A foundation model of cancer genotype enables precise predictions of therapeutic response

JungHo Kong[1], Ingoo Lee[1], Dean Boecher[2], Akshat Singhal[1], Marcus R. Kelly[1], Jimin Moon[3], Chang Ho Ahn[3], Chan-Young Ock[3], Dexter Pratt[1], Tannavee Kumar[4], Timothy J. Sears[5], David Laub[5], Sarah Wright[1], Patrick Wall[2], Hannah Carter[1], Zhen Wang[6*] and Trey Ideker[1,2,7,8*]

[1] Department of Medicine, School of Medicine, University of California, San Diego, USA

[2] Department of Bioengineering, University of California, San Diego, USA

[3] Lunit Incorporated, Seoul, South Korea

[4] Department of Biostatistics, University of California, San Diego, USA

[5] Bioinformatics and Systems Biology Program, University of California, San Diego, USA

[6] Halicioglu Data Science Institute, University of California, San Diego, USA

[7] Department of Computer Science and Engineering, University of California, San Diego, USA

[8] Big Data Institute, University of Oxford, Oxford, United Kingdom

* Corresponding Authors: Zhen Wang, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0688. E-mail: zhw085@ucsd.edu, Phone number: +1 (614) 805-9913; Trey Ideker, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0688. E-mail: tideker@ucsd.edu, Phone number: +1 (858) 472-3421

Running title

A tumor genome foundation model for precision oncology.

Conflict-of-interest

T.I. is a co-founder, member of the advisory board, and has an equity interest in Data4Cure and Serinus Biosciences. T.I. is also a consultant for and has an equity interest in Ideaya Biosciences and Eikon Therapeutics. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict-of-interest policies. J.M. and C.A. are employees at Lunit. C.O. holds a leadership role and is a stockholder at Lunit.

Abstract

While genetic sequencing is routine in cancer care, translating a tumor's complex mutation profile into actionable treatment decisions remains a central challenge. MutationProjector is pre-trained from a large corpus of genomic alterations across 30,000+ tumors, integrated with extensive molecular knowledge. The resulting projection reveals a tumor's altered molecular pathways, facilitating model interpretation, and it accurately reconstructs held-out mutations, demonstrating model generalization. When applied to predict immunotherapy or chemotherapy resistance across multiple cancer types and cohorts, MutationProjector achieves or exceeds state-of-the-art performance in all contexts. It identifies unexpected biomarkers, including *KMT2D* mutation in immunotherapy sensitivity and joint alteration of *SMARCA4* and *STK11* in immunotherapy resistance. These results establish a unifying framework for connecting tumor genotypes to biological mechanisms and therapeutic outcomes.

Statement of significance

This work describes MutationProjector, a tumor genomic foundation model pre-trained from 30,000+ tumors. It maps mutations to biological states for diagnosis and treatment, achieving state-of-the-art performance in predicting therapy resistance and identifying biomarkers such as *KMT2D* (sensitivity) and *SMARCA4/STK11* co-alterations (resistance).

Introduction

Tumor genetic profiling is a cornerstone of precision oncology (1), enabling patient assessment and treatment decisions across a range of cancer types (2). For this purpose, DNA sequencing panels — and in particular those that broadly identify alterations in cancer-associated genes — have been widely adopted in the clinic due to their relatively low cost, rapid turnaround, and established relevance to treatment outcomes (1,3). Some of the most common clinical gene panels include MSK-IMPACT (4,5) (400+ cancer-associated genes), FoundationOne CDx (6) (324 genes), Tempus xT (7) (648 genes), Thermo Fisher Oncomine (8) (161 genes), Caris Molecular Intelligence (>700 genes) (9), and Guardant360 (70 – 500+ genes depending on version), with one or more of these tests deployed by the vast majority of oncologists and cancer centers (10). As of 2025, more than 60 different Federal Drug Administration (FDA)-approved therapies employ cancer gene sequencing as a companion diagnostic. For example, mutation in the *EGFR* gene indicates treatment with gefitinib, erlotinib and others; the BRAF V600E protein variant indicates treatment with vemurafenib or sorafenib, and high tumor mutation burden (TMB) can indicate treatment with anti-PD1/PDL1 immunotherapies.

Despite such progress, the information from genetic sequencing that is clinically actionable remains limited to a few well-studied genes/biomarkers within specific tumor types or therapeutic contexts. A match is made to an FDA-approved targeted therapy in only about 8% of cases currently (11), usually on the basis of alteration in a single gene. While this situation may reflect the incomplete scope of genes covered by current sequencing panels, it clearly also reflects a fundamental lack of knowledge about how gene mutations should be interpreted. The average tumor has approximately 11 distinct genetic alterations identified by clinical sequencing (12,13), a potentially rich source of molecular information if it could be tapped for therapy selection. One

reason that cancer mutations have been difficult to associate with treatment outcomes is that most mutation events are rare (14). Another is that individual biomarkers do not function in isolation but act combinatorially to influence a drug response. For instance, while mutations in *ERCC2* (Excision Repair Cross-Complementation Group 2) have been associated with cisplatin sensitivity (15–17), this effect varies depending on mutations to other DNA repair genes including *BRCA2*, *ATM*, *RB1* and *FANCC* (18,19).

At least two strategies can be applied to address these challenges. First are approaches that integrate cancer mutations with knowledge of molecular networks, since rare and/or co-occurring genetic alteration events are often interconnected within common hallmark pathways (14,20). Although network analysis is generally based on a single network or pathway resource (21), it has been shown repeatedly that integration of multiple networks improves prediction performance in various biological tasks (21–26). This improvement can be largely explained by the importance of different types of molecular interactions in tumor pathogenesis and drug response (27,28), including physical protein-protein binding, protein-DNA transcriptomic regulation and transient signaling modifications including phosphorylation and ubiquitination.

A second approach is to leverage recent developments in artificial intelligence (AI), which present exciting opportunities to harness large-scale cancer genomics data for broader, multi-purpose clinical applications like tumor subtyping and drug response prediction. Foundation models, which are pre-trained on large datasets and then applied to solve diverse new challenges with relatively few samples (29), are especially well positioned to advance precision oncology. Such modeling has been enormously successful in natural language processing (e.g. GPT-5) and computer vision (29) and is being actively investigated in life sciences research, including models that analyze DNA (30,31), RNA (32–36), or protein (37–39) sequences, or microscopy and

radiology images (40–44). In oncology in particular, collections of histopathology slides, gene expression profiles, or tumor genomes have been used to train foundation models that were subsequently used to predict somatic mutations (40,45), cancer type (40,45), metastatic outcomes (46), survival (46) or immunotherapy response (47,medRxiv 2025.05.01.25326820).

Here we describe MutationProjector, a foundation model for decoding a cancer's complex landscape of genetic alterations. This model is trained from the genetic profiles of more than 30,000 tumor samples gathered across a spectrum of different cancer types (Fig. 1A), integrated with knowledge of cancer molecular networks (Fig. 1B). Once trained, MutationProjector translates the genome of a tumor into quantitative coordinates representing its altered biological state, enabling broad downstream clinical applications.

Results

Comprehensive curation of tumor genetic data and molecular networks

We accessed tumor genetic alterations from Project GENIE (48) and The Cancer Genome Atlas (TCGA) (49), encompassing 30,328 solid cancers of 10 distinct types (Fig. 1A). From these samples we extracted genetic alterations for 468 genes represented on clinical gene panels, for each gene considering the presence of somatic mutation, copy number amplification or copy number deletion. In addition to these primary genetic data, we included three types of covariates for each tumor: tumor mutation burden (50), aneuploidy (51) and genome mutational signatures (52,53).

The average numbers of genes per tumor with somatic mutations, copy number amplifications (CNA) or copy number deletions (CND) were 9.1, 15.9 or 14.0, respectively (sequenced genes only; Supplementary Fig. 1A). Guided by previous reports (54,55), we verified

that certain cancer types display a high rate of somatic mutations and low rate of copy number alterations (e.g. melanoma, bladder cancer) while other cancer types (e.g. ovarian cancer) demonstrate the opposite (Supplementary Fig. 1B). A total of 14 genes had pan-cancer alteration frequencies greater than 10% (considering all types of alterations; Supplementary Fig. 1C).

In parallel to these tumor data, we obtained seven different types of molecular networks, covering general protein-protein physical binding (56–58), transcriptional regulation (57–59), kinase-substrate phosphorylation (57,58,60), ubiquitin ligase-substrate ubiquitination (57,58,61,62), genetic interactions (63,64), specific protein-protein interactions among DNA damage repair factors (65), and global integrated gene-gene networks including STRING (66) and PCNet (23) (Fig. 1B; Methods). Network resources were complementary with respect to pairwise interactions (Supplementary Table 1), altogether contributing 19,789 interactions among the 468 genes (union of all networks).

An integrative foundation model of tumor genotype accurately reconstructs genetic alterations

MutationProjector uses the graph attention network (arXiv 1710.10903, arXiv 2105.14491) architecture to translate high dimensional data from tumor gene alterations and covariates into a unified and quantitative representation of tumor biological state (Fig. 1C, Supplementary Fig. 2A, B). This representation is created by learning associations among genes, implemented by message-passing within each of the input interaction networks, as well as between genes and covariates (Methods). Associations are learned through attention mechanisms similar to the Transformer model (arXiv 1706.03762).

Model optimization was based primarily on self-supervised learning, whereby a model learns to reconstruct masked components of a dataset using the rest of that dataset only, without

external labels (arXiv 1810.04805). Here, MutationProjector was trained to optimally recover genetic alteration states of masked genes, given the states of all other sequenced genes in the cancer genome (Fig. 2A). Following training, the accuracy of masked gene prediction was scored using area under the precision recall curve (AUPRC) as a primary measure. In cross-validation, we observed an overall AUPRC of 0.21 for predicting gene mutation status (Fig. 2B) as compared to 0.02 for random guessing, a 9.7 fold improvement. This performance was stable across cancer types (Supplementary Table 2) and similar when assessing an external lung adenocarcinoma cohort (Fig. 2B). Further inspection indicated that the model's favorable performance was due to the pervasive mutational dependencies captured by cancer gene networks, leading to high model attention on thousands of gene-gene interactions (Fig. 2C). For instance, genes *CCND1* and *CDKN2A* show frequent co-occurrence of copy number alterations, such that the alteration state of one is accurately recovered given information about the other; conversely, *BRAF* and *NRAS* show strong mutual exclusivity of somatic mutations, with similar cross-recovery capability (Fig. 2D, E).

We also trained the model in two auxiliary supervised tasks, based on prediction of cancer type and status of tumor infiltrating lymphocytes (TILs; Methods, Supplementary Fig. 3). Such a mixed self-supervised/supervised training strategy has been used in many foundation models, including AlphaMissense (67) and DNAGPT (arXiv 2307.05628). In these auxiliary tasks, we found a strong potential to recognize cancer subtypes (Supplementary Table 2), with AUPRCs ranging from 0.36 (esophageal cancer) to 0.94 (breast cancer). Recovery of TIL status was also significant (Supplementary Table 2), with AUPRCs of 0.42 for TIL-inflamed, 0.64 for TIL-excluded, or 0.45 for TIL-desert. These performance outcomes generally exceeded those observed

across alternative model configurations using different architectures and training procedures, further supporting the MutationProjector design (Supplementary Table 2, Methods).

Mutation projections capture complex genotypes and pathway alterations

We next sought to examine and interpret the compact tumor embedding which MutationProjector had learned from the cancer genomic data (derived from the final model layer; dimension = 10). Examination of the embedding coordinates (Fig. 3A, UMAP projection) (arXiv 1802.03426) indicated that tumors were approximately stratified by tissue type, with substantial admixture of certain types. For example, tumors from lung squamous cell carcinoma, head-neck squamous cell carcinoma or esophageal cancer were highly similar to one another (Fig. 3B) as part of a larger squamous cluster of tumors (Fig. 3C). Cancer types from both of the major tumor datasets (MSK-IMPACT and TCGA) clustered closely in the latent space, suggesting that batch effects are minimal (Supplementary Fig. 4A-C). Even when the cancer type prediction task was removed during pre-training, the embedding continued to stratify the major cancer types (Supplementary Fig. 4D). Beyond tissue type, the MutationProjector embedding conveyed substantial additional information regarding the alteration status of key cancer drivers (Supplementary Fig. 4E, F), including *APC* mutation in colorectal cancer (68), *ARID1A* mutation in a distinct bladder cancer subtype (69), *GATA3* mutation in a distinct breast cancer subtype (70), and *NFE2L2* mutation in a distinct set of squamous cell carcinomas (71) (Fig. 3D). Also discernable were distinct subtypes of tumors characterized by co-alterations of certain gene pairs involved in genetic or functional interactions, including (i) *FGFR3* mutation and *CDKN2A* deletion (involved in the proliferation-versus-apoptosis axis); (ii) *TP53* and *RB1* mutations (involved in cell-cycle regulation) (69); or (iii) *STK11* and *KEAP1* mutations (involved in cellular metabolic and oxidative stress) (71) (Fig.

3E). Further inspection of the *FGFR3-CDKN2A* interaction in bladder cancer tumors showed that alterations to these genes were mutually exclusive; the same was true for *TP53-RB1* (Fig. 3F).

Beyond genetic alteration profiles, we found that the MutationProjector embedding had significant correspondence with other molecular information not explicitly provided during model training. For example, although HPV+ status was not an input to the model, this status was easily discernable in the embedding (Fig. 3G, H and Supplementary Fig. 4G) based on a coordinated profile of alterations in genes involved in apoptosis regulation and Wnt signaling (e.g. *TP53*, *CDKN2A/B*, *CCND1*; Supplementary Fig. 4H) and consistent with prior studies (72–78). We also found that MutationProjector captured classical tumor subtypes previously defined by mRNA sequencing, e.g. basal versus luminal transcriptional subtypes in bladder cancer (69) (Fig. 3I) or breast cancer (70) (Fig. 3J).

MutationProjector excels in prediction of therapeutic responses

Next, we evaluated the utility of the MutationProjector tumor representation across a range of clinical applications related to prediction of immunotherapy response, prediction of chemotherapy response, or prediction of metastatic outcomes (Fig. 4A). For each of these tasks, we accessed multiple independent cohorts which were apportioned into training and test sets (49,79–90), comprising n = 3,362 patients in all (Table 1). MutationProjector coordinates of each tumor were provided as inputs to a Random Forest classifier for predicting the relevant outcomes (Methods).

For immunotherapy response prediction, the classifier was trained on anti-PD1/PDL1 treated patients (n = 94) encompassing multiple solid tumor types, then tested on three independent cohorts representing bladder (n = 130), lung (n = 229) or melanoma (n = 144) cancers (Table 1). In each of these contexts, MutationProjector demonstrated significant prediction of progression-

free and/or overall survival, with hazard ratios for bladder cancer of HR = 0.71 ($p = 1.6\times10^{-2}$), for lung cancer of HR = 0.74 ($p = 1.7\times10^{-5}$), and for melanoma of HR = 0.59 ($p = 5.8\times10^{-3}$) (Fig. 4B). This performance compared favorably to that of current biomarkers (91–93) including PD-L1 expression, tumor mutation burden, microsatellite status and previously nominated genetic alterations in *KRAS* and other genes (Fig. 4C, Supplementary Fig. 5A-F). Mutation Projector was also the leading approach in comparison to classical supervised learning methods (Fig. 4C).

For chemotherapy response prediction, the classifier was trained using data from a collection of cisplatin-treated tumors ($n = 237$, 8 solid tumor types), then tested on an independent cohort of bladder tumors ($n = 42$) (Table 1). Tumors classified as chemotherapy-responsive displayed substantially longer survival times than those predicted as chemotherapy-resistant, with HR = 0.43 (Fig. 4D, $p = 2.5\times10^{-2}$). Here too, MutationProjector performance compared favorably to that of current biomarkers (19) (Supplementary Fig. 5G, H) and classical machine learning methods (Fig. 4E).

For predictions related to metastatic outcomes, we first examined a cancer cohort with primary lung adenocarcinoma ($n = 1,974$) for which a subset of patients ($n = 389$) had been associated with the presence of distal metastases (Table 1). MutationProjector embedding coordinates of the primary tumor were used to train a classifier for patients with metastasis versus those determined to be metastasis-free (Supplementary Fig. 6A). When testing this classifier in an independent cohort ($n = 128$), we observed a high predictive AUPRC of 0.84 which outperformed alternative machine learning models (Supplementary Fig. 6B). Second, we studied the complementary challenge of classifying tissue-of-origin for metastatic cancers (Supplementary Fig. 6C), based on the tumor genetic alteration profiles of distal metastases for which the primary tissue-of-origin is known ($n = 272$) (Table 1). When testing this classifier on an independent cohort

of metastatic tumors (n = 112), MutationProjector showed high predictive AUPRC scores which again outperformed alternative models (Supplementary Fig. 6D-F).

Interpreting the model for biomarker discovery

Finally, we performed an analysis of model attention to identify the specific genetic biomarkers that had been most important for each of the clinical applications (Supplementary Fig. 7, Methods). Important features for prediction of a successful immunotherapy response included high TMB (92), a top feature in all cohorts including bladder (Fig. 5A), lung (Fig. 5B) and melanoma (Fig. 5C) cancers. Unexpectedly, model predictions relied on genetic alterations in numerous chromatin remodelers, such as SWI/SNF transcription factors (e.g. *ARID1A, ARID2*, *SMARCA4*), as well as in chromatin modifiers, such as histone lysine methyltransferases (*KMT2D*) (Fig. 5D). In addition to single biomarkers, we also found high model attention on interacting pairs of genes (Fig. 5E), some of which were strongly predictive of non-response including *KRAS-STK11* (14% of predicted non-responders), *KEAP1-STK11* (13%) and *SMARCA4-STK11* (10%) (Fig. 5D).

For chemotherapy response prediction (Supplementary Table 3) a variety of gene alterations had been deemed important including numerous DNA repair genes, several of which were already well-known markers of chemotherapy response, e.g. *BRCA1* mutation (94). Other important features for this task were less well known, including expanded components of cell cycle regulation (e.g. *CHEK1* and *CDKN2A*) as well as receptor tyrosine kinase pathways (e.g. *ERBB2/3*, *FGFR1*, *IGF1R, JAK2* and *PTCH1*).

Important features for prediction of metastatic outcomes included genetic alterations in β-catenin (*CTNNB1*) *(95)*, cell-cycle genes, e.g. *CDK12 (96)* and *CCNE1 (97)*, angiogenesis-promoting genes (e.g. *VEGFA*) *(98)*, APOBEC mutational signatures (99), and aneuploidy (100);

all of these markers were already well-known predictors of metastatic tumors (Supplementary Table 3). Unexpected features of metastasis included components of cell cycle and DNA damage repair (e.g. *CDKN2A*, *CDKN2B*, *RAD21*), a transcription factor (*FOXA1*), and RNA splicing (*U2AF1*).

Discussion

Here we have explored a strategy for cancer genome translation, whereby a model is first pretrained from the wealth of tumor DNA sequencing data in the public domain, then subsequently tuned for specific clinical applications using smaller well-annotated cohorts. Following large-scale task-agnostic training on over 30,000 tumor genomes, MutationProjector embeddings were used to accurately predict responses to immunotherapy with further exposure to fewer than 100 new tumor samples, achieving leading performance across multiple independent cohorts (Fig. 4). Notably, a single pre-trained tumor model was applied in multiple tasks (1-model / N-tasks), in contrast to the conventional approach of building separate models for each task (N-models / N-tasks). These findings reflect the current movement towards general-purpose AI systems capable of addressing a broad range of challenges (arXiv 2108.07258, arXiv 2302.09419).

Pretraining the MutationProjector on large genomic data allowed the model to capture nuanced cancer subtypes in its embedding representation (Fig. 3). For example, it yielded similar embedding coordinates for tumor tissue types that harbor squamous cell carcinomas, bridging head-neck cancers, lung squamous cell carcinomas and esophageal cancers (Fig. 3B, C) consistent with a previous report (101). It distinguished HPV status in cervical and head-neck cancers (Fig. 3G) although such status was not explicitly provided during training. The embedding also distinguished basal/luminal transcriptional subtypes despite not being provided with any mRNA

profiling data (Fig. 3I, J). Notably, basal bladder tumors have been associated with improved survival following neoadjuvant chemotherapy compared to luminal tumors (102–104), supporting the use of MutationProjector representations for clinical tasks. Beyond these anecdotes, the model embedding contributed to the vast majority of genes in predicting mutation patterns beyond what is explained by tissue type alone (95.7% of genes; Supplementary Fig. 4E, F).

Investigation of MutationProjector's internal logic revealed a spectrum of biomarkers in which mutational patterns are predictive of immunotherapy response (Fig. 5). These included lesser-known alterations in *KMT2D* and *SMARCA4* (Fig. 5D), which might promote a favorable immune response based on regulation of gene expression through DNA methylation (105) or chromatin remodeling (106,107). Indeed, previous studies have highlighted the role of *SMARCA4* in regulating the activity of adaptive cytokines such as interferons (106,108,109). Additionally, epigenetic modifications in tumors can alter the activation of genes involved in antigen presentation machinery, including the Major Histocompatibility Complex (MHC) (105,110).

The MutationProjector model was also able to capture combinatorial markers of immunotherapy resistance, such as *KEAP1-STK11 (111)* or *KRAS-STK11 (112)* co-alterations, which were used to predict non-responders (Fig. 5D). While mutations in *KRAS*, *STK11* or *KEAP1* were not individually more frequent in non-responders (Supplementary Table 4), their pairwise co-alterations were (Supplementary Table 5). Notably, the vast majority of gene pairs with high attention by the model (Fig. 2C) would not have been prioritized by a standard co-mutation analysis (94% of high attention gene pairs; Supplementary Fig. 8A, B). Rather, the present study offers a complementary approach in which the embeddings learned by a self-supervised model can be interpreted to recover complex, multi-gene mutational patterns (Supplementary Fig. 8C-E) underlying clinical outcomes (Fig. 5D).

While 30,000+ genomes representing 10 solid tumor types were considered in our study, numerous additional tumor samples are available for expansion of MutationProjector to tumor types such as pancreatic cancer, prostate cancer or sarcomas (113–115). Further cancer genomic resources, such as those provided by The International Cancer Genome Consortium (ICGC) (116), may yet further improve predictive performance and biomarker identification. Finally, in addition to genomic data, complementary information from other relevant modalities – such as annotations in the electronic health record, images from radiological computed tomography scans and mRNA transcriptomic profiles – may well benefit the pre-training step to increase model performance and interpretability. Other future studies should explore the extent to which the MutationProject concept can be applied to further clinical tasks of interest, including application to liquid biopsies of circulating tumor DNAs for early cancer detection.

Materials and Methods

Data preparation

Clinical datasets for pre-training were retrieved from Project GENIE 15.0 (48) and TCGA (49) databases. To remove potential data leakage, any duplicate samples were removed from analysis. We focused on 10 solid tumor types, centered on breast (n=8,163), colorectal (n=6,152) and non-small-cell lung cancer (adenocarcinoma n = 650 and squamous cell carcinoma n = 1,454) and additional cancers that are genomically adjacent to these (101), including bladder cancer (n=2,734), cervical cancer (n=526), esophageal cancer (n=887), head and neck cancer (n=650), melanoma (n=1,230) and ovarian cancer (n=87). We used the following OncoTree IDs (117) (recorded in the TCGA database) to select patients harboring the 10 tumor types: BLCA for bladder cancer, BRCA, BRCNOS, IDC, ILC, IMMC and MBC for breast cancer, CEEN, CEMU and CESC for cervical cancer, COAD, MACR, READ for colorectal cancer, ESCA, ESCC for esophageal cancer, HNSC for head and neck cancer, LUAD for lung adenocarcinoma, LUSC for lung squamous cell carcinoma, SKCM for melanoma and SCO for ovarian cancer. TILs were computed from hematoxylin and eosin stained images in the TCGA dataset using a previously developed ResNet-based deep learning model to predict the status of immune infiltrating lymphocytes (Supplementary Fig. 3) (118), constituting immune inflamed (n=1,052), immune excluded (n=1,666) or immune desert (n=1,015) phenotypes (Supplementary Table 6).

As done in previous studies (119,120), somatic mutations were prepared from 468 genes included in the MSK-IMPACT clinical gene sequencing panel (4,5). A gene was marked as mutated (“1”) in a tumor sample if it had the following mutation types: missense/nonsense mutations, frameshift insertions/deletions, splice site regions, in-frame insertions/deletions. For

TMB we used the reported values; if not available, we calculated TMB using the Maftools R package (50). To calculate arm-level aneuploidy score, we used the ASCETS R package (51). To compute mutational signatures, we used the results of the MESiCA algorithm (53) for patients profiled using clinical gene sequencing, whereas we used SigProfiler (52) for patients with whole-exome/-genome sequencing. As done in a previous study (53), we defined seven dominant signatures corresponding to: APOBEC (Apolipoprotein B mRNA editing enzyme catalytic subunit), Clock Signature SBS1, Clock Signature SBS5, Mismatch Repair (MMR), DNA Polymerase Epsilon (POLE), Tobacco, and Ultraviolet light (UV). For APOBEC, Tobacco, UV and MMR, a relative contribution of 30% was labeled as a dominant signature ("1"). For Homologous Recombination Deficiency (HRD) and POLE signatures, relative contributions of at least 50% and 20% were used, respectively. For Clock SBS1 and Clock SBS5 signatures, samples with no other signature and at least 40% relative contribution were labeled as having a dominant signature.

Molecular networks

We considered eight molecular networks in total, encompassing several different molecular interaction types: physical interaction, transcriptional regulation, phosphorylation, ubiquitination, genetic, and functional interaction. For physical interactions, we integrated BioPlex ('shared' interaction resource) (56), SIGNOR (58) and SignaLink (57). For transcriptional regulation, we integrated TRRUST (59) v2, SIGNOR (58) and SignaLink (57). For phosphorylation, we integrated PhosphoSitePlus (60), SIGNOR (58) and SignaLink (57). For ubiquitination, we integrated UbiNet 2.0 (61), UbiBrowser 2.0 (62), SIGNOR (58) and SignaLink (57). For genetic interactions, we integrated ISLE (63) and SynLethDB (64). Functional interaction networks

included the DDRAM network (65), PCNet (23) v1.3 and STRING (66) v12. The NDEx (121) repository was used to download PCNet, STRING, BioPlex and DDRAM networks. The number of interactions for each network type were 1,239 for physical interaction, 3,487 for transcriptional regulation, 774 for phosphorylation, 152 for ubiquitination, 1,245 for genetic interactions, 359 for DNA damage repair, 3,454 for STRING and 15,632 for PCNet.

Model architecture and pre-training

The MutationProjector design loosely follows the Transformer encoder architecture (arXiv 1706.03762), with extensions to encode tumor gene alterations and incorporate molecular networks. Specifically, we used the graph attention network (GAT) (arXiv 1710.10903, arXiv 2105.14491) to encode network-specific attention knowledge across all samples, enabling biologically-informed decisions (Fig. 1C). GAT layers were implemented via *GATv2Conv* from the pytorch-geometric (arXiv 1903.02428) python library. To propagate the impact of local gene-level alteration patterns across various molecular interaction types, we used multi-head attention with eight attention heads per encoder unit, where each attention head addressed one of the eight distinct interaction networks. This aspect enforces each attention head to capture biologically unique signals; for instance, the physical protein-protein interaction network head will have captured different signals compared to the phosphorylation network head. This configuration was similar to that of a previous study on driver gene prediction (25). For genomic covariates (TMB, aneuploidy, mutational signatures), we made connections from these features to all other nodes in the network in all attention heads, similar to how special tokens or nodes are used in BERT (arXiv 1810.04805), the Vision Transformer (arXiv 2010.11929) and Graphormer (arXiv 2106.05234). This configuration enables the model to aggregate and broadcast genomic covariate information

globally, which strengthens long-range interactions via attention patterns. We disabled self-loops in the message passing layer (GAT layer) and applied residual connections as in the original Transformer (arXiv 1706.03762) and ResNet (122) architectures (Fig. 1C). We used two encoder units, a depth commonly used in graph neural networks (25) (arXiv 1710.10903).

Similar to previous self-supervised methods (32,33,35) (arXiv 1810.04805), we conducted masked gene prediction by hiding (masking) alteration status of certain genes and asking the model to complete the masked gene's mutation and copy number alteration status (Fig. 2A). Each gene was represented as two distinct learnable vectors: a Gene Token embedding that encodes the gene's identity (e.g., *TP53*, *KRAS*), and a Mutation Embedding that encodes different combinations of somatic mutations and copy number alternations (Fig. 1C). The Mutation Embedding is a learnable parameter initialized using PyTorch's *nn.Embedding* module. We then randomly selected 15% of genes and replaced their Mutation Embedding with a special learnable masked token, while the Gene Token was left untouched. This approach provided the model with knowledge of precisely which gene is being predicted while remaining unaware of its actual alteration status. We took element-wise sums of the two token embeddings to generate gene embeddings (Fig. 1C). For TMB and aneuploidy, in each cohort, we binned the continuous values into five bins, as done similarly in scBERT (33). We used 10-dimensional vectors for gene and covariate embeddings ($d_{features}$; Supplementary Fig. 2B). To improve prediction performance in downstream tasks, we borrowed ideas from multi-task learning and jointly pre-trained our model on additional supervised labels (i.e. cancer type and levels of TILs) as done in BERT (arXiv 1810.04805), AlphaMissense (67) and DNAGPT (arXiv 2307.05628). Both cancer types and levels of TILs were categorical labels.

The MutationProjector objective function (Loss) aggregates the prediction errors of self-supervised and supervised learning using binary cross entropy (BCE):

$$Loss = \frac{1}{N_{masked\ genes}} BCE_{masked\ gene} + \frac{1}{N_{cancer\ types}} BCE_{cancer\ types} + \frac{1}{N_{TIL}} BCE_{TIL}$$

where $N_{masked\ genes}$, $N_{cancer\ types}$ and N_{TIL} are the number of genes masked during pre-training, the number of samples with cancer types, and the number of samples with TIL information, respectively. To avoid the tendency to predict abundant classes for the BCE losses, we added weights (inversely proportional to the frequency of a positive label) to label prediction. Specifically, we used the *pos_weight* parameter of the *torch.nn.BCEWithLogitsLoss* to account for class imbalance. The hyperparameters of batch size, number of epochs, learning rate, and dropout rate were set at 64, 100, 0.001, and 0.1, respectively. We optimized the objective function using the AdamW optimizer (arXiv 1711.05101) with a weight decay of 0.0001. All MutationProjector models were implemented in PyTorch and used NVIDIA V100 GPUs. To assess the model's pretraining performance, we randomly split the tumor dataset into 80% training (n=24,262) and 20% held-out testing (n=6,066) samples. Except for the evaluation of pre-training performance (Fig. 2), all 30,328 tumors were used to pre-train the MutationProjector for downstream applications.

We conducted several ablation studies to evaluate the importance of the major aspects of the model architecture and of the training procedure (Supplementary Table 2 and Supplementary Table 7). First, we explored substituting the three-task training procedure with simpler training protocols, in which we removed each of the three tasks from training. Second, we benchmarked MutationProjector against graph-free implementations, including (i) a naive transformer encoder with a matched number of layers but without any access to network knowledge or (ii) replacing graph attention layers with feed forward neural network layers. Third, we explored substituting the

graphs with randomized graphs by shuffling node labels. Fourth, we explored using only one graph for model training and testing.

Transfer learning and clinical task evaluations

Downstream prediction tasks implemented a transfer learning schema in which the MutationProjector embedding (dimension = 180; Supplementary Fig. 2B) was used as an input for a random forest classifier model (Fig. 4, 100 trees, max depth 10, class weight = balanced). This random forest using the MutationProjector embedding was compared to benchmark models including logistic regression, regular random forest and a supervised graph neural network (similar to MutationProjector but without pretraining), using gene alterations (mutations and copy number alterations) and genomic covariates as inputs. For the supervised graph neural network baseline, binary cross entropy loss was used to optimize the model.

For drug response prediction tasks, we used the response of each patient as measured according to standard clinical RECIST categories (123) (Response Evaluation Criteria in Solid Tumors). We binarized the RECIST categories into responders (Complete Response and Partial Response) and non-responders (Stable Disease and Progressive Disease). For the metastasis prediction task, we used tumor sequencing data from primary tumor biopsies and labeled each patient as metastatic if the pathologic tumor stage was IV, and as non-metastatic if the stage was I (124). Stage II and III tumors were not included in the analysis. For tissue-of-origin prediction for metastatic cancers, we used the reported primary tissue-of-origin (i.e. breast cancer, colorectal cancer or lung adenocarcinoma).

For drug response tasks, we labeled patients with the top 20% predicted sensitivity scores as ‘responder’ and otherwise as ‘non responder’ following thresholds set in prior studies (92,119).

To quantify the performance for predicting drug response, we used the log-rank test (Fig. 4B, D). Additionally, to assess performance of drug response predictions independent of any cutoff threshold, we computed Harrell's C-index (Supplementary Table 7) and hazard ratios (Fig. 4C, E) by running univariate Cox proportional hazards regression using the predicted drug response scores as continuous variables (Fig. 4). Specifically, the hazard ratio is reported per one standard deviation increase in the predicted probability. Of note, immunotherapy bladder cancer cohort was sequenced using a gene panel (i.e. FoundationOne) that is different from the panel used for pre-training dataset (i.e. MSK-IMPACT panel). For the metastasis and tissue-of-origin prediction tasks, we used the area under the precision recall curve ([AUPRC], Supplementary Fig. 6).

Identification of important features

We used attention weights to quantify the importance of each feature on model predictions. To calculate feature importance, as done in a previous study (33) we aggregated multi-head attention weight matrices into a single attention weight matrix by taking an element-wise maximum of the attention weights between a pair of features (Supplementary Fig. 7A, B). This max attention weight matrix was used to identify mutually exclusive attention patterns, for example between *TP53-RB1* and *FGFR3-CDKN2A* in bladder cancer (Fig. 3F). Then, we applied a simplified version of the attention-based linear probing approach that was originally used to investigate BERT (arXiv 1906.04341). In detail, to determine the importance of feature *i*, we used a linear regression on the input values of *i*, weighted by incoming and outgoing attention weights, in order to predict model outcomes (Supplementary Fig. 7C). The feature importance was measured by calculating the Spearman correlation between observed outcome and the predicted outcome of the linear regression model.

Data Availability

All pharmacogenomics datasets used in this study are publicly accessible through the sources listed (Table 1). Codes used in this study are available at: https://github.com/idekerlab/MutationProjector#

Research Resource Identifiers

Antibodies: Not applicable. No antibodies were used.

Cell lines: Not applicable. No cell lines were used.

Core Facilities: School of Medicine, University of California San Diego (RRID:SCR_010634), National Resource for Network Biology (RRID:SCR_004259)

Instruments: Not applicable. Computational study only.

Online Resources/Databases: Synapse (RRID:SCR_006307), The Cancer Genome Atlas (RRID: SCR_003193), cBioPortal (RRID:SCR_014555), Network Data Exchange (RRID:SCR_003943)

Organisms: Not applicable. Only genomic sequencing from human cancer tumors were used.

Services: Not applicable. Only publicly available datasets were used.

Software Tools: Python (RRID:SCR_008394), R (RRID:SCR_001905)

Acknowledgements

We thank members of the Ideker, Carter and Alexandrov laboratories for helpful discussions. We also acknowledge the data provided by the TCGA Research Network, cBioportal and Project Genie. The authors would like to acknowledge the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of the study authors. This work was supported by grants from the National Institutes of Health T32CA121938 to J.K., U54CA274502 to T.I., R01ES014811 to T.I. Additionally, this work was supported by NRNB (U.S. National Institutes of Health Division of Biomedical Technology, Bioinformatics, and Computational Biology, P41 GM103504). We gratefully acknowledge funding support from the Advanced Research Projects Agency for Health (ARPA-H) Advanced Analysis for Precision Cancer Therapy (ADAPT) Program (OT 140D042590013; T.I.).

References

1. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Genome Med. Springer Science and Business Media LLC; 2020;12:8.

2. Sicklick JK, Kato S, Okamura R, Schwaederle M, Hahn ME, Williams CB, et al. Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. Nat Med. Springer Science and Business Media LLC; 2019;25:744–50.

3. Yip S, Christofides A, Banerji S, Downes MR, Izevbaye I, Lo B, et al. A Canadian guideline on the use of next-generation sequencing in oncology. Curr Oncol. MDPI AG; 2019;26:e241–54.

4. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23:703–13.

5. Jibiki T, Nishimura H, Sengoku S, Kodama K. Regulations, open data and healthcare innovation: A case of MSK-IMPACT and its implications for better cancer care. Cancers (Basel). MDPI AG; 2021;13:3448.

6. Milbury CA, Creeden J, Yip W-K, Smith DL, Pattani V, Maxwell K, et al. Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. PLoS One. Public Library of Science (PLoS); 2022;17:e0264138.

7. Beaubier N, Tell R, Lau D, Parsons JR, Bush S, Perera J, et al. Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. Oncotarget. Impact Journals, LLC; 2019;10:2384–96.

8. Vestergaard LK, Oliveira DNP, Poulsen TS, Høgdall CK, Høgdall EV. Oncomine™ Comprehensive Assay v3 vs. Oncomine™ Comprehensive Assay Plus. Cancers (Basel). MDPI AG; 2021;13:5230.

9. Homicsko K, Kenneth R, Voss A, Michielin O. 3356 Triple wild type melanoma profiling in the Caris Molecular IntelligenceTM registry. Eur J Cancer. Elsevier BV; 2015;51:S687.

10. Freedman AN, Klabunde CN, Wiant K, Enewold L, Gray SW, Filipski KK, et al. Use of next-generation sequencing tests to guide Cancer Treatment: Results from a nationally representative survey of oncologists in the United States. JCO Precis Oncol. American Society of Clinical Oncology (ASCO); 2018;2:1–13.

11. Marquart J, Chen EY, Prasad V. Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. JAMA Oncol. 2018;4:1093–8.

12. Takeda M, Takahama T, Sakai K, Shimizu S, Watanabe S, Kawakami H, et al. Clinical application of the FoundationOne CDx assay to therapeutic decision-making for patients with advanced solid tumors. Oncologist. Oxford University Press (OUP); 2021;26:e588–96.

13. Pinet S, Durand S, Perani A, Darnaud L, Amadjikpe F, Yon M, et al. Clinical management of molecular alterations identified by high throughput sequencing in patients with advanced solid tumors in treatment failure: Real-world data from a French hospital. Front Oncol. Front Oncol; 2023;13:1104659.

14. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 2013;339:1546–58.

15. Van Allen EM, Mouw KW, Kim P, Iyer G, Wagle N, Al-Ahmadie H, et al. Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. Cancer Discov. 2014;4:1140–53.

16. Sawant A, Kothandapani A, Zhitkovich A, Sobol RW, Patrick SM. Role of mismatch repair proteins in the processing of cisplatin interstrand cross-links. DNA Repair (Amst). 2015;35:126–36.

17. Li Q, Damish AW, Frazier Z, Liu D, Reznichenko E, Kamburov A, et al. ERCC2 helicase domain mutations confer nucleotide excision repair deficiency and drive cisplatin sensitivity in muscle-invasive bladder cancer. Clin Cancer Res. American Association for Cancer Research (AACR); 2019;25:977–88.

18. Taber A, Christensen E, Lamy P, Nordentoft I, Prip F, Lindskrog SV, et al. Molecular correlates of cisplatin-based chemotherapy response in muscle invasive bladder cancer by integrated multi-omics analysis. Nat Commun. Springer Science and Business Media LLC; 2020;11:4858.

19. Plimack ER, Dunbrack RL, Brennan TA, Andrake MD, Zhou Y, Serebriiskii IG, et al. Defects in DNA repair genes predict response to neoadjuvant cisplatin-based chemotherapy in muscle-invasive bladder cancer. Eur Urol. 2015;68:959–67.

20. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144:646-74.

21. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18:551–62.

22. Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. Springer Science and Business Media LLC; 2013;14:719–32.

23. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst. 2018;6:484–95.e5.

24. Wright SN, Colton S, Schaffer LV, Pillich RT, Churas C, Pratt D, et al. State of the interactomes: an evaluation of molecular networks for generating biological insights. Mol Syst Biol. Springer Science and Business Media LLC; 2025;21:1–29.

25. Chatzianastasis M, Vazirgiannis M, Zhang Z. Explainable Multilayer Graph Neural Network for cancer gene prediction. Bioinformatics. 2023;39:btad643.

26. Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet 8, 699–710 (2007). https://doi.org/10.1038/nrg2144.

27. Avila PU, Padvitski T, Leote AC, Chen H, Saez-Rodriguez J, Kann M, et al. Gene regulatory networks in disease and ageing. Nat Rev Nephrol 20, 616–633 (2024). https://doi.org/10.1038/s41581-024-00849-7.

28. Tognetti M, Gabor A, Yang M, Cappelletti V, Windhager J, Rueda OM, et al. Deciphering the signaling network of breast cancer improves drug sensitivity prediction. Cell Syst. Elsevier BV; 2021;12:401–18.e12.

29. Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, et al. Pre-trained models: Past, present and future. AI Open. 2021;2:225–50.

30. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics. 2021;37:2112–20.

31. Benegas G, Batra SS, Song YS. DNA language models are powerful predictors of genome-wide variant effects. Proc Natl Acad Sci U S A. 2023;120:e2311219120.

32. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618:616–24.

33. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nature Machine Intelligence. Nature Publishing Group; 2022;4:852–66.

34. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21:1470–1480.

35. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024;21:1481–1491.

36. Sheinin R, Sharan R, Madi A. scNET: learning context-specific gene and cell embeddings by integrating single-cell gene expression data with protein-protein interactions. Nat Methods. Springer Science and Business Media LLC; 2025;22:708–16.

37. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871–6.

38. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure

prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.

39. Baek M, McHugh R, Anishchenko I, Jiang H, Baker D, DiMaio F. Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. Nat Methods. 2024;21:117–21.

40. Wagner SJ, Reisenbüchler D, West NP, Niehues JM, Zhu J, Foersch S, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell. 2023;41:1650–61.e4.

41. Ma C, Tan W, He R, Yan B. Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. Nat Methods. 2024;21:1558–1567.

42. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023;622:156–63.

43. Kim C, Gadgil SU, DeGrave AJ, Omiye JA, Cai ZR, Daneshjou R, et al. Transparent medical image AI via an image–text foundation model grounded in medical literature. Nat Med. Nature Publishing Group; 2024;30:1154–65.

44. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun. Springer Science and Business Media LLC; 2024;15:654.

45. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024;630:181–8.

46. Azizi S, Culp L, Freyberg J, Mustafa B, Baur S, Kornblith S, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nat Biomed Eng. 2023;7:756–79.

47. Arango-Argoty G, Kipkogei E, Stewart R, Sun GJ, Patra A, Kagiampakis I, et al. Pretrained transformers applied to clinical studies improve predictions of treatment efficacy and associated biomarkers. Nat Commun. Springer Science and Business Media LLC; 2025;16:2101.

48. Pugh TJ, Bell JL, Bruce JP, Doherty GJ, Galvin M, Green MF, et al. AACR Project GENIE: 100,000 Cases and Beyond. Cancer Discov. 2022;12:2044–57.

49. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

50. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28:1747–56.

51. Spurr LF, Touat M, Taylor AM, Dubuc AM, Shih J, Meredith DM, et al. Quantification of aneuploidy in targeted sequencing data using ASCETS. Bioinformatics. 2021;37:2461–3.

52. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.

53. Yaacov A, Ben Cohen G, Landau J, Hope T, Simon I, Rosenberg S. Cancer mutational signatures identification in clinical assays using neural embedding-based representations. Cell Rep Med. 2024;5:101608.

54. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. Springer Science and Business Media LLC; 2013;45:1127–33.

55. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science. 2017;355:eaaf8399.

56. Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell. 2021;184:3022–40.e28.

57. Csabai L, Fazekas D, Kadlecsik T, Szalay-Bekő M, Bohár B, Madgwick M, et al. SignaLink3: a multi-layered resource to uncover tissue-specific signaling networks. Nucleic Acids Res. 2022;50:D701–9.

58. Lo Surdo P, Iannuccelli M, Contino S, Castagnoli L, Licata L, Cesareni G, et al. SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. Nucleic Acids Res. 2023;51:D631–7.

59. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46:D380–6.

60. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2012;40:D261–70.

61. Li Z, Chen S, Jhong J-H, Pang Y, Huang K-Y, Li S, et al. UbiNet 2.0: a verified, classified, annotated and updated database of E3 ubiquitin ligase-substrate interactions. Database (Oxford). 2021;baab010.

62. Wang X, Li Y, He M, Kong X, Jiang P, Liu X, et al. UbiBrowser 2.0: a comprehensive resource for proteome-wide known and predicted ubiquitin ligase/deubiquitinase-substrate interactions in eukaryotic species. Nucleic Acids Res. 2022;50:D719–28.

63. Lee JS, Das A, Jerby-Arnon L, Arafeh R, Auslander N, Davidson M, et al. Harnessing synthetic lethality to predict the response to cancer treatment. Nat Commun. 2018;9:2546.

64. Wang J, Wu M, Huang X, Wang L, Zhang S, Liu H, et al. SynLethDB 2.0: a web-based

knowledge graph database on synthetic lethality for novel anticancer drug discovery. Database (Oxford). 2022;baac030.

65. Kratz A, Kim M, Kelly MR, Zheng F, Koczor CA, Li J, et al. A multi-scale map of protein assemblies in the DNA damage response. Cell Syst. 2023;14:447–63.e8.

66. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51:D638–46.

67. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381:eadg7492.

68. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, et al. Genetic alterations during colorectal-tumor development. N Engl J Med. New England Journal of Medicine (NEJM/MMS); 1988;319:525–32.

69. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2018;174:1033.

70. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.

71. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.

72. Poeta ML, Manola J, Goldwasser MA, Forastiere A, Benoit N, Califano JA, et al. TP53 mutations and survival in squamous-cell carcinoma of the head and neck. N Engl J Med. Massachusetts Medical Society; 2007;357:2552–61.

73. Gillison ML, Koch WM, Capone RB, Spafford M, Westra WH, Wu L, et al. Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. J Natl Cancer Inst. Oxford University Press (OUP); 2000;92:709–20.

74. Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman Research Institute of City of Hope, et al. Integrated genomic and molecular characterization of cervical cancer. Nature. pmc.ncbi.nlm.nih.gov; 2017;543:378–84.

75. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. Springer Science and Business Media LLC; 2015;517:576–82.

76. Sano T, Oyama T, Kashiwabara K, Fukuda T, Nakajima T. Expression status of p16 protein is associated with human papillomavirus oncogenic potential in cervical and genital lesions. Am J Pathol. Elsevier BV; 1998;153:1741–8.

77. Rischin D, Young RJ, Fisher R, Fox SB, Le Q-T, Peters LJ, et al. Prognostic significance of p16INK4A and human papillomavirus in patients with oropharyngeal cancer treated on TROG 02.02 phase III trial. J Clin Oncol. American Society of Clinical Oncology (ASCO); 2010;28:4142–8.

78. Seiwert TY, Zuo Z, Keck MK, Khattri A, Pedamallu CS, Stricker T, et al. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. Clin Cancer Res. American Association for Cancer Research (AACR); 2015;21:632–41.

79. Miao D, Margolis CA, Vokes NI, Liu D, Taylor-Weiner A, Wankowicz SM, et al. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. Nat Genet. Springer Science and Business Media LLC; 2018;50:1271–81.

80. Jee J, Fong C, Pichotta K, Tran TN, Luthra A, Waters M, et al. Automated real-world data integration improves cancer outcome prediction. Nature. Springer Science and Business Media LLC; 2024;636:728–36.

81. Pleasance E, Titmuss E, Williamson L, Kwan H, Culibrk L, Zhao EY, et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. Nat Cancer. Springer Science and Business Media LLC; 2020;1:452–68.

82. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, et al. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature. 2018;554:544–8.

83. Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. Nat Cancer. Springer Science and Business Media LLC; 2022;3:1151–64.

84. Liu D, Schilling B, Liu D, Sucker A, Livingstone E, Jerby-Arnon L, et al. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. Nat Med. Springer Science and Business Media LLC; 2019;25:1916–27.

85. Kim PH, Cha EK, Sfakianos JP, Iyer G, Zabor EC, Scott SN, et al. Genomic predictors of survival in patients with high-grade urothelial carcinoma of the bladder. Eur Urol. Elsevier BV; 2015;67:198–201.

86. Lengel HB, Mastrogiacomo B, Connolly JG, Tan KS, Liu Y, Fick CN, et al. Genomic mapping of metastatic organotropism in lung adenocarcinoma. Cancer Cell. Elsevier BV; 2023;41:970–85.e3.

87. Robinson DR, Wu Y-M, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548:297–303.

88. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer

genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. American Association for Cancer Research (AACR); 2012;2:401–4.

89. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6:l1.

90. de Bruijn I, Kundra R, Mastrogiacomo B, Tran TN, Sikina L, Mazor T, et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. Cancer Res. Cancer Res; 2023;83:3861–7.

91. Ferreiro-Pantín M, Anido-Herranz U, Betancor YZ, Cebey-López V, León-Mateos L, García-González J, et al. Clinical, molecular, and immune correlates of the Immunotherapy Response Score in patients with advanced urothelial carcinoma under atezolizumab monotherapy: analysis of the phase II IMvigor210 trial. ESMO Open. Elsevier BV; 2023;8:101611.

92. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet. 2019;51:202–6.

93. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015;372:2509–20.

94. Byrski T, Dent R, Blecharz P, Foszczynska-Kloda M, Gronwald J, Huzarski T, et al. Results of a phase II open-label, non-randomized trial of cisplatin chemotherapy in patients with BRCA1-positive metastatic breast cancer. Breast Cancer Res. Springer Science and Business Media LLC; 2012;14:R110.

95. Pacheco-Pinedo EC, Durham AC, Stewart KM, Goss AM, Lu MM, Demayo FJ, et al. Wnt/β-catenin signaling accelerates mouse lung tumorigenesis by imposing an embryonic distal progenitor phenotype on lung epithelium. J Clin Invest. American Society for Clinical Investigation; 2011;121:1935–45.

96. Liu H, Liu K, Dong Z. Targeting CDK12 for cancer therapy: Function, mechanism, and drug discovery. Cancer Res. American Association for Cancer Research (AACR); 2021;81:18–26.

97. Yao S, Meric-Bernstam F, Hong D, Janku F, Naing A, Piha-Paul SA, et al. Clinical characteristics and outcomes of phase I cancer patients with CCNE1 amplification: MD Anderson experiences. Sci Rep. Springer Science and Business Media LLC; 2022;12:8701.

98. Qin S, Yi M, Jiao D, Li A, Wu K. Distinct roles of VEGFA and ANGPT2 in lung adenocarcinoma and squamous cell carcinoma. J Cancer. Ivyspring International Publisher; 2020;11:153–67.

99. Li Q, Jiang M, Hong S, Yang J, Wu X, Pang J, et al. Comprehensive genomic and clinical analyses identify APOBEC mutational signatures as a brain metastasis risk factor in lung

adenocarcinoma patients. Transl Oncol. Elsevier BV; 2024;43:101921.

100. Vasudevan A, Baruah PS, Smith JC, Wang Z, Sayles NM, Andrews P, et al. Single-chromosomal gains can function as metastasis suppressors and promoters in colon cancer. Dev Cell. Elsevier BV; 2020;52:413–28.e6.

101. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173:291–304.e6.

102. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer Cell. Elsevier BV; 2014;25:152–65.

103. McConkey DJ, Choi W, Shen Y, Lee I-L, Porten S, Matin SF, et al. A prognostic gene expression signature in the molecular classification of chemotherapy-naïve urothelial cancer is predictive of clinical outcomes from neoadjuvant chemotherapy: A phase 2 trial of dose-dense methotrexate, vinblastine, doxorubicin, and cisplatin with bevacizumab in urothelial cancer. Eur Urol. 2016;69:855–62.

104. Seiler R, Ashab HAD, Erho N, van Rhijn BWG, Winters B, Douglas J, et al. Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy. Eur Urol. 2017;72:544–54.

105. Zhong F, Lin Y, Zhao L, Yang C, Ye Y, Shen Z. Reshaping the tumour immune microenvironment in solid tumours via tumour cell and immune cell DNA methylation: from mechanisms to therapeutics. Br J Cancer. 2023;129:24–37.

106. Chaudhri A, Lizee G, Hwu P, Rai K. Chromatin remodelers are regulators of the tumor immune microenvironment. Cancer Res. 2024;84:965–76.

107. Krishnamurthy N, Kato S, Lippman S, Kurzrock R. Chromatin remodeling (SWI/SNF) complexes, cancer, and response to immunotherapy. J Immunother Cancer. BMJ; 2022;10:e004669.

108. Huang M, Qian F, Hu Y, Ang C, Li Z, Wen Z. Chromatin-remodelling factor BRG1 selectively activates a subset of interferon-alpha-inducible genes. Nat Cell Biol. Springer Science and Business Media LLC; 2002;4:774–81.

109. Dudek AH, Pfaff F, Bolte H, Waguia Kontchou C, Schwemmle M. Partial inactivation of the chromatin remodelers SMARCA2 and SMARCA4 in virus-infected cells by caspase-mediated cleavage. J Virol. J Virol; 2018;92:e00343–18.

110. Abou El Hassan M, Yu T, Song L, Bremner R. Polycomb repressive complex 2 confers BRG1 dependency on the CIITA locus. J Immunol. Oxford University Press (OUP); 2015;194:5007–13.

111. Papillon-Cavanagh S, Doshi P, Dobrin R, Szustakowski J, Walsh AM. STK11 and

KEAP1 mutations as prognostic biomarkers in an observational real-world lung adenocarcinoma cohort. ESMO Open. Elsevier BV; 2020;5:e000706.

112. Knetki-Wróblewska M, Wojas-Krawczyk K, Krawczyk P, Krzakowski M. Emerging insights into STK11, KEAP1 and KRAS mutations: implications for immunotherapy in patients with advanced non-small cell lung cancer. Transl Lung Cancer Res. Transl Lung Cancer Res; 2024;13:3718–30.

113. Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu, Cancer Genome Atlas Research Network. Integrated genomic characterization of pancreatic ductal adenocarcinoma. Cancer Cell. 2017;32:185–203.e13.

114. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. Cell. 2015;163:1011–25.

115. Cancer Genome Atlas Research Network. Electronic address: elizabeth.demicco@sinaihealthsystem.ca, Cancer Genome Atlas Research Network. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. Cell. 2017;171:950–65.e28.

116. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The international cancer genome consortium data portal. Nat Biotechnol. Springer Science and Business Media LLC; 2019;37:367–9.

117. Kundra R, Zhang H, Sheridan R, Sirintrapun SJ, Wang A, Ochoa A, et al. OncoTree: A Cancer Classification System for Precision Oncology. JCO Clin Cancer Inform. 2021;5:221–30.

118. Park S, Ock C-Y, Kim H, Pereira S, Park S, Ma M, et al. Artificial Intelligence-Powered Spatial Analysis of Tumor-Infiltrating Lymphocytes as Complementary Biomarker for Immune Checkpoint Inhibition in Non-Small-Cell Lung Cancer. J Clin Oncol. 2022;40:1916–28.

119. Kong J, Zhao X, Singhal A, Park S, Bachelder R, Shen J, et al. Prediction of immunotherapy response using mutations to cancer protein assemblies. Sci Adv. 2024;10:eado9746.

120. Park S, Silva E, Singhal A, Kelly MR, Licon K, Panagiotou I, et al. A deep learning model of tumor cell architecture elucidates response and resistance to CDK4/6 inhibitors. Nat Cancer. 2024;5:996–1009.

121. Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, et al. NDEx, the Network Data Exchange. Cell Syst. 2015;1:302–5.

122. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016;770–778.

123. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–47.

124. Gress DM, Edge SB, Greene FL, Washington MK, Asare EA, Brierley JD, et al. Principles of cancer staging. AJCC Cancer Staging Manual. Cham: Springer International Publishing; 2017. page 3–30.

125. Zhao R, Shu Y, Xu W, Jiang F, Ran P, Pan L, et al. The efficacy of immunotherapy in non-small cell lung cancer with KRAS mutation: a systematic review and meta-analysis. Cancer Cell Int. Springer Science and Business Media LLC; 2024;24:361.

Tables

Table 1. Summary of downstream clinical tasks and corresponding datasets.

Datasets used for training

Task	Tissue	No. patients	Source
Immunotherapy	Bladder, Head-Neck, Lung, Melanoma	94	PMID: 30150660
Chemotherapy	Bladder, Breast, Cervical, Esophagogastric, Head-Neck, Lung, Ovarian, Melanoma	237	PMID: 24071849
Metastasis	Lung adenocarcinoma	1,974	PMID: 39506116
Tissue-of-origin	Bladder, Breast, Cervical, Colorectal, Esophagogastric, Head-Neck, Lung, Ovarian, Melanoma	272	PMID: 35121966

Datasets used for testing

Task	Tissue	No. patients	Source
Immunotherapy	Bladder	130	PMID: 29443960
Immunotherapy	Lung	229	PMID: 36038778
Immunotherapy	Melanoma	144	PMID: 31792460
Chemotherapy	Bladder	42	PMID: 25092538
Metastasis	Lung adenocarcinoma	128	PMID: 37084736
Tissue-of-origin	Breast, Colorectal, Lung adenocarcinoma	112	PMID: 28783718

Figures

Fig. 1. Pre-training on large datasets for accurate transfer learning. (A, B) Datasets collected to pre-train MutationProjector, including **(A)** tumor genetic alterations and **(B)** molecular networks. CNA: Copy Number Amplifications, CND: Copy Number Deletions, DDRAM (65): DNA Damage Response Assemblies Map, ISLE (63): Identification of clinically relevant Synthetic LEthality, PCNet (23): Parsimonious Composite Network, SIGNOR (58): SIGnaling Network Open Resource, STRING (66): Search Tool for Retrieval of Interacting Genes/Proteins, TMB: Tumor Mutation Burden, TRRUST (59): Transcriptional Regulatory Relationships Unravelled by Sentence-based Text-mining. **(C)** Model configuration for pre-training. Solid black mutation embedding indicates alteration profiles that are masked. Updated embeddings are generated via network-based message passing using graph attention networks. Size of the gene and covariate embeddings are noted as $d_{features}$. CNA: Copy Number Amplification, CND: Copy Number Deletion, Emb: Embedding, Norm: Normalization, TIL: Tumor Infiltrating Lymphocytes, TMB: Tumor Mutational Burden.

Fig. 2. Self-supervised learning and model pre-training performance. (A) Schema for self-supervised learning. At each epoch, 15% of the genes were randomly selected and masked. Reconstruction loss was computed to measure the accurate reconstruction of the masked genes by MutationProjector models. CNA: Copy Number Amplifications, CND: Copy Number Deletions. **(B)** Self-supervised learning performance in held-out samples and external datasets (n = 10 folds of cross validation, Methods). Performance shown separately for prediction of somatic mutations, CNAs or CNDs. Area under the precision recall curve (AUPRC) was used to quantify performance. The red bar shows the mean across the ten folds. **(C)** Waterfall plot of gene-gene interactions sorted in decreasing order of model attention (average across tumor samples). **(D)** Subnetwork of cell cycle regulatory genes. **(E)** Oncoprints (rows: genes; columns: tumors) showing co-occurrence of *CCND1* and *CDKN2A* copy number events in head-and-neck cancers (top) versus mutual exclusivity of *BRAF* and *NRAS* mutations in melanomas (bottom). Both interactions are significant by Fisher's exact text ($p < 0.05$).

Fig. 3. Exploration of the MutationProjector tumor projection. (A) Representation of tumors using the final layer embedding of MutationProjector, reduced to two dimensions using the UMAP method (arXiv 1802.03426). The points show tumor samples from TCGA (n=3,733), colored based on their annotated cancer type. **(B)** Similarity of MutationProjector embeddings across cancer types. Shown for each pair of cancer types is the average of all cosine similarities between the tumors of each pair. Red boxes indicate three tumor type pairs with highest cross-similarity. **(C)** UMAP highlighting squamous cell carcinomas. **(D)** UMAP highlighting tumors with representative genetic alterations. Mut: somatic mutation. **(E)** Representative tumors with co-alteration of gene pairs learned from graph attention. CND: Copy Number Deletion. (**F**) Mutually exclusive attention weights in bladder cancer involving (left column) *FGFR3* and *CDKN2A* or (right column) *TP53* and *RB1*. **(G)** UMAP highlighting the documented Human Papilloma Virus (HPV) status. **(H)** Similarities of MutationProjector representations. The red box indicates the tumor subtype pair with highest similarity. **(I), (J)** Similarities of MutationProjector embeddings for TCGA tumors annotated as basal versus luminal subtypes by mRNA analysis. **(I)** Bladder cancer (BLCA). **(J)** Breast cancer (BRCA).

Fig. 4. Prediction of drug response and metastasis across cohorts. (A) Schema for performing clinical tasks in test cohorts. **(B)** Predictive performance of immunotherapy response prediction in (left) bladder cancer, (middle) lung cancer and (right) melanoma. Hazard ratio and its 95% confidence interval is shown, with significance p-value measured using the log-rank test. OS: Overall Survival, PFS: Progression-Free Survival. **(C)** Summary of MutationProjector predictive performance results (red) in comparison to selected other modeling architectures and cancer-specific drug biomarkers (dark grey) in the three immunotherapy test cohorts. Hazard ratio and its corresponding 95% confidence interval are shown. Dashed vertical line indicates hazard ratio of 1. PD-L1 expression levels and MicroSatellite Instability status (MSI) are available only for bladder and lung cancer cohorts, respectively. *KRAS* mutation is shown for lung cancer due to its importance in lung cancer immunotherapy (125). TMB: Tumor Mutation Burden. **(D)** Predictive performance of chemotherapy response predictions in bladder cancer patients. **(E)** Summary of predictive performance in the chemotherapy test cohort, as in (C).

Fig. 5. Analysis of key biomarkers used for immunotherapy predictions. (A-C) Top 20 important gene alterations and co-variates for anti-PD1/PDL1 response prediction in **(A)** bladder cancer, **(B)** lung cancer and **(C)** melanoma. Spearman correlation coefficient was used to measure feature importance. Gene and covariate features are colored teal and orange, respectively. **(D)** Alteration patterns in the immunotherapy treated lung cancer cohort. Rows are features (genes or gene pairs), columns are patients. Features with significant alterations in the predicted responders and non-responders are shown in blue and orange, respectively (FDR corrected $P < 0.2$). Mutational frequencies in each group are shown as percentages. **(E)** Top 30 gene pairs with high model attention [average attention weights between genes in (B)] in the immunotherapy treated lung cancer cohort. Red arrows indicate gene pairs with co-alteration patterns observed in non-responders in (D).

Figure 1

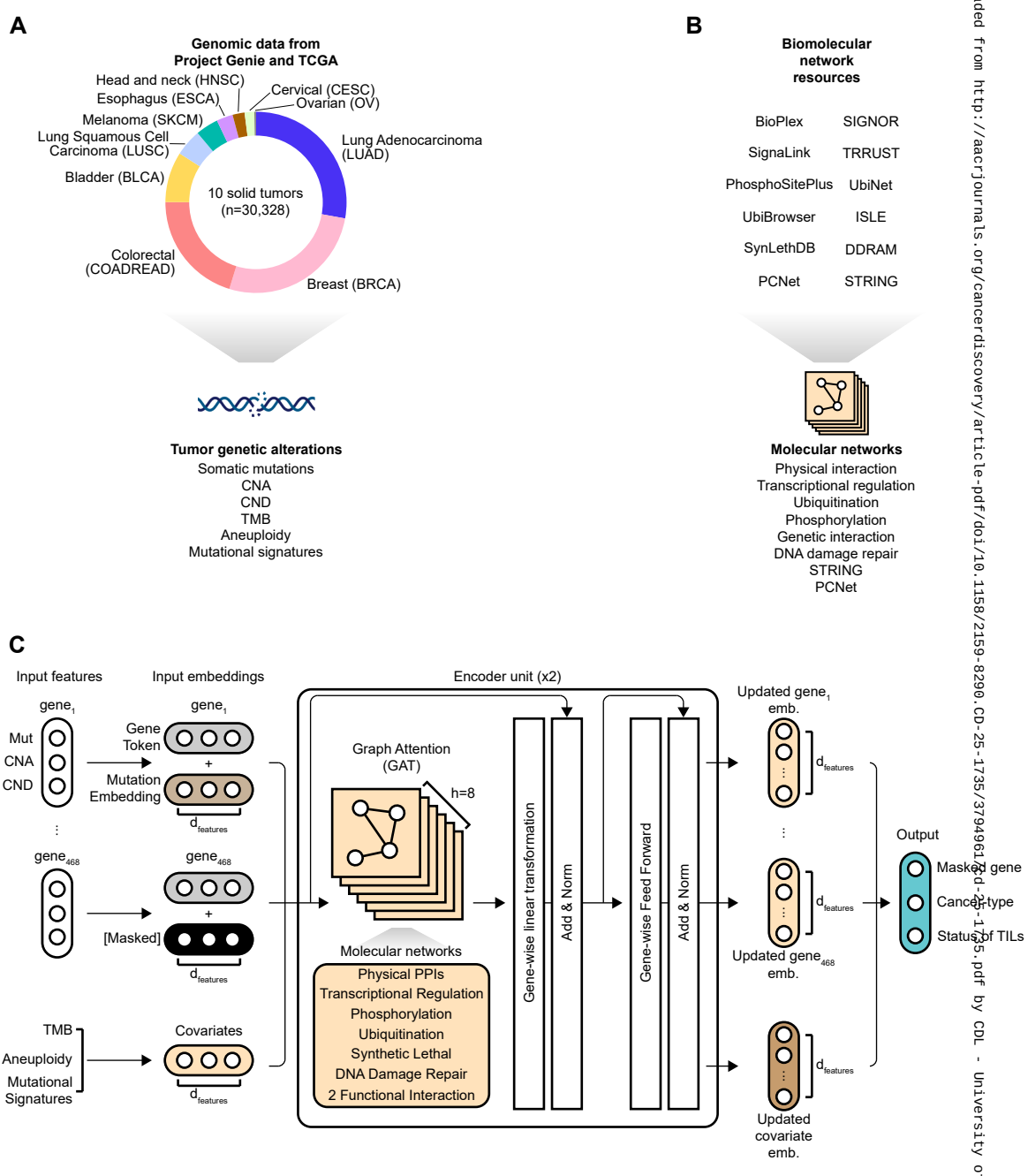
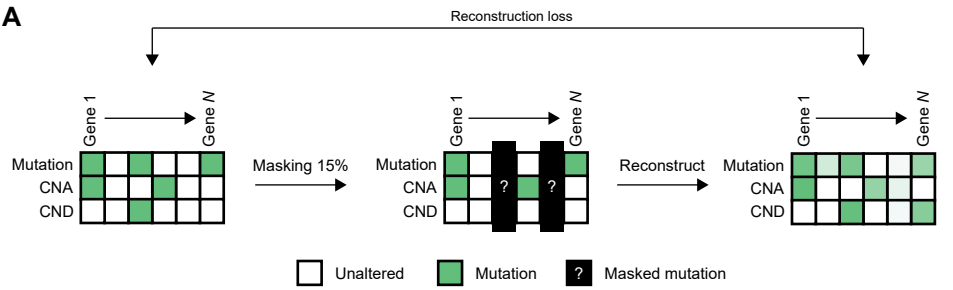
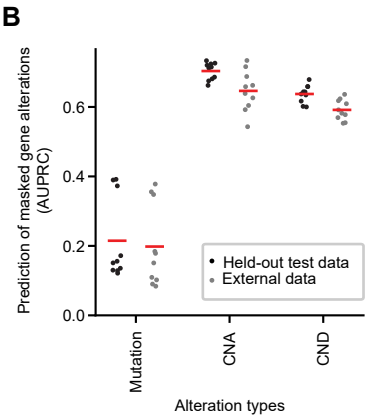
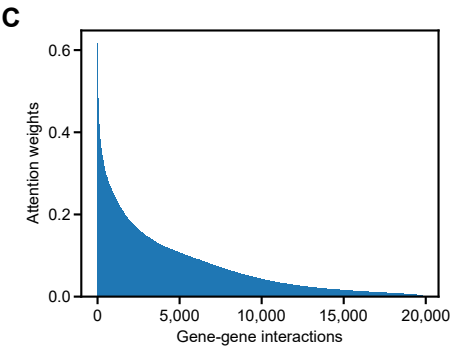
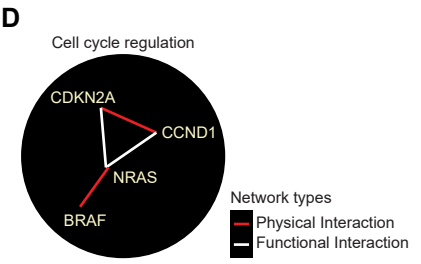


Figure 2









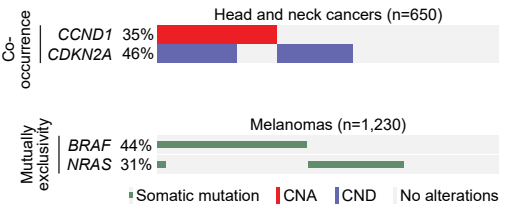


Figure 3

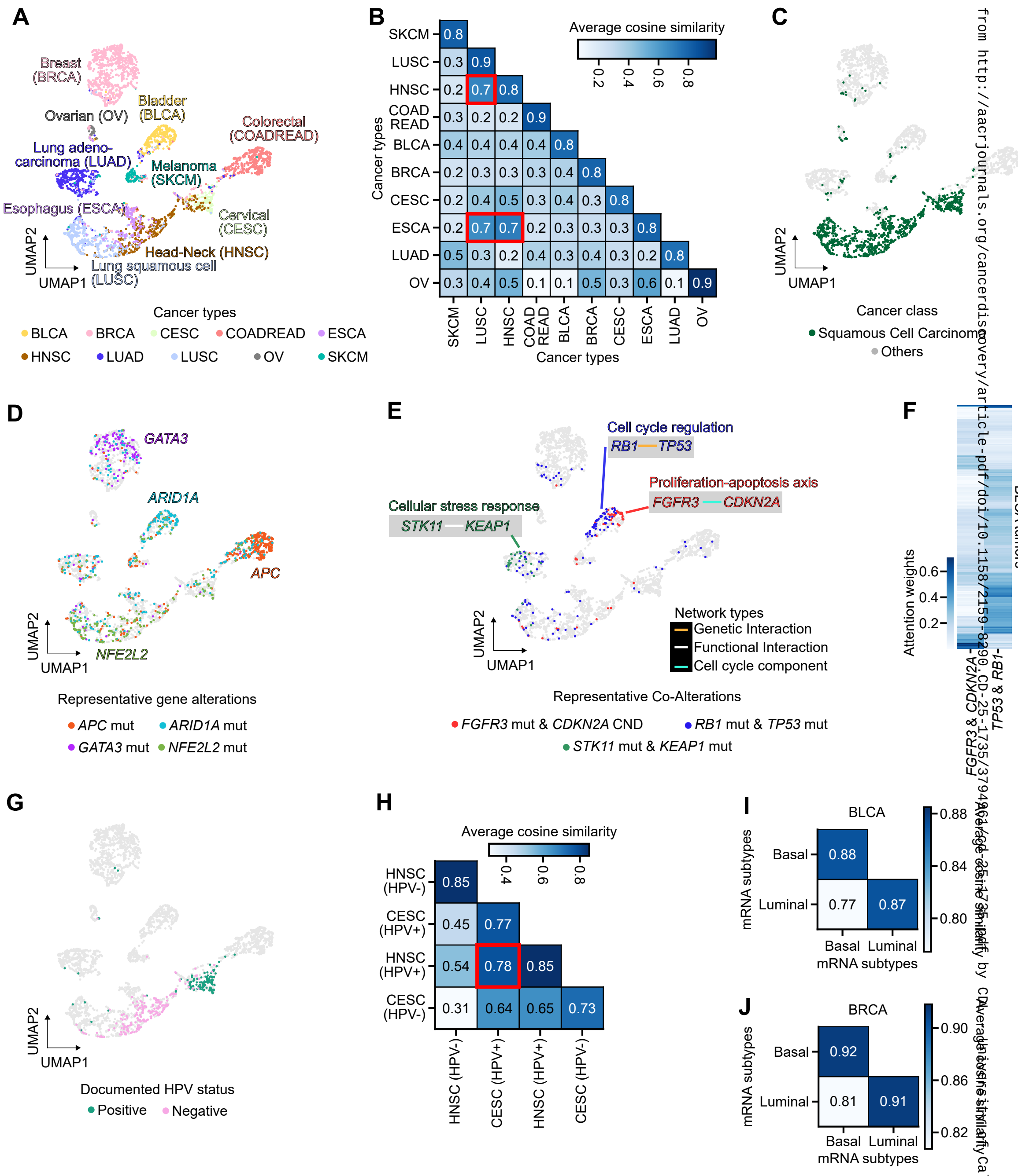


Figure 4

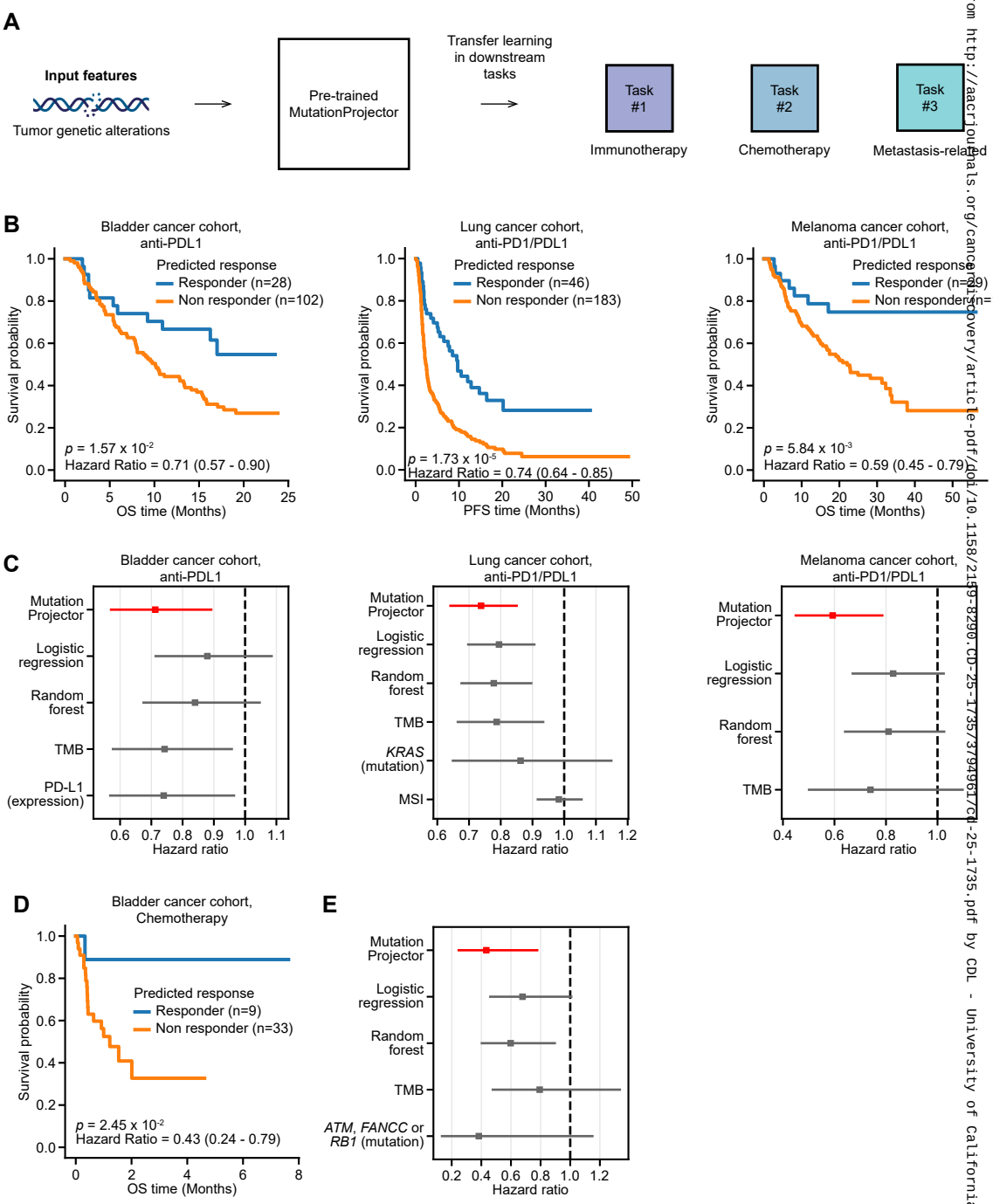


Figure 5