# Supplementary Materials
# NABP-LSTM-Att: Nanobody-Antigen Binding Prediction using bidirectional LSTM and soft attention mechanism

## 1 Data and Preprocessing

A variety of nanobody structures and their corresponding antigens are compiled in the SAbDab-nano database. As of September 2024, the database comprises 1608 complexes. We selected nanobodies based solely on heavy chain information and excluded complexes with antigenic sequences shorter than 50 amino acids. Following the application of this restriction, 1569 nanobody-antigen complexes were retained. Subsequently, we selected structures with a resolution superior to 3 Å, where the antigen type is either a protein or a peptide, and each antibody binds exclusively to a single antigen, with the reverse also being applicable. Following the application of these rules, 206 nanobody-antigen complexes were retained.

Due to the high specificity of nanobodies, we removed sequence redundancy by applying CD-HIT with a strict sequence identity threshold of 0.98, yielding 197 complexes. To further refine based on antigen specificity, we categorized these 197 complexes into 150 subgroups according to antigen sequences, using a 0.90 sequence identity threshold. This approach assumes that similar antibodies within each subgroup could interact effectively with similar antigens, while antibodies and antigens from different subgroups would likely not bind effectively. This process generated 1152 nanobody-antigen pairs, resulting in 1349 positive samples. To create negative samples, antigens and nanobodies from different subgroups were randomly paired, maintaining a 1:1 ratio of positive to negative samples. (Fig 1A).

Given the significant variation in the sequences of nanobodies that bind to different antigens, models may exhibit suboptimal performance in predicting antigens that differ markedly from those utilized during training. ClustalOmega was used to construct a phylogenetic tree for the antigens within the 150 subgroups, which were then organized into five clusters (Fig 1B). To prevent bias due to antigen type variability, we divided each cluster into a training set and an independent test set at a 4:1 ratio. Each nanobody sequence comprises three CDR regions (CDRH1, CDRH2, CDRH3), and the proposed model predicts the binding interactions between each CDR region and the antigen independently. Consequently, the final dataset consists of 8,094 pairs,

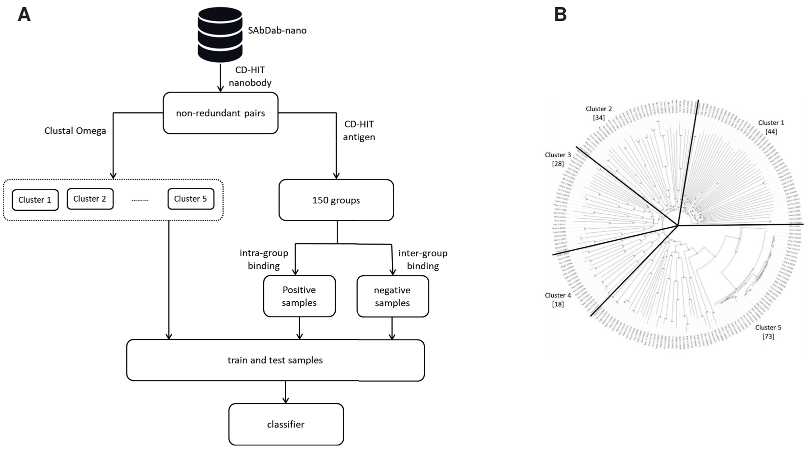

**Fig. 1**: Overview of NABP.(A) Flowchart of the prediction model of NABP. (B) Clustering tree of 197 antigens in 150 subgroups.

with half classified as positive and the other half as negative. 5% of the training set is allocated for validation purposes. The mean and median lengths of the nanobody CDR sequences are 8.7 and 7.0, whereas those of the antigen sequences are 529.7 and 438.0, respectively. The maximum lengths of the CDR and antigen sequences are 24 and 2371, respectively (see Fig 2).

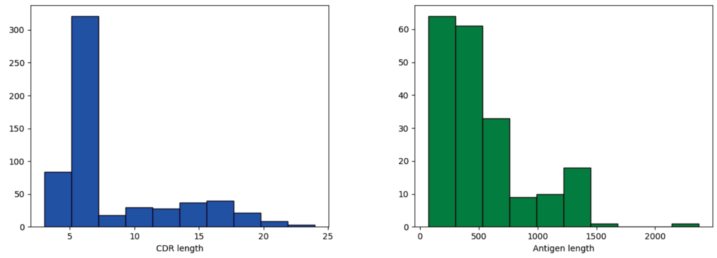

**Fig. 2**: Histograms for CDR and antigen length.