Enhancing Pharmaceutical Target Prediction Through Intelligent Feature Optimization and Ensemble Classification

*Ajay Kumar[1], Shashikant Gupta[2]
[1,2] Lincoln University College Malaysia
*Corresponding Email ID : ajay.phdcse@gmail.com

Abstract: The pharmaceutical industry has been in big trouble as far as accelerating the drug discovery process is concerned. These issues involve high costs of development, a time-consuming development process and numerous failures. The context-aware mechanisms with metaheuristic-guided feature selection and hybrid classification can also be combined in this paper to describe a new computational method that enhances drug-target interaction prediction. The pharmaceutical data is processed by using several stages of preparation in our methodology that involve the standardisation of the text, the division into linguistic tokens and the extraction of semantic features. The overall new concept is to use feature refinement with Ant Colony Optimisation and a hybrid classifier that combines the Elements of a Random Forest and a Logistic Regression. Pharmaceutical dataset tests indicate that this approach is much more effective than the prior machine learning approaches, with a 98.6% accuracy and a 0.985 F1-score. This approach facilitates the process of discovering candidates more quickly, reduces development times, and finds real-world applications in precision medicine and drug optimization.

Keywords: Drug discovery; Feature optimization; Ensemble learning; Metaheuristic algorithms; Pharmaceutical prediction; Machine learning classification

Introduction

Discovery of therapeutic agents remains highly relevant to enhance human health and manage the issues associated with chronic illnesses[1]. The traditional pharmaceutical development cycle is a series of steps, which involve the selection of a target, preclinical testing, clinical validation, and regulatory approval. This strategy is time and money-consuming, and the attrition rates are high[2]. Recent computational methods have demonstrated potential in enhancing the efficiency of candidate screening and refining predictions of binding between medicinal drugs and biological targets[3].

The paradigms of machine learning have been increasingly used to aid the pharmacological development, enabling researchers to navigate through large chemical space and predict molecular interactions with greater accuracy[4]. However, the existing computational approaches have been shown to be lacking in contextual sensitivity, semantic reasoning of pharmacological information, and adaptive feature identification in varied datasets. These limitations impede the discovery pipeline and it is necessary to develop more sophisticated prediction systems that gradually incorporate different optimization approaches using ensemble classification frameworks.

Related work

Development, merging and connecting were employed in order to incorporate the constituents into the study compound [5]. We investigated the VAE and the reinforcement learning models. AI-enabled fragment-based drug discovery advances are useful in exploring the large chemical world in an efficient way. The Black Box problem of interpretation of the operation of the DL model limited the scope of the investigation. As it is observed in [6], inventive methods of using Deep Learning (DL) in a low-data environment gained traction. Since de novo design, protein structure prediction, and synthesis planning, DL has become more and more significant in drug discovery. Based on low-data training, the findings took the risk of predicting future directions of drug discoveries. The studies have failed to produce sufficient standards and data to standardize the process of assessment, selection and development of specialized drug discovery strategies.

Plasmodium parasites are characterized by their resistance to treatment and clear-cut incapacity to stop spreading malaria in humans, making this disease a serious social health issue [7]. The experiment compared the ML and DL cycles and discovered that Fingerprints and Graph Neural Networks (FP-GNN) model was suitable to depict the structural features of drug discovery. The limitation of the study was related to the complexity of computing large data. The two most notable are target and phenotype-based experimental screening, which are tedious, time-consuming, and costly [8]. 832 Fingerprint GNN (FP-GNN) models projected the inhibitory effects of medicines on targeted and cancerous cells. It was not done using tumor cells and DeepCancer is upgraded to attack new targets.

Key Contribution

The authors have shown that deep learning systems like VAEs, reinforcement learning and FP GNN can be effective at searching large chemical space to predict the inhibitory activity of drug candidates, even when limited data is provided. Nevertheless, AI-based drug discovery continues to be challenged in several major ways such as black-box interpretability, the lack of standard benchmarks, computational requirements, and enhanced integration to address urgent diseases such as malaria and cancer.

The proposed methodology is encouraging but has concerns of generalizability as well as computational cost as well as partial interpretability since it is yet to be tested on pharmaceutically interesting data with different distributions, in addition, there are concerns that metaheuristic optimization is too resource intensive to be used in clinical decision support in real-time and even though feature importance can be ranked using ACO, the model is less interpretable compared to simpler statistical methods.

Methodologies

We conducted research using a large pharmaceutical database containing over 11,000 medicinal entries. It contained organised attributes such as the name of pharmaceuticals, active ingredients, therapeutic purposes, manufacturing companies, visual documentation references, and consumer satisfaction measures. The data was obtained from publicly obtainable pharmaceutical databases, which ensured that the research could be done again.

Preparation of the data involved systematically processing the data in order to transform unstructured language into forms that could be processed. The procedure consists of a number of steps that occurred consecutively:

Text Standardization: They were subjected to case normalization in order to simplify pharmaceutical descriptions by eliminating punctuation marks and other odd characters and also including numbers that did not have any meaning. These operations ensure that the clinical terminology remains intact and they are yet to be in a position to interact with natural language processing activities that follow them.

Language Processing: Removing stop words eliminated otherwise insignificant words that did not contribute much to discerning this or that thing. Following tokenization, pharmacological descriptions were discontinued into their respective lexical components and specific features became easier to extract. Lemmatization was a kind of morphological reduction, as inflexed word forms were converted to their canonical forms, and thus feature cohesion improved.

New feature extraction methods converted processed pharmaceutical descriptions:

N-gram Representation: Unigram and bigram analysis have been employed to discover consecutive pattern of words. This developed numeric codes which retain data on the setting of therapeutic applications and pharmaceutical properties. This approach assists the model to identify patterns in the drugs and diseases that occur repeatedly.

Assessing Semantic similarity Cosine similarity scores were used to assess the semantic associations between pharmacological descriptions, and the angular distances between representations of vectors were computed. Medications that are highly similar in their similarity ratings share similarity in their therapeutic uses and pharmacological properties thus it is easier to locate molecules that are structurally or functionally related.

The proposed ensemble system is a combination of three diversity optimization and classification concepts which are complementary:

Contextually-aware integration: Domain-specific contextual factors (i.e., patient medical history, symptom manifestation, time, and ambient conditions) enhance prediction personalization and therapeutic relevance. This contextual layer ensures that the recommendations are specific to the treatment needs of the respective person rather than giving blanket forecasts.

A hybrid classification model: The system is a weighted ensemble model of the Random Forest and Logistic Regression. Random Forest employs combination of decision trees to establish non-linear relationships whereas Logistic Regression employs finding linear relationships and potential outcomes. A mixture of these classifiers with various weights exploits their strengths and enables them to perform in a broader pharmacological context.

Metaheuristic feature selection is used to select the best feature subsets in classification using the Ant Colony Optimization (ACO) as a guide. The algorithm imitates the manner in which pheromones are deposited in a manner that presents how significant a specific characteristic is. It then varies the

concentration levels depending on its anticipatory of contribution of the feature. Adaptive processes include:

- *Dynamic Pheromone Adjustment*: Higher concentrations let you tell the difference between features better, and the strength of the pheromone changes with time.
- *Evaporation Dynamics*: Systematic pheromone degradation keeps the algorithm from getting stuck, which makes it easier to try out different feature combinations while still keeping good performance.
- *Boundary Mutation Strategy*: Mutational operations that stay within predetermined optimal boundaries improve the use of possible feature areas without needing a lot of extra research.

Results

An extensive performance analysis was done between the proposed integrated strategy with the established baseline algorithms such as transformer-based algorithms, gradient boosting algorithms, and traditional machine learning classifiers. The suggested system got:

The accuracy of the proposed algorithm, CA-HACO-LF, in classification is 98.6% in comparison to other classifiers, BERT 97%, XGBoost 80%, and a random forest baseline 72%. Besides that, the proposed model has precision measurement parameter score of 0.985 with as BERT 0.968 and traditional RF 0.63. See figure 1 illustrates summary of the comparative accuracy and precision of the proposed framework and baseline models.

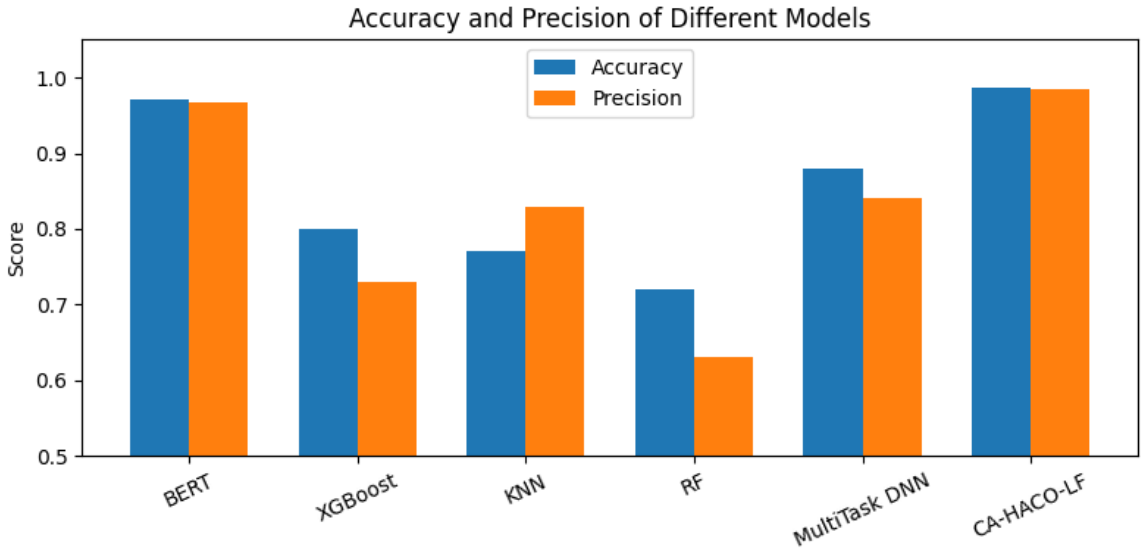

Figure 1. summarizes the comparative accuracy and precision of the proposed framework against baseline models.

The proposed model achieved a recall of 0.986, outperforming BERT with 0.963 and the traditional Random Forest model with 0.87. Additionally, the model attained a harmonic F1-score of 0.985, while the baseline methods showed scores ranging from 0.73 to 0.965 illustrated in figure 2.

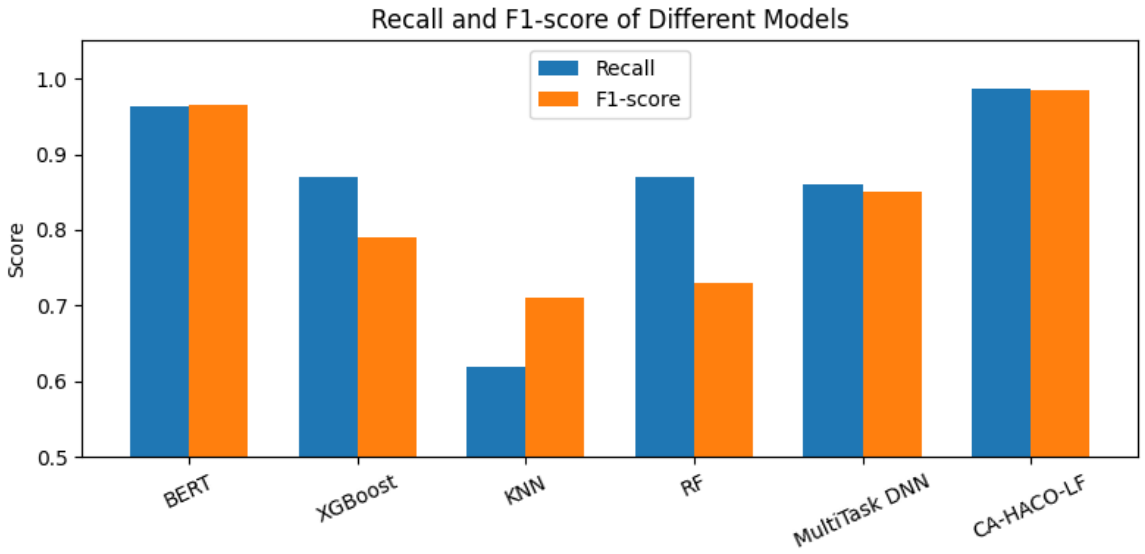

Figure 2. comparative result of Precision and Recall score achievements

Figure 3 reports Cohen's Kappa values, highlighting the stronger beyond-chance agreement achieved by the proposed framework. The error magnitude analysis further indicates reduced prediction variance, with an RMSE of 0.1446 (BERT: 0.1785), MSE of 0.0209 (BERT: 0.0318), and MAE of 0.0162 (BERT: 0.0318), reflecting improved calibration and reliability.

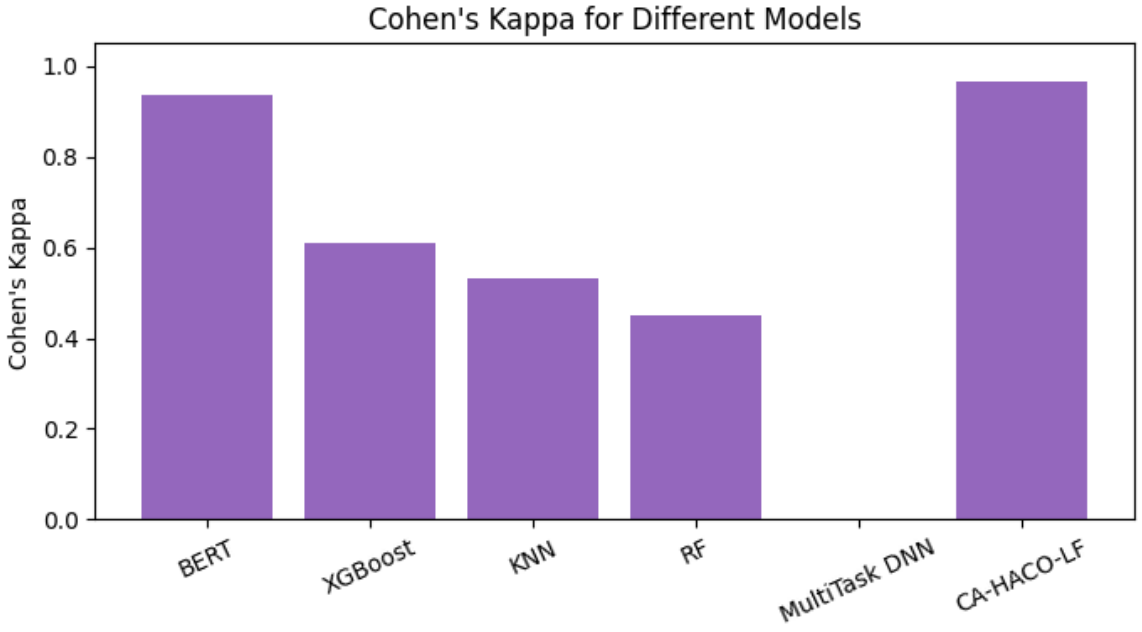

Figure 3. Cohen's Kappa and Error value measurement

Figure 4 depicts the confusion matrix for five representative disease categories, illustrating strong diagonal dominance and limited cross-class confusion.

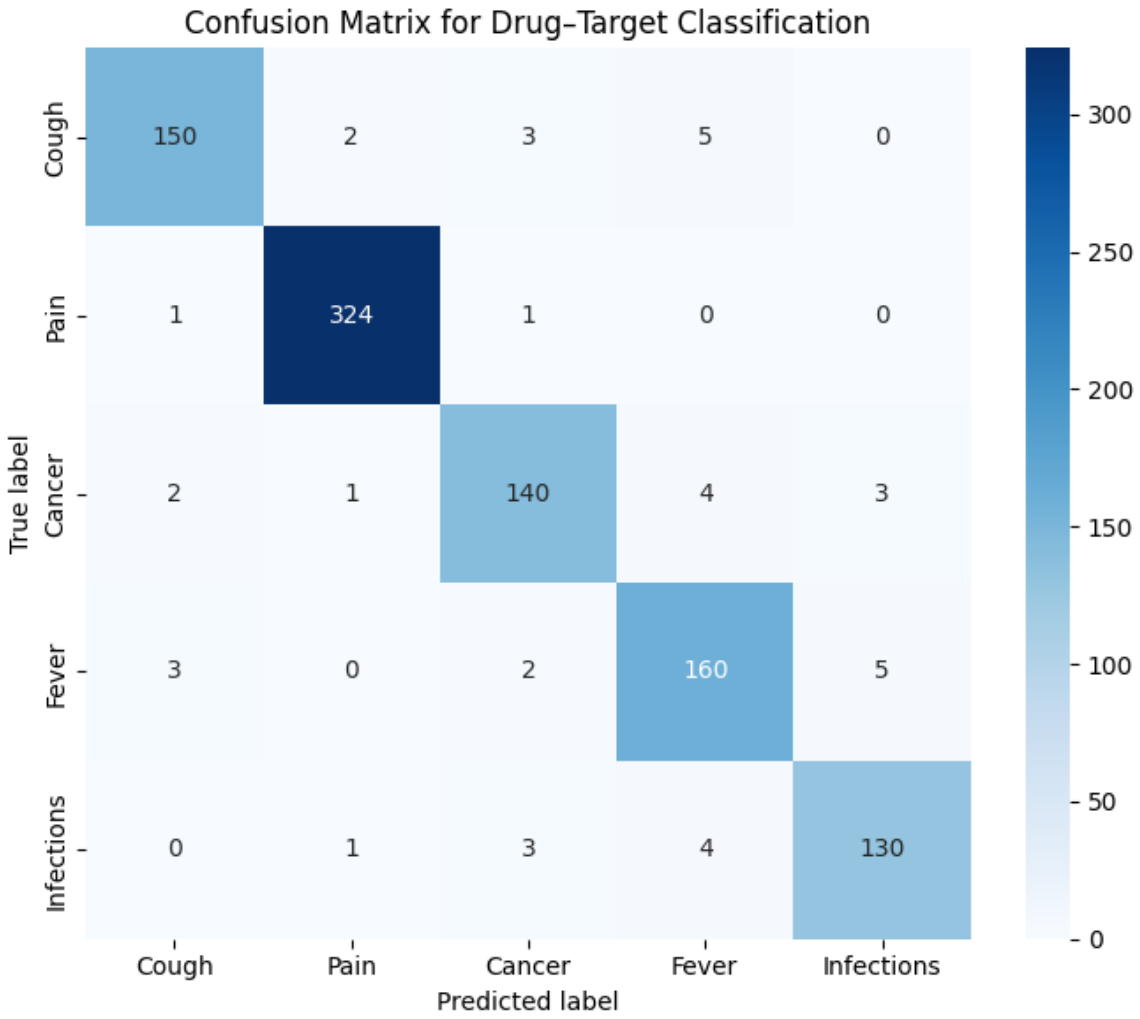

Figure 4. Confusion matrix for disease categories

Comparative Optimization Analysis

The comparative analysis of ACO and competing metaheuristic algorithms (Genetic Algorithm, Particle Swarm Optimization, Simulated Quantum Annealing Optimization) showed that ACO was more effective in the choice of features. As used in the hybrid categorization framework:

- The accuracy, precision, recall, and F1-score of ACO-based selection were 99%.
- The accuracy of the GA optimised strategy was 91, the PSO optimised strategy was 95 and SQAO optimised strategy was 96.

This performance is enhanced by the fact that, ACO could hit a balance between the refinement of promising subsets of features and the exploration of alternative feature combinations, and this keeps it out of local optima too early.

Discussions

The proposed design is significantly more effective than the traditional methods since it contains so many new ideas. To begin with, ACO guided feature selection finds the best pharmaceutical descriptors and eliminates noise and redundancy, which are the problems of most feature selection methods which operate on high-dimensional pharmaceutical data sets. Second, probabilistic as well as tree-based methods can be combined to provide hybrid classification, which would reveal both non-linear and linear pharmacological patterns. This renders the model more articulate. Third, contextual integration can contribute to the benefit of recommendations in the clinic based on the consideration of the personal characteristics of every patient and time of day. This renders them more applicable to every patient.

It has been found that specialized deep learning approaches (BERT) are equally well-functioning, though they require significant computing capabilities and do not easily integrate biological data. Stock ensemble techniques such as XGBoost and random forest are more likely to overfit medical data and appear to lack an understanding of what is going on. The suggested approach addresses these problems by utilizing adaptive feature selection and considering pharmacological context in a large number of dimensions.

Practical applications incorporate several pharmaceutical applications, such as in accelerating the screenings of prospective medications, in identifying the optimum clinical trial subjects, in discovering new methods of repositioning medications and in generating individualized therapy recommendations. The model is also effective in a variety of disorders and includes infectious diseases, pain management, cancer, and fever-related issues. This implies that it may be applicable in most areas of medicine.

Based on the confusion matrix, the CA HACO LF model is robust with regard to all five categories of illnesses. Cough and cancer are only slightly confused. This is illustrated by the high diagonal dominance particularly on the case of Pain.

However, the hybrid optimization as well as ensemble learning consumes a lot of computer power and it is not easy to apply in areas where resources are limited. Future applications may exploit model compression, distributed processing models or cloud deployment models. It should also have the capability to work with a broad spectrum of biomedical data formats and comply with all the regulations to be utilized in the pharmaceutical pipelines.

Conclusions

The methodology proposed has much potential, however, it has some issues that should be mentioned. It has not been demonstrated that the method can be applied in the context of pharmaceutical datasets that have much different distributions. The cost associated with computation of metaheuristic optimization can be a limitation to its use in real-time clinical decision-support systems. This model is interpretable, with the importance of features ranked by ACO selection, although not as interpretable as simplified statistical analysis.

Future studies ought to involve: (1) testing on larger and more diverse pharmaceutical databases to determine the degree to which the findings can be generalized; (2) combining with molecular docking simulations to determine how accurately predictions can be made; (3) creating efficient variants of the model which are applicable to environments with less than ideal computing capability; (4) long term

clinical validation comparing computer predictions with real-life findings; and (5) extending to multi-target drug discovery settings which involve compounds that have more than one therapeutic activity.

The article presents a unified computational paradigm that considerably improves the prediction of drug-target interactions with the joint utilization of the context-sensitive mechanism, metaheuristic features optimization, and ensemble classification. The empirical evaluation reveals 98.6% classification accuracy with improved results on a number of evaluation parameters compared to well-known approaches to the baseline. The technology can be applied to have new drugs discovered faster, reducing the time and cost of drug development, and improving hit rates.

The piece has a methodological contribution to computational pharmacology by overcoming the essential weaknesses of the current methods with the help of intelligent feature optimization and adaptive ensemble classification. It also demonstrates actual performance gains which are directly applicable to precision medicine and streamlining of drug development.

References

1. Adelusi, T.I., Oyedele, A.Q.K., Boyenle, I.D., Ogunlana, A.T., Adeyemi, R.O., Ukachi, C.D., Idris, M.O., Olaoba, O.T., Adedotun, I.O., Kolawole, O.E. and Xiaoxing, Y., 2022. Molecular modeling in drug discovery. *Informatics in Medicine Unlocked*, *29*, p.100880. https://doi.org/10.1016/j.imu.2022.100880
2. Shaker, B., Ahmad, S., Lee, J., Jung, C. and Na, D., 2021. In silico methods and tools for drug discovery. *Computers in biology and medicine*, *137*, p.104851. https://doi.org/10.1016/j.compbiomed.2021.104851
3. Walters, W.P. and Barzilay, R., 2021. Critical assessment of AI in drug discovery. *Expert opinion on drug discovery*, *16*(9), pp.937-947. https://doi.org/10.1080/17460441.2021.1915982
4. Vijayan, R.S.K., Kihlberg, J., Cross, J.B. and Poongavanam, V., 2022. Enhancing preclinical drug discovery with artificial intelligence. *Drug Discovery Today*, *27*(4), pp.967-984. https://doi.org/10.1016/j.drudis.2021.11.023
5. Yoo, J., Jang, W. and Shin, W.H., 2025. From part to whole: AI-driven progress in fragment-based drug discovery. *Current Opinion in Structural Biology*, *91*, p.102995. https://doi.org/10.1016/j.sbi.2025.102995
6. van Tilborg, D., Brinkmann, H., Criscuolo, E., Rossen, L., Özçelik, R. and Grisoni, F., 2024. Deep learning for low-data drug discovery: Hurdles and opportunities. *Current Opinion in Structural Biology*, *86*, p.102818. https://doi.org/10.1016/j.sbi.2024.102818
7. Lin, M., Cai, J., Wei, Y., Peng, X., Luo, Q., Li, B., Chen, Y. and Wang, L., 2024. MalariaFlow: A comprehensive deep learning platform for multistage phenotypic antimalarial drug discovery. *European Journal of Medicinal Chemistry*, *277*, p.116776. https://doi.org/10.1016/j.ejmech.2024.116776
8. Wu, J., Xiao, Y., Lin, M., Cai, H., Zhao, D., Li, Y., Luo, H., Tang, C. and Wang, L., 2023. DeepCancerMap: a versatile deep learning platform for target-and cell-based anticancer drug discovery. *European Journal of Medicinal Chemistry*, *255*, p.115401. https://doi.org/10.1016/j.ejmech.2023.115401