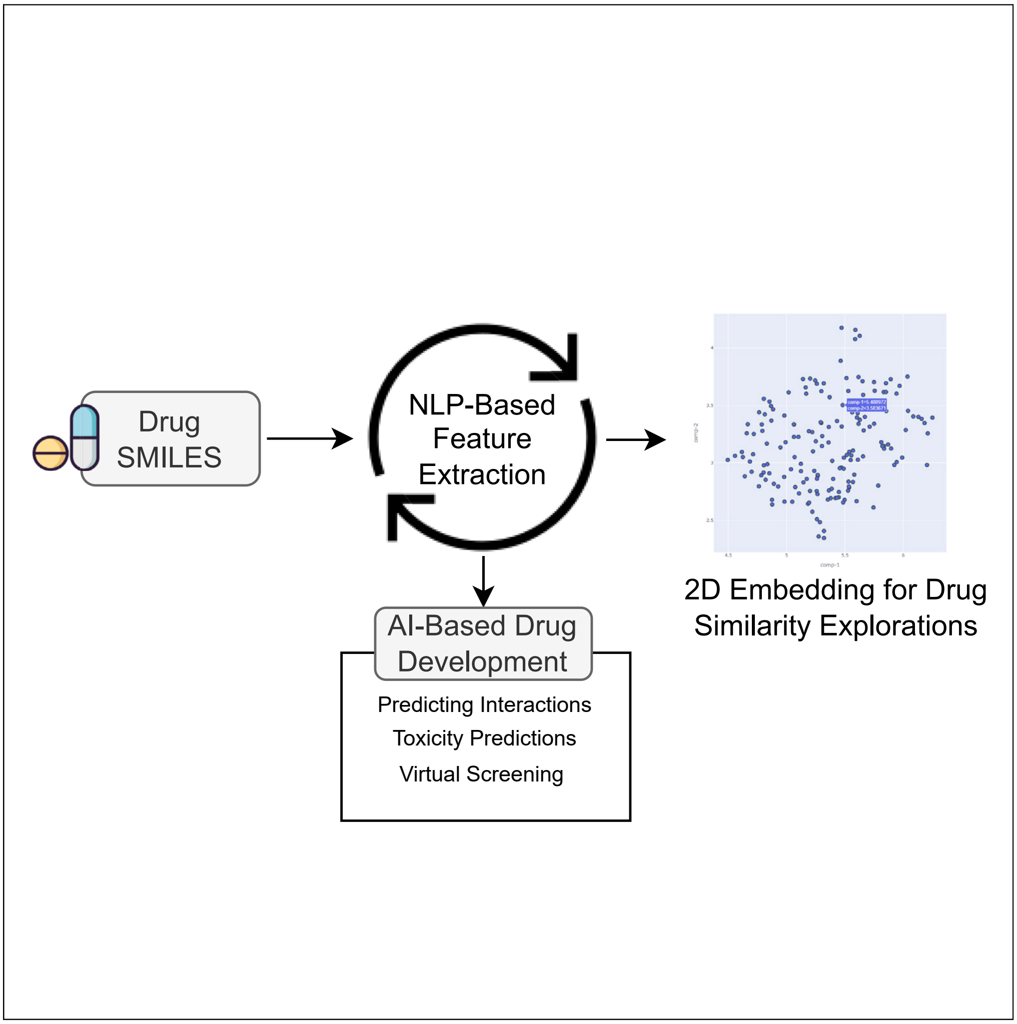
iScience

CellPress
OPEN ACCESS

Article

An NLP-based technique to extract meaningful features from drug SMILES



Rahul Sharma,
Ehsan Saghapour,
Jake Y. Chen

rsharma3@uab.edu (R.S.)
jakechen@uab.edu (J.Y.C.)

Highlights

NLP to extract meaningful features from drug SMILES

Python library for feature extraction from drug SMILES

Drug-drug similarity from NLP feature embeddings

In-silico personalized drug screening using NLP-based drug features

Sharma et al., iScience *27*, 109127
March 15, 2024 © 2024 The Author(s).
https://doi.org/10.1016/j.isci.2024.109127

iScience

CellPress
OPEN ACCESS

Article

An NLP-based technique to extract meaningful features from drug SMILES

Rahul Sharma,[1,*] Ehsan Saghapour,[1] and Jake Y. Chen[1,2,*]

SUMMARY

NLP is a well-established field in ML for developing language models that capture the sequence of words in a sentence. Similarly, drug molecule structures can also be represented as sequences using the SMILES notation. However, unlike natural language texts, special characters in drug SMILES have specific meanings and cannot be ignored. We introduce a novel NLP-based method that extracts interpretable sequences and essential features from drug SMILES notation using N-grams. Our method compares these features to Morgan fingerprint bit-vectors using UMAP-based embedding, and we validate its effectiveness through two personalized drug screening (PSD) case studies. Our NLP-based features are sparse and, when combined with gene expressions and disease phenotype features, produce better ML models for PSD. This approach provides a new way to analyze drug molecule structures represented as SMILES notation, which can help accelerate drug discovery efforts. We have also made our method accessible through a Python library.

INTRODUCTION

The representation of drug compounds through the simplified molecular-input line-entry system (SMILES) notation is a common practice in AI-based drug discovery. SMILES notations of drugs can be transformed into molecular fingerprints such as Morgan fingerprints, enabling the construction of machine learning (ML) models for virtual screening to predict a spectrum of drug properties, including toxicity,[1–3] drug-drug interaction,[4–7] and drug-target interactions.[8–11] Such molecular fingerprints have also been instrumental in precision medicine, guiding personalized drug screening (PSD) using gene expressions and multi-omics data.

Table 1 lists studies that have utilized drug SMILE molecular descriptors to build ML models for PSD. The table offers a comparative analysis of the molecular descriptor and drug SMILES-based PSD, demonstrating the superior performance of ML models built using drug SMILES features, including our proposed method.

The drug SMILES can be interpreted as natural language sequences, wherein special characters hold specific meanings that cannot be overlooked. However, unlike natural language texts, these sequences have an inherent association among atoms and should not be treated as isolated entities. Several deep learning methodologies have been developed to glean valuable features from drug SMILES strings. Despite their effectiveness in predicting various drug properties, most of these methods disregard the atomic associations and the need for interpretability. As a result, a lacuna exists for novel methods that can extract essential and explainable features from drug SMILES.[12–14]

In this realm, our paper introduces an innovative feature extraction method inspired by natural language processing (NLP) for PSD using SMILES notation. Our approach leverages N-grams, a tool derived from NLP, to isolate significant and interpretable features from drug SMILES sequences. Figure 1 shows a simplified view of our novel method for NLP-based feature extraction from Drug SMILES. Our method showcases promising results in virtual screening when contrasted with the widely utilized Morgan fingerprint bit-vector technique.

To establish our method's efficacy, we implemented it in two case studies focused on pan-cancer data and aggressive brain cancer, Glioblastoma Multiforme. The outcomes revealed that the NLP-based features significantly improve the prospects of developing more effective ML models for PSD.

The vast corpus of AI research underscores the utility of SMILES notations in feature extraction and building predictive ML models in drug discovery. For instance, Seq2seq fingerprint[15] uses an RNN[16] based approach with Long Short-Term Memory (LSTM)[17] to extract features from drug SMILES. SMILES2vec,[18] another RNN-based method, learns a vector-based representation for a Drug SMILES string.

Besides RNNs, transformers[19] (another deep-learning method for natural language modeling) is also used to extract features from drug SMILES. SMILES-transformer extracts feature from drug SMILES by using the latent vector generated at the output of the encoder as molecular fingerprints. SMILE-BERT[20] was built using a pre-trained BERT[21] on SMILE and is also a transformer used to extract features from drug SMILES. The molecule attention transformer (MAT)[22] used features extracted from both graph-based methods and features obtained

[1]Informatics Institute, School of Medicine, The University of Alabama at Birmingham, Birmingham, AL, USA
[2]Lead contact
*Correspondence: rsharma3@uab.edu (R.S.), jakechen@uab.edu (J.Y.C.)

https://doi.org/10.1016/j.isci.2024.109127

Table 1. List of some articles that used Molecular Descriptors to Build ML models for Personalized and Comparative Study

Article	Personalized Drug Screening Data Specs	R2
Menden et al.[28]	Omics (GE) + Molecule Descriptor	0.72
Ammad-ud-din et al.[29]	Omics (GE) + Molecular Descriptor	0.32
Liu et al.[14]	Omics (GE) + Drug SMILES	0.82
Xia et al.[30]	Multi-omics (GE) + Molecule Descriptor	0.6
Li et al.[31]	Multi-omics (GE) + Molecular Descriptor	0.78
Chang et al.[32]	Multi-omics (GE) + Molecular Descriptor	0.85
Our Method	Omics (GE) + Drug SMILES	0.82 (±0.04)

from pre-trained Transformers. CHEM-BERT[23] is another transformer that extracts features of drug SMILES and the chemical context of molecules. TransGRU[24] is an innovative approach that integrates Transformers and a bidirectional LSTM name BiGRU[25] to capture local and global information about the atomic positions of functional groups and their relative positions and further extract features from the SMILES. Though successful in extracting the complex nature of atomic associations in drug SMILES, these models lack interpretability.

Recognizing this gap, our proposed method extends beyond one-atom sequences in SMILES strings by incorporating their counts and considering local and global associations among the atoms. The key differentiator of our method lies in its use of simple N-grams for enhanced interpretability, thereby making the feature vectors obtained from the NLP operation more explainable (see Figure 2).

In conclusion, our novel NLP-based feature extraction technique using SMILES notation presents an effective and promising solution for developing more refined ML models for PSD. With its availability through a Python library and the focus on interpretability, this method holds substantial potential for advancing personalized medicine and drug development.

Following this introduction, the paper is structured as follows: The "Results" section offers a comprehensive analysis, starting with a comparative study of Morgan fingerprints bit-vectors and NLP-based features through 2-D Embeddings, followed by specific case studies centered on pan-cancer and glioblastoma multiforme (GBM) for PSD. Next, the "Discussion" section provides an in-depth exploration of the implications and interpretations of the results. The "Limitations of the study" section candidly discusses the possible shortcomings and areas of improvement. Finally, the "STAR Methods" section details our proposed method and furnishes information on data and code availability.

RESULTS

This section is divided into two parts: first, a comparative study of Morgan fingerprints bit-vectors with NLP-based features through 2-D embeddings and reasoning; second, pan-cancer and GBM-specific case studies for PSD.

Embeddings analysis: NLP-based features vs. Morgan fingerprints

The first graph (a) for Figure 3 is the embedding of NLP-based features, and the remaining five are the embeddings of the features obtained from the Morgan fingerprints molecular descriptor. These embeddings are produced using the UMAP[26] algorithm using the standard configurations. This embedding case study aims to show the discrepancies among the 173 drugs, which is crucial in building effective ML models for PSD.

The sparsity of the NLP-based features, shown in (a), is an essential aspect as it highlights the distinctiveness of each cancer drug. This means that the differences between the drugs are captured in the features, and thus, the ML models built using these features can capture the precise effects of the drugs.

Sparse features blended with dense features can help develop better-performing[27] models. The results of the PSD case studies in this section provide further evidence for the effectiveness of the NLP-based features in building effective ML models for drug screening. This

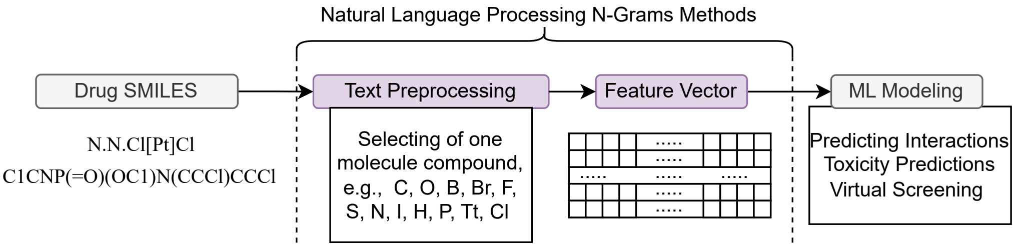

Natural Language Processing N-Grams Methods
Drug SMILES
N.N.Cl[Pt]Cl
C1CNP(=O)(OC1)N(CCCl)CCCl
Text Preprocessing
Selecting of one molecule compound, e.g., C, O, B, Br, F, S, N, I, H, P, Tt, Cl
Feature Vector
ML Modeling
Predicting Interactions
Toxicity Predictions
Virtual Screening

Figure 1. Overview of Drug SMILE Feature Extraction Process using NLP

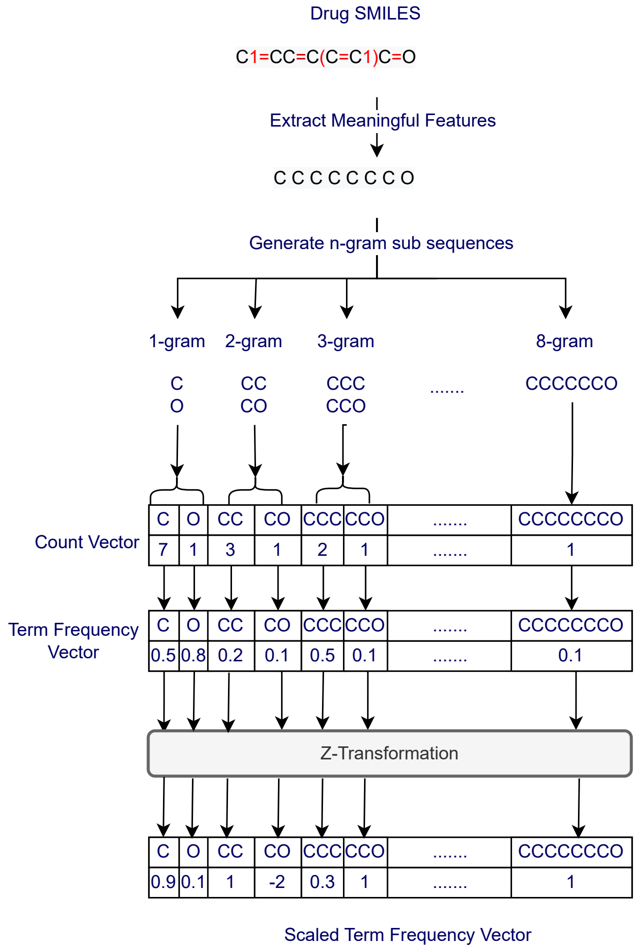


Figure 2. Feature Extraction from Drug SMILE using N-grams

will help to strengthen the conclusion that the NLP-based features are a viable option for developing better ML models for PSD using drug SMILES.

Machine learning-driven personalized drug screening: A comparative study of NLP-based and Morgan fingerprint features in pan-cancer and GBM-specific case studies

The PSD case studies aim to use gene expression data to improve cancer treatment personalization. ML models are built to predict the efficacy of cancer drugs, specifically the LN(IC50) value, which measures the drug's ability to inhibit cell growth. The prediction is based on the gene profile of the patient, the type of cancer being treated, and the drug's features. The input to the model includes gene expression data from 657 genes, the cancer type, and drug features.

(Note: the data preparation for the NLP-based experiment, the NLP-based drug features were combined with GE and cancer type, see Figure 4. And in Morgan fingerprint experiments, the Fingerprint bit vectors were joined with the same GE and cancer types.).

The data were divided into a training set (80% of the data) and a test set (the remaining 20%) to build the ML model. The training data were used to create the model, with 10-fold cross-validation to improve its accuracy. The prediction of drug efficacy was treated as a regression problem and evaluated using various metrics, including mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R-square score (R2), root-mean-square logarithmic error (RMSLE), and mean absolute percentage error (MAPE). The model's predictions were tested on the test data, and the results were analyzed to determine the accuracy and reliability of the predictions.

The Drug Efficacy values (LN(IC50)) in the original data range from −10 to +10. Hence, a lower value of MAE, MSE, RMSE, RMSLE, and MAPE indicates that the model performs better. A value of zero for R2 means the model is not significant, and the model's result is highly accurate when it's one. While all the metrics are important, the observations are mainly based on MAE, as its error values are close to the actual LN(IC50) and are least affected by outliers.

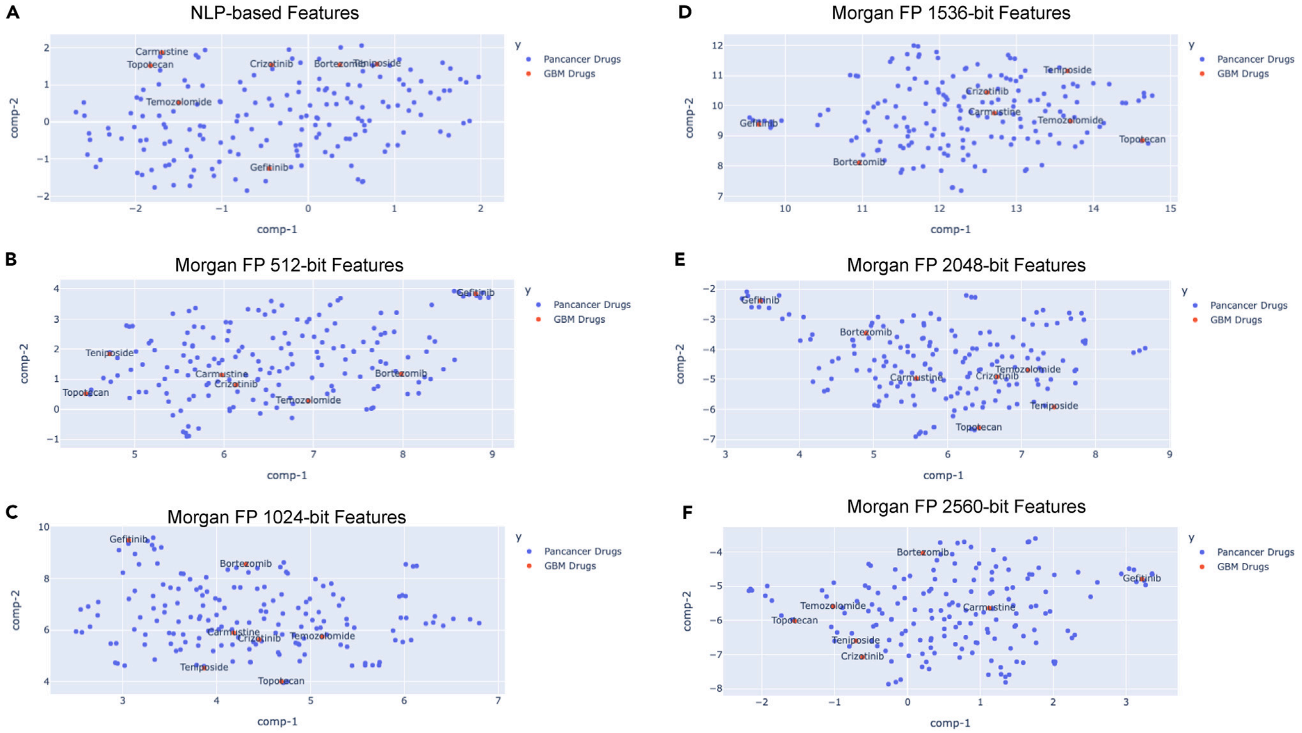


Figure 3. UMAP-derived 2D representations of drug features

(A) Visualization of 2D embeddings for novel NLP-based features derived from the SMILES notations of 173 drugs, represented in a feature space comprising 3196 dimensions.

(B) Two-dimensional representation of Morgan fingerprints encapsulated within a 512-bit vector space.

(C) Depiction of 2D embeddings for Morgan fingerprints, illustrated in a 1024-bit vector format.

(D) Spatial representation in 2D of Morgan fingerprints, expressed in a 1536-bit vector dimension.

(E) Graphical illustration of 2D embeddings corresponding to Morgan fingerprints with a vector size of 2048 bits.

(F) Two-dimensional layout of Morgan fingerprints, detailed in a 2560-bit vector framework.

Each panel (A–F) showcases the diversity and complexity of the drug features through the lens of UMAP's dimensional reduction capabilities.

Pan-cancer drug efficacy prediction: Morgan fingerprints vs. NLP features

The presented Table 2 reveals results from a pan-cancer case study encompassing 31 types of cancer. The data used for model training and testing was sourced from the Genomics of Drug Sensitivity of Cancer (GDSC2), encompassing information about 173 cancer drugs across these cancer types.

This study's primary objective (hypothesis) was to develop a LightGBM machine learning regressor that could predict drug sensitivity based on several factors: the type of drug used, the specific disease in question, and the gene expression of the patient's cancer cells.

The broader goal was to test if a LightGBM regressor trained on such a diverse dataset could generalize well and accurately predict drug efficacy across different types of cancer. The intent was to enhance personalized treatment by tailoring it to individual patient profiles - their specific type of cancer and gene expression.

The Table 2 presents the performance of ML models using different configurations of Morgan fingerprints (bit lengths of 512, 1024, 1536, 2048, and 2560) and NLP-based Features. These models were built to predict drug sensitivity in 31 types of cancer based on the drug, disease, and gene expression of the patient's cancer cells.

- Morgan FP 512-bits: This model had an MAE of 0.8993, MSE of 1.4348, RMSE of 1.1978, R2 score of 0.8211, RMSLE of 0.3274, and MAPE of 4.3281. This means the model was reasonably accurate but had a higher percentage error (MAPE) than other models.
- Morgan FP 1024-bits: The performance of this model was quite like the 512-bit model. However, it improved in the R2 score (0.8231), indicating a better fit, and had a significantly lower MAPE (2.9007), suggesting a reduction in percentage error.
- Morgan FP 1536-bits: This model showed improved performance across almost all metrics. It had the best R2 score (0.8264) among the Morgan FP models, indicating the best fit to the data and lower errors in terms of MAE, MSE, RMSE, and RMSLE. Its MAPE was a bit higher but still lower than the 512-bit model.
- Morgan FP 2048-bits and 2560-bits: The performances of these models were similar to the 512 and 1024-bit models, indicating no significant advantage in using a higher bit length in this case.

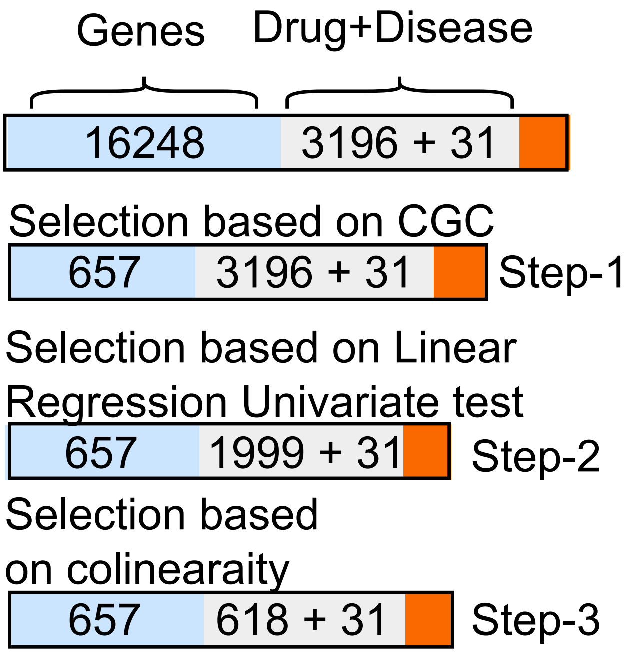


Figure 4. Feature Engineering for ML modeling using NLP-based features

- NLP-based Features: This model outperformed all Morgan FP models, with the lowest errors across all metrics: MAE of 0.8738, MSE of 1.363, RMSE of 1.1675, RMSLE of 0.3192, and MAPE of 1.31. Its R2 score of 0.8221 was competitive, indicating an excellent fit to the data.

In conclusion, while all models showed potential in predicting cancer drug efficacy, the NLP-based Features model provided the most accurate and reliable predictions, suggesting it might be the best choice for this application. This aligns with the study's hypothesis that a model considering a broad range of cancer types can effectively predict drug efficacy based on drug, disease, and gene expression data. This is a promising indication for the enhancement of personalized cancer treatment strategies.

Predicting drug sensitivity in glioblastoma: A comparative case study

This case study focuses on glioblastoma (GBM) cancer and represents a departure from the pan-cancer approach in the previous study. Instead of using a diverse dataset, the test data, in this case, is exclusively limited to GBM cancer samples. However, the model was still trained using the remaining pan-cancer data, which includes GBM cancer samples not part of the test set. This mirrors the training and validation procedure of the pan-cancer study. The objective (hypothesis) was to assess if a model trained in this way could accurately predict the LN(IC50) values (drug sensitivity) based on the drug, disease, and gene expression of the GBM cancer cells of the patients.

All 31 cancer types were considered in the training samples, with all 173 drugs used. The test samples were specifically from GBM cancer, randomly selected from all available GBM samples in the dataset.

The results (see Table 3) of the GBM cancer case study are as follows.

- Morgan FP 512-bits: This model showed improvement from the pan-cancer study, with an MAE of 0.8392, MSE of 1.1791, RMSE of 1.0859, R2 score of 0.8568, RMSLE of 0.2832, and MAPE of 0.5184.

Table 2. Pan-cancer comparative study

	MAE	MSE	RMSE	R2	RMSLE	MAPE
Morgan FP 512-bits	0.8993	1.4348	1.1978	0.8211	0.3274	4.3281
Morgan FP 1024-bits	0.8999	1.4321	1.1967	0.8231	0.3266	2.9007
Morgan FP 1536-bits	0.8892	1.4048	1.1852	0.8264	0.323	4.1501
Morgan FP 2048-bits	0.8962	1.413	1.889	0.8212	0.3285	3.486
Morgan FP 2560-bits	0.9075	1.457	1.2071	0.8183	0.3303	2.8912
NLP-based Features	0.8738	1.363	1.1675	0.8221	0.3192	1.31

Table 3. Comparative study for glioblastoma multiforme samples

	MAE	MSE	RMSE	R2	RMSLE	MAPE
Morgan FP 512-bits	0.8392	1.1791	1.0859	0.8568	0.2832	0.5184
Morgan FP 1024-bits	0.8158	1.0798	1.0391	0.8602	0.2783	0.5343
Morgan FP 1536-bits	0.8615	1.2453	1.1159	0.8371	0.2906	0.8687
Morgan FP 2048-bits	0.8231	1.1023	1.0499	0.8548	0.2818	0.6729
Morgan FP 2560-bits	0.8354	1.1527	1.0736	0.8329	0.2837	0.7954
NLP-based Features	**0.7527**	0.9135	0.9558	0.8557	0.26	0.615

- Morgan FP 1024-bits: The performance of this model improved considerably, especially in terms of error rates (MAE, MSE, RMSE, and RMSLE), while maintaining a solid R2 score of 0.8602 and a reasonable MAPE of 0.5343.
- Morgan FP 1536-bits to 2560-bits: These models demonstrated similar performance to the 512 and 1024-bit models, with minor fluctuations in error rates and R2 score.
- NLP-based features: Like in the pan-cancer study, this model outperformed all others. It had the lowest errors across all metrics: MAE of 0.7527, MSE of 0.9135, RMSE of 0.9558, RMSLE of 0.26, and MAPE of 0.615. The R2 score of 0.8557 was competitive, suggesting a perfect fit for the data.

In conclusion, the GBM-specific study aligns with the hypothesis that a pan-cancer model, when trained on a diverse set of cancer data excluding some GBM samples, can effectively generalize and accurately predict drug sensitivity for GBM cancer samples. As in the pan-cancer study, the NLP-based Features model offered the most precise predictions, indicating its reliability and effectiveness. The results further suggest the potential of ML models in enhancing personalized treatment strategies for specific cancer types like GBM.

DISCUSSION

The research underscores the robust potential of the NLP-based feature extraction method for enhancing ML models in PSD. This study enormous maximum possible feature based on the 11 atoms using their SMILES codes, which consisted of 11 single-atom elements: O, S, B, F, I, C, Br, Cl, Pt, N, and P.

Employing the NLP-based feature extraction method led to discovery of 3196 unique features. These features can each have "n" unique values, leading to an expansive sample space of up to n^{3196} (see Table 4). This sample space offers a broader capacity for pattern recognition, facilitating better differentiation between similar and dissimilar drugs. The method's effectiveness is backed by the total number of potential features (8712156) and the number of features extracted using various NLP methods for the 173 drugs, as presented in Table 5.

The larger sample space and unique sparsity of NLP-based features highlighted the distinctiveness of each drug, enabling the models to capture their precise effects. This level of differentiation is crucial for personalized medicine, as it informs precise predictions of drug efficacy.

The NLP-based features' performance outperformed traditional Morgan fingerprint bit-vectors consistently across two case studies: the pan-cancer and GBM-specific studies. The lower mean absolute error (MAE) and other error metrics achieved by the NLP-based features models in both studies demonstrate the effectiveness of the proposed feature extraction method.

The NLP-based feature extraction method further stands out in capturing meaningful relationships among atoms locally (atom-level) and globally (molecule-level). This high degree of explainability, combined with the results from the case studies and the UMAP embedding-based comparative study, confirms the potential of NLP-based features in building effective ML models for PSD.

Table 4. Feature and sample space for the drug features extraction methods

Method	Feature Count	Sample Count
Morgan Fingerprint 512-bits	512	2^{512}
Morgan Fingerprint 1024-bits	1024	2^{1024}
Morgan Fingerprint 1536-bits	1536	2^{1536}
Morgan Fingerprint 2048-bits	2048	2^{2048}
Morgan Fingerprint 2560-bits	2560	2^{2560}
NLP-based features	3196	n^{3196}

"n" is the count of feature values (i.e., atom sequence count); "n" is integer ≥ 0

Table 5. Count of maximum possible features and features extracted from the NLP-based methods

NLP Method	Possible Features	Features Extracted for 173 Drugs
1-Gram	$^{11}P_1$ = 11	11
2-Gram	$^{11}P_2$ = 55	46
3-Gram	$^{11}P_3$ = 990	115
4-Gram	$^{11}P_4$ = 7920	224
5-Gram	$^{11}P_5$ = 55440	367
6-Gram	$^{11}P_6$ = 332640	561
7-Gram	$^{11}P_7$ = 1662300	800
8-Gram	$^{11}P_8$ = 6652800	1072
Total	8712156	3196

In conclusion, the study provides compelling evidence that NLP-based features significantly improve efficacy predictions in drug discovery and personalized medicine. The distinct advantage of the method in pattern recognition, interpretability, and precision illustrates its worthiness in ongoing research and practical applications in AI-based drug development.

Limitations of the study

The limitation of the NLP-based feature extraction is the ample feature space. Table 5 depicts the enormous maximum possible feature based on the 11 atoms, i.e., 8712156. If an exhaustive list of drugs is used, we may obtain petabytes of data only related to drug features. This would require vast storage space and large computational capacity.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ NLP-based feature extraction from drug SMILES
  - ○ ML modeling for virtual drug screening

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109127.

ACKNOWLEDGMENTS

The research reported in this article was partially supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR003096, of which J.Y.C. is a co-investigator. The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

R.S. was responsible for implementing the research work, conducting experiments, validating results, drafting the manuscript, and implementing the Python library. E.S. helped review the manuscript, prepare the supplemental information, and perform exploratory analysis. J.Y.C. provided the scientific design, overall supervision, technical feedback, manuscript revision, and financial support to the research.

DECLARATION OF INTERESTS

There is no conflict of interest.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

Statement: During the preparation of this work the author(s) used ChatGPT in order to improve the language of the article. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Received: February 15, 2023
Revised: September 30, 2023
Accepted: February 1, 2024
Published: February 8, 2024

REFERENCES

1. Seal, S., Carreras-Puigvert, J., Trapotsi, M.A., Yang, H., Spjuth, O., and Bender, A. (2022). Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. Commun. Biol. *5*, 858. https://doi.org/10.1038/s42003-022-03763-5.
2. Banerjee, P., and Preissner, R. (2018). BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds. Front. Chem. *6*, 93. https://doi.org/10.3389/fchem.2018.00093.
3. Zhang, J., Mucs, D., Norinder, U., and Svensson, F. (2019). LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets. J. Chem. Inf. Model. *59*, 4150–4158. https://doi.org/10.1021/acs.jcim.9b00633.
4. Vo, T.H., Nguyen, N.T.K., and Le, N.Q.K. (2023). Improved prediction of drug-drug interactions using ensemble deep neural networks. Med. Drug Discov. *17*, 100149. https://doi.org/10.1016/j.medidd.2022.100149.
5. Luo, Q., Mo, S., Xue, Y., Zhang, X., Gu, Y., Wu, L., Zhang, J., Sun, L., Liu, M., and Hu, Y. (2021). Novel deep learning-based transcriptome data analysis for drug-drug interaction prediction with an application in diabetes. BMC Bioinf. *22*, 318. https://doi.org/10.1186/s12859-021-04241-1.
6. Pang, S., Zhang, Y., Song, T., Zhang, X., Wang, X., and Rodriguez-Patón, A. (2022). AMDE: a novel attention-mechanism-based multidimensional feature encoder for drug–drug interaction prediction. Brief. Bioinform. *23*, bbab545. https://doi.org/10.1093/bib/bbab545.
7. Zhang, J., Chen, M., Liu, J., Peng, D., Dai, Z., Zou, X., and Li, Z. (2023). A Knowledge-Graph-Based Multimodal Deep Learning Framework for Identifying Drug–Drug Interactions. Molecules *28*, 1490. https://doi.org/10.3390/molecules28031490.
8. Chen, Z.H., You, Z.H., Guo, Z.H., Yi, H.C., Luo, G.X., and Wang, Y.B. (2020). Prediction of Drug–Target Interactions From Multi-Molecular Network Based on Deep Walk Embedding Model. Front. Bioeng. Biotechnol. *8*, 338. https://doi.org/10.3389/fbioe.2020.00338.
9. Song, T., Zhang, X., Ding, M., Rodriguez-Paton, A., Wang, S., and Wang, G. (2022). DeepFusion: A deep learning based multi-scale feature fusion method for predicting drug-target interactions. Methods *204*, 269–277. https://doi.org/10.1016/j.ymeth.2022.02.007.
10. Lee, I., Keum, J., and Nam, H. (2019). DeepFusionDTA: DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput. Biol. *15*, e1007129. https://doi.org/10.1371/journal.pcbi.1007129.
11. Pu, Y., Li, J., Tang, J., and Guo, F. (2022). Drug-Target Binding Affinity Prediction With Information Fusion and Hybrid Deep-Learning Ensemble Model. IEEE/ACM Trans. Comput. Biol. Bioinform. *19*, 2760–2769. https://doi.org/10.1109/TCBB.2021.3103966.
12. Shao, J., Gong, Q., Yin, Z., Pan, W., Pandiyan, S., and Wang, L. (2022). S2DV: converting SMILES to a drug vector for predicting the activity of anti-HBV small molecules. Brief. Bioinform. *23*, bbab593. https://doi.org/10.1093/bib/bbab593.
13. Monteiro, N.R.C., Ribeiro, B., and Arrais, J.P. (2019). Deep Neural Network Architecture for Drug-Target Interaction Prediction. In International Conference on Artificial Neural Networks, pp. 804–809. https://doi.org/10.1007/978-3-030-30493-5_76.
14. Liu, P., Li, H., Li, S., and Leung, K.S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinf. *20*, 408. https://doi.org/10.1186/s12859-019-2910-6.
15. Xu, Z., Wang, S., Zhu, F., and Huang, J. (2017). Seq2seq Fingerprint. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics (ACM), pp. 285–294. https://doi.org/10.1145/3107411.3107424.
16. DiPietro, R., and Hager, G.D. (2020). Deep learning: RNNs and LSTM. In Handbook of Medical Image Computing and Computer Assisted Intervention (Elsevier), pp. 503–519. https://doi.org/10.1016/B978-0-12-816176-0.00026-0.
17. Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput. *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
18. Goh, G.B., Hodas, N.O., Siegel, C., and Vishnu, A. (2017). SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. Preprint at arXiv. https://doi.org/10.48550/arXiv.1712.02034.
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, *30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
20. Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. (2019). SMILES-BERT. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM), pp. 429–436. https://doi.org/10.1145/3307339.3342186.
21. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, *Volume 1*, J. Burstein, C. Doran, and T. Solorio, eds (Association for Computational Linguistics), pp. 4171–4186. Long and Short Papers. https://doi.org/10.18653/v1/n19-1423.
22. Maziarka, L., Danel, T., Mucha, S., Rataj, K., Tabor, J., and Jastrzebski, S. (2020). Molecule Attention Transformer. CoRR. https://arxiv.org/abs/2002.08264.
23. Kim, H., Lee, J., Ahn, S., and Lee, J.R. (2021). A merged molecular representation learning for molecular properties prediction with a web-based service. Sci. Rep. *11*, 11028. https://doi.org/10.1038/s41598-021-90259-7.
24. Jiang, J., Zhang, R., Ma, J., Liu, Y., Yang, E., Du, S., Zhao, Z., and Yuan, Y. (2023). TranGRU: focusing on both the local and global information of molecules for molecular property prediction. Appl. Intell. *53*, 15246–15260. https://doi.org/10.1007/s10489-022-04280-y.
25. Chakrabarty, A., Pandit, O.A., and Garain, U. (2017). Context Sensitive Lemmatization Using Two Successive Bidirectional Gated Recurrent Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 1481–1491. https://doi.org/10.18653/v1/P17-1136.
26. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.
27. Moldovanu, S., Toporaș, L.P., Biswas, A., and Moraru, L. (2020). Combining Sparse and Dense Features to Improve Multi-Modal Registration for Brain DTI Images. Entropy *22*, 1299. https://doi.org/10.3390/e22111299.
28. Menden, M.P., Iorio, F., Garnett, M., McDermott, U., Benes, C.H., Ballester, P.J., and Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS One *8*, e61318. https://doi.org/10.1371/journal.pone.0061318.
29. Ammad-ud-din, M., Georgii, E., Gönen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., Poso, A., and Kaski, S. (2014). Integrative and Personalized QSAR Analysis in Cancer by Kernelized Bayesian Matrix Factorization. J. Chem. Inf. Model. *54*, 2347–2359. https://doi.org/10.1021/ci500152b.
30. Xia, F., Allen, J., Balaprakash, P., Brettin, T., Garcia-Cardona, C., Clyde, A., Cohn, J., Doroshow, J., Duan, X., Dubinkina, V., et al. (2022). A cross-study analysis of drug response prediction in cancer cell lines. Brief. Bioinform. *23*, bbab356. https://doi.org/10.1093/bib/bbab356.
31. Li, M., Wang, Y., Zheng, R., Shi, X., Li, Y., Wu, F.X., and Wang, J. (2021). DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Trans. Comput. Biol. Bioinform. *18*, 575–582. https://doi.org/10.1109/TCBB.2019.2919581.
32. Chang, Y., Park, H., Yang, H.J., Lee, S., Lee, K.Y., Kim, T.S., Jung, J., and Shin, J.M. (2018). Cancer Drug Response Profile scan

(CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Sci. Rep. *8*, 8857. https://doi.org/10.1038/s41598-018-27214-6.

33. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961. https://doi.org/10.1093/nar/gks1111.

34. Gao, H., Korn, J.M., Ferretti, S., Monahan, J.E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat. Med. *21*, 1318–1325. https://doi.org/10.1038/nm.3954.

35. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.

36. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2021). PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. *49*, D1388–D1395. https://doi.org/10.1093/nar/gkaa971.

37. Ali, M. (2020). PyCaret: An open source, low-code machine learning library in Python. https://www.pycaret.org.

38. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, *30*, I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and Garnett, eds. (Curran Associates, Inc). https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Python Library drug-smile-fet	PyPI	https://pypi.org/project/drug-smile-fet/
Python Library Source Code	GitHub	https://github.com/rahulsharma-rs/drug-smile-fet
Experiment Code	UAB Box	https://uab.app.box.com/v/iscience-dsfet

RESOURCE AVAILABILITY

Lead contact

Dr. Jake Y. Chen (jakechen@uab.edu).

Materials availability

With this article, we are providing a supplemental information file that provides information about the following:

(1) Cancer-wise drug screening results, see Table S1 in the supplemental information.
(2) Cancer-wise plots of Real and Predicted LN(IC_{50}) values, see Table S5 in the supplemental information.
(3) The result of the Model Comparison, which formed the basis for choosing the LightGBM model, see Table S2 in the supplemental information.
(4) Table 3 provides the features extracted from the NLP-based n-gram method.
(5) Table 5 lists the 657 genes used in the personalized drug screening experiments.
(6) Section 1.5 of the supplemental information details the data used in this article.

Data and code availability

Data availability

Genomics of Drug Sensitivity of Cancer (GDSC2)[33] is the primary data source from which we get 135242 perturbation information, i.e., log(IC_{50}) values, of 198 drug sensitivities on 809 cancer cell lines. The data sets also provide The Cancer Genome Atlas (TCGA)[34] classification of the samples depicting 31 types of cancer. The GDSC database provides the sample ID (or COSMIC_ID) of the 809 cell lines from the Catalogue of Somatic Mutations (COSMIC) database.[35] Each cell line provides the Gene Expression (GE) values of 16248 genes normalized through Z-transformation for each gene. GDSC also provides the PubChem ID of 173 out of 198 drugs; therefore, 173 cancer drug smiles are obtained from the PubChem database.[36] The gene expressions from the COSMIC database and drug SMILES from the PubChem database are integrated into the GDSC dataset for our research work.

Code availability

The code related to the Embedding and Personalized drug screening experiments is provided through the box platform. The resource contains seven notebooks in the data folder, the raw, intermediate, and processed data for both experiments. The results are stored in the results and plots folders. The code and experiment-related materials can be obtained from the following Box link https://uab.box.com/v/iscience-dsfet.

Python library

The name of the python library is drug-smile-fet (it means drug SMILES feature extraction tool). This library aims to provide users a low-code interface to extract meaningful features from Drug SMILES for ML modeling. Currently, the library has only one method, i.e., the NLP-based feature extract extraction method proposed in this article. We are building to provide the implementation of all scientifically proven methods that extract features from drug SMILES for machine learning operation.

The Python library is publicly available through the Python Package Index (PyPI), which is a repository of software for the Python programming language (see https://pypi.org/project/drug-smile-fet/). The library is easy to install using the command ***pip install drug-smile-fet*** and requires a library named RdKit for Drug SMILE pre-processing. The source of the Python library is available on GitHub under an MIT license (see https://github.com/rahulsharma-rs/drug-smile-fet.git). The installation instructions are provided both at PyPI and GitHub descriptions.

METHOD DETAILS

NLP-based feature extraction from drug SMILES

This method is one of the main contributions of this article. The motivation for using the NLP-based feature extraction from drug SMILES is the sequence representation of the SMILES, which is like the Natural Language sequences such as words, sentences, etc. Figure 2 shows our approach to creating a feature vector from drug SMILES. The features are extracted using the following steps:

(1) Obtain all drug SMILES as a list.
(2) For each character in a SMILE, create a sub-SMILE consisting of a One-character legitimate molecule. After this, we get a sub-sequence of the SMILE consisting of a One-character molecule.
(3) The second step is performed for all the SMILES in the list to create a Bag-of-words.
(4) In this step, the N-gram method is used to create a count vector of 8-types: 1-character sequence, e.g., "C"; 2-character sequence, e.g., "CC"; 3-character e.g., "CCO"; 4-character sequence, e.g., "CCCC"; 5-character sequence, e.g., "CCCC"; 6-character sequence e.g. "CCCCCC"; 7-character sequence e.g. "CCCCCCC"; and 8-character sequence e.g. "CCCCCCCO". Figure 2 shows the vector created through this process.
(5) We can also use the count vector from step 4 for modeling, but when the SMILE sequences are longer, they will have higher average count values than the shorter SMILE sequence, and hence the vector will be biased. To overcome this issue, we use Term Frequency by dividing the occurrence of the sub-sequences in a SMILE by the total number of characters in the SMILE sequence. Consequently, we obtain a normalized Term Frequency vector, as shown in Figure 2.
(6) Since gene-expressing values are scaled through Z-transformation, therefore, the Term Frequency vectors are also scaled using Z-transformation as depicted by the following Equation 1:

$$z = \frac{feature - \mu}{\sigma} \qquad \text{(Equation 1)}$$

Where "μ" mean of the training samples, and "σ" standard deviation of training samples.

ML modeling for virtual drug screening

Data Preparation

In our research work, we used three datasets where: GDSC is a multivariate dataset consisting of numerical and categorical information; the COSMIC dataset is a univariate dataset providing numeric gene expression values; drug SMILE is a text dataset. For modeling, we need to integrate these datasets, but due to their heterogeneity, they are processed individually.

From the GDSC database, we use two data sets, the drug perturbation dataset and the Drug information dataset. The drug information database consists of the PubChem ID of the cancer drugs used to extract drug SMILES from the PubChem database. From the drug perturbation dataset, the following features are selected:

(1) COSMIC_ID: The identification of cell line sample id from the COSMIC database.
(2) Drug_Id: Drug Identification information.
(3) Drug_name: Name of Drug.
(4) TCGA_DESC: TCGA is the name of the cancer type.
(5) LN_IC50: Drug perturbations information.

Besides these five features, all the features are dropped from the perturbation dataset. The TCGA_DESC feature is On-Hot coded to create 31 columns depicting the binary coded feature for each cancer type. The Drug data is obtained using our proposed NLP-based method and Morgan Fingerprints. The Drug_Id and Drug_name features integrate the drug data into the perturbation dataset. The COSMIC_ID feature integrates the gene-expressing data obtained from the COSMIC database. At last, all the data samples consisting of missing values are dropped from the combined dataset. Having processed the data, we obtain a dataset that contains dependent features related to Drugs, Diseases, and Genes, along with the drug perturbations as an independent feature.

Feature Engineering

After NLP-based feature extraction, we obtained 3196 features for drug molecules. After processing the GDSC dataset, we got 31 features about 31 cancer types. The COSMIC database provides the expression values of 16248 genes. Altogether we had 19321 dependent features and one Target feature. We used three steps to select the best features: In step-1, Based on the Cancer Gene Census (CGS) from the COSMIC database, we selected 657 out of the 16248 genes (see Figure 4); In Step-2, we used all features after Step-1 and calculated feature scores based on the Linear Regression Univariate test. In Step-2, all the features with a score >= 70 were selected; In Step-3, collinearity among all the features was determined, and all collinear features were removed from the feature sets. Consequently, we obtained 1306 features where most drug features were removed, and both gene and disease-related features were retained, see Figure 4.

Model Selection

The drug perturbation prediction is a regression problem. For this research, we used a low code ML library named PyCaret[37] to perform regression modeling using the following methods: Linear Regression, Random Forest, ADA Boost, Gradient Boosting regressor, Extra Tree regressor, decision tree, ridge regression, K nearest regressor, and LightGBM. Based on the observation, the LightGBM regressor was the best and was selected for modeling (see Table S2 provides the comparative study of Model Selection).

Model development and evaluation

For modeling, the pre-processed data were randomly divided into two parts: 1) a training set consisting of 80% of the data; 2) a test set consisting of 20% of the data. The Light Gradient Boosting Machine (LigntGBM)[38] regressor is used for modeling in our research. The model was trained using the training set and 10-fold cross-validation. The model was evaluated on a test set using six evaluation metrics, i.e., Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-square score (R2), Root Mean Square Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE).