



A Neuro-Symbolic AI Approach to Identifying Potent DPP-4 Inhibitors for Diabetes Treatment

Delower Hossain, Ehsan Saghapour and Jake Chen

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

March 5, 2024

A Neuro-symbolic AI Approach to Identifying Potent DPP-4 Inhibitors for Diabetes Treatment

Delower Hossain^{1,2}, Ehsan Saghapour², Jake Y. Chen^{2*}

¹ Department of Computer Science, The University of Alabama at Birmingham, AL 35294, USA

² Department of Biomedical Informatics and Data Science, School of Medicine, The University of Alabama at Birmingham, AL 35205, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Diabetes Mellitus (DM) is the most widespread category within metabolic disorders and finding a potential therapeutic Dipeptidyl peptidase-4 (DPP-4) inhibitor agent is crucial. This study aims to uncover the efficacy of DPP-4 inhibitors utilizing a Neuro-symbolic approach, a new branch of artificial intelligence, and RoBERTa (NLP-transformer model). We employ the LTN (Logical Tensor Networks), a novel machine learning technique, procuring data from ChEMBL and BindingDB databases. After curation, each database consists of 3918 and 3285 for the classification task. We experimented with 14 molecular feature extraction approaches, including descriptors fingerprints such as AtomPairs2DCount, AtomPairs2D, EState CDKextended, CDK, CDKgraphonly, KlekotaRoth, KlekotaRothCount, MACCS, Substructure, PubChem, SubstructureCount, PubChemPy, Lipinski's Rule (RDKit). The LTN model yields a groundbreaking Accuracy incorporating an CDKextended fingerprint of 0.978, an F1-score of 0.978, an ROC AUC of 0.966, and an MCC of 0.931. Conversely, RoBERTa resulted in 0.9493 Accuracy, F1 score of 0.9491, ROC AUC 0.9174, and MCC 0.8423. Our findings show that integrating the neuro-symbolic strategy (neural network-based learning and symbolic reasoning) has immense potential to discover the drugs that have the potential to inhibit diabetes mellitus and classify biological activities that inhibit it. Overall, the LTN model exhibits interpretable reasoning and learning, which enables the discovery of novel insights into type 2 diabetes mellitus inhibitors.

Contact: jakechen@uab.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Diabetes, also identified as Diabetes Mellitus (DM), is a chronic metabolic syndrome characterized by elevated levels in the bloodstream, and it's considered a global epidemic. As per the World Health Organization (WHO) Report 2019, DM. has been included in the list of the top ten leading causes of mortality [1], reporting an estimated 1.6 million people globally died because of diabetes [2], page 22]. In the United States, diabetes has a substantial role in one of the primary causes of death. According to the national health institution Centers for Disease Control and Prevention (CDC), it is stated that approximately 37.3 million individuals, constituting around 11.3% of the entire U.S. population, are affected by diabetes, and total medical costs, lost wages are \$327 billion [3]. In addition to that, there are several crucial risk factors and health complications involved—for instance, blindness, stroke, kidney failure, and heart disease might occur due to diabetes [3].

Diabetes Mellitus primarily can be classified as Type-1 Diabetes Mellitus (T1DM) and Type-2 Diabetes Mellitus (T2DM). The severity and

impact of T2DM surpasses T1DM, with over 90% of diabetes cases attributed to T2DM and Dipeptidyl peptidase-4 (DPP-4), Glucagon-like peptide-1 (GLP-1) inhibitors are imperative therapeutic pathway for T2DM. Dipeptidyl Peptidase-4 is an enzyme involved in glucose metabolism. In diabetes, medications known as DPP-4 inhibitors are treated to help regulate blood sugar levels by inhibiting the activity of this enzyme, and this inhibitor is recognized as FDA-approved (Appendix A). Current DPP-4 inhibitors have several adverse effects. However, several studies have used AI methods to reveal the potential antidiabetic drugs. To illustrate, conventional artificial intelligence (AI) technologies began to be incorporated into diabetes management and research in the early 2000s. Researchers initiated the exploration of AI techniques, encompassing machine learning and data mining, to delve into the analysis of diabetes-associated data. In recent years, AI has become more widely integrated into numerous elements of diabetes management, including glucose prediction, insulin dose recommendations, early detection, new drug design and development utilizing millions of compounds data, identifying new

inhibitors and complex relationships among gene, protein, and cost reduction developing building AI-based Clinical Decision-Support System (CDSS).

Studies identified various QSAR (Quantitative Structure-Activity Relationship) modeling that employed machine learning strategies, auto-QSAR modeling incorporating AI algorithms, and models to predict the DPP-4 inhibitor’s efficacy. For instance, Conv1D-LSTM [44], Random Forest, Bayesian, Support Vector Machine, Rotation Forest, XGBoost, Recursive Partitioning, Generalized Linear Model, Gradient Boosting Machine, PSO, Rotation Forest, Genetic function approximation (GFA), Ada-boost, Bagging, ExtraTree, K Nearest Neighbor, Ridge, ElasticNet Model, SGD, Transformer, Multi Linear Regression, Deep Neural Network, and Artificial Neural Network [4]-[9]. Although they have shown outstanding performance, none of the conventional AI strategies were integrated data and knowledge-driven approaches. Although deep learning, a black box system, contributions are breakthroughs, various flaws exist, such as poor reasoning ability, massive data consumption, and lack of explainability, transparency, and interpretability.

In recent years, the new dimension of AI emergence, named Neuro-symbolic (NeSy) approaches [10]-[13], has gained immense attention for their ability to blend the métiers of neural networks and symbolic reasoning, leading to more interpretable and insightful predictions. Most importantly, reasoning, explainability, and interpretability is crucial in healthcare. Our study uncovered the following innovative NeSy models that were implemented with healthcare & non-healthcare domain: (Diabetes) FES [14], (Protin Function) MultipredGO [15], (Gene Sequence) KBANN [16], (Diabetic Retinopathy) ExplainDR [17], (Link Prediction) NeuralLP [18], (Ontology) RRN [19], NSRL [20], Neuro-Fuzzy [21], FSKBANN [22], DeepMiRGO [23], NS-VQA [24], DFOL-VQA [25], LNN [26], NofM [27], PP-DKL [28], FSD [29], CORGI [30], NeurASP [31], XNMs [32], Semantic loss [34], NS-CL [35], LTN [36].

However, this study aimed to investigate the role of the hybrid (LTN) and advanced pre-trained language model RoBERTa [37] in the DPP-4 bioactivity prediction. Specifically, finding potential therapeutic DPP-4 Inhibitor agents for type 2 Diabetes Mellitus. and developing a molecular compound classification predictive Neuro-symbolic model utilizing more diverse compound instances. To achieve this goal, we determined the ChEMBL and BindingDB databases with 14 distinct molecular feature extraction approaches, including descriptors fingerprints (PaDEL [41], RDKit [42], and PubChemPy [43]). The LTN model gained a ground-breaking accuracy, incorporating an PaDEL-CDKextended fingerprint of 0.9778 compared to RoBERTa, which had 0.9493 accuracy. Overall, the finding of this study exhibits that integrating the Neuro-symbolic strategy (neural network-based learning and symbolic reasoning) has immense potential in predicting and classifying biological activities.

The significant contribution of this study is that; 1) we built a highly scalable, robust AI predictive model with immense accuracy improvement for DM DPP-4 inhibitors. 2) A novel representation integrating data and rules (Knowledge) for DPP-4 inhibitor bio-activity classification 3) Acquired and utilized more diverse compound datasets and fingerprints than previous studies. 4) The RoBERTa (meta-AI) pre-trained model was also experimented with to compare the performance with LTN for DPP-4 chemical substance classification.

The remainder of the article elaborated as Section II outlines the Methodology. Afterward, the simulation result of this study is presented in Section III. Finally, the Conclusion and Future direction are given in Section IV.

2 Materials & Methods

This segment presents a set of methodology procedures to determine the performance of Logic Tensor Networks (LTN) [36] and an advanced language model known as RoBERTa [37] that we employed on ChEMBL and BindingDB Dataset related to DPP-4 inhibitors. LTN framework was retained to address the limitations of traditional deep learning systems, which are not well-suited to tasks that require reasoning, interpretability, symbolic manipulation, and knowledge integration. This section covers the entire pipeline, including the materials, data preprocessing, feature extraction, simulation environment, network architecture, LTN knowledge-based Setting, the training and inferencing phases, and the evaluation metrics used to measure the model’s performance.

2.1 Dataset

2.1.1 Data source and acquisition

The study utilized two publicly available databases: ChEMBL [38] & BindingDB [39]. The ChEMBL Database contains more than 2 million compounds. We retrieved a total of 5098 molecular canonical SMILES related to DPP-4 inhibitor with the target organism Homo Sapiens using ID: ChEMBL284 and standard type IC50 (Table 1). The data was extracted using the ChEMBL Database’s Python API (chembl_webresource_client). In addition, we procured 7331 data from BindingDB manually using DPP-4 string keywords from their official site.

Table 1: ChEMBL and BindingDB collected data distribution

Inhibitors	Content	ChEMBL	BindingDB	Total
DPP-4	Raw	5098	7331	12429
	After Curated	3918	3285	7203
	Final (active classes)	3080	2659	5739
	Final (inactive Classes)	838	626	1464

2.1.2 Data Preprocessing and Descriptors Calculation

We collected subsets from raw data focused on the IC50 biological activity standard value, Canonical SMILES, and similar steps for the BindingDB (Fig. 1). Afterwards, we combined both datasets, removed duplicates, and conducted preprocessing to eliminate irrelevant information. Standard value (IC50) is categorized into two groups: "Inactive = 0" and "active = 1". Those with less than 1000 nM were considered active, and those with values more than 10,000 nM were inactive. Intermediate values are those that fall between 1,000 and 10,000 nM. The Intermediate classes needed to be disregarded [40].

We computed several descriptors/fingerprints during the feature extraction phase, particularly by leveraging the PaDEL Descriptor tool. A total number of 12 descriptors gather (Table 2) and accumulate their corresponding attributes. Additional fingerprints are obtained using PubChemPy and RDKit (Lipinski’s rules). We experimented with 14 types of descriptors and post-feature extraction; we applied Standardization.

Fig. 1 The Fig illustrates the sample of final preprocess data.

	Name	smiles	standard_value	standard_type	source	classes
0	CHEMBL99558	N[C@@H](C1CC1)C(=O)N1CCCC1	217000.00	IC50	CHEMBL	0
1	CHEMBL443622	C[C@@H](N)C(=O)N1CCCC1	41000.00	IC50	CHEMBL	0
2	CHEMBL403892	O=C(C@H)1CCCN1)N1CCCC1	15000.00	IC50	CHEMBL	0
3	CHEMBL328655	S=C(C1CCCN1)N1CCCC1	500000.00	IC50	CHEMBL	0
4	CHEMBL328795	NC(=O)CC(N)C(=O)N1CCCC1	188000.00	IC50	CHEMBL	0
...
7198	228304	N[C@@H]1C[C@@H](COC1c1cc(F)c1)F1C2c1nc2C	1.90	IC50	BindingDB	1
7199	50306956	O=C(C[C@@H]1C[C@@H](CN1)NCCc1ccccc1)N1CC[C@H]1CN	1.90	IC50	BindingDB	1
7200	50434742	COCOC1c1cc2nc(sc2c1)C1(CCOCC1)NCC(=O)C[C@H]1CN	1.90	IC50	BindingDB	1
7201	50497008	Cc1cc(nc2c3CN(Cc3m12))C[C@@H]1CC[C@H](C[C@@H]1)N	1.90	IC50	BindingDB	1
7202	968880	N[C@@H]1C[C@@H](C)[C@@H]1c1cc(F)ccc1F1n1cc1n1N	1.91	IC50	BindingDB	1

7203 rows x 6 columns

Fig. 2: CDKextended fingerprint dataset that was collected based on corresponding smiles.

Name	ExtFP1	ExtFP2	ExtFP3	ExtFP4	ExtFP5	...	ExtFP1019	ExtFP1020	ExtFP1021	ExtFP1022	ExtFP1023	ExtFP
0	293465	0	0	0	0	0	0	0	0	0	0	0
1	51320348	1	0	1	0	1	...	0	0	0	0	0
2	50381388	0	0	1	0	0	...	0	0	0	0	0
3	50439684	0	0	0	1	0	...	0	0	0	0	0
4	225601	0	0	0	0	0	...	0	0	0	0	0

5 rows x 1025 columns

Table 2: List of Descriptors and no. of features	
PADEL Descriptors [41]	
Name of the Fingerprints/ Descriptors	No of Features
AtomPairs2DCount	780
AtomPairs2D	780
CDKextended	1024
CDK	1024
CDKgraphonly	1024
EState	79
KlekotaRothCount	4860
KlekotaRoth	4860
MACCS	166
PubChem	881
SubstructureCount	307
Substructure	307
RDKit [42]	
Lipinski	4
PubChem [43]	
PubChemPy	39

2.2 LTN Classification model:

LTNs were architected using two key components: a logic component and a neural network. The visual architecture of the classification model can be found in Appendix C. The logical mechanism contains a set of axioms or rules (explained in detail in the Knowledge-based setting); during the backpropagation, weights are updated based on LTN loss functions, which are calculated based on hypotheses. In this work, we built the DPP-4 LTN Classifier Constructing MLP, which consists of 4 layers and input units (1024) since CDKextended descriptors number of features are 1024, hidden layers units (1024,512,256), ReLU activation, batch size 32, optimizer Adam with learning rate 0.00001, seed 42. LTN knowledge-based setting and other significant components are discussed in the following section.

2.2.1 LTN Knowledge Base Setting

Knowledge-based was defined based on domains (features and labels), variables, Constants (classes), and Predicates (p). The true potential of

predicate logic $p(x, l)$ Neural-Symbolic systems, specifically Logic Tensor Networks, lie in their ability to represent and reason over complex logical relationships having domain-specific knowledge. The benefit of predicate logic is that it enables the training of neural networks with the domain knowledge (i.e., in this case, First-order logic and Real Logic). In addition, reasoning and interpretability are achievable with predicate logic. The concept of building this structure was adopted from the official LTN framework [36]. Table 3: denoting the significant components of LTN knowledge-based setting as well as representing the learning and loss function and detail setting up components can be found in Appendix B sections.

Table 3: LTN Knowledge-based Setting for DPP-4 Classification	
Contents	Classification
Define Axioms	<ul style="list-style-type: none"> $\forall x_A, p(x_A, l_A)$: all the examples of class A ($Inactive = 0$) should have a label l_A $\forall x_B, p(x_B, l_B)$: all the examples of class B ($Active = 1$) should have a label l_B
Axioms (rules, knowledge base)	$\mathcal{K} = \forall x_A P(x_A, l_A), \forall x_B P(x_B, l_B)$
SatAgg is given by	$SatAgg_{\phi \in \mathcal{K}} \mathcal{G}_{\theta, x \leftarrow D}(\phi)$
Learning & Loss	$L = \left(1 - SatAgg_{\phi \in \mathcal{K}} \mathcal{G}_{\theta, x \leftarrow D}(\phi) \right)_{\phi \in \mathcal{K}}$

Note: This table was developed inspired by the official LTN [36]

Here;

SatAgg is defined using the $pMeanError$ aggregator.

$$pME(u_1, \dots, u_n) = 1 - \left(\frac{1}{n} \sum_{i=1}^n (1 - u_i)^p \right)^{\frac{1}{p}} \geq 1 \quad (1)$$

$$SatAgg_{\phi \in \mathcal{K}} \mathcal{G}_{\theta, x \leftarrow D}(\phi)$$

- SatAgg: This stands for "Satisfaction Aggregator"
- $\phi \in \mathcal{K}$: This part indicates that ϕ (phi) belongs to the set \mathcal{K} . ϕ is often used to represent a predicate.
- $\mathcal{G}(\theta)$: This is denoted by grounding (\mathcal{G}) with parameters θ . θ represents a set of parameters or weights in a model.
- $x \leftarrow D$: D the data set of all examples
- the input to the functions SatAgg and $\mathcal{G}(\theta)$.

2.3 RoBERTa Classification Model

Keras NLP packages were used to develop the DPP-4 Finetuned RoBERTa[37] (base model) classifier. Here, we define the model hyperparameters, vocabulary size 50265, num layers 12; num heads 12, hidden dim 768, intermediate dim 3072, dropout 0.1, max sequence length 512, optimizer RMSprop with learning rate 0.00005, and batch size 16.

2.4 Model Training and Validation Phase

LTN was trained and tested using TensorFlow 2.15.0 Python 3.10.12 on Google Colab laboratory. Conversely, pre-trained RoBERTa was trained on Kaggle for 6.5 hours (Table 5) with GPUP100. For training and testing, we did partition 80: 20 ratios over 310 epochs during LTN training, while RoBERTa was trained 80 epochs. We stopped here as overfitting occurred. The following metrics, such as Accuracy, F-score (F), ROC AUC Score, and Mathew Correlation Coefficient (MCC), were used to assess the model's performance, and the misclassified classes can be seen in Fig.

4. Additionally, both models’ comparison efficiencies and other parameters are presented in table 5.

Equation 1: Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Equation 2: F1 Score

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Equation 3: ROC AUC Score

$$\text{ROC AUC} = \int_0^1 \text{TPR} \, d(\text{FPR}) \quad (4)$$

Equation 4: MCC

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (5)$$

3 Results

In this section, we illustrate the performance of two approaches, LTN (knowledge Integration into the neural network) and RoBERTa, an advanced language model, for revealing DPP-4 potential inhibitors. The

Table 4 shows LTN model achieved the highest Accuracy (0.9778), F1 score (0.9778), ROC AUC (0.9657), and MCC (0.9315) with the CDKextended 1024 features (Fig. 3). Notably, among the three descriptors (PaDEL Tool, PubChemPy, and RDKit) PaDEL consistently performed well with rules and neural network integration system (LTN), except AtomPairs2D and followed by the PubChemPy descriptors (Accuracy: 0.8820, F1 score: 0.8340, ROC AUC: 0.8688, MCC: 0.6777). Although RDKit (Lipinski) took last place. The RoBERTa model, on the other hand, achieved a competitive Accuracy (0.9493) and F1 score (0.9491) ROC AUC (0.9174), MCC (0.8423) using only tokenized features that indicated the potential of natural language processing approaches for this DPP-4 classification. However, training time was much higher (Table 5) than LTN. In addition, Table 6 showcases five random sample predictions from unseen test data. LTN accurately classified all, while RoBERTa misclassified one. The total misclassified report and ROC AUC curve can be visible in Figs. 4 and 6.

Table 4: LTN & RoBERTa DPP-4 Classification Result Summary

Model	Category	Descriptors	Features	Accuracy	F1 Score	ROC AUC	MCC	
LTN	PaDEL Tool	CDKextended	1024	0.9778	0.9778	0.9657	0.9315	
		CDK	1024	0.9722	0.9723	0.9610	0.9150	
		CDKgraphonly	1024	0.9702	0.9702	0.9546	0.9080	
		KlekotaRothCount	4860	0.9695	0.9535	0.9592	0.9071	
		KlekotaRoth	4860	0.9653	0.9474	0.9553	0.8952	
		PubChem	881	0.9625	0.9436	0.9549	0.8880	
		MACCS	166	0.9604	0.9398	0.9459	0.8798	
		AtomPairs2DCount	780	0.9500	0.9502	0.9254	0.8466	
		SubstructureCount	307	0.9389	0.9102	0.9312	0.8234	
		Substructure	307	0.9146	0.8746	0.8943	0.7522	
		EState	79	0.9063	0.9097	0.8980	0.7400	
		AtomPairs2D	780	0.8397	0.7902	0.8511	0.6106	
		PubChem	PubChemPy	33	0.8820	0.8340	0.8688	0.6777
		RDKit	Lipinski	4	0.7627	0.7796	0.7392	0.4133
		RoBERTa	Tokenize	RoBERTa Tokenization	768	0.9493	0.9491	0.9174

Table 5: LTN & RoBERTa Performance and Efficiency Comparison

Model	Input	Features	Training Times	Trainable Params	Total Params	Accuracy
LTN	Fingerprint (CDKextended)	1024	15 minutes	1,706,242	1,706,242	0.9778
RoBERTa	SMILES (Tokenization)	768	6.5 hours	592,130	124,644,866	0.9493

Table 5 demonstrates the comparison performance and efficiency of two models, LTN and ROBERTa, for predicting the properties of molecules associated with DM inhibitors. The LTN model consumed a fingerprint representation of molecules as input, while the ROBERTa (pertained finetune method) model uses a SMILES representation as a string and afterward processes RoBERTa tokenization. The LTN model has a higher accuracy (0.9778) than the RoBERTa model (0.9493) NLP-transformer model. It took approximately 15 minutes to train the LTN model and 6.5 hours to train the RoBERTa model. The RoBERTa has fewer trainable parameters than the LTN model since it was a pre-trained NLP; therefore, the parameters are immense.

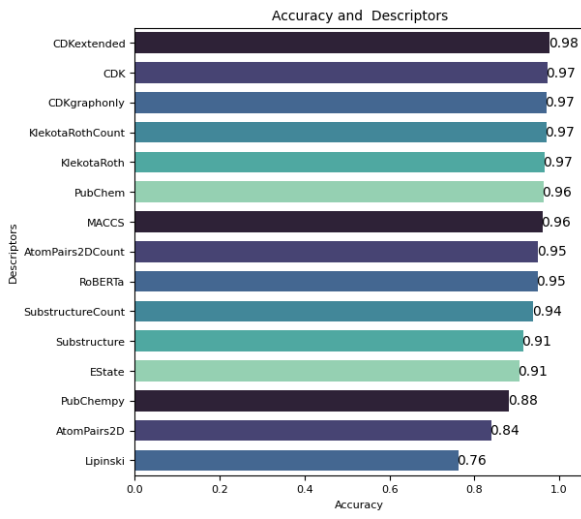


Fig 3: Descriptors and Accuracy (Rounded up)

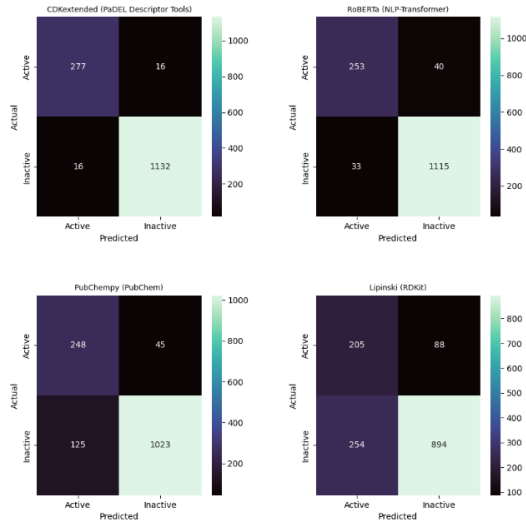


Fig 4: Confusion Matrix for LTN, RoBERTa

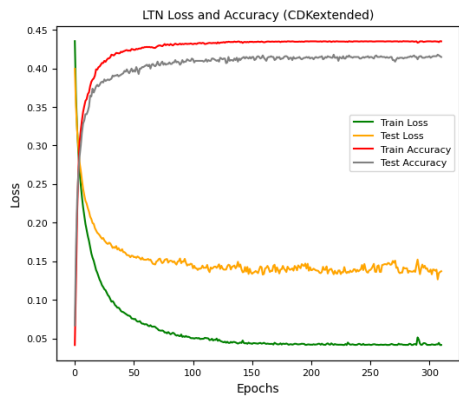


Fig 5: LTN Loss and Accuracy Graph (Training and Testing)

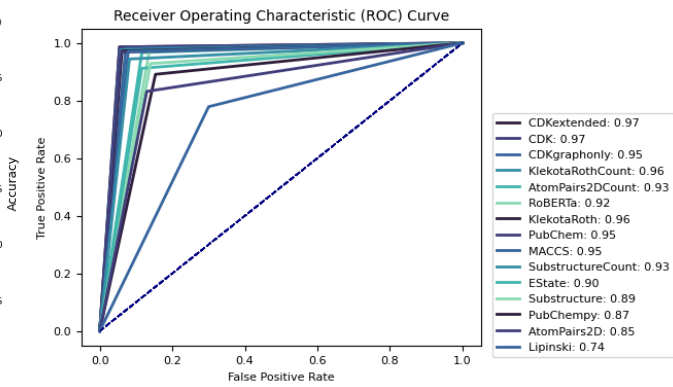
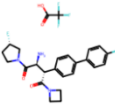
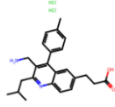
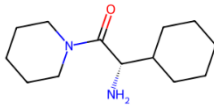
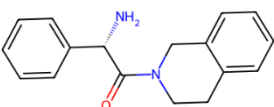
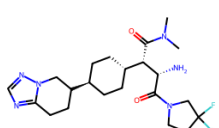


Fig 6: LTN and RoBERTa ROC AUC Curve

Table 6: LTN and RoBERTa model five random prediction samples based on the test dataset

S.No	Source ID	Smiles (Input)	Molecule (2D Structure)	Actual	RoBERTa	LTN
					Predicted (Output)	Predicted (Output)
1.	CHEMBL18 8043	<chem>N[C@H](C(=O)N1CC[C@H](F)C1)[C@H](C(=O)N1CC1)c1ccc(-c2ccc(F)cc2)cc1.O=C(O)C(F)(F)F</chem>		Active	Active	Active
2.	CHEMBL32 16493	<chem>Cc1ccc(-c2c(CN)c(CC(C)C)nc3ccc(CC(=O)O)cc23)cc1.C1Cl</chem>		Active	Active	Active
3.	22091	<chem>N[C@@H](C1CCC(CC1)C(=O)N1CCC(CC1</chem>		Inactive	Inactive	Inactive
4.	CHEMBL36 1758	<chem>N[C@H](C(=O)N1CCc2ccccc2C1)c1ccccc1</chem>		Inactive	Active	Inactive
5.	CHEMBL23 5199	<chem>CN(C)C(=O)[C@@H]([C@H]1CC[C@H](C2CCc3ncnn3C2)CC1)[C@H](N)C(=O)N1CCC(F)(F)C1</chem>		Active	Active	Active

Note: The RoBERTa model had input as SMILE string where LTN input is fingerprint/descriptor (i.e CDKextended)

Overall, this study suggests that the LTN model with the PaDEL Tool, CDKextended, is a promising, more robust, and efficient approach for predicting the properties of molecules classification of DPP-4 adverse and potential inhibitors. The findings also recommend using pre-trained transformer-based natural language approaches, such as the RoBERTa model, which also shows promise for DPP-4 classification. However, further research is needed to improve its performance, and the Neuro-symbolic approach is more scalable and robust than conventional AI methods (Table 5).

Conclusion

Diabetes Mellitus is a vital global health concern, and discovering effective chemical substances is crucial to tackling this epidemic. This study explored the therapeutic potential of DPP-4 inhibitors employing a novel approach called the LTN (Neuro-symbolic AI) that integrates domain-specific knowledge into neural networks. The study is a pioneer in applying Neuro-symbolic strategy in the DM domain and provides new insights showing groundbreaking performance for revealing DPP-4 potential inhibitors. The root cause of achieving such performance could be upholding learning and reasoning principles and training neural networks with rules. Furthermore, we experimented with the RoBERTa, an NLP pre-trained Transformer model, which also attained prominent Accuracy, although training performance and consumption of the resources are higher than LTN.

In conclusion, the findings of this study demonstrated that LTN is among the state-of-the-art models for uncovering potential DPP-4 inhibitors. We aim to deploy the model within a real-time prediction application to identify the right therapeutic agent that could promptly benefit ML practitioners, academics, and industry researchers. However, an

ideal next step could involve integrating additional potential Neuro-symbolic strategies, such as Semantic Loss, DeepProbLog on GLP-1, IDO, and PTP1B DM inhibitors extracting a variety of new descriptors, and fingerprints with different datasets (PubChem, Protein Data Bank) focusing regression task.

Acknowledgments

The authors acknowledged the biomedical data science infrastructure and staff support provided by the UAB U-BRITE program.

Conflict of Interest

The authors declare that they have no conflicts of interests in this work.

Author Contributions

The author, Delower Hossain, designed, implemented, and wrote the manuscripts, and Ehsan Saghapour worked together to edit and review. Jake Chen has been personally and actively guided as project administrator.

Funding

The work is partly supported by NIH grant R21DK129968 and research startup funding awarded to Dr. Jake Chen.

References

- [1] The WHO, "The top 10 causes of death" <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death#:~:text=Of%20the%2056.9%20million%20deaths%20world-wide%20in%202016%2C,of%20death%20globally%20in%20the%20last%2015%20years.>
- [2] World Health Statistics 2020, <https://apps.who.int/iris/bitstream/handle/10665/332070/9789240005105-eng.pdf>.
- [3] CDC, National diabetes report 2021, <https://www.cdc.gov/diabetes/data/statistics-report/index.html>
- [4] A. Yanuar, O. Hermansyah, A. Bustamam, "QSAR Modeling for Prediction of pIC50 DPP-4 Inhibitors with Machine Learning Method", 2019, conference
- [5] O. Hermansyah, A. Bustamam, A. Yanuar, "Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure-activity relationship-based artificial intelligence and molecular docking of hit compounds", 2021 Elsevier
- [6] O.A.Ojo, A.B.Ojo, C.E.Okolie, "Elucidating the interactions of compounds identified from Aframomum melegueta seeds as promising candidates for the management of diabetes mellitus: A computational approach," 2021, Elsevier
- [7] I. Kurniawan, R. R. Septiawan, B.H. Prakoso, "DPP IV Inhibitors Activities Prediction as An Antidiabetic Agent using Particle Swarm Optimization-Support Vector Machine Method," 2022, journal.
- [8] A. Bustamam, H. Hamzah, N. A. Husna, et al., "Artificial intelligence paradigm for ligand-based virtual screening on the drug discovery of type 2 diabetes mellitus", 2021, Springer
- [9] A. Basiru, O. Iwaloye, O. Owolabi, et al., "Screening of potential antidiabetic phytochemicals from Gongronema latifolium leaf against therapeutic targets of type 2 diabetes mellitus: multi-targets drug design", 2022, Springer.
- [10] D. Yu, B. Yang, D. Liu, et al., "Recent Advances in Neural-symbolic Systems: A Survey," ArXiv, 2022
- [11] Wenguan Wang, Yi Yang, Fei Wu "Towards Data-and Knowledge-Driven Artificial Intelligence: A Survey on Neuro-symbolic Computing", Harvard ADS, 2022 <https://ui.adsabs.harvard.edu/abs/2022arXiv221015889W/abstract>
- [12] M. Hassan, H. Guan, A. Melliou, et al., "Neuro-symbolic Learning: Principles and Applications in Ophthalmology", arXiv, 2022.
- [13] Zachary Susskind, Bryce Arden, Lizy K. John, "Neuro-symbolic AI: An Emerging Class of AI Workloads and their Characterization", DBLP, 2021 <https://dblp.org/rec/journals/corr/abs-2109-06133.html>
- [14] H. Thakkar, V. Shah, H. Yagnik, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," Elsevier, 2021
- [15] S. J. Giri, P. Dutta, P. Halani, et al., "Multipredgo: Deep Multi-Modal Protein Function Prediction By Amalgamating Protein Structure, Sequence, And Interaction Information," PubMed, 2021
- [16] G. G. Towell, J. W. Shavlik, "Knowledge-Based Artificial Neural Networks," ScienceDirect, 1994
- [17] S. I. Jang, M. J. A. Girard, A. H. Thiery, "Explainable Diabetic Retinopathy Classification Based On Neural-Symbolic Learning," 2021.
- [18] F. Yang, Z. Yang, W. W. Cohen, "Differentiable Learning of Logical Rules for Knowledge Base Reasoning," Neurips, 2017
- [19] P. Hohenecker, T. Lukasiewicz, "Ontology Reasoning with Deep Neural Networks," IJCAI, 2020
- [20] Z. Yang, A. Ishay, J. Lee, "Neurasp: Embracing Neural Networks into Answer Set Programming," IJCAI, 2020 (NSRL)
- [21] P. Kora, K. Meenakshi, K. Swaraja, "Detection of Cardiac arrhythmia using fuzzy logic," Elsevier, 2019
- [22] R. Maclin, J. W. Shavlik, "Refining Algorithms With Knowledge-Based Neural Networks: Improving The Chou-Fasman Algorithm For Protein Folding," ACM, 1992.
- [23] J. Wang, J. Zhang, Y. Cai, et al., "Deepmir2Go: Inferring Functions Of Human MicroRNAs Using A Deep Multi-Label Classification Model," PubMed, 2019
- [24] K. Yi, J. Wu, C. Gan et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding," Harvard University, 2018
- [25] S. Amizadeh, H. Palangi, A. Polozov, et al., "Neuro-symbolic Visual Reasoning: Disentangling Visual from Reasoning"-Microsoft, 2020
- [26] R. Riegelet, A. Gray, F. Luu's al, "Logical Neural Networks," 2020
- [27] G. G. Towell, J. W. Shavlik, "Interpretation of Artificial Neural Networks: Mapping Knowledge-Based Neural Networks into Rules" 1991
- [28] A. Lavin, "Neuro-symbolic Neurodegenerative Disease Modeling as Probabilistic Programmed Deep Kernels," Springer, 2021
- [29] K. Dobosz, W. Duch, "Fuzzy Symbolic Dynamics For Neurodynamical Systems," Springer, 2008.
- [30] Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, "Conversational Neuro-symbolic Commonsense Reasoning," AAAI, 2021
- [31] Z. Yang, A. Ishay, J. Lee, "Neurasp: Embracing Neural Networks Into Answer Set Programming," Ijcai, 2020
- [32] J. Shi, H. Zhang, J. Li, "Explainable and Explicit Visual Reasoning over Scene Graphs," IEEE Xplore, 2019
- [33] K. K. Teru, E. Denis, W. L. Hamilton, "Inductive Relation Prediction by Subgraph Reasoning," ICML, 2020
- [34] J. Xu, Z. Zhang, T. Friedman, et al., "A Semantic Loss Function for Deep Learning with Symbolic Knowledge," MLR Press, 2018
- [35] J. Mao, C. Gan, P. Kohli, et al., "The Neuro-symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision," DBLP, 2023
- [36] S. Badreddine, A. A. Garcez, L. Serafini, et al. "Logic Tensor Networks", 2020.
- [37] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv, 2019
- [38] A. Gaulton, L.J. Bellis, A.P Bento, et al., "ChEMBL: a large-scale bioactivity database for drug discovery," OUP, <https://academic.oup.com/nar/article/40/D1/D1100/2903401?login=false>, 2011
- [39] M.K. Gilson, T. Liu, M. Baitaluk, et al., "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology", <https://academic.oup.com/nar/article/44/D1/D1045/2502601>, 2015
- [40] C. Selvaraj, S.K. Tripathi, K.K. Reddy, et al., "Tool development for Prediction of pIC50 values from the IC50 values-A pIC50 value calculator", indianjournals, 2011
- [41] Twenty-five descriptors: 12 PaDEL XML Files: [padel/finger-prints_xml.zip at main · dataprofessor/padel · GitHub](#)
PaDEL Main Function: [GitHub - ecri/padelpy: A Python wrapper for PaDEL-Descriptor software](#)
- [42] RDKit: <https://www.rdkit.org/docs/Install.html>
- [43] PubChemPy: <https://pubchempy.readthedocs.io/en/latest/guide/introduction.html>

- [44] A. Ulfa, A. Bustamam, A. Yanuar et al., "Model QSAR Classification Using Conv1D-LSTM of Dipeptidyl Peptidase-4 Inhibitors", IEEE, 2021
- [45] FDA approved Dipeptidyl Peptidase IV (DPP IV) Inhibitors <https://www.ncbi.nlm.nih.gov/books/NBK542331/#:~:text=DPP%2D4%20inhibitors%2C%20known%20as,sax-agliptin%2C%20linagliptin%2C%20and%20alogliptin>
- [46] Drug Bank DPP-4 Inhibitors <https://go.drugbank.com/categories/DBCAT002653>
- [47] Wikipedia DPP-4 Inhibitors https://en.wikipedia.org/wiki/Dipeptidyl_peptidase-4_inhibitor

Appendix A: A list of FDA, E.U., EMA (European Medicines Agency), JAPAN, and KOREN body approved DPP-4 inhibitor's structure and respective 2D compound structures images as below.

ChEMBL ID	Target	Approved Body	Smiles	Ref
CHEMBL376359	Alogliptin	FDA	<chem>Cn1c(=O)cc(N2CCC[C@@H](N)C2)n(Cc2ccccc2C#N)c1=O</chem>	[45]
CHEMBL1929396	Anagliptin	Japan	<chem>Cc1cc2ncc(C(=O)NCC(C)(C)NCC(=O)N3CCC[C@@H]3C#N)cn2n1</chem>	[47]
CHEMBL3707235	Gemigliptin	Korea	<chem>N[C@@H](CC(=O)N1CCc2c(nc(C(F)(F)F)nc2C(F)(F)F)C1)CN1CC(F)(F)C</chem> <chem>CC1=O</chem>	[47]
CHEMBL237500	Linagliptin	FDA	<chem>CC#CCn1c(N2CCC[C@@H](N)C2)nc2c1c(=O)n(Cc1nc(C)c3ccccc3n1)c(=O)n2C</chem>	[45][47]
CHEMBL385517	Saxagliptin	FDA	<chem>N#C[C@@H]1C[C@@H]2C[C@@H]2N1C(=O)[C@@H](N)C12CC3CC(C</chem> <chem>C(O)(C3)C1)C2</chem>	[45][47]
CHEMBL1422	Sitagliptin	FDA	<chem>N[C@@H](CC(=O)N1CCn2c(nnc2C(F)(F)F)C1)Cc1cc(F)c(F)cc1F</chem>	[45],[47]
CHEMBL2147777	Teneligliptin	Japan	<chem>Cc1cc(N2CCN([C@@H]3CN[C@H](C(=O)N4CCSC4)C3)CC2)n(-</chem> <chem>c2ccccc2)n1</chem>	[45]
CHEMBL142703	Vildagliptin	EMA	<chem>N#C[C@@H]1CCCN1C(=O)CNC12CC3CC(CC(O)(C3)C1)C2</chem>	[45],[47]

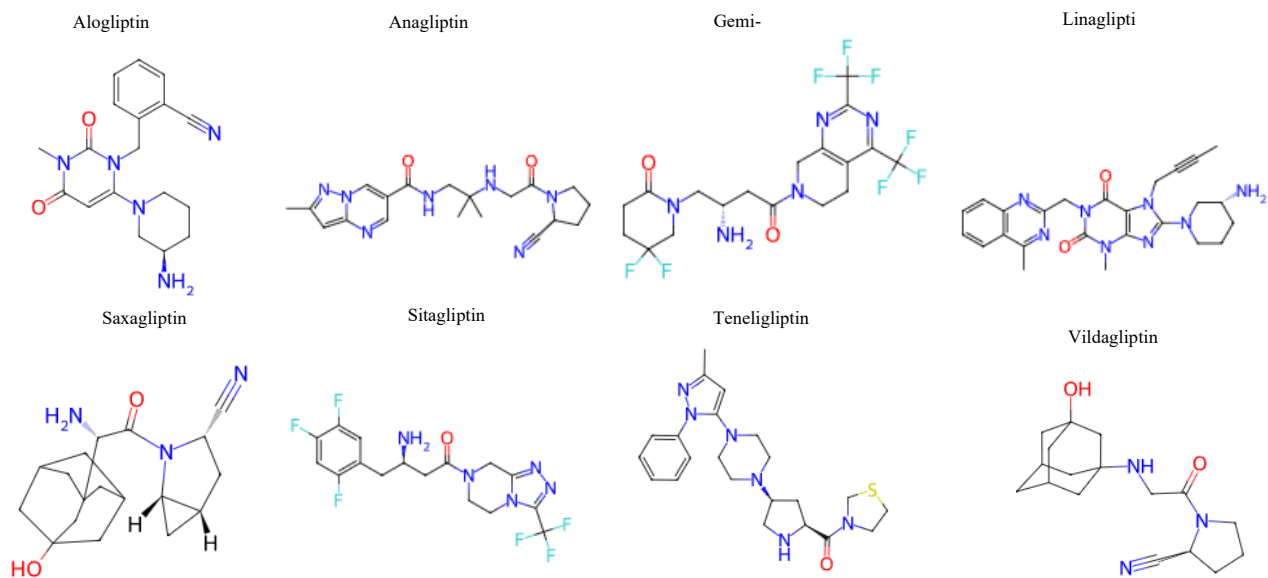


Fig. DPP-4 Inhibitors

Additional approved inhibitors can be found in ChEMBL [38], Drug Bank [46], and Wikipedia [47].

Article short title

Appendix B: LTN / Knowledge-based Setting

The construction of all the axioms components conceived from the official LTN framework [36]

Classification:

- *Domains*
 - *items*, denoting the examples from the DPP-4 dataset
 - *labels*, representing the class labels (IC50 values)

- *Define Variables*
 - $x_{inactive}, x_{active}$, for the positive examples of classes *A* and *B*
 - x for all examples
 - $D(x_A) = D(x_B) = D(x) = items$

- *Define Constants*
 - $L_{inactive}, L_{active}$ the labels of classes *A*(0) and *B*(1) Respectively.
 - $D(l_A) = D(l_B) = labels$

- *Define the P predicate.*
 - $\rho(x, l)$ Denoting the fact that item x is classified as l ;
 - $D_{in}(P) = items, labels$.

- *Connectives:*
 - *For All* \forall
 - *And* \wedge
 - *Not* \neg
 - *Or* \vee
 - *Implies* \Rightarrow

- *Axiom*
 - $\forall x_A, p(x_A, l_A)$: all the examples of class *A*(*inactive*) should have a label l_A
 - $\forall x_B, p(x_B, l_B)$: all the examples of class *B* (*active*) should have a label l_B

- *Grounding:*
 - $\mathcal{G}(items) = \mathbb{R}^N$, items are described by N features:
 - *Example*
 - DPP-4(AtomPairs2DCount descriptors): $\mathcal{G}(items) = \mathbb{R}^{780}$
 - $\mathcal{G}(labels) = \mathbb{N}^2$, We use an encoding to represent classes.
 - $\mathcal{G}(x_{inactive}) \in \mathbb{R}^{m_1 \times N}$, that is, $\mathcal{G}(x_{inactive})$ is a sequence of m_1 examples of class *inactive*;
 - $\mathcal{G}(x_{active}) \in \mathbb{R}^{m_2 \times N}$, that is, $\mathcal{G}(x_{active})$ is a sequence of m_2 examples of class *active*;
 - $\mathcal{G}(x) \in \mathbb{R}^{(m_1+m_2) \times N}$, that is, $\mathcal{G}(x)$ It is a sequence of all the examples.
 - $\mathcal{G}(l_A) = 0, \mathcal{G}(l_B) = 1$;
 - $\mathcal{G}(P | \theta): x, l \mapsto l^T \cdot softmax(MLP_\theta(x))$, where MLP has two output neurons corresponding to as many classes, notably in our cases, two classes as we explained earlier, and \cdot denotes the dot product as a way of selecting an output for $\mathcal{G}(P | \theta)$. Multiplying the MLP output by the probability. l^T Gives the probability corresponding to the class denoted by l .

Appendix C: LTN Model Architecture for multiclass classification.

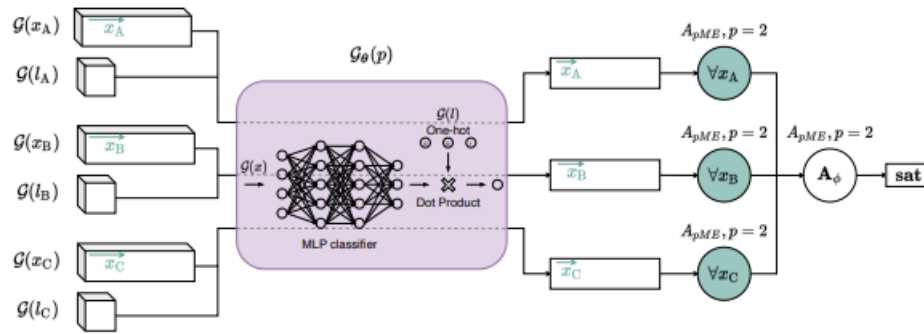


Fig. 7: LTN Classification Architecture [36]