ChemBERTaDDI: Transforming Drug-Drug Interaction Prediction with Transformers and Clinical Insights

Anastasiya A. Gromova
School of Computing and Informatics,
University of Louisiana at Lafayette
Lafayette, Louisiana 70503
Email: anastasiya.gromova1@louisiana.edu

Anthony S. Maida
School of Computing and Informatics,
University of Louisiana at Lafayette
Lafayette, Louisiana 70503
Email: anthony.maida@louisiana.edu

***Abstract*—The problem of polypharmacy arises when two or more drugs taken in combination cause adverse side effects, even when the use of the drugs individually causes no harm. Drug-drug interactions (DDIs) are a major cause of these reactions, contributing to increased morbidity and mortality. As the potential for harmful DDI grows exponentially, the prediction of drug-drug interactions is increasingly critical for patient safety and effective healthcare management. In this paper, we develop the *ChemBERTaDDI* framework, which effectively combines clinical domain data, represented by the mono side-effect features with the enriched chemical molecular representations derived from *ChemBERTa-77M-MLM*, a transformer-based language model. Experiments performed on a benchmark data set show superior performance compared with five state-of-the-art methods: Decagon, DeepWalk, DEDICOM, NNPS, and RESCAL. The evaluation demonstrates that *ChemBERTaDDI* achieves an F1 score of 0.94 and an AUROC of 0.97 outperforming the baseline architectures and generalizing to new introduced drug compounds.**

***Index Terms*—Transformer, Natural Language Processing, Molecular Representation Learning, Drug-Drug Interactions, Polypharmacy, ChemBERTa, Neural Networks, Deep Learning, Attention Mechanism, Bioinformatics**

I. Introduction

Polypharmacy, the use of multiple drugs by patients, has become a widespread practice in modern healthcare. Although often medically necessary and offering therapeutic benefits, polypharmacy often leads to harmful drug-drug interactions (DDI). The challenge of predicting potential DDIs becomes increasingly complex as the number of possible drug combinations grows exponentially, far outpacing the capacity of traditional clinical trials to identify rare and complex interactions. Existing computational methods for DDI prediction rely on extracting chemical and biological features from diverse drug properties. However, some drug properties are hard and costly to discover and may not be available in many cases.

In this paper, we present *ChemBERTaDDI* framework, which enhances drug-drug interaction (DDI) prediction by integrating transformer-based language model embeddings with clinical mono side effect data. Our approach leverages *ChemBERTa-77M-MLM* embeddings derived from transformer attention mechanisms and concatenates them with individual drug clinical side effect features. This combined feature set is then fed into a multi-layer neural network for binary classification task. By maintaining frozen-layer embeddings, *ChemBERTaDDI* efficiently captures high-quality molecular representations without additional computational overhead. The framework predicts whether a pair of drugs will result in specific adverse side effects. Benchmark comparisons with DeepWalk, Decagon, and NNPS demonstrate that *ChemBERTaDDI* outperforms existing models in DDI prediction and understanding. Additionally, a single neural network model is trained across 964 polypharmacy side effects, utilizing a comprehensive $52K$ token SMILES vocabulary to ensure scalability and adaptability to new drug compounds. In practice, *ChemBERTaDDI* framework utilizes a single neural network model that is trained across 964 polypharmacy side effects while preserving ChemBERTa's pre-trained knowledge (reducing computational overhead). This approach allows the framework to focus on learned molecular representations within a comprehensive $52K$ token SMILES chemical vocabulary, ensuring adaptability to new compound data and scalability. The key contributions of this paper are summarized as follows:

- Developed *ChemBERTaDDI* framework (Section 2) that jointly leverages drug clinical data with the external chemical structure embeddings features (Section 3) for the DDI prediction task.
- Demonstrate that incorporating molecular structure information from the SMILES encoding along with drug mono side effect clinical information significantly improves DDI prediction performance (Section 5). The choice of ChemBERTa model size affects the performance, thus we conduct ablation studies using *ChemBERTa-77M-MLM* and *ChemBERTa-10M-MLM* models (Section 6).

Given a set of 645 drugs, denoted by $D = \{d_1, d_2, \ldots, d_{645}\}$ and 964 polypharmacy side effects, denoted by $S = \{s_1, s_2, \ldots, s_{964}\}$, our objective is to determine whether an interaction occurs for a pair of drugs (d_i, d_j) and to identify the specific side effect(s) if present.

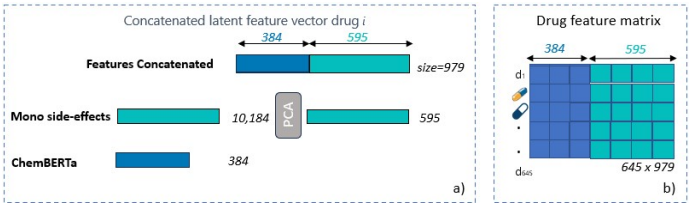

Fig. 1. ChemBERTaDDI Feature Representations. **(a)** Concatenated feature vector for drug d_i, combining ChemBERTa drug–chemical structure embeddings (384 dimensions, shown in blue) with Principal Component Analysis (PCA) representations of mono side effects (595 dimensions, shown in green). **(b)** Drug feature matrix formed by concatenating ChemBERTa and mono feature vectors for each drug, resulting in a 645×979 matrix.

For each side effect $s \in S$, we construct an adjacency matrix $A \in \mathbb{R}^{m \times m}$, where $m = 645$ (the total number of drugs). Each entry a_{ij} is defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } d_i \text{ and } d_j \text{ interact under } s, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Our primary objective is to develop a robust predictive model that leverages both high-dimensional clinical data with generalizable rich chemical structure representations to accurately predict DDIs.

II. Methodology

A. ChemBERTaDDI Framework

The **ChemBERTaDDI** framework is designed to integrate clinical mono side effect features with transformer-based molecular embeddings to predict potential DDIs. The framework operates in the following steps, refer to Figure 2.

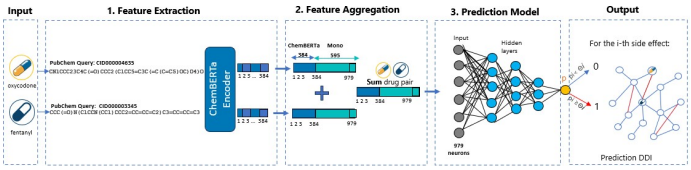

Fig. 2. ChemBERTaDDI Framework. Step 1: Extract 645 unique drugs and retrieve their SMILES from PubChem. Encode each SMILES with ChemBERTa into 384-dimensional features. Step 2: Aggregate features by concatenating the ChemBERTa embeddings (blue) with PCA-reduced mono side-effect vectors (green) to 99% variance. For each drug pair, sum the individual 979-dimensional vectors into a single representation. This final vector is passed to a prediction model that outputs a sigmoid neuron result.

Building on the motivations and problem settings identified by Zitnick et al. (Decagon) [1] and Masumshah et al. (NNPS) [2], the objective is to predict drug-drug interactions (DDIs) associated with 964 side effects. Two key drug-related characteristics are explored (see Figure 1 section (a) for details):

1) Drug mono-side effects clinical information: Clinical information is represented as categorical data with the total of 10,184 unique categories. One-hot encoding technique is utilized to transform categorical variables into a binary matrix, where each unique category is represented by a distinct binary feature to preserve the categorical nature. Principal component analysis (PCA) is applied for dimensionality reduction. By retaining 99% of the variance, PCA projects the data into a lower-dimensional space with 595 dimensions, effectively capturing the most significant patterns while minimizing computational complexity.

2) Chemical structure information: Motivated by the intuition of Huang et al. (CASTER) [3], which models drug pairs as a collection of chemical functional sub-structures to predict DDIs, we utilize chemical feature representations. The *ChemBERTaDDI* framework leverages the Simplified Molecular Input Line Entry System (SMILES) [4] strings to capture the intricate chemical structures of drugs. SMILES serve as a textual notation for chemical compounds, encapsulating their molecular composition and structural nuances. SMILES are collected from the external PubChem database [5] and are tokenized using `AutoTokenizer` from Hugging Face's Transformers library. `AutoTokenizer` is the sub-word tokenization technique that splits SMILES strings into tokens, allowing the transformer model to process them effectively. Below is the summary of the *ChemBERTaDDI* components, refer to Figure 2 for details.

- **Feature Extraction:** The `AutoTokenizer` tokenizer converts SMILES strings into token IDs, which are then fed into the *ChemBERTa-77M-MLM* transformer model developed by Chithranda et al. [6]. The idea is similar to CASTER [3], instead of frequent sequential pattern mining algorithm, we leverage `AutoTokenizer`, which generates the recurring chemical sub-structures of molecular representations of drugs. Inspired by the advancements in self-supervised pretraining for molecular property prediction [6], *ChemBERTaDDI* employs *ChemBERTa-77M-MLM* embeddings derived from transformer attention mechanisms. The *ChemBERTa* model processes tokens through multiple encoder layers to generate contextualized molecular embeddings. Serving as a robust feature extractor, *ChemBERTa-77M-MLM* transforms SMILES inputs into generalizable embeddings that effectively capture the chemical structure and properties of each drug based on its molecular composition.
- **Feature Aggregation:** The PCA-reduced clinical mono side effect features are concatenated with *ChemBERTa* embeddings to form a unified 979-dimensional feature vector for each drug. Subsequently, for each drug pair, their respective individual feature vectors are aggregated using an element-wise summation operation. The summation gate effectively combines the attributes of both drugs, resulting in a comprehensive feature vector that encapsulates the joint clinical and chemical characteristics necessary for predicting DDIs.
- **Prediction Model:** Inspired by the elegant, yet simple neural network-based predictor (NNPS) introduced by Masumshah et al. [2], the aggregated feature vectors are fed into a single neural network model configured for binary classification. The model is trained to predict the likelihood of adverse DDIs by leveraging the combined clinical and chemical information within feature vectors.

III. Drug Feature Representations

Each drug d_k is represented by two complementary feature matrices, encoding both clinical information (*mono-side effects*) and chemical structure (via *ChemBERTa*). Specifically, for each side effect, two types of feature matrices are considered, refer to Figure 1: (1) the drug mono side effects matrix 645×10,184 (2) the drug chemical feature matrix 645×384 derived from ChemBERTa.

A. Mono-side Effect Features

The individual drug mono side effects clinical information is represented via a one-hot encoded vector of length $10{,}184$, indicating the presence or absence of specific mono side effects. Because the mono-side effect feature matrix is high-dimensional and sparse, the Principal Component Analysis (PCA) is applied to reduce dimensionality while retaining 99% of the variance. As the result, the reduced mono side effect matrix yields:

$$\mathbf{z}_k^{\text{Mono side effects}} \in \mathbb{R}^{595}.$$

B. Molecular Feature Extraction

For the purpose of modeling detailed molecular semantics of each drug, the proposed idea is to adapt *ChemBERTa-77M-MLM*, a transformer-based language model pretrained on unlabeled SMILES data. The decision to utilize *ChemBERTa-77M-MLM* over latent variable models stems from its ability to provide robust and contextually rich embeddings with fewer architectural complexities. By leveraging the strengths of transformer-based models, *ChemBERTa* offers a powerful tool for capturing the nuanced semantics of molecular data, facilitating more accurate and efficient predictions in drug-drug interaction (DDI) task. In the *ChemBERTaDDI* framework, *ChemBERTa-77M-MLM* acts as an encoder from which a continuous latent representation is obtained, where molecular representation of drugs has been used as both input and output, thus learning a $52{,}000$-token vocabulary of molecular fragments. Chithrananda et al. [6] converted the discrete SMILES representations of the drug molecules into continuous multi-dimensional representation using *ChemBERTa-77M-MLM* transformer model. The primary objective of *ChemBERTa-77M-MLM* is to learn meaningful and robust representations of molecular structures encoded as SMILES strings. This is achieved through the **Masked Language Modeling (MLM)** paradigm, a foundational training objective borrowed from natural language processing models, such as BERT [7] and RoBERTa [8]. The *ChemBERTa-77M-MLM* model [6], employs the MLM paradigm, where a portion 15% of the SMILES tokens in the input SMILES string are randomly masked/hidden, and the model is trained to predict these masked tokens based on their surrounding context. Formally, given a SMILES string represented as a sequence of tokens:

$$\mathbf{X} = (x_1, x_2, \ldots, x_n)$$

a subset of tokens is masked to produce:

$$\mathbf{X}_{\text{masked}} = (x_1, x_2, \ldots, \texttt{[MASK]}, \ldots, x_n)$$

The objective is to maximize the likelihood of the masked tokens given the unmasked context:

$$\mathcal{L} = \sum_{i \in \mathcal{M}} \log P(x_i \mid \mathbf{X}_{\text{masked}} \setminus \{x_i\})$$

where:

- $\mathcal{M}$ denotes the set of masked token indices.
- $P(x_i \mid \cdot)$ represents the probability of token x_i being the correct token given the surrounding context.

Let $f(\cdot) : X \to H$ denote ChemBERTa's encoder mapping a tokenized SMILES X (length ≤ 512) to a final hidden state $H \in \mathbb{R}^{384}$. Using a *masked language modeling* objective [7], [8], ChemBERTa learns chemically meaningful latent representations from large-scale SMILES corpora. In our framework, we freeze ChemBERTa trained weights and extract the hidden representation molecular knowledge for each drug represented by a 384-dimensional embedding vector:

$$\mathbf{z}_k^{\text{ChemBERTa}} \in \mathbb{R}^{384}.$$

C. Drug Pair Feature Representation

To facilitate DDI prediction, we concatenate the reduced mono-side effect features and the ChemBERTa embeddings for each drug, resulting in a following unified feature vector:

$$\mathbf{z}_k = \left[\mathbf{z}_k^{\text{ChemBERTa}};\ \mathbf{z}_k^{\text{Mono Side Effects}}\right] \in \mathbb{R}^{979}.$$

For a drug pair (d_i, d_j), we aggregate their respective feature vectors by summation:

$$\mathbf{z}_{ij} = \mathbf{z}_i + \mathbf{z}_j \in \mathbb{R}^{979}.$$

This aggregated feature vector serves as the input to the neural network model, enabling the prediction of potential DDIs based on the combined clinical and chemical characteristics of the drug pair.

IV. Prediction Model

In the *ChemBERTaDDI* framework, drug-drug interactions (DDIs) are predicted using a single feedforward neural network for binary classification. The model processes an aggregated feature vector $\mathbf{z}_{ij} \in \mathbb{R}^{979}$ that combines chemical and clinical attributes. It consists of three fully connected hidden layers with 250, 300, and 200 neurons respectively, each initialized with the Glorot normal initializer and followed by Batch Normalization, ReLU activation, and Dropout for regularization. The output layer contains a single neuron with a sigmoid activation function, producing a binary classification output indicating the presence (1) or absence (0) of an adverse DDI.

Let $F(\cdot)$ denote our NN-based predictor. For side effect s, the predicted interaction probability between (d_i, d_j) is:

$$\hat{y}_{ij}^s = F(\mathbf{z}_{ij}),$$

where $\hat{y}_{ij}^s$ is compared against a threshold Θ to decide whether s occurs.

V. Experimental Setup

This section details the dataset and the experiments conducted using *ChemBERTaDDI*. Table I presents the statistical overview of the databases, while Table II summarizes the pre-processed feature dataset. The experiments aim to evaluate the performance of *ChemBERTaDDI* by addressing the following two questions: **Q1)** Does *ChemBERTaDDI* achieve more accurate predictions of drug-drug interactions compared to other strong baseline models? **Q2)** Can *ChemBERTaDDI* provide accurate DDI predictions using only drug chemical structure embeddings from *ChemBERTa*, without incorporating drug clinical characteristics?

A. Databases

The annotated drug pairs with polypharmacy side effects are sourced from the TWOSIDES database [9], a comprehensive database with drug-drug interactions (DDIs) containing 1,317 polypharmacy side effects across 645 drugs and 63,473 drug pairs extracted from the FDA Adverse Event Reporting System (FAERS). In total the TWOSIDES database contains 4,576,785 associations, which include multiple instances of interactions between the same pairs of drugs. Next, OFFSIDES and SIDER databases are integrated to derive the side effects of individual drugs (mono side effects). The SIDER database [10] compiles side effects from drug labels, while the OFFSIDES database [9] includes off-label side effects observed during clinical trials. The SIDER information is extracted from drug labels and includes $1,556$ drugs and $5,868$ side effects compiled from public documents. The OFFSIDES information is collected from clinical trials and includes $1,332$ drugs and $10,097$ off-label side effects. By merging these databases and eliminating synonym side effects, $10,184$ mono side effects are obtained for the 645 drugs which are in TWOSIDES database. In addition, during the pre-processing, the chemical structure information for 645 drugs is obtained from the *PubChem* database [5], the largest publicly available database for chemical structure information. The Simplified Molecular Input Line Entry System (SMILES) [4] notation is used to represent the chemical structures of the 645 drugs in a string format. PubChem is leveraged to obtain the corresponding SMILES notation, which encodes molecular structure information as a string.

TABLE I
Databases details

Database	Details
TWOSIDES	No. of side effects = 1,317 No. of drugs = 645 No. of drug pairs = 63,473
SIDER, OFFSIDES	No. of SIDER drugs = 1,556 No. of SIDER side effects = 5868 No. of OFFSIDES drugs = 1,332 No. of OFFSIDES side effects = 10,097
STITCH	No. of compounds = 519,022 No. of protein targets = 8,934 No. of interactions = 8,083,600
PubChem	No. of SMILES = 645

B. DB1 Dataset

The above databases were adapted and pre-processed to construct the *DB1* dataset. The *DB1* dataset is structured based on TWOSIDES annotated DDIs consisting of Polypharmacy Side Effects and a set of Drug Pairs. The polypharamacy side effects filtered to the 964 most frequently occurring side effects, which occurred in at least 500 DDIs. There are $63,473$ unique drug pairs, resulting meaning unique drug pair interactions. Each drug pair may have multiple associated adverse effects, resulting in 645 drugs with $63,473$ unique drug-drug interactions. In summary, DB1 can be represented as a triplet (Stitch 1, Stitch 2, Polypharmacy Side Effect). Further detailed information on the *DB1* dataset is available in Table II. Following standard practice [1], [2] the known drug-drug interaction pairs from TWOSIDES are used as positive samples to build the positive set, and the remaining unlabeled drug-drug pairs are considered as negative set.

TABLE II
DB1 Dataset details

Dataset	Details	Features
DB1	No. of polypharmacy effects = 964 No. of drugs = 645 No. of interactions = 63,473 No. of mono side effects = 10,184 No. of SMILES = 645	No. PCA mono side effects = 595 No. ChemBERTa = 384 No. Drug feature = 979

TABLE III
Baseline Model Performance Comparison

Method	ROC	AUPR	Recall	Precision	F1 Score
ChemBERTa77M-MLM, Mono	0.965	0.954	0.968	0.909	**0.937**
ChemBERTa77M-MLM, Mono, Protein	0.965	0.954	0.968	0.909	**0.937**
Decagon	0.874	0.825	0.950	0.771	0.850
NNPS	0.966	0.953	0.976	0.901	0.936

TABLE IV
Case Study: Results of Dangerous Polypharmacy Side Effects in ChemBERTaDDI and Decagon on AUPR

Polypharmacy	AUPR (ChemBERTaDDI)	AUPR (Decagon)	Evidence
Sarcoma	0.978	0.789	Serban et al. [14]
Carcinoma of the Cervix	0.952	0.810	Arbyn et al. [15]
Malignant Hypertension	1.00	0.858	Januszewicz et al. [16]
Oophorectomy	0.974	0.911	Evans et al. [17]
Leukopenia	0.960	0.695	Flanagan et al. [18]

TABLE V
ChemBERTaDDI Performance with various Feature Sets

Method	ROC	AUPR	Recall	Precision	F1 Score
ChemBERTa10M-MLM	0.9281	0.9106	0.9375	0.8485	0.8887
ChemBERTa10M-MLM, Protein	0.9311	0.9144	0.9387	0.8530	0.8917
ChemBERTa10M-MLM, Mono	0.9630	0.9499	0.9680	0.9032	0.9336
ChemBERTa10M-MLM, Mono, Protein	0.9626	0.9491	0.9665	0.9048	**0.9338**
ChemBERTa77M-MLM	0.9308	0.9137	0.9377	0.8513	0.8909
ChemBERTa77M-MLM, Protein	0.9414	0.9272	0.9402	0.8685	0.9012
ChemBERTa77M-MLM, Mono	0.9653	0.9539	0.9681	0.9089	**0.9368**
ChemBERTa77M-MLM, Mono, Protein	0.9658	0.9538	0.9688	0.9081	0.9366

TABLE VI
PROBING THE CHEMICAL STRUCTURE EMBEDDINGS

Models	Method	Average F1
ChemBERTa-10M-MLM	ChemBERTaDDI	0.8887
ChemBERTa-77M-MLM	ChemBERTaDDI	**0.8909**

TABLE VII
THE AVERAGE OF THE BEST RESULTS FOR CHEMBERTADDI, NNPS, AND DECAGON METHODS FOR 964 SIDE EFFECTS

Method	TP	FP	TN	FN	Precision	Recall	F-score
NNPS	105,153	12,013	95,955	2,608	0.901	**0.976**	0.936
Decagon	208,963	54,855	163,988	9,880	0.771	0.950	0.850
ChemBERTaDDI	107,790	12,817	99,061	4,116	**0.909**	0.970	**0.938**

C. Training Procedure

In the *ChemBERTaDDI* framework, pre-trained contextual embeddings *ChemBERTa10M-MLM* and *ChemBERTa77M-MLM* are are utilized to initialize the chemical representations of drug structures. These representations reside in a latent space, enabling generalization to novel drugs. We train a neural predictor via supervised learning on labeled drug pairs, using their chemical representations and mono side-effect information. Binary cross-entropy loss quantifies errors, optimized by Adam. Dropout and early stopping (10 epochs without validation improvement) mitigate overfitting. We measure performance (ROC, AUPR, Recall, Precision, F1-Score) for 964 side effects. The model is a feedforward network with three hidden layers and a single output neuron. Its input layer has 979 neurons (a 384-dimensional *ChemBERTa* embedding plus 595-dimensional mono side effects). We split drug pairs by side effect into training (90%), validation (5%), and test (5%) sets, following a 5-fold cross-validation. We explored three- and five-layer architectures, using both *ChemBERTa-10M-MLM* and *ChemBERTa-77M-MLM*, with learning rates (0.001, 0.005, 0.01) and dropout rates (0.1, 0.2, 0.3). The best configuration uses a batch size of 1024, 50 epochs, dropout 0.3, and early stopping at 10 stalled epochs. Our final model (three fully connected layers of 250, 300, and 200 neurons) with *ChemBERTa-77M-MLM, Mono* features achieves an F1-score of 0.938, AUPR of 0.954, and AUROC of 0.965. For training, the selected compute was 64 cores, 440 GB RAM, 2816 GB disk space.

VI. RESULTS AND DISCUSSION

In this section, we present a comprehensive analysis of the experimental results and findings obtained during our study. To address **Q1)**, we compared the *ChemBERTaDDI* framework with five well-established methods: Decagon [1], NNPS [2], DeepWalk [11], DEDICOM [12], and RESCAL [13]. The performance metrics for each method are as follows: Decagon achieved an ROC of 0.874 and an AUPR of 0.825. DeepWalk attained an ROC of 0.761 and an AUPR of 0.737. DEDICOM obtained an ROC of 0.705 and an AUPR of 0.637. RESCAL reached an ROC of 0.693 and an AUPR of 0.613. NNPS achieved an ROC of 0.966 and an AUPR of 0.953. Our experiments demonstrate that the *ChemBERTaDDI* framework significantly outperforms these baseline methods, as shown in Table III, achieving an F1-score of 0.94, an AUPR of 0.95, and an AUROC of 0.97. In addition, Table IV presents the results of a case study evaluating the performance of *ChemBERTaDDI* against Decagon in predicting dangerous polypharmacy side effects. The evaluation is measured by the Area Under the Precision-Recall Curve (AUPR) and includes supporting literature evidence. Across all examined conditions, *ChemBERTaDDI* consistently outperforms Decagon, demonstrating superior predictive accuracy.

To further investigate **Q2)**, we conducted a series of experiments to explore the specific contributions of the chemical structure embeddings within *ChemBERTaDDI*. These experiments utilized all possible feature sets as detailed in Table V. Based on the results, the inclusion of protein features did not enhance the predictive capabilities, suggesting that the *ChemBERTa, Mono* feature set is sufficiently informative for drug-drug interaction prediction. Furthermore, we explored the possibility of using only chemical structure embeddings without incorporating mono clinical information, as shown in Table VI, by comparing two variations of *ChemBERTa* embedding models. The *ChemBERTa-77M-MLM* model, which is pre-trained on 77 million unique SMILES, boosts performance achieving an average F1-score of 0.891. However, as indicated by the above results, using only drug chemical structure embeddings is insufficient.

As shown in Table VII, ChemBERTaDDI (derived from *ChemBERTa-77M-MLM, Mono* features) uses a single neural network trained on 964 side effects. The evaluation metrics indicate that ChemBERTaDDI achieves the highest overall F1-score 0.938 and precision 0.909, while NNPS yields the best recall 0.976. Although Decagon demonstrates a relatively high recall 0.950, its precision 0.771 remains lower than the other two methods.

VII. CONCLUSION

This paper developed a new computational *ChemBERTaDDI* framework for drug-drug interaction prediction that effectively integrates high-dimensional clinical data with transformer-based chemical molecular embeddings. The *ChemBERTaDDI* framework leverages a SMILES-based *ChemBERTa* encoder to extract a 384-dimensional embedding, which is then concatenated with a 595-dimensional mono side-effect feature vector before being passed into a feed-forward neural network. Experiments reveal two main findings: (i) combining molecular structure information from SMILES with drug mono side-effect data substantially improves DDI prediction performance; and (ii) relying solely on chemical embeddings is insufficient. By achieving state-of-the-art results on benchmark dataset, *ChemBERTaDDI* demonstrates strong generalization capability in scenarios where both chemical and clinical features are available. Future work will focus on incorporating additional modalities and exploring deeper transformer model. Datasets and code for ChemBERTaDDI algorithm and framework are freely accessible at https://github.com/anastasiyagromova/ChemBERTaDDI.

REFERENCES

[1] M. Zitnik, M. Agrawal, and J. Leskovec, *Modeling polypharmacy side effects with graph convolutional networks,* in *Bioinformatics,* vol. 34, no. 13, 2018, pp. i457–i466. doi: 10.1093/bioinformatics/bty294

[2] R. Masumshah, R. Aghdam, and C. Eslahchi, *A neural network-based method for polypharmacy side effects prediction,* in *BMC Bioinformatics,* vol. 22, 2021, pp. 1–17. doi: 10.1186/s12859-021-04116-6

[3] K. Huang, C. Xiao, T. Hoang, L. Glass, and J. Sun, *Caster: Predicting drug interactions with chemical substructure representation,* in *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 34, no. 01, 2020, pp. 702–709.

[4] D. Weininger, A. Weininger, and J. L. Weininger, *SMILES (simplified molecular input line entry system),* in *J. Chem. Inf. Comput. Sci.,* vol. 28, 1988, pp. 31–36.

[5] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, *PubChem 2023 update,* in *Nucleic Acids Res.,* vol. 51, no. D1, 2023, pp. D1373–D1380. doi: https://doi.org/10.1093/nar/gkac956

[6] S. Chithrananda, G. Grand, and B. Ramsundar, *ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction,* arXiv preprint arXiv:2010.09885, 2020. [Online]. Available: https://arxiv.org/abs/2010.09885

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,* Proceedings of the NAACL-HLT, vol. 1, pp. 2, 2019, Minneapolis, Minnesota.

[8] Y. Liu, *RoBERTa: A Robustly Optimized BERT Pretraining Approach,* arXiv preprint arXiv:1907.11692, vol. 364, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[9] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, *Data-driven prediction of drug effects and interactions,* in *Science Translational Medicine,* vol. 4, no. 125, 2012, pp. 125ra31–125ra31. doi: 10.1126/scitranslmed.3003061

[10] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, *The SIDER database of drugs and side effects,* in *Nucleic Acids Res.,* vol. 44, no. D1, 2016, pp. D1075–D1079. doi: 10.1093/nar/gkv1075

[11] B. Perozzi, R. Al-Rfou, and S. Skiena, *DeepWalk: Online learning of social representations,* in *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining,* vol. 1, no. 1, 2014, pp. 701–10. doi: https://doi.org/10.1145/2623330.2623732. arXiv:1403.6652.

[12] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, *Tensors for data mining and data fusion: Models, applications, and scalable algorithms,* in *ACM Trans Intell Syst Technol,* vol. 8, no. 2, 2016, pp. 1–44. doi: https://doi.org/10.1145/2915921

[13] M. Nickel, V. Tresp, and H. P. Kriegel, *A three-way model for collective learning on multi-relational data,* in *Proceedings of the 28th international conference on machine learning, ICML 2011,* vol. 1, 2011, pp. 809–16.

[14] B. Serban, Z. Panti, M. Nica, M. Pleniceanu, M. Popa, R. Ene, and C. Cîrstoiu, *Statistically based survival rate estimation in patients with soft tissue tumors,* in *Rom J Orthop Surg Traumatol,* vol. 1, no. 2, 2019, pp. 84–89. doi: https://doi.org/10.2478/rojost-2018-0085

[15] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay, and F. Bray, *Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis,* in *Lancet Glob Health,* vol. 8, no. 2, 2020, pp. 191–203. doi: https://doi.org/10.1016/S2214-109X(19)30482-6

[16] A. Januszewicz, T. Guzik, A. Prejbisz, T. Mikołajczyk, G. Osmenda, and W. Januszewicz, *158_Prejbisz_ONLINE,* in *PALSKIE,* vol. 126, no. 1, 2016, pp. 86–93.

[17] E. C. Evans, K. A. Matteson, F. J. Orejuela, M. Alperin, E. M. Balk, S. El-Nashar, J. L. Gleason, C. Grimes, P. Jeppson, C. Mathews, T. L. Wheeler, and M. Murphy, *Salpingo-oophorectomy at the time of benign hysterectomy: a systematic review,* in *Obstet Gynecol,* vol. 128, no. 3, 2016, pp. 476–85. doi: https://doi.org/10.1097/AOG.0000000000001592

[18] R. J. Flanagan and L. Dunk, *Haematological toxicity of drugs used in psychiatry,* in *Human Psychopharmacology: Clinical and Experimental,* vol. 23, no. S1, 2008, pp. S27–S41.