Computational and Structural Biotechnology Journal 20 (2022) 788–798

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Evaluation of protein descriptors in computer-aided rational protein engineering tasks and its application in property prediction in SARS-CoV-2 spike glycoprotein

Hocheol Lim [a,b], Hyeon-Nae Jeon [b], Seungcheol Lim [c], Yuil Jang [a], Taehee Kim [a], Hyein Cho [a], Jae-Gu Pan [d], Kyoung Tai No [a,b,*]

[a] The Interdisciplinary Graduate Program in Integrative Biotechnology and Translational Medicine, Yonsei University, Incheon, Republic of Korea
[b] Department of Biotechnology, Yonsei University, Seoul, Republic of Korea
[c] Department of Safety Engineering, Chungbuk National University, Cheongju, Republic of Korea
[d] Infectious Disease Research Center (Superbacteria Group), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon, Republic of Korea

ARTICLE INFO

Article history:
Received 13 August 2021
Received in revised form 18 January 2022
Accepted 27 January 2022
Available online 31 January 2022

Keywords:
Quantum mechanics
Fragment molecular orbitals
Protein engineering
Machine learning
Protein descriptor

ABSTRACT

The importance of protein engineering in the research and development of biopharmaceuticals and biomaterials has increased. Machine learning in computer-aided protein engineering can markedly reduce the experimental effort in identifying optimal sequences that satisfy the desired properties from a large number of possible protein sequences. To develop general protein descriptors for computer-aided protein engineering tasks, we devised new protein descriptors, one sequence-based descriptor (PCgrades), and three structure-based descriptors (PCspairs, 3D-SPIEs_5.4 Å, and 3D-SPIEs_8Å). While the PCgrades and PCspairs include general and statistical information in physicochemical properties in single and pairwise amino acids respectively, the 3D-SPIEs include specific and quantum–mechanical information with parameterized quantum mechanical calculations (FMO2-DFTB3/D/PCM). To evaluate the protein descriptors, we made prediction models with the new descriptors and previously developed descriptors for diverse protein datasets including protein expression and binding affinity change in SARS-CoV-2 spike glycoprotein. As a result, the newly devised descriptors showed a good performance in diverse datasets, in which the PCspairs showed the best performance ($R^2 = 0.783$ for protein expression and $R^2 = 0.711$ for binding affinity). As a result, the newly devised descriptors showed a good performance in diverse datasets, in which the PCspairs showed the best performance. Similar approaches with those descriptors would be promising and useful if the prediction models are trained with sufficient quantitative experimental data from high-throughput assays for industrial enzymes or protein drugs.

1. Introduction

Protein engineering is a progressive process to design or develop proteins with properties valuable for scientific, industrial, or medical applications [1]. Protein amino acid sequences determine protein properties and functions, including expression level and catalytic activity [1]. Protein engineering involves a premeditated process to investigate the relationship between the amino acid sequence and protein function and to identify amino acid sequences with improved function. As a powerful protein engineering technique, directed evolution has been successful in producing enzymes and binding proteins by emulating the natural evolution process in the laboratory [2]. Directed evolution leads to an accumulation of beneficial mutations through an iterative process that is coupled to sequence diversification methods and selection strategies [1,2]. However, this approach is time-consuming and resource-intensive due to multiple high-throughput iterations [2]. In directed evolution, it is difficult to learn lessons from failure because valuable information regarding unimproved sequences is discarded [1].

Machine learning can utilize the information of unimproved sequences to differentiate protein properties. Prediction models with machine learning can speed up the evolution and optimization of protein properties by evaluating and selecting new variants to screen [1]. The models can guide the design of future experi-

* Corresponding author at: The Interdisciplinary Graduate Program in Integrative Biotechnology and Translational Medicine, Yonsei University, Incheon, Republic of Korea.
E-mail address: ktno@yonsei.ac.kr (K.T. No).

https://doi.org/10.1016/j.csbj.2022.01.027

mental rounds to escape local optima by learning efficiently and synthesizing the most promising variants [1,3]. Even in the cases where the underlying biophysical mechanisms are not well explained, machine learning models can be applied and be powerfully predictive [1]. Machine learning models in protein engineering require descriptors, also known as features, that are suitable for obtaining information in proteins [3].

Many protein descriptors have been developed and applied to diverse protein engineering tasks [3,4]. The descriptors are generally based on mutation indicators, protein sequences, and protein structures. Mutation indicator (MutInd) is a binary vector of elements 0 or 1, indicating whether specific mutations in sequence exist or not [3]. The MutInd can directly utilize experimental values, but it is too simple to explain all protein functions and has a limitation on extrapolation for new single mutations.

Attempts to utilize protein sequences have been made because they may possess valuable information regarding protein expression, binding affinity, stability, and other properties. In a bottom-up approach, many single amino acid property descriptors were developed and most of them are listed in the AA-index database [5]. In a top-down approach, attempts were made to study representations from raw sequences. Some examples of this approach include the word embedding model 'doc2vec' and natural language processing-based continuous vector representation 'BioVec' [6,7]. Recently, a statistical unified representation (UniRep) was developed to summarize protein sequences not equal in length to fixed-length vectors via recurrent neural network methods [4,8,9]. UniRep may connote fundamental features of protein sequences because clustering with UniRep can distinguish biophysical single amino acid properties, secondary structural helix-sheet properties, and evolutional proteome properties [4]. However, protein sequence descriptors have a sequence-function gap because protein sequences must be translated to the accurately folded protein structures for protein functions.

Protein structures form the basis for the structure–activity relationship (SAR) and the analysis of SAR enables a prediction for protein activity of new mutated proteins in protein engineering. Protein structures can be generally made use of topological and biophysical descriptors. As a popular topological descriptor in protein structure, the distance matrix between amino acids in a protein structure can be used to extract the spatial arrangement in a protein structure. A structure-based descriptor derived from amino acid pairwise contact potentials (sPairs) used the distance matrix to filter the amino acid pairs and employed the AA-index amino acid pairwise contact potential descriptors [3]. Recently, graph convolutional networks (GCN) are developed to generalize convolutional operations on the graph-like molecular representation of protein structures [10]. DeepFRI is a GCN-based model for predicting protein functions by leveraging sequence features extracted from a protein language model and protein structures [11].

In biophysical descriptors, it is important to simulate molecular phenomena and accurately describe their physical, chemical, and biological properties. Energy calculations in molecular simulation have been used to predict the properties of biomolecules especially in the field of computer-aided drug discovery. Although quantum mechanics (QM) based molecular orbital calculation methods provide an accurate description of molecular phenomena, QM methods require huge computational costs and could not be easily applied to large biological systems [12]. Fragment molecular orbitals (FMO) method was developed for QM calculations of large molecular systems in 1999 [13]. The FMO method dramatically reduced computational cost without compromising the accuracy compared to the traditional QM method, which has been successfully applied to the protein–ligand interactions and protein–protein interactions [12,14–17]. Compared to traditional QM methods, the FMO method provides inter-fragment interaction energies, the map of which contains secondary structural and stability information in protein structure [18]. A 3-dimensional scattered pair interaction energies (3D-SPIEs) method extracts significant pair interactions in the map, which can be used to find a hot spot region in the protein–protein interface of the spike glycoprotein from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [12,14].

SARS-CoV-2 has been categorized as a human pathogen that caused the global pandemic of coronavirus disease that began in December 2019 [19,20]. The receptor binding domain (RBD) of the spike glycoprotein in SARS-CoV-2 binds to human angiotensin-converting enzyme 2 (hACE2), allowing the viral membrane of SARS-CoV-2 to fuse with the host cell membrane [21]. Analysis of the quantitative deep mutational scanning data obtained using yeast-surface display methods showed how RBD amino acid mutations affect the binding affinity with hACE2 and protein expression of RBD [22]. These experimental validations for single amino acid variants of RBD are valuable data for assessing whether the viral mutation is likely to be deleterious. The ongoing evolution of SARS-CoV-2 variants has provided critical insights for preparing for and preventing future outbreaks [23]. Accurately predicting the effects of amino acid changes on the ability of RBD, such as protein expression and binding affinity, can help assess the implications for public health in the ongoing evolution of SARS-CoV-2.

In this work, we devised one sequence-based (PCgrades) and three structure-based protein descriptors (PCspairs, 3D-SPIEs_5.4 Å, and 3D-SPIEs_8Å). And then we applied the descriptors to make prediction models for seven datasets. To evaluate the protein descriptors, we compared the performance of the prediction models trained with the newly devised descriptors and the known three protein descriptors (PCscores, sPairs, and UniRep fusion). The PCscores and sPairs were the top-ranked descriptors in the comparison of protein descriptors in diverse datasets by Xu et al [3]. The UniRep fusion is a newly devised descriptor with natural language processing methods and has the potential for diverse protein engineering tasks [4]. Because the PCscores, sPairs, and UniRep fusion can be applied to general property prediction in diverse datasets, we made baselines with the PCscores, sPairs, and UniRep fusion to evaluate the newly devised general descriptors in this work.

2. Methods

2.1. Data sets

The seven datasets in this work are summarized in Table 1. Reference protein sequences in all datasets were obtained from UniProt [24]. In the absorption wavelength shift dataset, *Gloeobacter violaceus* bacteriorhodopsin (GR) was used (UniProt ID: Q7NP59) and the dataset was from Engqvist et al [25]. In the enantiomeric selectivity dataset, *Aspergillus niger* epoxide hydrolase (ANEH) was used (UniProt ID: Q9UR30) and the dataset was from Gumulya et al and Reetz et al [26–28]. In the enantiomeric excess dataset, *Rhodothermus marinus* nitric oxide dioxygenase (RmaNOD) was used (UniProt ID: D0MGT2) and the dataset was from Wu et al and Wittmann et al [29,30]. In binding affinity and protein expression data sets, the spike glycoprotein of SARS-CoV-2 was used (UniProt ID: P0DTC2) and the datasets were from Starr et al [22]. In all datasets, only the substitution mutations were selected. The insertion-deletion mutations were removed.

2.2. Protein single amino acid property descriptor (PCscores and PCgrades)

Single amino acid property descriptors infer various physicochemical and biochemical properties of single amino acids [5].

Table 1
Summary of data sets in this work.

Protein Name	Species	Abbreviation	Observable variable	All set	Training set	Test set
Bacteriorhodopsin	*Gloeobacter violaceus*	GR-wave	Max absorption wavelength	71	56	15
		GR-shift	Max absorption wavelength shift			
Epoxide hydrolase	*Aspergillus niger*	ANEH-evalue	Enantiomeric selectivity (e-value)	163	130	33
		ANEH-ddG	Enantiomeric selectivity (ddG‡)			
Nitric oxide dioxygenase	*Rhodothermus marinus*	RmaNOD-ee	Enantiomeric excess	552	441	111
Spike glycoprotein	SARS-CoV-2	SARS2-expr	Protein expression	3799	3039	760
		SARS2-bind	Binding affinity with hACE2	3803	3042	761

AA-index is a large database of numerical indices of single amino acids and pairs [5]. The zScales and VHSE are based on principal component analysis to represent the scalar values of amino acids. While the zScales has 5 rows and it is based on the amino acid position information (size, hydrophobicity, charge, and so on.) [31], the VHSE has 50 rows and is vectors of hydrophobic, steric, and electronic properties [32]. The PCscores and PCgrades were used as a representative of protein single amino acid property descriptors (Fig. 1). The min–max scaling was introduced in principal component analysis. While the PCscores is the first 11 principal components of the 533 sets of single amino acid descriptors in the AA-index by Xu et al [3], we used the 553 sets in the AA-index to make the PCscores in this work. The PCgrades is firstly introduced in this work and is the first 11 principal components of the 606 sets of single amino acid descriptors in the AA-index, VHSE, and zScales [31,32]. The explained variance ratios of principal components in the PCscores and PCgrades are shown in Fig. S1, in which the first 11 principal components of PCscores accounted for about 91.3% of variances and the 11 components of PCgrades accounted for about 91.5% of variances.

2.3. Statistical unified representation descriptor (UniRep)

UniRep is based on a 1900-unit multiplicative long-/short-term memory recurrent neural network (mLSTM/RNN) model and was trained with approximately 24 million protein sequences (UniRef50) by Alley et al [4]. We used globally pre-trained weights with UniRef50 and calculated all descriptors with the performant reimplementation of UniRep in Jax [33]. The mLSTM/RNN-1900-unit model in UniRep provided all three representations of protein sequence: average hidden state, final hidden state, and the last internal cell state of the single 1900-dimensional layer. Alley et al devised the UniRep fusion by concatenating the three representations and they made prediction models with the UniRep fusion [4]. We also concatenated the three representations to obtain the UniRep fusion in this work (Fig. S2).

2.4. Protein structure preparation

All experimental protein structures were collected from the Protein Data Bank (PDB) [34]. In the GR dataset, we used chain A of a wild-type GR (PDB ID: 6NWD), the resolution of which is 2.00 Å [35]. In the ANEH dataset, we used chain A of a wild-type ANEH (PDB ID: 1QO7), the resolution of which is 1.80 Å [36]. In the RmaNOD dataset, we used chain A of a wild-type RmaNOD (6WK3), the resolution of which is 2.45 Å [30]. In the SARS-CoV-2 expression dataset, we used an unbound form of RBD in chain A of SARS-CoV-2 glycoprotein (PDB ID: 6ZGE), the resolution of which is 2.60 Å [37]. In the SARS-CoV-2 binding affinity dataset, we used a complex form of RBD in chain A of SARS-CoV-2 glycoprotein with chain B of hACE2 (PDB ID: 7KMS), the resolution of which is 3.64 Å [38]. All the missing side chains of the protein structures were filled with Prime implemented in the Schrödinger program [39]. Hydrogen atoms were added to the protein structures at pH 7.0 and their positions were optimized with the PROPKA implemented in the Schrödinger program [40]. The restrained energy minimization was performed with OPLS3 in the Schrödinger program within 0.3 Å root-mean-squared deviation [41]. Each mutant structure was generated with residue scanning, in which the mutated side-chain rotamers were searched for all mobile residues with Prime implemented in the Schrödinger program [42]. The residue scanning method is to predict the structures of residues in mutants with homology modeling, which adjusts the side-chain rotamers for repacking and minimizes the side-chain atoms. In this step, the backbone minimization of the mutated residues was not performed. Rather, the predicted mutant structures were re-prepared with the same protocol, and all hydrogen atoms are removed and re-added at pH 7.0. In generating all structure-based descriptors, one data point shares one protein structure in seven datasets.

2.5. Structure-based amino acid pairwise descriptors (sPairs and PCspairs)

AA-index database includes amino acid pairwise contact potentials for statistical analysis of protein sequences and protein structures [5]. The generation workflow for the sPairs and PCspairs is shown in Fig. 1. The sPairs is a structure-based descriptor employing the AA-index amino acid pairwise contact potential, which used statistical contact potential derived from 25 X-ray protein structures (TANS760101) [3]. While only a single 3D protein structure was used to make sPairs in Xu et al [3], we used each mutant structure to make sPairs for each mutant in this work. The PCspairs is firstly introduced in this work, and is the first principal component of the 135 sets of amino acid pairwise contact potentials (Fig. 1). The min–max scaling was introduced in principal component analysis. The explained variance ratios of the first principal components in each amino acid are shown in Fig. S3, in which each first principal component accounted for more than 50% of the variance in 135 sets. The 135 sets of amino acid potentials include the substitution matrices, contact potentials in X-ray structures, the transfer energy of amino acids from water to the protein environment, and potentials in the protein–protein interfacial regions [5]. The PCspairs can include effective information not only on contact potentials but also water-mediated and protein–protein interactions. In paired positions of two residues within 8 Å in the structure, the values were derived from the contact potential but otherwise were set to zero. The distance between two residues was measured with a single-linkage distance.

2.6. Quantum mechanical energy descriptors (3D-SPIEs_5.4 Å and 3D-SPIEs_8Å)

All FMO calculations were performed with the version of Feb 14, 2018 GAMESS [43], and with FMO2-DFTB3/D/PCM level. They include the two-body FMO method (FMO2), a self-consistent-charge density-functional tight-binding method derived via a third-order expansion (DFTB3) with the 3OB parameter set

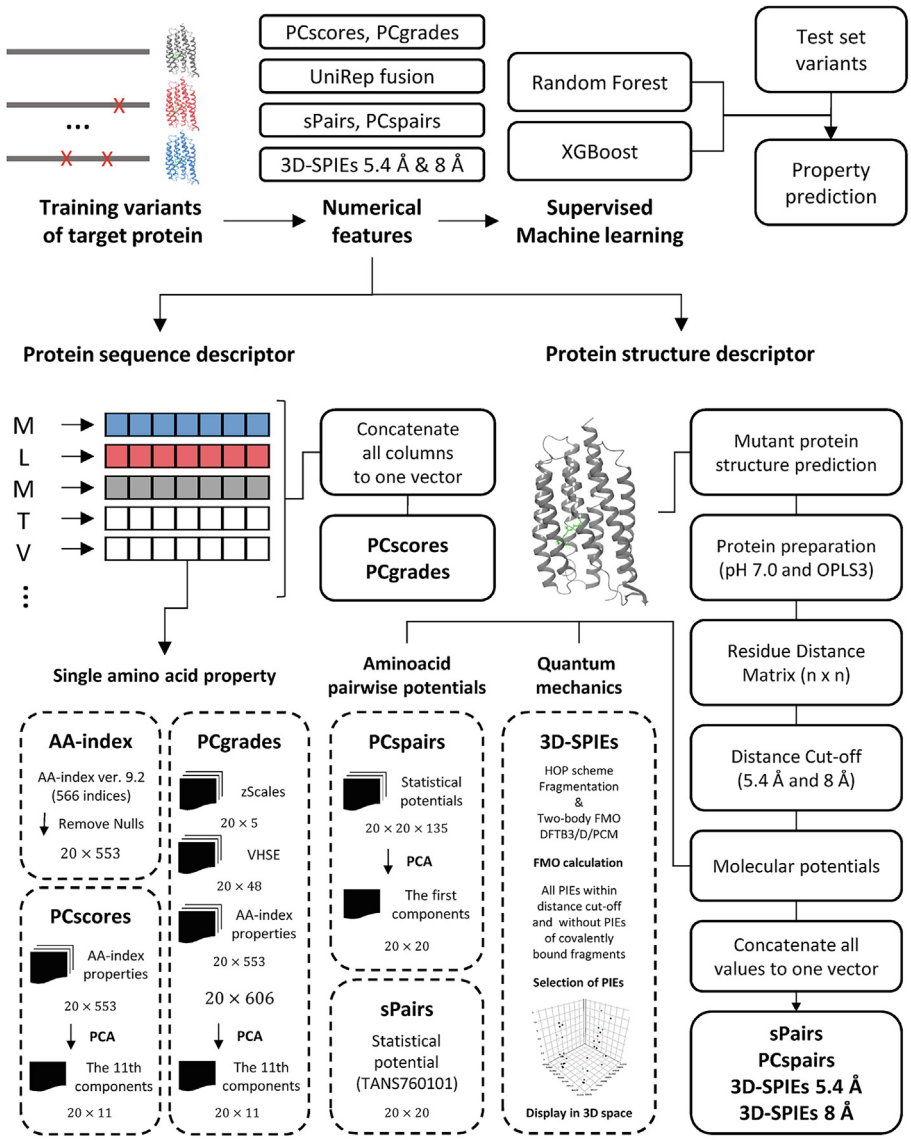

Fig. 1. Workflow of protein descriptor generation for computer-aided rational protein engineering tasks in this work.

[44,45], the UFF-type dispersion correction (D) [46,47], and implicit polarizable continuum model (PCM) [44]. The two-body FMO calculation consists of four steps with fragmentation, fragment self-consistent field calculation, fragment-pair self-consistent field calculation, and total property evaluation [12]. Firstly, all input files were prepared in compliance with the hybrid orbital projection (HOP) scheme fragmentation [48]. In the HOP scheme, each residue was defined as one fragment, and two cysteine residues forming the disulfide bridge were defined as one fragment. In the GR, the retinal and its covalently bound lysine were defined as one fragment. In RmaNOD, we removed the heme group and the heme-coordinated residues were terminated with hydrogen atoms because the 3OB parameter does not support the 'Fe' atom type. Secondly and thirdly, all the molecular orbitals in fragment and fragment-pair are optimized by self-consistent field theory in the whole electrostatic field from the other fragments. The difference between the second and third steps is just the size of the fragment, where the fragment pair is the combination of two fragments. Fourthly, all results from the second and third steps are pieced together to generate the whole picture of the system. In this step, FMO provides the pair interaction energies between two fragments. All 3D-SPIEs results were generated with a similar protocol in previous studies (Fig. 1) [12,14]. In this work, we did not apply the energy cut-off criteria and used all values within the specific distance (5.4 Å and 8.0 Å). The distance between two fragments was measured with a single-linkage distance.

2.7. Performance metrics and Train-test splits

Predictions on test sets were evaluated using R^2, RMSE, MAE, and Spearman's rank correlation. R^2 is a square of a measure of linear correlation between predicted and observed values. RMSE is a root-mean-square-error and a measure of the difference between the predicted and observed values. MAE is a mean absolute error and a measure of the absolute errors between the predicted and observed values. Spearman's rank correlation is a nonparametric measure of rank correlation. The Scikit-learn package was used to calculate R^2, RMSE, and MAE [49], and the SciPy package in Python was used to calculate Spearman's rank correlation coefficients (SCC) [50].

For all supervised tasks in this study, we prepared a split with 80% training and 20% test sets in Python using the Scikit-learn

package with a fixed random seed [49]. In training, we used 10-fold cross-validation with GridSearchCV in the Scikit-learn package [49]. To improve machine learning, we removed the columns with all same values in each descriptor feature and introduced a min–max scaling by fitting the scaler with training data and applying the scaler on training and test data, respectively.

2.8. Machine learning algorithms and hyperparameter tuning

Prediction models were constructed using the random forest (RF) regression model and the extreme gradient boosting (XGB) model. The RF regression model is an ensemble method based on many classifying decision trees, and it uses averaging to improve stability and accuracy by reducing variance and avoiding overfitting problems [3]. Decision trees use a flowchart-like tree structure and observe features in descriptors to provide a useful continuous output. The XGB is an ensemble learning method based on gradient tree boosting, which builds each tree sequentially and each tree fits the residuals of predictions of all previous trees [3].

The hyperparameter tuning procedure is summarized in Table S1. Some hyperparameters in RF were incorporated for tuning the models; the number of decision trees (n_estimators) and the number of features to consider when looking for the best split (max_features) [49]. Some hyperparameters in XGB were incorporated for tuning the models; the number of the gradient boosted trees (n_estimators), the maximum tree depth for based learners (max_depth), the boosting learning rate (learning_rate), the minimum loss reduction required to make a further partition on a leaf node of the tree (gamma), the minimum sum of instance weight needed in a child (min_child_weight), and the subsample ratio of the training instance (subsample) [51]. The 10-fold cross-validation was used for hyperparameter tuning of all machine learning algorithms. Grid-search was performed to find optimal values under each set of descriptors. The final model was selected with the best performance of R^2 in the validation set of cross-validation. The optimal hyperparameter values are summarized in Table S2.

3. Results

Predicting protein properties is important in computer-aided rational protein engineering tasks. To improve prediction performance, we devised new protein descriptors, one sequence-based descriptor (PCgrades) and three structure-based descriptors (PCspairs, 3D-SPIEs_5.4 Å, and 3D-SPIEs_8Å). The generation workflow of protein descriptors is shown in Fig. 1. To evaluate the new protein descriptors, we collected seven datasets and made prediction models from the combination of two machine learning models (RF and XGB) and seven protein descriptors (PCscores, PCgrades, sPairs, PCspairs, UniRep fusion, 3D-SPIEs_5.4 Å, and 3D-SPIEs_8Å).

All data sets in this study are summarized in Table 1 and illustrated in Fig. 2. Hyperparameter tuning and 10-fold cross-validation of all machine learning algorithms were performed with grid-search. The performance metrics of mean R^2 in training sets and 10-fold cross-validation test sets are summarized in Table S3 and Table 2. The performance metrics of R^2, RMSE, MAE, and Spearman's rank correlation in test sets are summarized in Table 3 and Tables S4–S6. The correlation plots from the test set predictions of the best model in each dataset are shown in Fig. 3.

3.1. Gloeobacter violaceus rhodopsin (GR)

GR is a valuable engineering target protein in light-harvesting to capture photons of solar light in the bioenergy production and bio-sensing industry [25]. The maximum absorption wavelength levels of various mutants of GR were obtained from the study by Engqvist et al [25]. In the GR dataset, we made prediction models separately for wavelength (GR-wave) and wavelength shift from wild-type (GR-shift).

In GR-wave, the prediction model from XGB and 3D-SPIEs_8Å showed the best performance in the test set prediction ($R^2 = 0.947$). The best model was trained with optimal hyperparameter (learning_rate = 0.01, max_depth = 10, and n_estimators = 1000). In the test set of the best model, the RMSE is 7.948, the MAE is 6.609, and the SCC is 0.950. The second-best model is the RF/PCgrades and XGB/3D-SPIEs_5.4 Å, the performance of which in the test set is $R^2 = 0.934$. The best prediction model by Xu et al showed $R^2 = 0.934$ with the VHSE and multilayer perceptron method [3].

In GF-shift, the prediction model from XGB and PCspairs showed the best performance in test prediction ($R^2 = 0.950$). The best model was trained with optimal hyperparameter (learning_rate = 0.05, max_depth = 15, and n_estimators = 500). In the test set of the best model, the RMSE is 10.554, the MAE is 8.972, and the SCC is 0.968. The second-best model is the RF/PCgrades, the performance of which in the test set is $R^2 = 0.934$.

3.2. Aspergillus niger epoxide hydrolase (ANEH)

Enantiomeric selectivity of an enzyme is the ability of an enzyme to selectively distinguish one enantiomer from its counter isomer in an enzymatic reaction. This property plays a critical role in fine chemical production and bioindustry. The enantiomeric selectivity levels of various ANEH mutants were obtained from the studies by Gumulya et al and Reetz et al [26–28]. ANEH has hydrolytic kinetic activity with glycidyl phenyl ether, and wild-type ANEH has selectivity in favor of the (S)-glycidyl phenyl ether ($\Delta\Delta G^{\ddagger} = -0.85$). In the ANEH dataset, we made prediction models separately for e-value (ANEH-evalue) and dd $G^{\ddagger}$ (ANEH-ddG).

In ANEH-evalue, the prediction model from RF and PCspairs showed the best performance in the test set prediction ($R^2 = 0.859$). The best model was trained with optimal hyperparameter (max_features = 'sqrt' and n_estimators = 500). In the test set of the best model, the RMSE is 12.862, the MAE is 8.628, and the SCC is 0.956. The second-best model is the RF/PCgrades, the performance of which in the test set is $R^2 = 0.848$. The best prediction model by Xu et al showed $R^2 = 0.754$ with sPairs and elastic-net regularized generalized linear model method [3].

In ANEH-ddG, the prediction model from RF and PCgrades showed the best performance in the test set prediction ($R^2 = 0.938$). The best model was trained with optimal hyperparameter (max_features = 'log2' and n_estimators = 1000). In the test set of the best model, the RMSE is 0.143, the MAE is 0.108, and the SCC is 0.935. The second-best model is the RF/PCscores, the performance of which in the test set is $R^2 = 0.935$.

3.3. Rhodothermus marinus nitric oxide dioxygenase (RmaNOD)

Enantiomeric excess is a measure of the purity of one enantiomer in a sample with mixed enantiomers. The enantiomeric excess levels of various mutants of RmaNOD were obtained from the study by Arnold et al [29,30]. RmaNOD catalyzes carbon-silicon bond formation between ethyl 2-diazopropanoate and phenyldimethyl silane and produces two possible product enantiomers of carbine Si-H insertion reaction [29]. In the RmaNOD dataset, we used enantiomeric excess as an objective variable (RmaNOD-ee).

In RmaNOD-ee, the prediction model from RF and PCscores showed the best performance in the test set prediction

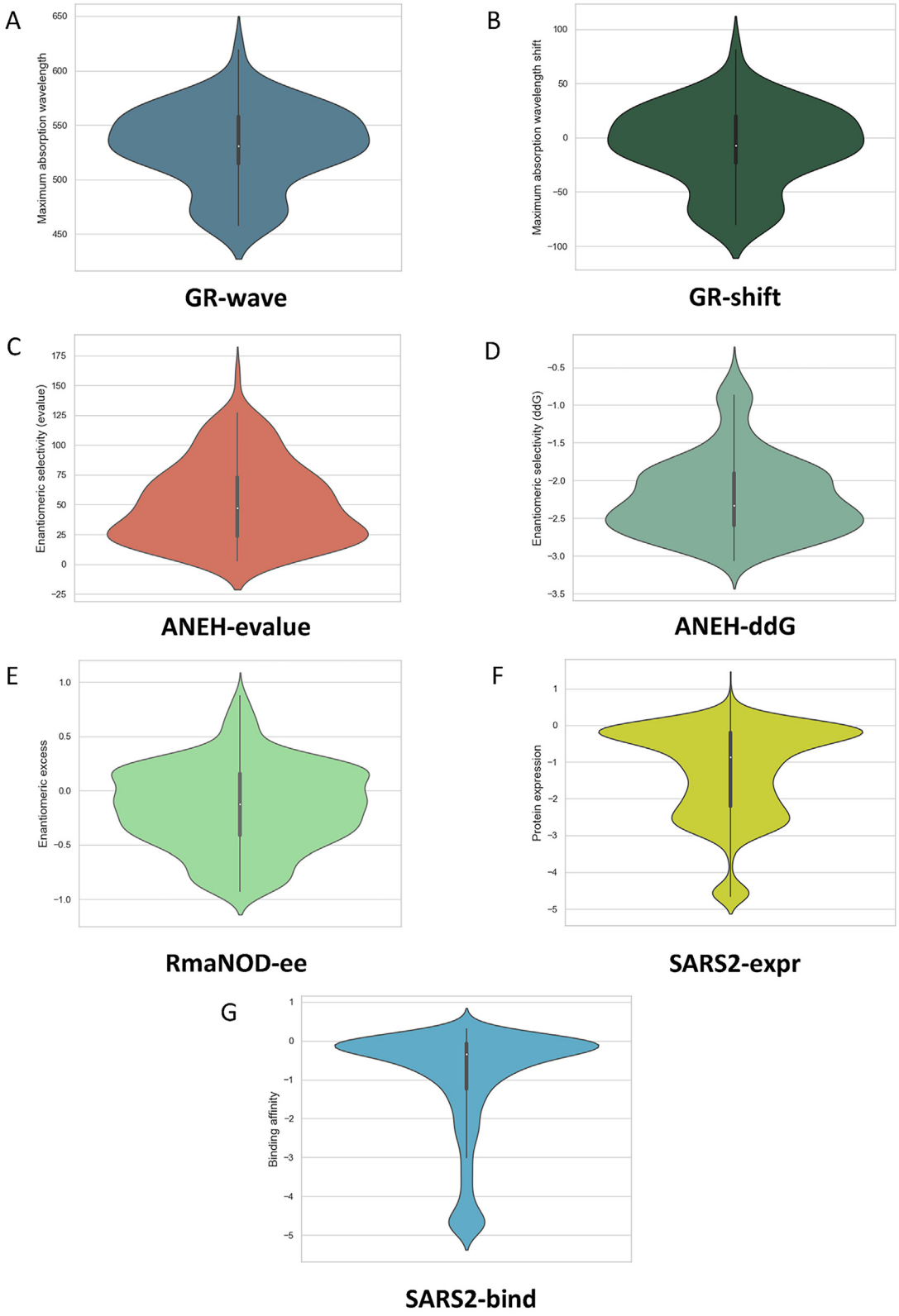

Fig. 2. Violin plots of each dataset in this work. (A) Maximum absorption wavelength of bacteriorhodopsin (GR-wave). (B) Maximum absorption wavelength shift of bacteriorhodopsin from wild-type (GR-shift). (C) Enantiomeric selectivity (e-value) of epoxide hydrolase (ANEH-evalue). (D) Enantiomeric selectivity (ddG) of epoxide hydrolase (ANEH-ddG). (E) Enantiomeric excess of nitric oxide dioxygenase (RmaNOD-ee). (F) Protein expression of spike glycoprotein of SARS-CoV-2 (SARS2-expr). (G) Binding affinity between spike glycoprotein of SARS-CoV-2 and human angiotensin converting enzyme 2 (SARS2-bind).

Table 2
Mean R-Squared of 10-fold Cross-validation sets Predictions.

Data set	method	PCscores	PCgrades	sPairs	PCspairs	UniRep fusion	3D-SPIEs_5.4 Å	3D-SPIEs_8Å
GR-wave	RF	0.822 ± 0.166	0.813 ± 0.150	0.754 ± 0.188	0.799 ± 0.196	0.677 ± 0.340	0.766 ± 0.140	0.761 ± 0.164
	XGB	0.821 ± 0.126	0.823 ± 0.125	0.756 ± 0.226	0.761 ± 0.208	0.739 ± 0.183	0.748 ± 0.166	0.718 ± 0.203
GR-shift	RF	0.822 ± 0.166	0.813 ± 0.151	0.755 ± 0.188	0.799 ± 0.196	0.677 ± 0.340	0.767 ± 0.140	0.761 ± 0.164
	XGB	0.757 ± 0.216	0.766 ± 0.220	0.743 ± 0.190	0.733 ± 0.227	0.772 ± 0.155	0.745 ± 0.171	0.729 ± 0.179
ANEH-evalue	RF	0.712 ± 0.161	0.713 ± 0.153	0.706 ± 0.159	0.706 ± 0.187	0.528 ± 0.202	0.555 ± 0.176	0.570 ± 0.119
	XGB	0.630 ± 0.349	0.632 ± 0.348	0.666 ± 0.139	0.693 ± 0.198	0.568 ± 0.258	0.555 ± 0.165	0.568 ± 0.138
ANEH-ddG	RF	0.802 ± 0.127	0.800 ± 0.129	0.818 ± 0.099	0.809 ± 0.112	0.706 ± 0.141	0.673 ± 0.115	0.685 ± 0.116
	XGB	0.763 ± 0.173	0.756 ± 0.180	0.751 ± 0.101	0.768 ± 0.141	0.701 ± 0.186	0.696 ± 0.133	0.701 ± 0.125
RmaNOD-ee	RF	0.707 ± 0.080	0.717 ± 0.068	0.706 ± 0.073	0.719 ± 0.074	0.641 ± 0.097	0.666 ± 0.080	0.678 ± 0.071
	XGB	0.696 ± 0.088	0.695 ± 0.087	0.676 ± 0.089	0.709 ± 0.078	0.659 ± 0.105	0.667 ± 0.086	0.691 ± 0.082
SARS2-expr	RF	0.724 ± 0.032	0.728 ± 0.031	0.736 ± 0.025	0.790 ± 0.025	0.517 ± 0.019	0.635 ± 0.034	0.649 ± 0.028
	XGB	0.705 ± 0.043	0.708 ± 0.039	0.718 ± 0.028	0.760 ± 0.035	0.614 ± 0.024	0.662 ± 0.032	0.673 ± 0.027
SARS2-bind	RF	0.701 ± 0.040	0.695 ± 0.048	0.695 ± 0.055	0.752 ± 0.045	0.484 ± 0.019	0.650 ± 0.033	0.660 ± 0.037
	XGB	0.686 ± 0.032	0.680 ± 0.027	0.696 ± 0.050	0.748 ± 0.045	0.602 ± 0.022	0.680 ± 0.032	0.689 ± 0.036

Table 3
R-Squared of Test sets Predictions of Best-found parameter models.

Data set	method	PCscores	PCgrades	sPairs	PCspairs	UniRep fusion	3D-SPIEs_5.4 Å	3D-SPIEs_8Å
GR-wave	RF	0.931	0.934	0.926	0.906	0.795	0.877	0.862
	XGB	0.892	0.896	0.894	0.928	0.834	0.934	0.947
GR-shift	RF	0.931	0.934	0.926	0.906	0.795	0.877	0.862
	XGB	0.901	0.892	0.922	0.950	0.849	0.915	0.921
ANEH-evalue	RF	0.844	0.848	0.831	0.859	0.685	0.685	0.660
	XGB	0.836	0.837	0.833	0.851	0.780	0.747	0.732
ANEH-ddG	RF	0.935	0.938	0.926	0.929	0.830	0.751	0.783
	XGB	0.923	0.929	0.915	0.923	0.886	0.845	0.839
RmaNOD-ee	RF	0.723	0.708	0.701	0.718	0.637	0.659	0.659
	XGB	0.691	0.693	0.706	0.702	0.637	0.675	0.675
SARS2-expr	RF	0.708	0.724	0.739	0.783	0.490	0.588	0.608
	XGB	0.690	0.712	0.689	0.743	0.606	0.607	0.630
SARS2-bind	RF	0.651	0.651	0.653	0.711	0.464	0.590	0.600
	XGB	0.648	0.671	0.638	0.702	0.576	0.629	0.628

($R^2 = 0.723$). The best model was trained with optimal hyperparameter (max_features = 'sqrt' and n_estimators = 1500). In the test set of the best model, the RMSE is 0.194, the MAE is 0.133, and the SCC is 0.838. The second-best model is the RF/PCspairs, the performance of which in the test set is $R^2 = 0.718$. The best prediction model by Xu et al showed $R^2 = 0.288$ with PCscores and XGB method [3].

3.4. Protein expression in the spike glycoprotein of SARS-CoV-2

Protein expression is affected by protein stability and solubility, which are the primary common causes of protein production failure. The mean protein expression levels of various mutants of RBD were obtained from the study by Starr et al [22]. In the SARS-CoV-2 protein expression dataset, we used mean protein expression levels as an objective variable (SARS2-expr).

In SARS2-expr, the prediction model from RF and PCspairs showed the best performance in the test set prediction ($R^2 = 0.783$). The best model was trained with optimal hyperparameter (max_features = 'log2′ and n_estimators = 1500). In the test set of the best model, the RMSE is 0.490, the MAE is 0.355, and the SCC is 0.891. The second-best model is the XGB/PCspairs, the performance of which in the test set is $R^2 = 0.743$. The general protein expression prediction model showed the prediction performance (R^2) between 0.504 and 0.698 [52].

3.5. Binding affinity between the spike glycoprotein of SARS-CoV-2 and hACE2

Binding affinity is a measure of the strength of the interaction between a protein and a ligand. The mean binding affinity levels of various mutants of RBD with hACE2 were obtained from the study by Starr et al [22]. In the SARS-CoV-2 binding affinity dataset, we used mean binding affinity levels as an objective variable (SARS2-bind).

In SARS2-bind, the prediction model from RF and PCspairs showed the best performance in test prediction ($R^2 = 0.711$). The best model was trained with optimal hyperparameter (max_features = 'sqrt' and n_estimators = 1500). In the test set prediction of the best model, the RMSE is 0.735, the MAE is 0.412, and the SCC is 0.873. The second-best model is the XGB/PCspairs, the performance of which in the test is $R^2 = 0.702$. The general binding affinity prediction models have the prediction performance (Pearson's linear correlation) between 0.61 and 0.76 [53–56], which can be converted to the prediction performance (R^2) between 0.3721 and 0.5776. The specific binding affinity prediction models for SARS-CoV-2 and hACE2 have the performance (Pearson's linear correlation) of 0.73 and 0.82, which can be converted to the prediction performance (R^2) of 0.5329 and 0.6724 [57,58].

3.6. Protein descriptor comparison in seven datasets.

To compare evaluation metrics, the performance of 98 final models (the 98 combinations from seven protein descriptors, seven datasets, and two machine learning models) was ranked according to the median value in increasing order for RMSE and decreasing order for R-squared in the test prediction (Fig. 4). The ranking results of RMSE and R-squared are almost identical. The PCspairs won both in RMSE and R-squared ranking, followed by PCgrades, PCscores, sPairs, 3D-SPIEs_8Å, 3D-SPIEs_5.4 Å, and UniRep fusion. In the model ranking, the combination of XGB and PCspairs won in all combinations, followed by RF/PCspairs, RF/PCgrades, RF/PCscores, and XGB/PCgrades.

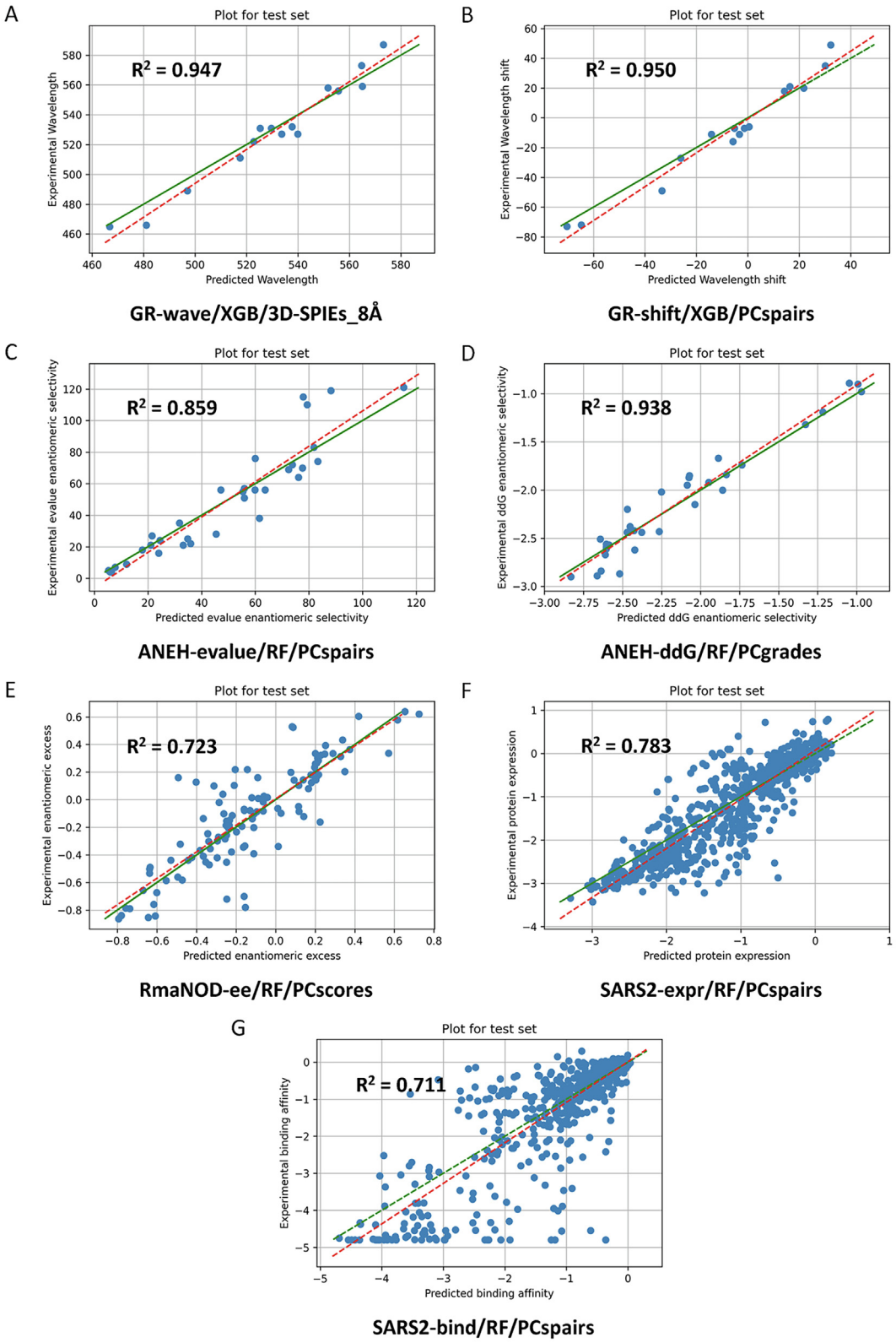

Fig. 3. The correlation plots from the test set prediction of the best models in seven datasets. (A) XGB model trained with the 3D-SPIEs_8Å in the GR-wave dataset. (B) XGB model trained with the PCspairs in the GR-shift dataset. (C) RF model trained with the PCspairs in the ANEH-evalue dataset. (D) RF model trained with the PCgrades in the ANEH-ddG dataset. (E) RF model trained with the PCscores in the RmaNOD dataset. (F) RF model trained with the PCspairs in the SARS2-expr dataset. (G) RF model trained with the PCspairs in the SARS2-bind dataset. The green dotted line indicates the identity function and the red dotted line indicates the trend line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

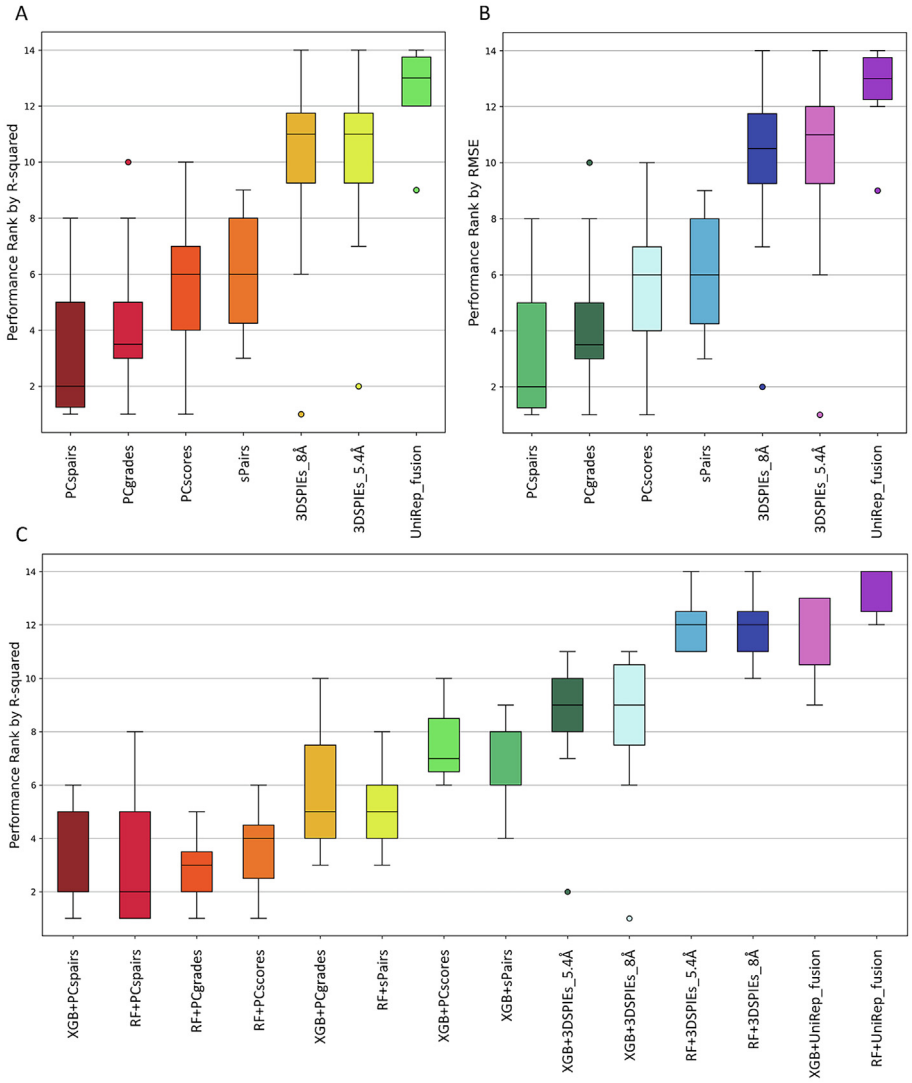

Fig. 4. The Box plot comparison rank. (A) Boxplot comparison rank of descriptors by R-squared metric in the test set, (B) Boxplot comparison rank of descriptors by RMSE metric in the test set, and (C) Boxplot comparison rank of models (machine learning method and descriptor combination) by R-squared in the test set.

4. Discussion

Given the successful application of machine learning methods in directed protein evolution, machine learning methods have been broadly applied in protein engineering [3,29]. The broader applications in protein engineering require sufficient data quantity, quality, and protein descriptors. Generating high quantity and quality data from high-throughput assays improves machine-learning-guided protein engineering in principle. But, most cases of protein engineering tasks have a data shortage in the desired properties of target proteins. For example, phage display technologies are superb and powerful for therapeutic and industrial projects but require tedious optimization and time-consuming test cycles. Therefore, many good protein descriptors are inevitably required in supervised machine learning with small data sets from the feedback between machine learning prediction models and experimental assays.

Protein descriptors are usually based on mutation indicators, protein sequences, and protein structures. As valuable information in databases including UniProt [24] and PDB [34] has increased, the necessity of utilizing protein sequences and unlabeled structures has increased. For example, the UniRep was developed from unlabeled protein sequences and distinguished physicochemical, secondary structural, and evolutionary information [4]. The UniRep is the mLSTM/RNN based on the statistical descriptor to extract fundamental features in protein sequences from a large unlabeled protein sequence database (UniRef50). Although the UniRep fusion underperformed compared to other protein descriptors in this work, the UniRep fusion has the potential in comparing protein sequences not equal in length. Although mutation indicators and protein sequence-based descriptors are powerful and effective tools for constructing prediction models to achieve diverse desired properties in protein engineering, protein structural descriptors are also promising because of the close relationship between structure and activity.

In this work, we devised one sequence-based protein descriptor (PCgrades) and three structural protein descriptors (PCspairs, 3D-SPIEs_5.4 Å, and 3D-SPIEs_8Å). To evaluate the newly devised protein descriptors, we made prediction models with the newly devised descriptors and the previously developed descriptors (PCscores, sPairs, and UniRep fusion). In the seven datasets, the PCspairs generally showed a better performance than other protein descriptors. The PCgrades and 3D-SPIEs showed the best performance in the ANEH-ddG and GR-wave datasets, respectively. The

PCgrades has more information on single amino acid descriptors than the PCscores, and PCspairs has more information on amino acid pairwise potential than the sPairs. Although the newly devised protein descriptors showed a good performance in the seven datasets, it still has a scope to improve the model performance for future studies. The combination of sequence and structure-based descriptors would be promising because they include valuable information from a different angle. For example, while sequence-based descriptors are easy to compare evolutional information, structure-based descriptors are easy to consider biophysical environments.

The 3D-SPIEs are based on quantum mechanical free energy calculations, and they can be improved with appropriate simulation systems, higher calculation levels, and more accurate homology models. The 3D-SPIEs only outperformed in GR-wave, because quantum mechanical simulations require more rigorous and detailed simulation systems. In GR-wave, we included a retinal molecule in FMO calculation, which improved the descriptor quality. Because molecular simulations mainly depend on appropriate simulation systems, it is important to construct specific simulation systems for the specific biological phenomena. Moreover, the molecular simulations also mainly depend on protein structures, so it is important to predict mutant protein structures accurately. As the homology modeling methods have been dramatically improved with deep learning methods [59,60], structure-based descriptors would be more accurate and powerful. New protein structure-based descriptors from rational protein analysis in the pharmaceutical industry would be promising in the future.

The newly devised protein descriptors (PCgrades, PCspairs, and 3D-SPIEs) contain different information in proteins. The PCgrades include the physicochemical property information of single amino acids in protein sequences and the PCspairs include the pairwise statistical potential information between two residues in protein structures. The two descriptors effectively compressed the information with principal component analysis and can be trained for the general and statistical properties of proteins. On the other hand, the 3D-SPIEs contain specific and mechanical information in protein structures. The 3D-SPIEs utilized quantum mechanical methods to quantify the interaction energies between two residues. Because the three descriptors can include different information in protein sequences and structures, they can be trained in response to diverse information from proteins. Therefore, similar approaches with those descriptors would be promising and useful for industrial enzymes and protein drugs.

In addition to the protein descriptors, various combinations with state-of-the-art machine learning algorithms and optimization of model architectures in deep learning would also improve the model performance [3]. Taken together, it has permitted the development of more accurate and powerful prediction models, which in turn would enable the computational exploration of enormous sequence space and suggestions for better variants in therapeutic or industrial research and development. In this work, we developed the prediction models for seven diverse datasets. The prediction models for the GR, ANEH, and RmaNOD would be applied to find optimal sequences satisfying the desired properties from many possible variants. Our prediction models in the three datasets outperformed the top-ranked models by Xu et al [3]. On the other hand, the prediction models for the protein expression and binding affinity of SARS-CoV-2 would be used to predict host adaption of SARS-CoV-2 variants with higher protein expression and binding affinity in the ongoing evolution of SARS-CoV-2. Our prediction models in the two datasets outperformed the general-purpose prediction models in protein expression [52] and binding affinity [53–56] and outperformed the specific prediction models for binding affinity in SARS-CoV-2 [57,58].

5. Conclusion

Protein engineering is a progressive process to find proteins with valuable properties in a tremendous possible protein sequence space. Machine-learning-guided protein engineering can speed up the identification of variants with optimal properties and has been expanded the predictions for diverse properties in many proteins. In this work, we developed one protein sequence-based and three structure-based descriptors and applied them to diverse protein engineering tasks. Similar approaches with those descriptors would be promising and useful if the prediction models are trained with sufficient quantitative experimental data from high-throughput assays for industrial enzymes or protein drugs.

CRediT authorship contribution statement

Hocheol Lim: Conceptualization, Methodology, Validation. **Hyeon-Nae Jeon:** Methodology. **Seungcheol Lim:** Conceptualization, Methodology. **Yuil Jang:** Investigation. **Taehee Kim:** Investigation. **Hyein Cho:** Writing - review & editing. **Jae-Gu Pan:** Conceptualization and Supervision. **Kyoung Tai No:** Conceptualization and Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.01.027.

References

[1] Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nat Methods 2019;16(8):687–94.
[2] Chowdhury R, Maranas CD. From directed evolution to computational enzyme engineering—a review. AIChE J 2020;66(3). https://doi.org/10.1002/aic.v66.310.1002/aic.16847.
[3] Xu Y et al. Deep dive into machine learning models for protein engineering. J Chem Inf Model 2020;60(6):2773–90.
[4] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019;16(12):1315–22.
[5] Kawashima S et al. AAindex: amino acid index database, progress report 2008. Nucl Acids Res 2007;36:D202–5.
[6] Le Q, Mikolov T. in International conference on machine learning. 1188-119 (PMLR).
[7] Asgari E, Mofrad MRK, Kobeissy FH. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE 2015;10 (11):e0141287.
[8] Favor A, Jayapurna I. Evaluating eUniRep and other protein feature representations for in silico directed evolution. Authorea Preprints 2020.
[9] Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-N protein engineering with data-efficient deep learning. Nat Methods 2021;18 (4):389–96. https://doi.org/10.1038/s41592-021-01100-y.
[10] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016).
[11] Gligorijević V et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 2021;12(1). https://doi.org/10.1038/s41467-021-23303-9.

[12] Lim H et al. Investigation of protein-protein interactions and hot spot region between PD-1 and PD-L1 by fragment molecular orbital method. Sci Rep 2019;9(1). https://doi.org/10.1038/s41598-019-53216-z.
[13] Kitaura K, Ikeo E, Asada T, Nakano T, Uebayasi M. Fragment molecular orbital method: an approximate computational method for large molecules. Chem Phys Lett 1999;313(3-4):701–6.
[14] Lim H et al. Hot spot profiles of SARS-CoV-2 and human ACE2 receptor protein protein interaction obtained by density functional tight binding fragment molecular orbital method. Sci Rep 2020;10(1). https://doi.org/10.1038/s41598-020-73820-8.
[15] Lim H et al. Investigation of hot spot region in XIAP inhibitor binding site by fragment molecular orbital method. Comput Struct Biotechnol J 2019;17:1217–25.
[16] Fedorov DG, Nagata T, Kitaura K. Exploring chemistry with the fragment molecular orbital method. PCCP 2012;14(21):7562. https://doi.org/10.1039/c2cp23784a.
[17] Tanaka S, Mochizuki Y, Komeiji Y, Okiyama Y, Fukuzawa K. Electron-correlated fragment-molecular-orbital calculations for biomolecular and nano systems. PCCP 2014;16(22):10310–44.
[18] Kurisaki I et al. Visualization analysis of inter-fragment interaction energies of CRP–cAMP–DNA complex based on the fragment molecular orbital method. Biophys Chem 2007;130(1-2):1–9.
[19] Chan J-W et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. The Lancet 2020;395(10223):514–23.
[20] Huang C et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 2020;395(10223):497–506.
[21] Wrapp D et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367(6483):1260–3.
[22] Starr TN et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 2020;182(5):1295–1310.e20.
[23] Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. Exp Mol Med 2021;53(4):537–47.
[24] Consortium U. UniProt: a hub for protein information. Nucl Acids Res 2015;43:D204–12.
[25] Engqvist MKM et al. Directed evolution of Gloeobacter violaceus rhodopsin spectral properties. J Mol Biol 2015;427(1):205–20.
[26] Gumulya Y, Sanchis J, Reetz MT. Many pathways in laboratory evolution can lead to improved enzymes: how to escape from local minima. ChemBioChem 2012;13(7):1060–6.
[27] Reetz MT et al. Directed evolution of an enantioselective epoxide hydrolase: uncovering the source of enantioselectivity at each evolutionary stage. J Am Chem Soc 2009;131(21):7334–43.
[28] Reetz MT, Sanchis J. Constructing and analyzing the fitness landscape of an experimental evolutionary process. ChemBioChem 2008;9(14):2260–7.
[29] Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad Sci 2019;116(18):8852–8.
[30] Wittmann BJ et al. Diversity-oriented enzymatic synthesis of cyclopropane building blocks. ACS Catal 2020;10(13):7112–6.
[31] Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 1998;41(14):2481–91.
[32] Mei Hu, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Peptide Science: Original Research on Biomolecules 2005;80(6):775–86.
[33] Ma E, Kummer A. Reimplementing Unirep in JAX. bioRxiv (2020).
[34] Berman HM et al. The protein data bank. Nucl Acids Res 2000;28:235–42.
[35] Morizumi T et al. X-ray crystallographic structure and oligomerization of Gloeobacter rhodopsin. Sci Rep 2019;9(1). https://doi.org/10.1038/s41598-019-47445-5.
[36] Zou J et al. Structure of Aspergillus niger epoxide hydrolase at 1.8 Å resolution: implications for the structure and function of the mammalian microsomal class of epoxide hydrolases. Structure 2000;8(2):111–22.
[37] Wrobel AG et al. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol 2020;27(8):763–7.
[38] Zhou T et al. Cryo-EM structures of SARS-CoV-2 spike without and with ACE2 reveal a pH-dependent switch to mediate endosomal positioning of receptor-binding domains. Cell Host Microbe 2020;28(6):867–879.e5.
[39] Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 2002;320(3):597–608.
[40] Olsson MHM, 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(2):525–37.
[41] Harder E et al. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J Chem Theory Comput 2016;12(1):281–96.
[42] Beard H, Cholleti A, Pearlman D, Sherman W, Loving KA. Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. PLoS ONE 2013;8(12):e82849.
[43] Alexeev Y, P Mazanetz M, Ichihara O, G Fedorov D. GAMESS As a Free Quantum-Mechanical Platform for Drug Research. Curr Top Med Chem 2012;12(18):2013–33.
[44] Nishimoto Y, Fedorov DG. The fragment molecular orbital method combined with density-functional tight-binding and the polarizable continuum model. PCCP 2016;18(32):22047–61.
[45] Gaus M, Lu X, Elstner M, Cui Q. Parameterization of DFTB3/3OB for sulfur and phosphorus for chemical and biological applications. J Chem Theory Comput 2014;10(4):1518–37.
[46] Zhechkov L, Heine T, Patchkovskii S, Seifert G, Duarte HA. An efficient a posteriori treatment for dispersion interaction in density-functional-based tight binding. J Chem Theory Comput 2005;1(5):841–7.
[47] Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc 1992;114(25):10024–35.
[48] Nakano T et al. Fragment molecular orbital method: application to polypeptides. Chem Phys Lett 2000;318(6):614–8.
[49] Pedregosa F et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 2011;12:2825–30.
[50] Virtanen P et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261–72.
[51] Brownlee J. XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn. (Machine Learning Mastery, 2016).
[52] Fernandes A, Vinga S, Stultz CM. Improving protein expression prediction using extra features and ensemble averaging. PLoS ONE 2016;11(3):e0150369.
[53] Huang X, Zheng W, Pearce R, Zhang Y, Valencia A. SSIPe: accurately estimating protein–protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. Bioinformatics 2020;36(8):2429–37.
[54] Li G et al. SAAMBE-SEQ: a sequence-based method for predicting mutation effect on protein–protein binding affinity. Bioinformatics 2021;37(7):992–9.
[55] Zhang N et al. MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions. Iscience 2020;23(3):100939. https://doi.org/10.1016/j.isci.2020.100939.
[56] Wang M, Cang Z, Wei G-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. Nature Machine Intelligence 2020;2(2):116–23.
[57] Chen C. et al. Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike RBD and human ACE2. Proceedings of the National Academy of Sciences 118, e2106480118, doi:10.1073/pnas.2106480118 (2021).
[58] Buratto D, Saxena A, Ji Q, Yang G, Pantano S, Zonta F. Rapid assessment of binding affinity of SARS-COV-2 spike protein to the human angiotensin-converting enzyme 2 receptor and to neutralizing biomolecules based on computer simulations. Front Immunol 2021;12. https://doi.org/10.3389/fimmu.2021.73009910.3389/fimmu.2021.730099.s00110.3389/fimmu.2021.730099.s00210.3389/fimmu.2021.730099.s00310.3389/fimmu.2021.730099.s00410.3389/fimmu.2021.730099.s005.
[59] Jumper J et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;1–11.
[60] Yang J et al. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci 2020;117(3):1496–503.