Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

XevaDB enables AI-driven longitudinal pharmacogenomic modeling in patient-derived xenografts

Benjamin Haibe-Kains

benjamin.haibe-kains@uhn.ca

University Health Network

Guanqiao Feng

Princess Margaret Cancer Centre, University Health Network

Matthew Boccalon

University Health Network

Michael Tran

Princess Margaret Cancer Centre, University Health Network

Sisira Kadambat Nair

University Health Network

Samah El Ghamrasni

University Health Network

Mitchell Elliot

Princess Margaret Cancer Centre https://orcid.org/0000-0001-6711-7287

Hellen Kuasne

McGill University

Meghan McGuire

Princess Margaret Cancer Centre

Quan Li

University Health Network, Princess Margaret Cancer Centre

Jennifer Silvester

Princess Margaret Cancer Centre

Chantal Tobin

Princess Margaret Cancer Centre

Aaron Dou

Princess Margaret Cancer Centre and University of Toronto https://orcid.org/0009-0007-8293-2635

Taylor Bird

Princess Margaret Cancer Centre

Jennifer Cruickshank

University Health Network

Nikolina Radulovich

University Health Network, Princess Margaret Cancer Centre

Arvind Mer

University Of Ottawa https://orcid.org/0000-0001-8510-7963

Geneviève Deblois

University of Montreal https://orcid.org/0000-0002-3067-2807

Ming-Sound Tsao

University Health Network, Princess Margaret Cancer Centre https://orcid.org/0000-0002-9160-5405

Morag Park

McGill University https://orcid.org/0000-0001-5400-606X

Nhu-An Pham

University Health Network, Princess Margaret Cancer Centre https://orcid.org/0000-0003-3063-0292

David Cescon

Princess Margaret Cancer Centre, University Health Network https://orcid.org/0000-0002-1080-0998

Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11356792/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. D.W.C. has received consulting or advisory fees from AstraZeneca, Exact Sciences, Eisai, Gilead, GlaxoSmithKline, Inflex, Inivata/NeoGenomics, Eli Lilly, Merck, Novartis, Pfizer, Roche, and Saga. His institution has received research funding from AstraZeneca, Guardant Health, Gilead, GlaxoSmithKline, Inivata/NeoGenomics, Knight, Merck, Pfizer, ProteinQure, and Roche. The remaining authors declare no competing interests.

XevaDB enables AI-driven longitudinal pharmacogenomic modeling in patient-derived xenografts

Guanqiao Feng[1], Matthew Boccalon[1], Michael Cao-Anh Tran[1], Sisira Kadambat Nair[1], Samah El Ghamrasni[1], Mitchell J. Elliott[1,2], Hellen Kuasne[3,4], Meghan McGuire[1], Quan Li[1], Jennifer Silvester[1], Chantal Tobin[1], Aaron Dou[1], Taylor Bird[1], Jennifer Cruickshank[1], Nikolina Radulovich[1], Arvind Singh Mer[6], Genevieve Deblois[7], Ming Tsao[1,5], Morag Park[3,4,*], Nhu-An Pham[1,*], David W. Cescon[1,8,*], Benjamin Haibe-Kains[1,9,10,11,12,*]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada

[2]Division of Medical Oncology, Yale Cancer Center, New Haven, CT, United States

[3]Rosalind and Morris Goodman Cancer Institute, McGill University, Montreal, QC, Canada

[4]Departments of Oncology, Biochemistry, McGill University, Montreal, QC, Canada

[5]Department of Laboratory Medicine, University of Toronto, Toronto, ON, Canada

[6]Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, ON, Canada

[7]Institute for Research in Immunology and Cancer (IRIC), Department of Molecular Biology, Faculty of Pharmacy, Université de Montréal Montréal, Quebec, QC, Canada

[8]Division of Medical Oncology & Hematology, Department of Medicine, University Health Network, Toronto, ON, Canada

[9]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

[10]Vector Institute for Artificial Intelligence, Toronto, ON, Canada

[11]Structural Genomics Consortium, Toronto, ON, Canada

[12]Department of Computer Science, University of Toronto, Toronto, ON, Canada

Corresponding authors:

Benjamin Haibe-Kains: Benjamin.Haibe-Kains@uhn.ca; David W. Cescon: Dave.Cescon@uhn.ca; Nhu-An Pham: Nhu-An.Pham@uhn.ca; Morag Park: Morag.Park@mcgill.ca

Abstract

Patient-derived xenografts (PDXs) are critical models for translational oncology, but their computational use is limited by fragmented pharmacogenomic data, inconsistent response annotations, and poor access to raw longitudinal tumour-volume measurements. Here we present XevaDB (xevadb.ca), a standardized platform designed to support machine learning (ML) and artificial intelligence (AI) applications through the integration of longitudinal treatment and control-arm tumour growth curves with harmonized drug and gene annotations, response metrics, replicate-level experimental design, clinical annotation, and matched molecular profiles. XevaDB provides access to data from 689 PDX models across six tumour types, 69 drugs, 4,828 treatment-control experiments, and matched RNA-seq, copy-number, and mutation data for 337 models. Through four case studies, we show that preserving raw longitudinal response data enables analyses beyond conventional static endpoints: transcriptomic prediction of intrinsic tumour growth, image-based deep learning of drug-response trajectories, replicate-aware transfer learning from cell-line-trained models, and multi-omics prediction of paclitaxel response. These results establish XevaDB as a resource for sharing standardized longitudinal pharmacogenomic data and enabling reproducible AI/ML modeling, demonstrating how integrated PDX data can accelerate computational approaches in preclinical cancer research.

Introduction

Patient-derived xenografts (PDXs) preserve many histologic, genomic, and therapeutic features of the tumours from which they are derived[1–3]. They have therefore become a key model system for translational cancer research and functional precision oncology[4,5]. Large-scale PDX studies have shown that these models can reproduce clinically relevant drug-response patterns and provide experimentally tractable systems for evaluating anticancer therapies[6,7]. Community initiatives such as PDX-MI[8], PDX Finder[9], PDCM Finder[10], EurOPDX[11], PDMR[12], and PDXNet[13] have improved the standardization, discoverability, and reproducibility of PDX resources[14]. Despite these advances, the full analytical value of PDX pharmacogenomic datasets remains largely unrealized.

A major limitation is that most pharmacogenomic resources emphasize model metadata and summary response annotations, while raw longitudinal tumour-volume data are often unavailable, inconsistently formatted, or disconnected from matched molecular profiles[9–11,13]. Conventional response descriptors such as RECIST or mRECIST categories, best response, tumour growth

inhibition (TGI), and area under the tumour growth curve (AUC) summaries are valuable but incomplete: they reduce complex tumour trajectories into static labels or scalar measurements[6,14–16]. This compression can obscure potentially clinically relevant patterns such as delayed response, transient stabilization, regrowth after initial regression, and differences between intrinsic slow growth and true treatment sensitivity[2,17]. By discarding the underlying trajectories, these summaries also limit dataset harmonization, meta-analysis, and the ability to train machine learning models that leverage temporal response dynamics.

These limitations are particularly important for ML/AI analysis. Modern predictive models, including ensemble methods, convolutional neural networks, recurrent architectures, and multi-omics integration frameworks, require standardized, high-dimensional input data and carefully defined endpoints[18–23]. In pharmacogenomics, model performance is often constrained not only by algorithmic choice, but also by cohort size, endpoint definition, replicate quality, feature leakage, and cross-study heterogeneity[24–27]. Thus, the field requires resources that preserve the full temporal structure of drug-response experiments while linking these measurements to molecular and experimental data, thereby supporting reproducible analyses, cross-study data integration, and larger pooled cohorts for ML/AI model development[16,17,28].

Here we introduce XevaDB (xevadb.ca), a standardized longitudinal PDX pharmacogenomic platform that harmonizes treatment and control-arm tumour-volume curves, response metrics, replicate-level experimental annotations, clinical metadata, and matched omics profiles. We demonstrate the utility of XevaDB through four case studies that use data extracted from the platform and analyzed using custom computational workflows. First, we used control-arm tumour growth curves to model intrinsic proliferation kinetics from transcriptomic features. Second, we developed an image-based deep-learning classifier that learns drug-response patterns directly from standardized tumour-volume plots. Third, we tested cross-platform transferability of a neural network trained on cell-line pharmacogenomic data. Finally, we integrated RNA expression, copy-number variation, and mutation features to predict paclitaxel response in breast cancer PDXs. These case studies illustrate how standardized longitudinal data broaden the computational questions that can be addressed using PDX pharmacogenomic cohorts.

Results

XevaDB: a standardized pharmacogenomic resource with longitudinal tumour-response data

XevaDB was developed to address a central barrier in PDX pharmacogenomics: the lack of harmonized raw longitudinal drug-screening data linked to molecular and experimental annotations[17,28]. Dataset curation was guided by established recommendations for PDX reporting and pharmacogenomic dataset standardization, including PDX-MI and recent PDXNet consensus recommendations[8,16], with the goal of preserving detailed clinical, experimental, molecular, and drug annotation fields wherever available. The current release contains four datasets (Figure 1a) including a total of 689 PDX models spanning six tumour types, 69 therapeutic agents, 4,828 treatment-control experiments (Figure 1b). Matched omics profiles include RNA-seq for 342 models, copy-number variation for 337 models, and mutation calls for 337 models (Figure 1b, c). Breast cancer is the most abundant cancer type, with a total of 419 models (60%), among which 189 are parental models and 230 are resistant models derived through serial *in vivo* drug exposure (Figure 1b, c). Drug annotations were harmonized at the compound level by mapping standardized drug names to PubChem compound identifiers using AnnotationGx[29]. Among 69 drugs, 9 are shared in more than one dataset (Figure 1d), supporting cross-dataset and cross-cancer-type drug response validation. Replicate-level treatment information is available for all datasets, enabling users to distinguish replicated experiments from single-observation screens during downstream analysis.

Unlike resources focused primarily on model discovery or static annotations (Figure 1e), XevaDB standardizes longitudinal measurements, compound annotations, response metrics, replicate structure, and molecular profiles into a unified format (Figure 1f). Across all 4,828 experiments, treatment start, sampling time, tumour volume, treatment arm, control arm, replicate identifiers, and response metrics are linked through shared model and experiment identifiers (Figure 1b). This structure enables direct modeling of raw tumour-volume trajectories, computation of multiple response endpoints, replicate-aware filtering, and integration with molecular predictors (Figure 1a). In summary, XevaDB provides a foundation for ML/AI workflows, including growth-rate modeling and drug sensitivity classification, transfer learning, and multi-omics prediction.

Transcriptomic signatures of intrinsic PDX growth kinetics

In PDX drug-screening studies, control-arm tumour growth curves are routinely collected but are rarely analysed beyond response normalization. We hypothesized that these measurements capture intrinsic differences in tumour growth kinetics. To investigate this possibility, control-arm tumour-volume trajectories from UHN breast and lung PDX cohorts were used to estimate baseline PDX growth kinetics for each model (Figure 2a). Tumour volume was modelled as an exponential function of time during the first 35 days. Doubling time was derived from the fitted exponential growth rate (Figure 2b; Methods). Models with insufficient follow-up or poor exponential fit were excluded, yielding doubling time estimates for 96 breast and 36 lung models.

Of these models, 75 breast and 27 lung PDXs had matched RNA profile data available for downstream analysis. Using the upper and lower tertiles of doubling time to define slow- and fast-growing classes (Figure 2c), we trained ElasticNet, Random Forest and Lassoed Forest classifiers using baseline transcriptomic profiles, with nested cross-validation for hyperparameter selection (Methods). ElasticNet showed the highest performance in the breast cohort (AUROC = 0.78 ± 0.17), whereas Lassoed Forest performed best in the lung cohort (AUROC = 0.79 ± 0.28; Figure 2d).

Gene-set enrichment analysis identified five Hallmark pathways with concordant enrichment across the breast and lung cohorts. Fast-growing tumours were consistently enriched for proliferative and growth-associated programmes, including E2F Targets, MYC Targets V1, G2M Checkpoint, and mTORC1 Signalling, as well as pathways related to cellular metabolism, including Oxidative Phosphorylation (Figure 2e). In contrast, no Hallmark gene set met the criteria for concordant enrichment in slow-growing tumours. Overall, transcriptomic correlates of growth kinetics converged on shared proliferative and metabolic programmes across cohorts. These findings demonstrate that baseline tumour growth kinetics can be inferred from control-arm measurements and linked to reproducible pathway-level transcriptomic programmes.

Image-based deep learning captures non-linear drug-response dynamics

Drug sensitivity categorization underpins biomarker discovery and predictive modeling in pharmacogenomics. However, commonly used response metrics such as mRECIST, TGI, angle, and area-based measurements capture different aspects of drug response and may lead to different interpretations of treatment activity[16,17]. In practice, interpretation of drug activity often requires manual inspection of tumour growth curves to resolve complex response patterns. We therefore asked whether deep learning could directly learn these dynamics from longitudinal tumour-volume trajectories. To address this, we developed an image-based workflow using standardized tumour-volume plots from the PDXE dataset[6] (Figure 3a).

Model-drug pairs were selected from the PDXE dataset if they included both a treatment-arm and a matched control-arm growth curve. A ground-truth dataset of 1,300 model-drug response curves (fixed duration of 0-50 days, normalized tumour volume to day 0) was independently labeled by three oncology experts into five ordinal response categories (Highly Resistant to Highly Sensitive; Supplementary Figure 2a; Methods). Labels were reconciled by majority vote; 1,267 curves (97.5%) reached rater consensus and were retained for modeling, while 33 (2.5%), mostly borderline cases between adjacent categories, had no majority agreement and were excluded (Supplementary Figure 2b). We trained a Residual Network 18 (ResNet18) convolutional neural network (CNN)[30] on standardized response-curve images after removing textual labels and annotations, ensuring that the model learned from trajectory shape rather than explicit metadata (Figure 3b; Supplementary Figure 2c; Methods). Performance was compared with a Random Forest model trained on 13 drug-sensitivity matrices computed using the same 50-day maximum-duration cutoff as the response-curve images. Both models were evaluated with identical cross-validation fold assignments, ensuring a matched, leakage-free comparison (Methods).

The image-based ResNet18 model achieved a macro-averaged F1-score of 0.843 ± 0.016 (mean ± SD across folds), outperforming the Random Forest model (F1 = 0.798 ± 0.039) with both a higher mean and lower fold-to-fold variance (Figure 3b; Supplementary Figure 2d). ResNet18's advantage was consistent across all five response categories, with the largest gains in the Highly Sensitive category (Figure 3c). To improve model interpretability and identify the response metrics driving classification, we applied SHAP analysis to the Random Forest model. The treatment-control divergence angle emerged as the most influential feature overall, particularly for the Highly

Resistant, Resistant, and Intermediate categories, whereas slope.treatment and best.response contributed most strongly to the Sensitive and Highly Sensitive categories, respectively (Supplementary Figure 2e). Gradient-weighted class activation mapping (Grad-CAM) confirmed that ResNet18's predictions were driven primarily by the treatment-arm trajectory (Figure 3d). For the Sensitive and Highly Sensitive categories specifically, attention concentrated where the treatment curve approached complete tumour disappearance, consistent with this threshold distinguishing partial shrinkage (Sensitive) from near-complete response (Highly Sensitive).

Together, these results indicate that raw curve geometry contains predictive information not fully captured by scalar summaries, and support the use of computer-vision approaches for pharmacogenomic response classification, demonstrating how longitudinal formatting enables new modeling strategies for PDX drug-screening data.

Cross-platform validation of cell-line-trained neural networks

Cross-platform generalization remains a major challenge in predictive oncology. Models trained on cancer cell-line pharmacogenomic resources often fail to transfer robustly to PDX cohorts because of differences in experimental systems, tumour microenvironment, endpoint definitions, and data quality[24,31]. To investigate whether predictive signals learned in cell-line models generalize to PDXs, we tested NeST-VNN, a visible neural network trained on cell-line drug-response data[32,33], for prediction of platinum response in PDX cohorts.

We evaluated NeST-VNN across two breast cancer PDX cohorts: a McGill cisplatin cohort (n = 29) and a UHN carboplatin cohort (n = 68), each with matched omics and drug-response data (Figure 4a). The McGill cohort consisted entirely of single-replicate experiments, whereas the UHN cohort (UHN complete) included both single- and multi-replicate studies (Figure 4b). To assess whether experimental replication influences model performance, we further constructed a subset (UHN strict) of the UHN cohort containing only multi-replicate experiments. NeST-VNN uses mutation and copy-number profiles from a predefined panel of 718 genes as input[32,33]. CNV coverage of this panel was nearly complete in both cohorts (99.0% in UHN and 99.9% in McGill), whereas mutation coverage was substantially lower in McGill (32.7%) than in UHN (93.3%) (Figure 4c; Supplementary Figure 3a; Supplementary Table 8). These differences likely reflect

cohort-specific sequencing and variant-calling pipelines. To provide a consistent endpoint across experiments with varying replication designs, response was defined using treatment-arm slope over a capped 28-day response window (Methods; Supplementary Figure 3b) and dichotomized at 0° into responder and non-responder classes (Methods; Supplementary Figure 3e).

Using this response definition, NeST-VNN demonstrated predictive performance only in the UHN strict cohort (AUROC = 0.84, 95% CI 0.63-0.98), whereas performance was indistinguishable from chance in the McGill cisplatin cohort (AUROC = 0.49, 95% bootstrap CI 0.26-0.70) and the UHN complete cohort (AUROC = 0.58, 95% CI 0.37-0.77) (Figure 4d; Supplementary Figure 3f; Supplementary Table 5).

Because the response window, response definition and cohort-selection strategy are analyst-defined choices rather than fixed properties of the data, we assessed the robustness of this finding to alternative parameterizations. Recomputing the response label using the uncapped (full treatment-duration) treatment-arm slope left the McGill and UHN complete cohorts essentially unchanged (AUROC = 0.46 and 0.59, respectively) but reduced the UHN strict AUROC to 0.76 (Figure 4d; Supplementary Table 6). Similar trends were observed when response thresholds were defined using cohort-specific slope percentiles rather than the fixed 0° cutoff, with UHN strict outperforming UHN complete in 8 of the 9 percentile cutoffs tested (Figure 4e; Supplementary Table 7).

Together, these results suggest that replicate-level data quality may influence the apparent transferability of cell-line-trained models to PDX cohorts. Although the UHN strict cohort consistently outperformed the complete cohort across response-window definitions and most threshold percentiles, the magnitude of this improvement varied substantially across analytic choices, and the analyses were not corrected for multiple comparisons. Notably, NeST-VNN failed to show significant predictive performance even in the McGill cisplatin cohort, possibly due to the limited mutation coverage. By preserving replicate-level metadata and harmonized response endpoints, XevaDB enables systematic evaluation of the conditions under which external predictive models can be reliably applied to PDX datasets.

Multi-omics integration for paclitaxel response prediction

XevaDB integrates longitudinal drug-response measurements with matched RNA-seq, copy-number and mutation profiles, providing a framework for evaluating multi-omics predictors of therapeutic response. As the predictive value of individual omics layers and their integration may vary across drugs and disease contexts, we used paclitaxel-treated breast cancer PDX models as an illustrative example.

We assembled a cohort of 80 paclitaxel-treated breast cancer PDX models from the UHN collection and integrated response measurements with available RNA-seq, copy-number, and mutation data (Figure 5a). Response was defined using the treatment-arm growth-curve slope, classifying models with net tumour regression as sensitive and those showing net growth as resistant (Methods). Among these models, 67 had RNA-seq data and 62 had matched RNA-seq, copy-number, and mutation profiles for multi-omics analyses. This cohort was predominantly composed of triple-negative breast cancer (TNBC) models (48/67; Supplementary Table 14).

Among individual omics modalities, RNA expression provided the strongest predictive signal (Figure 5b). In the matched 62-model cohort, the RNA-only random forest achieved an AUROC of 0.711 (95% CI, 0.671–0.751), outperforming copy-number (random forest AUROC = 0.449; logistic L2 regression AUROC = 0.482) and mutation profiles (random forest AUROC = 0.410; logistic L2 regression AUROC = 0.359). Integrating RNA expression with copy-number and mutation data did not improve predictive performance, with early-fusion and late-fusion models achieving AUROCs of 0.705 (95% CI, 0.661–0.748) and 0.641 (95% CI, 0.578–0.705), respectively. These results indicate that transcriptomic features captured the dominant predictive signal for paclitaxel response in this cohort.

To identify biological programmes associated with response, we performed gene-set enrichment analysis of baseline RNA expression (Figure 5c). Sensitive PDXs were enriched for proliferative and mitotic programmes, including E2F Targets and G2M Checkpoint. The enrichment of cell-cycle and mitotic programmes in sensitive models is consistent with paclitaxel's mechanism as a microtubule-stabilizing agent that disrupts mitosis[34–36]. In contrast, resistant PDXs showed enrichment for inflammatory and interferon-related pathways, including IL-6/JAK/STAT3 Signaling and Interferon Alpha Response. Similar inflammatory, interferon, and STAT1-

associated transcriptional states have been reported in paclitaxel-resistant breast cancer models and have been implicated in mechanisms of therapeutic resistance[37,38]. However, whether these associations extend to clinical cohorts remains unclear. To assess the clinical relevance of the response-associated programmes identified in PDXs, we compared selected Hallmark pathways with predefined transcriptomic signatures evaluated in the I-SPY2 neoadjuvant breast cancer trial[39]. Proliferation-associated signatures showed substantial overlap with the corresponding PDX Hallmark programmes and were positively associated with pathological complete response, supporting the direction of the proliferation-associated sensitivity observed in PDXs (Figure 5d). Interestingly, Interferon Alpha Response signalling showed no clear association in clinical response[39], despite enrichment in the PDX resistant models. These findings suggest that proliferation-associated paclitaxel sensitivity is consistent across PDX and clinical settings, whereas interferon-related associations may differ between tumour-intrinsic PDX biology and the immune-competent patient microenvironment, potentially reflecting the lack of a functional adaptive immune system in immunodeficient mouse hosts.

Together, these results demonstrate how XevaDB can be used to evaluate multi-omics predictors of therapeutic response and to connect predictive modelling with biological interpretation. In this example, transcriptomic features provided the strongest predictive signal for paclitaxel response, while pathway analysis highlighted proliferation-associated sensitivity that was directionally supported by an independent clinical cohort.

Discussion

XevaDB addresses a key gap in translational pharmacogenomics: the lack of standardized PDX datasets that preserve raw longitudinal tumour-volume data. Existing PDX resources have been essential for model discovery, metadata standardization, and community coordination[8–11,13,14,16]. XevaDB builds on these efforts by emphasizing the computational structure required for modern predictive modeling: harmonized treatment and control-arm curves, matched omics, replicate-level metadata, and multiple response metrics within a unified data model (Figure 1f).

The four case studies illustrate the contribution of raw longitudinal data. First, control-arm curves contain biological information about intrinsic growth state, enabling transcriptomic modeling of tumour proliferation. Second, image-based deep learning can extract response patterns directly

from curve morphology, capturing non-linear dynamics that may be missed by scalar endpoints. Third, raw longitudinal curves enable replicate-aware filtering, data-derived response-window selection, and flexible choice of response metric, supporting more rigorous and adaptable cohort comparisons than fixed summary endpoints. Fourth, multi-omics paclitaxel modeling shows both the potential and the constraints of molecular integration in realistic PDX cohort sizes.

We showed that improved data structure can expand the modeling space. Without raw longitudinal measurements, analyses are largely restricted to static labels or precomputed summaries. With standardized trajectories, investigators can model growth kinetics, derive alternative response metrics, evaluate time-dependent endpoints, perform image-based classification, benchmark scalar versus trajectory-based predictors, and test transfer learning under controlled inclusion criteria. Thus, the value of XevaDB is not only the standardized aggregation of data, but the preservation of experimental structure needed for reproducible computational oncology.

Our results also highlight important limitations. Predictive performance in PDX cohorts remains constrained by sample size, missing omics layers, heterogeneous treatment designs, and variable replicate availability. In some cases, molecular profiles and drug-response experiments were not generated from the same passage, introducing uncertainty due to potential molecular evolution across serial transplantation. In addition, replicate availability and treatment representation may reflect historical study-design priorities, including the preferential expansion of initially drug-sensitive models for acquired-resistance investigations, potentially introducing systematic biases into downstream analyses. The modest predictive performance observed in the paclitaxel multi-omics case study should not be interpreted as a failure of multi-omics modeling, but rather as a realistic benchmark under current PDX data constraints. More complex models are not inherently superior, particularly when training cohorts are small relative to feature dimensionality. These observations align with broader recommendations for rigorous ML/AI validation in biology and predictive oncology[26,27,31,40,41].

In future releases, XevaDB will expand tumour-type coverage, incorporate additional molecular modalities, and support richer longitudinal modeling, including treatment-regrowth dynamics and uncertainty-aware replicate aggregation. Integration with pathology images, spatial profiling, and humanized PDX models may further broaden the platform's utility. The standardized data model also facilitates the integration of future PDX pharmacogenomic datasets from diverse sources, enabling XevaDB to evolve as a continuously expanding resource for preclinical pharmacogenomics. By making longitudinal pharmacogenomic data more accessible,

standardized, and computationally tractable, XevaDB provides a foundation for reproducible ML/AI research in preclinical oncology.

Methods

Data retrieval

XevaDB (xevadb.ca) provides interactive exploration and visualization of PDX pharmacogenomic data, while ORCESTRA[42] (orcestra.ca) distributes versioned XevaSet objects and associated metadata. XevaSet objects corresponding to the four cohorts used in this study, UHN breast[43], UHN lung[44], McGill breast[45], and PDXE pan-cancer[46], were downloaded from ORCESTRA (orcestra.ca/xevaset). Each data type was extracted from the corresponding XevaSet using the Xeva[28] package and exported in CSV/Excel format for downstream analyses. For case study 2, tumour growth curve images were generated using Xeva package and stored in WebP format.

Case study 1: Transcriptomic modeling of intrinsic PDX growth kinetics

PDX cohorts and control-arm data. Breast and lung PDX models and their associated control-arm (untreated) tumour-volume measurements were obtained from XevaDB. For each PDX model, longitudinal tumour-volume measurements from replicate control-arm mice were used to estimate an intrinsic growth rate independent of any treatment effect.

Estimation of intrinsic tumour doubling time. For each control-arm mouse, tumour-volume measurements within the first 35 days of the experiment, following study enrollment (Day 0), defined as the time at which tumours reached the protocol-specified size threshold (approximately 150 mm3) and mice were assigned to treatment or control arms. tumour volume was assumed to follow exponential growth, $\ln(Vt) = \ln(V_0) + kt$, and the growth-rate coefficient k was estimated by ordinary least-squares regression of ln(volume) on time. Doubling time was calculated as ln(2)/k (Figure 2b). A mouse-level fit was retained only if it had at least 3 time points, a positive growth-rate coefficient, and a goodness-of-fit $R2 \geq 0.8$. For PDX models with more than one retained mouse-level fit, growth-rate coefficients (k) were averaged across mice, and model-level doubling

time was calculated as ln(2) divided by this mean k. The 35-day window was chosen because follow-up duration in this dataset is confounded with growth rate (fast-growing tumours are censored earlier by ethical size limits), so extending the fit window is not a neutral choice (Figure 1a). Sensitivity analyses varying the minimum time-point and goodness-of-fit thresholds independently, and characterizing the follow-up-duration/growth-rate confound, are reported in Supplementary Figure 1a and 1b, respectively.

RNA-seq processing. Bulk RNA expression matrices (TPM for breast cohort; log2 microarray intensity for lung cohort) were obtained from XevaDB for each cohort. Expression matrices were restricted to protein-coding genes and to PDX models with matched doubling-time estimates. For the breast cohort, genes were retained if expression was ≥ 1 TPM in at least 20% samples, and then log2(x + 1)-transformed. Genes with near-zero variance (variance ≤ 1×10^{-8}) after transformation were removed. For the lung cohort, genes with lowest 25% mean log2-intensity or near-zero variance (variance ≤ 1×10^{-8}) after transformation were removed. This yielded 75 breast and 27 lung PDX models with matched RNA profiles and doubling-time estimates (Figure 2c), out of 96 and 36 total evaluable control-arm models, respectively.

Definition of fast- and slow-growing classes. Doubling time was log(1 + x)-transformed prior to all downstream modeling. Within each outer cross-validation training fold (see below), the 33rd and 67th percentiles of log-transformed doubling time were computed from the training fold only and used to define fast-growing (≤ 33rd percentile) and slow-growing (≥ 67th percentile) extreme groups; both the training and held-out test fold were then restricted to samples falling into these training-derived extreme groups, and this training-only cutoff was used to binarize both the training and test labels (slow growth = 1, fast growth = 0). Models in the middle third were excluded from that fold's classification task. Cutoffs were re-derived independently in every fold from training data only.

Fold-safe feature selection. Within each training fold, highly variable genes (HVGs) were selected as the N genes with the highest variance in that fold's training-extreme expression matrix (computed on the training fold only). N was fixed per cohort (100 for breast and 600 for lung) and shared identically across all three classifiers described below; it was not tuned as part of any model's hyperparameter search. This choice was empirically justified by comparing classifier

performance across a grid of candidate values, holding all other hyperparameters fixed (Supplementary Figure 1c and d).

Cross-validation design. Model performance was assessed using repeated k-fold cross-validation (RepeatedKFold, scikit-learn v1.6.1), with 5 splits × 10 repeats for breast cohort (50 outer folds) and 3 splits × 10 repeats for lung cohort (30 outer folds; random_state = 42). An outer fold was excluded if only one class was present among test-extreme samples. All 50 breast and 30 lung outer folds were successfully evaluated. For hyperparameter selection within each completed outer fold, an inner stratified 4-fold split (StratifiedKFold; shuffled, random_state = 42) was performed using the training-extreme samples only. Three classifiers were trained independently within each completed outer fold, all using the same fold-specific HVG set and the same outer training/test split.

Elastic Net. Features were standardized (zero mean, unit variance) using a scaler fit on the training fold only. A logistic regression classifier with elastic-net penalty was fit using LogisticRegressionCV (solver = "saga", scoring = "roc_auc", max_iter = 20,000), with C and l1_ratio jointly selected via the inner cross-validation splits from a grid of l1_ratio ∈ {0.1, 0.5, 0.9} and C ∈ {0.01, 0.1, 1, 10}.

Random Forest. A random forest classifier (RandomForestClassifier; 500 trees, max_features = "sqrt") was trained with max_depth and min_samples_leaf selected via a nested grid search over max_depth ∈ {3, 6} × min_samples_leaf ∈ {1, 2}, choosing the combination with the highest mean AUROC across the inner cross-validation splits (fit and scored using only the outer fold's training-extreme data).

Lassoed Forest. A random forest (same architecture and hyperparameter grid as above) was first fit on the training-extreme data; its own max_depth/min_samples_leaf were selected independently via the same nested inner-CV procedure, but scored on the post-lasso AUROC of the pipeline described below (using a fixed regularization strength, C = 1, for both the ridge and lasso steps during this hyperparameter search, to avoid a third level of nested cross-validation on already-small fold sizes). Each of the forest's 500 trees' predicted positive-class probability was

used as a feature, producing a per-sample "tree design matrix." After standardizing this matrix, an initial ridge-penalized logistic regression (LogisticRegressionCV, penalty = "l2", solver = "lbfgs", scoring = "roc_auc") was fit with C selected via inner cross-validation from C ∈ {0.001, 0.01, 0.1, 1, 10, 100}. Adaptive-lasso weights were computed as $1/(|\beta_{ridge}| + \varepsilon)^{\gamma}$ ($\varepsilon = 10\text{-}3$, $\gamma = 1$), and a weighted L1-penalized logistic regression (LogisticRegressionCV, penalty = "l1", solver = "saga", scoring = "roc_auc") was fit on the reweighted tree-probability matrix, again with C selected via the same inner cross-validation and grid, yielding a sparse, adaptive-lasso-selected subset of trees for the final probability prediction.

Model evaluation. For each completed outer fold and each classifier, predicted probabilities for the held-out test-extreme samples were evaluated against the binary fast/slow labels using the area under the receiver-operating-characteristic curve (AUROC), and against the continuous log1p(doubling time) of the same test-extreme samples using Spearman and Pearson correlation. Per-classifier performance is reported as the mean ± population standard deviation (ddof = 0) across all completed outer folds for each cohort (Supplementary Table 1).

Gene-set enrichment analysis. GSEA was performed independently of the classifiers above, using the full protein-coding expression matrix for each cohort. Fast- and slow-growth groups were defined globally using the 33rd and 67th percentiles of log1p(doubling time) across all RNA-matched models in that cohort. For every gene, a ranking statistic was computed as the limma moderated t-statistics for slow-growth versus fast-growth mean expression[47]. Genes were ranked by this statistic and analyzed using pre-ranked GSEA implemented in GSEApy (v1.1.11; 5,000 permutations, minimum gene-set size = 15, maximum gene-set size = 500, random_state = 42) against the MSigDB Hallmark (2020) gene-set collection. Cross-cohort concordance was assessed by matching gene sets tested in both cohorts within each collection; a gene set was considered concordantly enriched if it showed the same sign of normalized enrichment score (NES) and FDR $q < 0.05$ in both breast and lung (Supplementary Table 2).

Case study 2: Deep-learning classification of drug-response trajectories

PDX dataset and case selection. tumour-volume response data were obtained from XevaDB (the PDXE pan-cancer dataset[6]). Model-drug batches were eligible for inclusion if the batch had

a defined control arm (a non-missing control field in the Xeva experimental design) and the corresponding treatment model had a non-missing mRECIST classification (CR, PR, SD, or PD). Because complete and partial responses are comparatively rare in PDXE, all eligible CR and PR curves were retained; the remaining sampling budget was allocated first to stable disease (SD, up to half of the remaining slots) and then progressive disease (PD, the remainder), sampled at random with a fixed seed (seed = 1), to assemble a stratified set of 1,300 model-drug pairs. The realized mRECIST composition of the selected set was 124 CR (9.5%), 300 PR (23.1%), 438 SD (33.7%), and 438 PD (33.7%).

tumour-volume plot generation. For each selected batch, a tumour-volume plot was generated with Xeva's "plotPDX" function, displaying volume-normalized control- and treatment-arm growth curves (control: dark orange #a6611a; treatment: dark green #018571) over a fixed 50-day window, with a horizontal reference line marking complete tumour regression. Plots were rendered at 6.2 × 6 in, 120 dpi, saved as PNG, and compressed to WebP. Per-batch engineered response statistics (mRECIST call, best response and time-to-best-response, AUC, growth slope at the model level; slope, angle, AUC, and area-between-curves at the batch level; tumour growth inhibition, TGI) were exported from XevaSet alongside each image as a manifest for downstream feature-based modeling.

Ground-truth response labeling. All 1,300 curve images were independently reviewed by three raters with oncology expertise, blinded to each other's ratings, and classified into five ordinal response categories (Highly Resistant, Resistant, Intermediate, Sensitive, Highly Sensitive; Supplementary Figure 2a). Consensus labels were assigned by majority vote (≥2 of 3 raters in agreement); curves with three-way rater disagreement (n = 33, 2.5%) were excluded, leaving 1,267 consensus-labeled curves for modeling. Individual-rater agreement with the consensus label was 86.2%, 77.2%, and 84.5% for the three raters, respectively. Consensus category frequencies were 305 (Highly Resistant), 359 (Resistant), 299 (Intermediate), 213 (Sensitive), and 91 (Highly Sensitive) (Supplementary Figure 2b).

Image preprocessing for deep learning. To prevent the model from learning from textual metadata rather than curve shape, the top 100 pixels of each raw plot (containing the batch title)

were cropped, yielding uniform 744 × 620-pixel images with no textual annotation. All subsequent image-based modeling used these cropped images exclusively.

Cross-validation design. All models were evaluated using stratified 5-fold cross-validation on the 1,267 consensus-labeled curves (StratifiedKFold, scikit-learn v1.8.0; shuffle = True, random_state = 42), stratified on the five-category consensus label. The fold assignment was generated once and used identically by both the image-based and feature-based models, so that both were evaluated on exactly the same held-out partitions. For each of the 5 outer folds, the held-out fold (~253–254 curves) served as the test partition and was scored exactly once; the remaining four folds (the training pool, ~1,013–1,014 curves) were used for model fitting and internal model selection, as described separately for each model below.

Image-based model (ResNet18). A ResNet18 convolutional neural network (torchvision v0.26.0, ImageNet-pretrained weights) was used, with the final fully connected layer replaced by a linear classifier mapping 512 features to 5 output classes. Images were resized to 224 × 224 pixels and normalized using standard ImageNet channel statistics. Within each outer fold's training pool, a further stratified 85/15 split (random_state = 42) separated a training set from a validation set used only for learning-rate scheduling, early stopping, and checkpoint selection. Training images were augmented with random-resized cropping (scale 0.8–1.0) and brightness/contrast jitter. Class imbalance was addressed with a class-weighted random sampler (inverse class frequency) applied to the training set and matching class weights in the cross-entropy loss. Models were optimized with Adam (weight decay = 1×10^{-4}) using layer-group-specific learning rates (1×10^{-6} for layer1/layer2, 1×10^{-5} for layer3/layer4, 1×10^{-4} for the final classifier), with learning rate reduced on validation-loss plateau (factor = 0.1, patience = 3 epochs). Training used a batch size of 32 for up to 50 epochs with early stopping (patience = 7 epochs without validation balanced-accuracy improvement); the checkpoint with the best validation balanced accuracy was reloaded and used for the single evaluation on that fold's held-out test partition (Supplementary Figure 2c). Training was performed using PyTorch (v2.11.0). All relevant random seeds (Python, NumPy, PyTorch CPU/GPU, and the data sampler) were fixed, and cuDNN determinism flags were enabled where applicable; end-to-end reproducibility of fold-level results was empirically confirmed across independent reruns.

Feature-based model (random forest). Thirteen drug sensitivity matrices were calculated using Xeva with 50 day cutoff (model-level: mRECIST call, best response, time to best response, average best response and its timing, AUC, growth slope; batch-level: control- and treatment-arm slope, divergence angle, control- and treatment-arm AUC, and area between curves). These matrices were used to train a random forest classifier (scikit-learn v1.8.0, class_weight = "balanced"). Hyperparameters (number of trees ∈ {100, 200, 300}; maximum depth ∈ {None, 10, 20}; minimum samples per split ∈ {2, 5, 10}; maximum features ∈ {$\sqrt{}$, $\log_2$}) were selected independently for each outer fold via grid search with an inner 5-fold cross-validation on that fold's training pool, scored by macro-averaged F1. For model-interpretation analyses only (feature importance and SHAP), a final random forest was fitted using the same grid-search procedure on all 1,267 consensus-labeled curves.

Performance evaluation. For each model and outer fold, macro-averaged precision, recall, and F1-score, and balanced accuracy (mean per-class recall) were computed on that fold's held-out test partition (Supplementary Table 3 and 4). Summary performance is reported as the mean ± SD across the 5 outer folds. A pooled out-of-fold confusion matrix was constructed by concatenating each fold's test-set predictions, such that every one of the 1,267 curves contributed exactly one prediction, made by the model that had not seen it during training.

Model interpretability. For the random forest, feature attributions were computed with TreeSHAP on the final all-data model; mean absolute SHAP values were computed per feature and per response category. For ResNet18, gradient-weighted class activation mapping (Grad-CAM)[7] was computed on the final convolutional block (layer4) via forward/backward hooks. For each response category, one correctly classified, genuinely held-out example was selected from the out-of-fold predictions and visualized using the checkpoint from the fold that had held it out during training, so that each explanation corresponds to a true generalization prediction rather than a training example.

Case study 3: Cross-platform validation of NeST-VNN

PDX cohorts and data extraction. Breast cancer PDX models and associated drug-response, experimental-design, and molecular-profile data were obtained from XevaDB for two independent cohorts: a McGill cohort[48] treated with cisplatin (dose: 4.0 mg/kg, frequency: twice per week, median treatment duration: 38 days, treatment interquartile range: 28-52 days) and a University Health Network (UHN) cohort treated with carboplatin (dose: 7.5 mg/kg, frequency: weekly, median treatment duration: 38 days, treatment interquartile range: 27-70 days).

Cohorts construction. For the McGill cohort, batches information is used to extract the 32 unique cisplatin-treated models (13 carboplatin-treated batches were also present but not used, owing to insufficient sample size for robust model testing). The modToBiobaseMap is used to identify the mapping between batch and omics. This yielded 29 cisplatin models with matched sensitivity and omics data. For the UHN cohort, a similar method was applied and identified 84 carboplatin batches. Restricting to models with an available CNV column yielded 68 carboplatin models.

Replicate-based cohort stratification (UHN). The number of replicates ranged from 1–7 across the 68 UHN carboplatin models (32 models with exactly one treatment-arm mouse), whereas all 29 McGill cisplatin models had exactly one (Figure 4b). We therefore defined a "complete" UHN cohort (all 68 models) and a "strict" UHN cohort restricted to models with n_replicates ≥ 2 (36 models), to evaluate whether replicate-supported response estimates altered downstream model-validation results.

Genomic feature processing. Mutation and copy-number matrices were restricted to each cohort's models and to the 718-gene panel used by NeST-VNN. Mutation calls were binarized as 1 if any of a whitelist of functional variant classifications was present at that gene (missense, nonsense, frameshift insertion/deletion, in-frame insertion/deletion, splice site, translation start site, nonstop mutation; and multiple classifications) and 0 otherwise. For CNV, amplification and deletion were defined as CN == 2 and CN == −2, respectively.

Cohort genomic comparability analyses. We assessed whether the two cohorts' molecular input features were broadly comparable prior to modeling by computing (i) the fraction of the 718 NeST-VNN panel genes with any recorded data in each cohort's raw mutation and CNV matrices,

(ii) the per-model number of nonzero mutation/CNA/CND events (two-sided Mann–Whitney U test, McGill vs. UHN complete), and (iii) a 2-component principal component analysis (scikit-learn v1.6.1) of each binarized feature matrix, fit jointly across both cohorts. CNV coverage was comparable between cohorts (UHN: 711/718 genes, 99.0%; McGill: 717/718, 99.9%), as was per-model CNA/CND event burden (CNA: U = 858.5, p = 0.317; CND: U = 1175.5, p = 0.136; both not significant; Supplementary Table 9). Mutation coverage differed substantially (UHN: 670/718, 93.3%; McGill: 235/718, 32.7%): UHN mutation and CNV calls were generated by whole-exome sequencing (WES) and McGill's by whole-genome sequencing (WGS). Per-model mutation event burden was significantly higher in UHN than McGill using the full 718-gene panel (U = 1824.5, p = 3.8×10^{-11}; Supplementary Figure 3a); restricting the comparison to only the 232 genes with recorded mutation data in both cohorts substantially weakened, but did not eliminate, this difference (U = 1348.0, p = 4.2×10^{-3}; Supplementary Figure 3a). PCA plot was used to confirm no clear batch effect observed between the cohort (Supplementary Figure 3d).

NeST-VNN input preparation and inference. For each cohort (McGill cisplatin, n = 29; UHN carboplatin complete, n = 68; UHN carboplatin strict, n = 36), binarized mutation, amplification, and deletion matrices were reoriented to model × gene format matching NeST-VNN's expected input files. A single gene index file provided by NeST-VNN was shared across all three cohorts to guarantee identical gene ordering. NeST-VNN's publicly released 5-fold pretrained cisplatin ensemble, trained on drug-response data for genomically characterized cell lines harmonized from the Cancer Therapeutics Response Portal (CTRP) and Genomics of Drug Sensitivity in Cancer (GDSC) databases, with cell-line genomic profiles from CCLE[33], was applied to each cohort's prepared input via the model's own inference script (PyTorch v1.13.1, Python 3.9.6; CPU inference, batch size 2,000), producing one continuous predicted-response score per model per fold. All downstream analyses used the mean of the five per-fold predictions as each model's predicted response.

Definition of a cross-cohort-comparable response endpoint. We used the batch-level growth-curve slope (an arctan-transformed regression slope of log-tumour-volume on time, in degrees, bounded ±90° by construction), computed identically for both cohorts from each batch's raw longitudinal tumour-volume measurements using Xeva's setResponse function restricted to a common 28 day response window (max.time = 28). This window was applied uniformly to both cohorts so that response is assessed over the same early treatment period regardless of each

model's total observed follow-up duration (McGill: 18–150 days; UHN: 8–286 days). The specific cutoff (max.time = 28) was chosen by sweeping candidate cutoffs in weekly steps from 14 to 56 days. McGill's median treatment-arm slope reaches its most negative (deepest, least regrowth-contaminated response) value at 28 days and erodes back toward non-responding at wider windows (Supplementary Figure 3c), while the uncensored (full duration) treatment-arm slope is itself confounded with each model's total observed follow-up duration in both cohorts (two-sided Spearman $\rho = -0.54$, $p = 2.4 \times 10^{-3}$ McGill; $\rho = -0.85$, $p = 1.2 \times 10^{-19}$ UHN). Control-arm slope distributions were broadly similar between cohorts (median 86.7° McGill vs. 83.8° UHN; two-sided Kolmogorov–Smirnov $D = 0.322$, $p = 2.2 \times 10^{-2}$; Supplementary Figure 3e), supporting comparable untreated growth behavior and the use of this metric as a cross-cohort baseline; treatment-arm slope distributions differed significantly ($D = 0.410$, $p = 1.4 \times 10^{-3}$; Supplementary Table 10), reflecting differential drug response rather than a measurement-scale artifact. Response was defined using a fixed, cohort-independent threshold on treatment-arm slope: net tumour regression (slope < 0°) as responding, net growth (slope ≥ 0°) as non-responding.

Model validation. For each cohort, NeST-VNN's average predicted response was evaluated against the fixed-threshold treatment-arm-slope response label using AUROC, together with a 95% percentile bootstrap confidence interval (2,000 resamples of models with replacement within each cohort, discarding any resample containing only one class; NumPy v1.26.4 Generator, seed = 0).

Window-sensitivity (robustness) analysis. Because the 28 day response window is itself a methodological choice (see "Definition of a cross-cohort-comparable response endpoint"), we recomputed the treatment-arm-slope label using the uncensored (uncapped, full observed treatment duration) slope and repeated the primary AUROC analysis for each cohort.

Threshold-sensitivity (robustness) analysis. Because the fixed-threshold definition of response is one of several reasonable choices, we repeated the AUROC analysis using a systematic sweep of cutoffs at every 10th percentile of each cohort's own slope distribution from the 10th to 90th percentile, in each case with the same bootstrap procedure as above (Supplementary Figure 3a). No correction for multiple comparisons was applied across the more than 30 threshold-, window-, and cohort-combinations evaluated across this analysis, the window-

sensitivity analysis above, and the primary fixed-cutoff analysis (fixed 0° cutoff evaluated on both the capped and uncapped slope, per-cohort median split, and the percentile sweep, each evaluated in three cohorts); given this number of comparisons, the UHN-strict result reported above should be read as suggestive rather than a robust, standalone finding.

Case study 4: Multi-omics prediction of paclitaxel response

PDX cohort and paclitaxel response phenotyping. PDX tumour volume longitudinal records and multiomics data were drawn from the extracted UHN breast cancer XevaSet. Drug sensitivity matrices were recalculated using Xeva with a 49-day cutoff. The specific cutoff (max time 49 days) was chosen by sweeping candidate cutoffs in weekly steps from 14 to 56 days. The dataset's median treatment-arm slope reached its most negative (deepest, least regrowth-contaminated response) value at 49 days and returned toward non-responsive values at wider windows (Supplementary Figure 4a). Batch-level mRECIST categories were derived from the per-mouse mRECIST calls recorded by majority vote across each batch's treatment-arm mice, with ties broken toward the worse response (PD > SD > PR > CR). The primary response label used for all downstream modeling was a fixed threshold for growth-curve slope metric: Resistant: slope ≥ 0°; Sensitive: slope < 0° (Supplementary Figure 4b).

Multi-omics data processing. RNA-seq data were processed as described in Case Study 1. Briefly, only protein-coding genes were considered. Genes with TPM ≥ 1 in at least 20% of samples were retained, expression values were log2(TPM + 1)-transformed, and genes with near-zero variance (variance ≤ 1×10^{-8}) were removed. The resulting expression matrix contained 14,020 genes across 67 models. For CNV, amplification and deletion were defined as CN == 2 and CN == −2, respectively. This yielded a matrix of 62 models × 10,265 genes. Mutation calls were binarized per gene per sample using an inclusive functional-variant whitelist (Missense_Mutation, Nonsense_Mutation, Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Splice_Site, Translation_Start_Site, Nonstop_Mutation). This yielded a matrix of 62 models × 12,630 genes, covering an identical model set to the CNV data. The intersection of all three modalities comprised 62 models and was used for all downstream multi-omics analyses.

Fold-safe feature selection. To avoid information leakage, feature selection was performed independently within each cross-validation training fold and applied unchanged to the corresponding held-out test fold. RNA features were selected by a variance prefilter (top 2,200 genes), a pairwise-correlation filter (Pearson $|r| > 0.85$, retaining the earlier-ranked gene of each correlated pair), and a final variance filter (top 2,000 genes). CNV features were selected by a positive-variance filter followed by a correlation filter ($|r| > 0.95$). Mutation features were selected by a prevalence filter (present in ≥10 training samples and ≤90% of training samples). For all three omics layers, any gene from a curated 69-gene paclitaxel-prior list (Supplementary Table 11) present in the training data was force-retained in addition to the filter-selected genes.

Cross-validation design. Model performance was assessed using repeated stratified k-fold cross-validation (RepeatedStratifiedKFold, scikit-learn v1.6.1), with 5 splits × 5 repeats (25 outer folds; random_state = 42). Hyperparameter selection within each outer fold was performed using stratified 4-fold inner cross-validation on the outer training data only, optimizing AUROC.

Logistic Regression. An L2-penalized logistic-regression classifier (solver = "liblinear", class_weight = "balanced") was trained. The regularization strength was selected using grid search ($C \in \{0.01, 0.1, 1, 10\}$) across the inner cross-validation folds, optimizing AUROC.

Random Forest. A random forest classifier (500 trees, max_features = "sqrt", class_weight = "balanced") was trained. Hyperparameters max_depth $\in \{3, 6\}$ and min_samples_leaf $\in \{1, 2\}$ were selected using grid search across the inner cross-validation folds, optimizing AUROC.

Multi-omics integration strategies. Three modeling architectures were evaluated independently. Single-omics models were trained independently on RNA, CNV, or mutation features using the paired 62-model cohort. Early-fusion models were trained by concatenating feature-selected RNA, CNV, and mutation features prior to classification. Late-fusion (stacking) models were constructed by training an omics-specific base classifier for each data layer using out-of-fold predictions from the training fold (inner 3-fold cross-validation), followed by an L2-penalized logistic-regression meta-model. A sensitivity analysis compared logistic-regression and random-forest base learners.

Model evaluation. For each completed outer fold and model configuration, predicted probabilities for the held-out test samples were evaluated using AUROC. Performance is reported as the mean AUROC across the 25 repeated cross-validation folds together with a 95% confidence interval derived from the t-distribution. Statistical significance was assessed using a one-sample t-test against chance performance (AUROC = 0.5). Because repeated cross-validation folds are not fully statistically independent, p-values should be interpreted alongside the reported confidence intervals. Benjamini-Hochberg false-discovery-rate correction was applied across all 10 model/omics/analysis combinations tested (Supplementary Table 12).

Pathway enrichment analysis. To identify biological pathways associated with response, all 14,020 profiled genes were ranked by limma moderated t-statistics contrasting Sensitive and Resistant models, and tested by pre-ranked gene set enrichment analysis (GSEApy v1.1.11) against the MSigDB Hallmark, KEGG and GO Biological Process libraries (Supplementary Figure 4c), using 5,000 permutations, a minimum/maximum gene-set size of 15/500, and a fixed random seed (42). Gene sets were considered significant at a within-collection Benjamini–Hochberg FDR < 0.05.

I-SPY2 external validation. To assess whether the Hallmark pathways associated with paclitaxel response in the PDX cohort generalized to a clinical setting, the two retained Hallmark pathways (Proliferation, defined as the median NES of the E2F Targets and G2M Checkpoint pathways, and Interferon Alpha Response) were evaluated against pathologic complete response (pCR) in the paclitaxel-monotherapy control arm of the I-SPY2 trial (I-SPY2-990 Data Resource[39]). Each PDX pathway was matched to a predefined I-SPY2 biomarker based on gene-level overlap between the pathway's GSEA leading-edge genes and the biomarker's published gene set. For each retained biomarker, logistic regression (pCR ~ standardized biomarker + hormone-receptor status) was fitted to patient-level control-arm data to estimate the odds ratio (OR) associated with a one-standard-deviation increase in biomarker score, together with its 95% confidence interval. Benjamini-Hochberg FDR correction was applied across the two reported comparisons. A PDX pathway-I-SPY2 biomarker pair was considered concordant when the direction of association in the PDX cohort (indicated by the sign of the NES) agreed with the direction of association in the clinical cohort (OR > 1 or OR < 1) (Figure 5d; Supplementary Table 13).

Statistics and reproducibility.

The statistical results for all reported tests are provided in the Supplementary Tables.

Data Availability

All datasets used in this study are publicly available through XevaDB (xevadb.ca/) and ORCESTRA platform[42] (orcestra.ca/xevaset). UHN breast cancer PDX Xevaset[43]: https://orcestra.ca/xevaset/6abe5faca5cd60d97f7278ca. UHN lung cancer PDX Xevaset[44]: orcestra.ca/xevaset/6abd09115ba06cb2f6aaab4c. PDXE pan-cancer PDX Xevaset[6,46]: orcestra.ca/xevaset/5ff4bfe7dcd1230de0416499. McGill breast cancer PDX Xevaset[45,48]: orcestra.ca/xevaset/6a19deaf6bbfd95deb76d216.

Code Availability

The complete code, input data, intermediate results, output figures, and tables used to generate the four case studies are available on GitHub (github.com/bhklab/XevaDB_case_study).

Reference

1. Blanchard, Z., Brown, E. A., Ghazaryan, A. & Welm, A. L. PDX models for functional precision oncology and discovery science. *Nat. Rev. Cancer* **25**, 153–166 (2025).
2. Bruno, R. *et al.* Progress and Opportunities to Advance Clinical Cancer Therapeutics Using Tumor Dynamic Models. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **26**, 1787–1795 (2020).
3. Woo, X. Y. *et al.* Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. *Nat. Genet.* **53**, 86–99 (2021).
4. Bruna, A. *et al.* A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. *Cell* **167**, 260-274.e22 (2016).
5. Guillen, K. P. *et al.* A human breast cancer-derived xenograft and organoid platform for drug discovery and precision oncology. *Nat. Cancer* **3**, 232–250 (2022).

6. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).

7. Savage, P. *et al.* Chemogenomic profiling of breast cancer patient-derived xenografts reveals targetable vulnerabilities for difficult-to-treat tumors. *Commun. Biol.* **3**, 310 (2020).

8. Meehan, T. F. *et al.* PDX-MI: Minimal Information for Patient-Derived Tumor Xenograft Models. *Cancer Res.* **77**, e62–e66 (2017).

9. Conte, N. *et al.* PDX Finder: A portal for patient-derived tumor xenograft model discovery. *Nucleic Acids Res.* **47**, D1073–D1079 (2019).

10. Perova, Z. *et al.* PDCM Finder: an open global research platform for patient-derived cancer models. *Nucleic Acids Res.* **51**, D1360–D1366 (2023).

11. Dudová, Z. *et al.* The EurOPDX Data Portal: an open platform for patient-derived cancer xenograft data sharing and visualization. *BMC Genomics* **23**, 156 (2022).

12. Evrard, Y. A. *et al.* Abstract 3840: The National Cancer Institute's patient-derived models repository (PDMR). *Cancer Res.* **77**, 3840 (2017).

13. Koc, S. *et al.* PDXNet portal: patient-derived Xenograft model, data, workflow and tool discovery. *NAR Cancer* **4**, zcac014 (2022).

14. Evrard, Y. A. *et al.* Systematic Establishment of Robustness and Standards in Patient-Derived Xenograft Experiments and Analysis. *Cancer Res.* **80**, 2286–2297 (2020).

15. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).

16. Meric-Bernstam, F. *et al.* Assessment of Patient-Derived Xenograft Growth and Antitumor Activity: The NCI PDXNet Consensus Recommendations. *Mol. Cancer Ther.* **23**, 924–938 (2024).

17. Forrest, W. F. *et al.* Generalized Additive Mixed Modeling of Longitudinal Tumor Growth Reduces Bias and Improves Decision Making in Translational Oncology. *Cancer Res.* **80**, 5089–5097 (2020).

18. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501–i509 (2019).

19. Kuenzi, B. M. *et al.* Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* **38**, 672-684.e6 (2020).

20. Baptista, D., Ferreira, P. G. & Rocha, M. Deep learning for drug response prediction in cancer. *Brief. Bioinform.* **22**, 360–379 (2021).

21. Partin, A. *et al.* Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images. *Front. Med.* **10**, 1058919 (2023).

22. Liu, F., Beck, S., Yang, L., Luo, H. & Zhang, K. Advancing AI for multi-omics and clinical data integration in basic and translational cancer research. *Nat. Rev. Cancer* **26**, 497–512 (2026).

23. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).

24. Xia, F. *et al.* A cross-study analysis of drug response prediction in cancer cell lines. *Brief. Bioinform.* **23**, bbab356 (2022).

25. Feizi, N. *et al.* PharmacoDB 2.0: improving scalability and transparency of in vitro pharmacogenomics analysis. *Nucleic Acids Res.* **50**, D1348–D1357 (2022).

26. Walsh, I. *et al.* DOME: recommendations for supervised machine learning validation in biology. *Nat. Methods* **18**, 1122–1127 (2021).

27. Attafi, O. A. *et al.* DOME Registry: implementing community-wide recommendations for reporting supervised machine learning in biology. *GigaScience* **13**, giae094 (2024).

28. Mer, A. S. *et al.* Integrative Pharmacogenomics Analysis of Patient-Derived Xenografts. *Cancer Res.* **79**, 4539–4550 (2019).

29. Tran, M., Joseph, J., Eeles, C. & Haibe-Kains, B. AnnotationGx: a package for building, updating and querying an annotation database for pharmacogenomic data. (2026).

30. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). doi:10.1109/CVPR.2016.90.

31. Singhal, A. *et al.* The Hallmarks of Predictive Oncology. *Cancer Discov.* **15**, 271–285 (2025).

32. Park, S. *et al.* A deep learning model of tumor cell architecture elucidates response and resistance to CDK4/6 inhibitors. *Nat. Cancer* **5**, 996–1009 (2024).

33. Zhao, X. *et al.* Cancer Mutations Converge on a Collection of Protein Assemblies to Predict Resistance to Replication Stress. *Cancer Discov.* **14**, 508–523 (2024).

34. Reeder, A. *et al.* Stress hormones reduce the efficacy of paclitaxel in triple negative breast cancer through induction of DNA damage. *Br. J. Cancer* **112**, 1461–1470 (2015).

35. Oshi, M. *et al.* E2F cell cycle pathway score as a predictive biomarker of ER+/HER2- breast cancer response to neoadjuvant chemotherapy. *J. Clin. Oncol.* **38**, e12593–e12593 (2020).

36. Datkhile, K. D. *et al.* Genetic Polymorphisms of DNA Repair Genes and their Influence on Paclitaxel based Chemotherapy Induced Toxicity Reactions in Breast Cancer Patients. *Asian Pac. J. Cancer Prev. APJCP* **25**, 4281–4292 (2024).

37. Gao, W., Sun, L., Gai, J., Cao, Y. & Zhang, S. Exploring the resistance mechanism of triple-negative breast cancer to paclitaxel through the scRNA-seq analysis. *PLOS ONE* **19**, e0297260 (2024).

38. Abouzeid, H. A., Kassem, L., Liu, X. & Abuelhana, A. Paclitaxel resistance in breast cancer: Current challenges and recent advanced therapeutic strategies. *Cancer Treat. Res. Commun.* **43**, 100918 (2025).

39. Wolf, D. M. *et al.* Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies. *Cancer Cell* **40**, 609-623.e6 (2022).

40. Moons, K. G. M. *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).

41. de Hond, A. A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Npj Digit. Med.* **5**, 2 (2022).

42. Mammoliti, A. *et al.* Orchestrating and sharing large multimodal data for transparent and reproducible research. *Nat. Commun.* **12**, 5797 (2021).

43. Feng, G. *et al.* UHN Cescon Lab Breast Cancer PDX Drug-Response Dataset (XevaSet). https://doi.org/10.5281/zenodo.23045950 (2026) doi:10.5281/zenodo.23045950.

44. Feng, G., Haibe-Kains, B., Pham, N.-A. & Tsao, M. UHN Tsao Lab Lung Cancer PDX Drug-Response Dataset (XevaSet). https://doi.org/10.5281/zenodo.22757395 (2026) doi:10.5281/zenodo.22757395.

45. Haibe-Kains, B., Tran, M., Bannon, J. & Kadambat Nair, S. McGill_Breast_PDX. https://doi.org/10.5281/zenodo.20447111 (2026) doi:10.5281/zenodo.20447111.

46. Haibe-Kains, B. Xeva_PDXE. https://doi.org/10.5281/zenodo.4302463 (2020) doi:10.5281/zenodo.4302463.

47. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

48. Savage, P. *et al.* Chemogenomic profiling of breast cancer patient-derived xenografts reveals targetable vulnerabilities for difficult-to-treat tumors. *Commun. Biol.* **3**, 310 (2020).

Acknowledgements

The authors thank members of the M.T. lab, D.W.C. lab, B.H.K. lab and Princess Margaret Cancer Center Breast Cancer Working Group for their fruitful discussions and feedback. We thank Dr. Mona Saad and Dr. Jinfeng Zou for helpful comments on the manuscript. Figs. 1A, 1F, 2A, 3A, 4A were created with BioRender.com, and publication licenses were obtained.

Author contributions

G.F., B.H.K., D.W.C., N.A.P and M.P. conceived and designed the study. M.M., C.T. and A.D. generated the patient-derived xenograft (PDX) models and performed drug screening. G.F., M.C.A.T., S.K.N., S.E.G., H.K., J.S., T.B. and J.C. curated the resulting datasets. M.B., M.C.A.T., and A.S.M. developed and maintained XevaDB. S.E.G., M.J.E., H.K., Q.L., N.R. and M.T. provided advisory input and oversight. G.F. designed and performed the case studies and data analyses, and generated the figures, tables and supplementary information. G.F. and B.H.K. wrote the manuscript. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the Canadian Cancer Society Data Transformation Grant (Grant No. 708539), the Princess Margaret Cancer Foundation, the Canadian Institutes of Health Research (CIHR), and the U.S. Department of Defense.

Competing interests

D.W.C. has received consulting or advisory fees from AstraZeneca, Exact Sciences, Eisai, Gilead, GlaxoSmithKline, Inflex, Inivata/NeoGenomics, Eli Lilly, Merck, Novartis, Pfizer, Roche, and Saga. His institution has received research funding from AstraZeneca, Guardant Health, Gilead, GlaxoSmithKline, Inivata/NeoGenomics, Knight, Merck, Pfizer, ProteinQure, and Roche. The remaining authors declare no competing interests.

Additional Information

Supplementary Figure 1-4

Supplementary Tables 1-14


a
McGill Breast
UHN Breast
UHN Lung
PDXE pan-cancer
Curation
Standardized XevaSets
XevaDB
1. Growth kinetics
Control-arm curves
RNA-seq
ML Models
Intrinsic proliferation signature
2. Image-based response classification
Standardized tumor-volume curves
CNN Model
Drug sensitivity classification
3. Cell-line to PDX transfer learning
Response metrics
Mutation CNV
Cell-line-trained Model
Transferability assessment
4. Multi-omics paclitaxel response prediction
Response metrics
RNA-seq + CNV + mutation
Fusion Model
Fusion-model benchmarking
b
Patient
PDX
Omics
Drug
Experiment
McGill breast
UHN breast
UHN lung
PDXE pan-cancer
Total
n=464
n=689
n=342/337/337
n=69
n=4,828
Parental model
Resistant model
RNAseq
Mutation
CNV
c
Cutaneous melanoma (33)
Pancreatic (42)
Colorectal (59)
Gastric (64)
Breast (419)
Lung (72)
UHN breast (n=342)
PDXE breast (n=43)
McGill breast (n=34)
UHN lung (n=36)
PDXE lung (n=36)
PDXE gastric (n=64)
PDXE colorectal (n=59)
PDXE pancreatic (n=42)
PDXE cutaneous melanoma (n=33)
d
McGill breast
UHN breast
UHN lung
PDXE pan-cancer
Paclitaxel
Gemcitabine
BGJ398
BKM120
Carboplatin
Cisplatin
Cetuximab
Erlotinib
Palbociclib
e
Model metadata
Molecular profiles
Drug response
Longitudinal tumour volume
Analysis-ready dataset
PDX Finder / PDCM Finder (retired)
EMBL-EBI / JAX
EurOPDX
European consortium
PDMR
NCI
PDXNet
NCI
XevaDB
PMCC
f
model metadata
dataset information
Omics profiles
drug annotation
XevaSet integrated object
Omics-response mapping
Longitudinal response
Replicate design
Response metrics

Figure 1. Overview of XevaDB, a standardized longitudinal PDX pharmacogenomic resource.
(**a**) Data curation and analysis pipeline. Raw longitudinal response, omics, and clinical metadata from four source cohorts (McGill Breast, UHN Breast, UHN Lung, PDXE pan-cancer) are curated into standardized XevaSets and integrated into XevaDB, which supports four downstream case studies: (1) growth-kinetics modeling from control-arm curves and RNA-seq, (2) image-based drug-response classification from standardized tumour-volume curves, (3) cell-line-to-PDX transfer learning using response metrics and mutation/CNV data, and (4) multi-omics prediction of paclitaxel response. (**b**) Summary counts across the four cohorts for patients (n=464), PDX models (n=689; parental and resistant models shown separately), matched omics profiles (RNA-seq/mutation/CNV; n=342/337/337), drugs (n=69), and treatment-control experiments (n=4,828). (**c**) Distribution of PDX models by cancer type (n=689 total across six tumour types), with each wedge group further broken down by contributing dataset (shown in the legend); slice groups are radially separated to indicate cancer type. (**d**) Drugs shared across more than one dataset (9 of 69 total drugs), shown as a dot-matrix across the four cohorts; connected dots indicate the same drug tested in multiple datasets. (**e**) Comparison of XevaDB with four existing PDX-related resources — PDX Finder/PDCM Finder (retired), EurOPDX, PDMR, and PDXNet — across five data-availability capabilities: model metadata, molecular profiles, drug response, longitudinal tumour volume, and analysis-ready dataset. Solid dots indicate full availability, semi-transparent dots indicate limited/partial availability, and the absence of a dot indicates the capability is not available. The organization or consortium responsible for each resource is shown below its name (EMBL-EBI, European Molecular Biology Laboratory–European Bioinformatics Institute; JAX, The Jackson Laboratory; NCI, National Cancer Institute; PMCC, Princess Margaret Cancer Centre). (**f**) XevaDB's internal data architecture. The XevaSet integrated object unifies model metadata, dataset information, omics profiles, drug annotation, omics-response mapping, longitudinal response, replicate design, and response metrics within a single standardized structure.

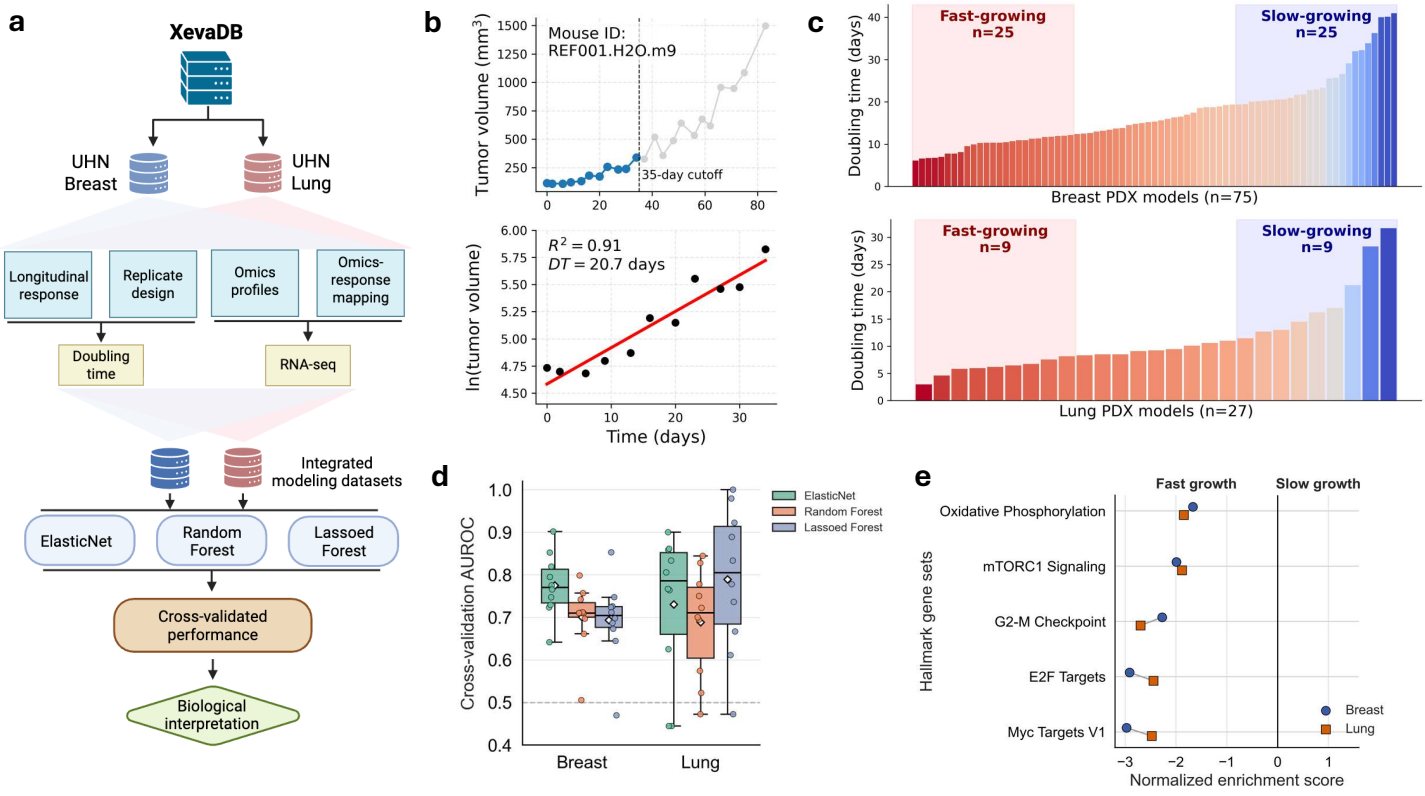


Figure 2. Transcriptomic modeling of intrinsic PDX growth kinetics.
(**a**) Case study one workflow: control-arm longitudinal tumour-volume curves and matched RNA-seq from breast and lung PDX cohorts are used to estimate model-level doubling time, modeled by ElasticNet, Random Forest, and Lassoed Forest classifiers under repeated cross-validation. (**b**) Representative doubling-time estimation for one control-arm mouse (REF001.H2O.m9): raw tumour-volume trajectory (top; points within the 35-day fitting window in blue, excluded points in gray) and the corresponding exponential-growth fit (ln(tumour volume) vs. time, bottom), with goodness-of-fit (R^2 = 0.91) and doubling time (DT = 20.7 days) annotated. (**c**) Distribution of estimated doubling times across breast (n = 75) and lung (n = 27) PDX models (integrated modeling datasets shown in a), with the fastest- and slowest-growing tertiles used to define the classification task highlighted (breast: n = 25/25; lung: n = 9/9). (**d**) Cross-validated AUROC of ElasticNet, Random Forest, and Lassoed Forest classifiers distinguishing fast- from slow-growing models, breast and lung; dashed line, chance performance. (**e**) Cross-cohort comparison of Hallmark gene-set enrichment (GSEA NES) associated with fast- vs. slow-growing tumours in the breast and lung cohorts.

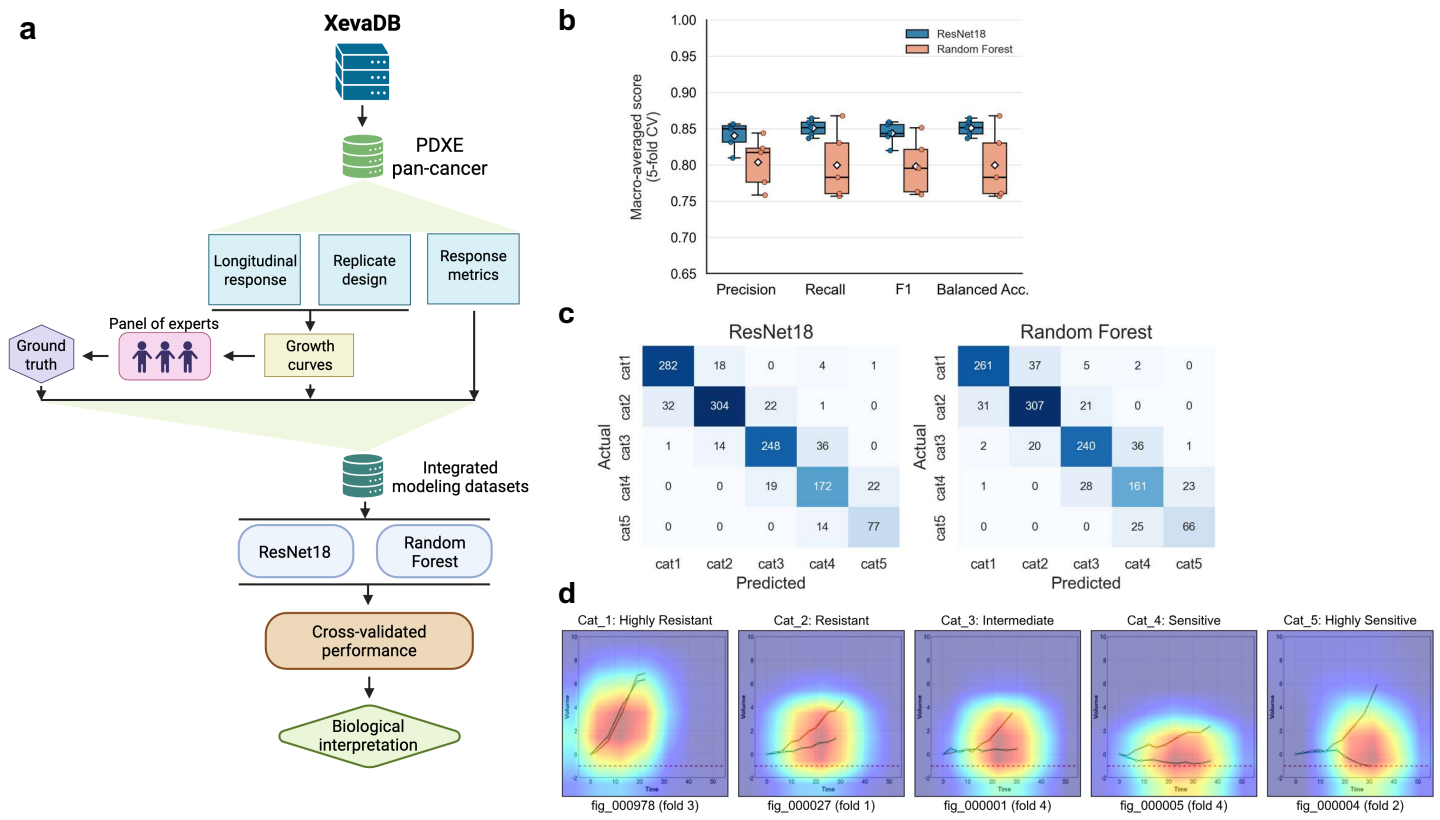


Figure 3. Image-based deep learning captures non-linear drug-response dynamics.
(**a**) Case study two workflow: tumour-volume growth curves for model-drug batches from the PDXE pan-cancer cohort are rendered as standardized images, scored by a three-rater expert panel into five ordinal response categories, and used to train and compare an image-based ResNet18 classifier against a feature-based Random Forest classifier under matched 5-fold cross-validation. (**b**) Cross-validated performance (macro-averaged precision, recall, F1-score, and balanced accuracy) of ResNet18 versus Random Forest across 5 outer folds. (**c**) Pooled out-of-fold confusion matrices for ResNet18 and Random Forest across all 1,267 curves. (**d**) Representative Grad-CAM visualizations for one correctly classified, held-out example per response category.

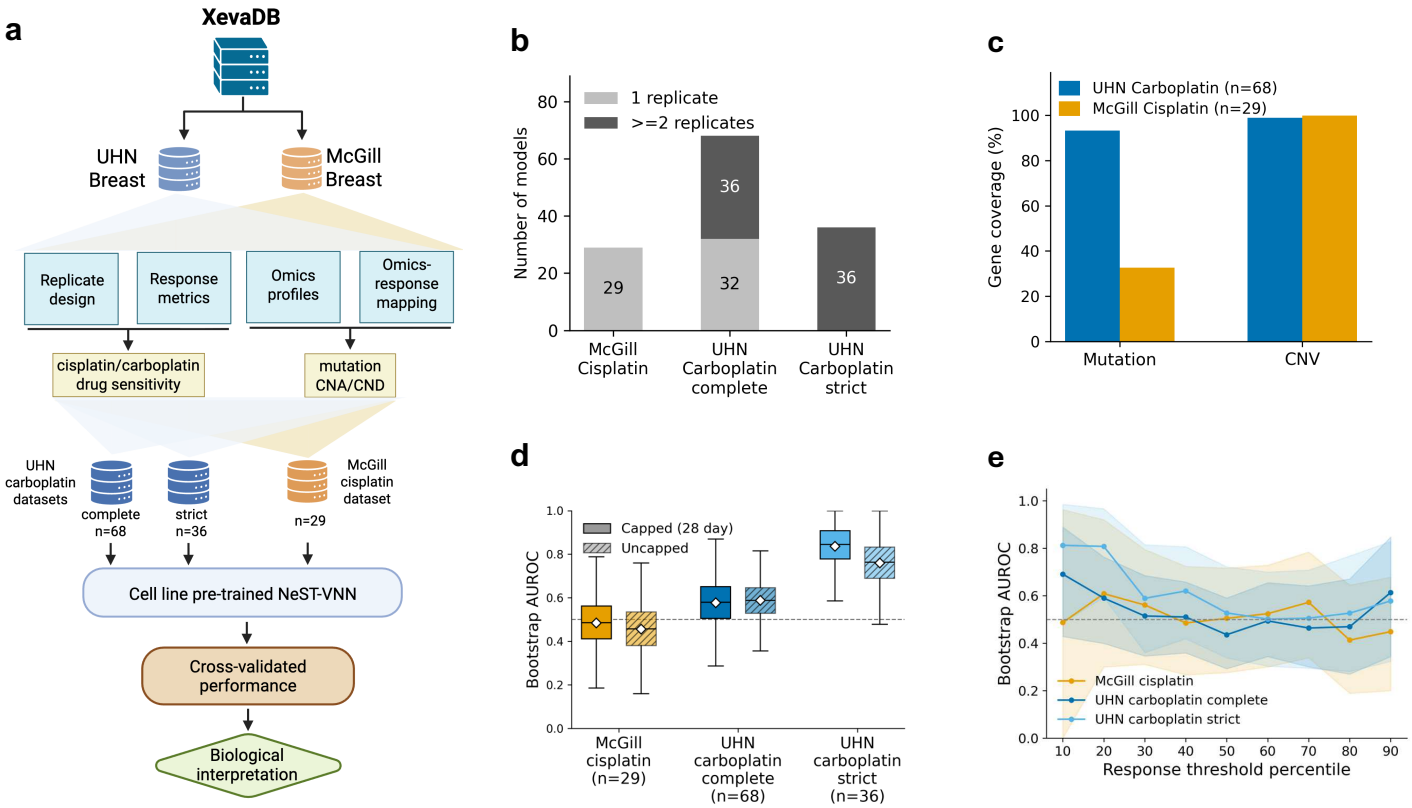


Figure 4. Cross-platform validation of NeST-VNN on independent PDX cohorts.
(**a**) Case study three workflow: drug-response, mutation, and copy-number data from the McGill (cisplatin, n = 29) and UHN (carboplatin; complete cohort n = 68, replicate-supported "strict" cohort n = 36) breast PDX cohorts are used as external input to a cell-line-pretrained NeST-VNN ensemble, validated against a growth-curve response label. (**b**) Number of models with exactly one versus two or more treatment-arm replicate mice, for each cohort/cohort definition. (**c**) Fraction of the 718-gene NeST-VNN input panel with recorded mutation and copy-number (CNV) calls in each cohort's raw molecular profiles. (**d**) Bootstrap 95% confidence intervals for NeST-VNN's AUROC under the capped (28-day window, primary) and uncapped (full observed duration) response-label definitions, for each of the three cohort definitions; dashed line, chance performance. (**e**) NeST-VNN performance across alternative response definitions. AUROC was evaluated across cohort-specific treatment-arm slope percentile cutoffs, with models below each cutoff classified as sensitive. Lines represent the McGill cisplatin, UHN carboplatin complete and UHN carboplatin strict cohorts, shaded regions indicate bootstrap 95% confidence intervals, and the dotted line indicates chance performance (AUROC = 0.5).

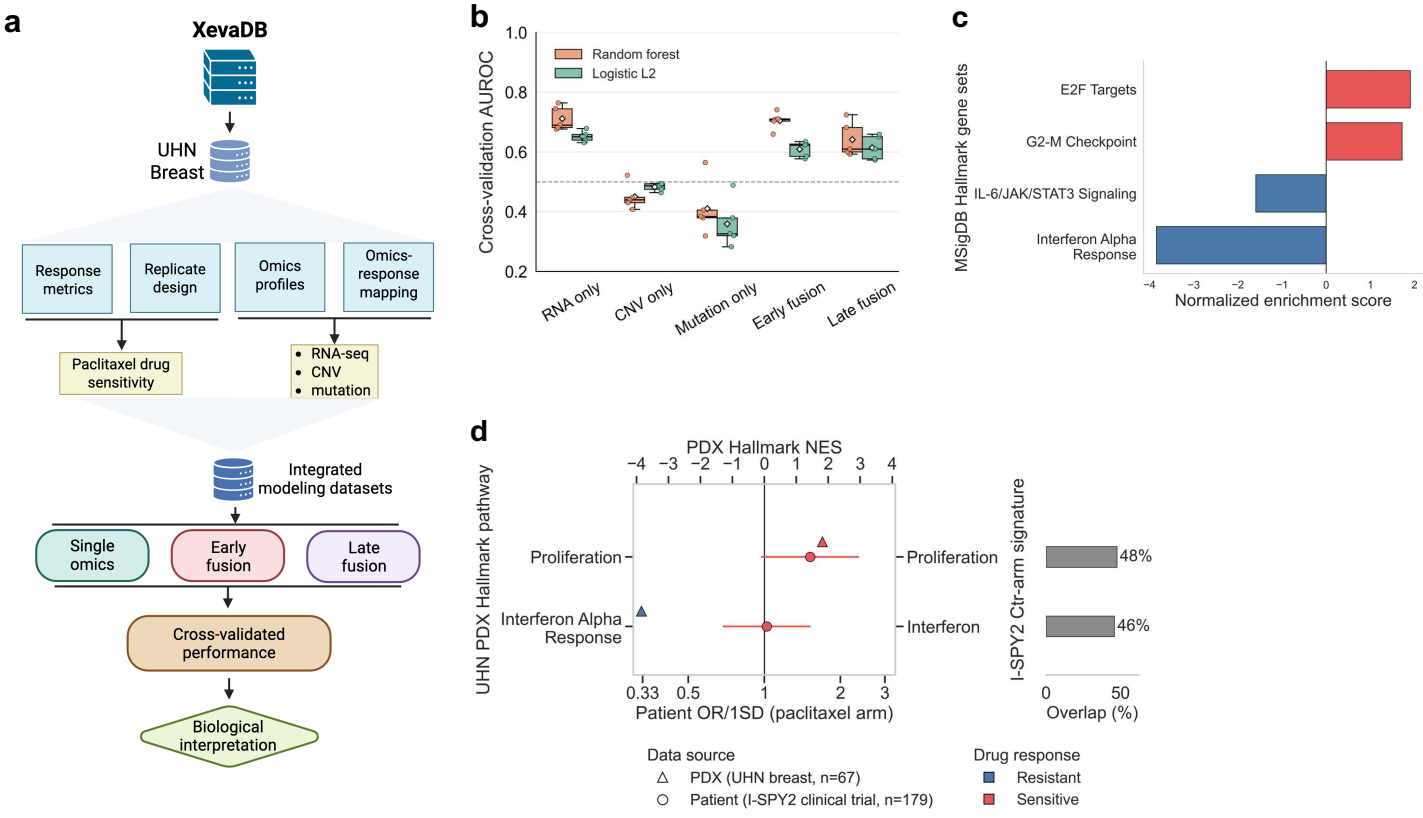


Figure 5. XevaDB-based multi-omics prediction of paclitaxel response in breast cancer PDX models. (**a**) Case study four workflow: drug-sensitivity metrics and RNA-seq/CNV/mutation profiles for the UHN breast paclitaxel cohort are extracted from XevaDB and assembled into single-omics, early-fusion, and late-fusion modeling datasets, evaluated by cross-validation. (**b**) Cross-validation AUROC of the random forest and logistic regression classifiers across single-omics (RNA-only, CNV-only, mutation-only) and multi-omics (early fusion, late fusion) feature sets; boxes show median/IQR and whiskers across five repeats, diamonds mark the mean, dashed line marks chance (AUROC = 0.5). (**c**) Pre-ranked GSEA of the full RNA transcriptome against MSigDB Hallmark gene sets; significant pathways per direction shown (FDR q < 0.05). Positive NES: enriched in Sensitive models; negative NES: enriched in Resistant models. (**d**) Cross-cohort comparison of transcriptomic programmes associated with paclitaxel response in UHN PDXs (n=67) and pathological complete response (pCR) in the I-SPY2 paclitaxel control arm (n=179 patients). Triangles, PDX normalized enrichment scores (NES) for MSigDB Hallmark pathways; circles, patient odds ratios (OR) for pCR per one-standard-deviation (1SD) increase in the corresponding predefined I-SPY2 signature, with 95% confidence intervals (refit from patient-level I-SPY2 data; see Methods). Markers are colored by their own direction of association (blue, Resistant; red, Sensitive), so a row is concordant when both markers share a color and discordant when they differ. Positive PDX NES indicates enrichment in sensitive PDXs, whereas patient OR values above 1 indicate higher odds of pCR. "Proliferation" reflects the median NES across two Hallmark gene sets significantly enriched in sensitive PDXs (E2F Targets, G2-M Checkpoint from c). Percentages indicate gene overlap between each I-SPY2 signature and its corresponding Hallmark gene set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- XevaDBsupptables.pdf
- XevaDBsuppfigures.pdf