WHEN SMILES HAVE LANGUAGE: DRUG CLASSIFICATION USING TEXT CLASSIFICATION METHODS ON DRUG SMILES STRINGS

Azmine Toushik Wasi[1], **Karlo Serbetar**[2], **Raima Islam**[3], **Taki Hasan Rafi**[4] **& Dong-Kyu Chae**[4*]
[1]Shahjalal University of Science and Technology, Bangladesh azmine32@student.sust.edu
[2]University of Cambridge, United Kingdom ks798@cantab.ac.uk
[3]BRAC University, Bangladesh raima.islam@g.bracu.ac.bd
[4]Hanyang University, Republic of Korea {takihr, dongkyu}@hanyang.ac.kr

ABSTRACT

Complex chemical structures, like drugs, are usually defined by SMILES strings as a sequence of molecules and bonds. These SMILES strings are used in different complex machine learning-based drug-related research and representation works. Escaping from complex representation, in this work, we pose a single question: What if we treat drug SMILES as conventional sentences and engage in text classification for drug classification? Our experiments affirm the possibility with very competitive scores. The study explores the notion of viewing each atom and bond as sentence components, employing basic NLP methods to categorize drug types, proving that complex problems can also be solved with simpler perspectives. The data and code are available here: https://github.com/azminewasi/Drug-Classification-NLP.

1 INTRODUCTION

Classifying drug types plays a pivotal role in drug discovery research, aiding in the categorization of established drugs and enhancing our understanding of the distinctive features of newly identified or synthesized drugs. It is necessary to ensure that a drug is used safely and that you get the greatest possible benefit with the lowest possible risk. Different deep generative models have demonstrated efficacy in addressing various drug discovery challenges (Pandey et al., 2022), mostly with the capabilities of utilizing complex chemical structural data.

Simplified Molecular Input Line Entry System (SMILES) is a text-based representation of a chemical molecule (Kong et al., 2022). They provide a standardized language for encoding molecular information, facilitating analysis and machine learning applications in drug-related research. One example of a drug structure and corresponding SMILES is shown in figure 2.

In this study, we explore the drug classification challenge from a simple perspective using drug SMIILES. Given that drug chemical structures are conventionally denoted through SMILES strings, an opportunity arises to avoid complex chemical representations by considering drug SMILES as simple text sentences. In this analogy, the individual atoms and bonds within the molecule serve as the constituent words, forming a coherent sentence using the sequential arrangement of SMILES, word after word. Experimental results show that applying a basic bag-of-n-grams model can achieve very competitive scores, showing proof that simple NLP approaches can be applied to complex problems too, without using any complex chemical embedding or pre-trained models.

2 METHOD

In the process, a given SMILES string undergoes encoding via the bag-of-n-grams model, where ‘n’ signifies the number of letters within each token. Multiple tokens are generated from the dataset,

[*] Corresponding author.

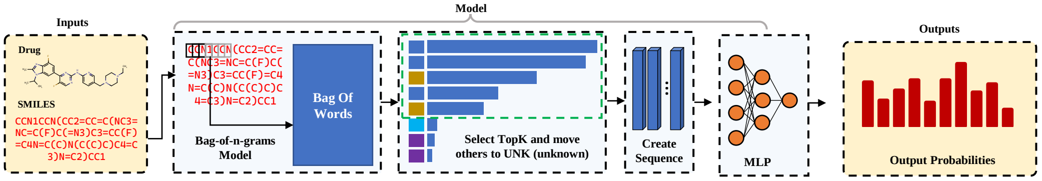
Model
Inputs
Drug
SMILES
CCN1CCN(CC2=CC=C(NC3=NC=C(F)C(=N3)C3=CC(F)=C4N=C(C)N(C(C)C)C4=C3)N=C2)CC1
Bag Of Words
Bag-of-n-grams Model
Select TopK and move others to UNK (unknown)
Create Sequence
MLP
Outputs
Output Probabilities

Figure 1: Overview of our approach for Drug Classification using Text Classification Methods on Drug SMILES Strings

with the top 'K' (the most frequently occurring) tokens selected. The remaining tokens are amalgamated into an 'unknown' token ('UNK'). Subsequently, a sequence is constructed based on the frequency of token appearances. This string is then fed into a Multilayer Perceptron (MLP) to obtain logits, from which a classification label is determined by selecting the one with the highest probability. A visual overview of the approach is presented in figure 1.

3 Experiment

In our experimental setup, we utilized a drug dataset obtained from Meyer et al. (2019), partitioned into 70% for training, 10% for development, and 20% for testing. The dataset has 12 classes; they are: dermatologic, antiinfective, antineoplastic, CNS, hematologic, lipidregulating, antiinflammatory, cardio, gastrointestinal, respiratory system, reproductive control, and urological. More dataset information is included in Appendix B.1. The modeling approach involved considering combinations of 1 to 5 grams, representing sequences of 1 to 5 letters in the tokens, while encoding each model. Each experiment is performed multiple times with different seeds, and an average is taken. Also, we have performed fine-tuning only on $TopK$, and the best value is 1250. Experimental setup details with comparison and ablation studies are presented in Appendix B. Here in table 1, we present comparative analysis between different methods For fingerprints, we use these fingerprints: Atom Pair Fingerprint, MACCS Fingerprint, Morgan Fingerprint.

Table 1: Performance Metrics for Different Configurations

Model	**Accuracy**	**Precision**	**Recall**	**F1 (Weighted)**	**F1 (Macro)**	**ROC-AUC**
1-gram+MLP	0.622	0.610	0.622	0.604	0.406	0.760
2-gram+MLP	0.669	0.700	0.669	0.672	0.445	0.810
3-gram+MLP	**0.737**	**0.764**	**0.737**	**0.744**	0.553	**0.848**
4-gram+MLP	0.726	0.758	0.726	0.731	0.524	0.841
5-gram+MLP	0.728	0.740	0.728	0.730	**0.563**	0.838
AtomPair+MLP	0.799	0.804	0.800	0.799	0.702	0.876
MACCS+MLP	0.797	0.801	0.797	0.796	0.702	0.873
Morgan+MLP	**0.800**	**0.804**	**0.800**	**0.799**	**0.703**	**0.876**

Table 1 illustrates the performance of several drug classification models. Among the ngram models, 3-gram models achieve around 73.7% accuracy and 76.4% precision in our experimental setup. Most of the ROC-AUC scores are also above 0.835, suggesting good performance. Molecular fingerprint models like AtomPair+MLP, MACCS+MLP, and Morgan+MLP exhibit improved accuracy and noteworthy precision, recall, and F1 scores. Remarkably, Morgan+MLP excels in various metrics, showcasing its effectiveness despite the advantage of molecular fingerprints. In summary, 3-gram+MLP emerges as the optimal solution among ngrams, showcasing competitive scores with fingerprint-based models. This demonstrates the feasibility of treating drug SMILES as strings and employing basic NLP methods for classification tasks. While molecular fingerprints are intended to capture particular molecular characteristics and are therefore anticipated to score higher, it is noteworthy that n-gram models, such as 3-gram+MLP, surprisingly hold their own and do reasonably well in drug classification. This observation emphasizes the versatility and competitive performance of n-gram approaches, even when compared to specialized molecular fingerprinting techniques. Additional details on our model's correlation with fingerprint-based models, our work's practical significance, and its limitations are discussed in Appendix C.

Acknowledgements

We appreciate the reviewers' insightful feedback, which helped us to improve the paper's quality. The work of Dong-Kyu Chae and Taki Hasan Rafi was supported by (1) the National Research Foundation of Korea (NRF) grant funded by the Korea government (*MSIT) (No.2018R1A5A7059549) and (2) the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01373,Artificial Intelligence Graduate School Program (Hanyang University)). *Ministry of Science and ICT

URM Statement

Authors Azmine Toushik Wasi and Raima Islam meet the URM criteria of the ICLR 2024 Tiny Papers Track.

References

Sarwan Ali, Prakash Chourasia, and Murray Patterson. When biology has chemistry: Solubility and drug subcategory prediction using SMILES strings, 2023. URL https://openreview.net/forum?id=28si4RXwDt1.

Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1), June 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00445-4. URL http://dx.doi.org/10.1186/s13321-020-00445-4.

Jianyuan Deng, Zhibo Yang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. Artificial intelligence in drug discovery: applications and techniques. *Briefings in Bioinformatics*, 23(1):bbab430, 11 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab430. URL https://doi.org/10.1093/bib/bbab430.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

Lal Khan, Ammar Amjad, Noman Ashraf, and Hsien-Tsung Chang. Multi-class sentiment analysis of urdu text using multilingual bert. *Scientific Reports*, 12(1):5436, 2022.

Weiya Kong, Yuejuan Hu, Jiao Zhang, and Qiaoyin Tan. Application of smiles-based molecular generative model in new drug design. *Frontiers in Pharmacology*, 13, 2022. ISSN 1663-9812. doi: 10.3389/fphar.2022.1046524. URL https://www.frontiersin.org/articles/10.3389/fphar.2022.1046524.

Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8464–8476. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9054-n-gram-graph-simple-unsupervised-representation-for-graphs-with-applications-pdf.

Jesse G. Meyer, Shengchao Liu, Ian J. Miller, Joshua J. Coon, and Anthony Gitter. Learning drug functions from chemical structures with convolutional neural networks and random forests. *Journal of Chemical Information and Modeling*, 59(10):4438–4449, Oct 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00236. URL https://doi.org/10.1021/acs.jcim.9b00236.

Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C Stern, and Artem Cherkasov. The transformational role of gpu computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4(3):211–221, 2022.

Sereina Riniker and Gregory A Landrum. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, 5(1), September 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-43. URL http://dx.doi.org/10.1186/1758-2946-5-43.

Benedikt Winter, Clemens Winter, Johannes Schilling, and André Bardow. A smile is all you need: predicting limiting activity coefficients from smiles with natural language processing. *Digital Discovery*, 1(6):859–869, 2022.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46(D1):D1074–D1082, January 2018.

Cheng-Kun Wu, Xiao-Chen Zhang, Zhi-Jiang Yang, Ai-Ping Lu, Ting-Jun Hou, and Dong-Sheng Cao. Learning to smiles: Ban-based strategies to improve latent representation learning from molecules. *Briefings in Bioinformatics*, 22(6):bbab327, 2021.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

A Related Works

Deep learning and NLP: Deep learning (DL) models are utilized in drug development for the following purposes: quantitative structure-activity relationship, virtual screening, and drug design (Deng et al., 2021). Therefore, in recent years, we have seen the application of various DL systems being employed for drug tasks where (Xiong et al., 2019) incorporated graph attention methods to Graph Neural Networks and constructed Attentive FP, a function capable of preserving the interactions between topologically adjacent atoms. (Gómez-Bombarelli et al., 2018) devised a Variational Autoencoder model to facilitate the automated design of molecules to transform the SMILES input strings into a representation of continuous vectors. Based on SeqGAN (Yu et al., 2017), (Guimaraes et al., 2017) constructed objective-reinforced Graph Adversarial Network (ORGAN) to produce molecules from SMILES sequences while optimizing a variety of domain-specific metrics. For text classification, n-gram modeling is one of the most basic and fundamental model. (Khan et al., 2022) have employed several n-gram features, including unigrams, bigrams, trigrams, and numerous combinations thereof to train DL classifiers and perform sentiment analysis, achieving a good score. In our work, we have also presented an example of how effective it can be in modeling very complex scenarios like drug classification.

SMILES-drug representation: While recent deep learning advances have resulted in accelerating drug discovery processes, not many of them address the generalization problem due to lack of labeled data. (Wu et al., 2021) develop a bidirectional long-short-term memory attention network (BAN) with a multi-step attention framework, extracting the important characteristics from SMILES strings and capturing latent representations of molecules. Tackling the same problem of data availability, (Winter et al., 2022) leverage an NLP mechanism called SMILES-to-properties-transformer (SPT) for predicting binary limiting activity coefficients from SMILES codes for thermodynamic property prediction. Through the integration of synthetic and experimental data, the fine-tuning process achieved computational efficiency.

Fingerprint-based representation: Ali et al. (2023) presented a detailed study of fingerprint-based feature extraction for drug subcategory classification. This work explored the prediction of drug subcategories by employing traditional molecular fingerprints and sequence-based embedding methods, specifically focusing on SMILES strings in the bioinformatics domain. The study evaluates five types of embeddings, including Morgan fingerprint, MACCS fingerprint, k-mers, and minimizer-based spectrum. Furthermore, a weighted variant of k-mers, incorporating inverse document frequency for assigning weights to individual k-mers within the spectrum, is also investigated.

Graph-based Drug representation: Graph ML has established its usefulness in biomedical applications, especially in classification tasks. Liu et al. (2019) developed N-Gram Graph, an unsupervised molecular representation. By embedding vertices in the molecule graph and constructing compact representations through short walks, this method successfully represents molecular properties; empirical experiments and theoretical analyses confirm the method's strong representation and prediction capabilities. In contrast, our model adopts a simpler perspective, leveraging drug SMILES as text sentences for drug classification directly without any complexities.

B Experiments

B.1 Dataset

In our study, we utilized a drug dataset retrieved from Meyer et al. (2019), partitioning it into training (70%), development (10%), and testing (20%) subsets for analysis. We removed all multi-label options and kept only single labels for this work to simplify the model. The dataset is available in the supplementary materials, and the distribution of labels across the 12 classes is outlined in Table 2. ”Antiinfective” drugs are most common, followed by ”antineoplastic” and ”CNS”. There is a class-imbalance between different classes. The most common type ”antiinfective” is 83 times more present than the least common type ”urological”.

Table 2: Dataset Classes

Type	Count	Type	Count
antiinfective	2412	gastrointestinal	259
antineoplastic	1175	lipidregulating	164
cns	1149	reproductivecontrol	148
cardio	797	dermatologic	115
antiinflammatory	372	respiratorysystem	100
hematologic	266	urological	29

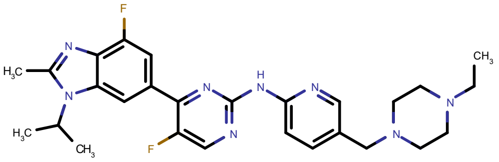

Figure 2: Molecular structure for a drug named **Abemaciclib**, with following SMILES string - CCN1CCN(CC2=CC=C(NC3=NC=C(F)C(=N3)C3=CC(F)=C4N=C(C)N(C(C)C)C4=C3)N=C2)CC1 (Wishart et al., 2018).

Here we present the n-gram modeling of a drug molecule SMILES, named **Abemaciclib**. The SMILES string is: CCN1CCN(CC2=CC=C(NC3=NC=C(F)C(=N3)C3=CC(F)=C4 N=C(C)N(C(C)C)C4=C3)N=C2)CC1. Table 3 shows the number of unique n-grams for each n, which is denoted by K for this particular SMILES.

Table 3: n and K in **Abemaciclib**

n	K	Some Examples
1	10	‘N’, ‘C’, ‘F’, ‘=’
2	31	‘NC’, ‘C2’, ‘C3’, ‘C=’, ‘CN’, ‘N=’
3	58	‘C4N=’, ‘4=C3’, ‘=C4’, ‘N1C’
4	67	‘C3=C’, ‘CCN1’, ‘C3=N’, ‘C4N=’, ‘=NC=’
5	69	‘(=N3)’, ‘CC2=C’, ‘C(F)C’

B.2 Experimental Details

Table 4 presents experimental parameters and hyperparameters. In all experiments, the parameters are kept the same. Experimental codes are included in supplementary materials.

Table 4: Parameter Configuration for the Experiment

NAME	VALUE
Number of Epochs	200
TopK	1250
Dev Set Ratio	0.1
Test Set Ratio	0.2
Number of ngrams considered (TopK+1)	1251
Batch Size	32
MLP Hidden Size	[512, 256, 128, 32]
MLP Dropout	0.1
MLP Learning Rate	3×10^{-5}

B.3 Ablation Study: TopK

In Table 5, we present an ablation study on the hyperparameter topK using 3-grams. The results indicate optimal performance at $TopK = 1250$, followed closely by $TopK = 1500$ as the second-best configuration.

Table 5: Performance at Different Values of TopK

TopK	Accuracy	Precision	Recall	F1 (W.)	F1 (Macro)	ROC-AUC	AUPRC
500	0.705	0.744	0.705	0.713	0.525	0.830	0.587
750	0.695	0.715	0.695	0.697	0.435	0.825	0.574
1000	0.711	0.740	0.711	0.717	0.477	0.834	0.593
1250	**0.737**	**0.764**	**0.737**	**0.744**	0.553	**0.848**	**0.620**
1500	0.732	0.758	0.732	0.737	**0.561**	0.846	0.612

C Discussion

C.1 Discussion on Experimental Findings

Drug classification is challenging, considering molecular structures are complex and multifaceted. The task is further complicated by the broad spectrum of compounds, minor variations in chemical composition, and the constantly changing interactions that occur within biological systems. Furthermore, the dynamic landscape of drug design and discovery demands flexible classification models that are capable of navigating novel chemical spaces, which makes the task inherently demanding.

Molecular fingerprints and n-gram modeling serve as indispensable tools in Cheminformatics and language modeling, respectively, enabling the efficient comparison and analysis of complex structures. In order to extract features, molecular fingerprints transform structural elements into bits in a bit vector or counts in a count vector, such as substructures for small molecules or atom-pairs for larger molecules Riniker & Landrum (2013). Morgan fingerprints, a type of molecular fingerprint, transform intricate molecular structures into unique bit vectors by identifying circular substructures within a specified radius around each atom Capecchi et al. (2020). On the other hand, n-gram modeling operates as a probabilistic language model, predicting the likelihood of word sequences by breaking down input into chunks (n-grams) and assessing their probabilities based on occurrence frequency. The goal of both methods is to reduce complex structures—like molecules or strings—to simpler, more similar forms. The correlation is also observed in our model's scores, as presented in Tables 1 and 5. This interesting phenomenon suggests a relationship between the chemical aspects of our model and the corresponding parameters in Morgan fingerprints. For instance, the optimal performance observed at $TopK = 1250$ and $n = 3$ mirrors the characteristics of Morgan fingerprints, where $TopK$ is associated with the standard number of bits and n with the specified radius. The close alignment of our model's optimal settings with Morgan fingerprints exemplifies how chemical intuition can fine-tune simple n-gram models in Cheminformatics intuitively. Further research could delve into these rationales and explore these intriguing relationships.

Apart from these shared characteristics, there are also some differences. Morgan fingerprints are susceptible to bit collisions when different substructures map to the same bit, which could lead to ambiguity. On the contrary, n-gram models perform very well in representing each n-gram distinctively, which neutralizes the risk of collisions. Furthermore, molecular fingerprints only account for the presence of specific substructures, potentially leaving out vital information in some cases, whereas n-gram models are context-sensitive and take letter order of SMILES into consideration, which can induce noise.

Our work also suggests that capturing some key substructures is sufficient for drug classification. Table 5 shows that the score increases until $TopK$ reaches 1250 and declines thereafter, indicating that only 1250 tokens are sufficient for accurate classification. This proves that, by focusing on essential molecular motifs, models are able to distill meaningful information, striking a balance between computational efficiency and predictive accuracy. This approach acknowledges the practical constraints of analyzing large chemical spaces while still yielding valuable insights for effective drug classification by proposing a simple baseline.

C.2 Discussion on Practical Impact and Scalability

We believe that using a basic NLP model in drug classification has significant potential impact and utility. This approach streamlines the representation of complex chemical structures into a format analogous to natural language, thereby simplifying the drug classification process. The model's success in achieving competitive scores without relying on intricate chemical embeddings or pre-trained models demonstrates its effectiveness in addressing complexity through simplicity. This technique has the potential to transform drug classification by providing a more accessible and interpretable framework, potentially enhancing collaboration between experts in diverse fields. Its simplicity not only promotes ease of implementation but also contributes to democratizing drug discovery processes, making them more approachable for researchers and practitioners without extensive expertise in chemoinformatics.

C.3 Discussion on Limitations and Future Works

In addition to drug classification, NLP-based methods can play a pivotal role in drug Quantitative Structure-Activity Relationship (QSAR) research by enabling the extraction of meaningful information from textual data, enhancing the understanding of drug properties and interactions through advanced linguistic analysis. Also, we can observe the class imbalance in the table B.1. Future works in this area may explore strategies like oversampling, undersampling, or synthetic data generation to address this issue. Additionally, leveraging advanced transfer learning models may enhance adaptability, presenting promising avenues for further investigation into robust drug classification methods using NLP techniques. The interpretability of this model can be utilized to clarify the decision-making process involved in drug classification, assisting researchers and healthcare providers in comprehending the variables affecting forecasts. The transparency of the model makes it easier to identify the critical characteristics that contribute to each drug type, offering insightful information about the classification choices made.

D Conclusion

This study showcases the application of fundamental NLP models to intricate challenges like drug classification by treating drug SMILES as strings. Our experimental findings reveal that our basic NLP model, typically defined by a bag-of-n-grams approach, attains highly competitive scores in drug classification tasks.