Supplementary: Deep Learning-Driven TCRβ Repertoire Analysis Enhances Diagnosis and Enables Mining of Immunological Biomarkers in Systemic Lupus Erythematosus

Tongfei Shen[1†], Yifei Sheng[1†], Wan Nie[1], Shuo Yang[1], Kaiqi Li[1], Ziwei Ma[1], Zhao Ling[1], Bowen Tan[1], Xikang Feng[2*], Miaozhe Huo[1*]

*Corresponding author(s). E-mail(s): fxk@nwpu.edu.cn; miaozhhuo2-c@my.cityu.edu.hk;
Contributing authors: timshen2-c@my.cityu.edu.hk; yifesheng2-c@my.cityu.edu.hk; wannie2-c@my.cityu.edu.hk; syang58-c@my.cityu.edu.hk; kaiqili2-c@my.cityu.edu.hk; ziweima3-c@my.cityu.edu.hk; zhaoling2-c@my.cityu.edu.hk; bowentan3-c@my.cityu.edu.hk;
†These authors contributed equally to this work.

1 Supplemental Methods

1.1 Data collection

The dataset utilized in this investigation was procured from publicly accessible databases and comprises TCR sequences extracted from the Peripheral Blood Mononuclear Cells (PBMCs) of 439 healthy individuals (HIs) and 877 patients diagnosed with SLE) [1]. This data was employed for both the training and cross-validation phases of our framework.

To rigorously assess the predictive efficacy of the model on independent samples, we included 37 samples from patients with Juvenile Idiopathic Arthritis (JIA) and 33 samples from individuals with Autoimmune Arthritis (AutoA). Both JIA and AutoA are autoimmune disorders that share fundamental pathophysiological similarities with SLE [2, 3]. To establish negative controls, we selected three distinct cohorts of healthy individuals, all sourced from public datasets [4]. The TCR sequences for the independent testing cohort and the healthy controls were also derived from PBMCs. This methodology ensures a comprehensive and robust evaluation of our model's performance across varied clinical contexts.

1.2 Dataset construction

All the repertoires were split into training and validation sets at a ratio of 4:1. In the training dataset, we conducted a random selection to choose an equal number of individuals from both healthy donors and SLE patients. Within the TCR repertoire of each selected individual, each valid TCR CDR3 sequence underwent rigorous filtering based on the following criteria: (i) the sequence length must be between 10 and 24 amino acids; (ii) the sequence must consist only of standard amino acids; (iii) the sequence should start with cysteine (C) and end with phenylalanine (F); and (iv) the variable gene locus must be identifiable. Considering computational efficiency and the richness of information, we retained the 2,000 most frequent TCR clone types from each individual to represent their immunological profile.

Representative CDR3 sequences were categorized as positive samples from SLE patients (labeled "1") or negative samples from HIs (labeled "0"). Sequences that appeared in both patient and HI groups were classified as negative TCRs. After this classification, excess negative sequences were discarded to balance the quantity of positive and negative samples. The curated sequences were then shuffled to create a training set with a 1:1 ratio of positive to negative sequences, ensuring a randomized and duplicate-free dataset.

We then constructed two additional datasets, named CDR3-V gene and CDR3-V gene family dataset, by following the procedures used for the CDR3-only dataset, but with a modification: the datasets integrate V gene or V gene family information alongside the CDR3 amino acid sequences. During the calculation of clonotype frequencies and the deduplication process of TCR data, we considered both the V gene type and the CDR3 sequences, rather than focusing solely on the CDR3. Consequently, we obtained a training dataset that consists of a sequence set $S = \{s_1, s_2, s_3, \ldots, s_n\}$, a gene set $G = \{g_1, g_2, g_3, \ldots, g_n\}$, and the corresponding label set $T = \{t_1, t_2, t_3, \ldots, t_n\}$

which contains the TCR-level labels. The sequence set S, gene set G, and label set T each comprise n elements, ensuring precise one-to-one correspondence among them.

Similarly, we construct independent datasets JIA and AutoA datasets.

1.3 Model architecture

We trained the TCR classifier model using three datasets with different feature combinations: the first uses only amino acid sequences, the second combines these with V gene features, and the third integrates V gene family features. The model integrates a convolutional neural network (CNN) combined with long short-term memory (LSTM) layers and residual connections.

The model accepts two types of inputs: CDR3 amino acid sequences and V gene features (optional). Each input type is tokenized using respective dictionaries. An embedding layer transforms these sequences into fixed-length vectors of 24 dimensions. To standardize sequence lengths, a zero-padding method is applied, padding shorter sequences at the beginning. For the V gene features, which are categorized either by mutation type or V gene family, a linear layer converts these features into embeddings with a dimensionality of 20.

The CNN architecture comprises four convolutional blocks, each consisting of two convolutional layers followed by a max pooling layer. The convolutional layers use ReLU activation and are equipped with 256 filters each. The max pooling layers have a pool size of 5 and a stride of 1. The convolution operations are defined by:

$$c_{i,j} = \sum_{k=0}^{K-1} W_{k,j} \cdot s_{i+k} + b_j$$

where $W_{k,j}$ is the weight of the j-th filter, b_j is the bias, and K is the filter size. Residual connections allow each block's input to derive from both its predecessor and all prior blocks. Outputs from the convolutional layers are flattened and then passed to a fully connected layer with 16 neurons and Sigmoid activation.

The LSTM component includes three bidirectional layers with a dropout rate of 0.2 to mitigate overfitting.

Outputs from the CNN and LSTM feature extractors are concatenated, either with or without the V gene feature embedding, and then passed through a connected layer with Sigmoid activation for binary classification to fit the assigned TCR labels.

1.4 Model training and evaluation

The TCR classifiers were trained using the RMSprop optimizer, initiated with a learning rate of 0.0001. Binary cross-entropy was employed as the loss function, and the models achieved convergence after five epochs.

For model validation, we employed 5-fold cross-validation using 439 HI samples and 439 SLE patient samples, which were randomly selected from the total patient population. For each individual, the DeepTAPE classifier aggregated the results from the top 2,000 most frequent representative TCRs in an individual's repertoire to compute an autoimmune risk score (ARS), which indicates the probability of the individual having SLE. The ARS is calculated using the following expression:

$$\text{ARS} = \frac{\sum_{i=1}^{n} \text{TCR classifier}(\text{tcr}_i)}{n}. \tag{1}$$

Here, tcr_i denotes the i-th TCR in the individual's set of n representative TCRs, where n equals 2,000. The TCR classifier represents a well-trained classifier at the TCR level, which assesses each TCR for its association with SLE.

The threshold for ARS was determined using the training dataset. In the validation phase, individuals with ARS values exceeding this threshold were classified as having SLE, while those below were considered healthy.

The model's performance metrics, including accuracy, precision, recall, F1-score, and the area under the curve (AUC), were calculated using the following formulas:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{F1 Score} &= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{AUC} &= \frac{1}{P \times N} \sum_{i=1}^{P} \sum_{j=1}^{N} \mathbf{1}(p_i > n_j)
\end{aligned} \tag{2}$$

In these formulas, TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives, respectively. P and N are the counts of positive and negative individuals. Indices i and j refer to these individuals, with p_i denoting the predicted probability that the i-th individual is positive, and n_j the probability that the j-th individual is negative.

1.5 Baseline models

To establish the effectiveness of TCR classifier, we compared its performance against several benchmark classifiers, each also utilizing amino acid sequences and gene frequency information.

CNN-LSTM: Similar in structure and hyperparameters to DeepTAPE, but lacks residual connections.

CNN: Comprising four convolutional layers paired with max-pooling layers, this model utilizes a kernel size of 5.

Bi-LSTM: This model features three layers of bidirectional LSTM, each with a dropout rate of 0.2.

SimpleRNN: Composed of two layers of recurrent neural networks (RNN), with a dropout rate of 0.2.

The diagnostic performance of DeepTAPE was also evaluated in comparison to the Random Forest (RF) classifier, as proposed in a previous study [1]. The RF classifier

has demonstrated effectiveness in distinguishing between SLE patients and healthy individuals by analyzing variations in individual V gene frequencies. We implemented the RF model adhering to the parameters specified in the original publication to ensure a consistent and fair comparison.

1.6 Independent test on non-SLE autoimmune diseases

The model underwent independent testing using external datasets from other autoimmune diseases, including JIA and AutoA datasets. Negative samples used for independent testing came from the external HI repertoires [5–7]. Considering the potential differences in the V gene and V gene family distributions of TCRs from other diseases compared to SLE, a self-adaptive mechanism based on the Pearson Correlation Coefficient (PCC) was introduced. This mechanism determines whether to utilize a specific gene feature based on the ratio of the gene frequency distribution correlation coefficient between the target disease and SLE samples versus the correlation coefficient between HI samples. The standard frequency distributions are derived from the average frequency distributions of 10 samples each from HI and SLE patients. For a sample's gene frequency distribution f_{SA} and a standard frequency distribution f_{ST}, the Pearson correlation coefficient formula is given by:

$$\mathrm{cor}(f_{\mathrm{SA}}, f_{\mathrm{ST}}) = \frac{\sum_{i=1}^{n}(f_{\mathrm{SA},i} - \bar{f}_{\mathrm{SA}})(f_{\mathrm{ST},i} - \bar{f}_{\mathrm{ST}})}{\sqrt{\sum_{i=1}^{n}(f_{\mathrm{SA},i} - \bar{f}_{\mathrm{SA}})^2}\sqrt{\sum_{i=1}^{n}(f_{\mathrm{ST},i} - \bar{f}_{\mathrm{ST}})^2}} \tag{3}$$

Here, $\bar{f}_{\mathrm{SA}}$ and $\bar{f}_{\mathrm{ST}}$ represent the mean of f_{SA} and f_{ST} respectively, and the summation is performed over all n gene types. The Correlation Ratio (CR) is defined as the ratio of the correlation coefficient between f_{SA} and the standard SLE frequency distribution f_{SLE} to the correlation coefficient between f_{SA} and the standard HI frequency distribution f_{HI}:

$$\mathrm{CR} = \frac{\mathrm{cor}(f_{\mathrm{SA}}, f_{\mathrm{SLE}})}{\mathrm{cor}(f_{\mathrm{SA}}, f_{\mathrm{HI}})} \tag{4}$$

Here, f_{SA}, f_{SLE}, and f_{HI} represent the gene frequency distributions of the sample, standard SLE, and standard HI respectively. CR determines which frequency distribution, SLE or HI, is closer to the sample. Similarly, an individual's disease probability was also assessed though ARS.

2 Supplemental Results

2.1 Diagnostic efficacy of CDR3 amino acid sequences and gene frequencies in SLE

Previous investigations have established the feasibility of utilizing the gene information within TCR repertoire for SLE diagnosing. This was achieved through a Random Forest (RF) algorithm that effectively distinguished between SLE patients and healthy individuals based on V gene frequencies, yielding promising outcomes. Moreover, several effective models have been developed for diagnosing various diseases by leveraging

the amino acid sequences of the CDR3 in TCR's β chain. In light of these findings, we adopted a neural network approach to diagnose autoimmune diseases, including SLE, by integrating information from both amino acid sequences and gene frequencies.

When compared to other diagnostic models that utilize TCR β CDR3 amino acid sequences and gene frequency information (see Method), the DeepTAPE model, built on a CNN-LSTM architecture with residual connections, demonstrated superior diagnostic performance across five metrics. The minimal performance variance observed during the 5-fold cross-validation indicates the model's stability (Tab. 1). According to the confusion matrix, all three DeepTAPE input modes, which utilize different feature combinations, exhibited commendable diagnostic capabilities. Notably, the DeepTAPE input mode that additionally incorporated V gene and V gene family features displayed a lower misdiagnosis rate compared to the model utilizing only amino acid sequences, indicating enhanced accuracy. Most misdiagnoses across the three models were categorized as false negatives (Fig. 1F, G, H).

2.2 Assessing the cross-disease generalizability of DeepTADE in autoimmune disorders

The three input versions of the DeepTAPE model were initially trained using data derived from SLE patients and HIs. The performance of these models varies markedly depending on the feature combinations implemented. While they exhibit strong diagnostic capabilities for SLE, they also demonstrate some level of diagnostic proficiency for other autoimmune disorders. In this context, samples from juvenile idiopathic arthritis (JIA) and autoimmune arthritis (AutoA), alongside samples from healthy individuals (HIs), were utilized to establish an external independent test set.

Equal quantities of patient samples and HI samples were randomly selected and input into the model for prediction, with this procedure repeated five times to derive average results. The findings revealed that for JIA, the AUC of the DeepTAPE model employing exclusively amino acid sequences achieved a value of 86.98%. However, when gene frequency data was included, the average AUC for the DeepTAPE model that integrated amino acid sequences with V genes and those combined with V gene families decreased to 64.22% and 63.33%, respectively (shown in Fig. 2A). A similar trend was noted for AutoA, where the average AUC utilizing amino acid sequences was higher at 85.78%, compared to average AUCs of 74.89% and 77.62% for the models that incorporated amino acid sequences with V genes and V gene families, respectively (Fig. 2B and Tab. 2). Moreover, the confusion matrix results indicated that the diagnostic performance of the DeepTAPE model based solely on amino acid sequences was significantly superior for both JIA and AutoA (Fig. 2C, D).

Despite the DeepTAPE model being primarily trained on SLE data, it retains a degree of classification capability for both JIA and AutoA. This observation suggests that SLE and other autoimmune diseases, such as JIA and AutoA, may share common pathogenic mechanisms, given their categorization as autoimmune disorders. Consequently, analogous features within the CDR3 amino acid sequences of TCR's β chain may be relevant for diagnostic purposes. However, the addition of gene frequency data appeared to detract from diagnostic performance, indicating that the incorporation of gene frequency features might have introduced confusion in the diagnosis of JIA

and AutoA. Analytical assessments revealed significant discrepancies in the frequency distributions of V genes and V gene families between the TCR β CDR3 of JIA and AutoA patients and that of SLE patients, adversely impacting diagnostic accuracy (Fig. 2E, F).

To resolve this challenge, the DeepTAPE model incorporated a self-adaptive mechanism based on the Pearson correlation coefficient (PCC), which facilitates a comparison between the frequency distributions of V genes or V gene families in the independent test set and those of SLE patients and HIs. The results indicated that the correlation coefficients for the V gene or V gene family distributions between JIA patients and HIs were significantly greater than those between JIA patients and SLE patients ($p < 0.001$), a similar observation is noted for AutoA, which complicates effective diagnosis (Fig. 2G, H). The self-adaptive mechanism utilizing PCC enables the model to autonomously select the most appropriate feature combinations tailored to each specific disease. Thus, this mechanism is essential for the automatic determination of diagnostic feature compositions for each condition, as evidenced by the model's choice to rely solely on amino acid sequences for diagnosing JIA and AutoA.

3 Supplemental Tables

Table 1: Comparison of the SLE diagnosis performance using 5-fold cross-validation

Model	AUC	Accuracy	Precision	Recall	F1-score
Only V Gene Frequency					
VGene-RF[1]	96.11% ± 1.70%	88.97% ± 3.38%	91.23% ± 4.49%	86.41% ± 4.26%	88.68% ± 3.47%
CDR3 Amino Acid Sequence					
CNN-LSTM	97.46% ± 0.69%	92.82% ± 1.53%	91.21% ± 2.68%	93.87% ± 1.81%	**92.97% ± 1.43%**
CNN	95.65% ± 2.10%	90.39% ± 3.01%	90.71% ± 4.78%	90.26% ± 4.85%	90.37% ± 3.02%
Bi-LSTM	97.30% ± 0.80%	92.69% ± 1.40%	91.44% ± 3.77%	91.03% ± 5.59%	92.53% ± 1.67%
SimpleRNN	96.25% ± 1.64%	91.41% ± 1.73%	90.07% ± 3.63%	93.33% ± 4.29%	91.56% ± 1.74%
DeepTAPE	**97.52% ± 0.68%**	**92.82% ± 1.39%**	**91.84% ± 4.74%**	**94.36% ± 3.46%**	92.95% ± 1.21%
CDR3 Amino Acid Sequence + V Gene					
CNN-LSTM	97.43% ± 0.83%	92.95% ± 4.33%	92.30% ± 3.26%	93.85% ± 2.07%	93.02% ± 3.45%
CNN	94.36% ± 5.94%	88.85% ± 2.74%	88.86% ± 5.08%	89.49% ± 7.77%	88.86% ± 2.98%
Bi-LSTM	96.85% ± 2.43%	92.05% ± 9.02%	91.43% ± 2.93%	92.05% ± 3.92%	92.05% ± 9.40%
SimpleRNN	97.67% ± 0.89%	93.33% ± 2.06%	92.01% ± 4.09%	**95.28% ± 9.02%**	93.48% ± 8.45%
DeepTAPE	**97.70% ± 0.86%**	**93.56% ± 1.80%**	**92.38% ± 3.69%**	95.15% ± 1.78%	**93.69% ± 1.57%**
CDR3 Amino Acid Sequence + V Gene Family					
CNN-LSTM	97.85% ± 0.76%	93.46% ± 1.05%	92.82% ± 2.91%	94.36% ± 2.95%	93.52% ± 1.04%
CNN	94.16% ± 1.71%	88.46% ± 2.48%	89.32% ± 4.17%	87.69% ± 5.26%	88.35% ± 2.58%
Bi-LSTM	97.37% ± 1.19%	93.33% ± 2.01%	92.80% ± 3.48%	94.10% ± 2.14%	93.40% ± 1.87%
SimpleRNN	97.26% ± 0.83%	92.95% ± 1.63%	92.08% ± 3.13%	94.10% ± 1.46%	93.05% ± 1.52%
DeepTAPE	**97.99% ± 0.82%**	**93.97% ± 1.61%**	**93.70% ± 2.57%**	**94.36% ± 2.14%**	**94.00% ± 1.57%**

The table outlines the effectiveness of various predictive models in diagnosing SLE on the validation dataset. Bold values represent the best performance, while underlined values indicate the second-best performance within each metric across methods.

Table 2: Performance of the DeepTAPE model across different input modes on non-SLE autoimmune disease datasets.

Input modes	AUC	Accuracy	Precision	Recall	F1-score
AutoA					
DeepTAPE (AA Seqs + V Gene Families)	77.62% ± 1.25%	80.33% ± 0.67%	**93.44% ± 1.91%**	65.33% ± 2.67%	76.83% ± 1.26%
DeepTAPE (AA Seqs + V Genes)	74.89% ± 0.78%	76.00% ± 0.82%	90.86% ± 3.21%	58.00% ± 2.67%	70.70% ± 1.43%
DeepTAPE (Only AA Seqs)	**95.78% ± 0.19%**	**90.33% ± 1.94%**	92.18% ± 4.83%	**88.67% ± 5.42%**	**90.14% ± 2.04%**
JIA					
DeepTAPE (AA Seqs + V Gene Families)	63.33% ± 1.84%	62.00% ± 2.45%	71.47% ± 8.13%	44.67% ± 15.00%	52.50% ± 9.65%
DeepTAPE (AA Seqs + V Genes)	64.22% ± 1.26%	71.00% ± 1.33%	**86.40% ± 4.14%**	50.00% ± 0.00%	63.31% ± 1.08%
DeepTAPE (Only AA Seqs)	**86.98% ± 1.47%**	**82.67% ± 3.09%**	82.67% ± 2.08%	**82.67% ± 6.46%**	**82.55% ± 3.81%**

Bold values represent the best performance, while underlined values indicate the second-best performance within each metric across three encoding modes. AA seq, amino acids sequence.

4 Supplemental Figures

References

[1] Liu, X. *et al.* T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Annals of the rheumatic diseases* **78**, 1070–1078 (2019).

[2] Henderson, L. A. *et al.* Next-generation sequencing reveals restriction and clonotypic expansion of treg cells in juvenile idiopathic arthritis. *Arthritis & rheumatology* **68**, 1758–1768 (2016).

[3] Bonami, R. H. *et al.* Bruton's tyrosine kinase supports gut mucosal immunity and commensal microbiome recognition in autoimmune arthritis. *Frontiers in immunology* **13**, 748284 (2022).

[4] Wong, C. & Li, B. Autocat: automated cancer-associated tcrs discovery from tcr-seq data. *Bioinformatics* **38**, 589–591 (2022).

[5] Lee, B. *et al.* Distinct immune characteristics distinguish hereditary and idiopathic chronic pancreatitis. *The Journal of clinical investigation* **130**, 2705–2711 (2020).

[6] Ramien, C. *et al.* T cell repertoire dynamics during pregnancy in multiple sclerosis. *Cell reports* **29**, 810–815 (2019).

[7] Mitchell, A. M. *et al.* Temporal development of t cell receptor repertoires during childhood in health and disease. *JCI insight* **7** (2022).

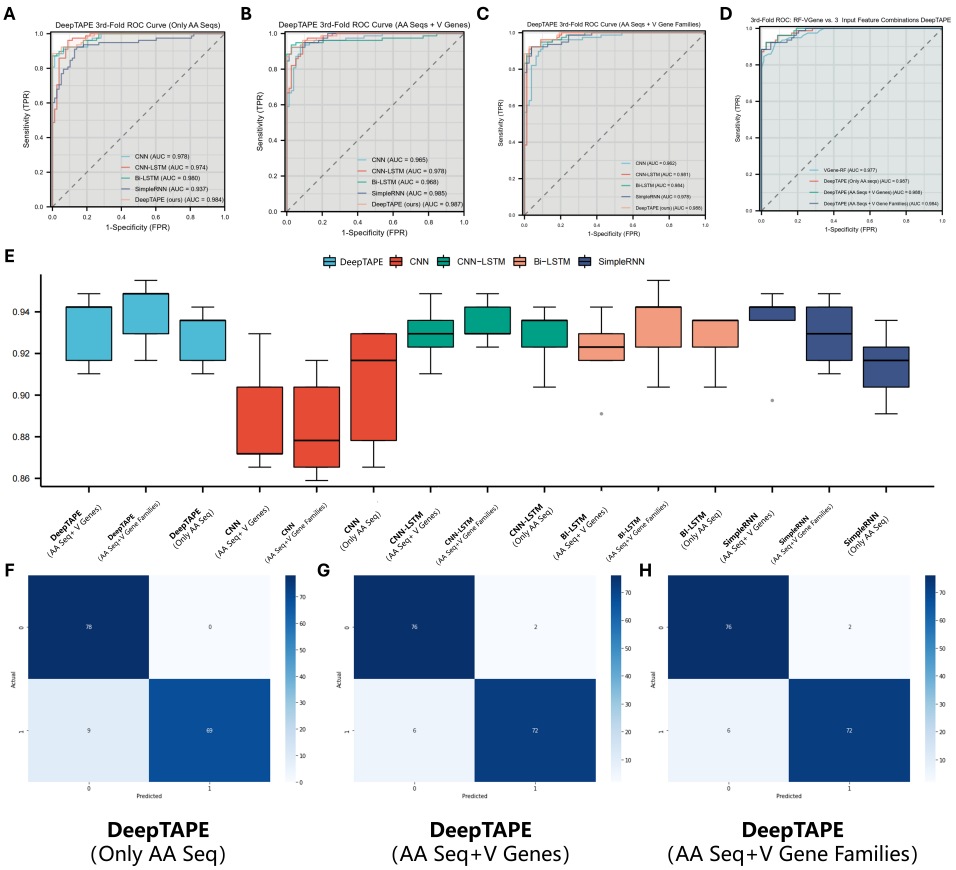

Fig. 1: **Diagnostic efficacy of TCRβ CDR3 amino acid sequences and gene frequencies in SLE.** (A) The ROC curve of DeepTAPE only input AA seqs, which performs better than other models with the same features. (B) The ROC curve of DeepTAPE using the combination of AA seq and V-genes, which performs better than other models with the same features. (C) The ROC curve of DeepTAPE using the combination of AA seq and V-gene families, which performs better than other models with the same features. (D) The ROC curves of the DeepTAPE on three input combinations, which use different features and perform better than the comparison RF-VGene. (E) Box plots comparing the performance of DeepTAPE and other models under various feature combinations.(F) Confusion matrix of DeepTAPE applied with only AA Seq features. (G) Confusion matrix of DeepTAPE applied with the combination of AA Seq and V Genes. (H) Confusion matrix of DeepTAPE applied with the combination of AA Seq and V Gene families. ROC, receiver operator curve; AA seq, amino acids sequence.

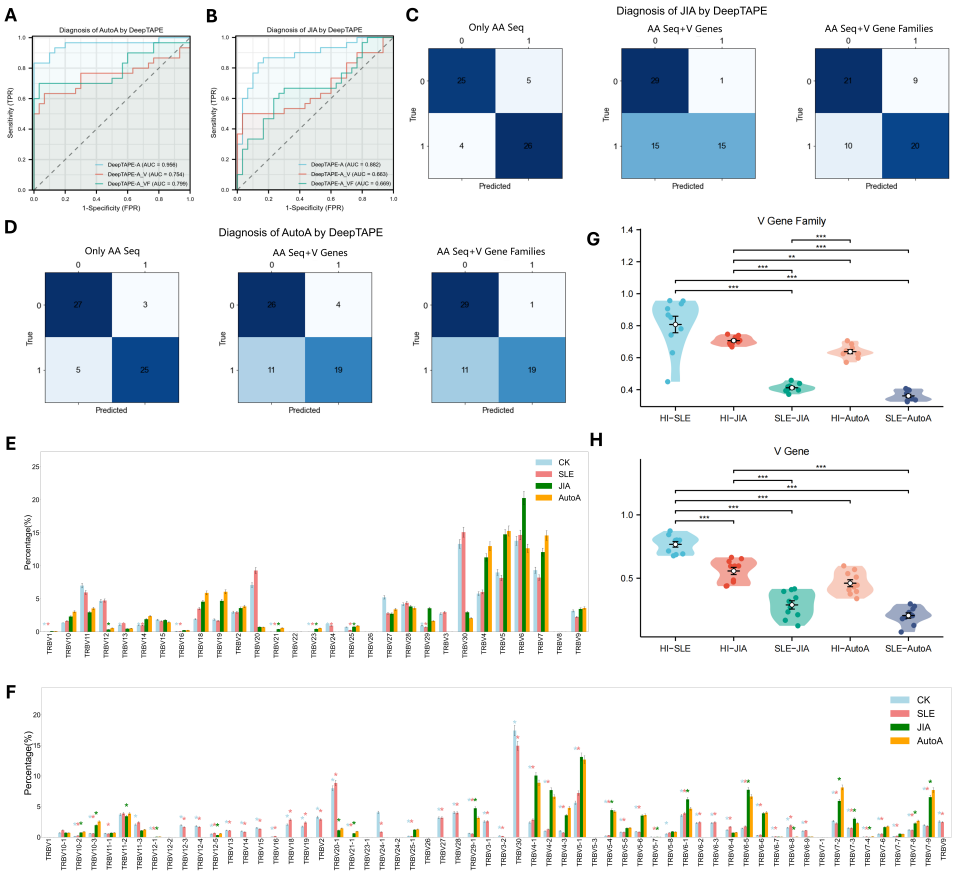

Fig. 2: Independent testing performance of TCRβ CDR3 Diagnostic Efficacy on exogenous autoimmune disease datasets and presentation of gene differences in autoimmune diseases. (A) ROC curves of the three versions of DeepTAPE for JIA diagnosis, with the version using only amino acid sequences performing the best. (B) ROC curves of the three versions of DeepTAPE for AutoA diagnosis, with the version using only amino acid sequences performing the best. (C) Heatmap of confusion matrices of the three versions of DeepTAPE for JIA diagnosis. (D) Heatmap of confusion matrices of the three versions of DeepTAPE for AutoA diagnosis. (E) Clustered bar chart showing significant frequency differences in the V gene family of CDR3 in TCR among SLE, HI, JIA, and AutoA samples. (F) Clustered bar chart showing significant frequency differences in the V gene of CDR3 in TCR among SLE, HI, JIA, and AutoA samples. (G) Violin plot comparing the Pearson correlation coefficients of V gene family frequencies between JIA and SLE, JIA and HI, and between AutoA and SLE, AutoA and HI. The correlations with HI are significantly stronger for both JIA and AutoA ($p < 0.001$). (H) Violin plot comparing the Pearson correlation coefficients of V gene frequencies between JIA and SLE, JIA and HI, and between AutoA and SLE, AutoA and HI. The correlations with HI are significantly stronger for both JIA and AutoA ($p < 0.001$).