Computational and Structural Biotechnology Journal 21 (2023) 5538–5543

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method article

SIGANEO: Similarity network with GAN enhancement for immunogenic neoepitope prediction☆

Yilin Ye [a,1], Yiming Shen [a,1], Jian Wang [a,1], Dong Li [a], Yu Zhu [a], Zhao Zhao [a], Youdong Pan [a], Yi Wang [a], Xing Liu [b], Ji Wan [a,*]

[a] Shenzhen Neocura Biotechnology Co. Ltd., Shenzhen 518055, China
[b] The Center for Microbes, Development and Health, Key Laboratory of Molecular Virology and Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai 200031, China

ARTICLE INFO

Keywords:
Deep learning
Neoepitope
Neoantigen
Immunogenicity
Cancer immunotherapy

ABSTRACT

Target selection of the personalized cancer neoantigen vaccine, which is highly dependent on computational prediction algorithms, is crucial for its clinical efficacy. Due to the limited number of experimentally validated immunogenic neoepitopes as well as the complexity of neoantigens in eliciting T cell response, the accuracy of neoepitope immunogenicity prediction methods requires persistent efforts for improvement. We present a deep learning framework for neoepitope immunogenicity prediction – SIGANEO by integrating GAN-like network with similarity network to address issues of missing values and limited data concerning neoantigen prediction. This framework exhibits superior performance over competing machine-learning-based neoantigen prediction algorithms over an independent test dataset from TESLA consortium. Particularly for the clinical setting of neoantigen vaccine where only the top 10 and 20 predictions are selected for vaccine production, SIGANEO achieves significantly better accuracy for predicting experimentally validated neoepitopes. Our work demonstrates that deep learning techniques can greatly boost the accuracy of target identification for cancer neoantigen vaccine.

1. Introduction

Neoantigen vaccines have emerged as a promising type of cancer immunotherapy through augmenting cancer-specific cytotoxic T cells [1,2]. The development of neoantigen vaccines is commonly initiated by computational prediction of candidate neoepitopes, the accuracy of which greatly impacts the efficacy of vaccine. The rationale for computational neoantigen prediction is retrieving neoepitopes derived from diversified events of genetic alteration and RNA dysregulation through bioinformatics analyses, which is followed by applying computational models to infer the likelihood of neoepitopes for eliciting T cell responses [3]. According to a recent survey, there are seven major categories of features being employed by a variety of academic and industrial groups for predicting tumor neoantigen immunogenicity [4]. To date, prediction accuracies of immunogenic neoepitopes are at a relatively low level and vary greatly, implying that persistent efforts are required for improvement [4,5]. One of the bottlenecks of machine-learning-based algorithms lies in the availability of experimentally validated immunogenicity results of candidate neoantigens. Only hundreds of neoantigens have been validated as immunogenic by a variety of independent studies. Furthermore, there is a great degree of inconsistency in features used by different studies. As a result, there was a great number of missing values after merging different datasets for model training. For instance, many biological features such as VAF and gene expression are not available due to the unavailability of raw sequencing data. Thus, to enhance the accuracy of computational neoantigen prediction and to optimally utilize experimentally validated data for model training, it is crucial to develop a novel machine learning framework for neoantigen prediction. This framework should manage to handle both limited positive samples and prevalent missing values in the training dataset.

Here we present a neoepitope immunogenicity prediction framework – SIGANEO by integrating a generative adversarial network (GAN) with a similarity network. We demonstrate that GAN-imputed data exhibits a

☆ Trained model is available at https://github.com/NCTool/SIGANEO-GAN.

* Corresponding author.
E-mail address: jiw@neocura.net (J. Wan).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2023.10.050
Received 19 July 2023; Received in revised form 26 October 2023; Accepted 27 October 2023
Available online 31 October 2023

higher level of agreement compared to other data-filling methods. In the meantime, our framework outperforms other machine-learning-based methods, especially in the top 10 or 20 predictions that are critical for the therapeutic application of neoantigen-based cancer vaccine [4,6,7].

2. Materials and methods

2.1. Datasets

To perform unsupervised training on our generative adversarial network (GAN) module, we collected paired-end whole-exome sequencing and RNA-sequencing data from 117 loci of 83 patients across six different cancer types [8–13]. After data processing and filtering, a total of 279,672 mutated peptides without labels were obtained for the GAN training process.

A total of 1589 experimentally validated peptides (104 were validated as immunogenic neoepitopes), which span 8 different cancer types, were collected from 12 distinct studies for similarity network training (Table 1 and Supplementary Table 1).

To evaluate the performance of SIGANEO, we compiled a test dataset by analyzing raw data of eight NSCLC and Melanoma patients obtained from the Tumor NeoEpitope Selection Alliance (TESLA) [4]. As a result, a total of 678 peptides that have been experimentally validated peptides were included in the dataset, out of which 34 were confirmed as immunogenic.

2.2. Data processing and feature extraction

All the whole exome sequencing data were firstly filtered using FASTP (v0.20.1) to remove low-quality reads [26]. The resulting high-quality reads were aligned to the hg38 human genome reference by BWA-MEM (v0.7.17) [27]. Alignment preprocessing, including duplicates marking and base quality score recalibration, was performed using the high-performance tool elPrep (v4.0) [28]. After preprocessing, somatic variants were identified by Mutect2 and FilterMutectCalls from GATK (v4.1.6) [29]. Only variants with PASS FILTER tag were retained. The effects of these variants were determined by VEP (v94) [30]. The identified variants were processed to extract mutant peptides (8- to 11-mer) with a sliding window size of 1. All mutant peptides were screened against the reference proteome and only tumor-specific peptides (without perfect matches in the reference proteome) were retained. Class I human leukocyte antigen (HLA) subtyping was inferred by running HLA-LA against the WES data of control samples [31]. Additionally, the whole transcriptome sequencing reads (RNA-seq) were also filtered using FASTP, followed by being aligned to the hg38 using STAR aligner (v2.7.2) [32].

Variant allele frequency (VAF) was calculated as the ratio of alternative allele depth (AD) to the total depth at a given variant locus. Gene expression quantification measures (TPM and FPKM) were retrieved using RSEM (v1.3.1) [33]. RNA alt/ref read counts were defined as the number of RNA-seq reads matching the variant/reference at a given locus.

To get a comprehensive binding affinity feature representation, binding affinities between HLA molecules and peptides were predicted using netMHCpan (v4.0), MHCflurry and MATHLA respectively [34–36]. Apart from binding affinity, binding stability of the pMHC complex is another critical feature for characterizing the pMHC. Accordingly, netMHCstabpan(v1.0) was applied to predict the pMHC binding stability [37]. Both Thalf (predicted half-life of the pMHC complex) and rank metrics were taken into consideration. Peptide HLA pairs with predicted binding rank larger than 2 % or IC50 value larger than 500 were filtered out to remove low affinity results. The recognition score represents the TCR recognition probability which was defined by Marta Łuksza et al. [38]. The sequence similarity was determined by aligning peptides to all the epitopes in the Immune Epitope Database [39]. The BLOSUM62 matrix was applied during alignment with gap open penalty 11 and gap extension penalty 1 [40]. In the recognition score model, both a and k are free parameters adjusted during the calculation of binding energies. Here, a denotes the horizontal shifting of the binding curve, while k represents the steepness. The values of a and k were set to 26 and 4.86936, respectively.

Many previous studies have uncovered that the strength of amino acid hydrophobicity plays an important role in activating T cell responses [41,42]. Therefore, we calculated two different amino acid hydrophobicity scores for each peptide with the following equations:

$$acid\_score_1(acid) = \begin{cases} 1, acid \in \{F, K, M, W, S\} \\ 0, otherwise \end{cases}$$

$$ah_1 = \sum_{i=0}^{peptide\ length} acid\_score_1(acid_i)$$

$$ah_2 = \sum_{i=0}^{peptide\ length} KD(acid_i) * W_{MATHLA}(HLA, i)$$

where KD refers to Kyte-Doolittle hydrophobicity, W_{MATHLA} represents the amino acid weight obtained by the multi-head attention mechanism from MATHLA, and HLA refers to the HLA allele of the pMHC complex [36,43].

3. Model

The whole SIGANEO framework is composed of five modules – GAN, pMHC encoder, representation encoder, similarity calculator and similarity rank net. Firstly, to address the issue of missing features in the public datasets used for model training, we employed a generative-adversarial-network-like (GAN-like) structure to impute missing values, which is an unsupervised learning module that combines a generator module and a discriminator module to execute data imputation (Fig. 1B). Next, the pMHC encoder module is designed to eliminate redundant information from pMHC-associated features, thus obtaining relevant pMHC embeddings. The representation encoder module integrates pMHC embeddings with original biological features to generate the final embeddings. All of these three modules are trained using unlabelled data. Finally, the similarity calculator and similarity rank net use labelled data to calculate similarity matrix and predict neoantigen immunogenicity.

For each pMHC complex, we combined the biological features (RNA_ref_read_count, RNA_alt_read_count, Gene_FPKM, Gene_TPM, VAF) with the pMHC-associated features (ah_1, ah_2, MHCFlurry_aff, MHCFlurry_rank, NetMHCPan_aff, NetMHCPan_rank, NetMHCstabpan_Thalf(h), NetMHCstabpan_rank, MATHLA_aff, Recognition_score) as the input of the GAN module (Fig. 1B). The generator module is a symmetrical four-layer MLP (Tanh function as activation function), where the original neoepitope representation is encoded into 16 and 32-

Table 1
Summary of the training dataset.

Studies	Number of Samples	Number of positive samples
[14]	23	1
[15]	45	5
[16]	112	25
[17]	17	3
[18]	699	9
[19]	60	3
[20]	308	10
[21]	37	10
[22]	28	7
[23]	21	9
[24]	71	4
[25]	168	18
Total	**1589**	**104**

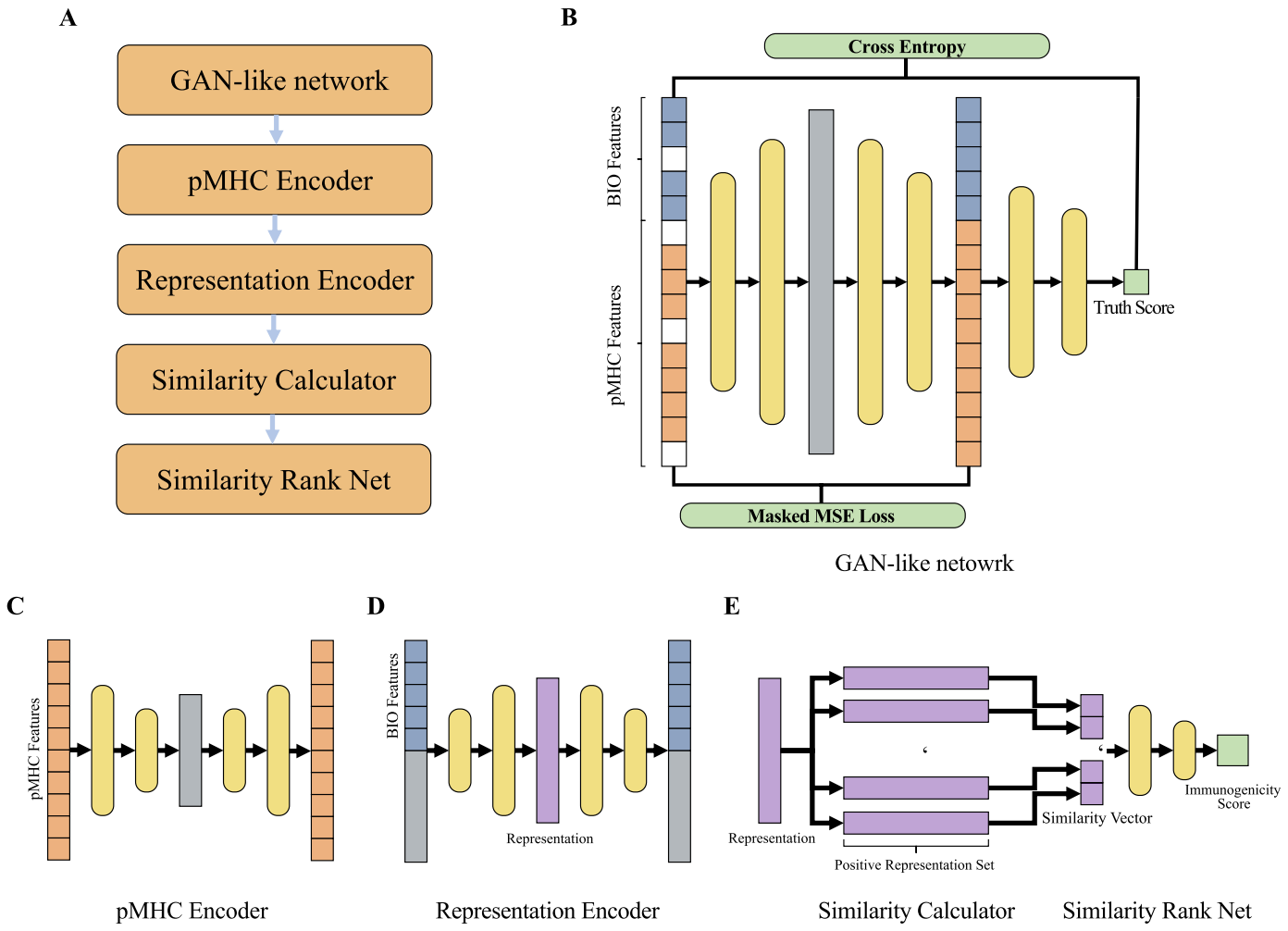

Fig. 1. The structure of SIGANEO. A. The overview of SIGANEO. B. The structure of GAN-like network. This network is composed of a generator module and a discriminator module. The output of the generator module is later used by the pMHC encoder and the representation encoder. The uncolored squares depicted in the figure correspond to the missing features within the original data. The differently sized, yellow rounded rectangles in the figure refer to linear layers with different dimensions. C. The structure of pMHC Encoder. This is an auto-encoder-decoder module for encoding pMHC-associated features imputed by preceding GAN-like module. The pMHC representation is the output of the second layer of the network. D. The structure of representation Encoder. This is an auto-encoder-decoder module to obtain sample representation by encoding biological features (imputed by GAN-like module) and pMHC representation in Fig. 1C. The sample representation is the output of the second layer of the network. E. The structure of similarity network. Similarity calculator is used to calculate the representation similarity between query sample and positive samples, which obtains a new sample embedding for predicting immunogenicity by similarity rank net.

dimensional vectors respectively which are later restored into the original dimensionality.

To ensure that the training process of generator module is not impacted by missing values, we used MSE loss with mask to evaluate the loss between the original representation and the restored representation, as shown in Equations 1 and 2 [44,45].

$$loss_{ij} = \frac{x_{ij} - \bar{x}_{ij}}{num} \times mask_{ij} \tag{1}$$

$$mask_{ij} = \begin{cases} 1, if \quad x_{ij} \quad is \quad not \quad None \\ 0, if \quad x_{ij} \quad is \quad None \end{cases} \tag{2}$$

In this generative adversarial network, a two-layer MLP serves as the discriminator module to distinguish real data from imputed data. The representation of each sample is transformed to a 32-dimensional vector which is further reduced to a 2-dimensional vector. The performance of discriminator is evaluated by the Cross Entropy Loss [46]. Once the adversarial training process of generator and discriminator modules is completed, the final data generated by the GAN are fed into downstream embedding modules.

We encoded and compressed pMHC features into pMHC binding representation through an auto-encoder-decoder module which consists of four fully-connected layers with dimensions of 16, 8, 8 and 16 sequentially (Fig. 1C) [47]. Furthermore, another auto-encoder-decoder network, consisting of four fully-connected layers with sequential dimensions of 16, 64, 64 and 16, is employed to automatically integrate and compress the pMHC binding representation with other biological features to obtain a 64-dimensional query representation, which is then used for subsequent prediction and classification tasks (Fig. 1D).

Due to the limited number of positive samples, the cosine similarity between the embedding of each sample and the positive samples was calculated, resulting in a new similarity vector according to the following equation [48]:

$$SV_i = \left[\frac{x_i \bullet \check{x}_1}{\|x_i\| \bullet \|\check{x}_1\|}, \frac{x_i \bullet \check{x}_2}{\|x_i\| \bullet \|\check{x}_2\|}, \ldots, \frac{x_i \bullet \check{x}_j}{\|x_i\| \bullet \|\check{x}_j\|} \right]$$

where x_i is the embedding of the neoepitope, SV_i is the similarity vector of x_i, and $\check{x}_j$ is the embedding of the j-th positive sample; $\| \bullet \|$ is the norm of the vector.

The embedding reconstruction can offset the systematic bias in previous steps. Meanwhile, the final embedding vector is not

constructed directly from the original features, so the network is more fault tolerant to imputed values [49].

Ultimately, a two-layer MLP structure was used in the final classification network. This facilitates the training of the similarity rank net, enabling it to achieve better performance even with a limited number of labelled training sets [50].

$$S_{Immunogenicity} = Sigmoid(Tanh(SV_i \bullet w_1) \bullet w_2)$$

Where $S_{Immunogenicity}$ is the predicted value of immunogenicity, $SV_i \in R^{1*104}$ is the Similarity Vector of the *i*-th sample, Sigmoid and Tanh are activation functions, and $w_1 \in R^{104*32}$ and $w_2 \in R^{32*1}$ are weight matrices of MLP.

During the training process of SIGANEO, the Adam optimization function is utilized with a learning rate of 0.001. The batch size of the GAN module is set as 256 and the epoch is set as 1000. In the similarity module, the batch size of the pMHC encoder, Representation encoder, and Similarity Rank Net are set to 64, 64, and 32 respectively. For each of these networks, the epoch is set as 200. The early stop technique is used, which stops the training process is if there is no decrease in loss over five consecutive epochs.

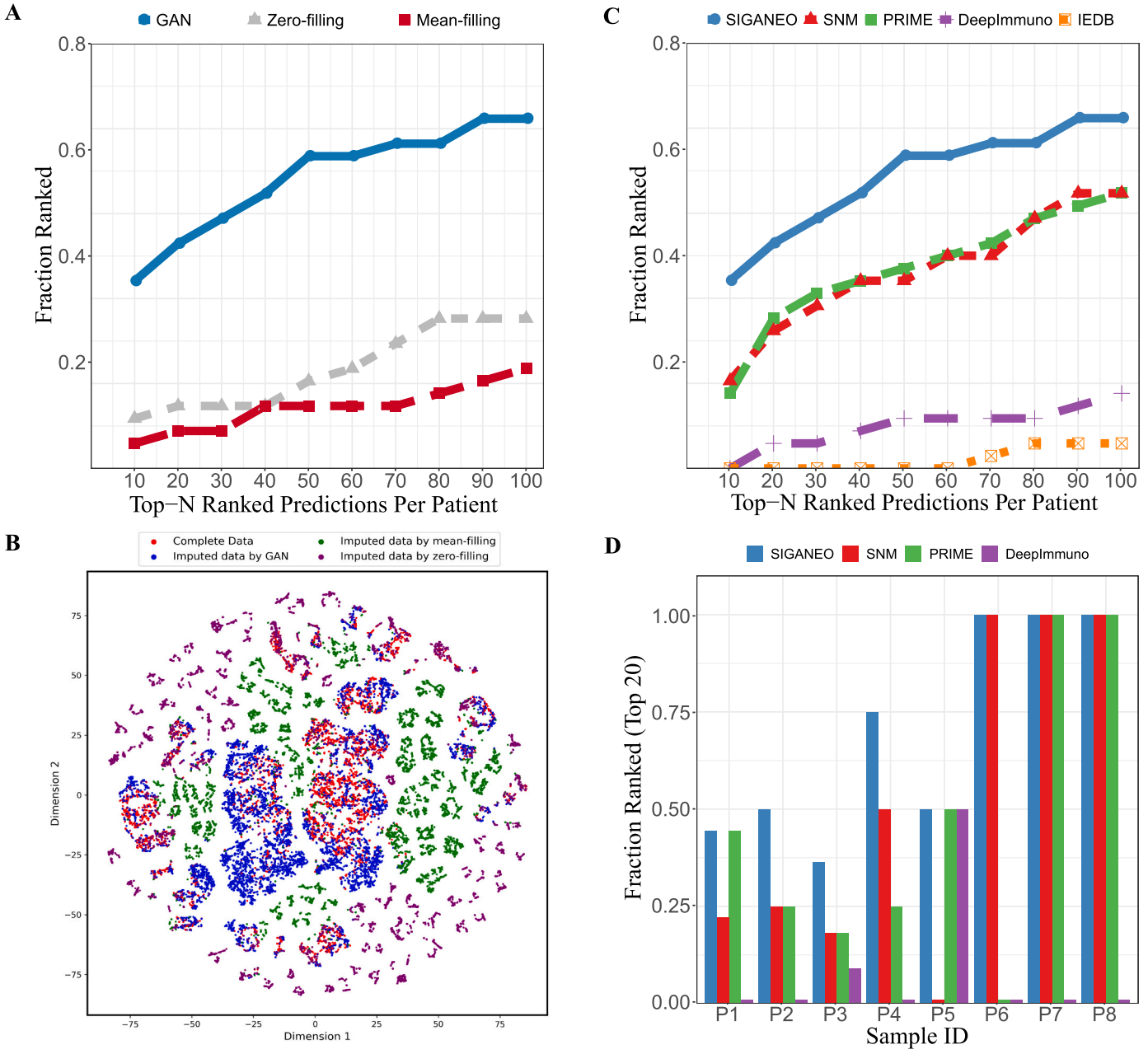

Fig. 2. Model performance. A. Line plot of fraction ranked values (y axis) at the top-N ranked predictions per sample (x axis). Each dot denotes the fraction of validated immunogenic peptides in the top N ranked predictions. B. Visualization of data imputation by SIGANEO's GAN module, zero-filling and mean-filling methods. The t-SNE results are visualized along dimensions 1 and 2. The complete TESLA dataset is denoted by red dots. Blue, green and purple dots represent data points imputed by GAN, mean-filling and zero-filling methods, respectively. C. Line plot of fraction ranked values (y axis) at the top-N ranked predictions per sample (x axis). Different colors represent different machine-learning-based immunogenicity methods. D. Histogram of FR values corresponding to each tool at the top 20 predictions among different patients (x axis). Bar colors represent different prediction tools. Patient IDs of P1 to P8 correspond to patient ID 1, 2, 3, 12, 16, 4, 8 and 9 from the TESLA consortium respectively. Methods with FR of 0 were assigned with a small value of 0.01 to avoid blank bar in the figure.

4. Results

To select the appropriate method for handling missing values in the training data, we compared three missing value imputation algorithms – zero-filling (filling all the missing values with 0), mean-value-filling (filling all the missing values with mean value of a given feature), and GAN-based filling methods. Initially, we utilized five-fold cross-validation to determine the optimal hyperparameters for the GAN-based model and to assess its performance (Fig. S1A). Next, the benchmarking metric – fraction ranked (FR), which was specifically designed to reflect accuracy of neoantigen prediction algorithm by a previous study, is utilized to compare different filling methods [4]. Fraction ranked is defined as the ratio of the number of experimentally validated neoepitopes in the top N predictions to the total number of validated immunogenic neoepitopes in the test set.

It can be apparently seen that FRs at the top-N predicted neoepitopes (N ranges from 10 to 100 with step size of 10) of the similarity network trained on the GAN-imputed data are significantly higher than the other two data filling methods (paired t-test p-values: $2.72*10^{-9}$ and $5.65*10^{-9}$ as compared to zero-filling and mean-filling respectively) (Fig. 2A). The average improvement of GAN-coupled similarity network of each top-N setting are 197 % and 376 % over zero-filling-coupled and mean-value-filling-coupled networks, respectively. It is also worth noting that the average performance of zero-filling-coupled network is higher than mean-value-filling-coupled network. One possible explanation is that zero-filling method can mask missing values by 0, whereas mean-filling method introduces more inaccurate information to the network [51]. To gain a more intuitive understanding of the impacts of missing-value imputation methods on neoantigen prediction performance, we randomly removed 42.5 % of values from the test dataset according to the rate of missing value in the training dataset. Given that t-SNE (t-Distributed Stochastic Neighbor Embedding) is better at preserving the local structures in high-dimensional data, it was employed to reduce the data dimensions for visualization before and after imputation. It is evident that the data imputed by GAN (blue dots) align more closely with the original complete data (red dots) compared to the other two methods (Fig. 2B).

To further evaluate the performance of our method, we compared SIGANEO to another four neoepitope immunogenicity prediction tools – SNM, PRIME, DeepImmuno and IEDB immunogenicity prediction tool [39,52–54]. We initiated a comprehensive model comparison using conventional evaluation metrics – the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area under the Precision-Recall Curve (AUPRC). This comparison aimed to provide an insight into the performance of our model over the entire range of prediction results. It is revealed that SIGANEO significantly outperforms all the other competing methods by achieving AUROC of 0.94 and AUPRC of 0.338 (Fig. S1B and S1C). In the field of neoantigen cancer vaccine, FR metric is a more important metric due to the fact that it can effectively capture the accuracy of the top-ranked predicted neoepitopes, which are subsequently utilized in the formulation of cancer vaccines. As a result, FRs of SIGANEO are distinctly higher than all the other four tools at every setting of the top N ranked neoepitopes (Fig. 2C). The mean FR of SIGANEO is 0.685 which is significantly higher than PRIME and SNM by 45 % and 47 % respectively (paired t-test p-values: $1.79*10^{-8}$ and $3.19*10^{-8}$). In real world clinical application of neoantigen vaccine, the top 10 or 20 predicted neoepitopes are usually selected for vaccine development [4,6,7]. Therefore, FRs at the top 10 and 20 predicted neoepitopes garner more weights during assessment of immunogenicity prediction tools. FRs at the top 10 and 20 predicted neoepitopes reaches 0.441 and 0.529 which are 114 % and 50 % higher than the second highest tools (SNM at the top 10 with FR of 0.206 and PRIME at the top 20 with FR of 0.353), demonstrating that SIGANEO possesses a significant advantage in identifying immunogenic peptides in the context of therapeutic application.

To further examine the performance of SIGANEO, we scrutinized the top 20 predicted neoepitopes at the level of individual patients as shown in Fig. 2D. The results of IEDB were excluded from the comparison because none of validated immunogenic peptide was included in its top 20 predictions of any patient. Overall, FRs of the top 20 predictions of SIGANEO are greater than or equal to FRs of all the other tools in all 8 patients. Particularly, SIGANEO identified 18 validated immunogenic neoepitopes for 8 patients which was much higher than competing tools (PRIME: 12, SNM: 11, DeepImmuno: 2). Notably, for the top 20 prediction results of all the 3 patients in the independent cohort of TESLA, namely P6, P7 and P8, two machine learning based tools, SIGANEO and SNM, both managed to identify 100 % of validated immunogenic epitopes, surpassing the approach proposed by the TESLA consortium [4]. Last but not least, SIGANEO was the only method being able to include at least one validated immunogenic neoepitope in the top 20 results of all 8 samples, showing its robustness in the ability to identify immunogenic peptides across different individuals.

5. Discussion

The efficacy of the individualized cancer neoantigen vaccine is heavily dependent on the accuracy of the neoantigen prediction. Therefore, accurate computational prediction of immunogenic neoepitopes is highly desired and has attracted great attention from both academic and industrial endeavors. However, the limited availability of experimentally validated neoepitope data has posed a challenge for machine learning algorithms to achieve high accuracy in neoantigen prediction. The superior performance of our model integrating autoencoder, generative adversarial network (GAN) and similarity network have provided several insights into further development of machine-learning-based neoepitope immunogenicity prediction frameworks. Firstly, unsupervised deep learning model such as GAN has a great potential in handling missing values within training dataset as well as reconciling heterogeneity across different studies. The superior performance of GAN imputation can likely be attributed to the inherent correlations among the features used for neoantigen prediction. In contrast to univariate imputation methods like zero-filling and mean-filling, GAN leverages the intercorrelation of features to more accurately estimate the distribution of missing data, resulting in a more precise imputation algorithm. In addition, GAN framework can also augment immunogenic neoepitope data which could possibly boost the performance of epitope immunogenicity prediction significantly [55,56]. Secondly, the similarity network can, to an extent, mitigate the imbalance between positive (immunogenic peptide) and negative samples in the training set. In a similar situation with an imbalanced training dataset, one-class kernel extreme learning machine has been effectively utilized for the Covid-19-Pneumonia classification [57]. In this regard, one-class classification algorithms can be implemented to neutralize the impact of imbalanced data on neoepitope immunogenicity prediction.

Funding

None.

CRediT authorship contribution statement

Y.Y. designed and developed the framework as well as wrote the manuscript. Y.S. and J.W. conducted data processing, compiled datasets for model training and evaluation, and wrote the manuscript. D.L. developed method for feature extraction, Z.Z., Y.P., Y.W. and X.L. revised the manuscript. J.W. designed the project and wrote the final manuscript.

Declaration of Competing Interest

The authors declared that they have no competing interests.

Acknowledgement

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.050.

References

[1] Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nat Rev Clin Oncol 2021;18(4):215–29.
[2] Yarchoan M, Johnson 3rd BA, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. Nat Rev Cancer 2017;17(4):209–22.
[3] Lang F, Schrors B, Lower M, Tureci O, Sahin U. Identification of neoantigens for individualized therapeutic cancer vaccines. Nat Rev Drug Discov 2022;21(4):261–82.
[4] Wells DK, van Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KCF, Campbell KM, et al. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. Cell 2020;183(3):818–34. e813.
[5] The problem with neoantigen prediction. Nature Biotechnol 2017, 35(2):97.
[6] Rojas LA, Sethna Z, Soares KC, Olcese C, Pang N, Patterson E, et al. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. Nature 2023;618(7963):144–50.
[7] Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Lower M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature 2017;547(7662):222–6.
[8] Jia Q, Wu W, Wang Y, Alexander PB, Sun C, Gong Z, et al. Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. Nat Commun 2018;9(1):5361.
[9] He Y, Ramesh A, Gusev Y, Bhuvaneshwar K, Giaccone G. Molecular predictors of response to pembrolizumab in thymic carcinoma. Cell Rep Med 2021;2(9):100392.
[10] Kang HG, Hwangbo H, Kim MJ, Kim S, Lee EJ, Park MJ, et al. Aberrant transcript usage is associated with homologous recombination deficiency and predicts therapeutic response. Cancer Res 2022;82(1):142–54.
[11] Lee HW, Chung W, Lee HO, Jeong DE, Jo A, Lim JE, et al. Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient. Genome Med 2020;12(1):47.
[12] Restrepo P, Yong R, Laface I, Tsankova N, Nael K, Akturk G, et al. Tumoral and immune heterogeneity in an anti-PD-1-responsive glioblastoma: a case study. Cold Spring Harb Mol case Stud 2020;6(2).
[13] Codrich M, Dalla E, Mio C, Antoniali G, Malfatti MC, Marzinotto S, et al. Integrated multi-omics analyses on patient-derived CRC organoids highlight altered molecular pathways in colorectal cancer progression involving PTEN. J Exp Clin Cancer Res: CR 2021;40(1):198.
[14] Sneddon S, Rive CM, Ma S, Dick IM, Allcock RJN, Brown SD, et al. Identification of a CD8+ T-cell response to a predicted neoantigen in malignant mesothelioma. Oncoimmunology 2020;9(1):1684713.
[15] Chen F, Zou Z, Du J, Su S, Shao J, Meng F, et al. Neoantigen identification strategies enable personalized immunotherapy in refractory solid tumors. J Clin Investig 2019;129(5):2056–70.
[16] Peng S, Zaretsky JM, Ng AHC, Chour W, Bethune MT, Choi J, et al. Sensitive detection and analysis of neoantigen-specific T cell populations from tumors and blood. Cell Rep 2019;28(10):2728–38. e2727.
[17] Cafri G, Yossef R, Pasetto A, Deniger DC, Lu YC, Parkhurst M, et al. Memory T cells targeting oncogenic mutations detected in peripheral blood of epithelial cancer patients. Nat Commun 2019;10(1):449.
[18] van den Bulk J, Verdegaal EME, Ruano D, Ijsselsteijn ME, Visser M, van der Breggen R, et al. Neoantigen-specific immunity in low mutation burden colorectal cancers of the consensus molecular subtype 4. Genome Med 2019;11(1):87.
[19] Perumal D, Imai N, Lagana A, Finnigan J, Melnekoff D, Leshchenko VV, et al. Mutation-derived neoantigen-specific T-cell responses in multiple myeloma. Clin Cancer Res: J Am Assoc Cancer Res 2020;26(2):450–64.
[20] Cohen CJ, Gartner JJ, Horovitz-Fried M, Shamalov K, Trebska-McGowan K, Bliskovsky VV, et al. Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes. J Clin Investig 2015;125(10):3981–91.
[21] Tran E, Ahmadzadeh M, Lu YC, Gros A, Turcotte S, Robbins PF, et al. Immunogenicity of somatic mutations in human gastrointestinal cancers. Science 2015;350(6266):1387–90.
[22] Gros A, Parkhurst MR, Tran E, Pasetto A, Robbins PF, Ilyas S, et al. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. Nat Med 2016;22(4):433–8.
[23] Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science 2015;348(6236):803–8.
[24] Zacharakis N, Chinnasamy H, Black M, Xu H, Lu YC, Zheng Z, et al. Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. Nat Med 2018;24(6):724–30.
[25] Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature 2017;547(7662):217–21.
[26] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34(17):i884–90.
[27] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754–60.
[28] Herzeel C, Costanza P, Decap D, Fostier J, Verachtert W. elPrep 4: a multithreaded framework for sequence analysis. PloS One 2019;14(2):e0209523.
[29] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43(5):491–8.
[30] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome Biol 2016;17(1):122.
[31] Dilthey AT, Mentzer AJ, Carapito R, Cutland C, Cereb N, Madhi SA, et al. HLA*LA-HLA typing from linearly projected graph alignments. Bioinformatics 2019;35(21):4394–6.
[32] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.
[33] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinforma 2011;12:323.
[34] 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 2020;48(W1):W449–54.
[35] O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. Cell Syst 2018;7(1):129–32. e124.
[36] Ye Y, Wang J, Xu Y, Wang Y, Pan Y, Song Q, et al. MATHLA: a robust framework for HLA-peptide binding prediction integrating bidirectional LSTM and multiple head attention mechanism. BMC Bioinforma 2021;22(1):7.
[37] Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, Nielsen M, et al. Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. J Immunol 2016;197(4):1517–24.
[38] Luksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature 2017;551(7681):517–20.
[39] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res 2019;47(D1):D339–43.
[40] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89(22):10915–9.
[41] Mahmoudvand S, Shokri S, Makvandi M, Taherkhani R, Rashno M, Jalilian FA, et al. In silico prediction of T-cell and B-cell epitopes of human papillomavirus type 16 L1 protein. Biotechnol Appl Biochem 2022;69(2):514–25.
[42] Shao MM, Yi FS, Huang ZY, Peng P, Wu FY, Shi HZ, et al. T cell receptor repertoire analysis reveals signatures of T cell responses to human mycobacterium tuberculosis. Front Microbiol 2022;13:829694.
[43] Porto WF, Ferreira KCV, Ribeiro SM, Franco OL. Sense the moment: a highly sensitive antimicrobial activity predictor based on hydrophobic moment. Biochim Et Biophys Acta Gen Subj 2022;1866(3):130070.
[44] Xu Z, Elshamy S, Zhao Z, Fingscheidt T. Components loss for neural networks in mask-based speech enhancement. EURASIP J Audio Speech Music Process 2021;2021(1):24.
[45] Luo Y., Chen Z., Gao X.: Self-distillation Augmented Masked Autoencoders for Histopathological Image Classification. In.; 2022: arXiv:2203.16983.
[46] Haque A.: EC-GAN: Low-Sample Classification using Semi-Supervised Algorithms and GANs. In.; 2020: arXiv:2012.15864.
[47] Dincer AB, Janizek JD, Lee SI. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. Bioinformatics 2020;36(Suppl_2):i573–82.
[48] Xia P, Zhang L, Li F. Learning similarity with cosine similarity ensemble. Inf Sci 2015;307:39–52.
[49] Li S, Sung Y. INCO-GAN: variable-Length Music generation method based on inception model-based conditional GAN. Mathematics 2021;9(4):387.
[50] Liu M, Cai Z, Chen J. Adaptive two-layer ReLU neural network: I. Best least-squares approximation. Comput Math Appl 2022;113:34–44.
[51] Dong H. Improvement of the model by preprocessing big data of tapping temperature prediction industry. J Phys: Conf Ser 2022;2235(1):012089.
[52] Gartner JJ, Parkhurst MR, Gros A, Tran E, Jafferji MS, Copeland A, et al. A machine learning model for ranking candidate HLA class I neoantigens based on known neoepitopes from multiple human tumor types. Nat Cancer 2021;2(5):563–74.
[53] Schmidt J, Smith AR, Magnin M, Racle J, Devlin JR, Bobisse S, et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. Cell Rep Med 2021;2(2):100194.
[54] Li G, Iyer B, Prasath VBS, Ni Y, Salomonis N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. Brief Bioinforma 2021;22(6).
[55] Antoniou A., Storkey A., Edwards H.: Data Augmentation Generative Adversarial Networks. In.; 2017: arXiv:1711.04340.
[56] Van Oort CM, Ferrell JB, Remington JM, Wshah S, Li J. AMPGAN v2: machine learning-guided design of antimicrobial peptides. J Chem Inf Model 2021;61(5):2198–207.
[57] Khan MA, Kadry S, Zhang YD, Akram T, Sharif M, Rehman A, et al. Prediction of COVID-19 - pneumonia based on selected deep features and one class kernel extreme learning machine. Comput Electr Eng: Int J 2021;90:106960.