frontiers

NeSyDPP4-QSAR: A Neuro-Symbolic AI Approach for Potent DPP-4-Inhibitor Discovery in Diabetes Treatment

Delower Hossain[1], Ehsan Saghapour[2], Jake Y. Chen[1,2*]

[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

*** Correspondence:**
Corresponding Author
jakechen@uab.edu

Keywords: DPP-4, Neuro-symbolic AI, Deep Learning

Abstract

Diabetes Mellitus (DM) is a global epidemic and among the top ten leading causes of mortality (WHO, 2019), projected to rank seventh by 2030. The US National Diabetes Statistics Report (2021) states that 38.4 million Americans have diabetes. Dipeptidyl Peptidase-4 (DPP-4) is an FDA-approved target for type 2 diabetes mellitus (T2DM) treatment. However, current DPP-4 inhibitors are associated with adverse effects, including gastrointestinal issues, severe joint pain (FDA safety warning), nasopharyngitis, hypersensitivity, and nausea. Identifying novel inhibitors is crucial. Direct in vivo DPP-4 inhibition assessment is costly and impractical, making in silico IC50 prediction a viable alternative. Quantitative Structure-Activity Relationship (QSAR) modeling is a widely used computational approach for chemical substance assessment.

We employ LTN, a neuro-symbolic approach, alongside DNN and transformers as baselines. DPP-4-related data is sourced from PubChem, ChEMBL, BindingDB, and GTP, comprising 6,563 bioactivity records (SMILES-based compounds with IC50 values) after deduplication and thresholding. A diverse set of features including descriptors (CDK Extended-PaDEL), fingerprints (Morgan), chemical language model embeddings (ChemBERTa2), LLaMa 3.2, and physicochemical properties is used to train the NeSyDPP4-QSAR model.

The NeSyDPP4-QSAR model yielded the highest accuracy, incorporating CDKextended and Morgan fingerprints, with an accuracy of 0.9725, an F1-score of 0.9723, an ROC AUC of 0.9719, and an MCC of 0.9446. The performance was benchmarked against two standard baseline models: a deep neural network and a transformer. To ensure fair comparisons, DNN models used the equivalent attributes with the same dimension and network configuration as NeSyDPP4-QSAR. Our findings showed that integrating the Neuro-symbolic strategy (neural network-based learning and symbolic reasoning) holds immense potential for discovering drugs that can inhibit diabetes mellitus and classifying biological activities that inhibit it.

1 Introduction

Diabetes Mellitus (DM) is a chronic metabolic disorder characterized by elevated blood glucose levels, posing a significant global health burden. According to the World Health Organization (WHO) 2019 report, diabetes ranks among the top ten leading causes of mortality, with an estimated 1.6 million deaths worldwide [1-2]. In the United States, diabetes is a major public health challenge, affecting approximately 38 million people (11.3% of the population) and leading to $327 billion in medical expenses and lost wages annually [3]. Beyond economic costs, diabetes is associated with severe complications, including blindness, kidney failure, stroke, heart disease, and neuropathy.

DM is broadly classified into Type 1 Diabetes Mellitus (T1DM) and Type 2 Diabetes Mellitus (T2DM), with T2DM accounting for over 90% of all cases. One crucial therapeutic target for T2DM management is the Dipeptidyl Peptidase-4 (DPP-4) enzyme, which regulates glucose metabolism. DPP-4 inhibitors, a class of FDA-approved medications, help control blood sugar levels by inhibiting this enzyme. However, current DPP-4 inhibitors have been linked to adverse effects such as gastrointestinal issues, severe joint pain, nasopharyngitis, hypersensitivity, and nausea [4]. As a result, discovering safer and more effective DPP-4 inhibitors remains a critical research challenge.

Artificial Intelligence (AI) has revolutionized diabetes management and drug discovery over the past two decades. Early AI models focused on glucose level prediction, insulin dosage recommendations, and patient monitoring. In recent years, AI has expanded into de novo drug design, utilizing vast molecular datasets to identify new inhibitors and analyze complex relationships between genes, proteins, and disease mechanisms. In the field of DPP-4 inhibitor prediction, Quantitative Structure-Activity Relationship (QSAR) models have been widely employed using machine learning techniques such as Random Forest, Support Vector Machines, XGBoost, Gradient Boosting Machines, and Deep Neural Networks [5-10]. While these models have demonstrated high predictive performance, they suffer from limitations, including poor interpretability, data inefficiency, and a lack of reasoning capabilities. The black-box nature of deep learning models further complicates their use in critical healthcare applications, where transparency and explainability are essential.

To address these challenges, Neuro-Symbolic AI (NeSy) has emerged as a promising paradigm that combines neural networks with symbolic reasoning for more interpretable and data-efficient learning. Unlike traditional AI approaches, NeSy AI enables models to integrate domain knowledge and perform logical reasoning, making them particularly suited for bioactivity prediction in drug discovery. Several NeSy models have already demonstrated success in biomedical applications [11-13], such as: Protein Function Prediction (MultipredGO [14]), Gene Sequence Analysis (KBANN [15]), Diabetic Retinopathy Diagnosis (ExplainDR [16]), (Gene Sequence) KBANN [17], hERG-LTN [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 [33], NS-CL [34], LTN [35],. This study investigates the role of a hybrid Neuro-symbolic model integrating Logic Tensor Networks (LTN) for DPP-4 bioactivity prediction. Our objective is to identify potential DPP-4 inhibitors for T2DM treatment while improving prediction accuracy.

Key Contributions to This Study

The significant key contribution of this study is: 1) we built a scalable, robust AI predictive model with immense accuracy improvement for T2DM inhibitors potency prediction. 2) A novel representation integrating data and rules (Knowledge) for DPP-4 inhibitor bio-activity classification 3) Acquired and utilized more diverse compound datasets with chemical embedding, descriptor, fingerprints, physiochemical properties that previous studies have not utilized. The proposed NeSyDPP4 can be used to discover novel DPP-4 active drugs by scanning large molecular datasets like ZINC, and identification of novel candidate compounds, accelerates de novo drug design. Additionally, it facilitates QSAR model downstream applications such as virtual screening, contraindications, bioactivity indications, and other key elements of DPP-4 inhibitors therapy in the clinical setting including docking, affinity prediction, ADMET analysis, and molecular dynamics (MD) studies for DPP-4 clinical settings.

The remainder of this paper is structured as follows: Section II describes the methodology, Section III presents the experimental results, and Section IV Discussion, and finally concludes with key findings and future research directions. `

1.1 Data acquisition

The study constructed a new DPP-IV cohorts utilized four publicly available chemical compound databases: ChEMBL [36] & BindingDB [37], PubChem, and GTP, The ChEMBL Database contains more than 2 million compounds. We retrieved canonical SMILES related to DPP-4 inhibitor with the target organism Homo Sapiens using ID: CHEMBL284 and standard type IC50. The data was extracted using the ChEMBL Python API (chembl_webresource_client). The BindingDB manually uses DPP-4 string keywords (dipeptidyl peptidase-4) from their official site. In addition, PubChem in CSV format with following(link), and GTP via the corresponding (link).

1.2 Data preparation and feature extraction

The initial bioactivity data remains various irrelevant attributes. We collected subsets focused on the IC50 biological activity standard value, ChEMBL inhibitors ID, and Canonical SMILES. However, numerical IC50 measurements in nM were given in ChEMBL, BindingDB, and GTP, but those in μM were given in PubChem, and were harmonized all units into nanomolar (nM). Subsequently, we calculated pIC50 values from the IC50 values, applying a normalization step through log10 conversion (equ. 1). Active and inactive label determined based on pIC50 by following previous DPP-iV chemical research article [38]. Afterwards, a diverse array of features was extracted from SMILES representations, encompassing Morgan fingerprints (512, 1024, and 2048 bits), CDKextended descriptors utilizing PaDELPy [39], chemical embeddings generated via ChemBERTa2 and LLaMA3.2, as well as a comprehensive set of physicochemical properties using RDkit [40].

Finally, ML trainable data comprised a total of 6563 upon dropping duplicates and NaN values.

$$\mathrm{pIC}_{50} = -\log_{10}(\mathrm{IC}_{50} \times 10^{-9}) \qquad (1)$$

1.3 LTN classification model

LTNs [35] 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 A. The logical mechanism contains a set of axioms or rules (explained in detail in the Knowledge-based setting). It's important to note that LTN logical reasoning reveals through rules/axioms. In our context Table 1 represents the axioms and relevant knowledge base component. However, other network configuration parameters can be found in Table 1.

Table 1: LTN Knowledge-based Setting for DPP-IV Classification

Contents	**Classification**
Define Axioms	• $\forall \mathrm{x}_{\mathrm{A}}, p(\mathrm{x}_{\mathrm{A}}, \mathrm{l}_{A})$: all the examples of class $A(Active = 0)$should have a label l_A • $\forall x_B, p(x_B, l_B)$: all the examples of class B (Inactive = 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	$\text{SatAgg}_{\phi\in\mathcal{K}}\mathcal{G}_{\boldsymbol{\theta},x\leftarrow\boldsymbol{D}}(\phi)$
Learning & Loss	$\boldsymbol{L} = \left(1 - \underset{\phi\in\mathcal{K}}{\text{SatAgg}}\mathcal{G}_{\boldsymbol{\theta},x\leftarrow\boldsymbol{B}}(\phi)\right)$
Note: This table was developed inspired by the official LTN	

Here,

The pMeanError aggregator

$$pME(u_1, \dots, u_n) = 1 - (\frac{1}{n}\sum_{i=1}^{n}(1-u_i)^p)^{\frac{1}{p}}p \geq 1 \quad (2)$$

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

In addition to experimenting with LTN, we conducted the simulation with DNN, and transformer with keras integrated for the fair comparison of with LTN performance. Table 2 depicts the network configuration parameters.

Table 2: LTN, DNN, Transformer Models Parameters Summary Classification

Parameters	LTN	DNN
Activation	ReLU	ReLU
Units	(768,384,192,2)	(768,384,192,2)
No of Dense layers	4	4
Seed	42	42
Batch Size	128	128
Training Epochs	100	100
Learning Rate	0.00001	0.00001
Loss Function	LTN pMeanError	Sparse Categorical Crossentropy
Optimizer	Adam	Adam

Note: The input depends on selected features (e.g., fingerprint, embedding, etc).
Transformer configuration parameter can be found on this project Github.

1.4 Model Training and Validation Phase

LTN, DNN, and transformer models were trained and tested using TensorFlow 2.15.1 Python 3.10.16 on UAB server, NVIDIA A100 80GB PCIe, other dependency packages can be found on this project GitHub environment.yml. In the training phase, we did partition the data as 80:10:20 ratios over 100 epochs during. while following metrics, such as Accuracy, F-score (F), ROC AUC Score, and Mathew Correlation Coefficient (MCC), were used to assess the trained model's performance, and the misclassified classes can appear in the Fig. 2.

Equation 1 Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Equation 2 F1 Score

$$\mathrm{F}_1 = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \tag{4}$$

Equation 3 ROC AUC Score

$$\text{ROC AUC} = \int_0^1 TPR\ d(FPR) \tag{5}$$

Equation 4 MCC

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{6}$$

2 Result

Here, we describe the performance of the developed NeSyDPP4 Model, for revealing DPP4 potential inhibitors leveraging LTN architecture (rules Integration into the neural network). DNN, and transformer since raw data is string format. We computed diverse features with the respective smiles/drugs such as morgan fingerprint, CDKExtended descriptor, Chemical foundation language model embedding using ChemBERTa2, LLaMA3.2 embedding, Physiochemical properties using RDkit. There are three tables in this section. Table 3 shows all the features separated and combination input results, Table 4 exposes the fair comparison with baseline DNN, and transformer architecture performance.

In Table 3 depicts the different input performance of LTN. The best-performing feature set is combining CDKExtended + ECFP, which yielded the highest Accuracy (97.25%), F1-score (97.23%), AUC-ROC (97.19%), and MCC (94.46, while physicochemical features alone yield the lowest performance. ChemBERTa2 and Llama3.2 performed comparably but were lower than fingerprint-based methods of Accuracy (73.49%), F1-score (73.16%), AUC-ROC (73.09%), and MCC (46.38%), Overall, suggesting that physicochemical properties alone are insufficient for effective bioactivity classification.

Furthermore, Table 4 demonstrates the comparison performance and efficiency of three models, LTN and DNN, and transformer for predicting the properties of molecules associated with T2DM DPP-IV inhibitors.

Upon obtaining the accurate features of chemical compounds from Table 1 experiment, we proceeded experiment with DNN utilizing same architecture and similar input as shows Table. The LTN model with CDKExtended + ECFP features outperforms the others, achieving 97.25% accuracy and 94.46% MCC, demonstrating the effectiveness of neuro-symbolic reasoning. The DNN model, using the same features, performs slightly lower (96.95% accuracy, 93.85% MCC), which indicates that LTN's logical constraints enhance predictions. In contrast, the Transformer model with SMILES embeddings shows the lowest performance (78.21% accuracy, 56.41% MCC), suggesting that fingerprint-based features are more effective than SMILES-based embeddings for bioactivity classification. However, the Fig illustrated the highest misclassification occurred by transformers since performance is lowest compared to three model simulation.

Table 4: LTN DPP4 Bio-activity Classification Result Summary

Model	**Features**	**Input**	**Acc**	**F1**	**AUC ROC**	**MCC**
LTN	CDKExetended + ECFP	1024+(512+1024+2048)	**0.9725**	**0.9723**	**0.9719**	**0.9446**
	ECFP	1024	0.9687	0.9684	0.9680	0.9370
	ECFP	2048	0.9657	0.9654	0.9650	0.9308
	ECFP	512	0.9649	0.9646	0.9643	0.9293
	Combined All	7430	0.9634	0.9631	0.9632	0.9262
	CDKExetended	1024	0.9504	0.9499	0.9492	0.9001
	ChemBERTa2	768	0.8956	0.8944	0.8935	0.7892
	Llama3.2	2048	0.8933	0.8926	0.8933	0.7854
	Physiochemical	6	0.7349	0.7316	0.7309	0.4638

Table 4: LTN DPP4 Bio-activity Classification Result Summary

Model	**Features**	**Input Dimension**	**Acc**	**F1**	**AUC ROC**	**MCC**
LTN	CDKExetended + ECFP	1024+(512+1024+2048)	**0.9725**	**0.9723**	**0.9719**	**0.9446**
DNN	CDKExetended + ECFP	1024+(512+1024+2048)	0.9695	0.9692	0.9691	0.9385
Transformer	SMILES/Emb	212	0.7821	0.7306	0.8549	0.5641

Model	Author	Metrics	result	Ref
NeSyDPP4	Hossain *et al*	Accuracy	0.9725	
Random Forest	Oky Hermansyah *et al*	Accuracy	0.9221	42
DNN	Haris Hamzah *et al*	Accuracy	0.9060	43
QSAR-DNN	Alhadi Bustamam *et al*	Accuracy	0.9040	9
NB	Jie Cai *et al*	Accuracy	0.8720	44
Conv1D-LSTM	Adawiyah Ulfa *et al*	Accuracy	0.8618	38
XGBoost	Oky Hermansyah *et al*	Accuracy	0.8164	45

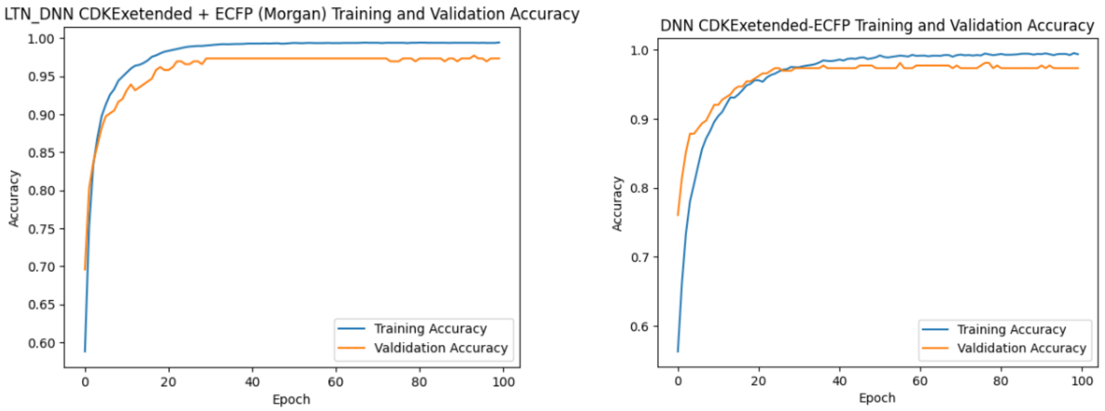

Fig.1, Epoch and Accuracy curve during the training and validation phase of LTN and DNN model

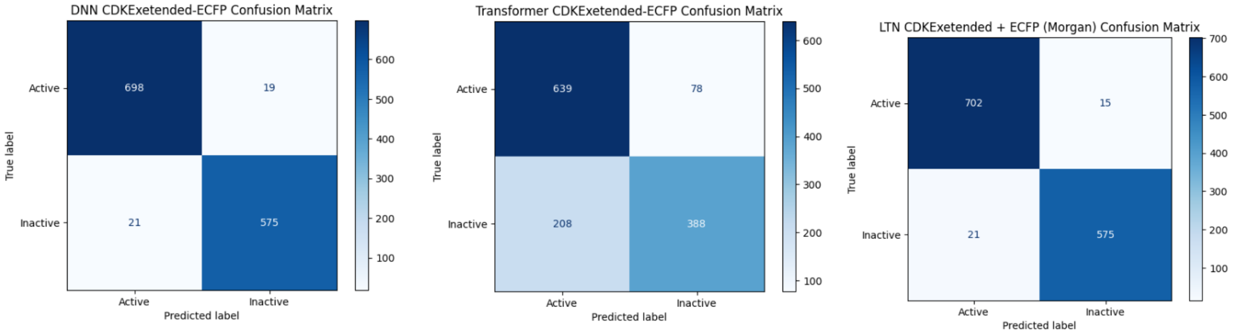

Fig.2, Classification matrix of DNN, Transformer, and LTN using CDKextended+Morgan all bit's features, it depicts that Transformer misclassified highest number of the samples.

3 Discussion

This article aimed to employ neuro-symbolic modeling (LTN), an integration of data and a logic-driven approach, for predicting diabetes mellitus DPP-4 inhibition. The study's findings provide valuable insights into the applicability and robustness of the LTN model in predicting inhibitor bioactivity behavior. As an illustration, the utilization of this advanced machine learning technique (LTN) surpassed the state-of-the-art performance compared to other models with classification tasks, the LTN model demonstrates superior accuracy of 0.9725 and an MCC score of 0.9446 for the DPP-4 inhibitors, while other studies shows the QSAR-DNN model by Bustamam et al. [1] achieved an accuracy of 0.9040, Ulfa et al. [2] reported an accuracy of 0.8618 using Conv1D-LSTM. Random Forest by Hermansyah et al. [3] yielded an accuracy of 0.9221. DNN by hamzah et al. [4] obtained an accuracy of 0.9060. NB by Cai et al. [5] gained an accuracy of 0.8720. XGBoost by Hermansyah et al. [6] achieved an accuracy of 0.8164.

The implications drawn from this research are profound. The utilization of neuro-symbolic modeling (LTN), blending data-driven and knowledge-driven methodologies has shown remarkable potential in predicting diabetes mellitus through DPP-4 inhibitors activity classification. Thus, this research tiles the way for advanced machine learning applications in diabetes prediction and marks a significant step forward in understanding inhibitor behavior and its implications for DM. These findings advocate for the transformative potential of LTN in diabetes prediction and emphasize the value of further exploration and implementation of neuro-symbolic strategies in healthcare research and applications.

Limitation

Acknowledging the limitations of our study, we state that while LTN has demonstrated significant promise, it may be uncapable to incorporate external biological additional knowledge with neural networks.

4 Conclusion

Diabetes Mellitus is a vital global health concern, and discovering effective chemical substances is crucial to tackling this epidemic. This study intend to develop QSAR system for 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 DNN, an NLP Transformer model, whereas LTN also attained prominent Accuracy.

In conclusion, the findings of this study prove 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.

5 Conflict of Interest

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

6 Author Contributions

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

7 Funding

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

8 Acknowledgments

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

9 Data Availability Statement

The dataset that utilized in this study can be found here link
And experimented code repo can be found here

10 References

1. World Health Organization: WHO. (2024, August 7). *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%20worldwide%20in%202016%2C,of%20death%20globally%20in%20the%20last%2015%20years.
2. World Health Organization. (2020). *World health statistics 2020: monitoring health for the SDGs, sustainable development goals*. https://iris.who.int/bitstream/handle/10665/332070/9789240005105-eng.pdf
3. *Methods for the National Diabetes Statistics Report*. (2024, May 15). Diabetes. https://www.cdc.gov/diabetes/php/data-research/methods.html?CDC_AAref_Val=https://www.cdc.gov/diabetes/data/statistics-report/index.html
4. Huang, J., Jia, Y., Sun, S., & Meng, L. (2020). Adverse event profiles of dipeptidyl peptidase-4 inhibitors: data mining of the public version of the FDA adverse event reporting system. *BMC Pharmacology and Toxicology*, *21*(1). https://doi.org/10.1186/s40360-020-00447-w
5. Yanuar, A., Hermansyah, O., & Bustamam, A. (2019). QSAR Modeling for Prediction of pIC50 DPP-4 Inhibitors with Machine Learning Method. *Conference*. https://conference.ui.ac.id/afps/AFPS-ICAPPS2019/paper/view/25310
6. Hermansyah, O., Bustamam, A., & Yanuar, A. (2021). Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure–activity relationship-based artificial intelligence and molecular docking of hit compounds. *Computational Biology and Chemistry*, *95*, 107597. https://doi.org/10.1016/j.compbiolchem.2021.107597
7. Ojo, O. A., Ojo, A. B., Okolie, C., Abdurrahman, J., Barnabas, M., Evbuomwan, I. O., Atunwa, O. P., Atunwa, B., Iyobhebhe, M., Elebiyo, T. C., Nwonuma, C. O., Adegboyega, A. E., Qusti, S., Alshammari, E. M., Hetta, H. F., & Batiha, G. E. S. (2021). Elucidating the interactions of compounds identified from Aframomum melegueta seeds as promising candidates for the management of diabetes mellitus: A computational approach. *Informatics in Medicine Unlocked*, *26*, 100720. https://doi.org/10.1016/j.imu.2021.100720

8. Septiawan, N. R. R., Prakoso, N. B. H., & Kurniawan, N. I. (2022). DPP IV Inhibitors Activities Prediction as An Anti-Diabetic Agent using Particle Swarm Optimization-Support Vector Machine Method. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(6), 974–980. https://doi.org/10.29207/resti.v6i6.4470

9. Bustamam, A., Hamzah, H., Husna, N. A., Syarofina, S., Dwimantara, N., Yanuar, A., & Sarwinda, D. (2021). Artificial intelligence paradigm for ligand-based virtual screening on the drug discovery of type 2 diabetes mellitus. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00465-3

10. Ajiboye, B. O., Iwaloye, O., Owolabi, O. V., Ejeje, J. N., Okerewa, A., Johnson, O. O., Udebor, A. E., & Oyinloye, B. E. (2021). Screening of potential antidiabetic phytochemicals from Gongronema latifolium leaf against therapeutic targets of type 2 diabetes mellitus: multi-targets drug design. *SN Applied Sciences*, *4*(1). https://doi.org/10.1007/s42452-021-04880-2

11. Hossain D, Chen JY. A Study on Neuro-Symbolic Artificial Intelligence: Healthcare Perspectives. arXiv preprint arXiv:2503.18213. 2025 Mar 23.

12. Yu, D., Yang, B., Liu, D., Wang, H., & Pan, S. (2023). A survey on neural-symbolic learning systems. *Neural Networks*, *166*, 105–126. https://doi.org/10.1016/j.neunet.2023.06.028

13. Wang, W., Yang, Y., & Wu, F. (2024). Towards Data-and Knowledge-Driven AI: A survey on Neuro-Symbolic Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–22. https://doi.org/10.1109/tpami.2024.3483273

14. Giri, S. J., Dutta, P., Halani, P., & Saha, S. (2020). MulTiPREDGO: Deep Multi-Modal Protein function prediction by amalgamating protein structure, sequence, and interaction information. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1832–1838. https://doi.org/10.1109/jbhi.2020.3022806

15. Towell, G. G., & Shavlik, J. W. (1994). Knowledge-based artificial neural networks. *Artificial Intelligence*, *70*(1–2), 119–165. https://doi.org/10.1016/0004-3702(94)90105-8

16. Jang, S.-I., Girard, M. J. A., & Thiery, A. H. (2021). Explainable Diabetic Retinopathy Classification Based on Neural-Symbolic Learning. *CEUR-WS*, 104–114. http://ceur-ws.org/Vol-2986/paper8.pdf

17. Maclin, R., & Shavlik, J. W. (1994). Refining algorithms with knowledge-based neural networks: improving the Chou-Fasman algorithm for protein folding. *Conference on Learning Theory*, 249–286. http://dl.acm.org/citation.cfm?id=188535

18. Hossain D, Chen JY, Abir FA. hERG-LTN: A New Paradigm in hERG Cardiotoxicity Assessment Using Neuro-Symbolic and Generative AI Embedding (MegaMolBART, Llama3. 2, Gemini, DeepSeek) Approach. bioRxiv. 2025:2025-02.

19. Yang, F., Yang, Z., & Cohen, W. W. (2017). Differentiable Learning of Logical Rules for Knowledge Base Reasoning. *NeurIPS*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/0e55666a4ad822e0e34299df3591d979-Abstract.html

20. Hohenecker, P., & Lukasiewicz, T. (2020). Ontology Reasoning with Deep Neural Networks. *Journal of Artificial Intelligence Research*, *68*. https://doi.org/10.1613/jair.1.11661

21. Yang, Z., Ishay, A., & Lee, J. (2020). NeurASP: Embracing Neural Networks into Answer Set Programming. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 1755. https://www.ijcai.org/proceedings/2020/0243.pdf

22. Kora, P., Meenakshi, K., Swaraja, K., Rajani, A., & Islam, M. K. (2019). Detection of Cardiac arrhythmia using fuzzy logic. *Informatics in Medicine Unlocked*, *17*, 100257. https://doi.org/10.1016/j.imu.2019.100257

23. Wang, J., Zhang, J., Cai, Y., & Deng, L. (2019). DEEPMIR2GO: Inferring functions of human MicroRNAs using a deep Multi-Label Classification model. *International Journal of Molecular Sciences*, *20*(23), 6046. https://doi.org/10.3390/ijms20236046

24. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018). Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. *arXiv (Cornell University)*, *31*, 1031–1042. http://arxiv.org/pdf/1810.02338.pdf

25. Amizadeh, S., Palangi, H., Polozov, O., Huang, Y., & Koishida, K. (2020). Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning." *International Conference on Machine Learning*, *1*, 279–290. http://proceedings.mlr.press/v119/amizadeh20a.html

26. Riegel, R., Gray, A. G., Luus, F. P. S., Khan, N., Makondo, N., Akhalwaya, I. Y., Qian, H., Fagin, R., Barahona, F., Sharma, U., Ikbal, S., Karanam, H., Neelam, S., Likhyani, A., & Srivastava, S. K. (2020). Logical neural networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2006.13155

27. Towell, G., & Shavlik, J. W. (1991). Interpretation of Artificial Neural Networks: Mapping Knowledge-Based Neural Networks into Rules. *Neural Information Processing Systems*, *4*, 977–984. http://papers.nips.cc/paper/546-interpretation-of-artificial-neural-networks-mapping-knowledge-based-neural-networks-into-rules.pdf

28. Lavin, A. (2022). Neuro-Symbolic neurodegenerative disease modeling as probabilistic programmed deep kernels. In *Studies in computational intelligence* (pp. 49–64). https://doi.org/10.1007/978-3-030-93080-6_5

29. Dobosz, K., & Duch, W. (2008). Fuzzy symbolic dynamics for neurodynamical systems. In *Lecture notes in computer science* (pp. 471–478). https://doi.org/10.1007/978-3-540-87559-8_49

30. Arabshahi, F., Lee, J., Gawarecki, M., Mazaitis, K., Azaria, A., & Mitchell, T. (2021). Conversational Neuro-Symbolic commonsense reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(6), 4902–4911. https://doi.org/10.1609/aaai.v35i6.16623

31. Shi, J., Zhang, H., & Li, J. (2019). Explainable and explicit visual reasoning over scene graphs. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8368–8376. https://doi.org/10.1109/cvpr.2019.00857

32. Teru, K., Denis, E., & Hamilton, W. (2020). Inductive relation prediction by subgraph reasoning. *International Conference on Machine Learning*, *1*, 9448–9457. http://proceedings.mlr.press/v119/teru20a/teru20a.pdf

33. Xu, J., Zhang, Z., Friedman, T., Liang, Y., & Broeck, G. (2018, July 3). *A Semantic Loss Function for Deep Learning with Symbolic Knowledge*. PMLR. https://proceedings.mlr.press/v80/xu18h.html

34. Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The Neuro-Symbolic Concept Learner: Interpreting scenes, words, and sentences from natural supervision. *International Conference on Learning Representations*. https://openreview.net/pdf?id=rJgMlhRctm

35. Badreddine, S., Garcez, A. D., Serafini, L., & Spranger, M. (2021). Logic Tensor networks. *Artificial Intelligence*, *303*, 103649. https://doi.org/10.1016/j.artint.2021.103649

36. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777

37. Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2015). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, *44*(D1), D1045–D1053. https://doi.org/10.1093/nar/gkv1072

38. Ulfa, A., Bustamam, A., Yanuar, A., Amalia, R., & Anki, P. (2021). Model QSAR Classification Using Conv1D-LSTM of Dipeptidyl Peptidase-4 Inhibitors. *IEEExplore*, *pp.160-163*, 1–6. https://doi.org/10.1109/aims52415.2021.9466083

39. *PaDEL*. (n.d.). GitHub. https://github.com/dataprofessor/padel.git

Ecrl. (n.d.). *GitHub - ecrl/padelpy: A Python wrapper for PaDEL-Descriptor software*. GitHub. https://github.com/ecrl/padelpy/tree/master

40. *Installation — The RDKit 2024.09.6 documentation*. (n.d.). https://www.rdkit.org/docs/Install.html

41. Hermansyah, O., Bustamam, A., & Yanuar, A. (2021). Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure–activity relationship-based artificial intelligence and molecular docking of hit compounds. *Computational Biology and Chemistry*, *95*, 107597. https://doi.org/10.1016/j.compbiolchem.2021.107597

42. Hermansyah, O., Bustamam, A., & Yanuar, A. (2020). Virtual screening of DPP-4 inhibitors using QSAR-Based artificial intelligence and molecular docking of HIT compounds to DPP-8 and DPP-9 enzymes. *Research Square (Research Square)*. https://doi.org/10.21203/rs.2.22282/v2

43. Hamzah, H., Bustamam, A., Yanuar, A., & Sarwinda, D. (2020). Predicting The Molecular Structure Relationship and The Biological Activity of DPP-4 Inhibitor Using Deep Neural Network with CatBoost Method as Feature Selection. *IEEEXplore*, 101–108. https://doi.org/10.1109/icacsis51025.2020.9263204

44. Cai, J., Li, C., Liu, Z., Du, J., Ye, J., Gu, Q., & Xu, J. (2017). Predicting DPP-IV inhibitors with machine learning approaches. *Journal of Computer-Aided Molecular Design*, *31*(4), 393–402. https://doi.org/10.1007/s10822-017-0009-6

45. Identification of DPP-4 inhibitor active compounds using machine learning classification. (2023). *ResearchGate*.

46. FDA approved Dipeptidyl Peptidase IV (DPP IV) Inhibitors https://www.ncbi.nlm.nih.gov/books/NBK542331/#:~:text=DPP%2D4%20inhibitors%2C%20known%20as,saxagliptin%2C%20linagliptin%2C%20and%20alogliptin

47. Wikipedia DPP-4 Inhibitors https://en.wikipedia.org/wiki/Dipeptidyl_peptidase-4_inhibitor

Appendix A: LTN Model Architecture for multiclass classification.

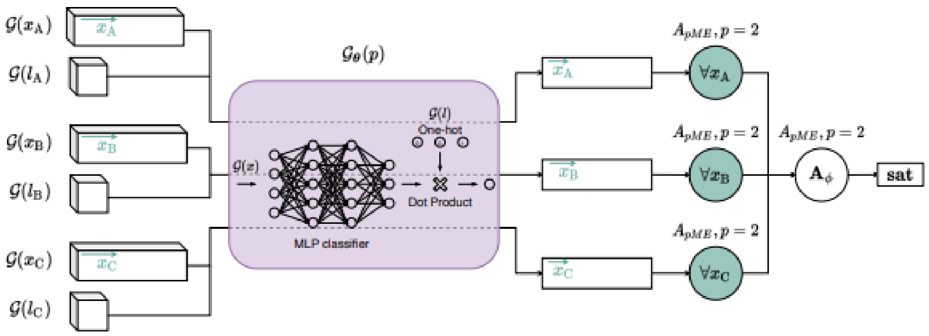

Fig. 7: LTN Classification Architecture [36]

Appendix B: A list of FDA, EU, EMA (European Medicines Agency), JAPAN, and KOREN BODY approved DPP4 inhibitor's structure and respective 3D compound structures images as below.

ChEMBL ID	Target	Approved Body	Smiles	Ref
CHEMBL376359	Alogliptin	FDA	Cn1c(=O)cc(N2CCC[C@@H](N)C2)n(Cc2ccccc2C#N)c1=O	[46]
CHEMBL1929396	Anagliptin	Japan	Cc1cc2ncc(C(=O)NCC(C)(C)NCC(=O)N3CCC[C@H]3C#N)cn2n1	[46-47]
CHEMBL3707235	Gemigliptin	Korea	N[C@@H](CC(=O)N1CCc2c(nc(C(F)(F)F)nc2C(F)(F)F)C1)CN1CC(F)(F)CCC1=O	[46-47]
CHEMBL237500	Linagliptin	FDA	CC#CCn1c(N2CCC[C@@H](N)C2)nc2c1c(=O)n(Cc1nc(C)c3ccccc3n1)c(=O)n2C	[46]
CHEMBL385517	Saxagliptin	FDA	N#C[C@@H]1C[C@@H]2C[C@@H]2N1C(=O)[C@@H](N)C12CC3CC(CC(O)(C3)C1)C2	[46]
CHEMBL1422	Sitagliptin	FDA	N[C@@H](CC(=O)N1CCn2c(nnc2C(F)(F)F)C1)Cc1cc(F)c(F)cc1F	[46]
CHEMBL2147777	Teneligliptin	Japan	Cc1cc(N2CCN([C@@H]3CN[C@H](C(=O)N4CCSC4)C3)CC2)n(-c2ccccc2)n1	**Error! Reference source not found.**[46-47]
CHEMBL142703	Vildagliptin	EMA	N#C[C@@H]1CCCN1C(=O)CNC12CC3CC(CC(O)(C3)C1)C2	[46-47]

Appendix C: LTN / Knowledge-based Setting

The construction of all the axioms components conceived from the official LTN framework [35].

Classification:

- *Domains*
  - $items$, denoting the examples from the DPP-4 dataset
  - $labels$, representing the class labels (IC50 values)
- *Define Variables*
  - $x_{active}, x_{inactive}$, , for the positive examples of classes A and B
  - x for all examples
  - $D(x_A) = D(x_B) = D(x) = items$
- *Define Constants*
  - $L_{active}, L_{inactive}$ the labels of classes $A(0)$ and $B(1)$ Respectively.
  - $D(l_A) = D(l_B) = labels$ (active inactive pic50 based)
- *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* ∀, *And* ∧, *Not* ¬, *Or* ∨, *Implies* ⟹
- *Axiom*
  - $\forall \mathrm{x_A}, p(\mathrm{x_A}, \mathrm{l}_A)$: all the examples of class $A(active)$should have a label l_A
  - $\forall x_B, p(x_B, x_B)$: all the examples of class B $(Inactive)$ should have a label l_B

Notice that rules about exclusiveness, such as $\forall\left(P(x, l_A) \Longrightarrow (\neg P(x, l_B) \wedge, \neg P(x, l_C))\right)$ They are omitted since such constraints are already imposed by the grounding of P, below, more specifically by the *softmax* function.

- Grounding:
  - $\mathcal{G}(\mathrm{items}) = \mathrm{R}^{\mathrm{N}}$, items are described by N features:
  - $\mathcal{G}(\mathrm{labels}) = \mathrm{N}^2$, We use an encoding to represent classes.
  - $\mathcal{G}(\mathrm{x_{active}}) \in \mathrm{R}^{\mathrm{m_1} \times \mathrm{N}}$, that is, $\mathcal{G}(\mathrm{x}_{active})$ is a sequence of $\mathrm{m_1}$ examples of class active;

- $\mathcal{G}(\mathrm{x}_{\mathrm{inactive}}) \in \mathrm{R}^{\mathrm{m}_2 \times \mathrm{N}}$, that is, $\mathcal{G}(\mathrm{x}_{\mathrm{inactive}})$ is a sequence of m_2 examples of class inactive;

- $\mathcal{G}(\mathrm{x}) \in \mathrm{R}^{(\mathrm{m}_1+\mathrm{m}_2)\times \mathrm{N}}$, that is, $\mathcal{G}(x)$ It is a sequence of all the examples.

- $\mathcal{G}(\mathrm{l}_\mathrm{A}) = 0, \mathcal{G}(\mathrm{l}_\mathrm{B}) = 1;$

$\mathcal{G}(\ \mathrm{P} \mid \theta\ ): x, l \mapsto \mathrm{l}^\top \cdot softmax\big(\mathrm{MLP}_\theta(\mathrm{x})\big)$, where MLP has two output neurons corresponding to as many classes, notably in our cases, two classes as we explained early, and $\cdot$ denotes the dot product as a way of selecting an output for $\mathcal{G}(\ P \mid \theta\ )$. Multiplying the MLP output by the probability. $l^\top$ Gives the probability corresponding to the class denoted by l.