A Machine Learning Driven Approach to Predicting the Effects of Drugs on Protein Expression

Kedar Chintalapati

With growing computational abilities, machine learning is becoming more applicable in drug discovery. However, the entire process of drug discovery can take over 12 years and cost approximately 2.6 billion dollars, rendering it intensive in both time and physical resources. One of the main steps in non-clinical development is hit identification and discovery, in which drug molecules with desired efficacies are identified. *In silico* modeling of protein expression changes from drugs could greatly speed this up, as changes in protein expression determine the effectiveness of drugs. In this study, a framework has been developed and validated to predict the impacts of drug molecules on certain proteins' expressions through a machine learning approach. Through the use of three distinct molecular featurization techniques – molecular fingerprints and numerical properties extracted using two different libraries – several machine learning models were trained to predict the impacts of drugs on protein expression of Caspase-3. The best result in this study was from a convolutional neural network (CNN) trained on molecular fingerprints data, with similar results from some other algorithms and featurization-methods. The CNN model made effective generalizations between drugs and protein expression, demonstrated by its accuracies on both balanced, (oversampled minority class) SMOTE-augmented data, and original, imbalanced test data. This framework can be applied to early stages of drug discovery to develop more machine learning models on more proteins and use them to speed up and cheapen the process through their abilities to generalize on the relations between drug molecules and changes in protein expression.

I. INTRODUCTION

Drug discovery/Hit identification

Finding new drugs is a complex problem with a low success rate [1]. The entire process can take an estimated 12 years and cost ~2.6 billion dollars [2], with only approximately 2.5% to 5% of drug candidates making it to preclinical testing, and less than 0.1% of candidates making it to the clinical phase [3]. Among the drug discovery process is hit identification and discovery, which involves finding and validating molecules that result in desired therapeutic effects [4]. In the case of the Caspase-3 protein, these desired effects can include increased sensitivity of cancer cells to chemotherapy and radiotherapy, and inhibition of cancer cell invasion and metastasis [5]. Existing approaches such as knowledge-based screening, fragment screening, physiological screening, and high throughput screening [4] can be effective but often require expensive, specialized facilities, or dependence on intuition and expert driven identification of drug candidates, which does not scale efficiently or cost-effectively when screening over 10,000 candidates. Moreover, hit identification is limited by availability or synthesis ability of potential candidates and the time and cost it takes to access and validate them in the lab. For instance, if a certain proposed drug candidate (i.e., through computational tools) is not commercially available, then the drug is often discarded unless the proposed tools/algorithms achieve high predictive accuracy in the success of the drug candidate.

In order to improve the efficiency of the hit discovery step, certain methods have been utilized in academic research and in industry. For instance, Lipinski's Rule-of-Five serves as a rule of thumb in preparing the initial list of drug candidates to be screened. Lipinski's Rule-of-Five states that an ideal drug candidate would have less than 5 H-bond donors, 10 H-bond acceptors, a molecular weight of less than 500, and a calculated LogP less than 5 [6]. When their properties are outside of these ranges, drugs are more likely to face poorer absorption and permeation. However, for more data-driven screening approaches, particularly with often "all-or-nothing" results [7], it is often challenging to identify relevant molecular descriptors. Moreover, a drug candidate could interact with many proteins [8], posing a great challenge in designing a selective drug candidate. As a result of these complexities, many drug candidates fail. It is often difficult to track and quantify the drugs' interactions with target proteins [8]. Predicting the impact of a drug candidate on protein expression is an important challenge in identifying hit molecules since they influence the therapeutic effects of the proteins, and thus an effective drug should achieve desired protein expression (i.e., increasing protein quantity or, in some cases, inhibiting proteins to deactivate the protein activity) to achieve desired treatment.

Protein Expression

In this project, the impact of drug candidates on Caspase-3 (CASP3) – a protein involved in cell apoptosis (cell death) – expression was studied [10]. Activation of CASP3 results in the cleavage of proteins, causing a process of changes in cells that eventually leads to their deaths. CASP3 is highly relevant to tumor-development, which is characterized by a lack of cell death. As a result, higher levels of activated CASP3 correlate to increased rates of recurrences and deaths in cancer patients [11]. Furthermore, CASP3 expression could predict the biochemical progression of cancer [11]. Essentially, the connections between CASP3 and cell death can influence the efficacies of certain treatments; for example, with cases such as MCF-7 in mice – a cell line observed with breast cancer – increased CASP3 presence can lead to significantly increased resistance to radiotherapy because of quicker radiation-induced apoptotic cell death and thus faster tumor cell repopulation [11].

Machine Learning Algorithms

Machine learning (ML) models can primarily be divided into supervised and unsupervised ML. In supervised machine learning, algorithms learn from labeled data, where algorithms are trained on a dataset to map inputs to outputs. By learning patterns to map inputs to outputs, the models can extrapolate to real-life applications with unseen samples. Unlike supervised learning, in unsupervised machine learning there are no labels, but rather only data with features. Through methods such as clustering, unsupervised algorithms find patterns, rather than necessarily mapping inputs and outputs, so there is no supervision correcting the models as they learn. In this project, several supervised machine learning algorithms were applied to determine the relationships between drug molecules and their impacts on protein expression.

II. METHODS AND MATERIALS

Data Processing

The Chemical-Gene Interactions dataset used in this project was collected from the Comparative Toxicogenomics Database (CTD) [13]. The dataset contains the following eleven columns: ChemicalName, ChemicalID, (MeSH identifier) CasRN (CAS Registry Number, if available), GeneSymbol, GeneID (NCBI Gene identifier), GeneForms, Organism, OrganismID (NCBI Taxonomy identifier), Interaction, InteractionActions, and PubMedIDs.

Data with the CASP3 protein was extracted due to the protein's best availability of data points after cleaning (i.e., ~6,000 of initial points, and ~400 after removing the duplicates). The CAS Registry Number of the drug molecule, the organism that the protein was from (e.g., Homo Sapiens), and the target (i.e., influence on the protein expression) were selected in the final dataset.

Data Cleaning

Cleaning of the dataset involved removal of rows with empty values for any of the features, and removal of duplicates. Since there were more variables in the original dataset than just the drugs, there were significant numbers of duplicates in the extracted data. Furthermore, there were samples of duplicates with the same features but different labels. This issue was approached through a voting system, where the label – which indicated either a decrease or an increase in expression of certain proteins – that had the highest number of instances among the same features, would be selected. For example, if the same drug candidate was reported six times in the dataset, with four instances being labeled with increased expression and two with decreased expression, then the label would be interpreted as increased expression.

Featurization

Three types of chemical features were generated for the drugs to convert the Drug IDs into a SMILES representation and then into molecular descriptors. These three types of features were molecular fingerprints, and all the numerical properties of the molecules available from the RDKit library [14] and the PubChemPy library [15].

The molecular fingerprints (Morgan fingerprints with radius=2, which is roughly equal to ECFP4) were matrices of dimensions 1x2048, within which each value was either 1 or 0 depending on the presence of certain substructures—fragments of the molecules. The dimensions for the RDKit- [14] and PubChemPy-extracted [15] descriptors were 1x43 and 1x33, respectively. Along with this, except for when training the CNN, all the extracted feature types included an extra column for the protein's original organism, numerically encoded. The labels – which were under the "InteractionActions" column – were either 0 or 1, based on whether protein expression decreased or increased, respectively.

Feature engineering

The original dataset was imbalanced in favor of increased expression. Specifically, after cleaning, ~87% of the samples had the label "1," while ~13% had the label "0." This means that a hypothetical naïve model that simply always predicts "1" would result in 87% accuracy on the data. Given that supervised machine learning algorithms learn data by minimizing the error on the predicted values versus actual values, an imbalanced dataset would make it very likely that ML models simply learn to always predict one output – in this case, "1" – like a naïve model, rather than actually learning a pattern to predict the protein expression. By predicting "1," the model would already be correct on the majority of the data thus achieving a low error, finishing its training without learning any real relationships. This problem was addressed using the SMOTE algorithm to oversample the minority class in the form of generating synthetic data points to be added to the minority class, and therefore balance the dataset. The process of implementing SMOTE involved splitting the data into train and test sets with 10% being the test data, except for certain algorithms where 20% or 30% test data size was used. Then, SMOTE was individually applied to each of the train and test sets by using the "Imbalanced-Learn" (imblearn) library [16]. This allowed for the trained models to be tested both on the original data and also on the augmented data with SMOTE.

III. RESULTS AND DISCUSSION

The results on the original and SMOTE-engineered data are provided in Table 1 and Table 2, respectively, comparing the three different featurization techniques and six different ML algorithms.

Out of all the models - based on evaluations of performance on the original, unaltered test data - the Convolutional Neural Network, Vanilla Neural Network, XGBoost, and Random Forest models achieved the best performances in terms of prediction accuracy, with SVM and Logistic Regression models also achieving notable performances when trained on molecular fingerprints data.

Table 1. Model Performances on original unaltered data. The values represent cross-validated model performances using Molecular Fingerprints, RDKit descriptors, and PubChemPy descriptors, respectively (e.g., 99.28/99.82/96.34) for drug molecule featurization. SVM (polynomial) degree of 11/200/200.

Model	**Train Accuracy [%]**	**Test Accuracy [%]**
CNN	98.03	77.50
ANN	99.44	75
XGB	82.82 / 95.57 / 95.25	65 / 79.75 / 83.52
RF	99.71 / 100 / 100	75 / 70 / 84.78
SVM (Linear)	96.62	70
SVM (RBF)	83.10 / 87.32 / 80.78	72.5 / 82.5 / 67.39
SVM (Polynomial)	55.21 / 87.04 / 40.625	35 / 82.5 / 50
LR	91.55 / 58.59 / 65.23	72.5 / 50 / 50

Table 2. Model Performances on SMOTE-engineered Data. The values represent cross-validated model performances using Molecular Fingerprints, RDKit descriptors, and PubChemPy descriptors, respectively (e.g., 99.28/99.82/96.34) for drug molecule featurization. SVM (polynomial) degree of 11/200/200.

Model	**Train Accuracy [%]**	**Test Accuracy [%]**
CNN	97.66	86.36
ANN	99.28 / 99.82 / 96.34	84.85 / 85.51
XGB	90.13 / 97.45 / 97.16	78.79 / 72.49 81.88
RF	99.84 / 100 / 100	84.84 / 60.61 63.75
SVM (Linear)	98.06	81.82
SVM (RBF)	89.00 / 51.78 / 60.65	81.82 / 50 / 41.25
SVM (Polynomial)	74.27 / 50 / 45.14	60.61 / 50 / 60
LR	95.15 / 63.11 / 71.06	83.33 / 56.06 28.75

SVM models with RBF and Polynomial kernels both achieved 82.5% accuracy on the test data when RDKit descriptors were used. However, these can likely be disregarded because these models resulted in 50% accuracies on the SMOTE-engineered data. The SMOTE-engineered data was perfectly balanced, meaning these models could have always predicted one output. The 82.5% accuracy would have been the result of 82.5% of the unaltered test data being labeled as the same one output that these SVM models always predicted. This suggests that these models did not learn the underlying patterns in the data, but rather simply performed similar to hypothetical naïve models. The 84.78% accuracy on the unaltered data using the Random Forest algorithm can also likely be disregarded for similar reasons to the seemingly high-performing SVMs.

Given the equally balanced dataset, many of the algorithms achieved higher than 80% prediction accuracy on the SMOTE-engineered test data, with the CNN trained on the Molecular Fingerprints data performing the best with an 86.36% accuracy. This result is especially promising because not only was the model not exposed to the data, but also that data was balanced. This performance shows that the models clearly identified underlying patterns in the data that included synthetic minority-class samples. SMOTE is used in real-life applications as an algorithm that allows for the creation of synthetic data based on existing samples, meaning that the generalizations the models learned to make on the augmented data can likely be applicable and relevant to real data. This also suggests that patterns in drug molecules can predict their impacts on protein expression.

Overall, the results had varied accuracies between 75% and 83.52% on the unaltered test data. Based on two key reasons – performance on unaltered test data and performance on SMOTE-engineered synthetic data – it can be determined that the algorithms implemented in this project have found significant results with real applications. For the unaltered test data, although 75% to 83.52% accuracies are technically lower than or similar to the 82.5% baseline accuracy that a hypothetical naïve model could achieve, the results are still significant given that the models clearly learned to make generalizations that worked well on the synthetic data. This means that the models were not simply naïve and only predicting one output, and thus their accuracies on the unaltered test data were authentic, making evident these models' potential for real-world applications.

IV. CONCLUSION

Overall, through evaluating several different ML models and using three different feature extraction methods, this project demonstrated the possibility of predicting drug effectiveness on protein expressions using machine learning. Specifically, it was found that ML algorithms, especially deep learning algorithms, have a real and promising applicative potential in making such predictions. Having a larger and more balanced dataset would allow for more potential use of this project's framework, and it can be replicated for different proteins where large experimental datasets are available. All three featurization techniques were also considerably high-level (these featurizations missed more specific details), and thus they may not have provided all relevant descriptors in determining protein expression. Along with using more balanced data, this project could also be expanded by using quantum mechanical methods for more descriptive molecular features, such as with Density Functional Theory (DFT), which would allow for electron density calculations. By discovering that drugs' impacts on protein expressions are predictable, this project offers a framework for potentially expanding on *in silico* drug discovery and thus helping to reduce the need of physical experimentation, which can often be far more cost-, resource-, and time-intensive. Overall, achieving this task computationally is significantly faster and cheaper than attempting it through only physical experimentation methods, and this approach offers a complementary tool to narrow down the numbers of experiments needed.

V. ACKNOWLEDGEMENTS

I would like to thank my mentor Perman Jorayev from the University of Cambridge, for all of his guidance and feedback in writing this paper.

VI. REFERENCES

[1] Altevogt, B. M., Davis, M., Pankevich, D. E., & Norris, S. M. P. (Eds.). (2014). *Improving and accelerating therapeutic development for nervous system disorders: workshop summary*. National Academies Press.

[2] Chan, H. S., Shan, H., Dahoun, T., Vogel, H., & Yuan, S. (2019). Advancing drug discovery via artificial intelligence. *Trends in pharmacological sciences*, *40*(8), 592-604.

[3] *Making a Medicine. step 1: Pre-discovery*. EUPATI Toolbox. (2021, April 16). Retrieved August 24, 2022, from https://toolbox.eupati.eu/resources/making-a-medicine-step-1-pre-discovery/.

[4] Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, *162*(6), 1239-1249.

[5] Zhou, M., Liu, X., Li, Z., Huang, Q., Li, F., & Li, C. Y. (2018). Caspase‐3 regulates the migration, invasion and metastasis of colon cancer cells. *International journal of cancer*, *143*(4), 921-930.

[6] Benet, Leslie Z., Chelsea M. Hosey, Oleg Ursu, and Tudor I. Oprea. "BDDCS, the Rule of 5 and drugability." *Advanced drug delivery reviews* 101 (2016): 89-98.

[7] Erlanson, D. A., & Jahnke, W. (Eds.). (2006). *Fragment-based approaches in drug discovery* (Vol. 3). Weinheim, Germany: Wiley-VCH.

[8] Zhou, H., Gao, M., & Skolnick, J. (2015). Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Scientific reports*, *5*(1), 1-13.

[9] Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there?. *Nature reviews Drug discovery*, *5*(12), 993-996.

[10] Porter, A. G., & Jänicke, R. U. (1999). Emerging roles of caspase-3 in apoptosis. *Cell death & differentiation*, *6*(2), 99-104.

[11] Sharma, A., Boise, L. H., & Shanmugam, M. (2019). Cancer metabolism and the evasion of apoptotic cell death. *Cancers*, *11*(8), 1144.

[12] Huang, Q., Li, F., Liu, X., Li, W., Shi, W., Liu, F. F., ... & Li, C. Y. (2011). Caspase 3–mediated stimulation of tumor cell repopulation during cancer radiotherapy. *Nature medicine*, *17*(7), 860-866.

[13] *Illuminating How Chemicals Affect Human Health. Comparative Toxicogenomics Database*. CTD. (n.d.). Retrieved July 28, 2022, from http://ctdbase.org/downloads/.

[14] *Getting Started with the RDKit in Python*. Getting Started with the RDKit in Python - The RDKit 2022.03.1 Documentation. (n.d.). Retrieved August 15, 2022, from https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors.

[15] *Properties*. PubChemPy 1.0.4 documentation. (n.d.). Retrieved August 15, 2022, from https://pubchempy.readthedocs.io/en/latest/guide/properties.html.

[16] *Imbalanced-Learn Documentation*. Imbalanced-Learn. (n.d.). Retrieved August 15, 2022, from https://imbalanced-learn.org/stable/.

[17] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[18] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Appendix

Code used for this study, hyperparameters of models, other metrics of models (e.g., precision and recall), and other information are all available upon request. Please contact the author for any questions or requests.