MedvAIsor: AI-powered Cancer Medicine Recommendation System (A GBM Case Study)

**Team Members:** Arsalan Ahmad (aahmad@asfaschool.org), DaVonte Curtis (davontecurtis4@gmail.com), Kermit Glenn Booker Jr (kermit.booker@bulldogs.aamu.edu), Rizwan Ahmad (rizwanahmad95@outlook.com), Mary Doamekpor (mdoam5@uab.edu), Radomir Slominski (rslom@uab.edu), Dr. Ehsan Saghapour (ehsan.saghapour@gmail.com), Dr. Rahul Sharma (rsharma3@uab.edu) [Team Lead], Dr. Jake Y. Chen (jakechen@uab.edu)[Mentor]

[1. Summary 1](#_Toc110710322)

[2. Introduction 2](#_Toc110710323)

[3. Project Components 2](#_Toc110710324)

[A. Data 2](#_Toc110710325)

[I. Data Sources 2](#_Toc110710326)

[II. Data Preprocessing 3](#_Toc110710327)

[III. Preprocessing of GDSC data 3](#_Toc110710328)

[IV. NLP-based Preprocessing of drug data 4](#_Toc110710329)

[B. ML Modeling for Drug Screening 4](#_Toc110710330)

[I. Feature Engineering 5](#_Toc110710331)

[II. Model Selection 5](#_Toc110710332)

[III. Model Validation 6](#_Toc110710333)

[IV. Proof of concept 6](#_Toc110710334)

[C. MedvAIsor tool 8](#_Toc110710335)

[D. Drug Repurposing Application 8](#_Toc110710336)

[I. Proof of concept 8](#_Toc110710337)

[4. Applications 9](#_Toc110710338)

[E. Impact of work in Medicine 9](#_Toc110710339)

[I. Drug Recommendations 9](#_Toc110710340)

[F. Impact of work in Drug Discovery 10](#_Toc110710341)

[I. Fixed drug varied by patient profile (Drug Screening & repurposing) 10](#_Toc110710342)

[5. Supplementary material 10](#_Toc110710343)

# Summary

The MedvAIsor acronym is a fusion of the word Medicine + (Advisor + AI). This hackathon project built a web-based tool that showcases a GBM case study where we can perform virtual drug screening for seven FDA-approved GBM drugs. The tool takes the GBM cancer cell line of 657 genes as input and predicts the efficacy of the seven drugs. The tool screens individual cancer cell lines and shows the effectiveness of each drug along with its threshold values. In this project, we also showcase the use of NLP to process the drugs SMILES in a novel way, and this method enables us to build a competent model that we used in our MedvAIsor tool. This project also shows that the ML model we developed for MedvAIsor generalizes performing drug repurposing. We proved this by predicting the efficiently repurposing of two drugs, Temozolomide and Carmustine, with MAE of (0.5 and 0.4, respectively). These two drugs are already approved for GBM treatment by FDA.

# Introduction

Personalized drug screening is one of the vital drug development phases for precision oncology. In this phase, each patient is screened for the efficacy of a given drug on a particular type of cancer. For this phase, the in-vivo screening is challenging; therefore, in-vitro screening is widespread. Cancer samples are screened for drug perturbations on tumors, and treatment can be recommended. The duration of in-vitro screening also varies from 6 to 72 hours (but can be longer). The drug perturbation screening facility is also unavailable in most clinics; therefore, logistics delay the screening time. One can view model-based virtual screening as a solution to such delays. Thus, in-silico drug screening is also becoming essential for precision oncology.

Graphical user interface, application

Description automatically generated

Figure Hackathon Project Overview

In this hackathon, we attempt to build a personalized drug screening tool called **MedvAIsor** that uses Gene Expression data of patients’ tumor cell lines and predicts the efficacy of the drugs in LN(IC50). The demonstrator of the tool is built using the Python Streamilit App. The tool is dedicated to showcasing the study of the Glioblastoma Multiform (GBM) case study. Besides the tool, the project also put forward an innovative way of processing the drug data (SMILES) using Natural Language Procession. We also show how the ML model used by MedvAIsor for drug screening can be used to perform drug repurposing for the approved FDA drugs for GBM cancer.

# Project Components

## Data

### Data Sources

Genomics of Drug Sensitivity of Cancer (GDSC2) (Yang et al., 2012) is the primary data source from which we get 135242 perturbation information, i.e., log(IC50) values, of 198 drug sensitivities on 809 cancer cell lines. The data sets also provide The Cancer Genome Atlas (TCGA) (Gao et al., 2015) 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 (Tate et al., 2019). 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 (Kim et al., 2021). For our research work, the gene expressions from the COSMIC database and drug SMILES from the PubChem database are integrated into the GDSC dataset.

### Data Preprocessing

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.

#### Preprocessing of GDSC data

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\_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, and we get a total of 119106 data samples.

Graphical user interface, text

Description automatically generatedHaving 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. Hence, we address this method as D2GNets (Drug, Disease, Gene Predictive Networks). Figure 1 shows an abstract view of the methodology displaying the data preparation pipeline (on the left side of the figure) and ML pipeline (on the right side of the figure). Section 2.2 describes the data preparation pipeline, whereas the following sections describe the ML approach adopted to build a predictive model. Besides the method description, we also provide the hypothesis of the drug screening and drug repurposing experiments.

#### NLP-based Preprocessing of drug data

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.

**Figure 2 Bag-of-words method to extract features from drug SMILES**

1. 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.
2. 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.
3. Since gene expressing values are scaled through Z-transformation, therefore, the Term Frequency vectors are also scaled using Z-transformation as depicted by following Equation (1):

-------(1)

## ML Modeling for Drug Screening

In this section, we describe how the model is built and evaluated. We also highlight the performance of the model. This section's main components are Feature Engineering, Model Selection, Model development, and model evaluation

### Feature Engineering

Graphical user interface, calendar

Description automatically generatedAfter NLP-based feature extraction, we obtained 3041 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 chose 657 out of the 16248 genes; 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,

Figure 3 Overview of Input and Output of ML modeling

### Model Selection

The drug perturbation prediction is a regression problem. For this research, we used a low code ML library named PyCaret (Moez Ali, 2020) 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 Supplementary Table 1 provides the comparative study of Model Selection).

Table 1 Model Selection

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE | TT (Sec) |
| Light Gradient Boosting Machine | 0.0473 | 0.0040 | 0.0631 | 0.8271 | 0.0414 | 0.0993 | 15.5250 |
| Random Forest Regressor | 0.0484 | 0.0043 | 0.0653 | 0.8151 | 0.0428 | 0.1005 | 524.7860 |
| Extra Trees Regressor | 0.0482 | 0.0043 | 0.0653 | 0.8148 | 0.0430 | 0.0997 | 323.4130 |
| Ridge Regression | 0.0502 | 0.0045 | 0.0672 | 0.8043 | 0.0441 | 0.1056 | 4.7370 |
| Linear Regression | 0.0503 | 0.0045 | 0.0673 | 0.8034 | 0.0442 | 0.1058 | 3.6330 |
| Gradient Boosting Regressor | 0.0602 | 0.0061 | 0.0780 | 0.7365 | 0.0512 | 0.1300 | 713.5630 |
| Decision Tree Regressor | 0.0688 | 0.0087 | 0.0930 | 0.6247 | 0.0607 | 0.1382 | 72.6800 |
| K Neighbors Regressor | 0.0840 | 0.0133 | 0.1153 | 0.4236 | 0.0762 | 0.1934 | 211.9050 |
| AdaBoost Regressor | 0.0982 | 0.0138 | 0.1173 | 0.4035 | 0.0757 | 0.1999 | 474.5170 |
| Lasso Regression | 0.1178 | 0.0231 | 0.1519 | -0.0001 | 0.1009 | 0.2888 | 0.9800 |

### Model Validation

To develop the model, the prepared dataset was divided into two types of samples:

1. Drug Screening samples:

These are 110488 data samples consisting of Multi-cancer and all drug data. For the drug screening experiment, 80% of this dataset is used to build the model, and the remaining 20% data is used for testing the model. This set of data does not contain the drug repurposing samples.

1. Drug Repurposing samples:

The drug repurposing dataset consists of GBM cancer-type samples having two drugs, Carmustine and Temozolomide. The entire Drug Screening dataset is used to build the model for the drug repurposing experiment. This model is tested using the Drug Repurposing dataset.

ML model training was performed with the training data samples, and we performed 10-fold cross-validation to evaluate the model using six evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2-score (R2), Root Mean Square Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE). Table 2 shows the result of the model training process.

Table 2 10-fold cross-validation of ML model with Training data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fold | MAE | MSE | RMSE | R2 | RMSLE | MAPE |
| 0 | 0.90 | 1.42 | 1.19 | 0.82 | 0.32 | 2.00 |
| 1 | 0.88 | 1.38 | 1.17 | 0.83 | 0.32 | 1.75 |
| 2 | 0.88 | 1.38 | 1.17 | 0.83 | 0.32 | 1.24 |
| 3 | 0.89 | 1.40 | 1.18 | 0.83 | 0.33 | 6.73 |
| 4 | 0.89 | 1.40 | 1.19 | 0.83 | 0.32 | 1.67 |
| 5 | 0.88 | 1.37 | 1.17 | 0.83 | 0.32 | 5.18 |
| 6 | 0.88 | 1.36 | 1.17 | 0.83 | 0.33 | 1.13 |
| 7 | 0.88 | 1.37 | 1.17 | 0.83 | 0.32 | 0.92 |
| 8 | 0.87 | 1.35 | 1.16 | 0.82 | 0.31 | 0.89 |
| 9 | 0.91 | 1.49 | 1.22 | 0.82 | 0.33 | 1.54 |
| Mean | 0.89 | 1.39 | 1.18 | 0.83 | 0.32 | 2.30 |
| SD | 0.01 | 0.04 | 0.02 | 0.00 | 0.01 | 1.89 |

**Chart, box and whisker chart

Description automatically generated**

Figure 4 Range of Real and Predicted LNIC50 values

### Proof of concept

The trained model is tested on unseen testing samples, and the results are shown in Table 3. All the metrics depict that model performed satisfactorily on the test data.

Table 3 Model validation with Test data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MAE | MSE | RMSE | R2 | RMSLE | MAPE |
| Test data | 0.77 | 1.0 | 1.0 | 0.85 | 0.28 | 1.31 |

Figure 4 shows a result of real and predicted LNIC50 values of our GBM case study, and we can observe that the range of all predicted LNIC50 values falls within the range of Real LNIC50 values of GBM data. Besides, Figure 5 shows the screening of GBM samples for all 173 drugs depicting how close the results of Real and Predicted LNIC50 values are.

Figure 5 Screening Result of GBM test data

## MedvAIsor tool

The MedvAIsor tool is the main contribution of this Hackathon project, where we developed the web-based software to screen GBM cancer cell lines (657 Genes; see Table 6). The tool deploys the ML model developed for Personalized drug screening (described in the previous section). The input to the MedvAIsor App is a CSV file containing a header depicting the ‘SAMPLE\_ID’ and 657 gene names (as mentioned in Table 6). To use the tool, get this link <https://esaghapour-bokeh-test-example-w7iblm.streamlitapp.com/>, upload the sample CSV (AllGE\_GBM\_657.csv) file in the GitHub of MedvAisor/App, and watch the video tutorial at <https://github.com/u-brite/MedvAIsor/blob/main/App/Tutorial_MedvAIsor.webm>.

For Drug screening, it was essential to build a mechanism to comment on the efficacy of the seven FDA-approved drugs. Table 4 shows the Min and Max thresholds for LNIC50 vales GDSC-base thresholds were calculated for each drug using the GDSC-data, whereas we performed a survey to get the Literature-based thresholds (see Table 7 for the survey).

When the app is used to predict the efficacy of individual samples for the seven drugs, each predicted LNIC50 value also shows the upper and lower thresholds for that drug based on Table 4 (Figure 8 shows the screenshot of the individual sample screening result). Figure 9 shows the screening results of the multiple GBM cancer cell line samples.

Table Minimum and Maximum Thresholds of GBM drugs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Literature-based Threshold | | GDSC-based Threshold | |
| GBM-drugs | Min | Max | Min | Max |
| Bortezomib | 2.30 | 3.00 | -5.55 | -5.16 |
| Carmustine | 5.77 | 6.22 | 5.51 | 6.20 |
| Crizotinib | 1.55 | 1.55 | 3.37 | 3.67 |
| Gefitinib | 2.15 | 2.77 | 3.35 | 3.79 |
| Temozolomide | 2.65 | 6.48 | 5.68 | 6.54 |
| Teniposide | 1.77 | 1.77 | 0.01 | 0.81 |
| Topotecan | 1.10 | 1.10 | -0.88 | 0.25 |

## Drug Repurposing Application

### Proof of concept

We also performed a case study of GBM cancer where we used two FDA-approved drugs, Temozolomide, and Carmustine, to show how our drug screening model generalizes and predicts the LNIC50 values very similar to the Real drug screening values. Table 4 shows the validation of the two drugs for repurposing, whereas Figure 6 & 7 shows the GBM sample-wise screening of the two drugs along with their real values.

**Computational proof of concept**

Table Drug Repurposing Evaluation for Carmustine and Temozolomide

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Drug Name | MAE | MSE | RMSE | R2 | RMSLE | MAPE |
| *Carmustine* | 0.53 | 0.41 | 0.64 | 0.55 | 0.09 | 0.09 |
| *Temozolomide* | 0.45 | 0.32 | 0.57 | 0.64 | 0.08 | 0.08 |



Figure 6 Predictions of drug repurposing GBM samples with Temozolomide drug features



Figure 7 Predictions of drug repurposing GBM samples with Carmustine drug features

# Applications

## Impact of work in Medicine

### Drug Recommendations

The GBM case study of the MedvAIsor web app shows the potential of our tool where the doctor can provide the gene expression (657 genes) of a patient’s tumor to the software and get the screening values of the seven drugs. This result will help the doctor make the Monotherapy treatment plan for the patient.

## Impact of work in Drug Discovery

### Fixed drug varied by patient profile (Drug Screening & repurposing)

The ML model built in this hackathon can be you to check the varying effects of the seven drugs on multiple patients. This helps select a cohort for clinical trials for de-novo drug discovery or drug repurposing (this screening is only possible with drugs that are structurally similar to the seven drugs)

# Supplementary material

Table Cancer Gene Census based on 657 Genes used in the modeling

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| NBEA | LSM14A | HIST1H3B | JAK3 | SMC1A | FGFR1 | DNAJB1 | TOP1 | FCGR2B |
| ZCCHC8 | ANK1 | PMS2 | BRCA1 | STK11 | IRS4 | AXIN1 | ELK4 | TGFBR2 |
| NOTCH2 | CTNNB1 | ABL2 | JAZF1 | NRAS | MEN1 | BAP1 | LHFPL6 | CCNB1IP1 |
| ID3 | BCL6 | MAFB | MAP2K4 | NONO | S100A7 | GNAS | DGCR8 | CTNND1 |
| PRCC | ERBB2 | PML | MUC1 | PTPN11 | MDM2 | CACNA1D | LMO1 | SMAD3 |
| SRC | PRPF40B | TRRAP | NCOA1 | H3F3B | NUP214 | MYC | TRIM33 | KRAS |
| RHOA | IDH1 | CDC73 | BAZ1A | ABL1 | CAMTA1 | GATA3 | POT1 | PBX1 |
| SSX1 | TAL1 | CSMD3 | MAP2K2 | CCR7 | MAPK1 | H3F3A | SDC4 | TAL2 |
| CDK12 | ETNK1 | ATRX | PBRM1 | CBFB | SF3B1 | MN1 | WIF1 | KLF6 |
| TP63 | MUTYH | COL3A1 | CNBP | ELL | SOX21 | KLF4 | ERCC2 | GRM3 |
| ARHGAP26 | NFIB | CDKN2A | ACVR2A | AFF1 | CD28 | COL1A1 | TFEB | ACSL3 |
| KNSTRN | PALB2 | HMGA1 | BCL9 | CD274 | CDKN2C | PLCG1 | PPM1D | FANCG |
| PCBP1 | KTN1 | HSP90AB1 | SS18L1 | APOBEC3B | CNOT3 | CLTC | LMNA | SALL4 |
| ZNF384 | NRG1 | PRDM2 | ZNF331 | DDIT3 | MGMT | NR4A3 | LARP4B | PLAG1 |
| CTCF | BTK | ZEB1 | NSD1 | RET | PSIP1 | XPO1 | BCL11A | FLI1 |
| LASP1 | SLC34A2 | SOCS1 | PAX3 | KMT2D | USP6 | YWHAE | BRD4 | MYCN |
| DCC | FBXO11 | CHCHD7 | SS18 | CRNKL1 | TCF12 | HLA-A | 6-Sep | KAT7 |
| ATIC | IGF2BP2 | TMEM127 | POU2AF1 | RARA | RAD51B | ZNF429 | USP44 | PTPRT |
| STAT6 | MLLT11 | REL | AFF4 | EPAS1 | CASP3 | HIP1 | ETV1 | KDM5C |
| CCNC | DNM2 | TRIP11 | STRN | SYK | CANT1 | TSC1 | GNA11 | SETD1B |
| ERCC4 | MAP3K1 | BCL3 | MYD88 | CD79B | PDGFRB | GOPC | FBXW7 | SIX2 |
| SGK1 | GPC5 | FCRL4 | PDGFB | NTRK3 | HNRNPA2B1 | SRGAP3 | BAX | PPP6C |
| IL2 | TCL1A | CDK6 | TFG | BRD3 | ROS1 | MLLT10 | SH3GL1 | SPEN |
| ERCC3 | JAK2 | DDX6 | PDE4DIP | RUNX1T1 | DICER1 | MYOD1 | MED12 | NFKBIE |
| KDM6A | SMAD2 | SIX1 | GATA2 | EXT2 | TEC | ARID1B | HOXC13 | PRRX1 |
| DDB2 | FANCD2 | SDHB | HOXC11 | TSC2 | BLM | MYH9 | FOXR1 | IRF4 |
| VHL | HOXA11 | FANCC | MYO5A | EPHA7 | RAD21 | PWWP2A | FH | FOXO3 |
| CBFA2T3 | MB21D2 | HNF1A | SPOP | CHIC2 | BIRC6 | PTPRC | BIRC3 | CUX1 |
| KLK2 | CSF1R | CDX2 | SFRP4 | CD79A | FAM135B | COX6C | CDKN1B | CREB3L2 |
| IL7R | MECOM | GPC3 | MUC16 | HOXD13 | ZFHX3 | SUZ12 | CCDC6 | CHD2 |
| POU5F1 | AFF3 | MUC4 | FAT4 | NCKIPSD | NFE2L2 | MSH6 | PPFIBP1 | CASP8 |
| MTOR | SMARCE1 | CEP89 | SND1 | TLX1 | SNX29 | SIRPA | ZRSR2 | TBX3 |
| BCL11B | PIK3CB | ABI1 | UBR5 | ARHGAP35 | BMP5 | RFWD3 | NCOR2 | EGFR |
| FGFR1OP | HOXA13 | PTPRD | RB1 | AKT1 | EXT1 | DAXX | MAP2K1 | PAX5 |
| RAD17 | NTHL1 | TNFRSF14 | CRLF2 | JAK1 | WWTR1 | POLD1 | EZR | FGFR3 |
| IKZF3 | FOXA1 | PAX7 | HIST1H4I | CREB1 | AKAP9 | SMAD4 | CXCR4 | POLE |
| MACC1 | B2M | BCL10 | PAFAH1B2 | TRIM24 | PRKACA | TNFRSF17 | CTNNA2 | MAF |
| EZH2 | STIL | MDM4 | TPR | FOXO4 | TP53 | FSTL3 | ZNRF3 | FLT3 |
| ALK | SETD2 | WT1 | ERC1 | VAV1 | NF2 | XPC | FAM131B | LEF1 |
| AXIN2 | GOLGA5 | CNTNAP2 | LMO2 | SETDB1 | MNX1 | HERPUD1 | EIF3E | ERBB4 |
| CDH11 | PTCH1 | MLLT6 | ACSL6 | COL2A1 | CNBD1 | GNAQ | DDX5 | GRIN2A |
| RNF213 | DDX3X | NDRG1 | RBM15 | SMARCA4 | CD74 | KIT | PRDM16 | MYH11 |
| BCL2 | JUN | TAF15 | TNC | CRTC1 | ARHGEF12 | PRKAR1A | PAX8 | SFPQ |
| PHF6 | TCF7L2 | FEV | RALGDS | HEY1 | RNF43 | AKT3 | FAM47C | CHD4 |
| EPHA3 | ECT2L | PIK3CA | WNK2 | CBL | PTPN13 | PDCD1LG2 | FANCF | PPARG |
| ARID2 | FUBP1 | CBLB | CHST11 | LRP1B | SDHA | PTEN | FLNA | ISX |
| ARHGEF10 | TBL1XR1 | TET1 | SMARCD1 | PRDM1 | LPP | 9-Sep | EIF4A2 | KDSR |
| ETV4 | MSH2 | TRIM27 | ZBTB16 | BCR | GAS7 | FUS | ATM | ARNT |
| RGS7 | TPM4 | ASPSCR1 | BCL7A | PIK3R1 | MYB | CLTCL1 | STAG2 | FLT4 |
| ALDH2 | NAB2 | ELF4 | RPN1 | ARID1A | FHIT | PRKCB | DNMT3A | HOXA9 |
| CNTRL | SLC45A3 | SRSF2 | EED | CCND3 | CRTC3 | MITF | ASXL1 | GLI1 |
| RBM10 | RMI2 | FOXP1 | ESR1 | MAX | KAT6A | PPP2R1A | RANBP2 | CALR |
| LIFR | MAML2 | CREBBP | TFPT | CD209 | NCOA2 | LCK | LRIG3 | ZNF521 |
| PMS1 | ROBO2 | CCNE1 | IDH2 | TLX3 | EPS15 | BARD1 | WAS | RECQL4 |
| PTPRK | USP8 | CDH17 | CSF3R | RPL5 | ZNF479 | MLF1 | NIN | OLIG2 |
| VTI1A | BMPR1A | LATS2 | 5-Sep | ATP2B3 | FOXL2 | ZMYM3 | ETV6 | HSP90AA1 |
| SBDS | KIF5B | P2RY8 | FAT3 | MLLT1 | WRN | SDHC | TPM3 | IKBKB |
| SETBP1 | KIAA1549 | CLP1 | LZTR1 | PTPRB | KDM5A | BRAF | PICALM | AKT2 |
| U2AF1 | BCL2L12 | RUNX1 | NUMA1 | ELN | ARHGEF10L | OMD | FOXO1 | ASXL2 |
| POLG | ARHGAP5 | ATP1A1 | NACA | MYCL | GPHN | FNBP1 | LYN | PDGFRA |
| NFKB2 | EP300 | SMARCB1 | DCAF12L2 | MSI2 | PTK6 | CARS | LYL1 | FIP1L1 |
| TERT | SUFU | ITGAV | SH2B3 | TFRC | STAT3 | BTG2 | RAP1GDS1 | MET |
| CUL3 | NUP98 | TRAF7 | LEPROTL1 | MALT1 | CASP9 | CCND1 | EML4 | CDH10 |
| FGFR4 | PATZ1 | USP9X | MLH1 | FANCE | PRF1 | TNFAIP3 | RHOH | ERBB3 |
| ERCC5 | FBLN2 | KAT6B | GATA1 | DROSHA | DDX10 | FANCA | THRAP3 | AMER1 |
| CDK4 | FES | MAP3K13 | PHOX2B | RSPO2 | TET2 | CTNND2 | EWSR1 | ITK |
| PTPN6 | NFATC2 | CCND2 | FAT1 | CIC | CLIP1 | QKI | ACVR1 | MLLT3 |
| CREB3L1 | A1CF | IL21R | N4BP2 | CYLD | STAT5B | PREX2 | BCLAF1 | ELF3 |
| LATS1 | MPL | DDR2 | IL6ST | NF1 | RPL22 | CIITA | NTRK1 | SKI |
| TSHR | SET | BCOR | EBF1 | XPA | RABEP1 | CPEB3 | TENT5C | ARAF |
| CYSLTR2 | TFE3 | NT5C2 | KEAP1 | NPM1 | CHEK2 | BCORL1 | KCNJ5 | NCOR1 |
| IKZF1 | HMGA2 | FGFR2 | NOTCH1 | TMPRSS2 | ZMYM2 | KDR | ERG | MSN |
| EIF1AX | CYP2C8 | HOOK3 | BCL9L | ETV5 | CARD11 | DEK | HRAS | CBLC |

Table Literature survey for the Min and Max IC50 value ranges of seven FDA approved GBM drugs

|  |  |  |  |
| --- | --- | --- | --- |
| Min | Max | Drug\_name | Source: |
| 3.529297384 | 6.476972363 | Temozolamide | <https://pubmed.ncbi.nlm.nih.gov/34794398/> |
| 5.147494477 | 6.214608098 | Temozolamide | <https://pubmed.ncbi.nlm.nih.gov/30258889/> |
| 2.646174797 | 5.457881936 | Temozolamide | <https://pubmed.ncbi.nlm.nih.gov/26447477/> |
| 5.768320996 | 5.768320996 | Carmustine | <https://pubmed.ncbi.nlm.nih.gov/29868469/> |
| 5.978885765 | 6.222576268 | Carmustine | <https://pubmed.ncbi.nlm.nih.gov/29706869/> |
| 1.098612289 | 1.098612289 | Topotecan | <https://pubmed.ncbi.nlm.nih.gov/24303074/> |
| 1.768149604 | 1.768149604 | Teniposide | <https://pubmed.ncbi.nlm.nih.gov/25151861/> |
| 1.547562509 | 1.547562509 | Crizotinib | <https://www.sciencedirect.com/science/article/pii/S0960894X19305207?via%3Dihub> |
| 2.148267733 | 2.772588722 | Gefitinib | <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6493376/> |
| 2.302585093 | 2.995732274 | Bortezomib | <https://cancercommun.biomedcentral.com/articles/10.1186/s40880-019-0424-2> |
| 2.302585093 | 2.302585093 | Bortezomib | <https://www.nature.com/articles/s41416-019-0551-1> |

Table Sample-wise drugs screening result of the seven FDA approved Drugs with the Thresholds

Chart

Description automatically generated

Table Screening Result display of four samples together on the MedvAIsor web-app

Chart, scatter chart

Description automatically generated

# References

Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., Balbin, O. A., Barbe, S., Cai, H., Casey, F., Chatterjee, S., Chiang, D. Y., Chuai, S., Cogan, S. M., Collins, S. D., … Sellers, W. R. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine*, *21*(11), 1318–1325. https://doi.org/10.1038/nm.3954

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, *49*(D1), D1388–D1395. https://doi.org/10.1093/nar/gkaa971

Moez Ali. (2020). *PyCaret: An open source, low-code machine learning library in Python* (1.0). https://www.pycaret.org. https://www.pycaret.org

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., & Garnett, M. J. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, *41*(D1), D955–D961. https://doi.org/10.1093/nar/gks1111