

Integrating Metagenomics and Immunoinformatics to Prioritize Antigens and Immune-Modulating Molecules from Environmental Microbiomes

Harsh Purohit* and Jignesh Kamdar

Department of Microbiology, School of Science, RK University, Rajkot - 360020, India

Abstract. Environmental microbiomes – especially diverse soil and terrestrial communities – harbor an immense and largely untapped reservoir of microbial genetic diversity. Integrating metagenomic sequencing of these habitats with immunoinformatics offers a new avenue to discover candidate antigens and microbial metabolites that modulate immunity. Metagenomics characterizes the mixed microbial genomes present in soil, water, or microplastic-associated biofilms, enabling recovery of genes and proteins (often via assembly and binning into metagenome-assembled genomes). Immunoinformatics tools can then predict B- and T-cell epitopes or antigenicity of these metagenome-derived proteins, prioritizing those likely to engage host immune receptors. Pipelines such as reverse vaccinology frameworks (e.g., ReVac, VaxiJen, Vaxign) screen genomes using features like surface localization, epitope content, and conservation. New Artificial Intelligence/Machine Learning approaches further refine candidate ranking by integrating multiple predicted features, including MHC-binding profiles and antigenicity. Beyond classical pathogen antigens, environmental microbes can provide innate immune stimuli (e.g., flagellin, LPS) and novel secondary metabolites (e.g., rapamycin-like immunosuppressants), which can be predicted or discovered via metagenomics. We review current sequencing and immunoinformatic workflows – from gene calling to epitope prediction and Machine Learning-based ranking – applied to soil and related microbiomes. We highlight examples of known immunomodulators from soil microbes and discuss how Artificial Intelligence-driven design can accelerate mining of environmental microbiomes for biomedical targets. This integration of metagenomics and immunoinformatics promises a One Health perspective that explicitly links human, animal, and environmental health, expanding antigen discovery beyond gut- or pathogen-centric views, while highlighting computational strategies, limitations, and future directions of this approach.

* Corresponding author: harshpurohit1998@gmail.com

1 Introduction

Microbial communities in diverse environments – notably soil, water, and built habitats – have profound but underexplored roles in ecosystem and human health [1]. Classical microbiome research has focused mainly on the human gut, revealing links to nutrition and immunity [2], but terrestrial and aquatic environments harbor vast microbial diversity. For example, the soil microbiome alone contains an estimated ~25% of Earth’s microbial biodiversity and acts as a “seed bank” for plant and even human microbiomes [3]. Exposure to natural soil microbes has been associated with enhanced immune resilience and reduced allergy risk [4]. These observations underscore that environmental microbes are not only crucial for ecology and agriculture but may also influence immunity through direct contact or ingestion [1,4]. However, most immunological research still centers on known pathogens or commensals; the broader immunological potential of environmental microbes remains largely untapped [1,5].

Metagenomics – the culture-independent sequencing of all DNA in an environmental sample – now enables comprehensive profiling of microbial communities [6]. Modern sequencing (shotgun NGS and long-read technologies) coupled with bioinformatics allows reconstruction of community composition, functional genes, and even draft genomes of uncultured organisms (metagenome-assembled genomes, MAGs) [7]. This wealth of sequence data opens the possibility of mining environmental microbiomes for molecules of immunological interest. In parallel, immunoinformatics – the application of computational methods to immune system questions – has matured with genomic medicine. Immunoinformatics encompasses epitope prediction for B-cell and T-cell responses, antigenicity scoring, and vaccine design tools. When applied to pathogen genomes, these methods accelerate candidate discovery [8].

Integrating metagenomics of environmental microbiomes with immunoinformatics, therefore, offers a novel paradigm: one can screen genes from soil or aquatic microbiomes to find candidate antigens or immune-active compounds [9]. Such a One Health perspective transcends conventional pathogen-centric vaccine design. In this review’s first half, we introduce the background and rationale for this integration: we review environmental microbiome diversity (with emphasis on soil and related niches) [10], the basics of immunoinformatic antigen discovery, and the emerging computational pipelines that marry the two. We aim to highlight how workflows — from metagenomic sequencing through to epitope prediction and ML-based prioritization — can be applied to discover vaccine candidates, diagnostics targets, and immunomodulators in environmental microbes. Examples from soil and other habitats will illustrate general principles. We also discuss existing tools, AI-enhanced methods, and the conceptual potential of this approach, setting the stage for specific applications and challenges.

2 Environmental microbiomes and metagenomics

2.1 Soil and terrestrial microbiomes

Soils and terrestrial habitats are among the richest microbial reservoirs on Earth. A single gram of soil can contain billions of bacterial and fungal cells spanning thousands of species [7]. This diversity arises from complex gradients of nutrients, moisture, and plant interactions. Notably, recent perspectives highlight a “soil–plant–human gut” axis, proposing that soil microbes can ultimately influence human gut communities and immunity [11]. For

example, microbes in edible plants and even incidental soil ingestion (geophagy) are thought to seed the gut microbiome and may enhance immunological resilience [12]. Beyond directly impacting health, soil microbes produce myriad bioactive molecules, many of which have immunological effects [13].

In addition to bulk soil, we consider related terrestrial niches. Microplastic pollution in soil and water has become a novel substrate for microbial biofilms; such plastisphere communities differ in composition and may concentrate different microbial taxa. Although still emerging, studies show microplastic-associated biofilms can perturb local microbial ecology [14]. Built environments (e.g., indoor air, dust) also form distinct microbiomes, though with less diversity than natural soil. Throughout, soil and terrestrial microbiomes often interface with animals, plants, and water systems, enabling gene flow of microbes and their molecules into broader ecosystems.

The ecological importance of soil microbiomes extends to immunity. The “biodiversity hypothesis” suggests that reduced exposure to environmental microbiota in urbanized settings may underlie rising allergic and autoimmune diseases [15]. Exposure to soil-derived microorganisms or their components can skew human immune responses toward regulatory (tolerogenic) pathways [16]. Notably, one study [17] argues that intentional soil exposure can modulate the immune system (e.g., through *Bacillus* species acting as TLR agonists). Collectively, the rich genetic and metabolic capacity of soil microbes makes them a promising target for immunological exploration.

2.2 Metagenomic sequencing and analysis

Characterizing environmental microbiomes relies heavily on sequencing-based approaches. Marker gene surveys, typically involving 16S rRNA amplicon sequencing—a method that targets conserved regions of the bacterial 16S rRNA gene to identify and classify microbes—provide taxonomic snapshots but do not reveal functional genes or novel proteins. In contrast, whole-community shotgun metagenomics sequences all DNA fragments, enabling analyses of genetic potential [6,17]. A typical metagenomic workflow begins with DNA extraction from environmental samples (requiring careful removal of inhibitors like humic acids in soil), followed by sequencing on short-read (Illumina) and/or long-read (PacBio, Nanopore) platforms [18]. Post-sequencing reads undergo quality filtering and removal of contaminant/human sequences. Assembly algorithms then stitch reads into contigs or scaffolds. Because environmental samples contain multiple organisms, read binning is often used: assembly-free binning (clustering reads) or contig binning (clustering assembled contigs by sequence composition and abundance) can group sequences into bins that represent draft genomes. These metagenome-assembled genomes (MAGs) can recover near-complete genomic sequences of previously uncultured bacteria and archaea [19]. Even without full genome bins, open reading frame (ORF) prediction tools (e.g., Prodigal) can identify genes on contigs [20].

Once genes are predicted, they are annotated via homology searches (e.g., using BLAST or HMMer against databases), but a large fraction of environmental ORFs have no known homologs. Nevertheless, predicted proteins can be translated into amino acid sequences for downstream analyses. Importantly for immunological applications, many environmental proteins will be novel, underscoring the need for *ab initio* prediction methods (e.g., antigenicity predictors) rather than relying solely on known reference antigens. Functional metagenomics is also used: in some studies, metagenomic libraries in model hosts (like *E. coli*) are screened for phenotypes. Such activity-based screens have uncovered new

antibiotics or enzymes [21]. A similar concept could be used to find immunologically active compounds (e.g., expressing environmental biosynthetic gene clusters in a host and assaying for cytokine modulation), though this is experimental rather than purely computational. For immunoinformatics integration, the focus is on sequence-driven analysis of metagenomic data.

2.3 Microplastics and anthropogenic niches

Microplastic particles provide novel surfaces for microbial colonization in soil and water [22]. Studies report that microplastic biofilms may harbor distinct taxa and differentially express genes (e.g., stress resistance) compared to surrounding soil [23]. These biofilm-associated shifts can change the local repertoire and concentration of microbe-associated molecular patterns (MAMPs), thereby modulating how host immune systems are triggered upon exposure. While research is still nascent, one can apply the same metagenomic and immunoinformatic approaches to these communities. For instance, if microplastics concentrate certain Gram-negative bacteria, their LPS content (a potent TLR4 agonist) might be enriched [24]. Environmental sequencing of built environments (homes, urban dust) similarly opens avenues: e.g., one could screen air microbiome data for fungal or bacterial antigens relevant to respiratory immunity [25].

3 Immunoinformatics approaches for antigen prediction.

3.1 Principles of epitope prediction and antigenicity

Immunoinformatics applies computational models to predict how immune systems recognize antigens. A core task is predicting epitopes – the specific parts of antigens (peptides) recognized by B-cell receptors (antibodies) or by T-cell receptors in the context of MHC presentation. T-cell epitopes are typically short linear peptides (8–25 amino acids) that bind to host MHC molecules. Algorithms like NetMHC and NetMHCIIpan use machine learning models trained on experimental peptide–MHC binding data to predict which peptides from a protein will bind to specific HLA (MHC) alleles. These tools cover both class I and class II MHC molecules and can accommodate host allele diversity [26]. B-cell epitope prediction is more challenging because most B-cell epitopes are conformational (discontinuous in sequence). However, linear B-cell predictors (e.g., BepiPred) and structural methods (e.g., ElliPro) exist to identify surface-exposed peptide segments likely to be antibody-accessible [27].

Beyond epitopes, antigenicity predictors like VaxiJen assess a protein's likelihood of being a protective antigen by analyzing its physicochemical properties. These do not rely on sequence similarity but use features (e.g., hydrophobicity, molecular weight) to score proteins against known antigens [28]. Such tools can flag proteins with high antigen potential even if sequence homologs are unknown. Immunoinformatics also leverages epitope and antigen databases. The Immune Epitope Database (IEDB) collects hundreds of thousands of experimentally validated epitopes and is widely used both as a tool suite (IEDB Analysis Resource) and as training data for new predictors. Specialized databases contain MHC allele sequences (IMGT) or epitope–MHC structural data (SYFPEITHI, MHCBN), further aiding model development. Recent reviews highlight that the availability of high-throughput sequencing and binding assay data has greatly improved epitope predictor accuracy [29].

3.2 Reverse vaccinology and antigen prioritization.

Reverse vaccinology refers to genome-based screening of potential vaccine candidates. The idea, pioneered in the 2000s, is to computationally scan a pathogen's genome for proteins with features of good antigens: typically, surface exposure (so the immune system can access them), presence of T- or B-cell epitopes, high conservation across strains, and appropriate expression profile [30]. Modern pipelines (e.g., Vaxign, ReVac) implement these criteria systematically. ReVac, for instance, analyzes multiple genomes to find proteins conserved among strains, not filtered out by adverse traits (like repetitive sequences), and then ranks them based on combined feature scores [31,32]. Many such tools follow either "filtering" approaches (stepwise elimination of proteins lacking a feature) or machine-learning classification (training on known antigens and non-antigens). For example, NERVE, Jenner-Predict, and VacSol are other reverse vaccinology pipelines that implement similar workflows [33]. In all cases, the output is a prioritized list of protein candidates for vaccine or diagnostic development. For bacterial pathogens, these pipelines often start with the predicted proteome (all proteins from a genome) and then apply subcellular localization predictors (e.g., SignalP, TMHMM, PSORTb) to find secreted or surface proteins and epitope predictors (MHC-binding and linear epitope content). Scoring schemes weight each feature; ReVac, for example, assigns points for antigenic signals and conservation, generating a ranked score. Such ranked lists greatly reduce experimental burden by focusing on the most promising antigens [32].

3.3 AI and machine learning enhancements

The field of immunoinformatics increasingly harnesses advanced AI/ML methods [34]. Many epitope predictors themselves use deep learning (e.g., NetMHCpan, MHCflurry), which improve as training data grow. Newer approaches integrate structure prediction (e.g., modeling peptide–MHC complexes) to refine epitope binding predictions. Importantly, in the context of vaccine design, recent work has explored AI-guided epitope focusing iterative algorithms to optimize epitope sets to maximize predicted immune coverage while minimizing issues like allergenicity [34]. For example, recent studies [35] describe "AI-driven antigen refinement," where machine learning is used to focus on immunodominant epitopes and even fuse conserved microbial motifs (MAMPs) with antigens to enhance broad immunity. Such methods illustrate how AI can not only predict epitopes but also design composite antigens (e.g., combining epitopes with bacterial adjuvant motifs) automatically. Machine learning is also used for multi-feature ranking of candidates. For instance, after predicting many possible epitopes, a classifier can rank the likelihood that an epitope will be immunogenic in vivo or integrate features like predicted allergenicity and toxicity to flag safe candidates. These AI-enhanced workflows are still emerging for environmental applications, but are key to handling the enormous data volume (millions of proteins) from metagenomes [37].

3.4 Integrating metagenomics with immunoinformatics

Bringing together metagenomics and immunoinformatics involves several computational steps. First, environmental DNA is sequenced and assembled, as described above. From assemblies or raw reads, ORF-finding tools predict protein-coding sequences. These metagenome-derived protein sequences (often millions, including many fragments) become inputs for immunoinformatic analysis. A typical workflow then applies epitope prediction to each protein. For T-cell epitopes, software like NetMHCpan or MHCflurry predicts binding affinities of all peptides (sliding windows) against a panel of HLA alleles representative of

target populations, flagging high-affinity binders as candidate epitopes [38,39]. Similarly, B-cell linear epitope predictors scan protein sequences to identify likely antibody-accessible regions, for example, using BepiPred-2.0 [40]. Proteins rich in predicted epitopes (for example, with multiple high-affinity peptides for common HLA alleles) are scored higher. Tools like VaxiJen can then compute an overall antigenicity score for each protein based on sequence-derived physicochemical features, independently of sequence homology [41]. As summarized in Figure 1, we outline an integrated metagenomic-immunoinformatic workflow that links environmental sampling, sequencing, and functional annotation to in silico antigen and immunomodulator prediction from environmental microbiomes.

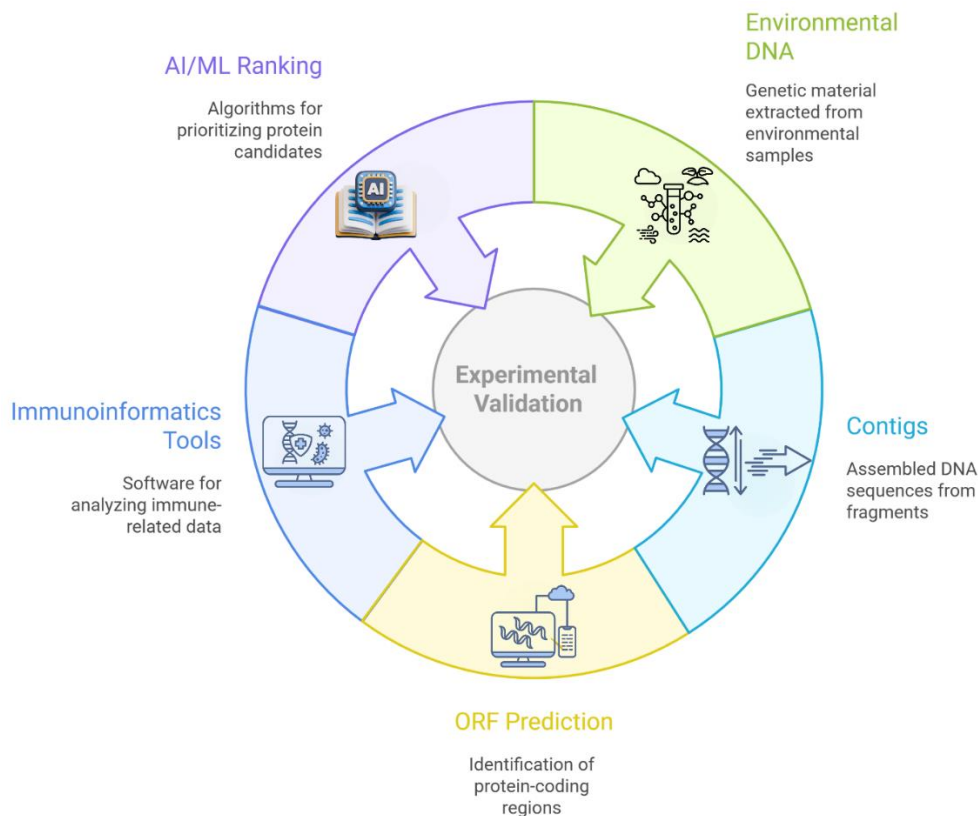


Fig. 1. Integrated metagenomic-immunoinformatic workflow for antigen and immunomodulator discovery from environmental microbiomes.

Parallel filters further reduce candidates. For vaccine antigens, one may require that proteins be non-homologous to the host proteome (to avoid autoimmunity) and not overly similar to microbiome commensals (unless a broad-spectrum vaccine is desired). Transmembrane helix prediction and signal peptide detection help prioritize secreted/surface proteins (typical antigenic targets). For immunomodulators, different criteria might apply: for example, gene clusters resembling known biosynthetic pathways, detected by secondary-metabolite genome-mining tools such as antiSMASH, can indicate putative natural product biosynthesis loci [42]. Sequences matching conserved innate immune ligands (e.g., flagellin-like motifs) are also noted. Importantly, multiple predictions are combined. Just as ReVac aggregates diverse features for each protein to prioritize bacterial vaccine candidates [43], an integrated

pipeline for metagenomes would score each candidate protein by a composite scheme. For instance, a scoring rubric might add points for each predicted strong MHC binder, each signal peptide, each broad epitope coverage across HLA supertypes, and subtract points if the protein has predicted human homology. Machine learning classifiers could be trained on known antigens to refine this scoring, although curated datasets for environmental antigens are currently sparse. The result is a prioritized list of proteins from the metagenome ranked by their vaccine potential.

Several “meta-immunoinformatics” pipelines are being developed (or could be developed) to automate these steps. While few are published specifically for environmental data, related concepts exist. For example, immunoinformatics studies often use IEDB’s TepiTool to perform standardized T-cell epitope screening and combine the output with antigenicity servers such as VaxiJen or ANTIGENpro [41,44]. A conceivable pipeline would take metagenomic ORFs, feed them through NetMHCpan (for a set of common HLA class I/II alleles), BepiPred, VaxiJen, and subcellular localization tools, then use a custom script to rank and output top candidates. Workflows can be parallelized on clusters to handle large datasets. In the context of gut or pathogen vaccines, web servers such as Vaxign/Vaxign2 already allow users to upload bacterial genomes and obtain predicted adhesins, subcellular localization, and epitopes in a reverse-vaccinology framework [45]. Adapting such tools to fragmented metagenomes would require assembling contigs into draft genomes (bins/MAGs) and running the pipeline genome-by-genome. ReVac’s strategy of analyzing many genomes in parallel is conceptually similar – in our case, each MAG or contig set is treated as an “isolate” in the dataset [43]. The key integration point is to treat metagenomic genes as input proteomes for immunoinformatics, rather than relying solely on single cultured genomes.

3.5 Immune-modulating molecule discovery

Besides classical antigens, environmental microbiomes are rich sources of immune-modulating compounds [46]. These include microbial-associated molecular patterns (MAMPs) and secondary metabolites with adjuvant or suppressive activity [46]. Immunoinformatic analysis can be extended to flag potential MAMPs in metagenomes. For example, sequence motifs characteristic of flagellin (TLR5 ligand) or unmethylated CpG DNA (TLR9 ligand) could be searched [47]. In one study, synthetic bacterial flagellin was shown to robustly activate innate immunity via TLR5 [48]. Thus, genes encoding flagellin or other TLR-stimulatory proteins in environmental microbes are intrinsically immunostimulatory and might be harnessed as adjuvants. Similar logic applies to lipopolysaccharide (LPS) biosynthesis genes (TLR4 agonist) and peptidoglycan-related motifs (TLR2), all identifiable via sequence homology [47].

For small molecule immunomodulators, pipelines would need to detect biosynthetic gene clusters (BGCs) in metagenomes. Tools like antiSMASH can find nonribosomal peptide synthetase (NRPS) or polyketide synthase (PKS) gene clusters [42]. Many known immunosuppressive natural products come from soil BGCs: for instance, rapamycin (an mTOR inhibitor) and tacrolimus (FK506) are macrolide antibiotics produced by soil *Streptomyces*, and cyclosporine comes from a fungus [49]. A metagenomic pipeline could flag similar polyketide clusters and compare them to databases of known immunomodulators. Once gene clusters of interest are identified, bioinformatics (including structure prediction and docking) could predict their immunological target (e.g., mTOR, calcineurin). Alternatively, one can scan metagenomic proteomes with epitope prediction tools to find peptides that themselves act as immune regulators. For example, some bacterial peptides induce regulatory cytokines (e.g. IL-10) or Th17 responses; predictors like IL10Pred or

IL17eScan (available online) could be used on environmental proteins to find such immunomodulatory motifs [50]. Although still experimental, this conceptually fits within immunoinformatics.

4 Practical Applications and Case Studies

Metagenomic-immunoinformatics integration has already begun yielding practical advances. In vaccine development, for example, candidate antigens can be discovered from previously uncharacterized pathogens or microbiome organisms. [51] emphasize that “integration of knowledge from Omics and Bioinformatics will certainly boost vaccine research and development, leading to novel therapeutic tools” [51]. Indeed, recent studies have used multi-omics data plus immunoinformatics to design experimental vaccines [52,53]. For instance, Shao et al. applied a multi-step antigen discovery pipeline to identify 13 potential vaccine proteins from the dog tapeworm *Echinococcus granulosus* [52]. They combined transcriptomics (stage-specific expression), proteomics (shared excretory/secretory proteins), literature curation, and *in silico* antigenicity to winnow thousands of candidates to a prioritized set [52]. This pipeline yielded six top proteins (including enolase and fatty-acid binding proteins) that were expressed as recombinant vaccine components, and a subsequent vaccination trial in dogs validated that the predicted cocktail induced partial protection and modulated the gut microbiota. This study illustrates how integrated analysis of environmental (canine fecal) microbiome data and immunoinformatics can rapidly prioritize novel antigens for vaccines [52].

In another case, Razzak et al. designed a self-amplifying RNA (saRNA) multi-epitope vaccine against *Helicobacter pylori* using an end-to-end computational pipeline [53]. Five antigenic bacterial proteins were selected (e.g., urease, adhesins) based on antigenicity, virulence, conservation, and lack of human homology [53]. B- and T-cell epitopes were predicted (IEDB tools), filtered for immunogenicity and non-allergenicity (AllerTOP), and assembled into a fusion peptide with linkers and adjuvant sequences; structural modeling and docking confirmed stable binding to immune receptors [53]. The resulting saRNA vaccine construct was predicted to be antigenic and non-allergenic, with broad population coverage, although the authors caution that “challenges remain in translating computational predictions into experimental efficacy” [53]. Nonetheless, this work underscores how immunoinformatics can convert omics-derived antigen information into concrete vaccine designs [51,53].

Beyond vaccines, integrated pipelines have diagnostic and therapeutic uses. Environmental metagenomes can reveal novel immune biomarkers: for example, surveillance of hospital wastewater or air by shotgun sequencing can detect emerging pathogens and potential allergens. In one improved environmental surveillance workflow, metagenomic sampling was used to track nosocomial pathogens; untargeted sequencing “is advantageous in identifying novel or rapidly emerging pathogens”, although background contamination and low biomass pose challenges [54]. Such approaches could be applied to water or soil microbiomes to flag zoonoses or allergens before they spread. In diagnostics, predicted microbial antigens (from commensals or pathogens) may serve as biomarkers or vaccine targets for infection and allergy. For instance, immunoinformatics screening of environmental or commensal microbial proteomes can suggest epitopes to include in broad-spectrum diagnostics.

Microbiome-derived molecules with immunomodulatory activity are another frontier. Commensal bacteria often secrete peptides or metabolites that tune host immunity, and these

can be predicted from metagenomic data [55]. For example, [56] identified peptides from *Bifidobacterium longum* and *Bacteroides fragilis* that modulate intestinal cytokines and antigen-presenting cell phenotypes. In healthy donors, a *B. longum*-derived peptide (B7) reduced chemokine receptor expression on dendritic cells, demonstrating immunomodulatory capacity. Such findings suggest pipelines could mine environmental (or gut) metaproteomes for small molecules with anti-inflammatory or immunostimulatory effects. These might become novel probiotic-derived therapies or adjunctive treatments, engineered through synthetic biology.

5 Limitations and Challenges

While promising, this integrated approach faces significant caveats. Data limitations: Environmental metagenomes are often incomplete or highly fragmentary, especially for rare or novel microbes. As a result, predicted proteins may be partial or misassembled, limiting the reliability of downstream epitope predictions. Many microbial taxa lack close reference genomes, so homology-based annotations (antigenicity, function) can fail. Additionally, bias in metagenomic sampling (e.g., DNA extraction, sequencing depth) may skew which organisms and genes are detected. Poor annotation of non-model organisms means candidate antigens might be incorrectly labeled “hypothetical protein” with unknown immune relevance.

5.1 Computational uncertainties

In silico predictions of epitopes and antigenicity carry false positives. Tools like VaxiJen and NetMHCpan have known limitations; predicted “antigenic” proteins might not be immunogenic in vivo, and vice versa. Allergenicity predictors (e.g., AllerTOP) reduce the risk of allergic reactions, but novel epitopes might have unanticipated cross-reactivity [57]. Over-reliance on sequence-based methods can overlook structural or post-translational context. As [53] notes, even a “comprehensive immunoinformatics pipeline” must be experimentally validated, since in silico findings do not guarantee real-world efficacy. There is also a risk of over-interpretation: without lab confirmation, predicted epitopes and modulators should be viewed as hypotheses, not proven candidates.

5.2 Safety and regulatory concerns

Predicted antigens from environmental microbes could potentially mimic human proteins, risking autoimmunity if not properly filtered. As part of good practice, pipelines typically remove peptides with high similarity to human sequences. Similarly, immunoinformatics workflows include multiple safety screens [58]. Nevertheless, computational checks are imperfect, and rare allergenic motifs might be missed. For vaccines, murine or in vitro validation is required before animal or human trials. Regulatory pathways for microbiome-based therapies are still evolving; introducing an “environmental antigen” may face novel oversight.

5.3 Biological complexity

Even if a peptide is predicted to bind MHC molecules or stimulate a T-cell, the host immune system’s response is influenced by many factors (adjuvants, epitope processing, HLA diversity). Environmental antigens often come from non-traditional hosts or commensals; their context of exposure will affect immunogenicity [59]. Immunosenescence, microbiota

interactions, and individual host genetics can greatly modulate outcomes. For example, a peptide that is immunomodulatory in a healthy gut (like *B. longum* B7) may behave differently in inflammatory disease or at another site. Thus, predictions provide leads, but validating safety and efficacy *in vivo* is essential [56].

6 Opportunities and Future Directions

Despite challenges, the horizon is bright for this combined approach. Vaccine innovation: Mining environmental microbiomes could uncover protective antigens from as-yet-uncultured pathogens or beneficial microbes. A reservoir of novel epitopes may broaden vaccine targets beyond classic pathogens. For emerging infectious diseases, metagenomic surveillance of wildlife and vectors combined with epitope prediction could accelerate “reverse vaccinology” for zoonoses [30,54]. In cancer immunotherapy, tumor-associated microbiota metabolites or peptides might also be discovered that synergize with checkpoint inhibitors [46].

6.1 Diagnostics and surveillance

High-throughput pipelines could generate libraries of predicted microbial antigens for diagnostic assays. For example, panels of synthetic peptides (from soil microbes or marine bacteria) might serve as controls or probes in serological surveys, broadening our ability to detect exposure to novel agents. Environmental sensors could couple metagenomic reads with on-the-fly immunoinformatics to raise alerts for novel harmful antigens.

6.2 Immune-modulating therapies

There is great interest in harnessing microbiome-derived molecules (postbiotics) to regulate immunity [46]. Computational pipelines can screen microbial genomes for peptides that, for instance, induce regulatory T cells or dampen inflammation. Such candidates could become novel probiotics or biologics. For instance, resources such as MAHMI-like peptide databases and related work on gut microbiota-derived bioactive peptides allow *in silico* screening of gut metagenomes for immunomodulatory molecules [56]. Extending this to soil or ocean microbiomes could lead to new anti-inflammatory compounds. Engineered microbes (e.g., probiotics carrying synthetic genes) might be designed based on pipeline predictions to secrete desired immune signals *in situ*.

6.3 Computational advances

The field will benefit from AI and machine learning that improve epitope prediction beyond current rule-based tools [36, 60]. Integrating multi-omics could refine target identification. For example, identifying which microbial peptides are presented by MHC on host cells via MS-eluted ligand or immunopeptidomics data would help validate *in silico* antigenicity calls [26]. Cloud-based platforms and web-accessible reverse vaccinology servers could automate end-to-end workflows for non-specialists [45].

7 Conclusion

Integrating metagenomics with immunoinformatics offers a powerful route to discovering novel antigens and immunomodulators in complex ecosystems. By sequencing environmental DNA and applying computational epitope prediction, researchers can

prioritize leads for vaccines, diagnostics, and therapies that would be difficult to identify using traditional methods alone. Case studies, from parasite vaccine targets in dogs to microbiome-derived peptides in human disease, demonstrate the promise of this approach. However, the pipeline is not foolproof. Data quality, computational limits, and safety concerns mean that rigorous validation remains essential. Predicted candidates must be tested experimentally to confirm immunogenicity and to rule out adverse reactions. Looking ahead, advances in sequencing technology, bioinformatics, and systems immunology will continue to sharpen these tools. Larger and more diverse genomic databases will expand the search space, while AI-driven predictors and high-throughput validation platforms (e.g., yeast display, high-content immunophenotyping) will help close the loop between prediction and experiment. The synergy of metagenomics and immunoinformatics could transform how we identify vaccine antigens and immune therapies, tapping the vast, largely unexplored reservoir of nature's microbiomes. This promising frontier has the potential to yield novel interventions for infectious disease, cancer, autoimmunity, and beyond, provided its use is guided by careful analysis and empirical testing.

References

1. Panthee, B., Gyawali, S., Panthee, P., & Techato, K. (2022). Environmental and Human Microbiome for Health. *Life* (Basel, Switzerland), 12(3), 456. <https://doi.org/10.3390/life12030456>
2. Shim, J. A., Ryu, J. H., Jo, Y., & Hong, C. (2023). The role of gut microbiota in T cell immunity and immune-mediated disorders. *International journal of biological sciences*, 19(4), 1178–1191. <https://doi.org/10.7150/ijbs.79430>
3. Bach, E. M., Ramirez, K. S., Fraser, T. D., & Wall, D. H. (2020). Soil biodiversity integrates solutions for a sustainable future. *Sustainability*, 12(7), 2662. <https://doi.org/10.3390/su12072662>
4. Roslund, M. I., Puhakka, R., Grönroos, M., Nurminen, N., Oikarinen, S., Gazali, A. M., Cinek, O., Kramná, L., Siter, N., Vari, H. K., Soininen, L., Parajuli, A., Rajaniemi, J., Kinnunen, T., Laitinen, O. H., Hyöty, H., Sinkkonen, A., & ADELE research group (2020). Biodiversity intervention enhances immune regulation and health-associated commensal microbiota among daycare children. *Science advances*, 6(42), eaba2578. <https://doi.org/10.1126/sciadv.aba2578>
5. Sun, X., Liddicoat, C., Tiunov, A., Wang, B., Zhang, Y., Lu, C., ... & Zhu, Y. G. (2023). Harnessing soil biodiversity to promote human health in cities. *npj Urban sustainability*, 3(1), 5. <https://doi.org/10.1038/s42949-023-00086-0>
6. Pérez-Cobas, A. E., Gómez-Valero, L., & Buchrieser, C. (2020). Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial genomics*, 6(8), mgen000409. <https://doi.org/10.1099/mgen.0.000409>
7. Purohit, H. V., & Chakraborty, J. (2025). Metagenomic approaches for studying ubiquitous yet diverse nucleoid-associated proteins in microbial communities: challenges and advances. *World journal of microbiology & biotechnology*, 41(10), 383. <https://doi.org/10.1007/s11274-025-04564-8>
8. Prawiningrum, A. F., Paramita, R. I., & Panigoro, S. S. (2022). Immunoinformatics Approach for Epitope-Based Vaccine Design: Key Steps for Breast Cancer Vaccine. *Diagnostics* (Basel, Switzerland), 12(12), 2981. <https://doi.org/10.3390/diagnostics12122981>

9. Leitão, J. H., & Rodríguez-Ortega, M. J. (2020). Omics and Bioinformatics Approaches to Identify Novel Antigens for Vaccine Investigation and Development. *Vaccines*, 8(4), 653. <https://doi.org/10.3390/vaccines8040653>
10. Purohit HV, Kanojia H, Pandya V, Nalla Y, Raval KY, Kapadiya KM, Kamdar JH (2023) Soil as a host to the biotic community. In: Gupta P, Shahnawaz M (eds) *Soil Microbiome of the cold habitats: trends and applications*. CRC, pp. 17–30. <https://doi.org/10.1201/9781003354031-2>
11. Ma, H., Cornadó, D., & Raaijmakers, J. M. (2025). The soil-plant-human gut microbiome axis into perspective. *Nature Communications*, 16(1), 7748. <https://doi.org/10.1038/s41467-025-62989-z>
12. Blum, W. E. H., Zechmeister-Boltenstern, S., & Keiblinger, K. M. (2019). Does Soil Contribute to the Human Gut Microbiome?. *Microorganisms*, 7(9), 287. <https://doi.org/10.3390/microorganisms7090287>
13. Banerjee, S., & van der Heijden, M. G. A. (2023). Soil microbiomes and one health. *Nature Reviews. Microbiology*, 21(1), 6–20. <https://doi.org/10.1038/s41579-022-00779-w>
14. Zhang, K., Hamidian, A. H., Tubić, A., Zhang, Y., Fang, J.K.H., Wu, C., & Lam, P. K. S. (2021). Understanding plastic degradation and microplastic formation in the environment: A review. *Environmental pollution (Barking, Essex: 1987)*, 274, 116554. <https://doi.org/10.1016/j.envpol.2021.116554>
15. Haahtela T. (2019). A biodiversity hypothesis. *Allergy*, 74(8), 1445–1456. <https://doi.org/10.1111/all.13763>
16. Kummola, L., González-Rodríguez, M. I., Marnila, P., Nurminen, N., Salomaa, T., Hiihtola, L., Mäkelä, I., Laitinen, O. H., Hyöty, H., Sinkkonen, A., & Junttila, I. S. (2023). Comparison of the effect of autoclaved and non-autoclaved live soil exposure on the mouse immune system: Effect of soil exposure on the immune system. *BMC immunology*, 24(1), 29. <https://doi.org/10.1186/s12865-023-00565-0>
17. Dekeukeleire, M., Vandenheuvel, D., Khondee, T., Delanghe, L., Van Rillaer, T., Thys, S., Timmermans, J. P., Lebeer, S., & Spacova, I. (2025). Immunostimulatory activity of inactivated environmental *Bacillus* isolates and their endospores. *Scientific reports*, 15(1), 30604. <https://doi.org/10.1038/s41598-025-12833-7>
18. Wnuk, E., Waško, A., Walkiewicz, A., Bartmiński, P., Bejger, R., Mielnik, L., & Bieganski, A. (2020). The effects of humic substances on DNA isolation from soils. *PeerJ*, 8, e9378. <https://doi.org/10.7717/peerj.9378>
19. Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., Weinstock, G. M., ... Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology*, 35(8), 725–731. <https://doi.org/10.1038/nbt.3893>
20. Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11, 119. <https://doi.org/10.1186/1471-2105-11-119>
21. Iqbal, H. A., Craig, J. W., & Brady, S. F. (2014). Antibacterial enzymes from the functional screening of metagenomic libraries hosted in *Ralstonia metallidurans*. *FEMS microbiology letters*, 354(1), 19–26. <https://doi.org/10.1111/1574-6968.12431>

22. Kanojia, H., Purohit, H., Joshi, M., Kamdar, J. H., & Chakraborty, J. (2024). Microplastics Accumulate Microbial Pathogens in the Terrestrial Environment. In *Microplastic Pollution* (pp. 351-362). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-8357-5_20
23. Zhu, D., Ma, J., Li, G., Rillig, M. C., & Zhu, Y. G. (2022). Soil plastispheres as hotspots of antibiotic resistance genes and potential pathogens. *The ISME journal*, 16(2), 521–532. <https://doi.org/10.1038/s41396-021-01103-9>
24. Zamyatina, A., & Heine, H. (2020). Lipopolysaccharide Recognition in the Crossroads of TLR4 and Caspase-4/11 Mediated Inflammatory Pathways. *Frontiers in immunology*, 11, 585146. <https://doi.org/10.3389/fimmu.2020.585146>
25. Fu, X., Ou, Z., & Sun, Y. (2022). Indoor microbiome and allergic diseases: From theoretical advances to prevention strategies. *Eco-Environment & Health*, 1(3), 133–146. <https://doi.org/10.1016/j.eehl.2022.09.002>
26. Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). 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 research*, 48(W1), W449–W454. <https://doi.org/10.1093/nar/gkaa379>
27. Galanis, K. A., Nastou, K. C., Papandreou, N. C., Petichakis, G. N., Pigis, D. G., & Iconomidou, V. A. (2021). Linear B-Cell Epitope Prediction for In Silico Vaccine Design: A Performance Review of Methods Available via Command-Line Interface. *International journal of molecular sciences*, 22(6), 3210. <https://doi.org/10.3390/ijms22063210>
28. Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC bioinformatics*, 8, 4. <https://doi.org/10.1186/1471-2105-8-4>
29. Bukhari, S. N. H., Jain, A., Haq, E., Mehbodniya, A., & Webber, J. (2022). Machine Learning Techniques for the Prediction of B-Cell and T-Cell Epitopes as Potential Vaccine Targets with a Specific Focus on SARS-CoV-2 Pathogen: A Review. *Pathogens (Basel, Switzerland)*, 11(2), 146. <https://doi.org/10.3390/pathogens11020146>
30. Khalid, K., & Poh, C. L. (2023). The Promising Potential of Reverse Vaccinology-Based Next-Generation Vaccine Development over Conventional Vaccines against Antibiotic-Resistant Bacteria. *Vaccines*, 11(7), 1264. <https://doi.org/10.3390/vaccines11071264>
31. Ong, E., Cooke, M. F., Huffman, A., Xiang, Z., Wong, M. U., Wang, H., Seetharaman, M., Valdez, N., & He, Y. (2021). Vaxign2: the second generation of the first Web-based vaccine design program using reverse vaccinology and machine learning. *Nucleic acids research*, 49(W1), W671–W678. <https://doi.org/10.1093/nar/gkab279>
32. Jaiswal, V., Chanumolu, S. K., Gupta, A., Chauhan, R. S., & Rout, C. (2013). Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC bioinformatics*, 14, 211. <https://doi.org/10.1186/1471-2105-14-211>
33. Dalsass, M., Brozzi, A., Medini, D., & Rappuoli, R. (2019). Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery. *Frontiers in immunology*, 10, 113. <https://doi.org/10.3389/fimmu.2019.00113>
34. Manvani, R., Purohit, H., Sahoo, C. R., Rajput, M., & Shah, S. (2024). Immunoinformatics: an interdisciplinary technique for designing and engineering vaccine antigen. In *Reverse Vaccinology* (pp. 87-99). Academic Press. <https://doi.org/10.1016/B978-0-443-13395-4.00012-5>

35. Sang, Y., Nahashon, S. N., & Webby, R. J. (2025). Microbiome-Immune Interaction and Harnessing for Next-Generation Vaccines Against Highly Pathogenic Avian Influenza in Poultry. *Vaccines*, 13(8), 837. <https://doi.org/10.3390/vaccines13080837>
36. Villanueva-Flores, F., Sanchez-Villamil, J. I., & Garcia-Atutxa, I. (2025). AI-driven epitope prediction: a system review, comparative analysis, and practical guide for vaccine development. *NPJ vaccines*, 10(1), 207. <https://doi.org/10.1038/s41541-025-01258-y>
37. Spiga, O., Visibelli, A., Pettini, F., Roncaglia, B., & Santucci, A. (2025). SHASI-ML: a machine learning-based approach for immunogenicity prediction in Salmonella vaccine development. *Frontiers in cellular and infection microbiology*, 15, 1536156. <https://doi.org/10.3389/fcimb.2025.1536156>
38. Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic acids research*, 47(D1), D339–D343. <https://doi.org/10.1093/nar/gky1006>
39. O'Donnell, T. J., Rubinsteyn, A., & Laserson, U. (2020). MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell systems*, 11(1), 42–48.e7. <https://doi.org/10.1016/j.cels.2020.06.010>
40. Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic acids research*, 45(W1), W24–W29. <https://doi.org/10.1093/nar/gkx346>
41. Zaharieva, N., Dimitrov, I., Flower, D. R., & Doytchinova, I. (2019). VaxiJen Dataset of Bacterial Immunogens: An Update. *Current computer-aided drug design*, 15(5), 398–400. <https://doi.org/10.2174/1573409915666190318121838>
42. Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., Van Wezel, G. P., Medema, M. H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic acids research*, 49(W1), W29–W35. <https://doi.org/10.1093/nar/gkab335>
43. D'Mello, A., Ahearn, C. P., Murphy, T. F., & Tettelin, H. (2019). ReVac: a reverse vaccinology computational pipeline for prioritization of prokaryotic protein vaccine candidates. *BMC genomics*, 20(1), 981. <https://doi.org/10.1186/s12864-019-6195-y>
44. Paul, S., Sidney, J., Sette, A., & Peters, B. (2016). TepiTool: a pipeline for computational prediction of T cell epitope candidates. *Current protocols in immunology*, 114(1), 18–19. <https://doi.org/10.1002/cpim.12>
45. Ong, E., Wang, H., Wong, M. U., Seetharaman, M., Valdez, N., & He, Y. (2020). Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. *Bioinformatics (Oxford, England)*, 36(10), 3185–3191. <https://doi.org/10.1093/bioinformatics/btaa119>
46. Zeng, L., Qian, Y., Cui, X., Zhao, J., Ning, Z., Cha, J., ... & Jian, Z. (2025). Immunomodulatory role of gut microbial metabolites: mechanistic insights and therapeutic frontiers. *Frontiers in Microbiology*, 16, 1675065. <https://doi.org/10.3389/fmicb.2025.1675065>
47. Akira, S., Uematsu, S., & Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell*, 124(4), 783–801. <https://doi.org/10.1016/j.cell.2006.02.015>
48. Yoon, S. I., Kurnasov, O., Natarajan, V., Hong, M., Gudkov, A. V., Osterman, A. L., & Wilson, I. A. (2012). Structural basis of TLR5-flagellin recognition and signaling. *Science*, 335(6070), 859–864. <https://doi.org/10.1126/science.1215584>

49. Sehgal, S. N. (2003, May). Sirolimus: its discovery, biological properties, and mechanism of action. In *Transplantation proceedings* (Vol. 35, No. 3, pp. S7-S14). Elsevier. [https://doi.org/10.1016/S0041-1345\(03\)00211-2](https://doi.org/10.1016/S0041-1345(03)00211-2)
50. Gupta, S., Mittal, P., Madhu, M. K., & Sharma, V. K. (2017). IL17eScan: a tool for the identification of peptides inducing IL-17 response. *Frontiers in immunology*, 8, 1430. <https://doi.org/10.3389/fimmu.2017.01430>
51. Leitão, J. H., & Rodríguez-Ortega, M. J. (2020). Omics and bioinformatics approaches to identify novel antigens for vaccine investigation and development. *Vaccines*, 8(4), 653. <https://doi.org/10.3390/vaccines8040653>
52. Shao, G., Zhu, X., Hua, R., Lu, Z., Chen, Y., Yang, A., & Yang, G. (2025). Cocktail vaccine induces immunoprotection and modulates the fecal microbiota in dogs against *Echinococcus granulosus* infection. *npj Vaccines*, 10(1), 214. <https://doi.org/10.1038/s41541-025-01275-x>
53. Razzak, A., Ahmed, F., & Mahmud, M. T. (2025). Development of a multi-epitope vaccine against *Helicobacter pylori* using a novel saRNA technology through an immunoinformatics approach. *Scientific Reports*, 15(1), 33753. <https://doi.org/10.1038/s41598-025-99512-9>
54. Shen, J., McFarland, A. G., Blaustein, R. A., Rose, L. J., Perry-Dow, K. A., Moghadam, A. A., ... & Hartmann, E. M. (2022). An improved workflow for accurate and robust healthcare environmental surveillance using metagenomics. *Microbiome*, 10(1), 206. <https://doi.org/10.1186/s40168-022-01412-x>
55. Takeuchi, T., Nakanishi, Y., & Ohno, H. (2024). Microbial metabolites and gut immunology. *Annual review of immunology*, 42(1), 153-178. <https://doi.org/10.1146/annurev-immunol-090222-102035>
56. Fernández-Tomé, S., Marin, A. C., Ortega Moreno, L., Baldan-Martin, M., Mora-Gutierrez, I., Lanás-Gimeno, A., ... & Bernardo, D. (2019). Immunomodulatory Effect of Gut Microbiota-Derived Bioactive Peptides on Human Immune System from Healthy Controls and Patients with Inflammatory Bowel Disease. <https://doi.org/10.3390/nul1112605>
57. Rehmani, M. B. I., Arshad, F., Khan, M. U., Ejaz, H., Nishan, U., Alotaibi, A., Ullah, R., Chen, K., Ojha, S. C., & Shah, M. (2025). Computational design of an mRNA vaccine targeting antifungal-resistant *Lomentospora prolificans*. *Scientific reports*, 15(1), 34157. <https://doi.org/10.1038/s41598-025-14907-y>
58. Nguyen, M. N., Krutz, N. L., Limviphuvadh, V., Lopata, A. L., Gerberick, G. F., & Maurer-Stroh, S. (2022). AllerCatPro 2.0: a web server for predicting protein allergenicity potential. *Nucleic acids research*, 50(W1), W36–W43. <https://doi.org/10.1093/nar/gkac446>
59. Liao, W. W., & Arthur, J. W. (2011). Predicting peptide binding to Major Histocompatibility Complex molecules. *Autoimmunity reviews*, 10(8), 469–473. <https://doi.org/10.1016/j.autrev.2011.02.003>
60. Kamdar, J. H., Jeba Praba, J., & Georrgge, J. J. (2020). Artificial intelligence in medical diagnosis: methods, algorithms and applications. In *Machine Learning with Health Care Perspective: Machine Learning and Healthcare* (pp. 27-37). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-40850-3_2