www.editada.org

*International Journal of Combinatorial Optimization Problems and Informatics, 17(1), Jan-April 2026, 196-214. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2026.v17i1.1030*

# Preprocessing of amino acid chains of antibody structure for machine learning analysis

*Manuel Erazo-Valadez*[1], *Yasmín Hernández*[1], *Elizabeth Ernestina Godoy-Lozano*[2], *Javier Ortiz-Hernández*[1], *Juan Téllez-Sosa*[2], *Juan Jose Flores-Sedano*[1], *Alejandra Cuevas-Chavez*[1]
[1] TecNM, Centro Nacional de Investigación y Desarrollo Tecnológico, México
[2] Centro de Investigación Sobre Enfermedades Infecciosas, Instituto Nacional de Salud Pública, México
{d20ce059, yasmin.hp, javier.oh, d20ce084, d18ce074}@cenidet.tecnm.mx, {elizabeth.godoy, jmtellez}@insp.mx

**Abstract.** Antibody classification represents a task of growing importance in bioinformatics. In recent years, the identification of antibodies capable of recognising and neutralising SARS-CoV-2 has become a central focus in immunological research and bioinformatics. Antibody representation presents several challenges, as antibody structure and function are highly variable, which complicates the development of a universal classification framework. Antibodies are composed of heavy and light chains that contain hypervariable complementarity-determining regions, which define their specificity. These structural variations create substantial challenges for sequence alignment, feature extraction, and classification. In this research, three methods for representing amino acid sequences were compared: TF–IDF, Atchley Factors, and ProtVec. These representations were evaluated using decision trees, logistic regression, and support vector machines. A separate dataset was generated for each representation. The results suggest that the representation based on Atchley Factors achieved comparatively stronger performance in the task of antibody classification.

**Keywords:** Antibody classification, SARS-CoV-2, amino acid sequence representation, Atchley Factors, TF-IDF, ProtVec

Article Info
*Received March 23, 2025*
*Accepted June 25, 2025*

## 1 Introduction

SARS-CoV-2 is a novel strain of coronavirus that was first detected in Wuhan, China, in 2019 and has since spread to approximately 200 countries (Yadav et al., 2022). The outbreak of this strain was subsequently labelled a global pandemic by the World Health Organization. In the current context, the scientific community has witnessed the contribution of bioinformatics tools and computational approaches, which have provided technical support and accelerated alternatives not only for identifying the characteristics and potential presence of the coronavirus, but also for supporting strategies to address COVID-19 infections (Pulendran & Davis, 2020). Bioinformatics tools have contributed to the identification of viral genome data, protein–peptide prediction, protein–protein docking, and the identification of antigenic epitopes and antibody structures relevant to vaccine development against SARS-CoV-2 (Yadav et al., 2022).

Research on the structural roles of heavy and light antibody chains is essential in immunology, particularly for the development of monoclonal antibody-based therapeutics (Pulendran & Davis, 2020). Antibodies are fundamental components of the adaptive immune system and possess the ability to identify and neutralise pathogens. Effective antigen binding requires the interaction between heavy and light chains. Advances in high-throughput sequencing technologies and computational biology have substantially increased the volume of available antibody-related data. However, most large-scale sequencing approaches analyse heavy and light chains independently, which may reduce the precision with which interactions between the two chains are characterised (Pulendran & Davis, 2020).

Machine learning provides a framework for analysing interactions between antibody chains, as it can process large datasets and identify complex structural patterns (Greiff et al., 2020). Nevertheless, representing amino acid chains presents several challenges, as inappropriate representations may lead to information loss and limit predictive accuracy. Antibody chains exhibit substantial variability in length, composition, and physicochemical properties, which complicates their modelling. Furthermore, it is necessary to capture both local interactions between neighbouring amino acids and long-range dependencies along the entire chain. Consequently, selecting an appropriate representation method constitutes a critical step in the computational study of antibody structure (Greiff et al., 2020). Computational analyses of antibodies face multiple challenges, among which sequence representation is particularly prominent. This difficulty arises from the high variability in amino acid sequence length, especially within the complementarity-determining regions (CDRs), which play a central role in antigen binding. Traditional representation approaches, such as one-hot encoding, have been explored; however, these methods produce high-dimensional representations and do not capture relationships between amino acids within the sequence. In contrast, approaches based on word embeddings or physicochemical properties offer lower-dimensional representations with uniform length, which may contribute to improved performance of machine learning models (Sapoval et al., 2022).

This article examines the principal challenges associated with representing amino acid sequences that form antibody structures. Three representation techniques are evaluated: TF–IDF, Atchley Factors, and ProtVec. Their expressiveness is assessed through classification performance, distinguishing antibodies that recognise SARS-CoV-2 from those that recognise other viruses. Decision trees, logistic regression, and support vector machines are employed as classification algorithms. Structural data for antibodies are obtained from specialised repositories, namely CoV-AbDab (Raybould et al., 2021) and OAS (Olsen et al., 2022).

The results indicate that the Atchley factor-based representation performs more favourably in the classification models when compared with TF–IDF and ProtVec representations. This finding suggests that incorporating physicochemical properties of amino acids constitutes a suitable approach for antibody classification. Among the evaluated models, the decision tree classifier demonstrated comparatively stronger performance when combined with this representation.

The main contribution of this study lies in the comparative evaluation of three encoding strategies—TF–IDF, Atchley Factors, and ProtVec—for representing antibody amino acid sequences in machine learning classification tasks. By examining their performance across three classifiers (decision trees, logistic regression, and support vector machines), this work offers insight into how representation choice influences model performance. The research seeks to address the following question: which representation method provides the most appropriate balance between dimensionality reduction and classification performance for identifying SARS-CoV-2-specific antibodies?

## 2 Background

The immune system protects the organism against harmful pathogens by identifying foreign elements and initiating appropriate biological responses (Punt et al., 2020). Antibodies constitute key components of adaptive immunity, as they identify and neutralise antigens with high specificity, which makes them central to immunological research. Nevertheless, the pronounced variability in antibody structure and function poses challenges for their classification and systematic analysis. Within this context, computational tools and machine learning models have enabled notable advances in the representation and classification of antibody amino acid sequences, thereby supporting the development of predictive models in bioinformatics and molecular biology. These approaches not only facilitate more detailed characterisation of antibodies but also may contribute to research on antibody-based therapies, drug discovery, and vaccine design.

### 2.1 Immune system

The immune system constitutes a complex network of organs, cells, and molecules that coordinate to defend the host against pathogens. Functionally, it is divided into two interconnected components: innate immunity, which provides an immediate but non-specific defence, and adaptive immunity, which is highly specific and capable of forming immunological memory. The adaptive response is initiated when innate mechanisms are insufficient, leading to the activation of B and T lymphocytes (Punt et al., 2020; Abbas et al., 2021). Antibodies, also referred to as immunoglobulins, are produced by B cells and play a central role in the adaptive immune response. Structurally, they consist of two identical heavy chains (IgH) and two identical light chains (IgL), forming a characteristic Y-shaped configuration. Each chain comprises constant and variable regions, with the variable regions forming the antigen-binding sites. These sites are located within three hypervariable loops, known as complementarity-determining regions (CDRs)—CDR1, CDR2, and CDR3—in both chains (Murphy et al., 2022; Parham, 2021).

The CDRs are essential for antigen recognition due to their pronounced sequence variability. Among these regions, CDR3 exhibits the greatest diversity in both length and amino acid composition, particularly within the heavy chain, as it is generated through V(D)J recombination. This diversity is fundamental to the generation of a broad antibody repertoire capable of recognising a wide range of pathogens (Ibero-American Cooperative Group on Transfusion Medicine, 2020; Parham, 2021). From a structural perspective, the specific arrangement of amino acids within the CDRs determines the shape, charge, and hydrophobicity of the antigen-binding site, which in turn influences specificity and affinity. Consequently, the representation of amino acid sequences—especially those within the CDRs—is widely regarded as a critical factor in computational modelling, particularly for tasks such as antibody classification, affinity prediction, and therapeutic design (Parham, 2021; Murphy et al., 2022).

## 2.2 Representation of amino acid sequences

The representation of amino acid sequences constitutes a key preprocessing step, as it transforms sequences into numerical vectors that can be processed by machine learning models. Because such algorithms require numerical input, different representation methods are employed to capture relevant information. In this study, three approaches are examined: TF–IDF, which quantifies the statistical frequency of amino acids within sequences; Atchley Factors, which encode physicochemical properties to reduce dimensionality while preserving biochemical characteristics; and ProtVec, which applies word-embedding techniques to capture contextual relationships between amino acids.

**Term Frequency–Inverse Document Frequency (TF–IDF).** TF–IDF is a method that converts textual content into vectorised numerical representations, enabling its application in tasks such as document categorisation, keyword extraction, and information retrieval (Jurafsky & Martin, 2021). The objective of this method is to estimate the importance of a term within a document relative to its relevance across a collection of documents (corpus). TF–IDF integrates three main components to achieve this transformation. Term Frequency (TF) reflects how often a term appears within a given document, whereas Inverse Document Frequency (IDF) down-weights terms that occur frequently across the dataset. Their combination highlights terms that are particularly informative within a corpus (Jurafsky & Martin, 2021). The resulting TF–IDF value combines these components to express the relevance of a term in a document in relation to the corpus as a whole. Each document is thus represented as a numerical vector in which each element corresponds to the relative importance of a specific term, allowing machine learning algorithms to operate on numerical data rather than raw text, which may improve computational efficiency (Jurafsky & Martin, 2021).

In the context of amino acid sequences, the term frequency of an amino acid is calculated as the ratio between its number of occurrences and the total number of amino acids in the sequence. The inverse document frequency is computed using the logarithm of the ratio between the total number of sequences and the number of sequences containing the given amino acid, with one added to the denominator to avoid division by zero. Finally, the TF–IDF value is obtained by multiplying the corresponding TF and IDF values (Jurafsky & Martin, 2021). Equation 1 presents the complete TF–IDF formulation.

$$TF - IDF\ (\alpha) = \frac{Number\ of\ times\ (\alpha) appears\ in\ de\ sequence}{Total\ number\ amino\ acid\ in\ the\ sequence} \times \log(\frac{N}{n_a + 1}) \quad \textbf{(1)}$$

**Atchley Factors** consist of five numerical values assigned to each of the 20 standard amino acids, derived from a principal component analysis of multiple physicochemical descriptors. These values enable a compact and biologically meaningful representation of amino acids for machine learning applications by capturing key biochemical characteristics. Specifically, Factor I reflects polarity and hydrophobicity, Factor II relates to secondary structure preferences, Factor III represents molecular volume or size, Factor IV captures amino acid composition and codon diversity, and Factor V accounts for electrostatic charge. This five-dimensional encoding is intended to simplify the integration of amino acid sequences into predictive models by reducing dimensionality while preserving relevant information. In this study, Atchley Factors were used to numerically represent the complementarity-determining regions (CDRs) of antibodies.

The use of Atchley Factors in sequence representation has received increasing attention due to their capacity to condense complex biochemical information into a low-dimensional format. Unlike approaches such as one-hot encoding or frequency-based methods, which often yield high-dimensional and sparse feature spaces, Atchley Factors provide a dense and interpretable alternative. By assigning each amino acid a five-dimensional vector that captures relevant physicochemical attributes, this approach allows key biological properties to be retained while potentially improving computational efficiency. This aspect is particularly relevant in antibody modelling, where preserving the functional characteristics of variable regions, such as the CDRs, is critical for accurate classification and prediction. Furthermore, this representation may reduce redundancy and limit noise introduced by less informative features, which is especially advantageous when working with limited datasets or models that are susceptible to overfitting. The compact nature of the Atchley encoding facilitates integration into a range of machine learning pipelines, supporting tasks such as clustering, classification, and regression. In the present work, this approach was applied to represent the six CDRs of antibodies prior to model training and evaluation.

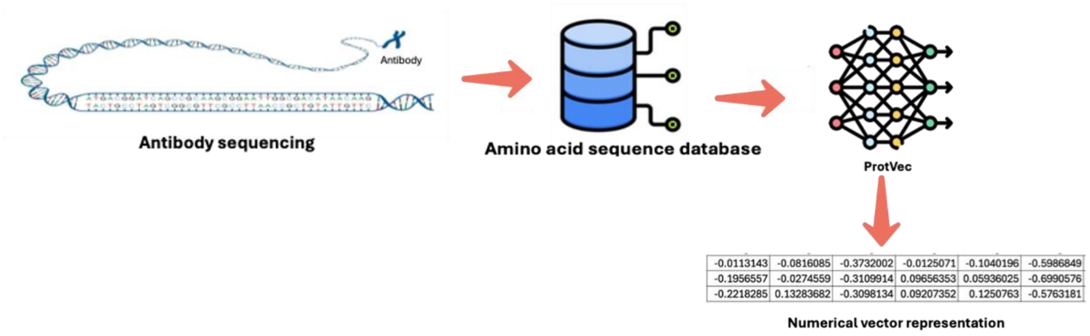
Table 1 lists the Atchley Factor values for the 20 standard amino acids, as originally derived by Atchley et al. (2005). These values constitute the basis of the encoding strategy employed in this research.

**Table 1.** Atchley Factors of amino acids (Atchley et al., 2005)

| Amino Acid | Symbol | Factor I | Factor II | Factor III | Factor IV | Factor V |
|---|---|---|---|---|---|---|
| Alanine | A | -0.591 | -1.302 | -0.733 | 1,570 | -0.146 |
| Cysteine | C | -1.343 | 0.465 | -0.862 | -1.020 | -0.255 |
| Aspartic Acid | D | 1.050 | 0.302 | -3.656 | -0.259 | -3.242 |
| Glutamic Acid | E | 1.357 | -1.453 | 1.477 | 0.113 | -0.837 |
| Phenylalanine | F | -1.006 | -0.59 | 1.891 | -0.397 | 0.412 |
| Glycine | G | -0.384 | 1.652 | 1.33 | 1.045 | 2.064 |
| Histidine | H | 0.336 | -0.417 | -1.673 | -1.474 | -0.078 |
| Isoleucine | I | -1.239 | -0.547 | 2.131 | 0.393 | 0.816 |
| Lysine | K | 1.831 | -0.561 | 0.533 | -0.277 | 1.648 |
| Leucine | L | -1.019 | -0.987 | -1.505 | 1.266 | -0.912 |
| Methionine | M | -0.663 | -1.524 | 2.219 | -1.005 | 1.212 |
| Asparagine | N | 0.945 | 0.828 | 1,299 | -0.169 | 0.933 |
| Proline | P | 0.189 | 2.081 | -1.628 | 0.421 | -1.392 |
| Glutamine | Q | 0.931 | -0.179 | -3.005 | -0.503 | -1.853 |
| Arginine | R | 1,538 | -0.055 | 1,502 | 0.44 | 2.897 |
| Serine | S | -0.228 | 1.399 | -4.76 | 0.67 | -2.647 |
| Threonine | T | -0.032 | 0.326 | 2.213 | 0.908 | 1.313 |
| Valine | V | -1.337 | -0.279 | -0.544 | 1.242 | -1.262 |
| Tryptophan | W | -0.595 | 0.009 | 0.672 | -2.128 | -0.184 |
| Tyrosine | Y | 0.26 | 0.83 | 3.097 | -0.838 | 1.512 |

**Word embeddings.** Word embeddings are a method widely used in natural language processing (NLP) that transforms words or entire documents into numerical vectors encoding semantic meaning and contextual relationships among terms. This representation supports a range of NLP tasks by enabling computational linguistic analysis of textual data (Birunda & Devi, 2021). Mapping words into a mathematical space can enhance the analysis and interpretation of terms within text, thereby potentially improving the performance of various machine learning techniques. In bioinformatics, word-embedding approaches have been applied to represent DNA, RNA, and protein sequences (Zhang et al., 2019). Within this domain, word embeddings may be used to encode amino acid sequences as numerical vectors, allowing evolutionary and structural patterns to be captured. Models such as ProtVec employ word embeddings to represent proteins based on amino acid subsequences, enabling machine learning algorithms to account for functional and structural similarities between proteins (Birunda & Devi, 2021).

**ProtVec.** ProtVec is a pre-trained model based on word-embedding techniques that represents protein sequences numerically using unsupervised learning. This approach segments sequences into overlapping tripeptides and encodes them as continuous vector representations that capture sequence patterns and contextual relationships. ProtVec has been applied in tasks such as protein classification, functional annotation, and structure prediction, as it considers both local and global sequence dependencies. One notable characteristic of this model is its bidirectional representation capacity, which allows the model to capture relationships between amino acids across the entire sequence (Asgari & Mofrad, 2015). In this study, ProtVec was applied to antibody amino acid sequences obtained through sequencing and stored in a dataset. These sequences served as input to ProtVec, which generated numerical vectors suitable for subsequent machine learning analysis. Figure 1 illustrates the ProtVec-based representation process.



| | | | | | |
|---|---|---|---|---|---|
| -0.0113143 | -0.0816085 | -0.3732002 | -0.0125071 | -0.1040196 | -0.5986849 |
| -0.1956557 | -0.0274559 | -0.3109914 | 0.09656353 | 0.05936025 | -0.6990576 |
| -0.2218285 | 0.13283682 | -0.3098134 | 0.09207352 | 0.1250763 | -0.5763181 |

**Fig. 1.** ProtVec process.

## 2.3 Antibody repositories

In the literature, several datasets are available for the study and analysis of antibodies. Among the most relevant repositories are the Coronavirus Antibody Database (CoV-AbDab) (Raybould et al., 2021) and the Observed Antibody Space (OAS) (Olsen et al., 2022). These repositories contain extensive and curated data on antibody structures, which are essential for immunological research and support the generation of knowledge regarding antibody structure and function.

**Coronavirus Antibody Database (CoV-AbDab).** CoV-AbDab is a repository specialised in the structural information of antibodies that recognise coronaviruses, with a particular focus on antibodies targeting SARS-CoV-2. Its primary objective is to provide structured data on antibody sequences, structures, binding affinities, and functional characteristics that have demonstrated activity against coronaviruses. The repository includes antibodies derived from multiple species, such as humans, mice, alpacas, and primates, and provides detailed structural information, including the complementarity-determining regions (CDRs) of both heavy and light chains. CoV-AbDab is commonly used to analyse antigen-recognition patterns using machine learning models and is widely regarded as a valuable resource for drug discovery research, the development of vaccines based on neutralising antibodies, and the structural modelling of interactions between SARS-CoV-2 and the immune system (Raybould et al., 2021).

**Observed Antibody Space (OAS).** The OAS repository was established in 2018 with the aim of providing clean, annotated, and translated antibody sequence data. Prior to the development of OAS, publicly available antibody data were predominantly distributed in unprocessed FASTQ files, which require substantial preprocessing before analysis. This limited the accessibility and effective use of antibody structural information. Currently, OAS represents one of the most comprehensive antibody repositories, containing millions of sequences organised according to biological source, collection methodology, and structural characteristics. In bioinformatics research, OAS is widely used for training machine learning models in tasks such as classification, affinity prediction, and therapeutic antibody optimisation. In addition, OAS provides antibody sequences that do not recognise SARS-CoV-2, which can support the construction of datasets for comparative antibody classification studies (Olsen et al., 2022).

# 3 Related work

As previously discussed, accurately representing amino acid sequences constitutes a critical step in antibody analysis, as it substantially influences the effectiveness of machine learning models applied to antibody classification and the prediction of antigen-binding interactions. Consequently, numerous research groups have focused on developing representations that account for sequential information, physicochemical properties, and structural relationships among amino acids.

Chen et al. (2020) employ machine learning to predict antibody developability using word embeddings and physicochemical features derived solely from antibody sequences. Their study uses antibody data collected from the SAbDab platform. Although the initial dataset comprised 3,816 antibodies, only 2,426 met the inclusion criteria: (1) availability of sequence structure files (FASTA) and Protein Data Bank (PDB) structure files; (2) presence of both heavy and light antibody chains; and (3) availability of a crystal structure with a resolution below 3 Å. Due to the relatively limited size of the final dataset, the resulting models exhibited overfitting and limited generalisation capability. Model performance was evaluated using the F1 score, and similar performance levels were observed when datasets with comparable attribute characteristics were used. The implemented machine learning models outperformed the reference model (mean F1 scores of 0.36 and 0.57), with Support Vector Machines achieving mean F1 scores of 0.52 and 0.65, and multilayer perceptrons achieving mean F1 scores of 0.50 and 0.64 (Chen et al., 2020).

Similarly, Li Xinmeng and Van Deventer (2020) describe the development of an antibody sequence analysis pipeline using statistical testing and machine learning (ASAP-SML), with the objective of identifying features that distinguish a target set of antibody sequences from those in a reference set (Li, Van Deventer, & Hassoun, 2020). The proposed pipeline comprises five stages: data preparation, antibody sequence numbering, feature extraction, analysis, and a final recommendation step. Three classification techniques—Support Vector Machines, Random Forests, and Adaptive Boosting (AdaBoost)—were applied. Classification performance was evaluated using the area under the receiver operating characteristic curve (AUC) with ten-fold cross-validation. Feature selection was conducted using a Random Forest algorithm implemented in Python with the scikit-learn library, while the implementation details for the remaining algorithms were not specified (Li, Van Deventer, & Hassoun, 2020).

Magar et al. (2021) applied multiple machine learning models to predict inhibitory antibodies against SARS-CoV-2. A dataset comprising 1,933 virus–antibody sequence pairs was collected, encompassing antibodies targeting viruses such as HIV, influenza, dengue, SARS-CoV-2, and Ebola. Using this dataset, both shallow and deep learning models were trained and evaluated, and the model with the highest performance was selected to predict potentially neutralising antibodies. The models were trained on data from 14 different viruses and achieved classification accuracies exceeding 90% across five evaluation tests. The authors report that the selected model achieved a 100% prediction rate for SARS-CoV-2 neutralisation and 84.61% for influenza. Using this model, thousands of hypothetical antibodies were evaluated, leading to the identification of 18 candidates with high neutralisation potential against SARS-CoV-2 (Magar et al., 2021).

Leem et al. (2022) present AntiBERT, a language model specifically designed for antibody sequences, which provides contextualised representations of B-cell receptor (BCR) sequences. The study reports that AntiBERT is capable of capturing information relevant to multiple antibody-related tasks. In a case study focused on paratope prediction, AntiBERT outperformed existing public tools across several evaluation metrics. The authors note that the model captures functional patterns of BCR sequences more effectively than alternative models, such as ProtBERT. The work demonstrates the model's capacity to anticipate antibody binding sites, thereby contributing to improved understanding of antibody function. Additional applications, including antibody engineering and other structure–function analyses, are also explored. Details regarding the datasets, pre-training process, and hyperparameter configuration are provided, outlining the model's applicability to a range of BCR repertoire analysis tasks (Leem et al., 2022).

In related work, Li et al. (2022) introduce deep learning approaches for predicting antibody properties and generating customised designs for drug discovery. Their study compares three classes of models and introduces a SARS-CoV-2 antibody-binding dataset. The authors discuss the impact of dataset size on model performance and outline several challenges and opportunities in therapeutic antibody design. The performance of a BERT-based Transformer language model is also evaluated for antibody-binding prediction, with comparisons across different datasets and training strategies (Li et al., 2022).

As demonstrated in prior research, machine learning represents one of the most widely adopted approaches for predicting structural characteristics of antibodies. Most studies frame these prediction tasks as classification problems, as this approach has been shown to be capable of processing large-scale datasets and capturing complex patterns in antibody structure. When implementing such algorithms, careful consideration of performance metrics is required, as these metrics are essential for interpreting results and determining whether models accurately capture antibody structural characteristics.

Natural language processing techniques are among the most commonly used approaches for representing antibody sequences, as they have demonstrated effectiveness in capturing structural patterns through statistical analysis. These techniques also enable antibody chains to be used directly as inputs for machine learning models. In parallel, Atchley Factors offer an alternative representation strategy that may preserve relevant structural characteristics while reducing dimensionality.

Although a substantial body of work has applied machine learning to antibody data, many studies have prioritised model architecture over systematic evaluation of input representations. However, the manner in which amino acid sequences are encoded directly affects model performance, generalisability, and interpretability, particularly in biomedical contexts where biological relevance is critical. Moreover, existing studies frequently rely on heterogeneous datasets, inconsistent evaluation metrics, or custom preprocessing pipelines, which complicates fair comparison across approaches. This research addresses this gap by implementing a controlled experimental framework to evaluate three widely used representation strategies—TF–IDF, Atchley Factors, and ProtVec—within a unified classification task. By isolating the effect of representation while maintaining a consistent classification framework, this study aims to provide empirical evidence to inform the selection of sequence-encoding methods in antibody modelling, thereby supporting downstream applications such as therapeutic design and antigen recognition.

## 4 Preprocessing of amino acid chains

Data preprocessing constitutes a critical stage in computational antibody analysis, as it ensures that heavy- and light-chain sequences are formatted appropriately for processing by machine learning algorithms. Biological datasets frequently contain redundancies, sequencing artefacts, and heterogeneous formats; therefore, thorough preprocessing is essential to maximise data quality and analytical reliability.

One of the principal challenges in amino acid sequence representation is the potential loss of sequential order inherent to certain encoding strategies. Amino acid order is fundamental to antibody structure formation and functional performance. If this information is not adequately preserved, the predictive accuracy of machine learning models may be adversely affected. A further challenge arises from sequence-length variability, particularly within the complementarity-determining regions (CDRs), which are responsible for antigen-binding specificity. Such variability complicates the generation of fixed-length numerical representations, a requirement for many machine learning algorithms.

The complementarity-determining regions of both heavy and light chains were extracted directly from the CoV-AbDab repository, which provides annotated sequences for CDR-H1, CDR-H2, CDR-H3, CDR-L1, CDR-L2, and CDR-L3 using the IMGT numbering scheme. The IMGT (International ImMunoGeneTics information system®) serves as the global reference in immunogenetics and immunoinformatics, offering standardised definitions and numbering for immunoglobulin and T-cell receptor sequences (Lefranc et al., 2009). For the OAS dataset, only entries with compatible CDR annotations were retained. In addition, all sequences included in this study were verified using IMGT/V-QUEST, an online tool provided by IMGT for the identification and analysis of variable (V), diversity (D), and joining (J) genes in immunoglobulin and T-cell receptor sequences. This platform enables accurate identification of CDR regions in accordance with IMGT standards, and no external automatic CDR prediction tools were applied. This preprocessing strategy ensured consistency and a high level of reliability in the representation of antibody variable regions (Lefranc et al., 2009).

### 4.1 Analysis of heavy and light chains

The amino acid sequences of the heavy and light chains obtained from CoV-AbDab were analysed using natural language processing techniques. The analysis was conducted in a segmented manner, focusing specifically on the complementarity-determining regions (CDRs), as these regions are responsible for pathogen binding. Moreover, CDRs exhibit the highest degree of variability within antibody chains, which makes them central to immune-response diversity and specificity (Pulendran & Davis, 2020).

The first step involved analysing the amino acids most frequently observed within each CDR. This analysis aimed to examine whether any meaningful associations exist among the amino acids present in these regions. Identifying frequently occurring amino acids in the CDRs can offer insight into potential patterns related to pathogen binding. The results indicate that, for both CDR1 and CDR2, the highest frequencies are concentrated in three amino acids. In CDR1, these amino acids are serine, glycine, and phenylalanine, whereas in CDR2 they are serine, isoleucine, and glycine. Notably, serine and glycine appear as the most frequent amino acids in both regions.

Figure 2 illustrates the segmentation of an antibody heavy chain according to the IMGT numbering scheme. The sequence is divided into framework regions (FR1, FR2, and FR3) and complementarity-determining regions (CDR-H1, CDR-H2, and CDR-H3), which are highlighted in red. These segments correspond to the variable loops that directly interact with the antigen, with CDR-H3 being the most diverse and structurally flexible region. The prefix “H” indicates that these CDRs belong to the heavy chain. The boundaries of each region were identified using the IMGT/V-QUEST platform, which applies standardised alignment rules and the IMGT numbering system. This figure represents the structure of the input used for feature extraction in the machine learning models developed in this study.

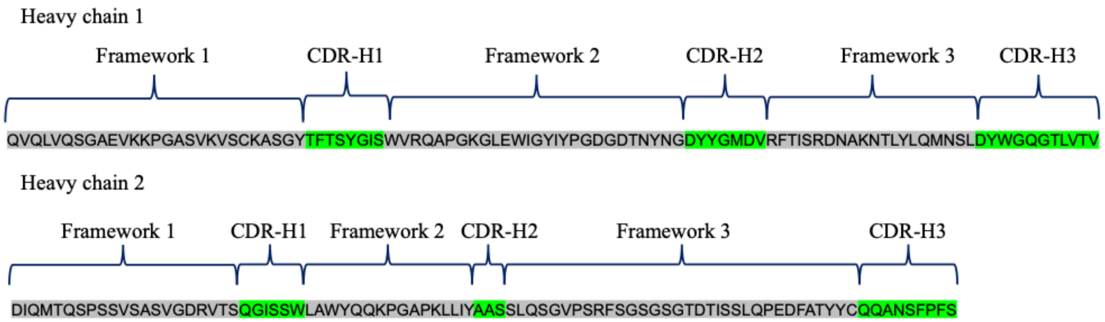

Heavy chain 1
Framework 1
CDR-H1
Framework 2
CDR-H2
Framework 3
CDR-H3
QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGKGLEWIGYIYPGDGDTNYNGDYYGMDVRFTISRDNAKNTLYLQMNSLDYWGQGTLVTV
Heavy chain 2
Framework 1
CDR-H1
Framework 2
CDR-H2
Framework 3
CDR-H3
DIQMTQSPSSVSASVGDRVTSQGISSWLAWYQQKPGAPKLLIYAASSLQSGVPSRFSGSGSGTDTISSLQPEDFATYYCQQANSFPFS

**Fig. 2.** Annotated example of a heavy chain sequence with CDR using IMGT numbering.

Compared to CDR1 and CDR2, CDR3 exhibits a greater variability in the frequency of amino acids, with this frequency being distributed among all twenty amino acids. This variability arises because CDR3 is the region with the greatest variation in both size and amino acid composition. Despite this variability, the amino acids that appear most frequently in CDR3 are Serine, Glycine, and Lysine. Table 2 shows the frequency of appearance of each of the amino acids in the three CDR of the chains that make up the amino acids.

A study of the same sections was conducted using tuples to determine the most frequently occurring amino acid pairs in the CDR. For heavy chains, it was discovered that, in CDR1, the most frequently occurring amino acid tuple is threonine and phenylalanine. For CDR2, it is threonine and isoleucine. Finally, the most frequently occurring amino acid tuple in CDR3 is tryptophan and cysteine. Fig. 3 shows the five most frequently occurring tuples in the CDR of heavy chains.

## 4.2 Antibody dataset

Data on anti–SARS-CoV-2 antibodies were obtained from the CoV-AbDab repository, which contains records of antibodies derived from multiple species. From this collection, only antibodies reported to recognise the SARS-CoV-2 coronavirus were selected. Subsequently, records corresponding to nanoantibodies were excluded. As a result, a total of 8,706 records representing complete anti–SARS-CoV-2 antibody structures were retained for analysis.

Seven attributes were extracted from these records: the Name attribute, which serves as a unique identifier; the sequences of the three complementarity-determining regions of the heavy chain (CDR-H1, CDR-H2, and CDR-H3); and the sequences of the three complementarity-determining regions of the light chain (CDR-L1, CDR-L2, and CDR-L3). Table 3 provides representative examples of the dataset records.

For the OAS dataset, the data search was conducted using the following criteria: samples had to originate from individuals who were not infected with any disease at the time of collection, and the samples had to be obtained prior to the identification of SARS-CoV-2, in order to increase the likelihood that the antibodies were not specific to SARS-CoV-2. Based on these criteria, a total of 24 CSV files were retrieved, containing sequences of heavy and light chains that together form complete antibodies lacking SARS-CoV-2 specificity. In total, 350,209 complete antibodies interacting with a range of antigens were collected.

**Table 2.** Frequency of Occurrence of Amino Acids in the OAS Dataset

| Amino Acid | Symbol | CDR1 % | CDR2 % | CDR3 % |
|---|---|---|---|---|
| Serine | S | 19.89% | 18.94% | 19.96% |
| Glycine | G | 17.75% | 17.37% | 17.75% |
| Lysine | K | 15.62% | 15.79% | 16.10% |
| Threonine | T | 12.78% | 13.42% | 13.31% |
| Tyrosine | Y | 12.06% | 12.63% | 12.21% |
| Alanine | A | 5.68% | 5.52% | 7.77% |
| Proline | P | 3.55% | 3.48% | 6.66% |
| Aspartic Acid | D | 2.84% | 2.84% | 6.11% |
| Valine | V | 2.56% | 2.53% | 5.55% |
| Isoleucine | I | 2.13% | 2.21% | 4.99% |

| | | | | |
|---|---|---|---|---|
| Asparagine | N | 1.42% | 1.42% | 4.44% |
| Leucine | L | 1.42% | 1.42% | 4.44% |
| Phenylalanine | F | 1.13% | 1.18% | 3.88% |
| Glutamic Acid | E | 0.99% | 1.03% | 3.33% |
| Arginine | R | 0.85% | 0.87% | 2.77% |
| Methionine | M | 0.71% | 0.71% | 2.22% |
| Histidine | H | 0.57% | 0.55% | 2.00% |
| Cysteine | C | 0.42% | 0.40% | 1.66% |
| Tryptophan | W | 0.50% | 0.47% | 1.66% |
| Glutamine | Q | 0.28% | 0.24% | 1.11% |

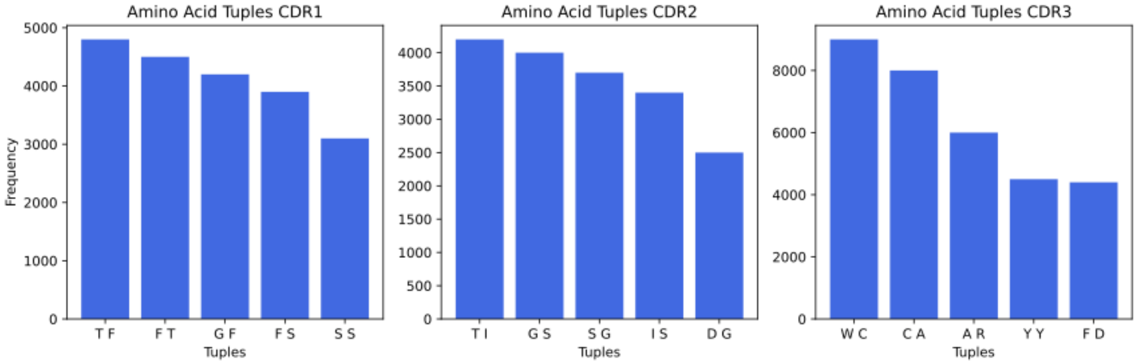


**Fig. 3.** Tuples with higher frequency in CDR of heavy chains.

**Table 3.** Example of data

| Name | CDR-H1 | CDR-H2 | CDR-H3 | CDR-L1 | CDR-L2 | CDR-L3 |
|---|---|---|---|---|---|---|
| Antibody_1 | GFTFSDY | INPYT | YFDY | QSVLTQPP | NQNSGGS | QHYSTPP |
| Antibody_2 | NFTLSWY | VKGQTP | YGSDV | QISLTYPD | NQNFGDS | QVYTPTP |

Each of these files contains 198 attributes, including the complete heavy- and light-chain sequences for each antibody. In this structure, each record represents an individual antibody. As with the anti–SARS-CoV-2 antibodies, only the amino acid sequences corresponding to the complementarity-determining regions (CDRs) were extracted for subsequent analysis.

Within the OAS dataset, each file includes 198 attributes per record, encompassing both metadata and full amino acid sequences of antibody chains. For the purposes of this study, amino acid sequences of the heavy and light chains were extracted primarily from the sequence_alignment_aa attribute, which contains the aligned amino acid sequence of the complete chain, and the chain attribute, which specifies whether the entry corresponds to a heavy or light chain. In addition, the cdr1, cdr2, and cdr3 attributes were used to retrieve the pre-annotated CDR regions. The sequence illustrated in Fig. 2 corresponds to the content of the sequence_alignment_aa attribute, segmented into framework and CDR regions according to the IMGT numbering scheme. These fields are curated and validated within OAS, ensuring that heavy and light chains are correctly paired and biologically consistent. Consequently, no additional pairing or prediction tools were required. This extraction procedure was intended to preserve the structural integrity of each antibody used in the machine learning analysis.

The two datasets were subsequently combined to construct a single dataset comprising two classes: SARS-CoV-2 and non–SARS-CoV-2 antibodies. The anti–SARS-CoV-2 dataset obtained from CoV-AbDab consisted of 8,706 records, whereas the non–SARS-CoV-2 antibody dataset derived from OAS contained 350,209 records. These datasets were merged into a unified dataset for downstream analysis and classification.

## 4.3 Representation based on TF-IDF

Although the TF–IDF representation yields vectors with fixed dimensionality—one feature per standard amino acid—the actual content and informativeness of these vectors vary depending on whether the complete antibody chain or only the CDR regions are considered. When TF–IDF is applied to full-length sequences, amino acid frequencies are distributed across both functional regions and conserved structural segments. This distribution results in more uniform vectors, which may dilute discriminative signals required for effective classification.

By contrast, restricting TF–IDF to the CDR regions focuses the representation on the most variable and antigen-specific components of the antibody, which are directly involved in antigen binding. This approach produces more compact and informative vectors, as structurally conserved but less relevant regions are excluded. From a methodological perspective, these observations suggest that a CDR-focused TF–IDF representation can offer a more specific and computationally efficient encoding strategy without increasing the dimensionality of the feature space. Table 4 presents illustrative examples of the TF–IDF representation.

**Table 4.** Dataset with Vectors generated with TF-IDF

| ID | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BA7054 | 0.28 | 0.05 | 0.18 | 0.10 | 0.08 | 0.42 | 0.03 | 0.07 | 0.12 | 0.14 | 0.11 | 0.16 | 0.22 | 0.18 | 0.20 | 0.30 | 0.33 | 0.26 | 0.08 | 0.10 |
| BA7125 | 0.26 | 0.06 | 0.17 | 0.12 | 0.09 | 0.40 | 0.02 | 0.08 | 0.11 | 0.15 | 0.10 | 0.15 | 0.20 | 0.17 | 0.21 | 0.28 | 0.31 | 0.25 | 0.09 | 0.11 |
| BA7208 | 0.25 | 0.05 | 0.19 | 0.11 | 0.07 | 0.43 | 0.04 | 0.06 | 0.13 | 0.13 | 0.12 | 0.17 | 0.21 | 0.19 | 0.18 | 0.29 | 0.32 | 0.27 | 0.07 | 0.09 |
| 6-2C | 0.27 | 0.05 | 0.16 | 0.12 | 0.10 | 0.41 | 0.02 | 0.07 | 0.10 | 0.14 | 0.11 | 0.14 | 0.23 | 0.16 | 0.22 | 0.31 | 0.34 | 0.24 | 0.10 | 0.12 |
| Jaiswal | 0.30 | 0.06 | 0.18 | 0.13 | 0.11 | 0.42 | 0.03 | 0.09 | 0.12 | 0.16 | 0.13 | 0.15 | 0.24 | 0.17 | 0.24 | 0.32 | 0.35 | 0.26 | 0.11 | 0.13 |
| IS-9A | 0.29 | 0.06 | 0.19 | 0.12 | 0.09 | 0.44 | 0.04 | 0.08 | 0.13 | 0.15 | 0.12 | 0.18 | 0.22 | 0.20 | 0.23 | 0.30 | 0.33 | 0.28 | 0.09 | 0.11 |
| CC68.109 | 0.31 | 0.07 | 0.20 | 0.13 | 0.10 | 0.45 | 0.05 | 0.07 | 0.14 | 0.17 | 0.13 | 0.16 | 0.23 | 0.21 | 0.22 | 0.31 | 0.34 | 0.27 | 0.10 | 0.12 |
| CC99.103 | 0.32 | 0.06 | 0.21 | 0.11 | 0.11 | 0.43 | 0.03 | 0.08 | 0.13 | 0.15 | 0.14 | 0.17 | 0.25 | 0.19 | 0.21 | 0.33 | 0.36 | 0.26 | 0.11 | 0.13 |
| sd1.040 | 0.33 | 0.06 | 0.19 | 0.10 | 0.12 | 0.41 | 0.02 | 0.07 | 0.12 | 0.14 | 0.11 | 0.15 | 0.21 | 0.18 | 0.23 | 0.29 | 0.32 | 0.25 | 0.12 | 0.14 |

Figure 4 illustrates the nine vectors from Table 4 generated using the TF–IDF representation. The highest peaks correspond to amino acids with the greatest presence, as derived from the analysis of the amino acid sequences of the heavy and light chains described in Section 4.1.

The same procedure was applied to the amino acid sequences of the CDRs. In this case, the highest peaks likewise correspond to the amino acids most frequently identified within each CDR. In addition, CDR3 exhibits greater variability, which is consistent with its known structural and functional diversity. Figure 5 presents the vector diagrams produced for the CDR regions.

## 4.4 Representation of chains by their Atchley Factors

In contrast to approaches that operate on the 20 standard amino acids and their individual properties, Atchley Factors provide a compact five-dimensional representation of each amino acid chain. This strategy reduces dimensionality while retaining key biochemical and structural characteristics.

The final values in the Atchley Factor representation are obtained by computing the average of each factor across all amino acids within a given sequence. Equation 2 presents the formulation used to calculate each Atchley Factor for the sequences.

$$F_i^{CDR} = \frac{1}{n}\sum_{j=1}^{n} F_i^{(j)} \tag{2}$$

where:
$F_i^{CDR}$ is the Atchley factor value for the entire sequence.
$F_i^{(j)}$ is the Atchley factor value for the j -th amino acid in the sequence.
$n$ is the total number of amino acids in the sequence.

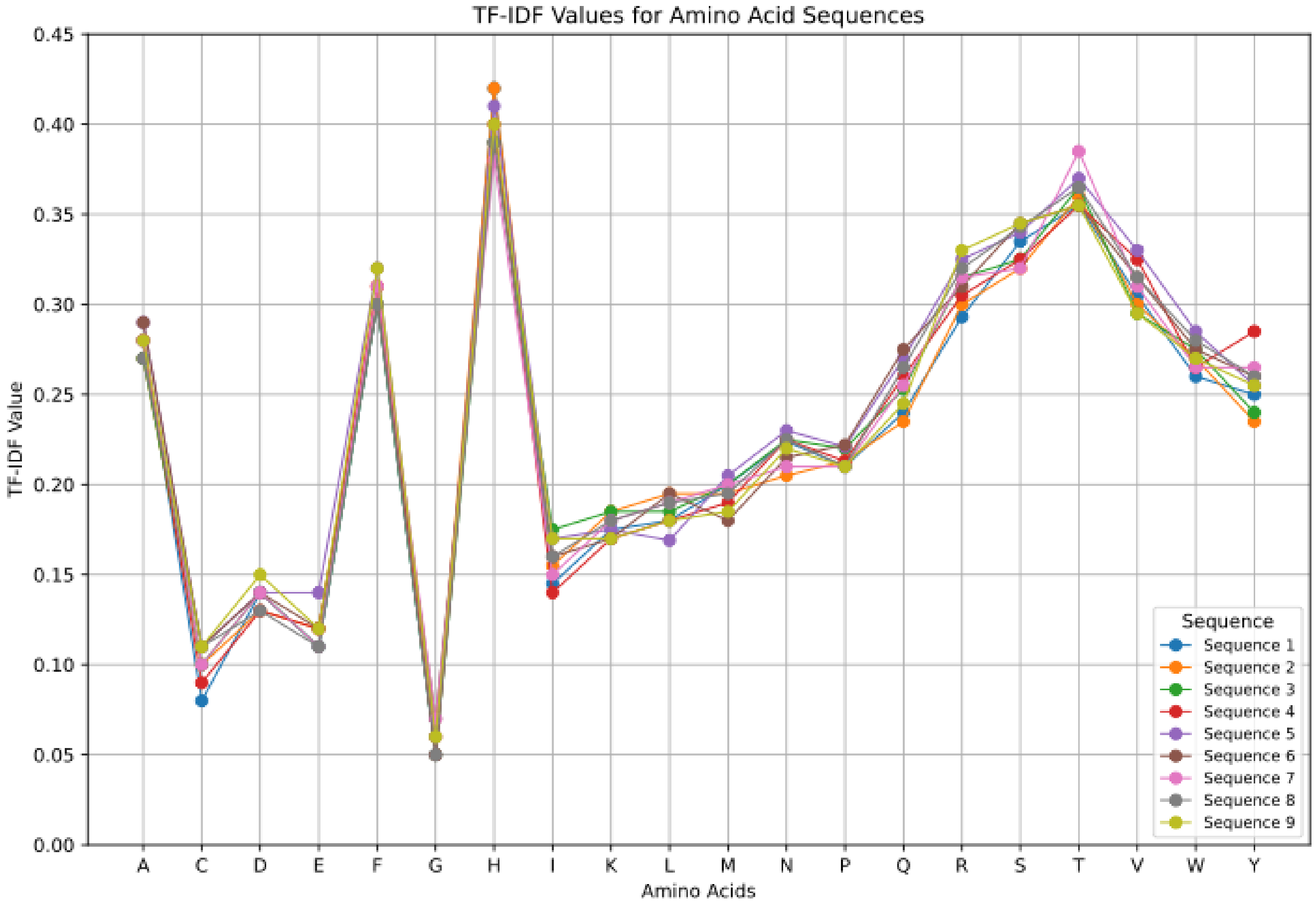


**Fig. 4.** Vectors generated with TF – IDF.

For this representation, the three CDR are encoded by means the five Atchley Factors, which results in 15 attributes per sequence. Table 5 presents an extract of the 15 attributes. The attribute F1CDR1 represents the factor I for the CDR1, the attribute F2CDR1 represents the factor II for the CDR2 and so on.

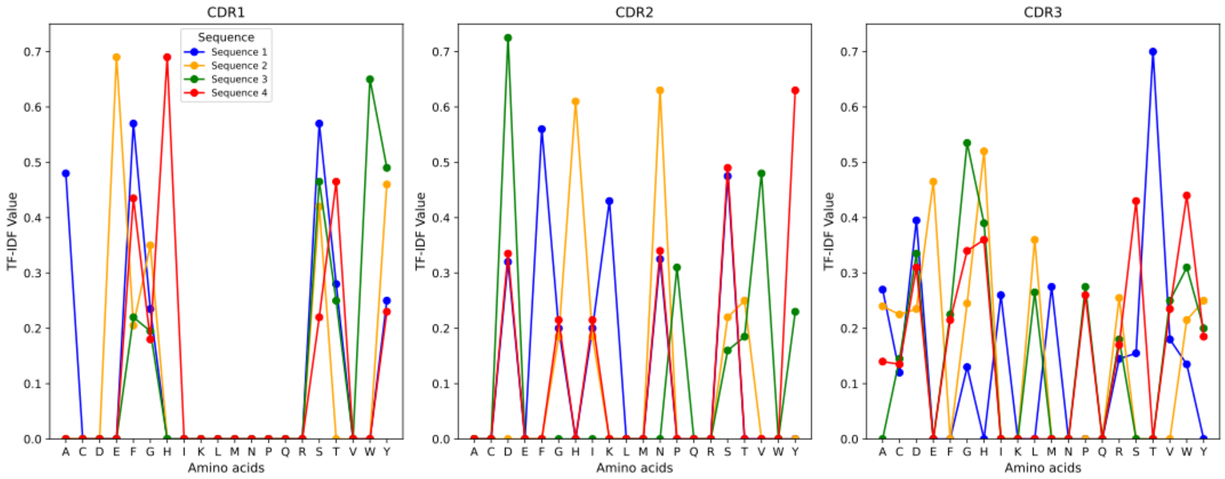


**Fig. 5.** CDR vectors.

**Table 5.** Representation of three CDR of the heavy chains with Atchley Factors

| F1CDR1 | F2CDR1 | F3CDR1 | F4CDR1 | F5CDR1 | …. | F1CDR3 | F2CDR3 | F3CDR3 | F4CDR3 | F5CDR3 |
|---|---|---|---|---|---|---|---|---|---|---|
| -0.401 | 0.390 | 0.021 | 0.403 | 0.034 | .... | -0.514 | 0.690 | 0.221 | 0.403 | 0.334 |
| -0.044 | 0.714 | 0.337 | 0.183 | 0.179 | …. | 0.195 | 0.514 | 0.537 | 0.383 | 0.379 |
| -0.244 | 0.731 | 0.347 | -0.113 | 0.166 | …. | 0.331 | 0.631 | 0.547 | -0.313 | 0.366 |
| -0.261 | 0.367 | 0.775 | 0.053 | 0.537 | …. | 0.423 | 0.367 | 0.775 | 0.253 | 0.737 |
| -0.401 | 0.390 | 0.021 | 0.403 | 0.034 | …. | 0.214 | 0.490 | 0.321 | 0.603 | 0.234 |
| -0.097 | 0.660 | 1.104 | -0.218 | 0.614 | …. | 0.297 | 0.560 | 1.204 | -0.418 | 0.614 |
| 0.119 | 0.628 | 2.054 | 0.056 | 1.249 | …. | 0.419 | 0.628 | 2.254 | 0.256 | 1.149 |
| -0.014 | 0.634 | 0.678 | 0.149 | 0.279 | …. | 0.254 | 0.634 | 0.878 | 0.349 | 0.479 |
| -0.271 | 0.522 | 1.371 | 0.132 | 0.736 | …. | 0.371 | 0.522 | 1.471 | 0.332 | 0.736 |

## 4.5 Representation using ProtVec

ProtVec is a technique used to numerically represent antibody sequences by transforming amino acid sequences into high-dimensional vectors through 3-gram word embeddings. The dimensionality of these embeddings plays a critical role, as it affects both the model's capacity to capture biologically relevant features and the associated computational cost during training.

According to Asgari and Mofrad (2015), ProtVec embeddings typically range from 50 to 300 dimensions, depending on the intended application. Lower-dimensional vectors (e.g., 50–100) are generally preferred for classification tasks, as they offer a balance between expressiveness and computational efficiency. By contrast, higher-dimensional representations (200–300) are more suitable for complex tasks such as structure prediction or evolutionary analysis. In this study, the 100-dimensional version of ProtVec was adopted as an initial configuration, following the original pre-trained model, and alternative dimensionalities were subsequently explored to assess their influence on classification performance.

To construct the antibody representation, each of the six complementarity-determining regions—CDR1, CDR2, and CDR3 from both the heavy and light chains—was processed independently. For each CDR, the mean of all ProtVec embeddings derived from its constituent 3-grams was computed, resulting in a 100-dimensional vector per region. These six vectors were then concatenated to form a single 600-dimensional representation per antibody (3 CDRs × 2 chains × 100 dimensions). This strategy captures local sequence features associated with antigen binding while maintaining a fixed input size compatible with standard machine learning models.

It should be noted that, although ProtVec embeddings capture statistically meaningful patterns learned from large protein corpora, their internal structure remains largely opaque. The contribution of individual amino acids or motifs to the final representation depends on the embedding context and training data, which makes direct interpretation of individual dimensions challenging. Table 6 summarises the dataset structure obtained using this representation.

**Table 6.** Representation of CDR1 with ProtVec

| ID | CDR1Light1 | CDR1Light2 | CDR1Light3 | CDR1Light4 | CDR1Light5 | …… | CDR1Light100 |
|---|---|---|---|---|---|---|---|
| BA7054 | -0.011314 | -0.081608 | -0.373200 | -0.012507 | -0.104019 | …… | 0.332234 |
| BA7125 | -0.195655 | -0.027455 | -0.310991 | 0.096563 | 0.059360 | …… | 0.502038 |
| BA7208 | -0.221828 | 0.132836 | -0.309813 | 0.092073 | 0.125076 | …… | 0.412048 |
| 6-2C | 0.046774 | -0.083658 | 0.150353 | -0.298076 | -0.252442 | …… | -0.039804 |
| Jaiswal_scFv | -0.229373 | 0.056932 | -0.328333 | 0.105182 | 0.106505 | …… | 0.474955 |
| IS-9A | 0.133214 | -0.084762 | 0.318797 | -0.499609 | -0.383408 | …… | -0.203989 |
| CC68.109 | 0.116688 | -0.199207 | -0.001692 | -0.065427 | -0.097210 | …… | 0.026315 |
| CC99.103 | 0.027772 | -0.127451 | 0.126028 | -0.122607 | -0.076175 | …… | 0.030137 |
| sd1.040 | 0.150177 | -0.585391 | 0.350960 | -0.307501 | -0.282573 | …… | -0.001659 |

After generating each of the amino acid sequence representations, three datasets were obtained, one for each representation. These datasets exhibit different characteristics, as each representation method results in a different number of attributes, leading to variations in the structure and dimensionality of the datasets. Table 7 presents a summary of the characteristics of each dataset.

**Table 7.** Datasets Summary

| Representation | Attributes per CDR | Total Attributes (6 CDR) | Classes |
|---|---|---|---|
| TF-IDF | 20 | 120 | Anti-SARS-CoV-2, Non-SARS-CoV-2 |
| Atchley Factors | 5 | 30 | Anti-SARS-CoV-2, Non-SARS-CoV-2 |
| ProtVec | 100 | 600 | Anti-SARS-CoV-2, Non-SARS-CoV-2 |

## 4.6 Class Balancing

Once the datasets were constructed for each representation, it became evident that they were highly imbalanced, with the majority class comprising antibodies that recognise viruses other than SARS-CoV-2. This class consisted of 350,209 records of complete antibody structures, whereas the class of antibodies recognising SARS-CoV-2 comprised only 8,706 records. To address this imbalance, a class-balancing strategy was adopted. The option of generating synthetic samples from the minority class was discarded, as there would be no certainty that the generated structures corresponded to antibodies that recognise SARS-CoV-2. Consequently, an undersampling approach was selected, whereby 8,706 records were randomly drawn from the majority class to obtain a balanced dataset while ensuring that all antibodies in both classes were biologically valid.

Following class balancing, each representation-specific dataset consisted of 8,706 anti–SARS-CoV-2 antibodies and 8,706 antibodies that do not recognise SARS-CoV-2, resulting in a total of 17,412 records per dataset. Each record contained the six complementarity-determining regions—three from the heavy chain and three from the light chain—that define the antibody structure.

To assess the impact of the undersampling procedure on model stability, ten independent undersampling iterations were conducted. In each iteration, a random subset of 8,706 antibody sequences was selected from the majority class to match the size of the minority class. For each sampled dataset, the full classification pipeline was executed, including feature extraction, sequence representation (TF–IDF, Atchley Factors, and ProtVec), model training, and evaluation. All classifiers were trained using stratified 10-fold cross-validation to ensure balanced and consistent evaluation across runs.

Performance metrics—including accuracy, recall, F1-score, and ROC AUC—were computed for each fold and then averaged across the ten undersampling iterations. Standard deviation values were calculated to quantify variability. The results indicate limited variation across the different random subsets, suggesting that classification performance was relatively stable and not strongly affected by sampling variability.

Figure 5 illustrates the average performance and variability associated with each representation across all models. The bars denote the mean value of each metric, while the error bars represent the corresponding standard deviation across the ten runs. As shown in the figure, the Atchley-based representation tended to achieve the highest performance with the lowest variability, whereas ProtVec exhibited greater fluctuations, which may be attributable to its higher dimensionality. This visualisation supports the robustness of the reported results and informs the selection of sequence representation based on both performance and stability.

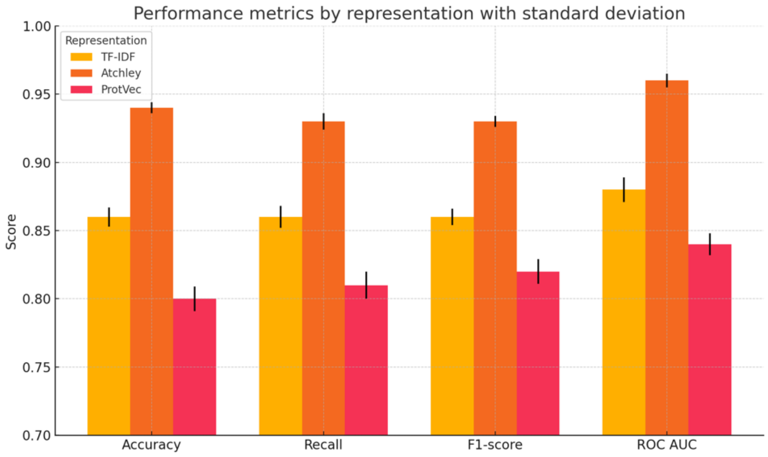


**Fig. 6.** Performance of classification models according to amino acid sequence representation.

## 5 Antibodies classification model

Before evaluating the classification models, a grid-search strategy was employed to tune key hyperparameters for each algorithm. Grid search involves the exhaustive evaluation of all predefined combinations of hyperparameter values and the selection of the configuration that maximises model performance, typically under cross-validation. This procedure helps ensure that model evaluation is not influenced by arbitrary parameter choices and that each algorithm operates under reasonably optimised conditions for the dataset.

The Decision Tree Classifier, Logistic Regression Classifier, and Support Vector Machine Classifier were selected based on their established suitability for supervised classification tasks. Decision Tree and Logistic Regression models are widely used in biomedical research due to their relative simplicity and interpretability, which make them appropriate for identifying influential features in biological datasets. By contrast, the Support Vector Machine classifier is particularly effective when handling high-dimensional data and is recognised for its robustness to overfitting, especially when kernel functions are employed. The evaluation therefore focused on examining how different amino acid sequence representations behave across models with distinct inductive biases and varying levels of complexity.

Hyperparameters for each model were defined in accordance with commonly accepted best practices and further refined using grid search on a validation subset. For example, the maximum depth of the Decision Tree was constrained to limit overfitting, while the regularisation parameter C in both Logistic Regression and Support Vector Machine models was adjusted to balance model complexity and generalisation. All models were trained and evaluated under identical conditions using stratified 10-fold cross-validation.

### 5.1 Decision trees

The experimental analysis was conducted with the objective of evaluating the performance of a decision tree–based model using datasets generated from three amino acid sequence representations: TF–IDF, Atchley Factors, and ProtVec.

Hyperparameter configuration plays a critical role in determining both model performance and generalisation capability. Certain parameters were manually specified to mitigate overfitting and improve classification accuracy, while others were retained at their default settings. Table 8 summarises the hyperparameters applied in this study. Using these configurations, experiments were carried out on each dataset corresponding to the different representations in order to assess model performance. The results obtained for each representation are reported in Table 9.

### 5.2 Logistic regression

Logistic regression is a widely applied algorithm for binary classification tasks in machine learning. It offers interpretable outcomes and adapts effectively to numerical feature representations. In this study, its hyperparameters were configured to enhance performance. Table 10 lists the hyperparameters used for the logistic regression model. The three amino acid sequence representations were evaluated to compare their effectiveness in antibody classification. The results for each representation are presented in Table 11. Among the three approaches, the representation based on Atchley Factors showed superior performance, suggesting that incorporating physicochemical properties of amino acids may contribute to improved antibody classification.

**Table 8.** Hyperparameters for Decision Trees

| Hyperparameter | Value | Description |
|---|---|---|
| *criterion* | *gini* | Function to measure the quality of a split |
| *max_depth* | 5 | Maximum depth of the tree to control overfitting |
| *random_state* | 42 | Seed value for random number generation to ensure reproducibility |
| *splitter* | best | Strategy to choose the split at each node |
| *min_samples_split* | 2 | Minimum number of samples required to split an internal node |
| *min_samples_leaf* | 1 | Minimum number of samples required to be at a leaf node |
| *min_weight_fraction_leaf* | 0 | Minimum weighted fraction of total samples required in a leaf node |
| *max_features* | None | Number of features to consider when looking for the best split |
| *max_leaf_nodes* | None | Maximum number of leaf nodes |
| *min_impurity_decrease* | 0 | A node will be split only if impurity decrease is greater than this value |
| *class_weight* | None | Weights associated with classes |
| *ccp_alpha* | 0 | Complexity parameter used for pruning |

**Table 9.** Results of representations with decision trees

| Representation | Accuracy | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|
| TF-IDF | 0.86 | 0.86 | 0.86 | 0.88 |
| Atchley Factors | 0.94 | 0.93 | 0.93 | 0.96 |
| ProtVec | 0.80 | 0.81 | 0.82 | 0.84 |

**Table 10.** Hyperparameters for Logistic regression

| Hyperparameter | Value | Description |
|---|---|---|
| *penalty* | l2 | Regularization type |
| *solver* | lbfgs | Algorithm used for optimization |
| *max_iter* | 200 | Maximum number of iterations for convergence |
| *random_state* | 42 | Seed value for random number generation to ensure reproducibility |
| *C* | 1 | Inverse of regularization strength |
| *tol* | 0.0001 | Stopping tolerance for optimization convergence |
| *fit_intercept* | True | Whether to calculate the intercept for the model |
| *intercept_scaling* | 1 | Scaling factor for intercept term |
| *warm_start* | False | Reuse previous solution to warm start |

**Table 11.** Logistic regression results

| Representation | Accuracy | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|
| TF-IDF | 0.85 | 0.84 | 0.84 | 0.87 |
| Atchley Factors | 0.86 | 0.85 | 0.85 | 0.88 |
| ProtVec | 0.80 | 0.79 | 0.78 | 0.81 |

## 5.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm used for classification tasks, particularly when dealing with high-dimensional datasets. The model was configured with appropriate hyperparameters to optimize its performance and evaluate its ability to distinguish between SARS-CoV-2 and non-SARS-CoV-2 antibodies. Table 12 presents the hyperparameters used in the SVM model. Table 13 presents the comparison of results for the different representations using the SVM algorithm.

**Table 12.** Hyperparameters for SVM

| Hyperparameter | Value | Description |
|---|---|---|
| *C* | 1 | Regularization parameter: higher values reduce regularization |
| *kernel* | rbf | Kernel type used in the algorithm |
| *degree* | 3 | Degree of the polynomial kernel function |
| *gamma* | scale | Kernel coefficient |
| *coef0* | 0 | Independent term in kernel function |
| *shrinking* | True | Whether to use the shrinking heuristic for optimization |
| *probability* | False | Enables probability estimates |
| *tol* | 0.001 | Tolerance for stopping criterion |
| *cache_size* | 200 | Size of the kernel cache in MB |

**Table 13.** SVM results

| Representation | Accuracy | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|
| TF-IDF | 0.87 | 0.86 | 0.86 | 0.89 |
| Atchley Factors | 0.88 | 0.87 | 0.87 | 0.90 |
| ProtVec | 0.82 | 0.80 | 0.79 | 0.83 |

## 6 Discussion

The superior performance observed for the Atchley Factor representation, relative to the other two approaches, can be attributed to its ability to encode the physicochemical properties of the amino acids that constitute antibody structures, combined with the low dimensionality of the resulting feature space. This compact representation is likely to facilitate improved model performance by reducing noise and limiting overfitting. The TF–IDF representation also yields competitive results; however, this technique does not preserve contextual or structural relationships within amino acid chains. By contrast, the ProtVec-based representation exhibited the lowest performance, which may be associated with the relatively high dimensionality it generates (100 dimensions per sequence for each chain), potentially increasing model complexity and variance.

Although the numerical results indicate that the Atchley-based representation achieves the highest overall performance, particular attention must be paid to the reproducibility of ProtVec-based features. As ProtVec embeddings rely on a pretrained model and the internal tokenisation of amino acid triplets, minor variations can occur unless a fixed random seed and a consistent preprocessing pipeline are enforced. In the present experiments, the random seed was fixed and input segmentation was standardised, which appeared to contribute to more stable output vectors across runs.

In addition to the tabulated results, Fig. 5 provides a visual comparison of model performance across representations, with error bars indicating variability over multiple undersampling iterations. This visualisation reinforces the numerical findings by highlighting the higher consistency and accuracy associated with Atchley-based features. The comparison summarises classification performance across different machine learning models when applied to amino acid sequence representations using the three evaluated methods. These results underscore the influence of representation choice on model accuracy and inform the selection of suitable approaches for antibody sequence classification tasks. Table 14 presents a consolidated comparison for each representation and highlights the best-performing models within each category.

When comparing model performance across representations, the Support Vector Machine classifier exhibited greater robustness and consistency, particularly when applied to high-dimensional representations such as TF–IDF and ProtVec. This observation is consistent with prior studies that highlight the capacity of SVMs to generalise effectively in complex feature spaces when appropriate regularisation is applied. In contrast, Decision Tree classifiers showed higher variability in performance, which may reflect their sensitivity to redundant or noisy features—an issue that is more pronounced in sparse representations such as TF–IDF. Logistic Regression demonstrated an intermediate performance profile, performing particularly well with more compact and interpretable encodings such as Atchley Factors. Collectively, these results suggest that classification effectiveness is not determined solely by model architecture, but also by its compatibility with the chosen feature representation. These findings emphasise the importance of jointly selecting representation strategies and learning algorithms in future antibody modelling applications.

**Table 14.** Comparison of representation of amino acid sequences

| Representation | ML Models | Accuracy | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|---|
| TF-IDF | Decision trees | 0.86 | 0.86 | 0.86 | 0.88 |
| | Logistic regression | 0.85 | 0.84 | 0.84 | 0.87 |
| | **SVM** | **0.87** | **0.86** | **0.86** | **0.89** |
| Atchley Factors | **Decision trees** | **0.94** | **0.93** | **0.93** | **0.96** |
| | Logistic regression | 0.86 | 0.85 | 0.85 | 0.88 |
| | SVM | 0.88 | 0.87 | 0.87 | 0.90 |
| ProtVec | Decision trees | 0.80 | 0.81 | 0.82 | 0.84 |
| | Logistic regression | 0.80 | 0.79 | 0.78 | 0.81 |
| | **SVM** | **0.82** | **0.80** | **0.79** | **0.83** |

Another aspect to consider is the class-balancing strategy employed. Using undersampling to equalise the number of records in each class helped mitigate bias towards the majority class. However, this approach inevitably reduced the amount of available data, which may constrain the model’s capacity to capture more complex patterns within the dataset. This limitation could potentially be addressed by exploring alternative strategies, such as generating synthetic samples for the minority class, which would increase the volume of information available for learning while preserving class balance.

When analysing the amino acid sequences of the antibody chains, it was observed that variability in sequence length—particularly within the complementarity-determining regions—poses a significant challenge for sequence representation. As expected, CDR3 exhibited the greatest diversity in both length and amino acid frequency. This observation reinforces its recognised role in determining antibody affinity and specificity.

Sequence-based representations alone do not incorporate structural or functional information, such as molecular interactions, conformational flexibility, or mutations that influence antibody activity. This limitation reduces the ability of models to identify complex relationships within the data. Data quality also represents an important consideration, as antibody sequence repositories may contain duplicate entries or incomplete annotations, which can adversely affect model performance. Such issues are likely related to differences in sequencing techniques and curation practices.

From a methodological perspective, this comparative evaluation of sequence representations provides a generalisable framework for bioinformatics research. The approach adopted here—evaluating representations within a consistent classification pipeline—can be extended to other tasks, including epitope prediction, protein–protein interaction modelling, and neoantigen identification in cancer immunotherapy. By isolating the effect of representation strategies, researchers can make more informed decisions regarding data preprocessing, thereby improving model performance and interpretability across diverse applications.

Given that amino acid sequence representation constitutes a critical preprocessing step, it is essential to identify techniques that preserve biologically relevant information. Doing so is expected to contribute positively to antibody analysis and design in biomedical contexts.

Moreover, the findings of this study carry practical implications for biomedical research. Accurate antibody sequence representations can enhance diagnostic pipelines by enabling the rapid identification of pathogen-specific antibodies, which is particularly important in the context of emerging infectious diseases. In addition, the methodologies explored here may support the rational design of therapeutic antibodies within drug development workflows. Beyond SARS-CoV-2, these approaches can be adapted to other pathogens, providing a robust framework for broader applications in immunology and bioinformatics. By improving the efficiency and reliability of antibody classification, this work has the potential to facilitate advances in immunodiagnostics, vaccine design, and precision medicine.

The TF–IDF representation assigns greater weight to less frequent terms in order to highlight distinctive amino acids within sequences, which can facilitate discrimination between antibody classes. It also offers the advantage of simplicity in both implementation and interpretation. However, it presents notable limitations. Most importantly, TF–IDF does not account for contextual or semantic relationships between amino acids. In addition, when applied to large corpora, the resulting feature space tends to be highly sparse, which can reduce the efficiency of certain classification algorithms.

Atchley Factors yielded the strongest performance across all models evaluated in this study. Nevertheless, this representation also involves limitations that must be considered. Dimensionality reduction may lead to the loss of information that could be relevant in specific analytical contexts. Furthermore, Atchley Factors treat amino acids independently, which means that interactions among residues within a sequence are not explicitly modelled.

Regarding feature importance, both Decision Tree and Logistic Regression models enabled the examination of influential features. In particular, Atchley Factor 3 from the CDR3 region consistently emerged as one of the most relevant predictors. This finding may be informative for immunologists interested in sequence–function relationships and for researchers designing feature-driven antibody screening pipelines. Future studies could extend this analysis by incorporating explainability techniques such as SHAP or LIME to provide deeper insights into individual model decisions.

In summary, the comparative analysis indicates that Atchley Factor–based representations perform effectively in low-dimensional settings, while TF–IDF and ProtVec offer higher expressiveness but require careful handling due to sparsity and complexity. Among the classifiers, the Support Vector Machine demonstrated robust generalisation across representations. A key methodological challenge involved ensuring fair evaluation across representations with differing dimensional characteristics.

## 7 Conclusions and future work

This study evaluated the effectiveness of three amino acid sequence encoding strategies—TF–IDF, Atchley Factors, and ProtVec—for antibody classification, using supervised learning algorithms including Decision Trees, Logistic Regression, and Support Vector Machines. The results demonstrate that the choice of representation plays a critical role in determining predictive performance.

The experiments show that preserving sequential context alone, as achieved by ProtVec, is not sufficient to guarantee superior classification. This suggests that representations may benefit from the integration of physicochemical features. In contrast, TF–IDF does not preserve amino acid order, which limits its ability to capture sequential patterns and structural motifs essential to antibody function. ProtVec, while retaining sequence order, exhibited the lowest performance in this study, which is likely related to its high dimensionality (600 features per record), increasing the risk of overfitting and reduced generalisation.

Undersampling was employed to balance classes and avoid bias towards non–SARS-CoV-2 antibodies. While effective, this approach reduced the volume of training data and may have limited the detection of complex patterns. Future work could explore synthetic data generation techniques, such as SMOTE, or cost-sensitive learning to mitigate this issue.

Variability in CDR length was identified as a further challenge, as machine learning models require fixed-length inputs. CDR3, in particular, exhibited substantial variability in length and amino acid composition, complicating feature normalisation.

Overall, the findings highlight the importance of selecting appropriate sequence representations to enhance classification performance. This work provides practical guidance for antibody modelling and offers potential benefits for biomedical applications, including therapeutic design and immunodiagnostics. Future research will explore hybrid representations that combine physicochemical features with advanced embeddings, as well as clustering approaches to analyse binding patterns and structural similarities between antibodies, supporting affinity prediction and antibody generation.

## References

Abbas, A. K., Lichtman, A. H., & Pillai, S. (2021). *Cellular and molecular immunology* (10th ed.). Elsevier.

Asgari, E., & Mofrad, M. R. K. (2015). ProtVec: A continuous distributed representation of biological sequences. *PLoS ONE*, 10(11), e0141287. https://doi.org/10.1371/journal.pone.0141287

Atchley, W. R., Zhao, J., Fernandes, A. D., & Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences*, 102(18), 6395–6400. https://doi.org/10.1073/pnas.0408677102

Birunda, S. S., & Devi, R. K. (2021). A review on word embedding techniques for text classification. In J. S. Raj, A. M. Iliyasu, R. Bestak, & Z. A. Baig (Eds.), *Innovative data communication technologies and application* (pp. 267–281). Springer. https://doi.org/10.1007/978-981-15-9651-3_23

Chen, X., Dougherty, T., Hong, C., Schibler, R., Zhao, Y. C., Sadeghi, R., Matasci, N., Wu, Y.-C., & Kerman, I. (2020). Predicting antibody developability from sequence using machine learning. *bioRxiv*. https://doi.org/10.1101/2020.06.18.159798

Greiff, V., Yaari, G., & Cowell, L. G. (2020). Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, 24, 109–119. https://doi.org/10.1016/j.coisb.2020.10.010

Ibero-American Cooperative Group on Transfusion Medicine. (2020). *Basic and applied immunohematology*. GCIAMT.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing* (3rd ed.). Pearson.

Leem, J., Mitchell, L. S., Farmery, J. H. R., Barton, J., & Galson, J. D. (2022). Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7), Article 100513. https://doi.org/10.1016/j.patter.2022.100513

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., … & Lefranc, G. (2009). IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, 37(Database issue), D1006–D1012. https://doi.org/10.1093/nar/gkn838

Li, L., Gupta, E., Spaeth, J., Shing, L., Bepler, T., & Caceres, R. S. (2022). Antibody representation learning for drug discovery. *arXiv*. https://doi.org/10.48550/arXiv.2210.02881

Li, X., Van Deventer, J. A., & Hassoun, S. (2020). ASAP-SML: An antibody sequence analysis pipeline using statistical testing and machine learning. *PLOS Computational Biology*, 16(4), e1007779. https://doi.org/10.1371/journal.pcbi.1007779

Magar, R., Yadav, P., & Barati Farimani, A. (2021). Potential neutralizing antibodies discovered for novel coronavirus using machine learning. *Scientific Reports*, 11(1), Article 5261. https://doi.org/10.1038/s41598-021-84637-4

Murphy, K. M., Weaver, C., & Berg, L. J. (2022). *Janeway's immunobiology* (10th ed.). W. W. Norton & Company.

Olsen, T. H., Boyles, F., & Deane, C. M. (2022). Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1), 141–146. https://doi.org/10.1002/pro.4205

Parham, P. (2021). *The immune system* (5th ed.). W. W. Norton & Company.

Pulendran, B., & Davis, M. M. (2020). The science and medicine of human immunology. *Science*, 369(6511), eaay4014. https://doi.org/10.1126/science.aay4014

Punt, J., Stranford, S. A., Jones, P., & Owen, J. (2020). *Kuby immunology* (8th ed.). McGraw-Hill.

Raybould, M. I. J., Kovaltsuk, A., Marks, C., & Deane, C. M. (2021). CoV-AbDab: The coronavirus antibody database. *Bioinformatics*, 37(5), 734–735. https://doi.org/10.1093/bioinformatics/btaa739

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., & Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1), Article 1728. https://doi.org/10.1038/s41467-022-29268-7

Yadav, D., Yadav, N., Kumar, A., Sharma, P., & Sood, D. (2022). Probing the immune system dynamics of the COVID-19 disease for vaccine designing and drug repurposing using bioinformatics tools. *Immuno*, 2(2), 172–191. https://doi.org/10.3390/immuno2020022

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1), Article 52. https://doi.org/10.1038/s41597-019-0055-0