BIOCOMPASS: INTEGRATING BIOMARKERS INTO TRANSFORMER-BASED IMMUNOTHERAPY RESPONSE PREDICTION

Sayed Hashim, Frank Soboczenski & Paul Cairns
University of York, UK
{sayed.hashim}@york.ac.uk

ABSTRACT

Datasets used in immunotherapy response prediction are typically small in size, as well as diverse in cancer type, drug administered, and sequencer used. Models often drop in performance when tested on patient cohorts that are not included in the training process. Recent work has shown that transformer-based models along with self-supervised learning show better generalisation performance than threshold-based biomarkers, but is still suboptimal. We present BioCOMPASS, an extension of a transformer-based model called COMPASS, that integrates biomarkers and treatment information to further improve its generalisability. Instead of feeding biomarker data as input, we built loss components to align them with the model's intermediate representations. We found that components such as treatment gating and pathway consistency loss improved generalisability when evaluated with Leave-one-cohort-out, Leave-one-cancer-type-out and Leave-one-treatment-out strategies. Results show that building components that exploit biomarker and treatment information can help in generalisability of immunotherapy response prediction. Careful curation of additional components that leverage complementary clinical information and domain knowledge represents a promising direction for future research.

1 INTRODUCTION

The immune system is responsible for the management of cancer and the identification of neoantigens produced by tumour cells that can trigger cellular immune responses (Grivennikov et al., 2010). However, tumour cells have devised ways to avoid immune surveillance (Rabinovich et al., 2007). To tackle this challenge, cancer immunotherapy emerged with the objective of reinstating the immune system's capacity to identify and destroy cancer cells (Li et al., 2024). Although immunotherapy has improved the prognosis for patients, its success is limited to a select, unpredictable fraction of individuals diagnosed with cancer (Drake et al., 2014). Thus, accurate characterisation of the tumour microenvironment (TME) in a patient with the ability to anticipate responses to immunotherapy is critical to enhance the efficiency of immunotherapy treatment strategies (Li et al., 2024).

Existing methods for predicting the efficacy of immunotherapy are predominantly dependent on specific biomarkers, including the level of immune cell infiltration (Simoni et al., 2018), the expression levels of programmed death 1 (PD-1) and programmed death-ligand 1 (PD-L1) (Garon et al., 2015), the expression of the cytotoxic T lymphocyte-associated protein 4 (CTLA-4) (Leach et al., 1996), as well as tumour mutational burden (TMB) (Rizvi et al., 2015). However, current clinical methodologies that rely on threshold-based approaches are often inadequate (Li et al., 2024). Many machine learning (ML)-based methods have been proposed to estimate biomarkers and treatment outcomes (Li et al., 2024). These models face challenges when tasked with new data that they were not previously trained on. When evaluated on new datasets, their performance tends to be mediocre or even inadequate, highlighting a gap in their ability to generalise (Li et al., 2024).

A recent work called COMPASS (Shen et al., 2025) used self-supervised learning (SSL) with a transformer-based encoder and a biologically grounded concept bottleneck layer to improve performance across cancer types and treatments. COMPASS is pre-trained on gene expression data from

33 types of cancers using a triplet loss based SSL method. Pre-training improves its generalisability, while the concept bottleneck enables interpretability. COMPASS is fine-tuned on clinical cohorts to predict immunotherapy response. In COMPASS, patient embeddings produced from the encoder are passed onto a concept bottleneck layer to generate scores for 44 biological concepts such as genome integrity, cell proliferation and immune checkpoint for each tumour. These are then passed onto a classifier module to generate treatment response probabilities.

The authors of COMPASS used Leave-one-cohort-out (LOCO) strategy to evaluate its generalisability. In this setting, all cohorts except one are used for training, and the left-out cohort is used for testing. Although the generalisability of COMPASS in this setting is better than methods that use single biomarkers, such as the expression of PD-1 or PDL-1, it is still suboptimal. For instance, the accuracy of COMPASS is about 65% across small cohorts (sample size less than 20) in LOCO setting as reported in their publication. Moreover, COMPASS does not make use of treatment information or external biomarkers. As treatment response could vary based on the treatment type, it is vital to feed this information into the model. COMPASS also does not have a way to validate its concepts with external biomarkers during training.

We present BioCOMPASS, a modified version of COMPASS that integrates external biomarkers and treatment information using components including treatment gating, concept alignment, pathway consistency, and auxiliary multi-task learning into COMPASS. Our contributions are the following.

1. A treatment gating layer to feed information about the target of the treatment (eg. PD-1, CTLA-4, combination) into the model so that it produces treatment-aware concepts.
2. Alignment between known external biomarker scores and concept scores produced by the model to ensure that concept scores are validated against biomarker scores during training.
3. Pathway consistency loss to ensure that embeddings produced by the model contain pathway relevant information and are biologically grounded.

In short, BioCOMPASS is an extension of COMPASS that exploits treatment information as well as external biomarker and pathway scores in order to make the model treatment-aware, biologically grounded, and thus more generalisable. This work also shows that rather than feeding clinical and biomarker data as input to the model, aligning the intermediate latent representations using them could be a good avenue to pursue in general medical applications, especially if such data is only available during training and not inference.

2 METHODS

2.1 DATA

The authors of COMPASS finetuned the model on a total of 16 cohorts. However, COMPASS does not provide access to these datasets; rather, they list the original publications of the cohorts, and access to these datasets needs to be requested from the original publications. Due to difficulties in accessing them, we visited the CRI iAtlas portal (Eddy et al., 2020), which contains preprocessed gene expression data, biomarkers, and treatment information for 8 out of the 16 immunotherapy cohorts. We downloaded data for 8 of these cohorts from CRI iAtlas. Due to issues in accessing data of all cohorts in COMPASS, we reproduced results for COMPASS using the 8 cohorts we could obtain and used the pretrained model from COMPASS with its default hyperparameters to finetune.

Information on the sample size of the cohorts, the drug used in them, and their publication is given in the Appendix A.1.1. The gene expression data was already normalised to Transcripts Per Million (TPM) units. A binary responder label derived from the labels based on response evaluation criteria in solid tumours (RECIST) was used for classification. BioCOMPASS is finetuned to predict this binary label from gene expression data.

2.2 MODEL ARCHITECTURE

We added biomarker and clinical components on top of the COMPASS architecture to build BioCOMPASS, as shown in Figure 1. The formulae for the components are in Appendix A.2. The implementation is available at https://github.com/hashimsayed0/BioCOMPASS.

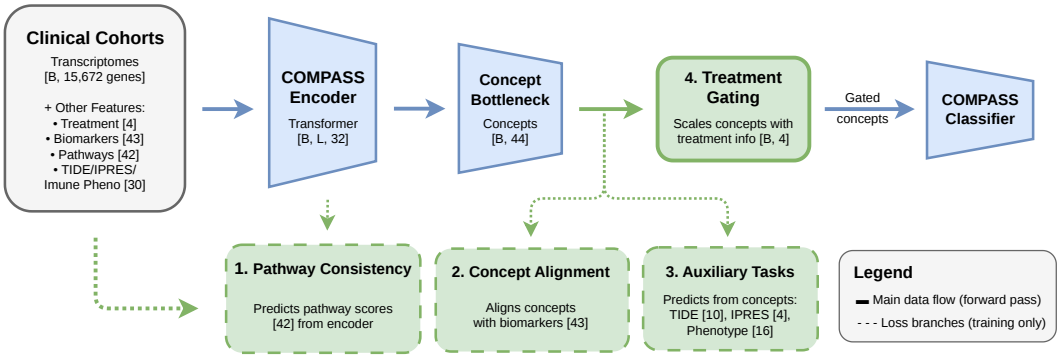

Figure 1: BioCOMPASS architecture: Gene expression data is first fed into the COMPASS encoder to generate embeddings. Minimising the pathway consistency loss makes sure that the pathway scores predicted from embeddings are aligned with external pathway scores. The embeddings are then fed into the COMPASS concept bottleneck to generate 44 biological concepts. These are aligned with cell-type biomarker scores using the concept alignment objective. They are also used to predict immunotherapy response prediction biomarkers such as TIDE & IPRES and other immune phenotypes. The concepts are also scaled based on the specific treatment type using the treatment gating module. The scaled concepts are then used to predict response using a classifier head. Components from COMPASS are in blue colour while BioCOMPASS components are in green.

Pathway Consistency: An auxiliary head containing fully connected layers is trained to predict external pathway activity scores (42 CTLA-4/PD-1 pathway features) from gene embeddings by minimising the mean-squared error (MSE) loss between them. This encourages the encoder to learn representations that are pathway relevant and not cohort-specific noise.

Concept Alignment: Concept alignment involves making the model align learnt concepts (eg. plasma cell, cytotoxic T-cell) and external biomarker scores (e.g., cell type abundances). A projection layer is used to bring them to the same latent dimension. The distance between concept projections and biomarker scores is then minimised.

Auxiliary Tasks: This component involves multi-task learning by predicting established biomarker scores (TIDE, IPRES, and immune phenotypes) from concepts alongside response prediction. This is done through separate decoder heads attached to the concept bottleneck layer. MSE loss between predictions from the auxiliary decoder heads and the actual scores is minimised.

Treatment Gating: Treatment gating scales biological concepts based on the target of immunotherapy treatment (PD-1, CTLA-4, combination). Treatment indicators are embedded and passed through a gating network to compute gate weights, which are then multiplied with the concepts. This allows the model to adaptively focus on concepts relevant to each treatment type and thus integrate treatment information into the concepts.

3 Results

We ran experiments to compare BioCOMPASS with COMPASS. We initialised both models with weights from the COMPASS model pretrained on The Cancer Genome Atlas (TCGA) data (Weinstein et al., 2013) and finetuned them in PFT (partial fine tuning) mode. This mode involves freezing the encoder and only training the concept bottleneck and classifier head. In BioCOMPASS, biomarker data is used during training, but is not required during inference. Each run was done 4 times with 4 different seeds to ensure robustness of results.

Table 1 shows the average performance on the left-out cohort across all 8 cohorts in LOCO setting across 4 seeds. BioCOMPASS excels over COMPASS in all metrics and settings except recall of large cohorts, which could be because BioCOMPASS might be more conservative in its predictions, evident from its higher precision. However, higher ROC-AUC and F1 score show its superior performance. Figure 2 shows the performance on each left-out cohort across 4 seeds. The metrics for COMPASS are obtained by reproducing on the 8 cohorts we could obtain and not all 16 cohorts

Table 1: LOCO validation of COMPASS (C) and BioCOMPASS (BC). This table shows the average performance (in %) across all left-out cohorts. The first two rows show results averaged across all 8 test cohorts; the ones below show the same across 4 small cohorts (less than 50 samples) and 4 large cohorts (more than 50 samples). The 95% confidence intervals (CI) show variation across 4 seeds.

Cohorts	Model	Accuracy	ROC-AUC	F1	Precision	Recall
All	C	63.10 ± 5.43	70.99 ± 2.88	46.65 ± 2.15	48.33 ± 6.24	56.93 ± 6.51
	BC	**70.00 ± 1.76**	**73.58 ± 1.29**	**54.01 ± 2.81**	**56.00 ± 2.64**	**58.55 ± 6.55**
Small (<50)	C	63.03 ± 6.15	72.05 ± 4.30	39.55 ± 3.98	40.82 ± 8.86	51.88 ± 3.91
	BC	**69.77 ± 1.86**	**74.44 ± 2.46**	**52.93 ± 1.67**	**52.51 ± 4.30**	**61.95 ± 6.96**
Large (>50)	C	63.18 ± 6.74	69.94 ± 2.75	53.74 ± 6.33	55.84 ± 6.25	**61.98 ± 15.36**
	BC	**70.24 ± 4.84**	**72.72 ± 0.80**	**55.08 ± 3.96**	**59.48 ± 5.83**	55.14 ± 9.82

as described in Section 2.1. Ablation studies showed that treatment gating is the most influential component, followed by pathway consistency. Results of ablation study are given in Appendix A.3.1. To further evaluate its generalisability, we also ran experiments in Leave-one-cancer-type-out (LOCTO) and Leave-one-treatment-out (LOTO) settings. BioCOMPASS excels over COMPASS in those settings as well as shown in Appendix A.3.2. We also trained logistic regression on biomarker-based baseline methods as well gene expression data. But they do not generalise well as shown in Appendix A.4.

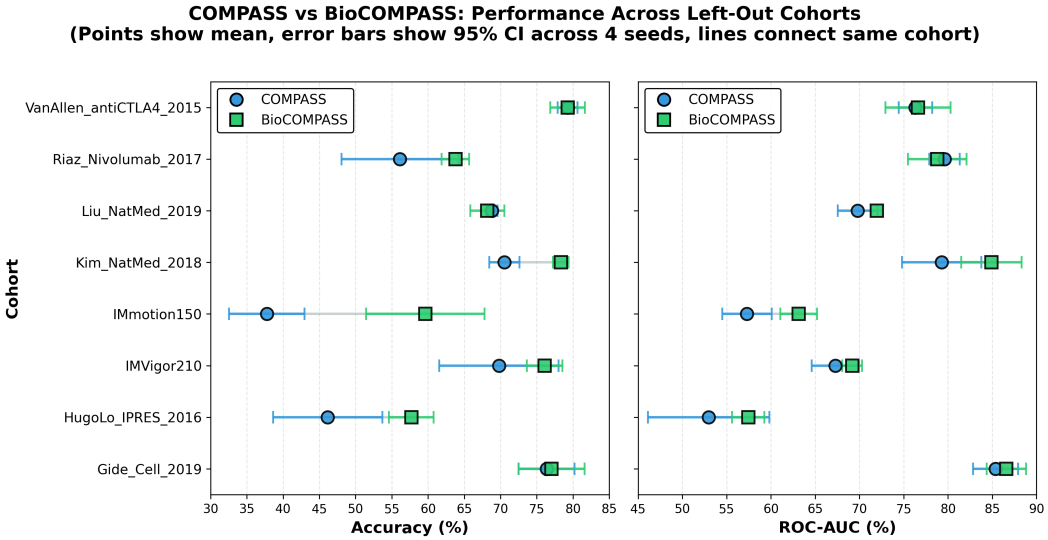

Figure 2: Points show mean performance across four random seeds with 95% CI error bars. Circles: COMPASS; Squares: BioCOMPASS. BioCOMPASS shows consistent improvements in accuracy and ROC-AUC across most cohorts. Please note that the subplots have different scales on x-axis.

4 Discussion

Results show that aligning representations using biomarkers and treatment information in the training process of BioCOMPASS helps it show better generalisation performance over COMPASS in LOCO, LOCTO and LOTO settings. Although the training process of BioCOMPASS requires biomarkers, only treatment information is required during inference. A possible future direction is to exploit domain information about concepts and treatment, such as by using biomedical text mining or biological knowledge graphs, to improve the latent representations. In conclusion, integrating richer clinical and domain knowledge into transformer-based architectures, through informed attention mechanisms or structured latent representation, could further enhance representation learning, robustness, and generalisation across heterogeneous patient cohorts.

5 ACKNOWLEDGMENTS

This work was supported by UKRI AI Centre for Doctoral Training in Safe Artificial Intelligence Systems (SAINTS) (EP/Y030540/1).

REFERENCES

Noam Auslander, Gao Zhang, Joo Sang Lee, Dennie T. Frederick, Benchun Miao, Tabea Moll, Tian Tian, Zhi Wei, Sanna Madan, Ryan J. Sullivan, Genevieve Boland, Keith Flaherty, Meenhard Herlyn, and Eytan Ruppin. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nature Medicine*, 24(10):1545–1549, October 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0157-9. URL https://www.nature.com/articles/s41591-018-0157-9.

Mark Ayers, Jared Lunceford, Michael Nebozhyn, Erin Murphy, Andrey Loboda, David R. Kaufman, Andrew Albright, Jonathan D. Cheng, S. Peter Kang, Veena Shankaran, Sarina A. Piha-Paul, Jennifer Yearley, Tanguy Y. Seiwert, Antoni Ribas, and Terrill K. McClanahan. IFN-**G**–related mRNA profile predicts clinical response to PD-1 blockade. *The Journal of Clinical Investigation*, 127(8):2930–2940, August 2017. ISSN 0021-9738. doi: 10.1172/JCI91190. URL https://www.jci.org/articles/view/91190.

Pei-Ling Chen, Whijae Roh, Alexandre Reuben, Zachary A. Cooper, Christine N. Spencer, Peter A. Prieto, John P. Miller, Roland L. Bassett, Vancheswaran Gopalakrishnan, Khalida Wani, Mariana Petaccia De Macedo, Jacob L. Austin-Breneman, Hong Jiang, Qing Chang, Sangeetha M. Reddy, Wei-Shen Chen, Michael T. Tetzlaff, Russell J. Broaddus, Michael A. Davies, Jeffrey E. Gershenwald, Lauren Haydu, Alexander J. Lazar, Sapna P. Patel, Patrick Hwu, Wen-Jen Hwu, Adi Diab, Isabella C. Glitza, Scott E. Woodman, Luis M. Vence, Ignacio I. Wistuba, Rodabe N. Amaria, Lawrence N. Kwong, Victor Prieto, R. Eric Davis, Wencai Ma, Willem W. Overwijk, Arlene H. Sharpe, Jianhua Hu, P. Andrew Futreal, Jorge Blando, Padmanee Sharma, James P. Allison, Lynda Chin, and Jennifer A. Wargo. 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(8):827–837, August 2016. ISSN 2159-8274, 2159-8290. doi: 10.1158/2159-8290.CD-15-1545. URL https://aacrjournals.org/cancerdiscovery/article/6/8/827/5660/Analysis-of-Immune-Signatures-in-Longitudinal.

Razvan Cristescu, Robin Mogg, Mark Ayers, Andrew Albright, Erin Murphy, Jennifer Yearley, Xinwei Sher, Xiao Qiao Liu, Hongchao Lu, Michael Nebozhyn, Chunsheng Zhang, Jared K. Lunceford, Andrew Joe, Jonathan Cheng, Andrea L. Webber, Nageatte Ibrahim, Elizabeth R. Plimack, Patrick A. Ott, Tanguy Y. Seiwert, Antoni Ribas, Terrill K. McClanahan, Joanne E. Tomassini, Andrey Loboda, and David Kaufman. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade–based immunotherapy. *Science*, 362(6411):eaar3593, October 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar3593. URL https://www.science.org/doi/10.1126/science.aar3593.

Teresa Davoli, Hajime Uno, Eric C. Wooten, and Stephen J. Elledge. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, 355(6322):eaaf8399, January 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaf8399. URL https://www.science.org/doi/10.1126/science.aaf8399.

Charles G. Drake, Evan J. Lipson, and Julie R. Brahmer. Breathing new life into immunotherapy: review of melanoma, lung and kidney cancer. *Nature Reviews Clinical Oncology*, 11(1):24–37, January 2014. ISSN 1759-4782. doi: 10.1038/nrclinonc.2013.208. URL https://www.nature.com/articles/nrclinonc.2013.208.

James A. Eddy, Vésteinn Thorsson, Andrew E. Lamb, David L. Gibbs, Carolina Heimann, Jia Xin Yu, Verena Chung, Yooree Chae, Kristen Dang, Benjamin G. Vincent, Ilya Shmulevich, and Justin Guinney. CRI iAtlas: an interactive portal for immuno-oncology research, August 2020. URL https://f1000research.com/articles/9-1028.

Louis Fehrenbacher, Alexander Spira, Marcus Ballinger, Marcin Kowanetz, Johan Vansteenkiste, Julien Mazieres, Keunchil Park, David Smith, Angel Artal-Cortes, Conrad Lewanski, Fadi Braiteh, Daniel Waterkamp, Pei He, Wei Zou, Daniel S Chen, Jing Yi, Alan Sandler, and Achim Rittmeyer. 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(10030):1837–1846, April 2016. ISSN 0140-6736. doi: 10.1016/S0140-6736(16)00587-0. URL https://www.sciencedirect.com/science/article/pii/S0140673616005870.

Samuel S. Freeman, Moshe Sade-Feldman, Jaegil Kim, Chip Stewart, Anna L. K. Gonye, Arvind Ravi, Monica B. Arniella, Irena Gushterova, Thomas J. LaSalle, Emily M. Blaum, Keren Yizhak, Dennie T. Frederick, Tatyana Sharova, Ignaty Leshchiner, Liudmila Elagina, Oliver G. Spiro, Dimitri Livitz, Daniel Rosebrock, François Aguet, Jian Carrot-Zhang, Gavin Ha, Ziao Lin, Jonathan H. Chen, Michal Barzily-Rokni, Marc R. Hammond, Hans C. Vitzthum von Eckstaedt, Shauna M. Blackmon, Yunxin J. Jiao, Stacey Gabriel, Donald P. Lawrence, Lyn M. Duncan, Anat O. Stemmer-Rachamimov, Jennifer A. Wargo, Keith T. Flaherty, Ryan J. Sullivan, Genevieve M. Boland, Matthew Meyerson, Gad Getz, and Nir Hacohen. Combined tumor and immune signals from genomes or transcriptomes predict outcomes of checkpoint inhibition in melanoma. *Cell Reports Medicine*, 3(2):100500, February 2022. ISSN 2666-3791. doi: 10.1016/j.xcrm.2021.100500. URL https://www.sciencedirect.com/science/article/pii/S2666379121003773.

Edward B. Garon, Naiyer A. Rizvi, Rina Hui, Natasha Leighl, Ani S. Balmanoukian, Joseph Paul Eder, Amita Patnaik, Charu Aggarwal, Matthew Gubens, Leora Horn, Enric Carcereny, Myung-Ju Ahn, Enriqueta Felip, Jong-Seok Lee, Matthew D. Hellmann, Omid Hamid, Jonathan W. Goldman, Jean-Charles Soria, Marisa Dolled-Filhart, Ruth Z. Rutledge, Jin Zhang, Jared K. Lunceford, Reshma Rangwala, Gregory M. Lubiniecki, Charlotte Roach, Kenneth Emancipator, and Leena Gandhi. Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 372(21):2018–2028, May 2015. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1501824. URL http://www.nejm.org/doi/10.1056/NEJMoa1501824.

Tuba N. Gide, Camelia Quek, Alexander M. Menzies, Annie T. Tasker, Ping Shang, Jeff Holst, Jason Madore, Su Yin Lim, Rebecca Velickovic, Matthew Wongchenko, Yibing Yan, Serigne Lo, Matteo S. Carlino, Alexander Guminski, Robyn P.M. Saw, Angel Pang, Helen M. McGuire, Umaimainthan Palendira, John F. Thompson, Helen Rizos, Ines Pires Da Silva, Marcel Batten, Richard A. Scolyer, Georgina V. Long, and James S. Wilmott. Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell*, 35(2):238–255.e6, February 2019. ISSN 15356108. doi: 10.1016/j.ccell.2019.01.003. URL https://linkinghub.elsevier.com/retrieve/pii/S1535610819300376.

Marilyn Giordano, Coralie Henin, Julien Maurizio, Claire Imbratta, Pierre Bourdely, Michel Buferne, Lukas Baitsch, Laurent Vanhille, Michael H. Sieweke, Daniel E. Speiser, Nathalie Auphan-Anezin, Anne-Marie Schmitt-Verhulst, and Grégory Verdeil. Molecular profiling of CD8 T cells in autochthonous melanoma identifies Maf as driver of exhaustion. *The EMBO Journal*, 34(15):2042–2058, August 2015. ISSN 1460-2075. doi: 10.15252/embj.201490786. URL https://doi.org/10.15252/embj.201490786.

Sergei I. Grivennikov, Florian R. Greten, and Michael Karin. Immunity, Inflammation, and Cancer. *Cell*, 140(6):883–899, March 2010. ISSN 0092-8674. doi: 10.1016/j.cell.2010.01.025. URL https://www.sciencedirect.com/science/article/pii/S0092867410000607.

Alexander C. Huang, Robert J. Orlowski, Xiaowei Xu, Rosemarie Mick, Sangeeth M. George, Patrick K. Yan, Sasikanth Manne, Adam A. Kraya, Bradley Wubbenhorst, Liza Dorfman, Kurt D'Andrea, Brandon M. Wenz, Shujing Liu, Lakshmi Chilukuri, Andrew Kozlov, Mary Carberry, Lydia Giles, Melanie W. Kier, Felix Quagliarello, Suzanne McGettigan, Kristin Kreider, Lakshmanan Annamalai, Qing Zhao, Robin Mogg, Wei Xu, Wendy M. Blumenschein, Jennifer H. Yearley, Gerald P. Linette, Ravi K. Amaravadi, Lynn M. Schuchter, Ramin S. Herati, Bertram Bengsch, Katherine L. Nathanson, Michael D. Farwell, Giorgos C. Karakousis, E. John Wherry, and Tara C. Mitchell. A single dose of neoadjuvant PD-1 blockade predicts clinical

outcomes in resectable melanoma. *Nature Medicine*, 25(3):454–461, March 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0357-y. URL https://www.nature.com/articles/s41591-019-0357-y.

Willy Hugo, Jesse M. Zaretsky, Lu Sun, Chunying Song, Blanca Homet Moreno, Siwen Hu-Lieskovan, Beata Berent-Maoz, Jia Pang, Bartosz Chmielowski, Grace Cherry, Elizabeth Seja, Shirley Lomeli, Xiangju Kong, Mark C. Kelley, Jeffrey A. Sosman, Douglas B. Johnson, Antoni Ribas, and Roger S. Lo. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, 165(1):35–44, March 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.02.065. URL https://www.sciencedirect.com/science/article/pii/S009286741630215X.

Peng Jiang, Shengqing Gu, Deng Pan, Jingxin Fu, Avinash Sahu, Xihao Hu, Ziyi Li, Nicole Traugh, Xia Bu, Bo Li, Jun Liu, Gordon J. Freeman, Myles A. Brown, Kai W. Wucherpfennig, and X. Shirley Liu. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nature Medicine*, 24(10):1550–1558, October 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0136-1. URL https://www.nature.com/articles/s41591-018-0136-1.

Johanna A. Joyce and Douglas T. Fearon. T cell exclusion, immune privilege, and the tumor microenvironment. *Science*, 348(6230):74–80, April 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa6204. URL https://www.science.org/doi/10.1126/science.aaa6204.

Seung Tae Kim, Razvan Cristescu, Adam J. Bass, Kyoung-Mee Kim, Justin I. Odegaard, Kyung Kim, Xiao Qiao Liu, Xinwei Sher, Hun Jung, Mijin Lee, Sujin Lee, Se Hoon Park, Joon Oh Park, Young Suk Park, Ho Yeong Lim, Hyuk Lee, Mingew Choi, AmirAli Talasaz, Peter Soonmo Kang, Jonathan Cheng, Andrey Loboda, Jeeyun Lee, and Won Ki Kang. Comprehensive molecular characterization of clinical responses to PD-1 inhibition in metastatic gastric cancer. *Nature Medicine*, 24(9):1449–1458, September 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0101-z.

JungHo Kong, Doyeon Ha, Juhun Lee, Inhae Kim, Minhyuk Park, Sin-Hyeog Im, Kunyoo Shin, and Sanguk Kim. Network-based machine learning approach to predict immunotherapy response in cancer patients. *Nature Communications*, 13(1):3703, June 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31535-6. URL https://www.nature.com/articles/s41467-022-31535-6.

Dana R. Leach, Matthew F. Krummel, and James P. Allison. Enhancement of Antitumor Immunity by CTLA-4 Blockade. *Science*, 271(5256):1734–1736, March 1996. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.271.5256.1734. URL https://www.science.org/doi/10.1126/science.271.5256.1734.

Yawei Li, Xin Wu, Deyu Fang, and Yuan Luo. Informing immunotherapy with multi-omics driven machine learning. *npj Digital Medicine*, 7(1):67, March 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01043-6. URL https://www.nature.com/articles/s41746-024-01043-6.

David Liu, Bastian Schilling, Derek Liu, Antje Sucker, Elisabeth Livingstone, Livnat Jerby-Arnon, Lisa Zimmer, Ralf Gutzmer, Imke Satzger, Carmen Loquai, Stephan Grabbe, Natalie Vokes, Claire A. Margolis, Jake Conway, Meng Xiao He, Haitham Elmarakeby, Felix Dietlein, Diana Miao, Adam Tracy, Helen Gogas, Simone M. Goldinger, Jochen Utikal, Christian U. Blank, Ricarda Rauschenberg, Dagmar von Bubnoff, Angela Krackhardt, Benjamin Weide, Sebastian Haferkamp, Felix Kiecker, Ben Izar, Levi Garraway, Aviv Regev, Keith Flaherty, Annette Paschen, Eliezer M. Van Allen, and Dirk Schadendorf. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nature Medicine*, 25(12):1916–1927, December 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0654-5. URL https://www.nature.com/articles/s41591-019-0654-5.

Sanjeev Mariathasan, Shannon J. Turley, Dorothee Nickles, Alessandra Castiglioni, Kobe Yuen, Yulei Wang, Edward E. Kadel III, Hartmut Koeppen, Jillian L. Astarita, Rafael Cubas, Suchit Jhunjhunwala, Romain Banchereau, Yagai Yang, Yinghui Guan, Cecile Chalouni, James Ziai,

Yasin Şenbabaoğlu, Stephen Santoro, Daniel Sheinson, Jeffrey Hung, Jennifer M. Giltnane, Andrew A. Pierce, Kathryn Mesh, Steve Lianoglou, Johannes Riegler, Richard A. D. Carano, Pontus Eriksson, Mattias Höglund, Loan Somarriba, Daniel L. Halligan, Michiel S. van der Heijden, Yohann Loriot, Jonathan E. Rosenberg, Lawrence Fong, Ira Mellman, Daniel S. Chen, Marjorie Green, Christina Derleth, Gregg D. Fine, Priti S. Hegde, Richard Bourgon, and Thomas Powles. TGFB attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature*, 554(7693):544–548, February 2018. ISSN 1476-4687. doi: 10.1038/nature25501. URL https://www.nature.com/articles/nature25501.

David F. McDermott, Mahrukh A. Huseni, Michael B. Atkins, Robert J. Motzer, Brian I. Rini, Bernard Escudier, Lawrence Fong, Richard W. Joseph, Sumanta K. Pal, James A. Reeves, Mario Sznol, John Hainsworth, W. Kimryn Rathmell, Walter M. Stadler, Thomas Hutson, Martin E. Gore, Alain Ravaud, Sergio Bracarda, Cristina Suárez, Riccardo Danielli, Viktor Gruenwald, Toni K. Choueiri, Dorothee Nickles, Suchit Jhunjhunwala, Elisabeth Piault-Louis, Alpa Thobhani, Jiaheng Qiu, Daniel S. Chen, Priti S. Hegde, Christina Schiff, Gregg D. Fine, and Thomas Powles. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nature Medicine*, 24(6): 749–757, June 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0053-3. URL https://www.nature.com/articles/s41591-018-0053-3.

Jane L. Messina, David A. Fenstermacher, Steven Eschrich, Xiaotao Qu, Anders E. Berglund, Mark C. Lloyd, Michael J. Schell, Vernon K. Sondak, Jeffrey S. Weber, and James J. Mulé. 12-Chemokine Gene Signature Identifies Lymph Node-like Structures in Melanoma: Potential for Patient Selection for Immunotherapy? *Scientific Reports*, 2(1):765, October 2012. ISSN 2045-2322. doi: 10.1038/srep00765. URL https://www.nature.com/articles/srep00765.

Martin Nurmik, Pit Ullmann, Fabien Rodriguez, Serge Haan, and Elisabeth Letellier. In search of definitions: Cancer-associated fibroblasts and their markers. *International Journal of Cancer*, 146(4):895–905, February 2020. ISSN 0020-7136, 1097-0215. doi: 10.1002/ijc.32193. URL https://onlinelibrary.wiley.com/doi/10.1002/ijc.32193.

Gabriel A. Rabinovich, Dmitry Gabrilovich, and Eduardo M. Sotomayor. Immunosuppressive Strategies that are Mediated by Tumor Cells. *Annual Review of Immunology*, 25(1):267–296, April 2007. ISSN 0732-0582, 1545-3278. doi: 10.1146/annurev.immunol.25.022106.141609. URL https://www.annualreviews.org/doi/10.1146/annurev.immunol.25.022106.141609.

Nadeem Riaz, Jonathan J. Havel, Vladimir Makarov, Alexis Desrichard, Walter J. Urba, Jennifer S. Sims, F. Stephen Hodi, Salvador Martín-Algarra, Rajarsi Mandal, William H. Sharfman, Shailender Bhatia, Wen-Jen Hwu, Thomas F. Gajewski, Craig L. Slingluff, Diego Chowell, Sviatoslav M. Kendall, Han Chang, Rachna Shah, Fengshen Kuo, Luc G. T. Morris, John-William Sidhom, Jonathan P. Schneck, Christine E. Horak, Nils Weinhold, and Timothy A. Chan. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, 171(4):934–949.e16, November 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.09.028. URL https://www.sciencedirect.com/science/article/pii/S0092867417311224.

Naiyer A. Rizvi, Matthew D. Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J. Havel, William Lee, Jianda Yuan, Phillip Wong, Teresa S. Ho, Martin L. Miller, Natasha Rekhtman, Andre L. Moreira, Fawzia Ibrahim, Cameron Bruggeman, Billel Gasmi, Roberta Zappasodi, Yuka Maeda, Chris Sander, Edward B. Garon, Taha Merghoub, Jedd D. Wolchok, Ton N. Schumacher, and Timothy A. Chan. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science*, 348(6230):124–128, April 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaa1348. URL https://www.science.org/doi/10.1126/science.aaa1348.

Whijae Roh, Pei-Ling Chen, Alexandre Reuben, Christine N. Spencer, Peter A. Prieto, John P. Miller, Vancheswaran Gopalakrishnan, Feng Wang, Zachary A. Cooper, Sangeetha M. Reddy, Curtis Gumbs, Latasha Little, Qing Chang, Wei-Shen Chen, Khalida Wani, Mariana Petaccia De Macedo, Eveline Chen, Jacob L. Austin-Breneman, Hong Jiang, Jason Roszik, Michael T. Tetzlaff, Michael A. Davies, Jeffrey E. Gershenwald, Hussein Tawbi, Alexander J. Lazar, Patrick Hwu, Wen-Jen Hwu, Adi Diab, Isabella C. Glitza, Sapna P. Patel, Scott E. Woodman, Rodabe N.

Amaria, Victor G. Prieto, Jianhua Hu, Padmanee Sharma, James P. Allison, Lynda Chin, Jianhua Zhang, Jennifer A. Wargo, and P. Andrew Futreal. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Science Translational Medicine*, 9(379):eaah3560, March 2017. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.aah3560. URL https://www.science.org/doi/10.1126/scitranslmed.aah3560.

Michael S. Rooney, Sachet A. Shukla, Catherine J. Wu, Gad Getz, and Nir Hacohen. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, 160 (1):48–61, January 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2014.12.033. URL https://www.sciencedirect.com/science/article/pii/S0092867414016390.

Wanxiang Shen, Thinh H. Nguyen, Michelle M. Li, Yepeng Huang, Intae Moon, Nitya Nair, Daniel Marbach, and Marinka Zitnik. Generalizable AI predicts immunotherapy outcomes across cancers and treatments, May 2025. URL http://medrxiv.org/lookup/doi/10.1101/2025.05.01.25326820.

Yannick Simoni, Etienne Becht, Michael Fehlings, Chiew Yee Loh, Si-Lin Koo, Karen Wei Weng Teng, Joe Poh Sheng Yeong, Rahul Nahar, Tong Zhang, Hassen Kared, Kaibo Duan, Nicholas Ang, Michael Poidinger, Yin Yeng Lee, Anis Larbi, Alexis J. Khng, Emile Tan, Cherylin Fu, Ronnie Mathew, Melissa Teo, Wan Teck Lim, Chee Keong Toh, Boon-Hean Ong, Tina Koh, Axel M. Hillmer, Angela Takano, Tony Kiat Hon Lim, Eng Huat Tan, Weiwei Zhai, Daniel S. W. Tan, Iain Beehuat Tan, and Evan W. Newell. Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature*, 557(7706):575–579, May 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0130-2. URL https://www.nature.com/articles/s41586-018-0130-2.

Eliezer M. Van Allen, Diana Miao, Bastian Schilling, Sachet A. Shukla, Christian Blank, Lisa Zimmer, Antje Sucker, Uwe Hillen, Marnix H. Geukes Foppen, Simone M. Goldinger, Jochen Utikal, Jessica C. Hassel, Benjamin Weide, Katharina C. Kaehler, Carmen Loquai, Peter Mohr, Ralf Gutzmer, Reinhard Dummer, Stacey Gabriel, Catherine J. Wu, Dirk Schadendorf, and Levi A. Garraway. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, 350(6257):207–211, October 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aad0095. URL https://www.science.org/doi/10.1126/science.aad0095.

John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics*, 45(10):1113–1120, October 2013. ISSN 1061-4036. doi: 10.1038/ng.2764. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3919969/.

Chia-Chin Wu, Y. Alan Wang, J. Andrew Livingston, Jianhua Zhang, and P. Andrew Futreal. Prediction of biomarkers and therapeutic combinations for anti-PD-1 immunotherapy using the global gene network association. *Nature Communications*, 13(1):42, January 2022. ISSN 2041-1723. doi: 10.1038/s41467-021-27651-4. URL https://www.nature.com/articles/s41467-021-27651-4.

A Appendix

A.1 Data

A.1.1 Cohorts

The cohorts used for LOCO evaluation are given in Table 2.

A.1.2 Preprocessing

The following preprocessing was applied to gene expression data on CRI iAtlas. ENST counts were generated by trimming FASTQ reads with TrimGalore (v0.6.2), aligning them to GRCh38 (gtf: v103; ref: p13) using STAR (v2.7.0f), and performing quantification with Salmon (v1.1.0)

Table 2: Cohorts used for LOCO evaluation along with the cancer type, drug used, number of samples and size category as well as a citation to the publication.

Cohort	Cancer Type	Drug	Number of Samples	Size Category
Mariathasan et al. (2018)	BLCA	Atezolizumab	298	Large (>50)
McDermott et al. (2018)	KIRC	Atezolizumab	247	Large (>50)
Liu et al. (2019)	SKCM	Nivolumab	121	Large (>50)
Gide et al. (2019)	SKCM	Nivolumab	73	Large (>50)
Riaz et al. (2017)	SKCM	Nivolumab	49	Small (<50)
Kim et al. (2018)	STAD	Pembrolizumab	45	Small (<50)
Van Allen et al. (2015)	SKCM	Ipilimumab	41	Small (<50)
Hugo et al. (2016)	SKCM	Pembrolizumab	26	Small (<50)

A.2 Model

A.2.1 Pathway Consistency

$$\mathcal{L}_{\text{pathway}} = \frac{1}{B}\sum_{i=1}^{B} \|\mathbf{p}_i - f_{\text{path}}(\text{mean}(\mathbf{E}_i))\|_2^2 \tag{1}$$

where $\mathbf{E}i \in \mathbb{R}^{L\times d_e}$ is the gene encoding for sample i, L is the number of gene tokens, d_e is the dimension of gene encoding, f_{path} is the pathway predictor head, $\mathbf{p}_i \in \mathbb{R}^{42}$ are target pathway scores and B is the batch size.

A.2.2 Concept Alignment

$$\mathcal{L}_{\text{align}} = \|\mathbf{C}W - \mathbf{B}\|_2^2 \tag{2}$$

where $\mathbf{C} \in \mathbb{R}^{B\times 44}$ are concepts, $W \in \mathbb{R}^{44\times d_b}$ is a learnable projection, $\mathbf{B} \in \mathbb{R}^{B\times d_b}$ are biomarkers, B is the batch size and d_b is the number of biomarkers. 44 is the number of biological concepts generated by COMPASS concept bottleneck.

A.2.3 Auxiliary Tasks

$$\mathcal{L}_{\text{aux}} = \sum_{k\in\{\text{TIDE,IPRES,pheno}\}} \frac{1}{B}\sum_{i=1}^{B} \|\mathbf{t}_i^k - f_k(\mathbf{c}_i)\|_2^2 \tag{3}$$

where f_k is the auxiliary decoder head for task k, $\mathbf{c}_i \in \mathbb{R}^{44}$ are concepts, and $\mathbf{t}_i^k$ are target scores for task k.

A.2.4 Treatment Gating

$$\mathbf{g} = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot \mathbf{e}_t + b_1) + b_2) \quad \mathbf{c}' = \mathbf{c} \odot \mathbf{g} \tag{4}$$

where $\mathbf{e}_t \in \mathbb{R}^{d_h}$ is the treatment embedding, $\mathbf{g} \in \mathbb{R}^{44}$ are gating weights, $\mathbf{c} \in \mathbb{R}^{44}$ are biological concepts, and $\odot$ denotes element-wise multiplication.

A.3 Results

A.3.1 Ablation Study

We conducted an ablation study to understand how generalisability changes with the absence of one component at a time and hence find out the most influential components.

Table 3: Ablation study of various components in BioCOMPASS. Each row shows performance when one component is disabled. Values are mean ± 95% CI margin across five random seeds across all cohorts in LOCO setting. Treatment gating contributes most to performance improvement, followed by pathway consistency. Lower is better here as it shows that taking way that component reduced performance the most.

Config	Acc (%)	AUC (%)	F1 (%)	Prec (%)	Recall (%)
No Gating	**72.13 ± 3.32**	**64.16 ± 4.52**	**48.23 ± 5.07**	**51.07 ± 6.48**	58.65 ± 8.37
No Pathway	72.24 ± 2.88	67.53 ± 3.51	52.05 ± 3.74	53.96 ± 5.88	**57.01 ± 5.52**
No Auxiliary	72.55 ± 3.13	68.32 ± 3.06	51.94 ± 4.04	53.58 ± 5.36	57.10 ± 6.15
No Alignment	72.52 ± 3.02	68.09 ± 3.48	52.50 ± 4.20	54.78 ± 5.76	58.83 ± 6.35

A.3.2 Additional generalisability experiments

The results of additional experiments in LOCTO and LOTO settings are in Table 4. The LOCTO setting involves leaving one of four types of cancer at a time: bladder urothelial carcinoma (BLCA), renal clear cell carcinoma (KIRC), cutaneous Cutaneous melanoma (SKCM), and Stomach adenocarcinoma (STAD). The LOTO setting includes leaving out one of four immunotherapy treatment targets: PD-1, PD-L1, CTLA-4 and CTLA-4 + PD-1. BioCOMPASS performs better in both settings in all metrics except recall which could be because BioCOMPASS might be more conservative in its predictions, as explained earlier. Figures 3 and 4 show average performance on each left-out cancer type and treatment target respectively across 4 seeds.

Table 4: Performance of COMPASS (C) and BioCOMPASS (BC) in LOCTO and LOTO settings. This table shows the average performance (in %) across 4 left-out cancer types (BLCA, KIRC, SKCM, STAD) and 4 immunotherapy treatment targets (PD-1, PD-L1, CTLA-4, CTLA-4 + PD-1). The 95% confidence intervals (CI) show variation across 4 seeds.

Setting	Model	Accuracy	ROC AUC	F1	Precision	Recall
LOCTO	C	61.69 ± 6.99	67.72 ± 3.87	42.41 ± 5.40	42.61 ± 13.69	**51.50 ± 8.44**
	BC	**70.05 ± 1.63**	**71.65 ± 3.02**	**49.16 ± 5.72**	**51.66 ± 2.38**	50.37 ± 10.22
LOTO	C	68.93 ± 2.95	74.49 ± 1.56	57.02 ± 2.16	55.44 ± 2.04	**63.53 ± 4.86**
	BC	**73.49 ± 1.36**	**76.85 ± 3.56**	**60.53 ± 2.66**	**61.01 ± 1.84**	63.36 ± 4.28

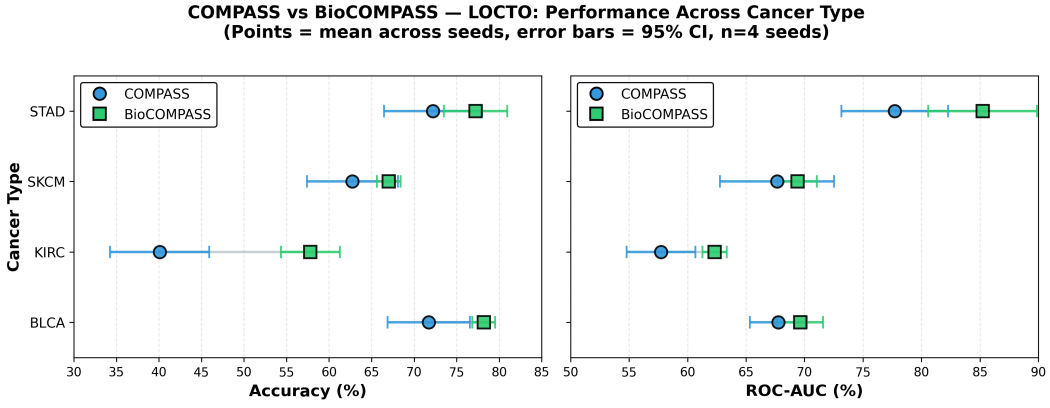

Figure 3: Points show mean performance across four random seeds with 95% CI error bars. Circles: COMPASS; Squares: BioCOMPASS. Please note that the subplots have different scales on x-axis.

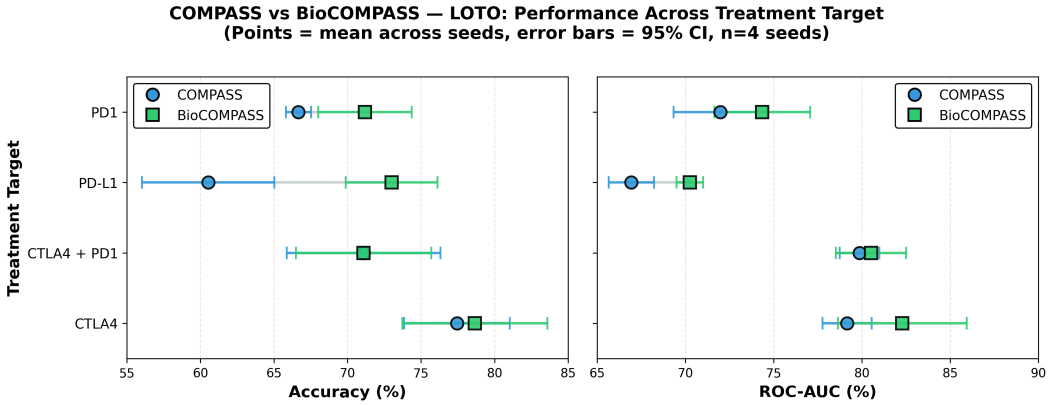

Figure 4: Points show mean performance across four random seeds with 95% CI error bars. Circles: COMPASS; Squares: BioCOMPASS. Please note that the subplots have different scales on x-axis.

A.4 Baseline Methods

We trained a logistic regression model on biomarker-based baseline methods to analyse their generalisability. The implementation from COMPASS was used for this purpose and are described in Table 5. We also trained logistic regression and other standard machine learning methods on gene expression data and biomarker data. Results are in Table 6. As can be seen from the results, these methods fail to generalise across unseen test groups.

A.5 Glossary

A.5.1 Immunotherapy & Treatment

PD-1 (Programmed Death 1) A checkpoint protein that regulates immune responses; target for immunotherapy.

PD-L1 (Programmed Death-Ligand 1) A protein that binds to PD-1; its expression level is used as a biomarker.

CTLA-4 (Cytotoxic T Lymphocyte-Associated Protein 4) An immune checkpoint protein that downregulates immune responses; target for immunotherapy.

Anti-PD-1 therapy Treatment that blocks PD-1 (e.g., Nivolumab, Pembrolizumab).

Anti-CTLA-4 therapy Treatment that blocks CTLA-4 (e.g., Ipilimumab).

Atezolizumab An anti-PD-L1 immunotherapy drug.

Nivolumab An anti-PD-1 immunotherapy drug.

Pembrolizumab An anti-PD-1 immunotherapy drug.

Ipilimumab An anti-CTLA-4 immunotherapy drug.

A.5.2 Cell Types & Immune Components

Cytotoxic T-cell (Cytotoxic T lymphocyte) Immune cells that kill cancer cells.

Plasma cell B cells that produce antibodies.

Immune cell infiltration The presence of immune cells within the tumour.

A.5.3 Biomarkers & Scoring Methods

TIDE Tumour Immune Dysfunction and Exclusion score; biomarker for immunotherapy response.

IPRES Innate anti-PD-1 Resistance signature.

Immune phenotypes Classifications of tumours based on immune cell composition.

Table 5: Description of biomarker-based baseline methods. These were used based on the implementation by COMPASS.

Method	Description	Reference
GeneBio	Combined score of immunotherapy target markers PD1/PDL1/CTLA4	Kong et al. (2022)
CTLA4	Expression of CTLA4 as a single ICI target marker	Kong et al. (2022)
PD1	Expression of PDCD1 as a single ICI target marker	Kong et al. (2022)
PDL1	Expression of CD274 as a single ICI target marker	Kong et al. (2022)
CD8	$CD8^+$ T cell score derived from average expression of CD8A and CD8B	Chen et al. (2016) Kong et al. (2022)
CIS	Cytotoxic immune signature score averaging cytotoxic immune genes	Davoli et al. (2017)
Teff	T-effector/IFN-γ signature score averaging T-effector genes	Fehrenbacher et al. (2016)
PGM	Prognostic gene-pair model; top pair:lymphocyte MAP4K1 and tumor TBX3	Freeman et al. (2022)
NRS	Neoadjuvant response signature score averaging NRS gene expression	Huang et al. (2019)
IFNG	IFN-γ response score based on an 18-gene signature	Ayers et al. (2017)
IMPRES	Sum of expression ratios across 15 immune/checkpoint gene pairs	Auslander et al. (2018)
TIDE	Tumor Immune Dysfunction and Exclusion composite score	Jiang et al. (2018)
CTL	Cytotoxic T lymphocyte score averaging CTL signature gene expression	Jiang et al. (2018)
TAM	Tumor-associated macrophage score averaging TAM signature genes	Joyce & Fearon (2015) Jiang et al. (2018)
Texh	T-cell exhaustion score averaging exhaustion signature genes	Giordano et al. (2015) Jiang et al. (2018)
CKS	12-chemokine signature score using PC1 of chemokine gene expression	Messina et al. (2012)
CAF	Cancer-associated fibroblast score averaging CAF signature genes	Nurmik et al. (2020)
IS	Immune score averaging expression of a panel of immune genes	Roh et al. (2017)
ICA	Immune cytolytic activity score based on GZMA and PRF1 expression	Rooney et al. (2015)
MIAS	MHC-I association immune score computed via ssGSEA on 100 genes	Wu et al. (2022)
GEP	T cell-inflamed gene expression profile score via ssGSEA on 18 genes	Cristescu et al. (2018) Wu et al. (2022)
NetBio	Score derived from the top 200 ICI target-proximal network genes	Kong et al. (2022)

Table 6: Performance of baseline methods in LOCO (C), LOCTO (CT) and LOTO (T) settings. It can be seen that all baseline methods show poor generalisability. Immune score based methods were used based on the implementation by COMPASS. LR: Logistic regression, GBM: Gradient boosting machine, RF: Random forest, PCA: Principal component analysis

Method	**AUC (%)**			**Accuracy (%)**			**F1 (%)**		
	C	**CT**	**T**	**C**	**CT**	**T**	**C**	**CT**	**T**
Logistic regression on baseline methods									
CKS	61.88	63.96	63.05	57.70	59.99	59.30	48.23	**47.95**	48.78
GEP	**62.83**	65.50	**65.42**	57.27	57.05	60.02	**48.30**	46.60	47.77
IFNG	61.97	64.82	62.89	65.57	68.51	65.23	38.85	38.36	39.31
CD8	61.13	64.39	61.99	57.27	59.74	62.65	44.61	42.89	48.80
Teff	62.39	**66.06**	63.98	56.71	60.33	63.70	44.04	46.35	39.79
IS	62.54	64.89	62.39	58.73	62.16	60.57	40.46	42.62	43.64
ICA	60.40	63.23	62.41	60.45	63.09	63.95	40.00	41.96	41.50
PDL1	59.72	62.92	59.71	59.24	60.59	58.40	42.13	46.94	45.76
CTL	61.21	64.35	63.61	60.83	64.95	62.16	37.55	39.25	35.84
CIS	60.81	62.98	61.59	64.56	67.81	53.65	36.12	40.75	41.41
GeneBio	60.75	64.20	61.57	57.31	62.69	61.59	39.04	41.11	40.67
CTLA4	59.53	60.37	62.16	58.19	57.07	55.69	46.12	40.73	**48.96**
PD1	60.43	60.93	59.13	56.79	56.67	58.70	45.00	42.37	47.32
MIAS	61.09	62.22	60.55	62.98	66.46	54.25	38.08	41.80	39.30
TAM	55.62	58.87	48.53	56.28	64.40	54.32	44.67	40.52	45.67
NRS	56.96	58.37	61.28	56.93	58.92	52.64	37.35	41.81	40.51
PGM	59.27	57.42	58.50	57.64	63.07	58.48	35.59	29.66	43.15
IMPRES	56.01	52.20	57.19	54.41	52.39	53.73	32.98	36.74	41.63
NetBio	56.36	50.01	53.98	56.52	45.41	53.71	30.58	33.69	39.69
Texh	49.14	48.05	45.56	43.67	45.06	46.20	45.93	40.51	48.89
CAF	54.45	56.73	41.94	45.04	41.09	42.10	43.24	41.22	39.56
TIDE	47.61	53.46	41.63	59.48	62.18	46.90	29.81	30.54	27.57
Standard ML models on biomarkers									
LR	59.61	61.85	57.14	58.95	62.85	54.26	40.57	39.04	41.27
GBM	53.04	52.44	60.10	55.96	61.39	60.97	33.28	31.01	40.75
RF	58.30	58.18	59.75	65.68	70.27	65.17	6.10	8.23	21.38
Standard ML models on gene expression									
PCA + LR	59.45	59.82	60.84	59.70	65.85	53.39	33.18	24.74	46.35
PCA + GBM	55.07	52.82	62.64	60.43	65.80	63.27	30.63	20.88	36.75
PCA + RF	55.60	54.22	58.20	**66.50**	**70.55**	**65.67**	0.95	0.40	9.95
LR	55.32	57.58	54.49	60.63	62.25	62.75	24.18	28.91	33.52
RF	56.67	57.82	55.47	65.62	70.10	63.83	2.29	2.18	12.33

Cell type abundances Quantification of different immune cell populations in tumours.

Pathway activity scores Measurements of biological pathway activation (e.g., CTLA-4/PD-1 pathways).