1. Descriptive statistics and data preprocessing

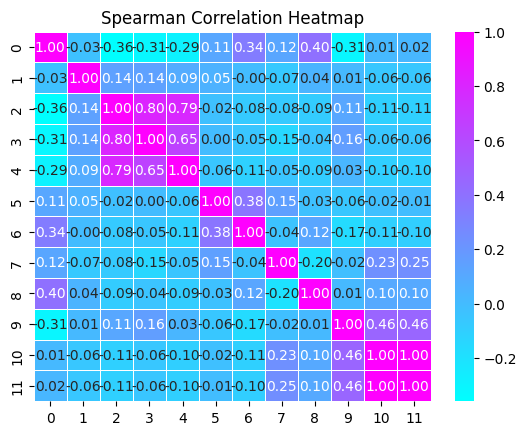
Based on the detailed explanation of VDJDB[1], we have summarized the meanings of each attribute. Due to the detailed explanations in the VDJDB documentation, the article will not repeat the meanings of each attribute. For the task of TCR specificity testing, the following 12 attributes are considered valuable: complex.id, cdr3, gene, v.segm, j.segm, antigen.epitope, antigen.gene,species, vdjdb.score, mhc.a, mhc.b, mhc.class.

We first need to detect missing values in the database, as shown in Table 1.1:

**Table 1.** Missing value detection

|  |  |  |
| --- | --- | --- |
| Index | Name | Missing values |
| 0 | complex.id | 0 |
| 1 | cdr3 | 0 |
| 2 | gene | 0 |
| 3 | v.segm | 0 |
| 4 | j.segm | 0 |
| 5 | antigen.epitope | 0 |
| 6 | antigen.gene | 0 |
| 7 | species | 0 |
| 8 | vdjdb.score | 0 |
| 9 | mhc.a | 0 |
| 10 | mhc.b | 0 |
| 11 | mhc.class | 0 |

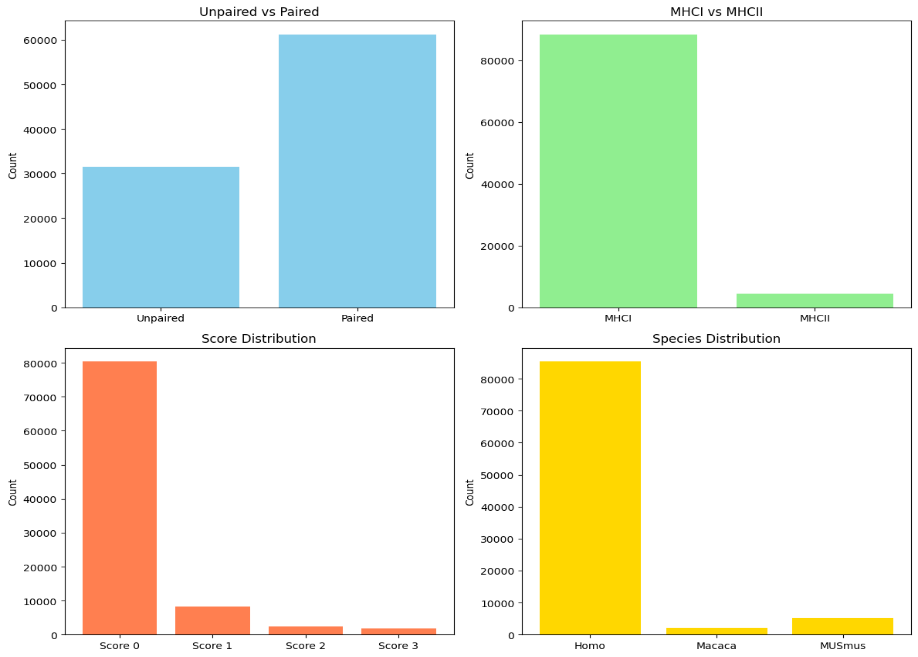
From Table 1.1, it can be seen that there are no missing values in the database for the attributes we need. Next, we conduct Spearman correlation analysis on these 12 attributes, and the results of the correlation analysis can help us screen for attributes with collinearity, thereby reducing the number of data attributes that the model needs to use. Our Spearman correlation analysis results are shown in Figure 1:



**Fig. 1.** Correlation Analysis

From the graph, it can be seen that mhc.class and mhc.b are completely correlated, and j.segm, v.segm, and gene are highly correlated. So in order to avoid the problem of high collinearity in the model, we need to avoid using highly correlated data simultaneously.

Next, we counted the data categories of some attributes as shown in the Figure 2:



**Fig. 2.** Data categories for certain attributes

Lu Tianshi's team[2] also focused on the data in the above figure when using deep learning to predict TCR specificity. When processing the data, they filtered out paired data that only selected MHC category as MHCI and belonged to human species. Rudolph, M. G [3] also pointed out in their article that the interaction process between TCR and MHC is different between MHCI and MHCII. Kim, S. M. [4] emphasized the importance of paired data in the article, pointing out that α-and β-chains play a significant role in the function of T cells.

Based on the above analysis, our preprocessing steps are as follows:

Step 1, retain only the complex.id, cdr3, gene, antigen.epitope, species, vdjdb.score, and mhc.class attributes.

Step 2, filter out data with mhc.class as MHCI.

Step 3, filter out paired data with complex.id not 0.

Step 4, filter out the data with species as HomoSapien.

Step 5, filter out data with vdjdb.score not being 0.

Step 6, rearrange the paired datapoint and merge them into the same datapoint(cdr3 with TRA and cdr3 with TRB).

1. References
2. Antigenomics. (2023, June). vdjdb-db. GitHub. <https://github.com/antigenomics/vdjdb-db/blob/2023-06-01/README.md>.
3. Lu, T., Zhang, Z., Zhu, J., Wang, Y., Jiang, P., Xiao, X., ... & Wang, T. (2021). Deep learning-based prediction of the T cell receptor–antigen binding specificity. Nature machine intelligence, 3(10), 864-875.
4. Rudolph, M. G., & Wilson, I. A. (2002). The specificity of TCR/pMHC interaction. Current opinion in immunology, 14(1), 52-65.
5. Kim, S. M., Bhonsle, L., Besgen, P., Nickel, J., Backes, A., Held, K., ... & Prinz, J. C. (2012). Analysis of the paired TCR α-and β-chains of single human T cells. PloS one, 7(5), e37338.