# MimicNeoAI: An integrated pipeline for identifying microbial mimicry antigens and tumor neoantigen partners

Tao Chen[1,2†], Wei Wang[1,3†], Xiao Zuo[1,4†], Yuxin Zhang[1,5†], Zhilei Li[1], Yin He[1], Yanfei Zhou[1,4], Fang Ye[1,4], Bin Zhang[1,4], Qionghui Jiang[1,3], Huimin Liu[1,3], Lu Zhang [5*‡], Jinman Fang[1,4*‡], Yuanwei Zhang[1,4*‡]

1. Hefei Cancer Hospital of CAS; Anhui Province Key Laboratory of Medical Physics and Technology; Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences (CAS), Hefei, China
2. Institutes of Physical Science and Information Technology, Anhui University, Hefei, China.
3. Anhui Medical University, Hefei, China.
4. Science Island Branch of Graduate School, University of Science and Technology of China, Hefei, China.
5. Department of Computer Science，HongKong Baptist University，Kowloon Tong，Hongkong,China

*Corresponding author. Email: ericluzhang@hkbu.edu.hk (ZL);

docfjm@mail.ustc.edu.cn (FJM); zyuanwei@cmpt.ac.cn (ZYW)

## Abstract

Recent studies underscore the significant role of the tumor microbiome, defined as the microbial community within the tumor microenvironment, in influencing cancer development, progression, and treatment response. Notably, epitopes derived from these tumor-resident microbial species can mimic tumor neoantigens through cross-reactive mechanisms, a phenomenon termed molecular mimicry. Unlike neoantigens, which are highly patient-specific, microbial antigens exhibit lower personalization, as

shared microbial species are frequently enriched across various patients and cancer types. This conservation enables the identification of recurrent microbial antigen mimics that are common to multiple cancers, thereby providing potential targets for broad-spectrum immunotherapies. Despite the growing recognition of microbial antigens in cancer immunity, the field currently lacks dedicated computational tools for systematically characterizing tumor-associated microbial antigens and their potential mimicry candidates. To address this gap, we present MimicNeoAI, a computational pipeline that integrates sequencing data to concurrently identify microbial antigens and tumor neoantigens. Additionally, it predicts the immunogenic potential of peptide-Human Leukocyte Antigen (HLA) complexes using a Bidirectional Long Short-Term Memory (BiLSTM) network model, which is trained on immunogenic epitope datasets from both hosts and microbes. Furthermore, this pipeline assesses microbial mimicry antigens and tumor neoantigen pairs through sequence similarity, allowing for the identification of candidates that may elicit cross-reactive immune responses and enhance immunotherapy. In cross-validation, MimicNeoAI achieved an average AUC of 0.90, successfully predicting 91% of microbial antigen-positive data. When applied to colorectal cancer (CRC), it identified five high-confidence microbial mimicry antigens associated with increased T cell clonal expansion, as evidenced by single-cell T Cell Receptor (TCR) sequencing, highlighting their biological relevance and potential as targets for immunotherapy.

## Introduction

Recent advances in immunotherapy have highlighted the essential role of tumor-specific antigens, particularly neoantigens derived from somatic mutations, in eliciting targeted immune responses[1-3]. However, the highly personalized nature of neoantigens restricts their broad applicability, as they vary significantly between patients, even within the same cancer type[4, 5]. In contrast, microbial antigens exhibit consistent enrichment patterns across patients, suggesting the potential for the development of broadly effective treatments[6, 7]. Emerging evidence indicates that tumor-resident

microbes can produce antigens that mimic neoantigens through a mechanism known as molecular mimicry, which may lead to the identification of a novel class of shared immunotherapy targets[8].

Various approaches have been explored in the field of antigen prediction[5, 9]. The complexity of neoantigen prediction primarily arises from its patient-specific nature and intricate biological conditions involved. Key challenges include the identification of somatic mutations, the prediction of peptide-HLA binding affinity, and assessment of immunogenicity[10]. In response, numerous researchers have developed algorithms to predict peptide-HLA binding affinities, facilitating antigen identification. Some methods employ machine learning and deep learning models to successfully identify tumor neoantigens, highlighting immunogenic peptides capable of activating immune response[11-14]. These advances have enhanced our understanding of patient-specific immune mechanisms and have supported personalized immunotherapy strategies to some extent. Additionally, researchers have created neoantigen identification and prioritization pipelines based on mutation analysis, enabling step-by-step processing of DNA-seq and RNA-seq data[15-17]. While these pipelines have gained popularity among non-computational researchers for their convenience, they often suffer from limitations such as incomplete antigen coverage. Current tools primarily focus on tumor neoantigens and lack an integrated approach for predicting microbial antigens, thus failing to systematically identify them. Moreover, there is a lack of comprehensive, user-friendly predictive tools in the emerging field of microbial antigen prediction that can seamlessly process sequencing data to generate relevant antigen results.

While algorithms such as NNalign[18], NetMHCpan[14], MHCflurry[19] and MHCnuggets[20] have made notable progress in predicting peptide-HLA binding affinities, they do not adequately capture the complex nature of immunogenicity. Not all peptide-HLA complexes with strong binding affinity are recognized by the immune system, underscoring the need for enhanced predictive models[21, 22]. Many existing tools are predominantly trained on human host-derived peptides and often overlook critical

physicochemical properties, such as hydrophobicity and polarity, which may influence immune recognition[6, 23]. This highlights the necessity for more comprehensive approaches that integrate both binding affinity and intrinsic peptide features to improve immunogenicity prediction.

Additionally, there is an urgent need to develop predictive models specifically for microbial antigens to address the diverse immune responses they elicit. Some studies have explored the potential of molecular mimicry between microbial and tumor antigens[24, 25]. These efforts have illuminated the possible cross-reactive immune responses elicited by such mimicry, offering a novel perspective for broad-spectrum immunotherapy target discovery. However, unaddressed mimicry mechanisms still exist. Currently, no existing toolset can analyze sequencing data to predict both tumor neoantigens and microbial antigens while calculating the sequence-level similarity of peptides to identify pairs of microbial mimicry antigens and tumor neoantigens.

To bridge these gaps, we developed MimicNeoAI, an integrated pipeline that unifies the identification of microbial antigens, tumor neoantigens, and microbial mimicry candidates from Whole-Transcriptome Sequencing (WTS) and Whole-Exome Sequencing (WES) data. Our framework represents the first comprehensive toolset designed to systematically identify microbial antigens while also predicting tumor neoantigens. To further enhance our predictions, we assess the immunogenic potential of peptide-HLA complexes, evaluating not only their binding affinities but also their ability to be recognized by T cells and activate anti-tumor immunity. To achieve this, we have developed an immunogenicity prediction model based on a BiLSTM[26] network, trained on datasets of host-derived immunogenic epitopes from Immune Epitope Database (IEDB[21]) and microbial immunogenic epitopes from MicroEpitope[27] database. Additionally, we evaluate microbial mimicry antigens based on sequence-level similarity to further extend our analysis. Identifying these antigens is crucial because they can elicit cross-reactive immune responses, potentially improving immunotherapy by targeting epitopes shared between tumors and microbial pathogens. Ultimately,

MimicNeoAI provides a comprehensive and integrated pipeline for advancing cancer immunotherapy, offering an all-in-one predictive tool for identifying broad-spectrum tumor immunotherapy targets.

## Results

### Overview of the MimicNeoAI Workflow

The MimicNeoAI framework is designed to facilitate the detection and prioritization of microbial and neoantigen candidates, comprising three primary components: microbial antigen candidate detection, neoantigen candidate detection, and the prediction of peptide-HLA complexes immunogenicity (Fig. 1). The microbial antigen detection sub-pipeline initiates with the removal of host-derived sequences, followed by the quantification of microbial taxa, ultimately leading to the identification of microbial peptides. In parallel, the neoantigen detection sub-pipeline begins with the identification of somatic variants, progresses through mutation annotation, and concludes with the generation of tumor-specific peptides. These foundational workflows set the stage for HLA-restricted epitope prediction by integrating assessments of binding affinity and immunogenicity. Following these steps, a unified post-analysis filtering phase is employed to systematically prioritize of epitopes, thus ensuring a comprehensive approach to antigen profiling.

### Performance Evaluation of the Immunogenicity prediction model

MimicNeoAI predicts the immunogenicity of peptide-HLA complexes by evaluating the probability that epitopes with high HLA binding affinity will be recognized by T cell receptors. To validate the performance of MimicNeoAI, we conducted five-fold cross-validation on both the host-derived and microbial dataset. The receiver operating characteristic (ROC) curve showed an area under the curve (AUC) value of at least 0.90 in each fold, while the precision-recall (PR) curve also demonstrated an average precision (AP) value of at least 0.90 across all folds (Fig. 2a and 2b). These results indicate that MimicNeoAI effectively learn features from the training data and

generalizes well to unseen test sets. These robust predictive performances were further corroborated by the parallel trajectories of training and test accuracy curves across all five validation folds (Fig. S1a-S1e), with the minimal metric divergence providing strong evidence against overfitting.

On the microbial antigen validation dataset, MimicNeoAI successfully predicted 91% of positive and 83% of negative instances (Fig. 2c). The performance metrics of accuracy, precision, recall, and F1 score were recorded as 0.87, 0.84, 0.91, and 0.88, respectively (Fig. 2d). An ablation study indicated that a model trained solely on the host-derived dataset performed less effectively on the microbial antigen validation dataset compared to those trained on the microbial dataset or the combined host and microbial dataset. Notably, the model trained exclusively on the microbial dataset outperformed the one trained on the combined dataset, suggesting inherent differences in amino acid composition or structure features between microbial and host immunogenic epitopes (Fig. S1f and S1g).

A comparative analysis with other immunogenicity prediction algorithms, including IEDB[21], DeepNeo[28] and DeepImmuno[29], demonstrated that MimicNeoAI, when trained on the Microbial Dataset, achieved superior AUC and AP values (Fig. 2g and 2h). Furthermore, while distinguishing among HLA types, MimicNeoAI exhibited a prediction accuracy of 100% for HLA-A 01:01, HLA-A 02:01 and HLA-B 54:01, selected due to their representation in the intersection of HLA types supported by the various algorithms on the microbial antigen validation dataset (Fig. 2i).

**Validation of MimicNeoAI performance in CRC**

To assess the clinical applicability of MimicNeoAI, we analyzed a CRC case involving a single tumor sample and a matched adjacent normal sample, both subjected to WTS and WES sequencing. For microbial antigen detection, MimicNeoAI processed 68 million reads from the tumor WTS data, identifying 5,782 high quality non-host reads. After aligning these with microbial reference genomes, 4,500 microbial reads remained.

Subsequent in-silico translation and protein database screening yielded 7,593 peptides with confirmed protein evidence, leading to the identification of 383 epitopes through binding affinity and immunogenicity predictions (Fig. 3a). The identical analysis process was applied to the normal sample's WTS data.

For host-derived antigen detection, MimicNeoAI analyzed 93 million reads from the tumor WES data, revealing 439,739 initial mutation sites. After filtering out those shared with the normal sample, 274,383 tumor-specific mutations remained. Following further analysis, including HLA binding affinity and immunogenicity predictions, 90,971 peptides and 848 epitopes were identified (Fig. 3b). To facilitate unified analysis, we standardized the lengths of the peptides: HLA class I peptides were set to a length of 9, while HLA class II peptides were set to a length of 15.

Despite a notable overlap in host-derived epitopes between the tumor and normal tissue, no overlap was observed in microbial epitopes when analyzing the tumor sample in relation to the normal sample. This finding indicates distinct tissue-specific distributions of microbial epitopes and underscores the selective enrichment of microbial epitopes in the tumor microenvironment (Fig. 3c). Quantitative analysis revealed significant differences in microbial composition between the tumor and normal samples, with high-abundance microbes in the tumor being rarely found in the normal sample or present in considerably lower abundance (Fig. 3h and 3i).

To better assess the immunogenicity prediction results of identified epitopes, MimicNeoAI classified host-derived epitopes into 734 weakly immunogenic, 112 moderately immunogenic and 2 highly immunogenic categories (Fig. 3d). Similarly, microbial epitopes were classified as 260 weakly immunogenic, 108 moderately immunogenic, and 15 highly immunogenic (Fig. 3e). The microbial mimicry analysis demonstrated a higher similarity between HLA-I microbial and host-derived epitopes compared to HLA-II counterparts (Fig. 3f and 3g). We focused on HLA-I pairs with similarity scores exceeding 0.4. For scores ≥0.5, all immunogenic ranks (weakly,

moderately, and highly) were retained; for scores between 0.4 and 0.5, only moderately and highly immunogenic ranks were considered. This process ultimately identified 15 microbial mimicry epitopes alongside their corresponding host-derived epitopes (Fig. 3j and Table 6).

**Validation of Immunogenicity via TCR Specificity**

To further validate the recognition of microbial epitopes, host-derived epitopes and microbial mimicry epitopes by TCRs, we conducted single-cell TCR sequencing on peripheral blood mononuclear cells (PBMCs) from the same patient. The pMTnet[30] algorithm was employed to predict TCR-epitope-HLA binding specificity (Table 7). Epitopes with a binding probability exceeding 0.8 for over 20% of complementarity-determining region 3 (CDR3) sequences were selected as candidate antigens. As the algorithm only predicts HLA-I epitopes, we screened 64 microbial antigens from 202 microbial HLA-I epitopes and 93 host antigens from 388 host-derived epitopes (Fig. 4a and 4b). Additionally, we identified 5 microbial mimicry antigens along with their matched host-derived epitopes.

Both immunogenicity prediction algorithms and TCR binding prediction algorithms extract features from epitope amino acid sequences. Consequently, epitopes and antigens derived from microbes and hosts that successfully pass through these two screenings are likely to exhibit shared similarities in amino acid composition. Sequence logo plots of host-derived epitopes and antigens (Fig. 4c) and microbial epitopes and antigens (Fig. 4d) reveal that high-frequency sites predominantly originate from position 2 and 9 of 9-mer epitopes.

For the identified antigens, we evaluated the correlation between their binding probability to TCRs and the clonal expansion ratio of T cells. Results demonstrated a positive correlation for both host-derived and microbial antigens (Fig. 4e and 4f), suggesting that T cells exhibiting higher clonal expansion possess greater antigen recognition capabilities, thereby validating the reliability of our identified antigens.

Notably, microbial mimicry antigens showed a stronger correlation with T cell clonal expansion than their matched host-derived epitopes (Fig. 4g and 4h), indicating that microbial mimicry antigens may be more effective in activating T cell responses.

Further analysis revealed a positive correlation between the immunogenicity prediction probability of microbial mimicry antigen-HLA complexes and the predicted probability of TCR recognition for these complexes (Fig. 4j). This reinforces the accuracy of our immunogenicity prediction model. In contrast, the host-derived epitopes paired with microbial mimicry antigens did not exhibit such a correlation, further supporting the notion that microbial mimicry antigens are more likely to elicit T cell responses (Fig. 4i). Finally, single-cell TCR sequencing data analysis results illustrated a significant proportion of T cell clones and the distribution of T cells based on their binding specificity to microbial mimicry antigen-HLAs, indicating a high likelihood of T cells activation by these complexes (Fig. 4k and 4l).

## Discussion

MimicNeoAI presents a comprehensive pipeline that integrates the identification of microbial antigens, tumor neoantigens and microbial mimicry candidates from WTS and WES data. A significant advantage of our approach is its ability to overcome the limitations of existing tools, which often overlook the complete antigen landscape. By systematically detecting not only tumor neoantigens but also microbial antigens and their mimicry candidates, MimicNeoAI enables a thorough profiling of potential immune targets. This dual focus is critical, as microbial epitopes can provoke cross-reactive responses, an aspect frequently neglected by conventional neoantigen-centric methodologies.

Our model significantly enhances immunogenicity prediction by integrating BiLSTM networks to capture long-range amino acid interactions within peptide-HLA complexes, complemented by FCNN networks that assess physicochemical properties. By

incorporating both host-derived and microbial epitope training datasets, MimicNeoAI achieves stable and accurate training outcomes, outperforming other algorithms on an independent validation set of microbial antigens. This performance underscores the effectiveness of our approach in addressing the unique immunogenic characteristics of microbial epitopes.

Moreover, MimicNeoAI is pioneering in its systematic evaluation of microbial mimicry antigens through sequence-level similarity analysis, filling the needed of mechanistic research in understanding cross-reactive immunity. Our validation of these microbial mimicry antigens through single-cell TCR sequencing in CRC samples confirms the biological relevance to our predictions. The observed positive correlation between TCR binding probabilities and clonal expansion ratio for both host-derived and microbial antigens further substantiates the validity of our identified antigens. Notably, microbial mimicry antigens demonstrated a stronger correlation, suggesting their potential for more effective T cell activation.

Despite these promising advancements, several avenues for future improvement and research remain. Firstly, while MimicNeoAI effectively integrates patient-specific HLA typing and physicochemical characteristics of candidate peptides into its prediction model, incorporating additional biological factors, such as the three-dimensional structures of peptides, HLA molecules, and TCR, could further enhance the prediction accuracy[31]. Secondly, the performance of any machine learning model, including MimicNeoAI, is dependent on the quality and quantity of training data. The continuous accumulation of training data across various cancer types and patient populations will be essential for further enhancing the model's capabilities. Lastly, leveraging advancements in deep learning, particularly transformer-based models like ProteinBERT[32] and ESM-2[33], could provide novel methodologies for improving predictive capabilities.

In summary, MimicNeoAI represents a notable advancement in identifying and prioritizing immunogenic candidates from tumor and microbial sources. By integrating microbial mimicry analysis, this model offers valuable methodological tools that facilitate the evaluation of immunogenicity. While it contributes to our understanding of immune responses, it primarily serves as a practical resource for future experimental research in targeted immunotherapies.

## Methods

### In-house Sequencing Data Generation

Tumor tissue, adjacent normal tissue, and PBMCs were obtained from a single CRC patient. Tumor tissue was collected via endoscopic biopsy, with 2–3 pieces excised, each approximately 3 mm in diameter, depending on tumor size. The specimens were rinsed with saline to remove blood and mucus, immediately snap-frozen on dry ice for 2 hours during transportation, and subsequently stored at -80 °C. PBMCs were isolated from the patient's blood using density gradient centrifugation. Blood samples were mixed with an equal volume of phosphate-buffered saline (PBS), centrifuged at 800 × g for 20 minutes, and the PBMC layer was collected. The isolated PBMCs were washed twice with PBS, suspended in freezing medium, and stored at -80 °C.

DNA for whole exome sequencing was extracted using the HiPure Universal DNA Kit (D3018-02, Magen). Libraries were prepared with the Hieff NGS® OnePot Pro DNA Library Prep Kit for Illumina® (12205, Yeasen) and sequenced on the DNBSEQ-T7 platform to generate 150 bp paired-end reads.

For whole transcriptome sequencing, total RNA was extracted using the Hieff NGS® MaxUp Human rRNA Depletion Kit (12257ES24, Yeasen). RNA integrity was assessed using the Agilent 2100 Bioanalyzer. Libraries were constructed with the Hieff NGS® Ultima Dual-mode mRNA Library Prep Kit (12301ES96, Yeasen) and sequenced on the DNBSEQ-T7 platform.

For single-cell TCR sequencing, PBMCs from CRC patient were isolated by density centrifugation using Ficoll-Paque plus (04-03-9391/03, Stemcell). Libraries were prepared using Chromium Single Cell 5' v2 and V(D)J Reagent kits (10x Genomics). The libraries were quantified using a High Sensitivity DNA Chip on the Agilent Bioanalyzer 2100 and the Qubit High Sensitivity DNA Assay (Thermo Fisher Scientific), followed by sequencing on the DNBSEQ-T7 platform with a PE150 read length.

**Host Sequences Removal**

Host genomic sequences were removed using a two-step alignment strategy against the GRCh38[34] and CHM13v2.0[35] reference genomes. Quality-controlled sequencing data (processed with Fastp[36] v0.22.0) were first aligned to GRCh38 using BWA-MEM[37] (v0.7.17). Unmapped reads were extracted with SAMtools[38] (v1.5), converted to FASTQ format, and realigned to CHM13v2.0 with BWA-MEM. Reads unmapped to both genomes were classified as non-host sequences. Alignment statistics (total reads, mapped reads and paired-end reads) were recorded at each step using SAMtools. For paired-end sequencing data, unmapped reads are separated into Rl and R2 FASTQ files per sample; single-end sequencing data were consolidated into one FASTQ file per sample.

**Custom Microbial Reference Database Construction**

***Taxonomic Curation*** Taxonomic data relevant to human tumor was initially curated from the Human Microbiome Project (HMP[39]) , published reference genomes [Thomas et al.][40], and literature (Table 1). We identified 355 bacterial genera and established their taxonomic hierarchy using NCBI taxonomy ID (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz, updated on Nov. 7, 2024.).

***Genome Selection*** For these genera, 8,315 high-quality reference genomes (species or strain level) were retrieved from NCBI RefSeq using assembly summary data (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/assembly_summary_refseq.txt, updated on

Nov. 14, 2024.). These genomes, manually curated by the NCBI, met high-quality standards. Viral reference sequences were sourced from Thomas et al[40]. and NCBI RefSeq.

***Microbial Functional Annotation*** To assess colonization potential, tumorigenic associations, and effects on anti-tumor treatments, we employed a large language model (LLM, ChatGPT 4.0) with the prompt:

"The list below contains microbiome taxa at the species or strain level. Please provide the following information for each and organize it in a table: Brief Description, Potential to Colonize Human Tissues, Association with Tumorigenesis and Progression, and Impact on Anti-Tumor Treatments."

Taxa were queried in randomized batches of 50. LLM outputs underwent manual validation via literature review and expert curation, retaining 3,199 bacterial and 1,262 viral taxa (Table 2). Criteria included:

Colonization Potential: Low (minimal tissue presence), Moderate (tissue-specific, non-dominant), High (prevalent in tissues).

Tumorigenic Association: Established (e.g., Porphyromonas gingivalis), Opportunistic (e.g., Enterobacter asburiae), or Probiotic-linked (limited evidence; e.g., Limosilactobacillus panis).

Treatment Impact: Positive (e.g., Streptomyces spp. as drug sources), neutral, or negative (e.g., opportunistic pathogens complicating therapy).

***Database Construction*** Genome sequences were downloaded from NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/*.fna.gz, ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/viral.1.1.genomic.fna.gz, updated on Sep. 7, 2024.). A reference database was built using BWA-MEM and GATK[41] PathSeqBuildReferenceTaxonomy (v4.6). Protein sequences retrieved from NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/*.fna.gz, updated on Sep. 7, 2024) and formatted with BLAST[42] makeblastdb (v2.14.0).

**Microbial Sequence Alignment**

Host-depleted reads were aligned to the custom microbial database using GATK PathSeqPipelineSpark[43] (BWA-MEM aligner). Parameters included retention of primary/suboptimal alignments and a minimum clipped read length of 50 bp. Taxonomic assignments were validated against a curated hierarchy file. Ambiguous alignments were resolved via: Reads with alignment lengths <95% of the best match length were discarded; Suboptimal alignments with the number of mismatches >1 were excluded.

**Microbial Peptide Identification**

Taxon-filtered reads were queried against the custom protein database using BLASTX (v2.12.0). Parameters: E-value ≤5e-2, top 5 hits per query (-max_target_seqs 5). Only reads with perfect identity to reference proteins (Pident = 100) were retained. Output fields included: qseqid, sseqid, qseq, sseq, stitle, Pident, length, mismatch, gapopen, evalue, and bitscore. Validated peptides were converted to FASTA format, and taxonomic origins were confirmed via NCBI taxonomy IDs.

**HLA Typing**

HLA alleles were called using HLA-HD (v1.7.0). Class I loci (A, B, C, E, F, G, H, J, K, L, V) and Class II loci (DRA, DRB1-9, DQA1, DQB1, DPA1, DPB1, DMA, DMB, DOA, DOB) were assessed. FASTQ files were aligned to HLA references via Bowtie2 (v2.2.8; parameters: -x [index] -p [threads]). Mapped reads were extracted (SAMtools view -h -F 4), converted to FASTQ, and standardized using AWK. Haplotypes were called via hlahd.sh (-m 50, population frequency data -f, HLA_gene references).

**HLA Binding Prediction**

Peptides with 100% identity to microbial proteins were analyzed using pVACtools (v4.2.1)[17, 44, 45] with HLA alleles from HLA-HD. Epitope lengths: 8–10-mers (Class I), 15-mers (Class II). Prediction algorithms: NNalign, NetMHCpan, NetMHCIIpan, NetMHCpanEL, NetMHCIIpanEL, SMM, SMMPMBEC, and SMMalign. Binding

affinity (IC50) thresholds were set at ≤500 nM (IEDB tools; --iedb-install-directory). Alleles without typing data were excluded.

**Immunogenicity Prediction Model Construction**

A Bidirectional LSTM (BiLSTM) ensemble model predicted immunogenicity for peptide-HLA pairs. Input encoding included:

Sequence embedding: Peptide and HLA sequences were tokenized using a 22-token vocabulary comprising a padding token ("<pad>"), the 20 standard amino acid residues, and "-" representing the ambiguous amino acid X. The padding token was mapped to a 12-dimensional zero vector. The 20 standard residues and "-" were mapped to 12-dimensional vectors through a simplified PCA feature space derived from 566 physicochemical properties in the AAindex1 database[29, 46].

Physicochemical features: Eight peptide-specific properties (Composition, polarity, volume, charge, hydrophobicity, Boman index, aliphatic index, and isoelectric point) were computed using the R Peptides package.

Features were processed through a 256-node dense layer. HLA sequences (retrieved from IPD-IMGT/HLA[47] database) were aligned to paratopes. Outputs were calibrated via softmax activation, and the probability value at index 1 was taken to represent the predicted probability of the positive class (immunogenic). Alleles without paratope matches received a neutral score (0.5).

**Model Training and Performance Evaluation**

Model parameters were optimized using the Adam optimizer (learning rate: $5\times10^{-5}$) with a cross-entropy loss. We performed 5-fold cross-validation, where each fold was trained for 200 epochs, requiring approximately 20-30 minutes of computation time per fold on a 48GB NVIDIA L20 GPU. In addition to the cross-validation, we conducted formal model training. For this training, we utilized all the data from the 5-fold cross-validation, including its test sets, but excluding the microbial antigen validation dataset. All implementations leveraged GPU acceleration with jupyterlab-pytorch:2.0.1-py3.9-cuda11.7-ubuntu20.04. for improved efficiency.

## Candidate Microbial Antigens Prioritization

***Taxonomic Validation*** To resolve taxonomic ambiguities, peptide-protein matches were cross-referenced against microbial taxonomic hierarchies and annotated protein databases. Only peptides that exhibited species-level consistency between aligned sequences and validated microbial taxon IDs were retained. This ensured unambiguous linkage of antigen candidates to specific microbial species through both sequence alignment and taxonomic lineage validation.

***Annotation*** Comprehensive annotations were integrated to support antigen evaluation, including functional metrics (Amino acid composition, polarity, etc.), taxonomic classification (Tax ID, scientific name, and lineage at species, genus, and family levels), sequence context (epitope sequence and extended peptide information), immunogenicity prediction, binding validation (MHC allele, median and best IC50 scores, and median and best percentile ranks), abundance metrics (microbial abundance, eRPKM values, and normalized eRPKM values), and alignment quality (E-value, pident, etc.), with full details in Table 3.

***Prioritization*** Weakly immunogenic: Score >0.5, IC50 <500 nM, %rank <2, abundance >0%. Moderately immunogenic: Score >0.6, IC50 <300 nM, %rank <1, abundance >1%. Highly immunogenic: Score >0.8, IC50 <50 nM, %rank <0.5, abundance >5%.

## Identification and Annotation of Host Variants

Raw sequencing data in FASTQ format underwent quality control using Fastp to remove low-quality reads and adapter sequences. Filtered reads were aligned to the human reference genome (GRCh38) using BWA-MEM with multithreading (-t parameter), and the resulting SAM files were converted to sorted BAM format and indexed using SAMtools with parallel sorting (-@ threads). Duplicate reads were identified and marked with GATK Spark MarkDuplicatesSpark, utilizing local multithreading (--spark-master local[threads]).

Base quality score recalibration (BQSR) was performed in two steps: first, recalibration models were generated using BaseRecalibratorSpark based on known variant resources (Mills_and_1000G_gold_standard.indels.hg38.vcf.gz, dbsnp_138.hg38.vcf.gz, and 1000G_phase1.snps.high_confidence.hg38.vcf.gz); second, adjustments were applied with ApplyBQSRSpark. Somatic variants were called with GATK Mutect2 with BQSR-processed BAM files, and candidate mutations were filtered using FilterMutectCalls to improve specificity and reliability. Variants were annotated using VEP[48] (v110) within Singularity[49] containers (v3.9.0-rc.3), leveraging the human reference genome GRCh38 (--assembly GRCh38 --species homo_sapiens) with input/output directories (/data_dir/, /output_dir/), VEP cache (/vep_data/), and reference genome (/fasta_data/hg38.fa) mounted via -B flags. Key parameters included: --offline mode with local cache, --vcf output format, transcript annotations (--symbol, --hgvs, --tsl, --biotype), and plugins (Frameshift, Wildtype) from /vep_data/VEP_plugins-release-110/. Population allele frequencies (--af, --af_1kg) and SIFT predictions (--sift b) were incorporated. Post-annotation filtering removed transcript-reference mismatches using ref-transcript-mismatch-reporter with strict mode (-f hard).

**Transcriptomic Analysis**

STAR[50] (v2.5.2b) aligned reads to GRCh38 (parameters: --outSAMunmapped Within, --quantMode TranscriptomeSAM). RSEM[51] (v1.2.28; rsem-calculate-expression --paired-end --alignments)[51] quantified expression.

**Candidate Neoantigen Prioritization**

***Annotation*** Comprehensive annotations were integrated to support antigen evaluation. These included: Functional Metrics (Amino acid composition, polarity, and related properties), Sequence Context (Mutant or wild-type epitope sequences, extended peptide sequences, mutation positions, amino acids, and wild-type protein information), Immunogenicity Prediction, Binding Validation (MHC allele, median and best mutant/wildtype IC50 scores, median and best mutant/wildtype percentile ranks, and

best IC50 fold change), Tumor-Specific Evidence (gene/transcript expression, tumor DNA depth, tumor DNA variant allele frequency, and tumor RNA depth), Transcript and Gene Annotation (Gene name, Ensemble gene ID, transcript details, biotype, transcript support level, consequence, and HGNC ID), Population and Clinical Relevance (Allele frequency, clinical significance, and somatic mutation status) and known existing variations. For detailed annotations, refer to Table 3.

***Prioritization*** Weakly immunogenic: Score >0.5, IC50 <500 nM, FC >2, %rank <2, TPM >0. Moderately immunogenicity: Score >0.6, IC50 <300 nM, FC >4, %rank <1, TPM >1. Highly immunogenicity: Score >0.8, IC50 <50 nM, FC >10, %rank <0.5, TPM >10.

**Single-cell RNA Sequencing Data Analysis and Cell Identity Determination**

Raw single-cell RNA sequencing (scRNA-seq) data were subjected to quality control using Fastp (v0.23.2) to remove low-quality reads. The cleaned reads were then aligned to the GRCh38 human using STAR (v2.7.10a) with default parameters. UMI count matrices were processed with Seurat[52] (v4.4.0) for quality control, normalization, and feature selection. Low-quality cells were excluded. Batch effects were corrected using Harmony (v1.2.0), and potential doublets were removed using DoubletFinder (v1.10.2). Clustering and dimensionality reduction visualization were performed using Scanpy[53] (v1.10.2). The first 10 principal components (PCs) were used to construct a k-nearest neighbor (KNN) graph, with each cell connected to its 10 nearest neighbors. Clustering was performed using the Leiden algorithm, and dimensionality reduction was achieved via UMAP.

**Single-cell TCR Sequencing Data Analysis and Antigen-peptide Affinity Assessment**

Raw TCR sequencing data were obtained in FASTQ format and processed using Cell Ranger VDJ (v7.2.0) with default parameters to identify TCR α- and β-chain sequences for each individual cell, along with corresponding clonotype, sequence, and abundance information. Cells with clonal isoform information were extracted from the TCR

sequencing data for subsequent analysis and then were analyzed using the same pipeline as described for scRNAseq. TCR-peptide binding affinity was predicted using pMTnet with an HLA-specific prediction model to evaluate the likelihood of TCR-peptide-MHC interactions. The distribution of different TCRs was visualized using UMAP to generate a spatial representation of TCR diversity and clustering patterns.

## References


1. Zhong, L. et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct Target Ther* **6**, 201 (2021).
2. Turajlic, S. et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol* **18**, 1009-1021 (2017).
3. Lang, F., Schrörs, B., Löwer, M., Türeci, Ö. & Sahin, U. Identification of neoantigens for individualized therapeutic cancer vaccines. *Nature Reviews Drug Discovery* **21**, 261-282 (2022).
4. Sawyers, C. Targeted cancer therapy. *Nature* **432**, 294-297 (2004).
5. Xie, N. et al. Neoantigens: promising targets for cancer therapy. *Signal Transduct Target Ther* **8**, 9 (2023).
6. Kalaora, S. et al. Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature* **592**, 138-143 (2021).
7. Naghavian, R. et al. Microbial peptides activate tumour-infiltrating lymphocytes in glioblastoma. *Nature* **617**, 807-817 (2023).
8. Ma, Y.-J. et al. Dissection of intratumor microbiome–host interactions at single-cell level in lung cancer. *hLife* (2024).
9. Fotakis, G., Trajanoski, Z. & Rieder, D. Computational cancer neoantigen prediction: current status and recent advances. *Immunooncol Technol* **12**, 100052 (2021).
10. Cai, Y. et al. Artificial intelligence applied in neoantigen identification facilitates personalized cancer immunotherapy. *Frontiers in oncology* **12**, 1054231 (2022).
11. Bulik-Sullivan, B. et al. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat Biotechnol* (2018).
12. Chen, B. et al. Predicting HLA class II antigen presentation through integrated deep learning. *Nat Biotechnol* **37**, 1332-1343 (2019).
13. Wu, J. et al. DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. *Front Immunol* **10**, 2559 (2019).

14. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* **48**, W449-w454 (2020).
15. Schubert, B., Brachvogel, H.P., Jürges, C. & Kohlbacher, O. EpiToolKit--a web-based workbench for vaccine design. *Bioinformatics* **31**, 2211-2213 (2015).
16. Bjerregaard, A.M., Nielsen, M., Hadrup, S.R., Szallasi, Z. & Eklund, A.C. MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol Immunother* **66**, 1123-1130 (2017).
17. Hundal, J. et al. pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol Res* **8**, 409-420 (2020).
18. Nielsen, M. & Andreatta, M. NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res* **45**, W344-w349 (2017).
19. O'Donnell, T.J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst* **11**, 42-48.e47 (2020).
20. Shao, X.M. et al. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunol Res* **8**, 396-408 (2020).
21. Vita, R. et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* **43**, D405-412 (2015).
22. Kim, S. et al. Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann Oncol* **29**, 1030-1036 (2018).
23. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-864 (1974).
24. Boesch, M. et al. Tumour neoantigen mimicry by microbial species in cancer immunotherapy. *Br J Cancer* **125**, 313-323 (2021).
25. Yao, P. et al. Proteogenomic analysis identifies neoantigens and bacterial peptides as immunotherapy targets in colorectal cancer. *Pharmacol Res* **204**, 107209 (2024).
26. Schuster, M. & Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* **45**, 2673-2681 (1997).
27. Li, D. et al. MicroEpitope: an atlas of immune epitopes derived from cancer microbiomes. *Nucleic Acids Res* **53**, D1435-d1442 (2025).
28. Kim, J.Y., Bang, H., Noh, S.J. & Choi, J.K. DeepNeo: a webserver for predicting immunogenic neoantigens. *Nucleic Acids Res* **51**, W134-w140 (2023).
29. Li, G., Iyer, B., Prasath, V.B.S., Ni, Y. & Salomonis, N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Briefings in bioinformatics* **22** (2021).
30. Lu, T. et al. Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat Mach Intell* **3**, 864-875 (2021).

31. Schmidt, J. et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Reports Medicine* **2**, 100194 (2021).
32. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102-2110 (2022).
33. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
34. Schneider, V.A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research* **27**, 849-864 (2017).
35. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44-53 (2022).
36. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107 (2023).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
38. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
39. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).
40. Battaglia, T.W. et al. A pan-cancer analysis of the microbiome in metastatic cancer. *Cell* **187**, 2324-2335.e2319 (2024).
41. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).
42. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
43. Kostic, A.D. et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**, 393-396 (2011).
44. Hundal, J. et al. Accounting for proximal variants improves neoantigen prediction. *Nature genetics* **51**, 175-179 (2019).
45. Hundal, J. et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome medicine* **8**, 11 (2016).
46. Kawashima, S. et al. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* **36**, D202-205 (2008).
47. Robinson, J. et al. IPD-IMGT/HLA Database. *Nucleic Acids Res* **48**, D948-d955 (2020).
48. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome biology* **17**, 122 (2016).
49. Kurtzer, G.M., Sochat, V. & Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PloS one* **12**, e0177459 (2017).
50. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

51. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
52. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e3529 (2021).
53. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 15 (2018).

## Code availability

MimicNeoAI is packaged, and distributed as an open-source, publicly available repository at https://github.com/BioStaCspublic/MimicNeoAI

## Data availability

The train and test datasets, custom microbial reference database, packed pipeline code, as well as the data and code for visualization, are all available on Zenodo at 10.5281/zenodo.15582924

## Funding

This work is supported by the National Natural Science Foundation for Young Scholars of China (82171601) , National Key R&D Program of China (2021YFC2700103) and Anhui Provincial Key Research and Development Plan (2022e07020054).

## Acknowledgements

We sincerely express our gratitude to the Gastrointestinal oncology Integrated Research Team of Hefei Cancer Hospital of CAS for their assistance in sample extraction, data organization, standardization of clinical cases and pathological prognosis evaluation. We gratefully acknowledge the Hefei Advanced Computing Center for providing computational resources and technical support. We thank the colleagues in our research group for inspiring discussion and their contributions.

**Author contributions**

Y.Z., T.C., J.F., and L.Z conceptualized the study. T.C. was responsible for the model design and pipeline implementation. T.C., Y.Z, W.W and X.Z. performed data analysis and model evaluation. Z.L., Y.H., Y.Z., F.Y. B.Z Q. J and H.L. provided clinical samples. T.C., W.W. and Y.Z. wrote the manuscript. L.Z., J.F., and Y.Z. supervised the study.

## Figures

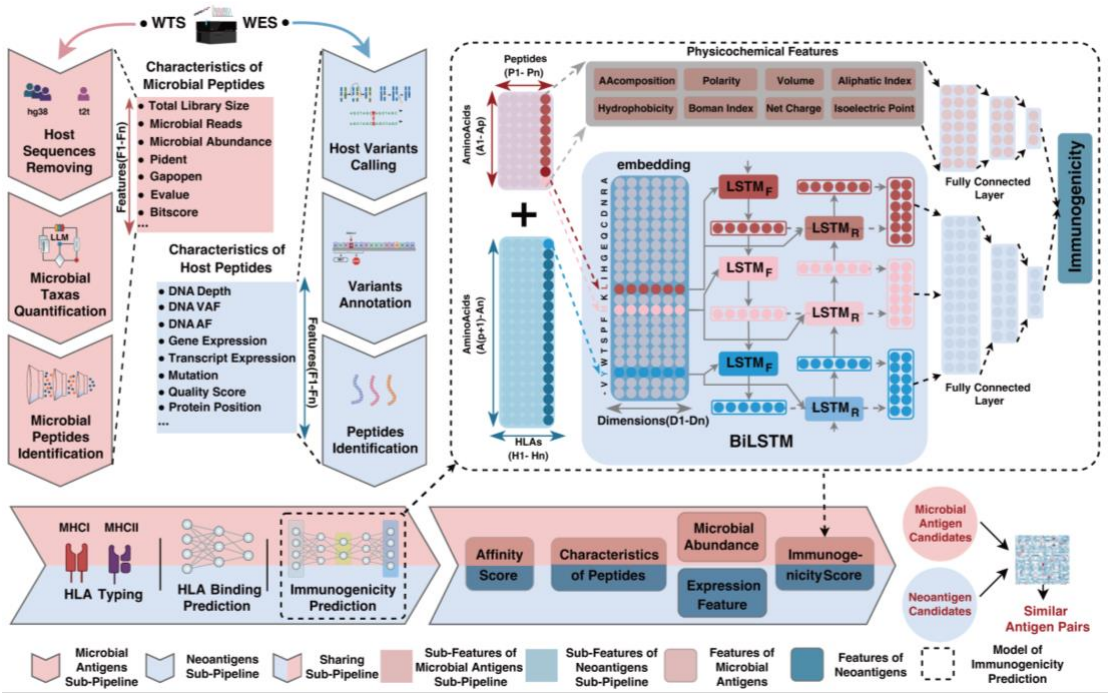



**Fig. 1 | Overview of the MimicNeoAI Framework.** MimicNeoAI comprises three core modules: microbial antigen candidate detection, which involves removing host-derived sequences, quantifying microbial taxa, and identifying microbial peptides; neoantigen candidate detection, which starts with somatic variant detection, proceeds through mutation annotation, and concludes with the generation of tumor-specific peptides; and immunogenicity prediction of peptide-HLA complexes, which integrates binding affinity assessment with immunogenicity evaluation.

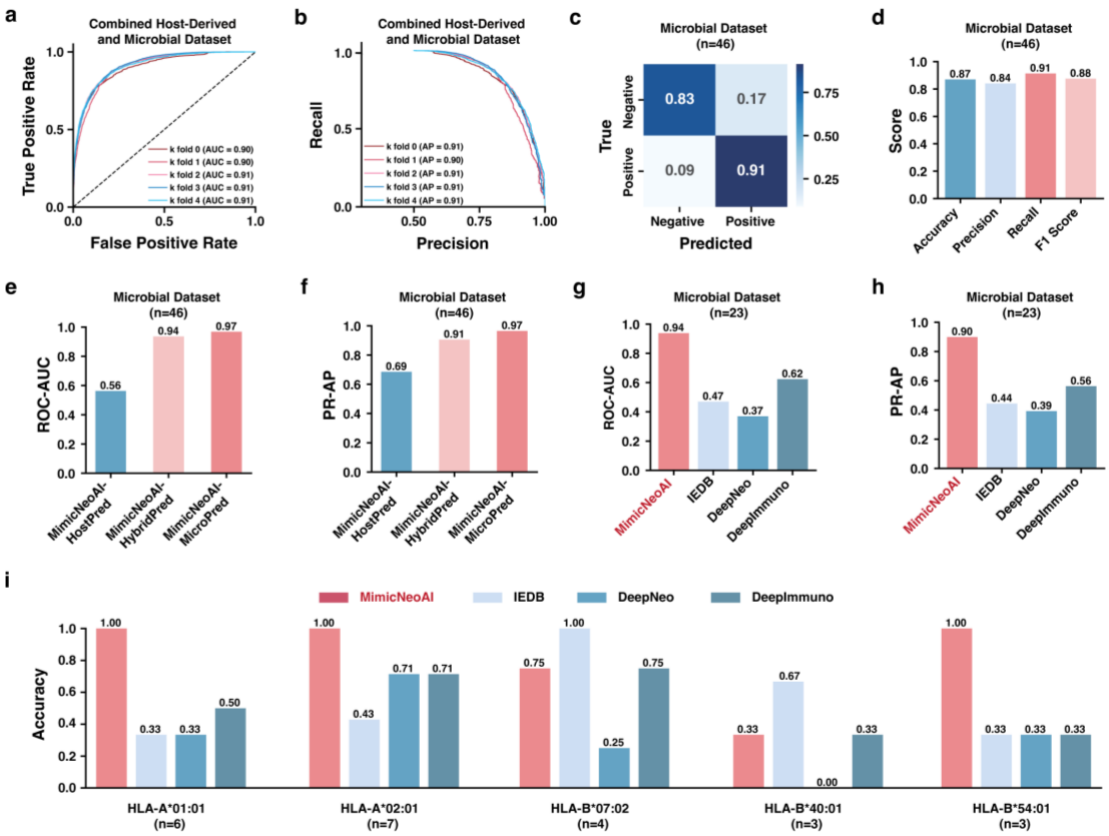
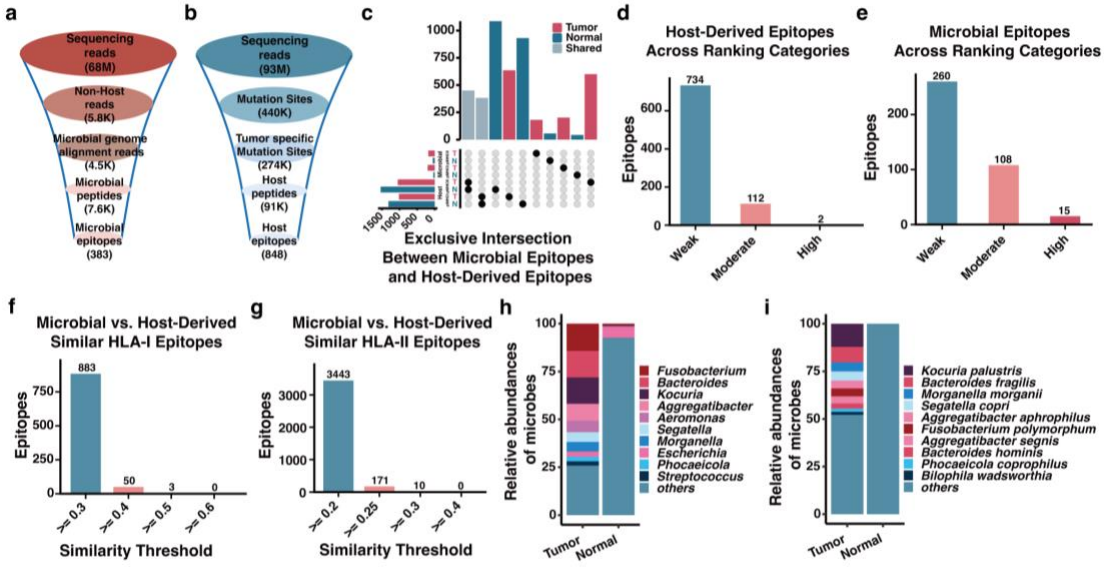



**Fig. 2 | Evaluation of Immunogenicity Prediction Performance.** Performance metrics of the immunogenicity prediction algorithm are illustrated through the Receiver Operating Characteristic (ROC) curve (a) and Precision-Recall (PR) curve (b) using a combined host-derived and microbial test dataset. A confusion matrix (c) and associated performance metrics (d) for immunogenicity predictions on an empirical microbial antigen test dataset. Models of MimicNeoAI trained on different datasets were assessed using the empirical microbial antigen test dataset. AUC values are presented in (e), while PR values are shown in (f). The models evaluated include HostPred (trained exclusively on the host-derived dataset), MicroPred (trained solely on the microbial dataset), and HybridPred (trained on the combined host-derived and microbial dataset). All model-related sections in this study utilized HybridPred. AUC (g) and PR (h) performance metrics for MimicNeoAI are compared against IEDB, DeepNeo, and DeepImmuno using the empirical microbial antigen test dataset, which represents the intersection of predictions supported by these algorithms. (i) Prediction accuracy of MimicNeoAI compared to IEDB, DeepNeo, and DeepImmuno across

specific HLA types in the empirical microbial antigen validation dataset, including HLA-A01:01, HLA-A02:01, HLA-B07:02, HLA-B40:01, and HLA-B*54:01.




j

| Epitope | | Attributes of Microbial Epitope | | | | | | | Attributes of Host-Derived Epitope | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Microbial Epitope | Host Epitope | MHC Allele | Microbe | Rank | Immun-ogenicity Score | Best IC50 Score | Best Percen-tile | Microbe Abundance | MHC Allele | Gene Name | Rank | Immun-ogenicity Score | Best MT IC50 Score | Best IC50 Fold Change | Best MT Percen-tile | Gene Express-ion |
| FLMIGDSAI | SLIPDSAIV | HLA-A*02:01 | Bacteroides fragilis NCTC 9343 | High | 0.90 | 9.54 | 0.08 | 8.18 | HLA-A*02:01 | ANKZF1 | Moderate | 0.95 | 27.44 | 736.33 | 0.10 | 7.99 |
| FLFFAAFAA | LFFAAVVGF | HLA-A*02:01 | Phocaeicola vulgatus ATCC 8482 | Weak | 0.88 | 7.12 | 0.05 | 0.65 | HLA-B*15:01 | ANO5 | Weak | 0.82 | 118.57 | 69.03 | 0.89 | 0.15 |
| MLFLFFAAF | | HLA-B*15:01 | | | 0.74 | 12.50 | 0.06 | | | | | | | | | |
| WMLFLFFAA | | HLA-A*02:01 | | | 0.56 | 16.15 | 0.15 | | | | | | | | | |
| AAREAVRIV | VTAEAVRSF | HLA-C*06:02 | Bacteroides fragilis NCTC 9343 | Moderate | 0.99 | 33.48 | 0.20 | 8.18 | HLA-B*15:01 | ALDH1L1 | Weak | 0.96 | 50.29 | 121.71 | 0.17 | 0.44 |
| RLLKLPAPI | KLPAAIEQL | HLA-A*02:01 | Bacteroides fragilis NCTC 9343 | Moderate | 0.94 | 27.45 | 0.25 | 8.18 | HLA-A*02:01 | DNAH5 | Weak | 0.96 | 116.61 | 13.25 | 0.08 | 0.15 |
| LMIGDSAIL | SLIPDSAIV | HLA-B*15:01 | Bacteroides fragilis NCTC 9343 | Moderate | 0.92 | 113.31 | 0.49 | 8.18 | HLA-A*02:01 | ANKZF1 | Moderate | 0.95 | 27.44 | 736.33 | 0.10 | 7.99 |
| RLASHKLPV | YRLASLYLY | HLA-A*02:01 | Granulicatella adiacens | Moderate | 0.89 | 24.69 | 0.22 | 1.58 | HLA-C*06:02 | IQCA1L | Weak | 0.96 | 48.04 | 6.56 | 0.04 | 0.13 |
| | CSSTLASHF | | | | | | | | HLA-B*15:01 | PPL | | 0.53 | 224.39 | 2.90 | 0.87 | 2.50 |
| TLVVAIETY | FLITLVVVV | HLA-B*15:01 | Granulicatella adiacens | Moderate | 0.86 | 112.91 | 0.07 | 1.58 | HLA-A*02:01 | SMIM29 | Weak | 0.85 | 3.44 | 3.26 | 0.02 | 6.91 |
| NMWIQLVAV | NLDKLIQLV | HLA-A*02:01 | Bacteroides fragilis NCTC 9343 | Moderate | 0.80 | 23.61 | 0.22 | 8.18 | HLA-A*02:01 | MAN2B1 | Weak | 0.91 | 47.88 | 5.12 | 0.05 | 18.69 |
| LLFFIIGSL | LLFFLLFLA | HLA-A*02:01 | Staphylococcus equorum | Moderate | 0.78 | 48.69 | 0.71 | 1.02 | HLA-A*02:01 | CD19 | Weak | 0.70 | 8.25 | 3.59 | 0.16 | 1.40 |
| MVSAFIVLF | SMMVSAGPL | HLA-B*15:01 | Bacteroides fragilis NCTC 9343 | Moderate | 0.74 | 133.25 | 0.56 | 8.18 | HLA-B*15:01 | NBPF11 | Weak | 0.84 | 42.76 | 3.26 | 0.23 | 1.95 |
| | | | | | | | | | HLA-A*02:01 | | | 0.83 | 85.23 | 15.95 | 0.76 | |
| LKLPAPIQM | KLPAAIEQL | HLA-C*07:02 | Bacteroides fragilis NCTC 9343 | Moderate | 0.73 | 54.40 | 0.24 | 8.18 | HLA-A*02:01 | DNAH5 | Weak | 0.96 | 116.61 | 13.25 | 0.08 | 0.15 |
| KSRNNAVSY | AVMKSRNVL | HLA-B*15:01 | Bacteroides fragilis NCTC 9343 | Moderate | 0.63 | 93.53 | 0.08 | 8.18 | HLA-B*13:02 | LAMA5 | Moderate | 0.85 | 57.19 | 7.63 | 0.18 | 44.70 |
| IMLLILFIA | TLLILSVAV | HLA-A*02:01 | Bacteroides fragilis NCTC 9343 | Moderate | 0.63 | 23.77 | 0.50 | 8.18 | HLA-A*02:01 | ERVH48-1 | Weak | 0.92 | 45.47 | 6.16 | 0.41 | 0.03 |
| | LLILSVAVL | | | | | | | | HLA-B*15:01 | | | 0.93 | 86.29 | 3.28 | 0.60 | |

**Fig. 3 | Detection of Microbial Mimicry Epitopes in CRC Samples.** Data volumes associated with the identification of microbial epitopes (a) and host-derived epitopes (b) in a CRC clinical case analyzed using MimicNeoAI. (c) Exclusive intersections between microbial and host-derived epitopes, depicted separately for tumor and normal samples, highlighting distinctions between HLA-I and HLA-II epitopes. Counts of host-derived epitopes (d) and microbial epitopes (e) categorized by immunogenicity levels: weakly, moderately, and highly, encompassing both HLA-I and HLA-II

epitopes. The number of HLA-I (f) and HLA-II (g) peptide pairs with varying similarity degrees identified by comparing microbial and host-derived epitopes across different similarity thresholds. Relative abundance differences of the top 10 detected microorganisms in tumor versus normal samples at the genus (h) and species (i) levels. (j) Pairs of microbial mimicry antigens and tumor neoantigens identified through sequence similarity analysis. Each entry is annotated with HLA allele, immunogenicity score, HLA binding metrics, mutated gene, microbial antigen source, gene expression, and microbe abundance.

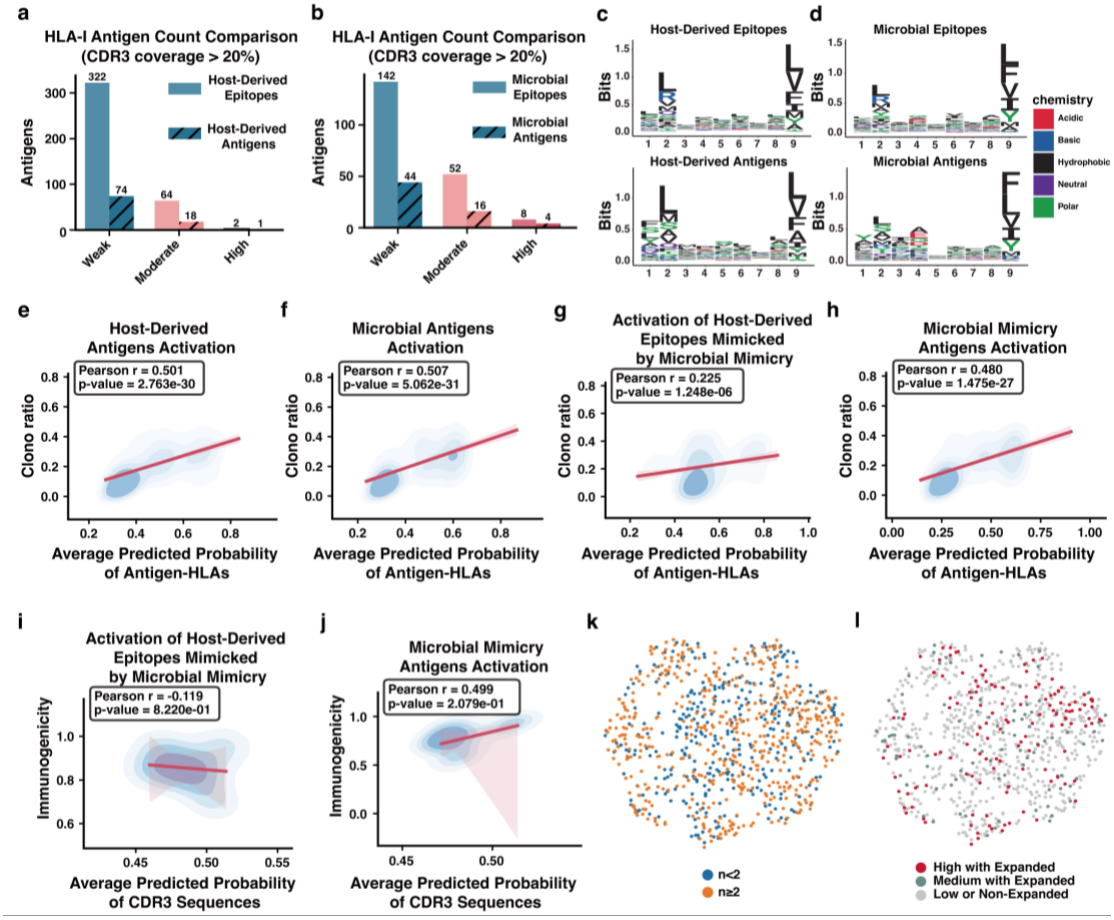



**Fig. 4 | Validation of Epitope-HLA Immunogenicity Through TCR-Epitope-HLA Binding specificity.** Comparison of counts for HLA-I epitopes and antigens derived from host (a) and microbe (b) sources. Epitopes classified as weak, moderate, and high were further evaluated for TCR-epitope-HLA binding specificity. Epitopes with a binding probability >0.8 for over 20% of CDR3 sequences were selected as candidate

antigens. Logarithmic frequency distribution of host-derived (c) and microbial (d) epitopes and antigens at critical amino acid positions. Correlation between T cell receptor clonal expansion ratios and average predicted binding probabilities to antigen-HLA complexes for T cells activated by various antigen types: Host-derived antigen (e), Microbial antigen (f), Host-derived antigens mimicked by microbial mimicry antigens (g), and microbial mimicry antigens (h). The average predicted probability reflects the mean likelihood of each CDR3 sequence binding to the entire antigen-HLA collection. Correlation between the immunogenicity of epitope-HLA complexes and average predicted probabilities of CDR3 sequence binding for host-derived epitopes mimicked by microbial mimicry antigens (i) and for microbial mimicry antigens (j). The average predicted probability indicates the predicted binding likelihood of each epitope-HLA with all T cell clones' CDR3 sequences. Distribution of T cells categorized by their binding specificity to microbial mimicry antigen-HLAs, with the left panel (k) depicting T cell clonal expansion and the right panel (l) illustrating the average probability of T cell activation by microbial mimicry antigen-HLAs, classified into High (⩾0.6) with Expanded, Medium (0.4–0.6) with Expanded, and Low (<0.4) or Non-Expanded binding probability categories.

## Tables

**Table 1. Compilation of literature-reported microbes enriched in different types of tumors.**

**Table 2. Custom database of RefSeq genomes derived from human-associated microbial catalogues for creating PathSeq reference.**

**Table 3. The features considered by our pipeline and their descriptions.**

**Table 4. Peptide-HLA complexes used for model construction and performance evaluation.**

**Table 5. Performance comparison of MimicNeoAI and other algorithms on microbial antigen validation dataset.**

**Table 6. Microbial mimicry antigens and tumor neoantigen partners detected from CRC cohort.**

**Table 7. Predicted binding affinity of epitope-MHCs with TCRs detected in CRC tumor sample.**

## Supplementary information

### Generation of Epitope candidates

Central to the MimicNeoAI framework, the microbial antigen detection sub-pipeline is adept at identifying low-abundance tumor-associated microbial antigens from WTS data. This methodology enables the detection of microbial epitope candidates even in low-biomass environments, with compatibility for large public datasets facilitating scalable, pan-cancer analyses. Initially, cleaned sequencing reads undergo processing to eliminate host-derived sequences, after which the remaining non-host reads are aligned to a custom-built microbial genome reference, minimizing contamination and enhancing microbiome identification. This reference is constructed through a

combination of LLM-based annotation and manual curation, ensuring the inclusion of microorganisms capable of colonizing the human body or influencing cancer progression or anti-tumor treatment response. Subsequently, microbial reads were translated into proteins and matched against protein databases. A rigorous filtering process is applied of the alignment results, allowing for the precise identification of microbial peptides that may serve as antigens, thereby enhancing the overall accuracy of antigen detection.

The neoantigen candidate detection sub-pipeline of MimicNeoAI is designed to detect neoantigens from WES data. This process begins with the alignment of cleaned sequencing reads to the human reference genome, identifying somatic mutations that are subsequently filtered to exclude low-quality variants absent in normal control samples. Additionally, expression quantification from WTS data is utilized, retaining only mutations in genes with detectable expression, thereby ensuring that identified mutations are actively transcribed and more likely to be presented as neoantigens.

**Prediction and Prioritization of Epitope immunogenicity**

To enhance the prioritization of peptides with significant immunogenic potential, MimicNeoAI incorporates HLA typing from WTS or WES data. Building on the peptide candidates generated by the aforementioned sub-pipelines, the framework transcends basic peptide-HLA binding affinity predictions. Utilizing a variety of algorithms to calculate binding affinities, MimicNeoAI further integrates an immunogenicity prediction model. This model leverages BiLSTM networks to interpret long-range amino acid interactions within peptide-HLA complexes, complemented by a Fully Connected Neural Networks (FCNN) that assessing eight physicochemical properties (amino acid composition, polarity, residue volume, net charge, hydrophobicity, Boman index, aliphatic index, and isoelectric point).

During the post-model screening steps, MimicNeoAI employs an extensive selection and ranking process, considering characteristics of both microbial and host-derived

peptides generated from the aforementioned pipeline. The microbial peptides characteristics include metrics such as total library size, microbial reads ratio, and microbial abundance derived from nucleic acid-level microbial alignment, along with protein reference sequence alignment results (e.g., percentage identity, gap opening, e-value, and bitscore). In contrast, the host-derived peptides characteristics encompass DNA-level metrics (e.g., depth, variant allele frequency, and allele frequency), gene expression levels, mutation details (positions, quality scores), and protein-level metrics (split protein positions). By integrating affinity and immunogenicity scores, MimicNeoAI effectively ranks the microbial antigen and tumor neoantigen candidates, categorizing them into non-immunogenic, weakly immunogenic, moderately immunogenic, and highly immunogenic classes.

**Identification of Microbial Mimicry Antigens**

Following the detection of microbial antigen candidates and tumor neoantigen candidates, the next phase involves identifying microbial mimicry antigens. This is accomplished by analyzing the length of the longest common substring (LCS) in relation to the maximum length of the two peptides, thereby normalizing the comparison.

**Materials of immunogenicity prediction model**

To train MimicNeoAI's immunogenicity prediction model, we integrated peptide-HLA complexes from multiple sources (Table 4). The host-derived (human) dataset, compiled from the IEDB, includes 117 HLA-I alleles and 12,039 peptides, resulting in 15,482 peptide-HLA complexes, primarily sourced through two specialized algorithms: DeepHLApan and DeepImmuno. This dataset was balanced with 7,741 immunogenic and 7,741 non-immunogenic complexes. The microbial dataset, derived from MicroEpitope, includes 4,556 HLA-I and 650 HLA-II alleles, with 17,894 HLA-I-binding and 1,018 HLA-II-binding peptides. These yielded 28,246 peptide-HLA-I complexes (14,123 immunogenic, 14,123 synthetic non-immunogenic) and 1,754 peptide-HLA-II complexes (877 immunogenic, 877 synthetic non-immunogenic),

ensuring balanced positive/negative representation. Additionally, we compiled a separate microbial antigen validation set of 46 peptide-HLA complexes, entirely distinct from the training data. This set combines 23 literature-verified immunogenic microbial complexes and 23 randomly generated non-immunogenic complexes for unbiased evaluation (Table 5).

## Supplementary figures

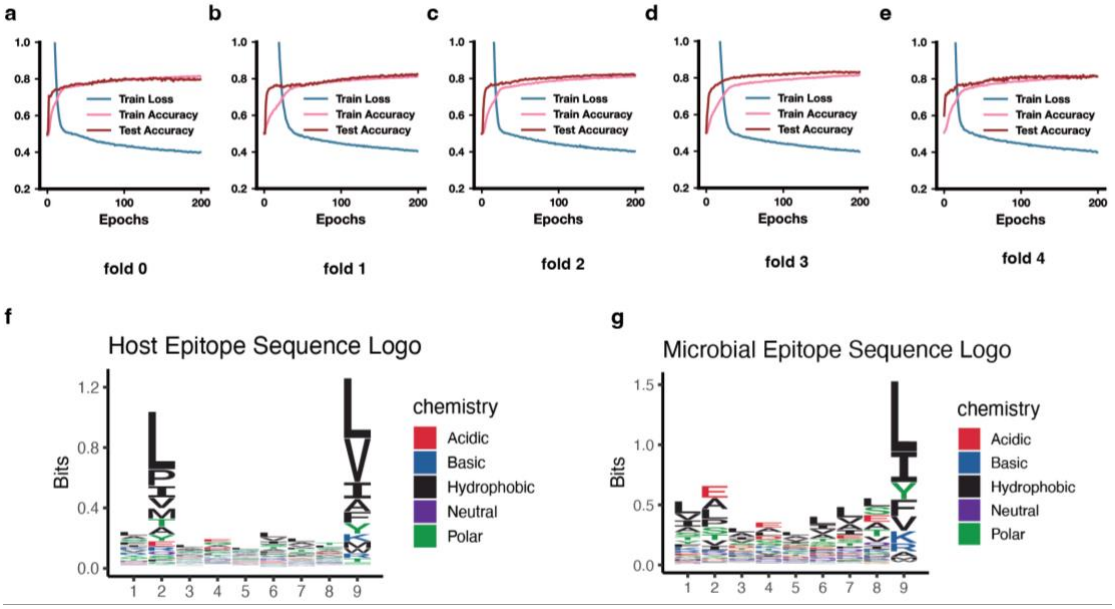



**Fig. S1 Evaluation of Training Performance in Five-Fold Cross-Validation and Analysis of Peptide Frequency in the Training Dataset.** Plots of train loss, train accuracy, and test accuracy for each fold of the five-fold cross-validation, reflecting the model's robustness, as shown in (a-e). Sequence logo plots of immunogenic peptides with length 9 from the host-derived dataset (f) and Microbial Dataset (g), showing the amino acid frequency distribution in positive training samples.