# ReadMe

The set of m-files and data files in this repository enables any user to run the two-step modelling routine in Matlab. More specifically, the folder *Code* contains the files needed to run the code, whereas the folder *Data* holds the preprocessed data from the GDSC data base and related publications that the models can be trained and tested on [1-4].

## Training two-step models

To apply the two-step modelling approach on any particular drug compound of the GDSC data set, the function twostepmodel.m is called with the index of the compound of interest. Said index can be determined by loading the file DrugOrder.mat or DrugOrder.csv and is an integer between 1 and 265. The following command runs the ansatz for the compound Erlotinib, for instance:

drug = 1;

[ models1, models2, prediction\_training, prediction\_test, response, trainingstats, teststats, AUCs, var\_sets, coeffs, ablstudies\_results, cv\_part ] = twostepmodel( drug );

The values that are returned by the function are explained in the following table:

|  |  |  |
| --- | --- | --- |
| Name | Quantity of interest | Details |
| models1 | First-step models | Predicted response for all clusters calculated by the discrete first-step models; pathway activation- and gene expression-based model; feature values associated with the genetic feature-based clusters |
| models2 | Second-step models | 13 multi-omics models |
| prediction\_training | Predictions on the training set | Binarised predictions of all first-step models and second-step models; pre-binarised predictions of all first-step models and all second-step models minus the naïve Bayes model |
| prediction\_test | Predictions on the test set | Binarised predictions of all first-step models and second-step models; pre-binarised predictions of all first-step models and all second-step models minus the naïve Bayes model |
| response | Measured response | Measured response data, both unprocessed and binarised, in the training and in the testing set |
| trainingstats | Predictive performance on the training set | Evaluation metrics – accuracy, precision, recall, f1-score, FDR – for all first- and second step models in training |
| teststats | Predictive performance on the test set | Evaluation metrics – accuracy, precision, recall, f1-score, FDR – for all first- and second step models in testing |
| AUCs | ROC-AUCs | ROC-AUCs of all first- and second-step models, with the exception of the naïve Bayes classifier, in training and testing |
| var\_sets | Significant predictive features | Sets of relevant mutation, CNV, and methylation events as well as tissue types used in the discrete first-step models,  in addition to the three extended sets of genetic features, including redundant features |
| coeffs | Weights and importance scores for first-step models | Indices of all non-constant first-step models; input weights as calculated by the linear and logistic regression models  and the SVMs as well as input importance scores calculated by the ensemble models |
| ablstudies\_results | Ablation studies | Up to 41 ablation models for each second-step model; ROCAUCS, if applicable, and accuracy values of all ablation models in training and testing |
| cv\_part | Cross-validation | Partition object used for the 10-fold cross validation |

Any output variables pertaining to the first-step single-omics models are ordered as follows:

1. somatic mutation-based model,
2. CNV-based model,
3. hypermethylation-based model,
4. tissue descriptor-based model,
5. pathway activation-based model,
6. gene expression-based model.

The corresponding ordering of variables containing information about second-step multi-omics

models is as follows:

1. Neural network
2. LASSO-regularised linear regression
3. Elastic net-regularised linear regression
4. Ridge-regularised linear regression
5. LASSO-regularised logistic regression
6. Elastic net-regularised logistic regression
7. Ridge-regularised logistic regression
8. Ridge-regularised SVM
9. LASSO-regularised SVM
10. Bagged decision tree ensemble
11. Boosted decision tree ensemble
12. Random forest

## Additional material

This repository additionally contains a Matlab-script and additional m-files and data files that enable the user to recreate the results discussed in the doctoral thesis of Nina Kusch, titled *Two-Step Models for Tumour-Drug Response Using Heterogeneous High-Dimensional Assays*. The contents of the folders *Code*, *Data* and *Script* ought to be saved into one common folder to run through the m-file Script.m. In order to recreate the portion of results pertaining to a study-internal benchmark, it is necessary to temporarily switch to the subfolder *one-step models*. For the last part of the findings, it is also necessary to include additional data from a study conducted by Jang et al.[5]. Details are included in the script itself.

## Further reading

For more details on the two-step modelling routine and the data in this repository, we refer to:

* Kusch, N. & Schuppert, A. (2020). Two-step multi-omics modelling of drug sensitivity in cancer cell lines to identify driving mechanisms, PLoS One, e0238961, <https://doi.org/10.1371/journal.pone.0238961>
* Kusch, N., (2021). Two-Step Models for Tumour-Drug Response Using Heterogeneous High-Dimensional Assays (Doctoral thesis, RWTH Aachen University, Aachen, Germany)

## References

[1] M. Garnett, E.J. Edelman, S.J. Heidorn, C.D. Greenman, A. Dastur, K.W. Lau, P. Greninger, I.R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R.J. Milano, G.R. Bignell, A.T. Tam, H. Davies, J.A. Stevenson, S. Barthorpe, S.R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O’Brien, J.L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H.G. Choi, J.W. Chang, J. Baselga, I. Stamenkovic, J.A. Engelman, S.V. Sharma, O. Delattre, J. Saez-Rodriguez, N.S. Gray, J. Settleman, P.A. Futreal, D.A. Haber, M.R. Stratton, S. Ramaswamy, U. McDermott, and

C.H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483:570–575, 2012.

[2] The Cancer Genome Project at the Wellcome Sanger Institute and the Center for Molecular Therapeutics, Massachusetts General Hospital Cancer Center. Online data repository containing additional supplementary data related to A landscape of pharmacogenomic interactions in cancer. https://www.cancerrxgene.org/gdsc1000/GDSC1000\_WebResources/Home.html, Retrieved 03.2017

[3] F. Iorio, T.A. Knijnenburg, D.J. Vis, G.R. Bignell, M.P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R.K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald,

M. Esteller, N.S. Gray, D.A. Haber, M.R. Stratton, C.H. Benes, L.F.A. Wessels, J. Saez-Rodriguez, U. McDermott, and M.J. Garnett. A landscape of pharmacogenomic interactions in cancer. Cell, 166(3):740–754, 2016.

[4] F. Iorio, T.A. Knijnenburg, D.J. Vis, G.R. Bignell, M.P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R.K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang,

T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N.S. Gray, D.A. Haber, M.R. Stratton, C.H. Benes, L.F.A. Wessels, J. Saez-Rodriguez, U. McDermott and M.J. Garnett. Supplementary data of A landscape of pharmacogenomic

interactions in cancer. https://www.cell.com/fulltext/S0092-8674(16)30746-2#supplementaryMaterial, Retrieved 12.2018. DOI: <https://doi.org/10.1016/j.cell.2016.06.017>.

[5] I.S. Jang, E.C. Neto, J. Guinney, S.H. Friend, and A.A. Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pacific Symposium on Biocomputing, pages 63–74, 2014.