Title: Harnessing AI-detectable molecular patterns to diagnose and treat cancers

Objectives

Cancer poses a significant global health challenge, claiming around 10 million lives each year. Cancer cell heterogeneity and plasticity act as major obstacles to detecting and managing the disease effectively. Our approach employs artificial intelligence (AI) to overcome these hurdles, focusing on the following directions.

1. Detecting and characterizing circulating tumor cells (CTCs): Early cancer detection is crucial to improve life expectancy of the patients. We combine microfluidics, single cell omics and AI to detect circulating tumor cells from patient blood with high accuracy. Our approaches are not contingent on canonical markers alone, and therefore are capable of recognising CTCs of diverse phenotypes.
2. Genome-Informed Personalized Therapeutics: Despite substantial accumulation of genome perturbation, high throughput screening, and cancer genome profiling data, cancer genome - therapy interactions remain elusive. We leverage public data to develop chemo genomic/transcriptomic approaches to model drug response in cell-lines, xenografts, and humans. We also experimentally validate our key findings.

Figure 1 provides a schematic description of both the works.

Methodology details

As mentioned above, our translational cancer research has two principal themes – a. blood based diagnosis, and b. personalized therapy recommendation. Here we discuss methods related to two closely related studies from our lab. First one exploits CTCs for blood based cancer detection, and the next leverages a chemo-transcriptomics approach for drug response prediction in cancer cell lines, xenografts and humans. In both cases, we significantly rely on transcriptomics and AI (Figure 1).

Identification and characterization of single circulating tumor cells from patient blood: CTCs represent cancer cells that lead to cancer metastasis. CTCs are rare in blood (one cell per 10^9 haematologic cells (Nagrath et al. 2007)) and therefore difficult to detect. Existing technologies rely on size and/or marker based enrichment techniques, thereby potentially missing on atypical CTC phenotypes such as ones under epithelial to mesenchymal transition. We combined the power of microfluidics, single cell transcriptomics and AI to address this. Our proposed unCTC workflow (Poonia et al. 2023) projects rather noisy single cell transcriptomes into a meta-space spanning cancer related pathways. This can be attained by computing pathway enrichment scores based on single cell gene expression levels. Further, we developed a tailored clustering algorithm called deep dictionary learning using k-means clustering cost (DDLK) for robust clustering of CTCs and blood cells. unCTC outperformed existing best practice single cell analysis softwares such as Seurat in discriminating CTCs and blood cells. The software was also enabled to approximate Copy Number Variation (CNV) patterns across

subjected cells, thereby highlighting the malignant ones. We validated the utility of unCTC on scRNA-seq profiles of breast CTCs from six patients, captured and profiled using an integrated ClearCell FX and Polaris workflow that works by the principles of size-based separation of CTCs and marker-based immune cell depletion (Figure 2). Further details of the methodology can be found in our recent publications (Iyer et al. 2020; Poonia et al. 2023) .

A tailored chemo-transcriptomics framework to infer cancer drug response: Unpredictable nature of patient specific response of anticancer therapy has attracted significant AI interventions. Most methods based on gene expression profiles fail to tackle the inherent noise of transcriptomic data, and factor in the structural essence of drug compounds. We developed Precily (Chawla et al. 2022), a software that uses deep neural networks to model drug response, using pathway enrichment scores (computed based on gene expression profiling data of cancer samples) and drug descriptors as feature vectors. Large scale cell-line based high throughput screening datasets (GDSC, CCLE, and CTRP) were used for model training, validation and prediction on unseen samples. We compared Precily with a number of state of the art methods such as CaDRReS-Sc (Suphavilai et al. 2021). Precily's performance was rigorously benchmarked and validated against existing methods and scRNA-seq data. In particular, we assessed the predictability of treatment outcomes using our in-house bulk RNA-seq datasets of prostate cancer cell lines and xenografts, exposed to differential treatment conditions. Further, we demonstrated the applicability of our approach on patient drug response data from The Cancer Genome Atlas (TCGA) and an independent clinical study describing the treatment journey of three melanoma patients. We highlighted the utility of Precily in predicting response to (unseen) anti-cancer compounds that are not part of the training data.

Key results

Harnessing Circulating Tumor Cells and Tumor Educated Platelets for blood-based probe into cancers: Most scRNA-seq studies of CTCs rely on marker-based approaches, with few marker-agnostic methods capable of confirming malignancy. Major challenges include high intra- and inter-tumoral molecular diversity (Li et al. 2017; Tirosh et al. 2016), low CTC concentration in peripheral blood (0-10 CTCs/mL) (Alix-Panabières and Pantel 2013), EMT-induced loss of epithelial markers (Mikolajczyk et al. 2011; Iyer et al. 2020), and batch effects across scRNA-seq studies (Kiselev, Andrews, and Hemberg 2019; Büttner et al. 2019).

We used machine learning based approaches to circumvent the aforesaid challenges. We formulated single cell transcriptomics based CTC characterization problems as both unsupervised (Iyer et al. 2020) and supervised (Poonia et al. 2023) learning tasks. In the supervised setting we curated publicly available gene expression profiles of single CTCs and WBCs and trained multiple alternative classifiers on the same, which could accurately recognise unseen CTC gene expression profiles. The classifier used a myriad of feature transcripts, selected following data science principles. We could even detect CTCs that did not express canonical epithelial markers such as EPCAM and CK (Iyer et al. 2020). In this work, we also demonstrated that CTCs lie on a continuum of epithelial to mesenchymal transition. To reduce the dependency on annotated single CTC and WBC transcriptomes, we developed unCTC, an

unsupervised, deep dictionary learning based approach for low dimensional clustering of the cellular expression profiles. We validated the utility of unCTC on scRNA-seq profiles of breast CTCs from six breast cancer patients, captured and profiled using an integrated ClearCell® FX and PolarisTM workflow (Figure 2) that works by the principles of size-based separation of CTCs and marker-based WBC depletion (Poonia et al. 2023).

The lack of typical surface markers in triple negative breast cancer cells (TNBCs) makes it difficult to detect them in a patient's blood, leading to most CTC studies focusing on other categories of breast cancer instead. The lack of specific surface markers on TNBC cells has hampered the development of CTC isolation and analysis strategies for this aggressive subtype. Our machine learning approach, unCTC could recognise single cell RNA-sequencing (scRNA-seq) profiles of the CTCs (Figure 3), captured by the integrated ClearCell® FX and PolarisTM workflow (Poonia et al. 2023).

A chemo-transcriptomics framework for cancer drug response inference: The significant extent of inter and intra-tumoral heterogeneity observed across cancer types continue to pose substantial challenges to the medical management of the disease. Precision oncology offers much-needed solutions by leveraging advancements in genomic platforms, high throughput drug screening technologies, and artificial intelligence. The past few years have seen a considerable rise in studies reporting omics plus predictive modeling approaches in inferring drug responses in cancer cell lines, animal models, and patient tumor clones (Chawla et al. 2022; Adam et al. 2020)

We developed Precily, a deep neural network (DNN) based framework to predict the response to cancer therapies based on gene expression profiles and drug descriptors (Figure 4). Precily uses pathway enrichment scores to highlight the underlying biological mechanisms contributing to drug resistance. With Precily, we studied how drug sensitivity prediction indicated a switch of cellular states into drug-responsive and resistant states in LNCaP cells and xenografts. In the presence of androgens, LNCaP cells were predicted to be more sensitive to cancer therapeutics that target highly proliferative cells. This is expected as androgens drive proliferation in androgen receptor positive (AR positive) PCa (prostate cancer) cell lines and xenografts. AR antagonists are the primary treatment option for metastatic PCa, which antagonize cellular androgen response pathways on a molecular level.

With AR antagonist treatments, LNCaP cells were predicted to have pronounced similarities and differences for AR antagonists bicalutamide (BIC), enzalutamide (ENZ), and apalutamide (APA) treatments. A strong reversal by an AR antagonist of dihydrotestosterone (DHT)-conferred sensitivity was observed for drugs targeting the PI3K/mTOR pathway, with ENZ showing the most profound effect. In contrast, ENZ treatment was predicted to further increase the DHT-conferred sensitivity to selected drugs, including cisplatin, docetaxel and paclitaxel, while other AR antagonists may not (Figure 5). These predictions suggest that patients on active treatment with ENZ may still benefit from added chemotherapy using cisplatin, while patients on treatment with BIC or APA may not.

To further evaluate the applicability of the drug sensitivity predictions with Precily, we used data derived from our well-annotated PCa xenograft model, following the progression from early androgen-responsive to CRPC (Castrate-Resistant Prostate Cancer), and then to ENZ treatment responsive and ENZ resistant states. The sensitivity predictions highlighted changing vulnerabilities of the tumors in different stages of progression and treatment. Notably, we observed that enzalutamide resistant (ENZR) tumors are predicted to develop a susceptibility to

selected therapeutics providing a new window of opportunity for therapeutic strategies. For example, our model predicted sensitivity to specific drugs targeting the EGFR signaling pathway in case of ENZ treatment, highlighting sapitinib as a potential therapeutic for ENZ resistant patients (Figure 6).

Precily was also able to successfully predict LNCaP sensitivity to metformin and orlistat, which are drugs that are not part of the modeling task. This suggests that Precily could be potentially used to evaluate the efficacy of drugs that are not well-studied in cancer.

Impact

Given the substantial global burden of cancer, with millions of lives affected each year, our approach strives to address key challenges using artificial intelligence (AI). By combining microfluidics, single cell omics, and AI, we have developed a method for accurate detection and characterization of circulating tumor cells (CTCs) in patient blood. Importantly, our approach is not confined to conventional markers, enabling the identification of diverse CTC phenotypes, including those undergoing epithelial to mesenchymal transition. Our efforts extend to the detection of triple negative breast cancer (TNBC) cells, a significant achievement considering the lack of specific surface markers for this subtype. Furthermore, we have introduced a personalized therapeutic strategy that integrates genomics and AI to model drug responses across diverse scenarios, spanning from cell lines to human samples. Our innovative Precily framework leverages deep neural networks to predict drug responses by analyzing gene expression profiles and drug descriptors. This approach demonstrates its potential to guide the development of more effective treatment plans. Collectively, the methodologies we have developed, as illustrated by unCTC and Precily, show significant promise in reshaping both cancer detection and treatment approaches.

Reference

Adam, George, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. 2020. “Machine Learning Approaches to Drug Response Prediction: Challenges and Recent Progress.” *NPJ Precision Oncology* 4 (June): 19.

Alix-Panabières, Catherine, and Klaus Pantel. 2013. “Real-Time Liquid Biopsy: Circulating Tumor Cells versus Circulating Tumor DNA.” *Annals of Translational Medicine* 1 (2): 18.

Büttner, Maren, Zhichao Miao, F. Alexander Wolf, Sarah A. Teichmann, and Fabian J. Theis. 2019. “A Test Metric for Assessing Single-Cell RNA-Seq Batch Correction.” *Nature Methods* 16 (1): 43–49.

Chawla, Smriti, Anja Rockstroh, Melanie Lehman, Ellca Ratther, Atishay Jain, Anuneet Anand, Apoorva Gupta, et al. 2022. “Gene Expression Based Inference of Cancer Drug Sensitivity.” *Nature Communications* 13 (1): 5680.

Iyer, Arvind, Krishan Gupta, Shreya Sharma, Kishore Hari, Yi Fang Lee, Neevan Ramalingam, Yoon Sim Yap, et al. 2020. “Integrative Analysis and Machine Learning Based Characterization of Single Circulating Tumor Cells.” *Journal of Clinical Medicine Research* 9 (4). https://doi.org/10.3390/jcm9041206.

Kiselev, Vladimir Yu, Tallulah S. Andrews, and Martin Hemberg. 2019. “Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data.” *Nature Reviews. Genetics* 20 (5): 273–82.

Li, Huipeng, Elise T. Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin

Goh, Say Li Kong, et al. 2017. “Reference Component Analysis of Single-Cell Transcriptomes Elucidates Cellular Heterogeneity in Human Colorectal Tumors.” *Nature Genetics* 49 (5): 708–18.

Mikolajczyk, Stephen D., Lisa S. Millar, Pavel Tsinberg, Stephen M. Coutts, Maryam Zomorrodi, Tam Pham, Farideh Z. Bischoff, and Tony J. Pircher. 2011. “Detection of EpCAM-Negative and Cytokeratin-Negative Circulating Tumor Cells in Peripheral Blood.” *Journal of Oncology* 2011 (April): 252361.

Nagrath, Sunitha, Lecia V. Sequist, Shyamala Maheswaran, Daphne W. Bell, Daniel Irimia, Lindsey Ulkus, Matthew R. Smith, et al. 2007. “Isolation of Rare Circulating Tumour Cells in Cancer Patients by Microchip Technology.” *Nature* 450 (7173): 1235–39.

Poonia, Sarita, Anurag Goel, Smriti Chawla, Namrata Bhattacharya, Priyadarshini Rai, Yi Fang Lee, Yoon Sim Yap, et al. 2023. “Marker-Free Characterization of Full-Length Transcriptomes of Single Live Circulating Tumor Cells.” *Genome Research* 33 (1): 80–95.

Suphavilai, Chayaporn, Shumei Chia, Ankur Sharma, Lorna Tu, Rafael Peres Da Silva, Aanchal Mongia, Ramanuj DasGupta, and Niranjan Nagarajan. 2021. “Predicting Heterogeneity in Clone-Specific Therapeutic Vulnerabilities Using Single-Cell Transcriptomic Signatures.” *Genome Medicine* 13 (1): 189.

Tirosh, Itay, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth 2nd, Daniel Treacy, John J. Trombetta, Asaf Rotem, et al. 2016. “Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq.” *Science* 352 (6282): 189–96.

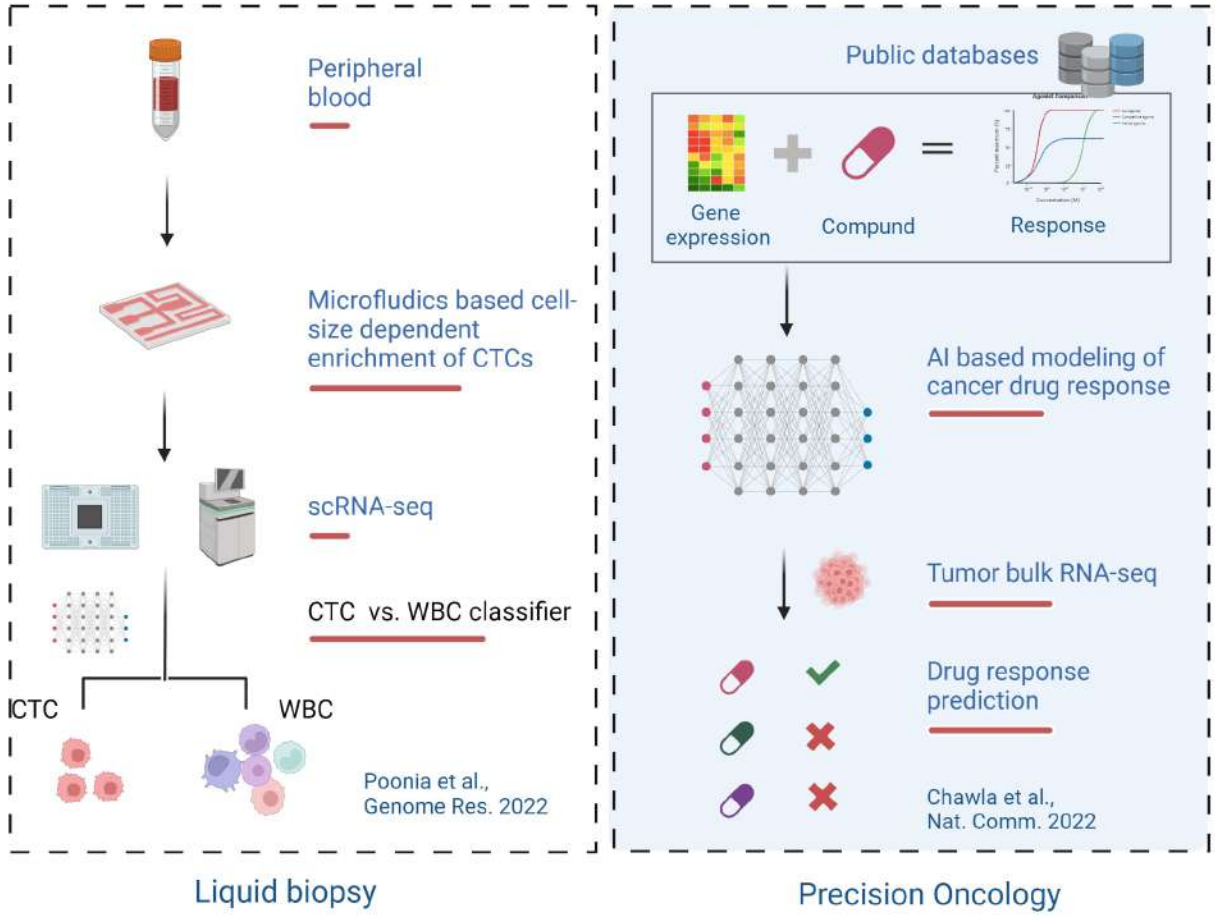

Figure 1: Schematic diagram depicting three major works discussed. On the left, schematic of marker-free characterization of full-length transcriptomes of single live circulating tumor cells; On the right, schematic of a deep learning backed chemo-transcriptomics framework to infer cancer drug response.

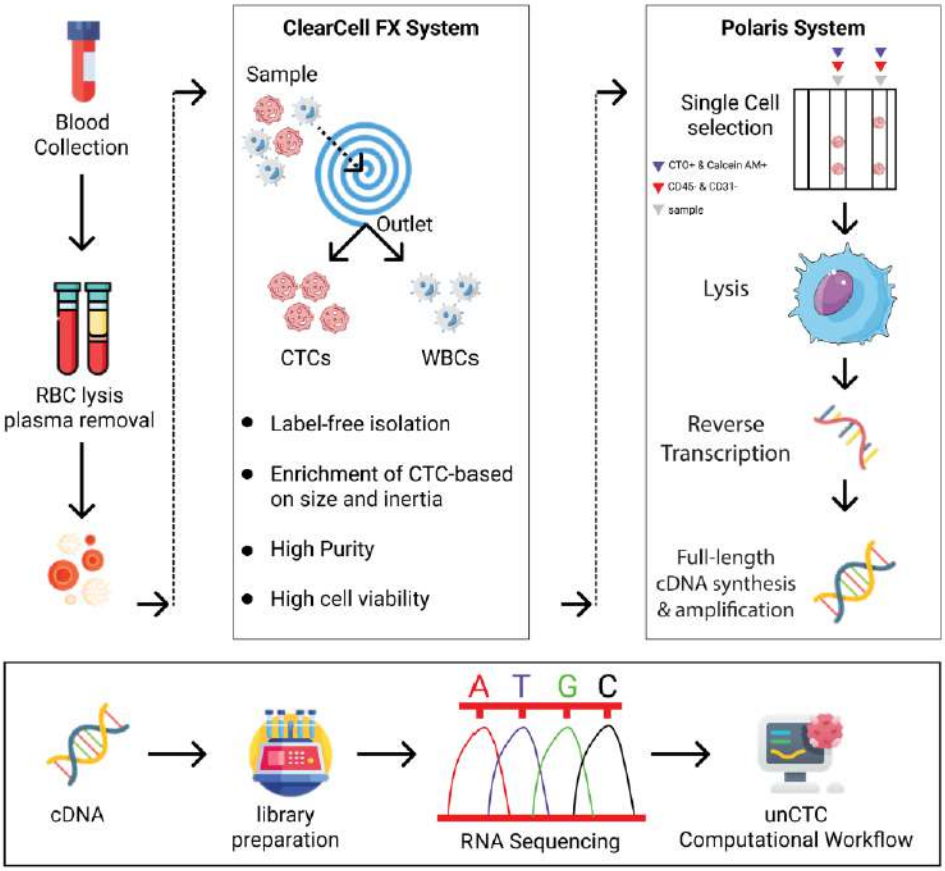

Figure 2: ClearCell FX and Polaris workflow for marker-free enrichment of CTCs. The schematic diagram depicts the key steps involved in the capture and isolation of CTCs using a two-pronged system. ClearCell FX uses a spiral chip to size-sort CTCs. Polaris performs single-cell capture and cDNA synthesis of potential CTCs after depletion of cells that are PTPRC/CD31-positive. Finally, cDNA thus received is subjected to library preparation and RNA sequencing. For further details please see: https://genome.cshlp.org/content/early/2022/11/22/gr.276600.122.

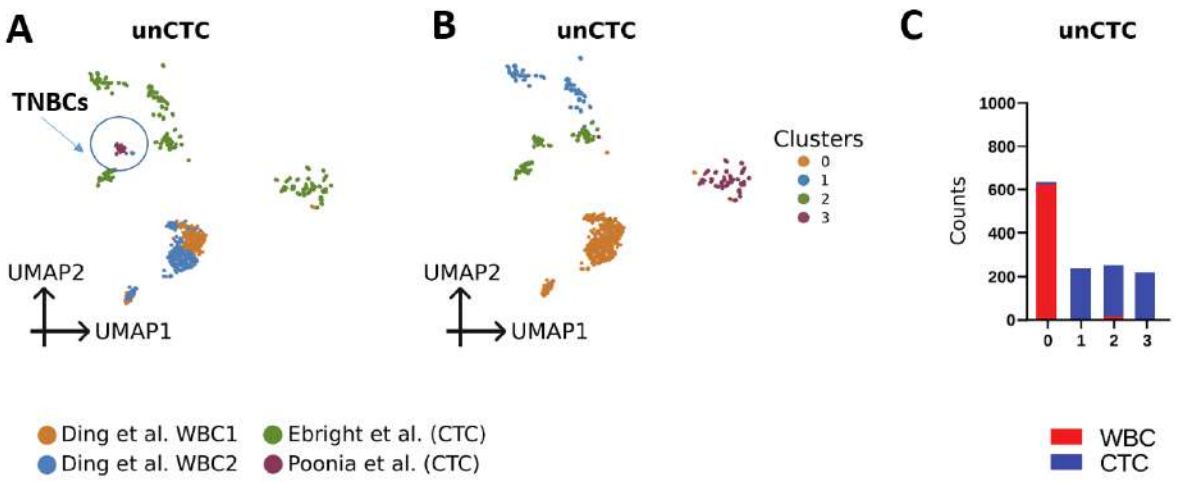

Figure 3: Clustering of CTCs obtained from ClearCell FX–Polaris system. unCTC accurately segregates CTCs and WBCs, obtained from multiple studies. CTCs obtained from ClearCell FX–Polaris system (brown colored TNBC cells) co-cluster with breast CTCs from Ebright et al. data (Ebright et al. 2020). (A) This subfigure colors cells by studies of origin; (B) This subfigure depicts different clusters as obtained by applying unCTC; (C) This subfigure depicts cluster purity. For further details please see: https://genome.cshlp.org/content/early/2022/11/22/gr.276600.122.

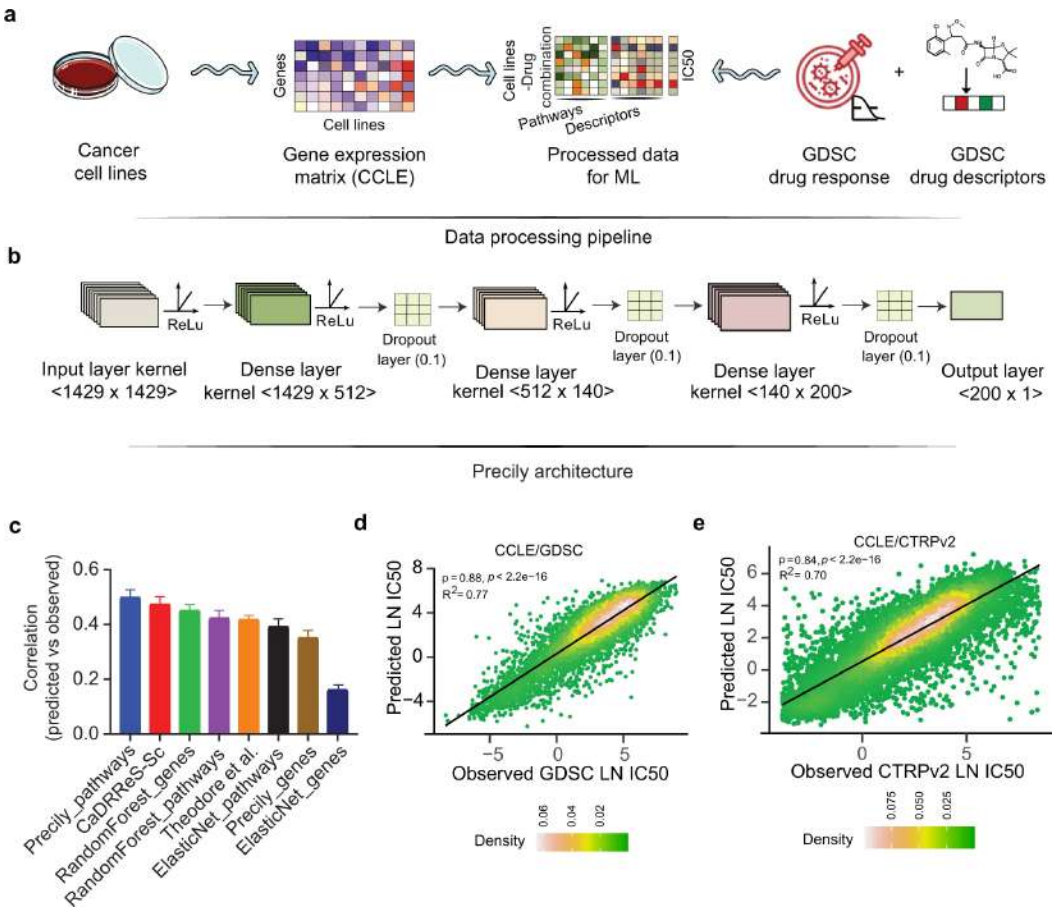

Figure 4: (a) Schematic workflow depicting the data processing pipeline of Precily. The first step involved the processing of training data. The RNA-seq gene expression (RSEM TPM) profiles from Cancer Cell Line Encyclopedia (CCLE) were subjected to pathway score transformation using GSVA. This GSVA score matrix was integrated with the drug descriptors obtained in the form of SMILES embedding for each compound. (b) Model architecture. The second step was the training of the ML model on this data, comprising GSVA scores and drug descriptors as an explanatory variable set and natural log-transformed IC50 values sourced from the GDSC database as the response variable. A deep neural network (DNN) from the Keras platform was used to perform the regression task of predicting drug response. (c) Comparison of drug response prediction across different approaches. Barplot shows the distribution of Pearson's correlation coefficients for predicted vs. observed LN IC50 values for individual drugs (n = 173). Data are presented as mean values + /− SEM (Standard Error of the Mean). (d) Scatter plot demonstrating the performance of Precily across all cell line-drug pairs in the CCLE/GDSC test data. P-value was calculated using a two-sided t-test. (e) Scatter plot demonstrating the performance of Precily across all cell line-drug pairs in the CCLE/CTRPv2 test data. P-value was calculated using a two-sided t-test. For further details please see: https://www.nature.com/articles/s41467-022-33291-z.

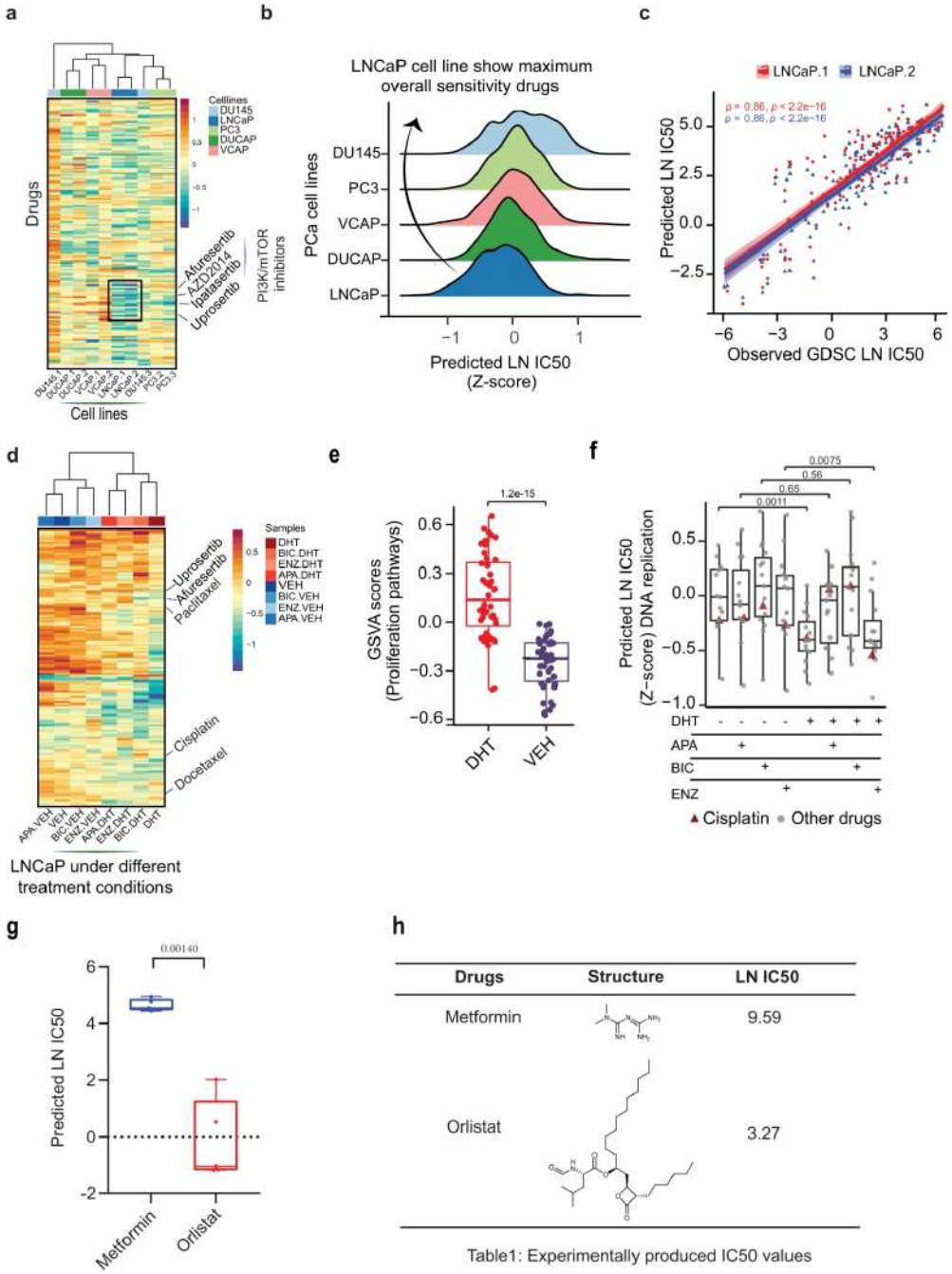

Figure 5: (a) Heatmap showing predicted LN IC50 (Z-score) for 155 drugs across five PCa baseline cell lines highlighting PI3K/mTOR signaling targeting drugs. The lower the LN IC50, the more sensitive a sample is predicted to a drug. Color bars indicate different cell lines. Euclidean distance was used for grouping cell lines. (b) Ridgeplot showing the overall distribution of predicted LN IC50 (Z-score) across five PCa cell lines, while pooling their predicted drug sensitivities across biological replicates. (c) Scatterplot depicting Pearson correlation (ρ) between observed and predicted LN IC50 for the LNCaP cell line. The line color indicated two biological replicates of the LNCaP cell line. P-value was calculated using a two-sided t-test.

Shaded areas depict a 95% confidence interval. (d) Heatmap of predicted LN IC50 (Z-score) of LNCaP cells in the presence and absence of androgens (DHT) and AR antagonists (ENZ, BIC, and APA) to 155 drugs. Euclidean distance was used for grouping samples. (e) Boxplots depicting the distribution of GSVA scores of proliferation-related pathways in the presence and absence of DHT. Notably, n = 12 pathways and n = 48 samples originating from n = 8 treatment groups (DHT, BIC.DHT, ENZ.DHT, APA.DHT, VEH, BIC.VEH, ENZ.VEH and APA.VEH) have been considered for this analysis. P-values were obtained from the two-sided Wilcoxon rank-sum test. (f) Boxplot depicting predicted LN IC50 for DNA replication targeting drugs (n = 15) across all treatment conditions (n = 8). Cisplatin is denoted using darkred colored filled triangle and other drugs are represented using grey filled circles. P-values were obtained from the two-sided Wilcoxon rank-sum test. (g) Boxplot showing the distribution of predicted LN IC50 values by n = 5 pre-trained models (based on cross-validation and hyperparameter tuning) for two drugs, metformin and orlistat. P-value was obtained by using a two-sided t-test. As expected, the direction of the relative difference in drug sensitivity is captured correctly even at the log scale. The structure of these two drugs, along with observed IC50 values is also depicted in (h). For further details please see: https://www.nature.com/articles/s41467-022-33291-z.

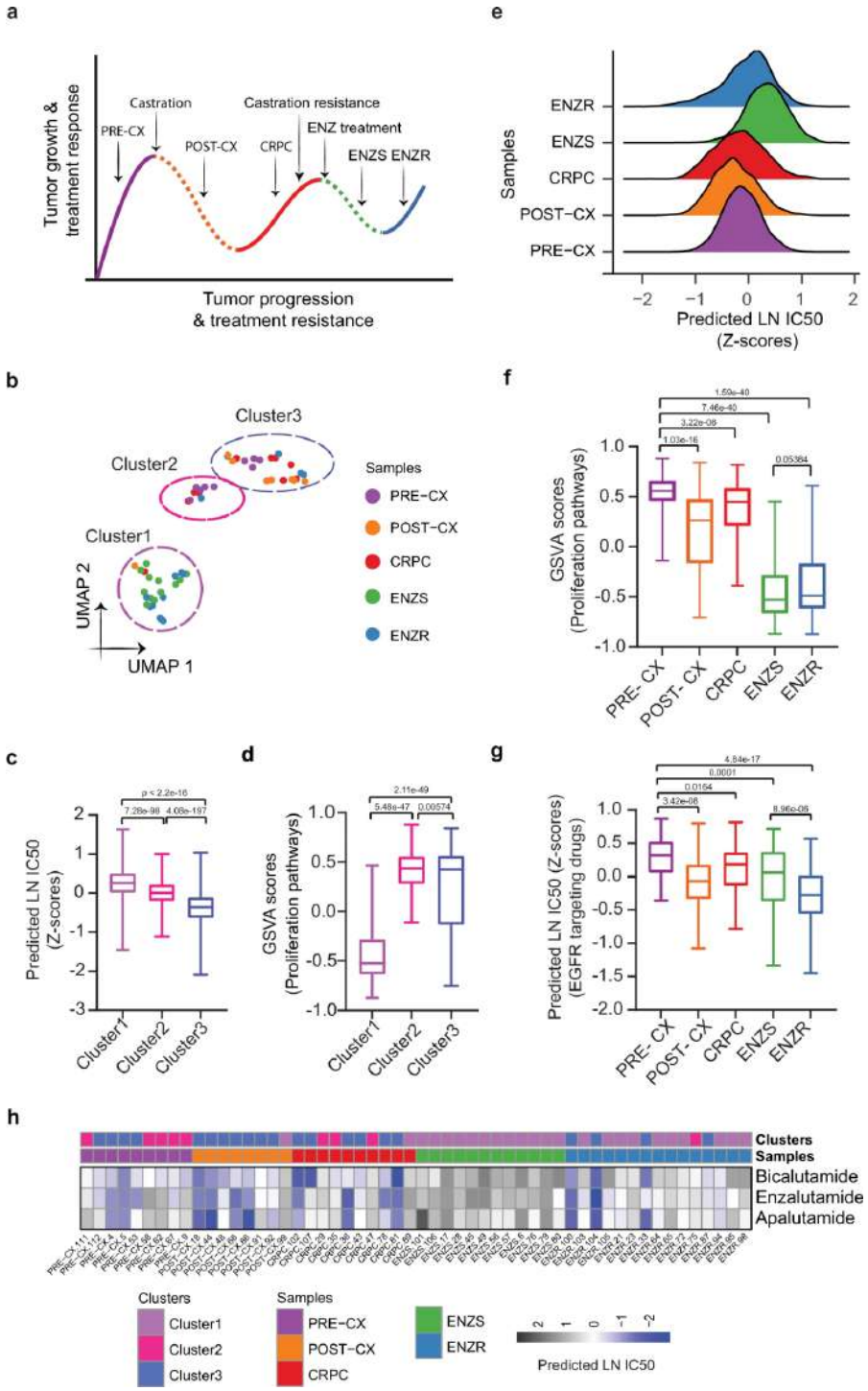

Figure 6: (a) Overall schematics of the experimental setup of LNCaP xenograft-based PCa progression study with indicated treatments, therapeutic response, and therapeutic resistance stages. Solid lines represent growth and treatment resistance; dotted lines represent treatment responsiveness. (b) Uniform Manifold Approximation and Projection (UMAP) based projections

of predicted LN IC50 showing three separate clusters. We subjected predictions to principal component analysis (PCA) and used the first 10 principal components as input for UMAP based embedding. (c) Boxplots depicting the distribution of predicted LN IC50 (Z-score) across three clusters. Each data point relates to a xenograft tumor sample-drug pair, wherein a sample belongs to one of the clusters as shown in Fig. 4b (n = 24, n = 9 and n = 21 samples from cluster 1, cluster 2 and cluster 3, respectively) and a GDSC drug (n = 155). P-values were obtained from two-sided Wilcoxon rank-sum test. (d) Boxplots showing the distribution of GSVA scores of proliferation-related pathways (n = 12) across three clusters (n = 24, n = 9 and n = 21 samples from cluster 1, cluster 2 and cluster 3, respectively). P-values were obtained from the two-sided Wilcoxon rank-sum test. (e) Ridgeplot showing the overall distribution of predicted LN IC50 (Z-score) across tumor types. (f) Boxplots showing the distribution of GSVA scores of proliferation-related pathways (n = 12) across tumor types (n = 9 PRE-CX, n = 8 POST-CX, n = 10 CRPC, n = 12 ENZS and n = 15 ENZR). P-values were obtained from the two-sided Wilcoxon rank-sum test. (g) Box plot depicting predicted LN IC50 (Z-score) of EGFR signaling pathway targeting drugs. We examined n = 7 drugs against 54 tumor bulk RNA-seq samples (n = 9 PRE-CX, n = 8 POST-CX, n = 10 CRPC, n = 12 ENZS and n = 15 ENZR), in every possible combination. P-values were obtained from a two-sided Wilcoxon rank-sum test. (h) Heatmap showing predicted LN IC50 for three unseen drugs not present in GDSC (BIC, APA, and ENZ). Color bars indicate tumor types and clusters as acquired through UMAP projections. For further details please see: https://www.nature.com/articles/s41467-022-33291-z.

I hereby declare that the information presented above is accurate and has not been previously awarded.

Sincerely,

Debarka Sengupta