LaSalle Universidad
DIRECCIÓN DE INVESTIGACIÓN

Universidad Católica San Pablo

Principales estrategias y métodos basados en deep learning para la detección de neo antígenos en el marco del desarrollo de vacunas personalizadas en la inmunoterapia del cáncer

Proyecto interno en colaboración con la UCSP

PhD(c). Vicente Machaca Arceda

Contenido

Introducción
- Proyecto
- Mutaciones y neo antígenos
- Problema y Objetivos

Revisión Sistemática
- Metodología
- Resultados

Discusión

Trabajos futuros

Contenido

Introducción

- Proyecto
- Mutaciones y neo antígenos
- Problema y Objetivos

Revisión Sistemática

- Metodología
- Resultados

Discusión

Trabajos futuros

Proyecto
Financiamiento

Proyecto

Duración y equipo de trabajo

- Seis (06) meses.
- Equipo:
  - Vicente Machaca Arceda (ULaSalle).
  - Valeria Goyzueta (ULaSalle).
  - Yván Tupac (UCSP).
  - Maria Cruz (UCSP).

Variantes y Mutaciones

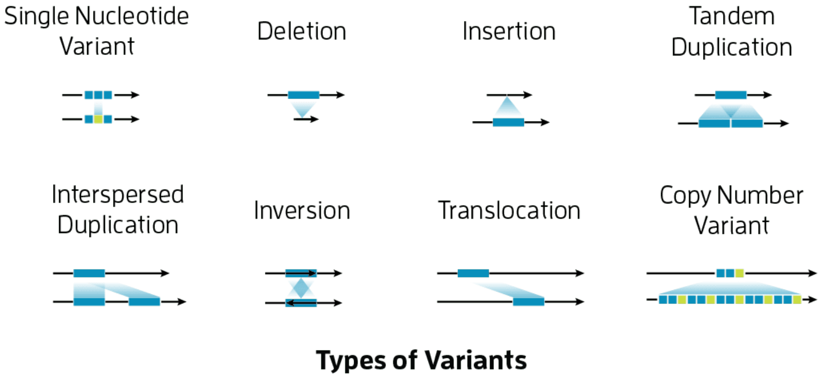


Figure: Example of structural variants. Source: [1]

Inmunoterapia del Cáncer

Es un tipo de tratamiento contra el Cáncer que estimula las defensas naturales del cuerpo para combatir el Cáncer [2].

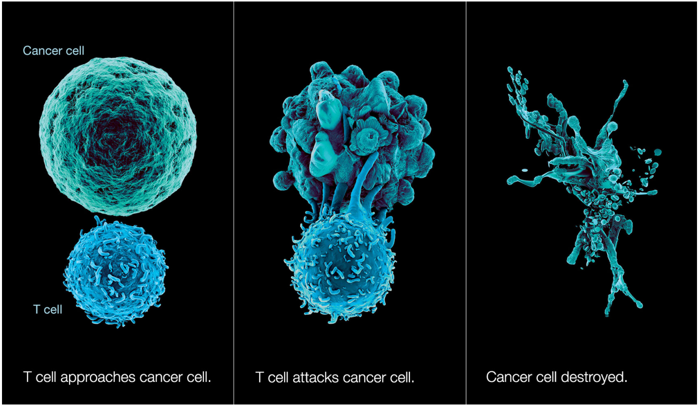


Figure: Ejemplo de como una célula T destruye células del cancer [3].

Inmunoterapia del Cáncer

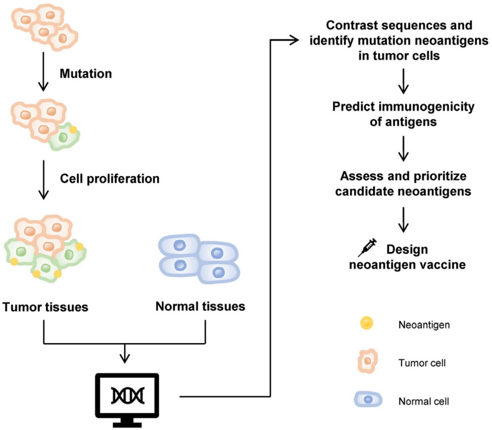
Neo antígenos

Es una **proteína** que se forma en las células de Cáncer cuando ocurre mutaciones en el DNA, cumplen un rol importante al **estimular una respuesta inmune** [4, 5].

En la actualidad hay varios métodos para detectar a predecir neo antígenos, pero **solo una pequeña cantidad de ellos** logran estimular al sistema inmune [6, 7].

Inmunoterapia del Cáncer

Generación de vacunas



Mutation

Cell proliferation

Tumor tissues

Normal tissues

Contrast sequences and identify mutation neoantigens in tumor cells

Predict immunogenicity of antigens

Assess and prioritize candidate neoantigens

Design neoantigen vaccine

● Neoantigen

Tumor cell

Normal cell

Figure: Proceso para la generación de vacunas personalizadas [8].

MHC-I

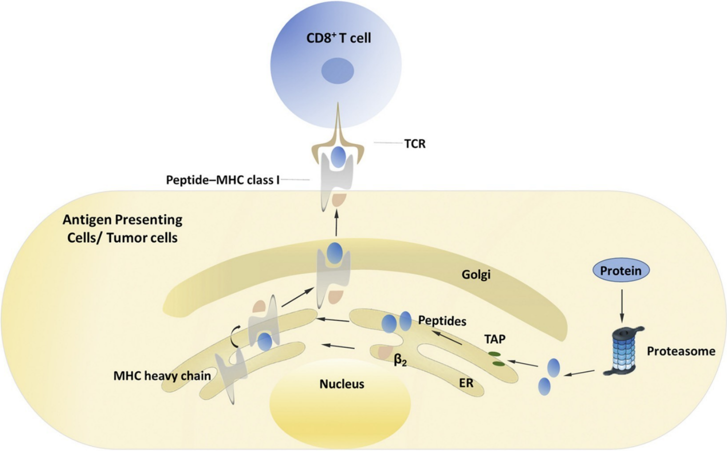


Figure: Presentación de antígenos por MHC-I. Fuente: [9]

Motivación

El cáncer representa el mayor problema de salud mundial, pero lamentablemente los métodos basados en cirugías, radioterapias, quimioterapias tienen baja efectividad [8].

La inmunoterapia del cáncer es una alternativa para el desarrollo de vacunas personalizadas, pero este proceso depende de una correcta detección de neo antígenos [10, 8].

Problema

Menos del 3% de los neoantígenos detectados logran activar a las células T (sistema inmune) [10].

Objetivo general

Desarrollar una revisión sistemática de métodos que utilizan *deep learning* para la detección de neo antígenos.

Contenido

Introducción

- Proyecto
- Mutaciones y neo antígenos
- Problema y Objetivos

Revisión Sistemática

- Metodología
- Resultados

Discusión

Trabajos futuros

Revisión Sistemática

Preguntas de investigación

Table: Research questions used in SLR.

Preguntas de investigación
Q1. Como se utilizan las técnicas de *deep learning* para la detección de neo antígenos?
Q2. Que tipos de datos y pre procesamiento es utilizado en la detección de ne antígenos?
Q3. Que bases de datos son utilizadas en la detección de neo antígenos?
Q4. Que método basado en *deep learning* tiene los mejores resultados?

Table: Cadenas de busqueda utilizadas en la RSL.

Cadena de busqueda
neoantigen AND (detection OR pipeline) AND deep learning
(MHC OR HLA) AND binding AND deep learning
(MHC-I OR MHC-II OR MHC OR HLA) AND (peptide OR epitope) AND (binding OR affinity OR prediction OR detection OR presentation)
TCR interaction prediction

Table: Bases de datos utilizadas en la RSL.

Bases de datos
IEEE Xplore
Science Direct
Springer
ACM Digital Library
PubMed
BioRxiv

Table: Criterios de inclusión y exclusión de artículos utilizados en la RSL.

Criterios de inclusión	Criterios de exclusión
Artículos con categoría ERA (A, B o C) si son conferencias y Journals Q1, Q2 o Q3.	Trabajos de baja calidad, que no esten rankeados.
Sobre *deep learning*	
La metodología es detallada.	
Tiene repositorio de código fuente y base de datos (deseable).	

Revisión Sistemática

Metodología

Table: Cantidad de artículos encontrados y seleccionados según los criterios de inclusión y exclusión en la RSL.

Año	Artículos encontrados	Artículos seleccionados
2018	57	21
2019	72	31
2020	86	29
2021	61	34
2022	58	19
Total	**334**	**134**

Revisión Sistemática

Convolutional Neural Networks

Table: List of research since 2018 that uses CNNs for peptide-MHC binding and presentation.

Year	Ref.	Approach	Name	MHC	Encoding
2022	[11]	pMHC(b)	DeepMHCII	II	PFR
2021	[12]	pMHC(b)	DeepImmuno	I	AAindex1
2021	[13]	pMHC(p)	APPM	I	One-hot
2021	[14]	pMHC(p)	MHCfovea	I	One-hot
2021	[15]	pMHC(b)	CNN-PepPred	II	BLOSUM
2020	[16]	pMHC(b)	IConMHC	I	PCA and AAindex3
2020	[17]	pMHC(b)	OnionMHC	I	BLOSUM and structural features
2020	[18]	pMHC(p)	MINERVA	I	Physicochemical properties
2019	[19]	pMHC(b)	CNN-NF	I	Sequence, Hydropathy, Polarity, Length
2019	[20]	pMHC(b)	DeepSeqPan	I	One-hot
2018	[21]	pMHC(b)	ConvMHC	I	Contact side HLA.peptide

Table: List of research since 2018 that uses CNNs s with RNN or attention mechanisms for peptide-MHC binding and presentation. MHCherryPan uses CNN with RNN, the other uses CNN witn Attention mechanims.

Year	Ref.	Approach	Name	MHC	Encoding
2021	[22]	pMHC(b)	DeepNetBim	I	BLOSUM
2021	[23]	pMHC(b)	Deep Attention Pan	I	BLOSUM
2019	[24]	pMHC(b)	ACME	I	BLOSUM
2020	[25]	pMHC(b)	MHCherryPan	I	BLOSUM

Table: List of research since 2018 that uses RNNs for peptide-MHC binding and presentation. MATHLA, DeepSeqPanII and DeepHLApan uses RNN with attention mechanims, meanwhile the other focus on GRU and LSTM.

Year	Ref.	Approach	Name	MHC	Encoding
2021	[26]	pMHC(b)	MATHLA	I	BLOSUM
2021	[27]	pMHC(b)	DeepSeqPanII	II	One-hot and BLOSUM
2021	[28]	pMHC(b)	GRU-based RNN	II	Embeding layer
2021	[29]	pMHC(b)	BVLSTM-MHC	I	One-hot and BLOSUM
2020	[30]	pMHC(b)	MHCnuggets	I, II	One-hot
2019	[31]	pMHC(b)	DeepHLApan	I	One-hot

Revisión Sistemática

Transformers

Table: List of research since 2018 that uses Transformers (self-attention) for peptide-MHC binding and presentation.

Year	Ref.	Approach	Name	MHC	Encoding
2022	[32]	pMHC(b)	MHCRoBERTa	I	Tokenized from a pre-trained model
2022	[33]	pMHC(b)	TransPHLA	I	Character embedding model
2021	[34]	pMHC(b)	BERTMHC	II	Embeding layer
2021	[35]	pMHC(p)	ImmunoBERT	I	Embeding layer

Revisión Sistemática

Bases de datos

Table: Public databases of *pMHC binding*, *pMHC presentation*, pMHC-TCR interaction, and 3D structures of proteins.

Name	Year ref.	Description
VDJdb	2018 [36]	TCR binding to pMHC, contains 5491 samples.
IEDB	2018 [37]	The bigger database, contains information *T-cell epitopes*
TSNAdb	2018 [38]	It contains 7748 samples of mutations and HLA of 16 types of cancer.
NeoPeptide	2019 [39]	It contains samples of neoantigens resulting from somatic mutations and related items. 1818137 epitopes of more than 36000 neoantigens.
pHLA3D	2019 [40]	Presents 106 3D structures of the alpha, *beta*2*M* chains, and peptides of HLA-I molecules
dbPepNeo	2020 [41]	It has validated samples of the *peptide-MHC* bond, from MS. It contains 407794 low-quality samples, 247 medium-quality, and 295 high-quality samples.
dbPepNeo2.0	2022 [42]	It gathers a list of neoantigens and HLA molecules. It presents 801 high-quality and 842,289 poor-quality HLAs. Also, 55 class II neoantigens and 630 TCR-bound neo antigens.
IntroSpect	2022 [43]	Tool for building databases on *peptide-MHC binding*. It uses data from *Mass Spectrometry*.
IPD-IMGT/HLA	2022 [44]	With 25000 MHC molecules and 45 alleles.

Revisión Sistemática

Pipelines

Table: List of *pipelines* since 2018 for the detection of neoantigens.

Name	**Year ref.**	**Input**	**Output**
Neopepsee	2018 [45]	RNA-seq, somatic mutations (VCF), HLA type (optional)	Neoantigens and gene expression levels
PGV Pipeline	2018 [46]	DNA-seq	Neoantigens
ScanNeo	2019 [47]	RNA-seq	Neoantigens
NeoPredPipe	2019 [48]	Mutations (VCF) y HLA type	Neoantigens and variant annotation
pVACtools	2020 [49]	Mutations (VCF)	Neoantigens
ProGeo-neo	2020 [50]	RNA-seq y somatic mutations (VCF)	Neoantigens
Neoepiscope	2020 [51]	Somatic mutations (VCF) and BAM files	Neoantigens and mutations
NeoANT-HILL	2020 [52]	RNA-seq y somatic mutations (VCF)	Neoantigens and gene expression levels
NAP-CNB	2021 [53]	RNA-seq	Neoantigens
PEPPRMINT	2021 [54]	DNA-seq	Neoantigens
Valid-NEO	2022 [55]	Somatic mutations (VCF), HLA type (optional)	Neoantigens

Contenido

Introducción

- Proyecto
- Mutaciones y neo antígenos
- Problema y Objetivos

Revisión Sistemática

- Metodología
- Resultados

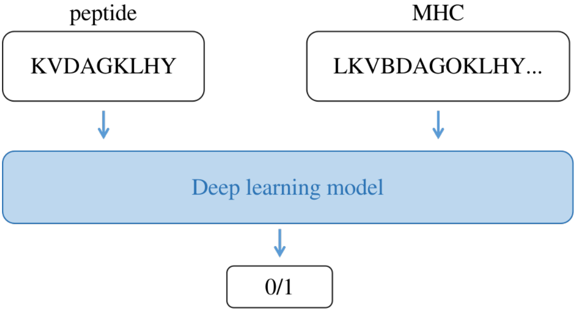
Discusión

Trabajos futuros

Discusión

Como se utilizan las técnicas de *deep learning* para la detección de neo antígenos?

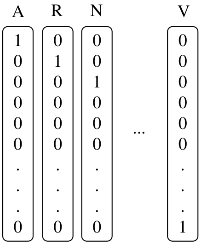
La detección de neo antígenos es visto como un problema de clasificación binaría.



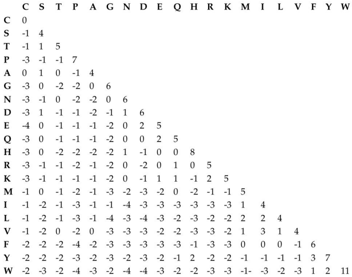
Discusión

Que tipos de datos y pre procesamiento es utilizado en la detección de ne antígenos?

Se consideran las cadenas de aminoacidos, como pre procesamiento se utiliza *one-hot encoding* y *BLOSUM*.



$$\begin{array}{ccccc} A & R & N & & V \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \end{array}$$



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	0																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	-1	0	-2	-2	0	6													
D	-3	1	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	0	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	-1	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	0	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-1	-4	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1-	-3	-2	-3	1	2	11

Discusión

Que bases de datos son utilizadas en la detección de neo antígenos?

Tabla descrita anteriormente.

Discusión

Bases de datos

Name	Year ref.	Description
VDJdb	2018 [36]	TCR binding to pMHC, contains 5491 samples.
IEDB	2018 [37]	The bigger database, contains information *T-cell epitopes*
TSNAdb	2018 [38]	It contains 7748 samples of mutations and HLA of 16 types of cancer.
NeoPeptide	2019 [39]	It contains samples of neoantigens resulting from somatic mutations and related items. 1818137 epitopes of more than 36000 neoantigens.
pHLA3D	2019 [40]	Presents 106 3D structures of the alpha, *beta*2*M* chains, and peptides of HLA-I molecules
dbPepNeo	2020 [41]	It has validated samples of the *peptide-MHC* bond, from MS. It contains 407794 low-quality samples, 247 medium-quality, and 295 high-quality samples.
dbPepNeo2.0	2022 [42]	It gathers a list of neoantigens and HLA molecules. It presents 801 high-quality and 842,289 poor-quality HLAs. Also, 55 class II neoantigens and 630 TCR-bound neo antigens.
IntroSpect	2022 [43]	Tool for building databases on *peptide-MHC binding*. It uses data from *Mass Spectrometry*.
IPD-IMGT/HLA	2022 [44]	With 25000 MHC molecules and 45 alleles.

Discusión

Que método basado en *deep learning* tiene los mejores resultados?

NetMHCpan4.1 es considerado el método con mejor desempeño. Pero, recientemente, **HLAB** [56] utilizado para *pMHC-I binding prediction*, utiliza BiLSTM, proBERT y *transfer learning* de UniRef100 [57] y BFD [58]; este modelo ha superado a NetMHCpan4.1.

Contenido

Introducción

- Proyecto
- Mutaciones y neo antígenos
- Problema y Objetivos

Revisión Sistemática

- Metodología
- Resultados

Discusión

Trabajos futuros

Trabajos futuros

Recientemente un trabajo [59] tambien propone el uso de *transfer learning* pero de un modelo pre-entrenado con 250 millones de proteínas. Entonces, se plantea utilizar la propuesta de HLAB, aumentar la cantidad de muestras y evaluar los resultados.

Actualmente se cuenta con una base de datos de proteínas MHC [40], entonces utilizando AlphaFold de Google, se plantea predecir la estructura de varios péptidos y analizar el enlace péptido-MHC desde un punto de vista de la computación gráfica.

References I

[1] PacBio, “Two review articles assess structural variation in human genomes,” https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes/, 2021, accessed: 2021-05-07. [Online]. Available: https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes/

[2] Cancer.net. (2022) Qué es la inmunoterapia. [Online]. Available: https://www.cancer.net/es/desplazarse-por-atencion-del-cáncer/como-se-trata-el-cáncer/inmunoterapia/qué-es-la-inmunoterapia

[3] NortShore, “Immunotherapy,” 2022. [Online]. Available: https://www.northshore.org/kellogg-cancer-center/our-services/immunotherapy/

References II

[4] NCI, “National cancer institute dictionary,” 2022. [Online]. Available: https://www.cancer.gov/publications/dictionaries/genetics-dictionary

[5] E. S. Borden, K. H. Buetow, M. A. Wilson, and K. T. Hastings, “Cancer neoantigens: Challenges and future directions for prediction, prioritization, and validation,” *Frontiers in Oncology*, vol. 12, 2022.

[6] I. Chen, M. Chen, P. Goedegebuure, and W. Gillanders, “Challenges targeting cancer neoantigens in 2021: a systematic literature review,” *Expert Review of Vaccines*, vol. 20, no. 7, pp. 827–837, 2021.

[7] Q. Hao, P. Wei, Y. Shu, Y.-G. Zhang, H. Xu, and J.-N. Zhao, “Improvement of neoantigen identification through convolution neural network,” *Frontiers in immunology*, vol. 12, 2021.

References III

[8] M. Peng, Y. Mo, Y. Wang, P. Wu, Y. Zhang, F. Xiong, C. Guo, X. Wu, Y. Li, X. Li *et al.*, “Neoantigen vaccine: an emerging tumor immunotherapy,” *Molecular cancer*, vol. 18, no. 1, pp. 1–14, 2019.

[9] X. Zhang, Y. Qi, Q. Zhang, and W. Liu, “Application of mass spectrometry-based mhc immunopeptidome profiling in neoantigen identification for tumor immunotherapy,” *Biomedicine & Pharmacotherapy*, vol. 120, p. 109542, 2019.

[10] L. Mattos, M. Vazquez, F. Finotello, R. Lepore, E. Porta, J. Hundal, P. Amengual-Rigo, C. Ng, A. Valencia, J. Carrillo *et al.*, “Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the esmo precision medicine working group,” *Annals of oncology*, vol. 31, no. 8, pp. 978–990, 2020.

References IV

[11] R. You, W. Qu, H. Mamitsuka, and S. Zhu, “Deepmhcii: a novel binding core-aware deep interaction model for accurate mhc-ii peptide binding affinity prediction,” *Bioinformatics*, vol. 38, no. Supplement_1, pp. i220–i228, 2022.

[12] G. Li, B. Iyer, V. S. Prasath, Y. Ni, and N. Salomonis, “Deepimmuno: deep learning-empowered prediction and generation of immunogenic peptides for t-cell immunity,” *Briefings in bioinformatics*, vol. 22, no. 6, p. bbab160, 2021.

[13] F. Lang, P. Riesgo-Ferreiro, M. L¨ower, U. Sahin, and B. Schr¨ors, “Neofox: annotating neoantigen candidates with neoantigen features,” *Bioinformatics*, vol. 37, no. 22, pp. 4246–4247, 2021.

References V

[14] K.-H. Lee, Y.-C. Chang, T.-F. Chen, H.-F. Juan, H.-K. Tsai, and C.-Y. Chen, “Connecting mhc-i-binding motifs with hla alleles via deep learning,” *Communications Biology*, vol. 4, no. 1, pp. 1–12, 2021.

[15] V. Junet and X. Daura, “Cnn-peppred: an open-source tool to create convolutional nn models for the discovery of patterns in peptide sets—application to peptide–mhc class ii binding prediction,” *Bioinformatics*, vol. 37, no. 23, pp. 4567–4568, 2021.

[16] B. Pei and Y.-H. Hsu, “Iconmhc: a deep learning convolutional neural network model to predict peptide and mhc-i binding affinity,” *Immunogenetics*, vol. 72, no. 5, pp. 295–304, 2020.

References VI

[17] S. Saxena, S. Animesh, M. J. Fullwood, and Y. Mu, “Onionmhc: A deep learning model for peptide—hla-a* 02: 01 binding predictions using both structure and sequence feature sets,” *Journal of Micromechanics and Molecular Physics*, vol. 5, no. 03, p. 2050009, 2020.

[18] F. S. Ng, M. Vandenberghe, G. Portella, C. Cayatte, X. Qu, S. Hanabuchi, A. Landry, R. Chaerkady, W. Yu, R. Collepardo-Guevara *et al.*, “Minerva: Learning the rules of hla class i peptide presentation in tumors with convolutional neural networks and transfer learning,” *Available at SSRN 3704016*, 2020.

[19] T. Zhao, L. Cheng, T. Zang, and Y. Hu, “Peptide-major histocompatibility complex class i binding prediction based on deep learning with novel feature,” *Frontiers in Genetics*, vol. 10, p. 1191, 2019.

References VII

[20] Z. Liu, Y. Cui, Z. Xiong, A. Nasiri, A. Zhang, and J. Hu, “Deepseqpan, a novel deep convolutional neural network model for pan-specific class i hla-peptide binding affinity prediction,” *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[21] Y. Han, “Deep convolutional neural networks for peptide-mhc binding predictions,” 2018.

[22] X. Yang, L. Zhao, F. Wei, and J. Li, “Deepnetbim: deep learning model for predicting hla-epitope interactions based on network analysis by harnessing binding and immunogenicity information,” *BMC bioinformatics*, vol. 22, no. 1, pp. 1–16, 2021.

[23] J. Jin, Z. Liu, A. Nasiri, Y. Cui, S.-Y. Louis, A. Zhang, Y. Zhao, and J. Hu, “Deep learning pan-specific model for interpretable mhc-i peptide binding prediction with improved attention mechanism,” *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 7, pp. 866–883, 2021.

References VIII

[24] Y. Hu, Z. Wang, H. Hu, F. Wan, L. Chen, Y. Xiong, X. Wang, D. Zhao, W. Huang, and J. Zeng, "Acme: pan-specific peptide–mhc class i binding prediction through attention-based deep neural networks," *Bioinformatics*, vol. 35, no. 23, pp. 4946–4954, 2019.

[25] X. Xie, Y. Han, and K. Zhang, "Mhcherrypan: a novel pan-specific model for binding affinity prediction of class i hla-peptide," *International Journal of Data Mining and Bioinformatics*, vol. 24, no. 3, pp. 201–219, 2020.

[26] Y. Ye, J. Wang, Y. Xu, Y. Wang, Y. Pan, Q. Song, X. Liu, and J. Wan, "Mathla: a robust framework for hla-peptide binding prediction integrating bidirectional lstm and multiple head attention mechanism," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–12, 2021.

References IX

[27] Z. Liu, J. Jin, Y. Cui, Z. Xiong, A. Nasiri, Y. Zhao, and J. Hu, "Deepseqpanii: an interpretable recurrent neural network model with attention mechanism for peptide-hla class ii binding prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[28] Y. Heng, Z. Kuang, W. Xie, H. Lan, S. Huang, L. Chen, T. Shi, L. Xu, X. Pan, and H. Mei, "A simple pan-specific rnn model for predicting hla-ii binding peptides," *Molecular Immunology*, vol. 139, pp. 177–183, 2021.

[29] L. Jiang, H. Yu, J. Li, J. Tang, Y. Guo, and F. Guo, "Predicting mhc class i binder: existing approaches and a novel recurrent neural network solution," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab216, 2021.

References X

[30] X. M. Shao, R. Bhattacharya, J. Huang, I. Sivakumar, C. Tokheim, L. Zheng, D. Hirsch, B. Kaminow, A. Omdahl, M. Bonsack *et al.*, "High-throughput prediction of mhc class i and ii neoantigens with mhcnuggetshigh-throughput prediction of neoantigens with mhcnuggets," *Cancer immunology research*, vol. 8, no. 3, pp. 396–408, 2020.

[31] J. Wu, W. Wang, J. Zhang, B. Zhou, W. Zhao, Z. Su, X. Gu, J. Wu, Z. Zhou, and S. Chen, "Deephlapan: a deep learning approach for neoantigen prediction considering both hla-peptide binding and immunogenicity," *Frontiers in Immunology*, p. 2559, 2019.

[32] F. Wang, H. Wang, L. Wang, H. Lu, S. Qiu, T. Zang, X. Zhang, and Y. Hu, "Mhcroberta: pan-specific peptide–mhc class i binding prediction through transfer learning with label-agnostic protein sequences," *Briefings in Bioinformatics*, vol. 23, no. 3, p. bbab595, 2022.

References XI

[33] Y. Chu, Y. Zhang, Q. Wang, L. Zhang, X. Wang, Y. Wang, D. R. Salahub, Q. Xu, J. Wang, X. Jiang *et al.*, “A transformer-based model to predict peptide–hla class i binding and optimize mutated peptides for vaccine design,” *Nature Machine Intelligence*, vol. 4, no. 3, pp. 300–311, 2022.

[34] J. Cheng, K. Bendjama, K. Rittner, and B. Malone, “Bertmhc: improved mhc–peptide class ii interaction prediction with transformer and multiple instance learning,” *Bioinformatics*, vol. 37, no. 22, pp. 4172–4179, 2021.

[35] H.-C. Gasser, G. Bedran, B. Ren, D. Goodlett, J. Alfaro, and A. Rajan, “Interpreting bert architecture predictions for peptide presentation by mhc class i proteins,” *arXiv preprint arXiv:2111.07137*, 2021.

References XII

[36] M. Shugay, D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov *et al.*, “Vdjdb: a curated database of t-cell receptor sequences with known antigen specificity,” *Nucleic acids research*, vol. 46, no. D1, pp. D419–D427, 2018.

[37] R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters, “The immune epitope database (iedb): 2018 update,” *Nucleic acids research*, vol. 47, no. D1, pp. D339–D343, 2018.

[38] J. Wu, W. Zhao, B. Zhou, Z. Su, X. Gu, Z. Zhou, and S. Chen, “Tsnadb: a database for tumor-specific neoantigens from immunogenomics data analysis,” *Genomics, proteomics & bioinformatics*, vol. 16, no. 4, pp. 276–282, 2018.

References XIII

[39] W.-J. Zhou, Z. Qu, C.-Y. Song, Y. Sun, A.-L. Lai, M.-Y. Luo, Y.-Z. Ying, H. Meng, Z. Liang, Y.-J. He *et al.*, “Neopeptide: an immunoinformatic database of t-cell-defined neoantigens,” *Database*, vol. 2019, 2019.

[40] D. M. T. Oliveira, R. M. S. de Serpa Brandão, L. C. D. da Mata Sousa, F. d. C. A. Lima, S. J. H. do Monte, M. S. C. Marroquim, A. V. de Sousa Lima, A. G. B. Coelho, J. M. S. Costa, R. M. Ramos *et al.*, “phla3d: An online database of predicted three-dimensional structures of hla molecules,” *Human Immunology*, vol. 80, no. 10, pp. 834–841, 2019.

[41] X. Tan, D. Li, P. Huang, X. Jian, H. Wan, G. Wang, Y. Li, J. Ouyang, Y. Lin, and L. Xie, “dbpepneo: a manually curated database for human tumor neoantigen peptides,” *Database*, vol. 2020, 2020.

References XIV

[42] M. Lu, L. Xu, X. Jian, X. Tan, J. Zhao, Z. Liu, Y. Zhang, C. Liu, L. Chen, Y. Lin *et al.*, “dbpepneo2. 0: A database for human tumor neoantigen peptides from mass spectrometry and tcr recognition,” *Frontiers in immunology*, p. 1583, 2022.

[43] L. Zhang, G. Liu, G. Hou, H. Xiang, X. Zhang, Y. Huang, X. Zhang, B. Li, and L. J. Lee, “Introspect: Motif-guided immunopeptidome database building tool to improve the sensitivity of hla i binding peptide identification by mass spectrometry,” *Biomolecules*, vol. 12, no. 4, p. 579, 2022.

[44] J. Robinson, D. J. Barker, X. Georgiou, M. A. Cooper, P. Flicek, and S. G. Marsh, “Ipd-imgt/hla database,” *Nucleic acids research*, vol. 48, no. D1, pp. D948–D955, 2020.

References XV

[45] S. Kim, H. S. Kim, E. Kim, M. Lee, E.-C. Shin, and S. Paik, “Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information,” *Annals of Oncology*, vol. 29, no. 4, pp. 1030–1036, 2018.

[46] A. Rubinsteyn, J. Kodysh, I. Hodes, S. Mondet, B. A. Aksoy, J. P. Finnigan, N. Bhardwaj, and J. Hammerbacher, “Computational pipeline for the pgv-001 neoantigen vaccine trial,” *Frontiers in immunology*, vol. 8, p. 1807, 2018.

[47] T.-Y. Wang, L. Wang, S. K. Alam, L. H. Hoeppner, and R. Yang, “Scanneo: identifying indel-derived neoantigens using rna-seq data,” *Bioinformatics*, vol. 35, no. 20, pp. 4159–4161, 2019.

References XVI

[48] R. O. Schenck, E. Lakatos, C. Gatenbee, T. A. Graham, and A. R. Anderson, “Neopredpipe: high-throughput neoantigen prediction and recognition potential pipeline,” *BMC bioinformatics*, vol. 20, no. 1, pp. 1–6, 2019.

[49] J. Hundal, S. Kiwala, J. McMichael, C. A. Miller, H. Xia, A. T. Wollam, C. J. Liu, S. Zhao, Y.-Y. Feng, A. P. Graubert *et al.*, “pvactools: a computational toolkit to identify and visualize cancer neoantigens,” *Cancer immunology research*, vol. 8, no. 3, pp. 409–420, 2020.

[50] Y. Li, G. Wang, X. Tan, J. Ouyang, M. Zhang, X. Song, Q. Liu, Q. Leng, L. Chen, and L. Xie, “Progeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection,” *BMC medical genomics*, vol. 13, no. 5, pp. 1–11, 2020.

References XVII

[51] M. A. Wood, A. Nguyen, A. J. Struck, K. Ellrott, A. Nellore, and R. F. Thompson, “Neoepiscope improves neoepitope prediction with multivariant phasing,” *Bioinformatics*, vol. 36, no. 3, pp. 713–720, 2020.

[52] A. C. M. Coelho, A. L. Fonseca, D. L. Martins, P. B. Lins, L. M. da Cunha, and S. J. de Souza, “neoant-hill: an integrated tool for identification of potential neoantigens,” *BMC Medical Genomics*, vol. 13, no. 1, pp. 1–8, 2020.

[53] C. Wert-Carvajal, R. Sánchez-García, J. R. Macías, R. Sanz-Pamplona, A. M. Pérez, R. Alemany, E. Veiga, C. Ó. S. Sorzano, and A. Muñoz-Barrutia, “Predicting mhc i restricted t cell epitopes in mice with nap-cnb, a novel online tool,” *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

References XVIII

[54] L. Y. Zhou, F. Zou, and W. Sun, “Prioritizing candidate peptides for cancer vaccines by pepprmint: a statistical model to predict peptide presentation by hla-i proteins,” *bioRxiv*, 2021.

[55] Y. L. Terai, C. Huang, B. Wang, X. Kang, J. Han, J. Douglass, E. H.-C. Hsiue, M. Zhang, R. Purohit, T. deSilva *et al.*, “Valid-neo: A multi-omics platform for neoantigen detection and quantification from limited clinical samples,” *Cancers*, vol. 14, no. 5, p. 1243, 2022.

[56] Y. Zhang, G. Zhu, K. Li, F. Li, L. Huang, M. Duan, and F. Zhou, “Hlab: learning the bilstm features from the protbert-encoded proteins for the class i hla-peptide binding prediction,” *Briefings in Bioinformatics*, 2022.

References XIX

[57] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium, “Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches,” *Bioinformatics*, vol. 31, no. 6, pp. 926–932, 2015.

[58] M. Steinegger and J. Söding, “Clustering huge protein sequence sets in linear time,” *Nature communications*, vol. 9, no. 1, pp. 1–8, 2018.

[59] N. Hashemi, B. Hao, M. Ignatov, I. Paschalidis, P. Vakili, S. Vajda, and D. Kozakov, “Improved predictions of mhc-peptide binding using protein language models,” *bioRxiv*, 2022.

Questions?