EPiC Series in Computing

Volume 83, 2022, Pages 116–125

Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

A Deep Dimensionality Reduction Method based on Variational Autoencoder for Antibody Complementarity Determining Region Sequences Analysis

Saeed Khalilian[2], Mohammad N. Isfahani[2], Zahra Moti[3], Arian Baloochestani[4], Alireza Chavosh[1], and Zahra Hemmatian[1,*]

[1] MarWell Bio Inc., California, United States,
[*]Corresponding Author. zara@marwell.bio
[2] Independent Researcher, Iran,
[3] Independent Researcher, The Netherlands
[4] Independent Researcher, Norway

Abstract

An essential task in antibody/nanobody therapeutics discovery is to rapidly identify whether an antibody/nanobody has specificity and cross-reactivity to one or multiple targets. Multiple target specificity and cross-reactivity of antibodies can be demonstrated by screening the third Complementarity Determining Region on the heavy chain (CDR-H3) of antibody sequences. However, the existing methods are costly and labor-intensive as repetitive wet-lab experimentation is required to explore the sequences space. Here, we present a deep learning dimensionality reduction model based on Variational Autoencoder (VAE) and Residual Neural Network (Resnet), which we named VAEResDR. Our VAEResDR can efficiently learn the sequences' key features while scaling down high-dimensional antibody sequences into a two-dimensional visualization representation for coherent analysis and rapid screening. We demonstrate that our VAEResDR can provide a tool to precisely analyze CDR-H3 sequences within the hidden patterns and effectively improve antibody/-nanobody CDR-H3 sequence clustering.

1 Introduction

Antibody-based therapeutics have increasingly fewer adverse effects due to their high specificity to given targets. Next-generation antibodies are offering promising classes of therapeutics for treating various diseases, particularly those requiring long-term treatment plans, e.g., cancer [14]. The third complementarity determining region of the heavy chains (CDR-H3) of an antibody/nanobody plays an essential role in recognizing and binding an antibody to its targets [26]. High variation in CDR-H3 amino acid sequences allows exploring antibody binding to one or multiple targets amongst a big pool of CDR-H3 sequences. However, this requires costly and laborious conventional wet-lab experimentation to screen and analyze CDR-H3 sequences for their specificity and binding responses to one or multiple targets [5]. In-silico analysis of antibody amino acid sequences can significantly shorten the time and reduce the cost of specificity

screening and cross-reactivity studies [14]. Such analysis may also improve the utility of antibodies/nanobodies in multiple model organisms, offering potential therapeutic candidates for multiple targets [11]. Machine learning-based (ML) approaches such as clustering and dimensionality reduction (DR) have been developed to analyze the sequence databases. Clustering algorithms such as K-means [8], and Gaussian Mixture Model (GMM) [30] have been utilized for sequence analysis. Nonetheless, these algorithms work reasonably well for low-dimensional databases; however, for high-dimensional databases such as antibody sequences, their clustering performance degrades considerably due to the curse of dimensionality [17]. To reduce this computational complexity on large and high-dimensional antibody sequence databases, dimension reduction (DR) methods are used alongside clustering. Machine learning-based DR methods such as Principal Component Analysis (PCA) [29], t-distributed Stochastic Neighbor Embedding (t-SNE) [27], and Uniform Manifold Approximation and Projection (UMAP) [19] are utilized to reduce the sequences' dimensions. Yet, compared to original features, using clustering alongside these DR methods often degrades clustering quality since important features are lost during dimension reduction. Thus, a DR model capable of holding key features is required to obtain a more accurate and efficient sequence analysis. Deep learning (DL) empowers finding the high-value bio-physicochemical parameters hidden within the sequences space resulting in improved sequence analysis and clustering due to their keen learning ability. DL methods have shown improvement in the clustering of bioimaging, biomedical text mining, and genomics sequences [10] due to their advanced featurization and DR ability. Antibody amino acid sequences analysis has been explored by ML [16, 13] and DL methods [22, 4]. Nevertheless, DL-based DR methods for multi-target specificity and cross-reactivity analysis of antibody sequences have not been explored yet. Here, we present a DL model which acts as a dimensionality reduction method and named it VAEResDR. Our VAEResDR is based on a Resnet-adopted convolutional Variational Autoencoder (VAE) we described previously [23]. We demonstrate that our scalable and generalizable VAEResDR reduces sequences' high-dimensional space to a two-dimensional visualization space while keeping the important features. The key extracted features are utilized to accurately analyze antibody CDR-H3 sequences and effectively improve clustering. These analyses may suggest whether CDR-H3 sequences have specificity or cross-reactivity to one or multiple targets.

2 Methodology

2.1 Datasets and Data Pre-processing

We utilized four large datasets on four different antigenic targets with 183,284 CDR-H3 sequences as training data to train the algorithms. Trastuzumab (Tras) dataset contains 38,839 sequences, %29 are classified as strong binders (Tras-Positive), and %71 are classified as weak binders (Tras-Negative) [18]. Ranibizumab (Rani) dataset consists of 67,769 sequences, %22 are classified as strong binders (Rani-Positive), and %78 are classified as weak binders (Rani-Negative) [12]. Yeast display scFv (Yeast) dataset contains 11,038 sequences [1]. Chicken Ovalbumin (OVA) contains 65,638 sequences [7].

We used the one-hot encoding method described previously [23] to convert amino acid sequences letters into an integer representation in a two-dimensional matrix. The amino acid CDR-H3 sequences' length varies between 8-20 among different databases. Therefore, we used ext-padding, a sparse padding, [12], [23] to have the sequences with the fixed lengths of 20.

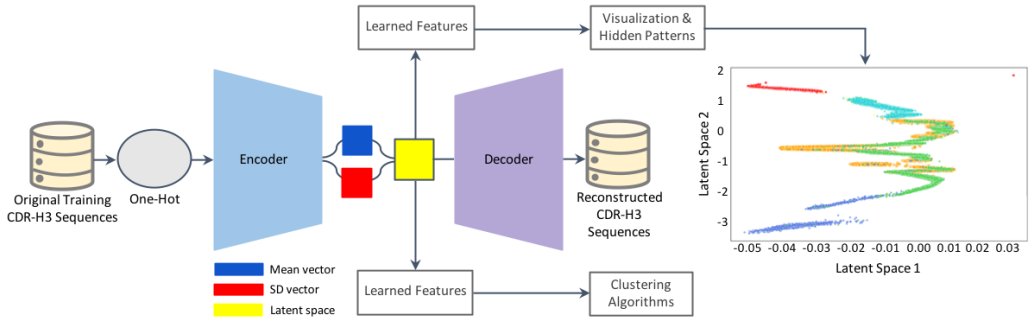

Figure 1: Overview of the VAEResDR model for the dimensionality reduction of antibody CDR-H3 sequences.

2.2 VAEResDR: Dimensionality Reduction (DR) Model

We present a Resnet-adopted convolutional VAE which we named VAEResDR. Our VAEResDR architecture illustrated in Figure 1 is based on our previously developed VAEResTL model [23] and consists of two parts: an encoder and a decoder. The encoder scales down high-dimensional input sequences into a two-dimensional feature space, while the decoder reconstructs the sequences from these two-dimensional features. To empower the encoder's ability to learn and extract the CDR-H3 sequences' key features more competently, we incorporated an optimized number of Resnet blocks into the encoder. Resnet blocks extend the depth of our neural network to expand the learning ability of our algorithm while preventing vanishing gradient problems. Moreover, to demonstrate the effectiveness of Resnet on improving the DR application of our proposed model, we compared VAEResDR with VAEDR, which is the VAEResDR without Resnet blocks.

To train our VAEResDR model, we used all four datasets containing 183,284 CDR-H3 sequences. Our optimized VAEResDR successfully extracts key features as it scales down the CDR-H3 sequences into two-dimension spaces. The extracted key features are also fed into clustering algorithms to find the similarity and diversity of CDR-H3 sequences and group them into different clusters resulting in improved clustering. The efficient feature extraction by our proposed method also enables us to map CDR-H3 sequences better and recognize the hidden patterns within the sequences space for a more efficient sequence analysis.

3 Experiments

3.1 Experimental Setup

To assess the impact of our DL model in DR tasks, improved visualizations, and hidden patterns finding, we compared three classic DR methods of PCA [29], t-SNE [27], UMAP [19] and two previously studied VAE [2], and scCCESS (Autoencoder-based) models [6] as baseline methods with our VAEDR and VAEResDR models. We executed two series of experiments. We utilized the Rani, Tras, and Yeast databases in the first series of experiments. In this experiment, we expected our proposed deep learning DR model to visually recognize CDR-H3 sequences in 5 groups of Rani-Positive, Rani-Negative, Tras-Positive, Tras-Negative, and Yeast. Our initial

biophysical and ML analysis showed high similarity between OVA and Rani CDR-H3 sequences. Therefore, we included the OVA CDR-H3 sequences in the second series of experiments to better illustrate our VAEResDR's performance in DR and biophysical feature extraction within the hidden patterns. In this experiment, we expected our DR model to visually illustrate CDR-H3 sequences in 6 groups of Rani-Positive, Rani-Negative, Tras-Positive, Tras-Negative, Yeast, and OVA and recognize whether there are CDR-H3 sequences with similar biophysical properties between different targets.

For benchmarking the DR performance and its effectiveness in clustering, we utilized two clustering algorithms of K-means [15] and GMM [30]. The VAEResDR reported results obtained on a P100 GPU with 25GB memory and a run time of 5h.

3.2 Experimental Metrics

We used visualization plots and clustering quality metrics to evaluate the impact of our VAEResDR model on antibody amino acid CDR-H3 sequences' DR and clustering as compared to baseline methods. We further measured biophysical metrics and ML visualization heatmaps to better demonstrate the biological validity of hidden patterns. The clustering quality metrics comprise five metrics:*Mutual Information (MI)* [25], *Adjusted Rank Index (ARI)* [9], *Homogeneity Score (HS)* [28], *Completeness Score (CS)* [28], and *V-measure Score (VMS)* [28]. The MI, ARI, HS, CS, and VMS values are confined in [0,1], equaling 1 when the two clusters are identical and 0 when they are independent. The larger the values (closer to 1), the better is the performance.

We used biophysical metrics of Charge, Hydrophobicity (H), and Isoelectric Point (ISO) to screen CDR-H3 amino acid sequences as they are highlighted to be the key features in antibody specificity. We used the bio and modLAMP library [20] to calculate the Charge, the H, and the ISO [24]. We used the Pairwise sequence similarity method of Needleman-Wunch (NW) [21] to evaluate the similarity of sequences. The higher the NW score, the more similar are the two sequences. We further estimated the sequences diversity by measuring the number of shared n-grams for different values of n between CDR-H3 sequences sharing the same latent space compared to CDR-H3 sequences in different latent spaces, referred to as Sn [3]. Therefore, a value of $S_n^{Condition1}/S_n^{Condition2} < 1$ indicates more diversity of sequences in latent space Condition 1 at a particular n than Condition 2. We refer to Condition 1 for sequences in different latent spaces and Condition 2 for sequences in the same latent space.

4 Results and Discussion

4.1 Dimensionality Reduction and Visualization Performance

We present the two-dimensional visualizations for five clusters (series one experiment) (Figure 2A-2G) and that of six clusters (series two experiment) (Figure 2H-2N) by our VAEResDR model compared to the baseline models. In the series one experiments, the visualization plots demonstrate moderately weak separation in PCA (Figure 2A and 2H), t-SNE (Figure 2B and 2I), UMAP (2C and 2J), VAE (Figure 2D and 2K), and scCCESS (AE) (Figure 2E and 2L). Albeit our VAEDR model could recognize the three target datasets of Yeast, Tras, and Rani in different latent spaces, it could not recognize the two classes of Rani-Positive and Rani-Negative sequences and Tras-positive and Tras-Negative sequences (2F). In the series two experiments, when we included the OVA sequences in the input datasets, for ML methods of PCA, t-SNE,

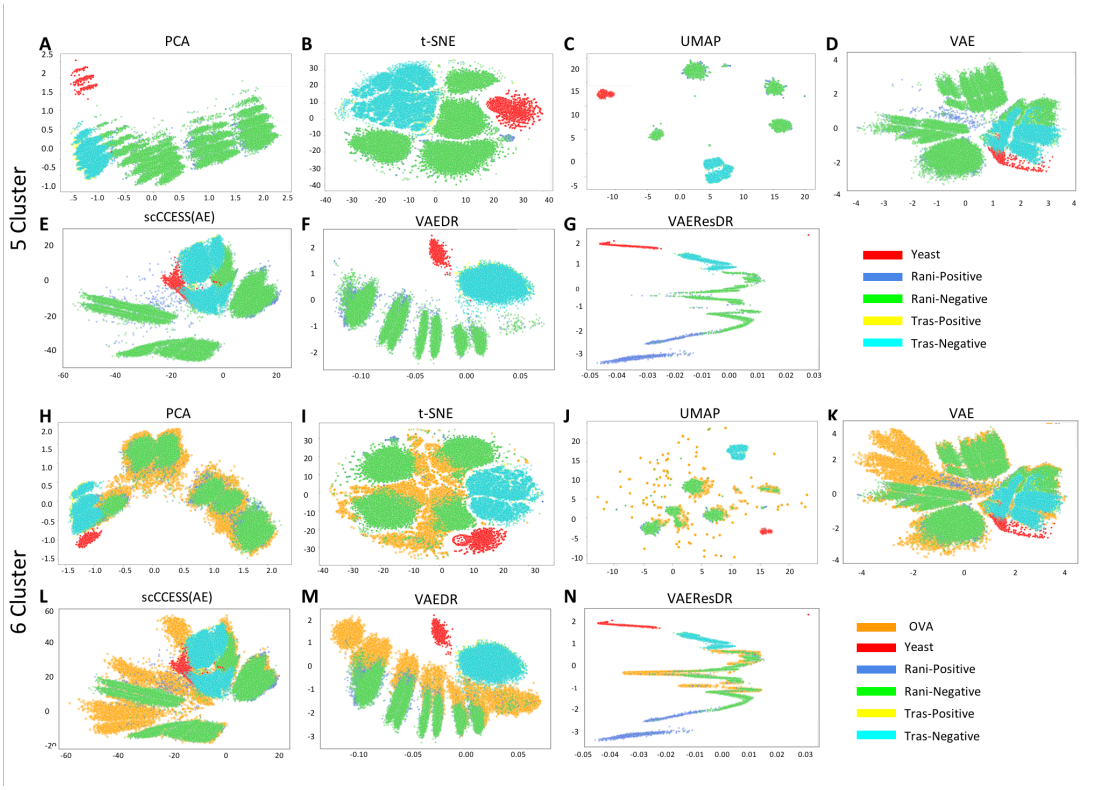

Figure 2: Visualization: 2-dimensional representation plots. Figure A to G representing PCA, t-SNE, UMPA, VAE, scCCESS (AE), VAEDR and VAEResDR for three targets in 5 clusters consecutively. Figure H to N representing PCA, t-SNE, UMPA, VAE, scCCESS (AE), VAEDR and VAEResDR for four targets in 6 clusters consecutively.

Table 1: Clustering Results: Quality metrics comparison for K-Means and GMM on CDR-H3 sequences

Method	Clustering Algorithm	5 Cluster					6 Cluster				
		ARI	MI	HS	CS	VMS	ARI	MI	HS	CS	VMS
PCA	**K-Means**	0.316	0.752	0.536	0.502	0.518	0.142	0.504	0.324	0.297	0.310
	GMM	0.344	0.786	0.560	0.521	0.540	0.124	0.485	0.312	0.310	0.311
t-SNE	**K-Means**	0.253	0.602	0.429	0.380	0.403	0.227	0.649	0.363	0.363	0.388
	GMM	0.300	0.666	0.475	0.422	0.447	0.239	0.685	0.441	0.384	0.411
UMAP	**K-Means**	0.343	0.780	0.556	0.501	0.527	0.200	0.634	0.408	0.366	0.386
	GMM	0.342	0.774	0.552	0.498	0.523	0.269	0.767	0.494	0.442	0.467
VAE	**K-Means**	0.283	0.489	0.316	0.499	0.387	0.104	0.424	0.255	0.273	0.263
	GMM	0.368	0.469	0.338	0.478	0.396	0.092	0.409	0.277	0.263	0.270
scCCESS (AE)	**K-Means**	0.285	0.468	0.295	0.477	0.365	0.128	0.458	0.261	0.295	0.277
	GMM	0.297	0.495	0.313	0.505	0.386	0.111	0.400	0.237	0.257	0.247
VAEDR	**K-Means**	0.449	0.681	0.443	0.695	0.541	0.328	0.840	0.484	0.541	0.511
	GMM	0.457	0.684	0.449	0.698	0.546	0.306	0.794	0.468	0.511	0.506
VAEResDR	**K-Means**	0.457	0.686	0.450	0.700	0.548	0.306	0.835	0.485	0.538	0.510
	GMM	**0.530**	**0.915**	**0.683**	**0.652**	**0.667**	**0.426**	**0.949**	**0.552**	**0.611**	**0.580**

UMAP, and DL methods of VAE, and scCCESS (AE) Figure 2H-2L, OVA (orange dots) sequences spread in all latent spaces. However, although VAEDR (Figure 2M) was able to

well-separate the OVA sequences (orange dots) from Yeast and Tras sequences, it was unable to separate OVA sequences from Rani sequences.

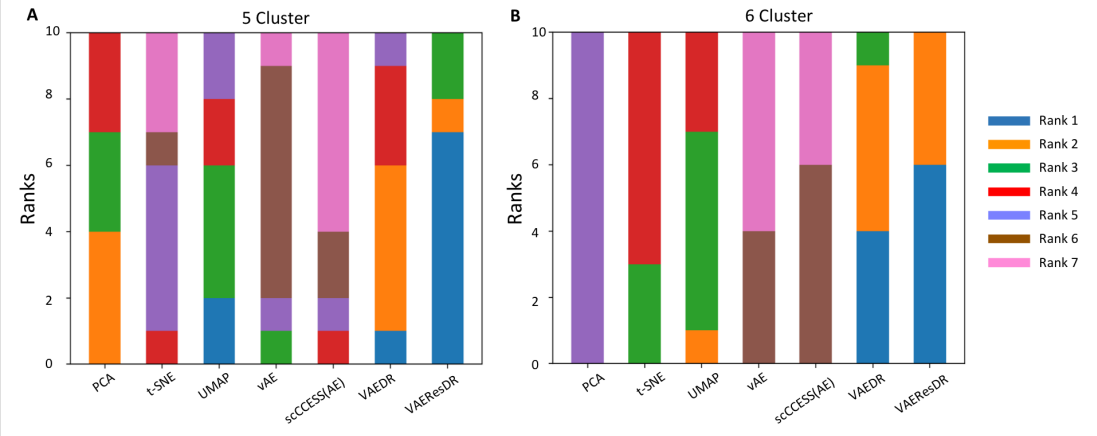

Figure 3: Bar chart clustering comparison to demonstrate the DR methods' performance. The best performance is ranked 1 and the worst performance is ranked 7. The color codes for rank 1, 2, 3, 4, 5, 6, and 7 are blue, orange, green, red, purple, brown, and pink consecutively.

In VAEResDR visualization plots (Figure 2G and 2N), we observed moderately well-separated patterns between 2 target datasets of Rani-Positive, Rani-Negative, and Yeast in both 5 and 6 clusters except the Tras-Positive and Tras-Negative sequences. When including the OVA sequences in the input datasets, OVA (orange dots) sequences share similar latent space with a portion of Rani-Negative and are well separated from Rani-Positive, Yeast, and Tras sequences (Figure 2N). Comparing the VAEDR performance with that of VAEResDR, we demonstrated that Resnet could improve the learning ability of our VAEResDR by separating the two classes of Rani-Positive and Rani-Negative, and could recognize the overlapping patterns of OVA with Rani-Negative sequences. With well-separated groups and valuable biophysical parameters recognized within the two-dimensional sequences space by our VAEResDR model, Tras-Positive and Tras-Negative sequences are not yet separated. These observations might be associated with the small range of weak and strong binders of Tras CDR-H3 sequences. The slight range difference between weak and strong binders might be due to the high accuracy of the CRISPR-Cas9 gene-editing method, a wet-lab antibody production technique used to produce Tras CDR-H3 sequences [18].

4.2 Benchmark Analysis for Improved Clustering

We tested K-Means and GMM on the extracted features by our VAEDR and VAEResDR models compared to the baseline models. We reported quality metrics of ARI, MI, HS, CS, and VMS for five clusters (series one experiment) and six clusters (series two experiments) (Table 1), in which our VAEResDR model with GMM shows its best performance. In particular, according to the quantitative analysis, we observed MI of 0.915 and CS of 0.652 for five cluster experiments and MI of 0.949 and CS 0.611 for six cluster experiments. When OVA sequences were added to the input datasets, the quantitative values for both K-Means and GMM degraded slightly due to the existing OVA sequences overlaps. Nevertheless, our VAEResDR outperformed PCA,

t-SNE, UMAP, and DL methods of VAE, scCCESS (AE), and our VAEDR model (Table 1) in combination with clustering methods in both five cluster and six cluster experiments.

To better illustrate the performance of DR methods in improving clustering analysis, we further reported the ranks statistics of the compared methods bar chart according to the quality metrics values (Figure 3A, 5 clusters and 3B, 6 clusters). VAEResDR ranked 1, 2, and 3 in 5 cluster experiments (Figure 3A) and ranked 1 and 2 in 6 cluster experiments (Figure 3B), demonstrating its best performance. Our VAEDR has a second place for its performance as it contains a distribution of ranks 1, 2, and 3 in 6 cluster experiments (Figure 3B). The scCCESS (AE) and VAE showed the lowest performance, with a distribution performance rank of 6 and 7 in the 6 cluster experiments (Figure 3B). The overall bar chart performance rank plots confirmed clustering performance degraded when OVA sequences were included in the input datasets. However, our VAEResDR model alongside GMM effectively showed high performance with improved antibody CDR-H3 sequences clustering. In our future work, testing other recently developed clustering approaches could be explored to find the impact of our proposed VAEResDR method in improving the clustering's performance.

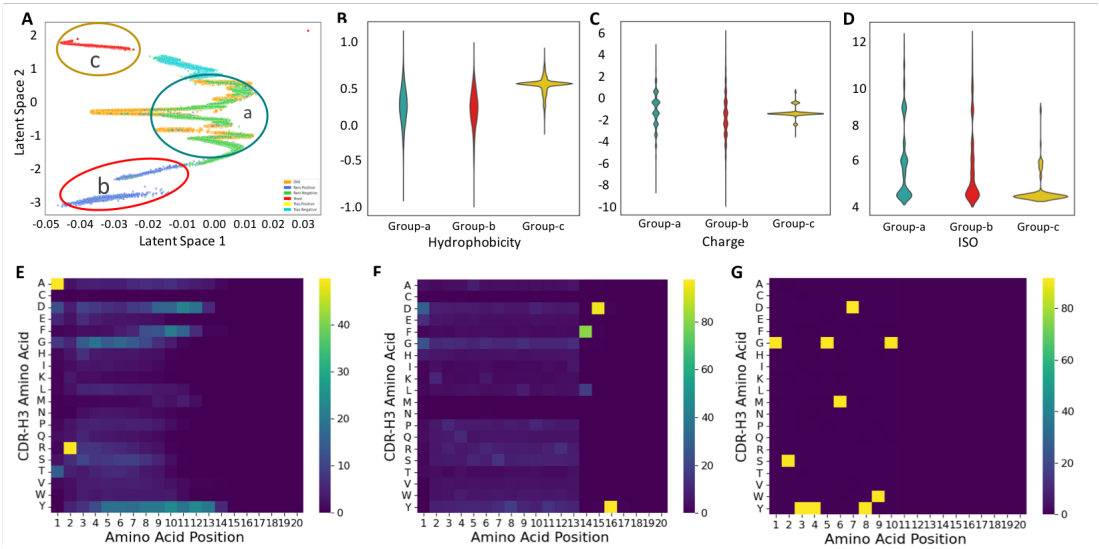

Figure 4: Hidden Patterns Analysis: Biophysical and ML visualization Heatmap analysis. (A) 2-dimensional visualization by VAEResDR method for four targets, 6 clusters. CDR-H3 Sequences sharing the same latent space are marked as group (a) and CDR-H3 sequences in different latent spaces are marked as group (b) and (c). (B) biophysical metrics of hydrophobicity (H) (C) net-charge, and (D) ISO. (E), (F), and (G) are heatmap plots for group (a), (b), and (c) sequentially.

4.3 Hidden Patterns Analysis

We quantitatively assessed whether the learned representation by our VAEResDR could prompt the biological characteristics of the CDR-H3 sequences on their specificity and cross-reactivity. We used the two-dimensional visualization plot provided by our VAResDR model for further analysis (Figure 4A). We selected CDR-H3 sequences from different regions of the two-dimensional visualization plot, which we marked as group (a), group (b), and group (c) (Figure

4A). CDR-H3 sequences in group (a) and group (b) have H (Figure 4B), Charge (Figure 4C), and ISO (Figure 4D) values closer to each other that differentiate them from the group (c). We also found from the ML visualization heatmaps (Figure 4E-4G) that sequences in group (a) and group (b) have higher similarity and the greatest difference with the sequences in the group (c). We further analyzed the diversity of CDR-H3 sequences in the three mentioned groups in terms of their shared n-grams (Sn). $S_n^{group(b\&c)}/S_n^{group(a)} < 1$, for n< 2, 2-gram, 3-gram, 4-gram, and 5-gram are 0.76, 0.64, 0.54 and 0.25 respectively. These results imply strong diversity between CDR-H3 sequences in group (a) with group (b) and (c). The average pairwise sequence similarity of NW for randomly selected sequences within group (a) was 44.2. The NW values for group (b) and group (c) were 32.5. These results may suggest a higher biological similarity between sequences within group (a). The higher similarity in ISO, H, and Charge confirmed by ML heatmap visualization plots for sequences in group (a) and group (b) demonstrates that the CDR-H3 sequences in group (a) and (b) may have specificity to multiple targets. Moreover, the higher NW score for sequences within-group (a) demonstrates that the CDR-H3 sequences within group (a) can have specificity to multiple targets. The color codes indicate group (a) is marked where OVA (Figure 4A, orange dots) and Rani-Negative (Figure 4A, green dots) are present. Group (b) is marked where Rani-Positive (Figure 4A, blue dots) occupies the latent space and group (c) is marked where Yeast sequences (Figure 4A, red dots) are accumulated. A close-up of the CDR-H3 sequences in group (a) illustrates that Rani-Negative CDR-H3 sequences share the same latent space with OVA CDR-H3 sequences. These results may suggest that Rani-Negative sequences with weak binding to target Rani may have specificity to target OVA, and OVA sequences may have specificity to target Rani. The multi-target specificity of OVA sequences to Rani and OVA targets and Rani-Negative to OVA and Rani targets, as a case study, can expand the CDR-H3 sequences library for sequence screening in therapeutic developments of either targets. Our sequences analysis by VAEResDR could also demonstrate that OVA sequences may have cross-reactivity with Rani target and Rani-Negative CDR-H3 sequences may have cross-reactivity with OVA target. With these analyses, we could prevent any false-positive or false-negative results when performing sequence screening for either of the targets. Our VAEResDR can expand the library of hit candidates to design the next generation of antibody-based therapeutics in-silico, especially for bispecific and trispecific antibody-based therapeutics.

Conclusion and Outlook

In this work, we developed a deep learning-based dimensionality reduction method, VAEResDR, for antibody amino acid CDR-H3 sequence analysis. Our quantitative analysis demonstrates that VAEResDR achieved superior DR performance in the two designed experimental series and has significant generalizability for different datasets with diverse structure in the original and high-dimensional spaces. Our VAEResDR further demonstrate effective improvement in CDR-H3 sequences' clustering. Our proposed model provides a readily usable tool for screening of CDR-H3 sequences' specificity and for identifying highly specific CDR-H3 sequences that bind to one or multiple targets. For our future work, we plan to expand the antibody analysis to other antibody/nanobody sequences' fractions as well as other targets. We further plan to implement our VAEResDR analysis to design bispecific and trispecific antibody therapeutics.

Acknowledgments

This wok was supported by MarWell Bio Inc. No external grants or funding contributed to the completion of this work. All present and future rights to any intellectual property arising from this work will be the sole property of MarWell Bio Inc.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Rhys M Adams, Thierry Mora, Aleksandra M Walczak, and Justin B Kinney. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife*, 5:e23156, 2016.

[2] Dongmei Ai, Yuduo Wang, Xiaoxin Li, and Hongfei Pan. Colorectal cancer prediction based on weighted gene co-expression network analysis and variational auto-encoder. *Biomolecules*, 10(9):1207, 2020.

[3] Payel Das, Kahini Wadhawan, Oscar Chang, Tom Sercu, Cicero Dos Santos, Matthew Riemer, Vijil Chenthamarakshan, Inkit Padhi, and Aleksandra Mojsilovic. Pepcvae: Semi-supervised targeted design of antimicrobial peptide sequences. *arXiv preprint arXiv:1810.07743*, 2018.

[4] Youyi Fong and Jun Xu. Forward stepwise deep autoencoder-based monotone nonlinear dimensionality reduction methods. *Journal of Computational and Graphical Statistics*, pages 1–10, 2021.

[5] Norbert Furtmann, Marion Schneider, Nadja Spindler, Bjoern Steinmann, Ziyu Li, Ingo Focken, Joachim Meyer, Dilyana Dimova, Katja Kroll, Wulf Dirk Leuschner, et al. An end-to-end automated platform process for high-throughput engineering of next-generation multi-specific antibody therapeutics. In *Mabs*, volume 13, page 1955433. Taylor & Francis, 2021.

[6] Thomas A Geddes, Taiyun Kim, Lihao Nan, James G Burchfield, Jean YH Yang, Dacheng Tao, and Pengyi Yang. Autoencoder-based cluster ensembles for single-cell rna-seq data analysis. *BMC bioinformatics*, 20(19):1–11, 2019.

[7] Leonard D Goldstein, Ying-Jiun J Chen, Jia Wu, Subhra Chaudhuri, Yi-Chun Hsiao, Kellen Schneider, Kam Hon Hoi, Zhonghua Lin, Steve Guerrero, Bijay S Jaiswal, et al. Massively parallel single-cell b-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. *Communications biology*, 2(1):1–10, 2019.

[8] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[9] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[10] Md Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, and Stefan Decker. Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22(1):393–415, 2021.

[11] David LaFleur, Donara Abramyan, Palanisamy Kanakaraj, Rodger Smith, Rutul Shah, Geping Wang, Xiao-Tao Yao, Spandana Kankanala, Ernest Boyd, Liubov Zaritskaya, et al. Monoclonal antibody therapeutics with up to five specificities: functional enhancement through fusion of target-specific peptides. In *MAbs*, volume 5, pages 208–218. Taylor & Francis, 2013.

[12] Ge Liu, Haoyang Zeng, Jonas Mueller, Brandon Carter, Ziheng Wang, Jonas Schilz, Geraldine Horny, Michael E Birnbaum, Stefan Ewert, and David K Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36(7):2126–2133, 2020.

[13] Hao Lou and Michael J Hageman. Machine learning attempts for predicting human subcutaneous bioavailability of monoclonal antibodies. *Pharmaceutical Research*, 38(3):451–460, 2021.

[14] Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science*, 27(1):1–30, 2020.

[15] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[16] Rishikesh Magar, Prakarsh Yadav, and Amir Barati Farimani. Potential neutralizing antibodies discovered for novel corona virus using machine learning. *Scientific reports*, 11(1):1–11, 2021.

[17] Rosalind B Marimont and Marvin B Shapiro. Nearest neighbour searches and the curse of dimensionality. *IMA Journal of Applied Mathematics*, 24(1):59–70, 1979.

[18] Derek M Mason, Simon Friedensohn, Cédric R Weber, Christian Jordi, Bastian Wagner, Simon M Meng, Roy A Ehling, Lucia Bonati, Jan Dahinden, Pablo Gainza, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, 5(6):600–612, 2021.

[19] Leland McInnes, John Healy, and James Melville. Umap: uniform manifold approximation and projection for dimension reduction. 2020.

[20] Alex T Müller, Gisela Gabernet, Jan A Hiss, and Gisbert Schneider. modlamp: Python for antimicrobial peptides. *Bioinformatics*, 33(17):2753–2755, 2017.

[21] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[22] David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny Asher Bitton. Biophi: A platform for antibody design, humanization and humanness evaluation based on natural antibody repertoires and deep learning. *bioRxiv*, 2021.

[23] Khalilian Saeed., Moti Zahra., Baloochestani Arian., Hallaj Yeganeh., Chavosh AliReza., and Hemmatian Zahra. Vaerestl: A novel generative model for designing complementarity determining region of antibody for sars-cov-2. In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOINFORMATICS,*, pages 107–114. INSTICC, SciTePress, 2022.

[24] Vikas K Sharma, Thomas W Patapoff, Bruce Kabakoff, Satyan Pai, Eric Hilario, Boyan Zhang, Charlene Li, Oleg Borisov, Robert F Kelley, Ilya Chorny, et al. In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability. *Proceedings of the National Academy of Sciences*, 111(52):18601–18606, 2014.

[25] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[26] Yuko Tsuchiya and Kenji Mizuguchi. The diversity of h 3 loops determines the antigen-binding tendencies of antibody cdr loops. *Protein Science*, 25(4):815–825, 2016.

[27] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[28] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

[29] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[30] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 28–31. IEEE, 2004.