


# Reusability report: Evaluating reproducibility and reusability of a fine-tuned model to predict drug response in cancer patient samples

Received: 28 February 2023

Accepted: 9 June 2023

Published online: 10 July 2023

 Check for updates

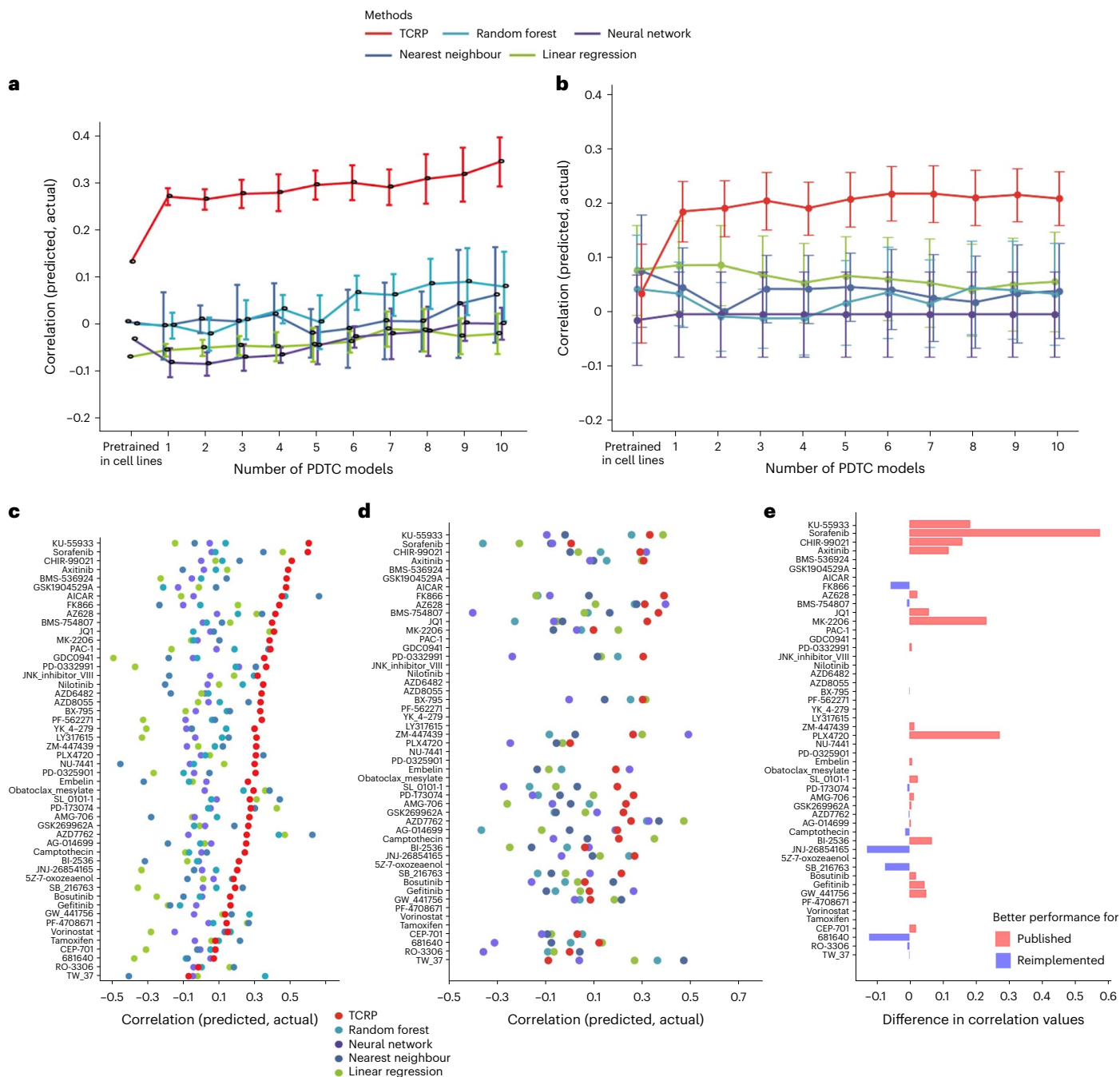
Emily So<sup>1,2</sup>, Fengqing Yu<sup>1,2</sup>, Bo Wang<sup>1,2,3,4</sup> & Benjamin Haibe-Kains<sup>1,2,3,4,5,6</sup> 

Machine learning and artificial intelligence methods are increasingly being used in personalized medicine, including precision oncology. Ma et al. (*Nature Cancer* 2021) have developed a new method called ‘transfer of cell line response prediction’ (TCRP) to train predictors of drug response in cancer cell lines and optimize their performance in higher complex cancer model systems via few-shot learning. TCRP has been presented as a successful modelling approach in multiple case studies. Given the importance of this approach for assisting clinicians in their treatment decision processes, we sought to independently reproduce the authors’ findings and improve the reusability of TCRP in new case studies, including validation in clinical-trial datasets—a high bar for drug-response prediction. Our reproducibility results, while not reaching the same level of superiority as those of the original authors, were able to confirm the superiority of TCRP in the original clinical context. Our reusability results indicate that, in the majority of novel clinical contexts, TCRP remains the superior method for predicting response for both preclinical and clinical settings. Our results thus support the superiority of TCRP over established statistical and machine learning approaches in preclinical and clinical settings. We also developed new resources to increase the reusability of the TCRP model for future improvements and validation studies.

With recent advances in molecular profiling and computational technologies, there has been increasing interest in developing and using machine learning (ML) and artificial intelligence (AI) methods for personalized medicine and precision oncology. One active area of research focuses on computational models capable of predicting the therapy response in patients with cancer. Given the challenges present in obtaining a compendium of clinical genomic data that is sufficiently large for multivariable analysis, the majority of studies train drug-response predictors on large-scale preclinical pharmacogenomic

data and assess the performance of the resulting predictive models using limited patient data. Preclinical models, however, do not always faithfully recapitulate the therapy response observed in patients. Possible underlying reasons for this limitation include the lack of tumour microenvironment components in cell lines and the clear differences in the evolutionary pressures in human physiological environments and cell cultures<sup>1</sup>. Translating predictors trained on preclinical data to achieve good accuracy on clinical data remains an open challenge. If the translation is successful, these predictors hold

<sup>1</sup>Princess Margaret Cancer Centre, University Health Network, Toronto, Canada. <sup>2</sup>Medical Biophysics, University of Toronto, Toronto, Canada. <sup>3</sup>Vector Institute for Artificial Intelligence, Toronto, Canada. <sup>4</sup>Department of Computer Science, University of Toronto, Toronto, Canada. <sup>5</sup>Ontario Institute for Cancer Research, Toronto, Canada. <sup>6</sup>Department of Biostatistics, Dalla Lana School of Public Health, Toronto, Canada. ✉e-mail: [benjamin.haibe-kains@uhn.ca](mailto:benjamin.haibe-kains@uhn.ca)



**Fig. 1 | Reproducibility attempt for Challenge #2 from the original paper**

**by Ma et al. a.** The original published results of Challenge 2. The results show that, across 50 priority drugs, the TCRP model had higher Pearson's correlation than baseline methods in predicting drug response in PDTCs when trained on GDSC1 cell lines. The error bars indicate mean performance across all drugs  $\pm$  95% confidence interval (CI). **b.** Our reproducibility attempt using resources and code provided in the original publication. The plot demonstrates that we were not able to reproduce the same objective value as the published TCRP performance, but were able to prove that it performs better than baselines. **c.** Dot plot depicting the

original plot of published TCRP versus baseline performance, separated by drug. **d.** Dot plot depicting our reimplementations' TCRP versus baseline performance, separated by drug. Once again there were differences in objective values of TCRP between our reimplementations and the original publication, but the overall TCRP performance is better than the baselines. **e.** Bidirectional plot showing the difference between the published objective value and the implementation objective value. Red bars (positive difference) indicate that the published model performed better for a drug's prediction and blue bars (negative difference) that the reimplemented model performed better for a drug's prediction.

great potential for improving the selection of anticancer therapies, a key challenge in precision oncology.

In a recent paper in *Nature Cancer*, Ma et al. introduced 'transfer of cell line response prediction' (TCRP)<sup>2</sup>, a new method for transferring drug response prediction based on few-shot learning. The authors showed that their approach enabled the development of computational models able to learn from immortalized cancer cell-line data

and predict response in more complex in vitro patient-derived cell cultures and in vivo patient-derived xenografts. Given the impressive nature of the original paper's results, this Reusability Report aims to address two issues: (1) confirming the performance of the TCRP model in its published context and (2) expanding its application on a larger compendium of preclinical pharmacogenomic and clinical-trial data. Following successful testing and deployment, modelling approaches

such as TCRP will help improve personalized medicine by facilitating the matching of a patient's molecular profile to optimal therapy.

## Reproducibility

We first aimed to fully reproduce the results outlined in the original publication: training drug response predictors using the GDSC1<sup>3</sup> immortalized cancer cell-line dataset and test its predictive value in patient-derived breast tumour cells (PDTCs)<sup>4</sup>. This task was called Challenge #2 in the original publication and was split into four stages: (1) extracting gene expression and drug-response features from the dataset (GDSC1); (2) training the model-agnostic meta-learning (MAML) algorithm to predict the drug response in cell lines, as well as hyperparameter tuning; (3) fine-tuning these models using a small subset of the PDTc dataset using few-shot learning; (4) validation on the remaining PDTc samples.

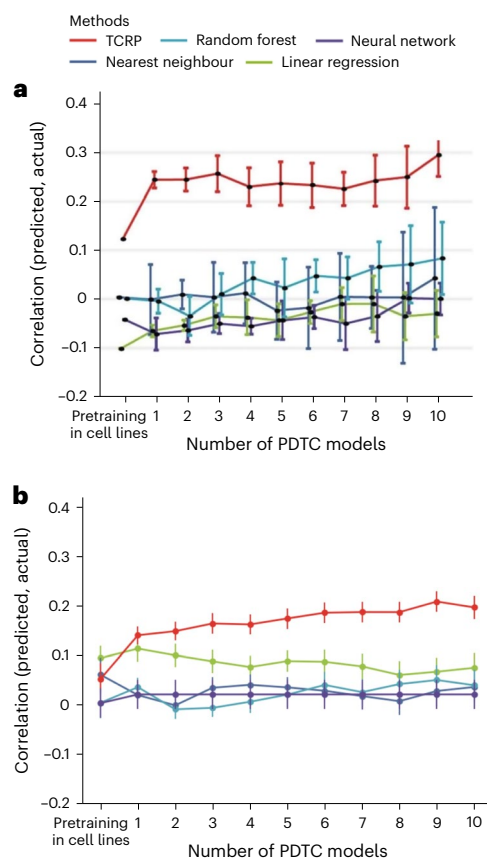
The authors shared the links to the GDSC1 and PDTc datasets, which are both publicly available. Their computer code was shared via two GitHub repositories, 'original codebase' and 'tcp-reproduce' (both provided by the original authors). In the original codebase, a download link for the TCRP model is provided, with one line command to run the model on a single drug, sorafenib. However, there is no instruction on model training or hyperparameter tuning on other drugs. In the 'tcp-reproduce' repository there are instructions for TCRP model training and baseline comparison to established predictive modelling approaches, but the instructions are incomplete. The software versions and dependencies are also listed.

In stage (1) we could only extract the input features (cell lines' expression and mutational profiles, drug response values) for 32 out of 50 priority drugs investigated in the original publication due to missing reference files. Possible causes of this issue are discussed in the Methods.

In stage (2) we were unable to reproduce the deep neural network (MAML) results exactly due to the missing drugs from stage (1), and the performance comparison was incomplete due to the lack of code for the baseline models' implementation. Of the four baseline models, only code for random forest, linear regression and k-nearest neighbours was provided by the authors. The neural network was described as a 'simplified version' of the TCRP model, and the architecture was not provided in the baseline code. Consequently, we had to reimplement the training and testing procedure of this important baseline model.

In stage (3), the hyperparameters for the fine-tuning process were not released with the original publication, forcing us to conduct a brute-force search to test all possible combinations of values listed by the authors to find the best-performing set of parameters for each drug–tissue combination. Although the original authors provided methods for configuring the tuning for a subset of hyperparameters, we were still unable to achieve the exact objective values published, most probably due to hyperparameter configurations we did not explore during cross-validation (Fig. 1).

In stage (4), the authors showed that their TCRP model outperformed established statistical (linear regression) and ML models (nearest-neighbour, random forest and neural network) on the set of 50 priority drugs on the PDTc dataset (Fig. 1a). Using our reimplemented TCRP model, we were able to confirm the superiority of the TCRP model on average (Fig. 1b). We did not achieve the same level of predictive value, however, with a Pearson correlation between predictions and measured drug response of over 0.3 in the original publication, whereas the reimplemented model only reached a peak predictive value of 0.2 (Fig. 1b). We also observed better predictive value for the simple linear regression model in our reanalysis, which reached a positive Pearson correlation, contrary to the published results (Fig. 1b). At the drug level there were consistent differences in our correlation values in comparison to the original publication (Fig. 1c,d). We observed the biggest differences for the drugs sorafenib and MK-2206. This could be due to the lack of features for extraction (common genes between



**Fig. 2 | Equivalent performance calculations on TCRP and baseline methods, using Spearman's correlation as previously done in the original paper by Ma et al.** **a**, Original supplemental results using Spearman correlation for performance calculations.

The results show that, although Spearman's correlation was consistently lower than the reported Pearson's correlation, the TCRP model remained superior to baseline methods in predicting drug response in PDTcs when trained on GDSC1 cell lines. **b**, Performance calculation using Spearman's correlation of our reproducibility attempt. We observed the same trend in performance calculation as the original authors. Using the 32 priority drugs for which we were able to obtain features, the results show that the TCRP model was also superior to baseline methods, with Spearman correlation values on average 0.01–0.04 lower than the Pearson correlation across the number of samples.

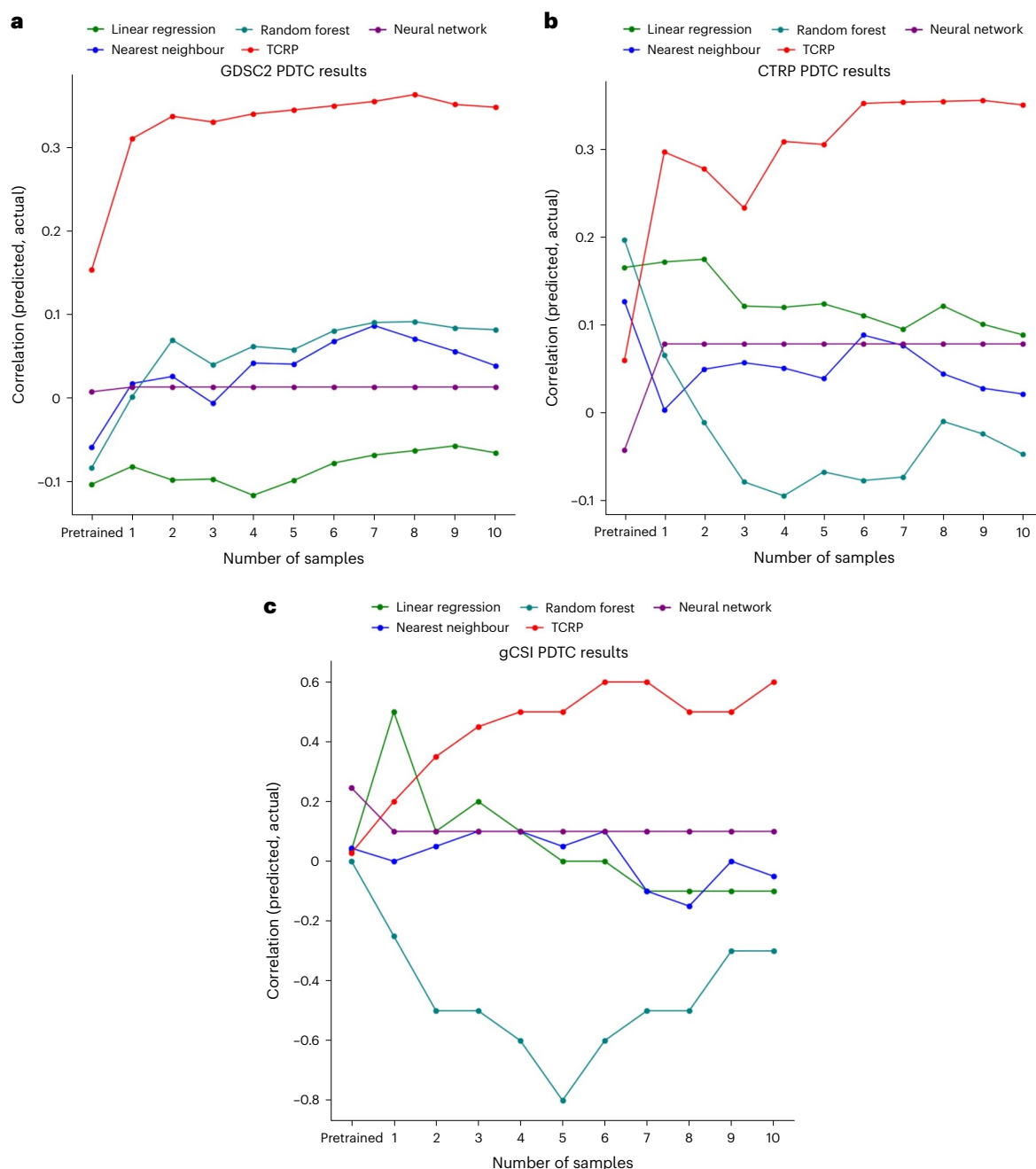
reference and transfer datasets), inconsistent names in drug targets or errors in the hyperparameter tuning (Fig. 1e).

In the original publication, the authors also included, as a supplementary image, the performance of the TCRP method in comparison to baseline methods using Spearman correlation instead of Pearson correlation (Fig. 2a). When performing the same experiment using our reproducibility attempt, we achieved similar results. Although minimal, the Spearman correlation of TCRP performance was on average 0.01–0.04 lower than the Pearson correlation across the number of samples (Fig. 2b). Although this is an interesting comparison to include in our reproducibility attempt, in previous comparisons of Pearson versus Spearman correlations, the former has proven to be the more powerful statistic<sup>5</sup> and therefore the results using that metric can take precedence in this context.

Overall, we were able to reimplement the TCRP model and to confirm its superiority compared to established statistical and ML models using the same training (GDSC1) and validation (PDTc) datasets.

## Reusability

To assess whether the TCRP model can be applied to preclinical and clinical datasets that were not explored in the original publication,



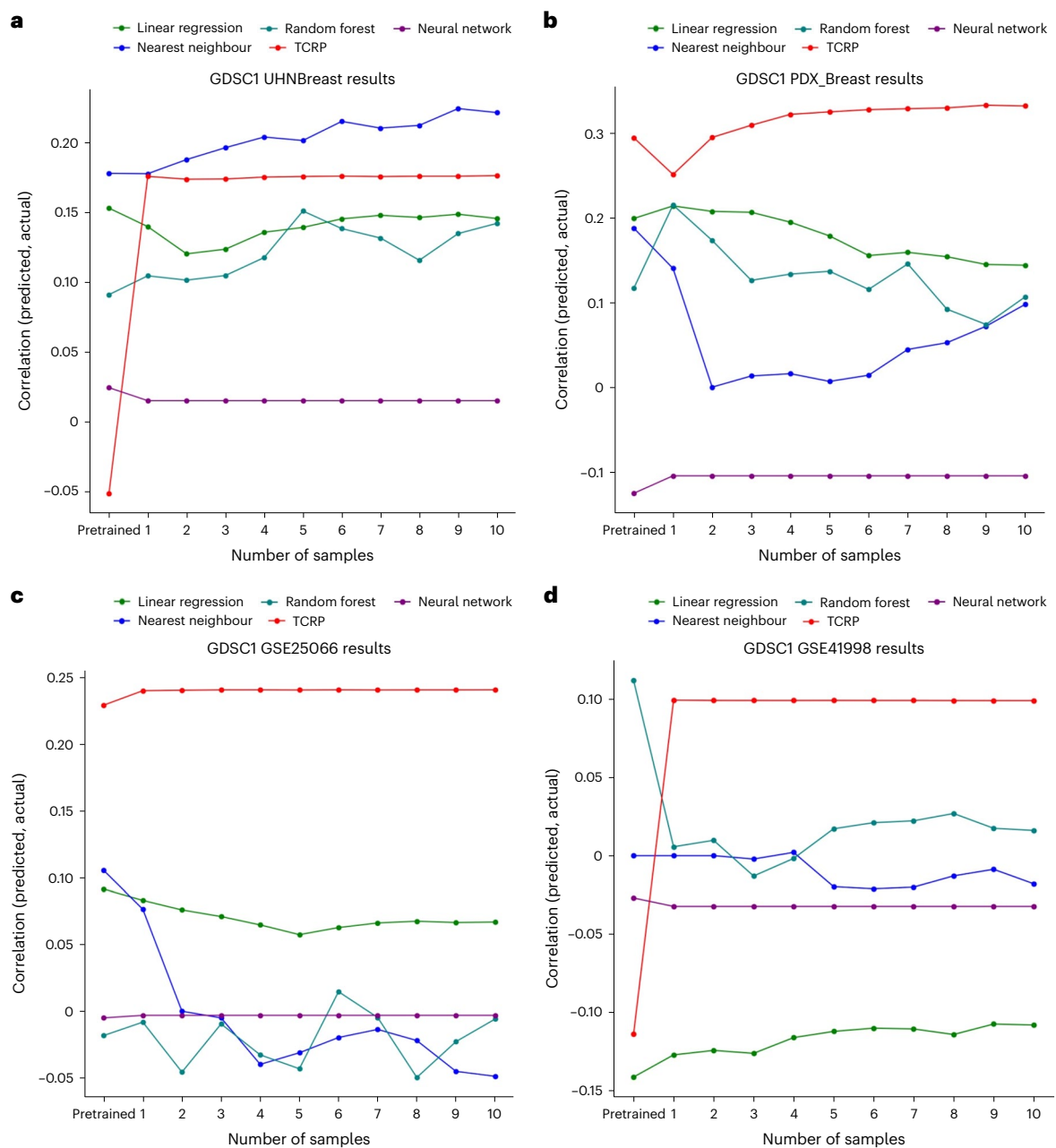
**Fig. 3 | Performance of TCRP versus common baselines on new reference datasets.** For each new reference dataset, instead of using the original 50 priority drugs, any drugs with a sufficient number of cell-line response values that were also present in the PDTC data were used. The drug sets are different across all three datasets. **a**, Performance of TCRP versus other baselines on predicting

drug response in PDTCs, trained on GDSC2. **b**, Performance of TCRP versus other baselines on predicting drug response in PDTCs, trained on CTRPv2. **c**, Performance of TCRP versus other baselines on predicting drug response in PDTCs, trained on gCSI.

we first investigated its performance when trained on different cell-line datasets available on our ORCESTR platform<sup>6</sup>, namely the CTRP<sup>7</sup>, gCSI<sup>8</sup> and GDSC2<sup>9</sup> datasets. During the training process for each dataset, we again tuned the hyperparameter of the model for each drug. We observed that, for drugs shared among the different training sets, the hyperparameters used for optimal TCRP performance on one dataset cannot seamlessly be used for another (Supplementary Information). Retuning hyperparameters for the TCRP model must occur any time a new dataset (reference or validation) is introduced, or when different sets of drugs are used to extract features. The newly trained TCRP models still outperformed the baseline predictive models, irrespective of

the training set used (Fig. 3), confirming the robustness of the approach with respect to cell-line data. TCRP was the only model that was consistently the best across training datasets. Interestingly, for the only drug common to all cell-line datasets—paclitaxel—we observed that training TCRP on the CTRPv2 dataset achieved lower predictive value on PDTC than the reference datasets (Pearson correlations of 0.48, 0.45 and 0.15 for the gCSI-, GDSC2- and CTRPv2-trained TCRP models, respectively), suggesting that the choice of training data may be important for developing the best drug-response predictors. Independently of TCRP performance, when evaluating performance on new reference datasets, we observed that the performance of baseline methods varied





**Fig. 4 | Performance of TCRP versus other baselines in three new validation contexts. a–d,** Performance was evaluated using only the drug response from paclitaxel on all four datasets, and all four models were trained on GDSC1: TCRP performance versus baselines on predicting the drug response on a new dataset of immortalized breast cancer cell lines (UHNBreast) (**a**); TCRP performance

versus baselines on predicting the drug response on breast cancer patient-derived xenografts models treated with paclitaxel (Novartis PDX Encyclopedia) (**b**); TCRP performance versus baselines on predicting the drug response on two clinical trials (**c,d**).

greatly. Possible factors related to this observation could be the differences in study design, such as sample normalization or the size of the reference training dataset. It could also be related to the authors' code design, such as the neural network performance being affected by the missing original implementation.

To further challenge the reusability of the TCRP approach, we assessed the predictive value of the TCRP model on datasets that are representative of diverse preclinical and clinical scenarios. Using GDSC1 as the training set, we tested the transfer of predictors for response to paclitaxel in breast cancer on (1) immortalized cell lines (UHNBreast<sup>10,11</sup>), (2) patient-derived xenografts (Novartis PDX Encyclopedia<sup>12</sup>) and (3) two clinical trials (GSE25066<sup>13–15</sup> and GSE41998<sup>16</sup>). This was challenging due to the fact that the drug response was assessed

differently in each setting. For in vitro models, such as immortalized cancer cell lines or patient-derived tumour cell cultures, the drug response is measured via drug dose–response curves. For in vivo models, such as patient-derived xenografts, the drug response is measured via tumour-growth curves. In clinical trials, drug responses are evaluated using the Response Evaluation Criteria in Solid Tumors (RECIST) measure<sup>17</sup>, categorizing the tumour response as complete response (CR), partial response (PR), stable disease (SD) or progressive disease (PD). To achieve alignment between these different drug response readouts, it is possible to map all drug responses to categorical variables. For example, Gao et al. introduce the modified RECIST (mRECIST) criterion to map tumour-growth curves to the RECIST categories used in clinical trials<sup>12</sup>. We<sup>18</sup> and others<sup>19</sup> have investigated ways to dichotomize

the drug response from dose–response curves in cell lines. To transfer drug-response prediction across the *in vitro*, *in vivo* and clinical settings, we binarized the drug-response readout as responders (R) versus non-responders (NR). Specifically, we grouped CR and PR as R, and SD and PD as NR. In cell lines, we enforce a threshold cutoff on the area above the dose response curve (AAC)<sup>18</sup>, with  $AAC > 0.2$  classified as R, and  $AAC \leq 0.2$  as NR.

On training and validating TCRP in this new analysis framework, we observed that the TCRP model showed an overall better performance than baseline models in most settings, including the two clinical trials (Fig. 4). Notably, the TCRP model failed to outperform the nearest-neighbour modelling approach in the UHNBreast cell-line dataset (Fig. 4a), while outperforming all other methods in the more complex patient-derived xenograft and clinical-trial datasets (Fig. 4b–d).

Altogether, our reusability study strongly supports the capacity of the TCRP modelling approach to transfer its predictive value by learning from a few samples generated in a new context. Our results show that TCRP can learn the complex relationship between molecular features of cancer cells and their drug response using large cell lines' pharmacogenomic datasets, and fine-tune the resulting predictors to achieve higher performance in more complex preclinical models or in clinical trials.

## Discussion

In the course of reproducing and reusing the TCRP models, some persistent challenges were encountered. The primary issue is the limited number of drugs that we could recover from the reference datasets. On utilizing the same datasets as employed in the original paper, only 32 out of 50 drugs were successfully recovered. Furthermore, on reapplying the model to alternative reference datasets (that is, baseline cell-line drug-screening studies), there was a substantial reduction in the number of drugs that the model was able to predict due to the smaller number of drugs commonly investigated.

One of the larger issues we faced in the reproducibility process was tuning the TCRP model for increased performance. The hyperparameters, including the number of few-shot samples and learning rate, must be uniquely adjusted for each specific drug and validation dataset combination. However, this process appeared to be worthwhile, as it appeared that, with tuning, the TCRP model performed well in various contexts for reusability. There was consistent success in our experiments in predicting response in both cell-line, patient-derived cancer models and clinical datasets.

Corroborating the original publication, our results support the TCRP model as a successful approach for predicting clinical drug response. However, there are additional measures that can be taken to ensure the full reproducibility and reusability of the TCRP approach. Direct replication of a published model through code cloning and reimplementing can be a time-intensive and error-prone process due to factors such as changes in path and directory and software environment requirements. It is thus critical to ensure that the model is implemented in a manner that is easily shareable and executable<sup>20</sup>. To this end, all our experiments were conducted using a Code Ocean capsule, which offers an end-to-end workflow for the model and enables straightforward sharing and execution by other researchers<sup>21</sup>. Additionally, proper preprocessing of the data is crucial for the model's application to other datasets. Our reference and testing datasets were sourced from ORCESTR, which provides a curated dataset with detailed versioning and a unique digital object identifier (DOI), ensuring full transparency and reproducibility of dataset-processing and subsequent analysis<sup>6</sup>. Through the integration and automation of the model construction process with data querying and processing, computational models such as TCRP can be readily evaluated and utilized by researchers, facilitating a greater impact of AI in medicine.

Our reproducibility and reusability study further cements the published few-shot learning model as a ML approach that could be applied to different contexts, including both cell-line and clinical-trial datasets. The success of TCRP in the clinical context suggests the great potential of this approach to improve precision oncology. In addition to introducing new contexts, there are improvements that can be made to the availability of this method for the scientific community. Methods such as the few-shot model can be presented in a containerized fashion, either in a Python package or in an interactive containerized repository on platforms such as Code Ocean.

## Methods

### Reproducibility of the original results

**Stage (1), extracting drug-response data.** The TCRP model first needs to be trained on large cell-line data, or the reference dataset, before transfer learning on PDTs. In the data processing, we needed to find the drug features common to the two datasets. Although we obtained the datasets using the provided link, after running the code for data processing we only obtained the features for 32 out of 50 drugs in the original paper. Further investigation showed that this was caused by inconsistencies among the names of genes at which the drugs were targeted. For example, 'CDKs' were not matched with 'CDK9' and 'CDK8', and the drug with mappable targets 'CDKs' was thus lost. The original authors may have reconciled the differences, but there is no clear indication on where and how they implemented this. It would be helpful if there were reference files or instructions to convert between names and resolve the inconsistencies. Otherwise, manual tracing of each drug and gene name is required to recover more drugs.

When running the TCRP model on other datasets, the number of drugs that can be predicted was reduced. Upon further investigation, we discovered that the assignment and distribution of training and testing samples was suboptimal, leading to instances where only a single testing sample was available. To ensure the accurate calculation of correlation coefficients, we manually removed drugs with only one sample.

**Stage (2), MAML.** The implementation of the baseline model was absent from the pipeline. As the authors did not explicitly define the architecture for the convolutional neural network (CNN) model, we had to implement it ourselves. We decided to follow the implementation of the MAML model architecture, using one linear layer, a ReLU activation layer and a batch normalization layer.

**Stage (3), fine-tuning via few-shot learning.** The hyperparameters used for the fine-tuning process are as follows:

- Tissue number evolved in the inner update: (6, 12, 20)
- Meta learning rate and inner learning rate: (0.1, 0.01, 0.001)
- Number of hidden neurons: (5, 10, 15, 20)
- Number of layers: (1, 2)

For every drug we trained the model using every possible combination of the hyperparameters with their values listed above. The hyperparameter set with the highest correlation was chosen for the specific drug model. This is costly in terms of computation, as a brute-force optimization method was done for each drug and each tissue, and for each time we trained on a new dataset.

**Stage (4), validation on independent data.** With the selected hyperparameters from stage (3), we ran the drug-specific model and made a comparison with the original publication. We listed the drugs in the sequence of their predictive values, as in the original publication. The trend of the predictiveness of the drugs we identified mostly aligned with the original publication. The top predictive drugs were KU-55933 and CHIR-99021. However, there were some drugs, such as Sorafenib and MK-2206, that had surprisingly low correlation values

in our reimplementation, despite having substantially higher values in the original publication.

## Data availability

All datasets used in our study are available on the data platform [ORCES-TRA](#), under dataset names [GDSC\\_2020\(v1-8.2\)](#), [GDSC\\_2020\(v2-8.2\)](#), [gCSI\\_2019](#), [CTRPv2\\_2015](#), [PDTX\\_2019](#), [PDXE](#), [UHNBreast\\_2019](#), [scRNA\\_GSE25066\\_Breast](#) and [scRNA\\_GSE41998\\_Breast](#).

## Code availability

The code to run and visualize the exact results of our reproducibility attempt as well as all our novel analyses for reusability are available in the corresponding Code Ocean capsule (Version 2.5)<sup>22</sup>. Code used for the analysis is available at [github.com/bhklab/TCRP\\_Reusability\\_Report](https://github.com/bhklab/TCRP_Reusability_Report).

## References

1. Trastulla, L., Noorbakhsh, J., Vazquez, F., McFarland, J. & Iorio, F. Computational estimation of quality and clinical relevance of cancer cell lines. *Mol. Syst. Biol.* **18**, e11017 (2022).
2. Ma, J. et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2**, 233–244 (2021).
3. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
4. Bruna, A. et al. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell* **167**, 260–274 (2016).
5. Smirnov, P. et al. Evaluation of statistical approaches for association testing in noisy drug screening data. *BMC Bioinf.* **23**, 188 (2022).
6. Mammoliti, A. et al. Orchestrating and sharing large multimodal data for transparent and reproducible research. *Nat. Commun.* **12**, 5797 (2021).
7. Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
8. Haverty, P. M. et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333–337 (2016).
9. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
10. Safikhani, Z. et al. Gene isoforms as expression-based biomarkers predictive of drug response in vitro. *Nat. Commun.* **8**, 1126 (2017).
11. Thu, K. L. et al. Disruption of the anaphase-promoting complex confers resistance to TTK inhibitors in triple-negative breast cancer. *Proc. Natl Acad. Sci. USA* **115**, E1570–E1577 (2018).
12. Gao, H. et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
13. Hatzis, C. et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305**, 1873–1881 (2011).
14. Itoh, M. et al. Estrogen receptor (ER) mRNA expression and molecular subtype distribution in ER-negative/progesterone receptor-positive breast cancers. *Breast Cancer Res. Treat.* **143**, 403–409 (2014).
15. Baldasici, O. et al. Circulating small EVs miRNAs as predictors of pathological response to neo-adjuvant therapy in breast cancer patients. *Int. J. Mol. Sci.* **23**, 12625 (2022).
16. Horak, C. E. et al. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. *Clin. Cancer Res.* **19**, 1587–1595 (2013).

17. RECIST 1.1 (EORTC); <https://recist.eortc.org/recist-1-1-2/>
18. Safikhani, Z. et al. Revisiting inconsistency in large pharmacogenomic studies. *F1000Res.* **5**, 2333 (2016).
19. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
20. Raff, E. Research reproducibility as a survival analysis. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 35, 469–478 (AAAI, 2021).
21. Clyburne-Sherin, A., Fei, X. & Green, S. A. Computational reproducibility via containers in psychology. *Meta-Psychology* <https://doi.org/10.15626/MP.2018.892> (2019).
22. Reusability Report: Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients (Version 2.5) (Code Ocean); <https://codeocean.com/capsule/8411716/tree/v2>

## Acknowledgements

This work was supported by Canadian Cancer Society Data Transformations (707609) and by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2021-02680).

## Author contributions

E.S. and F.Y. conducted the reproducibility and reusability experiments, curated datasets, drafted the first version of the manuscript and integrated all the edits. B.W. supervised the manuscript. B.H.-K. contributed ideas and supervised the manuscript writing and final edits.

## Competing interests

B.H.K. is a shareholder and paid consultant for Code Ocean Inc. E.S. is a paid consultant for Code Ocean Inc. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42256-023-00688-4>.

**Correspondence and requests for materials** should be addressed to Benjamin Haibe-Kains.

**Peer review information** *Nature Machine Intelligence* thanks Bo Yuan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023