Artificial intelligence for drug response prediction in disease models

Ballester, Pedro J.

2022-01-17

Ballester , P J , Stevens , R , Haibe-Kains , B , Huang , R S & Aittokallio , T 2022 , ' Artificial intelligence for drug response prediction in disease models ' , Briefings in Bioinformatics , vol. 23 , no. 1 , 450 . https://doi.org/10.1093/bib/bbab450

http://hdl.handle.net/10138/353596
https://doi.org/10.1093/bib/bbab450

unspecified
acceptedVersion

Artificial intelligence for drug response prediction in disease models

Pedro J. Ballester[1,2,3,4,*], Rick Stevens[5,6], Benjamin Haibe-Kains[7,8,9,10,11,12], R. Stephanie Huang[13], Tero Aittokallio[14,15,16]

[1] Cancer Research Center of Marseille, INSERM U1068, F-13009 Marseille, France

[2] Institut Paoli-Calmettes, F-13009 Marseille, France

[3] Aix-Marseille Université UM105, F-13009 Marseille, France

[4] CNRS UMR7258, F-13009 Marseille, France

[5] Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont IL, 60439, USA

[6] Department of Computer Science, University of Chicago, Chicago IL 60637, USA

[7] Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

[8] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

[9] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

[10] Ontario Institute of Cancer Research, Toronto, Ontario, Canada

[11] Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada

[12] Dalla Lana School of Public Health, Toronto, Ontario, Canada

[13] Department of Experimental and Clinical Pharmacology, College of Pharmacy, University of Minnesota, Minneapolis, MN 55455 USA

[14] Institute for Molecular Medicine Finland (FIMM), Nordic EMBL Partnership for Molecular Medicine, HiLIFE, University of Helsinki, Finland

[15] Institute for Cancer Research (ICR), Oslo University Hospital, Oslo, Norway

[16] Oslo Centre for Biostatistics and Epidemiology (OCBE), University of Oslo, Oslo, Norway

* correspondence to: pedro.ballester@inserm.fr

Accumulated preclinical data are increasingly being re-used to build and validate predictive models generated by Artificial Intelligence (AI)[1] algorithms. Such *in silico* models have a range of applications in biomedicine and healthcare, including drug discovery[2] and precision oncology[3]. Here, we focus on predictive modelling of phenotypic activities of molecules in non-molecular targets, e.g., in cancer cell lines or bacterial cultures, serving as preclinical disease models. Machine learning (ML), including deep learning (DL), is by far the most commonly employed AI subarea to tackle this important healthcare problem.

This Special Issue (SI) covers various aspects of cancer pharmaco-omic modeling, such as the diverse ways in which preclinical datasets to re-use for ML are generated[4], how molecules are characterized numerically to enable the use of ML algorithms[5], which ML-based drug response prediction models have been built to date[6], in which scenarios DL-based models tend to work better than models built with less recent ML algorithms[7],

which datasets lead to the most predictive ML models[8], what types of modeling challenges are caused by the diversity of cell lines[9] or how software implementing preclinical models can facilitate their application to patients[10].

More concretely, Piyawajanusorn et al.[4] provide a primer on understanding preclinical data for cancer pharmaco-omic modeling (both drug response and drug combination synergy). The paper covers the following topics: experimental models of cancer tumors (tumor samples, *in vitro* tumor models, *in vivo* tumor models), multi-omics tumor profiles (genomic, transcriptomic, epigenomic, proteomic, and bulk *versus* single-cell profiling), drug-sensitivity profiling of tumors (*in vitro*, *ex vivo* and *in vivo*) and primary data resources (for each resource, the number of drugs, cell lines and cancer types, the employed assay and the first ML-modelling studies in those datasets).

The application of ML to drug response prediction requires a way to characterize molecules numerically. There is a range of such molecular representations and An et al. [5] presents an entry-level review covering those most widely used. More precisely, this review comprises various linear notations (InChI, SMILES), molecular fingerprints (structural keys, circular fingerprints), graph notations (MPNN, GCN, AGBT) and related topics (pre-training, multi-task learning). For each representation, its generation mechanism, implementation aspects and application examples are outlined.

Firoozbakht et al.[6] provides a systematic review of monotherapeutic drug response prediction built from preclinical molecular profiles and responses to various drugs. Summarized here are over 70 ML methods as well as commonly used data sources for training, input and output data types, and evaluation methods. Introductory material to both ML principles as well as the types of biological data used for model generation is provided for those new to the subject.

Complex DL methods are increasingly being used in the prediction of drug responses in cancer cell lines. In addition to reviewing the state-of-the-art in the field, Chen and Zhang[7] compared the performance of recent DL methods, and pointed out a number of limitations in their current performance, especially in a blind test scenario where a non-DL method worked substantially better. This indicates potential over-fitting of DL to limited training data, which has also been observed when modelling clinical data[11,12]. The authors also

recommend several improvements that complement those suggested in the other recent reviews in the field[13,14].

Xia et al.[8] focus on the generalizability of drug response prediction models for cancer cell lines. They observe that cross-validations tend to overestimate model performance and are thus inadequate for practical scenarios that involve more than one dataset. They review the factors that contribute to the differences in feature and response data, highlighting the effects of assay variability and drug diversity in raising prediction upper bounds. Through ML model comparison and simulation experiments, they also provide some general recommendations for future drug screening experiments.

Sharifi-Noghabi et al.[9] explore the different aspects of the development of univariate and multivariate predictors of monotherapy response *in vitro*. They found that the use of specific datasets led to more generalizable predictors, and that generalizability must be computed on fully independent datasets to test the models' capacity to overcome the inevitable biases originating from different experimental protocols. The authors also found that the diversity of cell lines in these datasets also presented its own challenges, as lymphoid cell lines grown in suspension exhibited higher drug response across on average and their inclusion during model training may lead to suboptimal models.

Maeser et al.[10] present a new R package, *oncoPredict*, to facilitate drug discovery from drug response predictions. This package unites three separate methodologies to (1) predict clinical drug response in patients; (2) associate the predicted drug response with clinical features for *in vivo* drug biomarker discovery; and (3) correct for general levels of drug sensitivity to enable drug-specific biomarker discovery. The authors show how the package can be applied to various *in vitro* and *in vivo* datasets to enable the generation of translational research hypotheses.

In addition to these monotherapy (single-drug) response prediction articles, the SI also extents to drug combination synergy prediction[15] by looking at how to account for experimental noise in drug combination synergy estimation[16], and how single-cell data can be exploited to identify patient-specific drug combinations that selectively co-inhibit only malignant cell populations[17].

Rønneberg et al.[16] went beyond monotherapies to develop a novel approach to estimate the effects of the combinations of drugs on cancer cell viability by accounting for the

inevitable experimental noise in drug screening data[18,19]. The authors developed the open-source *bayesynergy* package which implements a probabilistic model where the drug combination surface is modelled using a Bayesian approach. The *bayesynergy* package is likely to enable future research as the compendium of drug combination datasets quickly grows over time[20,21].

To address the intra- and inter-tumoral heterogeneity when identifying combinatorial treatment regimens for cancer patients, He et al.[17] demonstrated in an ovarian cancer case study how a ML-based platform enables prediction of drug combinations that selectively co-inhibit only malignant cell populations in individual patient samples. The platform makes use of data from single-cell imaging drug response assay, combined with genome-wide transcriptomic and genetic profiles, and it is widely applicable also to other cancer types to predict cancer-selective and patient-specific combinations.

Last but not least, the SI includes a study presenting a new antimicrobial pharmaco-omic dataset and its ML models[22]. Antimicrobial resistance (AMR) is a global health threat impacting millions of people each year. Understanding the transmission of AMR, and rapidly predicting resistance in pathogens, is important for reducing disease burden and improving patient outcomes. In this context, VanOeffelen et. al.[22] describe a curated collection of over 60,000 bacterial genomes paired with laboratory-derived antimicrobial susceptibility test data for use in both traditional bioinformatics and ML studies. To demonstrate the utility of the collection, they build a set of ML models for classifying susceptible and resistant phenotypes as well as predicting antimicrobial minimum inhibitory concentrations.

We hope this SI will provide a concrete and practical guidance to both experienced modellers and newcomers about the data and methods being used and their impact in this research area. We would like to thank the journal staff, reviewers and authors for their work to make this SI possible.

Biographies

Pedro J. Ballester is a Group Leader in machine learning for healthcare at INSERM (France). Postdoctoral research at EMBL-EBI, Cambridge University and Oxford University (UK). PhD in computational geophysics at Imperial College London (UK).

Rick Stevens is an Associate Laboratory Director at Argonne National Laboratory and a Professor in Computer Science at University of Chicago (USA).

Benjamin Haibe-Kains is a Senior Scientist at the Princess Margaret Cancer Centre and Associate Professor in the departments of Medical Biophysics and Computer Science of the University of Toronto (Canada). PhD in Computer Science from the Université Libre de Bruxelles (Belgium).

R. Stephanie Huang is an Associate Professor at the Department of Experimental and Clinical Pharmacology, University of Minnesota (USA). The Huang laboratory's main research focus is translational pharmacogenomics, with particular interest in the pharmacogenomics of anti-cancer agents.

Tero Aittokallio is a Group Leader at FIMM and ICR, and Professor at OCBE (Finland and Norway). His groups use network-centric and machine learning-based approaches and preclinical models to predict optimal treatment regimens for cancer patients.

References

1. Jordan MI. Artificial Intelligence—The Revolution Hasn't Happened Yet. Harvard Data Sci. Rev. 2019; 1–9

2. Freedman DH. Hunting for New Drugs with AI. Nature 2019; 576:S49–S53

3. Ballester PJ, Carmona J. Artificial intelligence for the next generation of precision oncology. npj Precis. Oncol. 2021; 5:1–3

4. Piyawajanusorn C, Nguyen LC, Ghislat G, et al. A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling. Brief. Bioinform. 2021; bbab312

5. An X, Chen X, Yi D, et al. Representation of molecules for drug response prediction. Brief. Bioinform. 2021;

6. Firoozbakht F, Yousefi B, Schwikowski B. An Overview of Machine Learning Methods for Monotherapy Drug Response Prediction. Brief. Bioinform. 2021;

7. Chen Y, Zhang L. How much can deep learning improve prediction of the responses to drugs in cancer cell lines? Brief. Bioinform. 2021; 2021:1–8

8. Xia F, Allen J, Balaprakash P, et al. A cross-study analysis of drug response prediction in cancer cell lines. Brief. Bioinform. 2021;

9. Sharifi-Noghabi H, Jahangiri-Tazehkand S, Smirnov P, et al. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. Brief. Bioinform. 2021;

10. Maeser D, Gruener RF, Huang RS. oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data . Brief. Bioinform.

2021;

11. Bomane A, Gonçalves A, Ballester PJ. Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data. Front. Genet. 2019; 10:1041

12. Chen D, Liu S, Kingsbury P, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. npj Digit. Med. 2019; 2:43

13. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. Brief. Bioinform. 2021; 22:360–379

14. Chiu Y-C, Chen H-IH, Gorthi A, et al. Deep learning of pharmacogenomics resources: moving towards precision oncology. Brief. Bioinform. 2019;

15. Menden MP, Wang D, Mason MJ, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat. Commun. 2019; 10:2674

16. Rønneberg L, Cremaschi A, Hanes R, et al. bayesynergy: flexible Bayesian modelling of synergistic interaction effects in *in vitro* drug combination experiments. Brief. Bioinform. 2021; 2021:1–12

17. He L, Bulanova D, Oikkonen J, et al. Network-guided identification of cancer-selective combinatorial therapies in ovarian cancer. Brief. Bioinform. 2021;

18. Niepel M, Hafner M, Mills CE, et al. A Multi-center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. Cell Syst. 2019; 9:35-48.e5

19. Haibe-Kains B, El-Hachem N, Birkbak NJ, et al. Inconsistency in large pharmacogenomic studies. Nature 2013; 504:389–93

20. Zheng S, Aldahdooh J, Shadbahr T, et al. DrugComb update: A more comprehensive drug sensitivity data repository and analysis portal. Nucleic Acids Res. 2021; 49:W174–W184

21. Seo H, Tkachuk D, Ho C, et al. SYNERGxDB: An integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. Nucleic Acids Res. 2020; 48:W494–W501

22. VanOeffelen M, Nguyen M, Aytan-Aktug D, et al. A genomic data resource for predicting antimicrobial resistance from laboratory-derived antimicrobial susceptibility

phenotypes. Brief. Bioinform. 2021;