BMC Bioinformatics

RESEARCH Open Access

Prediction of polyreactive and nonspecific single-chain fragment variables through structural biochemical features and protein language-based descriptors

Hocheol Lim[1,2] and Kyoung Tai No[1,2,3*]

*Correspondence:
ktno@yonsei.ac.kr

[1] The Interdisciplinary Graduate Program in Integrative Biotechnology and Translational Medicine, Yonsei University, Incheon 21983, Republic of Korea
[2] Bioinformatics and Molecular Design Research Center (BMDRC), Incheon 21983, Republic of Korea
[3] Baobab AiBIO Co., Ltd., Incheon 21983, Republic of Korea

Abstract

Background: Monoclonal antibodies (mAbs) have been used as therapeutic agents, which must overcome many developability issues after the discovery from in vitro display libraries. Especially, polyreactive mAbs can strongly bind to a specific target and weakly bind to off-target proteins, which leads to poor antibody pharmacokinetics in clinical development. Although early assessment of polyreactive mAbs is important in the early discovery stage, experimental assessments are usually time-consuming and expensive. Therefore, computational approaches for predicting the polyreactivity of single-chain fragment variables (scFvs) in the early discovery stage would be promising for reducing experimental efforts.

Results: Here, we made prediction models for the polyreactivity of scFvs with the known polyreactive antibody features and natural language model descriptors. We predicted 19,426 protein structures of scFvs with trRosetta to calculate the polyreactive antibody features and investigated the classifying performance of each factor for polyreactivity. In the known polyreactive features, the net charge of the CDR2 loop, the tryptophan and glycine residues in CDR-H3, and the lengths of the CDR1 and CDR2 loops, importantly contributed to the performance of the models. Additionally, the hydrodynamic features, such as partial specific volume, gyration radius, and isoelectric points of CDR loops and scFvs, were newly added to improve model performance. Finally, we made the prediction model with a robust performance (AUC = 0.840) with an ensemble learning of the top 3 best models.

Conclusion: The prediction models for polyreactivity would help assess polyreactive scFvs in the early discovery stage and our approaches would be promising to develop machine learning models with quantitative data from high throughput assays for antibody screening.

Keywords: Antibody design, Nonspecificity, Polyreactivity, Single-chain fragment variable, Machine Learning, Artificial Intelligence

Introduction

Monoclonal antibodies (mAbs) have been important biological research tools and therapeutic agents due to their attractive properties, such as specific binding, conformational stability, safety for a human, and manufacturability [1, 2]. One of the most important properties of mAbs is specificity and binding affinity through complementary-determining regions (CDRs) to a specific antigen unique to its target. To discover novel mAbs, animal immunization has been traditionally used and it has limited control over specificity and binding affinity because of the difficulty in controlling antigen presentation to the animal immune system [2]. Advanced in vitro technologies, such as phage and yeast surface display, have enabled the rapid isolation of mAbs and improved control over antigen presentation [2]. However, the antibodies initially identified via either immunization or these display methods are not suitable for therapeutic use and usually have some unfavorable biophysical characteristics such as stability, solubility, viscosity, polyreactivity, and so on [1].

Therapeutic mAbs must have the desired biophysical properties. The use of mAbs as therapeutics needs to optimize several important properties, such as binding affinity, specificity, folding stability, solubility, pharmacokinetics, effector functions, and other compatibilities with additional antibody or cytotoxic drugs [2]. Although each property can be addressed through screening large antibody libraries, it is difficult to simultaneously optimize multiple properties of mAbs with the screening methods. Attempts to divide and conquer the properties sequentially are limited by the fact that optimizing one property can worsen other properties. The computational antibody-design methods have been developed to overcome the complexity of optimizing multiple properties of mAbs [2].

Many computational tools have been developed to predict the developability of mAbs at an early stage, which includes high aggregation, and poor stability. The predictive tools for aggregation risk have been developed through the identification of chemical modifications in CDRs [3, 4], semi-empirical methods based on spatial-aggregation-propensity [5–7], and machine learning methods predicting the hydrophobic chromatography retention time [8, 9] and the physicochemical properties such as viscosity and isoelectric point [10]. Many computational tools for predicting stability-enhancing mutations have been developed through the phylogenetic information from multiple sequence alignments, the biomolecular simulations for thermodynamic energies [11–13], CDR-dependent position-specific-scoring-matrix [14], and other machine learning methods [15–17] trained with large databases such as ProTherm [18, 19] and TS50 [17]. Furthermore, the developability scores of mAbs have been quantified to help eliminate undesirable mAbs at an early stage through the Developability Index [5] and Therapeutic Antibody Profiler [20], which integrated the aggregation propensity, CDR lengths, and distributions of hydrophobicity and charges on a surface.

Although current antibody discovery methods focused on the generation of mAbs or fragments with high specificity on target, some mAbs can strongly bind to one target and weakly bind to additional antigens. Polyreactivity is called nonspecificity and is an important property because polyreactive mAbs can show reactivity for diverse off-target epitopes. Early assessment of polyreactivity is important in clinical development, which can allow for the prevention of potentially poor candidates. Because polyreactivity of mAbs is linked to poor antibody pharmacokinetics [21], the experimental assessments

of polyreactivity have been performed with an enzyme-linked immunosorbent assay (ELISA) [22], protein biochip-based ELISAs [23], and Fluorescence-Activated Cell Sorting (FACS)-based high-throughput selections [24–26]. Because the experimental assessments are usually time-consuming, expensive, and tedious, the computational prediction for polyreactivity help assess mAbs and narrow the search space early. The computational analysis tools for polyreactivity have been developed by identifying the biochemical features in CDRs of polyreactive mAbs, which showed the enrichment of glycine, tryptophan, valine, and arginine motifs [25], an increase in inter-loop crosstalk [27], neutral binding surface [27], high isoelectric points [28], constrained β-sheet structures [29], longer CDRH/L3 loops [30], and the occurrence of glutamine residues [30]. Recently, Harvey et al. showed machine learning (ML) models to assess the polyreactivity of nanobodies from protein sequences [26]. The ML models would help the antibody design and diminish experimental efforts because the ML models can allow the quantifications of the polyreactivity of mAbs.

Machine learning (ML) in protein engineering learns the information from data to predict the protein properties of new variants. Prediction models with ML can accelerate the optimization of protein properties by evaluating the new variants, separating the grain from the chaff, and diminishing experimental efforts. To build the ML models, suitable protein descriptors are required to obtain information in protein sequences. For example, one-hot-encoding of amino acids and single amino acid properties can be used to describe protein sequence as a bottom-up approach [31]. However, obtaining meaningful labels and annotations from the explosively growing protein sequences databases needs expensive and tedious experimental resources [32]. Advanced natural language processing (NLP) techniques are applied to self-supervised learning with unlabeled protein sequences in large protein databases, which may extract evolutionary information from protein sequences [32–35]. In addition to protein sequences, extracting useful information from protein structures is important because a protein function is directly related to and depends on its unique 3D protein structure. Figuring protein structures out is known as the folding problem and the prediction of protein structures from protein sequences has been a long-lasting challenge [36]. In the biennial Critical Assessment of protein Structure Prediction conference, deep learning methods such as AlphaFold and trRosetta outperformed other traditional methods [37, 38], and the more advanced methods such as AlphaFold2 and RoseTTA fold showed a better performance with three-track network architectures [36, 39].

In this work, we made prediction models for the polyreactivity of single-chain fragment variables (scFv) with antibody features and NLP descriptors. First, we calculated sequence- and structure-based antibody features of scFvs. In the sequence-based features, we calculated net charges and lengths of CDR loops. To obtain structure-based features, we predicted protein structures of scFvs with trRosetta. And then we calculated aggregation scores, solvent-accessible surface area, and hydrodynamic properties of scFvs. We investigated the classifying performances of the antibody features for polyreactivity with the area under the curve (AUC) and p-values. Second, we made 20 prediction models for the polyreactivity with the antibody features (F46) and NLP descriptors (UniRep, TAPE, ESM-1b, and ESM-1v) using four machine-learning algorithms (GBM, LGBM, RF, and XGB). Third, we made 16 prediction models with the concatenated

descriptors of the antibody features and NLP descriptors to improve model performance. Fourth, we made 14 ensemble models with average- and linear regression-based methods using the 36 prediction models. The prediction models for polyreactivity would help detect the polyreactive scFvs in an early stage and our approaches would help develop machine learning models with high throughput data for antibody screening.

Methods

Dataset construction

The polyreactivity dataset of single-chain fragment variables was derived from the PotaBank [40] and the high-throughput nonspecificity assays by Kelly et al. [25], where a FACS was used to sort depending on whether the scFvs bind to either the soluble membrane preparations or soluble cytosolic preparations in HEK or Sf9 cells, or not. We performed dataset preparation with three steps. First, we removed the duplicate sequences with the same sequences and identical annotations (nonspecific or not). Second, we removed the ambiguous sequences with the same sequences and different annotations. Third, we added pre-gene and post-gene overhang sequences to make full sequences of the scFvs. Finally, We obtained 19,426 sequences, containing the 8867 polyreactive scFvs and 10,559 non-polyreactive scFvs. For the supervised classification task, we prepared a stratified split with an 80% training set (15,540) and a 20% test set (3886) in Python using the Scikit-learn package with a fixed random seed [41].

Performance metrics and statistical analysis

Performances of the prediction models were evaluated using the area under the receiver operating characteristics curve (AUC), accuracy, precision, recall, and F1-score metrics. The AUC in the ROC curve is a performance measurement for classification problems at various threshold settings and indicates how much the prediction model can distinguish the polyreactivity of scFvs. The accuracy is the ratio of the correctly predicted polyreactive and non-polyreactive scFvs to all the experimental polyreactive and non-polyreactive scFvs in the given data set, which represents how the model can correctly classify the polyreactive and non-polyreactive scFvs out of the data set. The precision score is the ratio of correctly predicted polyreactive scFvs to the total predicted polyreactive scFvs in the given data set, which represents the ability to identify all polyreactive scFvs without any non-polyreactive scFvs. The recall score is the ratio of correctly predicted polyreactive scFvs to all the experimental polyreactive scFvs in the given data set, which represents the ability to correctly predict the polyreactive scFvs out of the experimental polyreactive scFvs. F1-score is the harmonic mean of precision and recall scores, which is an alternative to accuracy. In accuracy, precision, recall, and F1-score metrics, we used the criteria of 0.5.

Statistical difference between two groups of polyreactive and non-polyreactive mAbs in each factor was analyzed by the Student's *t*-test and two-tailed tests. The p value was used to indicate a statistically significant difference, where *p value < 0.05, **p value < 0.01, and ***p value < 0.001 are considered in this work.

Homology modeling and protein structure preparation

Homology modeling was performed with transform-restrained Rosetta (trRosetta) [38]. The trRosetta is a deep residual-convolutional network from multiple sequence alignments to make information on the relative distances and orientation of all residue pairs in the protein [38]. And then the restrained minimization was performed to make a protein structure with a fast Rosetta model building protocol with the information from the network [38].

All protein structures from homology modeling were prepared in the following steps. All hydrogen atoms in the protein structures were removed and re-added to the protein structures at pH 7.0. Their positions were optimized with the PROPKA3 implemented in the Maestro program [42]. And then the restrained energy minimization was performed on all protein structures with OPLS3 in the Maestro program within 0.3 Å root mean square deviation [43].

2–4. Aggregation propensities and solvent-accessible surface area.

AggScore [7] is the prediction model for protein aggregation, which is one of the most routinely encountered developability issues [44]. Because the AggScore uses the distribution of hydrophobic and electrostatic patches on the surface of the 3D protein structures and uses the intensity and relative orientation of the surface patches [7], the application domain includes the antibody. Zyaggregator predicts the effects of mutations on the protein aggregation propensity with the physicochemical properties of amino acids [45]. The Zyaggregator score is the sum of Zyaggregator profile Z-scores, whereas the Zyaggregator_p is the normalized score for comparing the proteins which have different lengths. Solvent-accessible surface area (SASA) is the surface of a protein that solvent molecules (water molecules) can access and a probe with the van der Waals radius of a solvent molecule sweeps by rolling over a protein. We calculated the SASA of all hydrophobic atoms (All Hydrophobic SASA) and the exposed hydrophobic atoms (Exposed Hydrophobic SASA). The AggScore, Zyaggregator, and SASA were calculated with the command-line script 'calc_protein_descriptors.py', implemented in Schrodinger suite ver. 2018–3.

Hydrodynamic properties

Hydrodynamic properties of scFvs were calculated with HullRad [46], which uses a convex hull to calculate the smallest convex envelopes with a set of points and to model a hydrodynamic volume of a protein [46]. The 13 factors for hydrodynamic properties are partial specific volume (v_{bar}, mL/g), anhydrous volume sphere radius (R_o, Å), the anhydrous radius of gyration (R_g, Å), maximum dimension (D_{max}, Å), axial ratio, frictional ratio (f/f_0), translational diffusion coefficients (D_t, cm^2/s), translational hydrodynamic radius (R_{trans}, Å), sedimentation coefficients (s, sec), rotational diffusion coefficients (D_r, s^{-1}), rotational hydrodynamic radius (R_{rot}, Å), tumbling correlation time (tauC, ns), and asphericity. The detailed mathematical equations of the factors are well described in Flemin et al. [46].

The lengths and G/Q/R/V/W motifs of CDR loops

Delimitation and numbering of CDR regions in all scFv antibodies were performed with AbRSA [47], where the 40% similarity and Chothia scheme [48] were used. In the CDR lengths, the lengths of the whole CDR regions and only CDR3 regions were calculated with the concatenated sequences of all CDR regions and only CDR3 regions. In the CDR3-G/Q/R/V/W motifs, the occurrences of the glycine, glutamine, arginine, valine, and tryptophan residues in CDR3 regions were calculated with the counts and count-to-length ratio using the concatenated CDR3 regions.

Isoelectric points (IEP)

Isoelectric points (IEP) are the pH, where a molecule has no net electric charge or the statistical mean of the electricity of a molecule is neutral. To estimate the effects of CDR on IEP, we subdivided IEPs into three classes (whole-IEP, CDR-IEP, and CDR3-IEP). The IEP values were predicted with the DTASelect algorithm [49] implemented in pIR [50]. To calculate whole-IEP, CDR-IEP, and CDR3-IEP, the linear approximations were performed with the 25 experimental antibodies' IEPs [51] and 41,943 experimental peptides' IEPs [52] through Eqs. (1) and (2).

$$IEP_{antibody} = 2.0306 * IEP_{DTASelect} - 7.8541 \tag{1}$$

$$IEP_{peptide} = 1.1552 * IEP_{DTASelect} - 0.8839 \tag{2}$$

We applied $IEP_{antibody}$ to the calculation of whole-IEP and $IEP_{peptide}$ to the calculations of CDR-IEP and CDR3-IEP. The concatenated sequences were used to calculate the CDR-IEP and CDR3-IEP.

Nanobody polyreactivity

The two prediction models for the polyreactivity of nanobodies were developed by Harvey et al. [26], which are one-hot embedding (OneHot-CDRS) and 3-mer embedding (3MER-CDRS) logistic regression models. The OneHot-CDRS model learned weights for each amino acid type at each position in the CDR sequences, whereas the 3MER-CDRS model learned weights for each motif of polyreactive nanobodies [26]. Although the models were applied to 19,426 scFvs, the polyreactivity scores of 16,337 scFvs were predicted. The AUC scores of OneHot-CDRS and 3MER-CDRS models were calculated with only 16,337 scFvs.

Protein language-based descriptors (UniRep, TAPE, ESM-1b, and ESM-1v)

Natural language processing (NLP) techniques have been applied to extracting useful evolutionary information from unlabeled protein sequences with self-supervised learning [32–35]. We used UniRep, TAPE, ESM-1b, and ESM-1v descriptors for the NLP-based protein sequence descriptors. The UniRep used the UniRef50 database and was based on a four-layer multiplicative LSTM with 256 hidden units, leading to 18.2 M parameters and 1900 features [33]. The Tasks Assessing Protein Embeddings (TAPE) used the Pfam database and was based on a 12-layer Transformer with a hidden size of 512 units and 8 attention heads, leading to 38 M parameters and 768 features [32]. The Evolutionary Scale Modeling-1b (ESM-1b) used the high-diversity sparse UniRef50

dataset and was based on a 33-layer Transformer with a hidden size of 5120 units and 20 attention heads, leading to 650 M parameters and 1280 features [34]. The Evolutionary Scale Modeling-1v (ESM-1v) had the same architecture as ESM-1b (650 M parameters and 1280 features), but ESM-1v are ensembles of 5 models, used the UniRef90 dataset, and employed zero-shot inference to predict a new class unseen in training sets [35]. The UniRep descriptors were constructed with a concatenation of average hidden unit outputs, the final hidden unit, and the final cell, whereas the TAPE, ESM-1b, and ESM-1v descriptors were constructed with average hidden unit outputs.

Machine learning algorithms, hyperparameter tuning, and ensemble learning

Prediction models for polyreactivity of single-chain antibody fragments were constructed using the gradient boosting (GBM) classifier, the random forest (RF) classifier, the light GBM (LGBM) classifier, and the extreme gradient boosting (XGB) classifier models. The GBM combines many weak-leaning models, such as a decision tree, to make a strong prediction model and is based on additive expansions in a forward stage-wise fashion [41, 53]. The RF is a meta-classifier with many classifying decision trees on various subsamples of the dataset, and it uses averaging to improve the predictive accuracy and to control overfitting problems [41]. The LGBM and XGB use the gradient boosting framework, but the difference is how to grow decision trees. The LGBM builds each decision tree in a leaf-wise fashion [54], whereas the XGB builds each decision tree in a depth-wise fashion [55].

In training, we used tenfold cross-validation with GridSearchCV in the Scikit-learn package [41] using hyperparameter settings (Additional file 1: Table S1). The best hyperparameters from the GridSearchCV were selected with the performance in cross-validation sets. And then the final model with the best hyperparameters was re-trained with all training sets. Some hyperparameters in GBM were incorporated for tuning the models; the number of boosting stages (n_estimators), the maximum depth of the individual regression estimators (max_depth), the number of features to consider when finding the best split (max_features), and the boosting learning rate (learning_rate) [41, 53]. Some hyperparameters in LGBM were incorporated for tuning the models; n_estimators and learning_rate [54]. Some hyperparameters in RF were incorporated for tuning the models; n_estimators and max_features [41]. Some hyperparameters in XGB were incorporated for tuning the models; n_estimators, max_depth, and learning_rate [55].

Ensemble learning through a meta-learning classifier

Ensemble learning through a meta-learning classifier was performed with average-based (AVG) and linear regression-based (LR) methods. In the AVG method, we calculated the average of the probabilities from the pre-trained selected models. In the LR method, we trained a simple linear regression model without an intercept term using the probabilities from the pre-trained selected models in the training set. The difference between the AVG and LR strategies is that the contribution to the final probability of each pre-trained model is the same in the AVG strategy, whereas the contribution of each pre-trained model is not the same in the LR strategy. Machine learning learns how to best use input features and information to predict nonspecificity, whereas ensemble learning learns how to best use the machine learning models to predict nonspecificity.

Results

Computational prediction of the polyreactivity of antibodies is important in evaluating the developability of antibodies at an early stage. The workflow to make computational prediction models for the polyreactivity in scFvs is shown in Fig. 1.

Each performance of biochemical patterns in scFvs for polyreactivity

Most prediction methods for polyreactivity have focused on the identification of the biochemical features which the polyreactive antibodies [25–30]. The biochemical patterns in polyreactive antibodies have been analyzed through an increase in neutral binding surface [27], longer CDRH/L3 loops, an increase of glycine, tryptophan, valine, and arginine motifs [25], the occurrence of glutamine residues [30], and high isoelectric points [28]. To investigate each classifying performance of the sequence-based and structure-based features, we predicted 19,426 scFv antibody structures with trRosetta and calculated the area under the receiver operating characteristic curves (AUC) with the 51 biochemical features (Additional file 1: Table S3 and Additional file 1: Table S4). The distributions of the statistically significant features (p value < 0.001) are illustrated in Additional file 1: Figure S1 and Additional file 1: Figure S2.

A slightly hydrophilic and neutral-charged binding surface can have weak interactions with various ligands [27]. To investigate the classifying performance of the increased neutral binding surfaces, we calculated the area under the ROC curves (AUC) with the

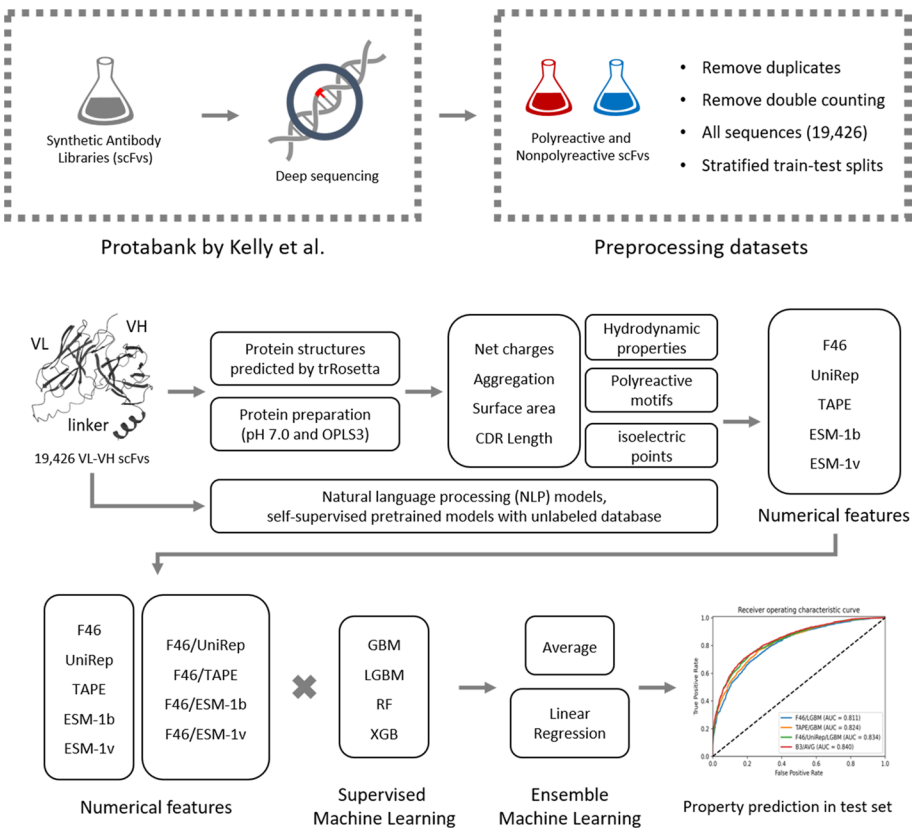

Fig. 1 Workflow to make prediction models for polyreactivity in this work

net charges of CDR loops, spatial aggregation propensities (SAP), and solvent-accessible surface area (SASA) with the predicted structures. Firstly, we delimited the CDR of scFvs with the Chothia scheme and calculated the net charges of all CDRs, CDR1, CDR2, and CDR3 loops. The AUC scores of the net charges of all CDRs and CDR1, CDR2, and CDR3 loops are 0.521, 0.556, 0.413, and 0.521, respectively. Secondly, we predicted the SAP of scFvs with AggScore and Zyggregator with the predicted structures. The AUC scores of AggScore, Zyggregator, and Zyggregator_p are 0.506, 0.489, and 0.541, respectively. Thirdly, we calculated the SASA of all hydrophobic atoms and the exposed hydrophobic atoms with the predicted protein structures. The AUC scores of the hydrophobic SASA of exposed residues and all residues are 0.512 and 0.502.

In addition to the SAP and SASA, the molecular-scale hydrodynamic effects are related to the cavity-ligand binding due to the capillary fluctuations [56]. Hydrodynamic properties can be used to estimate the size and shape of the proteins in solution [46] because the hydrodynamic radius in a protein involves the motion of the protein relative to the aqueous solvent where the protein is dissolved [57]. We calculated the 14 hydrodynamic properties and measured the classifying performance for polyreactivity in scFvs. In the 14 properties, the AUC scores of the anhydrous radius of gyration, asphericity, and frictional ratio are 0.628, 0.621, and 0.600, respectively. The three factors of the gyration radius, frictional ratio, and asphericity are related to the hydration effect [46, 58], they are relatively better predictive of the polyreactivity in scFvs than other hydrodynamic properties.

Lecerf et al. reported that the hydrophobicity and propensity for aggregation of mAbs are associated with the longer CDRH/L3 loops, but there is no significant correlation between the size of hypervariable loops and the polyreactivity [30]. To investigate the correlation between the length of CDR loops and polyreactivity, we measured the lengths of all CDRs and CDR-1/2/3 loops in scFvs. The AUC scores of the lengths of all CDRs, CDR1, CDR2, and CDR3 are 0.482, 0.393, 0.530, and 0.545. The lower AUC scores mean that longer CDR lengths cannot distinguish the nonspecificity in scFvs. On the other hand, shorter lengths of CDR1 showed a relatively better classifying performance (AUC = 0.607). The results were in agreement with Lecerf et al. [30], where the mAbs with shorter CDR loops might tend to reduce the risk of polyreactivity.

Enrichment of the glycine-, glutamine-, arginine-, valine-, and tryptophan motifs in CDR-H3 is associated with polyreactivity [25, 30]. To investigate the classifying performance of each motif, we calculated the AUC scores of the number and ratio of each motif in CDR3 and CDR-H3 (G, Q, R, V, W, VV, and WW motifs). Most motifs showed low AUC scores between 0.450 and 0.550, but the Trp motifs even misled the classification (AUC = 0.383).

Isoelectric points (IEP) of mAbs are important in solution behavior and related to viscosity [10]. Because therapeutic antibodies need to be positively charged for efficient fluid-phase endocytosis at the physiological pH of 7.4, an IEP in the range of 8–9 is desirable. To investigate the correlation between IEPs and polyreactivity, we predicted the IEPs of scFvs based on the full sequences and concatenated CDR loops. Because most IEP prediction methods have been developed for peptides and proteins [49, 52], we corrected the predicted IEP with experimental IEPs from peptides and antibodies and prediction models for peptides and antibodies (Additional file 1: Figure S3). To predict

the IEPs for CDR loops, we collected the experimental IEPs of 41,943 peptides and made a prediction model of $R^2 = 0.9845$ and RMSE = 0.2743. Whereas, to predict the IEPs for antibodies, we collected the experimental IEPs of 25 antibodies and made a prediction model of $R^2 = 0.9603$ and RMSE = 0.1669. The AUC scores of the predicted IEPs of CDRs and scFvs are 0.504 and 0.535, respectively.

Although the length of CDR1, anhydrous gyration radius, frictional ratio, and asphericity showed relatively high AUC over 0.6, the 40 biochemical features in scFvs showed low AUC if each pattern was used to classify the polyreactive scFvs alone. Because a single biochemical feature in scFvs is not enough to distinguish and predict the polyreactive scFvs, it is necessary to build machine learning models to predict the polyreactivity of the scFvs.

Machine learning models with biochemical features

Machine learning models can utilize the information of the scFvs to predict the polyreactivity of the scFvs. We developed machine learning models with the combination of the biochemical features and four NLP-based descriptors (UniRep, TAPE, ESM-1b, and ESM-1v). To compare the performance of the models, we used the AUC scores in the test set after tenfold cross-validation and refitting for our-own models. And then we built baselines with the best AUC of the single features and the AUC for the two previously developed models for antibody fragments by Harvey et al. [26] (one-hot-CDRS and 3mer-CDRs). The ROC plots of the two models are illustrated in Fig. 2A.

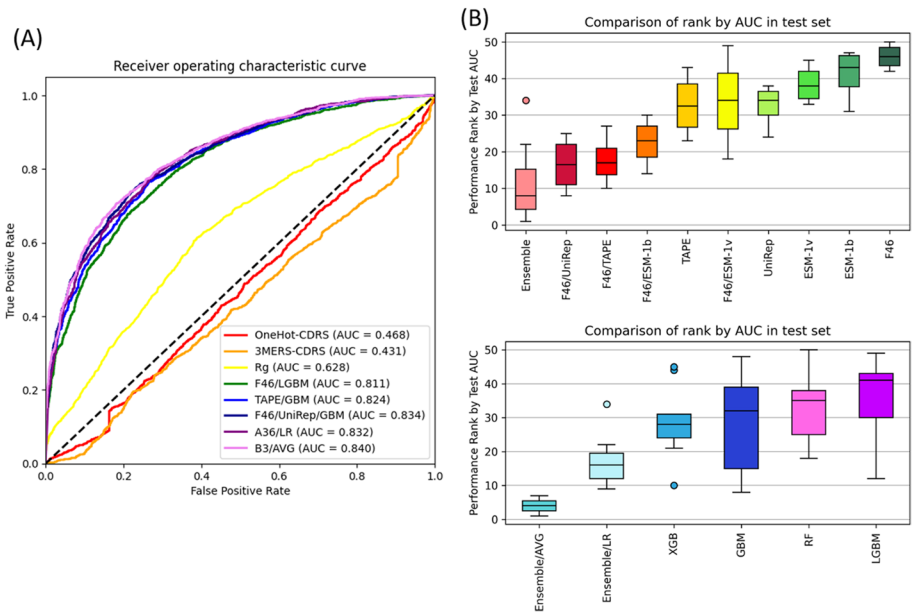

Fig. 2 Performance metrics of the models in this work. **A** ROC plots for the best models from the combinations of descriptors and algorithms. The 'OneHot-CDRS' and '3MERS-CDRS' are the reported models by Harvey et al. (ref. 26) for nanobody and we used them as baselines. The 'Rg' is the anhydrous radius of gyration and the best single factor in this work. The best models were trained with the optimal hyperparameter after grid search. The AUC metric was measured in the test set. **B** Boxplot plot for the relative comparison rank of descriptors and algorithms. We put all the models together and ranked the descriptors and algorithms by the AUC metric in the test set

The anhydrous radius of gyration showed the best performance (AUC = 0.628) in the single biochemical features for predicting polyreactive scFvs. To build the baselines with the previously developed models, we applied the two polyreactive prediction models for antibody fragments by Harvey et al. [26] (one-hot-CDRS and 3mer-CDRS) to the scFvs. Due to the IMGT numbering scheme with ANARCI in Harvey et al. [26], the polyreactivities of only 16,337 scFvs were used to calculate the AUC values. The AUC values of one-hot-CDRS and 3mer-CDRS for scFvs are 0.467 and 0.428.

To find the best model with the biochemical features, we made four machine learning models (GBM, LGBM, RF, and XGB) with 46 biochemical patterns (F46) except for two SASA and three aggregation factors. The performance metrics of the four models are summarized in Table 1. The optimal hyperparameters of the models are summarized in Additional file 1: Table S2. The prediction model from F46 and LGBM showed the best performance in the test set prediction (AUC = 0.811). The best model was trained with the optimal hyperparameter (learning_rate = 0.01 and n_estimators = 500). Accuracy, precision, recall, and F1-score of the best model (F46/LGBM) are 0.762, 0.732, 0.756, and 0.744 in the training set, whereas those are 0.731, 0.700, 0.720, and 0.710 in the test set. The ROC plot of the best model (F46/LGBM) is illustrated in Fig. 2A.

We calculated the feature importance in the best model to investigate the contributions of the 46 biochemical features to the best model (F46/LGBM). The top 11 important features in the best model had 50.87% contributions to the performance, which are the net charge of the CDR2 loop (7.25%), partial specific volume (6.00%), the isoelectric point of CDR loops (5.90%), the anhydrous radius of gyration (4.45%), the ratio of tryptophan residues in CDR-H3 (4.41%), the isoelectric point of scFv (4.21%), the CDR1 length (4.01%), the anhydrous volume sphere radius (3.97%), the CDR2 length

Table 1 Performance of models with different ML algorithms and single descriptors

Descriptor	Method	Train AUC	Valid AUC	Test AUC	Accuracy	Precision	Recall	F1-score
F46	GBM	0.917 ± 0.006	0.705 ± 0.129	0.805	0.730	0.753	0.607	0.672
	LGBM	0.858 ± 0.009	0.704 ± 0.124	0.811	0.731	0.700	0.720	0.710
	RF	1.000 ± 0.000	0.697 ± 0.123	0.795	0.728	0.722	0.658	0.689
	XGB	0.951 ± 0.003	0.701 ± 0.125	0.810	0.732	0.705	0.710	0.708
UniRep	GBM	1.000 ± 0.000	0.591 ± 0.177	0.821	0.747	0.741	0.686	0.712
	LGBM	0.926 ± 0.006	0.606 ± 0.178	0.816	0.734	0.704	0.722	0.713
	RF	1.000 ± 0.000	0.575 ± 0.178	0.815	0.736	0.728	0.671	0.699
	XGB	0.999 ± 0.000	0.596 ± 0.181	**0.824**	0.740	0.718	0.709	0.713
TAPE	GBM	1.000 ± 0.000	0.647 ± 0.160	**0.824**	0.745	0.741	0.679	0.709
	LGBM	0.919 ± 0.005	0.657 ± 0.155	0.810	0.731	0.703	0.713	0.708
	RF	1.000 ± 0.000	0.638 ± 0.160	0.815	0.745	0.747	0.666	0.704
	XGB	0.998 ± 0.000	0.651 ± 0.156	0.822	0.746	0.729	0.706	0.717
ESM-1b	GBM	0.979 ± 0.003	0.603 ± 0.182	0.807	0.727	0.742	0.616	0.673
	LGBM	0.922 ± 0.006	0.608 ± 0.176	0.807	0.730	0.697	0.722	0.710
	RF	1.000 ± 0.000	0.593 ± 0.176	0.814	0.736	0.730	0.669	0.698
	XGB	0.998 ± 0.000	0.604 ± 0.177	0.821	0.741	0.718	0.713	0.716
ESM-1v	GBM	1.000 ± 0.000	0.594 ± 0.150	0.819	0.743	0.761	0.639	0.695
	LGBM	0.919 ± 0.005	0.602 ± 0.172	0.813	0.730	0.694	0.729	0.711
	RF	1.000 ± 0.000	0.582 ± 0.171	0.816	0.740	0.735	0.674	0.703
	XGB	0.949 ± 0.002	0.597 ± 0.175	0.808	0.728	0.702	0.701	0.701

The bold means the best performance, the AUC score in the test set

(3.85%), the count of glycine residues in CDR-H3 (3.61%), and sedimentation coefficients (3.21%). The net charge of the CDR2 loop and the counts and ratios of tryptophan and glycine residues are associated with the known biochemical features of the neutral binding surface [27] and the enrichment of tryptophan and glycine motifs [25]. The four hydrodynamic properties (partial specific volume, the anhydrous radius of gyration, the anhydrous volume sphere radius, and sedimentation coefficients) had 17.63% contributions to the performance of the best model, indicating that the solution behavior is important in the polyreactivity of the scFvs.

Machine learning models with NLP-based descriptors

Natural language processing methods have been applied to large unlabeled protein sequence data sets through self-supervised learning, which leads to language model-based descriptors. Because the language model-based descriptors can extract evolutionary information from protein sequences, we used four language model-based descriptors (UniRep, TAPE, ESM-1b, and ESM-1v) to extract information from scFvs' sequences.

To find the best model with the language model-based descriptors, we made 16 machine learning models with the combination of four machine learning algorithms (GBM, LGBM, RF, and XGB) and four language model-based descriptors (UniRep, TAPE, ESM-1b, and ESM-1v). The performance metrics of the 16 models are summarized in Table 1. The ROC plot of the best model (TAPE/GBM) is illustrated in Fig. 2A.

In the UniRep, the prediction model from XGB showed the best performance in the test set (AUC = 0.824). The best model (UniRep/XGB) was trained with the optimal hyperparameter (learning_rate = 0.01, max_depth = 10, and n_estimators = 500). The accuracy, precision, recall, and F1-score of the best model (UniRep/XGB) are 0.740, 0.718, 0.709, and 0.713 in the test set, respectively. In the TAPE, the prediction model from GBM showed the best performance in the test set (AUC = 0.824). The best model (TAPE/GBM) was trained with the optimal hyperparameter (learning_rate = 0.01, max_depth = 10, max_features = 'sqrt', and n_estimators = 1000). The accuracy, precision, recall, and F1-score of the best model (TAPE/GBM) are 0.745, 0.741, 0.679, and 0.709 in the test set, respectively. In the ESM-1b, the prediction model from XGB showed the best performance in the test set (AUC = 0.821). The best model (ESM-1b/XGB) was trained with the optimal hyperparameter (learning_rate = 0.01, max_depth = 10, and n_estimators = 500). The accuracy, precision, recall, and F1-score of the best model (ESM-1b/XGB) are 0.741, 0.718, 0.713, and 0.716 in the test set, respectively. In the ESM-1v, the prediction model from GBM showed the best performance in the test set (AUC = 0.819). The best model (ESM-1v/GBM) was trained with the optimal hyperparameter (learning_rate = 0.05, max_depth = 15, max_features = 'log2', and n_estimators = 3000). The accuracy, precision, recall, and F1-score of the best model (ESM-1v/GBM) are 0.743, 0.761, 0.639, and 0.695 in the test set, respectively.

To compare the performance of the machine learning models from biochemical features and language model-based descriptors, we used the AUC in tenfold cross-validation test sets and five metrics in the test set (AUC, accuracy, precision, recall, and F1-score metrics).

In the AUC metric in tenfold cross-validation sets, the F46/LGBM model showed the best performance (0.704 ± 0.124), whereas the other models had relatively low mean

and high standard deviations of AUC in tenfold cross-validation sets. It indicated that the language model-based descriptors are more sensitive to the data splits than the 46 biochemical features. In the AUC metric in the test set, the TAPE/GBM model showed the best performance (0.824), whereas the other models also showed a robust performance over 0.8. The ROC plot of the best model (TAPE/GBM) is illustrated in Fig. 2A. In the accuracy in the test set, the TAPE/GBM model showed the best accuracy (0.745), whereas the other models also had similar accuracy over 0.730. In the precision metric in the test set, the ESM-1v/GBM showed the best precision score (0.761), whereas the other models also had a robust precision score of over 0.7. In the recall metric in the test set, the F46/LGBM model showed the best recall score (0.720), whereas the TAPE/GBM and ESM-1v/GBM models showed relatively low recall scores (0.679 and 0.639, respectively). In the F1-score metric in the test set, the ESM-1b/XGB showed the best F1-score (0.716), whereas the ESM-1v/GBM model showed the worst F1-score (0.695) and the other models showed a robust F1-score over 0.7.

Machine learning models with both biochemical features and language model-based descriptors

To improve model performance, we concatenated the 46 biochemical features and four language model-based descriptors and made the four descriptors (F46/UniRep, F46/TAPE, F46/ESM-1b, and F46/ESM-1v). And then we made the 16 machine learning models with the combinations of the four descriptors and four machine learning algorithms (GBM, LGBM, RF, and XGB), the performance metrics of which are summarized in Table 2. The optimal hyperparameters of the models are summarized in Additional file 1: Table S2.

In the F46/UniRep, the prediction model from GBM showed the best performance in the test set (AUC = 0.834). The ROC plot of the best model (F46/UniRep/GBM) is illustrated in Fig. 2A. The best model (F46/UniRep/GBM) was trained with the optimal

Table 2 Performance of models with different ML algorithms and the concatenated descriptors

Descriptor	Method	Train AUC	Valid AUC	Test AUC	Accuracy	Precision	Recall	F1-score
F46/UniRep	GBM	1.000 ± 0.000	0.624 ± 0.163	**0.834**	0.759	0.751	0.706	0.728
	LGBM	0.928 ± 0.006	0.638 ± 0.166	0.830	0.746	0.719	0.729	0.724
	RF	1.000 ± 0.000	0.594 ± 0.174	0.823	0.747	0.742	0.683	0.711
	XGB	0.971 ± 0.002	0.624 ± 0.166	0.826	0.752	0.728	0.729	0.729
F46/TAPE	GBM	1.000 ± 0.000	0.669 ± 0.137	0.829	0.748	0.745	0.680	0.711
	LGBM	0.919 ± 0.006	0.676 ± 0.147	0.826	0.746	0.720	0.726	0.723
	RF	1.000 ± 0.000	0.654 ± 0.153	0.823	0.749	0.752	0.671	0.709
	XGB	0.997 ± 0.000	0.668 ± 0.151	0.831	0.748	0.727	0.718	0.722
F46/ESM-1b	GBM	1.000 ± 0.000	0.643 ± 0.154	0.830	0.755	0.746	0.704	0.724
	LGBM	0.882 ± 0.008	0.654 ± 0.157	0.821	0.744	0.709	0.746	0.727
	RF	1.000 ± 0.000	0.622 ± 0.169	0.826	0.750	0.748	0.680	0.713
	XGB	0.954 ± 0.004	0.648 ± 0.153	0.823	0.750	0.721	0.738	0.729
F46/ESM-1v	GBM	0.985 ± 0.002	0.643 ± 0.139	0.814	0.738	0.741	0.654	0.695
	LGBM	0.834 ± 0.010	0.657 ± 0.134	0.802	0.727	0.677	0.766	0.719
	RF	1.000 ± 0.000	0.611 ± 0.165	0.827	0.751	0.750	0.681	0.714
	XGB	0.951 ± 0.003	0.643 ± 0.151	0.822	0.743	0.709	0.742	0.725

The bold means the best performance, the AUC score in the test set

hyperparameter (learning_rate = 0.01, max_depth = 10, max_feature = 'auto', and n_estimators = 1000). The accuracy, precision, recall, and F1-score of the best model (F46/UniRep/GBM) are 0.759, 0.751, 0.706, and 0.728 in the test set, respectively. In the F46/TAPE, the prediction model from XGB showed the best performance in the test set (AUC = 0.831). The best model (F46/TAPE/XGB) was trained with the optimal hyperparameter (learning_rate = 0.01, max_depth = 10, and n_estimators = 500). The accuracy, precision, recall, and F1-score of the best model (F46/TAPE/XGB) are 0.748, 0.727, 0.718, and 0.722 in the test set, respectively. In the F46/ESM-1b, the prediction model from GBM showed the best performance in the test set (AUC = 0.830). The best model (F46/ESM-1b/GBM) was trained with the optimal hyperparameter (learning_rate = 0.01, max_depth = 10, max_feature = 'auto', and n_estimators = 500). The accuracy, precision, recall, and F1-score of the best model (F46/ESM-1b/GBM) are 0.755, 0.746, 0.704, and 0.724 in the test set, respectively. In the F46/ESM-1v, the prediction model from RF showed the best performance in the test set (AUC = 0.827). The best model (F46/ESM-1v/RF) was trained with the optimal hyperparameter (max_feature = 'auto', and n_estimators = 500). The accuracy, precision, recall, and F1-score of the best model (F46/ESM-1b/GBM) are 0.751, 0.750, 0.681, and 0.714 in the test set, respectively.

Ensemble models to improve the performance

To improve model performance, we performed ensemble learning with the 36 machine learning models from the previous hyperparameter optimization steps. We made 14 ensemble models with the two combination methods and two ensemble learnings (average-based and linear regression-based methods). We ranked the 36 models with the AUC metric in the test set. In one combination, we selected all, the top 10, top 5, and top 3 models in all 36 models, which led to A36, A10, A5, and A3 models, respectively. In the other combination, we selected all models, the top 5, and top 3 models in the best models of nine protein descriptor sets, which led to B9, B5, and B3 models, respectively. The performance metrics are summarized in Table 3.

In the average-based ensemble learning (AVG), the prediction model from B3 showed the best performance in the test set (AUC = 0.840). The accuracy, precision, recall, and F1-score of the best model (B3/AVG) are 0.765, 0.755, 0.717, and 0.735 in the test set, respectively. The prediction models from A10/AVG, A5/AVG, A3/AVG, and B5/AVG tied for the second performance in the test set (AUC = 0.839). The prediction model from A36/AVG also showed a slightly better performance than the base model (F46/UniRep/GBM), but it showed the worst performance in the average-based ensemble models. In the linear regression-based ensemble learning (LR), the prediction model from A36 showed the best performance in the test set (AUC = 0.832). The accuracy, precision, recall, and F1-score of the best model (A36/LR) are 0.754, 0.773, 0.653, and 0.708 in the test set, respectively. The models from LR-based ensemble learning showed worse performance than the baseline (F46/UniRep/GBM), which may be from the overfitting problem.

Performance comparison of the prediction models.

To compare evaluation metrics in protein descriptors and machine learning algorithms, the performance of the 50 final models was ranked according to the mean value of decreasing order for the AUC in the test set in Fig. 2B. The Ensemble won in protein

Table 3 Performance of ensemble models with the trained model combinations

Name	Method	Test AUC	Accuracy	Precision	Recall	F1-score
A36	Average-based Ensemble (AVG)	0.836	0.758	0.750	0.703	0.726
A10		0.839	0.764	0.757	0.710	0.733
A5		0.839	0.760	0.755	0.701	0.727
A3		0.839	0.764	0.751	0.724	0.737
B9		0.838	0.754	0.751	0.688	0.718
B5		0.839	0.759	0.748	0.711	0.729
B3		**0.840**	0.765	0.755	0.717	0.735
A36	Linear Regression-based Ensemble (LR)	0.832	0.754	0.773	0.653	0.708
A10		0.828	0.748	0.745	0.680	0.711
A5		0.828	0.748	0.745	0.680	0.711
A3		0.831	0.759	0.751	0.706	0.728
B9		0.819	0.743	0.761	0.639	0.695
B5		0.830	0.758	0.750	0.705	0.727
B3		0.825	0.757	0.748	0.705	0.726
TAPE/GBM		0.824	0.745	0.741	0.679	0.712
F46/UniRep/GBM		0.834	0.759	0.751	0.706	0.728

The bold means the best performance, the AUC score in the test set

descriptor ranking, followed by F46/UniRep, F46/TAPE, F46/ESM-1b, TAPE, F46/ESM-1v, UniRep, ESM-1v, ESM-1b, and F46. The Ensemble/AVG won in machine learning algorithm ranking, followed by Ensemble/LR, XGB, GBM, RF, and LGBM.

Discussion

Antibodies have been successful biological drugs with over 100 molecules approved for therapeutic use and hundreds more in clinical development. Improved high-throughput technologies enable to find of very specific antibodies against targets, but some can show polyreactivity with low affinity for multiple epitopes. Because polyreactive antibodies have potential side effects through multiple epitopes, previous studies have identified the biochemical patterns and characteristics of polyreactive antibodies. Here, we constructed machine learning models to predict the polyreactivity of scFvs with antibody features and NLP descriptors. The computational frameworks to predict the polyreactivity of a given scFv would be useful by evaluating the potential fate of a therapeutic antibody and the potential efficacy with natural immune systems in the early discovery stage. Many studies have focused to identify the biochemical polyreactive features in antibodies [25, 27–30], but the single factors have low AUC performance to classify the polyreactive scFvs. Recently, the in silico method has been developed to predict the polyreactivity of antibody fragment [26], but it focused on nanobodies and the application to scFvs showed low AUC performance. Therefore, computational prediction models for polyreactivity in scFvs in this work could be used to support the isolation of the potential polyreactive scFvs in the process of therapeutic antibody screening. Moreover, similar approaches with the protein structural features from protein structure prediction methods and NLP descriptors would be promising and useful to make machine learning models for industrial enzymes and protein drugs.

Advanced natural language processing technologies have enabled us to learn statistical representations and evolutionary information of protein sequences with continuously increased unlabeled protein sequence databases from advanced sequencing technologies. The NLP-based protein descriptors (UniRep, TAPE, ESM-1b, and ESM-1v) can capture the polyreactive features of scFv sequences to classify the polyreactive scFvs. The machine learning models with the NLP-based descriptors showed moderate performance, but they are more sensitive to data splits than structural features. Many language models have been proposed with large-scale databases of protein sequences over the families of related protein sequences [32–34, 59]. Although we used four NLP descriptors in this work, the more advanced NLP descriptors with large-scale databases would improve the model performance. The NLP protein descriptors have the potential for diverse protein engineering tasks [31] because it enables us to compare the protein sequences too diverse to perform multiple sequence alignment analysis. However, there is an inevitable limitation of the sequence-function gap, because protein functions are from the accurately folded protein structures.

Protein structures form the basis of the structure–activity relationship. Although the protein structural features help the analysis of the relationship, the relatively small number of experimentally determined protein structures has set a limit on wide applications of machine learning and deep learning approaches with the determined protein structures. Protein structure prediction methods such as AlphaFold [37, 39], trRosetta [38], and RoseTTAFold [36], have enabled the rapid generation of protein structural features for machine learning approaches. Although we used the trRosetta method to predict the protein structures of scFvs due to the computational cost, AlphaFold 2 and RoseTTAFold have been known to outperform trRosetta to predict antibody structures [39, 60], which can improve model performance with the structural features from the more accurate protein structures. The more accurate and rapid protein structure prediction methods would accelerate the applications of machine learning and deep learning approaches with protein structures.

Not only the development of protein descriptors but also the advance of machine learning algorithms would help make prediction models. The decision tree models such as GBM, LGBM, RF, and XGB used in this work are ensemble methods with weak learners and are computationally efficient models and have been used in various classification tasks [61, 62]. Although we use only the decision-tree-based ensemble models, there are alternative or possibly better machine learning algorithms. Kernel methods, such as support vector machine (SVM), calculate the similarity between inputs and implicitly project the features into a high dimensional space. The SVM has been successfully applied to various classification tasks [63, 64] and worked well when there is a clear margin of separation between two classes. However, the SVM requires high computational cost and is not suitable for large data sets, because the SVM needs to calculate the similarity between input features. Recently, a canonical deep neural network architecture for tabular data (TabNet) was developed and showed better performance than the decision tree models for some supervised learning and semi-supervised learning tasks [65]. Therefore, to find a suitable combination for the training data, it would be helpful to build models with

diverse machine learning algorithms and compare the performance their performance in the future.

Conclusion

Monoclonal antibodies have been essential biological therapeutic agents, which require optimizing many physical properties for clinical development after in vitro library screening. Polyreactivity is one of the most important properties in clinical development because it leads to poor pharmacokinetic properties and potential poor candidates. We made prediction models with the known polyreactive antibody features and NLP descriptors, where we predicted all scFv protein structures with trRosetta to calculate structure-based features. The best model in this work showed a robust performance (AUC = 0.840) with 76.5% accuracy and 75.5% precision rates. Therefore, computational prediction for polyreactivity with our models would help detect the polyreactive mAbs and allow for the prevention of potentially poor candidates in the early discovery stage. Furthermore, our approaches would be promising to make machine learning models with quantitative data from high throughput assays for industrial enzyme and antibody screening.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-05010-4.

Additional file 1. Supplementary figures and tables.

Acknowledgements
This research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE), Korea, under the "Infrastructure Support Program for Industry Innovation" (reference number P0014714) supervised by the Korea Institute for Advancement of Technology (KIAT).

Author contributions
H.L. conceptualized this study. K.N. advised on the study. All authors reviewed the manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets used in the present research are available on the Github repository (https://github.com/hclim0213/Pred_nonspecificity_scFvs).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 September 2022 Accepted: 26 October 2022
Published online: 05 December 2022

References

1. Rabia LA, Desai AA, Jhajj HS, Tessier PM. Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. Biochem Eng J. 2018;137:365–74.
2. Tiller KE, Tessier PM. Advances in antibody design. Annu Rev Biomed Eng. 2015;17:191.
3. Lu X, Nobrega RP, Lynaugh H, Jain T, Barlow K, Boland T, Sivasubramanian A, Vásquez M, Xu Y. Deamidation and isomerization liability analysis of 131 clinical-stage antibodies. MAbs. 2019;11:45–57.

4. Xu A, Kim HS, Estee S, ViaJar S, Galush WJ, Gill A, Hötzel I, Lazar GA, McDonald P, Andersen N. Susceptibility of antibody CDR residues to chemical modifications can be revealed prior to antibody humanization and aid in the lead selection process. Mol Pharm. 2018;15:4529–37.
5. Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. J Pharm Sci. 2012;101:102–15.
6. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Prediction of aggregation prone regions of therapeutic proteins. J Phys Chem B. 2010;114:6614–24.
7. Sankar K, Krystek SR Jr, Carl SM, Day T, Maier JK. AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. Proteins Struct Funct Bioinform. 2018;86:1147–56.
8. Hanke AT, Klijn ME, Verhaert PD, van der Wielen LA, Ottens M, Eppink MH, van de Sandt EJ. Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties. Biotechnol Prog. 2016;32:372–81.
9. Jain T, Boland T, Lilov A, Burnina I, Brown M, Xu Y, Vásquez M. Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. Bioinformatics. 2017;33:3758–66.
10. Thorsteinson N, Gunn JR, Kelly K, Long W, Labute P. Structure-based charge calculations for predicting isoelectric point, viscosity, clearance, and profiling antibody therapeutics. mAbs. 2021;13:1981805.
11. Seeliger D, De Groot BL. Protein thermostability calculations using alchemical free energy simulations. Biophys J. 2010;98:2309–16.
12. Buß O, Rudat J, Ochsenreither K. FoldX as protein engineering tool: Better than random based approaches? Comput Struct Biotechnol J. 2018;16:25–33.
13. Wang B, Qi Y, Gao Y, Zhang JZ. A method for efficient calculation of thermal stability of proteins upon point mutations. Phys Chem Chem Phys. 2020;22:8461–6.
14. Warszawski S, Borenstein Katz A, Lipsh R, Khmelnitsky L, Ben Nissan G, Javitt G, Dym O, Unger T, Knop O, Albeck S. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. PLoS Comput Biol. 2019;15: e1007207.
15. Montanucci L, Capriotti E, Frank Y, Ben-Tal N, Fariselli P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinform. 2019;20:1–10.
16. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: predicting the stability change of protein point mutations using neural networks. J Chem Inf Model. 2019;59:1508–14.
17. Harmalkar A, Rao R, Honer J, Deisting W, Anlahr J, Hoenig A, Czwikla J, Sienz-Widmann E, Rau D, Rice A. Towards generalizable prediction of antibody thermostability using machine learning on sequence and structure features. bioRxiv. 2022.
18. Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A. ProTherm: thermodynamic database for proteins and mutants. Nucleic Acids Res. 1999;27:286–8.
19. Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res. 2021;49:D420–4.
20. Raybould MI, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi J, Deane CM. Five computational developability guidelines for therapeutic antibody profiling. Proc Natl Acad Sci. 2019;116:4025–30.
21. Hötzel I, Theil FP, Bernstein LJ, Prabhu S, Deng R, Quintana L, Lutman J, Sibia R, Chan P, Bumbaca D. A strategy for risk mitigation of antibodies with fast clearance. mAbs. 2012;4:753–60.
22. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. Science. 2003;301:1374–7.
23. Lueking A, Beator J, Patz E, Müllner S, Mehes G, Amersdorfer P. Determination and validation of off-target activities of anti-CD44 variant 6 antibodies using protein biochips and tissue microarrays. Biotechniques. 2008;45:i–v.
24. Xu Y, Roach W, Sun T, Jain T, Prinz B, Yu T-Y, Torrey J, Thomas J, Bobrowicz P, Vásquez M. Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool. Protein Eng Des Sel. 2013;26:663–70.
25. Kelly RL, Le D, Zhao J, Wittrup KD. Reduction of nonspecificity motifs in synthetic antibody libraries. J Mol Biol. 2018;430:119–30.
26. Harvey EP, Shin JE, Skiba MA, Nemeth GR, Hurley JD, Wellner A, Shaw AY, Miranda VG, Min JK, Liu CC. An in silico method to assess antibody fragment polyreactivity. bioRxiv. 2022.
27. Boughter CT, Borowska MT, Guthmiller JJ, Bendelac A, Wilson PC, Roux B, Adams EJ. Biochemical patterns of antibody polyreactivity revealed through a bioinformatics-based analysis of CDR loops. Elife. 2020;9: e61393.
28. Rabia LA, Zhang Y, Ludwig SD, Julian MC, Tessier PM. Net charge of antibody complementarity-determining regions is a key predictor of specificity. Protein Eng Des Sel. 2018;31:409–18.
29. Kelly RL, Zhao J, Le D, Wittrup KD. Nonspecificity in a nonimmune human scFv repertoire. MAbs. 2017;9:1029–35.
30. Lecerf M, Kanyavuz A, Lacroix-Desmazes S, Dimitrov JD. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. Mol Immunol. 2019;112:338–46.
31. Lim H, Jeon H-N, Lim S, Jang Y, Kim T, Cho H, Pan J-G, No KT. Evaluation of protein descriptors in computer-aided rational protein engineering tasks and its application in property prediction in SARS-CoV-2 spike glycoprotein. Comput Struct Biotechnol J. 2022. https://doi.org/10.1016/j.csbj.2022.01.027.
32. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, Abbeel P, Song Y. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689–701.
33. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16:1315–22.
34. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci. 2021;118: e2016239118.
35. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. Adv Neural Inf Process Syst. 2021;34:29287–303.
36. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373:871–6.

37. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
38. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci. 2020;117:1496–503.
39. Evans R, O'Neill M, Pritzel A, Antropova N, Senior AW, Green T, Žídek A, Bates R, Blackwell S, Yim J. Protein complex prediction with AlphaFold-Multimer. BioRxiv. 2021.
40. Wang CY, Chang PM, Ary ML, Allen BD, Chica RA, Mayo SL, Olafson BD. ProtaBank: a repository for protein design and engineering data. Protein Sci. 2018;27:1113–24.
41. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
42. Olsson MH, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions. J Chem Theory Comput. 2011;7:525–37.
43. Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J Chem Theory Comput. 2016;12:281–96.
44. M Redington J, Breydo L, N Uversky V. When good goes awry: the aggregation of protein therapeutics. Protein Pept Lett. 2017;24:340–7.
45. Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev. 2008;37:1395–401.
46. Fleming PJ, Fleming KG. HullRad: fast calculations of folded and disordered protein and nucleic acid hydrodynamic properties. Biophys J. 2018;114:856–69.
47. Li L, Chen S, Miao Z, Liu Y, Liu X, Xiao ZX, Cao Y. AbRSA: a robust tool for antibody numbering. Protein Sci. 2019;28:1524–31.
48. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol. 1987;196:901–17.
49. Tabb DL, McDonald WH, Yates JR. DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res. 2002;1:21–6.
50. Audain E, Ramos Y, Hermjakob H, Flower DR, Perez-Riverol Y. Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. Bioinformatics. 2016;32:821–7.
51. Kingsbury JS, Saini A, Auclair SM, Fu L, Lantz MM, Halloran KT, Calero-Rubio C, Schwenger W, Airiau CY, Zhang J. A single molecular descriptor to predict solution behavior of therapeutic antibodies. Sci Adv. 2020;6:eabb0372.
52. Kozlowski LP, IPC 2.0. Prediction of isoelectric point and p K a dissociation constants. Nucleic Acids Res. 2021;49:285–92.
53. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232.
54. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. LightGBM: A highly efficient gradient boosting decision tree. In: Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3147–3155.
55. Brownlee, J., XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn. Machine Learning Mastery: 2016.
56. Setny P, Baron R, Michael Kekenes-Huskey P, McCammon JA, Dzubiella J. Solvent fluctuations in hydrophobic cavity–ligand binding kinetics. Proc Natl Acad Sci. 2013;110:1197–202.
57. Harding S. Protein hydrodynamics. Protein: a comprehensive treatise. In: Allen G, editor. Greenwich: Jai Press, Incorporated; 1997. p. 271–305.
58. Perkins SJ. X-ray and neutron scattering analyses of hydration shells: a molecular interpretation based on sequence predictions and modelling fits. Biophys Chem. 2001;93:129–39.
59. Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. arXiv preprint arXiv:2007.06225. 2020.
60. Liang T, Jiang C, Yuan J, Othman Y, Xie XQ, Feng Z. Differential performance of RoseTTAFold in antibody modeling. Brief Bioinform. 2022;23(5):bbac152. https://doi.org/10.1093/bib/bbac152.
61. Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics. 2020;36:3350–6.
62. Hasan MM, Alam MA, Shoombuatong W, Deng H-W, Manavalan B, Kurata H. NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning. Brief Bioinform. 2021;22:bbab167.
63. Xiong Y, Wang Q, Yang J, Zhu X, Wei D-Q. PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. Front Microbiol. 2018;9:2571.
64. Chen X, Xiong Y, Liu Y, Chen Y, Bi S, Zhu X. m5CPred-SVM: a novel method for predicting m5C sites of RNA. BMC Bioinformatics. 2020;21:1–21.
65. Arik SÖ, Pfister T. Tabnet: attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2021; 35:6679–6687.

Publisher's Note