User Guide

1. Introduction

1.1 Model Overview

Generative adversarial network (GAN) has been successfully applied to the generation of functional protein sequences. However, traditional GAN often suffer from inherent randomness, resulting in a lower probability of obtaining the desired sequences. Moreover, due to the high cost of wet lab experiments, the primary objective of computer-aided antibody optimization is to filter a few high-quality candidate antibodies from a vast search space. Therefore, improving the ability of GAN to generate the desired antibodies is a pressing challenge. In this study, we propose a Language Model Guided Antibody Generative Adversarial Network (AbGAN-LMG) model. This model utilizes the embedding representations acquired from language model as input, thereby harnessing the powerful representational capabilities of language model to enhance its ability to generate high-quality antibodies.

The workflow of this study is shown in Figure 1, which consists of three main steps: Training Model, Generating Antibody Library, and Evaluating Generated Library and Sequences.

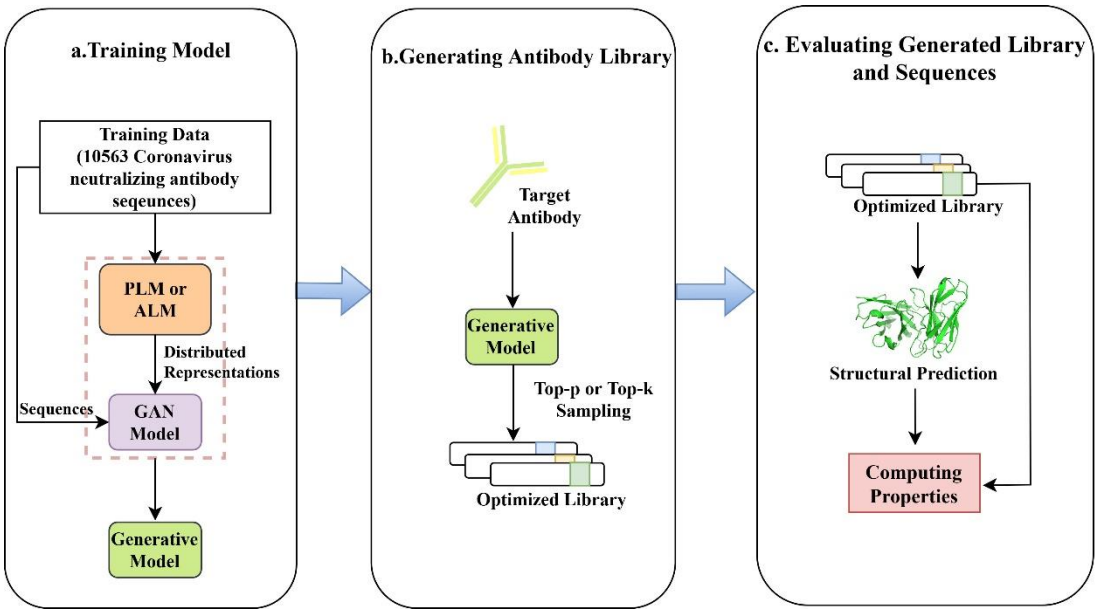


Figure 1. Workflow. a. The processed data serves as input to extract distributed representations using a PLM or ALM. distributed representations and sequence information are then utilized as input for training the generative model. **b.** The distributed representation of the target antibody sequence is extracted using a PLM or ALM, which is employed as input to the GAN model. The generative model produces sequences that are sampled using either the Top-p or Top-k sampling method. **c.** The physicochemical properties, developability and affinity of optimized antibodies are evaluated.

The training data was sourced from the CoV-AbDab database, which aggregates information on 12,021 anti-coronavirus antibodies, including both light chain and

heavy chain variable region sequences. Sequences in the database that exceeded a length of 128 amino acids or contained non-standard amino acids were excluded, resulting in a total of 11,205 antibody sequences remaining. Given the presence of relatively high sequence homogeneity within the dataset, we used MMseq2 [31] to cluster the antibody sequences at a 70% similarity threshold, a strategy employed to enhance the diversity of the training data. In clusters with fewer than 3 sequences, one sequence was randomly selected, while in other clusters, 5% of the sequences were randomly chosen to constitute the test set (642 sequences). The remaining sequences formed the training set (10,563 sequences).

We leveraged language models to obtain distributed representations for each antibody sequence in the dataset, using these representations as feature vectors for input to GAN model. ESM2-150M, ProtBERT, BERTAb2D, AntiBERTy, and AbLang were employed to characterize antibody sequences to assess the performance of different language models in antibody optimization tasks.

1.2 Applicable Scenarios

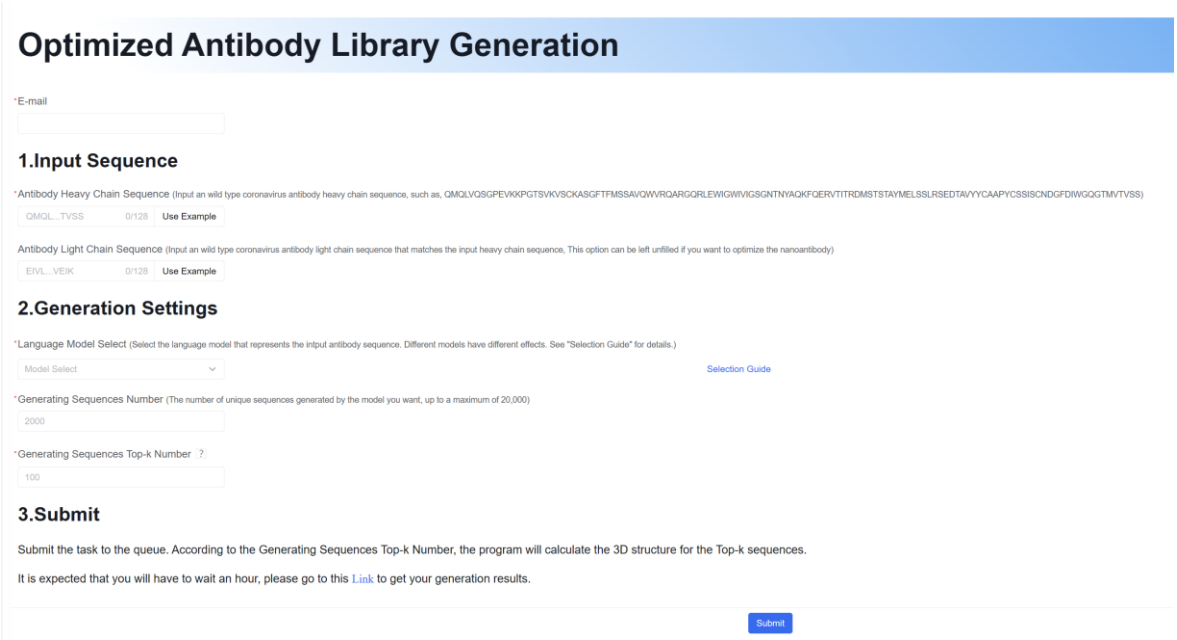
This software is designed for researchers, bioengineers, and biotechnology companies aiming to generate antibody libraries with superior development potential. The intended audience for this user manual includes individuals with a fundamental understanding of biology and laboratory skills, such as researchers and professionals in the field of biology. You can use this software to generate antibody libraries and evaluate the development of generated libraries, and we will demonstrate the specific steps below.

2. Antibody Library Generation

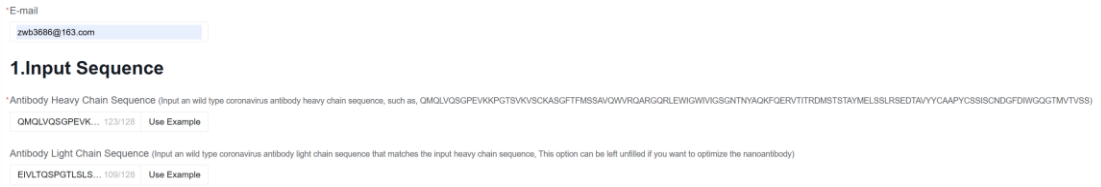
2.1 Data Input

2.1.1 Inputting Antibody Sequences

1. You need to click on "Antibody Library Genertion" on the main page to access the antibody library generation page.



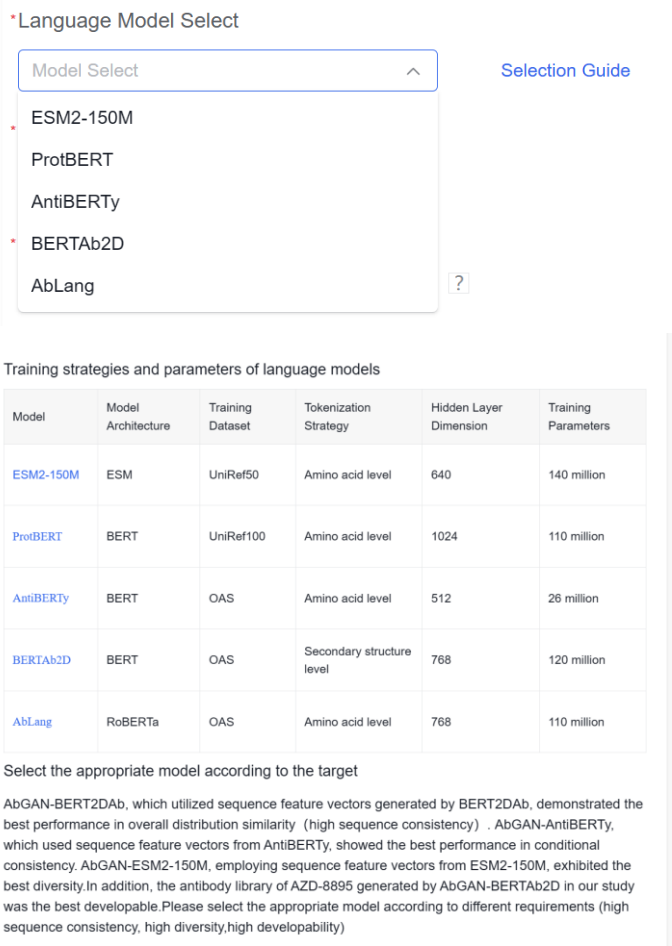
2. To begin, please provide your email address as your unique authentication identifier. Subsequently, enter the sequence of the antibody's Fv region that you intend to use for library generation. Please note that the sequence length should not exceed 128 amino acids. Here we provide you with a sample, you can click the Use sample button for AZD-8895 antibody library generation. If you require library generation for nanobodies, there is no need to input the light chain sequence.



2.1.2 Parameter Configuration

In this section, you are required to set three parameters essential for the generation process. These parameters include the choice of language model, the number of generated sequences, and the top-k sequences to be generated.

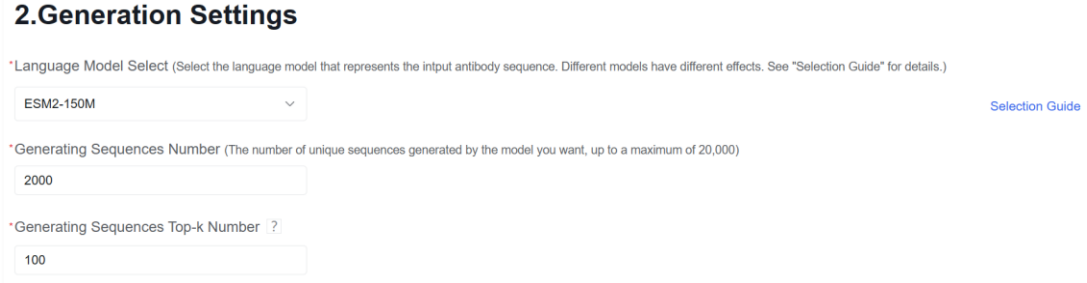
Language Model Select: There are five available language models to select from for representing your input sequences. The choice of the model depends on your specific requirements, such as the need for high diversity in the library or a focus on improved developability. To assist you in this selection, we offer a guidance button that provides recommendations.



Generating Sequences Number: It pertains to the quantity of generated sequences. It denotes the number of distinct antibody sequences that the model will generate. Please note that due to resource limitations, the maximum number of sequences that can be generated is capped at 20,000.

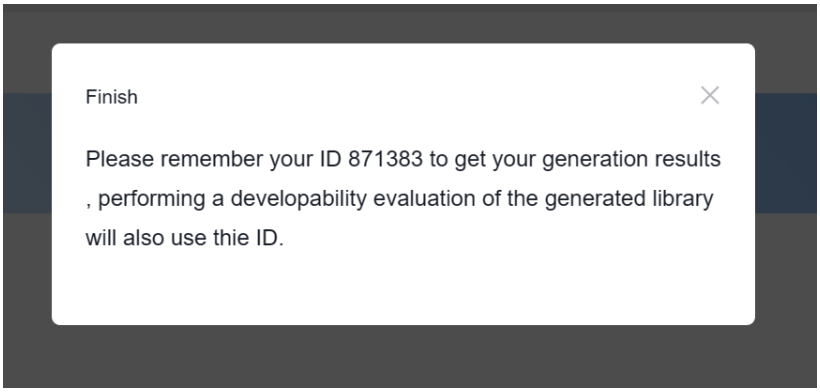
Generating Sequences Top-k Number: It involves the selection of the top-k sequences generated. During the antibody library generation process, the model may produce duplicate sequences. In such cases, we record the frequency of occurrence of these duplicate sequences and present you with the top-k sequences, which are the ones the model generates most frequently. Additionally, we calculate the three-dimensional structure of these sequences for your reference. It is important to be mindful that the computation of three-dimensional structures is a time-consuming process. Therefore, we recommend keeping the value for this parameter below 300, as exceeding this limit may result in longer processing times for obtaining your results.

Here we select these configurations to test.

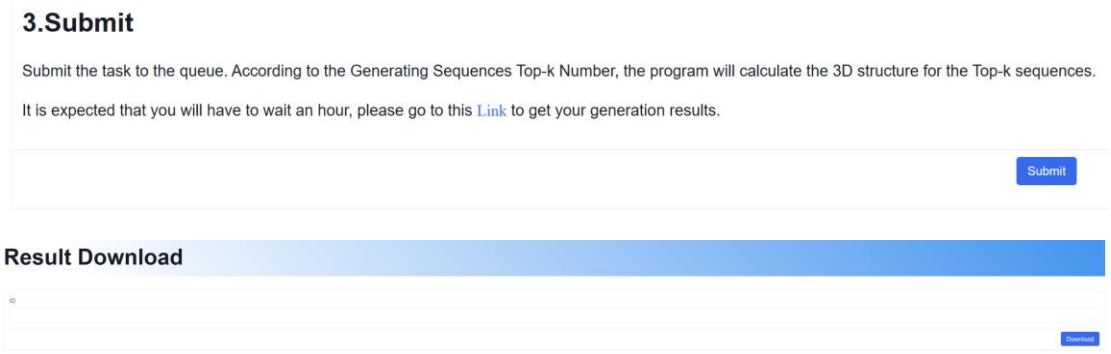


2.3 Results

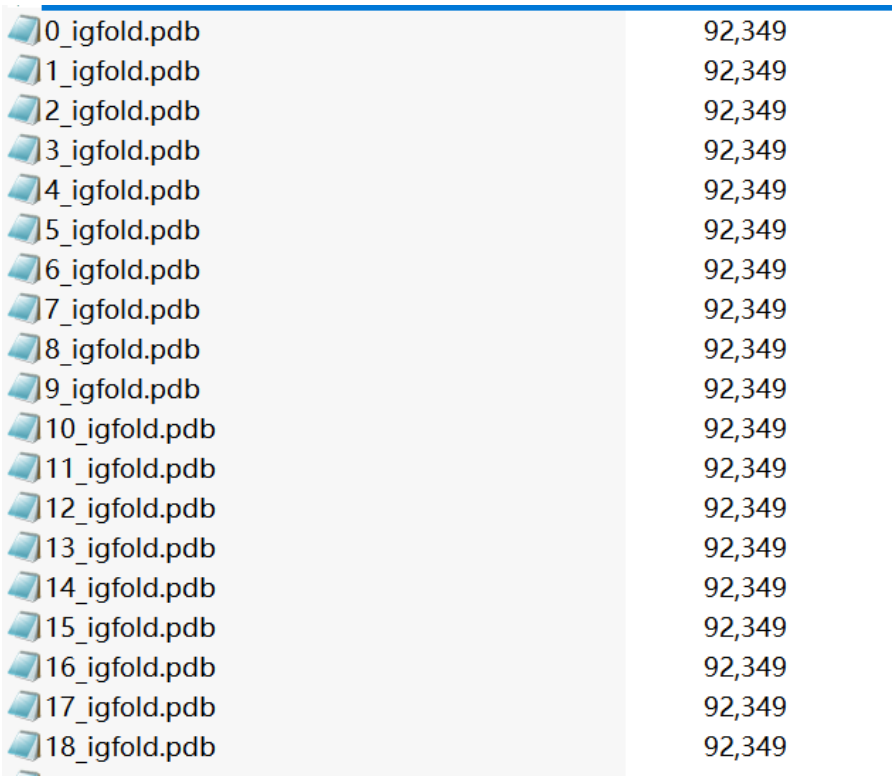
When you submit, there will be a pop-up where you need to record the ID number.



Click Link under the Submit text box to jump to the result download interface, after a certain amount of time, enter your ID number to get the generated result.



The result is a zip file that contains three parts: igfold_result.tar.gz, TargetAntibody_Library_871383.fasta, and TargetAntibody_Library_top-k871383.fasta.



3. Library Developability Evaluation

3.1 Parameter Configuration

To initiate the process, you are required to input the unique ID number obtained during the antibody library generation. The generated antibody library from the previous step will undergo a thorough assessable evaluation for developability. However, please be advised that due to limitations in computational resources, we do not offer unrestricted uploading of custom libraries.

In addition, in the calculation settings, there are two main categories, Biophysical Properties and Developability. The following six indicators are described in detail.

Isoelectric Point:

The isoelectric point, often abbreviated as pI, refers to the pH value at which a substance (such as proteins or amino acids) in a solution reaches a state of charge neutrality. At the isoelectric point, the substance exhibits a net charge of zero, meaning that the ionization levels of acidic and basic groups within the molecule are equal, resulting in the cancellation of negative and positive charges, thereby rendering the molecule in a neutral state.

For large molecules like proteins and amino acids, their charge state in a solution depends on the ionization levels of their amino acid residues. In solutions with varying pH values, protein or amino acid molecules exhibit different levels of ionization for acidic and basic groups, leading to distinct electric charges. When the pH value equals the isoelectric point, the ionization levels of acidic and basic groups within the molecule are balanced, resulting in a charge-neutral state of the molecule.

Hydrophobicity:

Antibody hydrophobicity refers to the quantity and distribution of hydrophobic groups within the antibody molecule. Hydrophobic groups typically consist of non-polar, uncharged chemical moieties that tend to aggregate with other hydrophobic groups, forming hydrophobic regions, thereby reducing contact with polar solvents. In the context of antibody antigen-binding sites, hydrophobicity plays a crucial role in recognizing and binding antigen molecules. This is because many key binding sites in antigen molecules also exhibit higher hydrophobicity.

Aggregation Propensity:

Protein aggregation refers to the property of proteins to interact with each other and form aggregates or precipitates under specific conditions such as high concentration, low pH, high temperature, and others. This aggregation tendency can lead to the loss of a protein's biological activity or stability, thereby affecting its applications in pharmaceuticals, biotechnology, and biomedical research, among other fields. Protein aggregation is typically undesirable as it can result in product quality issues and adverse clinical reactions. Therefore, in protein engineering and biopharmaceuticals, studying and controlling protein aggregation is of paramount importance.

Solubility:

CamSol intrinsic solubility profile, or solubility of the unfolded state. The CamSol method yields a solubility profile (one score per residue in the protein sequence) where regions with scores larger than 1 denote highly soluble regions, while scores smaller than -1 poorly soluble ones.

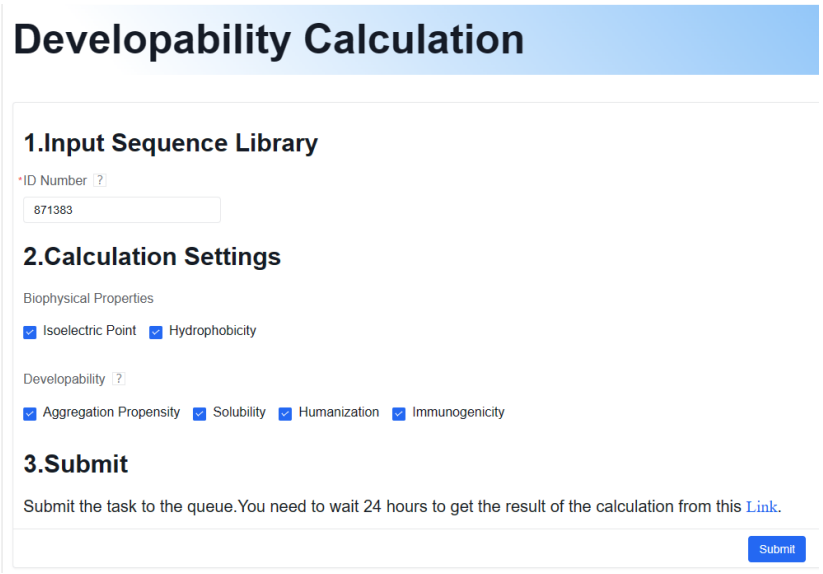
An overall solubility score will be assigned to the whole sequence. This score can be used to rank with high accuracy different protein variants (i.e. proteins with similar

sequences) according to their solubility. It may however perform poorly when ranking widely different proteins.

Humanization:
The humanization score refers to the similarity between the amino acid sequence of an antibody and the corresponding sequence found in human antibodies. Antibodies derived from non-human sources, such as mice or rabbits, may trigger immune reactions in the human body, leading to potential complications in clinical applications.

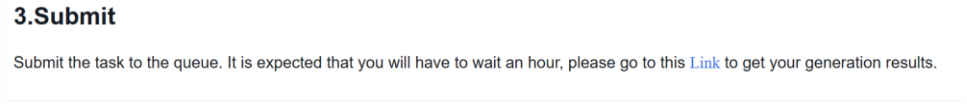
Immunogenicity:
Antibody immunogenicity refers to the ability of an antibody (immunoglobulin) to induce an immune response within the human body. Major Histocompatibility Complex (MHC) molecules play a crucial role in the immune system by presenting antigens to facilitate their recognition, subsequently triggering immune responses. During immune reactions, peptide segments bind to MHC molecules, forming peptide-MHC complexes. These complexes are then presented to T cells, and if T cells can recognize and bind to these complexes, it leads to the initiation of an immune response. The immunogenicity of antibodies is closely associated with their susceptibility to recognition and clearance by the human immune system.

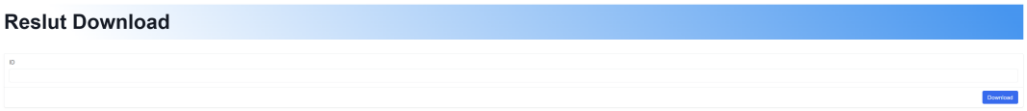
Here we select all the indicators for calculation.



3.2 Results

Click Link under the Submit text box to jump to the result download interface, after a certain amount of time, enter your ID number to get the generated result.





The result is a xlsx file that contains the exploitability calculation item of your choice, and in addition, it has a Total_Sum column which is the score function for the increase in exploitability of the produced library compared to the original library, the sequence of exploitability generated with regular scores is better than the original sequence.

	A	B	C	D	E	F	G	H	I
1	Name	VH	Isoelectric Point	Hydrophobicity	Aggregation Propensity	CamSol Score	Humanization	Immunogenicity	Total_Sum
2	Generated Sequence_71	QVQLVQSGPEVKK	8.337132072	0.551421037	-94.4754	-0.273486	0.800592593	74.70674632	6.778908441
3	Generated Sequence_6	QVQLVQSGPEVKK	8.416428185	0.567200242	-95.9412	-0.249365	0.813666667	74.9778125	6.138550473
4	Generated Sequence_91	QMQLVQSGPEVK	8.416428185	0.572938689	-95.5009	-0.224722	0.78762963	74.9778125	5.406845386
5	Generated Sequence_25	QMQLVQSGPEVN	7.830475044	0.572325071	-101.5336	-0.191101	0.827555556	75.16948529	4.584144962
6	Generated Sequence_2	QVQLVQSGPEVKK	8.416428185	0.553492485	-94.395	-0.178958	0.800592593	75.06422794	4.128142393
7	Generated Sequence_92	QVQLVQSGPEVKK	8.416428185	0.542204866	-97.0051	-0.169333	0.920333333	79.96944853	4.112232413
8	Generated Sequence_77	QVQLVQSGPEVKK	8.416428185	0.547362285	-98.0579	-0.151429	0.813666667	74.70674632	3.407495596
9	Generated Sequence_34	QVQLVQSGPEVKK	8.416428185	0.536085536	-99.2852	-0.138948	0.827555556	80.12134191	3.161464588
10	Generated Sequence_28	QMQLVQSGPEVK	8.663406563	0.528171049	-96.0317	-0.130164	0.933814815	76.39205882	2.971772084
11	Generated Sequence_9	QMQLVQSGPEVK	8.416428185	0.548901595	-95.368	-0.11827	0.933814815	75.16948529	2.614158965

4. Sample Demonstrations

We conducted a comprehensive evaluation of the antibody libraries and sequences generated by AbGAN-LMG for the novel coronavirus (SARS-CoV-2) and Middle East Respiratory Syndrome (MERS-CoV).

4.1 Antibody Library Generation Examples

4.1.1 Anti-SARS-CoV-2 Neutralizing Antibody Library Generation

AZD-8895 was used as the input for AbGAN-LMG to generate a library (2000 sequences). AZD-8895 was chosen as the target antibody for optimization to assess the performance of AbGAN-LMG in the context of antibody optimization scenarios (where specific antibodies need to be improved), which is a monoclonal antibody targeting the RBD of the SARS-CoV-2 spike protein, designed to against the SARS-CoV-2. In this specific scenario, only the sequence of AZD-8895 was embedded as a distributed representation using language model.You can Download the generated AZD-8895 library.

1	VH	Isoelectric Point	Hydrophobicity	CamSol Score	Heavy OASis Per	Average NetMHCIIpan	Aggregation Propensity
2	QMQLVQSGI	8.42738781	0.549716333	−0.035608	0.746833333	77.00659926	−92.1466
3	QMQLVQSGI	8.42738781	0.519453519	−0.001318	0.643	79.28702206	−104.4204
4	QMQLVQSGI	8.42738781	0.544042998	−0.066198	0.78762963	77.00659926	−96.6293
5	QMQLVQSGI	8.42738781	0.535383697	0.06671	0.746833333	77.00659926	−83.7718
6	QMQLVQSGI	8.42738781	0.549716333	−0.035608	0.746833333	77.00659926	−100.2291
7	QMQLVQSGI	8.42738781	0.529710361	0.046226	0.705046512	77.69761029	−103.8235
8	QMQLVQSGI	8.42738781	0.510840511	0.13159	0.746833333	79.28702206	−98.6387
9	QMQLVQSGI	8.42738781	0.520690682	−0.036414	0.693962963	78.59336397	−89.4969
10	QMQLVQSGI	8.42738781	0.526347127	−0.005825	0.746833333	78.59336397	−92.2325
11	QMQLVQSGI	8.42738781	0.518222222	0.079646	0.746833333	77.69761029	−92.2531
12	QMQLVQSGI	8.42738781	0.562894026	−0.016718	0.599740741	77.69761029	−102.4449
13	QMQLVQSGI	8.42738781	0.506450645	0.220854	0.733407407	80.41959559	−103.7973
14	QMQLVQSGI	8.42738781	0.524452445	0.088119	0.6555	80.41959559	−105.5063

ESM2-150M | BERT2DAb | ProtBERT | AntiBERTy | AbLang

4.1.2 Anti-MERS-CoV Neutralizing Nanobody Library Generation

Additionally, we applied AbGAN-LMG to generate and evaluate a library of nanobody VHH-01 against MERS-CoV. You can Download the generated VHH-10 library.

	A	B	C	D	E	F
1	VHH	soelectric Poin	Hydrophobicit	CamSol Score	Aggregation Propen	Average Ne
2	EVVLQESGC	8. 45929966	0. 37377388	1. 212815	−84. 1176	58. 2123
3	EVELVESGC	9. 40105381	0. 37063472	1. 567552	−73. 4301	55. 3205
4	EVELCESGC	7. 83111973	0. 34278351	1. 58813	−75. 1038	64. 4
5	EVELVESGC	8. 38645039	0. 43460034	1. 286735	−73. 6375	69. 1
6	EVELCESGC	8. 31695347	0. 3553378	1. 413892	−63. 1917	59. 25
7	EVELQESGC	8. 97717419	0. 44005935	1. 038004	−76. 2888	67. 9649
8	EVVLQESGC	8. 62472553	0. 4109749	1. 37368	−71. 9645	64. 6842
9	EVELVESGC	8. 56206226	0. 38720807	1. 311613	−82. 2475	76. 1254
10	EVELQESGC	8. 9977396	0. 38362843	1. 04974	−68. 7332	69. 8626
11	EVSLQESGC	8. 84527187	0. 39580974	0. 949947	−66. 6538	64. 8366
12	EVELQESGC	8. 77687092	0. 36748927	1. 3421	−75. 3217	75. 8577
13	EVELQESGC	8. 85932598	0. 37926367	1. 364696	−77. 9531	60. 3
14	EVELQESGC	8. 72316875	0. 37894458	1. 246799	−55. 9227	62. 9539
15	EVQLQESGC	9. 40156956	0. 38266332	1. 267096	−50. 206	53. 4674
16	EVELQESGC	8. 56206226	0. 35797562	1. 389343	−73. 7939	75. 64
17	EVELCESGC	9. 07980785	0. 36250647	1. 487924	−75. 0078	57. 70
18	EVSLCESGC	8. 5534235	0. 45150894	1. 097117	−77. 7954	72. 6782
19	EVELQESGC	8. 97717419	0. 38632163	1. 148656	−68. 9471	67. 9075
20	EVNLQESGC	9. 04099789	0. 33745723	1. 566257	−66. 7247	75. 6196

ESM2-150M | AntiBERTy | ProtBERT | BERT2DAb

4.2 Library Developability Calculation Examples

4.2.1 Anti-SARS-CoV-2 Neutralizing Antibody Library Developability Calculation

when generating sequences using AZD-8895 as the target for optimization, over 50% of the generated sequences exhibited better developability than AZD-8895. Moreover, through molecular docking, we identified 70 candidate antibodies that demonstrated higher affinity for the wild-type receptor-binding domain (RBD) of the SARS-CoV-2 compared to AZD-8895.

The developability of the antibodies generated by AbGAN-LMG based on AZD-8895 is assessed. We calculated various indicators, including aggregation, solubility, humanization potential, and immunogenicity for the generated antibodies. As shown in Figure2 (a) and Table 1, based on the distribution of the bar chart and the proportion of sequences that outperformed AZD-8895, the majority of models based on AbGAN-LMG achieved over 50% of generated sequences with improved properties compared

to AZD-8895. This indicates that AbGAN-LMG can generate antibodies with higher developability than the original antibody to be optimized. Furthermore, we found that the model using BERT2DAb-guided feature vectors (AbGAN-BERT2DAb) exhibited the best or second-best performance across all the evaluated indicators. We analyzed the affinity of the antibodies generated by AbGAN-LMG based on AZD-8895 for the wild-type SARS-CoV-2 virus(RBD). As shown in Figure2 (b), AbGAN-LMG identified 70 antibodies out of 500 generated sequences with affinity superior to that of the wild-type antibody for the antigen. The top 100 sequences, ranked from highest to lowest by exploitability, can be downloaded here.

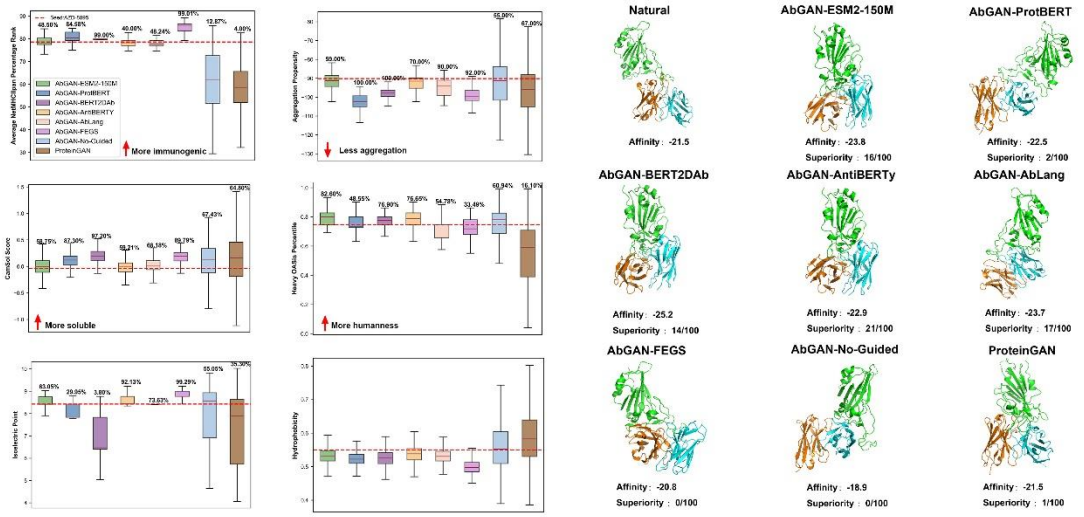


Figure 2. Evaluation of library quality of neutralizing antibodies against SARS-CoV-2. a. The amino acid conversions in the CDR of the antibody library were measured. A larger letter size indicates a higher probability of occurrence at this position. **b.** The RMSD value of the 3D structure of the generated antibody compared to the 3D structure of the wild-type antibody is displayed. The color transitions from gray to pink, where pinker represents larger RMSD value, indicating significant structural changes. The yellow box highlights cases where the 3D structure has gained or lost β-strands compared to the wild-type antibody sequence. **c.** Biophysical properties, as well as the developability of the libraries, were measured. In this context, a lower aggregation propensity value is considered better, while higher values for solubility, humanization, and immunogenicity indicators are preferred. **d.** The 3D structure of the antibody with the highest affinity among the 100 sequences is visualized and labeled with its affinity value. where the green represents the antigen, orange represents the antibody heavy chain, and blue represents the antibody light chain. Also, the count of the 100 docked antibodies with superior affinity to the original antibody is indicated.

Table 1. Developability of the generated anti-SARS-CoV-2 neutralizing antibody library

Model	Isoelectric Point	Hydrophobicity	Aggregation Propensity	CamSol Score	Heavy OASis Percentile	Average NetMHCIIpan Percentage
AbGAN-ESM2-150M	8.480±0.270	0.532±0.023	-91.466±4.439	-0.005±0.016	0.795±0.045	78.922±2.900
AbGAN-ProtBERT	7.799±0.713	0.526±0.021	-102.307±3.974	0.111±0.127	0.757±0.042	80.667±2.333

AbGAN-BERT2DAb	6.933±1.017	0.524±0.024	-98.030±2.810	0.198±0.124	0.779±0.063	79.928±0.574
AbGAN-AntiBERTy	8.575±0.256	0.538±0.025	-92.242±4.305	-0.019±0.130	0.778±0.051	78.359±1.720
AbGAN-AbLang	8.352±0.329	0.531±0.021	-94.929±4.634	0.027±0.124	0.724±0.059	78.133±1.566
AbGAN-No-Guided	7.993±1.295	0.558±0.067	-92.419±13.185	0.098±0.338	0.750±0.117	62.205±12.803
AbGAN-FEGS	8.803±0.233	0.499±0.021	-98.593±4.753	0.156±0.187	0.719±0.065	85.116±2.492
ProteinGAN	7.310±1.536	0.585±0.081	-96.048±14.335	0.4790±-0.005	0.537±0.221	59.301±10.639
Seed:AZD-8895	8.427	0.550	-90.145	-0.036	0.746	78.593

4.2.2 Anti-MERS-CoV Neutralizing Nanobody Library Developability Calculation

Additionally, we applied AbGAN-LMG to generate and evaluate a library of nanobody VHH-01 against MERS-CoV. As shown in Figure3 (a) and Table 2, remarkably, AbGAN-BERT2DAb showed lower developability in generating nanobody libraries, possibly due to BERT2DAb being trained without nanobody data, resulting in the model not learning relevant features of nanobodies. The other four AbGAN-LMG models were able to maintain the biophysical properties consistent with the wild-type antibodies while generating libraries with improved developability and identifying more nanobody sequences with higher affinity. The top 100 sequences, ranked from highest to lowest by exploitability, can be downloaded here.

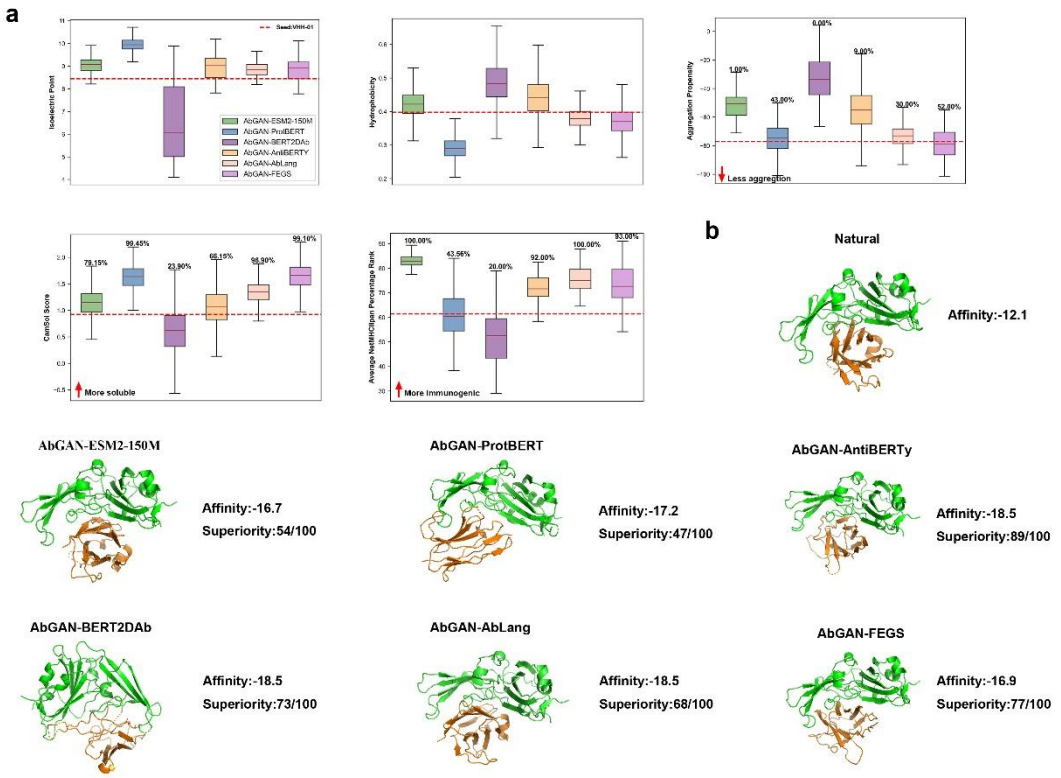


Figure 3. Evaluation of library quality of neutralizing nanobodies against MERS -CoV. a. Biophysical propertie, as well as developability properties of the libraries, were measured. In this context, a lower aggregation propensity value is considered better, while higher values for solubility, humanization, and immunogenicity indicators are preferred. **b.** The 3D structure of antigen-nanobody docking is visualized, where the green represents the antigen, orange represents the nanobody heavy chain. The antibody with the highest affinity for the antigen among the 100 sequences is selected and labeled with its affinity value. Additionally, the number of the 100 docked antibodies that had better affinity than the original antibody is indicated.

Table 2. Developability of the generated anti-MERS-CoV neutralizing nanobody library

Model	Isoelectric Point	Hydrophobicity	Aggregation Propensity	CamSol Score	Average NetMHCIIpan Percentage
AbGAN-ESM2-150M	8.966±0.510	0.421±0.041	-51.803±9.582	1.141±0.271	83.365±2.869
AbGAN-ProtBERT	9.954±0.309	0.291±0.033	-74.804±11.002	1.627±0.240	61.608±9.549
AbGAN-BERT2DAb	6.540±1.631	0.489±0.064	-32.831±15.696	0.604±0.431	52.407±10.729
AbGAN-AntiBERTy	8.717±0.965	0.447±0.059	-54.555±17.211	1.044±0.345	71.868±6.284
AbGAN-AbLang	8.723±0.618	0.379±0.031	-73.301±9.181	1.338±0.212	76.018±5.547
AbGAN-FEGS	6.540±1.631	0.373±0.043	-77.327±12.012	1.636±0.262	73.497±8.483
Seed:VHH-01	8.448	0.398	-77.001	0.925	61.425

5. Model Advanced Functionality Examples

In addition to being able to generate libraries for a single antibody, You can generate an antibody library with a similar distribution based on an existing antibody library. Due to limited computing resources, this function cannot be used in our web server, but you can use this function locally by downloading the GitHub open-source code.

After you download it locally, please follow the steps in README.md to install and deploy, when the installation is complete, You can use this code to get your generated antibody library：

```
Python Antibody_Library(fasta)_Generation.py --inputfasta your_antibody_library.fasta --num 2000 --top_num 100 --model BERT2DAb
```

Here we performed library generation and evaluation of CoV-AbDab: Antibody sequences from the CoV-AbDab were used as the input for AbGAN-LMG to generate a library (12021 sequences). In this scenario, distributed representations of each antibody sequence in the CoV-AbDab extracted by the language model are used as input to the generator. Each antibody in CoV-AbDab corresponds to a unique distributed representation and generates a new antibody, resulting in a library size of 12021. You can Download the generated antibody library.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	QVELIESGGSLVQPGGSLRLSPAISGFTLDYYAIGWFRQAAGKEREVSCIDPSGTMTTWARFVKGRFTISKDNTKNTVYLQMNLLKEEDTAMYYCAANKVTGWCGPGCWCYEWYPEYEYWGQGTEVT													
2	QVQLVESGGGVVQPGRSLRLSCAASGFTFSSYAMHWVRQAPGKGLEWVAVISYDGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCARADMTTTDDAMDVWGQGTTVTVSS													
3	EVQLQQWGAGLLKPSETLSLTCAVYGDSFSGYYWTWIRQSPGKGLEWIGQINHSGATNYNPSLMSRVTMEVDTSKNQFSLKLSSVTAADTAVYYCARGLGLDYWGQGTLVTVSS													
4	EVQLVQSGAEVKKPGESLKISCKGSGYSFTSYWIGWVRQMPGKGLEWMGAIYPGDSDTTYVRSFQGEVTISASKSISTAYLQWSSLTSAESAAMYYARRRETGYTEYFDLWGQGSVVTSSS													
5	QVQLVESGGGVVQPGRSLRLSCEASGFTFSTYHMHWVRQPPGKGLEWVAFISYDGSNYYYADSVKGRFTISRDNSKNTVYLQMNSLRAEDTALYYCARDESGQWHDGHFDYWGQGTLVTVSS													
6	EVQLQQSGSEMVRPGSSVKISCKTSGYTMTNYYEMHWVKQSHKKSLEWIGGNNPNNGSTINNQKFKGLATVTAVKSSSTAYMELRSLTSEDSAVFFCARDGYSYSYSGYWGQGTSLTVSS													
7	QVALQEAGFAEGLAGAKVINCCKGSGYTFSFNQRVWVKFSPGQGLRWMIYMFSGDGSNTPKGKLRAKDSVYADKSSATAYMQLSSLSSDSSTACFVKGRKYYDAAYGGYWGETNTVTSS													
8	QVQLQQSGPELVKPGASVKLSCKASGYTFTHYYIHWVKQRPGQGLEWIGWIYPGNGKSNYNEKFKGKATLTADKSSNTAYMQLSSLTSEDSAVYFCARGEDGAYFHYWGQGTFTTVSQ													
9	EVQLVETGGGLIQPGGSLRLSCAASGFMVSSNYMSWVRQAPGKGLEWVSVIYSGGGTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCARGTEGGTGTTGPWGQGTLVTVSS													
10	EVQLVESGGGLVQPGGSLRLSCAASGFSVSNNYMIWVRQAPGKGLEWVSIIYANGNTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCSTSNYEGSKENFNNWGQGTMVTVSS													
11	EVQLVESGGGLVQPGGSLRLSCAASGFNVSNNYMIWVRQAPGKGPEWVSVIYSGGTTYYADSVKGRFTISRDNSKNTVYLQMNSLRVEDTAVYYCARGEEGGTGGMGTWGQGTLVTVSS													
12	EVQLVESGGGLVQPGGSLRLSCAASFSFFSSYYHWVRVQPGKGLEWWARRTSSKSSTYYADDVKGRLTSSRHNSKRSALLQWNSLLAAETTTYYYSASRMDVRGNYYNHWGQGLVLVTSS													
13	EVQLVESGGALIQPGGSLRLSCAVSGFTVSSNYMSWVRQAPGKGLEWVSVIYSGGSTFYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCARGRPFGELPPFDWWGRGTLTTVS													
14	QVQLVESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIYYRGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVGAADTAVYYCARQGRGYYYYGSGSYRWAAFALDGWGQGTTV													
15	QVQLQEAGGGLVQPGGSLRLSCAASGFSFSIYTMGWYRQAPGKEREFVAIISASGDSESYADSVKGRFTISRDNSKNTVYLQMDSLRPEDTAVYYCAAPGLNYYDSTSDDDFDYWGQGTQVTVSS													
16	EVQLLESGGGLVQPGKSLRLSCAASGFTFSSYAMSWVRQAPGKGLEYVSTISARGNSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTALYYCAKSTATFGYWGQGTLVTVSS													
17	QVQLVEWGAGLLKPSETLSLTCAVYGGSLSGYYWSWIRQPPGKGLEWIGEIFHSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADVAVYYCARGLILFVGVVGMDSWGQGTTVTVSS													
18	QVQLVQSGAEVKKPGESLKISCKGSGYSFTNYWISWVRQMPGKGLEWMGIIDPSDSYTNYSPSFQGHVTISRDKSISNAYLQWSSLKASDTAMYYCARRGRFYDSGGSGNYNGGQGLLVTSS													
19	QVQLVESGGGVVQPGRSLRLSCAASGFTFSSYGMHWVRQAPGKGLEWVAVISYEGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAGYGGYYYYYGMDVWGQGTTVTVSS													
20	EVQLVESGGGLVKPRGSLRLSCAASGFNFNTYSMNWVRQAPGKGLEWVSSISGNTNSAASYNDYSDGFKTRTTRNANDLFQSLFVQMESTRICDTASYICSRSRSSSVSCTFWIPPNDCLDLWGLVT													
21	EVQLVESGGGLVQPGGSLRLSCAASGLTVSSNYMNWVRQAPGKGLEWVSVLYSGGSTFYADSVKGRFTISRHNSKNTLYLQMNSLRAEDTAVYYCARGLGDLLNNWGQGTLVTVSS													
22	QVQLVQSGAEVKKPGASVKVSCEASGYTFTNYYIHWVRQAPGQGLEWMGIENPSGDGTSYAQKFQGRVTMTRDTSTSTVYMELSSLRSEDTAVYYCARESPWNPFDPWGQGTLTTVSS													
23	QVQLVQSGTEVKRPGASVRVSCKASGYTFTDNYIHWVRQAPGQGLEWMGIINPSGGNTRYAQRFQGRVAMTRDTSTSTVYMELSSLRSEDTAVYYCARDRRSGGTDWGQGTLVTTSS													
24	QVQLVQSGAALKKPGASVKISCKASGYTFTSDYIHWVRQAPGQGLEWMGIINPSGGATRYAQKFQGRVTMTRDTSVSTFYMELTRLKSDDTAVYYCATGVLKALFDEWGQGTLVTVSS													
25	EVQLVQSGAEVKKPGASMKVSCKASGYTFTNYYMHWVRQAPGQGLEWMGIIHPSAGITSYAQKFQGRVTMTRDTSASTVYMELSSLRSDDTAVYYCAREDRKAFDWWGQGTLVTTSS													
26	QVQLVQSGAEVKKPGASVKVSCKASGYTFTSFYIHWVRQAPGQGLEWMGIIDPSGGSTSYAQKFQGRVTMTRDTSTSTVYMDLSSLRSEDTAVYYCTRDSDAGAADFFDSWGQGTLVTVSS													
27	QVQLQQPGAELVKPGASVKVSCKASGYTFTSYWMHWVKQRPGQGLEWIGEIHPSDGDTNYNQKFKVKATLTVDKSSNTAYMQLSSLTSEDSAVYYCAISDYAWWLDYWGSGTLVTVSS													
28	QVQLQQSGAEAVRPGASVKLSCKASGYIFTSTSMHWVKQRTGQGLEWIGEIYPNNGDTTYNEKFKDKATLTADKSSSTAYMELSSLTSEDSAVYFCARDFYAEVYDYWGQGTTLTVSS													
29	QVQLQQGAELAMPGTSVKLSCKASGYTFTSYWMHWVKQRPGQGLEWIGEIYPSDSYTRYNQKFKSKATLTVDKSSSTAYMQLSSLTSEDSAVYYCARGGYYYYYSAGYYGVWGSGTSVSSSS													
30	QVQLQQSAAELVNPGASVKLSCKASGYTFTSFWMHWVKQTPGQGLEWIGEIHPSSSYTRYNQKFKSKATLTVDKSSSTAYMQLSSLTSEDSAVYYCARGGYYYYYGYYYMYIWTTGTTTSDS													
31	QVQLQQPGAELVKPGASVKVSCKASGYTFTSYWMHWVKQRPGQGLEWIGRIYPSDSDTNYNGKFKVKATLTVDKSSSTAYMQLSSLTSEDSAVYSCAIDEDDDWWWGTTTLTITTT													
32	QVQLQQSGAELVKPGASVKVSCKASGYTFTSYWMHWVKQRPGQGLEWIGEIYPSDSDTNYNQKFKAKATLTVDKSSSTAYMQLSSLTSEDSAVYSCTMDEYDDWWWGTTTLTVTT													
33	QVQLQQSGAELVRPGASVKLSCKASGYTFTDYYINWVKQRPGQGLEWIGRIYPNSGDTNYNEKFESKATLTADKSSSTAYMQLSRLTSEDSAVYFCARYEYEYYDDYFYYDAWGDGTRLTDLS													
34	QVQLQQSGAELVRPGASVKLSCKASGYTFTDYYINWVKQRPGQGLEWIGRIYPGSGETNYNEKFEDKATLTVDKSSSTAYMQLSSLTSEDSAVYFCARDYGSLYYYYFYYWGTGTTLTVSS													
35	QVQLQQSGAEAVRPGASVKLSCKTSGYTFTSISMHWVKQRTGQGLEWIGEIYPNNGDTTYNEKFKDKATLTADKSSSTAYMELSSLTSEDSAVYFCARDEYAEYFNDWGRGTTLTVSS													
36	QVWLQQSGAELARPGASVKLSCKASGYTFTSYSINWLKQSHGQGLEWIGEIYPNTGDTYYNEKFEDKATVTADKSSSTAEMELSSLTSEDSAVYFCAREGYGGSYADYWGQGTTVTVSS													
37	AVKLVESGGDLVKPGGSLRLSCAASGFTFSDYYMSWIRQAPGKSLEWLGYIRPEANTYATAYSASVKGRFIISRDNSDSILYLMMNTLKAEDAASYYCQRAMTSWFDEYWGKTTSTTTVS													
38	QIQLVQSGPELKKPGETVKISCKASGYTFSKYPMYWVKQAPGKGFKWMGMIYTDTGESTYAENFQGRFAFSLELSASTAYLQINNLKNEDTSTYFCVRAGPEFPMEYWGQGTSLTVSS													
39	QIQLVQSGPELKKPGETVKISCKASGYTFSEYPMYWVKQAPGKGDKWMGMIYTDTGEPTYSEDFIGRFAPSLETSASTAYLQINNLKNEDTATYFCVRHGPWFPMDYWGWTTSVTASS													
40	QIQLVQSGPELKKPGETVKISCKASGYTFTDYPMYWVKQAPGKGKWLGMIYTDTGETTYAEDFTGRFAFSLETRASTAYLQINNLKNEDTAQYFTSSWKYEFEYWWQQGTLTTSS													
41	QIQLVQSGPELKKPGETVKISCKASGYTFTEYPMYWVKQAPGKNYKWMGMIYTDTGEPTYAEDFIDRFAFSLETSASTAYYQINNLKNDDTATQFCIRGGPYWAMDDWGATTILTVSS													
42	QIQLVQSGPELKKPGETVKISCKASGYTFSEYPMYWVKQAPGKGFKWMGMIYTDTGEPTYAEDFTGRFAFSLETSASTAYLQINNLKNEDTATYFCARAGPWYAMDYWGQTTSLTVSS													

the proportions of different amino acid types and secondary structures in the generated library are analyzed. As shown in Figure 4 (a) (b), the proportions of different amino acid types (polar and non-polar, positively charged, and negatively charged) and secondary structures (α-helix, β-strand, and irregular coil) in the generated sequences were similar to those in the wild-type antibody sequences (training set). This similarity indicates that AbGAN-LMG can learn both the features and structural characteristics of the wild-type antibody sequences. However, the figures also reveal that compared to the wild-type antibody sequences, the generated sequences exhibit a long-tail phenomenon in the proportions of different amino acids and secondary structures. This suggests that the model may produce a small number of sequences with higher or lower proportions of certain amino acids or secondary structures, leading to relatively extreme sequences.

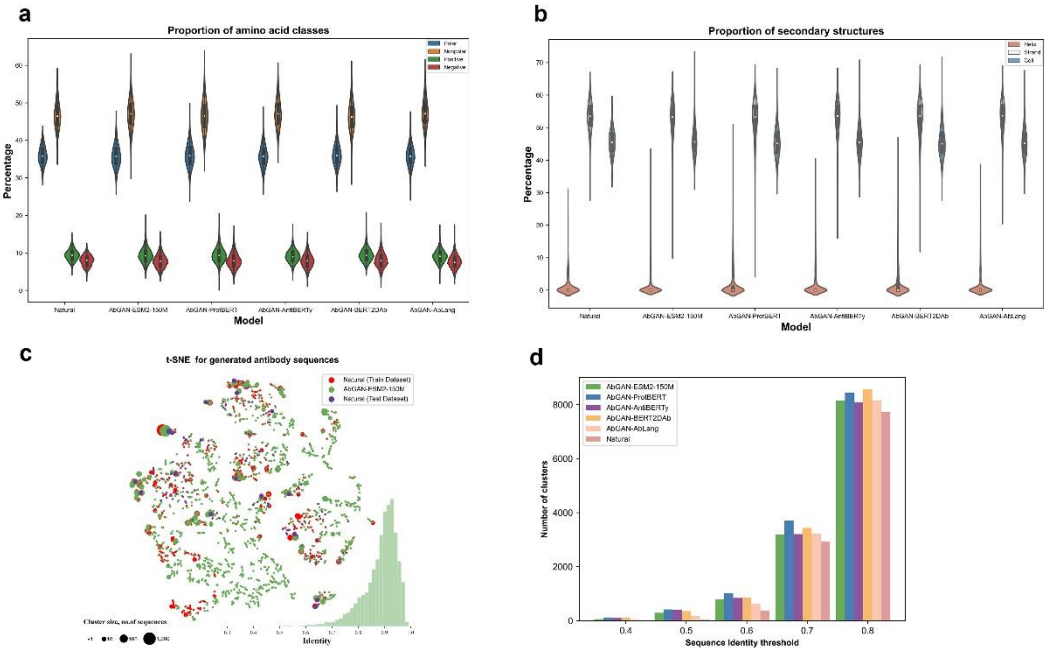


Figure 4. AbGAN-LMG learns essential features of antibodies and generates libraries with high diversity. a. Distribution of amino acid types in the generated library and wild-type antibody sequences. **b.** Distribution of 2D structural in the generated library and wild-type antibody sequences. **c.** The visualization of the generated sequences using t-SNE. **d.** The generated antibody sequences and wild-type antibody sequences were clustered in different sequence identities, and the number of clusters was counted.

t-SNE is employed for dimensionality reduction to visualize the sequences in the generated library to analyze the diversity of the generated library. As shown in Figure 4 (c), the generated antibody sequences by AbGAN-ESM2-150M can partially overlap with the large clusters of wild-type antibody sequences, indicating that the model has learned the features of the wild-type antibodies. Moreover, there are numerous smaller clusters surrounding the large cluster, suggesting that the generated library has higher diversity compared to the training and test sets. As shown in Figure 4 (f), the generated antibody sequences are clustered based on different sequence identity levels, and observed that AbGAN-LMG generated more clusters of sequences under different sequence identity levels than the wild-type sequences, indicating greater diversity in the generated antibody sequences.

6. Technical Support and Contact Information

For technical support and assistance, please feel free to reach out to our dedicated team of experts. We are committed to providing you with the best possible support to ensure a seamless experience with our products and services.

Contact Details:
Email: zwb3585@163.com

7. Acknowledgement

AbGAN-LMG is based on antibody repertoires from the CoV-AbDab database. IgFold is used for 3D structural modeling of antibodies. Developability Calculation: Solubility is calculated using CamSol, humanness is calculated using BioPhi, ease of aggregation is calculated using Aggrescan3D, and immunogenicity is calculated using NetMHC-3.0.

8. Authors' Contributions to the Tool

The following individuals made significant contributions to both the research paper and the development of the associated tool:

Dongsheng Zhao, Academy of Military Medical Sciences: Conceptualization, Methodology, Resources, Formal analysis, Writing - Review & Editing, Supervision, Project administration.

Wenbin Zhao, Academy of Military Medical Sciences: Methodology, Software, Validation, Formal analysis, Data Curation, Investigation, Writing - Original Draft, Visualization Preparation.

Xiaowei Luo, Academy of Military Medical Sciences: Methodology, Software, Validation, Formal analysis, Data Curation, Investigation, Writing - Original Draft, Writing - Review & Editing.

Fan Tong, Academy of Military Medical Sciences: Methodology, Software, Formal analysis, Writing - Review & Editing.

Xiangwen Zheng, Academy of Military Medical Sciences: Methodology, Software, Formal analysis, Writing - Review & Editing.

Jing Li, Beijing Institute of Microbiology and Epidemiology, State Key Laboratory of Pathogen and Biosecurity: Methodology, Resources, Writing - Review & Editing.

Guangyu Zhao, Beijing Institute of Microbiology and Epidemiology, State Key Laboratory of Pathogen and Biosecurity: Methodology, Resources, Writing - Review & Editing.