Bioinformatics: Drug discovery using ML models

Anupriya Palanisamy, Manisha Panda, Sanjana P., Silmi Ali Jariwalla

Professor: Sangwha Cha

Harrisburg University of Science and Technology

CISC 525

Date: 11/23/2023

Contents

Introduction 3
Drug discovery and how it works 3
Dataset 4
System Implementation 6
Experiment 8
Results 11
Jaccard Score and the Drug-Drug Similarity 11
Visualization of spread of scores 13
Unique Counts 14
Correlation Heatmap 15
Conclusion 17
References 18

Introduction :

The prevailing field of data science and its' applications has positively influenced the healthcare industry. Machine learning has helped improve medical diagnosis, treatment, and healthcare delivery by increasing the speed at which the data is processed and analyzed (Yoon & Amadiegwu, 2023). In this paper, we will gain an in-depth understanding of the field of Machine Learning in Drug Discovery. This is an important field to navigate through as it will help us understand the influence of Machine Learning in the Big data field such as Drug discovery in healthcare system.

Drug discovery and how it works :

Drug discovery involves the identification and the characterization of molecules which possess the ability to address or modulate safely, various diseases and one of its main purposes is to alleviate issues faced by patients and improve their lives by ensuring that new medicines can be discovered.

The healthcare industry generated almost 2000 exabytes of data in 2020 (Yoon & Amadiegwu, 2023) and this number will keep increasing in upcoming years. It is important for healthcare professionals to efficiently examine a patient's data and diagnose them in a timely manner. With the data boom, analyzing this data would be impossible without the help of Artificial Intelligence and applications of Machine Learning algorithms.

In this project, we will be researching how AI can be implemented in the field of drug discovery. This is an important topic because researchers and scientists need to constantly design and predict the efficiency of potential drug candidates to be manufactured and released for human use (Qureshi et al., 2023).

Drug discovery is the process of inventing new drugs from prior knowledge of drug composition and computational models (Databricks). Quantitative Structure-Activity Relationship (QSAR) is a machine learning technique used to study the relationship between chemical structures and biological activities.

In QSAR, the three-dimensional molecular structure of a chemical compound will be represented in binary notation and stored as molecular descriptors. The molecular descriptors of many chemical compounds will be stored in a large database such as the Opentargets database.

The dataset obtained from Opentargets will be subjected to pre-processing and be used to train a machine learning model that will help predict the relationship between chemical/drug structure and its biological activity. For example, when a new drug with a known chemical composition is tested against the algorithm, the model can predict if the molecule is active or inactive. Therefore, using machine learning on big data we can achieve accurate prediction and feature selection of important variables that will help with efficient medical diagnosis.

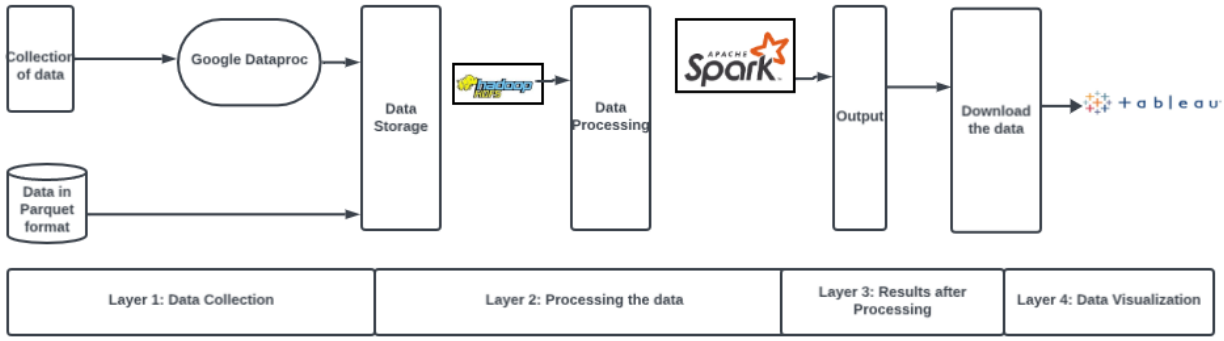
Dataset :

The datasets available on Opentargets.com are either in JSON or parquet formats. We have used the parquet format since it allows us to view nested information in a way that is machine readable (Open Targets). The parquet format that was used refers to an open source data format that is particularly useful for large volumes of data and for handling columnar data formats.

The variables in the dataset include chembl_id which represents a ChEMBL identifier that, in turn refers to the specific targets of the bioactive molecules associated with specific drug. The ‘event’ refers to any specific outcome or observations from that drug that could refer to a side effect, clinical trial outcome or any other even that is significant based on the use or the development of

the particular drug. The ‘actionType’ variable refers to the pharmacological or biological action that is produced by the relevant specific drugs on the relevant targets. The variable ‘targets’ refers to the specific molecular targets of the drugs. The ‘pathwayCategory’ variable refers to the biological pathway associated with the drug. The ‘approvedIndications’ variable refers to the indications or medical conditions associated with the drug or what the drug is being used to treat or address.

System Implementation



For this study, the authors relied on Google Cloud Platform, especially Dataproc. Dataproc is a part of the Google Cloud Platform, and it acts as a service for Apache Spark as well as Apache Hadoop clusters. The Secure Shell (SSH) within Dataproc allows us to connect to virtual machines that help to run and manage the Dataproc clusters.

For the data storage, the authors used Hadoop. Hadoop is a great tool for data storage because it processes large amounts of data and does so, quite quickly. Since it uses open-source technologies, it makes it cost-effective solution for data storage. Within Hadoop, HDFS (Hadoop Distributed File System), the files are divided into blocks and then used stored on a node (Databricks).

For the data processing, Spark was used since it supports Python programming language, which is what the authors relied on along with Dataproc and Hadoop and other tools as well. Spark processes data very quickly and handles massive amounts of data across clusters of physical or virtual servers and also has a large number of developer libraries (McDonald, 2020).

The visualization for this study was done using Tableau. Tableau is a widely used data analytics and visualization tool. It was an easy tool for the authors to use for the data visualizations

represented in this project which are discussed in further detail in the results section of this study, especially because of its user-friendly interface.

Experiment

The data was collected and was based on the Open Targets platform (Opentargets). The open targets platform provides human genetics as well as genomic data that is especially helpful for drug identification and prioritization.

For this project, the main database files used are ‘adverseDrugReactions’, ‘indications’, ‘mechanismofAction’, ‘molecule’, ‘significantAdverseDrugReactions’ and ‘targets’. Each of these files contain variables that would have to be collaborated to obtain the final output file that we aim to have. The final output file we generated had the unique identifier of each drug, also known as the CHEMBL ID. The CHEMBL ID is connected to the mechanism of action which refers to the specific biological pathway that each drug uses to perform its action. The target variable refers to the exact organ/tissue in the human body that the specific drug plays an effect on, and the Adverse Drug Reaction refers to any side effects that the drug could have such as headache or nausea. All the above variables put together contains the final raw data that we would then process to train the chosen machine algorithm model on.

The below steps outline how this project will be carried out:

1. Data Collection :

We will choose a target drug that treats a certain disease. For example, people with Alzheimer’s disease have a lower level of a protein called acetylcholinesterase than people without Alzheimer's disease. Research suggests the lower level of acetylcholinesterase protein could be a major cause of Alzheimer's (2). Therefore, we will choose a drug that increases the amount of acetylcholinesterase in the body. This drug will then be compared to other drugs with similar chemical composition and their similarity score will be computed.

We will collect data regarding this protein from the Opentarget.com database from a python library. The Opentargets.com database contains extensive genomic data that will be needed to convert the genomic information into effective drugs (ProjectPro, 2023). We obtained chemical, bioactivity, and genomic data from the Open Targets database and performed data pre-processing techniques such as efficient labelling, handling duplicates and outliers to clean the data. Additionally, converting unstructured data to readable and executable data format was also performed. All pre-processing was performed on Spark using python programming language. Genomic information on this platform will be used to identify effective new drugs (Chembl).

In this dataset, drugs are represented in SMILES strings. SMILES refers to Simplified Molecular Input Line Entry Specification which depicts the molecular structures of the drugs (Han et al., 2022).

2. Exploratory Data Analysis :

We performed EDA on the pre-processed data. The most important aspect in this process was cleaning of the Simplified Molecular Input Line Entry System (SMILES) notation. SMILES is how a three-dimensional chemical structure is represented in the form of a string of symbols. The computer is then able to understand these strings of data (ProjectPro, 2023).

Once this step was complete, we calculated the jaccard score. The Jaccard score or the Jaccard index is used in terms of measuring the similarities between the different sets of patterns, in this case the similarities between the drugs chosen from our dataset (Fletcher & Islam, 2018).

Following this, we performed statistical analysis and data visualization of the distribution and findings of the data which is explained in more detail in the results section of this report.

3. Descriptor Calculation :

Calculation of individual drugs were made using jaccard similarity score. The jaccard similarity score can be used, when dealing with two sets of data, to ascertain which elements are shared and which maintain a distinction (Karabiber, 2020).

The jaccard similarity score was especially useful to us because it has been used to measure the drug to drug similarity and distance in previous studies and is an established method of measuring and assessing drug-diagnosis association as well (Zeng et al., 2019). With respect to the drug-drug similarities, the best way to do it was to pick particular features of the drugs and then the similarities between such features would be quantified so as to assess similarity based on the whether these features are present or absent.

4. Data Modelling :

For data modelling, the data was split, trained, tested and then the jaccard similarity score was calculated.

Split Train, test, cv data into 60,20,20 ratio

```
train, validation, test = \
            np.split(known_drug_target_ae.sample(frac=1, random_state=42),
                     [int(.6*len(known_drug_target_ae)), int(.8*len(known_drug_target_ae))])  #  train =0.6, val = 1-0.8,  test = diff(0.6-0.8)
```

✓ 0.0s

5. Model deployment :

For the model deployment, we were able to successfully calculate the jaccard similarity scores for the two drugs.

Calculate Similarity -Jaccard

```
jac_sim = 1 - pairwise_distances(nd_array, metric='jaccard',n_jobs =-2)

jac_sim_df = pd.DataFrame(jac_sim, index= one_hot_encoded.index, columns=one_hot_encoded.index)
```

✓ 6.2s

Results

Jaccard Score and the Drug-Drug Similarity :

	drug_target_1	drug_target_2	jaccard_similarity
0	CHEMBL1000-ENSG00000196639	CHEMBL1000-ENSG00000196639	1.000000
1	CHEMBL1000-ENSG00000196639	CHEMBL100116-ENSG00000147955	0.000000
2	CHEMBL1000-ENSG00000196639	CHEMBL1002-ENSG00000169252	0.125000
3	CHEMBL1000-ENSG00000196639	CHEMBL1004-ENSG00000196639	0.500000
4	CHEMBL1000-ENSG00000196639	CHEMBL1008-ENSG00000006283	0.000000
...	...	...	...
23522495	CHEMBL99946-ENSG00000108576	CHEMBL9967-ENSG00000168539	0.000000
23522496	CHEMBL99946-ENSG00000108576	CHEMBL997-ENSG00000160752	0.142857
23522497	CHEMBL99946-ENSG00000108576	CHEMBL998-ENSG00000196639	0.000000
23522498	CHEMBL99946-ENSG00000108576	CHEMBL99946-ENSG00000103546	0.333333
23522499	CHEMBL99946-ENSG00000108576	CHEMBL99946-ENSG00000108576	1.000000

```
similar_drug_targets
✓ 0.0s
```

	drug_target_1	drug_target_2	jaccard_similarity
3	CHEMBL1000-ENSG00000196639	CHEMBL1004-ENSG00000196639	0.5
13275309	CHEMBL30-ENSG00000113749	CHEMBL1201356-ENSG00000184845	0.5
13275451	CHEMBL30-ENSG00000113749	CHEMBL1201747-ENSG00000196639	0.5
13275494	CHEMBL30-ENSG00000113749	CHEMBL1201759-ENSG00000196639	0.5
13275527	CHEMBL30-ENSG00000113749	CHEMBL1206-ENSG00000168539	0.5
...	...	...	...
700981	CHEMBL1088-ENSG00000102468	CHEMBL243712-ENSG00000157219	1.0
700982	CHEMBL1088-ENSG00000102468	CHEMBL243712-ENSG00000158748	1.0
16441504	CHEMBL404849-ENSG00000006283	CHEMBL1008-ENSG00000006283	1.0
9105332	CHEMBL1743082-ENSG00000167553	CHEMBL1743082-ENSG00000258947	1.0
13410072	CHEMBL304902-ENSG00000135914	CHEMBL908-ENSG00000147246	1.0

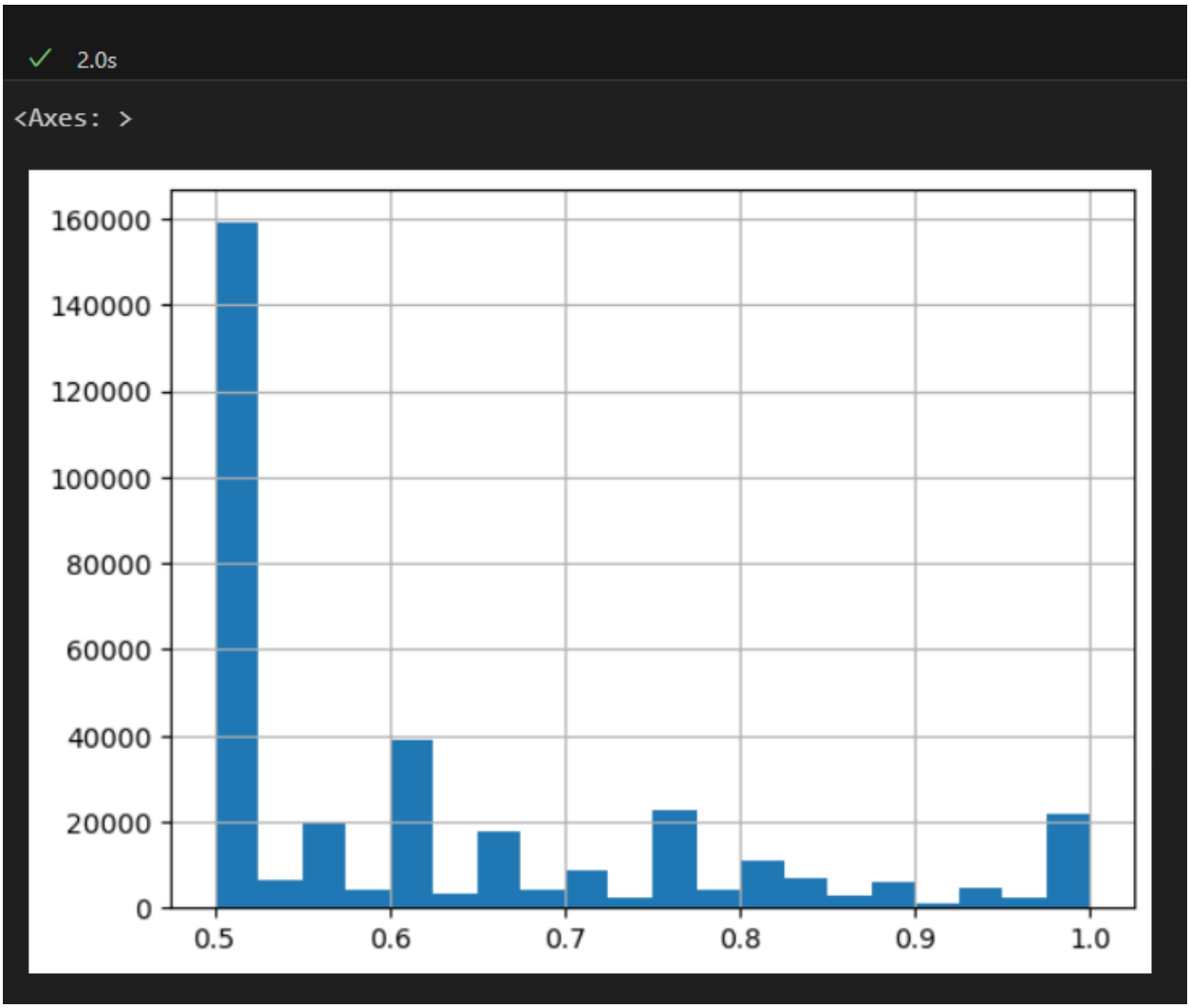
As explained in more detail in the Experiment section of this study, the jaccard similarity score can be used, when dealing with two sets of data, to ascertain which elements are shared and which

maintain a distinction (Karabiber, 2020). Specifically, with respect to this study, which analyzes the chemical structures of the drugs (SMILES notation) and based on a mathematical algorithm, the jaccard score analyzes the similarities or the lack thereof, between the drugs.

For example, based on the results in the screenshot above, CHEMBL1088-ENSG00000102468 and CHEMBL243712-ENSG00000157219 have a similarity score of 1.0 which means that, based on the chemical structure, both the drugs are 100% similar. Whereas CHEMBL30-ENSG00000113749 and CHEMBL1206-ENSG00000168539 have a jaccard similarity score of 0.5 which means that there is a 50% similarity between the two drugs based on all the variables used such as the molecular structure of the drugs.

The variables used in the jaccard score include, as mentioned earlier in this report, the 'actionType' variable refers to the pharmacological or biological action that is produced by the relevant specific drugs on the relevant targets. The variable 'targets' refers to the specific molecular targets of the drugs. The 'pathwayCategory' variable refers to the biological pathway associated with the drug. The 'approvedIndications; variable refers to the indications or medical conditions associated with the drug or what the drug is being used to treat or address.

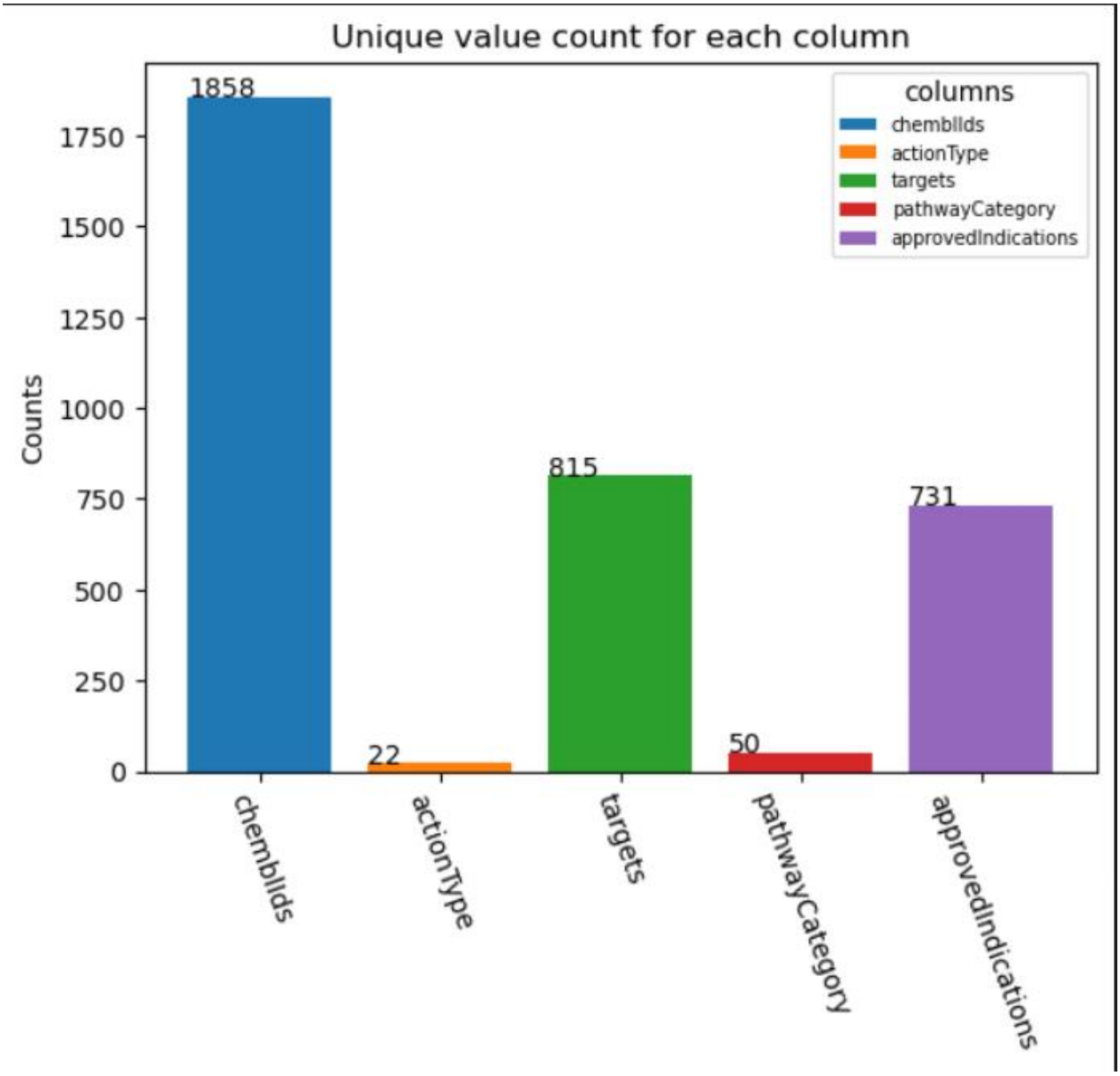
Visualization of spread of scores :



The visualization spread here represents the distribution of similarity scores for the various drug combinations. The visualization here shows the spread of data points that correspond to the jaccard scores between 0.5 and 1.0. Based on the drugs that are included in this similarity study, there are over 2 million pairs of drugs in the sample set. The visualization here shows how many pairs of drugs exhibit a particular similarity level. E.g., there are around 160,000 pairs of drugs

that have a similarity score of around 0.5. On the other hand, there are only around 1000-2000 drug combinations that have a similarity score in the 0.9-0.925 range.

Unique Counts:



The dataset contains information across various columns, each with a distinct number of unique values. The "Chemblids" column, with 1858 unique values, likely serves as a unique identifier for different chemical entities in the dataset. The "ActionType" column, boasting 22

unique values, indicates the diverse types of actions associated with these chemical entities. "Targets" with 815 unique values reflects the varied molecular targets associated with the entities, offering insights into their interactions. The "PathwayCategory" column, with 50 unique values, suggests the involvement of these entities in a diverse range of biological pathways.

Lastly, the "ApprovedIndications" column, featuring 731 unique values, provides information on the extensive range of approved medical indications or uses associated with the chemical entities. Together, these unique counts offer a comprehensive picture of the dataset's diversity and the distinctive characteristics associated with each variable, setting the stage for more in-depth analyses.

Correlation Heatmap :

```
craemrs_correlation.style.background_gradient(cmap='YlGnBu')
```

✓ 0.2s

	chembl_id	event	actionType	targets	pathwayCategory	approvedIndications
chembl_id	1.000000	0.360000	0.950000	0.600000	0.420000	0.500000
event	0.360000	1.000000	0.490000	0.250000	0.210000	0.230000
actionType	0.950000	0.490000	1.000000	0.760000	0.280000	0.600000
targets	0.600000	0.250000	0.760000	1.000000	0.530000	0.320000
pathwayCategory	0.420000	0.210000	0.280000	0.530000	1.000000	0.260000
approvedIndications	0.500000	0.230000	0.600000	0.320000	0.260000	1.000000

The correlation heatmap reveals the relationships between various variables in the dataset. Noteworthy correlations include a strong positive association (0.95) between "ActionType" and "Chembl_id," indicating a significant link between the type of action and chemical identifiers. "Targets" and "ActionType" also show a strong positive correlation (0.76), suggesting a connection between molecular targets and types of actions associated with chemical entities. Additionally, a moderate positive correlation (0.6) exists between "ApprovedIndications" and "ActionType,"

hinting at a relationship between approved medical indications and the types of actions associated with chemical entities. On the other hand, negative correlations are absent from the provided data. Overall, these correlation coefficients provide insights into potential associations among different features in the dataset, supporting further investigation into the interplay between chemical identifiers, action types, molecular targets, approved indications, and pathway categories.

Conclusion

The dataset used by us from Open Targets enables us to analyze data associated with the molecular structure of the drugs, the outcomes associated with them, what they are diagnosed for, the pathways associated with the drugs and the pharmacological action produced by the drugs. We were able to store and process our dataset, as well as compute the jaccard similarity scores using big data tools such as Hadoop, and Spark. We were able to use Datproc as well to run and manage clusters. Additionally, we learned about Tableau, a visualization tool.

The components of this dataset and the big data tools used in this study helped us calculate jaccard scores between the drugs which was our goal. Finding the similarity score between the drugs will enable future researchers to understand and analyze drug-drug interactions effectively as well as recognizing drug targets and predicting side-effects of the drugs.

References

EMBL-EBI-European Molecular Biology Laboratory.(n.d.). *ChEMBL.* ChEMBL Database (ebi.ac.uk)

Fletcher, S., & Islam, M. Z. (2018). Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems*, *22*. https://journal.acs.org.au/index.php/ajis/article/view/1538

García-Ayllón, M. S., Small, D. H., Avila, J., & Sáez-Valero, J. (2011). Revisiting the role of acetylcholinesterase in Alzheimer's disease: cross-talk with P-tau and β-amyloid. *Frontiers in molecular neuroscience*, *4*, 22. https://www.frontiersin.org/articles/10.3389/fnmol.2011.00022/full

Han, X., Xie, R., Li, X., & Li, J. (2022). Smilegnn: drug–drug interaction prediction based on the smiles and graph neural network. *Life*, *12*(2), 319. https://www.mdpi.com/2075-1729/12/2/319

Huang, S. M., Lertora, J. J., Vicini, P., & Atkinson Jr, A. J. (Eds.). (2021). *Atkinson's principles of clinical pharmacology*. Academic Press.

Karabiber, F. (December 8, 2020). *Jaccard Similarity*. LearnDataSci. https://www.learndatasci.com/glossary/jaccard-similarity/

Khamouli, S., Belaidi, S., Bakhouch, M., Chtita, S., Hashmi, M. A., & Qais, F. A. (2022). QSAR modeling, molecular docking, ADMET prediction and molecular dynamics simulations of some 6-arylquinazolin-4-amine derivatives as DYRK1A inhibitors. *Journal of Molecular Structure*, *1258*, 132659. https://www.sciencedirect.com/science/article/abs/pii/S0022286022003325

McDonald, C. (July 3, 2020). *Spark 101:What is it, What it does, and Why it Matters*. Hewlett Packard Enterprise Development LP. https://developer.hpe.com/blog/spark-101-what-is-it-what-it-does-and-why-it-matters/

Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., & Andrade, C. H. (2018). QSAR-based virtual screening: advances and applications in drug discovery. *Frontiers in pharmacology*, *9*, 1275. https://www.frontiersin.org/articles/10.3389/fphar.2018.01275/full

Open Targets. (n.d.). *Open Targets Platform*. https://www.opentargets.org/platform

Roskoski Jr, R. (2023). Rule of five violations among the FDA-approved small molecule protein kinase inhibitors. *Pharmacological Research*, 106774. https://www.sciencedirect.com/science/article/pii/S1043661823001305

United States Environmental Protection Agency.(n.d.). *Sustainable Futures / P2 Framework Manual 2012 EPA-748-B12-001 Appendix F. SMILES Notation Tutorial.* Appendix F SMILES Notation Tutorial (epa.gov)

Yoon, S. & Amadiegwu A. (2023, June 21). *Emerging tech, like AI, is poised to make healthcare more accurate, accessible and sustainable.* World Economic Forum. AI can make healthcare more accurate, accessible, and sustainable | World Economic Forum (weforum.org)

Zeng, X., Jia, Z., He, Z., Chen, W., Lu, X., Duan, H., & Li, H. (2019). Measure clinical drug–drug similarity using electronic medical records. *International journal of medical informatics*, *124*, 97-103. https://www.sciencedirect.com/science/article/pii/S1386505618305963

Zhou, S. F., & Zhong, W. Z. (2017). Drug design and discovery: principles and applications. *Molecules*, *22*(2), 279. https://www.mdpi.com/1420-3049/22/2/279/htm