CellNeighbor: A transcriptional atlas of patient tumors and cell line models to inform preclinical model selection

Caitlin M.A. Simopoulos, Gabrielle Persad, Otto Morris, Carlos A. Origel Marmolejo, Laura M. Richards, Kelly M. Biette; *Recursion*

Abstract #1466

1 INTRODUCTION

Preclinical drug development often relies on in vitro experiments using model cell lines selected for molecular profiles that support the program's hypothesis. This can lead to poor translatability into the clinic because: 1) extensive in vitro culturing of these cell lines can influence genetic changes, raising concerns about whether these models still represent the patient tumors from which they were derived, and 2) cell lines lack clinical data that may be relevant to inform patient population strategies.

To address this translational gap, we developed CellNeighbor, a computational framework that contextualizes cell line expression profiles within the landscape of real-world patient transcriptomic data from Tempus and TCGA[1] tumor samples (**Fig. 1**). Notably, these connected data allow better interpretation of in vitro assay results, with the addition of clinical covariates and outcomes information, and aid in objective cell model selection by prioritizing cell lines most resembling patients.

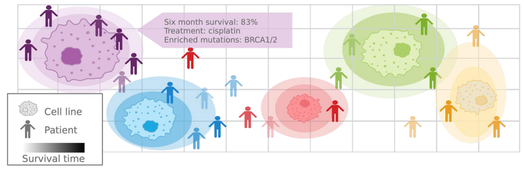


Fig 1. CellNeighbor places patient tumor samples and cell line models into the same transcriptional space. The patient neighborhoods centered around cell lines complete a holistic picture of the translational potential of in vitro models by linking them to real-world patient subpopulations and clinical covariates.

2 DATA

Model cell lines

CCLE

Baseline expression profiles of >1300 cancer cell line models. The CCLE[2,3] also includes genomic data and gene essentiality info.

Real world patient tumor data

Tempus

Tumor gene expression dataset complemented with rich clinical data useful for translational work.

TCGA

Standardized transcriptomic profiles serving as a peer-reviewed benchmark for cancer expression[1].

By integrating bulk RNA-seq across cell lines (Cancer Cell Line Encyclopedia[2,3]) and patient cohorts (Tempus, The Cancer Genome Atlas[1]), we bridge the gap between in vitro models and real-world clinical expression.

ACKNOWLEDGEMENTS AND REFERENCES

We thank Tempus for their partnership. In true OneRecursion fashion, this work would not be possible without the support of the entire Data Science & Computational Biology Team and our Value Hub Biology collaborators.

1. The results here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga
2. DepMap, Broad (2025). DepMap Public 25Q3. Dataset. depmap.org
3. Arafeh et al. *Nat Rev Cancer*. 2025
4. Warren et al. *Nat Commun*. 2021
5. Abid et al. *Nat Commun.* 2018
6. Korsunsky et al. *Nat Methods*. 2019

View poster online

Presented at 117th Annual Meeting of the American Association for Cancer Research (AACR), April 17-22, 2026; San Diego, California.

3 METHODS

A. Removing tumor microenvironment (TME) bias

An inherent challenge with combining cell line and patient tumor bulk RNA-seq samples is the influence of the TME-related gene expression from patient data. Building on Celligner's framework[4], we used contrastive PCA[5] (cPCA) to identify sources of transcriptional variation driving the differentiation of patient tumor samples from cell line profiles (**Fig. 2**). CellNeighbor explicitly identifies TME-related gene expression using pathway enrichment tests by identifying cPCs driven by genes enriched in functions related to immune activity and cell communication. These sources of TME-related variation were then regressed out of the combined dataset before dataset harmonization. The highlight of this method is the flexibility and interpretability. Notably, we can select *which* sources of variation to remove by identifying the specific cPCs enriched in functions of interest.

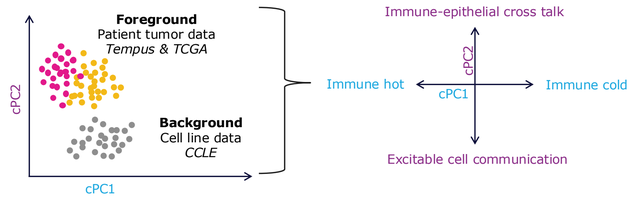
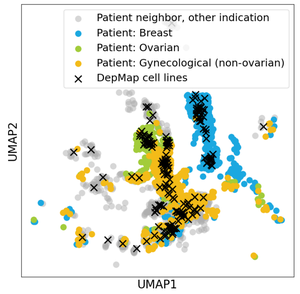


Fig 2. cPCA identifies variance unique to the foreground dataset (e.g. patient tumor data). This allows the identification of vectors of TME-related variation only found in patient tumor data that can be removed from their gene expression profiles to enable integration with in vitro cell lines models comprised of pure tumor cells.

B. Transcriptome harmonization and neighborhood mapping



After removing TME-related variation identified by cPCA, Harmony[6] was used to integrate all three cancer databases into a single transcriptional embedding space (**Fig. 3**). Within this harmonized embedding space, patient tumor neighborhoods were defined around each cell line using a symmetric nearest neighbor approach. **Each neighborhood is now a connected data source of both cell line and patient tumor molecular data.**

Fig 3. Harmonized transcriptomes visualized by UMAP projection of CellNeighbor embeddings highlighting the transcriptional alignment of CCLE cell lines (X) with their respective patient neighbors.

4 NOVEL NEIGHBORHOOD METRICS

We also developed novel metrics to quantify the confidence of cell line to patient tumor transcriptional similarity for use in objective and automated decision making. As one example, a 'homogeneity score' is used to express how representative a cell line's neighboring tumor samples are to its original tissue (**Table 1**) and is calculated by finding the proportion of patient neighbors with the same tissue of origin as the cell line. This score can be used to rank potential cell line models for in vitro experiments.

Cell line rank	Top neighbor tissue match	Homogeneity score
1	Breast	1
2	Breast	1
3	Breast	0.8
…	…	…
9	Lung	0.3
10	Ovarian	0.2

Table 1. Example theoretical ranking of breast cancer cell lines.

5 OPTIMIZING CELL LINE MODEL SELECTION USING CELLNEIGHBOR

Objective: Prioritize cell lines for in vitro assays that most accurately represent patient tumor biology

Breast cancer atlas

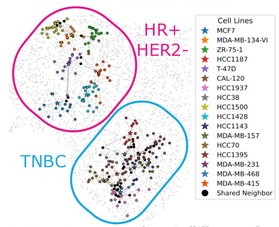


Fig 4. Breast cancer atlas: Cell line-patient tumor mapping reveals subtype-specific clustering (HR+/HER2- vs. TNBC).

Pan cancer atlas

Cell line rank	Cell Line Name	Top neighbor tissue match	Homogeneity score
1	T-47D	Breast	1
…	…	…	…
15	CAL-120	Mesothelioma	0.08
16	MDA-MB-231	Lung	0
17	MDA-MB-157	Gynecological	0

Poor translators

Table 2. Pan-cancer ranking: Identifying cell lines with low homogeneity scores and off-target tissue similarity.

Indication-Specific Mapping: A breast cancer-specific CellNeighbor atlas identified 17 candidate cell lines that clustered closely with HR+/HER2- and TNBC patient subtypes (**Fig. 4**).

Pan-Cancer Validation: These 17 lines were cross-referenced against a pan-cancer context and ranked by homogeneity score to ensure tissue-of-origin fidelity.

Identifying "Poor Translators": Analysis revealed three breast cancer cell lines (CAL-120, MDA-MB-231, and MDA-MB-157) with transcriptional profiles more like lung, mesothelioma, or gynecological cancers than their tumor site of origin, breast cancer (**Table 2**).

Outcome: By filtering out cell lines with low homogeneity scores, we prioritize models that most closely resemble the transcriptional profiles of real patient tumors, suggesting their utility in clinical translation.

6 PATIENT STRATIFICATION STRATEGY FROM IN VITRO ASSAYS

Objective: Define a patient stratification strategy for REC-XX using integrated data layers

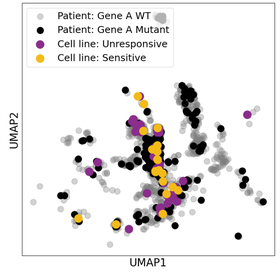


Fig 5. Sensitive cell lines cluster with patient tumors containing a pathogenic mutation in Gene A within CellNeighbor embedding space.

Data Integration: We layered CellTiter-Glo (CTG) sensitivity results onto the CellNeighbor embedding space of harmonized cell lines and patient tumors.

Neighborhood Analysis: Patient neighborhoods of sensitive cell lines were found to be enriched with samples harboring *Gene A* mutations (**Fig. 5**).

Validation: A secondary analysis confirmed that Gene A mutations are significantly associated with REC-XX sensitivity at the cell line level.

Outcome: These combined analyses support *Gene A* mutations as a potential predictive biomarker for REC-XX clinical programs with strong likelihood of clinical translatability.

7 CONCLUSIONS

CellNeighbor is a data-driven, translational-focused method developed to ensure that preclinical decisions are rooted in real-world patient biology. It offers a robust method to identify cell line models that closely resemble patient tumors, ultimately aiming to increase the translatability of preclinical discoveries into clinical applications. By adding an additional layer of clinical covariates onto in vitro model neighborhoods, this framework allows scientists to validate drug hypotheses in more clinically relevant contexts and better interpret assay results by considering real world data alongside traditional analysis methods.