# AI and Its Involvement in the Future of Neoantigen Discovery

**The overabundance of data, constantly developing data acquisition technologies, and the rise of applied AI are significantly accelerating cancer neoantigen immunotherapy research**

*Aleksandar Mihajlovic at APIS Assay Technologies*

Targeting specificity is one of the greatest barriers in the fight for curing cancer. Over the past decade, immunotherapies have experienced a considerable rise in attention due to their successful targeting abilities. One of the more promising immunotherapeutic methods is neoantigen therapy. This cancer cell targeting approach is based on discriminating cellular neoantigens or peptides that incite an immune response from those that do not. Cancer cells produce non-functional proteins, whose proteasome digested peptide by-products can be used by the immune system for pathogenic cell identification (1). The process of identifying peptides that can be used in this manner, i.e., neoantigens in cancer cells, identify the genes that are used to encode them.

## Neoantigens

A patient would start with exome sequencing of both a cancer biopsy and normal tissue to discriminate genes with tumour originating missense mutations and to determine the patient's human leukocyte antigen (HLA) alleles. The list of discriminant gene candidates is then cross-checked with corresponding gene expression data. A subset of highly transcribed genes is isolated and used as a starting point. The abundance of the target protein plays a key role in increasing the efficacy of the immunogenic response (2). The more abundant the transcript, the more abundant the protein, and the higher the sensitivity of the response.

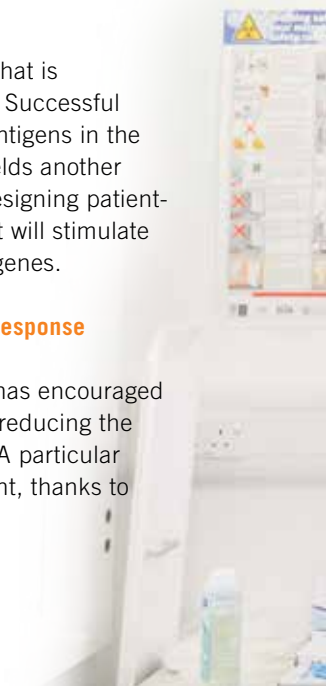
## Neoantigen Discovery

Having identified these gene candidates, the key to neoantigen discovery lies in the ability to determine which of the derived proteins are processed into 8-11 residue peptides that can bind into the crevice of an empty MHC-I complex for inspection by CD8+ T cells, and whether the CD8+ T cells will react with an immune response (3). Machine learning tools, such as NetChop, utilise algorithms for modelling how the proteasome will digest a candidate protein into 8-11 residue peptides (1). The results are millions of peptide candidates that can be evaluated by deep neural network algorithms for their binding affinities to HLA allele specific MHC-I complexes. NetMHC is a tool that extracts a variety of peptide features from the peptide set (4). Some of these features include categorical encodings of each residue, binary representations of certain functional groups, and continuous measurements of biophysical properties to model the binding affinities (5). The presence of the artificial intelligence (AI) determined peptide candidates in the tumour sample are confirmed via mass spectrometry. All of the peptides from the AI-derived list that intersect with the mass spectrometry detected peptides are then finally tested for CD8+ responsiveness *in vitro* and *in vivo* validation (6). The peptides that incite an immunological reaction are considered neoantigens and are ready for clinical trials.

Neoantigen discovery is only the first step that is accelerated by data and machine learning. Successful treatment relies on the abundance of neoantigens in the tumour cells for effective targeting. This yields another challenge that is being addressed by AI: designing patient-tailored transcription modulating drugs that will stimulate the transcription of neoantigen-producing genes.

## Increasing Sensitivity of the Immunogenic Response

The awesome cost and high risk of failure has encouraged a rapid embrace of AI methods, which are reducing the difficulty involved in developing drugs (7). A particular category of drugs that have seen recent light, thanks to



> *This yields another challenge that is being addressed by AI: designing patient-tailored transcription modulating drugs*

AI, are transcription modulating drugs and aptamers. It is foreseeable that in combination with the neoantigen therapies of the future, transcription modulating drugs that hyper-methylate neoantigen coding gene promoters will drastically increase therapeutic efficacy and the curative potential of cancer treatment (8). Determining the promoter sequences of such genes is instrumental in such upregulation. Some promoter sequences are determined easily while others are not. Special machine learning algorithms that carefully analyse myriad genomic features have been developed to detect downstream promoter elements (DPE) motifs, which are cryptic promoter sequences that are not easily identifiable like the TATA box, and that account for an equivalent number of gene promoters (9). By being able to identify these promoter sequences, one of the key barriers to precision transcription modulation in pharmacogenomics has been alleviated. Targeting the promoters of neoantigen encoding genes promises to greatly increase the sensitivity of the immunogenic response.

## Maximisation of the Immunogenic Response

In order for neoantigen therapy to take effect, not only is it necessary for the abundance of neoantigens to be sufficient, it is also necessary for the CD8+ T cell receptors to be activated. A variety of aptamers described to date have been shown to be reliable in modulating immune responses against cancer by either blocking or activating immune receptors (10). Such activity is intended to further improve the efficacy of the neoantigen immunogenic response by increasing receptor sensitivity to neoantigens. Aptamers are small single strand RNA or DNA oligonucleotides with 3D folding structures, which allow them to bind to their targets with high specificity. Their interactions with proteins result in geometrical changes of the substrate that can reverse the effects of the underlying non-functional protein particles and to stimulate metabolism (11).

The process by which they are manufactured is commonly referred to as systematic evolution of ligands by exponential enrichment (SELEX), which consists of several repeated rounds of binding, partitioning, and amplification (12). In the SELEX procedure, aptamers are identified by their abilities to bind to a protein of interest from libraries containing up to $10^{16}$ different RNA or DNA sequences (13). The design process of aptamers can be costly and time-consuming if performed manually. Therefore, tools that employ random forest

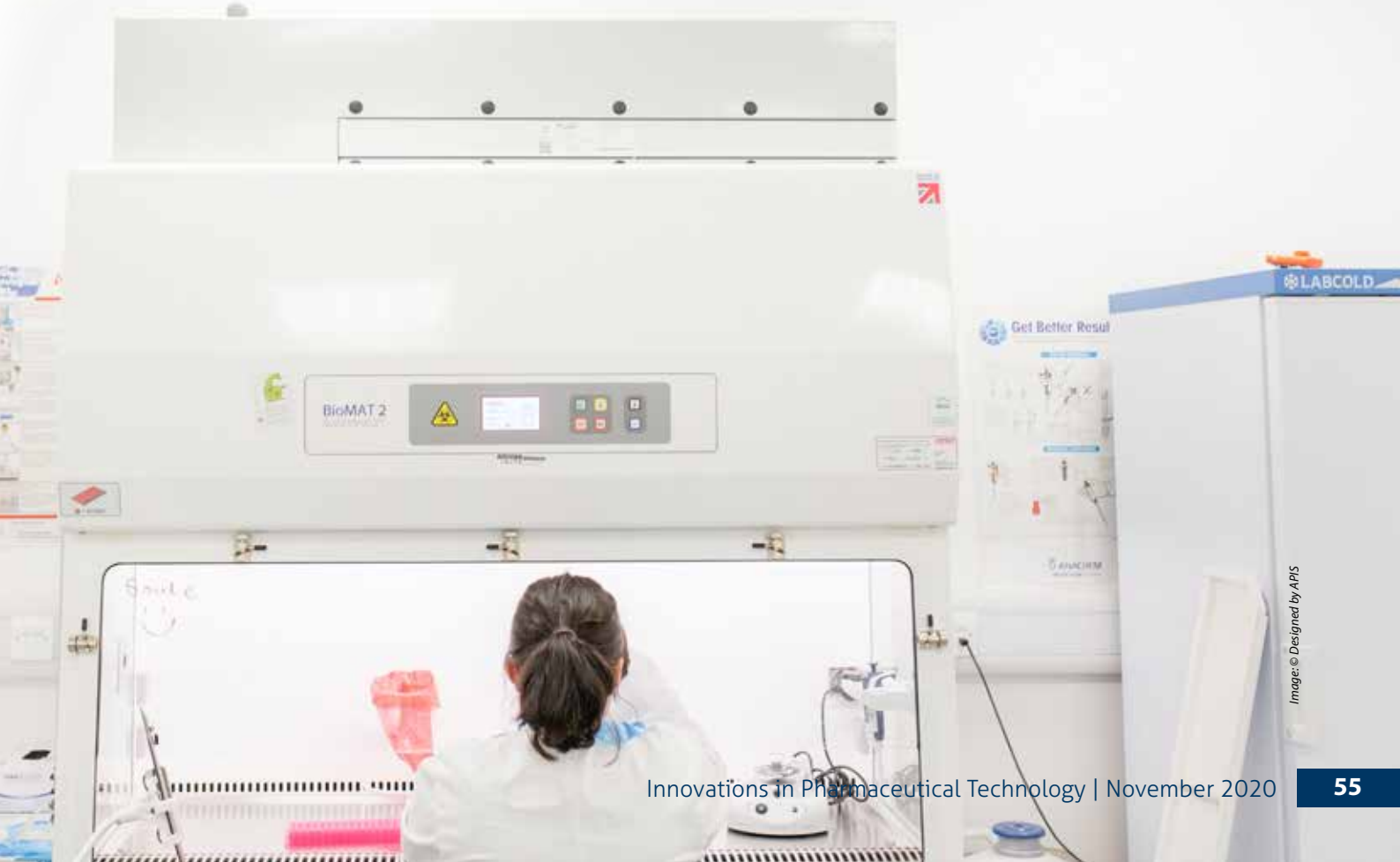



*Image: © Designed by APIS*

> *The development of neoantigen immunotherapies has been heavily stimulated by AI. The benefits of its applications can be seen particularly in the targeting stage of drug development*

machine learning algorithms are used to rapidly and effectively predict the aptamer-protein interacting pairs. The predictions are made using substrate information, such as patient-specific gene allele sequences, and/or messenger RNA transcript sequences. The prediction process involves an ensemble of algorithms, such as pseudo K-tuple nucleotide composition, discrete cosine transformation, disorder information, and bi-gram position specific scoring matrix (14). To facilitate even smoother interaction detection, databases cataloguing already documented aptamer-protein interactions are being used. Aptamer Base is a database that contains approximately 2,000 entries of interactions that provided detailed and structured information about the experimental conditions under which aptamers were selected and had their binding affinities quantified (15). Interestingly, the interaction data has been obtained by employing natural language processing algorithms – another category of machine learning algorithms – to identify approximately 3,000 PubMed articles from primary scientific literature that support and describe interactions (15). The database was initially developed using Freebase, which was subsequently acquired by Google, and, unfortunately, shut down in 2016. However, a resource description framework dump of the database is available online.

The development of neoantigen immunotherapies has been heavily stimulated by AI. The benefits of its applications can be seen particularly in the targeting stage of drug development. With the help of AI, immunotherapies, such as neoantigen therapies, will yield CD8+ T cells more sensitive to cancer cell detection. Rapid advancements in data acquisition and data processing technologies are creating a breeding ground for clinically relevant data. Their availability, specifically on the genomic level, are fuelling the personalised drug design industry not only for immunotherapies, such as neoantigens, but as a whole. Supporting this argument is the fact that in 2018 there were 76 companies actively using AI for drug discovery, and today, in 2020, there are over 230 (16-17). We have yet to see what more AI will contribute to the world of neoantigen immunotherapeutics.

*References*


1. *Nielsen M* et al, *The role of the proteasome in generating cytotoxic T-cell epitopes: Insights obtained from improved predictions of proteasomal cleavage,* Immunogenetics, *57(1-2): pp33-41, 2005*
2. *Visit: www.nature.com/articles/nbt.3800*
3. *Visit: www.onlinelibrary.wiley.com/doi/pdf/10.1002/1521-4141(200011)30:11%3C3089::AID-IMMU3089%3E3.0.CO;2-5*
4. *Andreatta M and Nielsen M, Gapped sequence alignment using artificial neural networks: application to the MHC class I system,* Bioinformatics *32(4): pp511-7, 2016*
5. *Visit: link.springer.com/article/10.1186/s12859-018-2561-z*
6. *Visit: www.nature.com/articles/nbt.3932*
7. *Visit: www.healthaffairs.org/doi/full/10.1377/hlthaff.25.2.420*
8. *Yi M* et al, *Immune pressures drive the promoter hypermethylation of neoantigen genes,* Experimental Hematology and Oncology, *8(32): 2019*
9. *Kutach AK and Kadonaga JT, The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters,* Mol Cell Biol *20(13): 2000*
10. *Visit: www.ncbi.nlm.nih.gov/pmc/articles/pmc4931050*
11. *Visit: www.nature.com/articles/s41598-018-33887-w*
12. *Visit: www.ncbi.nlm.nih.gov/pmc/articles/pmc5666824*
13. *Visit: www.ncbi.nlm.nih.gov/pmc/articles/pmc4666376*
14. *Visit: link.springer.com/article/10.1186/s12859-016-1087-5*
15. *Visit: www.ncbi.nlm.nih.gov/pmc/articles/pmc3308162*
16. *Visit: www.nbcnews.com/mach/science/why-big-pharma-betting-big-ai-ncna852246*
17. *Visit: blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery*


**Aleksandar Mihajlovic** has been working in the genomics field since 2011. He has designed and evaluated GWASs (genome-wide association study) of atherosclerosis patients, evaluated genotype-phenotype correlation between atherosclerosis plaque build-up and certain SNP alleles, and algorithmically imputed missing SNP calls at the Mathematical Institute of the Serbian Academy of Arts and Science. Aleksander has contributed to the development and optimisation of the cancer genomics cloud platform (presented results at ASCO conference in 2017), and on the development of preliminary version of CloudNeo, a neoantigen pipeline at Seven Bridges Genomics. Founder and CEO of Beogenomics, a small bioinformatics start-up, he is now working as Head of Operations at **APIS Assay Tech** in Belgrade, Serbia, overseeing bioinformatics business growth and development.