Deep Learning for Biomarker Discovery in Cancer Genomes

Michaela Unger (1), Chiara M. L. Loeffler (1, 2, 3), Laura Žigutytė (1), Srividhya Sainath (1), Tim Lenz (1), Julien Vibert (4), Andreas Mock (5, 6, 7), Stefan Fröhling (6, 7, 8, 9), Trevor A. Graham (10), Zunamys I. Carrero (1), Jakob Nikolas Kather (1, 2, 11, 12)

+ Correspondence to jakob-nikolas.kather@alumni.dkfz.de

1. Else Kroener Fresenius Center for Digital Health, University of Technology Dresden, Dresden, Germany
2. Medical Department 1, University Hospital and Faculty of Medicine Carl Gustav Carus, University of Technology Dresden, Dresden, Germany
3. National Center for Tumor Diseases Dresden (NCT/UCC), a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
4. Drug Development Department (DITEP), Gustave Roussy, Villejuif, France
5. Institute of Pathology, Ludwig-Maximilians-University München, Munich, Germany
6. Division of Translational Medical Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany
7. National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, Heidelberg, Germany
8. German Cancer Consortium (DKTK), Core Center Heidelberg, Heidelberg, Germany
9. Division of Translational Precision Medicine, Institute of Human Genetics, Heidelberg University, Heidelberg, Germany
10. Centre for Evolution and Cancer, Institute of Cancer Research, London, UK
11. National Center for Tumor Diseases Dresden (NCT/UCC), a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, and Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
12. Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Abstract

Background: Genomic data is essential for clinical decision-making in precision oncology. Bioinformatic algorithms are widely used to analyze next-generation sequencing (NGS) data, but they face two major challenges. First, these pipelines are highly complex, involving multiple steps and the integration of various tools. Second, they generate features that are human-interpretable but often result in information loss by focusing only on predefined genetic properties. This limitation restricts the full potential of NGS data in biomarker extraction and slows the discovery of new biomarkers in precision oncology.

Methods: We propose an end-to-end deep learning (DL) approach for analyzing NGS data. Specifically, we developed a multiple instance learning DL framework that integrates somatic mutation sequences to predict two compound biomarkers: microsatellite instability (MSI) and homologous recombination deficiency (HRD). To achieve this, we utilized data from 3,184 cancer patients obtained from two public databases: The Cancer Genome Atlas (TCGA) and the Clinical Proteome Tumor Analysis Consortium (CPTAC).

Results: Our proposed deep learning method demonstrated high accuracy in identifying clinically relevant biomarkers. For predicting MSI status, the model achieved an accuracy of 0.98, a sensitivity of 0.95, and a specificity of 1.00 on an external validation cohort. For predicting HRD status, the model achieved an accuracy of 0.80, a sensitivity of 0.75, and a specificity of 0.86. Furthermore, the deep learning approach significantly outperformed traditional machine learning methods in both tasks (MSI accuracy, p-value = 5.11×10^{-18}; HRD accuracy, p-value = 1.07×10^{-10}). Using explainability techniques, we demonstrated that the model's predictions are based on biologically meaningful features, aligning with key DNA damage repair mutation signatures.

Conclusion: We demonstrate that deep learning can identify patterns in unfiltered somatic mutations without the need for manual feature extraction. This approach enhances the

detection of actionable targets and paves the way for developing NGS-based biomarkers using minimally processed data.

Keywords: AI, Deep Learning, Biomarker, Microsatellite Instability, Homologous Recombination Deficiency, Cancer Genomics

Background

Bulk and targeted sequencing of cancer are being progressively integrated into clinical routine workflows. Today, knowledge about genetic variants or expression profiles are key pillars of personalized oncology [1–5]. Several efforts and programs have been established to analyze patients' cancer genomes, benefiting oncologists' decision-making [6,7]. In molecular tumor boards, genomic biomarkers play a substantial role when determining diagnosis of a patient, hereditary predispositions, treatment scheme and can monitor therapy response, or explain resistances. Two important and highly clinically relevant biomarkers that are characterized by sets of genomic alterations are microsatellite instability (MSI) [8] and homologous recombination deficiency (HRD) [9]. In MSI, the mismatch repair (MMR) system of a cell is impaired, leading to small-scale insertions and deletions (indels). MSI-high tumors are an ideal target for immunotherapy [8,10]. In contrast, HRD introduces a different mutational scar in the cancer genome [11,12]. This makes HRD tumors eligible for treatment with poly ADP ribose polymerase (PARP) inhibitors in certain situations [13,14]. Both MSI and HRD leave distinguishable and complex patterns in the cancer genome.

Since the start of the Human Genome Project [15], bioinformaticians have been involved in designing computational algorithms to analyze genomes. Traditionally, extensive multi-step pipelines process sequencing data for specific applications. For instance, gene expression is determined by mapping transcriptome reads to the human genome and assigning counts to the respective genes [16–20]. Similarly, small- and large-scale mutations are detected by variant callers that compare the sequence of a cancer genome to a reference [21–24]. Information on gene expression or variants in cancer-related genes is then interpreted in the clinical context. However, this approach is fundamentally limited by human expertise. By constraining the analysis of next generation sequencing (NGS) data to pre-defined genomic features, potentially relevant information is neglected. This limitation is particularly evident for compound biomarkers such as MSI and HRD which are reflected by a large spectrum of

changes in the genome and can be defined through a range of molecular assays [25–27]. Therefore, having an objective, unbiased end-to-end tool that directly relates a disease-specific genomic pattern to patient subgroups and outcomes could complement existing bioinformatics methods in precision oncology.

One instrument that could serve such a purpose is deep learning (DL). With increasing computational resources and decreasing sequencing costs [28], DL has been increasingly applied in cancer bioinformatics [29,30]. For instance, DL has been applied to variant calling, such as prognostication or drug response prediction [31–34]. Nevertheless, most of these models still rely on heavily handcrafted features or make strong prior assumptions about feature interactions. Here, we build upon previous attempts to use attention-based multiple instance learning (attMIL) for genomic data analysis with minimal human intervention [35,36]. We extend and apply this approach on unfiltered somatic mutations to predict MSI status in colon, rectum, gastric, and uterine cancers and HRD status in breast, ovarian, prostate, and pancreatic cancers. To our knowledge, this is the first study to apply the attMIL DL method for somatic mutations in multiple clinical cohorts with external validation. Finally, we propose a set of explainability methods to gain insights into the potential biomedical features the model uses for its predictions. In summary, our approach provides an interpretable, genomics-based tool applicable for oncological tasks that could support clinical decision-making.

Methods

Data acquisition

Small-scale somatic mutations in form of single base substitutions (SBSs) and insertions and deletions (indels) of 3,185 whole exome sequenced (WES) cancer patients from The Cancer Genome Atlas Program (TCGA) (n=3,080 patients) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (n=105 patients) were utilized. (**Fig. 1a**) TCGA controlled-access and CPTAC public mutation data were obtained in Mutation Annotation Format (.maf) files through the Genomic Data Commons (GDC) (**Fig. 1b**). In total we acquired 8.3 million mutations for TCGA and 70k mutations for CPTAC COAD.

To reconstruct the mutation sequence context, mutation entries were mapped back to their genomic locations using the 'Chromosome', 'Start_Position', and 'End_Position' columns to generate their sequence context. Human reference genome builds GRCh37 and GRCh38 were used for TCGA and CPTAC data respectively, with chromosome sequences obtained from the Ensembl genome database (**Fig. 1c**).

For prediction of MSI status, four common cancer types were selected: colon adenocarcinoma (COAD) (n=338), rectal adenocarcinoma (READ) (n=110), stomach adenocarcinoma (STAD) (n=432), and uterine corpus endometrial carcinoma (UCEC) (n=450) (**Fig. 1a & S1**), resulting in a dataset of 1,330 patients from TCGA. MSI status was obtained through cBioPortal (https://www.cbioportal.org/) and previous studies conducted on TCGA data [37]. The MSI status of all patients was determined using consensus calls from polymerase chain reaction (PCR) -based assays, MANTIS [38], and MSIsensor [39], with previously described thresholds set at binarized values of 0.4 and 3.5, respectively (**Additional File 1**). For validation, 104 COAD patients from CPTAC were used (**Supplementary Material 1**).

Furthermore, to simulate how our models perform on mutations from targeted sequencing panels, we filtered the WES-derived variant data using the gene lists from FoundationOne CDx (324 genes) [40,41] and TruSight Oncology 500 (523 genes) [42,43]. Variants were filtered by matching their HUGO nomenclature entries to the gene lists. Since the gene panels both include only coding exonic regions, we kept mutations with the following variant classification: Missense_Mutation, Silent, Nonsense_Mutation, Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins. The full CPTAC dataset contained 70,861 variants from 16,478 genes. After filtering by Trusight Oncology 2,718 variants from 432 genes remained after the FoundationOne Dx panel 1,758 variants from 280 genes remained (**Supplementary Material 2**). To match the mutation load of simulated CPTAC panel sequencing with WES, we imputed the mutation distribution by repeating the variants ten times.

For prediction of HRD status, three, cancer types where HRD is prevalent were selected [40]: breast (BRCA) (n=967), ovarian (OV) (n=182), pancreatic (PAAD) (n=139) and prostate cancer (PRAD) (n=462) from TCGA (**Fig. 1a & S2**). The genomic HRD scar score [41] consists of three numerical properties: loss of heterozygosity (LOH) [42], telomeric allelic imbalance (TAI) [43], and large-scale transitions (LST) [44] from which a sum HRD status was determined. Using tissue-specific cutoffs [45], we generated binarized prediction targets with cutoffs set at 45 for BRCA, 54 for OV, 37 for PAAD, and 21 for PRAD [45] (**Additional File 2**). Due to the limited number of HR-deficient samples in CPTAC for these cancer types, external validation was not feasible for HRD predictions.

Data Preparation

In our experiments we compared our DL model to a state-of-the-art (SOTA) machine learning (ML) tool to have a baseline reference regarding performance. Data preparation for both models differed in fundamental parts.

The input data for the DL model consisted of the sequence context surrounding each mutation. Four 20-nucleotide sequences were extracted: the mutation itself, the upstream sequence (5'), the downstream sequence (3'), and the reference sequence at the mutation site. For mutations shorter than 20 nucleotides - such as single-nucleotide substitutions - the remaining positions were padded with gap characters ('-'). Similarly, shorter reference sequences were also padded to ensure consistent length. For all experiments, a window size of 20 nucleotides was defined to capture the specific indel signatures associated with MSI and HRD mechanisms (**Fig. 1d**).

To account for the double stranded nature of DNA, the reverse complement of each mutation sequence was provided as well and processed as blocks as described above. Additionally, a binary indicator was included to specify the DNA strand on which the mutation occurred. All sequences were numerically encoded by mapping nucleotide bases (A, C, G, T) and gap characters ('-').

For the ML model, mutations were represented in their SBS and indel forms, comparable to Catalogue Of Somatic Mutations In Cancer (COSMIC) v3.4 [46] features. Therefore, we used 178 input features for the ML model, 96 for SBS mutations and 83 for indels. Feature values were normalized by mutation count of the respective patient (**Supplementary Methods**).

Data Splitting

To enhance generalizability and mitigate potential site-specific biases, a site-specific training split based on the Tissue Source Site (TSS) codes within the TCGA patient identifiers was conducted [47]. TSS codes were mapped to their respective institutions (**Supplementary Material 3**, **Additional File 3**). The test set for each model comprised patients from specific institutions excluded from the training set to ensure comparable results. The training data was divided for a 5-fold cross-validation in an approximate 70:15:15 split (**Fig. S1 & S2**). Additionally, to this source-specific split, for MSI experiments we validated our findings on

CPTAC. For the training data positive samples for MSI ranged from 20 to 22 % between folds, and from 15 to 17 % for HRD (**Table S1 & S2**).

Baseline Gradient Boosting

To establish a SOTA ML classifier on somatic mutations as baseline for performance, Extreme Gradient Boosting (XGB) models were employed using the scikit-learn v1.2.0 and xgboost v2.1.1 libraries in Python. For MSI and HRD classification tasks, default parameters of the xgb.XGBClassifier() and fit() function were used.

Deep Learning Model Setup

The DL model employed in this study was directly adopted from the implementation provided by Anaya et al. [35], in TensorFlow v2.12.0. The model comprises two main components: a trainable mutation encoder and an attention multiple instance learning (attMIL) module (**Supplementary Methods**) (**Fig. 1d & 1e**).

The mutation encoder consists of a 2D convolutional layer followed by a dense layer, transforming each mutation's sequence context into a feature vector. Convolutions are applied separately to the forward and reverse complement sequences, and the resulting features are concatenated and passed through the dense layer. The output dimensionality is set to 128 (**Fig. 1d**).

In the attMIL module [48], mutation-level feature vectors are aggregated into a patient-level representation using an attention mechanism. Weighted sum pooling is employed, where feature vectors are summed and weighted by their attention scores. The aggregated patient vector is then passed through two dense layers, with the final layer using a sigmoid activation function for binary classification tasks. For MSI and HRD, binary cross-entropy was used as a loss function (**Fig. 1e**).

Models were trained with the Adam optimizer and a batch size of 128. For MSI models, training was conducted for up to 300 epochs with a learning rate of 0.001. HRD models were trained for up to 500 epochs with a learning rate of 0.01. The training and deployment of the models were carried out on an NVIDIA RTX A6000 with 46 GB of RAM.

Model performance was evaluated using accuracy, F1 score, receiver operating characteristic (ROC) area under the curve (AUC), and precision-recall (PR) AUC, specificity, sensitivity and absolute values of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for classification tasks. To estimate significance between model ROC AUCs we utilized the DeLong's test, for accuracy McNemar, for count data (TP, FP, TN, FN) the Wilcoxon Signed-Rank test and for all other metrics a paired t-test was applied. Furthermore, for patient level significance tests, such as DeLong's test and McNemar we aggregated the p-value for all five folds with Fisher's method. To generate a binary classification from continuous prediction values we used the best cutoff of the ROC AUC on the training set.

Explainability Techniques

Several explainability techniques were applied to interpret the model's predictions. This approach allowed us to interpret how the attMIL model encodes and evaluates mutations, as well as how these encodings contribute to patient-level predictions. Dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP) (python library umap-learn v0.5.5), which facilitated the visualization of mutation- and patient-level feature vectors by projecting them into lower-dimensional space to identify sample clusters. Mutation-level features were extracted from a layer that encodes individual mutations into vector representations, while patient-level features were obtained from the aggregation layer, which integrates information across all mutations for a given patient. Mutation encodings were then stratified based on attention value and grouped using k-means clustering (k=7) with the sklearn.cluster library, allowing the analysis of the importance of similar mutations to the model. The choice of k=7 balanced sufficient granularity to capture distinct mutation patterns

with avoiding excessive fragmentation. To further understand the relationship between these clusters and biological processes, mutation catalogs from the clusters were generated and compared to known COSMIC mutational signatures associated with defective mismatch repair (dMMR) and HRD.

Finally, the relationship between the model's prediction scores and various tumor properties was examined for potential trends. For MSI, these included associations with cancer type, tumor mutational burden (TMB), indel mutation density, driver mutations in dMMR genes (e.g., *MLH1*, *MSH2*, *MSH3, MSH6*, *PMS2*) and *POLE*, as well as the methylation status of *MLH1* promoter. For this, TCGA patients included in the model's validation and test sets were selected, and the binarization threshold was determined based on the optimal cutoff for the ROC AUC derived from the training set. For HRD, correlations were evaluated in relation to cancer type, *BRCA1/2* status, continuous genomic scarHRD scores and their components (i.e., LOH, TAI, and LST), TMB, and mutations in additional HR pathway genes (e.g., *ATM, BRIP1, CHEK2, NBN, PALB2, RAD51C*).

For XGB models, Shapley additive explanations (SHAP) values were calculated with the python library shap (v0.45.1) for each input feature. By default, 20 most important features are ranked by SHAP value and displayed in increasing order.

Results

DL detects microsatellite instability from sequencing panels

MSI arises because of defects in the mismatch repair (MMR) pathway [49,50], and leads to characteristic small-scale indels in repetitive sequences as well as distinctive patterns of SBSs (**Fig. 2a**), which in principle are recognizable by DL-based models [11,51–53]. We trained an attMIL model to detect MSI status in 1,330 patients from TCGA, including patients from COAD, READ, STAD and UCEC. After training, we externally validated this model on the COAD cohort from CPTAC (n=105 patients). The attMIL model achieved an accuracy of 0.98±0.01, a ROC AUC of 1.00±0.00, sensitivity of 0.95±0.02 and a specificity of 1.00±0.00 confirming its ability to detect MSI-related patterns (**Table S3**). When analyzing the absolute number of misclassified patients across 5-fold cross-validation (ensemble of 5 models), we observed no false positive patients and only 1.20±0.45 false negatives in the total population of 105 patients. These results were based on an optimal threshold determined in the training set. For real-world application, the threshold could be adapted to yield a high positive predictive value. For example, in the external test set, a threshold corresponding to a sensitivity of 0.99 would yield a specificity of 0.99±01, which is much higher than for many clinically used tests.

Next, we compared the DL performance to that of a SOTA ML model to investigate whether DL is indeed more powerful than standard ML for this application. As expected, XGB also achieved excellent performance on the CPTAC datasets, with an accuracy of 0.89±0.07, a ROC AUC of 0.99±0.01, sensitivity of 0.99±0.02 and a specificity of 0.80±0.16 (**Fig. 2b, Table S3**). However, a significant difference in performance between DL and ML models was seen when comparing F1 scores (DL 0.97±0.01; ML 0.76±0.15) with a p-value of 0.04 (**Table S3**). Furthermore, when investigating misclassified patients, we found that the XGB model produces a large amount of false positive patients (16.40±12.93 of 105 patients). These data

show that MSI prediction is a relatively easy task which even classical ML solves well, however, our new DL approach was still substantially, and clinically meaningfully, better.

In the real world, most patients do not undergo whole exome sequencing (WES), but just panel sequencing of a few hundred genes. We evaluated the performance of DL and ML models on pseudo panel sequencing data. To this end, the number of somatic mutations was filtered to include only genes present in two targeted sequencing panels: FoundationOne Dx [54,55] and TruSight Oncology [56,57]. Reducing the mutation set to the TruSight Oncology 523 gene panel revealed a significant divergence in performance between DL and ML models, with the attMIL model achieving an accuracy of 0.96±0.04 outperforming XGB's 0.61±0.05 (p-value 1.67×10^{-38}) and a F1 score of attMIL of 0.96±0.04 compared to 0.50±0.02 of XGB (p-value 7.56×10^{-5}) (**Fig. 2b**). As expected, the task became more challenging with fewer mutations available, indicated by lower overall accuracy - however, the DL model maintained a reasonably good performance. In addition, we found that especially the XGB model had reduced performance regarding specificity, from 0.80±0.16 to 0.54±0.28, when lowering the number of variants for model input, due to high false positive number (24.60±15.26 of 105 patients) (**Table S4**). This suggests that the classical ML model is considerably biased towards positive predictions. When the genes were further narrowed to 324 from the FoundationOne Dx panel, performance differences persisted despite increased variance (accuracy attMIL 0.83±0.10 vs. XGB 0.59±0.06, p-value 7.94×10^{-21}; F1 scores attMIL 0.78±0.17 vs. XGB 0.47±0.11, p-value 0.06) (**Fig. 2b, Table S5**). These results indicate that the DL model outperforms classical ML in clinically relevant genomic subsets present in commercial sequencing panels.

DL detects biomedically relevant MSI patterns

Subsequently, we investigated if the feature representations obtained by the DL model result in a clinically relevant clustering of patients, which would provide further proof that the DL model learned clinically relevant patterns. We investigated a possible clustering of at both the

mutation and patient levels (**Fig. 2c**). While no substantial differences were apparent at the mutation level, a clear separation emerged at the patient level. MSI and MSS cancers segregated distinctively, indicating that the DL model was able to encode their respective patients differently (**Fig. 2c**).

We then evaluated which mutations were most influential for the model's predictions by extracting mutation encodings from the embedding layer. These encodings were clustered using k-means (k=7) and ranked by the mean attention values of their mutations to identify highly and minimally influential groups (**Fig. 2c**). When examining the mutation catalogs of the clusters with the highest attention scores (clusters II+III), they were enriched in C and G indels (**Fig. 2c**). Comparing this to the indel mutation signature ID7 [46] (**Fig. 2a**) associated with MSI, which depicts deletions of mononucleotide stretches but also of dinucleotide repeats, this suggests that the model is capturing relevant MSI-associated mutational patterns.

When analyzing a mutation cluster with average attention scores (cluster I), we observed several similarities to the SBS6 and SBS15 mutation signatures associated with defective MMR (dMMR) [46] (**Fig. S3a**). Similarities encompassed the features of the C>T mutation class, with comparable peaks in the sequence context pairs of an up- and downstream GG, AG and CG. This could further indicate that the model may have learned to group and assign weights to MSI-specific mutations. Other SBS clusters mostly contain similar mutations with C>A mutations in a TT context (**Fig. S3b**). Comparing this to the Shapley additive explanation (SHAP) analysis of the handcrafted features of XGB, XGB also utilizes deletions of a single C/G similar to the DL model (**Fig. S4**). However, other features that strongly influenced the XGB model's predictions, such as indels at repetitive sites or certain C>G/C>A mutations, did not resemble classical dMMR signatures, suggesting either the identification of novel patterns or a potential bias towards irrelevant features.

DL identifies MSI-associated patterns beyond mutation counts

Finally, we investigated whether the model's predictions were driven by subtle patterns within the mutations or if it was simply confounded by factors such as mutation counts. To this end, we investigated if the model's predictions are associated with other relevant properties of the patients' cancers related to MSI (Fig. 2d). These properties included the ground truth label, features of genomic instability, driver mutations, and the methylation status of MMR-related genes. The model's normalized prediction scores were highly indicative of MSI and MSS status across all cancer types, effectively reflecting the underlying distinction between the two groups. Next, as expected, a correlation trend between the prediction score and TMB and indel density was observed. However, patients with high TMB/indel count in an MSS context were not predicted as MSI by the DL model. This suggests that the model learns subtle, clinically relevant patterns and does not simply count mutations as indicators for MSI. We also observed that driver mutations in key MMR pathway proteins are not necessarily indicative of a positive model prediction, nor were they always correlated with consensus labels from MANTIS, MSIsensor or PCR (Fig. 2d). A close relationship was found between the model scores and the MLH1 promoter methylation, as silencing of this gene has a more significant impact on MMR function than mutations alone [58].

These results emphasize DL's potential as an alternative tool for analyzing NGS data, specifically for panel sequencing data. Our results demonstrate high performance generalizability, and explainability of the DL-based attMIL model.

DL predicts HRD and highlights mutational patterns in alternative repair pathways

We next investigated HRD, which arises from defects in key homologous recombination (HR) pathway genes, as well as from other associated genomic defects. When HR fails to repair double-strand breaks (DSBs), cells rely on alternative pathways, such as microhomology-

mediated end joining (MMEJ), which in turn induce specific mutational patterns we detected with our DL model. (**Fig. 3a**) [59,60]. We trained an attMIL model to detect HRD using data from 1,750 patients across BRCA, OV, PAAD, and PRAD cohorts from TCGA, with internal validation performed in a source-site-specific manner. After 5-fold cross-validation, the attMIL model performed slightly better than XGB (attMIL accuracy: 0.80±0.01, XGB accuracy: 0.75±0.02 (p-value 1.07×10^{-10}); attMIL ROC AUC: 0.88±0.01, XGB ROC AUC: 0.86±0.01 (p-value 0.72); DL PR AUC: 0.67±0.01, XGB PR AUC: 0.75±0.02) (**Fig. 3b, Fig. S6a**)). Regarding specificity, both models performed well with (attMIL 0.86±0.05, XGB 0.98±0.00 (p-value 0.049)), however sensitivity was comparably low with 0.74±0.06 for attMIL and 0.52±0.03 for XGB (p-value 0.003), indicating that predicting HRD is a more challenging task than predicting MSI (**Table S6**).

We then investigated the explainability of DL-based HRD predictions. The patient level UMAP shows a gradient between HRD and HR- proficient (HRP) patients, but no clear separation as in the MSI case (Fig. 3c). Furthermore, the mutation level features for HRD and HRP also cluster very much together. In the k-mean clustering heatmap of mutation features, we again identified groups of mutations with high attention scores. The second highest attention group (cluster II) consists of indel mutations, primarily monomer stretch insertions and deletions of C and G, as well as various other indels in repetitive regions. Notably, the highest attention scores (cluster III) were assigned to microhomologies and deletions at 5-nucleotide repeats, consistent with indel signatures ID6 [46] (Fig. 3a) and ID8 [46] (Fig. S4a). Both signatures are associated with alternative repair pathways to HRD, namely MMEJ and non-homologous end joining (NHEJ). A group of mutations with mean attention attribution (cluster I) displays mostly SBSs of various classes related to SBS3 [46], reflecting the general high mutation load of HRD cancers (Fig. S4a). These findings are in line with the highly ranked indel features of the XGB model, which mostly consist of microhomologies or deletions of single C/G nucleotides (Fig. S4b). Additionally, the XGB model identified several C>G mutations as important predictors, which, to our knowledge, previously have not been brought into context with HRD.

DL predicts HRD with biologically meaningful features

Following the same approach as before with a MSI, we again compared the model's predictions to target-specific tumor properties. (**Fig. 3d**). We observed that higher prediction scores correlated with a greater likelihood of patients being HRD-positive, suggesting that the model's prediction scores reflect a ranking of HRD probability among samples. Upon examining the model's prediction across cancer types, we noted that the proportion of patients predicted as HRD varied: approximately 1:4 BRCA patients and 3:5 OV patients were predicted as HRD positive, whereas there were no positive cases in PAAD and only 1:20 in PRAD, aligning with the distribution of the HRD status of patients in these cancer types (**Table S2**). It remains unclear whether this is a result of the model learning to differentiate tissue types or solely predicting HRD status. The model predictions were not correlated to LST, LOH, TAI individually, but only to the HRD scar signature, which is composed of all three components [45]

In summary, we show that deep learning models can effectively predict key biomarkers like MSI and HRD across cancer types while using biologically meaningful features and generalizing well to external cohorts, outperforming traditional machine learning approaches, especially in panel sequencing data.

Discussion

In this study, we demonstrated that DL models can learn to predict clinically relevant biomarkers solely from sequences of small-scale somatic mutation of cancer genomes, without using any prior assumptions or expert knowledge. By comparing an attMIL DL framework with a XGB ML model, we found that for both targets, MSI and HRD, DL outperforms ML. The DL model was explainable and specific, biologically plausible, genomic features were associated with known mutational patterns. Although this model still relies on variants called by classical bioinformatic tools as input, it represents a first step towards end-to-end DL workflows for cancer genomes.

Our findings indicate that DL models offer advantages over traditional ML approaches. For MSI prediction, the DL model outperformed an ML model, especially when applied to limited data during inference. While both models performed similarly on full datasets, the DL model maintained excellent performance even with targeted sequencing panels, such as the TruSight Oncology 500 gene panel. This suggests that DL models are more robust to data sparsity and could streamline clinical workflows by reducing the need for additional MSI testing. In addition, as sequencing panels are widely used in clinical routine, DL models performing well not only on WES/WGS data, could help bring genomic advancements to a broader population.

In predicting HRD, the DL model was slightly better than that of the ML model with a performance comparable to other tools [61–63]. HRD is primarily characterized by large-scale genomic alterations rather than small indels alone. This suggests that small-scale somatic mutations may not provide sufficient information for accurate HRD prediction. Incorporating further information about the sequencing data, such as variant allele frequency, genomic region or affected gene as well as data on structural variations and large-scale rearrangements could enhance model performance.

DL models excel in adapting to diverse datasets without predefined features, making them ideal for the heterogeneity of tumor genomes. Unlike classical ML models, they process sequencing data end-to-end, uncovering both known and novel patterns.

This study is not without limitations. External validation for HRD predictions was not feasible due to limited data availability, restricting the generalizability of our findings. Furthermore, in the case of MSI, DL predictions based on NGS and PCR labels should be compared to the histological approach of diagnosis - immunohistochemistry. Generally, small cohort sizes and class imbalances may have introduced biases to our predictions. Unlike most studies, we meticulously attempted to mitigate biases through site-aware data splitting [47], confounding factors cannot be entirely excluded. Future studies should aim to replicate and validate our findings in larger and more diverse patient cohorts to establish robustness.

As sequencing technologies advance, end-to-end DL approaches could become integral to clinical genomics, particularly when WES/WGS data is unavailable and analysis relies on routinely used targeted sequencing panels. These models hold the potential to streamline workflows by bypassing lengthy preprocessing pipelines, possibly accelerating time to diagnosis and treatment, and with this might reduce costs. For this purpose, future advancements could focus on incorporating raw sequencing reads directly into DL models, making them fully end-to-end. Finally, the incorporation of multi-omics data - whole genome/exome, RNA and bisulfite sequencing - could enable DL to fully capture the complexity and heterogeneity inherent in cancer genomes [64].

Conclusion

In summary, our study displays that DL models can predict clinically relevant biomarkers from genomic data and while capturing complex mutational patterns associated with MSI and HRD. By learning to identify and prioritize mutations influencing a given phenotype, DL models reduce the need for human intervention and complement human expertise in precision oncology. We provide an open-source toolkit to enable reproducibility and broader application of these methods across cancer types and other diseases. Future efforts should focus on expanding datasets, integrating raw multi-omics data, and refining model architectures to maximize the impact of DL in advancing personalized medicine.

Abbreviations

AI – Artificial Intelligence

AUC – Area under the curve

AUROC – Area under the receiver operating characteristic

BRCA – Breast invasive carcinoma

COAD – Colon adenocarcinoma

CRC – Colorectal carcinoma

DL – Deep learning

HRD – Homologous Recombination Deficiency

HRP – Homologous Recombination Proficiency

LUAD – Lung adenocarcinoma

LUSC – Lung squamous cell carcinoma

MLP – Multilayer perceptron

MMEJ – Microhomology-mediated end joining

(d)MMR – (Defective) mismatch repair

MIL – Multiple instance learning

MSI – Microsatellite instability

MSS – Microsatellite stability

MUT – Mutated

NHEJ – Non-homologous end joining

NGS – Next-generation sequencing

PAAD – Pancreatic adenocarcinoma

PRAD – Prostate adenocarcinoma

OV – Ovarian serous cystadenocarcinoma

TCGA – The Cancer Genome Atlas

READ – Rectal adenocarcinoma

ROC – Receiver operating characteristic

UCEC – Uterine corpus endometrial carcinoma

WGS – Whole genome sequencing

WT – Wildtype

WXS – Whole exome sequencing

XGB – Extreme gradient boosting

References

1. Andre F, Filleron T, Kamal M, Mosele F, Arnedos M, Dalenc F, et al. Genomics to select treatment for patients with metastatic breast cancer. Nature. 2022;610:343–8.

2. Mateo J, Steuten L, Aftimos P, André F, Davies M, Garralda E, et al. Delivering precision oncology to patients with cancer. Nat Med. 2022;28:658–65.

3. Hodder A, Leiter SM, Kennedy J, Addy D, Ahmed M, Ajithkumar T, et al. Benefits for children with suspected cancer from routine whole-genome sequencing. Nat Med. 2024;30:1905–12.

4. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463:191–6.

5. Horak P, Heining C, Kreutzfeldt S, Hutter B, Mock A, Hüllein J, et al. Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. Cancer Discov. 2021;11:2780–95.

6. Sosinsky A, Ambrose J, Cross W, Turnbull C, Henderson S, Jones L, et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. Nat Med. 2024;30:279–89.

7. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–93.

8. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science. 2017;357:409–13.

9. Chopra N, Tovey H, Pearson A, Cutts R, Toms C, Proszek P, et al. Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative

1 breast cancer. Nat Commun. 2020;11:2662.

2 10. Germano G, Lamba S, Rospo G, Barault L, Magrì A, Maione F, et al. Inactivation of DNA
3 repair triggers neoantigen generation and impairs tumour growth. Nature. 2017;552:116–20.

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

6 12. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, et al. A mutational
7 signature reveals alterations underlying deficient homologous recombination repair in breast
8 cancer. Nat Genet. 2017;49:1476–86.

9 13. Moore Kathleen, Colombo Nicoletta, Scambia Giovanni, Kim Byoung-Gie, Oaknin Ana,
10 Friedlander Michael, et al. Maintenance Olaparib in Patients with Newly Diagnosed
11 Advanced Ovarian Cancer. N Engl J Med. 2018;379:2495–505.

12 14. Tutt ANJ, Garber JE, Kaufman B, Viale G, Fumagalli D, Rastogi P, et al. Adjuvant
13 Olaparib for Patients with BRCA1- or BRCA2-Mutated Breast Cancer. N Engl J Med.
14 2021;384:2394–405.

15 15. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial
16 sequencing and analysis of the human genome. Nature. 2001;409:860–921.

17 16. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat
18 Rev Genet. 2009;10:57–63.

19 17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
20 universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

21 18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying
22 mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.

23 19. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq

quantification. Nat Biotechnol. 2016;34:525–7.

20. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.

21. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

22. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.

23. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics. 2011;27:1922–8.

24. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.

25. Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite instability: a review of what the oncologist should know. Cancer Cell Int. 2020;20:16.

26. Dedeurwaerdere F, Claes KB, Van Dorpe J, Rottiers I, Van der Meulen J, Breyne J, et al. Comparison of microsatellite instability detection by immunohistochemistry and molecular techniques in colorectal and endometrial cancer. Sci Rep. 2021;11:12880.

27. Stewart MD, Merino Vega D, Arend RC, Baden JF, Barbash O, Beaubier N, et al. Homologous Recombination Deficiency: Concepts, Definitions, and Assays. Oncologist. 2022;27:167–74.

28. Kris A. Wetterstrand MS. DNA sequencing costs: Data [Internet]. Genome.gov. NHGRI;

2019 [cited 2024 Sep 9]. Available from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

29. Unger M, Kather JN. Deep learning in cancer genomics and histopathology. Genome Med. 2024;16:44.

30. Unger M, Kather JN. A systematic analysis of deep learning in genomics and histopathology for precision oncology. BMC Med Genomics. 2024;17:48.

31. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.

32. Chiu Y-C, Chen H-IH, Zhang T, Zhang S, Gorthi A, Wang L-J, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genomics. 2019;12:18.

33. Elmarakeby HA, Hwang J, Arafeh R, Crowdis J, Gang S, Liu D, et al. Biologically informed deep neural network for prostate cancer discovery. Nature. 2021;598:348–52.

34. Friedman S, Gauthier L, Farjoun Y, Banks E. Lean and deep models for more accurate filtering of SNP and INDEL variant calls. Bioinformatics. 2020;36:2060–7.

35. Anaya J, Sidhom J-W, Mahmood F, Baras AS. Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. Nat Biomed Eng. 2024;8:57–67.

36. Sanjaya P, Maljanen K, Katainen R, Waszak SM, Genomics England Research Consortium, Aaltonen LA, et al. Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping. Genome Med. 2023;15:47.

37. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.

38. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, et al. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget. 2017;8:7452–63.

39. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics. 2014;30:1015–6.

40. Nguyen L, W M Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous recombination deficiency. Nat Commun. 2020;11:5584.

41. Sztupinszki Z, Diossy M, Krzystanek M, Reiniger L, Csabai I, Favero F, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. NPJ Breast Cancer. 2018;4:16.

42. Abkevich V, Timms KM, Hennessy BT, Potter J, Carey MS, Meyer LA, et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. Br J Cancer. 2012;107:1776–82.

43. Birkbak NJ, Wang ZC, Kim J-Y, Eklund AC, Li Q, Tian R, et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer Discov. 2012;2:366–75.

44. Popova T, Manié E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Res. 2012;72:5454–62.

45. Rempel E, Kluck K, Beck S, Ourailidis I, Kazdal D, Neumann O, et al. Pan-cancer analysis of genomic scar patterns caused by homologous repair deficiency (HRD). NPJ Precis Oncol. 2022;6:36.

46. Cosmic. COSMIC [Internet]. 2020 [cited 2024 Jul 30]. Available from:

https://cancer.sanger.ac.uk/signatures/id/

47. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021;12:4423.

48. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning. PMLR; 10--15 Jul 2018. p. 2127–36.

49. Cortes-Ciriano I, Lee S, Park W-Y, Kim T-M, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun. 2017;8:15180.

50. Dámaso E, Castillejo A, Arias MDM, Canet-Hermida J, Navarro M, Del Valle J, et al. Primary constitutional MLH1 epimutations: a focal epigenetic event. Br J Cancer. 2018;119:978–87.

51. Aaltonen LA, Peltomäki P, Leach FS, Sistonen P, Pylkkänen L, Mecklin JP, et al. Clues to the pathogenesis of familial colorectal cancer. Science. 1993;260:812–6.

52. Guo Q, Househam J, Lakatos E, Nowinski S, Al Bakir I, Grant H, et al. Long deletion signatures in repetitive genomic regions track somatic evolution and enable sensitive detection of microsatellite instability [Internet]. Bioinformatics. bioRxiv; 2024. Available from: https://www.biorxiv.org/content/10.1101/2024.10.03.616572v1

53. Meier B, Volkova NV, Hong Y, Schofield P, Campbell PJ, Gerstung M, et al. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. Genome Res. 2018;28:666–75.

54. TUMOR TYPES BIOMARKERS FDA-APPROVED THERAPY. Table 1: Companion diagnostic indications [Internet]. [cited 2024 Nov 15]. Available from: https://www.foundationmedicine.com/sites/default/files/media/documents/2024-

07/F1CDxTechnical_Specifications_Commercial_SPEC-01197_V3.0.pdf

55. 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. 2021;26:e588–96.

56. TruSight Oncology 500 [Internet]. [cited 2024 Nov 15]. Available from: https://emea.illumina.com/products/by-type/clinical-research-products/trusight-oncology-500.html#tabs-48377e6865-item-d4e16e33c1-documentation

57. Zhao C, Jiang T, Hyun Ju J, Zhang S, Tao J, Fu Y, et al. TruSight Oncology 500: Enabling Comprehensive Genomic Profiling and Biomarker Reporting with Targeted Sequencing [Internet]. Bioinformatics. bioRxiv; 2020. Available from: https://www.biorxiv.org/content/10.1101/2020.10.21.349100v1

58. Sinicrope FA, Sargent DJ. Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. Clin Cancer Res. 2012;18:1506–12.

59. Black SJ, Ozdemir AY, Kashkina E, Kent T, Rusanov T, Ristic D, et al. Publisher Correction: Molecular basis of microhomology-mediated end-joining by purified full-length Polθ. Nat Commun. 2020;11:1831.

60. Bennardo N, Cheng A, Huang N, Stark JM. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. PLoS Genet. 2008;4:e1000110.

61. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med. 2017;23:517–25.

62. Lee JJ, Kang HJ, Kim D, Lim SO, Kim SS, Kim G, et al. expHRD: an individualized, transcriptome-based prediction model for homologous recombination deficiency assessment

in cancer. BMC Bioinformatics [Internet]. 2024;25. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-024-05854-y#Sec2

63. Abbasi A, Steele CD, Bergstrom EN, Khandekar A, Farswan A, Mckay RR, et al. Detecting HRD in whole-genome and whole-exome sequenced breast and ovarian cancers. medRxiv [Internet]. 2024; Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11261949/

64. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell. 2022;40:1095–110.

Declarations

Ethics approval and consent to participate

This study was carried out in accordance with the Declaration of Helsinki. Datasets from CPTAC and TCGA do not require formal ethics approval for a retrospective study since samples are already anonymised. Furthermore, the analysis was approved by the Ethics commission of the Medical Faculty of the Technical University Dresden (BO-EK-444102022).

Consent for publication

Not applicable.

Availability of data and materials

The mutation data used in this study can be accessed through the following sources: TCGA data is available at https://gdc.cancer.gov/about-data/publications/mc3-2017, and CPTAC data can be found at https://portal.gdc.cancer.gov/. TCGA access can be granted through eRA Commons (https://public.era.nih.gov/commonsplus/home.era?menu_itemPath=600) and dbGaP (https://www.ncbi.nlm.nih.gov/gap/). Chromosome information is accessible at https://ftp.ensembl.org/pub/. Information regarding MSI and HRD status, as well as mutation statuses of driver genes, is available through cBioPortal at https://www.cbioportal.org/.

Scripts for data preparation and model implementation are hosted on GitHub and can be found at https://github.com/mxsunc/Biomarker-ATGC.

Competing Interests

JNK declares consulting services for Bioptimus, France; Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; AstraZeneca, UK; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI GmbH, Germany, Synagen GmbH, Germany; has

received a research grant by GSK; and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. No other competing financial interests are declared by any of the remaining authors. **TG** is named as a coinventor on patent applications that describe a method for TCR sequencing (GB2305655.9), and a method to measure evolutionary dynamics in cancers using DNA methylation (GB2317139.0). TG has received honorarium from Genentech and consultancy fees from DAiNA therapeutics. The remaining authors have no competing interests to declare.

Funding

JNK is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (TransplantKI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

MU and JNK conceptualized the study. MU performed the DL experiments and explainability analysis. MU prepared the original draft and created the figures. JNK acquired funding and

supervised the study. MU, CMLL, LZ, SS, TL, JV, AM, SF, TG, ZIC, JNK provided scientific input, reviewed and edited the manuscript.

Acknowledgements

We acknowledge the TCGA (https://www.cancer.gov/tcga) and the CPTAC (https://proteomics.cancer.gov/programs/cptac) research networks, which generated the data on which the results shown in this manuscript are based on.

In accordance with the COPE (Committee on Publication Ethics) position statement of 13 February 2023 (https://publicationethics.org/cope-position-statements/ai-author), the authors hereby disclose the use of the following artificial intelligence models during the writing of this article: GPT-4 (OpenAI) for checking spelling and grammar.

Figure 1

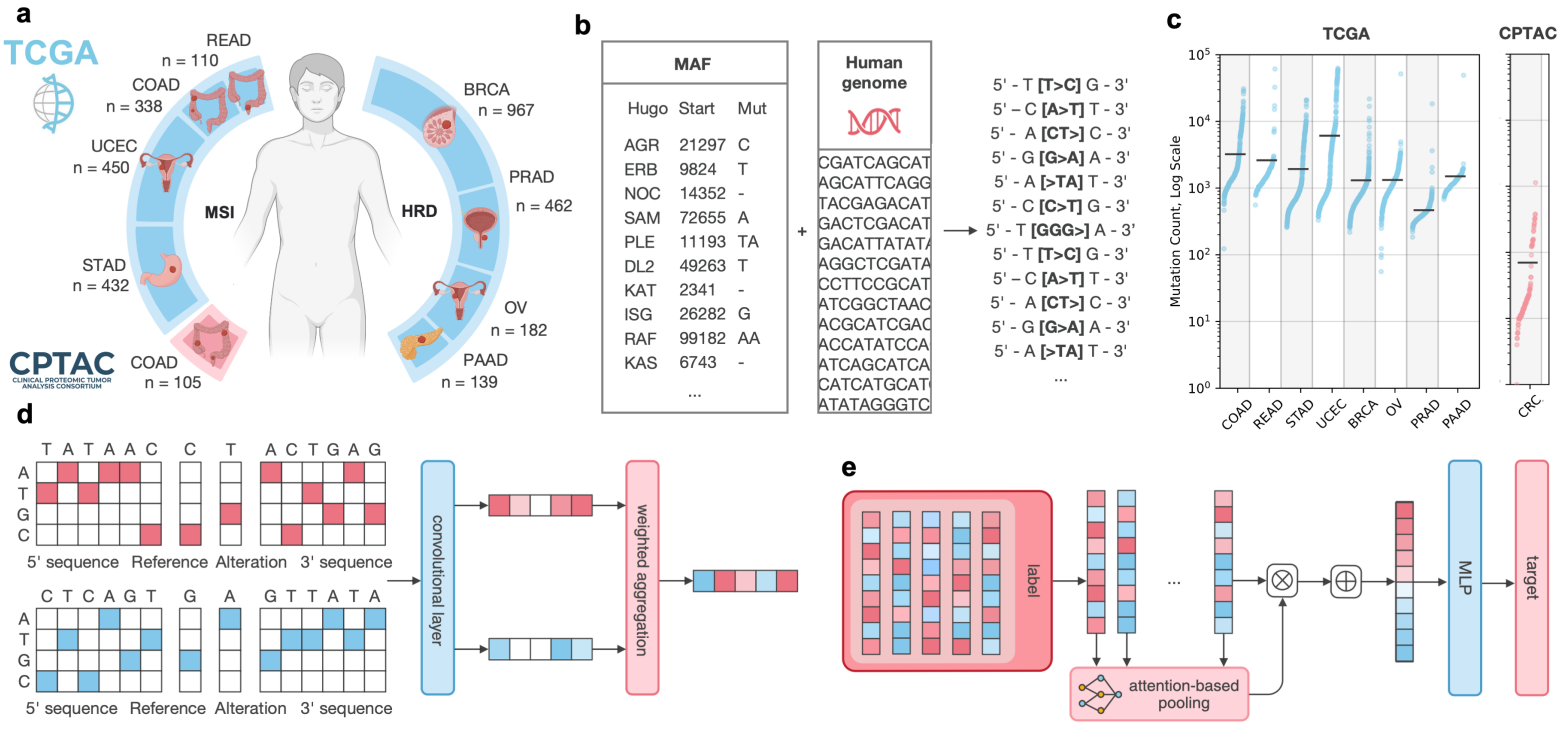
a
TCGA
READ
n = 110
COAD
n = 338
UCEC
n = 450
STAD
n = 432
COAD
n = 105
MSI
HRD
BRCA
n = 967
PRAD
n = 462
OV
n = 182
PAAD
n = 139
CPTAC
CLINICAL PROTEOMIC TUMOR ANALYSIS CONSORTIUM
b
MAF
Hugo Start Mut
AGR 21297 C
ERB 9824 T
NOC 14352 -
SAM 72655 A
PLE 11193 TA
DL2 49263 T
KAT 2341 -
ISG 26282 G
RAF 99182 AA
KAS 6743 -
...
Human genome
CGATCAGCAT
AGCATTCAGG
TACGAGACAT
GACTCGACAT
GACATTATAT
AGGCTCGATA
CCTTCCGCAT
ATCGGCTAAC
ACGCATCGAC
ACCATATCCA
ATCAGCATCA
CATCATGCAT
ATATAGGGTC
5' - T [T>C] G - 3'
5' – C [A>T] T - 3'
5' - A [CT>] C - 3'
5' - G [G>A] A - 3'
5' - A [>TA] T - 3'
5' - C [C>T] G - 3'
5' - T [GGG>] A - 3'
5' - T [T>C] G - 3'
5' – C [A>T] T - 3'
5' - A [CT>] C - 3'
5' - G [G>A] A - 3'
5' - A [>TA] T - 3'
...
c
TCGA
CPTAC
Mutation Count, Log Scale
COAD
READ
STAD
UCEC
BRCA
OV
PRAD
PAAD
CRC
d
5' sequence
Reference
Alteration
3' sequence
convolutional layer
weighted aggregation
e
label
attention-based pooling
MLP
target

Figure 2

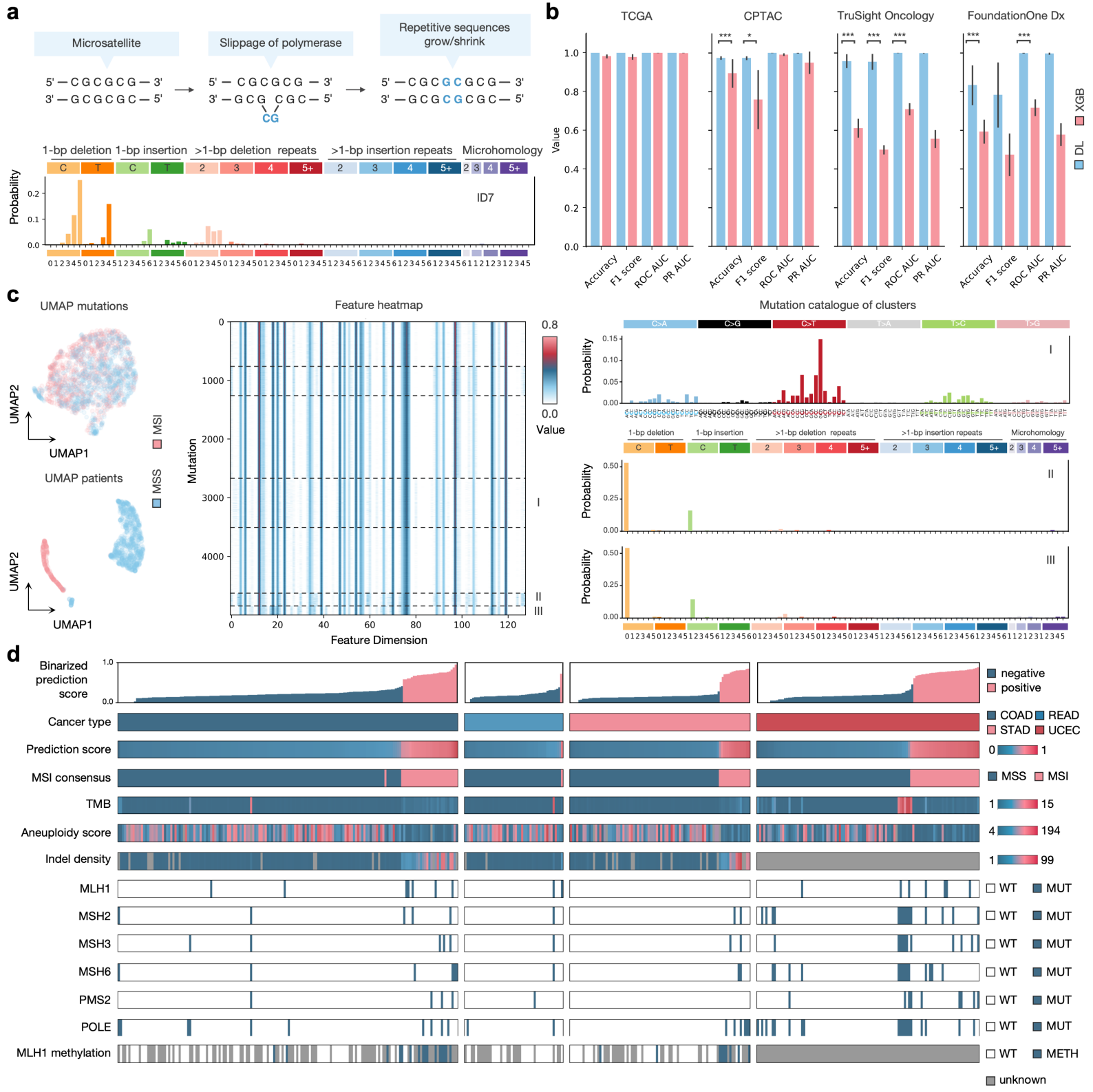
a
Microsatellite
Slippage of polymerase
Repetitive sequences grow/shrink
1-bp deletion
1-bp insertion
>1-bp deletion repeats
>1-bp insertion repeats
Microhomology
Probability
ID7
b
TCGA
CPTAC
TruSight Oncology
FoundationOne Dx
Value
Accuracy
F1 score
ROC AUC
PR AUC
XGB
DL
c
UMAP mutations
UMAP patients
UMAP1
UMAP2
MSI
MSS
Feature heatmap
Mutation
Feature Dimension
Value
Mutation catalogue of clusters
C>A
C>G
C>T
T>A
T>C
T>G
Probability
d
Binarized prediction score
negative
positive
Cancer type
COAD
READ
STAD
UCEC
Prediction score
MSI consensus
MSS
MSI
TMB
Aneuploidy score
Indel density
MLH1
MSH2
MSH3
MSH6
PMS2
POLE
MLH1 methylation
WT
MUT
METH
unknown

Figure 3

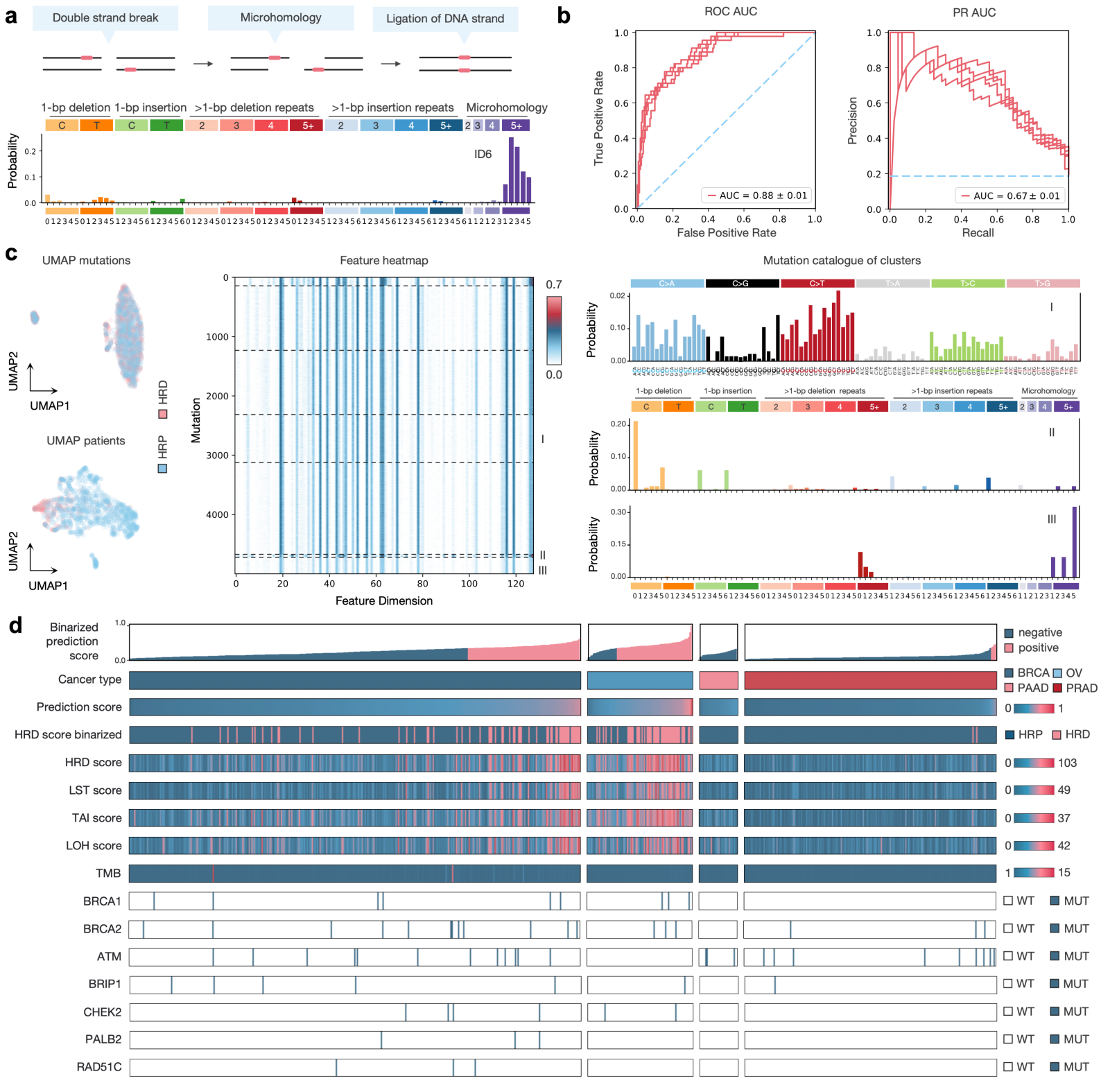
a
Double strand break
Microhomology
Ligation of DNA strand
1-bp deletion
1-bp insertion
>1-bp deletion repeats
>1-bp insertion repeats
Microhomology
Probability
ID6
b
ROC AUC
True Positive Rate
False Positive Rate
AUC = 0.88 ± 0.01
PR AUC
Precision
Recall
AUC = 0.67 ± 0.01
c
UMAP mutations
UMAP2
UMAP1
HRD
HRP
UMAP patients
Feature heatmap
Mutation
Feature Dimension
Mutation catalogue of clusters
C>A
C>G
C>T
T>A
T>C
T>G
Probability
I
II
III
d
Binarized prediction score
negative
positive
Cancer type
BRCA
OV
PAAD
PRAD
Prediction score
HRD score binarized
HRP
HRD
HRD score
LST score
TAI score
LOH score
TMB
BRCA1
BRCA2
ATM
BRIP1
CHEK2
PALB2
RAD51C
WT
MUT

Figure Captions

Fig. 1 | Overview of the study setup. a Datasets (TCGA and CPTAC) used in this study with cancer types and patient counts stratified by biomarkers. Cancer types from TCGA are accentuated in blue, red indicates CPTAC. **b** Preprocessing of .maf files and reference genome to mutation information. **c** Mutation counts by cohort and cancer type used in this study. The black line indicates the mean mutation count. **d** Encoder part of the attMIL model. Mutations from (c) are separated into four blocks: upstream, reference, alteration and downstream element. Each block contains 20 nucleotides or gaps. Reverse strand is modeled in the same manner by the reverse complement of the mutation. Both strand encodings are passed through a 2D-convolution and a dense layer producing a mutation feature vector. **e** Mutation vectors are gathered at the patient level and are aggregated by the attMIL mechanism to a patient feature vector. Patient features are used for the classification task. (COAD - colon adenocarcinoma, READ - rectal adenocarcinoma, STAD - stomach adenocarcinoma, UCEC - uterine corpus endometrial carcinoma, OV - Ovarian serous cystadenocarcinoma, BRCA - Breast invasive carcinoma, PRAD - Prostate adenocarcinoma, PAAD - Pancreatic adenocarcinoma, MLP - multilayer perceptron)

Fig. 2 | DL predicts MSI with high accuracy in panel sequencing. a Mutation mechanism and COSMIC Indel signature of MSI. **b** Bar chart of performance metrics to compare attMIL and XGB models in the internal test dataset of TCGA, the external full CPTAC dataset and CPTAC data filtered by targeted sequencing panels of FoundationOne and Trusight Oncology. Significance is indicated by * (p-value < 0.05), ** (p-value < 0.005) and *** (p-value < 0.0005). **c** Explainability of the attMIL model. UMAP of mutation and patient level features. Heatmap of 5,000 random mutation features clustered by k-means (k=7) and ranked by mean attention score. Indel/SBS mutation catalogues of three mutation clusters. **d** Patient features sorted by increasing prediction score of the attMIL model and separated by cancer type. (COAD - colon adenocarcinoma, READ - rectal adenocarcinoma, STAD - stomach adenocarcinoma, UCEC

- uterine corpus endometrial carcinoma, MSI - Microsatellite instability, MSS - microsatellite stability, WT - wildtype, MUT - mutated, METH - methylated)

Fig. 3 | DL predicts HRD. a Mutation mechanism of MMEJ and COSMIC Indel signature of HRD. **b** ROC AUC and PR AUC plots for attMIL HRD predictions. **c** Explainability of the attMIL model. UMAP of mutation and patient level features. Heatmap of 5,000 random mutation features clustered by k-means (k=7) and ranked by mean attention score. Indel/SBS mutation catalogues of three mutation clusters. **d** Patient features sorted by increasing prediction score of the attMIL model and separated by cancer type. (OV - Ovarian serous cystadenocarcinoma, BRCA - Breast invasive carcinoma, PRAD - Prostate adenocarcinoma, PAAD - Pancreatic adenocarcinoma, HRD - Homologous Recombination Deficiency, HRP - Homologous Recombination Proficiency, WT - wildtype, MUT - mutated)

Supplementary Material

Supplementary Figures

Supplementary Fig. 1 | MSI data preparation. a Bar chart of site-specific data split for MSI. Dark blue indicates training data, light blue validation data and red the test set. The test set stays the same over all folds. **b** Doughnut plot of the data splits regarding class distribution of MSI vs MSS and number of patients. **c** Bar chart of number of patients per cancer type.

Supplementary Fig. 2 | HRD data preparation. a Bar chart of site-specific data split for HRD. Dark blue indicates training data, light blue validation data and red the test set. The test set stays the same over all folds. **b** Doughnut plot of the data splits regarding class distribution of HRD vs HRP and number of patients. **c** Bar chart of number of patients per cancer type**.**

Supplementary Fig. 3 | MSI features. a COSMIC SBS mutation signatures 6 and 15 associated with MSI. **b** SBS mutation catalogues of mutation feature clusters not displayed in Fig. 2c. Only mutation catalogues of the majority mutation class are displayed.

Supplementary Fig. 4 | XGB features used in MSI predictions. 20 most important features of XGB to classify MSI. Features are sorted by contribution to prediction (high to low SHAP values). SBS mutations are displayed as, for example, a C>T mutation with an upstream G and a downstream C (GC>TC). Indel mutations can be read as first the indel pattern length and then the number of repetitions. For example, an insertion pattern of at least five nucleotides without a repetition (5InsRep0) or a mononucleotide deletion of a C with the length one (1DelC1).

Supplementary Fig. 5 | HRD features. a COSMIC SBS mutation signature three and indel signature eight associated with HRD and NHEJ. **b** SBS and indel mutation catalogues of mutation feature clusters not displayed in Fig. 3c. Only mutation catalogues of the majority mutation class are displayed.

Supplementary Fig. 6 | XGB features used in HRD predictions. a ROC AUC and PR AUC of XGB predicting HRD in a five-fold cross validation. **b** 20 most important features of XGB to classify HRD. Features are sorted by contribution to prediction (high to low SHAP values). Microhomology mutations can be read as first the deletion length and then the length of the microhomology. For example, a deletion of at least five nucleotides containing a microhomology of two nucleotides to at least one of the flanking sites (5+DelMH2).

Supplementary Tables

Supplementary Table. 1 | MSI patient counts. MSI and MSS cases in the TCGA data split by cancer type and in total

	COAD	READ	STAD	UCEC	total
MSI cases	59	3	82	136	280
MSS cases	279	107	350	314	1050

Supplementary Table. 2 | HRD patient counts. HRD and HRP cases in the TCGA data split by cancer type and in total

	BRCA	OV	PAAD	PRAD	total
HRD cases	149	74	6	79	308
HRP cases	818	108	133	383	1442

Supplementary Table. 3 | MSI performance metrics. Performance values on the full CPTAC external test dataset of attMIL and XGB predicting MSI. The p-values for corresponding model comparison are stated below the performances. (accuracy - acc, sens - sensitivity, spec - specificity)

	acc	F1 score	ROC AUC	PR AUC	sens	spec	TP	FP	TN	FN
attMIL	0.98± 0.01	0.97± 0.01	1.00± 0.00	1.00± 0.00	0.95± 0.02	1.00± 0.00	22.80 ± 0.45	0.00± 0.00	82.00 ± 0.00	1.20± 0.45
XGB	0.89± 0.07	0.76± 0.15	0.99± 0.01	0.95± 0.06	0.99± 0.02	0.80± 0.16	22.80 ±0.45	16.40 ± 12.93	64.60 ± 12.93	0.20± 0.45
p-value	5.11×10^{-18}	0.04	0.32	-	0.04	0.047	1.00	0.06	0.06	0.06

Supplementary Table. 4 | MSI performance metrics on TruSight Oncology 500 CPTAC variants. Performance values on the TruSight Oncology 500 filtered CPTAC external test

dataset of attMIL and XGB predicting MSI. The p-values for corresponding model comparison are stated below the performances. (accuracy - acc, sens - sensitivity, spec - specificity)

	acc	**F1 score**	**ROC AUC**	**PR AUC**	**sens**	**spec**	**TP**	**FP**	**TN**	**FN**
attMIL	0.96± 0.04	0.96± 0.04	1.00± 0.00	1.00± 0.00	0.92± 0.07	1.00± 0.00	22.00 ± 1.73	0.00± 0.00	82.00 ± 0.00	2.00± 1.73
XGB	0.61± 0.05	0.50± 0.02	0.71± 0.03	0.56± 0.05	0.68± 0.19	0.54± 0.28	15.60 ±4.28	24.60 ± 15.26	29.40 ± 15.26	7.40± 4.28
p-value	1.67×10^{-38}	7.56×10^{-5}	4.66×10^{-16}	-	0.04	0.02	0.06	0.06	0.06	0.59

Supplementary Table. 5 | MSI performance metrics on FoundationOne Dx CPTAC variants. Performance values on the FoundationOne Dx filtered CPTAC external test dataset of attMIL and XGB predicting MSI. The p-values for corresponding model comparison are stated below the performances. (accuracy - acc, sens - sensitivity, spec - specificity)

	acc	**F1 score**	**ROC AUC**	**PR AUC**	**sens**	**spec**	**TP**	**FP**	**TN**	**FN**
attMIL	0.83± 0.10	0.78± 0.17	1.00± 0.00	1.00± 0.00	0.67± 0.20	1.00± 0.00	16.00 ± 4.95	0.00± 0.00	82.00 ± 0.00	8.00± 4.95
XGB	0.59± 0.06	0.47± 0.11	0.72± 0.04	0.58± 0.06	0.63± 0.26	0.55± 0.28	14.60 ±6.02	22.40 ± 14.21	27.60 ± 14.21	8.40± 6.02
p-value	7.94×10^{-21}	0.06	2.88×10^{-17}	-	0.85	0.02	0.6	0.06	0.06	0.06

Supplementary Table. 6 | HRD performance metrics. Performance values on the TCGA test dataset of attMIL and XGB predicting HRD. The p-values for corresponding model comparison are stated below the performances. (accuracy - acc, sens - sensitivity, spec - specificity)

	acc	**F1 score**	**ROC AUC**	**PR AUC**	**sens**	**spec**	**TP**	**FP**	**TN**	**FN**

attMIL	0.80± 0.01	0.65± 0.02	0.88± 0.01	0.67± 0.01	0.74± 0.06	0.86± 0.05	33.40 ± 2.88	25.00 ± 9.38	159.0 ± 9.38	11.60 ± 2.88
XGB	0.75± 0.02	0.66± 0.03	0.86± 0.01	0.75± 0.02	0.52± 0.03	0.98± 0.00	25.40 ± 1.52	2.80± 0.84	183.2 0± 0.84	23.60 ± 1.52
p-value	1.07×10^{-10}	0.51	0.72	-	0.003	0.049	0.06	0.06	0.06	0.06

Supplementary Methods

Feature Catalogues for ML model

To generate the SBS mutation catalogues, patient mutations were grouped by 96 SBS features, encompassing six mutation classes: C>A, C>G, C>T, T>A, T>C, T>G (including the reverse complement). Each class was further stratified by the adjacent 3' and 5' nucleotides. Since there are four possible options for the 3' and 5' positions, this results in 6 * 4 * 4 = 96 features.

For indel feature generation, mutations were grouped according to insertion or deletion length, repetition pattern, and a possible homology to flanking sites. For mononucleotide deletions of C or T (including complementary base), two classes were created, along with two classes for corresponding mononucleotide insertions. Additionally, four classes were defined based on repetition patterns for deletions and insertions. Each of these twelve initial subclasses was then subdivided into six subclasses based on the number of repeats deleted or inserted, resulting in 72 subclasses. Finally, four microhomology classes were assigned, describing mutations in which either the 3' or 5' flanking site has partial homology to the deletion sequence. These classes were further subdivided based on homology length, adding 11 additional classes to the 72 remaining classes, resulting in a total of 83 indel classes.

Multiple Instance Learning

Multiple Instance Learning (MIL) is a deep learning framework designed to handle scenarios where labels are associated with sets of instances, referred to as "bags," rather than individual instances. In MIL, only a subset of instances within a bag contribute to the overall class prediction, while others may be irrelevant or unrelated. This makes MIL particularly suitable for contexts where identifying which instances are informative is inherently challenging or unknown [48]

Here, we applied MIL to model patient-level predictions, where each patient is represented as a "bag" containing multiple somatic mutations ("instances") [35,36]. Not all mutations contribute to the phenotype of interest (e.g., MSI or HRD), as some mutations may not be associated with the underlying mutational mechanisms driving these biomarkers. MIL allows the model to autonomously learn which mutations are most relevant by assigning attention weights to individual instances, rather than relying on predefined rules or human-curated knowledge.

Additional Files

Additional file 1 | MSI status. Excel file in which the patient identifiers of TCGA and CPTAC patients with corresponding tissue type and MSI status is stored.

Additional file 2 | HRD status. Excel file in which the patient identifiers of TCGA patients with corresponding information is stored. Information includes: tissue type, scar HRD scores with subscores (LST, LOH, TAI) and HRD status with cutoffs at 42 and tissue specific cutoffs.

Additional file 3 | Tissue source sites. Excel file in which the tissue source site codes within the TCGA patient identifiers and their mapping to the actual source site is stored.