Abstract: Cancer researchers test thousands of potential drugs every year, yet only 5% of those that are selected for clinical testing succeed in the patient setting[1]. This is fundamentally a prediction problem: we know that initial results in cancer cell lines don't accurately reflect their efficacy in the clinic, yet we have to use results from various imperfect laboratory models to make expensive decisions about which potential drugs to invest in translating to clinical trials[2]. Predictive validity measures the accuracy of laboratory models' results relative to the same intervention's ultimate performance in the clinic[3,4]. However, it is not routinely measured in any form of drug development and it is absent from current datasets[5–7], both for common laboratory models and for patient data. This absence is expensive, both for drug developers and for machine learning (ML) research approaches, where the lack of clear predictive validity metrics restricts best-in-class ML predictions to the type of data on which they were trained. Here we propose the development of a vertically integrated colorectal cancer (CRC) dataset that characterizes patient samples and preclinical models over time to rigorously measure the predictive validity of each model for different drug perturbations. This dataset, as a common good for academia and industry alike, will enable clearer measurement of predictive value for both wet lab and ML model results, which in turn will empower researchers to develop new types of ML models that predict efficacy across multiple models, investors to assess clinical likelihood of success more rigorously, and clinicians to match patients to treatment regimens most likely to deliver curative results.

AI Task Definition: Current ML efforts in drug development and oncology are largely limited to within-data-type predictions, such as *in vitro* drug screening results or modeling a limited set of cancer progression drivers in mathematical models. In cases where there are datasets that track predictive validity of results in preclinical models, such as acute myeloid leukemia (AML), recent ML work has successfully modeled the genetic evolution of disease, predicted specific kinds of resistance, and created personalized treatment algorithms with high clinical success[8,9]. This quality of data is, however, currently available only for AML and not for any kind of solid cancers, where rising incidence rates in CRC and the recent glut of failed immunotherapy trials has highlighted how the absence of predictive validity metrics contributes to low clinical success rates for new drug development. The proposed dataset will knit together several currently disparate efforts in CRC to introduce measures of clinical predictive validity at each stage of drug development. This will enable multi-modal transformer architectures to learn representations across different experimental contexts, meta-learning approaches that predict experimental reliability in drug development investment, and dynamic flow matching models of tumor and model evolution through treatment. This will accelerate ML in oncology drug development via creating a field-wide standard for predictive validity, better calibrate translational model selection and drug candidate investment[10], and bring the personalized treatment prediction paradigm from AML to solid cancers[2].

Dataset Rationale: Novel datasets in oncology have been instrumental in developing revolutionary new drugs, such as the pan-cancer finding of NTRK mutations in The Cancer Genome Atlas enabling the rapid and successful development of Larotrectinib[11,12]. However, current best-in-class data collection efforts, such as the Human Tumor Atlas Network (HTAN), are prioritizing data breadth over model creation or measurement of predictive validity[13]. Meanwhile, use of live/die screens in cancer cell lines and animal studies is pervasive in drug development, also without quantification of predictive validity. Without this measurement, ML-driven efforts to improve drug development, model selection, or patient personalization are restricted to within-data-type predictions with low translational potential and low impact on drug development timelines or success. Colorectal cancer (CRC) was selected as a first solid cancer to focus on measurement of clinical validity for several reasons. First, it is rising in incidence in young populations with a large current patient population and is—relative to other solid cancers—accessible for longitudinal sampling and model creation. Second, there is established precedent of treatment results in patient derived CRC organoid systems, which retain key 3D and immune characteristics of the human setting, matching those seen in patients and delivering high clinical predictive validity to measure against[14,15]. Finally, third, in immunotherapy and other CRC clinical trials, there is frequently a mix of non-responders and hyper-responders, where differential characterization and comparison is immensely valuable for new model creation, developing more predictive models for treatment response, and understanding which molecular signals in patients are highest value for ML model development, preclinical model selection, and biotech investment. Ideally, the proposed dataset will be developed in 2 complementary phases. First, in an exploratory phase, multiple techniques and models will be tested in parallel on a small n of ~20 patient sources to identify which models and biological characterization techniques deliver the greatest predictive value, both in clinical success and in ML model results. Then, in a scaling phase that will also be applicable to other solid cancers, a narrower set of models and characterization techniques will be used to generate a dataset of 500+ longitudinal profiles matching patient data with preclinical model data given the same treatments. Measurement of predictive validity across timepoints will establish new standards for measuring experimental reliability in both wet lab and ML model results, enabling every stakeholder in drug development to make better decisions based on ML predictions they can trust.

Data Creation Pathways: The development of post-genomics characterization technologies such as 3D spatial transcriptomics has revealed new features of cancer biology with direct translational relevance. Many of these features are absent from common preclinical models[10]. As a result, there are ongoing efforts to use as many characterization technologies as possible on real patient samples, such as the Human Tumor Atlas Network, which includes a CRC cohort being characterized in a federally-funded academic consortium[13]. In parallel, developing patient-derived organoid models and tracking their responses to the same treatments given to patients has been piloted at several leading academic cancer centers, demonstrating both technical and logistical feasibility[14]. From these efforts, the proposed dataset can be built via a public-private consortium of academic labs and specialist companies via highly targeted “infill” studies that strategically add predictive validity metrics to studies already underway. These studies will focus on assessing predictive validity of results in each model type relative to that seen in patients, with an initial phase designed to determine if this is best measured at the proteomic, genetic, transcriptomic, or morphological layer. Results from this will be used to create a rich open dataset with standardized data attributes that can be used for the wide variety of ML tasks outlined above. Technical, logistical, and personnel resources developed over the course of generating this data can then be applied to other solid cancer types or even chronic inflammatory conditions.

Cost/Scalability: By leveraging current HTAN data generation efforts and ongoing academic clinical trials in CRC, which can cost upwards of $120,000/patient, the most expensive aspect of the proposed dataset—primary patient data—is already addressed. Generation and characterization of the various models in Phase 1 is estimated, based on comparable academic projects, to cost ~$50,000/patient source, for a total of $1M across 20 patients. The narrower set of models and characterization in techniques for Phase 2 is estimated on the same basis to cost $15,000/patient source for a total of $7.5M. Delivering even a modest 5% improvement in drug development success will, in turn, save billions of dollars currently spent on testing low validity predictions in poorly matched patients.

1. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
2. Mateo, J. *et al.* Delivering precision oncology to patients with cancer. *Nat. Med.* **28**, 658–665 (2022).
3. Scannell, J. W. *et al.* Predictive validity in drug discovery: what it is, why it matters and how to improve it. *Nat. Rev. Drug Discov.* **21**, 915–931 (2022).
4. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. *Diagnosing the Decline in Pharmaceutical R&D Efficiency*. *nature.com* www.nature.com/reviews/drugdisc (2012) doi:10.1038/nrd3681.
5. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
6. Home page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer. https://www.cancerrxgene.org/.
7. Chiu, Y. C. *et al.* Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med. Genomics* **12**, (2019).
8. Pino, J. C. *et al.* Mapping the proteogenomic landscape enables prediction of drug response in acute myeloid leukemia. *Cell Rep. Med.* **5**, (2024).
9. Posso, J. C. *et al.* Abstract 5768: Identifying signaling changes that give rise to drug resistance in acute myeloid leukemia through multiomic modeling across diverse model systems. *Cancer Res.* **85**, 5768 (2025).
10. Honkala, A., Malhotra, S. V., Kummar, S. & Junttila, M. R. Harnessing the predictive power of preclinical models for oncology drug development. *Nat. Rev. Drug Discov.* **21**, 99–114 (2022).
11. Berger, S., Martens, U. M. & Bochum, S. Larotrectinib (LOXO-101). in *Small Molecules in Oncology* (ed. Martens, U. M.) 141–151 (Springer International Publishing, Cham, 2018). doi:10.1007/978-3-319-91442-8_10.
12. Hong, D. S. *et al.* KRASG12C Inhibition with Sotorasib in Advanced Solid Tumors. *N. Engl. J. Med.* **383**, 1207–1217 (2020).
13. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236–249 (2020).
14. Liu, L., Yu, L., Li, Z., Li, W. & Huang, W. Patient-derived organoid (PDO) platforms to facilitate clinical decision making. *J. Transl. Med.* **19**, 40 (2021).
15. Cantrell, M. A. & Kuo, C. J. Organoid modeling for cancer precision medicine. *Genome Med.* **7**, (2015).