Antibody Therapeutics, 2023, Vol. 6, No. 3 **147–156**
https://doi.org/10.1093/abt/tbad007
Advance Access Publication on 12 April 2023

Research Article

AB-Amy: machine learning aided amyloidogenic risk prediction of therapeutic antibody light chains

Yuwei Zhou[1], Ziru Huang[1], Yushu Gou[1], Siqi Liu[1], Wei Yang[2], Hongyu Zhang[3], Anthony Mackitz Dzisoo[4],* and Jian Huang[1],*

[1]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China, [2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China, [3]Research and Development, Zhanyuan Therapeutics Ltd., Hangzhou, Zhejiang 310000, China, and [4]Bioinformatics, Data and Medical Reporting, Arcencsus GmbH, Rostock, Mecklenburg-Vorpommern 18055, Germany

Received: October 26, 2022; Revised: March 30, 2023; Accepted: April 6, 2023

ABSTRACT

Over 120 FDA-approved antibody-based therapeutics are used to treat a variety of diseases.However, many candidates could fail because of unfavorable physicochemical properties. Light-chain amyloidosis is one form of aggregation that can lead to severe safety risks in clinical development. Therefore, screening candidates with a less amyloidosis risk at the early stage can not only save the time and cost of antibody development but also improve the safety of antibody drugs. In this study, based on the dipeptide composition of 742 amyloidogenic and 712 non-amyloidogenic antibody light chains, a support vector machine–based model, AB-Amy, was trained to predict the light-chain amyloidogenic risk. The AUC of AB-Amy reaches 0.9651. The excellent performance of AB-Amy indicates that it can be a useful tool for the *in silico* evaluation of the light-chain amyloidogenic risk to ensure the safety of antibody therapeutics under clinical development. A web server is freely available at http://i.uestc.edu.cn/AB-Amy/.

Statement of Significance: Statement of Significance: The amyloidogenic propensity of light chains is not only associated with disease but also the developability of the antibody. A machine learning method for the assessment of therapeutic antibody amyloidosis is herein presented.

KEYWORDS: therapeutic antibody; amyloidosis; developability; support vector machine (SVM); prediction

INTRODUCTION

Therapeutic antibodies have become a major class of biopharmaceutical products because of their high specificity to the therapeutic targets that are undruggable for small-molecule drugs [1,2]. In the last decade, the therapeutic antibody has been widely used for the treatment of cancers, immune-related diseases and infections [3,4]. With the advancement of hybridoma, phage display and single B-cell platforms, a growing number of therapeutic antibodies have been discovered and proceed into clinical development. The number of therapeutic antibodies in the late clinical stage has been more than tripled in the last decade [5]. In July 2021, the US Food and Drug Administration approved the 100th monoclonal antibody therapy [6]. However, over 85% of human or humanised mAbs failed in preclinical development, with many cases due to unfavorable physicochemical properties leading to increased aggregation tendency and high viscosity and thus, poor manufacturability [7, 8]. As a result, the evaluation of aggregation propensity is a key part of antibody developability assessment [9].

Aggregation not only leads to bioprocessing failure but also causes safety risks in clinical trials. The majority of antibody-related amyloidosis is light-chain (AL) amyloidosis. In addition, heavy-chain (AH) amyloidosis and

*To whom correspondence should be addressed. Jian Huang, University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu 610054, China. Email: hj@uestc.edu.cn immunoglobulin heavy- and light-chain (AHL) amyloidosis are reported [10, 11]. AL, also known as primary amyloidosis, is a type of complex and incurable disease caused by abnormal clone plasma cells overproducing immunoglobulin light chains (LCs) that misfold and aggregate as amyloidogenic fibrils in certain organs [12, 13]. AL is often related to multiple myeloma [14] and often leads to cardiac amyloidosis or systemic amyloidosis that affects the kidney, peripheral and autonomic nervous system, liver and other organs [11]. As a result, the therapeutic antibody should not contain LCs with an amyloidogenic risk for the safety of patients. Therefore, the developability assessment of the therapeutic antibody candidates should involve the evaluation of the amyloidogenic risk of their LCs.

The characterisation of the amyloidogenic propensity of LCs *in vitro* was primarily by Congo Red staining. However, due to the limited sensitivity of the Congo Red assay [15], several methods with higher sensitivity have been developed recently [16]. Nevertheless, these assays are laborious, costly and thus not suitable for high-throughput characterisation. Besides, experimental assays only report the presence of the amyloid aggregation once it is significantly formed, rather than giving an early warning before the formation of the nucleus. Bioinformatics method could provide predictive tools to this problem. In recent years, several amyloidogenic sequence databases such as AMYPdb [17], AmyPro [18] and WALTZ-DB2.0 [19] have been constructed. Based on these databases, a few computational tools for the prediction of amyloidogenic proteins have been built [20,21,22–24]. These methods, however, only perform well in detecting hot-spot regions (about six residues) to find aggregation-prone regions. Thus, the results became less accurate when tested by longer sequences [25]. In addition, no predictive tool has been developed specifically for the AL amyloidogenic risk to assess the developability and safety of therapeutic antibodies.

In this study, we trained a novel SVM-based predictor called AB-Amy for the evaluation of the AL amyloidogenic risk of a therapeutic antibody. It enables researchers to exclude amyloidogenic candidates in early development, thereby saving the research and development cost and time. AB-Amy showed reliability in cross- validation and robustness in the test dataset. To our knowledge, this is the first investigation linking AL amyloidogenic propensity and the developability of therapeutic antibodies.

MATERIALS AND METHODS

Figure 1 shows the framework of AB-Amy. In brief, after building the datasets, 21 categories of features were extracted and selected, and unrelated features were reduced. Later on, an SVM-based predictive model was trained and evaluated. Finally, a web server and a standalone program were constructed, respectively.

Dataset construction

AL-base [26] is a curated database that contained LC amino acid sequences from patients with AL amyloidosis. We downloaded 527 AL amyloidosis LC sequences from AL-base and further cleaned these sequences with the following criteria:

a) excluded the sequences containing illegal characteristic “X”;
b) numbered the sequences using IMGT scheme [27];
c) extracted the VL regions of sequences;
d) excluded the sequences with missing or unmatched CDRs.

The other 263 amyloidogenic sequences were taken from the work by David *et al*. [28] and were cleaned as above. Combining the data of the two sources together, we obtained 742 unique sequences of the antibody VL region causing amyloidosis as the positive dataset. The sequences of the LCs of approved antibodies or those in clinical trials are considered as non-amyloidogenic. These LC sequences were extracted from “The Therapeutic Structural Antibody Database” (Thera-SAbDab) [29] and cleaned as above. We finally used the remaining 712 unique sequences as the negative dataset. From the positive and negative datasets, we randomly picked out 500 amyloidogenic and 500 non-amyloidogenic sequences, respectively, to make the training dataset. The remaining sequences (454 in total) were used as an independent test dataset to evaluate the performance of the SVM model.

Feature extraction and selection

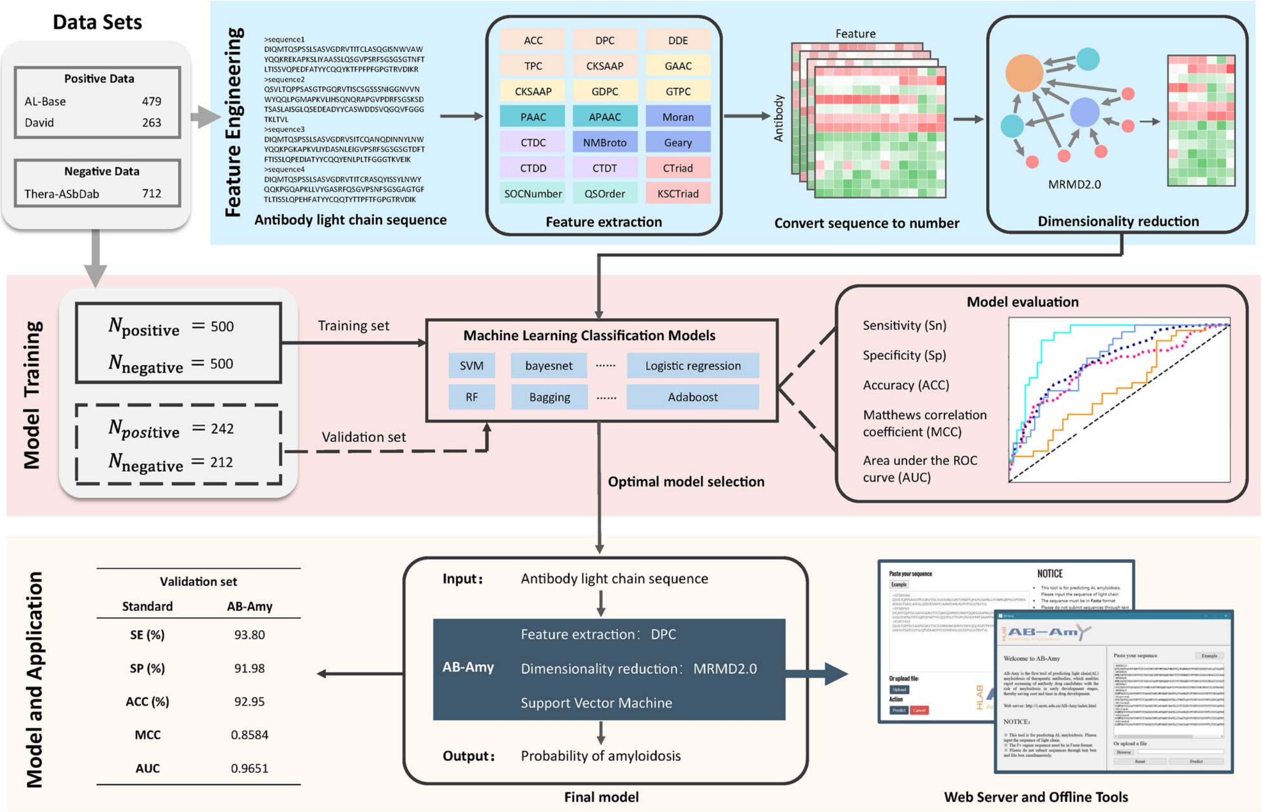
The most important step in building a reliable machine learning model is to extract features from data using an appropriate mathematical method such that LC sequences can be classified. The features extracted from amino acid sequences have displayed good performance on the classification of many proteins and peptides [30–32]. Based on a large number of trials, we found that DPC displayed the best performance on predicting amyloidogenic LCs. DPC encodes the frequency of amino acid pairs (i.e., AA, AC, AD, ..., YY) in a protein or peptide sequence. It is defined as

$$\mathrm{DPC}\,(\mathrm{r,s}) = \frac{N_{rs}}{N-1}, r, s \in \{A, C, D, \cdots, W, Y\} \quad (1)$$

where N_{rs} is the number of the corresponding amino acid pair *rs*. *N* is the length of a protein or peptide sequence. DPC has 400 (20 × 20) descriptors.

In addition, other 20 feature extraction methods were tested. They are AAC, APAAC, DPC, TPC, CKSAAP, CKSAAGP, GAAC, GDPC, GTPC, Moran, Geary, NMBroto, CTDC, CTDT, CTDD, CTriad, KSCTriad, SOCNumber, QSOrder and PAAC. All the feature extraction processes were performed using the iFeature Python package [33].

Exceeding the amount of features used in a model tends to overfit the training data, increase training time and reduce the robustness of the model. To avoid this, features should be reduced. We used an integrated feature selection algorithm MRMD2.0 developed by He et al [34]. MRMD2.0 first sorts all features using different methods



```
>sequence1
DIQMTQSPSSLSASVGDRVTITCLASQGISNWVAW
YQQKREKAPKSLIYAASSLQSGVPSRFSGSGSGTNFT
LTISSVQPEDFATYYCQQYKTFPFPFGPGTRVDIKR
>sequence2
QSVLTQPPSASGTPGQRVTISCSGSSSNIGGNVVN
WYQQLPGMAPKVLIHSQNQRAPGVPDRFSGSKSD
TSASLAISGLQSEDEADYYCASWDDSVQGQVFGGG
TKLTVL
>sequence3
DIQMTQSPSSLSASVGDRVSITCQANQDINNYLNW
YQQKPGKAPKVLIYDASNLEIGVPSRFSGSGSGTDFT
FTISSLQPEDIATYYCQQYENLPLTFGGGTKVEIK
>sequence4
DIQMTQSPSSLSASVGDRVTITCRASQYISSYLNWY
QQKPGQAPKLLVYGASRFQSGVPSNFSGSGAGTGF
TLTISSLQPEHFATYYCQQTYTTPFTFGPGTRVDIK
```

Validation set	
Standard	AB-Amy
SE (%)	93.80
SP (%)	91.98
ACC (%)	92.95
MCC	0.8584
AUC	0.9651

Figure 1. The framework of AB-Amy.

such as mRMR, LASSO and ANOVA . Then, the PageRank algorithm is used to obtain the new ranking of features. Finally, the optimal feature subset is selected by sequential forward selection (SFS). SFS is the one of the simplest feature subset selection methods in which feature set A is initially given one feature, and then, more features are orderly given to evaluate the model until any new feature does increase the performance. The corresponding set A is considered the best feature set.

Model establishment

Support vector machine (SVM) is a supervised machine learning framework for data classification and regression, which was first proposed by Vapnik et al. Because of its high accuracy in handing data characterised by small sample sizes, non-linearity and high dimensional patterns, SVM has been applied in many fields such as peptide identification [35], protein–protein interaction [36] and cancer prediction [37]. We employed LIBSVM [38] with radial basis function (RBF) kernel to construct the model. The SVM output score P, ranging from 0 to 1, is the probability of an LC sequence to be amyloidogenic. LC sequence is classified as amyloidogenic if the probability is greater than 0.5. Furthermore, we compared the SVM-based model with several other classification algorithms, which are Logistic Regression, Random Forest, Decision Tree, Naive Bayes, k-Nearest Neighbors and AdaBoost, using the “sklearn” Python package.

Performance evaluation

In machine learning, cross-validation has been widely used to facilitate model evaluation and hyperparameter selection. We implemented 5-fold cross validation to assess the performance of our predictive model and obtain the optimal kernel parameter γ and penalty parameter c. In our procedure for 5-fold cross validation, the dataset was randomly split into five equal parts, each of which was regarded as the test set in turn, whereas the remaining four parts were used as training set. Eventually, the average accuracy was taken as the final value of accuracy. To evaluate the performance of the predictive model, we employed five widely used classification metrics: sensitivity (Sn), specificity (Sp), accuracy (ACC), Matthews correlation coefficient (MCC) and AUC. These metrics can be defined as

$$\text{Sn} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN+FP}} \tag{3}$$

$$\text{ACC} = \frac{\text{TN+TP}}{\text{TP+FN+TN+FP}} \tag{4}$$

$$\text{MCC} = \frac{\text{TN}\times\text{TP}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{5}$$

where TP is the number of correctly predicted amyloidogenic sequences; FP is the number of falsely predicted amyloidogenic sequences; TN is the number of correctly predicted non-amyloidogenic sequences; and FN is the number of incorrectly predicted non-amyloidogenic sequences. In

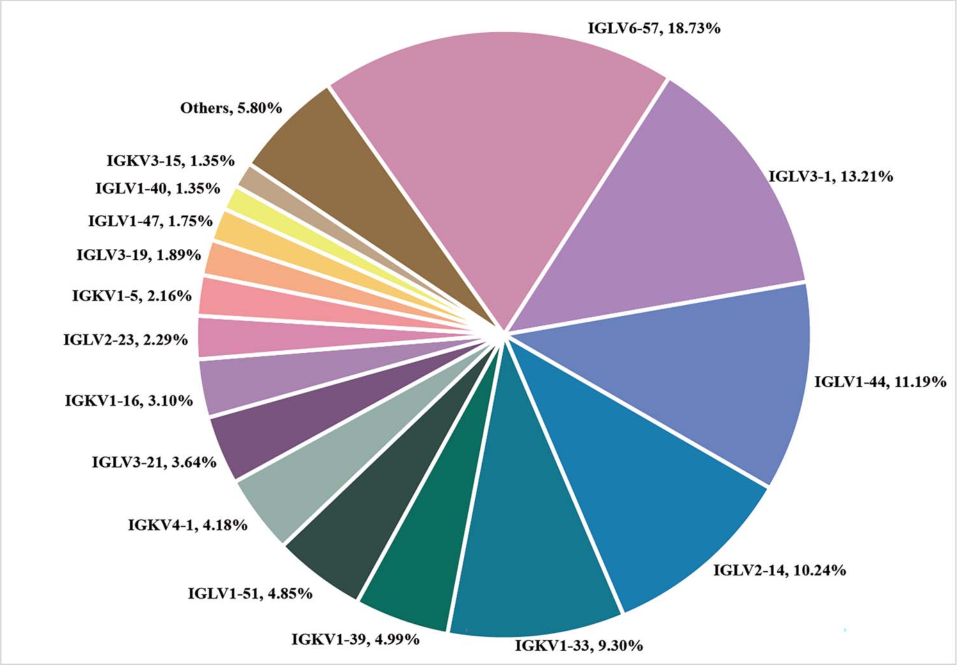


Figure 2. The usage frequency of germline gene in amyloidosis sequences. The other groups include IGLV2-8, IGKV2-28, IGLV2-11, IGKV1D-16, IGLV4-69, IGLV3-27, IGLV3-25, IGKV1-12, IGLV1-36, IGLV7-43, IGKV1-27, IGLV10-54, IGLV3-10, IGKV2-30, IGLV3-9, IGKV3-20 and IGKV3D-15.

addition, the true-positive rate (TPR) and false-positive rate (FPR) will be calculated to plot the receiver operating characteristic (ROC) curve. The AUC (area under the ROC curve) is used to illustrate the performance of the model because it is independent of the choice of the threshold for the predicted values. The AUC value of 0.5 represents a random prediction. A model with the AUC value of 1 means a perfect performance.

Online web server and standalone tool

To provide an easy and user-friendly interface, we constructed a web server for AB-Amy. The front-end of AB-Amy was developed and implemented using HTML, CSS and JavaScript. The back-end of AB-Amy was implemented with PHP and Python scripts. To ensure a stable and secure service, the standalone versions of AB-Amy can also be freely downloaded. The graphical user interface (GUI) of the standalone version of AB-Amy is implemented with "pyside2," a Python module, which provides an access to the complete Qt 5.12+ framework.

RESULTS

Germline gene usage in AL amyloidosis

The use of certain germline gene segments poses a risk for the development of amyloidosis [39]. We analyzed the germline gene usage of 742 AL amyloidosis LC sequences. The most frequently used germline genes were IGLV6-57 (6a), IGLV3-1 (3r), IGLV1-44 and IGLV2-14, accounting for 18.73, 13.21, 11.19 and 10.24 of total usage, respectively (Fig. 2). The germline gene IGLV6-57 is more common in AL amyloidosis than in the normal B cells and is associated with renal involvement [40]. Using IGLV1-44 and IGLV2-14 germline genes is linked to predominant cardiac involvement, whereas IGLV3-1 is frequently found in the amyloid infiltration of various organs [39]. In contrast, IGLV7-43, IGKV1-27, IGLV10-54, IGLV3-10, IGKV2-30, IGLV3-9, IGKV3-20 and IGKV3D-15 are used only once in AL-base. These genes were rarely reported to have usage bias in AL patients.

Feature selection on DPC

In our model, DPC has 400 features that were then selected by MRMD2.0. Figure 3 shows the variation of ACC using different numbers of features during the SFS process. Initially, the ACC increased significantly from 0.803 to 0.901, whereas the number of features increased from 1 to 37. Subsequently, with the increase of the number of features, the ACC fluctuated around 0.90 and reached the maximum value when the number of features was 45. Compared with the original 400 features, the ACC of optimal feature set has been raised from 0.896 to 0.907. The visual increase of ACC was not significant, but we suggested that the small sub-feature-set was better even if the ACC equaled to model trained with original features. The results of MRMD2.0 indicated that the 45 selected features were the most representative feature set of the original features, which nearly covered the characteristics of amyloidogenic LC sequences.

Performances of SVM-based models in identification of amyloidogenic antibody LCs

The SVM-based model was trained using the optimal sub-feature-set with 45 features selected in the previous step. We used 5-fold cross-validation to investigate the performance

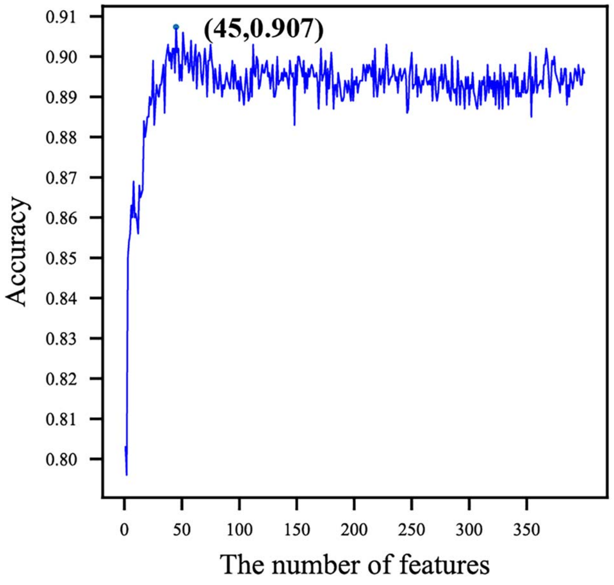


Figure 3. The ACC calculated by a different sub-feature-set in the sequential forward selection procession.

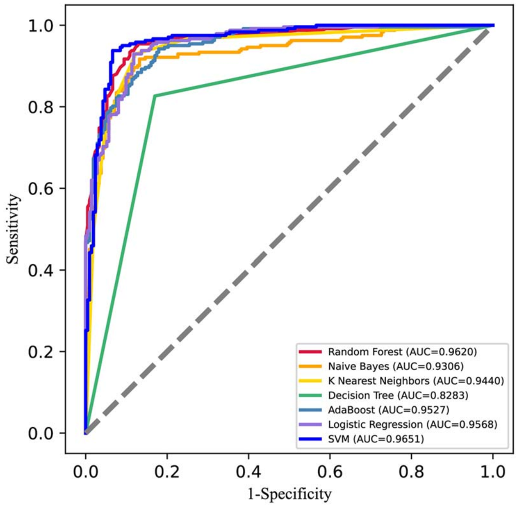


Figure 4. The ROC curve for AB-Amy and other six classifiers tested on the independent dataset.

of the model. The model reached the highest accuracy when $c = 2$ and $\gamma = 0.125$. The results from the 5-fold cross validation showed that the ACC of the predictive model was 90.60% when the threshold was set to 0.5. To test the robustness and generalisation of the proposed method, external validation is required to evaluate the developed predictive model. Therefore, we assessed AB-Amy on an independent test set, the ACC of which was 92.95% with an SE of 93.80%, an SP of 91.98% and an MCC of 0.8584. Figure 4 shows the ROC curve of AB-Amy, where the AUC reached 0.9651. The results of cross-validation and independent testing confirmed that our proposed predictor, AB-Amy, effectively recognised amyloidogenic LCs from therapeutic antibody LCs.

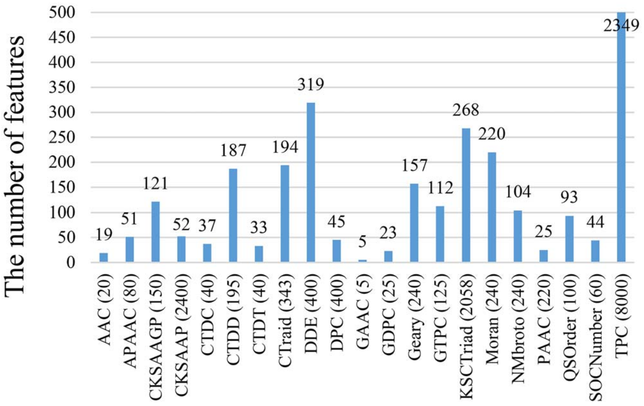


Figure 5. The number of features of 21 feature extraction methods processed by MRMD2.0. The original feature number of each method is marked in parentheses.

Performance of different feature extraction methods

As described previously, to verify the effectiveness of the proposed feature, we compared DPC with other 20 popular feature extraction methods. Figure 5 shows the comparison of the original and reduced feature dimension results of all the methods processed by MRMD2.0. The smaller representative feature subsets were obtained to reduce redundancy and correlation between features. The independent test results of all the methods are shown in Table 1. The performance of the two models was compared by five metrics: SE, SP, ACC, MCC and AUC. In the comparison, the SVM model based on DPC achieved the highest sensitivity, specificity, ACC and MCC of 93.80, 91.98, 92.95 and 0.8584, respectively (Table 1). In the AUC metric, TPC and DDE performed slightly higher than DPC. However, the feature number of TPC and DDE is 2349 and 319, respectively, which is strikingly more than that of DPC. Fewer features can reduce the training time and overfitting risk of the model. Therefore, DPC is more proper for predicting amyloidogenic antibody LC sequences in terms of feature dimension and the model performance in external validation.

Comparison with other classifiers

We compared the SVM model with other widely used classifiers including Logistic Regression, Random Forest, Decision Tree, Naive Bayes, *k*-Nearest Neighbors and AdaBoost. The results are shown in Table 2. The SVM-based method is obviously superior to the six other classifiers in SE, ACC and MCC and higher than them by about 0.0031–0.1368 in the AUC metric (Fig. 4). The *k*-Nearest Neighbors also achieve the highest specificity of 91.98%. However, its sensitivity is 83.47%, which is much lower than that of SVM.

Comparison with other published predictors

Quite a few computational tools have been developed for the prediction of amyloidogenic sequences in proteins. To further evaluate AB-Amy, we compared its performance

Table 1. The performance evaluation of the SVM model based on 21 feature extraction algorithms

Feature	SE (%)	SP (%)	ACC (%)	MCC	AUC
DPC	93.80	91.98	92.95	0.8584	0.9651
TPC	92.15	91.98	92.07	0.8408	0.9700
Geary	91.32	91.98	91.63	0.8322	0.9612
DDE	93.39	88.68	91.19	0.8231	0.9728
CTriad	90.50	91.04	90.75	0.8145	0.9608
Moran	91.32	90.09	90.75	0.8142	0.9571
QSOrder	90.91	90.57	90.75	0.8143	0.9684
GTPC	90.50	90.57	90.53	0.8100	0.9517
KSCTriad	91.32	89.62	90.53	0.8097	0.9635
NMBroto	89.67	91.04	90.31	0.8059	0.9575
CTDD	89.26	89.62	89.43	0.7880	0.9629
APAAC	89.67	88.21	88.99	0.7788	0.9519
CKSAAP	88.84	88.21	88.55	0.7701	0.9525
CTDC	87.60	89.62	88.55	0.7709	0.9427
SOCNumber	88.84	88.21	88.55	0.7701	0.9424
CKSAAGP	88.43	88.21	88.33	0.7658	0.9588
CTDT	88.43	88.21	88.33	0.7658	0.9403
AAC	88.43	86.32	87.44	0.7477	0.9359
PAAC	89.26	84.91	87.22	0.7432	0.9365
GDPC	84.30	88.21	86.12	0.7235	0.9296
GAAC	79.75	81.60	80.62	0.6124	0.8694

Table 2. Comparison of SVM with other classifiers

Model	SE (%)	SP (%)	ACC (%)	MCC	AUC
SVM	93.80	91.98	92.95	0.8584	0.9651
Random Forest	91.74	91.98	91.85	0.8365	0.9620
Naive Bayes	90.08	86.79	88.55	0.7698	0.9306
K-Nearest Neighbors	83.47	91.98	87.44	0.7533	0.9440
Decision Tree	81.82	83.96	82.82	0.6565	0.8283
AdaBoost	88.43	87.26	87.89	0.7567	0.9527
Logistic Regression	88.02	89.62	88.77	0.7752	0.9528

with previously published methods. We used VLAmY-Pred [41], Aggrescan [20], AmyloGram [42], APPNN [43], Pasta2.0 [44], Waltz [21] and iAMY-SCM [45] to predict the amyloidogenic patches or amyloidogenic propensity of the sequences with our independent test dataset. The outputs of some methods were processed further for properly comparing their performance. For example, Pasta2.0 and Waltz make prediction to the number of amyloidogenic hotspot regions. We assumed that an antibody is amyloidogenic if at least one amyloidogenic region was predicted in its sequence. Moreover, Aggrescan calculates the global protein aggregation propensity average score of input sequences. Because high Aggrescan score corresponds to low aggregation propensity, we sorted all the scores and classified all antibodies into amyloidogenic and non-amyloidogenic groups by the threshold of −12.96 [46]. The performances of all selected methods are listed in Table 3. It is shown that AB-Amy exhibited the best overall performance. The SE of Aggrescan, Waltz and iAMY-SCM are all significantly higher than SP, whereas APPNN and AmyloGram predicted all sequences as positive samples, which indicates that these methods have significant biases in predicting amyloidosis-prone sequences. VLAmY-Pred is an *in silico* tool for screening the potential amyloidogenic LCs. However, it achieved an unsatisfied ACC of 74.67%. It is obvious that AB-Amy is an efficient model for the accurate prediction of the amyloidogenic antibody LCs.

DPC features and amyloidogenic risk

We analyzed the dipeptide occurrence in amyloidogenic and non-amyloidogenic LC sequences. Figure 6 shows the dipeptides that are significantly different in their occurrence between the positive and negative samples (FDR < 0.05). The positive samples are rich in DE, WD, SI and NF sequences, whereas the frequency of EI, TF and LE is much higher in the negative samples. This suggests that amyloidogenic and non-amyloidogenic LCs have clearly distinguishable dipeptide composition [47]. The above result indicated that the selected dipeptide features

Table 3. Performance of AB-Amy and other published methods in amyloidogenic antibody sequence identification

Model	**SE (%)**	**SP (%)**	**ACC (%)**	**MCC**	**AUC**
AB-Amy	93.80	91.98	92.95	0.8584	0.9651
VLAmY-Pred	73.55	75.94	74.67	0.4939	0.7475
Aggrescan	78.93	0.94	42.51	−0.3127	0.3993
AmyloGram	100	–	53.3	–	0.5
APPNN	100	–	53.3	–	0.5
Pasta2.0	52.06	75	62.77	0.2763	0.6353
Waltz	97.52	0.47	52.2	−0.8129	0.4899
iAMY-SCM	98.76	28.3	53.96	0.0569	0.5079

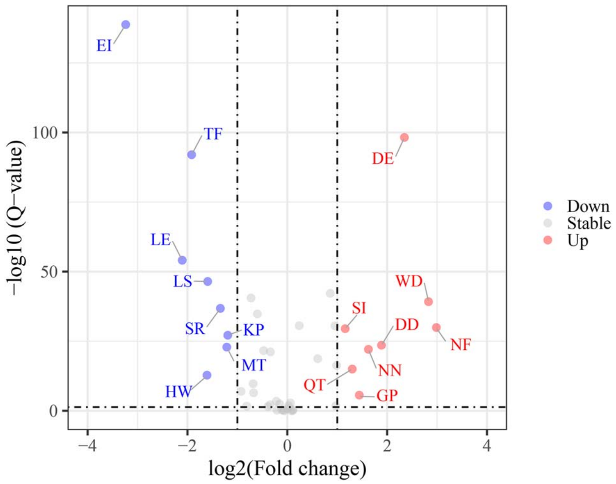


Figure 6. Volcano map of 45 dipeptides in the AB-Amy model showed that 16 dipeptides were significantly different between the positive and negative groups. The thresholds were |log2 (FC)| > 1 and FDR < 0.05. Blue and red points indicated the down- and up-regulated dipeptides, respectively.

are useful for characterising amyloidogenic sequences. The occurrence of up-regulated dipeptides probably enhances the amyloidogenic risk of LCs. Interestingly, we found that WD only existed in the CDR3 region of amyloidogenic LCs.

Web server and standalone tool

For the convenience of users, we implemented AB-Amy into a user-friendly web server, which is freely available at http://i.uestc.edu.cn/AB-Amy. Figure 7 shows the interface of AB-Amy. At the "submit" page (Fig. 7A), users can input or paste LC sequence in the text box or upload files in raw or FASTA format. AB-Amy only recognises 20 common amino acids in one letter code as legal characters. Users can download their results once the prediction is completed (Fig. 7C). In the downloaded file, the "Probability" column represents the probability of amyloidosis (the default threshold is 0.5). "1" in the "Result" column denotes that the submitted LC exhibits a high risk of amyloidosis and should be excluded from the development pipeline. For users who want keep their data private, they can download standalone versions of AB-Amy via http://i.uestc.edu.cn/AB-Amy/download.html. The interface style and usage of the GUI AB-Amy are consistent with that of the web server (Fig. 7B and D).

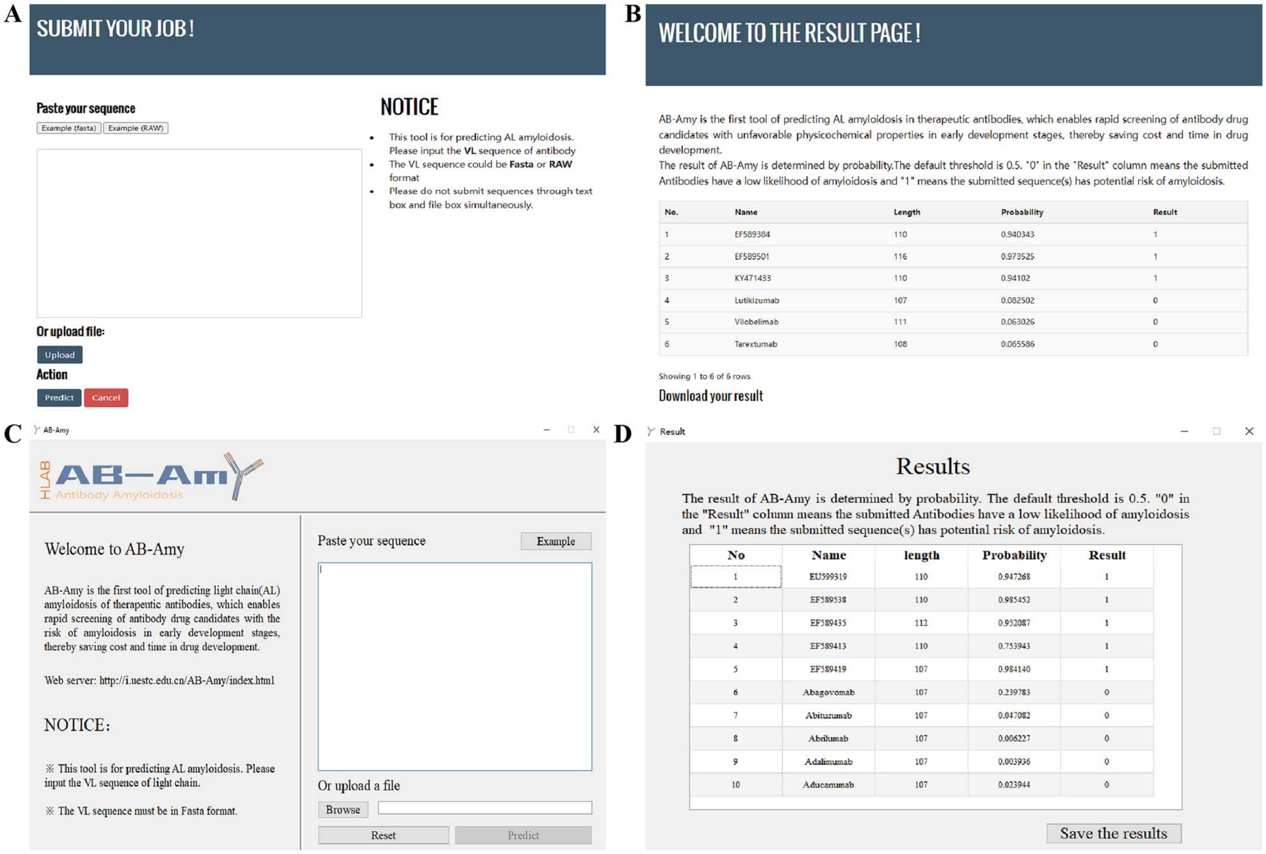
DISCUSSION

Relevant computational studies

Quite a few computational methods have been applied to decipher the properties leading to amyloidosis. The existing methods can be grouped into two categories: the structure-based and the sequence-based methods. The structure-based methods [48] rely on crystal structures. In contrast, the sequence-based methods [20, 21, 42–45] provide easier ways to predict the amyloidogenic proteins. However, most of the sequence-based methods are mainly applied to predict amyloidogenic hotspot regions in proteins. As a result, these methods show high false-positive rates when they are applied to longer protein sequences.

David et al. [28] explored a naive Bayesian classifier and a weighted decision tree for predicting the amyloidogenicity of immunoglobulin sequences, and the ACC of the best decision tree model reached 78.64% for the test set. Using the data from David, Liaw et al. [47] constructed a random forest model (AbAmyloid) to predict antibody amyloidosis. The accuracy of AbAmyloid for cross-germline prediction was 83.33%. The above-mentioned tools do not provide any web server, or their web server becomes unavailable. Rawat et al. [41] analyzed the sequence features of the amyloidogenic and non-amyloidogenic LC and proposed a machine learning model named "VLAmY-Pred." It does provide an available web server. However, its accuracy is only 79.70% on the complete dataset. LICTOR [49] was a random forest-based model for LC aggregation toxicity prediction that used somatic mutations as predictor variables. Tested on an independent set, LICTOR achieved a prediction accuracy of 83%. The authors further approved that toxic sequences were more prone to aggregate than non-toxic ones.

All the computational tools introduced above only aim to predict the amyloidogenic regions of proteins or more specifically, amyloidogenic antibody LCs, which can help to reveal the etiology of amyloidosis. Regretfully, the performance of the above tools is far from satisfying. In addition, most tools cannot be used conveniently.



NOTICE

- This tool is for predicting AL amyloidosis. Please input the **VL** sequence of antibody
- The VL sequence could be **Fasta** or **RAW** format
- Please do not submit sequences through text box and file box simultaneously.

AB-Amy is the first tool of predicting AL amyloidosis in therapeutic antibodies, which enables rapid screening of antibody drug candidates with unfavorable physicochemical properties in early development stages, thereby saving cost and time in drug development.
The result of AB-Amy is determined by probability.The default threshold is 0.5. "0" in the "Result" column means the submitted Antibodies have a low likelihood of amyloidosis and "1" means the submitted sequence(s) has potential risk of amyloidosis.

No.	Name	Length	Probability	Result
1	EF589384	110	0.940343	1
2	EF589501	116	0.973525	1
3	KY471433	110	0.94102	1
4	Lutikizumab	107	0.082502	0
5	Vilobelimab	111	0.063026	0
6	Tarextumab	108	0.065586	0

Showing 1 to 6 of 6 rows

Welcome to AB-Amy

AB-Amy is the first tool of predicting light chain(AL) amyloidosis of therapeutic antibodies, which enables rapid screening of antibody drug candidates with the risk of amyloidosis in early development stages, thereby saving cost and time in drug development.

Web server: http://i.uestc.edu.cn/AB-Amy/index.html

NOTICE:

※ This tool is for predicting AL amyloidosis. Please input the VL sequence of light chain.

※ The VL sequence must be in Fasta format.

Results

The result of AB-Amy is determined by probability. The default threshold is 0.5. "0" in the "Result" column means the submitted Antibodies have a low likelihood of amyloidosis and "1" means the submitted sequence(s) has potential risk of amyloidosis.

No	Name	length	Probability	Result
1	EU599319	110	0.947268	1
2	EF589538	110	0.985452	1
3	EF589435	112	0.952087	1
4	EF589413	110	0.753943	1
5	EF589419	107	0.984140	1
6	Abagovomab	107	0.239783	0
7	Abituzumab	107	0.047082	0
8	Abrilumab	107	0.006227	0
9	Adalimumab	107	0.003936	0
10	Aducanumab	107	0.023944	0

Figure 7. AB-Amy online web server (A and B) and the GUI of the standalone version (C and D).

Amyloidogenic LC and antibody developability

Therapeutic antibodies are highly specific and very effective drugs [2]. However, developing them suffers from unfavorable physicochemical properties, especially those leading to antibody aggregation during or after bioprocessing, storage and administration [50]. Amyloidosis is a group of severe diseases characterised by the deposition of misfolded proteins in the form of amyloid fibrils [51]. AL amyloidosis is the most common systemic amyloidosis. It is caused by extracellular deposition of circulating LCs as amyloid fibrils, resulting in the dysfunction of vital organs [40]. These circulating LCs are most commonly produced and secreted by a plasma cell clone [52]. However, free LCs are not necessarily amyloidogenic. The most widely known free LCs in blood are Bence–Jones proteins (BJPs). It is reported that only 15–20% of BJP are amyloidogenic and the amyloidogenic property is associated with their variable region [53]. In addition to AL amyloidosis, AH amyloidosis and AHL amyloidosis are reported, indicating the amyloidogenic risk of heavy chains or whole antibodies [51]. Therefore, we infer that therapeutic antibody candidates with amyloidogenic LCs might have a higher developability risk. Consequently, excluding therapeutic antibody candidates with an amyloidogenic risk in early development is necessary.

In our study, we combine 263 amyloidogenic LCs reported by David *et al*. with 527 amyloidogenic LCs from the AL-base database to compose the positive dataset. The LCs of approved antibodies or those in clinical trials are extracted from the Thera-SAbDab database to make the negative dataset as none of these LCs has been reported to be amyloidogenic. Based on the datasets, a novel SVM-based model called AB-Amy was built. Its accuracy and AUC reached 92.95% and 0.9651, respectively, in an independent test dataset. As shown in Table 3, AB-Amy shows a better performance than those of existing tools. However, we must agree that such comparison is not fair as some of those tools are trained on different datasets and some are even not based on machine learning models.

Nevertheless, this is the first study, at least to our knowledge, that aims to assess the therapeutic antibody developability from the point of the AL amyloidogenic risk. Furthermore, we provided not only a user-friendly web server but also standalone versions of AB-Amy. The latter allows to assess the AL amyloidogenic risk in a high-throughput and more secure way.

Future directions

In this study, the negative dataset is composed of the LCs of therapeutic antibodies from the Thera-SAbDab database. Thus, the current version of AB-Amy does not fit for assisting the AL amyloidosis diagnosis. In future, one direction is to make a new negative dataset from the B-cell receptor repertoire sequencing data of healthy control and then train a new model. In our opinion, this new model can be used to evaluate the B-cell receptor repertoire sequencing data. This might help to identify the abnormal B-cell or plasma cell clones that cause amyloidosis. By doing this, AB-Amy might show the potential to provide a good *in silico* diagnostic tool for the amyloidogenic propensity of *in vitro* samples containing the sequences of antibody LC. At present, AL amyloidosis is usually suspected on the basis of symptoms of organ involvement, which present very late in the disease course, and Congo red stain is still the gold standard for the demonstration of amyloid in tissue sections. Thus, such new models are valuable.

In this paper, we only study the amyloidogenic risk model for the LCs of therapeutic antibodies as there are adequate data of known amyloidogenic LCs. With the accumulation of data of amyloidogenic heavy chains or whole antibodies, we can expect to build new and more models in future and evaluate the amyloidogenic risk of therapeutic antibodies more comprehensively.

SUPPLEMENTARY DATA

Supplementary Data are available at ABT Online.

FUNDING

This work was supported by a grant from the National Natural Science Foundation of China (grant number 62071099).

CONFLICT OF INTEREST STATEMENT

H.Z. holds the position of Assistant Editor for *Antibody Therapeutics* and is blinded from reviewing or making decisions for the manuscript. H.Z. is also an employee of Zhanyuan Therapeutics Ltd and declares to have no conflicts of interest that might be relevant to the contents of this article. A.M.D. is an employee of Arcensus GmbH and declares to have no conflicts of interest that might be relevant to the contents of this article.

AUTHOR CONTRIBUTIONS

J.H. and A.M.D. proposed the initial idea. Y.W.Z. built the dataset, trained the model and wrote the manuscript. Z.R.H. and H.Y.Z. analyzed the features. Y.S.G. drew the figures. S.Q.L. and W.Y. wrote the interface scripts of web service. All authors read and revised the manuscript.

DATA AVAILABILITY STATEMENTS

No new data were generated or analyzed in support of this research. Publicly available datasets were analyzed in this study. These data can be found here: https://wwwapp.bumc.bu.edu/BEDAC_ALBase/(AL-base), http://opig.stats.ox.ac.uk/webapps/newsabdab/therasabdab/search/(Thera-SAbDab) and https://static-content.springer.com/esm/art%3A10.1186%2F1471-2105-11-79/MediaObjects/12859_2009_3536_MOESM2_ESM.PDF

ETHICS AND CONSENT STATEMENT

The consent is not required.

ANIMAL ETHICS STATEMENT

Not applicable.

REFERENCES

1. Graves, J, Byerly, J, Priego, E, Makkapati N., Parish S., Medellin B., Berrondo M. A review of deep learning methods for antibodies. *Antibodies (Basel)* 2020; **9** 2 12.
2. Rabia, LA, Desai, AA, Jhajj, HSet al. Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochem Eng J* 2018; **137**: 365–74.
3. Ning, L, Abagna, HB, Jiang, Q *et al.* Development and application of therapeutic antibodies against COVID-19. *Int J Biol Sci* 2021; **17**: 1486–96.
4. Lyu, X, Zhao, Q, Hui, J *et al.* The global landscape of approved antibody therapies. *Antib Ther* 2022; **5**: 233–57.
5. Kaplon, H, Muralidharan, M, Schneider, Z *et al.* Antibodies to watch in 2020. *MAbs* 2020; **12**: 1703531.
6. Mullard, A. FDA approves 100th monoclonal antibody product. *Nat Rev Drug Discov* 2021; **20**: 491–5.
7. Carter, PJ, Lazar, GA. Next generation antibody drugs: pursuit of the 'high-hanging fruit'. *Nat Rev Drug Discov* 2018; **17**: 197–223.
8. Jain, T, Sun, T, Durand, S *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci U S A* 2017; **114**: 944–9.
9. Xu, Y, Wang, D, Mason, B *et al.* Structure, heterogeneity and developability assessment of therapeutic antibodies. *MAbs* 2019; **11**: 239–64.
10. Manabe, S, Iwasaki, C, Hatano, M *et al.* AL amyloidosis with non-amyloid forming monoclonal immunoglobulin deposition; a case mimicking AHL amyloidosis. *BMC Nephrol* 2018; **19**: 337.
11. Kim, C, Brealey, J, Jobert, A *et al.* A case of monoclonal gammopathy of renal significance presenting as atypical amyloidosis with IgA lambda paraproteinemia. *J Pathol Transl Med* 2020; **54**: 504–7.
12. Blancas-Mejia, LM, Misra, P, Dick, CJ *et al.* Immunoglobulin light chain amyloid aggregation. *Chem Commun (Camb)* 2018; **54**: 10664–74.
13. Falk, RH, Alexander, KM, Liao, R *et al.* AL (light-chain) cardiac amyloidosis: a review of diagnosis and therapy. *J Am Coll Cardiol* 2016; **68**: 1323–41.
14. Incel Uysal, P, Akdogan, N, Bozdogan, O *et al.* Amyloid light-chain amyloidosis with haemorrhagic bullous eruption disclosing multiple myeloma. *Int Wound J* 2020; **17**: 510–3.
15. Li, T, Huang, X, Cheng, S *et al.* Utility of abdominal skin plus subcutaneous fat and rectal mucosal biopsy in the diagnosis of AL amyloidosis with renal involvement. *PloS One* 2017; **12**: e0185078.
16. Rahman, MM, Schmuck, B, Hansson, H *et al.* Enhanced detection of ATTR amyloid using a nanofibril-based assay. *Amyloid* 2021; **28**: 158–67.
17. Pawlicki, S, Le Bechec, A, Delamarche, C. AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinform* 2008; **9**: 273.
18. Varadi, M, De Baets, G, Vranken, WF *et al.* AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res* 2018; **46**: D387–92.
19. Louros, N, Konstantoulea, K, De Vleeschouwer, M *et al.* WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res* 2020; **48**: D389–93.
20. Conchillo-Sole, O, de Groot, NS, Aviles, FX *et al.* AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 2007; **8**: 65.
21. Maurer-Stroh, S, Debulpaep, M, Kuemmerer, N *et al.* Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods* 2010; **7**: 237–42.
22. Emily, M, Talvas, A, Delamarche, C. MetAmyl: a METa-predictor for AMYLoid proteins. *PloS One* 2013; **8**: e79722.

23. Kim, C, Choi, J, Lee, SJ *et al.* NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. *Nucleic Acids Res* 2009; **37**: W469–73.
24. Garbuzynskiy, SO, Lobanov, MY, Galzitskaya, OV. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics* 2010; **26**: 326–32.
25. Ahmed, AB, Kajava, AV. Breaking the amyloidogenicity code: methods to predict amyloids from amino acid sequence. *FEBS Lett* 2013; **587**: 1089–95.
26. Bodi, K, Prokaeva, T, Spencer, B *et al.* AL-base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. *Amyloid* 2009; **16**: 1–8.
27. Ehrenmann, F, Lefranc, MP. IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). *Cold Spring Harb Protoc* 2011; **2011**: 737–49.
28. David, MP, Concepcion, GP, Padlan, EA. Using simple artificial intelligence methods for predicting amyloidogenesis in antibodies. *BMC Bioinformatics* 2010; **11**: 79.
29. Raybould, MIJ, Marks, C, Lewis, AP *et al.* Thera-SAbDab: the therapeutic structural antibody database. *Nucleic Acids Res* 2020; **48**: D383–8.
30. He, B, Chen, H, Huang, J. PhD7Faster 2.0: predicting clones propagating faster from the Ph.D.-7 phage display library by coupling PseAAC and tripeptide composition. *PeerJ* 2019; **7**: e7131.
31. He, B, Kang, J, Ru, B *et al.* SABinder: a web service for Predicting Streptavidin-Binding Peptides. *Biomed Res Int* 2016; **2016**: 1–8.
32. Dzisoo, AM, Kang, J, Yao, P *et al.* SSH: a tool for predicting hydrophobic interaction of monoclonal antibodies using sequences. *Biomed Res Int* 2020; **2020**: 1–6.
33. Chen, Z, Zhao, P, Li, F *et al.* iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018; **34**: 2499–502.
34. He, SD, Guo, F, Zou, Q *et al.* MRMD2.0: a python tool for machine learning with feature ranking and reduction. *Curr Bioinforma* 2020; **15**: 1213–21.
35. Islam, SM, Sajed, T, Kearney, CM *et al.* PredSTP: a highly accurate SVM based model to predict sequential cystine stabilized peptides. *BMC Bioinformatics* 2015; **16**: 210.
36. Das, S, Chakrabarti, S. Classification and prediction of protein-protein interaction interface using machine learning algorithm. *Sci Rep* 2021; **11**: 1761.
37. Jiang, L, Zhou, L, Ai, Z, Xiao C., Liu W., Geng W., Chen H., Xiong Z., Yin X., Chen Y.C. Machine learning based on diffusion kurtosis imaging histogram parameters for glioma grading. *J Clin Med* 2022, **11** 9 2310.
38. Chang, C-C, Lin, C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Sys Technology* 2011; **2**(3): 1–27.
39. Berghaus, N, Schreiner, S, Granzow, M *et al.* Analysis of the complete lambda light chain germline usage in patients with AL amyloidosis and dominant heart or kidney involvement. *PloS One* 2022; **17**: e0264407.
40. Merlini, G. AL amyloidosis: from molecular mechanisms to targeted therapies. *Hematology Am Soc Hematol Educ Program* 2017; **2017**: 1–12.
41. Rawat, P, Prabakaran, R, Kumar, S *et al.* Exploring the sequence features determining amyloidosis in human antibody light chains. *Sci Rep* 2021; **11**: 13785.
42. Burdukiewicz, M, Sobczyk, P, Rodiger, S *et al.* Amyloidogenic motifs revealed by n-gram analysis. *Sci Rep* 2017; **7**: 12961.
43. Familia, C, Dennison, SR, Quintas, A *et al.* Prediction of peptide and protein propensity for amyloid formation. *PloS One* 2015; **10**: e0134679.
44. Walsh, I, Seno, F, Tosatto, SC *et al.* PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res* 2014; **42**: W301–7.
45. Charoenkwan, P, Kanthawong, S, Nantasenamat, C *et al.* iAMY-SCM: improved prediction and analysis of amyloid proteins using a scoring card method with propensity scores of dipeptides. *Genomics* 2021; **113**: 689–98.
46. de Groot, NS, Castillo, V, Grana-Montes, R *et al.* AGGRESCAN: method, application, and perspectives for drug design. *Methods Mol Biol* 2012; **819**: 199–220.
47. Liaw, C, Tung, CW, Ho, SY. Prediction and analysis of antibody amyloidogenesis from sequences. *PloS One* 2013; **8**: e53235.
48. Thompson, MJ, Sievers, SA, Karanicolas, J *et al.* The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci U S A* 2006; **103**: 4074–8.
49. Garofalo, M, Piccoli, L, Romeo, M *et al.* Machine learning analyses of antibody somatic mutations predict immunoglobulin light chain toxicity. *Nat Commun* 2021; **12**: 3532.
50. Obrezanova, O, Arnell, A, de la Cuesta, RG *et al.* Aggregation risk prediction for antibodies and its application to biotherapeutic development. *MAbs* 2015; **7**: 352–63.
51. Picken, MM. The pathology of amyloidosis in classification: a review. *Acta Haematol* 2020; **143**: 322–34.
52. Palladini, G, Milani, P, Merlini, G. Management of AL amyloidosis in 2020. *Blood* 2020; **136**: 2620–7.
53. Stone, MJ, Guirl, MJ. In: Johnson, LR (ed). *Encyclopedia of Gastroenterology*. San Diego: Academic Press, 2004, 59–69