**Supplementary material**

Figures and tables

Tables

|  |  |  |
| --- | --- | --- |
| S.No | Content | Count |
| **1** | Total Number of entries | 14963 |
| **2** | Cell lines | 51 |
| **3** | Target Pathways | 24 |
| **4** | Targets | 251 |
| **5** | Drugs | 304 |

Supplementary Table 1: Summary of Breast cancer-specific data from the GDSC (Genomics of Sensitivity in Cancer) Database

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S.No | Modalities | Test set | | | | |
| **R-Square** | **RMSE** | **MSE** | **MAE** | **Explained Variance** |
| 1 | CN+D | 0.964 | 0.413 | 0.171 | 0.270 | 0.940 |
| 2 | CN+KO+D | 0.948 | 0.418 | 0.175 | 0.281 | 0.935 |
| 3 | CN+KO+E+D | 0.922 | 0.468 | 0.219 | 0.325 | 0.923 |
| 4 | CN+FP+E+KO+D | 0.921 | 0.482 | 0.232 | 0.314 | 0.917 |
| 5 | CN+FP+E+KO+Mut+D | 0.928 | 0.677 | 0.459 | 0.394 | 0.838 |
| 6 | **CN+FP+E+KO+Mut+P+D**  **(MM-DNN)** | **0.917** | **0.3127** | **0.18301** | **0.2891** | **0.917** |

Supplementary Table 2: Validation Metrics of QSAR Models Developed Incrementally on the Test Set. Abbreviations: CN - Copy Number, D - Drug Response Data, KO - CRISPR Knock-Out Data, E - RNA Expression Data, FP - Fusion Protein Data, Mut - Mutational Data, P - Proteomic Data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.No | Modalities | Validation set | | | | | |
| **R-Square** | **RMSE** | **MSE** | **MAE** | **Explained Variance** | **Q-Square** |
| 1 | CN+D | 0.939 | 0.408 | 0.166 | 0.271 | 0.939 | 0.939 |
| 2 | CN+KO+D | 0.937 | 0.407 | 0.165 | 0.277 | 0.939 | 0.939 |
| 3 | CN+KO+E+D | 0.905 | 0.460 | 0.212 | 0.320 | 0.923 | 0.922 |
| 4 | CN+FP+E+KO+D | 0.926 | 0.460 | 0.212 | 0.304 | 0.922 | 0.922 |
| 5 | CN+FP+E+KO+Mut+D | 0.928 | 0.527 | 0.278 | 0.369 | 0.898 | 0.898 |
| 6 | **CN+FP+E+KO+Mut+P+D**  **(MM-DNN)** | **0.920** | **0.335** | **0.197** | **0.216** | **0.920** | **0.920** |

Supplementary Table 3: Validation Metrics of QSAR Models Developed Incrementally on the Test Set. Abbreviations: CN - Copy Number, D - Drug Response Data, KO - CRISPR Knock-Out Data, E - RNA Expression Data, FP - Fusion Protein Data, Mut - Mutational Data, P - Proteomic Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Fold | RMSE | MSE | MAE | Explained Variance | R-squared | Q squared |
| 1 | 0.486191 | 0.236382 | 0.31457 | 0.916778 | 0.916344 | 0.916344 |
| 2 | 0.523729 | 0.274292 | 0.331364 | 0.8972 | 0.897042 | 0.897042 |
| 3 | 0.456047 | 0.207979 | 0.302613 | 0.928788 | 0.925757 | 0.925757 |
| 4 | 0.525902 | 0.276573 | 0.33239 | 0.905006 | 0.903377 | 0.903377 |
| 5 | 0.416938 | 0.173837 | 0.27101 | 0.939609 | 0.934287 | 0.934287 |
| 6 | 0.496267 | 0.246281 | 0.295641 | 0.915213 | 0.913266 | 0.913266 |
| 7 | 0.447029 | 0.199835 | 0.31235 | 0.932581 | 0.927733 | 0.927733 |
| 8 | 0.458886 | 0.210577 | 0.290556 | 0.930056 | 0.930001 | 0.930001 |
| 9 | 0.449212 | 0.201791 | 0.280554 | 0.924295 | 0.923915 | 0.923915 |
| 10 | 0.785068 | 0.616332 | 0.630663 | 0.880927 | 0.784861 | 0.784861 |

Supplementary Table 4: 10-Fold Cross-Validation Metrics Results of MM-DNN-Based QSAR Model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Drug Name | PubChem Similarities (Count) | Tanimoto Coefficient  Outcomes (Count) | K-means Clustering outcomes (Count) |
| 1 | Alectinib | 640 | 553 | 291 |
| 2 | Amuvatin | 134 | 85 | 42 |
| 3 | AST-1306 | 719 | 602 | 350 |
| 4 | Axitinib | 570 | 570 | 380 |
| 5 | AZD1332 | 63 | 63 | 47 |
| 6 | AZD4547 | 198 | 131 | 89 |
| 7 | AZD6094 | 1000 | 708 | 560 |
| 8 | AZD8931 | 1000 | 540 | 330 |
| 9 | BIBF-1120 | 1000 | 818 | 652 |
| 10 | BMS-758 | 724 | 384 | 271 |
| 11 | Cabozantinib | 1000 | 427 | 225 |
| 12 | CI-1033 | 1000 | 780 | 445 |
| 13 | crizotinib | 1000 | 648 | 403 |
| 14 | Dasatinib | 1000 | 932 | 240 |
| 15 | Foretinib | 1000 | 617 | 264 |
| 16 | GSK1904529A | 310 | 295 | 172 |
| 17 | GW-2580 | 731 | 306 | 169 |
| 18 | GW44175 | 986 | 641 | 401 |
| 19 | JNJ38877605 | 1000 | 807 | 257 |
| 20 | Kobe2602 | 36 | 24 | 14 |
| 21 | Lapatinib | 1000 | 962 | 269 |
| 22 | Linifanib | 251 | 168 | 126 |
| 23 | MASITINIB | 403 | 241 | 151 |
| 24 | Motesanib | 693 | 475 | 175 |
| 25 | NVP-TAE684 | 451 | 410 | 219 |
| 26 | OSI-930 | 49 | 44 | 19 |
| 27 | Pazopanib | 503 | 256 | 179 |
| 28 | PD173074 | 175 | 126 | 106 |
| 29 | PF-002998 | 1000 | 499 | 354 |
| 30 | PHA-665752 | 494 | 446 | 412 |
| 31 | Quizartinib | 386 | 376 | 271 |
| 32 | SB505124 | 255 | 107 | 68 |
| 33 | SB52334 | 1000 | 73 | 62 |
| 34 | Sorafenib | 1000 | 908 | 616 |
| 35 | SU11274 | 173 | 131 | 98 |
| 36 | Sunitinib | 1000 | 703 | 520 |
| 37 | tivozanib | 117 | 82 | 76 |

Supplementary Table 5: Tabular Column Representing the Summary of the Structural Similarities Filtering Process

|  |  |  |
| --- | --- | --- |
| Drugs | Targets | Clusters |
| AZD4547 | FGFR1, FGFR2, FGFR3, FGRF1, FGFR2, FGFR3 | **Angiogenesis and Proliferation Hub** |
| Cabozantinib | VEGFR, MET, RET, KIT, FLT1, FLT3, FLT4, TIE2,AXL |
| Sunitinib | PDGFR, KIT, VEGFR, FLT3, RET, CSF1R |
| Alectinib | ALK |
| GSK1904529A | IGF1R, IR |
| PF-00299804 | EGFR, ERBB2, ERBB4 |
| Foretinib | MET, KDR, TIE2, VEGFR3/FLT4, RON, PDGFR, FGFR1, EGFR |
| PD173074 | FGFR1, FGFR2, FGFR3 |
| PHA-665752 | MET |
| Quizartinib | FLT3 |
| BIBF-1120 | VEGFR, PDGFR, FGFR |
| SU11274 | MET |
| GW441756 | NTRK1 | **Growth Factor and Receptor Modulation Hub** |
| JNJ38877605 | MET |
| GW-2580 | CSF1R |
| Kobe2602 | RAS effector |
| Motesanib | VEGFR, RET, KIT, PDGFR |
| BMS-754807 | IGF1R, IR |
| AST-1306 | EGFR, ERBB4 |
| Crizotinib | MET, ALK, ROS1 |
| OSI-930 | KIT, VEGFR2 |
| AZD1332 | NTRK1, NTRK2, NTRK3 |
| Pazopanib | CSF1R, KIT, PDGFRA, PDGFRB |
| Sorafenib | PDGFR, KIT, VEGFR, RAF |
| AZD6094 | MET |
| Linifanib | VEGFR1, VEGFR2, VEGFR3, CSF1R, FLT3, KIT |
| SB505124 | TGFBR1, ACVR1B, ACVR1C |
| Axitinib | PDGFR, KIT, VEGFR |
| Tivozanib | VEGFR1, VEGFR2, VEGFR3, PDGFR, KIT |
| SB52334 | ALK5 |
| Masitinib | KIT, PDGFRA, PDGFRB | **Tumor Proliferation and Adaptive Responses** |
| AZD8931 | EGFR, ERBB2, ERBB3 |
| Lapatinib | EGFR, ERBB2 |
| CI-1033 | EGFR, ERBB2, ERBB4 |
| Dasatinib | ABL, SRC, Ephrins, PDGFR, KIT |
| Amuvatinib | KIT, PDGFRA, FLT3 |
| NVP-TAE684 | ALK |

Supplementary Table 6: Tabular Column Representing Clusters Obtained, Along with Corresponding Drugs and Their Targets

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.No | Dataset Name | Repository Name | No. Of Samples | No. Of features | | Data type | Contents |
| 1 | GISTIC (Genomic Identification of Significant Targets in Cancer) copy number data | Cell Modal Passports | 81 | 20670 | Numerical data | | Copy number alterations of 81 breast cancer cell lines |
| 2 | CRISPR KNOCK OUT data - fitness scores | Cell Modal Passports | 81 | 17490 | Numerical data | | Fitness scores of the genes of 81 breast cancer cell lines |
| 3 | CRISPR KNOCK OUT data – Fold change values | Cell Modal Passports | 81 | 17490 | Numerical data | | Gene expression fold change values of 81 breast cancer cell lines |
| 4 | CRISPR Gene Dependency data and gene effect data | Cell Modal Passports | 81 | 16384 | Numerical data | | 81 breast cancer cell lines gene Dependency and effect profiles for survival and proliferation |
| 5 | FPKM (Fragments Per Kilobase of transcript per Million mapped reads) RNA Expression data | Cell Modal Passports | 81 | 37603 | Numerical data | | FPKM quantifies gene expression in RNA-Seq by considering gene length and mapped reads, facilitating comparison between samples |
| 6 | Fusion protein data | Cell Modal Passports | 81 | 982 | Categorical data | | Fusion protein data provides information on the kind of alterations because of protein fusion |
| 7 | Mutation Data | Cell Modal Passports | 81 | 18990 | Categorical data | | The dataset provides information on mutations, including their association with cancer driver genes, predisposition to cancer, functional effects, variant allele frequencies |
| 8 | Proteomic data | Cell Modal Passports | 81 | 8459 | Numerical data | | The dataset provides quantitative information on protein expression levels and modifications across breast cancer cell lines. |
| 9 | Drug response data | GDSC – Genomics of drug sensitivity in cancer | 14103 | 8 | Numerical and categorical data | | The dataset provides information on the drugs half maximal inhibitory concentration profiles on various breast cancer cell lines |
| 10 | Molecular descriptor data | PADELpy Library in Python v3.12.0 | 305 | 1870 | Numerical data | | Molecular descriptors quantitatively represent various properties of the drug molecule in a 2-dimensional and 3-dimensional manner |

Supplementary Table 7: The table presents a comprehensive summary of the datasets sourced from various repositories for training the model. Each entry includes details such as dataset contents, data types, and a detailed description elucidating their characteristics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | MODEL ID | CELL LINE NAME | S.No | MODEL ID3 | CELL LINE NAME4 |
| 1 | SIDM01037 | JIMT-1 | **42** | SIDM01056 | EFM-19 |
| 2 | SIDM00866 | HCC1143 | **43** | SIDM01042 | EVSA-T |
| 3 | SIDM00314 | ZR-75-1 | **44** | SIDM01062 | HDQ-P1 |
| 4 | SIDM01592 | Hs-274-T | **45** | SIDM01001 | DU-4475 |
| 5 | SIDM00135 | Hs-578-T | **46** | SIDM01187 | UACC-812 |
| 6 | SIDM00147 | KPL-1 | **47** | SIDM01186 | UACC-893 |
| 7 | SIDM00097 | T47D | **48** | SIDM00005 | MDA-MB-134-VI |
| 8 | SIDM00122 | BT-549 | **49** | SIDM01422 | SUM-52PE |
| 9 | SIDM00146 | MDA-MB-231 | **50** | SIDM01430 | SUM-185PE |
| 10 | SIDM00272 | MDA-MB-453 | **51** | SIDM01452 | SUM-159PT |
| 11 | SIDM00332 | MFM-223 | **52** | SIDM01441 | SUM-149PT |
| 12 | SIDM00529 | MDA-MB-157 | **53** | SIDM01461 | SUM-229PE |
| 13 | SIDM00528 | MDA-MB-361 | **54** | SIDM00241 | OCUB-M |
| 14 | SIDM00562 | MRK-nu-1 | **55** | SIDM01594 | Hs-281-T |
| 15 | SIDM00897 | SK-BR-3 | **56** | SIDM01590 | Hs-343-T |
| 16 | SIDM00881 | HCC1428 | **57** | SIDM01591 | Hs-606-T |
| 17 | SIDM00675 | HCC38 | **58** | SIDM01622 | HMC-1-8 |
| 18 | SIDM00673 | HCC70 | **59** | SIDM01702 | Hs-739-T |
| 19 | SIDM00631 | MDA-MB-330 | **60** | SIDM01708 | HMEL |
| 20 | SIDM00628 | MDA-MB-468 | **61** | SIDM01701 | Hs-742-T |
| 21 | SIDM01053 | EFM-192C | **62** | SIDM01826 | VP229 |
| 22 | SIDM01054 | EFM-192B | **63** | SIDM01836 | SUM-1315MO2 |
| 23 | SIDM00882 | HCC1419 | **64** | SIDM00629 | MDA-MB-436 |
| 24 | SIDM00630 | MDA-MB-415 | **65** | SIDM00884 | HCC1395 |
| 25 | SIDM00893 | BT-20 | **66** | SIDM01825 | VP267 |
| 26 | SIDM00874 | HCC1937 | **67** | SIDM01835 | SUM-44PE |
| 27 | SIDM00872 | HCC1954 | **68** | SIDM01837 | SUM-102PT |
| 28 | SIDM00870 | HCC202 | **69** | SIDM00879 | HCC1500 |
| 29 | SIDM00774 | HCC2157 | **70** | SIDM00878 | HCC1569 |
| 30 | SIDM00772 | HCC2218 | **71** | SIDM01900 | 21PT |
| 31 | SIDM00877 | HCC1599 | **72** | SIDM01902 | 21MT-1 |
| 32 | SIDM00898 | AU565 | **73** | SIDM01912 | 21NT |
| 33 | SIDM00963 | BT-474 | **74** | SIDM00933 | CAL-51 |
| 34 | SIDM00892 | BT-483 | **75** | SIDM01901 | 21MT-2 |
| 35 | SIDM00928 | CAL-85-1 | **76** | SIDM01809 | H16N2 |
| 36 | SIDM00920 | CAMA-1 | **77** | SIDM00971 | ZR-75-30 |
| 37 | SIDM00885 | HCC1187 | **78** | SIDM01261 | YMB-1-E |
| 38 | SIDM00875 | HCC1806 | **79** | SIDM00148 | MCF7 |
| 39 | SIDM01260 | YMB-1 | **80** | SIDM00633 | MDA-MB-175-VII |
| 40 | SIDM00940 | CAL-120 | **81** | SIDM00938 | CAL-148 |
| 41 | SIDM00954 | COLO-824 | **82** | SIDM01002 | EFM-192A |

Supplementary Table 8: The table displays the breast cancer cell lines and Unique ID’s.

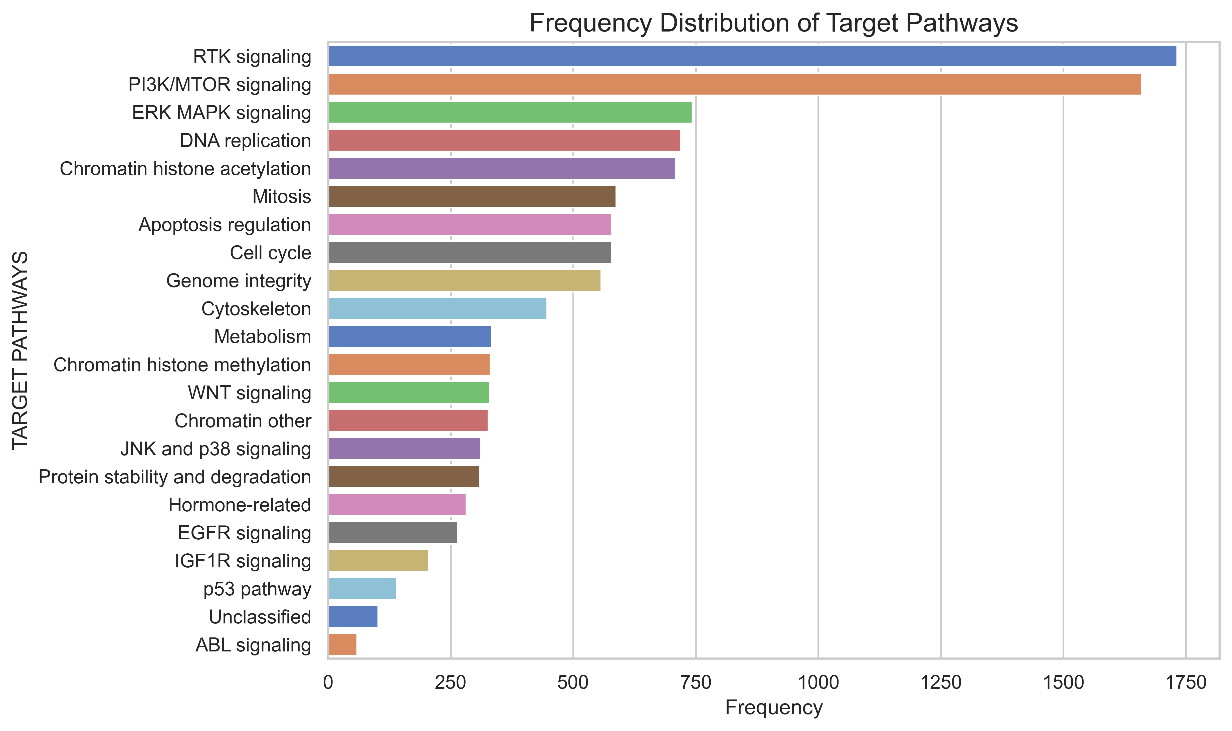
|  |  |  |
| --- | --- | --- |
| Database Name | Contents | Description |
| DrugMAP | Interacting molecules for >30,000 drugs/drug candidates. Differential expression patterns for >5,000 interacting molecules among different disease sites. ADME (absorption, distribution, metabolism, and excretion)-relevant organs and physiological tissues. A network containing >200,000 interactions among drugs and molecules | DrugMAP provides a comprehensive molecular atlas and pharma information on drugs, including a vast list of interacting molecules, expression patterns, and network of interactions, aiding drug discovery efforts. |
| Therapeutic Target Database (TTD) | Nine categories of established druggability characteristics for 426 successful, 1014 clinical trials, 212 preclinical/patented, and 1479 literature-reported targets | TTD collects druggability characteristics for various drug targets, aiding in the identification and validation of potential drug targets. |
| ADCdb | Information on antibody-drug conjugates (ADCs) including pharma-information and biological activities for over 6,500 ADCs | ADCdb focuses on providing critical data on ADCs, facilitating research in biopharmaceutical drug discovery. |
| Interactome of Drug-Metabolizing Enzymes (INTEDE) | Microbiome-DME, xenobiotics-DME, and host protein-DME interactions for 1047 unique DMEs | INTEDE offers comprehensive interaction data for drug-metabolizing enzymes, enabling insights into drug metabolism and precision medicine. |
| Drug Resistance Database (DRESIS) | Comprehensive information on drug resistance mechanisms across various diseases | DRESIS systematically provides data on molecular mechanisms underlying drug resistance, aiding in new drug discovery and treatment optimization. |
| TheMarker | Five types of therapeutic biomarkers (Pharmacodynamic, Safety, Monitoring, Predictive, Surrogate Endpoint) for a wide range of drugs | TheMarker database offers comprehensive information on therapeutic biomarkers, crucial for drug development and clinical practice. |

Supplementary Table 9: Comprehensive Overview of Drug-related Databases and Repositories Enriching Drug Discovery Efforts with Insights on Drug Interactions, Resistance Mechanisms, Druggability Characteristics, and Therapeutic Biomarkers

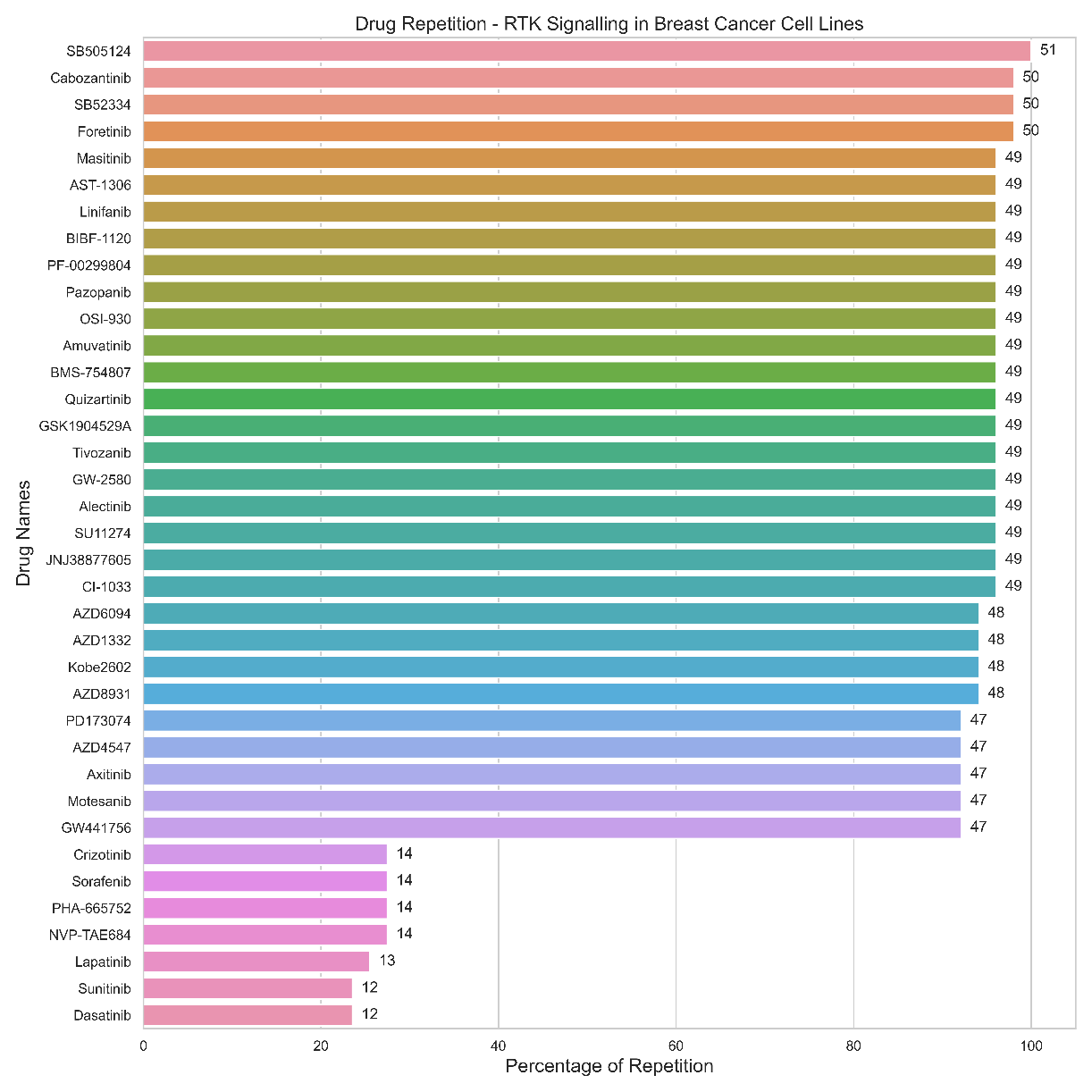
|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Reference compound | Structural similarity | Targets |
| 1 | AZD4547 - N-[5-[2-(3,5-dimethoxyphenyl)ethyl]-1H-pyrazol-3-yl]-4-[(3R,5S)-3,5-dimethylpiperazin-1-yl]benzamide | 1-chloro-N,N,2-trimethylprop-1-en-1-amine;N-[5-[2-(3,5-dimethoxyphenyl)ethyl]-1H-pyrazol-3-yl]-3-methoxybenzamide | FGFR1, FGFR2, FGFR3, FGRF1, FGFR2, FGFR3 |
| 2 | Cabozantinib - 1-N-[4-(6,7-dimethoxyquinolin-4-yl)oxyphenyl]-1-N'-(4-fluorophenyl)cyclopropane-1,1-dicarboxamide | 4-(6,7-dimethoxyquinolin-4-yl)oxy-N-[3-(4-fluorophenyl)propyl]aniline | VEGFR, MET, RET, KIT, FLT1, FLT3, FLT4, TIE2,AXL |
| 3 | Sunitinib - N-[2-(diethylamino)ethyl]-5-[(Z)-(5-fluoro-2-oxo-1H-indol-3-ylidene)methyl]-2,4-dimethyl-1H-pyrrole-3-carboxamide | N-[2-(diethylamino)ethyl]-5-[1-(5-fluoro-2-oxo-1,3-dihydroindol-3-yl)propyl]-2,4-dimethyl-1H-pyrrole-3-carboxamide | PDGFR, KIT, VEGFR, FLT3, RET, CSF1R |
| 4 | Alectinib - 9-ethyl-6,6-dimethyl-8-(4-morpholin-4-ylpiperidin-1-yl)-11-oxo-5H-benzo[b]carbazole-3-carbonitrile | 3-amino-6,6-dimethyl-9-[4-(oxetan-3-yl)piperazin-1-yl]-5H-benzo[b]carbazol-11-one | ALK |
| 5 | GSK1904529A - N-(2,6-difluorophenyl)-5-[3-[2-[5-ethyl-2-methoxy-4-[4-(4-methylsulfonylpiperazin-1-yl)piperidin-1-yl]anilino]pyrimidin-4-yl]imidazo[1,2-a]pyridin-2-yl]-2-methoxybenzamide | N-(2,6-difluorophenyl)-3-{3-[2-({5-methyl-2-(methyloxy)-4-[4-(methylsulfonyl)-1-piperazinyl]phenyl}amino)-4 pyrimidinyl]imidazo[1,2-a]pyridin-2-yl}benzamide | IGF1R, IR |
| 6 | PF-00299804 - (E)-N-[4-(3-chloro-4-fluoroanilino)-7-methoxyquinazolin-6-yl]-4-piperidin-1-ylbut-2-enamide | N-[4-(3-chloro-4-fluoroanilino)-7-(difluoromethoxy)quinazolin-6-yl]acetamide | EGFR, ERBB2, ERBB4 |
| 7 | Foretinib - 1-N'-[3-fluoro-4-[6-methoxy-7-(3-morpholin-4-ylpropoxy)quinolin-4-yl]oxyphenyl]-1-N-(4-fluorophenyl)cyclopropane-1,1-dicarboxamide | N-[4-[7-(3-bromopropoxy)-6-methoxyquinolin-4-yl]oxy-3-fluorophenyl]-1-formylcyclopropane-1-carboxamide | MET, KDR, TIE2, VEGFR3/FLT4, RON, PDGFR, FGFR1, EGFR |
| 8 | PD173074 - 1-tert-butyl-3-[2-[4-(diethylamino)butylamino]-6-(3,5-dimethoxyphenyl)pyrido[2,3-d]pyrimidin-7-yl]urea | tert-butyl N-[N'-[2-[amino-[(Z)-2-amino-3-[[7-(tert-butylcarbamoylamino)-6-(3,5-dimethoxyphenyl)pyrido[2,3-d]pyrimidin-2-yl]amino]prop-1-enyl]amino]ethyl]-N-[(2-methylpropan-2-yl)oxycarbonyl]carbamimidoyl]carbamate | FGFR1, FGFR2, FGFR3 |
| 9 | PHA-665752 - (3Z)-5-[(2,6-dichlorophenyl)methylsulfonyl]-3-[[3,5-dimethyl-4-[(2R)-2-(pyrrolidin-1-ylmethyl)pyrrolidine-1-carbonyl]-1H-pyrrol-2-yl]methylidene]-1H-indol-2-one | 5-[5-(2,6-Dichloro-phenylmethanesulfonyl)-2-oxo-1,2-dihydro-indol-(3Z)-ylidenemethyl]-2,4-dimethyl-1H-pyrrole-3-carboxylic acid amide | MET |
| 10 | Quizartinib - 1-(5-tert-butyl-1,2-oxazol-3-yl)-3-[4-[6-(2-morpholin-4-ylethoxy)imidazo[2,1-b][1,3]benzothiazol-2-yl]phenyl]urea | 1-[4-(6-Methoxyimidazo[2,1-b][1,3]benzothiazol-2-yl)phenyl]-3-(3-methyl-1,2-oxazol-5-yl)urea | FLT3 |
| 11 | BIBF-1120 - methyl 2-hydroxy-3-[N-[4-[methyl-[2-(4-methylpiperazin-1-yl)acetyl]amino]phenyl]-C-phenylcarbonimidoyl]-1H-indole-6-carboxylate | methyl 3-[N-[4-[(3-aminopropylamino) methyl]phenyl]-C-phenylcarbonimidoyl]-2-hydroxy-1H-indole-6-carboxylate | VEGFR, PDGFR, FGFR |
| 12 | SU11274 - (3Z)-N-(3-chlorophenyl)-3-[[3,5-dimethyl-4-(4-methylpiperazine-1-carbonyl)-1H-pyrrol-2-yl]methylidene]-N-methyl-2-oxo-1H-indole-5-sulfonamide | tert-butyl 4-[5-[(Z)-[5-[(3-chlorophenyl)-methylsulfamoyl]-2-oxo-1H-indol-3-ylidene]methyl]-2,4-dimethyl-1H-pyrrole-3-carbonyl]piperazine-1-carboxylate | MET |
| 13 | GW441756 - (3Z)-3-[(1-methylindol-3-yl)methylidene]-1H-pyrrolo[3,2-b]pyridin-2-one | (3Z)-3-ethylidene-1H-pyrrolo[3,2-b]pyridin-2-one;1-methylindole | NTRK1 |
| 14 | JNJ38877605 - 6-[difluoro-[6-(1-methylpyrazol-4-yl)-[1,2,4]triazolo[4,3-b]pyridazin-3-yl]methyl]quinoline | 6-[Difluoro-(6-iodo-[1,2,4]triazolo[4,3-b]pyridazin-3-yl)methyl]quinoline | MET |
| 15 | GW-2580 - 5-[[3-methoxy-4-[(4-methoxyphenyl)methoxy]phenyl]methyl]pyrimidine-2,4-diamine | 2,4-Diamino-5-[3,4-dimethoxy-5-(2-propynyloxy)benzyl]pyrimidine | CSF1R |
| 16 | Kobe2602 - 1-[2,6-dinitro-4-(trifluoromethyl)anilino]-3-(4-fluorophenyl)thiourea | 1-(4-Ethylphenyl)-3-[2-nitro-4-(trifluoromethyl)anilino]thiourea | RAS effector |
| 17 | Motesanib - N-(3,3-dimethyl-1,2-dihydroindol-6-yl)-2-(pyridin-4-ylmethylamino)pyridine-3-carboxamide | N-(4-pentylphenyl)-2-(pyridin-4-ylmethylamino)pyridine-3-carboxamide | VEGFR, RET, KIT, PDGFR |
| 18 | BMS-754807 - (2S)-1-[4-[(5-cyclopropyl-1H-pyrazol-3-yl)amino]pyrrolo[2,1-f][1,2,4]triazin-2-yl]-N-(6-fluoropyridin-3-yl)-2-methylpyrrolidine-2-carboxamide | N-[1-[4-[(5-cyclopropyl-1H-pyrazol-3-yl)amino]pyrrolo[2,1-f][1,2,4]triazin-2-yl]piperidin-4-yl]acetamide | IGF1R, IR |
| 19 | AST-1306 - N-[4-[3-chloro-4-[(3-fluorophenyl)methoxy]anilino]quinazolin-6-yl]prop-2-enamide | [1-[[4-[3-Chloro-4-[(3-fluorophenyl)methoxy]anilino]quinazolin-6-yl]amino]-2-(2-fluorobutoxycarbonylamino)-1-oxopropan-2-yl] acetate | EGFR, ERBB4 |
| 20 | Crizotinib - 3-[(1R)-1-(2,6-dichloro-3-fluorophenyl)ethoxy]-5-(1-piperidin-4-ylpyrazol-4-yl)pyridin-2-amine | 3-Phenylmethoxy-5-(1-piperidin-4-ylpyrazol-4-yl)pyridin-2-amine | MET, ALK, ROS1 |
| 21 | OSI-930 - sulfur monoxide | 3-(pyridin-4-ylmethylamino)-N-[4-(trifluoromethoxy)phenyl]thiophene-2-carboxamide | KIT, VEGFR2 |
| 22 | AZD1332 -  5-chloro-2-N-[1-(5-fluoropyridin-2-yl)ethyl]-4-N-(3-propan-2-yloxy-1H-pyrazol-5-yl)pyrimidine-2,4-diamine | 5-Chloro-N2-[(1S)-1-(5-fluoropyridin-2-yl)ethyl]-N4-(5-isopropoxy-1H-pyrazol-3-yl)pyrimidine-2,4-diamine phosphate | NTRK1, NTRK2, NTRK3 |
| 23 | Pazopanib - 5-[[4-[(2,3-dimethylindazol-6-yl)-methylamino]pyrimidin-2-yl]amino]-2-methylbenzenesulfonamide | 3-[4-[[4-[(2,3-dimethylindazol-6-yl)-methylamino]pyrimidin-2-yl]amino]phenyl]-N-methylpropane-1-sulfinamide | CSF1R, KIT, PDGFRA, PDGFRB |
| 24 | Sorafenib - 4-[4-[[4-chloro-3-(trifluoromethyl)phenyl]carbamoylamino]phenoxy]-N-methylpyridine-2-carboxamide | 4-[4-[[[[4-Chloro-3-(trifluoromethyl)phenyl]amino]carbonyl]amino]-3-fluorophenoxy]-N-methyl-2-pyridinecarboxamide hydrate | PDGFR, KIT, VEGFR, RAF |
| 25 | AZD6094 - 3-[(1S)-1-imidazo[1,2-a]pyridin-6-ylethyl]-5-(1-methylpyrazol-4-yl)triazolo[4,5-b]pyrazine | 3-(Imidazo[1,2-b]pyridazin-6-ylmethyl)-5-(1-methylpyrazol-4-yl)triazolo[4,5-b]pyrazine | MET |
| 26 | Linifanib - 1-[4-(3-amino-1H-indazol-4-yl)phenyl]-3-(2-fluoro-5-methylphenyl)urea | 1-[4-(3-amino-1H-indazol-4-yl)phenyl]-3-(2-fluoro-5-methylphenyl)-1-(2-hydroxyethyl)urea | VEGFR1, VEGFR2, VEGFR3, CSF1R, FLT3, KIT |
| 27 | SB505124 - 2-[4-(1,3-benzodioxol-5-yl)-2-tert-butyl-1H-imidazol-5-yl]-6-methylpyridine | 4-(1,3-benzodioxol-5-yl)-5-(6-methylpyridin-2-yl)-1H-imidazol-2-amine | TGFBR1, ACVR1B, ACVR1C |
| 28 | Axitinib - N-methyl-2-[[3-[(E)-2-pyridin-2-ylethenyl]-1H-indazol-6-yl]sulfanyl]benzamide | N-[5-(hydroxyamino)-5-oxopentyl]-2-[[3-(2-pyridin-2-ylethenyl)-1H-indazol-6-yl]sulfanyl]benzamide | PDGFR, KIT, VEGFR |
| 29 | Tivozanib - 1-[2-chloro-4-(6,7-dimethoxyquinolin-4-yl)oxyphenyl]-3-(5-methyl-1,2-oxazol-3-yl)urea | 1-[4-(6,7-Dimethoxyquinolin-4-yl)oxyphenyl]-3-(5-methyl-1,2-oxazol-3-yl)urea | VEGFR1, VEGFR2, VEGFR3, PDGFR, KIT |
| 30 | SB52334 - 2-phenylpyridine-3-carboxylic acid | 2-methylpyridine-3-carboxylic acid | ALK5 |
| 31 | Masitinib - 4-[(4-methylpiperazin-1-yl)methyl]-N-[4-methyl-3-[(4-pyridin-3-yl-1,3-thiazol-2-yl)amino]phenyl]benzamide | 4-(2-aminophenyl)-1-N-[4-methyl-3-[(4-pyridin-3-yl-1,3-thiazol-2-yl)amino]phenyl]cyclohexa-1,5-diene-1,4-dicarboxamide | KIT, PDGFRA, PDGFRB |
| 32 | AZD8931 - 2-[4-[4-(3-chloro-2-fluoroanilino)-7-methoxyquinazolin-6-yl]oxypiperidin-1-yl]-N-methylacetamide | 6-[4-(3-chloroanilino)-7-methoxyquinazolin-6-yl]oxy-N-hydroxyhexanamide | EGFR, ERBB2, ERBB3 |
| 33 | Lapatinib - N-[3-chloro-4-[(3-fluorophenyl)methoxy]phenyl]-6-[5-[(2-methylsulfonylethylamino)methyl]furan-2-yl]quinazolin-4-amine | Methyl 6-[[5-[4-[3-chloro-4-[(3-fluorophenyl)methoxy]anilino]quinazolin-6-yl]furan-2-yl]methylamino]hexanoate | EGFR, ERBB2 |
| 34 | CI-1033 - N-[4-(3-chloro-4-fluoroanilino)-7-(3-morpholin-4-ylpropoxy)quinazolin-6-yl]prop-2-enamide | (Z)-N-[4-(3-chloro-4-fluoroanilino)-7-(difluoromethoxy)quinazolin-6-yl]-4-(dimethylamino)but-2-enamide | EGFR, ERBB2, ERBB4 |
| 35 | Dasatinib - N-(2-chloro-6-methylphenyl)-2-[[6-[4-(2-hydroxyethyl)piperazin-1-yl]-2-methylpyrimidin-4-yl]amino]-1,3-thiazole-5-carboxamide | N-(2-chloro-6-methylphenyl)-2-[[6-[4-(2-hydroxyethyl)piperazin-1-yl]-2-methylpyrimidin-4-yl]amino]-1,3-thiazole-5-carboxamide;sulfuric acid | ABL, SRC, Ephrins, PDGFR, KIT |
| 36 | Amuvatinib - N-(1,3-benzodioxol-5-ylmethyl)-4-([1]benzofuro[3,2-d]pyrimidin-4-yl)piperazine-1-carbothioamide | 3-(1,3-Benzodioxol-5-yl)-1-[4-([1]benzofuro[3,2-d]pyrimidin-4-yl)piperazin-1-yl]propane-1-thione;methanesulfonic acid | KIT, PDGFRA, FLT3 |
| 37 | NVP-TAE684 - 5-chloro-2-N-[2-methoxy-4-[4-(4-methylpiperazin-1-yl)piperidin-1-yl]phenyl]-4-N-(2-propan-2-ylsulfonylphenyl)pyrimidine-2,4-diamine | 5-chloro-2-N-[2-methoxy-4-(4-piperazin-4-ium-1-ylpiperidin-1-yl)phenyl]-4-N-(2-propan-2-ylsulfonylphenyl)pyrimidine-2,4-diamine;2,2,2-trifluoroacetate | ALK |

Supplementary Table 10: Tabular column representing the structural comparison of the reference drug and its corresponding structural similarity

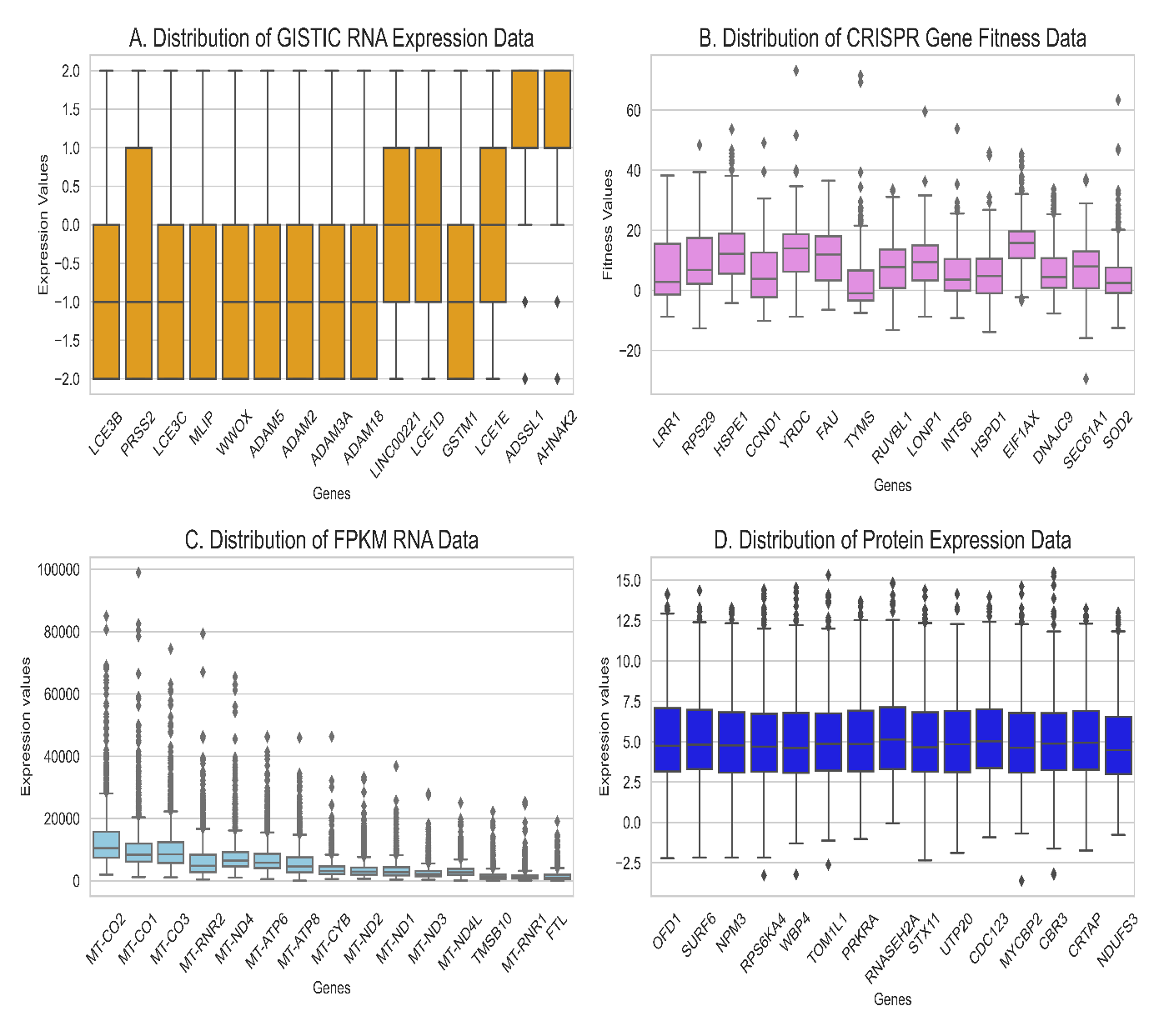
Figures



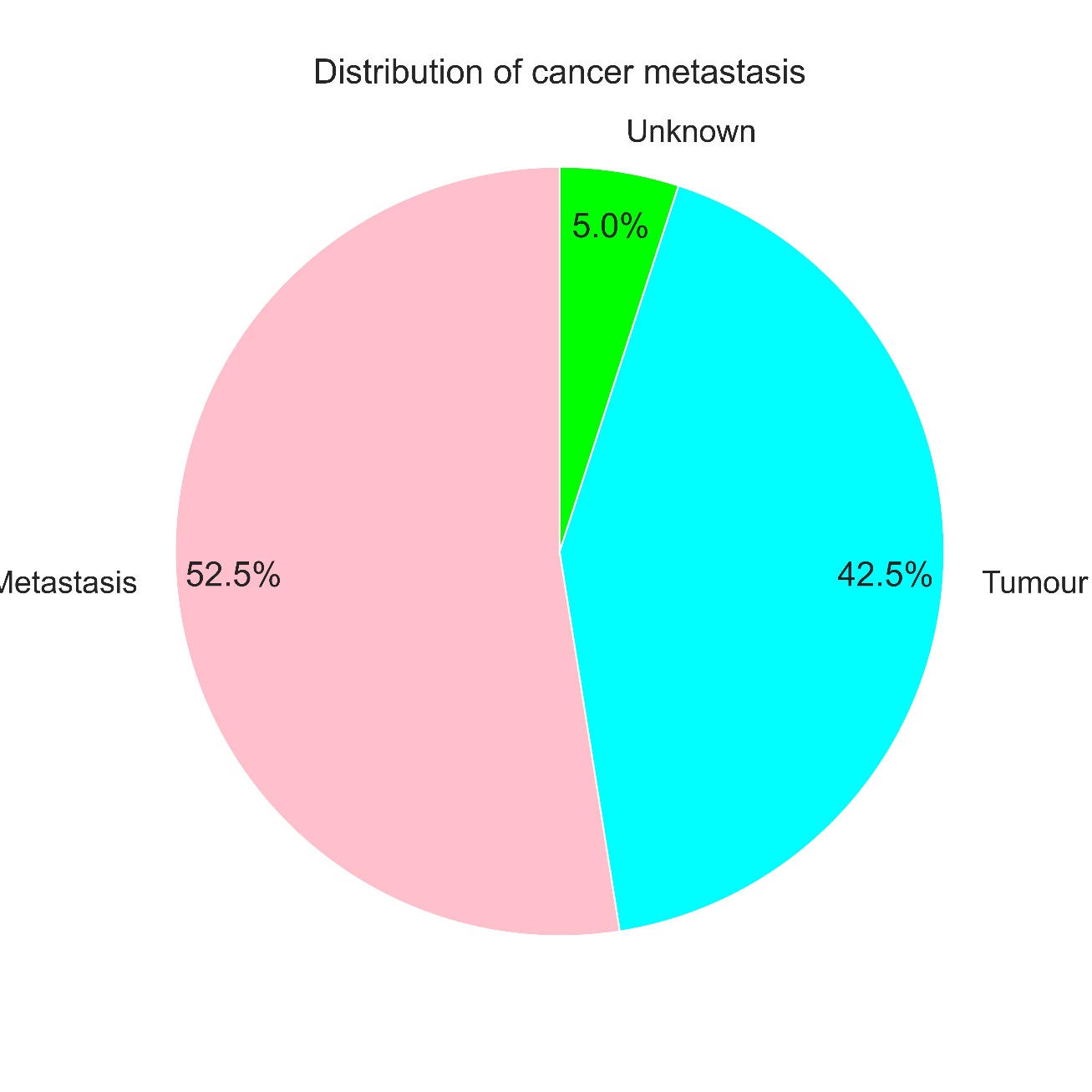
Supplementary Figure 1: Bar Plot Representing Target Pathways and Frequency Distribution (Data Sourced from GDSC Database)

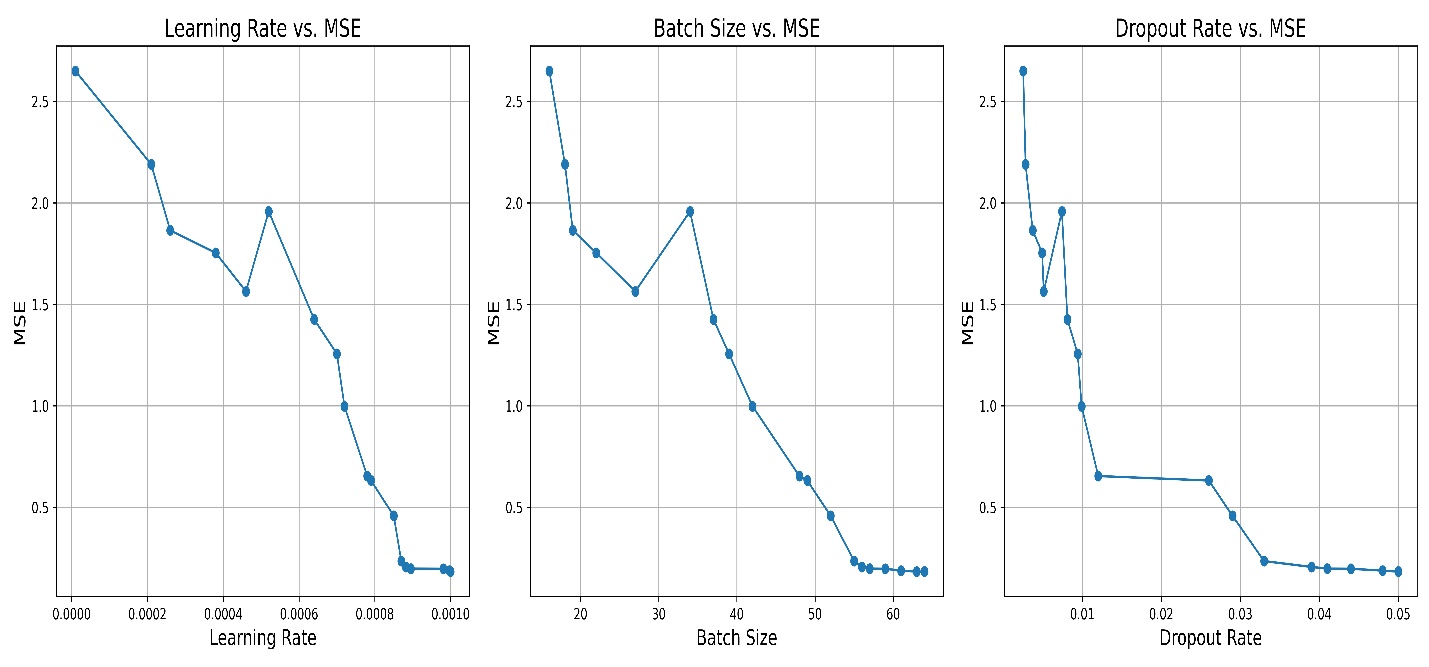


Supplementary Figure 2: Bar Plot Representing Drugs Specific to RTK Signaling in Breast Cancer Cell Lines. The Numerical Value Atop Each Bar Signifies the Impact of the Respective Drug on the Number of Breast Cancer Cell Lines.

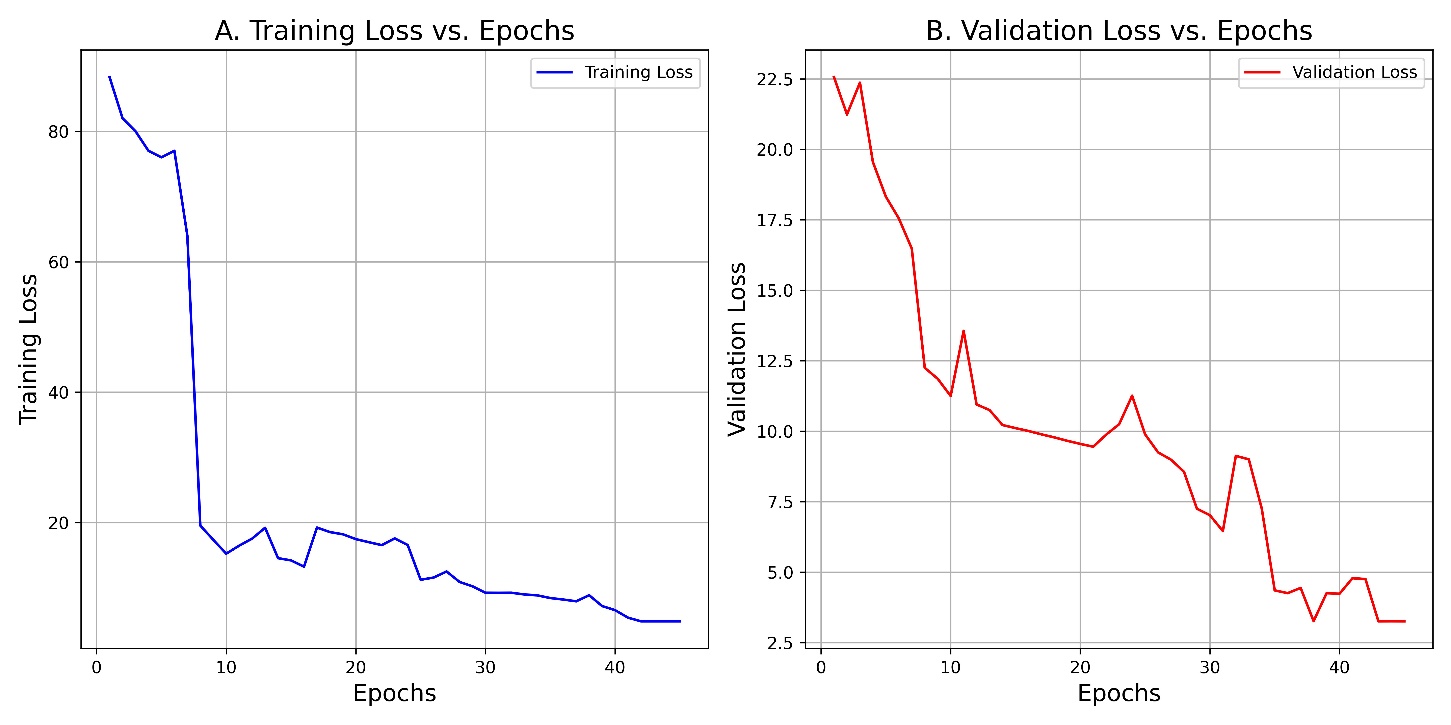


Supplementary Figure 3: Boxplot Visualization of Dataset Distribution: The image showcases the distribution of datasets through boxplots, with a focus on the top 15 genes exhibiting high variance. A depicts GISTIC RNA expression data, B represents CRISPR Gene Fitness Data, C illustrates FPKM RNA data, and D showcases Protein expression data. This visualization aids in presenting the variability within each dataset effectively.

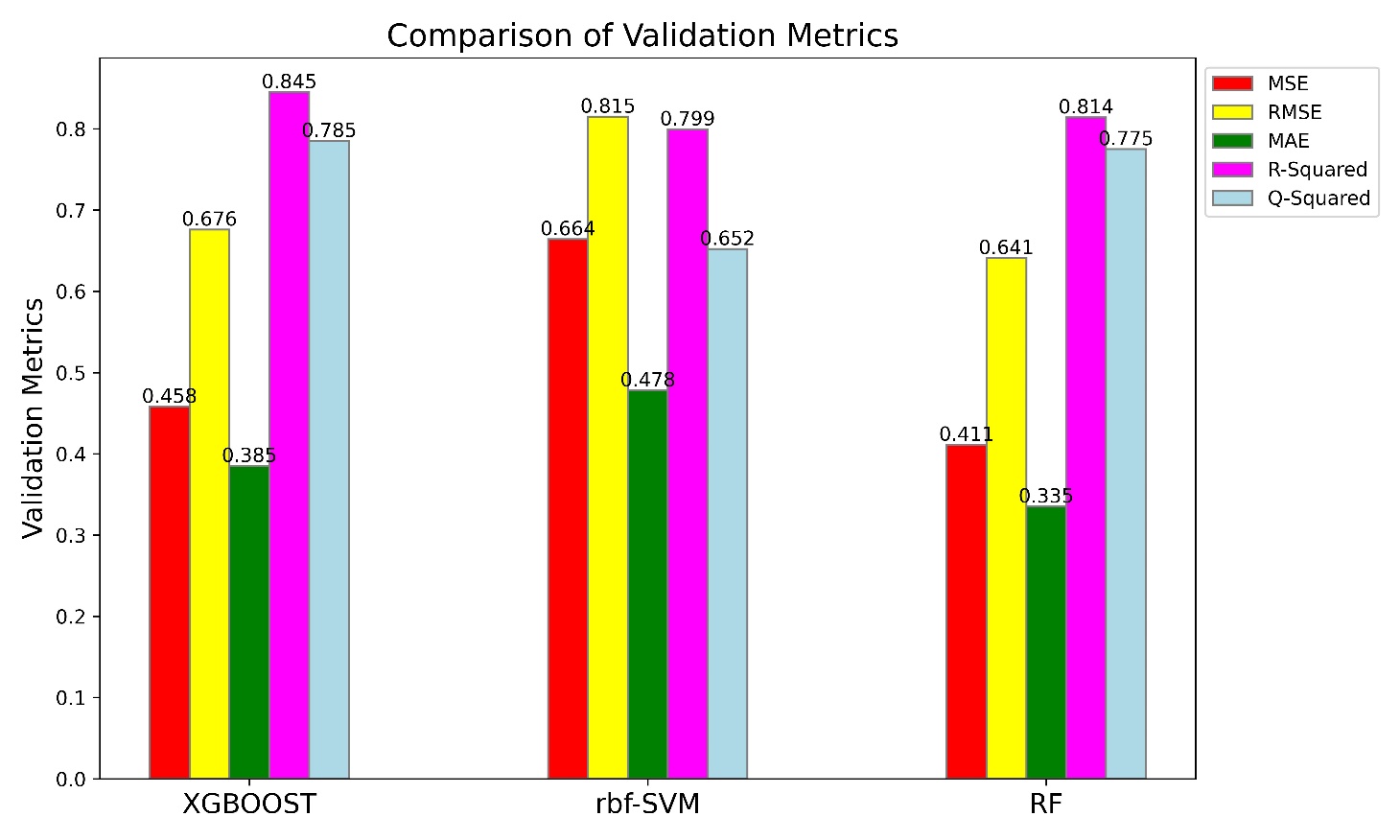
Supplementary Figure 4: The provided pie chart illustrates the distribution of metastatic behaviour among breast cancer cell lines. This visual representation offers insight into the relative proportions of different metastatic phenotypes observed within the dataset. Metastatic tumors are those that have spread from their original site to other parts of the body, while normal tumors remain localized at the site of origin.



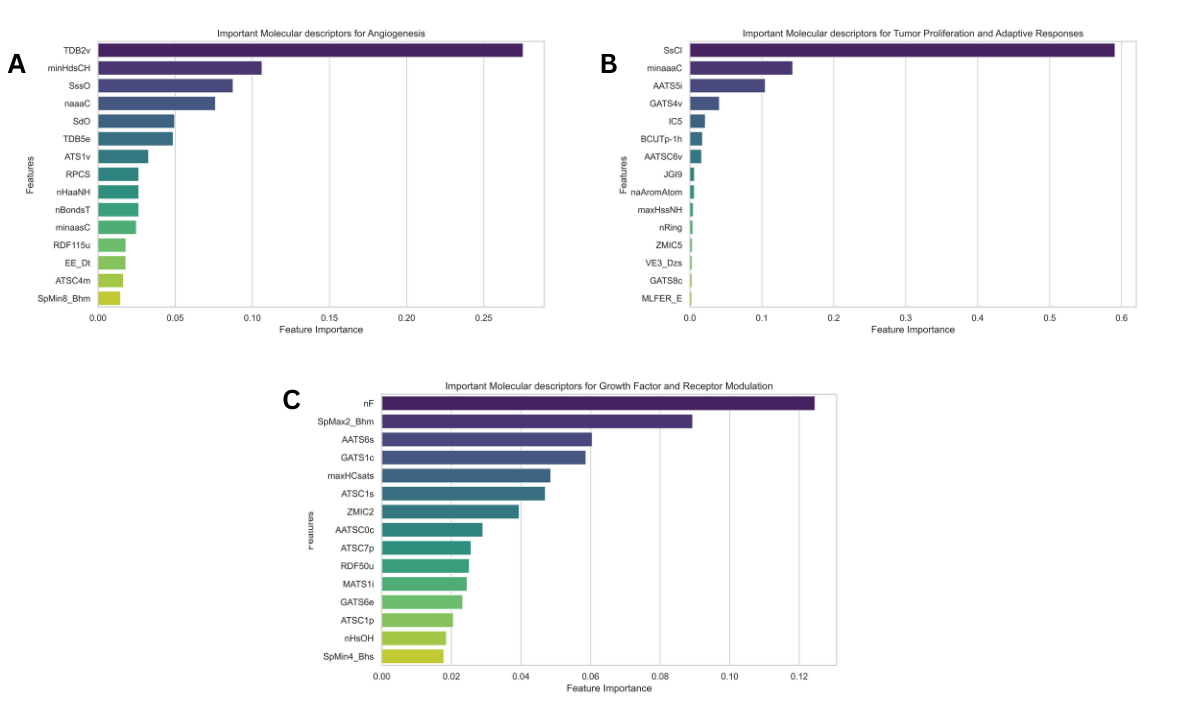
Supplementary Figure 5: Hyperparameter Optimization Curves for MM-DNN Using Hyperopt. This graph illustrates the hyperparameter optimization process conducted with Hyperopt for the MM-DNN model. Mean Square Error (MSE) was employed as the loss function during optimization. The curves represent the iterative refinement of parameters such as learning rate, dropout rate, and batch size, crucial for enhancing model performance and convergence. Each curve reflects the dynamic exploration of parameter space, aiming to minimize MSE and achieve optimal model efficiency.



Supplementary Figure 6: Training and Validation Loss Trends of MM-DNN Model with SGD Optimizer. This graph illustrates the training and validation loss patterns of the MM-DNN model trained using the Stochastic Gradient Descent (SGD) optimizer. While SGD exhibited a decrease in loss over iterations, the trend was not as consistent or optimal compared to the performance achieved with the Adam optimizer. Consequently, Adam was selected due to its superior performance and ability to achieve lower final loss values.



Supplementary Figure 7: Comparison of Validation Metrics Across Various Machine Learning Algorithms. This graph presents a comparative analysis of validation metrics for several machine learning algorithms trained on a complex multimodal dataset. The algorithms evaluated include Extra Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machines with radial basis function kernel (SVM-rbf). Despite rigorous training efforts, these algorithms failed to attain optimal or acceptable validation metrics due to the high complexity inherent in the dataset.

****

Supplementary Figure 8: Bar Plots Representing Important Molecular Descriptors. (A) Cell Proliferation and Angiogenesis, (B) Tumor Proliferation and Adaptive Responses, C: Growth Factor and Receptor Modulation

The 37 reference molecules were subjected to clustering into three groups using the K-means method. These clusters were named Angiogenesis and Proliferation Hub, Growth Factor and Receptor Modulation Hub, and Tumor Proliferation and Adaptive Responses Hub based on the downstream regulatory proteins they target. Additionally, structural similarities corresponding to the reference molecules were also clustered accordingly. Please refer to Supplementary Table 6.

Resultant clusters exhibit distinctions based on targeted proteins, with shared growth factor receptors responsible for cell proliferation. The Angiogenesis and Proliferation Hub cluster converges on critical tyrosine receptor kinases, including KIT, PDGFRA, PDGFRB, EGFR, ERBB2, ERBB3, FGFR1, FGFR2, FGFR3, MET, signaling pathways implicated in breast cancer. Whereas, The Growth Factor and Receptor Modulation Hub cluster emphasize targets related to receptor modulation, regulatory elements influencing cellular growth, and growth factor signaling pathways. Notably, NTRK1, MET, CSF1R, RAS effector, and receptor tyrosine kinases like VEGFR, RET, KIT, and PDGFR play pivotal roles in breast cancer progression, reflecting a broad spectrum of regulatory functions. The Tumor Proliferation and Adaptive Responses Hub centres on targets associated with tumor proliferation and adaptive responses. Including ABL, SRC, Ephrins, ALK5, and kinases like KIT and PDGFRA suggests involvement in adaptive responses to micro-environmental changes, cell survival, and proliferation.

In exploring angiogenesis and cell proliferation, the 15 molecular descriptors span a diverse range of properties crucial for comprehending molecular structure and behaviour. Encompassing 3D autocorrelation descriptors (TDB2V, TDB5E), Atom Type Electropological State descriptors (NaaaC, minHdSC, Sss0), 2D autocorrelation descriptor (ATS1V), charged surface area descriptor (RPCS), radial distribution function descriptor (RDF115U), hydrogen atom count descriptor (nHaaNH), eigenvalue descriptor (SpMin8\_BHM), and detour matrix descriptor (EE\_Dt), these descriptors play a pivotal role in drug discovery. They provide insights into spatial arrangement, electrophilicity, charge distribution, topological relationships, and electronic structure, facilitating the prediction and optimization of molecular properties crucial for effective drug-protein interactions. Please refer to Figure S3A.

In the context of growth factor and receptor modulation, the 15 molecular descriptors offer a comprehensive toolkit, ranging from atom counts to autocorrelation and burden-modified eigenvalues. Noteworthy features include atom counts focusing on fluorine atoms, shedding light on bioisosteres and metabolic stability. Autocorrelation descriptors such as ATSC1s and MATS1I reveal spatial arrangements and electronegativity patterns crucial for predicting molecular interactions. Burden-modified eigenvalues, exemplified by spmin4\_bhs, capture structural complexity and stability, influencing interactions with biological targets. Descriptors like GATS6E and RDF RDF50U provide valuable information on electronegativity and atomic mass distributions, enhancing bioactivity prediction and understanding three-dimensional molecular arrangements. Collectively, these descriptors contribute to advanced cheminformatics, enabling precise predictions and informed decision-making in drug discovery processes. Please refer to Figure S3B.

The top 15 key molecular descriptors of Tumor proliferation and Adaptive responses cover essential properties for understanding the molecular structure, reactivity, and interactions. Atom Type Electro-topological State descriptors (SsCl, minaaaC, maxHssNH) reveal electrophilic nature, charge distribution, and topological features crucial for predicting molecular reactivity and interactions. Autocorrelation descriptors (AATS5i, GATS4v, GATS8c) provide spatial distribution insights, aiding in understanding molecular shapes and potential binding interactions. Information Content descriptors (IC5, ZMIC5) offer insights into molecular complexity, contributing to a diverse understanding of drug design. BCUTp-1h provides unique information about Burden eigenvalues, offering a perspective on electronic structure. Descriptors like JGI9 and naAromAtom contribute to understanding topological charge and aromaticity, respectively. nRing assesses structural complexity, while VE3\_Dzs provides insights into three-dimensional features. MFER\_E, representing Molecular Linear Free Energy Relation, aids in understanding energetic aspects of molecular interactions. In drug discovery, these descriptors empower researchers to predict and optimize molecular properties, streamlining drug development. Please refer to Supplementary Figure 3C.

In the exploration of angiogenesis and cell proliferation, a diverse set of 15 molecular descriptors spanning 3D autocorrelation, Atom Type Electropological State, 2D autocorrelation, charged surface area, radial distribution function, hydrogen atom count, eigenvalue, and detour matrix descriptors were considered. These descriptors play a pivotal role in drug discovery, providing insights into spatial arrangement, electrophilicity, charge distribution, topological relationships, and electronic structure. They facilitate the prediction and optimization of molecular properties crucial for effective drug-protein interactions. Moving to the context of growth factor and receptor modulation, a separate set of 15 molecular descriptors offered a comprehensive toolkit, ranging from atom counts to autocorrelation and burden-modified eigenvalues. Noteworthy features included atom counts focusing on fluorine atoms, shedding light on bioisosteres and metabolic stability, while autocorrelation descriptors revealed spatial arrangements and electronegativity patterns crucial for predicting molecular interactions. In the sphere of tumor proliferation and adaptive responses, another distinct set of the top 15 key molecular descriptors covered essential properties for understanding the molecular structure, reactivity, and interactions. These included Atom Type Electro-topological State descriptors revealing electrophilic nature, charge distribution, and topological features. Autocorrelation descriptors provided spatial distribution insights, while Information Content descriptors offered insights into molecular complexity. Together, these descriptors contribute to advanced cheminformatics, enabling precise predictions and informed decision-making in drug discovery processes. Each group of descriptors presents a unique perspective, collectively empowering researchers in predicting and optimizing molecular properties for various facets of drug development. Please refer to Figures Supplementary Figure 3A, 3B, and 3C for a visual representation of these insights

Abbreviations

1. KIT - Kit Proto-Oncogene

2. PDGFRA - Platelet-Derived Growth Factor Receptor Alpha

3. PDGFRB - Platelet-Derived Growth Factor Receptor Beta

4. EGFR - Epidermal Growth Factor Receptor

5. ERBB2 - Erb-B2 Receptor Tyrosine Kinase 2 (HER2)

6. ERBB3 - Erb-B2 Receptor Tyrosine Kinase 3

7. FGFR1 - Fibroblast Growth Factor Receptor 1

8. FGFR2 - Fibroblast Growth Factor Receptor 2

9. FGFR3 - Fibroblast Growth Factor Receptor 3

10. MET - MET Proto-Oncogene, Receptor Tyrosine Kinase

11. NTRK1 - Neurotrophic Receptor Tyrosine Kinase 1 (TrkA)

12. MET - MET Proto-Oncogene, Receptor Tyrosine Kinase

13. CSF1R - Colony Stimulating Factor 1 Receptor

14. RAS - Rat Sarcoma Virus (RAS) Oncogene

15. VEGFR - Vascular Endothelial Growth Factor Receptor

16. RET - Rearranged During Transfection (RET) Proto-Oncogene

17. PDGFR - Platelet-Derived Growth Factor Receptor

18. ALK5 - Activin A Receptor Type I (ACVR1)

19. TDB2v - TPSA (Topological Polar Surface Area)

20. minHdsCH - Minimum Hybridization Distance to Unsaturation

21. SssO - Sum of SsO (Sum of Oxygen-centered Segments)

22. naaac - Number of Aromatic Atoms in Aliphatic Chains

23. SdO - Sum of double bond counts

24. TDB5e - TPSA (Topological Polar Surface Area) contribution from TDB5

25. ATS1v - Autocorrelation of lag 1 / weighted by van der Waals volumes

26. RPCS - Ring Penalty in ChemScore

27. nHaaNH - Number of H-bond acceptors in NH groups

28. nBondsT - Total Number of Bonds

29. minaasC - Minimum Atomic Partial Charge in Aliphatic Structures

30. RDF115u - Radial Distribution Function at a bond length of 1.15 Å

31. EE\_Dt - E-state Indices for Dubois total topological

32. ATSC4m - Average topological charge separation index of order 4 weighted by mass

33. SpMin8\_Bhm - Shortest Path Atom-Centric topological distance-based index of order 8 weighted by atomic masses

34. SsCl - Sum of SsCl (Sum of Chlorine-centered Segments)

35. minaaaC - Minimum atomic area of C atoms

36. AATS5i - Average absolute topological charge index of order 5

37. GATS4v - Geary autocorrelation of lag 4 / weighted by van der Waals volumes

38. IC5 - Information content of order 5

39. BCUTp-1h - Burden cluster-based topological index of order -1 / weighted by Sanderson electronegativities

40. AATSC6v - Average absolute atomic charge of order 6 / weighted by van der Waals volumes

41. JGI9 - Jurs graph information index of order 9

42. naAromAtom - Number of aromatic atoms

43. maxHssNH - Maximum absolute net charge on H atoms

44. nRing - Number of rings

45. ZMIC5 - Zagreb index of order 5 / weighted by atomic masses

46. VE3\_Dzs - Eigenvalue-based descriptor from 3D matrix weighted by Sanderson electronegativities

47. GATS8c - Geary autocorrelation of lag 8 / weighted by atomic masses

48. MLFER\_E - Modified linear free energy relationship descriptor (E-state)

49. nF - Number of Fluorine atoms

50. SpMax2\_Bhm - Second leading eigenvalue of Burden matrix / weighted by mass

51. AATS6s - Average absolute topological charge index of order 6 / weighted by mass

52. GATS1c - Geary autocorrelation of lag 1 / weighted by mass

53. maxHCsats - Maximum number of heavy atom-connected atoms in a saturated system

54. ATSC1s - Average topological charge separation index of order 1 / weighted by van der Waals surfaces

55. ZMIC2 - Zagreb index of order 2 / weighted by atomic masses

56. AATSC0c - Average absolute atomic charge of order 0 / weighted by atomic masses

57. ATSC7p - Average topological charge separation index of order 7 / weighted by polarizabilities

58. RDF50u - Radial Distribution Function at a bond length of 5.0 Å

59. MATS1i - Moran autocorrelation of lag 1 / weighted by ionization potentials

60. GATS6e - Geary autocorrelation of lag 6 / weighted by ionization potentials

61. ATSC1P - Average topological charge separation index of order 1 / weighted by polarizabilities

62. nHsOH - Number of hydroxyl groups

63. SpMin\_Bhs - Minimum topological atom-bond connectivity distance-based index / weighted by Sanderson electronegativities