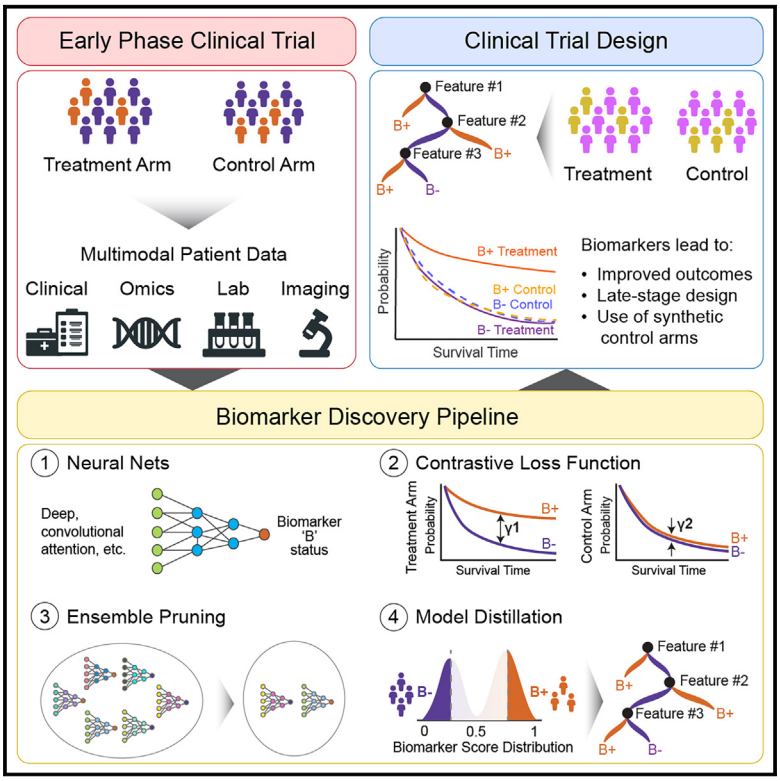
Cancer Cell

Article

AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes

Graphical abstract



Highlights

- AI-driven framework discovers predictive, rather than prognostic, biomarkers
- Framework outperforms existing approaches across real-world and clinical trial data
- Framework generates interpretable biomarkers to facilitate clinical actionability
- Retrospective improvement of patient selection for phase 3 immuno-oncology trials

Authors

Gustavo Arango-Argoty, Damian E. Bikiel, Gerald J. Sun, ..., Sebastian Carrasco Pro, Elizabeth Y. Choe, Etai Jacob

Correspondence

gustavo.arango@astrazeneca.com (G.A.-A.), etai.jacob@astrazeneca.com (E.J.)

In brief

Arango-Argoty et al. introduce the Predictive Biomarker Modeling Framework (PBMF), an AI-driven contrastive learning framework to identify putative predictive, rather than prognostic, biomarkers for clinical studies. The PBMF identifies biomarkers, particularly for hard-to-predict therapies, such as immuno-oncology treatments, that are shown to retrospectively improve treatment survival outcomes.

Arango-Argoty et al., 2025, Cancer Cell *43*, 1–16
May 12, 2025
https://doi.org/10.1016/j.ccell.2025.03.029

CellPress

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

Cancer Cell

CellPress
OPEN ACCESS

Article

AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes

Gustavo Arango-Argoty,[1,*] Damian E. Bikiel,[1] Gerald J. Sun,[1] Elly Kipkogei,[1] Kaitlin M. Smith,[1] Sebastian Carrasco Pro,[2] Elizabeth Y. Choe,[1] and Etai Jacob[1,3,*]

[1]Oncology Data Science, Oncology R&D, AstraZeneca, Waltham, MA, USA
[2]Life Sciences, Tempus AI, Boston, MA, USA
[3]Lead contact
*Correspondence: gustavo.arango@astrazeneca.com (G.A.-A.), etai.jacob@astrazeneca.com (E.J.)
https://doi.org/10.1016/j.ccell.2025.03.029

SUMMARY

Modern clinical trials can capture tens of thousands of clinicogenomic measurements per individual. Discovering predictive biomarkers, as opposed to prognostic markers, remains challenging. To address this, we present a neural network framework based on contrastive learning—the Predictive Biomarker Modeling Framework (PBMF)—that explores potential predictive biomarkers in an automated, systematic, and unbiased manner. Applied retrospectively to real clinicogenomic datasets, particularly for immuno-oncology (IO) trials, our algorithm identifies biomarkers of IO-treated individuals who survive longer than those treated with other therapies. We demonstrate how our framework retrospectively contributes to a phase 3 clinical trial by uncovering a predictive, interpretable biomarker based solely on early study data. Patients identified with this predictive biomarker show a 15% improvement in survival risk compared to those in the original trial. The PBMF offers a general-purpose, rapid, and robust approach to inform biomarker strategy, providing actionable outcomes for clinical decision-making.

INTRODUCTION

The promise of precision medicine lies in treating patients with therapies that precisely target their unique diseases.[1,2] Key to this approach are predictive biomarkers that identify individuals more likely to benefit from a specific therapy. Predictive biomarkers differ from prognostic biomarkers, as the latter relate to general disease outcomes regardless of treatment. While prognostic biomarkers provide insight into disease progression, predictive biomarkers are crucial for optimizing clinical trial design to evaluate treatment effectiveness and ensure patients receive therapies that maximize survival outcomes and quality of life. For example, breast cancer patients with HER2 overexpression experience improved progression-free and overall survival on anti–HER2 antibodies compared to those who are HER2-negative.[3,4] Similarly, the *BCR-ABL1* fusion gene is predictive of response to tyrosine kinase inhibitors in chronic myeloid leukemia.[3,5] Drug development programs integrating patient preselection biomarkers have a striking 2-fold increase in the likelihood of approval.[6] This is particularly impactful in oncology where 90% of therapies entering clinical development fail to reach market approval.[7]

Discovering predictive biomarkers is a complex and challenging endeavor due to the complex interactions between disease biology and treatments, especially for immunotherapies, which modulate the immune system rather than the tumor, and therefore lack an obvious molecular biomarker hypothesis. The advent of next-generation sequencing technologies providing large-scale profiling of gene mutations, transcript expression, and protein have both increased the opportunity to find predictive biomarkers as well as further complicated the task.[8] With this increasingly available high dimensional data, the redundancies inherent in biological systems can result in multiple features being predictive of response. Moreover, the presence of numerous prognostic factors may hinder pinpointing the predictive biomarker within the studied patient population. Traditional regression methods such as Cox proportional hazards (PH) modeling[9] have been widely employed to discover predictive biomarkers. However, these methods necessitate the explicit enumeration of covariates and interactions, a task that becomes impractical as the number of features increases, particularly in scenarios involving a diverse set of clinical and -omic features. Specifically for immuno-oncology (IO), validated predictive single biomarkers such as PD-L1 expression,[10] microsatellite instability,[11] and tumor mutation burden (TMB)[12] still imperfectly enrich for responsive patients.

Composite biomarkers (i.e., potentially nonlinear combinations of multiple clinical measurements) have recently been proposed to improve therapeutic outcome predictions for IO.[13–18] Methods have been developed to discover predictive, potentially

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

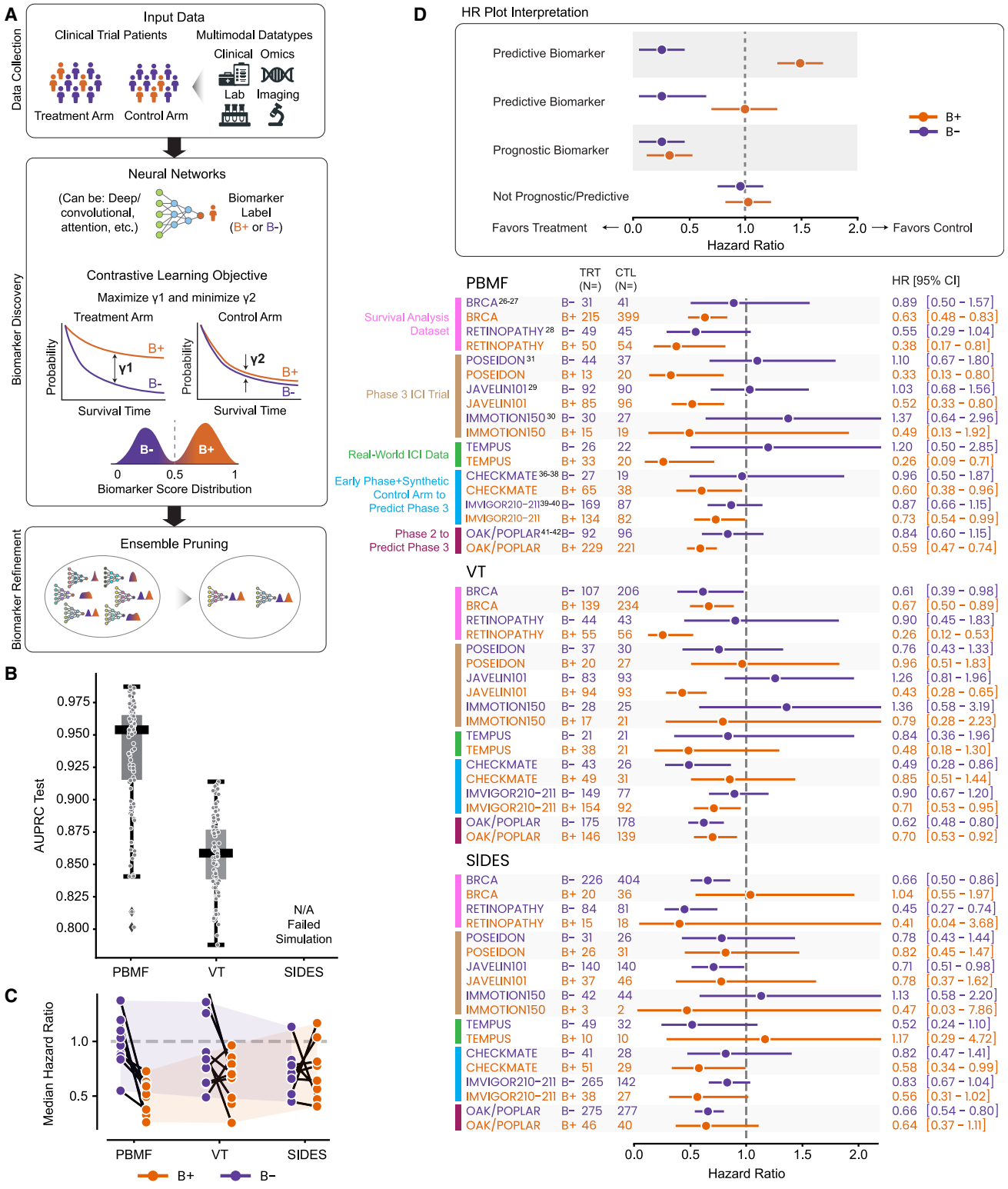


Figure 1. Schematic of the PBMF and its evaluation on simulated and real clinical datasets versus other methods

(A) The PBMF utilizes data of any modality collected for each sample from each of two treatment arms. The PBMF trains an ensemble of neural networks, each independently trained on clinical trial data with a contrastive loss function. The loss is designed to enhance the differential impact of B+ versus B– in the treatment group and concurrently minimize B+ influence over B– in the control arm. The ensemble of PBMF models is pruned to retain only those models whose performance is similar. The PBMF then outputs a consolidated single biomarker score that enriches for longer time-to-event times (e.g., survival) only on the treatment of interest.

(legend continued on next page)

composite biomarkers without requiring explicit specification of covariates and interactions. These approaches utilize algorithms designed to maximize the difference in target outcomes between subgroups with different treatments,[19,20] but still encounter challenges. For example, subgroup identification based on differential effect search (SIDES)[21] uses a decision tree to identify patient subgroups with significant differences in treatment effects. While interpretable, SIDES cannot identify composite biomarkers with nonlinear relationships. Virtual Twins (VT),[22] on the other hand, calculates treatment effects for each patient by training separate models for treatment and control outcomes. While capable of detecting nonlinear composite biomarkers, users must search for a post-hoc cutoff to determine subgroups.

To address these limitations, we present the Predictive Biomarker Modeling Framework (PBMF), a neural network-powered contrastive learning process guided by a novel training objective. Here, we provide a diverse body of empirical evidence showcasing the robust predictive biomarker discovery capability of the PBMF across various scenarios, including simulated biomarker discovery, well-established clinical datasets for survival analysis, and real-world and randomized controlled clinical trial data for various immunotherapies. Notably, the PBMF outperforms existing approaches in subgroup identification within both simulated and real datasets. We show how the PBMF may retrospectively contribute to patient selection for two phase 3 clinical trials, using only single-arm early phase trial data with synthetic control arms, leading to at least a 10% improvement in efficacy versus the original trials. Furthermore, we illustrate how the PBMF retrospectively contributes to patient selection in a phase 3 clinical trial by uncovering a predictive biomarker based solely on phase 2 trial data. This discovery leads to a 15% improvement in efficacy in the original trial, achieved through a straightforward decision tree generated via PBMF knowledge distillation.

RESULTS

A contrastive learning framework to identify predictive biomarkers

Neural networks can uncover complex, nonlinear relationships between features and support diverse data types. The PBMF employs a neural network-powered contrastive learning process guided by a novel training objective (Figure 1A). Specifically, the PBMF utilizes a ratio of treatment versus control effects that enables direct learning of treatment-specific predictive, rather than prognostic signals. The biomarker score cutoff and sample prevalence constraints are also components of the model's training objective, abrogating the need for post-hoc tuning. The framework takes an ensemble approach by training multiple neural networks to minimize overfitting. Finally, the PBMF includes an optional step to distill the neural network outputs into an interpretable decision tree, making the results clinically actionable. A publicly available web app (Zenodo: https://doi.org/10.5281/zenodo.14766044) contains a user-friendly PBMF framework and tools for generating simulated data to benchmark models. For method architecture and implementation details, see STAR Methods details.

PBMF effectively identifies predictive biomarkers across diverse clinical studies and simulations

We tested the ability of the PBMF to discover a composite predictive signal in the presence of a prognostic signal in synthetic datasets representing realistic combinations of features and time-to-event data (i.e., survival) that mirrored conditions commonly encountered in real-world scenarios. We benchmarked the PBMF against SIDES[21] and VT,[22] two established methods with publicly available software, and that are designed to identify patient subgroups more likely to respond to specific therapies (Table S1). SIDES failed to solve any of the simulated scenarios.

The synthetic data scenario comprised 3 features, 2 predictive and 1 prognostic, where the predictive signal was present only as a combination of the two predictive features (Figure S1A). The PBMF yielded an area under the precision-recall curve (AUPRC) of 0.918 ± 0.047 (mean ± standard deviation) and outperformed VT (AUPRC = 0.858 ± 0.029) (Figure 1B). Real-world scenarios often involve the presence of noninformative features, complicating the extraction of the underlying predictive signal. In our next benchmarking scenarios, we introduced additional features containing random noise (n_{noise} = 7, 17, 37, 99). The PBMF consistently outperformed VT (Figure S1B), and the performance differential between PBMF and VT was further widened by pruning suboptimal models within the PBMF ensemble or by increasing the training data sample size (Figures S1C–S1E). All subsequent benchmarking was therefore performed using a pruned PBMF ensemble (hyperparameters can be found in Table S2).

Having established the success of the PBMF in simulated scenarios, we benchmarked the PBMF, VT, and SIDES across a diversity of 9 clinical studies (Table 1), including real-world data, various cancer and non-cancer indications, and phase 1, 2, and 3 IO clinical trials. These datasets spanned a diversity of data modalities such as DNA, RNA, clinical, and demographic (Tables S3 and S4). Overall, the PBMF markedly outperformed all other methods by consistently identifying predictive biomarkers (Figure 1C). We detail the results of our clinical study benchmarking in the sections to follow.

(B) AUPRC for a simulated data test set comparing the PBMF, VT,[22] and SIDES[21] models trained on simulated training dataset containing 3 features (2 predictive, 1 prognostic; training performed on 1000 data points, with 100 training-test split replicates). Boxplot: centerline, median; box limits, quartile 1 and 3; box whiskers, 1.5x interquartile range; diamonds, outliers; dots, data points.

(C) Hazard ratios for PBMF, VT, and SIDES methods across all 9 test datasets and across treatments for each biomarker status, B+ and B–. Points are connected if they represent hazard ratios computed for biomarker groups within the same dataset. Shaded areas correspond to the bounding box defined by the maximum and minimum hazard ratios for each method, for a given biomarker status, B+ and B–.

(D) Forest plot illustrating the performance comparison of PBMF, VT, and SIDES methodologies, applied to test datasets. Shown are the hazard ratios (HR) and 95% confidence intervals (95% CI) from a Cox proportional hazards model fit to each treatment comparison within a biomarker status. Patient numbers (N) are shown to the left of the forest plot, where TRT = the treatment for which the predictive biomarker was desired (e.g., IO for TEMPUS) and CTL = the comparator treatment (e.g., chemotherapy for TEMPUS).

See also Figures S1–S5, Tables S1, S2, S3, and S4.

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

Table 1. Summary of clinical study and real-world data used in the present study

Training	Test	Disease	Treatment	Outcomes	Features
Survival analysis datasets					
Rotterdam study cohort[26] (*n* = 863)	German study cohort[27] (*n* = 686)	Breast cancer	Hormone therapy + chemo vs. chemo	Overall survival	• Age • Menopause • Tumor size • Tumor grade • Number of nodes • Progesterone receptor status • Estrogen receptor status
Blair et al.[28] study (randomized half)	Blair et al.[28] study (randomized half)	Diabetic retinopathy (DR)	Laser coagulation (*n* = 197) vs. no treatment (*n* = 197)	Time to visual acuity <5/200 for two successive visits	• Age • Diabetes type • Risk score
Immune checkpoint inhibitor trials					
JAVELIN Renal 101 phase 3 trial[29] (randomized half)	JAVELIN Renal 101 phase 3 trial[29] (randomized half)	Advanced renal cell carcinoma (aRCC)	Avelumab (anti-PD-L1) + axitinib (chemo) (*n* = 354) vs. sunitinib (SoC) (*n* = 372)	Overall survival	• PD-L1 status • Expression of 59 TME and pathway-related RNA signatures (Table S4) based on FFPE tumor tissue RNA-seq
POSEIDON phase 3 trial[31] (randomized half)	POSEIDON phase 3 trial[31] (randomized half)	Metastatic non-small-cell lung cancer (mNSCLC)	Durvalumab (anti-PD-L1) + chemo (*n* = 114) vs. chemo (*n* = 114)	Overall survival	• Expression of 35 TME-related RNA signatures based on peripheral blood RNA-seq (Table S4)
IMmotion150 phase 2 trial[30] (randomized half)	IMmotion150 phase 2 trial[30] (randomized half)	Metastatic renal cell carcinoma (mRCC)	Atezolizumab (anti-PD-L1) + bevacizumab (anti-VEGF) (*n* = 83) vs. sunitinib (anti-VEGF, SoC) (*n* = 81)	Overall survival	• Age • Sex • Liver metastasis • Previous nephrectomy • T cell effector signature score • Plasma IL8 • Sum of longest tumor diameter • Sample type (primary/metastatic)
Real world data					
Tempus NSCLC cohort (randomized half)	Tempus NSCLC cohort (randomized half)	NSCLC	Anti-PD-1 or anti-PD-L1 (*n* = 117) vs. chemo (*n* = 84)	Overall survival	• Expression of 50 cancer Hallmark gene sets (MSigDB C5)[45] based on RNA-seq of pre-treatment tumors

(*Continued on next page*)

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

Table 1. *Continued*

Training	Test	Disease	Treatment	Outcomes	Features
Synthetic control arm trials					
Checkmate-010 single-arm phase 2 trial[36] (n = 25) + random selection of control arm of Checkmate-025 phase 3 trial[37] (n = 25)	Checkmate-009 phase 1 trial[38] + Checkmate-025 phase 3 trial[37] (n = 149; excludes those used for training)	Clear cell renal carcinoma (ccRCC)	Nivolumab (anti-PD-1) vs. everolimus (mTOR inhibitor)	Overall survival	• Expression of 3 immune-related RNA signatures (Table S4) based on FFPE tumor tissue RNA-seq • Copy number variations of 24 DNA segments (Table S4) based on FFPE tumor tissue whole-exome sequencing • Mutations in 11 genes (Table S4) based on FFPE tumor tissue whole-exome sequencing
IMvigor210 single-arm phase 2 trial[39] (n = 119) + random selection of control arm of IMvigor211 phase 3 trial[40] (n = 100)	IMvigor211 phase 3 trial[40] (n = 472; excludes those used for training)	Metastatic urothelial carcinoma (mUC)	Atezolizumab (anti-PD-L1) vs. chemo	Overall survival	• Age • Sex • Liver metastasis • ECOG performance status • IL8 plasma level at baseline (C1D1) • IL8 plasma level after treatment (C3D1) • IL8 plasma ratio (C3D1/C1D1)
Phase 2 to predict phase 3 trials					
POPLAR phase 2 trial[41] (n = 206)	OAK phase 3 trial[42] (n = 638)	NSCLC	Atezolizumab (anti-PD-L1) vs. docetaxel	Overall survival	• Age • Sex • ECOG performance status • Sum of longest diameter of target lesions at baseline • Number of metastatic sites at enrollment • Histology • Smoking history • Maximum somatic allele frequency • Blood tumor mutational burden • ctDNA mutations of 20 most prevalent genes (Table S4)

See also Figure S1, Tables S1, S2, S3, and S4.

CellPress
OPEN ACCESS

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

Identification of predictive biomarkers in commonly used clinical datasets for survival analysis

We evaluated PBMF against VT and SIDES with well-characterized clinical datasets used in common practice for time-to-event statistical modeling (specifically survival analysis).[23–25] We utilized datasets from breast cancer[26,27] and diabetic retinopathy[28] studies, as these were the most feature-rich and appropriate for a predictive biomarker discovery task (see STAR Methods, experimental model and study participant details for details on source data origin).

First, we benchmarked the PBMF against VT and SIDES for identifying a biomarker predictive of longer survival with hormone therapy (tamoxifen) plus chemotherapy versus chemotherapy alone across the two available independent breast cancer datasets. Models were trained on the Rotterdam breast cancer cohort[26] and subsequently tested on the German breast cancer study cohort.[27] On the training dataset, the PBMF (B+: hazard ratio [HR] = 0.71, confidence interval [CI] = 0.54–0.94, p = 1.69e-2; B–: HR = 1.91 CI = 1.48–2.48, p = 9.37e-7) and VT (B+: HR = 0.56, CI = 0.44–0.70, p = 4.98e-7; B–: HR = 1.81, CI = 1.30–2.52, p = 4.32e-4) methods successfully identified a predictive biomarker, whereas SIDES found a prognostic biomarker (Figures 1D, S2, and 2A). On the test dataset, only the PBMF generalized as a predictive biomarker (B+: HR = 0.63, CI = 0.48–0.83, p = 1.02e-3; B–: HR = 0.89, CI = 0.50–1.57, p = 6.84e-1), whereas both VT and SIDES were prognostic.

We next benchmarked the PBMF against VT and SIDES for identifying a biomarker predictive of longer time to vision loss with laser therapy versus no treatment in a study for treating diabetic retinopathy.[28] On the training split of the data, the PBMF (B+: HR = 0.27, CI = 0.13–0.55, p = 3.67e-4; B–: HR = 0.69, CI = 0.38–1.24, p = 2.13e-1) identified the strongest predictive biomarker (Figure S2). VT (B+: HR = 0.38, CI = 0.21–0.70, p = 1.88e-3; B–: HR = 0.55, CI = 0.28–1.09, p = 8.81e-2), and SIDES (B+: HR = 0.38, CI = 0.09–1.52, p = 1.71e-1; B–: HR = 0.46, CI = 0.29–0.74, p = 1.51e-3) found mostly prognostic biomarkers (Figures S2A and S2B). In particular, for VT, the biomarker from the training data appears to enrich for reduced time to vision loss within each treatment, which is opposite to the desired behavior (Figure S2C). This therefore discounts the otherwise favorable generalization of VT on the test split of the data (Figures 1D and 2A). In contrast, the PBMF (B+: HR = 0.38, CI = 0.17–0.81, p = 2.26e-4; B–: HR = 0.55, CI = 0.29–1.04, p = 6.62e-2) identified a predictive biomarker, albeit with a prognostic component (Figures 1D and 2A).

Predictive biomarker identification in immune checkpoint inhibitor therapies

Encouraged by our results from simulated biomarker scenarios and well-established clinical datasets for survival analysis, we asked whether the PBMF would excel over VT and SIDES in the challenging predictive biomarker discovery space of immuno-oncology, specifically for immune checkpoint inhibitor (ICI) therapy. We trained and tested models on each of three phase 3 clinical trials (JAVELIN 101,[29] NCT02684006; IMmotion 150,[30] NCT01984242; POSEIDON,[31] NCT03164616; see STAR Methods, experimental model and study participant details for details on source data origin) for three different ICI therapies given in a first-line setting (avelumab, atezolizumab, durvalumab, respectively) for either renal cell carcinoma or non-small cell lung cancer (NSCLC). SIDES failed to find a predictive biomarker on the training data for IMmotion 150 and JAVELIN 101, whereas both the PBMF and VT consistently found a predictive biomarker on the training data for all three clinical trials (Figure S2).

On the test data for POSEIDON, only the PBMF identified a predictive biomarker that generalized (Figures 1D and 2B; B+: HR = 0.33, CI = 0.13–0.80, p = 1.4e-2; B–: HR = 1.10, CI = 0.67–1.80, p = 7.06e-1). When testing on JAVELIN 101, only the PBMF (B+: HR = 0.52, CI = 0.33–0.80, p = 3.32e-3; B–: HR = 1.03, CI = 0.68–1.56, p = 8.81e-1) generalized as a predictive biomarker. The PBMF identified a B+ group characterized by longer survivors in the avelumab + axitinib arm of interest versus all other groups and arms (Figures 1D and 2B). Although VT appears to have found a generalizable predictive biomarker as well (B+: HR = 0.43, CI = 0.28–0.65, p = 5.48e-5; B–: HR = 1.26, CI = 0.81–1.96, p = 3.10e-1), examination of the Kaplan-Meier plots suggests that it instead identified a B+ group treated with the control therapy, sunitinib, that had worse survival versus all other groups and arms (Figure 2B). Finally, when testing on IMmotion 150, the PBMF trended the best toward a predictive biomarker, as it enriched for both for patients that had better survival across treatments within the B+ group (HR = 0.49, CI = 0.13–1.92, p = 3.08e-1), as well as across biomarker status within the ICI treatment (Figure 2B). In contrast, although VT similarly trended toward a predictive biomarker (Figure 1D), the B+ group across treatments trended toward worse survival than the B– group (Figure 2B).

In summary, PBMF demonstrated superior performance in all three phase 3 clinical trials for immune checkpoint inhibitor therapies, consistently identifying predictive biomarkers where SIDES failed and VT misidentified beneficial groups. PBMF reliably pinpointed patient groups with improved survival outcomes, highlighting its potential as a robust tool for predictive biomarker discovery.

Predictive biomarker identification with real-world data

Randomized controlled phase 3 clinical trials are often considered the gold standard for tasks like predictive biomarker discovery analysis; these datasets often take a significant amount of time to accumulate and require substantial investments. With the increasing availability of real-world data (RWD), we chose to benchmark PBMF against VT and SIDES despite challenges associated with the use of RWD, including issues related to inconsistent data quality, comparability, and bias.[32,33] To facilitate this comparison, we curated an NSCLC real-world data cohort through a licensing agreement with Tempus to specifically to evaluate first-line ICI therapy versus chemotherapy.

On the training dataset, only the PBMF and VT yielded a biomarker with predictive value for ICI over chemotherapy, whereas SIDES exhibited a trend toward prognostic behavior (Figure S2). On the test dataset, only the PBMF (B+: HR = 0.26, CI = 0.09–0.71, p = 9.02e-3; B–: HR = 1.20, CI = 0.50–2.85, p = 6.86e-1) demonstrated enrichment for longer survivors specific to ICI therapy, indicating the discovery of a predictive biomarker that can generalize (Figures 1D and 2C). In contrast, VT failed to generalize in the test dataset (B+: HR = 0.48, CI = 0.18–1.30, p = 1.49e-1; B–: HR = 0.84, CI = 0.36–1.96, p = 6.83e-1), despite very strong predictive behavior observed in the training dataset. The

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

CellPress
OPEN ACCESS

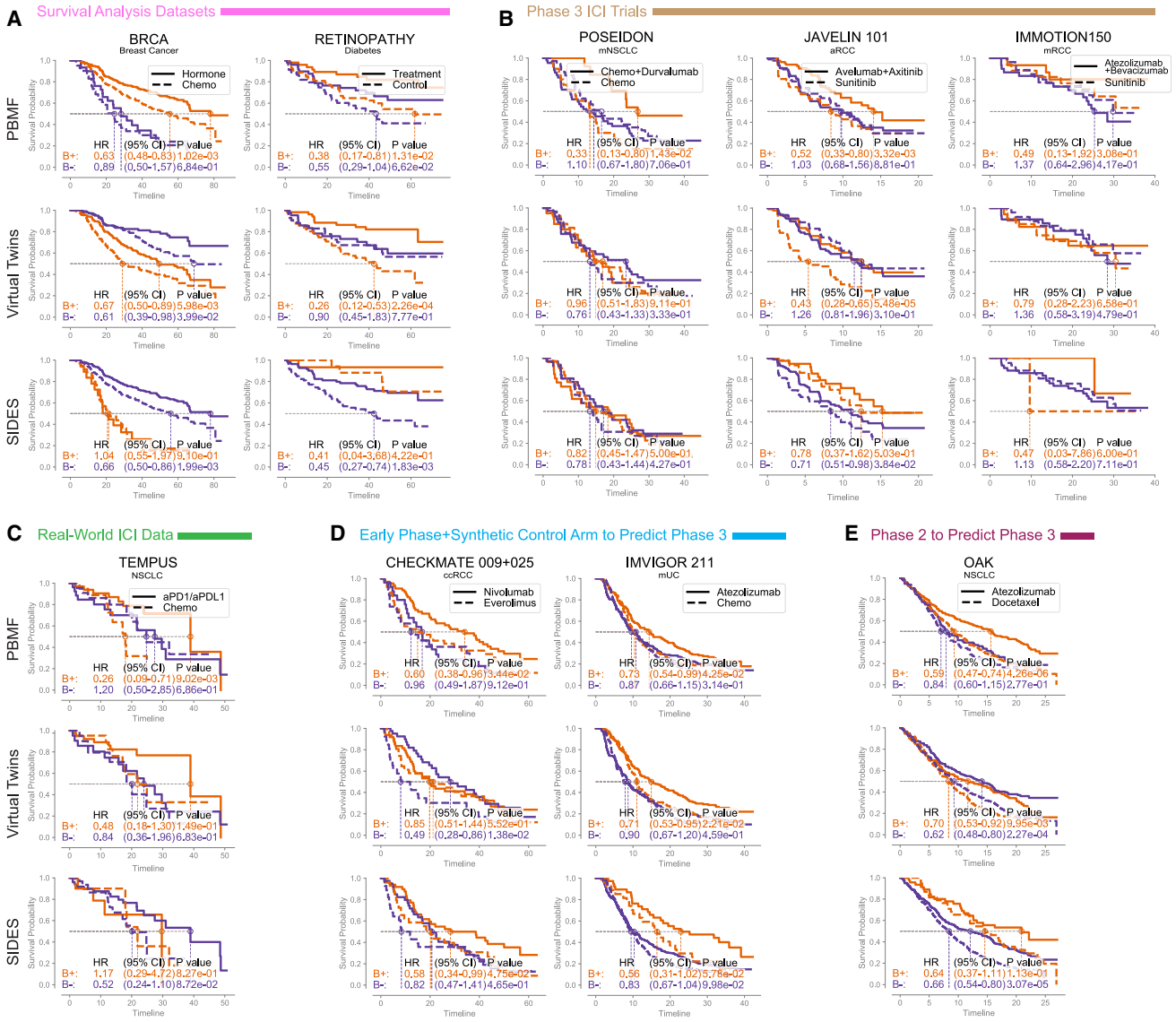


Figure 2. Kaplan-Meier curves for evaluation of PBMF for predictive biomarker identification on real clinical datasets against other methods

(A–E) Kaplan-Meier curves per treatment and biomarker status (from PBMF, VT, or SIDES), as evaluated on the (A) test data from well-established clinical datasets for survival analysis (breast cancer[26,27] and retinopathy[28]), (B) immuno-oncology clinical trial test data (POSEIDON,[31] JAVELIN 101,[29] and IMmotion 150[30]), (C) TEMPUS real-world data test set, (D) clinical trial test data that utilized synthetic control arms (CheckMate 009[38] + CheckMate 025[37] and IMvigor 211[40]), and (E) OAK[42] phase 3 clinical trial test dataset. Timeline is in months. Hormone, hormone + chemotherapy; chemo, chemotherapy; atezo, atezolizumab; B, bevacizumab; aPD1, anti–PD-1 immunotherapy; aPDL1, anti–PD-L1 immunotherapy. Sample sizes (N) for biomarker-positive/negative and treatment/control groups for each trial are provided in Figure 1D. *p*-values shown are derived from Wald tests.

See also Figures S2–S5.

trend toward prognostic behavior failed to generalize for SIDES (B+: HR = 1.17, CI = 0.29–4.72, p = 8.27e-1; B–: HR = 0.52, CI = 0.24–1.10, p = 8.72e-2).

Predictive biomarker discovery with synthetic control arms

Early phase trials are often single-arm studies, complicating efforts to derive biomarkers specific to a treatment of interest. Recent FDA guidance suggests common[34] or external[35] control arms might be used in certain settings to minimize redundancy, especially for and motivated in large part by oncology drug discovery. We therefore evaluated our approach in this "synthetic control arm" scenario, whereby we used a fraction of phase 3 control arm data exclusively for model training alongside phase 2 single-arm trial data (see STAR Methods, experimental model and study participant details for details on source data origin).

In the context of pre-treated advanced clear cell renal carcinoma (ccRCC), PBMF, VT, and SIDES all identified a predictive biomarker for ICI therapy on the training data from the nivolumab arm of phase 2 CheckMate 010[36] (NCT01354431) and a synthetic control arm from a random subset of patients receiving everolimus from phase 3 CheckMate 025[37] (NCT01668784; Figure S2).

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

However, only the PBMF generalized to the test dataset on the combined population from phase 1 CheckMate 009[38] (NCT01358721) and phase 3 CheckMate 025 trials (Figures 1A and 2D excluding those from CheckMate 025 used for training; B+: HR = 0.60, CI = 0.38–0.96, p = 3.44e-2; B–: HR = 0.96, CI = 0.49–1.87, p = 9.12e-1). SIDES trended toward a prognostic biomarker (B+: HR = 0.58, CI = 0.34–0.99, p = 4.75e-2; B–: HR = 0.82, CI = 0.47–1.41, p = 4.65e-1), whereas VT did not generalize, as it displayed a predictive biomarker for the control arm (B+ HR = 0.85, CI = 0.51–1.44, p = 5.52e-1; B–: HR = 0.49, CI = 0.28–0.96, p = 1.38e-2). Overall, the PBMF identified a B+ subpopulation with a 12% decrease in risk of death when treated with nivolumab versus everolimus, relative to the biomarker-evaluable population (BEP) in the combined CheckMate 009 and 025 trials (Figures 1D and S3; PBMF HR = 0.60; CheckMate 009 and 025 BEP HR = 0.68; CheckMate 025 BEP trial-reported HR = 0.69; CheckMate 025 intent-to-treat HR = 0.73).

The PBMF also generalized well in an additional independent cohort examining atezolizumab versus chemotherapy in locally advanced or metastatic urothelial carcinoma (mUC). In this analysis, we included all available input features at baseline (age, sex, ECOG, pIL-8 expression, and liver metastasis) and on-treatment (pIL-8 after 6 weeks) to evaluate their association with overall survival. On the training data from the atezolizumab arm from phase 2 IMvigor210[39] (NCT02951767, NCT02108652) and a synthetic control arm from a random subset of patients receiving chemotherapy from phase 3 IMvigor211[40] (NCT02302807), only the PBMF and VT but not SIDES yielded a biomarker with predictive value of atezolizumab over chemotherapy (Figure S2). Similarly, on the test dataset (IMvigor 211 excluding patients used for the training synthetic control arm), both PBMF (B+: HR = 0.73, CI = 0.54–0.99, p = 4.25e-2, B–: 0.87, CI = 0.66–1.15, p = 3.14e-1) and VT (B+: HR = 0.71, CI = 0.53–0.95, p = 2.21e-2; B–: HR = 0.90, CI = 0.67–1.20, p = 4.59e-1) generalized well as a predictive biomarker (Figures 1D and 2D). This corresponded to a 10% and 12% decrease in risk of death, respectively, when treated with atezolizumab versus chemotherapy, relative to the BEP in the IMvigor 211 trial (Figures 1D and S3; PBMF HR = 0.73; VT HR = 0.71; IMvigor 211 BEP HR = 0.81; IMvigor 211 intent-to-treat HR = 0.85). These trends recapitulate when, instead of using phase 1 data (one-arm, synthetic control), using phase 2 data (two-arms)[41] to predict phase 3 outcomes[42] (Figure 2E) (B+: HR = 0.59, CI = 0.47–0.74, p = 4.26e-6; B–: HR = 0.84, CI = 0.60–1.15, p = 2.27e-1). Both VT (B+: HR = 0.70, CI = 0.53–0.92, p = 9.95e-3; B–: HR = 0.62, CI = 0.48–0.80, p = 2.27e-4) and SIDES (B+: HR = 0.64, CI = 0.37–1.11, p = 1.13e-1; B–: HR = 0.66, CI = 0.54–0.80, p = 3.07e-5) yielded only prognostic biomarkers.

Knowledge distillation from the PBMF neural network produces a simple, interpretable decision tree

Clinical study biomarker strategy may require an interpretable "white-box" biomarker in order to be practically deployed. Utilizing a consensus score across the models within a given PBMF ensemble, we determined an optimal biomarker probability score cutoff to classify B+ and B– samples, subsequently referred to as pseudo-labels (Figure 3A, STAR Methods). These pseudo-labels were then used to distill the complex, original neural network ensemble PBMF model into a simple interpretable model—a decision tree. To further facilitate interpretability, and to explore how the decision tree recapitulates the original predictive biomarker, our framework includes an interactive web app. Distilled decision tree PBMF biomarkers were generated for all clinical datasets evaluated in the present study (Figures 3B, 3C, and S4A–S4F).

For the POSEIDON ICI study that utilized peripheral blood gene expression signatures,[43] B+ individuals were characterized by lower signature for tumor-associated macrophages (TAMs) or higher signature for B cells (Figure 3B). Although these signatures were computed on peripheral blood gene expression, the TAM signature level likely correlates with tumor macrophage levels, as only monocytes or traveling macrophages may be found in the blood. Low levels of TAMs may correlate with improved activity of ICI therapy.[44] High B cell signature may be indicative of overall immune health, which would permit a favorable ICI therapy response. Subsequent splits in the tree yielded a more strongly predictive biomarker (B+ HR = 0.21 at depth = 5 vs. HR = 0.48 at depth = 2), despite maintaining similar numbers of B+ individuals.

For the TEMPUS RWD cohort that utilized gene expression signature scores computed from tumor RNA-seq across 50 hallmark gene sets (mSigDB C5[45]; see STAR Methods), B+ individuals were low for E2F hallmark signature or high IFN-γ response hallmark signature (Figure 3C). The E2F signature contains cell cycle genes and therefore is expected to correlate with more aggressive disease. High IFN-γ has been previously associated with ICI favorable outcomes,[46] and helps steer the biomarker from being prognostic (B+ HR = 0.71, B– HR = 0.65 at depth = 1) to being predictive (B+ HR = 0.26, B– HR = 0.91 at depth = 2). As with POSEIDON, subsequent splits in the tree yielded a more strongly predictive biomarker (B+ HR = 0.21, B– HR = 1.23 at depth = 5), although the biological interpretation was less straightforward.

A discovery pipeline for predictive biomarker prototypes: Identification of individuals with improved survival outcomes in early-stage clinical trial data to inform phase 3 trial design

One critical application of predictive biomarker discovery is to inform the patient selection strategy for phase 3 clinical trials by using data from earlier phases. Given the consistent ability of the PBMF to identify a predictive biomarker, particularly in clinical trial settings, we devised an end-to-end biomarker discovery pipeline that generates a human-understandable predictive biomarker prototype, poised for translation into clinical settings. As a sample use case for this pipeline, we sought to guide patient selection for second-line atezolizumab therapy versus chemotherapy in NSCLC, relying solely on data from an earlier study. Specifically, we identified clinicogenomic phase 2 trial data (POPLAR,[41] NCT01903993; Figure 4A) with which we trained a PBMF model (Figures 1D, 2E, and 4B) to identify a predictive biomarker (Figures 4B–4D and S5A–S5E; STAR Methods) that was then tested on phase 3 trial data (OAK,[42] NCT02008227; Figure S5E); see STAR Methods, experimental model and study participant details for details on source data origin.

The PBMF identified a generalizable predictive biomarker (Figures 1D, 2E, and 4B–4D); B+: HR = 0.59, CI = 0.47–0.74, p = 4.26e-6; B–: HR = 0.84, CI = 0.60–1.15, p = 2.27e-1) when

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

CellPress
OPEN ACCESS

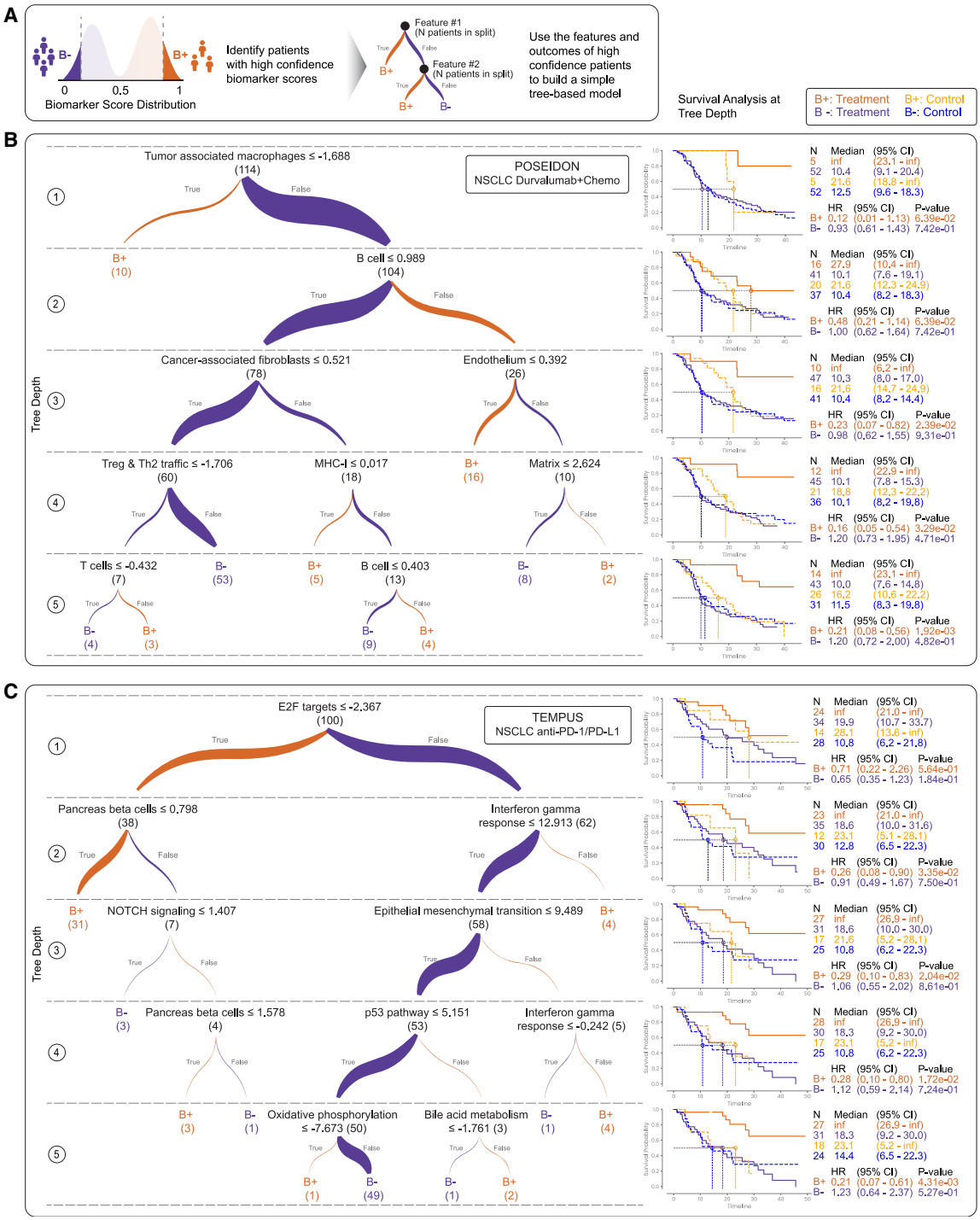


Figure 3. PBMF model distillation into interpretable decision trees facilitates understanding of biomarker

(A) High-confidence patient samples are identified through biomarker pseudo-labeling from biomarker scores generated from the pruned PBMF ensemble. These then serve to construct an interpretable, simplified decision tree model, categorizing patients as B+ or B–.

(B) Example distilled PBMF decision tree (depth = 5) generated from POSEIDON[31] training data that best recapitulated the original PBMF biomarker for the dataset. Pathways are from Bagaev et al. 2021.[43] Sample sizes (N) for biomarker-positive/negative and treatment/control groups for each trial are provided in the figure. *p*-values shown are derived from Wald tests.

(C) Example distilled PBMF decision tree (depth = 5) generated from TEMPUS training data that best recapitulated the original PBMF biomarker for the dataset. Pathways are MSigDB gene sets.[45] KM curves are shown to reflect the nature of the biomarker at successive levels of depth. Line thickness is proportional to number of patients in parentheses. Sample sizes (N) for biomarker-positive/negative and treatment/control groups for each trial are provided in the figure. *p*-values shown are derived from Wald tests.

See also Figure S4.

CellPress
OPEN ACCESS

Cancer Cell
Article

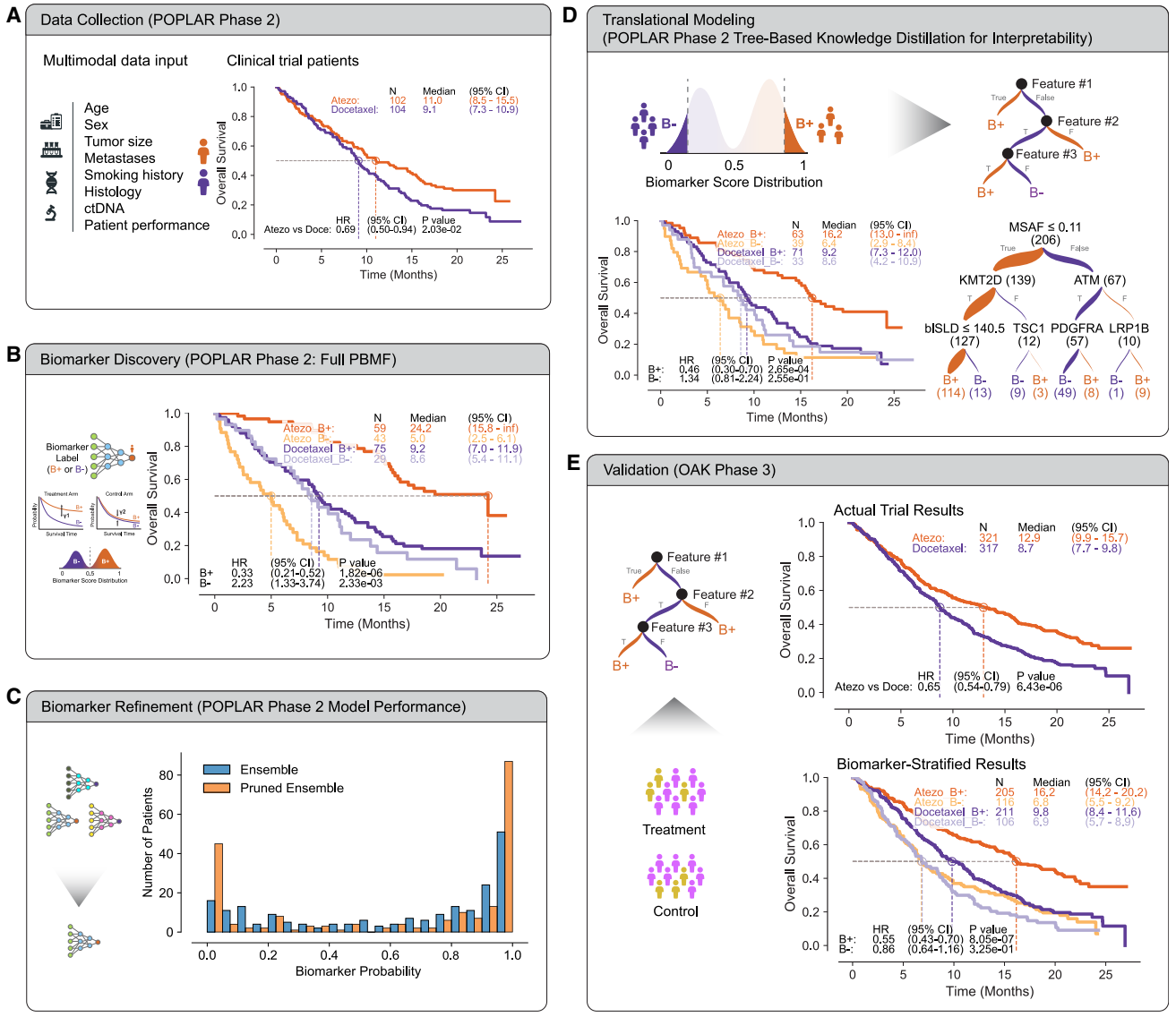


Figure 4. Application of PBMF in the design of biomarker-driven clinical trials

(A) Clinical trial data and endpoints collection: Kaplan-Meier curves for the discovery (POPLAR phase 2 clinical trial[41]) dataset. Sample sizes (N) for biomarker-positive/negative and treatment/control groups for each trial are provided in the figure. *p*-values shown are derived from Wald tests.

(B) Identification of predictive biomarker: using the discovery dataset (POPLAR trial) the PBMF successfully finds a biomarker that identifies which patients will survive longer on atezolizumab but not docetaxel. Sample sizes (N) for biomarker-positive/negative and treatment/control groups for each trial are provided in the figure. *p*-values shown are derived from Wald tests.

(C) Refinement of predictive biomarker: the enhancement of the predictive biomarker involves pruning to eliminate spurious models from the ensemble, and (D) subsequent derivation of a rule set (i.e., a decision tree) that encapsulates the biomarker's predictive power. Line thickness is proportional to number of patients in parentheses. Sample sizes (N) for biomarker-positive/negative and treatment/control groups for each trial are provided in the figure. *p*-values shown are derived from Wald tests.

(E) Independent validation set, OAK Phase 3 trial[42] (top), and patient stratification using the simplified predictive biomarker identified in the POPLAR trial and subsequently applied to the OAK trial (bottom). Such independent dataset validation of the PBMF model affirms the biomarker's predictive capacity, demonstrating the model's reliability from ensemble to simplified tree representation, thus reinforcing its utility in clinical trial stratification. Sample sizes (N) for biomarker-positive/negative and treatment/control groups for each trial are provided in the figure. *p*-values shown are derived from Wald tests.

See also Figures S2, S3, S5, and Table S5.

trained on POPLAR study data and subsequently applied as a hypothetical patient selection biomarker for the phase 3 OAK trial test data. Both VT (B+: HR = 0.70, CI = 0.53–0.92, *p* = 9.95e-3; B–: HR = 0.62, CI = 0.48–0.80, *p* = 2.27e-4) and SIDES (B+: HR = 0.64, CI = 0.37–1.11, *p* = 1.13e-1; B–: HR = 0.66, CI = 0.54–0.80, *p* = 3.07e-5) yielded only prognostic biomarkers (Figures 1D and 2E). This was despite PBMF (B+: HR = 0.30, CI = 0.19–0.48, *p* = 2.57e-7; B–: HR = 2.41, CI = 1.41–4.11, *p* = 1.25e-3) and VT (B+: HR = 0.38, CI = 0.24–0.60, *p* = 3.72e-5; B–: HR = 1.14, CI = 0.72–1.78, *p* = 5.76e-1) having

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

identified a predictive signal from the phase 2 POPLAR training data; and SIDES having identified a mixed predictive and prognostic signal within the training data (B+: HR = 0.42, CI = 0.14–1.21, p = 1.08e-1; B–: HR = 0.75, CI = 0.54–1.05, p = 9.51e-2) (Figure S2). Compared with the BEP in the OAK trial (Figure S3), the PBMF B+ subpopulation yielded a ∼9% decrease in risk of death for atezolizumab versus docetaxel treatment (PBMF, HR = 0.59; OAK BEP HR = 0.65). Thus, to hypothetically inform strategies for patient selection in phase 3 clinical trials, only the PBMF successfully identified a predictive, high-prevalence biomarker from phase 2 data that generalized to phase 3 results.

Having established a predictive biomarker with the pruned PBMF ensemble, we distilled the biomarker into an interpretable decision tree (Figure 4D). Like the original PBMF from which it was derived, the distilled decision tree PBMF biomarker was predictive when applied on all samples from the phase 2 trial training data (Figure 4D; B+: HR = 0.46, CI = 0.3–0.7, p = 2.6e-4; B–: HR = 1.34, CI = 0.8–2.2, p = 0.2). The distilled PBMF was also predictive for phase 3 OAK trial test data (Figure 4E; B+: HR = 0.55, CI = 0.43–0.7, p = 8.05e-7; B–: HR = 0.86, CI = 0.64–1.16, p = 0.3). Importantly, the OAK HR of the distilled decision tree was improved by approximately 7% compared with the original PBMF (original PBMF HR = 0.59; distilled decision tree PBMF HR = 0.55; Figures 4E and S5E), owing to the reduction in prevalence from 80% to 64%. Notably, the original PBMF had a ∼9% decrease in risk of death within the B+ atezolizumab versus docetaxel-treated subpopulation relative to the BEP in the OAK trial, and the distilled decision tree PBMF had a ∼15% decrease in risk of death (distilled PBMF HR = 0.55; original PBMF HR = 0.59; OAK BEP trial-reported HR = 0.65, OAK intent-to-treat HR = 0.73).

Upon scrutinizing the decision tree of the distilled PBMF, we observed that the predictive biomarker comprised a specific subset of clinical and genomic features: the maximum circulating tumor DNA (ctDNA) allele frequency (MSAF), sum of longest diameter of target lesions at baseline (bISLD), and mutation status on the *KMT2D*, *TSC1*, *ATM*, *PDGFRA* and *LRP1B* genes (Figure 4D). MSAF and bISLD likely reflect overall tumor burden, and low values for these would favor beneficial outcomes for both ICI and chemotherapy. *KMT2D* mutations have been previously associated with immune-inflamed tumor microenvironment[47] and improved benefit from ICI but not chemotherapy.[48] *ATM* mutations have also been previously described, albeit with equivocal results regarding ICI benefit.[49,50] With the exception of *ATM* mutations, which were both predictive and prognostic (POPLAR: mutation [Mut] B+ HR = 0.33, wild type [Wt] B– HR = 0.776; OAK: Mut B+ HR = 0.43, Wt B– HR = 0.68) but with a notably low prevalence (28 patients for *ATM* B+/Mut and 205 for the distilled PBMF B+), each individual feature fell short in matching the biomarker prevalence or the consistent, predictive signal of the collective (Figure S5F and Table S5). In comparison with a commonly described single-feature ICI biomarker, blood TMB,[51–53] the PBMF more robustly enriched for longer survival for both the training and test clinical trial datasets (Figures 4F, S5G, and Table S5).

Potential applications of the PBMF

The PBMF is an end-to-end application programming interface (API) capable of integrating with pretrained models across various data modalities; for example, genomic, radiomic, and clinical data (Figure 5A). The modularity of the PBMF, enabled by its compatibility with various differentiable models, allows it to process these embeddings to discover biomarkers for survival, adverse events, dosing strategies, and more (Figure 5B).

DISCUSSION

Across diverse, challenging benchmarks spanning simulated scenarios through informing strategies for patient selection in clinical trials, the PBMF outperformed other methods for discovering predictive biomarker signals. Among comparator methods, only the PBMF found signals that were consistently predictive across training and test datasets. Along with the PBMF's ability to accurately identify known IO biomarkers from phase 2/3 trials, we also showed that the PBMF can nominate a novel composite biomarker from a set of clinicogenomic features that outperformed blood TMB.

We emphasize here the importance of the predictive constraint embedded in the PBMF. A common pitfall in biomarker discovery is to focus only on identifying populations with enhanced responses to a specific treatment.[54] In these cases, one cannot distinguish between a biomarker that is prognostic versus one that enriches for better responses specifically in a treatment of interest. Thus, the PBMF contrastive loss function enforces the constraint that a biomarker must be considered in the context of a control treatment. Although VT[22] and SIDES[21] also enforce such a constraint, they do so by finding groups of patients that maximize survival difference in the treatment of interest and (in the opposite direction) the control (Table S1), and in the case of VT, require an independently trained model for each of the treatment and control. In contrast, the PBMF seeks groups of patients that *simultaneously* maximize survival difference in treatment and *minimize* difference in control. The PBMF therefore constrains the biomarker search space, potentially increasing the chance of finding a generalizable predictive biomarker; furthermore, formulating the learning objective as a contrastive learning task may help the model capture finer similarities and differences between the control and treatment groups.

In our patient selection strategy example, we successfully distilled a complex ensemble neural network model into a simple decision tree. In this regard, we can view the PBMF as a highly effective search function, as we required the complex model to discern whether a predictive signal exists and what features may drive it. While a multivariate Cox proportional hazards (Cox PH) model could theoretically achieve similar results,[9,24,55] it requires explicit and systematic testing of each potential variable and their combinations, making it impractical to implement. In contrast, the gradient descent within the PBMF implicitly traverses the vast expanse of potential feature combinations and interactions. Additionally, the PBMF simultaneously accounts for treatment effects within its loss function, whereas a Cox PH model requires enumeration of each hypothesized treatment-feature interaction.

Beyond its contrastive loss function and interpretable tree, the PBMF stands out as an end-to-end API for predictive biomarker discovery. Not only does the PBMF ensemble of fully connected neural networks outperform other methods but the API is also compatible with any differentiable model and therefore can be

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

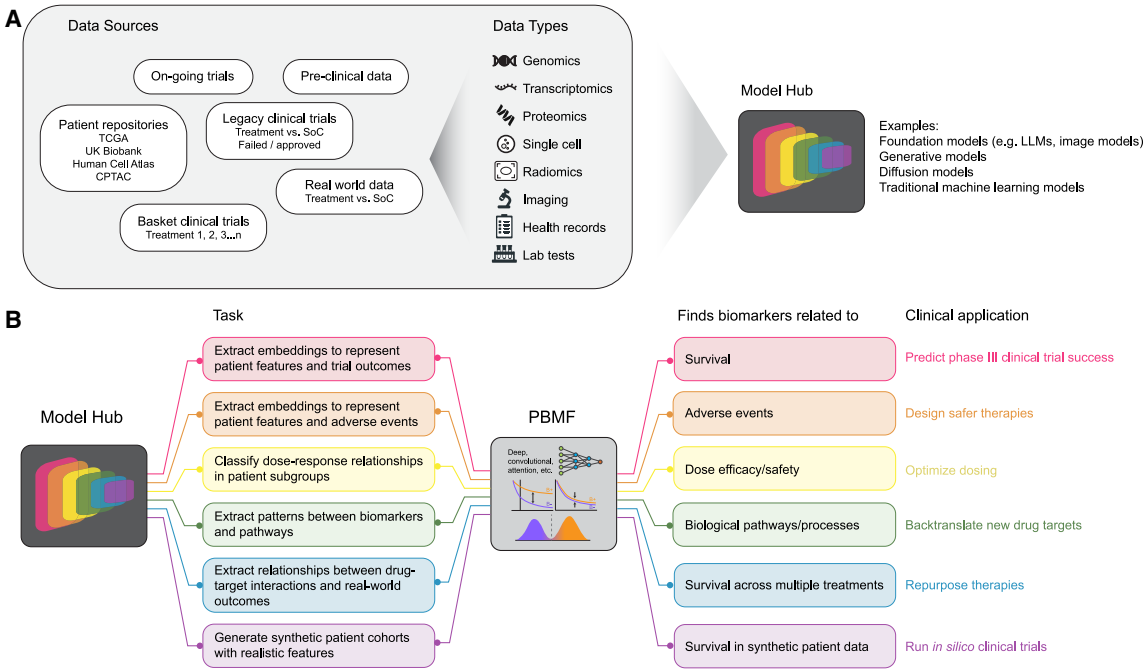


Figure 5. Potential future use cases of the PBMF

(A) Data sources for predictive biomarker discovery include patient repositories (e.g., TCGA and UK Biobank), various types of clinical trials (e.g., past failed trials, basket trials testing multiple treatments), and real-world data. These sources encompass diverse data types, such as genomics, radiomics, imaging, and health records, which can feed into a "Model Hub" of a variety of pretrained models, including large-language models, generative models, diffusion models, and traditional machine learning models.

(B) The PBMF leverages embeddings, patterns, and/or synthetic data generated by pretrained models fine-tuned for specific tasks.

incorporated within larger multi-component AI systems. This flexibility makes it possible to explore predictive biomarker signals using input features from single or multiple modalities, or diverse data representations. For instance, an attention-based transformer model could effectively model unstructured data such as clinical notes or histopathology images.[56,57] This capability allows the integration of pretrained models, such as foundation models like large-language models, generative models, and traditional machine learning models, and more, feeding prior knowledge into the PBMF and potentially enabling successful predictive biomarker discovery even in data-limited or noisy contexts.[58] Paired with pretrained models trained on ever-expanding sets of genomic, radiomic, clinical, and other datatypes, the PBMF could eventually extract complex embeddings to predict not only survival outcomes but also adverse events,[59] dose-response relationships, combination therapy effects, and even back-translate new drug targets. By leveraging synthetic patient cohorts generated through generative AI, the PBMF may further accelerate drug development by simulating theoretical trial designs, testing hypotheses before implementation, and optimizing investment decisions in resource-intensive phase 3 trials. These applications highlight the PBMF as a central component in a broader ecosystem for predictive biomarker discovery. With the integration of other future pretrained models, the PBMF has the potential to enhance precision medicine by supporting the discovery of clinically actionable biomarkers.

Specific considerations and limitations apply when using any predictive biomarker method to inform late-stage clinical trial decision-making. As alluded to earlier, data availability is often limiting. The success of the PBMF in identifying potential predictive biomarkers from real-world data and from using synthetic control arms is thus promising. Future work will be required to know whether synthetic control arms from non-randomized evidence (i.e., real-world data) could be used; any such exploration would need to carefully consider the substantial heterogeneity within patient populations. For instance, examination of established metrics such as the propensity score[60] may be required to evaluate comparability of arms with potentially disparate baseline characteristics. A related point is that it is often difficult to ensure that cohorts are comparable across studies, as the intent-to-treat clinical trial design guarantees only within-trial comparisons. Moreover, considering the rising trend of combination therapies, it will be crucial to investigate the PBMF's performance across various arms and their pairwise combinations. As our study is retrospective in nature, an important next step would be to validate the PBMF prospectively in a future clinical study. Finally, future work can explore the tradeoff between data maturity, ability to extract a predictive signal, and phase 3 trial investment decision timing.

Our benchmarks nonetheless demonstrate that with the availability of the appropriate data, the PBMF could nominate a predictive biomarker that is likely to outperform the original study design in selecting patients who would derive greater benefit from the new treatment in a phase 3 study. The use of the PBMF has the potential to improve strategies for patient selection over what can be achieved with conventional study designs.

Limitations of the study

Despite its strengths, the PBMF has limitations that are common to many biomarker discovery methods. First, there is no guarantee that a predictive signal exists among the available features in a given cohort. Indeed, many well-established clinical datasets for survival analysis contain only age and/or sex features, and only prognostic biomarkers can be found with any modeling approach. Related to the known challenge of limited datasets and high heterogeneity in patient populations, the PBMF cannot be used to determine whether the data are adequate and representative of the target population and biology. Nevertheless, it is noteworthy that the PBMF demonstrated superior performance in scenarios with small data sizes. In situations with substantial data, PBMF scaled with data size, whereas the performance of the VT method reached a plateau. Second, the ensemble PBMF may be unable to maintain its magnitude of predictive power when distilled into a simple model, as there is often a tradeoff between a biomarker's predictive power and its parsimony.[61] However, the enhanced interpretability of the model may contribute to a better understanding of the biological factors underpinning the predictive signal of the biomarker. More generally, with any biomarker nomination process, there is the risk of overfitting to the training data and lack of generalization when the biomarker is deployed prospectively. Encouragingly, at least within the scope of the current study, the PBMF provided concordant results between training and test sets and to a greater degree than the compared methods. Third, while the PBMF outperformed other methods in discerning predictive signals from noisy or prognostic features, we might still find that strongly prognostic features can impede the identification of predictive signals, and therefore our method could potentially gain more from prior feature selection. Fourth, the PBMF's contrastive loss function formulation tends to attenuate the discovery of biomarkers that show a modest positive effect in the control treatment but a more substantial benefit in the treatment of interest. Finally, the PBMF is a discovery tool, and any biomarker hypothesis requires prospective clinical validation.[62–64] Acknowledging these limitations, we recommend best practices for hyperparameters and usage of the PBMF in Table S6.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for materials should be addressed to and will be fulfilled by Etai Jacob etai.jacob@astrazeneca.com.

Materials availability

This study did not generate new unique reagents.

Data and code availability

The original clinical trial papers cited in our main text provide details on study design and patient demographics. With the exception of Tempus and POSEIDON data, patient-level and biomarker data necessary for our analyses were only available in follow-up publications. The data used for our analyses were obtained from the following sources. Data for breast cancer cohorts is available from the textbook by Royston and Sauerbrei[25] at the following URL: https://www.uniklinik-freiburg.de/imbi/stud-le/multivariable-model-building.html. Data for diabetic retinopathy cohort is available within the R survival package.[65] Data for POPLAR and OAK studies was accessed from Gandara et al.[51] Data from Tempus may be purchased for use (https://www.tempus.com). IMmotion150, IMVigor210 and IMVigor211 data can be obtained directly from Yuen et al.[66] supplementary material. CheckMate data can be downloaded from the supplementary information from Braun et al.[67] publication. JAVELIN 101 Renal can be obtained directly from Motzer et al.[68] publication. POSEIDON data underlying the findings described in this manuscript may be obtained in accordance with AstraZeneca's data sharing policy described at https://astrazenecagrouptrials.pharmacm.com/ST/Submission/Disclosure. Data for studies directly listed on Vivli can be requested through Vivli at www.vivli.org. Data for studies not listed on Vivli could be requested through Vivli at https://vivli.org/members/enquiries-about-studies-not-listed-on-the-vivli-platform/. Code for the PBMF is publicly available on Zenodo: https://doi.org/10.5281/zenodo.14766044.

Any additional information required to reanalyze data reported in this paper is available from the lead contact.

ACKNOWLEDGMENTS

We thank J.C. Barrett and A. Meier for discussions of this work. We thank D.J. Shuman for editing help. This study was funded by AstraZeneca.

AUTHOR CONTRIBUTIONS

G.A.-A. contributed to the conception of the study. G.A-A., D.E.B., G.J.S., E.K., and E.J. contributed to the design of the study. G.A.-A., D.E.B., G.J.S., E.K., K.M.S., and E.J. contributed to algorithm development. G.A.-A., D.E.B., G.J.S., K.M.S., and S.C.P. contributed to analysis of the data. G.A.-A., D.E.B., G.J.S., E.Y.C., and E.J. wrote the manuscript. E.J. supervised the work.

DECLARATION OF INTERESTS

G.A.-A., D.E.B., G.J.S., E.K., K.M.S., E.Y.C., and E.J. are current or former employees of AstraZeneca with stock ownership, interests, and/or options in the company. S.C.P. was an employee of Tempus during the study period, with stock ownership, interests, and/or options in the company.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT-4o to proofread sections of the main text for grammatical flow and to offer suggestions for brevity and clarity. The authors reviewed and edited the suggestions and take full responsibility for the content of the publication.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Real-world and clinical data sets
- METHOD DETAILS
  - Predictive biomarkers, contrastive learning, and model architecture
  - Model implementation and extensions
  - Predictive biomarker loss function
  - Biomarker scoring
  - Feature and patient subsetting during model training
  - PBMF ensemble model pruning
  - Model distillation: pseudo-labeling
  - Model distillation: tree-based model explainability
  - VT implementation
  - SIDES implementation
  - Synthetic data generation
  - Creating uncorrelated covariate matrix for simulations
  - Optimizing virtual twins for simulations
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2025.03.029.

Received: August 6, 2024
Revised: December 20, 2024
Accepted: March 26, 2025
Published: April 17, 2025

REFERENCES

1. Ciardiello, F., Arnold, D., Casali, P.G., Cervantes, A., Douillard, J.Y., Eggermont, A., Eniu, A., McGregor, K., Peters, S., Piccart, M., et al. (2014). Delivering precision medicine in oncology today and in future-the promise and challenges of personalised cancer medicine: a position paper by the European Society for Medical Oncology (ESMO). Ann. Oncol. *25*, 1673–1678. https://doi.org/10.1093/annonc/mdu217.
2. Schwartzberg, L., Kim, E.S., Liu, D., and Schrag, D. (2017). Precision Oncology: Who, How, What, When, and When Not? Am. Soc. Clin. Oncol. Educ. Book. *37*, 160–169. https://doi.org/10.1200/EDBK_174176.
3. Goossens, N., Nakagawa, S., Sun, X., and Hoshida, Y. (2015). Cancer biomarker discovery and validation. Transl. Cancer Res. *4*, 256–269.
4. Oldenhuis, C.N.A.M., Oosting, S.F., Gietema, J.A., and de Vries, E.G.E. (2008). Prognostic versus predictive value of biomarkers in oncology. Eur. J. Cancer *44*, 946–953. https://doi.org/10.1016/j.ejca.2008.03.006.
5. Kantarjian, H., Sawyers, C., Hochhaus, A., Guilhot, F., Schiffer, C., Gambacorti-Passerini, C., Niederwieser, D., Resta, D., Capdeville, R., Zoellner, U., et al. (2002). Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. N. Engl. J. Med. *346*, 645–652. https://doi.org/10.1056/NEJMoa011573.
6. (2021). Clinical Development Success Rates and Contributing Factors 2011–2020. Biotechnology Innovation Organization, Informa Pharma Intelligence, and QLS Advisors. https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf.
7. Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R., and Schacht, A.L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat. Rev. Drug Discov. *9*, 203–214. https://doi.org/10.1038/nrd3078.
8. McDermott, J.E., Wang, J., Mitchell, H., Webb-Robertson, B.J., Hafen, R., Ramey, J., and Rodland, K.D. (2013). Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. Expert Opin. Med. Diagn. *7*, 37–51. https://doi.org/10.1517/17530059.2012.718329.
9. Cox, D.R. (1972). Regression Models and Life-Tables. J. Roy. Stat. Soc. B *34*, 187–202.
10. Doroshow, D.B., Bhalla, S., Beasley, M.B., Sholl, L.M., Kerr, K.M., Gnjatic, S., Wistuba, I.I., Rimm, D.L., Tsao, M.S., and Hirsch, F.R. (2021). PD-L1 as a biomarker of response to immune-checkpoint inhibitors. Nat. Rev. Clin. Oncol. *18*, 345–362. https://doi.org/10.1038/s41571-021-00473-5.
11. Marabelle, A., Le, D.T., Ascierto, P.A., Di Giacomo, A.M., De Jesus-Acosta, A., Delord, J.P., Geva, R., Gottfried, M., Penel, N., Hansen, A.R., et al. (2020). Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study. J. Clin. Oncol. *38*, 1–10. https://doi.org/10.1200/jco.19.02105.
12. Marabelle, A., Fakih, M., Lopez, J., Shah, M., Shapira-Frommer, R., Nakagawa, K., Chung, H.C., Kindler, H.L., Lopez-Martin, J.A., Miller, W.H., et al. (2020). Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. Lancet Oncol. *21*, 1353–1365. https://doi.org/10.1016/S1470-2045(20)30445-9.
13. Havel, J.J., Chowell, D., and Chan, T.A. (2019). The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat. Rev. Cancer *19*, 133–150. https://doi.org/10.1038/s41568-019-0116-x.
14. Chang, T.G., Cao, Y., Sfreddo, H.J., Dhruba, S.R., Lee, S.H., Valero, C., Yoo, S.K., Chowell, D., Morris, L.G.T., and Ruppin, E. (2024). LORIS robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features. Nat. Can. (Ott.) *5*, 1158–1175. https://doi.org/10.1038/s43018-024-00772-7.
15. Chowell, D., Yoo, S.-K., Valero, C., Pastore, A., Krishna, C., Lee, M., Hoen, D., Shi, H., Kelly, D.W., Patel, N., et al. (2022). Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. Nat. Biotechnol. *40*, 499–506.
16. Litchfield, K., Reading, J.L., Puttick, C., Thakkar, K., Abbosh, C., Bentham, R., Watkins, T.B.K., Rosenthal, R., Biswas, D., Rowan, A., et al. (2021). Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cellule *184*, 596–614.e14. https://doi.org/10.1016/j.cell.2021.01.002.
17. Martens, A., Wistuba-Hamprecht, K., Geukes Foppen, M., Yuan, J., Postow, M.A., Wong, P., Romano, E., Khammari, A., Dreno, B., Capone, M., et al. (2016). Baseline Peripheral Blood Biomarkers Associated with Clinical Outcome of Advanced Melanoma Patients Treated with Ipilimumab. Clin. Cancer Res. *22*, 2908–2918. https://doi.org/10.1158/1078-0432.CCR-15-2412.
18. Swami, U., Nussenzveig, R.H., and Agarwal, N. (2020). Quest for Ideal Composite Biomarkers for Response to Immunotherapies. Clin. Cancer Res. *26*, 5059–5061. https://doi.org/10.1158/1078-0432.CCR-20-2321.
19. Loh, W.-Y., Cao, L., and Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. WIREs Data Min. &. Knowl. *9*, e1326. https://doi.org/10.1002/widm.1326.
20. Alemayehu, D., Chen, Y., and Markatou, M. (2018). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. Stat. Methods Med. Res. *27*, 3658–3678. https://doi.org/10.1177/0962280217710570.
21. Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat. Med. *30*, 2601–2621. https://doi.org/10.1002/sim.4289.
22. Foster, J.C., Taylor, J.M.G., and Ruberg, S.J. (2011). Subgroup identification from randomized clinical trial data. Stat. Med. *30*, 2867–2880.
23. Therneau, T.M. (2024). A Package for Survival Analysis in (R). https://CRAN.R-project.org/package=survival.
24. Therneau, T.M., and Grambsh, P.M. (2000). Modeling Survival Data: Extending the Cox Model (Springer).
25. Royston, P., and Sauerbrei, W. (2008). Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables (John Wiley & Sons, Ltd).
26. Foekens, J.A., Peters, H.A., Look, M.P., Portengen, H., Schmitt, M., Kramer, M.D., Brünner, N., Jänicke, F., Meijer-van Gelder, M.E., Henzen-Logmans, S.C., et al. (2000). The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. Cancer Res. *60*, 636–643.
27. Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R.L., and Rauschecker, H.F. (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. J. Clin. Oncol. *12*, 2086–2093. https://doi.org/10.1200/JCO.1994.12.10.2086.
28. Blair, A.L., Hadden, D.R., Weaver, J.A., Archer, D.B., Johnston, P.B., and Maguire, C.J. (1980). The 5-year prognosis for vision in diabetes. Ulst. Med. J. *49*, 139–147.
29. Motzer, R.J., Penkov, K., Haanen, J., Rini, B., Albiges, L., Campbell, M.T., Venugopal, B., Kollmannsberger, C., Negrier, S., Uemura, M., et al. (2019). Avelumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma. N. Engl. J. Med. *380*, 1103–1115. https://doi.org/10.1056/NEJMoa1816047.

30. McDermott, D.F., Huseni, M.A., Atkins, M.B., Motzer, R.J., Rini, B.I., Escudier, B., Fong, L., Joseph, R.W., Pal, S.K., Reeves, J.A., et al. (2018). Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. Nat. Med. *24*, 749–757. https://doi.org/10.1038/s41591-018-0053-3.

31. Johnson, M.L., Cho, B.C., Luft, A., Alatorre-Alexander, J., Geater, S.L., Laktionov, K., Kim, S.W., Ursol, G., Hussein, M., Lim, F.L., et al. (2023). Durvalumab With or Without Tremelimumab in Combination With Chemotherapy as First-Line Therapy for Metastatic Non-Small-Cell Lung Cancer: The Phase III POSEIDON Study. J. Clin. Oncol. *41*, 1213–1227. https://doi.org/10.1200/JCO.22.00975.

32. Liu, F., and Panagiotakos, D. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med. Res. Methodol. *22*, 287. https://doi.org/10.1186/s12874-022-01768-6.

33. Zisis, K., Pavi, E., Geitona, M., and Athanasakis, K. (2024). Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. J. Pharm. Pharmaceut. Sci. *27*, 12302. https://doi.org/10.3389/jpps.2024.12302.

34. (2022). Master protocols: efficient clinical trial design strategies to expedite development of oncology drugs and biologics. Guidance for Industry. U.S. Department of Health and Human Services (Food and Drug Administration).

35. (2023). Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products. Guidance for Industry. U.S. Department of Health and Human Services (Food and Drug Administration).

36. Motzer, R.J., Rini, B.I., McDermott, D.F., Redman, B.G., Kuzel, T.M., Harrison, M.R., Vaishampayan, U.N., Drabkin, H.A., George, S., Logan, T.F., et al. (2015). Nivolumab for Metastatic Renal Cell Carcinoma: Results of a Randomized Phase II Trial. J. Clin. Oncol. *33*, 1430–1437. https://doi.org/10.1200/JCO.2014.59.0703.

37. Motzer, R.J., Escudier, B., McDermott, D.F., George, S., Hammers, H.J., Srinivas, S., Tykodi, S.S., Sosman, J.A., Procopio, G., Plimack, E.R., et al. (2015). Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. N. Engl. J. Med. *373*, 1803–1813. https://doi.org/10.1056/NEJMoa1510665.

38. Choueiri, T.K., Fishman, M.N., Escudier, B., McDermott, D.F., Drake, C.G., Kluger, H., Stadler, W.M., Perez-Gracia, J.L., McNeel, D.G., Curti, B., et al. (2016). Immunomodulatory Activity of Nivolumab in Metastatic Renal Cell Carcinoma. Clin. Cancer Res. *22*, 5461–5471. https://doi.org/10.1158/1078-0432.CCR-15-2839.

39. Balar, A.V., Galsky, M.D., Rosenberg, J.E., Powles, T., Petrylak, D.P., Bellmunt, J., Loriot, Y., Necchi, A., Hoffman-Censits, J., Perez-Gracia, J.L., et al. (2017). Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. Lancet *389*, 67–76. https://doi.org/10.1016/S0140-6736(16)32455-2.

40. Powles, T., Durán, I., van der Heijden, M.S., Loriot, Y., Vogelzang, N.J., De Giorgi, U., Oudard, S., Retz, M.M., Castellano, D., Bamias, A., et al. (2018). Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. Lancet *391*, 748–757. https://doi.org/10.1016/S0140-6736(17)33297-X.

41. Fehrenbacher, L., Spira, A., Ballinger, M., Kowanetz, M., Vansteenkiste, J., Mazieres, J., Park, K., Smith, D., Artal-Cortes, A., Lewanski, C., et al. (2016). Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. Lancet *387*, 1837–1846.

42. Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., Von Pawel, J., Gadgeel, S.M., Hida, T., Kowalski, D.M., Dols, M.C., et al. (2017). Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Lancet *389*, 255–265.

43. Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., Osokin, N., Kozlov, I., Frenkel, F., Gancharova, O., et al. (2021). Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer. Cell. *39*, 845–865.e7. https://doi.org/10.1016/j.ccell.2021.04.014.

44. DeNardo, D.G., and Ruffell, B. (2019). Macrophages as regulators of tumour immunity and immunotherapy. Nat. Rev. Immunol. *19*, 369–382. https://doi.org/10.1038/s41577-019-0127-6.

45. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

46. Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D.R., Albright, A., Cheng, J.D., Kang, S.P., Shankaran, V., et al. (2017). IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. J. Clin. Investig. *127*, 2930–2940. https://doi.org/10.1172/JCI91190.

47. Wang, G., Chow, R.D., Zhu, L., Bai, Z., Ye, L., Zhang, F., Renauer, P.A., Dong, M.B., Dai, X., Zhang, X., et al. (2020). CRISPR-GEMM Pooled Mutagenic Screening Identifies KMT2D as a Major Modulator of Immune Checkpoint Blockade. Cancer Discov. *10*, 1912–1933. https://doi.org/10.1158/2159-8290.CD-19-1448.

48. Guan, X., Cai, S., Wu, X., Chen, Y., Deng, H., Zhong, X., Chen, T., and Huang, M. (2021). 995P A pan-cancer analysis of KMT2D as a potential biomarker for immune checkpoint therapy. Ann. Oncol. *32*, S847. https://doi.org/10.1016/j.annonc.2021.08.1379.

49. Thu, K.L., and Yoon, J.Y. (2024). ATM-the gene at the moment in non-small cell lung cancer. Transl. Lung Cancer Res. *13*, 699–705. https://doi.org/10.21037/tlcr-23-853.

50. Vokes, N.I., Galan Cobo, A., Fernandez-Chas, M., Molkentine, D., Treviño, S., 3rd, Druker, V., Qian, Y., Patel, S., Schmidt, S., Hong, L., et al. (2023). ATM Mutations Associate with Distinct Co-Mutational Patterns and Therapeutic Vulnerabilities in NSCLC. Clin. Cancer Res. *29*, 4958–4972. https://doi.org/10.1158/1078-0432.CCR-23-1122.

51. Gandara, D.R., Paul, S.M., Kowanetz, M., Schleifman, E., Zou, W., Li, Y., Rittmeyer, A., Fehrenbacher, L., Otto, G., Malboeuf, C., et al. (2018). Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. Nat. Med. *24*, 1441–1448.

52. Wang, Z., Duan, J., Cai, S., Han, M., Dong, H., Zhao, J., Zhu, B., Wang, S., Zhuo, M., Sun, J., et al. (2019). Assessment of blood tumor mutational burden as a potential biomarker for immunotherapy in patients with non–small cell lung cancer with use of a next-generation sequencing cancer gene panel. JAMA Oncol. *5*, 696–702.

53. Kim, E.S., Velcheti, V., Mekhail, T., Yun, C., Shagan, S.M., Hu, S., Chae, Y.K., Leal, T.A., Dowell, J.E., Tsai, M.L., et al. (2022). Blood-based tumor mutational burden as a biomarker for atezolizumab in non-small cell lung cancer: the phase 2 B-F1RST trial. Nat. Med. *28*, 939–945.

54. Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions. J. Clin. Oncol. *29*, 4718–4719. https://doi.org/10.1200/JCO.2011.38.3729.

55. Royston, P., and Altman, D.G. (2013). External validation of a Cox prognostic model: principles and methods. BMC Med. Res. Methodol. *13*, 33.

56. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med. *30*, 850–862. https://doi.org/10.1038/s41591-024-02857-3.

57. Wagner, S.J., Reisenbuchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., et al. (2023). Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell *41*, 1650–1661. https://doi.org/10.1016/j.ccell.2023.08.002.

58. Arango-Argoty, G., Kipkogei, E., Stewart, R., Sun, G.J., Patra, A., Kagiampakis, I., and Jacob, E. (2025). Pretrained transformers applied to clinical studies improve predictions of treatment efficacy and associated biomarkers. Nat. Commun. *16*, 2101. https://doi.org/10.1038/s41467-025-57181-2.

59. Sun, G.J., Arango-Argoty, G., Doherty, G.J., Bikiel, D.E., Pavlovic, D., Chen, A.C., Stewart, R.A., Lai, Z., and Jacob, E. (2024). Machine learning modeling of patient health signals informs long-term survival on immune checkpoint inhibitor therapy. iScience *27*, 110634. https://doi.org/10.1016/j.isci.2024.110634.

60. Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika *70*, 41–55. https://doi.org/10.1093/biomet/70.1.41.

61. Harrell, F.E.J. (2023). Biostatistics for Biomedical Research. https://hbiostat.org/bbr/.

62. Sun, X., Briel, M., Walter, S.D., and Guyatt, G.H. (2010). Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. Br. Med. J. *340*, c117. https://doi.org/10.1136/bmj.c117.

63. Dmitrienko, A., Muysers, C., Fritsch, A., and Lipkovich, I. (2016). General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. J. Biopharm. Stat. *26*, 71–98. https://doi.org/10.1080/10543406.2015.1092033.

64. Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., and Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. J. Biopharm. Stat. *26*, 99–119. https://doi.org/10.1080/10543406.2015.1092034.

65. Drysdale, E. (2022). SurvSet: An open-source time-to-event dataset repository. Preprint at arXiv. https://doi.org/10.48550/arXiv.2203.03094.

66. Yuen, K.C., Liu, L.F., Gupta, V., Madireddi, S., Keerthivasan, S., Li, C., Rishipathak, D., Williams, P., Kadel, E.E., 3rd, Koeppen, H., et al. (2020). High systemic and tumor-associated IL-8 correlates with reduced clinical benefit of PD-L1 blockade. Nat. Med. *26*, 693–698. https://doi.org/10.1038/s41591-020-0860-1.

67. Braun, D.A., Hou, Y., Bakouny, Z., Ficial, M., Sant' Angelo, M., Forman, J., Ross-Macdonald, P., Berger, A.C., Jegede, O.A., Elagina, L., et al. (2020). Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. Nat. Med. *26*, 909–918. https://doi.org/10.1038/s41591-020-0839-y.

68. Motzer, R.J., Robbins, P.B., Powles, T., Albiges, L., Haanen, J.B., Larkin, J., Mu, X.J., Ching, K.A., Uemura, M., Pal, S.K., et al. (2020). Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial. Nat. Med. *26*, 1733–1741. https://doi.org/10.1038/s41591-020-1044-8.

69. Michuda, J., Breschi, A., Kapilivsky, J., Manghnani, K., McCarter, C., Hockenberry, A.J., Mineo, B., Igartua, C., Dudley, J.T., Stumpe, M.C., et al. (2023). Validation of a Transcriptome-Based Assay for Classifying Cancers of Unknown Primary Origin. Mol. Diagn. Ther. *27*, 499–511. https://doi.org/10.1007/s40291-023-00650-5.

70. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

71. Breiman, L. (1996). Bagging Predictors. Mach. Learn. *24*, 123–140. https://doi.org/10.1023/A:1018054314350.

72. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations (PMLR), pp. 1597–1607.

73. van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1807.03748.

74. Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. (2020). Debiased contrastive learning. Adv. Neural. Inf. Process. Syst. *33*, 8765–8775.

75. Woolson, R.F. (1981). Rank tests and a one-sample logrank test for comparing observed survival data to a standard population. Biometrics *37*, 687–696.

76. Meier, A., Nekolla, K., Hewitt, L.C., Earle, S., Yoshikawa, T., Oshima, T., Miyagi, Y., Huss, R., Schmidt, G., and Grabsch, H.I. (2020). Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. J. Pathol. Clin. Res. *6*, 273–282.

77. Monti, S., Tamayo, P., Mesirov, J.P., and Golub, T.R. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Mach. Learn. *52*, 91–118.

78. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

79. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and Lauer, M.S. (2008). Random Survival Forests (Wiley StatsRef: Statistics Reference Online). https://doi.org/10.1002/9781118445112.stat08188.

80. Fotso, S. (2019). PySurvival: open source package for survival analysis modeling. https://www.pysurvival.io.

81. Crowther, M.J., and Lambert, P.C. (2013). Simulating biologically plausible complex survival data. Stat. Med. *32*, 4118–4134. https://doi.org/10.1002/sim.5823.

82. Davidson-Pilon, C. (2019). lifelines: survival analysis in Python. J. Open Source Softw. *4*, 1317. https://doi.org/10.21105/joss.01317.

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Rotterdam breast cancer cohort	Royston and Sauerbrei[25]	https://www.uniklinik-freiburg.de/fileadmin/mediapool/08_institute/biometrie-statistik/Dateien/Studium_und_Lehre/Lehrbuecher/Multivariable_Model-building/rotterdam_br_ca.zip
German breast cancer study group cohort	Royston and Sauerbrei[25]	https://www.uniklinik-freiburg.de/fileadmin/mediapool/08_institute/biometrie-statistik/Dateien/Studium_und_Lehre/Lehrbuecher/Multivariable_Model-building/gbsg_br_ca.zip
Diabetic retinopathy	Drysdale[65]	https://github.com/ErikinBC/SurvSet/blob/main/SurvSet/_datagen/output/retinopathy.csv
POSEIDON	Johnson et al.[31]	Data may be obtained in accordance with AstraZeneca's data sharing policy described at https://astrazenecagrouptrials.pharmacm.com/ST/Submission/Disclosure. Data for POSEIDON (NCT03164616) can be requested through Vivli at https://vivli.org/members/enquiries-about-studies-not-listed-on-the-vivli-platform/
JAVELIN 101	Motzer et al.[68]	Table S15; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-020-1044-8/MediaObjects/41591_2020_1044_MOESM3_ESM.xlsx
IMmotion 150	Yuen et al.[66]	Table S3; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-020-0860-1/MediaObjects/41591_2020_860_MOESM1_ESM.xlsx
Tempus	Tempus	AZ-Tempus NSCLC cohort; https://www.tempus.com
CheckMate 009	Braun et al.[67]	Table S1 and 4C; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-020-0839-y/MediaObjects/41591_2020_839_MOESM2_ESM.xlsx
CheckMate 010	Braun et al.[67]	Table S1 and 4C; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-020-0839-y/MediaObjects/41591_2020_839_MOESM2_ESM.xlsx
CheckMate 025	Braun et al.[67]	Table S1 and 4C; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-020-0839-y/MediaObjects/41591_2020_839_MOESM2_ESM.xlsx
IMvigor 210	Yuen et al.[66]	Table S3; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-020-0860-1/MediaObjects/41591_2020_860_MOESM1_ESM.xlsx
IMvigor 211	Yuen et al.[66]	Table S3; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-020-0860-1/MediaObjects/41591_2020_860_MOESM1_ESM.xlsx
POPLAR	Gandara et al.[51]	Table S8; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-018-0134-3/MediaObjects/41591_2018_134_MOESM3_ESM.xlsx
OAK	Gandara et al.[51]	Table S8; https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-018-0134-3/MediaObjects/41591_2018_134_MOESM3_ESM.xlsx
Software and algorithms		
Python	Python Software Foundation	Version 3.6.9; RRID: SCR_008394; http://www.python.org/
scikit-learn	Pedregosa et al.[78]	Version 0.24.1; RRID: SCR_002577; http://scikit-learn.org/
Lifelines	Davidson-Pilon[82]	Version 0.26.0; RRID: SCR_024899; https://lifelines.readthedocs.io/
Tensorflow	Google	Version 2.6.0; RRID: SCR_016345; https://www.tensorflow.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SIDES	Lipkovich et al.[21]	Version 12.21.2018; https://drive.google.com/file/d/1auc3GOspjWYFCgOuMgUfrqZwCHHkZdHy/view (previously available on https://biopharmnet.com/subgroup-analysis-software/)
pysurvival	Fotso et al.[80]	Version 0.1.2; https://square.github.io/pysurvival/
PBMF	This paper	Version 1.0.0; https://doi.org/10.5281/zenodo.14766044

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Real-world and clinical data sets

Original clinical trial publications provide details on study design and patient demographics. With the exception of Tempus and POSEIDON data, patient-level and biomarker data required for our analyses were only available in follow-up publications. Therefore, we have cited the original clinical trial papers to acknowledge the foundational studies and their design, while also referencing the follow-up publications where we obtained the necessary data for our analyses.

For clinical data derived from public sources (Rotterdam and German study cohorts, Diabetic retinopathy, JAVELIN Renal 101 [NCT02684006], IMmotion150[NCT01984242], CheckMate 09-010-025[NCT01358721, NCT01354431, NCT01668784], IMvigor210-211[NCT02108652, NCT02302807], POPLAR[NCT01903993], and OAK[NCT02008227]) study designs (where applicable) and information (where available) related to age, sex, ancestry, race, ethnicity, and socioeconomic status can be found in publications cited in Table S3. For POSEIDON (NCT03164616), such demographic information is also available in its source publication (Table S4). The POSEIDON protocol and any amendments were approved by Institutional Review Boards or Ethics Committees of participating centers, and all patients provided written informed consent. For TEMPUS, all Tempus data were de-identified in accordance to the Health Insurance Portability and Accountability Act (HIPAA). Available demographic information is listed in Table S3.

Prior to any modeling, features were selected as specified in Table S3. Hyperparameters (Table S2) were tuned for the PBMF, VT, and SIDES for each clinical dataset, using only training data. Total sample numbers for all studies are summarized in Table 1.

The Rotterdam breast cancer cohort[26] (863 patients) was used as a training data set, and the German breast cancer study cohort[27] (686 patients) was used as a test data set. We selected only patients treated with hormone-based treatments and chemotherapy. The 7 features used for training the PBMF are age, menopause, tumor size, tumor grade, number of nodes, pr (progesterone receptor status), and er (estrogen receptor status). We trained the model using overall survival and death. Data were downloaded from Royston and Sauerbrei.[25]

The diabetic RETINOPATHY study[28] evaluates the treatment of laser coagulation to delay diabetic retinopathy. In this study, 197 patients underwent treatment in one eye, while the other eye remained untreated. The treatment eye, right or left, was randomized. Treating each eye as an individual sample resulted in 394 observations in the dataset. The event of interest was the time from the start of treatment to the time when visual acuity dropped below 5/200 for two visits in a row. Censoring was caused by death, dropout, or the end of the study. Age of diabetes onset, diabetes type, and risk score were included as the features of this dataset. Diabetes type was a binary feature indicating juvenile diabetes (diagnosis before age 20) or adult. Risk score was defined by the Diabetic Retinopathy Study, and a score greater than 6 out of 12 indicates high risk. Only individuals with risk score ≥ 6 were available in the dataset. The dataset was split into training and testing at a prevalence of 50% (random seed = 0). Data were downloaded from SurvSet.[65]

The randomized phase 2 clinical trial IMmotion150[30] evaluated the efficacy of atezolizumab (anti-PD-L1) alone or in combination with bevacizumab (anti-VEGF) versus sunitinib (RTK inhibitor) in treatment-naive metastatic renal cell carcinoma (mRCC). Data from IMmotion150 was downloaded from Yuen et al.[66] and comprised a total of 248 patients with no missing values (84 atezolizumab, 81 sunitinib, and 83 atezolizumab + bevacizumab). Available features on this dataset: age, sex, liver metastasis, previous nephrectomy, T-cell effector signature score (binarized into high vs. low via median cutoff), Plasma IL8, SLD (sum of longest tumor diameter) and sample type (primary / metastatic). IMmotion150 dataset was split into training / testing with a 50% prevalence, stratified by treatment and overall survival event (random seed = 0). The PBMF was trained to discriminate between atezolizumab + bevacizumab against sunitinib using overall survival time and event as endpoints.

The JAVELIN Renal 101 trial[29] evaluated the effectiveness of avelumab (PD-L1) plus axitinib (chemotherapy) versus sunitinib in advanced renal cell carcinoma (aRCC). Clinical response, PD-L1 status and RNA derived signatures (pathway scores) were downloaded from the biomarker analysis publication reported by Motzer et al.[68] A total of 59 signatures were used, including tumor microenvironment-derived signatures (e.g., T-cells, B-cells, Macrophages), pathway-derived signatures (e.g., cell cycle, lipid metabolism, cell-cell signaling), PD-L1 status, and the 26-gene signature (B9991003_Javelin_Renal_101_genes26) reported in the publication (Table S4). In total 726 patients (372 sunitinib, 354 avelumab+axitinib) were retrieved. The data was split into training and testing with a 50% prevalence (random seed = 0) stratified by treatment and survival event. The PBMF was trained to identify a sub-population predictive of avelumab+axitinib against sunitinib using progressive free survival time and event as endpoints.

POSEIDON[31] is a phase 3 randomized clinical trial that evaluated the efficacy of durvalumab plus tremelimumab plus chemotherapy and durvalumab plus chemotherapy against chemotherapy alone in first-line metastatic non-small-cell lung cancer (mNSCLC).[31] In this study, we focused on peripheral blood RNA seq data for durvalumab + chemotherapy (114 patients) and chemotherapy alone

(114 patients) treatment arms. RNA seq data was Log2(TPM+0.001) transformed, and we extracted a set of custom and publicly available tumor microenvironment-related signatures[43] (Table S4) using the median score across genes. Dataset is split into training / testing with a 50% prevalence (random seed = 0) stratified by treatment and event. PBMF was trained using to identify predictive biomarker of durvalumab + chemotherapy against chemotherapy alone using overall survival time and event as endpoints.

Data from the Tempus NSCLC cohort were selected from the Tempus deidentified multimodal database.[69] Patients were included if they were diagnosed with a primary or metastatic NSCLC diagnosis on or after 2016, confirmed by histology, and received chemotherapy or ICIs as first treatment. For these patients, real-world overall survival was calculated using treatment start date as the index date. RNA expression (batch-corrected and transformed to transcripts per million) data was obtained for pre-treatment samples. In the case of patients with multiple biopsies, only the closest one to treatment start date was selected. ssGSEA (corto R package) was run per RNA sample for the 50 cancer hallmark gene sets (msigDB C5).[45,70] A total of 201 patients with stage 4 NSCLC undergoing chemotherapy (84) or immunotherapy (117) were selected. The data set was equally split into training and testing (50% each) and stratified by treatment (random seed = 0). The training set had 42 patients with chemotherapy and 58 with immunooncology treatment; and the testing set had 42 patients with chemotherapy and 59 with immunooncology treatment. We used overall survival and death as endpoints for training the PBMF model.

The POPLAR[41] and OAK[42] clinical trials were used to represent phases 2 and 3, respectively, to evaluate the efficacy of atezolizumab as a second-line therapy for patients unresponsive to first-line platinum-based chemotherapy in the NSCLC population. The therapeutic potential of atezolizumab was compared against that of docetaxel. The dataset, sourced from Gandara et al.,[51] encompasses ctDNA from blood samples in addition to patient demographics and clinical biomarkers, as detailed in Table S4. We conducted a prevalence-based ranking of ctDNA genes from patients in the POPLAR trial, identifying the top 20 genes that exhibit a minimum prevalence of 20% across the combined data set from both atezolizumab and docetaxel cohorts. The PBMF was not trained by using progression-free survival, and this outcome was used for testing only. POPLAR trial data were used for training the PBMF, and OAK was used for independent evaluation. We used the overall survival time and event as endpoints.

The CheckMate prospective clinical trials 009,[38] 010,[36] and 025[37] were designed to evaluate the efficacy of nivolumab (PD-1 blockade) against everolimus (mTOR inhibition) in advanced clear cell renal carcinoma (ccRCC). RNA sequencing (RNA-seq) and whole-exome sequencing (WES) derived features were obtained from Braun et al.[67] The PBMF was trained using the phase 2 CheckMate 010 clinical trial data and validated on the combined populations of CheckMate 025 and CheckMate 009. We included only patients with a complete set of features, excluding any with missing data. Consequently, 199 patients out of the available 311 had all complete features. Among these, 25 patients were from the Phase 2 (CheckMate 010) clinical trial. As CheckMate 010 did not have a control arm, we randomly selected 25 patients from the CheckMate 025 everolimus arm to match the number of patients treated with nivolumab. The remaining patients from CheckMate 009 and the Phase 3 CheckMate 025 trial were utilized for independent validation (i.e. test data set). Overall survival time and event status were used as endpoints for training the PBMF. The complete list of features used for training are shown in the Table S4.

IMvigor210[39] is a single-arm phase 2 clinical trial evaluating the efficacy of atezolizumab as a first (1L) or second (2+) line of treatment in locally advanced or metastatic urothelial carcinoma (mUC). IMvigor211[40] is a randomized phase 3 clinical trial that evaluated the efficacy of atezolizumab compared to chemotherapy in metastatic urothelial carcinoma as a second (2+) line of treatment. Data from IMvigor210 and 211 was downloaded from supplementary material of Yuen et al.[66] Both studies reported a total of 1222 patients. We only kept patients without missing values and filtered out all patients that were treated with Atezo as a first line of treatment in order to match the phase 3 (IMvigor211) population. In total we obtained 691 patients (422 atezolzumab and 269 chemotherapy). For training, we selected all patients from the IMvigor210 atezolizumab arm. As control, we selected 100 patients from the chemotherapy arm from the IMvigor211 phase 3 trial. For test data, we used all the patients on the phase 3 (IMvigor211), except the patients from chemotherapy that were used during training. The features in these cohorts include: age, sex, liver metastasis, ECOG, plasma IL8 at baseline (C1D1) and after treatment IL8 (C3D1) as well as plasma IL8 ratio (C3D1/C1D1). Therefore, this analysis is not limited to baseline measurements as on-treatment increased expression of plasma IL8 are known to be predictive of worse overall survival for atezolizumab and not for chemotherapy.[66] The PBMF was trained to identify predictive biomarkers of atezolizumab against chemotherapy using overall survival time and event in the IMvigor210 cohort and validated on the IMvigor211 trial.

METHOD DETAILS

Predictive biomarkers, contrastive learning, and model architecture

We define a predictive biomarker, B, as a tool categorizing a population into positive (B+) or negative (B–) for the biomarker, specific to a given treatment. B can encompass various patient measurements (e.g., age, blood counts, RNA gene expression). The biomarker is predictive if the B+ subpopulation is selectively enriched for individuals benefitting from a treatment of interest (“treatment”), but not a comparator one (“control”; Figure 1D, top). Similarly, the B– subpopulation should be selectively enriched for those not benefiting from any treatment, or perhaps benefiting instead from a comparator. In contrast, a prognostic biomarker is characterized by similar benefit irrespective of treatment.

With this definition, we formulated the PBMF to distinguish between two patient populations based on their differential response to treatments, i.e. contrastive learning. Specifically, the training objective of the PBMF (i.e. its loss function) actively maximizes the differences in outcomes for a given treatment (similar to pushing apart dissimilar items in contrastive learning) for B+ versus B– patients. Simultaneously, it minimizes the differences in outcomes for the control arm (similar to bringing similar items closer in contrastive

learning). By doing so, the network is trained to contrast the effects of two treatments across the biomarker-defined groups, effectively learning the distinctive features that separate patient responses. More formally from a technical perspective, the loss function is defined as the ratio between control and treatment log-rank test statistics (Figure 1A; see STAR Methods section "predictive biomarker loss function"). In plain terms, this has the effect of maximizing the separation of survival curves (or generally, for any time-to-event curves) between B+ and B– in the subpopulation receiving the treatment (i.e. large log-rank test statistic) while minimizing the separation for the subpopulation receiving the control. The model therefore optimizes for predictive biomarker behavior (Figures 1A and 1D). For applications requiring a particular biomarker prevalence, the PBMF can be run with an optional constraint (specifically, a penalization term) to encourage a predefined B+ prevalence proportion.

We designed the PBMF to be flexible and usable by the technical community (via an application programming interface). In particular, its modular design allows use of any neural network-based machine learning model, including deep, convolutional, and attention-based networks. The PBMF can use data from any modality (e.g., genomics, clinical, imaging), without restriction on the number or type (e.g., categorical or continuous). The PBMF outputs a "confidence" (i.e. probability) score from 0 to 1, which can be used (strictly speaking as a likelihood) to assign a sample to the B+ or B– subpopulation.

Model implementation and extensions

Overfitting poses a significant challenge in biomarker discovery, due to heterogeneity in patient populations and large numbers of features, particularly when attempting to predict the efficacy of one treatment over another rather than that of a single treatment. The PBMF therefore incorporates an established solution to increase model robustness by allowing training of a diverse collection of models (i.e. M independently trained neural networks), also known as an ensemble (Figure 1A), and then aggregating the ensemble predictions to yield a better prediction than any ensemble constituent. Model diversity is achieved by allowing each model to learn with a unique random subset of samples and features (akin to the machine learning principle of bagging[71]; Table S2). Following model training, we provide a solution whereby one can optionally remove poor performing models in the ensemble, i.e. model pruning, which can further enhance ensemble performance (Figure 1A).

Finally, an opaque neural network in the PBMF-generated biomarker may compromise confidence and hinder applicability in clinical settings. To address this, the PBMF incorporates an optional pipeline for simplifying the model ('model distillation' or 'knowledge distillation') into a parsimonious, interpretable decision tree. This is achieved by training a decision tree classifier on the subset of samples for which the ensemble had the highest confidence scores (Figure 3A). This decision tree thus transforms the candidate predictive biomarker into a simple set of rules, facilitating seamless integration into the design of future clinical studies (Figures 4A–4E, and 4F).

To facilitate usability, recommendations for PBMF hyperparameters are provided in Table S6.

Predictive biomarker loss function

The PBMF (Figure 1A) uses as input time-to-event data with censoring, a treatment label, and a feature matrix (S patients by F features). The feature matrix $X \in \mathbb{R}^F$ is used as the input to a fully connected neural network of user-defined depth and width. The PBMF was implemented in Tensorflow (https://www.tensorflow.org/).

The goal of the neural network is to assign patients to either the B+ or B– group. To refine this categorization, we employed a contrastive learning approach in which patients in the B+ group, when under treatment, show an improvement in survival times compared with those in the B– group. Conversely, in the control arm, the model aims to minimize the differences in survival times between the two biomarker groups according to the principle of contrastive learning.[72–74]

The distinction or similarity in survival times is quantified using log rank test statistics[75] within each treatment arm as follows:

$$TLogRank(a) = \frac{\left(E_a^+ - O_a^+\right)^2}{E_a^+} + \frac{\left(E_a^- - O_a^-\right)^2}{E_a^-}$$

where the E_a^+, E_a^- pair represents the expected number of events for the treatment a, under B+ and B–, respectively. The O_a^+, O_a^- pair depicts the observed events within the treatment a for B+ and B–, respectively.

Formally, the expected and observed events are defined as follows:

$$E_a^b = \sum_i^N B_i^b * I(A_i = a) * \lambda_i$$

$$O_a^b = \sum_i^N B_i^b * I(A_i = a) * I(C_i = 1)$$

$$\lambda_i = \sum_t \frac{\Omega_t}{N_t} I(T_i > t)$$

where the treatment arm is defined by $a \in \{Treatment\ (Tr), Control\ (CR)\}$ and the indicator function $I(A_i = a)$ determines whether the patient i is under treatment a or not. The biomarker group is defined by the output of the neural network where $b \in$ {positive (+),

negative (–)}. Therefore, each patient i has a probability of being labeled as being in the positive (B_i^+) or negative (B_i^-) group. C_i represents the censoring status of patient I, and λ_i is a scalar independent on the parameters of the neural network and can be precalculated (see Meier et al.[76]). Ω_t is the number of observed events at time t, and N_t is the number of subjects at risk at time t.

The log-rank test for the treatment and control is then defined as:

$$LR(Tr) = \frac{\left(\sum_i^N B_i^+ * I(A_i = Tr)[\lambda_i - I(C_i = 1)]\right)^2}{\sum_i^N B_i^+ * I(A_i = Tr) * \lambda_i} + \frac{\left(\sum_i^N B_i^- * I(A_i = Tr)[\lambda_i - I(C_i = 1)]\right)^2}{\sum_i^N B_i^- * I(A_i = Tr) * \lambda_i}$$

$$LR(Cr) = \frac{\left(\sum_i^N B_i^+ * I(A_i = Cr)[\lambda_i - I(C_i = 1)]\right)^2}{\sum_i^N B_i^+ * I(A_i = Cr) * \lambda_i} + \frac{\left(\sum_i^N B_i^- * I(A_i = Cr)[\lambda_i - I(C_i = 1)]\right)^2}{\sum_i^N B_i^- * I(A_i = Cr) * \lambda_i}$$

The contrastive nature of the loss function is evident in its formulation as follows:

- Treatment arm optimization: For patients receiving the actual treatment, the model maximizes the survival time difference between B+ and B– groups. This is quantified by the treatment log rank test score, $LR(Tr)$.
- Control arm optimization: For the control group, the model minimizes the survival time difference between the two biomarker groups. This is quantified by the control log rank test score, $LR(Cr)$.

The contrastive loss for the predictive biomarker is then defined as the ratio between the control log rank test score by the treatment log-rank test score:

$$loss_b = \frac{LR(Cr)}{LR(Tr)}.$$

The custom contrastive loss is the ratio of two log-rank tests computed over the time-to-event data, grouped by the treatment label, and stratified by the neural network output score. During optimization, the neural network learns a set of parameters that outputs scores to maximize the separation (i.e., larger log-rank test statistic) for the treatment while minimizing the separation (i.e., smaller log-rank test statistic) for the control. This ensures that the neural network will learn to generate a predictive biomarker score, since it will only stratify patients for a specific treatment.

We also integrated a population prevalence term to the loss to enable the model to identify a predictive biomarker given a specific desired minimal population (*minP*) such that:

$$prev(B^+) = \frac{\sum_i^N B_i^+}{\sum_i^N (B_i^+ + B_i^-)}$$

$$loss_p = \left(\frac{prev(B^+)}{minP} - 1\right)^2$$

The $loss_p$ will have a minimum value of 0 when *minP* is equal to the population of B^+. Finally, the composite PBMF loss function takes the following form:

$$Loss = \omega_1 * loss_b + \omega_2 * loss_p$$

where ω_1 and ω_2 dictate the contribution of each loss component. For example, when $\omega_2 = 0$, the PBMF finds a population with the best predictive power independent of the number of patients, and when $\omega_2 = 0.5$ the PBMF identifies a predictive biomarker of the treatment at a 50% patient prevalence.

Biomarker scoring

The output of the neural network ($B \in \mathbb{R}^2$) is composed of two units representing the B+ and B– scores $\{b^+, b^-\}$. Scores are then passed through a SoftMax activation to convert the network scores into probabilities. Thus, the biomarker scores for a given patient i can be expressed as:

$$B_i^+ = \frac{e^{b_i^+}}{e^{b_i^+} + e^{b_i^-}}, \; B_i^- = \frac{e^{b_i^-}}{e^{b_i^+} + e^{b_i^-}}$$

The probability of the negative biomarker can be written as $B^- = (1 - B^+)$. In this way, B^+ values close to 0 indicate B– and values close to 1 indicate B+. We assume the B+ to be contained within the neuron at index 0 from the output of the neural network. However, because the loss function does not have control of the directionality of the assignments, B+ can be arbitrary placed in neuron at the index 0 or 1. Therefore, after training and when making predictions, we corrected the B+ by computing the HR between the

B+ and B– within the treatment arm as $HR^{Treatment} = \frac{\sum O^+ / \sum O^-}{\sum E^+ / \sum E^-}$. Thus, an $HR^{Treatment} < 1$ defines the B+ in the neuron 0, whereas an $HR^{Treatment} > 1$ defines the biomarker positive in the neuron 1.

With ensemble of neural networks, for a given patient i and a total of M neural network models, we generated a set of scores $\{B^+_{i,1}, \ldots, B^+_{i,M}\}$ and computed a consensus score defined by the average score over all the models in the patient i such that $B^+_i = \frac{1}{M}\sum_{m=1}^{M} B^+_{i,m}$.

Since the final output of the PBMF assigns each sample to one of two groups (B+ and B–), the predictive biomarker identification task could be considered as a form of clustering. In this respect, our contrastive loss would serve as the clustering distance metric.

Feature and patient subsetting during model training

A random subset of patients and features can be specified (Table S2) to guard against model overfitting. Patient subsetting ('Ignore patients during loss computation, *ifrac*') is performed before model loss computation, and a different subset of patients will be excluded at each gradient update. Feature subsetting ('Use only *n* features [for each model if using an ensemble]') is performed before model training, and the given model will only train on the feature subset; when training an ensemble, each model will utilize its own unique random subset. During ensemble model evaluation, no patients or features are excluded.

PBMF ensemble model pruning

Under the assumption that some models in the ensemble perform poorly and damage the entire ensemble's performance, we implemented the following model pruning approach. We first binarized the set of scores, $\{B^+_{i,1}, \ldots, B^+_{i,M}\}$, generated from the trained ensemble, using the default 0.5 score threshold for the PBMF. Using this S patients by M models binary matrix, R, we then compute an $S \times S$ patient agreement matrix, A, by calculating the proportion of models that assigned two different patients to the same class[77]:

$$A_{ij} = \frac{1}{M}\sum_{k=1}^{M} I(R_{ik} = R_{jk})$$

A contains 1 along its diagonal, is symmetric, and contains values $\in$ [0,1]. Patients with similar scores across each model in the ensemble will tend to have higher values; those with dissimilar scores will have lower values. Each column or row of A represents how consistently patients were assigned to a particular class by the models in the ensemble, from the reference point of one patient.

We then computed the Pearson correlation between each column in A with each column in R to generate an $S \times M$ matrix, C, of correlation coefficients that represents how well the patient scores from an individual model in the ensemble correlate with the patient agreement matrix. We assumed that only a minority of models have poor performance, such that we should keep models that agree on how patients should be scored and discard models that disagree. This was done by selecting a percentile, p, e.g., the 90th percentile of all the correlations. By thresholding on the value in C associated with this percentile, the models were sorted by the number of times that each model exceeded the threshold, to generate a $1 \times M$ vector of counts. We then thresholded on the value associated with our percentile in this vector to return the final subset of models, M_S, that exceed this threshold. A new consensus score was then computed as the average score across the reduced set of models in the ensemble.

Model distillation: pseudo-labeling

The distribution of scores generated from the ensemble is used to identify patients with "high-quality" predictions, i.e., those whose distributions are heavily skewed toward 0 (strongly B–) or 1 (strongly B+).

To identify the patients with the best high-quality scores, we choose a 0.5 cut point and add an offset value ε, such that the biomarker label for a patient i i is defined as:

$$L_i = \begin{cases} B^+ \text{ if } Cs > 0.5 + \varepsilon \\ B^- \text{ if } cs < 0.5 - \varepsilon \\ \text{No biomarker other case} \end{cases}$$

We set $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.4\}$ and then fitted a Cox PH model to compute the hazard ratios between the treatment and the control arms for both the B+ and B–. The optimal ε score is extracted by determining the maximum difference between the absolute log of the B+ and B– hazard ratios.

$$\text{optimal } \varepsilon = \text{Max}_{\varepsilon_i \in \boldsymbol{\varepsilon}}\{|log(HR^+_{\varepsilon_i}) - log(HR^-_{\varepsilon_i})|\}$$

We then applied the optimal ε to compute a reduced set of patients with high-quality scores.

Model distillation: tree-based model explainability

Once the high-quality population is defined, a tree classifier (Python sklearn[78] tree classifier package, random_seed = 0) is fit, using the input features and the B+ and B– as the labels. The goal of the tree classifier is to define a simple rule that approximates the neural network-derived predictive biomarker. In practice, a tree classifier is fit for each of max_depth = 1, 2, …, 19 and a final tree classifier model for explainability is chosen as that with the smallest max_depth which maximizes the training set AUROC. Although the tree

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

model is trained on only a subset of patients with "high quality" scores, it is used to predict the biomarker status on all training samples or test data set samples.

VT implementation

We implemented the VT approach proposed by Foster et al.[22] as follows. We used a random survival forest model[79] to predict time-to-event based on the log-rank test loss (pySurvival[80]). We built two survival models $\{M_T, M_C\}$, where T and C refer to the population under treatment and under the control, respectively. Each model was trained using only its respective population. We then computed the difference in risk score between the treatment and control models to define the counterfactual risk score $r_i = M_T(i) - M_C(i)$ for any given patient i.

To stratify patients into B+ and B–, we computed the median value of the counterfactual risk score distribution across all patients and assigned to B+ those patients below the median score (low risk) and to B– those with a counterfactual risk score above the median. Consequently, this design choice intrinsically classified patients evenly, 50% being assigned to B+ and the remaining 50% to B–. This can potentially lead to an overestimation of favorable results in data sets where the predictive biomarker prevalence is 50%.

For simulations, model hyperparameters were tuned as described in Supplemental Information and Table S7. Model hyperparameters for identifying predictive biomarkers for clinical studies is described in Table S2.

SIDES implementation

The SIDES algorithm was set for survival analysis using the time and event features as the targets and the treatment versus control setting. The features used were the same as those used for PBMF and VT and depended on the analyzed data set. We used the R implementation of SIDES provided by the SIDES authors (sides.dylib, CSIDES.r, and stochSIDES_util.R). We selected the best biomarker sorted by the adjusted P value and assigned it as B+. The discovered predictive biomarker rule was then validated in a given independent test set. Model hyperparameters for identifying predictive biomarkers for clinical studies is described in Table S2.

Synthetic data generation

We generated 10,000 patients for each data set. For a given replicate, 2000 patients (20%) were randomly selected, without replacement. Among those selected, a 50-50 training/test split was performed. Evaluation metrics are reported only from the test set. Proportional hazard assumptions were imposed to induce each one of the behaviors (Figure S1A). The ability of each methodology to correctly call the biomarker was measured by recording the precision, recall, and AUPRC of a holdout test data set (2000 patients for each data set).

The generation of synthetic data sets involves three stages. Initially, a set of covariates with predetermined level of correlation and prevalence is defined (Figure S1A). These covariates establish subgroups for which desired hazard ratios will be generated. For the parametric model, the cumulative hazard is

$$H_i(t) = \lambda(t^\gamma)\exp\left(X_i^T\beta\right)$$

Where X_i is a vector of covariates associated to the parameters β. The β parameters used to sample survival times can be estimated after setting the HR requirements between groups. For example, assuming a treatment variable and a predictive biomarker, we can define the following hazard ratios:

$$HR^{Control,B+vs\ B-} = HR_1$$

$$HR^{Treatment,B+vs\ B-} = HR_2$$

$$HR^{B+,Treatment\ vs\ Control} = HR_3$$

$$HR^{B-,Treatment\ vs\ Control} = HR_4$$

The time-independent part of $H_i(t)$ can be expanded as:

$$H_i \sim \exp\left(\beta_{trt}trt_i + \beta_{x1}x1_i + \beta_{trt-x1}trt_ix1_i\right)$$

Replacing for each one of the cases in equation 1, we obtain the following equations:

$$\log(HR_1) = \beta_{x1}$$

$$\log(HR_2) = \beta_{x1} + \beta_{trt-x1}$$

Please cite this article in press as: Arango-Argoty et al., AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.03.029

$$\log(HR_3) = \beta_{trt} + \beta_{trt-x1}$$

$$\log(HR_4) = \beta_{trt}$$

Random survival times are then obtained using the technique outlined in Crowther and Lambert,[81]

$$t_i = \left(\frac{-\log(u)}{\lambda \, exp\left(X_i^T \beta\right)} \right)^{\frac{1}{\gamma}}$$

where λ and γ and are the scale and shape parameters, and u is a random variable sampled from the uniform distribution $U(0, 1)$. Note that additional censoring, not covered in this work, can also be introduced.

Creating uncorrelated covariate matrix for simulations

To properly control the relationship between covariates, we simulated random multivariate normal with specific covariance matrix. Although induced covariance may be interesting in some cases, we forced all the features to be fully uncorrelated by using an identity matrix as a covariance matrix. To ease the definitions of the hazard ratio group in the next stage, a binarization process was performed in each feature. To create a particular prevalence, the binarization was done at the specific percentile desired in each feature. This step was followed by multiplication of the features and treatment to the level needed (i.e., first-order interaction, two feature interactions or higher). The next step is the definition of the underlying hazard ratios structure, meaning the definition of how many different groups will be in the data and the relative hazard between them. Once the values of the parametric model are defined, the covariate extended matrix is used to create survival times by sampling from the inverted hazard function. Once the times are obtained, it is possible to increase the complexity of the data set by inducing extra noise in several ways. For example, while fully random features can be added at the beginning in the covariate matrix and then have their parameters in the hazard function set to 1, extra noise can be added to the predictive and prognostic features to transform the features from binary back to continuous.

Optimizing virtual twins for simulations

To perform a fair comparison between the performance of virtual twins and the PBMF, a grid search was conducted over various hyperparameters for the random forest used in the virtual twins model. Different values for the number of features chosen in each split, the maximum depth of each tree, and the number of trees in the forest were evaluated (Table S7). Similar data to the one used during the experiments was created. In this case a dataset containing the predictive biomarker (2 features) in addition to 3 random features was used to train and evaluate these models. For each combination of hyperparameters, 10 different splits of the training and test set were generated from 80% of the data. After evaluating the results for each model trained on a different split of the data, we recorded the mean and standard deviation of the AUPRC for both the training and test sets. The best-performing model from the search was trained with 4 features in each split, 200 trees in the ensemble, and a maximum depth of 5, which resulted in a training AUPRC of 0.946 ± 0.014 and a testing AUPRC of 0.933 ± 0.015.

QUANTIFICATION AND STATISTICAL ANALYSIS

Hazard ratios and 95% confidence intervals were computed by fitting a univariate Cox proportional hazards model (lifelines Python package[82]) to the survival data, within a given PBMF biomarker group, and using the treatment as the only covariate. P-values for hazard ratios were computed with a Wald test. When comparing survival distributions across treatments for a given biomarker group, a logrank test statistic and its associated p-value was computed and reported. Because our analyses are all retrospective, we avoid specifying statistical significance thresholds and instead faithfully report all p-values.

Model performance on synthetic datasets was evaluated using the AUPRC metric. This was chosen because we assume that identification of biomarker positive individuals is most important for biomarker discovery, and that a minority of individuals will be biomarker positive for any given real data cohort. Therefore, metrics that equally weight model performance in identifying biomarker positives and negatives, such as area under the receiving operator characteristic curve, may be poor choices. AUPRC was not reported for clinical datasets due to lack of ground truth.