Generalizable AI predicts immunotherapy outcomes across cancers and treatments

Wanxiang Shen[1], Thinh H. Nguyen[2], Michelle M. Li[1], Yepeng Huang[1], Intae Moon[1],
Nitya Nair[3], Daniel Marbach[4,‡], and Marinka Zitnik[1,5,6,7,‡]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Division of Immunology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[3]Roche Pharma Research and Early Development, Oncology Early Clinical Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland
[4]Roche Pharma Research and Early Development, Data & Analytics, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland
[5]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, MA, USA
[6]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[7]Harvard Data Science Initiative, Cambridge, MA, USA

‡Corresponding authors: daniel.marbach.dm1@roche.com, marinka@hms.harvard.edu

Immune checkpoint inhibitors have become standard care across many cancers, but most patients do not respond. Predicting response remains challenging due to complex tumor-immune interactions and the poor generalizability of current biomarkers and models. Predictors such as tumor mutational burden, PD-L1 expression, and transcriptomic signatures often fail across cancer types, therapies, and clinical settings. There is a clear need for a robust, interpretable model that captures shared immune response principles and adapts to diverse clinical contexts. We present COMPASS, a foundation model for predicting immunotherapy response from pan-cancer transcriptomic data using a concept bottleneck architecture. COMPASS encodes tumor gene expression through 44 biologically grounded immune concepts representing immune cell states, tumor-microenvironment interactions, and signaling pathways. Trained on 10,184 tumors across 33 cancer types, COMPASS outperforms 22 baseline methods in 16 independent clinical cohorts spanning seven cancers and six immune checkpoint inhibitors, increasing precision by 8.5%, Matthews correlation coefficient by 12.3%, and area under the precision-recall curve by 15.7%, with minimal or no additional training. The model generalizes to unseen cancer types and treatments, supporting indication selection and patient stratification in early-phase clinical trials. Survival analysis shows that COMPASS-stratified responders have significantly longer overall survival (hazard ratio = 4.7, $p < 0.0001$). Personalized response maps link gene expression to immune concepts, revealing distinct mechanisms of response and resistance. For example, among immune-inflamed non-responders, COMPASS identifies distinct resistance programs involving TGF-β signaling, endothelial exclusion, CD4+ T cell dysfunction, and B cell deficiency. By combining mechanistic interpretability with transfer learning, COMPASS provides mechanistic insights into treatment response variability, supports clinical decision-making, and informs trial design.

Main

Immune checkpoint inhibitors (ICIs) have revolutionized cancer treatment, yet clinical benefits remain heterogeneous across tumor types, with only a minority of patients achieving durable responses[1]. While tumor mutational burden (TMB) and PD-L1 expression represent clinically validated predictive biomarkers, their limited accuracy restricts reliable patient selection[2]. Individual response patterns vary dramatically, from complete and sustained remission to transient responses or primary resistance, reflecting fundamental differences in tumor-immune interactions between patients. In solid tumors, ICI responders typically exhibit an immune-inflamed phenotype, marked by CD8+ T cell infiltration. Non-responders generally fall into two distinct categories: immune-desert tumors, which lack immune infiltration, and immune-excluded tumors, where immune cells accumulate in the stroma but fail to penetrate tumor tissue[3,4]. A substantial subset of non-responders nevertheless maintains immune-inflamed characteristics, highlighting the biological complexity of resistance mechanisms. Advancing predictive capacity for ICI and delineating the mechanisms of response and resistance are critical for optimizing personalized treatments and improving patient outcomes[5].

Biomarkers, including TMB, PD-L1 IHC score, CD8+ T cell infiltration, and immune gene expression signatures, provide incomplete insights into ICI response mechanisms[6]. High TMB correlates with clinical benefit to ICI in some cancers, likely as a surrogate for increased neoantigen presentation[7–9], but fails to fully predict response: many high-TMB tumors remain refractory, while some low-TMB tumors respond robustly. Transcriptomic signatures, such as TIDE (T cell dysfunction)[10], IMPRES (immune checkpoint activity)[11], and others measuring IFN-γ, cytotoxicity, MHC-I presentation, or tertiary lymphoid structures, offer mechanistic insights[12–14]. However, these signatures exhibit variable predictive performance across cancer types. While TIDE performs well in melanoma, it shows limited efficacy in lung and bladder cancer, and IMPRES demonstrates poor generalizibility beyond melanoma[10,11]. A pan-cancer analysis of 27,810 ICI-treated patients further underscores these limitations, revealing weak or inconsistent associations between PD-L1, CD8+ T cells, immune gene scores, and TMB with response across tumors[15]. Machine learning models seek to enhance prediction accuracy but are often limited by small training datasets and lack sufficient biological interpretability[16,17]. Consequently, they may fall short in supporting clinical development objectives, including patient enrichment for novel immunotherapies or identifying mechanisms of resistance.

Here, we introduce COMPASS, a foundation model that predicts immunotherapy response from tumor transcriptomes and identifies resistance mechanisms using interpretable tumor–immune concepts. COMPASS combines self-supervised learning with a concept bottleneck architecture to map bulk RNA-seq profiles onto biologically grounded features, including immune cell states, tumor-microenvironment interactions, and signaling pathways (**Figure 1a**). Instead of relying on predefined markers, COMPASS learns hierarchical immune representations directly from data, supporting generalization across cancer types, treatment regimens, and patient populations. The model is pre-trained on transcriptomes from 33 cancer types, encompassing both primary and metastatic tumors, and fine-tuned on clinical cohorts to predict response to immune checkpoint inhibitors, including anti-PD1/PD-L1, anti-CTLA4, and combination therapies (**Figure 1b, c**). This two-stage training strategy captures shared immune response patterns while adapting to the specific characteristics of each cohort.

We evaluate COMPASS on 1,133 patients from 16 clinical cohorts spanning seven cancer types, using pre-treatment tumor RNA-seq profiles to predict response to ICIs. To accommodate varying cohort sizes and data availability, COMPASS supports four fine-tuning strategies: full fine-tuning, partial fine-tuning, linear probing, and zero-shot prediction. In leave-one-cohort-out evaluation, COMPASS outperforms 22 existing models, achieving 8.5% higher accuracy and 15.7% greater area under the precision-recall curve (AUPRC). Partial fine-tuning and linear probing emerge as the most robust strategies across cohorts of varying sizes. COMPASS generalizes in leave-one-patient-out evaluation, achieving 76% accuracy in predicting treatment response to ICIs for kidney renal cell carcinoma patients and 72% accuracy for lung adenocarcinoma patients, and demonstrates successful cross-cohort knowledge transfer in 163 out of 240 evaluations. It accurately predicts treatment responses in untested cancer types (83.7% accuracy for stomach adenocarcinoma), and across drug classes, including 76.1% accuracy for anti-CTLA4 following training on anti-PD1/PD-L1 cohorts. Multi-stage fine-tuning enables precise drug-specific predictions, achieving 73.7% accuracy for atezolizumab response in kidney cancer. In a held-out phase II trial of metastatic urothelial cancer, patients predicted as responders by COMPASS show significantly longer survival ($HR = 4.7$, $p = 1.7 \times 10^{-7}$), outperforming TMB and PD-L1 immunohistochemistry biomarkers.

COMPASS generates personalized response maps that link gene expression to immune concepts, allowing mechanistic interpretation for individual patients. For each case, COMPASS identifies critical drivers of response or resistance, such as T cell exhaustion, macrophage activity, and immunoregulatory signaling. For patients displaying response patterns inconsistent with conventional immune phenotype classifications (inflamed, excluded, desert), COMPASS identifies distinct functional immune states: for example, inflamed non-responders exhibit TGF-β-driven suppression, vascular exclusion, or CD4+/B-cell dysfunction, while non-inflamed responders display residual cytotoxic activity or TMB-associated pathways without immunosuppressive signals. These findings demonstrate that COMPASS can identify functional immune profiles beyond the classification of binary responses. By integrating transcriptomic data with tumor immune concept modeling, COMPASS supports cross-cancer and cross-ICI therapy applications, improves existing models, and offers biological insight into treatment outcomes.

Results

COMPASS model for treatment response prediction across cancers and immune checkpoint inhibitors.

Predicting the response to immunotherapy remains difficult due to the small cohort size, patient heterogeneity, and the complexity of tumor-immune interactions. To address this, we developed COMPASS, a concept bottleneck-based model that combines pan-cancer pre-training on bulk tumor transcriptomes with a biologically structured architecture grounded in immune and stromal concepts (**Figure 1a**). This design leverages large-scale self-supervised learning to improve generalizability and supports parameter-efficient, interpretable adaptation to clinical settings.

To ground the model in biologically meaningful features, we curated 132 gene signatures from the literature (**Supplementary Data S1**). These signatures capture a wide range of immune cell types (e.g., T cells, B cells), functional states (e.g., exhausted CD8 T cells, cytotoxic activity), and pathways (e.g., interferon-γ response, antigen presentation). We also included non-immune signatures relevant to tumor biology, such as stromal cell infiltration and DNA damage response. The 132 signatures were grouped into 43 high-level tumor immune microenvironment (TIME) concepts, reflecting established biology and integrating insights from both bulk and single-cell studies. These concepts form the basis of COMPASS's interpretable representation of tumor-immune dynamics. COMPASS uses a transformer-based gene language model to encode expression profiles of 15,672 protein-coding genes (**Figure 1d**, **Methods 2**). It projects gene embeddings onto 132 gene signatures, which are then aggregated into 43 high-level TIME concepts using a hierarchical projector. To capture cancer-specific immune context, the model incorporates a cancer-type token alongside gene tokens, treated as an additional concept. This process generates 44-dimensional patient embeddings that reflect patient-specific transcriptional programs.

We pre-trained COMPASS on transcriptomes from 10,184 tumors in The Cancer Genome Atlas (TCGA) using self-supervised contrastive learning (**Methods 3**). The model is trained on triplets consisting of an anchor tumor, a perturbed (augmented) version of the same tumor, and a tumor from a different patient. It learns to position the anchor and its perturbed version closer together than unrelated samples in the 44-dimensional concept space (**Figure 1d**). This approach allows the model to learn patient representations of TIME. After pre-training, COMPASS is fine-tuned on clinical cohorts to predict immunotherapy response (**Methods 4**). Patient embeddings are processed by a classifier adaptor to generate response probabilities based on 44-dimensional concept representations of patients. This adaptor maps biologically grounded concepts to treatment outcomes, enabling prediction while preserving interpretability.. The extent of fine-tuning depends on the cohort dataset size (**Figure 1d, e**). For large cohorts, the full model is fine-tuned. For medium-sized cohorts, only the concept projector and the classifier adaptor are updated. For small cohorts, training is limited to the classifier adaptor. When treatment response outcomes are unavailable for fine-tuning the classifier adaptor, the pre-trained model can still generate predictions in a zero-shot manner using similarity-based inference in the latent concept space.

COMPASS achieves state-of-the-art performance across 16 immunotherapy cohorts.

We consider pre-treatment bulk RNA sequencing (RNA-seq) and clinical outcomes for 1,133 patients from 16 clinical cohorts (**Figure 2a**, **Supplementary Table S1**). These cohorts span seven cancer types: bladder urothelial carcinoma (BLCA), glioblastoma (GBM), renal cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous carcinoma (LUSC), melanoma (SKCM), and stomach adenocarcinoma (STAD). They cover five immune checkpoint inhibitor treatment regimens (anti-PD1 therapies (nivolumab and pembrolizumab), anti-PD-L1 therapy (atezolizumab), anti-CTLA4 therapy (ipilimumab), and combination treatments) and represent clinical studies across early- to late-phase clinical trials (phases I-III). Cohorts are stratified by size: large cohorts ($n > 100$), including IMvigor210[18], IMmotion150[19], Ravi-1 (SU2C-MARK LUAD cohort)[20], and Liu[21]; medium-size cohorts ($n = 30$ to 100), including Freeman (MGH cohort)[22], Van Allen[23], Kim[24], Riaz[8], Gide[25], and Rose[26]; and small cohorts ($n < 30$), including Choueiri[27], Miao[28], Snyder[29], Zhao[30], Ravi-2 (SU2C-MARK LUSC cohort)[20], and Hugo[31]. Patients are classified as responders ($n = 346$, 30.5%; complete or partial remission) or non-responders ($n = 787$, 69.5%; progressive or stable disease) (**Methods 1.2**). This dataset captures heterogeneity arising from variation in treatments, cancer types, sequencing platforms, study designs, patient demographics, biopsy sites, and tumor-immune interactions.

To benchmark predictive performance, we compared COMPASS against 22 widely used ICI response prediction methods (**Supplementary Table S2**). These include algorithms that rely on single or composite biomarkers (PD1, PD-L1, GeneBio[16]), immune-related signature scoring methods (Teff[32], IS[33], IFN-γ[12], MIAS[13], GEP[13, 14]), and network-based or machine learning models (IMPRES[11], NetBio[16], TIDE[10]).

We used a leave-one-cohort-out evaluation strategy to assess generalizability in clinically relevant scenarios (**Figure 2b**). This approach mimics clinical situations in which a predictive model encounters entirely new patient cohorts. The models were trained with all but one cohort and then tested in the holdout cohort, thus evaluating the predictive performance on previously unseen patient populations. We evaluated performance on three metrics: precision, area under the precision-recall curve (AUPRC), and Matthews correlation coefficient (MCC) (**Methods 5**).

COMPASS substantially improved upon established methods across all cohort sizes and metrics (**Figure 2c, Supplementary Figure S2**). Partial fine-tuning (COMPASS-PFT) and linear probing (COMPASS-LFT) consistently delivered the highest overall predictive performance. Compared to the second-best-performing models, COMPASS-PFT and COMPASS-LFT improved precision by an average of 8.5%, AUPRC by 15.7%, and MCC by 12.3%, demonstrating their superior capability for response prediction.

For large ($n = 672$) and medium ($n = 331$) sized cohorts, COMPASS-PFT and COMPASS-LFT outperformed all other methods across each metric. In small cohorts ($n = 130$), COMPASS achieved the highest accuracy, although it ranked second in AUPRC and MCC behind TIDE and IS, respectively. However, neither TIDE nor IS showed consistently strong performance across all metrics and cohort sizes. In contrast, models such as NetBio demonstrated notable performance variability, excelling in larger cohorts but declining markedly in smaller datasets. COMPASS-PFT and COMPASS-LFT are the only methods with consistently high performance across cohorts of varying sizes.

Among the fine-tuning strategies evaluated, partial fine-tuning (COMPASS-PFT) and linear probing (COMPASS-LFT) provided the best balance between model adaptability and stability. These strategies consistently achieved high accuracy across different cohort sizes. In contrast, full fine-tuning (COMPASS-FFT), which updates all model parameters during training, performed worse when large cohorts were held out during leave-one-cohort-out evaluation, likely due to overfitting from limited training data, as the full model involves many trainable parameters (**Figure 2c**) (**Figure 2c**). The biologically informed architecture and parameter-efficient fine-tuning of COM-

PASS thus offer a practical approach to capture clinically relevant tumor-immune interactions for robust and generalizable predictions of immunotherapy outcomes.

COMPASS generalizes within and across immunotherapy cohorts.

We next assessed how accurately COMPASS predicts immunotherapy responses when training data are limited to a single cohort. This setting reflects early-phase drug development, where models must often be trained on a single cohort. Unlike the previous multi-cohort evaluations, this analysis tests performance when data are limited to one study. We used two benchmarking strategies. First, we used intra-cohort validation, using a leave-one-patient-out procedure. In this approach, we iteratively trained models on all patients within a cohort except one, then evaluated predictive accuracy on the withheld patient. This procedure evaluates the model's ability to generalize within relatively homogeneous patient populations. Second, we performed cohort-to-cohort transfer analyses, where models trained on one cohort predicted treatment responses in a completely different cohort. This approach tests model adaptability across patient populations.

In intra-cohort validation, COMPASS consistently outperformed previously developed methods across all cohort sizes (**Supplementary Figures S3, S4**). Partial fine-tuning (COMPASS-PFT), linear probing (COMPASS-LFT), and full fine-tuning (COMPASS-FFT) achieved the highest accuracy on medium and large cohorts. For small cohorts (fewer than 30 patients), the no fine-tuning variant (COMPASS-NFT) performed best. This observation likely results from the limited data available in small cohorts, making extensive fine-tuning less effective. Instead, COMPASS-NFT leverages pre-trained tumor-immune concepts, allowing effective patient stratification through similarity-based inference even without additional training. Among previously published approaches, NetBio performed best on medium and large cohorts. For small cohorts, methods with simpler architectures, such as PGM, Texh, and TIDE, demonstrated better accuracy, likely due to their reduced complexity and lower tendency to overfit limited datasets.

The cohort-to-cohort transfer analyses, models trained on one cohort predicted responses for patients in a separate cohort, resulting in 240 total pairwise transfer evaluations (16 cohorts × 15 possible transfers per cohort; **Figure 2e**, **Supplementary Figures S5, S6**). We defined successful transfer as achieving a prediction accuracy greater than the reference accuracy for the target cohort while accounting for multiple hypothesis testing (**Methods 5**). Linear probing (COMPASS-LFT) performed best, with successful transfer in 163 of 240 cases, followed closely by partial fine-tuning (COMPASS-PFT) with 155 successful transfers. These results substantially exceeded the top-performing previously published methods, including PGM (130/240), Teff (118/240), and

NetBio (117/240). The strong performance of COMPASS-LFT, which updates only the final classification layer, suggests it effectively prevents overfitting when trained on small datasets.

COMPASS predicts treatment response across cancer types, treatments, and drug targets.

To be clinically useful, predictive models must generalize beyond individual studies, meaning that they can generate accurate predictions for all types of cancer, treatment regimens, and immune checkpoint targets despite biological and technical heterogeneity[34,35]. We evaluated COMPASS's generalizability along these clinically important axes by stratifying patients by cancer type, therapy, and immune checkpoint targets (**Figure 3**).

We compared partial fine-tuning (COMPASS-PFT) against top-performing baseline methods (PGM, NetBio, Teff) in each setting. COMPASS-PFT consistently achieved higher accuracy and precision-recall metrics, particularly in training scenarios with more patients, where fine-tuning of pre-trained tumor-immune concepts was more effective. In cross-indication prediction (**Figure 3a**), a task critical for indication selection in clinical drug development, where response rates may be unknown in untested indications, COMPASS-PFT identified responders in cancer types excluded from training. For example, when trained on 1,031 patients across all cohorts except lung adenocarcinoma, COMPASS-PFT predicted responses of adenocarcinoma patients with 76.5% accuracy, outperforming NetBio (58.8%) and PGM (51.0%) (**Supplementary Table S4**). These results demonstrate that COMPASS can distinguish generalizable features of antitumor immunity from indication-specific transcriptional variation.

Cross-therapy evaluation showed that COMPASS-PFT generalized across treatment regimens targeting related immune pathways (**Figure 3b**). When trained without pembrolizumab-treated patients, COMPASS-PFT predicted pembrolizumab responses with 71% accuracy, exceeding Teff (53.7%), NetBio (59.5%), and PGM (55.2%) (**Supplementary Table S5**). The model demonstrated comparable performance in cross-target evaluation (**Figure 3c**), for example, achieving 70.8% accuracy in predicting anti-CTLA4 responses when trained exclusively on PD1/PD-L1 cohorts. These results show that COMPASS captures immune mechanisms relevant across molecular targets, not just treatment-specific signatures.

Another clinically relevant problem is predicting responses to combination therapies (ipilimumab plus pembrolizumab) using models trained on non-combination cohorts. Here, COMPASS-PFT achieved 85.3% accuracy when trained only on monotherapy cohorts, outperforming NetBio (76%) and PGM (72%) (**Figure 3b**, **Supplementary Figure S7**). In this task, the training set contained a lower proportion of responders than the test set (30% vs. 64%). Despite this inverse class balance, COMPASS-PFT maintained high predictive accuracy, indicating that its performance is driven by underlying biology rather than dataset-specific response rates.

We also tested generalizability across technical factors, including sequencing platforms and biopsy sites (**Supplementary Figure S8**). While other models showed notable performance drops across these cross-factor settings, COMPASS-PFT maintained robust accuracy. In contrast, signature-based models such as Teff were sensitive to platform and site differences.

Multi-stage fine-tuning improves prediction for new therapies and cancer types.

In clinical development, indication selection and patient enrichment—identifying populations most likely to benefit from a therapy—are critical for trial success. However, in early-stage trials for new or combination immunotherapies, predictive modeling is challenging because only limited data are available for training. To address this, COMPASS uses a multi-stage fine-tuning (MSFT) strategy to build robust, treatment-specific models from small clinical cohorts (**Figure 4a**). MSFT proceeds in three stages. First, the model is pre-trained on large-scale transcriptomic data to learn generalizable tumor-immune microenvironment representations. Second, the model is fine-tuned on pan-cancer immunotherapy cohorts to capture immune response patterns broadly associated with immunotherapy efficacy (independent of specific drugs or targets). Third, the model is refined using data from patients treated with the specific drug or combination of interest. This sequential design enables the model to capture broad immune response features while adapting to the molecular and clinical context of the target therapy. The goal is to support key clinical decisions, including trial design, inclusion criteria, and indication prioritization, even when only limited data are available for the therapy under development.

We evaluated MSFT by developing drug-specific models for three immune checkpoint inhibitors: atezolizumab (anti-PD-L1), pembrolizumab, and nivolumab (both anti-PD1) (**Methods 6**). For each drug, we used three training tiers: (1) pre-training on pan-cancer transcriptomes ($n =$ 10,184), (2) fine-tuning on clinical cohorts excluding the target drug or drug class ($n =$ 596–602), and (3) drug-specific refinement on a single cohort treated with the target drug ($n =$ 105–354) (**Supplementary Table S6**). We then tested the models on held-out cohorts treated with the same drug ($n =$ 63–167).

MSFT outperformed single-stage fine-tuning strategies and existing drug-specific models (**Figure 4b**). In predicting atezolizumab response in kidney cancer (KIRC, n = 89), MSFT achieved 73.7% accuracy, compared to 70.3% for single-stage fine-tuning on the drug-specific cohort alone (SSFT1) and 60.7% when fine-tuning models using only pan-cancer data (SSFT2).

SSFT2 also produced a high false-positive rate (41.8%) compared to MSFT (8.5%), illustrating the limitations of using general models without therapy-specific refinement.

We also applied MSFT to disease-specific prediction tasks (**Methods 6**). For lung adenocarcinoma (LUAD), we predicted pembrolizumab responses ($n = 33$) using only non-pembrolizumab-treated LUAD patients ($n = 69$) for fine-tuning (**Supplementary Figure S9**). MSFT (first fine-tuning a COMPASS model on pan-cancer cohorts and then refining it on LUAD) reached 91% accuracy and an MCC of 0.79. In contrast, SSFT1 trained only on the LUAD cohort achieved 67% accuracy and an MCC of 0.36. These results demonstrate that MSFT effectively leverages large datasets to improve prediction in small treatment- or disease-specific settings, addressing the pervasive "small n" challenge in early-phase oncology trials with limited target-specific data..

COMPASS identifies immune cell states and pathways linked to ICI response.

To understand the biological basis of COMPASS's predictions, we analyzed the 44 high-level TIME concepts that structure the model (**Figure 5a**). These concepts are derived from 132 gene signatures representing lymphoid, myeloid, and mesenchymal cell types, as well as key immune and stromal pathways. The signatures were designed to have minimal gene overlap (Jaccard index < 0.2; **Supplementary Figure S10**) to ensure each concept captures a distinct dimension of TIME biology. Unlike fixed gene set scoring methods, COMPASS generates dynamic concept embeddings during pre-training, allowing the model to learn context-specific transcriptional programs across diverse cancer types.

Concept representations learned by COMPASS are consistent across clinical datasets, with stable distributions observed across immunotherapy-treated cohorts (**Figure 5b**; **Supplementary Figures S11-S12**). In leave-one-cohort-out evaluations, `Exhausted Tcell`, `Macrophage`, `NKcell`, `Cytotoxic Tcell`, and `IFNg pathway` concepts, consistently distinguished responders from non-responders (**Figure 5c**). To quantify their predictive relevance, we applied SHAP feature importance analysis[36] to fine-tuned COMPASS models (**Methods 7**). `Exhausted Tcell` repeatedly ranked among the most informative concepts across diverse cancer types. In contrast, the `Reference` concept, based on housekeeping genes, and the `Cancer type` concept contributed minimally in single-cohort models. The `Cancer type` concept became more influential only in cross-cohort settings, where it helps model tissue-specific transcriptional context (**Supplementary Figures S13-S14**).

The `Stem cells` concept was most predictive in glioblastoma[37], reflecting the known role of stem-like populations in driving tumor progression. `Mesothelial cells` contributed strongly to predictions in gastric cancer[38], consistent with mesothelial-to-mesenchymal transition in the gastric tumor microenvironment. In melanoma, high relevance of the `Plasma cell` concept aligns with the role of antibody-producing cells in antitumor immunity[39]. The `Genome integrity` concept, capturing DNA damage response activity, was most predictive in non-small cell lung cancer[40], a cancer type characterized by high genomic instability. These associations are consistent with cancer-specific immunobiology (**Supplementary Figure S14**). Unlike fixed gene set scoring methods, such as ssGSEA and geometric mean aggregation (**Supplementary Figures S15-S16**, **Supplementary Note 1**), COMPASS generates dynamic concept embeddings that reflect cohort-specific transcriptional programs and reduce spurious associations between distinct immune states (`Exhausted Tcell` and `Cytotoxic Tcell`).

COMPASS predicts survival and reveals resistance mechanisms beyond immune phenotypes.

We evaluated whether COMPASS can predict long-term clinical outcomes and identify resistance mechanisms by applying it to IMvigor210, a Phase II trial of patients with metastatic urothelial cancer treated with atezolizumab (anti-PD-L1)[18,41]. A COMPASS-PFT model fine-tuned on all other ICI-treated cohorts, excluding IMvigor210, was used to prevent data leakage.

Patients classified by COMPASS as responders had significantly improved overall survival (OS) compared to non-responders (**Figure 6a**). Response probabilities outperformed concept-based risk scores in stratifying outcomes (**Methods 8**, **Supplementary Figure S17**). Among $n = 298$ patients, those with $P_{(R)} \geq 0.5$ ($n = 42$) showed a 1-year OS rate of 86% (95% CI: 71-93%), compared to 40% (95% CI: 34-46%) for those with $P_{(R)} < 0.5$ (n = 256), yielding a hazard ratio of 4.7 (log-rank $p = 1.7 \times 10^{-7}$). COMPASS outperformed TMB ($HR = 1.67$, $p = 0.0038$), PD-L1 IC2+ scoring ($HR = 1.75$, $p = 0.0018$), and IHC-based immune phenotype ($HR = 1.85$, $p = 0.0042$) (**Figure 6b-d**)[42].

To uncover the immunological features that inform COMPASS's predictions, we analyzed patient-specific TIME concept scores across immune phenotypes (**Methods 9**). Immune phenotype annotations of tumors are defined by CD8+ T cell infiltration patterns as inflamed (CD8+ T cells in direct contact with tumor cells), excluded (CD8+ T cells restricted to stroma near the tumor), and desert (low CD8+ T cell infiltration)[42]. Immune desert and excluded phenotypes are non-inflamed tumors that are usually non-responsive to ICI. Conversely, the immune-inflamed phenotype is strongly infiltrated with T cells and is more often responsive to ICI[43]. Inflamed tumors showed high activation of pro-inflammatory concepts, including `Cytotoxic Tcell`, `IFNg pathway`, `Immune checkpoint`, and `Macrophage`, as well as elevated `Genome` integrity and Cell proliferation (**Figure 6e**, **Supplementary Figure 21**). These concept scores positively correlated with the expression of genes underlying the learned concepts. For example, increased gene expression levels of *CD3E*, *CD8A*, *PRF1*, and *GZMB* positively correlated with Cytotoxic Tcell concept, while increased *CXCL9*, *CXCL10* and *IFNG* gene expression positively correlated with the IFNg pathway concept (**Supplementary Figure 19**). Desert tumors, defined by absent T cell infiltration, lacked activation of pro-inflammatory concepts (Cytotoxic Tcell) and exhibited elevated scores for dysfunctional or deficient immune components, including NKcell, Innate lymphoid cell, Bcell general, and Plasma cell (**Figure 6e**). These concept scores were negatively correlated with the expression of underlying genes (e.g., *CD19*, *MS4A1* (encoding *CD20*)), reflecting reduced infiltration or impaired functionality of the corresponding immune cells (**Supplementary Figure 19-20**). Immune-excluded tumors displayed intermediate features, with lower activation of inflammatory programs than inflamed tumors and weaker immunodeficiency signals than desert tumors (**Figure 6e**). These tumors showed elevated TGFb pathway and Endothelial concept scores, consistent with immunosuppressive stromal remodeling and vascular exclusion.

Next, we examine how COMPASS resolves the heterogeneity of ICI responses within and between immune phenotypes. Patients were stratified into four groups based on their response and phenotype: inflamed responders and non-inflamed non-responders (groups showing the expected clinical outcome based on conventional immune phenotype), and non-inflamed responders and inflamed non-responders (groups showing an unexpected clinical outcome based on immune phenotype classifications). We focus on correctly predicted cases and cluster their concept profiles (**Figure 6f**, **Supplementary Figure S22**, **Methods 9**).

Inflamed responders formed two major clusters, both characterized by a strong activation of pro-inflammatory concepts (Cytotoxic Tcell, IFNg pathway) and absence of immuno-suppression or immuno-deficiency signals. These clusters differed in TMB-associated concept activity, particularly Genome integrity and Cell proliferation (**Supplementary Figure S21b**), with one cluster showing strong activation of these concepts and the other lacking it. Non-inflamed responders clustered into two groups, exhibiting intermediate pro-inflammatory activation but lacking immunosuppressive signals, which distinguished them from non-responders. These patients comprised predominantly immune-excluded tumors (7 of 9) and a few immune-desert tumors (2 of 9), phenotypes generally associated with a lack of response to ICI, exemplifying how RNA-seq provides information on functional immune states beyond traditional IHC classification.

A greater degree of heterogeneity in grouping was observed among non-responders. Inflamed non-responders are of particular interest, because according to conventional immune phenotypes, they would be predicted to be responders, yet COMPASS correctly identified these patients as non-responders. These patients displayed variable pro-inflammatory levels but were unified by the presence of distinct resistance mechanisms. A first cluster showed strong activation of the `Endothelial` concept, reflecting angiogenesis and vascular remodeling processes that create a physical barrier to T cell infiltration, impairing immune cell trafficking. A second cluster exhibited high `TGFb pathway` concept activation; TGF-β signaling promotes stromal remodeling and fibrosis via cancer-associated fibroblast activation, excluding T cells from the tumor microenvironment[42,44]. The remaining three clusters were characterized by CD4+ T cell immunosuppression and B cell deficiency. The `CD4 Tcell` concept correlated with downregulation of cytokine receptors (*IL7R*, *IL2RA*, *IL2RB*, *IL21R*) and co-stimulatory molecules (*ICOS*, *CD40LG*), coupled with upregulation of *IL17RE* and *IL17A*, which suggests a CD4 Th17 profile (**Supplementary Figure S19**). CD4 Th17 cells can exhibit dual roles in tumor immunity[45] with TGF-β implicated in imprinting suppressive functional attributes of these cells. Indeed, `TGFb pathway` and `CD4 Th17cell` concepts were co-activated in 4 of 9 non-responders clusters in this dataset. The deficiency of B cells, a cell type associated with the presence of tertiary lymphoid structures and survival benefit to ICI[46–48], further suggests a loss or dampening of adaptive immune responses critical for anti-tumor immunity. Non-inflamed non-responders also exhibited heterogeneity, forming four distinct clusters. The largest cluster comprised patients without pro-inflammatory activation (16 immune-desert and two immune-excluded tumors). Two clusters showed minimal pro-inflammatory signals with strong TGF-β activation (3 immune-desert and eight immune-excluded tumors), while another cluster exhibited strong endothelial concept-driven immune exclusion (four immune-desert and four immune-excluded tumors). These results show that COMPASS stratifies patients by survival outcomes and suggests resistance mechanisms between and within immune phenotypes.

COMPASS generates personalized response maps to explain individual patient predictions.

We developed personalized response maps to interpret how COMPASS links patient-specific gene expression profiles to predicted immunotherapy outcomes. These maps trace the flow of information through COMPASS's concept bottleneck, capturing five hierarchical levels: (1) gene expression input, (2) gene-level representations from the transformer encoder, (3) projection onto granular immune-relevant concepts, (4) aggregation into high-level TIME concepts, and (5) the final response probability derived from concept activations. This structured visualization allows users to trace how individual genes (e.g., *IL21R*) contribute to intermediate features, such as `CD56dim NKcell`, and higher-order concepts, such as `NKcell`, which ultimately inform the model's prediction. We generated personalized response maps for representative patients from each immune-response cluster in **Figure 6f**. Four illustrative examples are shown in **Figure 7**, with additional cases provided in **Supplementary Figures S23-S27**.

The response maps reveal how diverse immune profiles influence predicted outcomes. One patient with an immune-inflamed phenotype exhibits broad activation of pro-inflammatory pathways (**Figure 7a**), including IFN-γ signaling and cytotoxic T cell activity, as well as minimal activation of immunosuppressive programs. In this case, COMPASS assigns a high response probability ($P_{(R)} = 1.0$). Another responder (**Figure 7b**), despite being classified as an immune-desert phenotype lacking classical immune activation, exhibits strong activation of the `Genome Integrity` concept and moderate IFN-γ signaling, suggesting a TMB-associated response mechanism, with a predicted probability of $P_{(R)} = 0.80$. In contrast, a third patient (**Figure 7c**) classified as inflamed does not respond to treatment. Although some pro-inflammatory activity is present, co-activation of immunosuppressive programs, including TGF-β signaling and B cell deficiency, reduces the predicted probability to $P_{(R)} = 0.22$. A fourth patient (**Figure 7d**), with an immune-desert phenotype and low expression of inflammatory genes, shows dominant immunodeficiency features and receives a response probability of $P_{(R)} = 0$. These examples show how COMPASS disentangles immune mechanisms at the individual level. By linking gene expression to concept activation, personalized response maps provide a mechanistic interpretation of model predictions and support hypothesis generation for patient-level and population-level analyses (www.immuno-compass.com/explore/IMvigor210/).

Discussion

Immune checkpoint inhibitors have improved outcomes across multiple cancer types, but response rates remain low and biomarkers such as tumor mutational burden and PD-L1 immunohistochemistry often fail to stratify patients accurately. Here, we present COMPASS, a concept bottleneck model that predicts immunotherapy response from transcriptomic data and provides mechanistic insight into response variability. COMPASS enables three key capabilities: accurate response prediction, identification of resistance mechanisms, and generation of patient-specific hypotheses for clinical interpretation and trial design.

COMPASS supports several applications in clinical development. It stratifies patients by mechanistic features of response, including cytotoxic T cell activity, interferon-γ signaling, and TGF-β pathway activation. For example, patients with elevated TGF-β or endothelial concept scores may benefit from combination therapies involving TGF-β inhibitors[49] or anti-angiogenic agents[50]. Personalized response maps generated by COMPASS trace the contributions of individual genes and immune concepts to predicted outcomes, enabling interpretable patient-level insights. These maps may support inclusion and exclusion criteria for early-phase trials and guide biomarker-driven patient enrichment strategies. Beyond prediction, COMPASS can monitor pharmacodynamic changes in immune concepts over time, aiding dose selection and validation of mechanism in clinical trials. It can also facilitate reverse translation by prioritizing immune cell states or cytokine pathways for targeted intervention[45,51].

COMPASS integrates pan-cancer transcriptomic pre-training with a concept bottleneck architecture that encodes 44 immune concepts derived from curated gene signatures. This design improves generalizability across cancer types and checkpoint therapies while maintaining interpretability. COMPASS performs robustly even in small datasets using parameter-efficient adaptation strategies such as partial fine-tuning and linear probing. These features make the model suitable for clinical settings with limited training data or computational resources. By modeling functional immune states, COMPASS resolves heterogeneity that conventional immune phenotypes cannot. It distinguishes responders within both inflamed and non-inflamed tumors and reveals resistance mechanisms in patients who fail treatment despite high immune infiltration. These mechanisms include TGF-β signaling, endothelial exclusion, CD4+ T cell dysfunction, and B cell deficiency. The findings extend prior results from the IMvigor210 study[42,46], which linked TGF-β activity and tertiary lymphoid structures to immunotherapy outcomes.

Limitations of COMPASS include reliance on bulk RNA-seq, which lacks spatial resolution and may obscure signals from rare immune cell populations. Integrating single-cell[52–55] or spatial[56] transcriptomic data could improve resolution of cell-specific and spatially restricted immune states. COMPASS complements emerging predictive models that leverage real-world data[57], clinico-genomic features[35,58], and multimodal biomarkers[59,60], which typically lack mechanistic interpretability. Future extensions could integrate genomic alterations or longitudinal immune profiling to improve prediction and deepen biological insight.

By linking tumor transcriptomes to interpretable immune representations, COMPASS supports treatment selection, biomarker development, and mechanistic analysis in immunotherapy. Its performance across multiple cancer types and checkpoint therapies establishes a foundation for incorporating immune modeling into clinical development and translational research.

Data availability. The pan-cancer TCGA datasets, including gene expression and mutation data, are available from the Genomic Data Commons data portal (https://portal.gdc.cancer.gov/, version 37). Clinical data for TCGA patients are provided in Liu *et al.*[61]. The datasets for the IMmotion150 cohort (EGA accession: EGAS00001002928)[19], IMvigor210 cohort (EGA accession: EGAS00001002556)[18] and IMvigor210CoreBiologies (v1.0.1, Ref[42]), Choueiri *et al.*[27], Zhao *et al.* (SRA accession: PRJNA482620)[30], Miao *et al.* (dbGAP accession: phs001493.v1.p1)[28], and Kim *et al.* (ENA accession: PRJEB25780)[24] are available from the Cancer Research Institute iAtlas (https://cri-iatlas.org/)[62] and Synapse (https://www.synapse.org/, accession: syn10337516). The Ravi *et al.* cohort 1 (LUAD) and cohort 2 (LUSC) are non-small cell lung cancer (NSCLC) cohorts from the SU2C-MARK study[20], available through dbGaP under accession number phs002822.v1.p1. The Snyder*et al.* cohort[29] is available on Zenodo (https://doi.org/10.5281/zenodo.546110). The Rose *et al.*[26] dataset is available on the Gene Expression Omnibus under accession ID GSE176307. The melanoma cohorts from Liu *et al.*[21] (dbGAP accession: phs000452.v3.p1), Gide *et al.*[25] (ENA accession: PRJEB23709), Riaz *et al.*[8] (BioProject accession: PRJNA356761), Van Allen *et al.*[23] (dbGAP accession: phs000452.v2.p1), Hugo *et al.*[31] (GEO accession: GSE78220), and Freeman *et al.* (the MGH cohort, dbGAP accession: phs002683.v1.p1)[22]. Additional information, including patient metadata, gencode annotations, cancer codes, and COMPASS-concepts, are available on the COMPASS website (https://www.immuno-compass.com/download).

Code availability. The Python implementation of 22 baseline methods for immunotherapy response prediction, COMPASS code, model training, response prediction, and feature extraction used in this study is available on GitHub at https://github.com/mims-harvard/COMPASS/. An interactive online response prediction server based on various COMPASS models is at https://www.immuno-compass.com/predict. The COMPASS-based gene, geneset, and concept feature extraction online tool is accessible at https://www.immuno-compass.com/extract. The data processing pipeline for preparing COMPASS input from raw FASTQ or raw count mRNA TPM data is at https://github.com/mims-harvard/COMPASS-web/tree/main/mRNA_pipeline. COMPASS tool to generate personalized response maps is available at www.immuno-compass.com/explore/IMvigor210/.

Acknowledgements. We gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH, Pfizer Research, Gates Foundation (INV-079038), Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas

Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. M.M.L. is supported by The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute. T.H.N. acknowledges the support of NIH-2T32AI007512. We thank Min Cong at Brigham and Women's Hospital for discussions on immune checkpoint treatments. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Authors contributions. W.X., D.M., and M.Z. designed the study. W.X. and M.Z. conceptualized the COMPASS models and algorithm. D.M., W.X., and T.H.N. provided and processed gene signatures. W.X., T.H.N., and M.M.L. collected and preprocessed the datasets. W.X. implemented the COMPASS code and conducted the benchmarking. W.X. and T.H.N. performed the patient survival analysis. Y.H. and I.M. provided suggestions on COMPASS model. All authors contributed to writing the manuscript.

Competing interests. D.M. and N.N are currently employed by F. Hoffmann-La Roche Ltd.

Figures

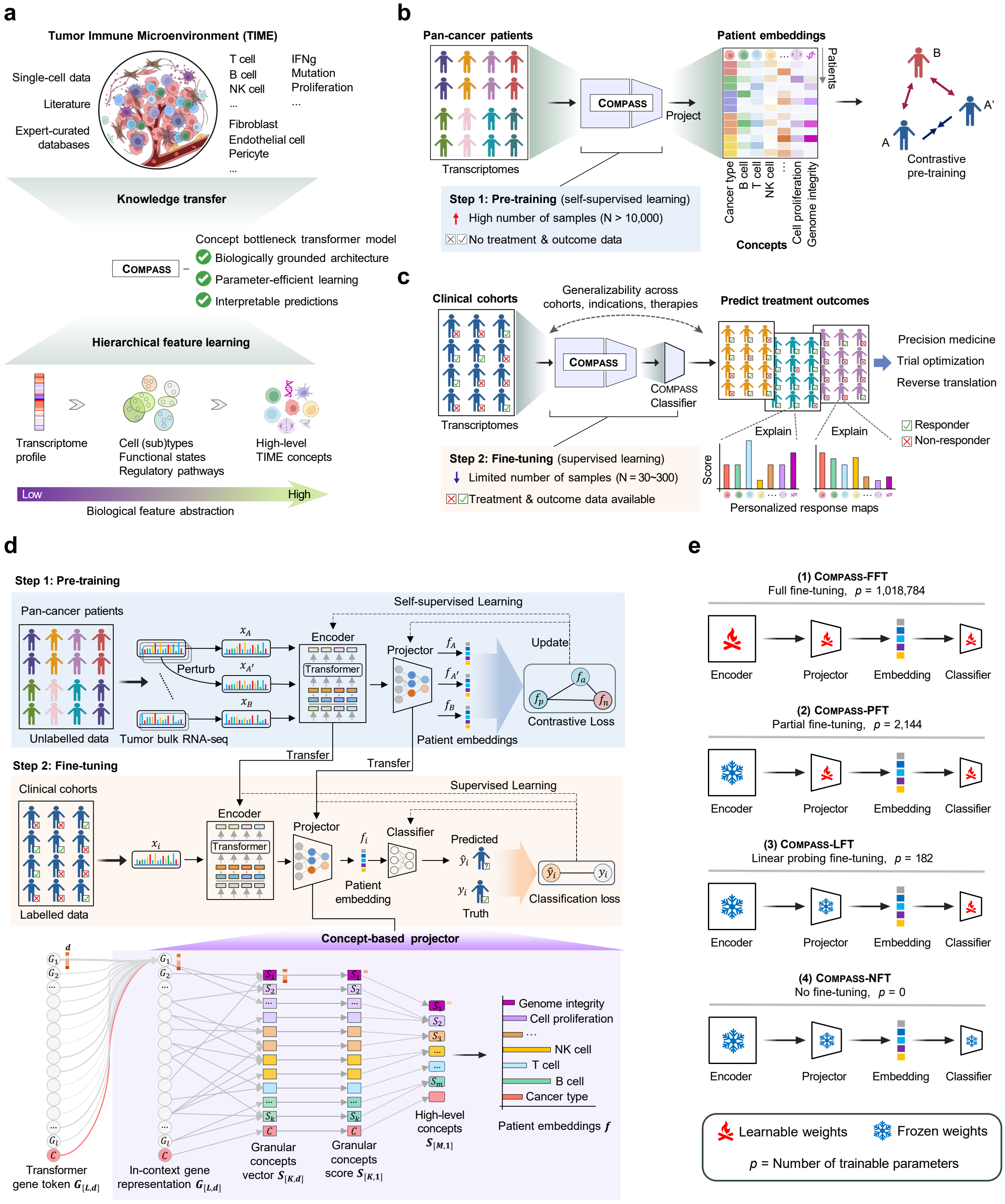
a
Tumor Immune Microenvironment (TIME)
Single-cell data
Literature
Expert-curated databases
T cell
B cell
NK cell
...
IFNg
Mutation
Proliferation
...
Fibroblast
Endothelial cell
Pericyte
...
Knowledge transfer
COMPASS
Concept bottleneck transformer model
Biologically grounded architecture
Parameter-efficient learning
Interpretable predictions
Hierarchical feature learning
Transcriptome profile
Cell (sub)types
Functional states
Regulatory pathways
High-level TIME concepts
Low
High
Biological feature abstraction
b
Pan-cancer patients
Transcriptomes
COMPASS
Project
Patient embeddings
Patients
Cancer type
B cell
T cell
NK cell
...
Cell proliferation
Genome integrity
Concepts
Step 1: Pre-training (self-supervised learning)
High number of samples (N > 10,000)
No treatment & outcome data
B
A
A'
Contrastive pre-training
c
Clinical cohorts
Transcriptomes
Generalizability across cohorts, indications, therapies
COMPASS
COMPASS Classifier
Predict treatment outcomes
Precision medicine
Trial optimization
Reverse translation
Responder
Non-responder
Explain
Explain
Score
Personalized response maps
Step 2: Fine-tuning (supervised learning)
Limited number of samples (N = 30~300)
Treatment & outcome data available
d
Step 1: Pre-training
Self-supervised Learning
Pan-cancer patients
Unlabelled data
Tumor bulk RNA-seq
Perturb
x_A
$x_{A'}$
x_B
Encoder
Transformer
Projector
f_A
$f_{A'}$
f_B
Update
f_a
f_p
f_n
Contrastive Loss
Patient embeddings
Transfer
Transfer
Step 2: Fine-tuning
Supervised Learning
Clinical cohorts
Labelled data
x_i
Encoder
Transformer
Projector
f_i
Patient embedding
Classifier
Predicted
$\hat{y}_i$
y_i
Truth
Classification loss
Concept-based projector
d
G_1
G_2
...
G_l
C
Transformer gene token $G_{[L,d]}$
In-context gene representation $G_{[L,d]}$
S_1
S_2
S_k
C
Granular concepts vector $S_{[K,d]}$
Granular concepts score $S_{[K,1]}$
S_1
S_2
S_3
S_m
High-level concepts $S_{[M,1]}$
Genome integrity
Cell proliferation
...
NK cell
T cell
B cell
Cancer type
Patient embeddings f
e
(1) COMPASS-FFT
Full fine-tuning, p = 1,018,784
Encoder
Projector
Embedding
Classifier
(2) COMPASS-PFT
Partial fine-tuning, p = 2,144
Encoder
Projector
Embedding
Classifier
(3) COMPASS-LFT
Linear probing fine-tuning, p = 182
Encoder
Projector
Embedding
Classifier
(4) COMPASS-NFT
No fine-tuning, p = 0
Encoder
Projector
Embedding
Classifier
Learnable weights
Frozen weights
p = Number of trainable parameters

Figure 1: Concept-bottleneck foundation model for interpretable prediction of immunotherapy response.
(a) Transfer of immuno-oncology knowledge via hierarchical concept learning. The COMPASS concept-bottleneck architecture integrates tumor immune microenvironment (TIME) gene signatures capturing cell types, functional states, and signaling pathways (top). Hierarchical feature learning transforms expression profiles into interpretable TIME representations through successive abstraction levels (bottom). This biologically grounded, parameter-efficient architecture enables generalizable and interpretable predictions of immunotherapy outcomes.
(b) Self-supervised pre-training on pan-cancer transcriptomes. Using contrastive learning, COMPASS projects bulk RNA-seq profiles into biologically grounded patient embeddings that capture essential cell-type and functional pathway information. Triplet training aligns perturbed (augmented) versions of each tumor (A, A') while distinguishing unrelated samples (B) in concept space. By leveraging large-scale, unlabeled pan-cancer cohorts, this process builds generalizable TIME representations that form the foundation for predictive tasks.
(c) Fine-tuning on clinical cohorts for explainable prediction of immunotherapy outcomes. The pre-trained model is fine-tuned on clinical cohorts using supervised learning, where a lightweight classifier is added to predict immune checkpoint inhibitor (ICI) patient responses. Through parameter-efficient transfer learning, COMPASS adapts to diverse treatment regimens and patient populations without overfitting, generating interpretable predictions grounded in the learned TIME concepts.
(d) Architecture of the COMPASS model. The model comprises three components: (1) a transformer-based gene encoder that transforms expression profiles into context-aware representations, (2) a hierarchical concept projector that progressively aggregates these into multi-scale TIME concepts, and (3) a task-specific classifier that outputs predictions from concept-level features. The projector's concept-bottleneck architecture (purple box) first maps gene embeddings to granular TIME concepts, then compresses these into higher-level representations. The classifier leverages these biologically grounded features to distinguish responders from non-responders. During pre-training (blue box), triplet contrastive learning updates the encoder and projector to construct discriminative TIME representations. This is achieved through contrastive loss minimization, which reduces the cosine distance between perturbed (augmented) views of the same tumor (f_A, $f_{A'}$) while increasing their separation from other samples (f_B) in the concept space. Fine-tuning (yellow box) employs supervised learning on clinical cohorts, selectively adapting components (encoder, projector, or classifier) while preserving the interpretable concept hierarchy.
(e) Flexible fine-tuning strategies for clinical adaptation. Four transfer learning modes balance stability with adaptation to new clinical cohorts of varying sizes: (1) **Full fine-tuning (COMPASS-FFT)**: Updates all components (encoder, projector, classifier; $p = 1{,}018{,}784$), refining model weights for task-specific alignment. (2) **Partial fine-tuning (COMPASS-PFT)**: Adjusts only projector and classifier ($p = 2{,}144$), preserving encoder knowledge while recalibrating TIME concepts. (3) **Linear probing (COMPASS-LFT)**: Updates classifier alone ($p = 182$), leveraging frozen pre-trained TIME concepts for small cohorts. (4) **No fine-tuning (COMPASS-NFT)**: Uses cosine similarity to reference patients with known response labels in pre-trained TIME-space ($p = 0$; **Supplementary Figure S1**). Trainable components are denoted by fire icons (learnable weights) versus ice icons (frozen weights).

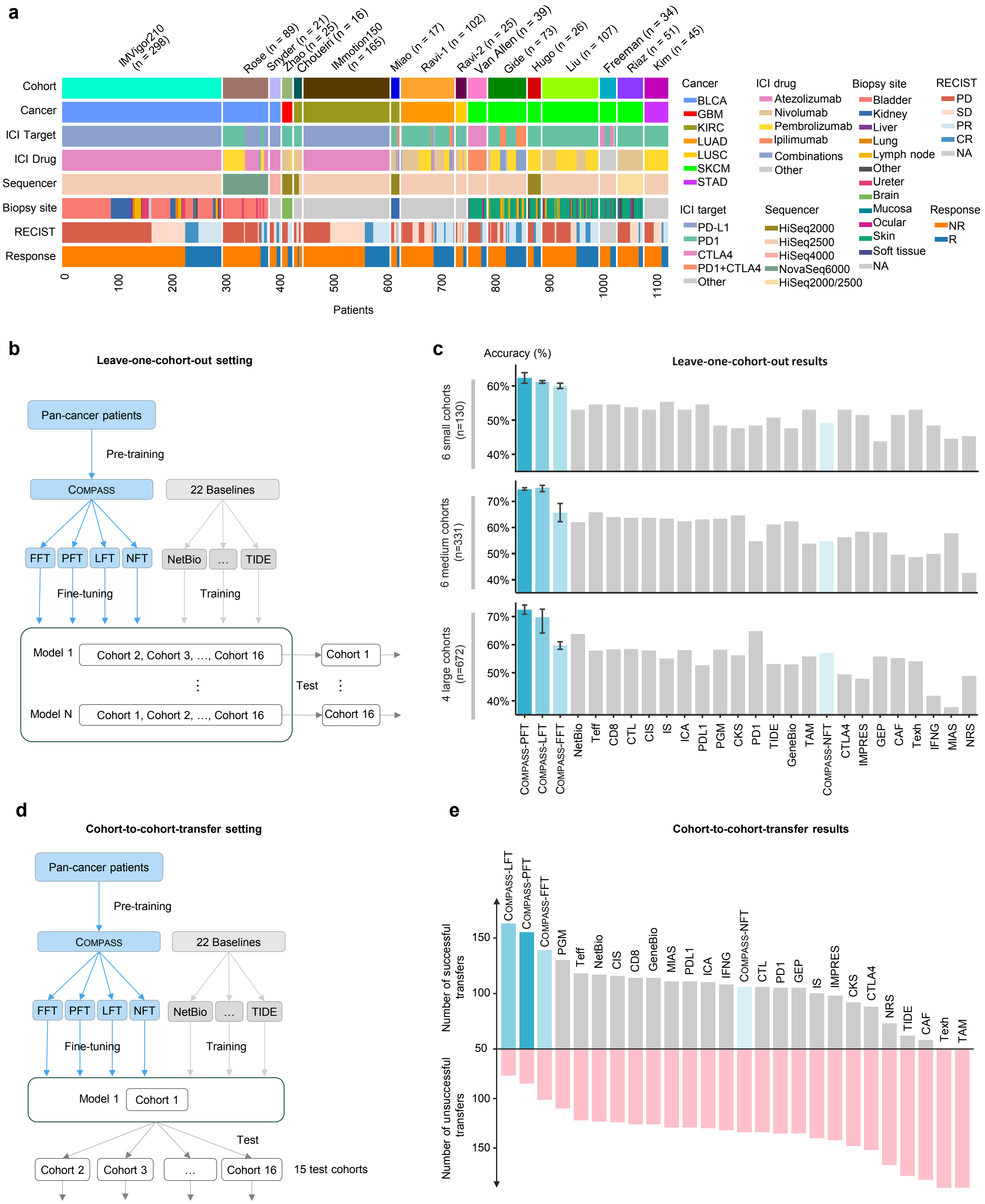
a
IMVigor210 (n = 298)
Rose (n = 89)
Snyder (n = 21)
Zhao (n = 25)
Choueiri (n = 16)
IMmotion150 (n = 165)
Miao (n = 17)
Ravi-1 (n = 102)
Ravi-2 (n = 25)
Van Allen (n = 39)
Gide (n = 73)
Hugo (n = 26)
Liu (n = 107)
Freeman (n = 34)
Riaz (n = 51)
Kim (n = 45)
Cohort
Cancer
ICI Target
ICI Drug
Sequencer
Biopsy site
RECIST
Response
0
100
200
300
400
500
600
700
800
900
1000
1100
Patients
Cancer
BLCA
GBM
KIRC
LUAD
LUSC
SKCM
STAD
ICI drug
Atezolizumab
Nivolumab
Pembrolizumab
Ipilimumab
Combinations
Other
Biopsy site
Bladder
Kidney
Liver
Lung
Lymph node
Other
Ureter
Brain
Mucosa
Ocular
Skin
Soft tissue
NA
RECIST
PD
SD
PR
CR
NA
ICI target
PD-L1
PD1
CTLA4
PD1+CTLA4
Other
Sequencer
HiSeq2000
HiSeq2500
HiSeq4000
NovaSeq6000
HiSeq2000/2500
Response
NR
R
b
Leave-one-cohort-out setting
Pan-cancer patients
Pre-training
COMPASS
22 Baselines
FFT
PFT
LFT
NFT
NetBio
…
TIDE
Fine-tuning
Training
Model 1
Cohort 2, Cohort 3, …, Cohort 16
Cohort 1
Test
Model N
Cohort 1, Cohort 2, …, Cohort 16
Cohort 16
c
Accuracy (%)
Leave-one-cohort-out results
6 small cohorts (n=130)
60%
50%
40%
6 medium cohorts (n=331)
70%
60%
50%
40%
4 large cohorts (n=672)
70%
60%
50%
40%
COMPASS-PFT
COMPASS-LFT
COMPASS-FFT
NetBio
Teff
CD8
CTL
CIS
IS
ICA
PDL1
PGM
CKS
PD1
TIDE
GeneBio
TAM
COMPASS-NFT
CTLA4
IMPRES
GEP
CAF
Texh
IFNG
MIAS
NRS
d
Cohort-to-cohort-transfer setting
Pan-cancer patients
Pre-training
COMPASS
22 Baselines
FFT
PFT
LFT
NFT
NetBio
…
TIDE
Fine-tuning
Training
Model 1
Cohort 1
Test
Cohort 2
Cohort 3
…
Cohort 16
15 test cohorts
e
Cohort-to-cohort-transfer results
Number of successful transfers
150
100
50
Number of unsuccessful transfers
100
150
COMPASS-LFT
COMPASS-PFT
COMPASS-FFT
PGM
Teff
NetBio
CIS
CD8
GeneBio
MIAS
PDL1
ICA
IFNG
COMPASS-NFT
CTL
PD1
GEP
IS
IMPRES
CKS
CTLA4
NRS
TIDE
CAF
Texh
TAM

Figure 2: Evaluation of COMPASS across clinical cohorts for immunotherapy response prediction.
(a) Overview of the diverse clinical cohorts. Each column represents one patient. The dataset is diverse in cohort size, patient indication, drug treatment, and response outcomes. It includes 16 cohorts across seven cancer types, categorized by size into large (four cohorts, n = 672), medium (six cohorts, n = 331), and small (six cohorts, n = 130) groups (**Supplementary Table S1**). All RNA-seq data originate from pre-treatment samples, with variability in biopsy sites, sequencing platforms, and analytical pipelines.. This heterogeneity enables rigorous benchmarking of COMPASS and baseline methods.
(b-c) Leave-one-cohort-out evaluation. Models are trained on 15 cohorts and tested on the held-out cohort. This process is iterated across all cohorts for 22 baseline methods (**Supplementary Table S2**) and COMPASS 's four fine-tuning modes. Performance is aggregated by cohort size groups (large/medium/small) to improve statistical power. For COMPASS 's fine-tuning modes, experiments are repeated three times with different random seeds; error bars show standard deviations. The numerical results are provided in **Supplementary Table S3**.
(d-e) Cohort-to-cohort-transfer evaluation. Models initially trained on one cohort are tested on the remaining 15 cohorts to evaluate generalization across different populations, indications, and treatment regimens. A cohort-to-cohort transfer is considered successful when prediction accuracy is greater than the baseline accuracy for the held-out target cohort (**Method 5**). The success rate across 240 pairwise cohort transfers (16×15) is and indicator of the model's robustness.
PD1/PDL1/CTLA4: drug target markers; NetBio[16]: Network-Based ICI Treatment Biomarkers; PGM[22]: Paired Gene Markers; GeneBio[16]: combination of PD1, PDL1, and CTLA4; CIS[63]: Cytotoxic Immune Signature; Teff[32]: T-effector-IFNg Signature; NRS[64]: Neoadjuvant Response Signature; IFNG[12]: IFNG Signature Score; CTL[10]: Cytotoxic T Lymphocytes Markers; TAM[10,65]: Tumor-Associated Macrophages; Texh[10,66]: T-cell Exhaustion; CKS[67]: Chemokine Signature Score; CAF[68]: Cancer-Associated Fibroblasts Signature Score; IS[33]: Roh Immune Score; ICA[69]: Immune Cytolytic Activity Score; CD8[16,70]: CD8 Signature Score (CD8A + CD8B); MIAS[13]: MHC I Association Immune Score; GEP[13,14]: the T Cell-Inflamed Gene Expression Profile Score (GEP); IMPRES[11]: the Immuno-Predictive Score; TIDE[10]: Tumor Immune Dysfunction and Exclusion Score.

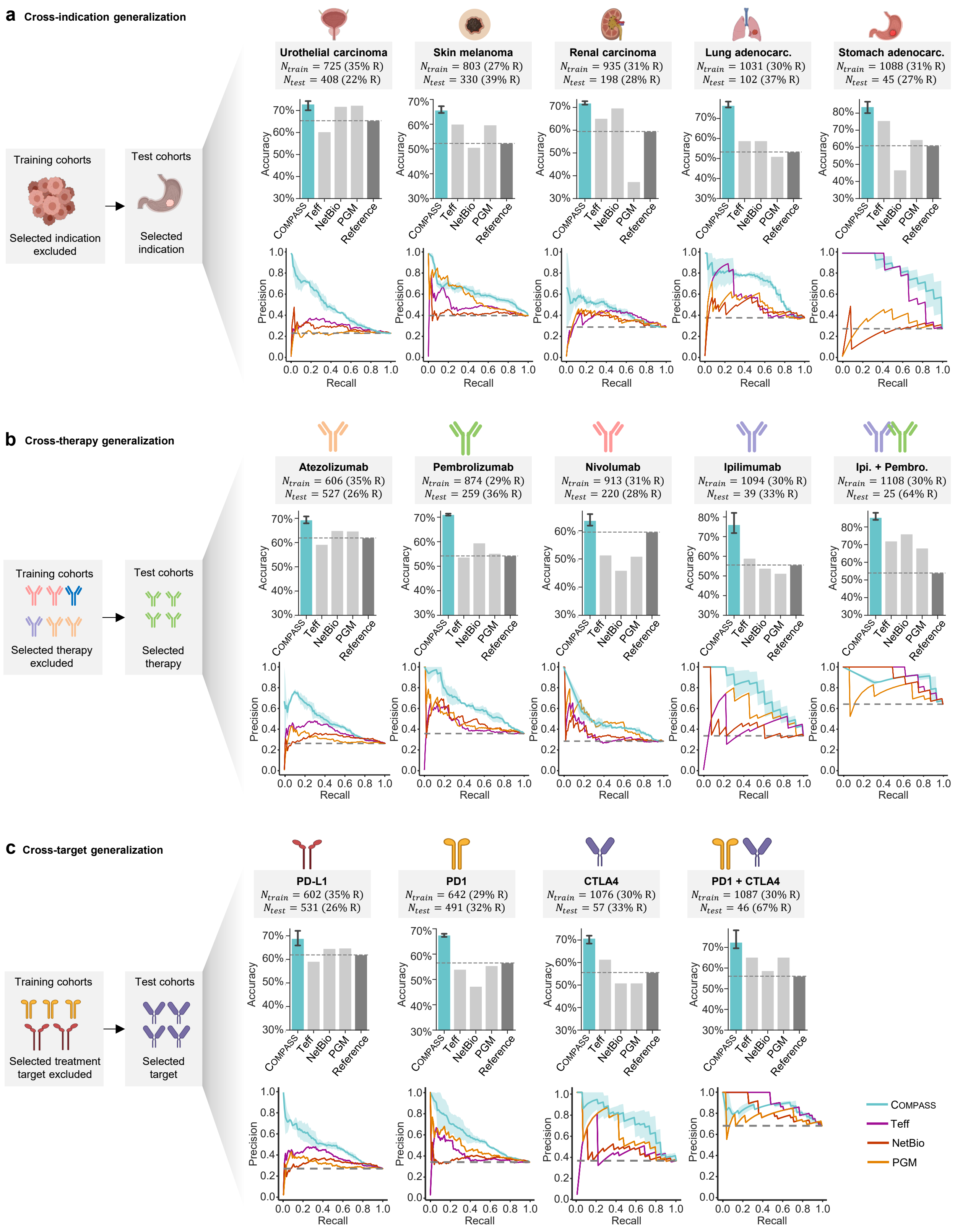
a Cross-indication generalization
Training cohorts
Selected indication excluded
Test cohorts
Selected indication
Urothelial carcinoma
N_{train} = 725 (35% R)
N_{test} = 408 (22% R)
Skin melanoma
N_{train} = 803 (27% R)
N_{test} = 330 (39% R)
Renal carcinoma
N_{train} = 935 (31% R)
N_{test} = 198 (28% R)
Lung adenocarc.
N_{train} = 1031 (30% R)
N_{test} = 102 (37% R)
Stomach adenocarc.
N_{train} = 1088 (31% R)
N_{test} = 45 (27% R)
Accuracy
30%
40%
50%
60%
70%
80%
COMPASS
Teff
NetBio
PGM
Reference
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
b Cross-therapy generalization
Training cohorts
Selected therapy excluded
Test cohorts
Selected therapy
Atezolizumab
N_{train} = 606 (35% R)
N_{test} = 527 (26% R)
Pembrolizumab
N_{train} = 874 (29% R)
N_{test} = 259 (36% R)
Nivolumab
N_{train} = 913 (31% R)
N_{test} = 220 (28% R)
Ipilimumab
N_{train} = 1094 (30% R)
N_{test} = 39 (33% R)
Ipi. + Pembro.
N_{train} = 1108 (30% R)
N_{test} = 25 (64% R)
c Cross-target generalization
Training cohorts
Selected treatment target excluded
Test cohorts
Selected target
PD-L1
N_{train} = 602 (35% R)
N_{test} = 531 (26% R)
PD1
N_{train} = 642 (29% R)
N_{test} = 491 (32% R)
CTLA4
N_{train} = 1076 (30% R)
N_{test} = 57 (33% R)
PD1 + CTLA4
N_{train} = 1087 (30% R)
N_{test} = 46 (67% R)
COMPASS
Teff
NetBio
PGM

Figure 3: Evaluation of model generalization across biological contexts. To assess the generalizability of the COMPASS-PFT model relative to best-performing methods (Teff, NetBio, and PGM), we evaluate the models under biologically meaningful data splits stratified by indication, drug, or ICI target. In the figure, t denotes the number of training samples, n represents the number of test samples, with the percentage of responders in each set annotated. For each evaluation, the models are trained on cohorts excluding a specific category and tested on the excluded cohort. Top panels show prediction accuracy (%), bottom panels precision-recall curves labeled with average precision (AU-PRC).
(a) Cross-disease generalization. Models are trained on ICI cohorts excluding one specific indication and then tested on the excluded indication. For example, the stomach adenocarcinoma cohorts are excluded during training and used for testing. The results show prediction accuracy and precision-recall curves for urothelial carcinoma, skin melanoma, renal carcinoma, lung adenocarcinoma, and stomach adenocarcinoma (**Supplementary Table S4**).
(b) Cross-therapy generalization. Models are trained on cohorts excluding those treated with a specific drug and then tested on the excluded therapy. For instance, the pembrolizumab treatment cohorts are excluded during training and used for testing. The results include prediction accuracy and precision-recall curves for atezolizumab, pembrolizumab, nivolumab, ipilimumab, and a combination of ipilimumab and pembrolizumab (**Supplementary Table S5**).
(c) Cross-target generalization. Models are trained on cohorts excluding those targeting a specific immune checkpoint and then tested on the excluded target. For example, the CTLA4 treatment cohort is excluded during training and used for testing. The results cover prediction accuracy and precision-recall curves for PDL1, PD1, CTLA4, and a combination of PD1 and CTLA4 (**Supplementary Table S6**).

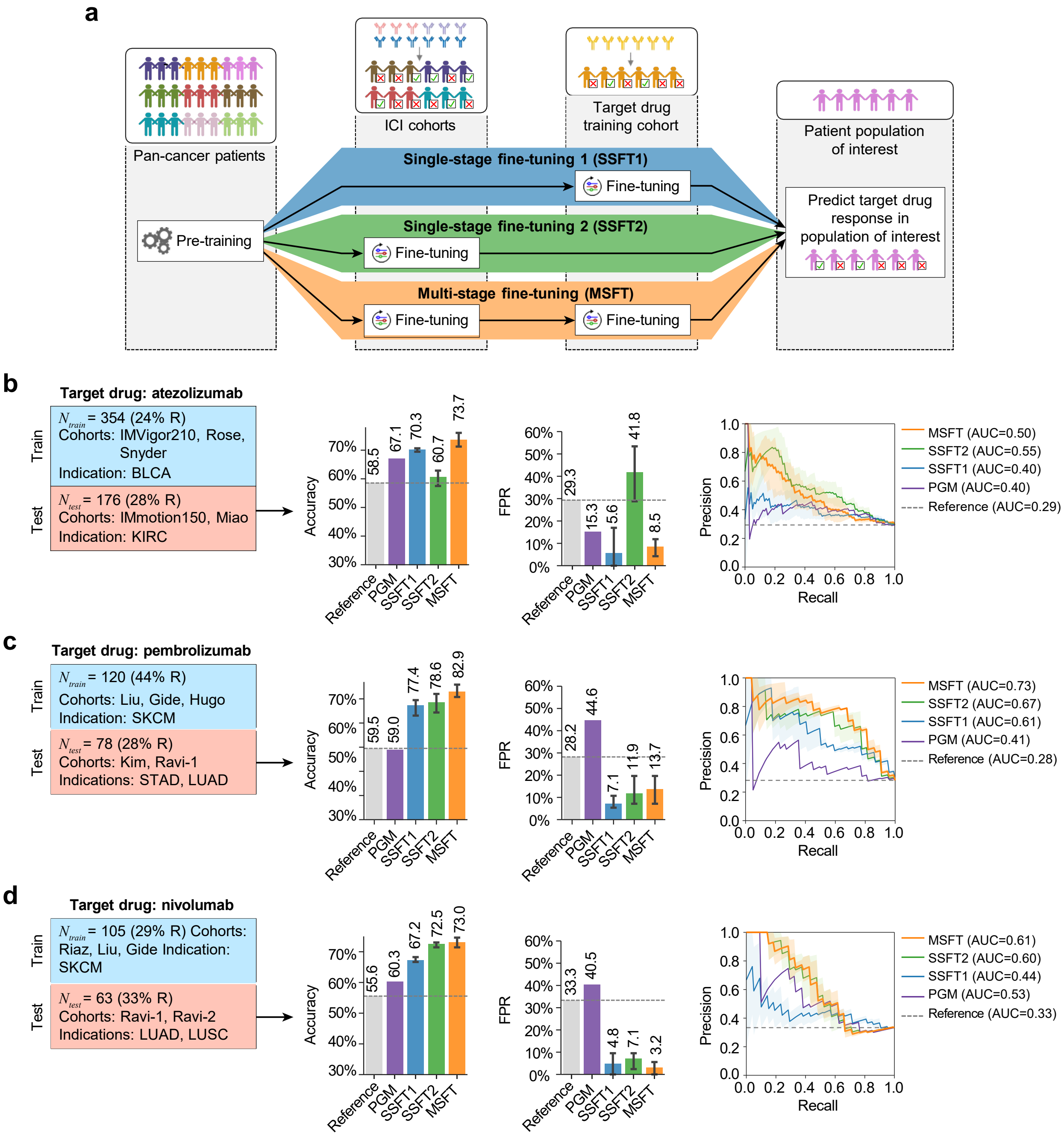
a
Pan-cancer patients
ICI cohorts
Target drug training cohort
Patient population of interest
Pre-training
Single-stage fine-tuning 1 (SSFT1)
Fine-tuning
Single-stage fine-tuning 2 (SSFT2)
Fine-tuning
Multi-stage fine-tuning (MSFT)
Fine-tuning
Fine-tuning
Predict target drug response in population of interest
b
Target drug: atezolizumab
Train
N_{train} = 354 (24% R)
Cohorts: IMVigor210, Rose, Snyder
Indication: BLCA
Test
N_{test} = 176 (28% R)
Cohorts: IMmotion150, Miao
Indication: KIRC
Accuracy
70%
60%
50%
40%
30%
58.5
67.1
70.3
60.7
73.7
Reference
PGM
SSFT1
SSFT2
MSFT
FPR
60%
50%
40%
30%
20%
10%
0%
29.3
15.3
5.6
41.8
8.5
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
MSFT (AUC=0.50)
SSFT2 (AUC=0.55)
SSFT1 (AUC=0.40)
PGM (AUC=0.40)
Reference (AUC=0.29)
c
Target drug: pembrolizumab
N_{train} = 120 (44% R)
Cohorts: Liu, Gide, Hugo
Indication: SKCM
N_{test} = 78 (28% R)
Cohorts: Kim, Ravi-1
Indications: STAD, LUAD
59.5
59.0
77.4
78.6
82.9
28.2
44.6
7.1
11.9
13.7
MSFT (AUC=0.73)
SSFT2 (AUC=0.67)
SSFT1 (AUC=0.61)
PGM (AUC=0.41)
Reference (AUC=0.28)
d
Target drug: nivolumab
N_{train} = 105 (29% R) Cohorts: Riaz, Liu, Gide Indication: SKCM
N_{test} = 63 (33% R)
Cohorts: Ravi-1, Ravi-2
Indications: LUAD, LUSC
55.6
60.3
67.2
72.5
73.0
33.3
40.5
4.8
7.1
3.2
MSFT (AUC=0.61)
SSFT2 (AUC=0.60)
SSFT1 (AUC=0.44)
PGM (AUC=0.53)
Reference (AUC=0.33)

Figure 4: Development and evaluation of drug-specific models using single-stage and multi-stage fine-tuning strategies.
(a) Fine-tuning strategies based on the pre-trained COMPASS model. Drug-specific models are developed using three strategies: single-stage fine-tuning 1 (SSFT1), which fine-tunes directly on the drug-specific cohort; single-stage fine-tuning 2 (SSFT2), which fine-tunes on general ICI-treated cohorts; and multi-stage fine-tuning (MSFT), which sequentially fine-tunes on both (i.e., general then drug-specific cohorts). PGM and reference baseline models are trained using only the drug-specific cohorts. Patient cohorts are mutually exclusive between SSFT1 and SSFT2 stages.
(b–d) Model performance across three drug settings. Drug-specific models were developed and evaluated for: (b) atezolizumab (anti–PD-L1), (c) pembrolizumab (anti–PD-1), and (d) nivolumab (anti–PD-1). Each subplot summarizes training and test cohort composition, including sample sizes (N_{train} and N_{test}) and responder proportions (**Supplementary Table S7**). Training schemes differ by fine-tuning strategy: SSFT1 uses only drug-specific cohorts (e.g., bladder cancer for atezolizumab); SSFT2 uses broader ICI cohorts, excluding treatments that target the same checkpoint pathway; MSFT combines both. All models are evaluated on held-out cancer types. Bar plots show accuracy and false positive rates (FPR), and line plots display precision-recall curves with area under the curve (AUC) values. MSFT consistently outperforms SSFT1, SSFT2, PGM, and reference baselines. Complete results in **Supplementary Table S8**.

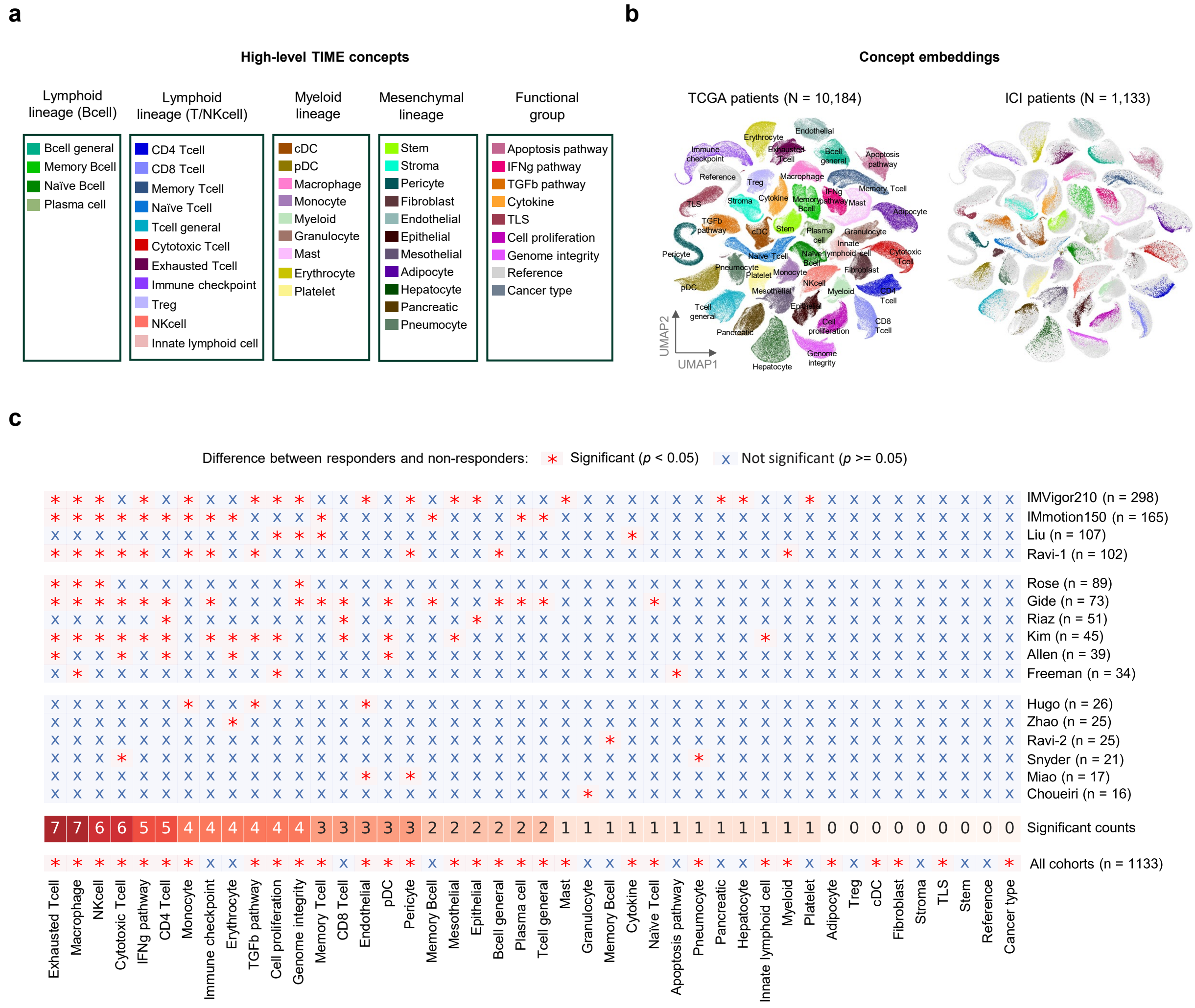

Figure 5: Analysis of COMPASS model concepts across cohorts and their association with treatment response. To investigate the biological relevance of the learned concept features, we analyze the 44 COMPASS-derived concepts and their associations with immunotherapy response.

(a) Overview of high-level TIME concepts in COMPASS. The 44 concepts are organized into broader biological categories: 4 B cell–related concepts, 11 T/NK cell concepts, 9 myeloid lineage concepts, 11 mesenchymal lineage concepts, and 9 pathway/function-related concepts. Complete gene sets underlying these concepts are provided in **Supplementary Data S1**. Concept similarity based on gene overlap (Jaccard index) is shown in **Supplementary Figure S10**, and variance across cancer types is analyzed in **Supplementary Figure S13**.

(b) UMAP embedding of COMPASS concept representations using TCGA and ICI patients. Each dot represents a patient's 32-dimensional representation for one of the 43 learned concepts from the pre-trained model COMPASS-PT. The UMAP visualizes the 2D embedding of all such patient-level concept vectors, with clusters corresponding to distinct concepts. The left panel shows embeddings from TCGA patients, while the right panel overlays ICI patient embeddings (colored) onto the TCGA-derived UMAP (gray), demonstrating cross-population consistency. This alignment highlights the robustness and domain generalizability of the learned concept features. Additional UMAP and PCA visualizations for both granular and high-level concept embeddings are shown in **Supplementary Figures S11-S12**.

(c) Predictive significance of COMPASS concepts. The heatmap displays the statistical significance of TIME concepts for distinguishing responders from non-responders across 16 cohorts. Using leave-one-cohort-out validation, COMPASS-PFT models were iteratively trained while holding out each cohort for testing. The bottom row ("All cohorts") shows significance when fine-tuned on the complete dataset. Significant associations ($p < 0.05$) are marked with red asterisks; non-significant results with blue crosses.

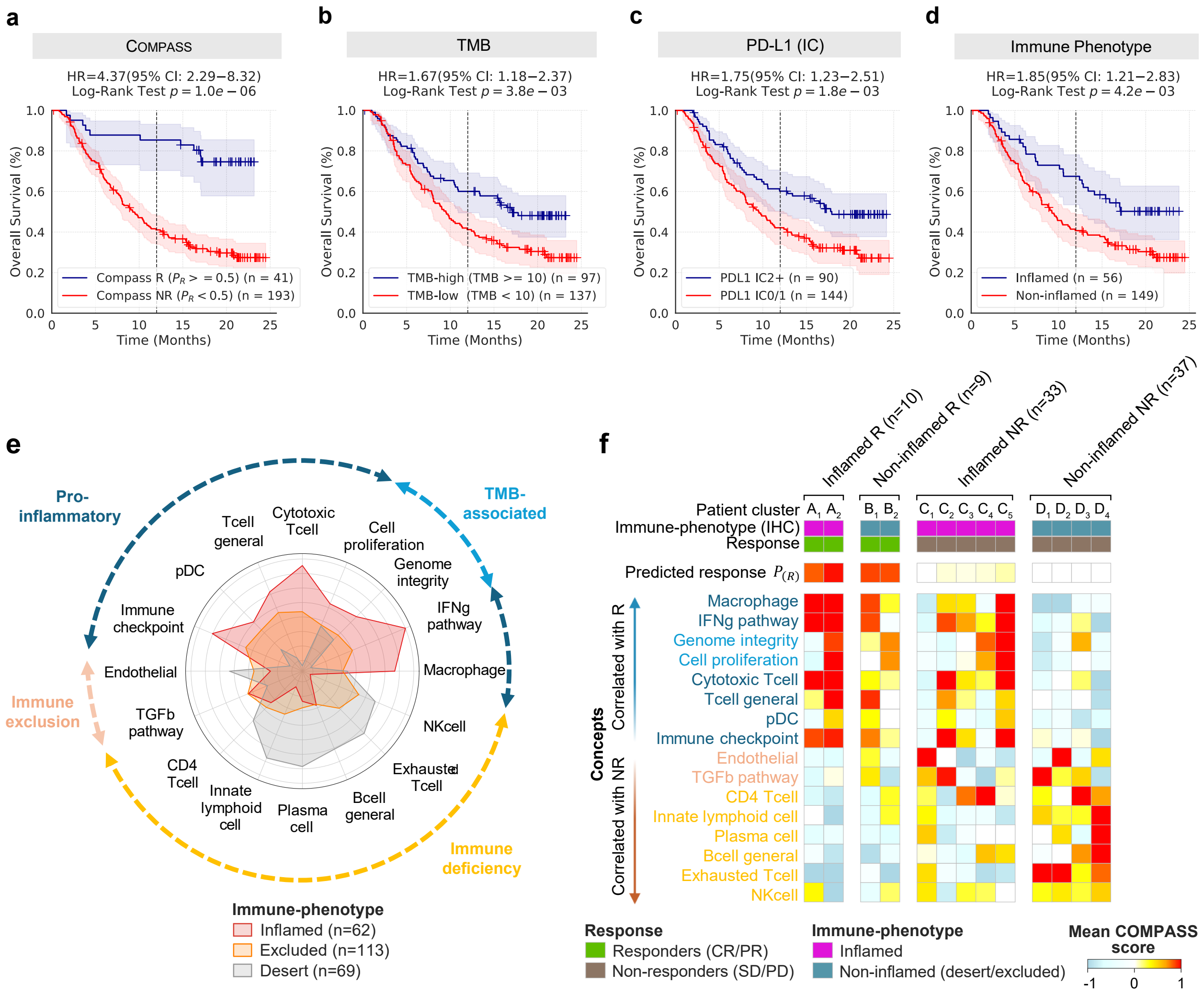

Figure 6: Survival prediction and tumor-immune profiling in atezolizumab-treated urothelial carcinoma. COMPASS evaluation of atezolizumab-treated urothelial carcinoma patients (IMvigor210 cohort), with partial fine-tuning on external ICI-treated cohorts to prevent data leakage.

(a–d) Kaplan–Meier survival analysis. Overall survival stratified by: (a) COMPASS-predicted response probability ($P_R \geq 0.5$ vs $P_R < 0.5$), (b) tumor mutation burden (TMB-high [$\geq$ 10 mut/Mb] vs TMB-low), (c) PD-L1 immune cell score (IC2+ vs IC0/1), and (d) immune phenotype (inflamed vs excluded/desert). Analysis restricted to patients with available TMB data ($n = 234$). COMPASS-predicted responders showed superior survival benefit (HR = 4.37, 95% CI: 2.29–8.32, $p = 1.0 \times 10^{-6}$) compared to conventional biomarkers.

(e) Immune phenotype-specific concept profiles. Radar plot displays average scores of 16 key COMPASS-derived TIME concepts across immune phenotypes: inflamed (red, $n = 62$), excluded (orange, $n = 113$), and desert (beige, $n = 69$). Concepts are grouped by functional category (**Method 9**): pro-inflammatory (dark blue), TMB-associated (blue), immune exclusion (peach), and immune deficiency (yellow). Inflamed tumors show elevated pro-inflammatory and TMB-associated concept scores, while excluded/desert phenotypes exhibit immunosuppressive features.

(f) Concept score patterns across immune phenotype and response subgroups. Heatmap shows average scores of the top 16 COMPASS-derived TIME concepts across four patient subgroups: inflamed responders, non-inflamed responders, inflamed non-responders, and non-inflamed non-responders. Patients are clustered within each subgroup by concept score similarity (**Supplementary Figure S22**), revealing distinct TIME concept profiles. The top row indicates mean COMPASS-predicted response probability (P_R) per cluster, demonstrating the relationship between concept activation patterns and clinical outcomes.

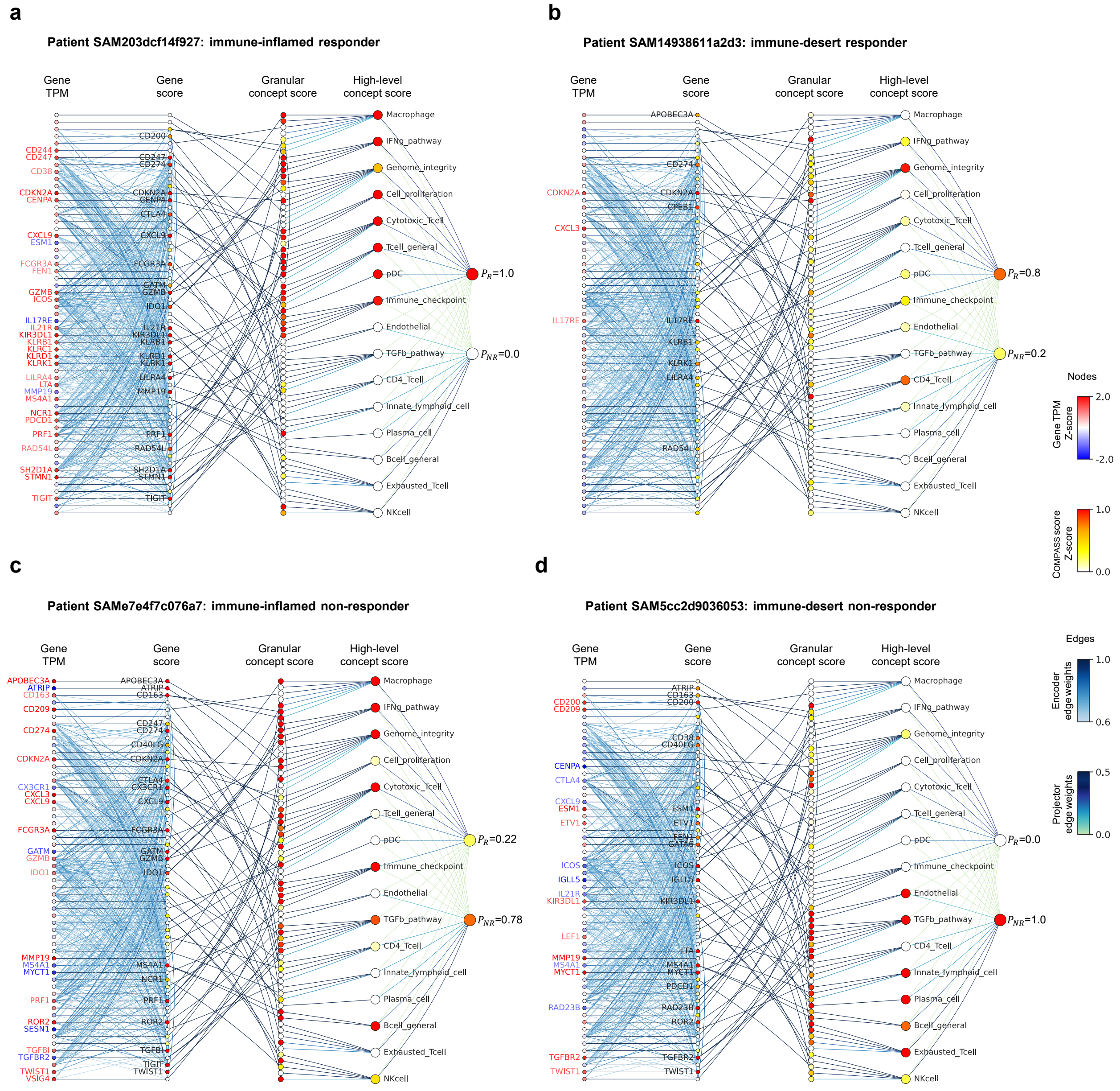

Figure 7: Personalized response maps explain ICI outcome predictions for individual patients. The maps trace how COMPASS connects input gene expression profiles to immunotherapy predictions through its interpretable concept hierarchy. Four representative patients from **Figure 6f** clusters are shown (**a-d**, additional examples in **Supplementary Figures S23–S26**). Each map displays: Z-score normalized gene expression levels (TPM), Z-score normalized gene and concept scores, predicted probabilities (response [P_R] and non-response [P_{NR}]), and model attention weights. For clarity, only the top-16 high-level concepts and the dominant contributing gene of each granular concept are shown (zoomed views: **Supplementary Figure S27**). An interactive tool for exploring personalized response maps is available at the COMPASS website. **(a)** Inflamed responder — Patient SAM203dcf14f927 (cluster A_3, $P_R = 1.0$); **(b)** Desert responder — Patient SAM14938611a2d3 (cluster B_2, $P_R = 0.8$); **(c)** Inflamed non-responder — Patient SAMe7e4f7c076a7 (cluster C_5, $P_R = 0.22$); **(d)** Desert non-responder — Patient SAM5cc2d9036053 (cluster D_2, $P_R = 0.0$).

Methods

The Methods describe: (1) dataset curation and pre-processing, (2) the COMPASS model, (3) self-supervised pre-training of COMPASS, (4) supervised fine-tuning for response prediction, (5) benchmarking COMPASS models against established methods, (6) multi-stage fine-tuning for drug- and disease-specific models, (7) SHAP analysis of important features, (8) overall survival analysis, (9) TIME concept analysis in the IMvigor210 cohort, and (10) personalized response maps generation.

1 Dataset curation and pre-processing

1.1 The Cancer Genome Atlas (TCGA) datasets

Pre-training datasets are acquired from The Cancer Genome Atlas (TCGA) via the Genomic Data Commons (GDC) portal (version 37; GDC Portal), using TCGAbiolinks[71] for data retrieval. To ensure cross-cohort compatibility with downstream ICI analyses, all RNA-seq data are uniformly processed through our standardized pipeline. Read alignment is performed against the GRCh38/hg38 reference genome using STAR[72] (v2.7.5c), with gene features annotated according to GENCODE v36.

Raw counts are normalized by gene effective length and converted to TPM:

$$\mathrm{TPM}_i = \left(\frac{\mathrm{RPK}_i}{\sum_{j=1}^{N} \mathrm{RPK}_j}\right) \times 10^6,$$

where N is the number of genes and

$$\mathrm{RPK}_i = \frac{\text{Normalized count of gene } i}{\text{Length of gene } i \text{ in kb}}$$

This normalization facilitates cross-sample comparability. Initial data included 60,660 genes across 11,274 samples. After excluding normal tissue samples, 10,534 samples remained. Further exclusions were applied for prior treatment samples and non-FFPE samples, resulting in 10,305 samples. Finally, aggregation to the patient level using the ”bcr patient barcode” key yielded 10,184 unique patient tumor samples. Protein-coding genes are selected (15,672 genes) based on overlap with the gene expression data from the clinical cohorts (see next section).

1.2 Immune checkpoint inhibitor (ICI) clinical cohorts

We curate 16 cohorts (**Figure 2a**) spanning 7 cancer types, categorized into three groups: large cohorts (>100 patients), medium-sized cohorts (30-100 patients), and small cohorts (<30 patients).

Publicly available RNA-seq data underwent uniform processing through the same standardized pipeline that was used to process the TCGA data (see previous section), converting raw sequencing data (FASTQ) to counts and TPM values. For cohorts with available raw data, FASTQ files were reprocessed; otherwise, TPM values were derived from the counts using TCGA-aligned gene lengths. To ensure cross-cohort consistency, all data were mapped to the same reference genome, correcting for potential differences in original genomic builds. To ensure reproducibility, researchers may download raw data using accession IDs in **Supplementary Table S1** and reprocess it via our publicly available code (https://github.com/mims-harvard/COMPASS-web/tree/main/mRNA_pipeline), which includes parameters for alignment, quantification, and TPM conversion. All steps rely on the GRCh38 reference genome and GENCODE v36 annotations to maintain cross-cohort consistency. Only pre-treatment samples are included. Responders are defined as patients achieving partial response (PR) or complete response (CR), while non-responders include those with stable disease (SD) or progressive disease (PD), per RECIST v1.1 BOR criteria as reported in source studies, unless otherwise noted. See **Supplementary Table S1** for cohort overview.

Large cohorts. The IMvigor210 cohort ($n = 298$) includes atezolizumab-treated bladder cancer patients (68 responders, 230 non-responders)[18], with data sourced from the IMvigor210CoreBiologies (v1.0.1) R package and CRI iAtlas[62]. The IMmotion150 cohort ($n = 165$) comprises atezolizumab-treated renal cell carcinoma patients (48 responders, 117 non-responders)[19]. For the cohort from Liu *et al.* ($n = 107$), post-treatment samples are excluded, retaining 41 melanoma patients (nivolumab/pembrolizumab) classified as responders and 66 as non-responders[21]. The Ravi-1 cohort ($n = 102$) is a sub-cohort of the SU2C-MARK NSCLC study[20], focusing on lung adenocarcinoma (LUAD) patients receiving PD-(L)1 ± CTLA4 inhibitors.

Medium-sized cohorts. The Rose *et al.* cohort ($n = 89$) includes bladder cancer patients treated with PD-(L)1 inhibitors (16 responders and 73 non-responders)[26]. The Gide *et al.* ($n = 73$)[25] cohort comprises melanoma patients receiving anti-PD-1 ± anti-CTLA4 (40 responders and 33 non-responders). Additional cohorts include: Riaz *et al.* ($n = 51$) with nivolumab-treated melanoma patients (10 responders and 41 non-responders)[8]; Kim *et al.* ($n = 45$) with pembrolizumab-treated stomach adenocarcinoma patients (12 responders and 33 non-responders)[24]; Van Allen *et al.* ($n = 39$) with ipilimumab-treated melanoma patients (26 responders [CR/PR or SD with overall survival > 1 year] and 13 non-responders [PD or SD with OS < 1 year])[23]; and Freeman *et al.* ($n = 34$) with melanoma patients from the MGH cohort treated with nivolumab, pembrolizumab, ipilimumab, or combination therapies (12 responders and 22 non-responders)[22].

Small cohorts. The Hugo *et al.* cohort ($n = 26$) involves pembrolizumab-treated melanoma patients (14 responders and 12 non-responders by irRECIST)[31]. The Ravi-2 cohort ($n = 25$), represents a SU2C-MARK NSCLC sub-study[20] of squamous cell carcinoma (LUSC) patients treated with PD-1 or PD-L1 inhibitors (8 responders and 17 non-responders). For the Zhao *et al.* cohort ($n = 25$), glioblastoma patients receiving nivolumab or pembrolizumab are classified as responders based on either: (1) post-treatment histopathology showing inflammatory response with minimal/no residual tumor cells, or (2) radiographic evidence of stable/shrinking tumor volume over six months[30]. The Snyder *et al.* cohort ($n = 21$) includes atezolizumab-treated bladder cancer (BLCA) patients (7 responders and 14 non-responders)[29]. For renal cell carcinoma (KIRC) patients: the Choueiri *et al.* cohort ($n = 16$) contains nivolumab-treated cases (3 responders and 13 non-responders)[27], while the Miao *et al.* cohort ($n = 17$) includes patients receiving PD-(L)1 ± CTLA4 inhibitors (5 responders and 12 non-responders)[28], both sourced from CRI iAtlas[62]. No new datasets are generated in this study.

2 The COMPASS model

The COMPASS model comprises three key components. The first component, a transformer-based Gene Language Model (GLM), serves as the **encoder** to generate contextualized representations of individual genes. Next, a hierarchical **projector** transforms these gene-level embeddings into high-level biological concepts, including immune cell types and pathways. The final component, a **classifier**, performs immunotherapy response prediction from the concept representations, employing either a multilayer perceptron (MLP) or a non-parametric, similarity-based method for zero-shot prediction.

2.1 GLM encoder

The Gene Language Model (GLM) adapts natural language modeling techniques to transcriptomic data[52, 54, 73, 74], where each gene is treated as a token. Unlike natural language where tokens follow a clear sequential structure, gene expression profiles are inherently unordered and represented in tabular format. This fundamental difference renders traditional positional encodings—such as fixed sinusoidal encodings used in NLP—suboptimal for capturing gene-gene relationships[73, 74]. Drawing inspiration from FT-Transformer[75], which is designed for tabular data, we introduce a learnable gene-specific positional bias that enables the model to capture contextual interactions between genes in a biologically informed, data-driven manner.

Gene abundance embedding. Let $\mathbf{X}_{\text{gene}} \in \mathbb{R}^{B \times L}$ denote the input gene expression matrix, where B is the batch size and L is the number of genes. Each gene is embedded into a d-dimensional latent space using a learnable embedding matrix $\mathbf{W} \in \mathbb{R}^{L \times d}$, initialized from a uniform distribution:

$$\mathbf{W} \sim \mathcal{U}\left(-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right)$$

To generate expression-aware embeddings, we scale each gene's embedding vector by its corresponding expression value. Specifically, the embedding for the l-th gene in the b-th sample is given by:

$$\mathbf{E}_{\text{gene}}[b, l, :] = \mathbf{X}_{\text{gene}}[b, l] \cdot \mathbf{W}[l, :]$$

for $b = 1, \ldots, B$ and $l = 1, \ldots, L$, resulting in the embedding tensor $\mathbf{E}_{\text{gene}} \in \mathbb{R}^{B \times L \times d}$. This design enables the model to capture both gene identity (via $\mathbf{W}$) and sample-specific abundance (via $\mathbf{X}_{\text{gene}}$) in the representation.

Learnable positional encoding. To inject gene-specific inductive biases into the model, we introduce a learnable positional encoding matrix $\mathbf{P} \in \mathbb{R}^{L \times d}$, initialized in the same manner:

$$\mathbf{P} \sim \mathcal{U}\left(-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right)$$

Each gene receives a unique, trainable positional vector $\mathbf{P}[l, :]$, which acts as a contextual bias. The final input embedding for each gene in each sample is computed by element-wise addition of the positional encoding:

$$\mathbf{E}_{\text{final}}[b, l, :] = \mathbf{E}_{\text{gene}}[b, l, :] + \mathbf{P}[l, :]$$

resulting in $\mathbf{E}_{\text{final}} \in \mathbb{R}^{B \times L \times d}$. Unlike fixed encodings, this learnable scheme allows the model to adaptively encode gene-level functional relevance during training, thereby serving as a gene-aware bias that enhances the transformer's capacity to model context-specific gene interactions.

Cancer type token embedding. To account for pan-cancer heterogeneity, COMPASS integrates a cancer type token that interacts with gene tokens through attention mechanisms and is separately projected as a `cancer type` concept in the model's hierarchy. To generate the cancer type token embedding, the 33 cancer types are first encoded as integers (0-32). This integer encoding serves as an index for looking up a learnable embedding matrix:

$$\mathbf{W}_{\text{cancer}} \in \mathbb{R}^{33\times d}$$

Given a batch of cancer type labels $\mathbf{X}_c \in \mathbb{R}^B$, we perform a lookup to obtain their embeddings:

$$\mathbf{E}_{\text{cancer}} = \mathbf{W}_{\text{cancer}}[\mathbf{X}_c] \in \mathbb{R}^{B\times d}$$

These embeddings are reshaped as $\mathbf{E}_{\text{cancer}} \in \mathbb{R}^{B\times 1\times d}$, and prepended to the gene embeddings along the sequence axis. This cancer-type token interacts with gene tokens via self-attention and is later projected into a dedicated `cancer type` concept node in the concept hierarchy.

Transformer encoder for contextual learning. The full input to the transformer encoder is constructed by concatenating the cancer type token and gene embeddings:

$$\mathbf{H}^{(0)} = \text{Concat}([\mathbf{E}_{\text{cancer}}, \mathbf{E}_{\text{final}}]) \in \mathbb{R}^{B\times(L+1)\times d}$$

This tensor passes through a multi-layer transformer encoder composed of stacked self-attention and feedforward layers:

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{H}^{(0)})$$

Within each layer, the self-attention mechanism enables each token to dynamically attend to all others, including gene-gene and gene–cancer-type interactions. The attention weights are computed via:

$$A(\mathbf{H}_i, \mathbf{H}_j) = \text{Softmax}\left(\frac{Q(\mathbf{H}_i)K(\mathbf{H}_j)^\top}{\sqrt{d_k}}\right)$$

Given the large number of genes typically used in the model, full attention becomes computationally expensive. To address this, we adopt the Performer architecture[76], which replaces standard attention with a linear approximation while preserving expressiveness. Additionally, the architecture supports FlashAttention[77] as an optional alternative to further reduce memory overhead and runtime. The output tensor $\mathbf{H} \in \mathbb{R}^{B\times(L+1)\times d}$ encodes the contextualized representations of both the gene expression profile and cancer type context, which are then used for downstream biological concept projection and prediction.

2.2 Concept-based hierarchical projector

To convert contextualized gene-level embeddings into interpretable biological features, we introduce a hierarchical projector that maps the transformer output onto two levels of concept representations: (i) granular concept scores (S_{Geneset}), each corresponding to a curated gene set (e.g., a pathway or immune cell signature); and (ii) high-level concept score (S_{Concept}), which integrates biologically related gene sets into functional modules (e.g., immune activation, suppression, or dysfunction). This hierarchical design enables multi-scale reasoning over biological processes within the tumor microenvironment.

Granular concept (gene set) aggregation Given a curated gene set $G = \{g_1, g_2, \ldots, g_k\}$ with k genes, we extract their contextualized embeddings from the GLM output $\mathbf{H}'_{\text{gene}} \in \mathbb{R}^{B\times L\times d}$:

$$\mathbf{H}_G = \mathbf{H}'_{\text{gene}}[:, G, :] \in \mathbb{R}^{B\times k\times d}$$

To aggregate the k gene vectors into a unified representation, we introduce a learnable attention vector $\mathbf{a}_G \in \mathbb{R}^k$, initialized from a normal distribution and normalized via softmax:

$$\mathbf{a}_G^{\text{softmax}} = \text{Softmax}(\mathbf{a}_G)$$

The attention-weighted gene set embedding is then computed as:

$$\mathbf{H}_G^{\text{agg}} = \sum_{i=1}^{k} \mathbf{a}_G^{\text{softmax}}[i] \cdot \mathbf{H}_G[:, i, :] \in \mathbb{R}^{B\times d}$$

This aggregated vector is passed through a linear layer to produce the scalar score for gene set G:

$$S_{\text{Geneset}}(G) = \text{Linear}(\mathbf{H}_G^{\text{agg}}) \in \mathbb{R}^B$$

The full set of gene sets yields a tensor $S_{\text{Geneset}} \in \mathbb{R}^{B\times K}$, where K denotes the total number of gene sets in the model. This mid-level representation captures modular biological information across curated pathways and cell types.

High-level concept aggregation Each high-level concept C consists of a subset of gene sets $\{G_1, G_2, \ldots, G_{k_C}\}$, where k_C is the number of gene sets associated with concept C. The corresponding gene set scores $\{S_{\text{Geneset}}(G_1), \ldots, S_{\text{Geneset}}(G_{k_C})\}$ are aggregated using a second-level attention mechanism. A learnable attention vector $\mathbf{a}_C \in \mathbb{R}^{k_C}$ is normalized via softmax:

$$\mathbf{a}_C^{\text{softmax}} = \text{Softmax}(\mathbf{a}_C)$$

The high-level concept score is then computed as:

$$S_{\text{Concept}}(C) = \sum_{j=1}^{k_C} \mathbf{a}_C^{\text{softmax}}[j] \cdot S_{\text{Geneset}}(G_j)$$

This attention-based aggregation allows each high-level concept to dynamically weight its constituent gene sets, enabling flexible and interpretable summarization of complex biological programs.

Final concept representation The COMPASS framework defines $M = 43$ high-level concepts derived from immune-related pathways, cell types, and functional groups. One additional concept represents the cancer type. The final concept representation for each sample is:

$$\mathbf{C}_{\text{final}} = \text{Concat}\left(S_{\text{Concept}}(C_1), \ldots, S_{\text{Concept}}(C_M), S_{\text{Concept}}(\text{Cancer})\right) \in \mathbb{R}^{B\times(M+1)}$$

Here, the cancer type score is computed by projecting the cancer token embedding $\mathbf{H}_{\text{cancer}} \in \mathbb{R}^{B\times d}$ through a linear layer:

$$S_{\text{Concept}}(\text{Cancer}) = \text{Linear}(\mathbf{H}_{\text{cancer}})$$

Together, the 44-dimensional vector $\mathbf{C}_{\text{final}}$ provides a biologically grounded and interpretable embedding of each patient's tumor microenvironment, suitable for downstream prediction and mechanistic analysis.

2.3 Task classifier module

The task classifier module in the COMPASS model performs the conversion of high-level concept representations into probabilistic predictions of immunotherapy response. To accommodate both standard supervised learning and generalization to new cohorts or cancer types, COMPASS supports two distinct classifier types: a parametric multilayer perceptron (MLP) and a non-parametric cosine similarity-based Prototypical Network. The latter is referred to as the NFT (Non-Fine-Tuned) classifier, as it operates without gradient-based optimization during inference. These classifier heads provide complementary strengths and can be selected based on the availability of training labels and the desired generalization behavior.

MLP classifier (parametric). The MLP classifier transforms high-level concept vectors into binary predictions via a trainable feedforward network. Given an input matrix $\mathbf{X} \in \mathbb{R}^{B\times 44}$, where B is the batch size and 44 is the number of concepts (43 biological concepts plus 1 cancer type), the inputs are first standardized:

$$\mathbf{X}_{\text{norm}} = \text{BatchNorm}(\mathbf{X})$$

The normalized vectors are then passed through fully connected layers, producing output logits $\mathbf{Z} \in \mathbb{R}^{B\times 2}$:

$$\mathbf{Z} = \text{Linear}(\mathbf{X}_{\text{norm}})$$

To allow calibration of the output confidence, the logits are scaled by a learnable temperature parameter τ, defined as $\tau = \exp(\text{log\_temperature})$:

$$\mathbf{Z}_{\text{scaled}} = \frac{\mathbf{Z}}{\tau}$$

These scaled logits are transformed into probabilities using the softmax function:

$$\mathbf{y} = \text{Softmax}(\mathbf{Z}_{\text{scaled}})$$

The learnable temperature τ allows the model to adjust the sharpness of its predictions and improves its ability to distinguish ambiguous cases, such as borderline responders.

Prototypical classifier (non-parametric, NFT). The NFT (no fine-tuning) classifier adopts a Prototypical Network architecture[78] that performs inference through similarity comparisons with labeled support examples, completely eliminating the need for model fine-tuning. This non-parametric approach is advantageous when limited training data prevents effective model adaptation.

As illustrated in **Supplementary Figure S1a**, the classifier begins with a support set of labeled patient embeddings, each represented by a high-level concept vector $\mathbf{f} \in \mathbb{R}^{44}$ and a binary response label.

The support examples are grouped by class $c \in \{\text{responder}, \text{non-responder}\}$, and a class prototype is computed by averaging the normalized vectors in each group:

$$\mathbf{p}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$$

where $\mathbf{x}_i \in \mathbb{R}^{44}$ is the concept vector of the i-th support sample in class c, and N_c is the number of support examples in that class. The resulting prototypes $\mathbf{p}_c$ are unit-normalized to enable cosine-based comparison.

Given a query patient with concept vector $\mathbf{q}$, the classifier computes the cosine similarity between the query and each class prototype:

$$\text{similarity}(\mathbf{q}, \mathbf{p}_c) = \frac{\mathbf{q} \cdot \mathbf{p}_c}{\|\mathbf{q}\| \cdot \|\mathbf{p}_c\|}$$

The similarity scores are scaled by a fixed temperature τ (typically 0.1), then passed through a softmax layer:

$$\text{logits}_c = \frac{\text{similarity}(\mathbf{q}, \mathbf{p}_c)}{\tau}, \quad \mathbf{y} = \text{Softmax}(\text{logits})$$

This produces a probability distribution over the binary response classes.

3 Pre-training COMPASS on TCGA

The COMPASS model was pre-trained on bulk RNA-seq data from 10,184 patients across 33 TCGA cancer types using a self-supervised triplet contrastive learning approach. This framework learns to map tumor transcriptomes (TPM values) into a 44-dimensional concept embedding space that captures TIME features. During training, each triplet consists of an anchor sample (a patient's transcriptome), a positive sample (an augmented version of the same transcriptome), and a negative sample (a transcriptome from a different patient within the same cancer type). The model optimizes the embedding space to minimize cosine distance between anchor-positive pairs while maximizing separation from negative samples.

To address imbalance in TCGA cohort sizes across cancer types, we implemented balanced sampling with replacement, upweighting underrepresented cancer types during training. This ensures all cancer types contribute proportionally to the learned representations.

Data augmentation. For contrastive learning, each anchor transcriptome undergoes stochastic transformation via one of two augmentation methods. Random masking independently zeros each gene's expression value x_i with probability $p_{mask} = 0.1$ according to:

$$x'_i = \begin{cases} 0 & \text{with probability } 0.1 \\ x_i & \text{otherwise} \end{cases}$$

Alternatively, Gaussian jittering adds normally distributed noise to each value:

$$x''_i = x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 0.1)$$

The augmentation method is selected randomly for each transformation event.

Self-supervised optimization. The model is trained using a margin-based triplet loss function:

$$L_{\text{CL}} = \max(1 - \cos(f(a), f(p)) - (1 - \cos(f(a), f(n))) + \text{margin}, 0)$$

where $f(a)$, $f(p)$, and $f(n)$ represent anchor, positive, and negative embeddings respectively, $\cos$ denotes cosine similarity, and a default margin of 1 is used.

Training was conducted on NVIDIA Tesla A100 80GB GPUs with batch size 128 and learning rate 1e-3 using the Adam optimizer. We reserved 1% of samples as a validation set for early stopping, with training halted if validation loss failed to improve for 10 consecutive epochs. For robustness, we performed three independent training runs with random seeds 24, 42, and 64, selecting the checkpoint with lowest validation loss for downstream use.

4 Fine-tuning COMPASS for response prediction

Following pre-training on TCGA, the COMPASS model is fine-tuned on ICI-treated cohorts to predict clinical response. To accommodate varying dataset sizes and quality across cohorts, we implemented four complementary fine-tuning strategies with differing levels of parameter adaptation: non-parametric zero-shot inference (COMPASS-NFT mode), linear probing (COMPASS-LFT mode), partial fine-tuning (COMPASS-PFT mode), and full model fine-tuning (COMPASS-FFT mode). These approaches provide a spectrum from maximal parameter efficiency (COMPASS-NFT) to full model adaptability (COMPASS-FFT).

All parametric modes (COMPASS-LFT, COMPASS-PFT, COMPASS-FFT) process the 44-dimensional concept vector through a single-layer dense classifier with 16 hidden units, generating logit outputs for binary response prediction. The trainable parameters vary substantially across modes: COMPASS-FFT updates all model parameters (approximately 1,018,784 total), including the GLM encoder and projector; COMPASS-PFT adapts only the classifier and projection layers (2,144 parameters); COMPASS-LFT modifies solely the classifier head (182 parameters); while COMPASS-NFT maintains frozen pre-trained weights with no trainable parameters, instead using prototypical inference based on cosine similarity in concept space (detailed in Method 2.3).

For parametric modes, models are optimized using cross-entropy loss with learning rates between 10^{-3} and 10^{-2} (slightly higher than the pre-training learning rate to promote faster adaptation to new domains), batch sizes of 8-16 (scaled to cohort size and GPU memory), and weight decay (10^{-2} and 10^{-4}) tuned per mode and dataset. Internal cross-validation determined optimal training epochs, with FFT typically converging faster but being more prone to overfitting compared to the more stable LFT and PFT approaches in small datasets. All experiments ran on NVIDIA Tesla

V100 GPUs, with final model selection based on cross-validated validation performance.

The COMPASS-NFT (no fine-tuning) mode represents a distinct non-parametric approach where labeled samples from the training cohort serve as a support set to compute responder/non-responder prototypes in the frozen 44-dimensional concept space. New samples are classified by cosine similarity to these prototypes (detailed in Method 2.3). This enables generalization to new domains without any additional gradient updates, making COMPASS-NFT ideal for low-data or zero-shot transfer scenarios.

5 Comparative evaluation of ICI response prediction models

Overview of baseline methods. We evaluated COMPASS against 22 established ICI response prediction methods (**Supplementary Table S2**), all implemented in our open-source Python module (GitHub repository). These methods encompass three main categories: (1) target gene markers (PD1, PDL1, CTLA4, and their combination [GeneBio][16]); (2) immune cell and functional signatures including Cytotoxic Immune Signature (CIS)[63], T-effector-IFNg Signature (Teff)[32], Neoadjuvant Response Signature (NRS)[64], IFN-γ Signature Score (IFNG)[12], Cytotoxic T Lymphocytes Markers (CTL)[10], Tumor-Associated Macrophages (TAM)[10,65], T-cell Exhaustion (Texh)[10,66], Chemokine Signature Score (CKS)[67], Cancer-Associated Fibroblasts Signature Score (CAF)[68], Roh Immune Score (IS)[33], Immune Cytolytic Activity Score (ICA)[69], CD8 Signature Score (CD8)[16,70], MHC I Association Immune Score (MIAS)[13], and the T Cell-Inflamed Gene Expression Profile Score (GEP)[13,14]; and (3) comprehensive integrative methods including Tumor Immune Dysfunction and Exclusion Score (TIDE)[10], Immuno-Predictive Score (IMPRES)[11], Paired Gene Markers (PGM)[22], and Network-Based ICI Treatment Biomarkers (NetBio)[16].

For benchmarking purposes, we implemented standardized logistic regression models using the marker gene or predictive score from each baseline method (detailed in **Supplementary Table S2**) as input features. For each model, we performed hyperparameter optimization via scikit-learn's `GridSearchCV`[79], employing L2 regularization with the LBFGS solver. All models incorporated balanced class weighting and were configured with a maximum of 10^{10} iterations to guarantee convergence. Through 5-fold cross-validation grid search across the regularization strength range ($C \in [0.1, 1]$), we identified the optimal parameter value using the area under the ROC curve (AUC) as scoring metric. This optimized C value was subsequently used to finalize each logistic regression model for comparative performance evaluation across all baseline methods.

Cross-cohort and within-cohort model validation. To rigorously assess model performance across clinically relevant scenarios, we implemented three complementary validation strategies. First, leave-one-cohort-out validation evaluates cross-cohort generalization by training models on all available cohorts except one held-out cohort, then testing performance exclusively on the excluded cohort. Second, cohort-to-cohort transfer prediction provides a more stringent assessment of cross-cohort generalizability by training models on a single complete cohort and directly predicting outcomes for patients from an entirely different cohort. Third, within-cohort leave-one-patient-out validation assesses intrinsic predictive performance through iterative training on all patients except one within individual cohorts, followed by testing on each excluded patient - this approach provides robust performance estimates in homogeneous clinical settings while effectively controlling for overfitting. These strategies serve distinct purposes: both leave-one-cohort-out validation and cohort-to-cohort transfer prediction examine external validity across diverse clinical populations, while leave-one-patient-out validation focuses on optimizing and evaluating within-study performance.

Two baseline methods, NetBio and TIDE, implement specialized prediction approaches tailored to specific therapies and cancer types respectively. During cross-cohort transfer evaluations, both methods were carefully applied in configurations matching the test cohort characteristics, following their original published implementations. NetBio operates by selecting the top 200 therapy-specific genes corresponding to the treatment regimen (‘PD1‘, ‘PDL1‘, ‘PD1_CTLA4‘, or ‘PD1_PDL1_CTLA4‘)[16], while TIDE employs distinct scoring models for melanoma and non-small cell lung cancer (NSCLC), with all other cancer types processed through a generalized ’Other’ category[10].

Evaluation metrics and reference performance. We evaluated model performance using three complementary metrics: accuracy, Matthews correlation coefficient (MCC), and precision-recall area under the curve (PR-AUC). MCC and PR-AUC were included as robust measures for class-imbalanced datasets. To properly interpret cross-cohort transfer results, we established rigorous baseline/reference performance metrics derived from the test cohort's response distribution. These baselines represent the expected performance if predictions were made knowing only the responder prevalence in the test cohort (information unavailable to the models during prediction):

$$\text{Reference Accuracy} = \left(\frac{R}{R+NR}\right)^2 + \left(\frac{NR}{R+NR}\right)^2$$

$$\text{Reference Precision} = \frac{R}{R+NR}$$

where R and NR denote the counts of responders and non-responders in the test cohort, respectively. These formulations naturally account for class imbalance, with perfectly balanced cohorts (R = NR) yielding reference values of 50% accuracy and 0.5 precision (see cohort-specific distributions in **Supplementary Figure S6**). For our 240 cohort-to-cohort transfer evaluations, we defined successful transfer as cases where model accuracy surpassed the target cohort's reference accuracy.

6 Multi-stage fine-tuning for drug- and disease-specific model development

The multi-stage fine-tuning (MSFT) approach enables COMPASS to adapt to new therapeutic contexts with limited data availability through a hierarchical training strategy. This process begins with coarse fine-tuning on large, heterogeneous ICI-treated cohorts spanning multiple cancer types, followed by precise fine-tuning on smaller drug- or disease-specific datasets **Figure 4**. The first fine-tuning stage establishes general ICI response prediction capabilities, capturing pan-cancer TIME features, while the second fine-tuning stage optimizes these features for specific drug mechanisms or clinical populations. This sequential approach enhances model transferability and robustness in data-limited settings where direct fine-tuning on small cohorts would risk overfitting.

We evaluated MSFT against two single-stage baselines (SSFT1: direct fine-tuning on drug- or indication-specific cohorts; SSFT2: fine-tuning on pan-cancer ICI cohorts) along with two reference models (PGM trained on SSFT1 data and baseline/reference performance for the test cohort as described in **Methods 5**).

For each assessment, clinical cohorts were split into two mutually exclusive groups: (1) pan-cancer

ICI cohorts for coarse fine-tuning stage 1, and (2) drug- or disease-specific cohorts. The drug- or disease-specific cohorts were further split into disjoint training and test cohorts for rigorous cross-cohort transfer assessment. See **Figure 4** and **Supplementary Table S7** for an overview of dataset configurations.

Importantly, when developing drug-specific models, drugs with the same target were excluded from the pan-cancer ICI cohorts. For example, when developing pembrolizumab-specific models, all other anti-PD1 drugs were excluded from the pan-cancer ICI dataset.

Dataset splits. To evaluate drug-specific adaptation, we applied MSFT to three immune checkpoint inhibitors: atezolizumab (anti-PD-L1), pembrolizumab (anti-PD-1), and nivolumab (anti-PD-1). For atezolizumab, the drug-specific training cohort comprised 354 patients (IMvigor210: n=298 bladder cancer; Rose: n=35; Snyder: n=21), with testing performed on 176 kidney cancer patients (IMmotion150: n=165; Miao: n=2). Pembrolizumab-specific models were trained on 120 melanoma patients (Liu: n=62; Gide: n=32; Hugo: n=26) and evaluated on 78 gastric/lung cancer cases (Kim: n=45; Ravi-1 LUAD: n=33). Nivolumab models utilized 105 melanoma patients for training (Riaz: n=51; Liu: n=45; Gide: n=9) and 63 non-melanoma patients for testing (Ravi-1 LUAD: n=49; Ravi-2 LUSC: n=14). Consistent with our exclusion criteria, the pan-cancer ICI cohorts used for initial coarse fine-tuning excluded any cohorts treated with drugs sharing the same target mechanism (e.g., all anti-PD-1 therapies were excluded for pembrolizumab/nivolumab studies).

For population-specific adaptation in lung adenocarcinoma (LUAD; **Supplementary Figure S9**), the training data consisted of 69 Ravi-1 cohort patients treated with non-pembrolizumab therapies, while testing used a held-out set of 33 pembrolizumab-treated LUAD patients from the same cohort. As with drug-specific models, the pan-cancer ICI fine-tuning stage excluded all LUAD patients to prevent data leakage.

7 SHAP concept importance analysis

We employed SHAP (SHapley Additive exPlanations) analysis[80] to quantify the contribution of each of the 44 high-level concepts to ICI response predictions. Using the Kernel SHAP implementation in the `shap` package (v0.46.0), we analyzed the partially fine-tuned (COMPASS-PFT) model after training on all available cohorts. The analysis was performed at both global (pan-cancer) and cancer-specific levels (BLCA, KIRC, SKCM, LUAD, STAD, GBM, and LUSC), as shown in **Supplementary Figure S14**.

Because our primary objective was to determine the relative importance of 44 high-level concepts for response prediction, the SHAP analysis focused on the classifier module of the COMPASS model. A key step in the SHAP workflow was the selection of a background dataset: to capture the underlying data distribution while ensuring computational feasibility, we applied K-means clustering to generate 100 centroid points from the input data. In cases where fewer than 100 samples were available, the entire dataset was used as the background. SHAP values were computed separately for responder (R) and non-responder (NR) classes, and the final ranking of the 44 concepts was derived from the mean absolute SHAP values across all patients within each dataset.

8 Patient survival analysis

We examined COMPASS prediction of long-term clinical outcomes using its learned concept representations and predicted response probabilities. Specifically, the COMPASS model functions in two capacities: (1) as a feature extractor generating gene (S_{Gene}), granular concept (S_{Geneset}), and high-level concept (S_{Concept}) features for downstream risk modeling, and (2) as a direct predictor of individual response probabilities. These approaches were evaluated on the IMvigor210 cohort, comparing COMPASS-derived features, COMPASS-predicted response probabilities, and established biomarkers (TMB, PD-L1 (IC) score, IHC immune phenotype). All survival analyses used overall survival (OS) as the endpoint, with censoring applied strictly according to the original study criteria[42]. For statistical analysis, we employed the `lifelines` package (v0.27.8)[81] to generate Kaplan-Meier survival curves, calculate log-rank test p-values, and compute hazard ratios (HRs) with 95% confidence intervals.

Survival analysis using COMPASS-derived features

For survival analysis based on COMPASS-derived features, we used the COMPASS-PFT model trained on all cohorts except the IMvigor210 cohort (using leave-one-cohort-out approach). From this model, we extracted 132 granular concept and 44 high-level concept features as inputs for ridge-regularized Cox proportional hazards models (RCOX).

The RCOX models were trained on a combined dataset of 562 patients with available survival data, excluding the IMvigor210 cohort. Feature values were standardized using z-score normalization. Implementation used the `scikit-survival` package (v0.20.0)[82], with the regularization parameter (alpha) optimized through five-fold cross-validation to maximize concordance index (C-index). Final risk scores were normalized to a 0-1 range using min-max scaling based on the training data.

For testing on the IMvigor210 cohort, feature standardization and risk score scaling were applied using the scalers fitted on the training data. Patients in the test set were stratified into high-risk and low-risk groups based on the top 10% risk score cutoff values derived from the training set. Using this procedure, the granular concept-based RCOX model classified 261 patients into the high-risk group and 37 patients into the low-risk group, while the high-level concept-based RCOX model classified 264 patients into the high-risk group and 34 into the low-risk group. Kaplan-Meier (KM) plots were generated based on these stratifications (**Supplementary Figure S17a**).

Survival analysis using COMPASS-predicted response probabilities

The COMPASS-PFT model (trained excluding IMvigor210 as described above) generated response probabilities ($P_{(R)}$), stratifying patients into responders ($P_R \geq 0.5$, $n = 42$) and non-responders ($P_R < 0.5$, $n = 256$). Kaplan-Meier analysis assessed survival differences between these groups (**Supplementary Figure S17b**).

Comparative analysis with established biomarkers

We evaluated COMPASS's performance relative to three clinical biomarkers in the IMvigor210 cohort (limited to patients with TMB data, n=234). Patients were stratified using standard cutoffs for each biomarker[42,83]:

- TMB level: High (≥ 10 mut/Mb, n=97) vs low (< 10 mut/Mb, n=137).
- PD-L1 (IC) score: IC2+ (n=90) vs IC0/1 (n=144)

- IHC immune phenotype: Inflamed (n=56) vs non-inflamed (combined desert and excluded, n=149)

9 Analysis of COMPASS concepts in the IMvigor210 cohort

To explore the contribution and biological relevance of COMPASS concepts to response prediction in the IMvigor210 cohort, we analyzed concept scores generated by the COMPASS-PFT model that was trained on all ICI cohorts except IMvigor210 (leave-one-cohort-out approach). We computed the Pearson correlation coefficient between each concept score S_{Concept} and the predicted probability of response $P_{(R|NR)}$ in the held-out IMvigor210 cohort. Concept scores were ranked based on the strength of their correlation with predicted responders (*R*) and non-responders (*NR*). The top 16 most strongly correlated scores were selected for downstream analysis. These included eight concepts positively associated with responder prediction (P_R): `Macrophage`, `IFNg pathway`, `Genome integrity`, `Cell proliferation`, `Cytotoxic Tcell`, `Tcell general`, `pDC`, and `Immune checkpoint`; and eight positively associated with non-responder prediction (P_{NR}): `NKcell`, `Exhausted Tcell`, `Bcell general`, `Plasma cell`, `Innate lymphoid cell`, `CD4 Tcell`, `TGFb pathway`, and `Endothelial` (**Supplementary Figure S18**).

Functional categorization and gene expression correlations of COMPASS concepts. To assess biological relevance of these TIME concepts, we analyzed associations between concept scores and expression levels of their constituent genes. Specifically, Pearson correlation coefficients were computed between each concept score and the TPM expression levels of its corresponding genes. Genes were then ranked from most negatively to most positively correlated, and the proportions of positive and negative correlations were summarized in **Supplementary Figure S19**. Based on their immunological roles and gene correlation patterns, the 16 concepts were grouped into four functional categories: pro-inflammatory (`Macrophage`, `IFNg pathway`, `Cytotoxic Tcell`, `Tcell general`, `pDC`, `Immune checkpoint`), TMB-associated (`Genome integrity`, `Cell proliferation`, see **Supplementary Figure S21b**), immune-exclusion (`TGFb pathway`, `Endothelial`), and immune-deficiency (`NK cell`, `Exhausted Tcell`, `B cell general`, `Plasma cell`, `Innate lymphoid cell`, `CD4 Tcell`).

Patient-specific COMPASS concept profiles and subgroup clustering. To characterize inter-patient heterogeneity, we stratified patients into four subgroups based on their COMPASS-predicted response probabilities and immune phenotypes: inflamed responders ($P_R \geq 0.5$, $n = 10$), non-inflamed responders ($P_R \geq 0.5$, $n = 13$), inflamed non-responders ($P_R < 0.5$, $n = 33$), and non-inflamed non-responders with strong non-response predictions ($P_R < 0.0001$, $n = 37$). For each group, the Z-score normalized COMPASS concept scores (top 16) were visualized in a heatmap. Patients were clustered using hierarchical clustering with cosine distance and complete linkage to reveal intra-group patterns (Inflamed R: cluster A_1, A_2, and A_3; Non-inflamed R: cluster B_1 and B_2; Inflamed NR: cluster C_1, C_2, C_3, C_4 and C_5; Non-inflamed NR: cluster D_1, D_2, D_3, and D_4). The average concept score across patient clusters is shown in **Figure 6f**. Each patient was annotated with their predicted P_R, immune phenotype (inflamed, excluded, desert), tumor mutational burden

(TMB-high vs. TMB-low, using a threshold of 10 mutations/Mb), and patient ID (**Supplementary Figure S22**).

10 Generation of personalized response maps

Personalized response maps provide a hierarchical visualization of information propagation within the COMPASS model, enabling interpretation of how molecular features contribute to response prediction. By tracing the flow of information from input gene expression (X_{GeneTPM}) through successive representational layers—gene scores (S_{Gene}), granular concept scores (S_{Geneset}), and high-level concept scores (S_{Concept})—to the final predicted probability $P_{(R|NR)}$, these maps reveal the biological reasoning underlying the model's output for individual patients (**Figure 7**, **Supplementary Figures S23-S26**).

Gene expression input layer. The first layer of personalized response maps represents the input gene expression matrix $X_{\text{GeneTPM}} \in \mathbb{R}^{B\times L}$, where B is the number of patients and L is the number of genes. All input expression values are z-score normalized within the cohort, such that the value for gene g in a given patient reflects its relative expression compared to other patients:

$$X_{\text{GeneTPM}}(g) = \frac{x_g - \mu_g}{\sigma_g}$$

where x_g is the $\log_2$(TPM + 1) value of gene g, and μ_g, σ_g are the cohort-level mean and standard deviation.

COMPASS gene score layer. Gene scores S_{Gene} are computed by treating each gene as a singleton granular concept and applying the same projection process used for curated gene sets. Specifically, the transformer-based Gene Language Model (GLM) produces contextualized gene embeddings $\mathbf{X}_{\text{GeneGLM}} \in \mathbb{R}^{B\times L\times d}$, where each gene embedding incorporates both the gene's expression and its interactions with other genes through self-attention. Each gene embedding is then mapped to a scalar gene score using a shared linear projection layer:

$$S_{\text{Gene}}(g) = \text{Linear}\left(\mathbf{X}_{\text{GeneGLM}}[g]\right) \in \mathbb{R}^{B\times 1}$$

This formulation ensures that gene scores are directly comparable to granular concept scores, as they are derived using the same scoring mechanism. The resulting S_{Gene} reflects both individual gene activity and its contextual importance.

COMPASS granular concept, high-level concept, and prediction layers. Gene scores S_{Gene} are aggregated into granular concept scores S_{Geneset} using attention-based weighting mechanisms that learn each gene's contribution to its associated gene set. Each granular concept is projected into a scalar score using a shared linear layer applied to the weighted combination of gene embeddings. The granular concept scores are further aggregated into high-level concept scores $S_{\text{Concept}} \in \mathbb{R}^{B\times 43}$, where each score represents a broader functional module (e.g., immune cell types or signaling pathways). A second-level attention step is used to learn the relevance of each granular concept to its parent high-level concept. Finally, these scores are used to compute the overall response probability $P_{(R|NR)}$, representing the likelihood that the patient will benefit from immunotherapy.

Interactive clinical exploration tool To generate personalized response maps for the IMvigor210 cohort, we employed the COMPASS-PFT model trained on all ICI cohorts except IMvigor210 (leave-one-cohort-out approach). These maps display z-score normalized features across hierarchical layers, highlighting inter-patient variation rather than absolute magnitude. Edge weights between layers represent the importance of each connection and are estimated by computing Pearson correlations between the z-scores of source and target nodes across the cohort.
An online personalized response map viewer (available at https://www.immuno-compass.com/explore/IMvigor210/) enables interactive exploration, highlighting the top 16 high-level concepts (8 associated with response and 8 with non-response, as shown in **Supplementary Figure S18**) and features exceeding user-specified thresholds (e.g., genes with $|z| > 1$ in the input layer or concepts with $|z| > 0.5$ in projection layers). Example maps in **Supplementary Figure S27** illustrate how the interactive tool enables users to zoom in on particular concepts of interest (e.g., cytotoxic T cell activation) that drive predictions in a given patient.

References

1. Gong, J., Chehrazi-Raffle, A., Reddi, S. & Salgia, R. Development of pd-1 and pd-l1 inhibitors as a form of cancer immunotherapy: a comprehensive review of registration trials and future considerations. *Journal for immunotherapy of cancer* **6**, 8 (2018).

2. Havel, J. J., Chowell, D. & Chan, T. A. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nature Reviews Cancer* **19**, 133–150 (2019).

3. Wu, B., Zhang, B., Li, B., Wu, H. & Jiang, M. Cold and hot tumors: from molecular mechanisms to targeted therapy. *Signal Transduction and Targeted Therapy* **9**, 274 (2024).

4. Wang, M. M., Coupland, S. E., Aittokallio, T. & Figueiredo, C. R. Resistance to immune checkpoint therapies by tumour-induced t-cell desertification and exclusion: key mechanisms, prognostication and new therapeutic opportunities. *British Journal of Cancer* **129**, 1212–1224 (2023).

5. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine* **24**, 541–550 (2018).

6. Topalian, S. L., Taube, J. M., Anders, R. A. & Pardoll, D. M. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nature Reviews Cancer* **16**, 275–287 (2016).

7. Goodman, A. M. *et al.* Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Molecular cancer therapeutics* **16**, 2598–2608 (2017).

8. Riaz, N. *et al.* Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell* **171**, 934–949 (2017).

9. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).

10. Jiang, P. *et al.* Signatures of t cell dysfunction and exclusion predict cancer immunotherapy response. *Nature medicine* **24**, 1550–1558 (2018).

11. Auslander, N. *et al.* Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nature medicine* **24**, 1545–1549 (2018).

12. Ayers, M. *et al.* Ifn-γ–related mrna profile predicts clinical response to pd-1 blockade. *The Journal of clinical investigation* **127**, 2930–2940 (2017).

13. Wu, C.-C., Wang, Y. A., Livingston, J. A., Zhang, J. & Futreal, P. A. Prediction of biomarkers and therapeutic combinations for anti-pd-1 immunotherapy using the global gene network association. *Nature communications* **13**, 42 (2022).

14. Cristescu, R. *et al.* Pan-tumor genomic biomarkers for pd-1 checkpoint blockade–based immunotherapy. *Science* **362**, eaar3593 (2018).

15. Hsiehchen, D. *et al.* Mutation burden and anti-pd-1 outcomes are not universally associated with immune cell infiltration or lymphoid activation. *Cancer cell* **42**, 1985–1987 (2024).

16. Kong, J. *et al.* Network-based machine learning approach to predict immunotherapy response in cancer patients. *Nature communications* **13**, 3703 (2022).

17. Wei, F. *et al.* Machine learning for prediction of immunotherapeutic outcome in non-small-cell lung cancer based on circulating cytokine signatures. *Journal for ImmunoTherapy of Cancer* **11**, e006788 (2023).

18. Balar, A. V. *et al.* Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *The Lancet* **389**, 67–76 (2017).

19. McDermott, D. F. *et al.* Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nature medicine* **24**, 749–757 (2018).

20. Ravi, A. *et al.* Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. *Nature genetics* **55**, 807–819 (2023).

21. Liu, D. *et al.* Integrative molecular and clinical modeling of clinical outcomes to pd1 blockade in patients with metastatic melanoma. *Nature medicine* **25**, 1916–1927 (2019).

22. Freeman, S. S. *et al.* Combined tumor and immune signals from genomes or transcriptomes predict outcomes of checkpoint inhibition in melanoma. *Cell Reports Medicine* **3** (2022).

23. Van Allen, E. M. *et al.* Genomic correlates of response to ctla-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).

24. Kim, S. T. *et al.* Comprehensive molecular characterization of clinical responses to pd-1 inhibition in metastatic gastric cancer. *Nature medicine* **24**, 1449–1458 (2018).

25. Gide, T. N. *et al.* Distinct immune cell populations define response to anti-pd-1 monotherapy and anti-pd-1/anti-ctla-4 combined therapy. *Cancer cell* **35**, 238–255 (2019).

26. Rose, T. L. *et al.* Fibroblast growth factor receptor 3 alterations and response to immune checkpoint inhibition in metastatic urothelial cancer: a real world experience. *British journal of cancer* **125**, 1251–1260 (2021).

27. Choueiri, T. K. *et al.* Immunomodulatory activity of nivolumab in metastatic renal cell carcinoma. *Clinical Cancer Research* **22**, 5461–5471 (2016).

28. Miao, D. *et al.* Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359**, 801–806 (2018).

29. Snyder, A. *et al.* Contribution of systemic and somatic factors to clinical response and resistance to pd-l1 blockade in urothelial cancer: an exploratory multi-omic analysis. *PLoS medicine* **14**, e1002309 (2017).

30. Zhao, J. *et al.* Immune and genomic correlates of response to anti-pd-1 immunotherapy in glioblastoma. *Nature medicine* **25**, 462–469 (2019).

31. Hugo, W. *et al.* Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell* **165**, 35–44 (2016).

32. Fehrenbacher, L. *et al.* Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (poplar): a multicentre, open-label, phase 2 randomised controlled trial. *The Lancet* **387**, 1837–1846 (2016).

33. Roh, W. *et al.* Integrated molecular analysis of tumor biopsies on sequential ctla-4 and pd-1 blockade reveals markers of response and resistance. *Science translational medicine* **9**, eaah3560 (2017).

34. Bergstrom, E. N. & Alexandrov, L. B. Enhanced precision in immunotherapy. *Nature Cancer* 1–3 (2024).

35. Chowell, D. *et al.* Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nature biotechnology* **40**, 499–506 (2022).

36. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

37. Lathia, J. D., Mack, S. C., Mulkearns-Hubert, E. E., Valentim, C. L. & Rich, J. N. Cancer stem cells in glioblastoma. *Genes & development* **29**, 1203–1217 (2015).

38. Tanaka, M. *et al.* Mesothelial cells create a novel tissue niche that facilitates gastric cancer invasion. *Cancer Research* **77**, 684–695 (2017).

39. Bosisio, F. M. *et al.* Plasma cells in primary melanoma. prognostic significance and possible role of iga. *Modern Pathology* **29**, 347–358 (2016).

40. Gu, W., Zhuang, W., Zhuang, M., He, M. & Li, Z. Dna damage response and repair gene mutations are associated with tumor mutational burden and outcomes to platinum-based chemotherapy/immunotherapy in advanced nsclc patients. *Diagnostic Pathology* **18**, 119 (2023).

41. Rosenberg, J. E. *et al.* Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *The Lancet* **387**, 1909–1920 (2016).

42. Mariathasan, S. *et al.* Tgfβ attenuates tumour response to pd-l1 blockade by contributing to exclusion of t cells. *Nature* **554**, 544–548 (2018).

43. Hegde, P. S. & Chen, D. S. Top 10 challenges in cancer immunotherapy. *Immunity* **52**, 17–35 (2020).

44. Roller, A. *et al.* Tumor-agnostic transcriptome-based classifier identifies spatial infiltration patterns of cd8+ t cells in the tumor microenvironment and predicts clinical outcome in early-phase and late-phase clinical trials. *Journal for Immunotherapy of Cancer* **12**, e008185 (2024).

45. Montauti, E., Oh, D. Y. & Fong, L. Cd4+ t cells in antitumor immunity. *Trends in cancer* (2024).

46. Griss, J. *et al.* B cells sustain inflammation and predict response to immune checkpoint blockade in human melanoma. *Nature communications* **10**, 4186 (2019).

47. Pagliarulo, F. *et al.* Molecular, immunological, and clinical features associated with lymphoid neogenesis in muscle invasive bladder cancer. *Frontiers in immunology* **12**, 793992 (2022).

48. Vanhersecke, L. *et al.* Mature tertiary lymphoid structures predict immune checkpoint inhibitor efficacy in solid tumors independently of pd-l1 expression. *Nature Cancer* **2**, 794–802 (2021).

49. Shou, M., Zhou, H. & Ma, L. New advances in cancer therapy targeting tgf-β signaling pathways. *Molecular Therapy-Oncolytics* **31** (2023).

50. Kuo, H.-Y., Khan, K. A. & Kerbel, R. S. Antiangiogenic–immune-checkpoint inhibitor combinations: Lessons from phase iii clinical trials. *Nature Reviews Clinical Oncology* **21**, 468–482 (2024).

51. Zhang, L. *et al.* Synergistic induction of tertiary lymphoid structures by chemoimmunotherapy in bladder cancer. *British Journal of Cancer* **130**, 1221–1231 (2024).

52. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

53. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nature Methods* 1–11 (2024).

54. Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods* 1–11 (2024).

55. Li, M. M. *et al.* Contextual ai models for single-cell protein biology. *Nature Methods* **21**, 1546–1557 (2024).

56. Wang, C. *et al.* scgpt-spatial: Continual pretraining of single-cell foundation model for spatial transcriptomics. *bioRxiv* 2025–02 (2025).

57. Becker, T. *et al.* An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Annals of Oncology* **31**, 1561–1568 (2020).

58. Chang, T.-G. *et al.* Loris robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features. *Nature Cancer* 1–18 (2024).

59. Vanguri, R. S. *et al.* Multimodal integration of radiology, pathology and genomics for prediction of response to pd-(l) 1 blockade in patients with non-small cell lung cancer. *Nature cancer* **3**, 1151–1164 (2022).

60. Yoo, S.-K. *et al.* Prediction of checkpoint inhibitor immunotherapy efficacy for cancer using routine blood tests and clinical data. *Nature Medicine* 1–12 (2025).

61. Liu, J. *et al.* An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).

62. Eddy, J. A. *et al.* Cri iatlas: an interactive portal for immuno-oncology research. *F1000Research* **9** (2020).

63. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).

64. Huang, A. C. *et al.* A single dose of neoadjuvant pd-1 blockade predicts clinical outcomes in resectable melanoma. *Nature medicine* **25**, 454–461 (2019).

65. Joyce, J. A. & Fearon, D. T. T cell exclusion, immune privilege, and the tumor microenvironment. *Science* **348**, 74–80 (2015).

66. Giordano, M. *et al.* Molecular profiling of cd 8 t cells in autochthonous melanoma identifies maf as driver of exhaustion. *The EMBO journal* **34**, 2042–2058 (2015).

67. Messina, J. L. *et al.* 12-chemokine gene signature identifies lymph node-like structures in melanoma: potential for patient selection for immunotherapy? *Scientific reports* **2**, 765 (2012).

68. Nurmik, M., Ullmann, P., Rodriguez, F., Haan, S. & Letellier, E. In search of definitions: Cancer-associated fibroblasts and their markers. *International journal of cancer* **146**, 895–905 (2020).

69. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).

70. Chen, P.-L. *et al.* Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer discovery* **6**, 827–837 (2016).

71. Colaprico, A. *et al.* Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research* **44**, e71–e71 (2016).

72. Dobin, A. *et al.* Star: ultrafast universal rna-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

73. Yang, F. *et al.* scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).

74. Szałata, A. *et al.* Transformers in single-cell omics: a review and new perspectives. *Nature Methods* **21**, 1430–1443 (2024).

75. Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* **34**, 18932–18943 (2021).

76. Choromanski, K. *et al.* Rethinking attention with performers. *arXiv preprint arXiv:2009.14794* (2020).

77. Dao, T., Fu, D., Ermon, S., Rudra, A. & Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems* **35**, 16344–16359 (2022).

78. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems* **30** (2017).

79. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

80. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017).

81. Davidson-Pilon, C. lifelines: survival analysis in python. *Journal of Open Source Software* **4**, 1317 (2019).

82. Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research* **21**, 1–6 (2020).

83. Ferreiro-Pantín, M. *et al.* Clinical, molecular, and immune correlates of the immunotherapy response score in patients with advanced urothelial carcinoma under atezolizumab monotherapy: analysis of the phase ii imvigor210 trial. *ESMO open* **8**, 101611 (2023).