

Five latent factors underlie response to immunotherapy

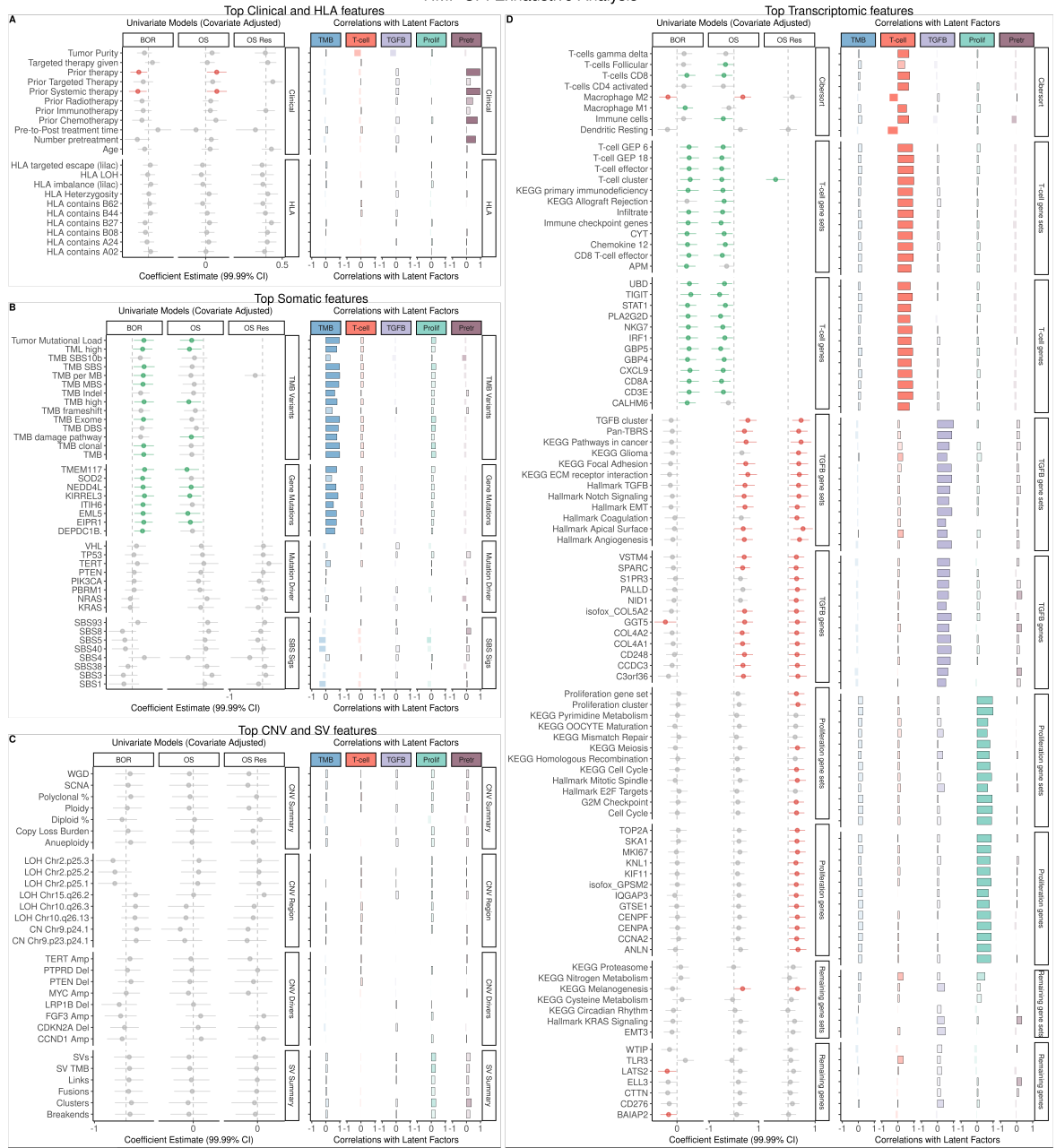
In the format provided by the
authors and unedited

Index

Supplementary Figures.....	2
Supplementary Note 1.....	11
0. Naming convention.....	11
1. Extraction of features from HMF-CPI cohort.....	11
2. Systematic identification of CPI biomarkers.....	14
2.1. Goal of Systematic Analysis and Main steps.....	14
2.2. Pre-processing.....	14
2.3. Testing Procedure for each feature.....	16
2.4. Tissue-specific regressions.....	17
2.5. Multiple testing threshold.....	17
2.6. Survival Residuals analysis.....	19
2.7. Systematic Analysis Summary.....	20
3. Discovery of the five latent factors.....	21
3.1. Cluster R1.....	21
3.1.1. Somatic Gene Mutations Analysis.....	21
3.2. Cluster R2.....	23
3.3. Cluster R3.....	23
3.4. Cluster S1.....	24
3.4. Cluster S2.1.....	24
3.6. Interchangeability of proxy measurements of the latent factors.....	25
3.7 Latent factors in non-CPI therapies.....	26
3.8 Conservation of the expression of genes in transcriptomics latent factors.....	27
3.8 Significant features outside clusters.....	27
4. Validation.....	28
5. Multivariate machine learning models of CPI response and survival.....	31
5.1. Model fitting.....	31
5.2.Evaluation of models with repeated re-sampling and cross-validation.....	31
5.3. Broad overview of HMF XGBoost studies.....	32
5.4. Study 1: Hybrid models.....	32
5.5. Study 2: Hyperparameter Settings Study.....	33
5.6. Study 3: Main Study comparing various models with different features.....	34
5.7. Relationship between predicted response and predicted hazard across patients in HMF-CPI.....	35
References.....	37

Supplementary Figures

HMFI CPI Exhaustive Analysis



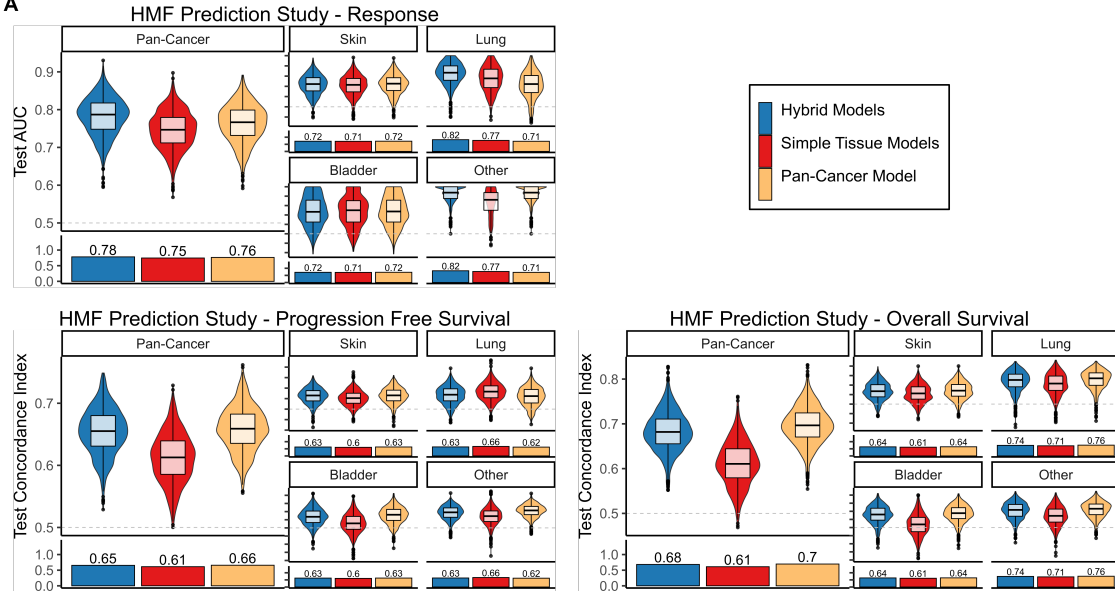
transcriptomic features in their associations with CPI response, overall survival and overall survival residuals, as well as their correlations to the five latent factors.

These data was derived from the unbiased analysis of the HMF-CPI cohort (N=479) described in the main text.

In the forest plots, the dots represent the strength (coefficients estimated through multivariate logistic or Cox regression) of the association between the latent factor and response or survival across cohorts. The horizontal bars across dots denote the 95% confidence intervals.

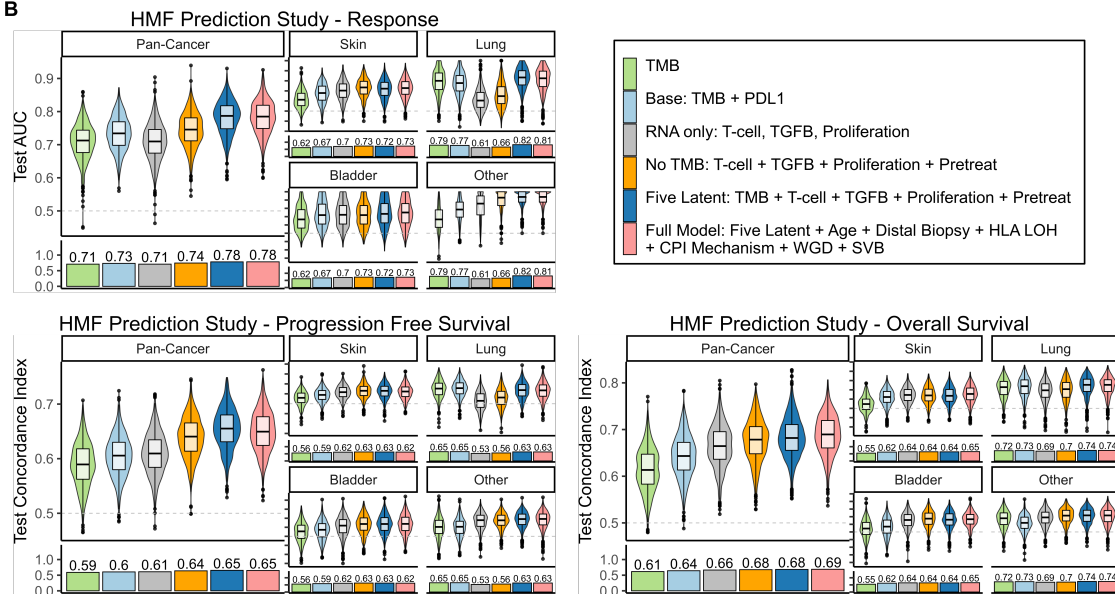
Modelling Strategies: Comparison of performance

A



Hybrid Models: Evaluation over different features

B



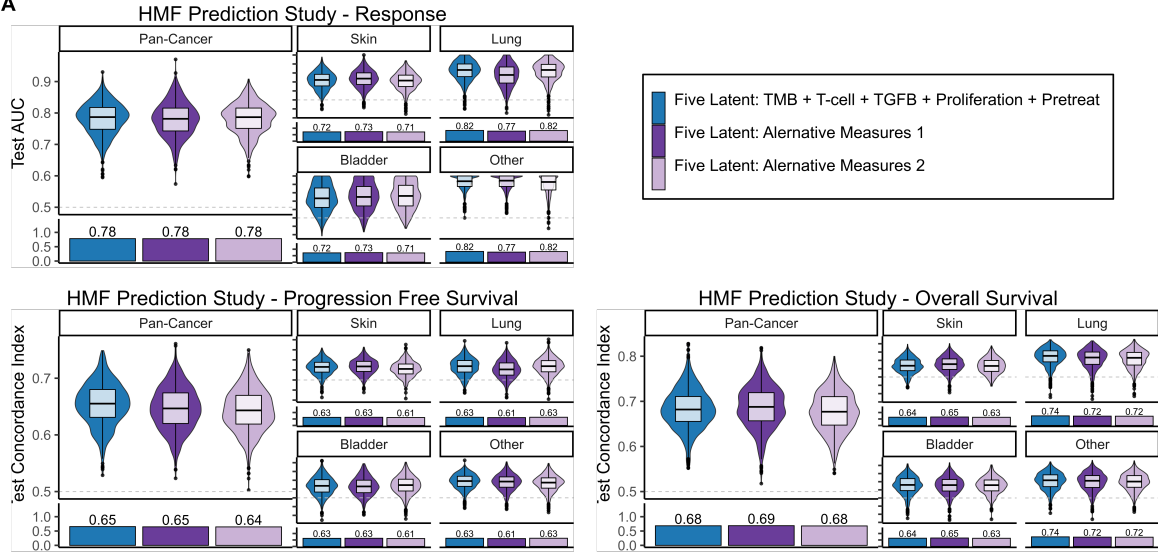
at specific groups of tumors, which highlights the importance of training from a cohort as representative as possible of different cases.

B) Comparison of hybrid multivariate models trained with five different sets of biomarkers. The five latent factors were incorporated into the training of these models, together with regularly used biomarkers (PDL1 expression), covariables of the regression analyses (age, biopsy site), or features that appeared as non-significant (HLA LoH, etc). The multivariate model incorporating the five latent factors outperforms other partial models. Moreover, the model incorporating non-significant features and covariables of the analysis shows the same performance as the one trained only on the five latent factors. This result supports the importance of the five latent factors as independent predictors of CPI response and survival and the notion that no other feature significantly contributes to this prediction in the HMF-CPI cohort.

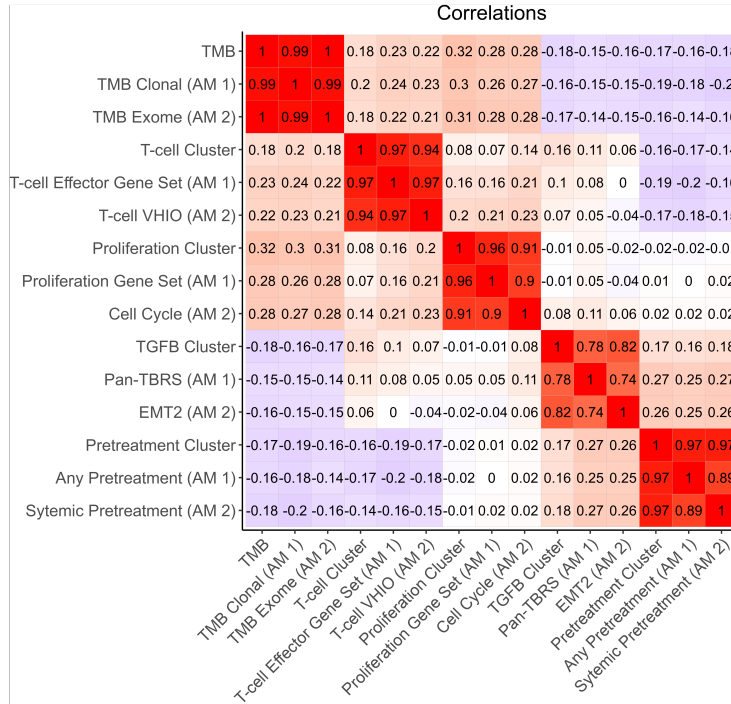
In the boxplots represented within the violin plots, the center line represents the median of the distribution, the box limits, the first and third quartiles, the whiskers, the lowest/highest data points at first quartile minus/plus $1.5 \times$ IQR, and the points, the outliers.

Evaluation of 3 models with different representations of 5 latent factors

A



B



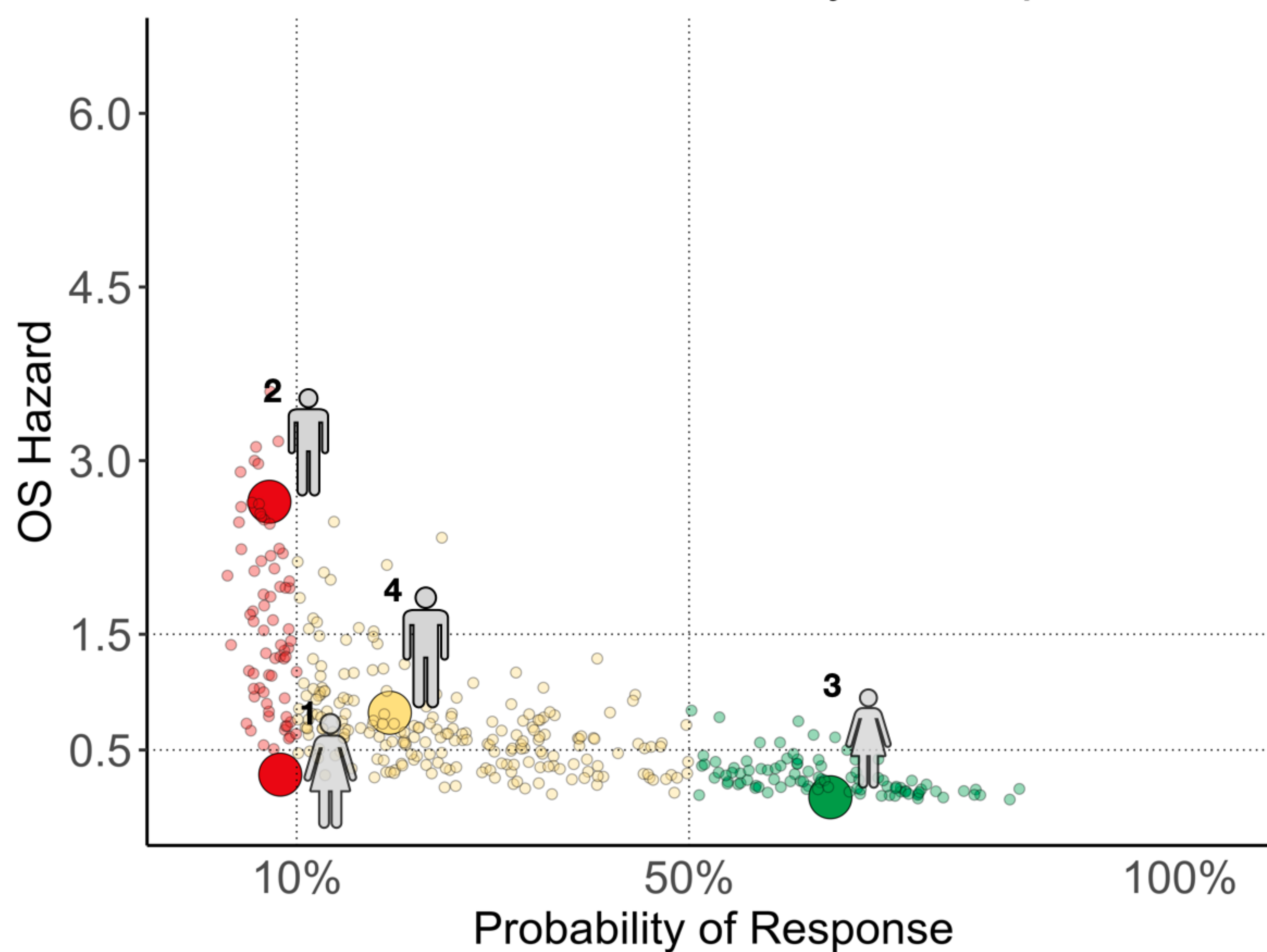
B) Correlation between the mean of the clusters representing the latent factors (as described in the manuscript) and two alternative measures of each latent factor.

Alternative measures of the latent factors. TMB: TMB, clonal TMB and exonic TMB; effective T-cell infiltration: mean of the cluster of T-cell-related features, mean of a geneset reported in the literature to represent T-cell infiltration, mean of a set of genes selected from the VHIO cohort as the most highly correlated with the mean of the cluster (see main manuscript Methods); tumor proliferative potential: mean of the cluster of proliferation related genes, mean of a geneset tracking proliferation, mean of a cell cycle geneset; TGF-beta activity in the microenvironment: mean of TGF-beta related features, mean of a geneset tracking the activity of TGF-beta on tumor-associated fibroblasts, mean of a geneset related with epithelial-to-mesenchymal transition; pretreatment: mean of the cluster of all pretreatment features, any type of pretreatment, any type of systemic pretreatment.

In the boxplots represented within the violin plots, the center line represents the median of the distribution, the box limits, the first and third quartiles, the whiskers, the lowest/highest data points at first quartile minus/plus $1.5 \times$ IQR, and the points, the outliers.

A

OS Hazard vs Probability of Response

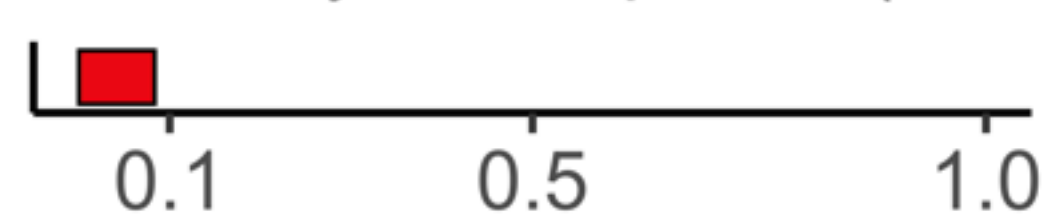


1

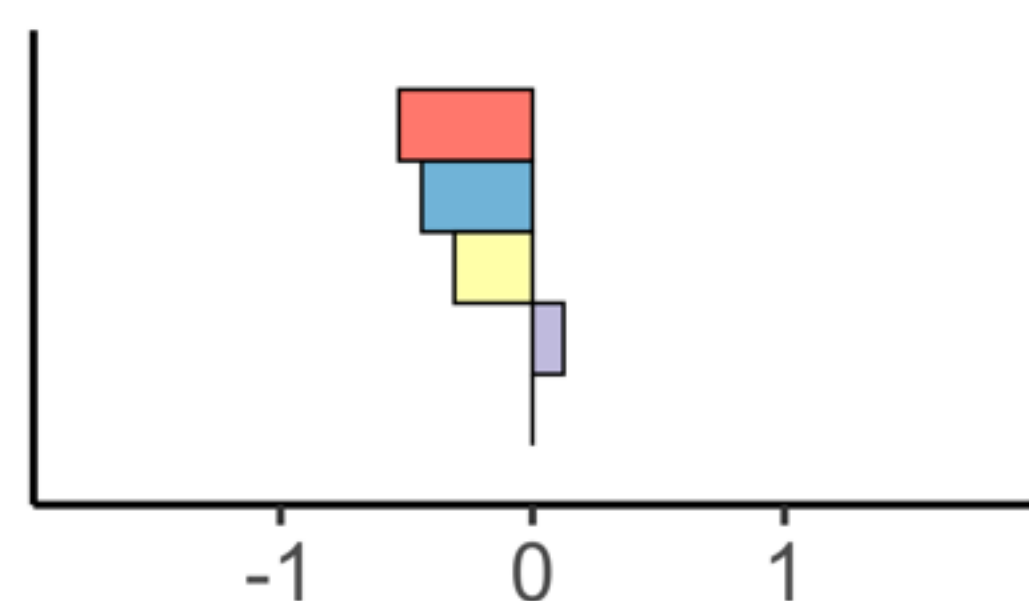


Tissue: Skin
Age: 73
Gender : Female

Probability of Response (0.08)

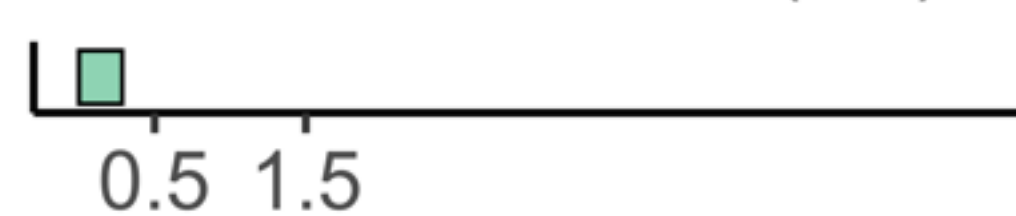


BOR Shapley Values

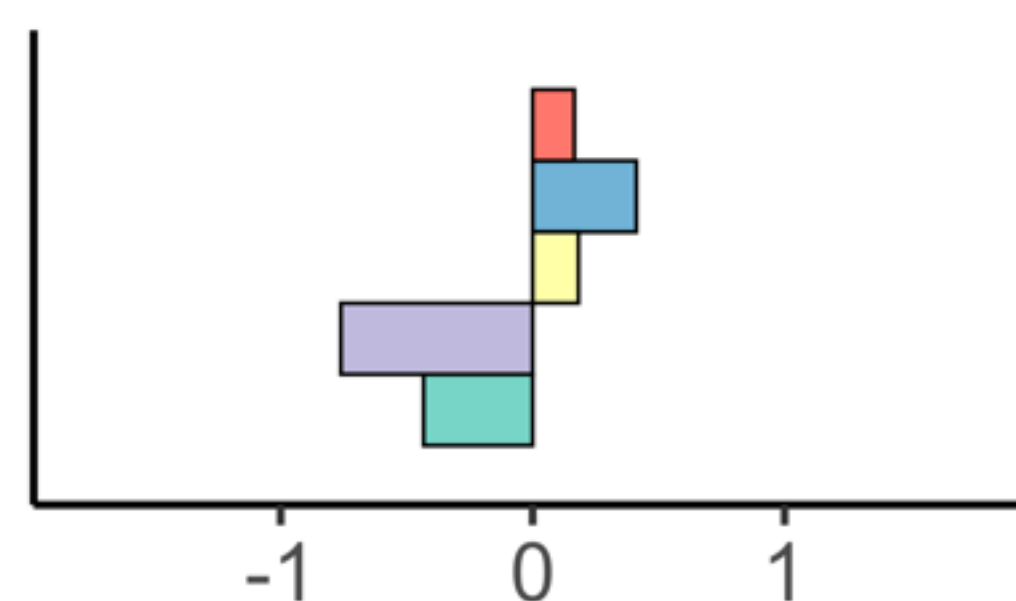


Log Odds of Response

Predicted Hazard (0.3)

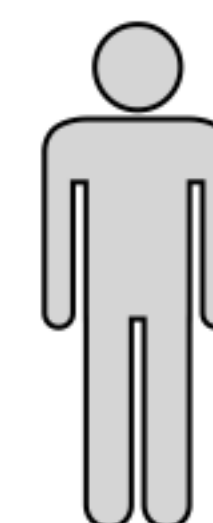


OS Shapley Values



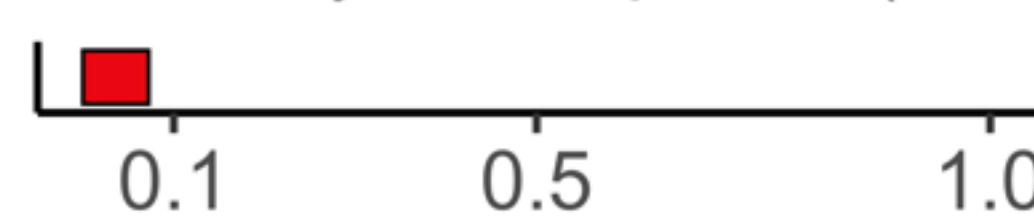
Log Hazard

2

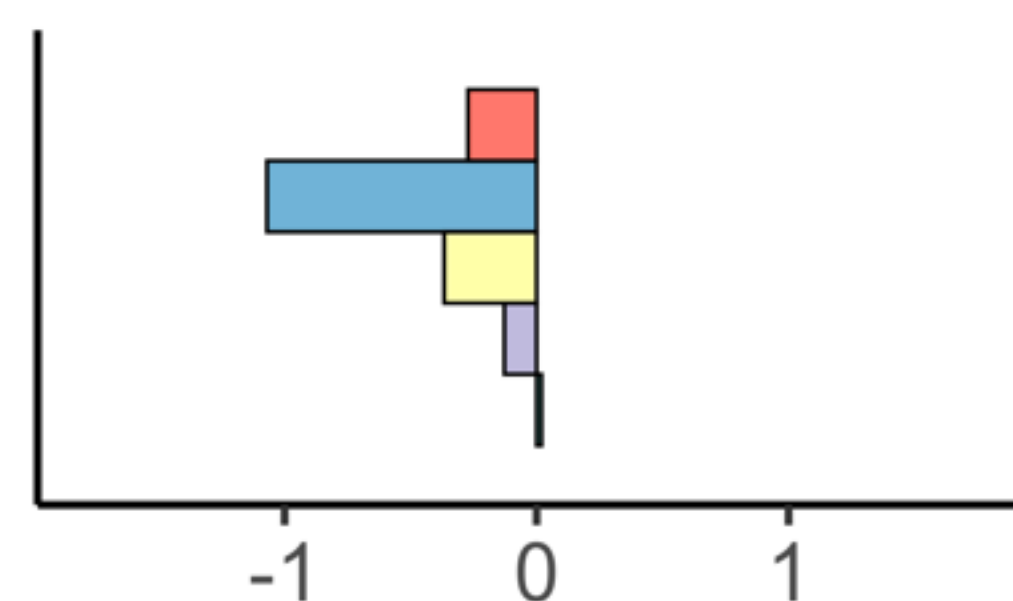


Tissue: Bladder
Age: 74
Gender : Male

Probability of Response (0.07)

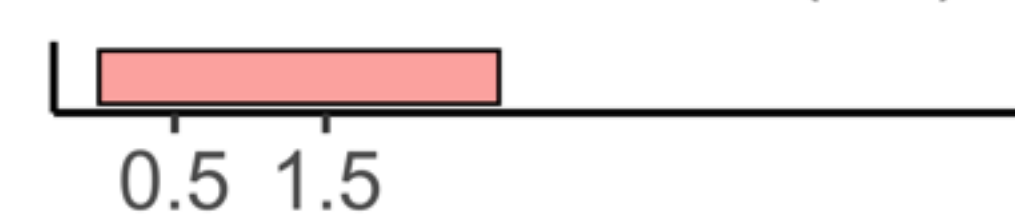


BOR Shapley Values

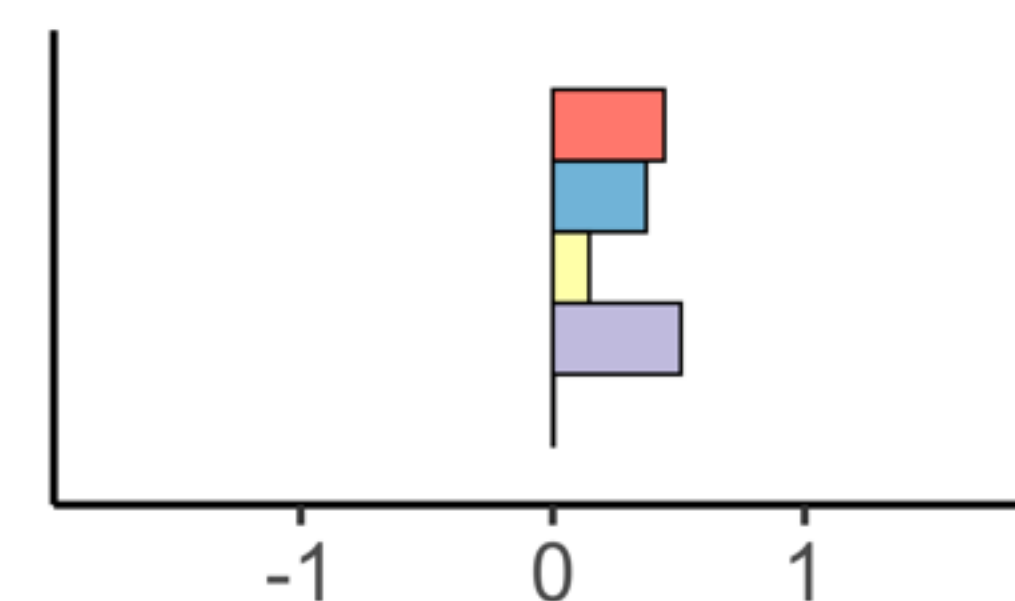


Log Odds of Response

Predicted Hazard (2.6)



OS Shapley Values



Log Hazard

3

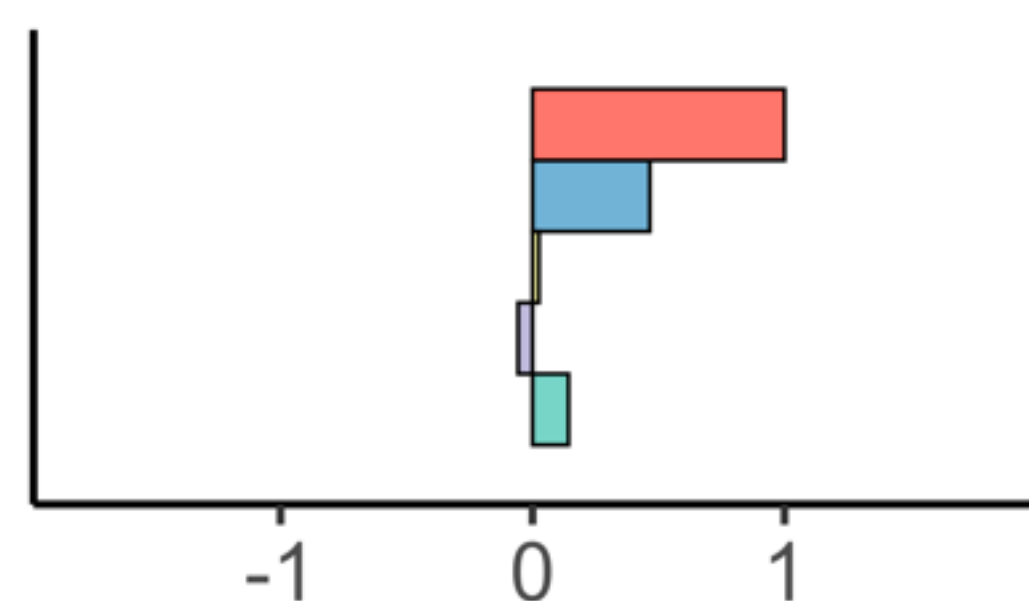


Tissue: Head, Neck
Age: 52
Gender : Female

Probability of Response (0.64)

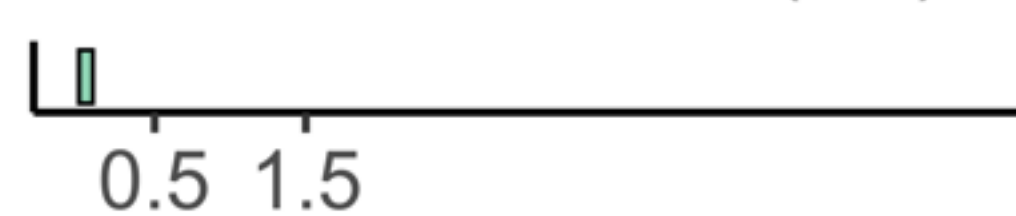


BOR Shapley Values

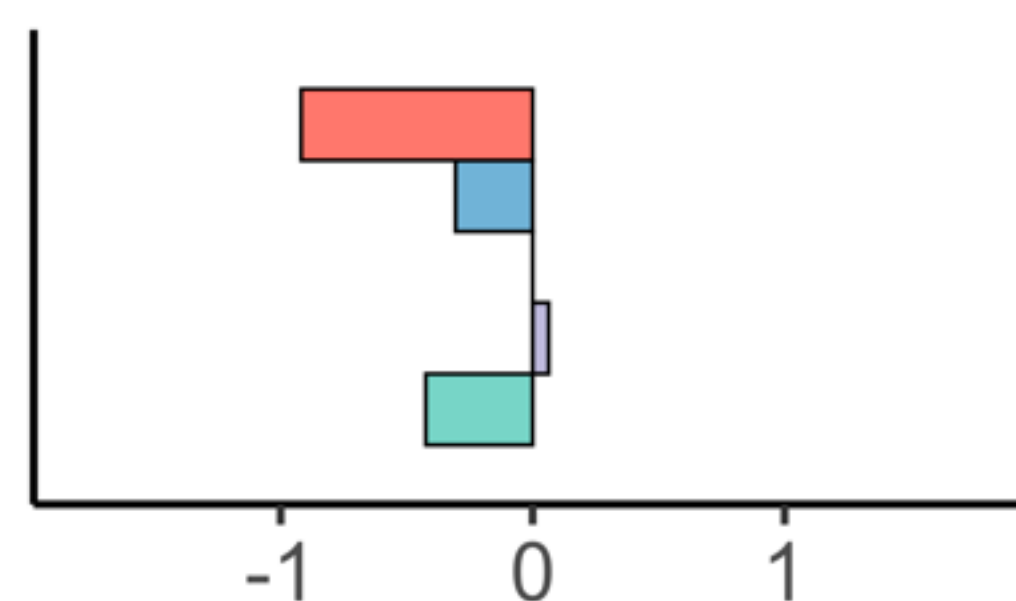


Log Odds of Response

Predicted Hazard (0.1)

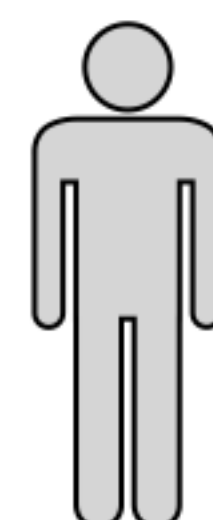


OS Shapley Values



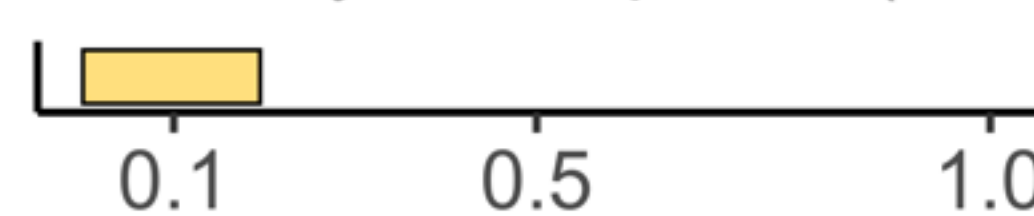
Log Hazard

4

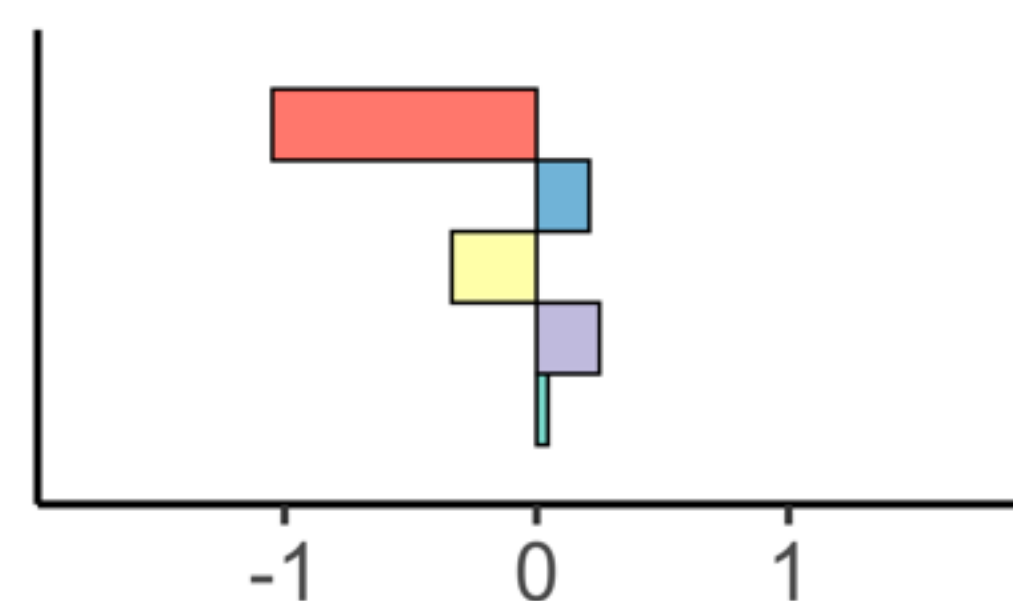


Tissue: Lung
Age: 70
Gender : Male

Probability of Response (0.19)

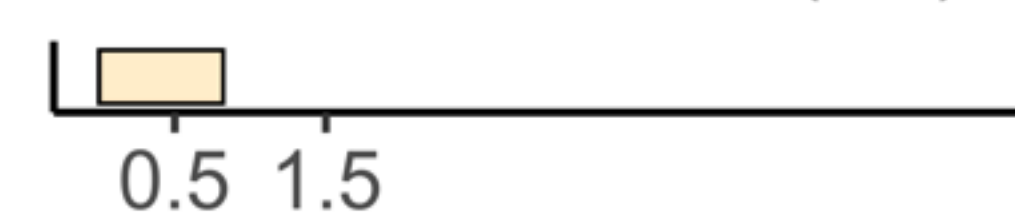


BOR Shapley Values

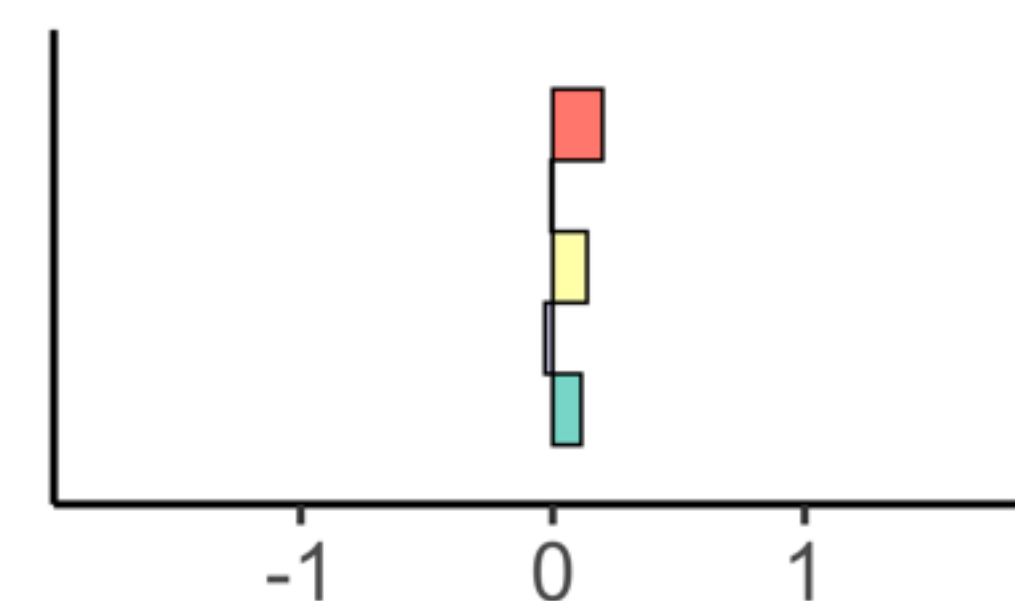


Log Odds of Response

Predicted Hazard (0.8)



OS Shapley Values



Log Hazard

22A) Same as Extended Data Figure 8, with dots representing all patients in the HMF-CPI cohort following the probability of response (x-axis) and the hazard ratios (y-axis) cast by the multivariate models of best overall response and overall survival, respectively.

All other panels present examples of the contribution of the five latent factors to the predicted probability of response and hazard ratios for patients in different locations of the plot.

Supplementary Note 1

0. Naming convention

For the sake of clarity, Tables and Figures that pertain only to this Supplementary Note are included within the text and named Table SNX and Figure SNX, to distinguish them from Extended Data Figures and Supplementary Figures mentioned in the main manuscript.

1. Extraction of features from HMF-CPI cohort

Features for the systematic analysis were extracted from the Hartwig Medical Foundation (HMF) dataset¹ (Supplementary Table 1). Some features, calculated by the HMF pipeline, were extracted directly from the dataset, and others were derived from these. Documentation on the computational pipeline to extract, compute, and combine all the features can be found in the publicly available repository (https://bitbucket.org/bbglab/hartwig_biomarkers/). While the repository provides documentation, here we share a brief overview of the pipeline.

Clinical features and outcomes were computed based on HMF metadata files, specifically, RECIST² longitudinal evaluations for each patient, and patient treatment history. Clinical outcomes were computed based on the survival fields in the metadata and longitudinal RECIST evaluations for each patient. Best overall response was computed based on the best RECIST evaluation per patient (best overall survival). Progression free survival was computed based on the start of treatment, longitudinal RECIST evaluations, as well as survival status and time of death or censoring. Overall survival was computed from the metadata based on start of treatment, survival status, and time to death or censoring. Demographic and clinical features including age, gender, tumor location, biopsy location, prior radiotherapy, and prior systemic therapy were extracted directly from HMF metadata files. The biopsy site feature was transformed into a numeric indicator of distal or proximal biopsy location. More granular pre-treatment and post-treatment features were computed based on the provided treatment history files.

HLA features were extracted from both germline and somatic copy number (HLA LoH) files. From the germline data, the HLA allotypes were derived for each patient. The HLA allotypes were used to detect HLA heterozygosity and to derive HLA supertypes. The HLA supertype features were transformed into indicator variables of presence or absence for each sample. Loss of heterozygosity at the HLA allele regions was procured from the copy number region data. The LILAC tool was used to further estimate somatic loss of HLA heterozygosity³.

Somatic features were computed in several ways, using data produced by purple tools in the Hartwig somatic pipeline (somatic pipeline: <https://github.com/hartwigmedical/pipeline5>, purple tools: <https://github.com/hartwigmedical/hmftools/tree/master/purple>). TMB variants including overall TMB, MSI status, and MSI indels burden were provided directly in summary files. Other TMB variants were computed by parsing the provided VCF files. These derived features

included TMB for clonal and subclonal mutations, exome TMB, single base substitutions burden, double base substitutions burden, multiple base substitutions burden, frameshift burden, and total indel burden. Similarly, the somatic mutant status of each gene across samples was determined by parsing the variant calling format (VCF) files provided by HMF. All mutational features including TMB variants and mutations per gene were log-transformed prior to analyses. Somatic driver features were taken directly from the driver catalogs output by the Hartwig driver catalog located with purple tools (driver catalog: <https://github.com/hartwigmedical/hmftools/blob/master/purple/DriverCatalog.md>).

Mutational signatures were calculated in two steps. First, per sample counts of single base substitutions and their tri-nucleotide contexts were calculated from the VCF files. Next, from the computed counts, mutational signatures were estimated in R using the deconstructSigs package⁴ and reference data frame from COSMIC⁵ version 3.2.

Copy number variant features were also procured from HMF purple pipeline output (purple tools: <https://github.com/hartwigmedical/hmftools/tree/master/purple>). Whole genome doubling, mean ploidy, and diploid proportion were provided directly in summary files. Copy number and loss of heterozygosity per genomic region were computed from provided copy number files, and overall somatic copy number alterations and copy loss burden were also calculated.

Structural variants were obtained directly from the output by the HMF linx tool (linx tool: <https://github.com/hartwigmedical/hmftools/tree/master/linx>).

RNA features were derived from output by the HMF isofox transcriptome pipeline (isofox tool: <https://github.com/hartwigmedical/hmftools/tree/master/isofox>). For each gene, the adjusted transcript per million (TPM) values were selected. All gene-level adjusted TPM values were log-transformed prior to analyses. Gene set features later used to interpret the clusters of features associated with CPI response and survival were calculated from the average log adjusted TPM values over all genes for each respective gene set (Supplementary Table 2). The relative proportion of immune populations computed via Cibersort were also procured from the adjusted TPM values per gene⁶.

Table SN1, below provides an overview of the feature extraction pipeline, with links to HMF tools used in the extraction.

Table SN1. HMF-CPI feature extraction process

Hartwig Source	File source	Feature Type	Example Features	Derived
Clinical data files	treatment_responses.tsv	Clinical outcomes	best overall response, progression free survival, overall survival	Yes
	metadata.tsv	Clinical Features	age, gender, primary tumor location, biopsy location	No
	*biopsy_drugs.tsv	Clinical Features	prior chemotherapy, prior targeted therapy, post chemotherapy	Yes

somatic.purple (version 3.4.2) https://github.com/hartwigmedical/pipeline5	{sample}.purity*	Somatic Summaries	TMB overall, MSI status, MSI indels whole genome doubling, mean ploidy Overall structural variant TMB	No
		TMB Variants	TMB overall, MSI status, MSI Indels	No
	{sample}.somatic.vcf*	TMB Variants	TMB clonal, TMB sub-clonal, TMB frameshift, TMB indels, TMB sbs	Yes
		Somatic gene mutations	TP53 mutations, MSH6 mutations	Yes
		Mutational Signatures	SBS1 exposure	Yes
	{sample}.cnv.gene*	CNV granular	CNV region, LOH region, copy loss burden	Yes
	{sample}.cnv.somatic	CNV summary	Somatic copy number alterations (SCNA)	Yes
	{sample}.driver.catalog*	Somatic Drivers	TERT somatic driver	No
		Copy Number Drivers	PTEN deletion driver MYC amplification driver	No
somatic.linx (version=1.19) https://github.com/hartwigmedical/hmftools/tree/master/linx	{sample}.breakend* {sample}.clusters* {sample}.svs* {sample}.fusion* {sample}.links*	Structural Variants	SV counts, Fusion counts, Breakends counts	No
Isofox https://github.com/hartwigmedical/hmftools/tree/master/isofox	{sample}.isf.gene_data*	Gene Expression	CXCL9 adjusted TPM	No
		Gene Sets	T-cell effect Gene Set	Yes
		Cibersort	macrophages M1 exposure	Yes
	{sample}.isf.neoepitope*	Neoepitopes	Number of predicted neoepitopes	No
germline	{sample}.vcf*	HLA supertypes	HLA contains B07	Yes
		HLA loss of heterozygosity	HLA LOH, HLA mutations	Yes
		Lilac Tool	Lilac HLA LOH	Yes

2. Systematic identification of CPI biomarkers

2.1. Goal of Systematic Analysis and Main steps

The overarching goal of the systematic analysis of features extracted in the HMF-CPI dataset was to identify in an unbiased way –i.e., without any prior information– the primary, independent underlying molecular and clinical aspects of the tumor and the host that explain CPI response and survival - which we refer to as latent factors. To meet this goal, the analysis we developed took three primary steps. First, we exhaustively tested all features extracted from the HMF-CPI dataset (see previous section) for association with CPI response and survival, and identified a few thousand significant features. Second, we used correlation-based analysis to cluster the significant features into 3 latent factors of CPI response. In a survival residuals analysis, i.e., modeling the survival outcome after adjusting for these latent factors of CPI response done in the same way and using the same pairwise correlation-based procedure to cluster them, we identified two further latent factors.

2.2. Pre-processing

The genomics, transcriptomics and clinical features were first pre-processed to prepare the exhaustive analysis. This initial pre-processing involved four main steps: 1) numerization, 2) transformation, 3) pre-filtering, and 4) feature standardization. The pre-processing aimed to curate and test features for interpretability, stable model fitting, adequate statistical power to identify associations (i.e. enough non-zero samples) and suitability for correlation based analysis (i.e. numeric features).

Numerization. All features were required to be numeric. This rule allowed for common comparisons of effect sizes across features and enabled the systematic correlation based analysis (e.g. Fig. 2). For clinical features, all pre-treatment and post-treatment information was coded into 0 or 1 indicator variables (e.g. “exposure to prior chemotherapy”, or “post-treatment contains targeted therapy”). Similarly, biopsy location was transformed into distal vs proximal and coded with an indicator variable. For the HLA features, HLA supertypes were coded as presence or absence indicator variables (e.g. contain B07 HLA supertype). All other binary character variables in the analysis (e.g. gender, HLA LOH, TMB high, MSI status) were encoded as numeric. The only non-numeric features in the main study were primary tissue and biopsy site (non-transformed to an indicator variable), both of which were used as categorical covariates in the exhaustive testing.

Transformation. Features were transformed to enhance model fitting and increase power for many subsequent association tests. For example, features within groups with a tendency towards highly right skewed values (e.g. somatic mutation, TPM values) were all log-transformed. The chosen transformations by feature groups are summarized in Table SN2, below.

Pre-filtering was performed to reduce multiple testing in the analysis. Features with possible unstable model fitting or low power (e.g. only few non-zero values) or low practicality (e.g. extremely low TPM expression values) were filtered away using pre-set criteria. The overwhelming majority of features that were pre-filtered out of the analysis were somatic gene mutation and RNA seq gene expression features. All pre-filtering rules and details are summarized below in Table SN2.

Table SN2: Transformation and Pre-filtering of input features

Feature group	Feature sub-group	Transformations	Filtering rules
Clinical	Clinical	Factors set to numeric	All kept
HLA	HLA	Factors set to numeric	All kept
Somatic	TMB variants	$\ln(x + 1)$	All kept
	Somatic gene mutations	$\ln(x + 1)$	At least 1 mutation in at least 1 of 20 patients
	Mutational signatures	none	At least .02 exposure in at least 1 in 20 patients
	CNV and SV summaries	none	All kept
	CNV region features	none	All kept
Transcriptome	Transcript expression	$\ln(x + 1)$	1. must be protein coding genes 2. log-transformed TPM have mean and standard deviation > .5
	Gene Sets	Average of transformed gene TPMs	All kept
	Cibersort	none	All kept
Predicted Drivers	Somatic	none	At least 1 predicted driver event in at least 1 of 30 patients
	Deletion and amplifications	none	

Standardization was the final pre-processing step. All features were transformed to have mean 0 and standard deviation 1. The feature standardization was done to facilitate common comparison of effect sizes in terms of their respective standard deviation units (e.g Fig. 2A). The standard deviations and means used for the standardization were computed across the 479 HMF-CPI patients.

After the pre-processing, 27,000 standard normalized features were kept for systematic analyses for their association with clinical outcomes.

In total, the HMF-CPI analyses included 479 patients that received PD1/PDL1 CPI therapies. However, only a subset of 396 had transcriptomic data available. Sample sizes for the HMF data are available as part of Supplementary Table 1.

2.3. Testing Procedure for each feature

Each feature was tested for association with 3 outcomes –response, progression free survival and overall survival. For all tests, fitting leveraged all 479 pan-cancer samples, with adjustments for primary tissue, biopsy location, age, and tumor purity as covariates in the main analysis. For all tests, generalized linear models and their native maximum likelihood based methods were used for parameter estimation, covariate adjustment, and p-value computation.

Response was modeled with logistic regression. For each feature, we assumed the number of responses to CPI treatment followed a binomial distribution with a probability of response p , and thus full and reduced models were fitted:

$$\text{Full Model: } \text{Logit}(p) = \beta_0 + X \beta_X + \sum_{j \in \text{tissue}} I_j \beta_j + \sum_{k \in \text{biopsy}} I_k \beta_k + X_{\text{purity}} \beta_{\text{purity}} + X_{\text{age}} \beta_{\text{age}}$$

$$\text{Reduced Model: } \text{Logit}(p) = \beta_0 + \sum_{j \in \text{tissue}} I_j \beta_j + \sum_{k \in \text{biopsy}} I_k \beta_k + X_{\text{purity}} \beta_{\text{purity}} + X_{\text{age}} \beta_{\text{age}}$$

Logistic regression models were fitted using the base R glm function with the family parameter set to “binomial”.

Progression-free survival and overall survival were modeled for each feature with Cox proportional hazards models. For each feature and survival outcome, the assumed proportional hazard rates, denoted $h(t)$, were modeled:

$$\text{Full Model: } \text{Log}(h(t)) = X \beta_X + \sum_{j \in \text{tissue}} I_j \beta_j + \sum_{k \in \text{biopsy}} I_k \beta_k + X_{\text{purity}} \beta_{\text{purity}} + X_{\text{age}} \beta_{\text{age}}$$

$$\text{Reduced Model: } \text{Log}(h(t)) = \sum_{j \in \text{tissue}} I_j \beta_j + \sum_{k \in \text{biopsy}} I_k \beta_k + X_{\text{purity}} \beta_{\text{purity}} + X_{\text{age}} \beta_{\text{age}}$$

In these equations, the β coefficients represent the effect of the feature X or of the covariates on the outcome measured. Since tissue and biopsy sites are categorical features with multiple levels, their respective terms are represented as a summation of indicator variables corresponding to the levels.

Survival models were fitted with the coxph function with the survival package in R.

For each model, p-values were calculated from likelihood ratio tests comparing the specified full model to the reduced models. Effect sizes were estimated with maximum likelihood on the full models. All effect sizes, standard errors, and corresponding p-values were stored for further analysis. Stored results are listed in Supplementary Dataset 1. Furthermore, all code documentation is accessible in the public repository.

In the discovery of latent factors associated with survival residuals, the same approach as the one described above for overall survival was used, but including among the covariates the three previously identified latent factors.

2.4. Tissue-specific regressions

While we chose to focus the presentation in the main manuscript on pan-cancer regression analyses adjusted for tissue of the primary tumor, we have also computed tissue specific regressions. For all four primary tissue groups –Skin, Lung, Bladder, Other– we exhaustively fit univariate models and stored their output. We focused the paper on the adjusted pan-cancer models for several reasons. First, tissue specific models leverage less data and therefore have smaller statistical power. Second, the statistical power problem is compounded as more models amplify multiple testing burdens. While tissue specific features likely exist in CPI response and survival, the statistical power of the exhaustive approach is limited at present to detecting them reliably. However, we note, the five main latent factors from the study all have qualitatively similar signals across tissues (Fig. 4 and Extended Data Fig. 6). Complete results from tissue specific models are stored in the shared Supplementary Dataset 1 output but should be interpreted with caution. Finally, it is worth noting that our multiple testing threshold took into consideration our added exploratory testing of the tissue specific models.

2.5. Multiple testing threshold

A challenge in any analysis involving many tests is magnified potential for both statistical false positives and false negatives based on the leniency or strictness of the specified significance threshold. Here, we discuss our approach to specify a multiple testing threshold used for the discovery of features associated with CPI response and survival.

Overall, we exhaustively tested more than 27,000 features, across 4 primary outcomes (best-overall response, progression free survival, overall survival, survival residuals), for 5 primary tumor location cohorts (pan-cancer, skin, lung, bladder, and other primary tumor). These regressions were carried out adjusting for different sets of covariates. Tumor type specific regressions were adjusted by no covariate or by tumor purity, biopsy site and age. The pan-cancer regressions were adjusted by tissue, tissue and purity, tissue, purity, biopsy location, and age). In total, the study included more than 1,100,000 tests (all documented in Supplementary Dataset 1). Adjusting for multiple testing appropriately is a tough task when many correlated tests –as is the case here– are conducted, whose precise mathematical correlation structure is difficult to define. In our case, the systematic testing included many highly correlated features, across highly correlated outcomes, with overlapping cohorts (i.e. pan-cancer includes all the tissues), and highly similar covariate vectors. We emphasize our multiple testing threshold was built based on all the tests run, and not only those from the pan-cancer analyses (adjusted for tissue, purity, biopsy location, and age) highlighted in the main paper.

A naive and extremely conservative approach is to assume all tests are independent and control for the family-wise error rate across the more than 1,100,000 tests with a Bonferroni (BF)

correction - making a multiple threshold of approximately 4.8×10^{-8} . However, the clearly dependent nature of the tests, and conservative family-wise error rate, make the BF approach too stringent.

Conversely, using the Benjamini Hochberg (BH) approach, which also controls the false discovery rate for independent tests, produces a p-value threshold of 4.5×10^{-3} , which is too lenient.

We aimed to control the false discovery rate while accounting for the high dependency of the tests. Therefore, we used the Benjamini-Yekutieli (BY) procedure⁷ to set the multiple testing threshold. The BY procedure is a widely cited method designed to control the false discovery rate under dependency of tests, it is designed to address our situation that involves both multiple correlated outcomes and many correlated features (see BY paper examples). The BY threshold was computed directly in R using the base function `p.adjust` with `method = "BY"`. The BY threshold is less conservative than the Bonferroni method (BF) but more conservative than Benjamini Hochberg (BH). Table SN3 below summarizes the multiple testing and computed number of features (for at least 1 outcome, given covariate adjustment) that pass the multiple test thresholds.

Table SN3. Multiple Testing Thresholds

Group	# Features tested	Threshold set based on all tests in 1.1 million tests in the exhaustive analysis					
		BH threshold = 4.5×10^{-3}		BY Threshold = 5.5×10^{-5}		BF Threshold = 4.8×10^{-8}	
		Pass threshold*	% Pass	Pass threshold*	% Pass	Pass threshold*	% Pass
Somatic mutations	16176	9188	56.8	1932	11.9	16	0.1
CNV/SVs	2500	28	1.1	0	0.0	0	0.0
RNA	9187	1319	14.4	293	3.19	37	0.4
Clinical	38	9	23.7	3	7.9	0	0.0
HLA	22	2	9.1	0	0.0	0	0.0
Total	27923	10546	37.8	2228	7.98	53	0.2

This table focuses only on the pan-cancer analyses run with the covariates used in the main text. It considers a feature passed the multiple testing threshold if it was significant for at least 1 outcome (BOR, PFS, OS, OS residuals). However, the testing thresholds are set based on all 1.1 million study tests (pan-cancer and per tumor type). The number of features that exceed the pass threshold corresponds to the main analysis that adjusts for tissue, biopsy, age, and purity. A feature is considered to pass if it exceeds the multiple testing threshold for at least 1 outcome (BOR, PFS, OS, OS residuals). BH - Benjamini Hochberg, BY - Benjamini-Yekutieli, BF - Bonferroni.

The exhaustive study with the BY multiple testing probably produces Type II statistical errors. While we highlight in the study the five independent factors (TMB, T-cell effective infiltration, prior treatment, TGF-beta activity in the microenvironment, tumor proliferative potential) with

sufficient statistical significance to pass the BY threshold, more independent biomarkers likely exist with association to CPI response and survival.

2.6. Survival Residuals analysis

The three independent latent factors associated with response were also significant univariate predictors of progression free survival, and overall survival. However, clearly, survival outcomes are largely determined by response to therapy. Moreover, the initial analyses neglected potential multivariate signals that could be identified through additional covariates. Therefore, we hypothesized that independent predictors of survival could be uncovered after accounting for the three previously identified latent factors –TMB, T-cell effective infiltration, and exposure to prior therapy– as additional covariates in the regression analysis. To test this hypothesis, we ran another exhaustive analysis to find features that may explain the survival residuals; that is, the unexplained variation in overall survival and progression-free survival after added adjustment for response signals.

For survival residuals analysis, we again used Cox proportional hazards models. For each feature, proportional hazards for overall survival were again modeled:

Full Model:

$$\text{Log}(h(t)) = X\beta_X + \sum_{j \in \text{tissue}} I_j \beta_j + \sum_{k \in \text{biopsy}} I_k \beta_k + X_{\text{purity}} \beta_{\text{purity}} + X_{\text{age}} \beta_{\text{age}} + X_{\text{tmb}} \beta_{\text{tmb}} + X_{\text{tcell}} \beta_{\text{tcell}} + I_{\text{pre}} \beta_{\text{pre}}$$

Reduced Model:

$$\text{Log}(h(t)) = \sum_{k \in \text{biopsy}} I_k \beta_k + X_{\text{purity}} \beta_{\text{purity}} + X_{\text{age}} \beta_{\text{age}} + X_{\text{tmb}} \beta_{\text{tmb}} + X_{\text{tcell}} \beta_{\text{tcell}} + I_{\text{pre}} \beta_{\text{pre}}$$

That is, we added the representative biomarkers of the three latent factors to the covariate vector in the full and reduced models. To remove redundant features, and avoid multicollinearity, all features with correlation > .5 with either TMB, effective T-cell infiltration and prior treatment were excluded from the residual analysis.

The residual analysis followed the same rationale used in the main analyses. We first identified several hundred features with statistical significance greater than the BY threshold. Then, following the analysis described in the main manuscript, we identified two new clusters of features associated with survival residual, which represented TGF-beta activity in the microenvironment and tumor proliferative potential.

In summary, through exhaustive and pragmatic analyses we identified TMB, T-cell effective infiltration, exposure to prior therapy, tumor proliferative potential, and TGF-beta activity in the microenvironment as key independent factors underlying response and survival. These latent factors were identified with different sets of covariables used to adjust the regressions (Extended Data Fig. 1A-C). Moreover, no sets of features significantly associated with CPI response or survival remain when features highly correlated with clusters representing the latent factors are removed (Extended Data Fig. 3A-E).

2.7. Systematic Analysis Summary

In summary, through exhaustive analysis we highlighted TMB, T-cell effective infiltration, prior therapy, tumor proliferative potential, and TGF-beta activity in the microenvironment as key independent factors underlying response and survival. All code and data output from the exhaustive analysis can be found in the project github repository (https://bitbucket.org/bbqlab/immune_biomarkers/src/main/).

3. Discovery of the five latent factors

3.1. Cluster R1

The overall TMB of tumors was the feature in cluster R1 with the highest correlation with the mean of the cluster. We reasoned that cluster R1 represents a latent factor of CPI response identified by TMB. Thus, we called this latent factor TMB and chose the overall TMB of tumors to represent it.

Interestingly, the top ranking of p-value of all features in Cluster R1 of their association with CPI response is occupied by the mutation burden of some individual genes. While we could have used any of them to represent this cluster, we decided to use the TMB for two reasons. First, it showed the highest correlation with the mean value of all features in the cluster. Secondly, we reasoned that TMB would be a more stable proxy metric of neoantigenic burden than any individual gene.

One important result of the analysis of the TMB latent factor is the finding that the mutated status of virtually all frequently mutated genes in the cohort show a positive correlation with the TMB (Fig. 2E; Extended Data Fig. 2A). This implies that analyses of the potential influence of these mutated genes on the response to CPI would be confounded by the TMB, and highlights the difficulties of identifying independent somatic mutation biomarkers. (We expand on this finding in the following sub-section.) For example, mutations, biallelic inactivation, or copy number variation affecting several genes, such as B2M, have been identified as associated with tumor immune escape pathways across studies³. In the HMF-CPI cohort, loss-of-heterozygosity of the genomic region containing the B2M gene appears significantly associated with CPI response in skin melanomas in the HMF cohort. Nevertheless, across the entire HMF-CPI cohort, and in other individual tumor types (lung and bladder) these appeared below significance (Supplementary Dataset 1).

While the association with CPI response is highly significant across the HMF-CPI cohort, it is not equally clear for all tumor types (Fig. 4 and Extended Data Fig. 4A,B). It is clearly significant for lung and bladder tumors, as well as for the 'other tumors' group, but not for melanomas, probably because both, responder and non-responder melanomas have high TMB.

3.1.1. Somatic Gene Mutations Analysis

Somatic gene mutations were a big part of the HMF systemic analysis, as roughly 60% of tested features (16,000 of 27,000) measured mutations within individual genes. To compute the features, somatic gene mutation counts were calculated by parsing the HMF VCF files, and further filtering was performed to remove genes with mutations in less than 1 in 20 samples. Since somatic gene mutations represent a preponderance of all features, and highlight the issue confounding with an underlying latent factor (TMB), we expand here in more detail.

For simplicity, let's focus on the HMF response analysis (logistic) that was un-adjusted for tissue. In this analysis, strikingly, 100% (15851/15851) of genes showed a positive correlation to

TMB; 99% (15665/15851) were positively associated with response to CPIs in the logistic regression analyses; and a clear positive relationship appeared between the correlation to TMB and the statistical significance of the association with CPI response. Given these observations, and also accepting that TMB is a biomarker for response to CPIs and that mutations are distributed along the genome, arguably the somatic mutations observed in *all* genes are likely weak positive biomarkers for response CPIs. Conceptually speaking, thus, the rate of mutations of any gene is a weak proxy of TMB with lower statistical power. We hypothesized then that longer genes (since they have higher chance to bear mutations) should tend to be better proxies with higher statistical power. To test this, we collected the gene transcript sizes (from *geneLenDataBase* in bioconductor⁸) and joined them with the exhaustive analysis results. As shown in Figure SN1, we do indeed observe that longer genes tend to show higher statistical significance, corroborating that gene mutation features are weak proxies of TMB.

Figure SN1

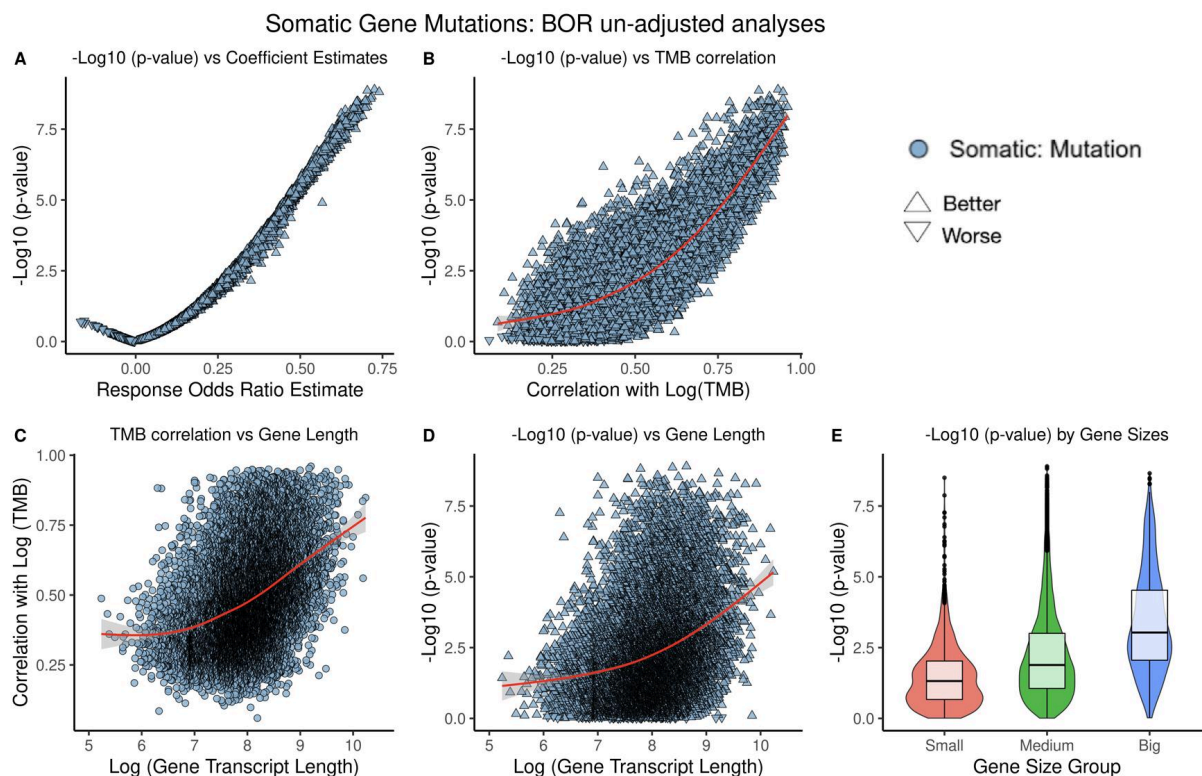


Figure SN1. Best overall response un-adjusted analysis of somatic gene features

- Individual univariate analyses for best overall response (logistic regression) across all somatic gene features. Almost all somatic gene features are positively associated with response to CPIs (x-axis), labeled "Better".
- Individual univariate analyses for best overall response (logistic) across all somatic gene features. We see a clear positive relationship between the correlation with TMB (x-axis) and the statistical significance of the association with response (y-axis).
- Demonstrates the positive relationship between gene size (x-axis) and correlation to TMB (y-axis).

d) Individual univariate analyses for best overall response (logistic) across all somatic gene features. Larger genes measured by log(gene transcript length) (x-axis) tended to show more statistical significance (y-axis).

e) Violin plots of distribution of statistical significance by gene size. On the x-axis, genes are grouped by size quantiles (x-axis): Small (0-10%), Medium (10-90%), and Big (90-100%) and the y-axis measures the $-\log_{10}$ (p-value). Large genes show the most statistical power.

All corrected p-values (BY) shown in the plots were computed via logistic (response) or Cox (survival) regressions, as in Figures 2 and 3 of the main manuscript. These are, by definition, two-sided

The plots represent all gene analyzed for their mutation burden (N=15,829)

In the boxplots represented within the violin plots, the center line represents the median of the distribution, the box limits, the first and third quartiles, the whiskers, the lowest/highest data points at first quartile minus/plus $1.5 \times$ IQR, and the points, the outliers.

3.2. Cluster R2

Cluster R2 includes only two clinical features, representing prior systemic treatment or any prior treatment received by patients for their primary tumor. Patients with prior treatments in the HMF-CPI cohort represented between 75% and 87% of those suffering from lung, bladder or 'other' tumors, but only 26% of those suffering from melanomas. The outcome of patients with exposure to prior treatment was worse, for all tumor type groups, than that of patients with no prior exposure. Across the HMF-CPI cohort patients with no prior treatment (n=195) had a 47% response rate, those with only prior radiotherapy (n=28) had a 36% response rate, those with only prior systemic therapy (n=144) had a 22% response rate, while the patients exposed to both prior systemic therapy and radiotherapy (n=112) responded at a rate of only 14%. We hypothesized that patients previously exposed to prior therapies may present in worse general condition, or with more aggressive metastatic tumors.

3.3. Cluster R3

The mean of the expression of all genes in cluster R3 exhibits a high correlation with gene sets representing T-cell infiltration (Fig. 2D), the overall cytolytic activity in the tumor, as well as individual genes related with cytotoxic T-cell activity, such as the CD8A and CXCL9 genes (Supplementary Dataset 1). The former is a membrane protein marking the lineage of cytotoxic T-cells⁹, while the latter is a chemokine involved in the growth and activation of cytotoxic T-cells¹⁰

Interestingly, gene sets representing other populations of immune cells (and population levels inferred using the Cibersort method⁶) for the most part appear either correlated or anti-correlated with the T-cell effective infiltration representative biomarker (Supplementary Dataset 1). In particular, populations of myeloid cells, such as M0 and M2 macrophages, which appear significantly negatively associated with CPI response are highly anticorrelated with T-cell effective infiltration. This raises the question of whether the latent factor represented by the T-cell effective infiltration gene set actually relates to a more general immune infiltration pattern. Irrespective of the true nature of this latent factor, it does appear that all immune populations that appear significantly associated with response are proxy biomarkers of the same aspect of

the immune infiltrate of the tumor, and are thus efficiently represented by the T-cell effective infiltration gene set.

3.4. Cluster S1

A number of genesets related with proliferation, cell cycle and other basic cellular processes exhibited high or moderate correlation with the mean expression of the genes in cluster S1 (Extended Data Fig. 4B,C). The highest correlation with the mean of this cluster was exhibited by a geneset first derived by Pabla et al. (2019)¹¹ as a group of genes showing differential expression between CPI responder and non responder lung tumors. Upon enrichment analysis, the authors of the original report concluded that these genes were all related with cell proliferation. Therefore, we have called the latent factor emerging from this cluster tumor proliferative potential. Tumors with higher proliferative potential have a significantly worse outcome across tumor types and cohorts (Fig. 4 and Extended Data Fig. 6). They appear to reflect a basal underlying factor influencing the outcome of any tumor, irrespective of the treatment (Extended Data Fig. 4A,B). As expected, gene sets that show a high correlation with the mean of this cluster also appear mutually correlated (Extended Data Fig. 3C).

3.4. Cluster S2.1

The geneset exhibiting the highest correlation with this cluster was related with TGF-beta signaling in fibroblasts¹². Other genesets also exhibiting a high correlation with the mean of this cluster represent, more generally, the activity of TGF-beta in the microenvironment of the tumor^{13,14}, or are related with the epithelial to mesenchymal transition of the tumor (Extended Data Fig. 3B,E). Thus, we decided to name the latent factor emerging from this cluster TGF-beta activity in the microenvironment. The influence of TGF-beta activity in the microenvironment on overall survival upon CPI treatment is also apparent across validation cohorts (Fig. 4 and Extended Data Fig. 6).

More fine-grained analyses reveal also a direct bearing on CPI response. The mean of the cluster, while not significantly different across responders and non-responders in the entire cohort or any group of tumors, is significantly different when the patients are stratified according to the level of effective T-cell infiltration (Fig. SN2). Despite this result, TGF-beta activity in the microenvironment and T-cell effective infiltration appeared in our analysis as truly independent latent factors in their association with response and survival.

Figure SN2

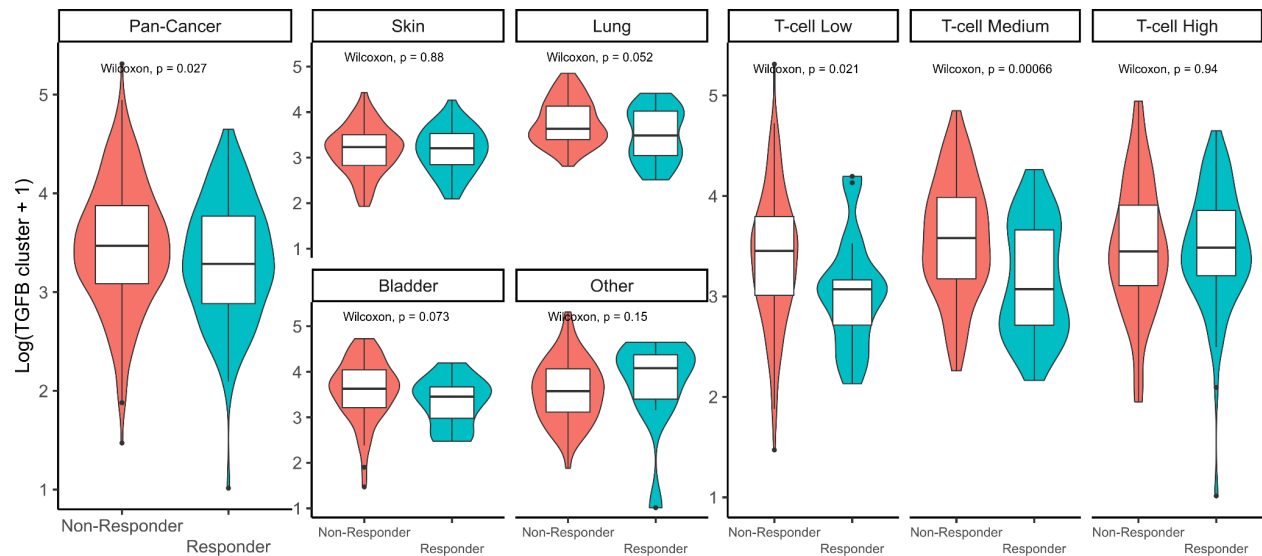


Figure SN2. The effect of the activity of TGF-beta in the microenvironment on CPI response depends on the degree of effective T-cell infiltration.

The violin plots represent all tumors in the HMF-CPI cohort with RNAseq data (Pan-cancer, N=396; bladder, N=77; lung, N=83; other tumors, N=72; melanoma, N=164)

In the boxplots represented within the violin plots, the center line represents the median of the distribution, the box limits, the first and third quartiles, the whiskers, the lowest/highest data points at first quartile minus/plus $1.5 \times \text{IQR}$, and the points, the outliers. The p-values were obtained via two-sided Wilcoxon-Mann-Whitney tests.

As shown in Figure 3B, clusters S2.1 and S2.2 contain genes with a certain degree of mutual correlation. The decision to split them into two clusters was purely data-driven, based on the dendrogram resulting from the hierarchical clustering. While the mean of the expression of genes in cluster S2.1 show a strong correlation with several genesets reflecting the activity of TGF-beta in the microenvironment of the tumor, the expression of genes in cluster S2.2 also show some degree of correlation with these genesets. Therefore, these two clusters do not represent two independent latent factors, but two sets of genes with different levels of correlation of their expression with just one latent factor. We have decided to keep them separately, following the initial data-driven split, under the assumption that genes in cluster S2.1 are closer to the latent factor and the inclusion of the genes in cluster S2.2 may affect its stability.

3.6. Interchangeability of proxy measurements of the latent factors

One important characteristic of the latent factors that becomes apparent from the way they are discovered (Fig. 2 and 3, main manuscript) is that they are integrated by multiple highly intercorrelated features. Actually, the behavior of these features, naturally clustering on the

basis of their pairwise correlations, directly leads to the discovery of latent factors. The idea comes intuitively, then, that the features in each cluster could be used interchangeably to represent the latent factors. This idea is represented in Supplementary Figure 1, where a number of features of different types (together with genesets) exemplify the association with CPI response and survival and the correlation with the different latent factors. This is an important point that implies that multivariate models alternative to the ones described in the manuscript could be trained with different arrays of features. We put this hypothesis to the test by training multivariate models using two alternative metrics of the five latent factors and comparing them with the models trained on the metrics described in the main manuscript. The results of this comparison demonstrate the equivalence between the three models trained on three different metrics of the latent factors (Supplementary Fig. 3).

3.7 Latent factors in non-CPI therapies

A comparative analysis of the association of the five latent factors with CPI and non-CPI therapies provides intriguing observations (Extended Data Fig. 4A,B). In each of the four graphs in the figure, the latent factors are represented as circles in a plane of statistical significance vs effect size of their association with the two outcomes measured (response, A or overall survival, B) across CPI (top) or non-CPI (bottom) treated patients in the HMF cohort. The vertical dotted line in each graph separates the positive (increased response or survival) and negative (decreased response or survival) effects of the latent factors on the outcome. The horizontal dotted line represents the threshold of statistical significance for the associations. Thus, circles in the top left or right quadrants represent latent factors significantly associated with a negative or positive outcome, respectively. Circles in either bottom quadrant represent non significant associations between the latent factor and the outcome in question.

The analysis of the results show that the effects of TMB, effective T-cell infiltration and TGF-beta activity in the microenvironment are specific for CPI response and survival. A (weaker) significant association with increased response to non-CPI therapies is still observed for effective T-cell infiltration, probably as a result of the general role of the immune system in the clearance of the tumor in these cases. However, no effect of survival upon treatment with non-CPI therapies is observed, contrary to the result for CPI therapies. The effect of TMB in survival upon treatment with non-CPI therapies is the opposite to that observed for CPI therapies. Tumors with a high TMB may be more aggressive, or evolve faster, accelerating the acquisition of resistance to therapies, thus resulting in decreased survival.

The results for prior treatment and tumor proliferative potential suggest that their effect on response and survival is not necessarily restricted to CPI therapies, but rather more general. Specifically, the exposure to prior therapy results is significantly associated with decreased response to both CPI and non-CPI therapies. The negative effect on survival is significant for CPI, but not for non-CPI therapies. The significant negative association of survival with tumor proliferative potential observed for CPI therapies is also apparent (more highly significant) for non-CPI therapies. Nevertheless, unlike in the case of CPI, the tumor proliferative potential is significantly associated with an increased response to non-CPI therapies, probably because

many targeted and cytotoxic therapies specifically target the proliferative capabilities of tumor cells.

3.8 Conservation of the expression of genes in transcriptomics latent factors

We aimed to explore the conservation of the expression of the genes representing T-cell effective infiltration, TGF-beta activity in the microenvironment, and tumor proliferative potential across four validation cohorts (Extended Data Fig. 6B). To do this, we computed the silhouette scores for each gene in each of the cohorts. This score tracks the degree of kinship of a gene with respect to a cluster, with higher silhouette scores designating genes that are close to the center of their cluster and well separated from other clusters. We then ranked the genes in the three clusters in descending order of their silhouette scores in each cohort, and finally, we computed a rank aggregated score tracking how much more than expected each gene tended to appear at the top of the ranking. We observed that genes with higher silhouette scores in HMF-CPI tend to have higher rank aggregation scores (Extended Data Fig. 6B). In other words, genes closer to the cluster center in HMF-CPI tend to also be closer across cohorts. This indicates that latent factors (and their association with CPI response and survival) are relatively resistant to the potential aforementioned perturbations.

3.8 Significant features outside clusters

When we clustered the features significantly associated with CPI response, three genes remained as a small cluster of features which were thus not included as part of any latent factors. These genes are BAIAP2, an Insulin Receptor Substrate, LATS2 a serine threonine kinase, and GGT5, a Gamma-Glutamyltransferase 5. We consider the significance of these genes spurious. In the study of the features significantly associated with survival residuals, three clusters were obtained, although closer inspection reveals that the features in the third cluster appear somewhat correlated with the second cluster (TGF-beta activity in the microenvironment). Therefore, while we have chosen to not include them, further studies could reveal that they are indeed part of this latent factor.

4. Validation

The discovery of the 5 latent factors was carried out in the HMF-CPI cohort, drawn from a specific population, and with a specific tumor type composition. The tumor samples obtained for the molecular profiling were freshly frozen, and they were profiled at whole-genome and whole-transcriptome levels. We wondered if these same factors are associated with the response to CPI therapies in other cohorts representing different arrays of tissues and with samples profiled using different technologies. This is a key question given the variety of technologies for sampling and profiling employed in the clinical practice. To answer it, we collected the data of five cohorts from different studies available in the public domain, and one real-life cohort of cancer patients treated at the Vall d'Hebron Institute of Oncology in Barcelona (summary in Table SN4, and complete information in Supplementary Table 1).

Table SN4. Summary description of cohorts used in the study

Cohort	N	BOR data	Survival data	Somatic data	Transcriptomics data	Pretreatment data	All data types	Tissues included
HMF-CPI	479	467	479	479	396	479	385	bladder, lung, skin, other
INSPIRE	76	76	76	64	76	76	64	skin, other
Lyon	315	178	315	0	315	184	0	head and neck carcinoma, lung
MARIATHASAN	348	298	348	272	348	0	0	bladder
PARKER ICI	315	315	0	212	259	315	0	skin
RAVI	362	352	352	297	120	359	57	lung
VHIO	75	75	74	75	75	59	58	bladder, colon, lung, skin

Overall, the cohorts of the six validation studies include 1491 tumor samples of the cancer types represented in the HMF-CPCT cohort and others, such as head and neck squamous cell carcinoma and colorectal cancer. The procedures used in sample collection and profiling in each of these studies represent the variety observed across the clinical practice (Table SN5).

Table SN5. Procedures employed across studies for tumor sample collection and profiling

Study	Sample collection	Genomic profiling	Transcriptomic profiling	Paired normal sample
HMF-CPI	Fresh frozen core needle biopsy	Whole-genome sequencing	RNA sequencing	Yes

INSPIRE	Fresh frozen pellet of cells obtained from dissociation of biopsy tissue	Whole-exome sequencing	RNA sequencing	Yes
Lyon	FFPE	-	Targeted (2559 genes) RNA sequencing	-
MARIATHASAN	FFPE	Whole-exome sequencing (FMone in a subset of cases)	RNA sequencing	Yes
PARKER ICI	FFPE	Whole-exome sequencing	RNA sequencing	Yes
RAVI	FFPE	Whole-exome sequencing	RNA sequencing	Yes (blood or adjacent FFPE)
VHIO	FFPE	Targeted sequencing	NanoString of 170 genes	No

To test the association of the five latent factors identified across the HMF-CPI cohort with CPI response and survival in the six validation cohorts, we used the clusters of features described above. In the case of the VHIO cohort, where only the expression of a small set of 170 genes is available, the effective T-cell infiltration, TGF-beta activity in the microenvironment and tumor proliferative potential were measured through groups of genes with expression correlated with those of the corresponding latent factors (in the HMF-CPI cohort).

We reasoned that the differences in sample collection protocols and tumor profiling could influence the signal of the association between the five latent factors and CPI response and survival. However, we found that, overall, in the six validation cohorts, the five latent factors show the same trend of association with CPI response or survival as in the HMF-CPI using either univariate (Extended Data Fig. 6) or multivariate regressions (Fig. 4). Still, some differences were observed for particular cohorts. For example, TMB in the VHIO cohort showed lack of significant association with response to CPI (with a trend opposite to that observed across all other cohorts), probably due to the difficulty in accurately calculating the TMB in a tumor-only sequencing scenario. In the univariate analysis TGF-beta activity in the microenvironment appears less significant in association with survival than in its multivariate counterpart. This is because in the second, the effect of each latent factor on CPI response and survival is corrected for the influence of the other four. This is closer to measuring the effect on survival residuals, as in the process of discovery.

Interestingly, despite differences in sample collection procedures (fresh frozen vs FFPE) and transcriptome profiling (RNAseq vs NanoString), the association of effective T-cell infiltration and TGF-beta activity in the microenvironment (based on gene expression features) are observed in the overall assessment of the validation cohorts. Tumor proliferative potential, although negatively associated with survival upon CPI treatment in the overall analysis of validation cohorts (as in the HMF-CPI cohort), does not reach statistical significance.

The comparison of cohorts composed of patients with the same tumor type reveals some interesting regularities and differences. For example, the PARKER ICI cohort integrated by patients with different subtypes of melanoma (from different individual subtypes) shows an overall behavior that is very similar to that of skin tumors (metastatic melanomas) in the HMF-CPI cohort. In both cohorts, association with response (with the same sign) is observed for TMB, effective T-cell infiltration and pretreatment. The effect of TMB on response is stronger (reaching significance) in the PARKER ICI cohort than in the HMF-CPI. In the case of lung, the latent factors show similar results in their association with CPI response and survival as in the HMF-CPI lung cohort. This includes a lack of association of effective T-cell infiltration with CPI response.

Despite these results showing the universality of the latent factors of CPI response and survival across tumor types in different cohorts, important challenges remain to guarantee that a multivariate model comprising them trained on data from one cohort could be readily transferred to other cohorts. Some of them have been mentioned previously. They include different sample collection and processing methods, as well as tumor profiling approaches, different composition of tumor types, as well as gaps of data annotation. A multivariate model may show an important decrease in performance from one cohort to another even as the latent factors are shown to be significantly associated with CPI response across both. Such a model is not only sensitive to fluctuations of individual latent factors, but also of the interactions between them.

Still, in this work we thought it was important to explore the construction of a multivariate model integrating the five latent factors. We believe this could be a proof-of-principle of a way, within a growing existing cohort, to stratify newly profiled patients based on their probability to respond to CPI, thus supporting clinical decision-making. The construction of this model is explained in the next section.

5. Multivariate machine learning models of CPI response and survival

Our broad goal was to create machine learning models that combined both high predictive power and interpretability. To meet this goal, we chose to use boosted decision trees as the underlying method for our model building. Decision trees are able to capture interactive and non-linear associations between features and outcome, while their fitting XGBoost is generally known to have strong predictive performance^{15–17}. Moreover, tree-based models facilitate both feature level interpretations in the form of relative influence measures and partial dependence plots, as well as patient level interpretations in the form of Shapley values¹⁸. In this section, we go into details about the XGBoost fitting procedure, our larger prediction study, and method for extracting Shapley values used for interpretation.

5.1. Model fitting

All models were fitted with the XGBoost package in R. For both the response and survival models we used default loss functions provided within XGBoost. For best overall response, the log-loss from logistic regression likelihood was selected, and for the survival models the Cox log-likelihood was optimized. Within XGBoost, various hyperparameters that include learning rate, model depth, minimum tree leaf size, and sub-sampling are specified, and their input values can affect model fitting. To address the hyperparameters, the XGBoost model fitting was designed to handle a parameter sweep across hyperparameters and choose the optimal parameters based on the XGBoost internal cross-validation evaluation.

5.2. Evaluation of models with repeated re-sampling and cross-validation

Xgboost models were internally evaluated within the Hartwig CPI samples using a simulation based cross-validation study. For the simulation study, the CPI data was split 1000 times into training and test sets comprising 80% and 20% of the data, respectively. Within each iteration, samples were randomly split in a stratified manner along two dimensions - primary tissue (skin, lung, bladder, and remaining mixed tissues) and completeness status (had transcriptomics profiled or not) - to maintain roughly 4 to 1, train to test proportions (or 80/20) within each combined sub-group (i.e. Lung, complete data). For each training set, various models were constructed for the 3 measured CPI outcomes - best overall response, progression free survival and overall survival. All constructed models from the training data were applied to the test data. From the test set predictions, best overall response models were evaluated with the area under the curve (AUC) metric, while progression free and overall survival models were evaluated with the concordance index (C-index). All model evaluation was restricted to subsets of the test data with completeness (i.e. no missing transcriptomic features). Evaluation metrics for all model predictions were collected across 1000 iterations, and results were aggregated.

5.3. Broad overview of HMF XGBoost studies

Our broad goal was to evaluate the predictive performance on XGBoost models based on the five factors - TMB, T-cell effective infiltration, TGF-beta- activity in the microenvironment (Pan-TBRS) gene set, tumor proliferative potential gene sets, and exposure to prior therapy. To meet this goal, we conducted 3 separate studies which each focused on evaluation of individual aspects of the XGBoost modeling. In 2 initial studies, we investigated the best basic modeling strategy and hyperparameter settings for the five factor model. In the first study, we probed whether pan-cancer, tissue specific, or hybrid models exhibit best the performance results. In this first study, hybrid models showed the best performance, and were the focus of all subsequent XGBoost modeling. In the second study, we examined the XGBoost hyperparameters and their influence on the model fits. The results from the second study suggested that simple additive XGBoost models show satisfactory predictive performance, as no appreciable gains are observed with depth greater than 1. In our final study - using hybrid models with simple additive hyperparameters - we compared the performance of models with various input features to the five factor model. All multivariate models built were, thus, based on trees of depth 1. A summary of the 3 studies is shown in Table SN6.

Table SN6. Overview of XGBoost studies

Study	Focus of Study
1	Hybrid models: compare Pan-Cancer vs Tissue Specific vs Hybrid Models
2	Hyperparameter study: compare fits with different hyperparameters
3	Main Study: compare performance with different input features

5.4. Study 1: Hybrid models

We wanted to build tissue specific models that leveraged learning from the pan-cancer data. We had observed five latent factors highlighted in the exhaustive study with similar effects across tissues, but the magnitude of their effects differed. We sought a modeling approach that balanced the superior stability of the pan-cancer models with the superior specificity of simple tissue cohort models, which we termed hybrid models. To build these hybrid models, the pan-cancer models were plugged as initial inputs into the training of the 4 tissue models (skin, lung, bladder, mixed) which were then further updated to better fit the tissue specific cohort.

To compare the 3 model building approaches we ran a simulation study. We compared predictive performance of the pan-cancer, tissue specific, and the hybrid models (Table SN7). In the simulation study, all models used the five factors as features and used simple additive hyperparameter settings.

Table SN7. Hybrid Models Study Settings

Model Features	Modeling Method	Settings
----------------	-----------------	----------

Five Factors: TMB + T-cell + Proliferation + TGF-beta + Prior therapy	Pan-Cancer	Simple Additive
	Tissue Specific	
	Hybrid	

Results from the study are shown in Supplementary Figure 2. Overall, the cross validation study results showed nuance and trade-offs between the pan-cancer, tissue, and hybrid models. Within tissues the optimal approach varied; within the bladder and lung cohorts the tissue specific models showed superior performance, whereas the pan-cancer models were best within the skin and other cohorts. Meanwhile, across each cohort, the hybrid models either equaled best performance or found a middle value between the pan-cancer and tissue specific models. For the survival outcomes, the hybrid and pan-cancer models were clearly best. While relative performance varied across tissue cohorts the pan-cancer and hybrid models tended to be superior for survival outcomes.

From the observed results, we concluded the hybrid models performed best when broadly viewed across both response and survival outcomes. Therefore, we chose to use the hybrid models for all comparisons of different feature inputs, and their subsequent validation.

5.5. Study 2: Hyperparameter Settings Study

To understand and ultimately select satisfactory hyperparameters we ran a second simulation study. Our goal for the hyperparameter study was to evaluate the impact of hyperparameter settings on predictive performance, and importantly, determine if interaction (depth parameter > 1) added predictive power to the models. For all evaluation in the hyperparameter setting study we focused only on the models built with the five latent factors and fit with the hybrid modeling method.

We studied hyperparameters over three pre-specified settings, described in Table SN8. For the simple interaction and full grid models, within each training iteration, the final specified parameters used were selected with internal XGBoost cross-validation across a grid search.

Table SN8. Hyperparameter settings studies

	Hyperparameter settings					
Model Features	Settings	Learning rate	Depth	Min leaf size	Subsampling	Others
Five Factors: TMB + T-cell + Proliferation + TGF-beta + Prior therapy	Simple Additive	.05	1	5	.75	default
	Simple Interaction	.05	1 or 2	5	.75	default
	Interaction + Full grid	.05 or .1	1 or 2	5 or 10	1	default

Results of the hyperparameter grid search are shown in Figure SN6. All the models provide highly similar results with no noted improvement with introduction of interaction or full grid

search. Overwhelmingly, the clear message from the study was that the models fit with different hyperparameter settings had similar predictive performance. No appreciable gains were observed from more complex hyperparameter settings. Therefore, we focused on fitting models with the simple additive settings, since these have the added value of simpler interpretation and very fast testing.

Figure SN6

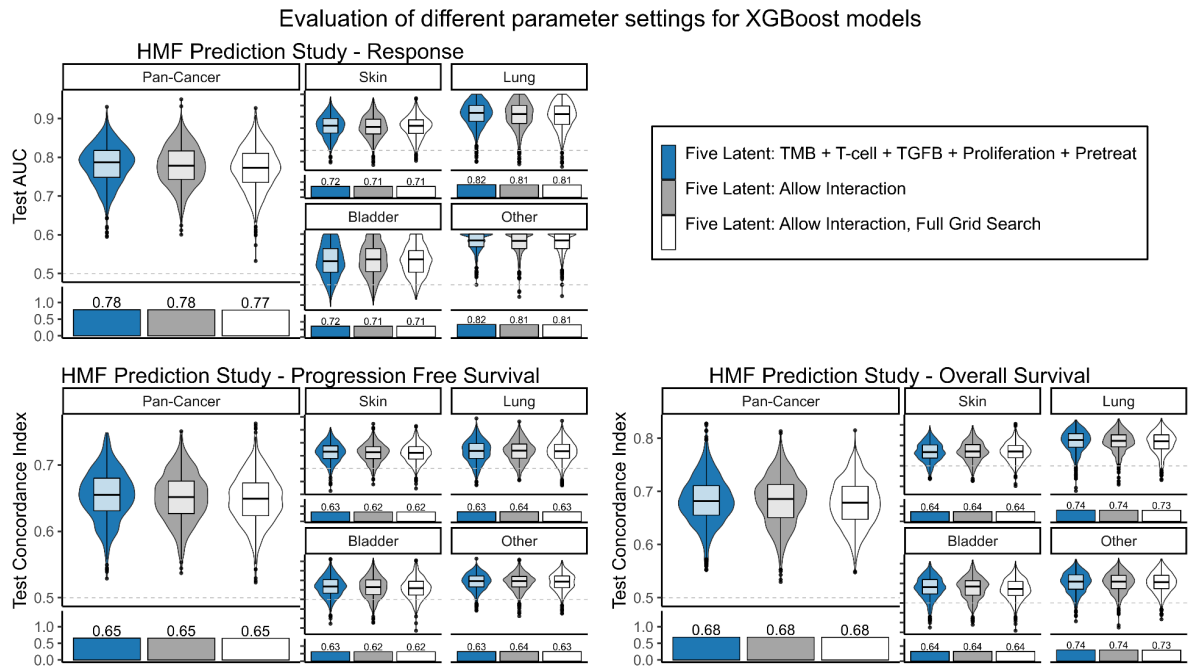


Figure SN6. Similar predictive performance across models fit with different hyperparameter settings.

We compare 3 XGBoost models of the 5 factors with different hyperparameters. For each cross-validation run, the XGBoost model selects the best hyperparameters based on internal cross-validation criteria. In the boxplots represented within the violin plots, the center line represents the median of the distribution, the box limits, the first and third quartiles, the whiskers, the lowest/highest data points at first quartile minus/plus 1.5× IQR, and the points, the outliers.

5.6. Study 3: Main Study comparing various models with different features

In our final XGBoost study, we evaluated the relative predictive performance of five factors models compared to models with alternative features. We compared 5 models with different sets of input features that are summarized in Table SN9, below. Results of the study can be found in Supplementary Figure 3.

Table SN9. XGboost feature study features study settings

Model	Model Features
TMB only	TMB

Approved Biomarkers	TMB + PDL1
No TMB	T-cell + Proliferation + TGF-beta + Prior therapy
Five Factors	TMB + T-cell + Proliferation + TGF-beta + Prior therapy
Full Model	<i>Five Factors</i> + purity + age + distal biopsy + CPI mechanism + LOH HLA + WGD + SV TMB

We first compared the five factor model to a model trained only on approved biomarkers (TMB + PDL1 expression). For response, overall the five factor model clearly outperformed the model fit only approved markers with average AUC values of .79 and .73 respectively. The five factor model outperformed the TMB + PDL1 model in all cohorts. The survival outcomes evaluated showed a similar story.

Next, we evaluated the importance of inclusion of TMB in the five factor model performance. Overall, the no TMB model (T-cell effective infiltration, tumor proliferative potential, TGF-beta activity in the microenvironment, pretreatment) showed decreased predictive power. By removing TMB from the five factor model, the pan-cancer average AUC only dropped to .74 (vs .78), while the C-index metrics for PFS decreased slightly (.64 vs .65), and did the OS metric stayed similar (.68 vs .68). The results varied across tissues.

We also compared an RNA only model to the five factor model. Overall, the RNA model (T-cell effective infiltration, tumor proliferative potential, TGF-beta activity in the microenvironment) showed decreased predictive power relative to the five factor model. The model showed similar performance to the TMB model for predicting responder with AUC of .71 (vs .71 for TMB only), and improved performance that was similar to TMB + PDL1 models for predicting survival outcomes. The results varied across tissues.

For our final comparison we investigated whether a full model with additional features that included age, distal biopsy, mechanism (PD1/PDL1 or PD1/PDL1 + CTLA4), LOH HLA, whole genome doubling, and structural variant TMB could improve prediction of the CPI outcomes. From this investigation, we observed no appreciable gain from the extra features.

In summary, in this HMF cross-validation study we observed that 1) the five factor XGBoost model was superior to a model built with only TMB + PDL1, 2) removing TMB from the five model did not diminish prediction power outside the lung cohort, 3) RNA only models may provide satisfactory predictions in some settings, and 4) a fuller model with more features did not appreciably improve model performance.

5.7. Relationship between predicted response and predicted hazard across patients in HMF-CPI

We explored the relationship between the predicted probability of response and hazard across the HMF-CPI cohort. Despite a good (expected) inverse correlation between predicted

probability of response and hazard, there is a wide distribution of the hazard values predicted for patients in the low probability of response group (Extended Data Fig. 8A). Actually, a few patients in this group show predicted hazard values similar to those of patients in the group of high response probability. To understand what prompts this divergence between response and survival multivariate models, we separated HMF-CPI patients with predicted hazard above (high) and below (low or medium) 1.5 hazard ratio. We observed that the tumor proliferative potential, TGF-beta activity in the tumor microenvironment and (to a lesser degree) the exposure to prior systemic treatment contributed (SHAP values; Methods) more than the other two latent factors to the separation between these two groups of patients (Supplementary Fig. 4). These differences are best illustrated through examples (Extended Data Fig. 8C). For example, patient 2 presents a case with very low predicted probability of response (0.07) and very high predicted hazard ratio (2.6), with all latent factors contributing in the same direction to both outcomes. In the case of patient 1, a similar predicted probability of response (0.08) is mostly driven by the analogous negative contribution of T-cell infiltration, TMB and exposure to prior treatment. However, an important contribution of the tumor proliferative potential and TGF-beta activity in the tumor microenvironment result in a much lower predicted hazard ratio (0.3). Finally, for patient 3, the multivariate models predict a high probability of response (0.64) and, coherently, a low hazard ratio (0.1).

References

1. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
2. Schwartz, L. H. *et al.* RECIST 1.1—Update and clarification: From the RECIST committee. *Eur. J. Cancer* **62**, 132–137 (2016).
3. Martínez-Jiménez, F. *et al.* Genetic immune escape landscape in primary and metastatic cancer. 2022.02.23.481444 Preprint at <https://doi.org/10.1101/2022.02.23.481444> (2022).
4. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
5. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
6. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
7. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
8. Young, M. geneLenDataBase: Lengths of mRNA transcripts for a number of genomes. R package version 1.34.0. *Bioconductor* <http://bioconductor.org/packages/geneLenDataBase/> (2022).
9. Taniuchi, I. CD4 Helper and CD8 Cytotoxic T Cell Differentiation. *Annu. Rev. Immunol.* **36**, 579–601 (2018).
10. House, I. G. *et al.* Macrophage-Derived CXCL9 and CXCL10 Are Required for Antitumor Immune Responses Following Immune Checkpoint Blockade. *Clin. Cancer Res.* **26**, 487–504 (2020).
11. Pabla, S. *et al.* Proliferative potential and resistance to immune checkpoint blockade in

- lung cancer patients. *J. Immunother. Cancer* **7**, 27 (2019).
12. Mariathasan, S. *et al.* TGF β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).
 13. Tauriello, D. V. F. *et al.* TGF β drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* **554**, 538–543 (2018).
 14. Cañellas-Socias, A. *et al.* Metastatic recurrence in colorectal cancer arises from residual EMP1+ cells. *Nature* **611**, 603–613 (2022).
 15. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
 16. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016). doi:10.1145/2939672.2939785.
 17. Horrell, M. T. Wide Boosting. Preprint at <http://arxiv.org/abs/2007.09855> (2022).
 18. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. *et al.*) 4765–4774 (Curran Associates, Inc., 2017).