely on physiochemical and structural properties, failing to fully utilize the vast amount of biomedical semantic information and functional implications that were presented within the human knowledgebase texts in molecular biology, medicine, and pharmacology fields.[37,38] In general, biomedical literature, among other knowledge sources, could offer extensive, multifaced textual descriptions, providing linguistic context absent in isolated representation.[39] Given their neglect of such ample linguistic data, traditional tools were hindered by generating representations involving a holistic understanding of biomedical functions, especially in the analysis of DBAs.[40]

In this study, LEDAP was proposed to apply LLM-based biotext encoding for DBA prediction. Owing to large-scale knowledge pre-training, LLMs demonstrated formidable comprehension of natural language and exhibited proficiency across a broad spectrum of human topics.[41] Therefore, the emergent development of LLMs allowed researchers to analyze biotexts taking natural language as carriers, thereby compensating for conventional inadequacies and generating human knowledge-integrated representations that contributed to DBA analysis.[42] Within the scope of this study, LEDAP represented biomedical entities solely via a pre-trained attention-based LLM Llama 2, and used classical ML methods for the DBA prediction, including drug-disease association (DDA), drug−drug interaction (DDI), and drug-side effect association (DSA). Experimental results indicated that even when merely cooperating with classical ML methods, LEDAP with LLM-based feature representations consistently facilitated the achievement of more competitive performance compared with those existing popular tools in various DBA task forms, involving binary classification, multiclass classification, and regression. These insights illustrated the considerable potential and application prospects of LLMs in drug development analysis, serving to catalyze further evolution in related fields. All source code and datasets are available at https://github.com/idrblab/LEDAP.

MATERIALS AND METHODS

Data Collection and Biofunctional Description Preprocess. In this study, three tasks, respectively, corresponding to DDA, DDI, and DSA, were prepared for experiments to validate the effectiveness of our proposed strategy in DBA analysis. Specifically, in the strategy, an indispensable step was to employ LLMs to conduct feature representation for biomedical entities with their functional descriptions. The background of the drugs was mainly collected from DrugBank v5.1.12 (https://go.drugbank.com/), an extensively referenced knowledgebase that encompasses comprehensive drug information, including chemical, pharmacological, and pharmaceutical data.[43] For the minority of drugs without expert-curated descriptions or DrugBank IDs, their descriptions were extracted from NCATS Inxight Drugs (https://drugs.ncats.io/), which incorporates and unifies a wealth of data, including manually curated data supplied by the FDA and private companies.[44] Regarding diseases, the Disease Ontology Knowledgebase (DO-KB, https://disease-ontology.org/) was adopted to map the diseases to their DO IDs as well as their corresponding definitions, encompassing disease mechanisms, etiology knowledge, symptoms, etc.[45] As for side effects, according to their Unified Medical Language System Concept Unique Identifiers, information was standardized based on SIDER 4.1 (http://sideeffects.embl.de/), a resource that contains marketed medicines and their recorded adverse drug reactions.[46]

Before being input for encoding, the collected initial biotexts should undergo preprocessing, including the removal of invalid punctuation and substitution of genuine names with the code name of 'this entity', segmentation, and tokenization. Taking 'ascorbic acid' as an example, all these processes are illustrated in Supporting Information, Figure S1.

Natural Language Biotext Encoding Based on Llama 2. LLM families were deep neural network systems with billions of parameters pre-trained on extensive natural language text corpora, incorporating massive online content, documents, literature, and other expert-collected materials.[47] The large-scale pre-training process endowed their human language topics comprehension and proficiency, in tandem with a capacity to resolve intricate tasks.[41] Such attributes paved the way for computational method construction in the biomedical field so that researchers could use LLMs to generate human knowledge-integrated representations based on biomedical semantic concepts.[42]

This study employed one of the most popular open-source LLMs, Llama 2, to transform the biotext of the biomedical entities into high-dimensional feature vectors. Llama 2 is a collection of pre-trained and fine-tuned LLMs developed and released by Meta,[48] including variants with 7 billion (Llama 2-7B), 13 billion (Llama 2-13B), and 70 billion (Llama 2-70B) parameters. It is an autoregressive model that uses an optimized Transformer architecture, representatively including RMSNorm,[49] SwiGLU activation function,[50] and rotary

https://doi.org/10.1021/acs.analchem.4c01793
Anal. Chem. 2024, 96, 12395−12403

positional embeddings.[51] Unlike the renowned ChatGPT series models, Llama 2 is fully open-source, making it accessible for safe and fair use under its use policy.[48] Researchers have the convenience of deploying and modifying Llama 2 locally according to various task-specific requirements. Llama 2 has demonstrated impressive performance in language-related tasks, and this new thing has been embraced with immense enthusiasm to tackle chemical and biological challenges.[52−54]

In practice, the biotext encoder was derived from the 7B model, followed by an average pooling layer enabling unified representation.[55] The pre-trained parameter weights used in this study were prepared according to the released tutorial (https://github.com/Meta-Llama/llama).

Model Employment for the Drug Association Prediction. Feature vectors obtained by biotext encoding were then concatenated and used as input for DBA predictors. Classical ML methods, typically including eXtreme Gradient Boosting (XGBoost) and Random Forest (RF), were employed to directly apply the LLM-based feature representations to conduct the DBA modeling and prediction. Using classical models could better reflect the role feature representation plays in the whole process. LEDAP adopted a suitable ML method for each specific task and optimized model hyperparameters using optuna.

XGBoost is a highly optimized distributed gradient boosting algorithm that constructs a strong learner by sequentially combining multiple weak learners (primarily decision tree models) using the ensemble learning strategy. With each step in the sequence, the model constructs a new weak learner based on the prediction residuals of the previous model to minimize the overall prediction loss.[12] Furthermore, XGBoost explicitly incorporates a regularization term during the model training to effectively prevent overfit, thereby enhancing the prediction accuracy while ensuring the model's generalization capability.[56]

RF is an algorithm that implements the bagging strategy, considering the prediction results from multiple decision trees to make the final decision. During its training phase, it applies bootstrap sampling to gather training samples to construct a decision tree and make splits based on a randomly selected subset of features. This process is repeated to generate a large, 'forest'-like collection of decision trees.[57] For the final prediction, the RF algorithm employs a voting system for classification tasks and averages the prediction results of all decision trees for regression tasks.[58,59]

Experimental Setup, Evaluation, and Implementation. For each experiment, to facilitate comparison, the prepared benchmark dataset remained consistent with the contrastive study. To be specific, Task-1 acted in accordance with DREAMwalk[12] to adopt equal-sized negative sampling for DDA prediction, and all samples were randomly split for 10-fold cross-validation. For DDI prediction, the same training, development, and testing set with SumGNN[22] was adopted for Task-2. As for Task-3 of DSA prediction, following NRFSE[23] and SDPred,[24] all known drug-side effect frequency entries[60] and the remaining unknown entries were randomly divided into 10-folds. Each fold was set as the test set in turn, while the remaining 9-folds were set as the train set.

Specific evaluation metrics were used to measure method performance in different forms of DBA tasks. In Task-1, accuracy (ACC), area under the receiver operating characteristic curve and precision-recall curve (AUROC and AUPR) were adopted for binary classification.[12] In Task-2, ACC, average F1 score over different classes (Macro-F1), and Cohen's Kappa (Kappa) were adopted for multiclass classification[22]

$$\text{Accuracy} = \frac{|Y \cap \hat{Y}|}{|Y|}$$

$$\text{Macro } F1 = \frac{1}{N}\sum_{n=1}^{N}\frac{2P_n \bullet R_n}{P_n + R_n}$$

$$\text{Cohen's Kappa} = \frac{p_o - p_e}{1 - p_e}$$

where Y was the predicted labels, and $\hat{Y}$ was the ground-truth labels. N was the number of classes, and P_n, R_n were the precision and recall for the n th class. p_o was the observed agreement (identical to accuracy), and p_e was the probability of randomly seeing each class.

In Task-3, method performances were, respectively, assessed based on identifying DSAs and estimating their frequency values.[23] On the one hand, AUROC and AUPR were adopted for association identification, while known frequency items were regarded as positive and unknown candidates as negative. On the other hand, root mean square error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (PCC) were adopted for the frequency value prediction, while only known frequency items were collected for metric value calculation[23,60]

$$\text{RMSE} = \sqrt{\frac{\sum_{ij}(P_{ij} - R_{ij})^2}{t}}$$

$$\text{MAE} = \frac{\sum_{ij}|P_{ij} - R_{ij}|}{t}$$

$$\text{PCC} = \frac{\sum_{ij}(P_{ij} - \overline{P})(R_{ij} - \overline{R})}{\sqrt{\sum_{ij}(P_{ij} - \overline{P})^2}\sqrt{\sum_{ij}(R_{ij} - \overline{R})^2}}$$

where t was the number of known drug-side effect frequency items, P_{ij} and R_{ij} were the predicted score and ground-truth frequency of the drug-side effect pair ij, and $\overline{P}$, $\overline{R}$ were their averages. It was worth noting that better performance was indicated by higher values for AUPR, AUC, and PCC, as well as lower values for RMSE and MAE. To enhance clarity and facilitate intuitive evaluation, reciprocals should also be calculated for RMSE and MAE (1/RMSE and 1/MAE).[61]

All the scripts were written in Python 3.8.8, and experiments were developed mainly based on scikit-learn 0.24.1 (https://scikit-learn.org/), xgboost 2.0.3 (https://xgboost.readthedocs.io/en/stable/), pytorch 1.10.1 (https://pytorch.org/), and optuna 2.10.0 (https://optuna.org/). The project was employed on the platform with Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz, NVIDIA(R) Quadro(R) RTX8000 48GB GPU and 754GB RAM on Red Hat Enterprise Linux release 8.4 (Ootpa).

RESULTS AND DISCUSSION

DBA Prediction with LLM-Based Representation. This study has proposed an effective technical strategy called LEDAP, which applies large-scale natural language models pre-trained on a massive corpus of human knowledgebase to

https://doi.org/10.1021/acs.analchem.4c01793
Anal. Chem. 2024, 96, 12395−12403

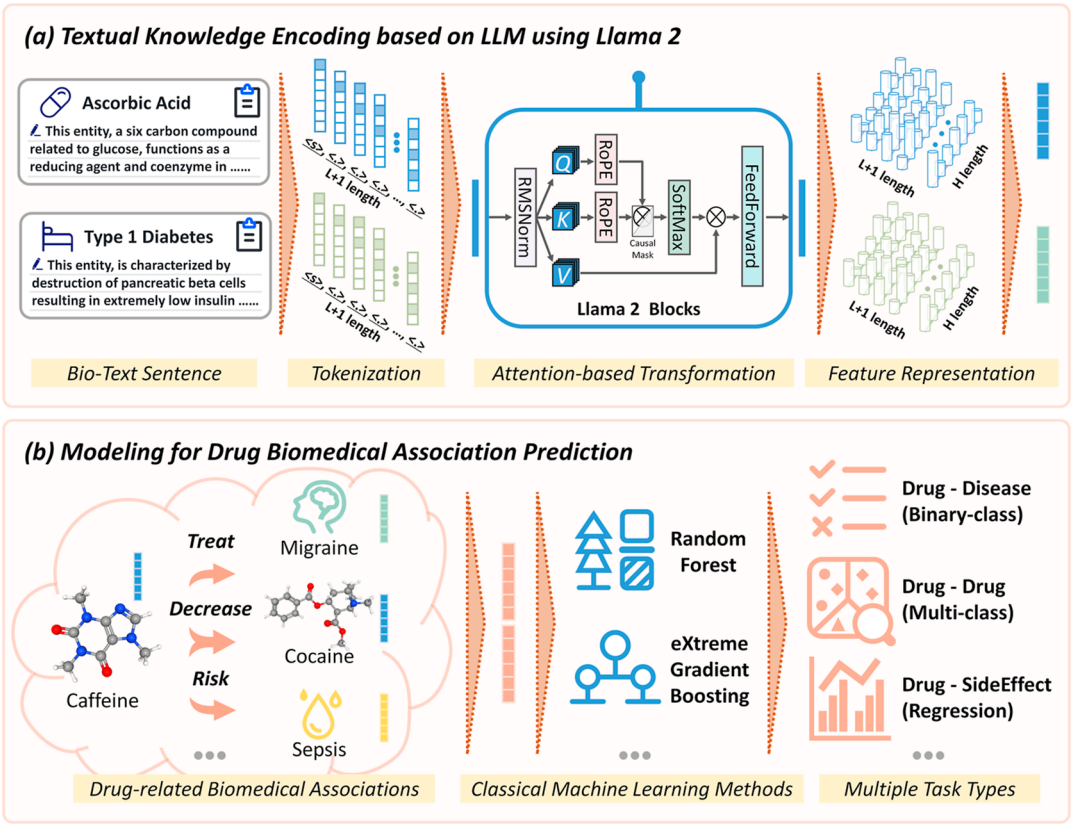


Figure 1. Framework of LEDAP for DBA analysis. (a) Textual knowledge encoding based on LLM using Llama 2: LEDAP followed several steps to produce LLM-based feature representations for biomedical entities. Biofunctional descriptions of the research objects, such as ascorbic acid and Type 1 diabetes, were first collected for tokenization. Biotext sentences were then segmented, initialized, and padded according to the Llama 2 vocabulary, resulting in $L + 1$ length tokenized sequences. After the attention-based nonlinear transformation through multilayers of Llama 2 blocks, the hidden embeddings ($L + 1 \times H$) were taken out and resized to one-dimensional feature vectors (H) by average pooling along the sequence axis. (b) Modeling for DBA prediction: classical ML methods were employed to build predictors for analyzing DBAs. Feature vectors of the involved biomedical entities were concatenated and used as model inputs for subsequent prediction tasks, such as DDA (binary classification), DDI (multiclass classification), and DSA (regression).

encode biomedical entities based on their biofunctional textual descriptions. This LLM-based encoding enabled the AI models to gain comprehensive insights from human knowledge to produce embeddings of biomedical semantic concepts, with the anticipation of effectively improving the computational prediction of DBAs.

Here, Meta's open-source LLM, Llama 2,[48] was used for textual encoding, and the LLM-based feature representations were then utilized as inputs for classical ML predictors, as illustrated in Materials and Methods. At the stage of textual knowledge encoding (Figure 1a), for a specific biomedical entity α, LEDAP implemented preprocessing and segmentation for its biotext sentence according to the Llama 2 vocabulary, producing L tokens. After a start token was prepended, the tokenized sequence of length $(L + 1)$ was input to multiple weighted blocks from Llama 2 (Supporting Information, Figure S1). Thus, the encoder performed the attention-based nonlinear transformation by query (Q), key (K), and value (V) embeddings. Then the hidden embeddings of shape $(L + 1 \times H)$ were taken out and subjected to average pooling, where the mean values were calculated across the sequence dimension, resulting in a fixed-size feature vector of length (H) that encapsulates biomedical semantic knowledge. At the stage of modeling for DBA prediction (Figure 1b), with the feature representations generated by LLM-based biotext encoding, classical ML methods such as XGBoost and RF were used to build predictors for analyzing DBAs. For modeling, feature vectors of the involved bioentities were concatenated and used as model inputs for subsequent prediction tasks.

To investigate the feasibility of our strategy from multiple angles, this study conducted various experiments corresponding to different task forms, mainly including DDA prediction (related to binary classification), DDI prediction (related to multiclass classification), and DSA prediction (related to regression). Based on the experimental results, LEDAP has demonstrated remarkable competitiveness in analyzing various DBAs across all types of tasks. Further details were elaborated on in the subsequent sections.

DDA Prediction of Binary Classification Problem. Drug repurposing aimed to shorten the drug discovery process dramatically and reduce the medication risk by reutilizing approved drugs to treat additional diseases.[62,63] Computational methods for screening all known drugs have been proposed to accelerate the process of drug repurposing.[64] Among them, representative AI-based computational methods were usually systematic data-driven, utilizing biomedical KG (BKG)-based embedding and completion,[12,65−67] which was inspired by the prosperity of expert-curated network analysis in novel DDA mining.[68]

Here, in Task-1, our strategy was applied to DDA prediction. A series of traditional BKG-based tools in the field of drug repurposing, including DREAMwalk, were used for compar-

https://doi.org/10.1021/acs.analchem.4c01793
Anal. Chem. 2024, 96, 12395−12403

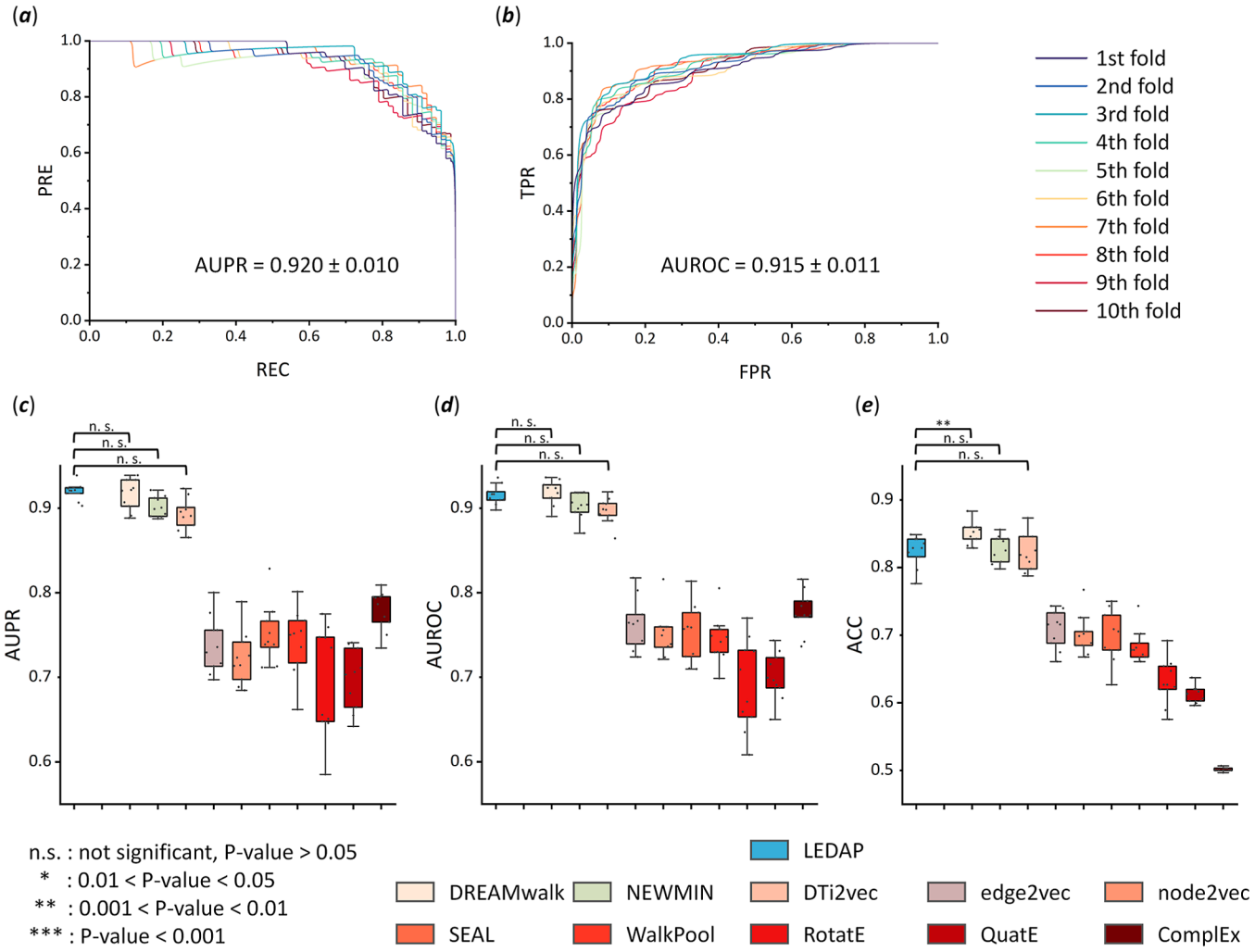


Figure 2. Performance evaluation of LEDAP in DDA prediction. (a) Test precision−recall curve among 10-folds, with an average AUPR of 0.920 ± 0.010. (b) Test receiver operating characteristic curve among 10-folds, with an average AUROC of 0.915 ± 0.011. (c−e) Prediction performance comparisons based on the AUPR, AUROC, and ACC in 10-fold cross-validation. LEDAP produced the best results at AUPR (always ranked the top 1) and demonstrated competitiveness at AUROC and ACC (surpassing edge2vec, node2vec, SEAL, WalkPool, RotatE, QuatE, ComplEx, and showing similar strength to DREAMwalk, NEWMIN, and DTi2vec).

ison with LEDAP.[12] Specifically, following DREAMwalk, the consistent dataset with 755 drug-disease pairs was used as the positive set, and an equal number of negative pairs were randomly sampled from Hetionet.[69] Meanwhile, the identical 10-fold cross-validation settings were curated for model training and performance evaluation. Details of biotext collection, feature encoding preparation, and experiment design are illustrated in Materials and Methods. Comparing our strategy with typical tools in this field, its competitiveness in binary classification for DBAs could be validated.

As a result, LEDAP achieved an average AUPR of 0.920 ± 0.010, AUROC of 0.915 ± 0.011, and ACC of 0.824 ± 0.023 in 10-fold cross-validation when used with RF predictor. The test PRC curves and ROC curves among 10-folds are shown in Figure 2a,b. Furthermore, the results of our strategy were compared with the other typical tools using nonparametric one-side Mann−Whitney U tests.[70,71] The alternative hypothesis was defined as our performance values being not higher than the others. As shown in Figure 2c−e, our strategy was especially advantageous at AUPR and always ranked top 1, and consistently ranked top 3 at AUROC and ACC (the results of the contrastive tools were from the reports in DREAMwalk[12] under identical experimental settings). Specifically, LEDAP significantly outperformed edge2vec, node2vec, SEAL, WalkPool, RotatE, QuatE, and ComplEx in all metrics. Meanwhile, except for the ACC value compared to DREAMwalk ($0.001 < P\text{-value} < 0.01$), LEDAP performed on the same level as DREAMwalk, NEWMIN, and DTi2vec in all metrics (P-value >0.05). The tiny standard deviation on different folds and the robust performance indicated the stability and reliability of our strategy. Supporting Information on the modeling is recorded in Supporting Information, Method S1. Source data of all experiment results are provided in Supporting Information, Table S1.

All in all, it has been shown that by combining LLM-based representation with classical ML methods, LEDAP could achieve robust and competitive performance in DDA prediction. This level of performance could stand in comparison with even the most representative DL methods presently used in this field, which were based on KG embedding.

Multi-Class Classification for Multityped DDI Events. The effect of one drug with another when administered together could pose a common and potentially dangerous

https://doi.org/10.1021/acs.analchem.4c01793
Anal. Chem. 2024, 96, 12395−12403

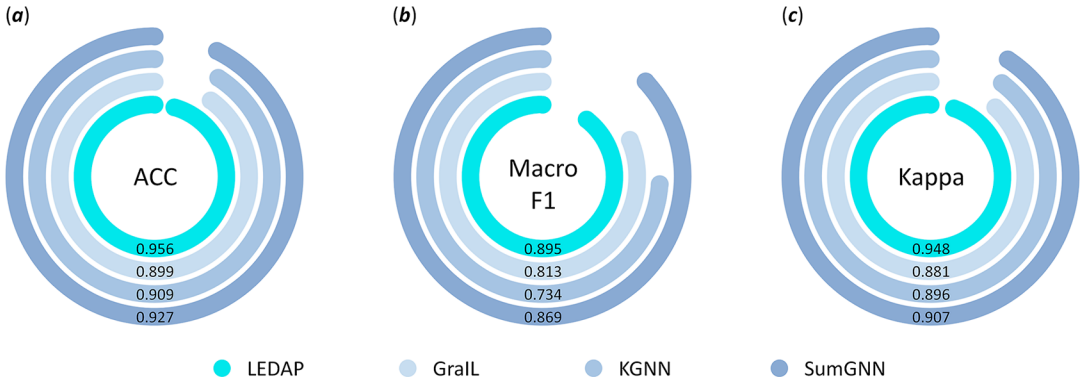


Figure 3. Performance evaluation of LEDAP in DDI prediction. (a−c) ACC, Macro F1, and Cohen's Kappa results of the competing tools on the test set. LEDAP always outperformed and showed significant improvements.

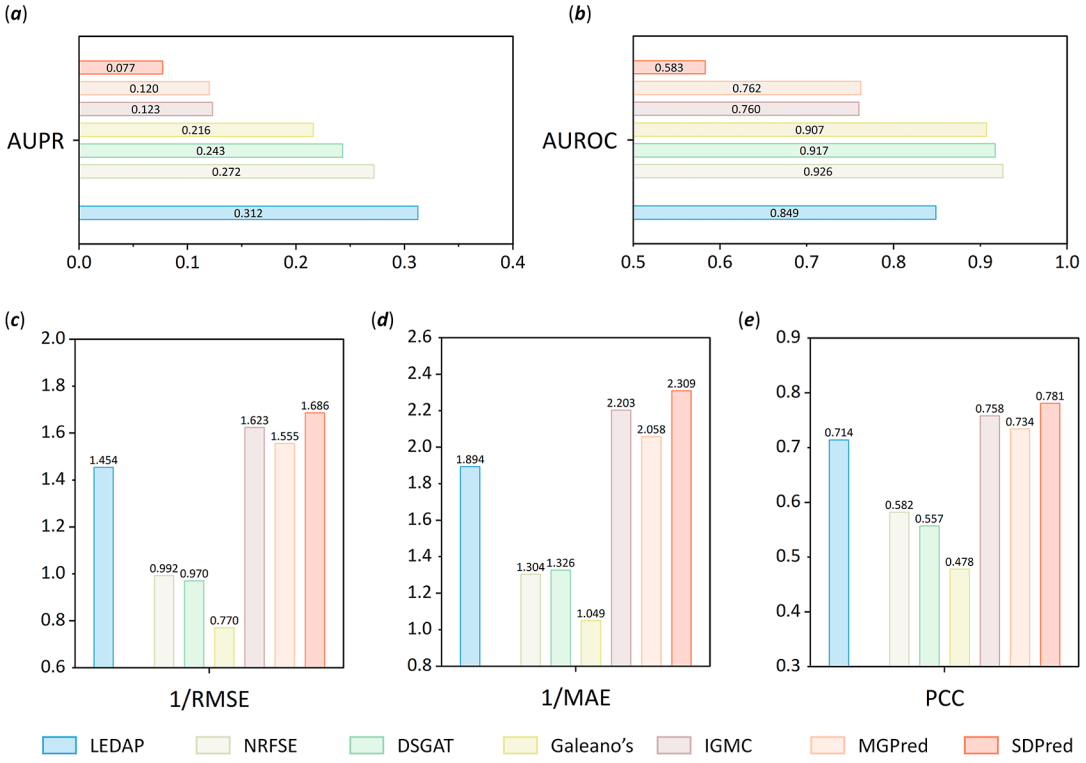


Figure 4. Performance evaluation of LEDAP in DSA prediction. (a,b) For the scenario of association presence identification, mean values of the AUPR and AUROC results in 10-fold cross-validation were collected for performance comparison. (c−e) For the scenario of frequency value estimation, mean values of the RMSE, MAE, and PCC results in 10-fold cross-validation were collected for performance comparison. Of note, for visual appeal, the reciprocals of mean RMSE and MAE were plotted, thus better performance could always be indicated by higher bars.

scenario for patients with complex conditions.[72] Numerous AI-based computational tools have already been released for DDI prediction, and several of them were specially designed for predicting multiple DDI types rather than simply predicting links between drug pairs.[73,74] Among these tools, SumGNN has gained popularity for its adoption of KG-based subgraph anchoring, self-attention-based summarization, and multi-channel knowledge integration, significantly improving multi-typed DDI predictions.[22]

Here, in Task-2, our strategy was applied to the DDI prediction. For our model training and performance evaluation, the DrugBank dataset utilized by SumGNN, as well as its training, development, and testing settings, were curated for our experiment.[22] Following the guidelines of SumGNN, drug pairs with more than one DDI type were excluded from the analysis, as more than 99.8% of pairs have only one type.[22,75] The DDIs in this dataset were associated with 1709 unique drugs and 86 types of pharmacological relations, such as an increase of cardiotoxic activity, decrease of serum concentration, and so on.[43] Details of biotext collection, feature encoding preparation, and experiment design are illustrated in Materials and Methods. Comparing our strategy with the popular tools in this field, its competitiveness in multiclass classification for DBAs could be validated.

As a result, LEDAP achieved test ACC of 0.956, Macro F1 of 0.895, and Cohen's Kappa of 0.948 when used with XGBoost predictor (as shown in Figure 3). LEDAP performed better than GraIL, KGNN, and SumGNN and made significant improvements, being at least 3.13% superior at ACC (0.956 vs 0.927), 2.99% at Macro F1 (0.895 vs 0.869), and 4.52% at

https://doi.org/10.1021/acs.analchem.4c01793
Anal. Chem. 2024, 96, 12395−12403

Kappa (0.948 vs 0.907) (the results of the contrastive tools were from the reports in SumGNN[22] under identical experimental settings). Supporting Information on the modeling is recorded in Supporting Information, Method S1. Source data of all experiment results are provided in Supporting Information, Table S2.

To sum up, combining LLM-based representation with classical ML methods has already enabled robust and competitive performance in multityped DDI classification. Analyzing multiple event classes, although more challenging, proved to be more rewarding than the binary classification of DDIs.[22]

DSA Identification and Frequency Regression. Galeano et al. first released a ML framework using an innovative matrix decomposition algorithm to predict drug-side effects frequencies.[60] Soon afterward, the Galeano-curated benchmark dataset was widely used in various studies for AI-based computational method construction. Incipiently, SDPred and MGPred were successively released by Zhao et al., respectively, based on graph attention networks and similarity-based DL, for determining the frequencies of drug side effects.[24,76] Xu et al. introduced DSGAT, characterized by using the drug molecular graph instead of the interaction graph for feature representation.[77] Lately, Wang et al. introduced NRFSE, a neighborhood-regularization method that leveraged multiview data on drugs and side effects, outperforming previous approaches.[23]

Here, in Task-3, our strategy was applied to the DSA prediction. It was worth mentioning that, two scenarios were highlighted for DSA prediction, including the identification of presence or absence and the estimation of frequency values. Method performances were, respectively, evaluated in terms of the distinct task forms, as detailed in Materials and Methods. The curated data for modeling and performance evaluation were derived from NRFSE,[23] which was initially extracted by Galeano's[60] from SIDER-4.1,[46] involving 759 drugs and 994 side effects. Therein, all known frequency DSA pairs (37,441) were formatted into five frequency classes and then assigned an integer value to each class: very rare (=1), rare (=2), infrequent (=3), frequent (=4), and very frequent (=5).[60] Notably, the other DSA candidates were treated as unknowns, as they did not fall under any frequency class and carried no corresponding integer values.[23] Details of biotext collection, feature encoding preparation, and experiment design are illustrated in Materials and Methods. Comparing our strategy with the latest tools in this field, its competitiveness in identification and regression for DBAs could be validated.

As a result, in 10-fold cross-validation, LEDAP used with XGBoost predictor achieved an average AUPR of 0.312 ± 0.020 and AUROC of 0.849 ± 0.022 for association presence identification, as well as an average RMSE of 0.688 ± 0.023, MAE of 0.528 ± 0.015, and PCC of 0.714 ± 0.014 for frequency value estimation (the reciprocals of mean RMSE and MAE values were 1.454 and 1.894, as shown in Figure 4). First and foremost, compared with the other tools mentioned above, LEDAP performed the best at AUPR (as shown in Figure 4a). Nevertheless, it was essential to not only correctly discriminate real DSAs (with higher AUPR and AUC values) but also accurately estimate frequency values (with higher 1/RMSE, 1/MAE, and PCC values).[77] According to the results of the other four metrics, each of the six methods for comparison had its own advantages and disadvantages in both scenarios (the results of the contrastive tools were from the reports in NRFSE[23] under identical experimental settings). As shown in Figure 4b−e, it could be observed that NRFSE, DSGAT, and Galeano's method generally outperformed IGMC, MGPred, and SDPred at association identification while performing oppositely at frequency estimation. In this regard, LEDAP had no serious problem of prejudice compared with the contrastive methods above. Specifically, LEDAP consistently outperformed NRFSE, DSGAT, and Galeano's method according to 1/RMSE, 1/MAE, and PCC, despite our slightly lower AUROC values. Similarly, LEDAP consistently outperformed IGMC, MGPred, and SDPred according to AUPR and AUROC, despite our marginally lower 1/RMSE, 1/MAE, and PCC values. Supporting Information on the modeling is recorded in Supporting Information, Method S1. Source data of all experiment results are provided in Supporting Information, Table S3.

All in all, the combination of LLM-based representation with classical ML methods has already proven to be successful in producing robust and competitive performance in DSA identification and frequency estimation. LEDAP was able to achieve the best AUPR and maintain good efficiency in both prediction scenarios, making it comparable to the latest methods such as matrix decomposition algorithm-based and graph neural network-based approaches.

CONCLUSIONS

This study proposed a novel strategy called LEDAP to analyze DBAs with LLM-based feature representations. The results have proven the convincing competitiveness of LEDAP in comparison to other existing popular tools for DDA, DDI, and DSA prediction. It is worth mentioning that using LLM-based natural language encoding in simple conjunction with classical ML methods could effectively facilitate the prediction of different DBA forms, such as binary classification, multiclass classification, and regression. Drawing upon these insights could guide the optimized implementation of LLMs in drug development analysis, potentially heightening performance in deciphering complex biomedical challenges. The ongoing advancements in LLMs and their ingenious applications in relevant fields foresee a future of vigorous progress in biomedical research.

ASSOCIATED CONTENT

Data Availability Statement

All source code and datasets are available at https://github.com/idrblab/LEDAP.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.4c01793.

Introduction of the modeling details and hyperparameter optimization process of the RF and XGBoost predictors; process diagram of preprocessing, segmentation, and tokenization for initial biotexts; and source data of evaluation results of LEDAP on DDA, DDI, and DSA predictions (PDF)

AUTHOR INFORMATION

Corresponding Authors

Feng Zhu − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, State Key Laboratory of Advanced Drug Delivery and Release Systems, Zhejiang University, Hangzhou 310058, China; Innovation Institute for Artificial Intelligence in Medicine of*

https://doi.org/10.1021/acs.analchem.4c01793
Anal. Chem. 2024, 96, 12395−12403

Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China; orcid.org/0000-0001-8069-0053; Email: zhufeng@zju.edu.cn

Chang-Yu Hsieh − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* Email: kimhsieh@zju.edu.cn

Honglin Li − *Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai 200062, China;* orcid.org/0000-0003-2270-1900; Email: hlli@ecust.edu.cn

Authors

Hanyu Zhang − *College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, State Key Laboratory of Advanced Drug Delivery and Release Systems, Zhejiang University, Hangzhou 310058, China; Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China;* orcid.org/0000-0002-4728-7702

Yuan Zhou − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0009-0002-0302-3913

Zhichao Zhang − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China*

Huaicheng Sun − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-1381-9571

Ziqi Pan − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-3883-4161

Minjie Mou − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0001-7619-2975

Wei Zhang − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0009-0007-8335-8663

Qing Ye − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China*

Tingjun Hou − *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0001-7227-2580

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.4c01793

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (82373790, 22220102001, 81872798, and U1909208); the Natural Science Foundation of Zhejiang Province (LR21H300001); the National Key R&D Program of China (2022YFC3400501); the Leading Talent of the 'Ten Thousand Plan'—National High-Level Talents Support Plan of China; the Fundamental Research Fund for Central University (2018QNA7023); the "Double Top-Class" University Project (181201*194232101); and the Key R&D Program of Zhejiang Province (2020C03010). This work was supported by Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; and the Information Technology Center of Zhejiang University.

■ REFERENCES

(1) Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; Ting, D. S. W. *Nat. Med.* **2023**, *29* (8), 1930−1940.
(2) Zhao, L.; Wang, J.; Hu, Y.; Cheng, L. *Mol. Ther. Nucleic Acids* **2020**, *22*, 198−208.
(3) Savage, N. *Nat. Biotechnol.* **2023**, *41* (5), 585−586.
(4) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; et al. *Nat. Biotechnol.* **2023**, *41* (8), 1099−1106.
(5) Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T. Y. *Briefings Bioinf.* **2022**, *23* (6), bbac409.
(6) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. *Bioinformatics* **2020**, *36* (4), 1234−1240.
(7) Xue, D.; Zhang, H.; Chen, X.; Xiao, D.; Gong, Y.; Chuai, G.; Sun, Y.; Tian, H.; Wu, H.; Li, Y.; et al. *Sci. Bull.* **2022**, *67* (9), 899−902.
(8) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. *Science* **2023**, *379* (6637), 1123−1130.
(9) Theodoris, C. V.; Xiao, L.; Chopra, A.; Chaffin, M. D.; Al Sayed, Z. R.; Hill, M. C.; Mantineo, H.; Brydon, E. M.; Zeng, Z.; Liu, X. S.; et al. *Nature* **2023**, *618* (7965), 616−624.
(10) Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. *Chem. Rev.* **2023**, *123* (6), 3089−3126.
(11) Cooper, M. A. *Nat. Rev. Drug Discov.* **2002**, *1* (7), 515−528.
(12) Bang, D.; Lim, S.; Lee, S.; Kim, S. *Nat. Commun.* **2023**, *14* (1), 3570.
(13) Yin, J.; Zhang, H.; Sun, X.; You, N.; Mou, M.; Lu, M.; Pan, Z.; Li, F.; Li, H.; Zeng, S.; et al. *IEEE J. Biomed. Health Inform.* **2023**, 1−12.
(14) Yang, Q.; Gong, Y.; Zhu, F. *Anal. Chem.* **2023**, *95* (13), 5542−5552.
(15) Zhang, Y.; Sun, H.; Lian, X.; Tang, J.; Zhu, F. *Adv. Sci.* **2023**, *10* (15), No. e2207061.
(16) Tang, J.; Mou, M.; Zheng, X.; Yan, J.; Pan, Z.; Zhang, J.; Li, B.; Yang, Q.; Wang, Y.; Zhang, Y.; et al. *Anal. Chem.* **2024**, *96* (12), 4745−4755.
(17) Yang, Q.; Chen, S.; Jiang, W.; Mi, L.; Liu, J.; Hu, Y.; Ji, X.; Wang, J.; Zhu, F. *Anal. Chem.* **2024**, *96* (4), 1410−1418.
(18) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. *Nat. Biotechnol.* **2011**, *29* (11), 1046−1051.
(19) Kauppi, K.; Rosenthal, S. B.; Lo, M. T.; Sanyal, N.; Jiang, M.; Abagyan, R.; McEvoy, L. K.; Andreassen, O. A.; Chen, C. H. *Am. J. Psychiatry* **2018**, *175* (7), 674−682.
(20) Gregori-Puigjané, E.; Setola, V.; Hert, J.; Crews, B. A.; Irwin, J. J.; Lounkine, E.; Marnett, L.; Roth, B. L.; Shoichet, B. K. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (28), 11178−11183.
(21) Ye, Q.; Hsieh, C. Y.; Yang, Z.; Kang, Y.; Chen, J.; Cao, D.; He, S.; Hou, T. *Nat. Commun.* **2021**, *12* (1), 6775.
(22) Yu, Y.; Huang, K.; Zhang, C.; Glass, L. M.; Sun, J.; Xiao, C. *Bioinformatics* **2021**, *37* (18), 2988−2995.
(23) Wang, L.; Sun, C.; Xu, X.; Li, J.; Zhang, W. *Bioinformatics* **2023**, *39* (9), btad532.
(24) Zhao, H.; Wang, S.; Zheng, K.; Zhao, Q.; Zhu, F.; Wang, J. *Briefings Bioinf.* **2022**, *23* (1), bbab449.
(25) Zong, N.; Kim, H.; Ngo, V.; Harismendy, O. *Bioinformatics* **2017**, *33* (15), 2337−2344.
(26) Li, F.; Yin, J.; Lu, M.; Mou, M.; Li, Z.; Zeng, Z.; Tan, Y.; Wang, S.; Chu, X.; Dai, H.; et al. *Nucleic Acids Res.* **2023**, *51* (D1), D1288−D1299.
(27) Fernández-Torras, A.; Duran-Frigola, M.; Bertoni, M.; Locatelli, M.; Aloy, P. *Nat. Commun.* **2022**, *13* (1), 5304.

(28) Su, Y.; Hu, Z.; Wang, F.; Bin, Y.; Zheng, C.; Li, H.; Chen, H.; Zeng, X. *Briefings Bioinf.* **2023**, *25* (1), bbad474.
(29) Hu, P.; Ye, Q.; Zhang, W.; Liu, J.; Ruan, T. *Bioinformatics* **2023**, *39* (11), btad689.
(30) Fu, T.; Li, F.; Zhang, Y.; Yin, J.; Qiu, W.; Li, X.; Liu, X.; Xin, W.; Wang, C.; Yu, L.; et al. *Nucleic Acids Res.* **2022**, *50* (D1), D1417−D1431.
(31) Zhang, H.; Wang, Y.; Pan, Z.; Sun, X.; Mou, M.; Zhang, B.; Li, Z.; Li, H.; Zhu, F. *Briefings Bioinf.* **2022**, *23* (6), bbac411.
(32) Cao, L.; Zhao, H.; Guan, R.; Jiang, H.; Dixneuf, P. H.; Zhang, M. *Nat. Commun.* **2021**, *12* (1), 4206.
(33) Islam, Y.; Leach, A. G.; Smith, J.; Pluchino, S.; Coxon, C. R.; Sivakumaran, M.; Downing, J.; Fatokun, A. A.; Teixidò, M.; Ehtezazi, T. *Adv. Sci.* **2021**, *8* (11), No. e2002085.
(34) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(35) Vatansever, S.; Schlessinger, A.; Wacker, D.; Kaniskan, H.; Jin, J.; Zhou, M. M.; Zhang, B. *Med. Res. Rev.* **2021**, *41* (3), 1427−1473.
(36) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (15), No. e2016239118.
(37) Dierickx, S.; Castelein, M.; Remmery, J.; De Clercq, V.; Lodens, S.; Baccile, N.; De Maeseneire, S. L.; Roelants, S.; Soetaert, W. K. *Biotechnol. Adv.* **2022**, *54*, 107788.
(38) Russo, J.; Tanaka, H. *Nat. Commun.* **2014**, *5*, 3556.
(39) Shen, Y.; Yuan, K.; Yang, M.; Tang, B.; Li, Y.; Du, N.; Lei, K. *J. Cheminform* **2019**, *11* (1), 22.
(40) Pei, Q.; Wu, L.; Gao, K.; Zhu, J.; Wang, Y.; Wang, Z.; Qin, T.; Yan, R. *arXiv* **2024**, arXiv:2403.01528.
(41) Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. *arXiv* **2023**, arXiv:2303.18223.
(42) Hakenberg, J.; Plake, C.; Royer, L.; Strobelt, H.; Leser, U.; Schroeder, M. *Genome Biol.* **2008**, *9* (Suppl 2), S14.
(43) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. *Nucleic Acids Res.* **2018**, *46* (D1), D1074−D1082.
(44) Siramshetty, V. B.; Grishagin, I.; Nguyen Đ, T.; Peryea, T.; Skovpen, Y.; Stroganov, O.; Katzel, D.; Sheils, T.; Jadhav, A.; Mathé, E. A.; et al. *Nucleic Acids Res.* **2022**, *50* (D1), D1307−D1316.
(45) Schriml, L. M.; Mitraka, E.; Munro, J.; Tauber, B.; Schor, M.; Nickle, L.; Felix, V.; Jeng, L.; Bearer, C.; Lichenstein, R.; et al. *Nucleic Acids Res.* **2019**, *47* (D1), D955−D962.
(46) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. *Nucleic Acids Res.* **2016**, *44* (D1), D1075−D1079.
(47) Mitchell, M.; Krakauer, D. C. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120* (13), No. e2215907120.
(48) Touvron, H.; Martin, L.; Stone, K. R.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. *arXiv* **2023**, arXiv:2307.09288.
(49) Zhang, B.; Sennrich, R. Root mean square layer normalization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.
(50) Shazeer, N. M. *arXiv* **2020**, arXiv:2002.05202.
(51) Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; Liu, Y. *Neurocomputing* **2024**, *568*, 127063.
(52) Keloth, V. K.; Hu, Y.; Xie, Q.; Peng, X.; Wang, Y.; Zheng, A.; Selek, M.; Raja, K.; Wei, C. H.; Jin, Q.; et al. *Bioinformatics* **2024**, *40* (4), btae163.
(53) Sadeghi, S.; Bui, A.; Forooghi, A.; Lu, J.; Ngom, A. *arXiv* **2024**, arXiv:2402.00024.
(54) Zhao, Z.; Ma, D.; Chen, L.; Sun, L.; Li, Z.; Xu, H.; Zhu, Z.; Zhu, S.; Fan, S.; Shen, G.; et al. *arXiv* **2024**, arXiv:2401.14818.
(55) Bojar, D.; Lisacek, F. *Chem. Rev.* **2022**, *122* (20), 15971−15988.
(56) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: San Francisco, California, USA, 2016.
(57) Breiman, L. *Mach. Learn.* **2001**, *45* (1), 5−32.
(58) Zhang, C.; Mou, M.; Zhou, Y.; Zhang, W.; Lian, X.; Shi, S.; Lu, M.; Sun, H.; Li, F.; Wang, Y.; et al. *Briefings Bioinf.* **2022**, *23* (5), bbac160.
(59) Strobl, C.; Malley, J.; Tutz, G. *Psychol. Methods* **2009**, *14* (4), 323−348.
(60) Galeano, D.; Li, S.; Gerstein, M.; Paccanaro, A. *Nat. Commun.* **2020**, *11* (1), 4575.
(61) Keller, A. S.; Pines, A. R.; Shanmugan, S.; Sydnor, V. J.; Cui, Z.; Bertolero, M. A.; Barzilay, R.; Alexander-Bloch, A. F.; Byington, N.; Chen, A.; et al. *Nat. Commun.* **2023**, *14* (1), 8411.
(62) Peyvandipour, A.; Saberian, N.; Shafi, A.; Donato, M.; Draghici, S. *Bioinformatics* **2018**, *34* (16), 2817−2825.
(63) Aghdam, R.; Habibi, M.; Taheri, G. *J. Cheminform* **2021**, *13* (1), 70.
(64) Liu, R.; Wei, L.; Zhang, P. *Nat. Mach. Intell.* **2021**, *3* (1), 68−75.
(65) Thafar, M. A.; Olayan, R. S.; Albaradei, S.; Bajic, V. B.; Gojobori, T.; Essack, M.; Gao, X. *J. Cheminform* **2021**, *13* (1), 71.
(66) Yu, L.; Xia, M.; An, Q. *Briefings Bioinf.* **2022**, *23* (1), bbab364.
(67) Gao, Z.; Fu, G.; Ouyang, C.; Tsutsui, S.; Liu, X.; Yang, J.; Gessner, C.; Foote, B.; Wild, D.; Ding, Y.; et al. *BMC Bioinf.* **2019**, *20* (1), 306.
(68) Zhou, Y.; Wang, F.; Tang, J.; Nussinov, R.; Cheng, F. *Lancet Digit Health* **2020**, *2* (12), e667−e676.
(69) Himmelstein, D. S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S. L.; Hadley, D.; Green, A.; Khankhanian, P.; Baranzini, S. E. *Elife* **2017**, *6*, No. e26726.
(70) Zhao, S.; Dai, G.; Li, J.; Zhu, X.; Huang, X.; Li, Y.; Tan, M.; Wang, L.; Fang, P.; Chen, X.; et al. *NPJ. Digit Med.* **2024**, *7* (1), 3.
(71) Mou, M.; Pan, Z.; Zhou, Z.; Zheng, L.; Zhang, H.; Shi, S.; Li, F.; Sun, X.; Zhu, F. *Research* **2023**, *6*, 0240.
(72) Vilar, S.; Uriarte, E.; Santana, L.; Lorberbaum, T.; Hripcsak, G.; Friedman, C.; Tatonetti, N. P. *Nat. Protoc.* **2014**, *9* (9), 2147−2163.
(73) Lin, X.; Quan, Z.; Wang, Z.-J.; Ma, T.; Zeng, X. KGNN: knowledge graph neural network for drug-drug interaction prediction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*: Yokohama, Yokohama, Japan, 2021.
(74) Teru, K. K.; Denis, E. G.; Hamilton, W. L. Inductive relation prediction by subgraph reasoning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.
(75) Ryu, J. Y.; Kim, H. U.; Lee, S. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115* (18), E4304−E4311.
(76) Zhao, H.; Zheng, K.; Li, Y.; Wang, J. *Briefings Bioinf.* **2021**, *22* (6), bbab239.
(77) Xu, X.; Yue, L.; Li, B.; Liu, Y.; Wang, Y.; Zhang, W.; Wang, L. *Briefings Bioinf.* **2022**, *23* (2), bbab586.

https://doi.org/10.1021/acs.analchem.4c01793
Anal. Chem. 2024, 96, 12395−12403