ProteinUnet2 for Fast Protein Secondary Structure Prediction: A Step Towards Proper Evaluation

Katarzyna Stapor (katarzyna.stapor@polsl.pl)
Silesian University of Technology

Krzysztof Kotowski
Silesian University of Technology

Tomasz Smolarczyk
Silesian University of Technology

Irena Roterman
Jagiellonian University Medical College

Research Article

Keywords: protein secondary structure prediction, U-Net, deep learning, PSSM, HHblits

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-900318/v1

License: ©  This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background: The importance of protein secondary structure (SS) prediction is widely known, its solution enables learning about the role of a protein in organisms. As the experimental methods are expensive and sometimes impossible, many SS predictors, mainly based on different machine learning methods have been proposed for many years. SS prediction as the imbalanced classification problem should not be judged by the commonly used Q3/Q8 metrics. Moreover, as the benchmark datasets are not random samples, the classical statistical null hypothesis testing based on the Neyman-Pearson approach is not appropriate. Also, the state-of-the-art predictors have usually relatively long prediction times.

Results: We present a new deep network ProteinUnet2 for SS prediction which is based on U-Net convolutional architecture. We also propose a new statistical methodology for prediction performance assessment based on the significance from Fisher-Pitman permutation tests accompanied by practical significance measured by Cohen's effect size. Through an extensive evaluation study, we report the performance of ProteinUnet2 in comparison with two state-of-the-art methods SAINT and SPOT-1D on benchmark datasets TEST2016, TEST2018, and CASP12.

Conclusions: Our results suggest that ProteinUnet2 has much shorter prediction times while maintaining (or outperforming) the mentioned predictors. We strongly believe that our proposed statistical methodology will be adopted and used (and even expanded) by the research community.

Background

The function of a protein is correlated with tertiary structure, also known as the native structure is a unique, stable, and kinetically accessible three-dimensional structure [1]. The first tertiary structure was determined for myoglobin by John Kendrew and his associates in 1957 [2]. For the studies on the structure of globular proteins, Kendrew received the Nobel Prize in Chemistry in 1962. More than 60 years later, there are 177,426 protein structures deposited in the Protein Data Bank [3] as of May 9th, 2021. For comparison, UniProtKB/Swiss-Prot database, which contains manually annotated and reviewed protein sequence (primary structure) has 564,638 sequences deposited and UniProtKB/TrEMBL, which contains automatically annotated and not reviewed sequences, has 214,406,399 sequences deposited as of May 9th, 2021 (The UniProt Consortium, 2021). The cost of determining sequence is significantly lower compared to the cost of determining the structures [4]. Hence researchers try to create statistical or machine learning that would predict the structure of the proteins.

For the secondary structure prediction, three generations of methods and algorithms are described in the literature [5]. The first generation, represented by Chou-Fasman's method, was leveraging statistical propensities of amino acids residues towards a specific secondary structure class [6]. The prediction accuracy of such methods was usually less than 60%.

The second generation of methods started in the 1980s and was leveraging sophisticated statistical methods, machine learning techniques as well as information about the neighboring residues usually using a sliding window approach [5]. It was represented by methods like GOR [7] or Lim [8], but the accuracy was still less than 65% [9].

The third generation of methods could be characterized especially by deep neural networks and additional features based on multiple sequence alignment profiles such as position-specific scoring matrices (PSSM) [10] or HHblits [11]. The accuracy of those methods reached 80% for models such as PSIPRED [12]. Given the growing number of known protein sequences, and more effective neural network architectures, recent methods are able to predict the secondary structure with more than 85% accuracy like SPOT-1D [13] based on long short-term memory (LSTM) bi-directional recurrent neural networks (BRNN) or SAINT [14] based on convolutions with self-attention mechanism.

In this study, we present ProteinUnet2, a significantly extended and improved version of ProteinUnet, our previous deep neural network architecture for SS3 and SS8 prediction from a single sequence [15]. It is now possible to feed any number of features to the input of the network (like PSSM or HHblits). We performed the analysis of the significance of the input features resulting in the selection of their best combination. The architecture has been improved with the addition of attention and dropout layers and training with a variable learning rate. This new architecture allowed us to significantly decrease the prediction times compared to the best SS predictors SAINT and SPOT-1D while maintaining similar or better performance on the benchmark datasets TEST2016,
For the first time (to our knowledge), we raised the problem of the incorrect methodology used for prediction efficiency assessment in the previously published works. The SS prediction is a heavily imbalanced classification problem and should not be judged using commonly used Q3/Q8 metrics. Instead, we proposed to use the Adjusted Geometric Mean (AGM) metric [16], which has been proven to be more appropriate for bioinformatics imbalanced classification problems [17]. Moreover, as the benchmark datasets are not random samples, the classical null hypothesis significance testing using the Neyman-Pearson inference approach should not be used. We propose the new assessment methodology based on Fisher-Pitman model of inference - statistical significance from the permutation tests. We also suggest supplementing such statistical significance with the practical significance measured by Cohen’s effect size. Using the proposed statistical methodology, we compared ProteinUnet2 with state-of-the-art predictors, SAINT and SPOT-1D.

Thus, we have made the following significant contributions: (i) we proposed a new U-Net-based deep architecture that enabled us to decrease the prediction times while maintaining similar or better performance than other state-of-the-art methods, (ii) introduced the new statistical methodology for prediction performance assessment, more appropriate in highly imbalanced SS8 prediction problem.

Results And Discussion

Like the authors of SAINT, we focused only on SS8 prediction analysis as it contains more useful information, does not depend on the SS3 mapping method, and is much more challenging to solve.

Predictors comparison

We compare ProteinUnet2 against the most recent and accurate SS8 predictors SPOT-1D and SAINT. These state-of-the-art methods have been shown to outperform other popular predictors like MUFOLD-SS [18] or NetSurfP-2.0 [19]. For the reasons stated in Methods section, in the comparison of performance, we focus mainly on the Adjusted Geometric Mean (AGM) metric for each structure (Table 2) as well as the macro-averaged AGM (Table 3) to assess the overall performance. The results for F1 score (Table 1), Q8 (Table 3), precision (Supplementary Table S3), and recall (Supplementary Table S4) are also presented.

Figure 1 presents the boxplots of macro-averaged F1 and AGM as well as Q8 metrics at the sequence level on TEST2016, TEST2018, and CASP12 datasets for 3 predictors: ProteinUnet2, SPOT-1D, and SAINT. These boxplots reveal small differences between the predictors’ medians and means (denoted by red triangles) for all presented metrics. Also, very high variability in all distributions is clearly visible. Moreover, the distributions of metrics on TEST2016/8 datasets contain many outliers. To compare quantitatively the observed slight difference we used the statistical methodology proposed in Methods section.

Table 2 and Table 3 report the performances obtained for AGM metric for 8 structures and the macro-average respectively. Table 4 presents the obtained p-values together with Cohen’s effect sizes for two separate comparisons between classifiers: ProteinUnet2 vs. SPOT-1D, and ProteinUnet2 vs. SAINT; on three test datasets: TEST2016, TEST2018, and CASP12.

The obtained macro-averaged AGM results in Table 3 and Table 4 prove that ProteinUnet2 has a statistically significantly higher mean than SAINT and SPOT-1D on TEST2016 dataset (p < 0.01). Accompanying Cohen’s effect sizes are 0.1399 (very small) and 0.290 (small), respectively. ProteinUnet2 has also statistically significantly better macro-averaged AGM than SPOT-1D on TEST2018 dataset (p < 0.01, very small effect 0.144) while on CASP12 dataset we observe a similar very small effect (0.186) but no significance (p = 0.102), probably because of the small sample size. Regarding the differences in performances on single classes (Table 2 and Table 4), ProteinUnet2 is significantly better (p < 0.01) than SAINT and SPOT-1D on rare class G on TEST2016 and TEST2018 datasets (small effect sizes 0.316, 0.327 and 0.280, 0.324, respectively). Very small effect sizes are observed on this class for both classifiers on CASP12 dataset, but with no statistical significance (small sample size). ProteinUnet2 is also significantly better than both compared classifiers on rare class G on TEST2016 dataset. It is worth emphasizing that despite the lack of significance, ProteinUnet2 obtains small effect sizes (0.259 and 0.199) on class E on small CASP12 dataset when comparing with other classifiers.

In summary, when the appropriate AGM metric is used for assessment of classifiers’ performance on imbalanced SS8 prediction problem ProteinUnet2 is significantly better in overall performance (macro-averaged AGM) than SPOT-1D and SAINT on TEST2016 dataset and achieves

Loading [MathJax]/jax/output/CommonHTML/jax.js
comparable results with SAINT on this dataset. The comparison of ProteinUnet2 on a relatively small CASP12 dataset leads to the conclusion that there is no significant difference between our predictor and SPOT-1D nor SAINT.

For the reasons stated in Methods section, we do not discuss and compare classifiers using F1 score or Q8. However, for easier comparison with the previous literature, we report the obtained performances in Table 1 and Table 3.

Table 1
The comparison of F1 score for each SS8 separately at the residue level on all test sets for ProteinUnet2 vs SPOT-1D (circle symbol) and SAINT (square symbol).

| SS8 | ProteinUnet2 | SPOT-1D (●) | SPOT-1D (●) | SPOT-1D (●) | SPOT-1D (●) | SPOT-1D (●) | SPOT-1D (●) |
|-----|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
|     | TEST2016     | TEST2018    | CASP12      | TEST2016    | TEST2018    | CASP12      | TEST2016    | TEST2018    | CASP12      |
| H   | 0.912        | 0.904       | 0.904       | 0.912       | 0.906       | 0.901       | 0.912       | 0.908       | 0.895       |
| B   | 0.228        | 0.241       | 0.191       | 0.170       | 0.182       | 0.113       | 0.183       | 0.180       | 0.182       |
| E   | 0.862        | 0.843       | 0.817       | 0.865       | 0.849       | 0.817       | 0.864       | 0.848       | 0.819       |
| G   | 0.455        | 0.441       | 0.400       | 0.445       | 0.431       | 0.420       | 0.467       | 0.451       | 0.426       |
| I   | 0.082        | 0.0          | 0.226       | 0           | 0           | 0           | 0.618       | 0           | 0           |
| T   | 0.623        | 0.596       | 0.563       | 0.626       | 0.604       | 0.562       | 0.639       | 0.619       | 0.573       |
| S   | 0.458        | 0.421       | 0.373       | 0.437       | 0.408       | 0.340       | 0.466       | 0.433       | 0.376       |
| C   | 0.677        | 0.638       | 0.635       | 0.682       | 0.647       | 0.632       | 0.687       | 0.652       | 0.646       |

Table 2. The comparison of AGM for each SS8 separately at the residue level on all test sets for ProteinUnet2 vs SPOT-1D (circle symbol) and SAINT (square symbol). The green/red symbols on the left/right side of the ProteinUnet2 score denote the statistical significance that it has a better/worse mean at the sequence level than other networks at p<0.01. The dash means the metric was impossible to calculate.

Table 3. The comparison of macro-averaged F1, macro-averaged AGM, and Q8 at the residue level on all test sets for ProteinUnet2 vs SPOT-1D (circle symbol) and SAINT (square symbol). The green/red symbols on the left/right side of the ProteinUnet2 score denote statistical significance that it has a better/worse mean at the sequence level than other networks at p<0.01.
Table 4

P-values from one-sided paired permutation tests and Cohen’s d effect sizes (after the backslash) for the difference in AGM between ProteinUnet2 and other networks using the alternative hypothesis that ProteinUnet2 has a greater mean. The dash means that there are not enough samples (< 20) to run tests.

| AGM | ProteinUnet2 versus | SPOT-1D | TEST2018 | CASP12 | TEST2016 | TEST2018 | CASP12 |
|-----|---------------------|---------|----------|--------|----------|----------|--------|
|     |                     | TEST2016 | TEST2018 | CASP12 | TEST2016 | TEST2018 | CASP12 |
| SS8 | H                   | 0.723/0.019 | 0.934/0.102 | 0.552/0.020 | 0.963/0.053 | 0.539/0.006 | 0.147/0.190 |
|     | B                   | 1e-4/0.316  | 2e-4/0.327  | 0.031/0.391  | 1e-4/0.280  | 2e-4/0.324  | 0.501/0.182 |
|     | E                   | 0.391/0.009  | 0.988/0.133  | 0.062/0.259  | 0.059/0.050  | 0.656/0.056  | 0.126/0.199 |
|     | G                   | 1e-4/0.163  | 0.014/0.164  | 0.770/0.124  | 0.002/0.098  | 0.140/0.082  | 0.582/0.046 |
|     | I                   | -          | -          | -          | -          | -          | -          |
|     | T                   | 0.544/0.003  | 0.663/0.031  | 0.352/0.056  | 0.928/0.042  | 0.792/0.056  | 0.834/0.144 |
|     | S                   | 1e-4/0.242  | 0.076/0.094  | 1e-4/0.306  | 0.087/0.040  | 0.397/0.017  | 0.269/0.088 |
|     | C                   | 1.000/0.139  | 0.716/0.038  | 0.510/0.225  | 1.000/0.266  | 0.801/0.056  | 0.940/0.003 |
| Macro avg | 1e-4/0.290  | 0.010/0.144  | 0.102/0.186  | 1e-4/0.139  | 0.077/0.091  | 0.308/0.073 |

Dependence on the number of homologs

Figure 2 shows the dependence of the macro-averaged AGM on the number of effective homologous sequences (Neff) for the TEST2016 set. Each point on the plot is an average of at least 20 proteins with the given Neff (rounded down to the nearest integer) calculated by HHBlits. This figure shows that metrics increase with the increasing Neff. AGM for all networks is much lower for sequences with less than 4 homologs (Neff < 4). Interestingly, ProteinUnet2 shows the highest scores for these sequences. The advantage of ProteinUnet2 over SPOT-1D is statistically significant for Neff = 1 and Neff in range 7 to 12, and over SAINT for Neff 6 and 8.

Analysis of incorrect predictions

We noticed that for particular sequences from TEST2016 (5doiE, 5dokA, 5d6hB) the performance of all networks is very poor (AGM < 0.3). It turned out that they are missing some amino acids in the original PDB files (5doiE – 4 gaps with 35 out of 128 AA missing, 5dokA – 1 gap with 34 out of 204 AA missing, 5d6hB – 8 gaps with 54 out of 152 AA missing). The gaps for 5d6hB chain are presented in Figure 3 generated using PDBsum web server [20] with 3D visualization by VMD software [21] in Figure 4. Even a single missing amino acid may change the secondary structure [22]. It may explain the very low performance for mentioned proteins. Thus, the problem lays in the dataset itself.

Running time

Table 5 presents the inference time of ProteinUnet2 and SPOT-1D. The times were measured on the PC with AMD Ryzen 9 3900X CPU with Nvidia RTX 3070 GPU. They do not include PSSM, HHBlits, or SPOT-Contact features extraction, just the network processing. ProteinUnet2 is orders of magnitude faster than SPOT-1D (up to 50 times faster for TEST2016 dataset) and around 10% faster than SAINT. One epoch of ProteinUnet2 training takes around 2 minutes which gives an average 30 minutes per model. The training times of SPOT-1D and SAINT were not reported, but are expected to be proportionally longer.
Table 5  
Prediction times in seconds for ProteinUnet2, SPOT-1D, and SAINT on all test sets

| Dataset   | ProteinUnet2 | SPOT-1D  | SAINT |
|-----------|--------------|----------|-------|
| TEST2016  | 229          | 11796    | 252   |
| TEST2018  | 88           | 3644     | 98    |
| CASP12    | 59           | 486      | 64    |

Conclusions

ProteinUnet2 significantly extends and improves our previous ProteinUnet deep architecture [15]. It introduces multiple inputs with evolutionary profiles like PSSM, HHblits, and contact maps. The performance is increased by additional mechanisms of attention and dropouts. ProteinUnet2 achieves comparable or better results to the state-of-the-art models – SPOT-1D based on LSTM-BRNN architecture and SAINT based on self-attention modules while running several times faster than SPOT-1D and 10% faster than SAINT. That makes it especially useful in large-scale predictions and applications on low-cost and embedded devices.

The proposed methodology for assessment of the performance of secondary structure predictors based on an appropriate measure for imbalanced classification together with permutation tests as well as analyzing significance of performance difference based on effect sizes may and should be further developed through, for example, other measures of effect size or its interpretations appropriately to the application domain.

Methods

Datasets

For a fair comparison, we use the same training, validation, and test sets as SPOT-1D and SAINT. The training set TR10029 contains 10 029 proteins, and the validation set VAL983 has 983 proteins. We benchmark our model on 3 test sets: TEST2016 with 1213, TEST2018 with 250, and CASP-12 with 49 proteins (see [13] and [14] for the details about these datasets).

Metric for secondary structure imbalance classification problem

Some protein secondary structures, e.g. alpha-helices, are much more frequent than others (Figure 5). This leads to the class imbalance problem [23] which is rarely mentioned or addressed in the literature about SS prediction. Assessing the performance of SS classifiers plays a vital role in their construction process. The most commonly used metrics of SS prediction performance are overall accuracies Q3 and Q8 [5,9,24] that are not appropriate for imbalance problems [25,26]. Using them may lead to the accuracy paradox where high accuracy is not necessarily an indicator of good classification performance [26], e.g., a classifier that always predicts class H will have ten times better accuracy than a classifier that always predicts class G (see Figure 5).

The existing popular measures proposed for imbalanced learning like the geometric mean or F-score can still result in suboptimal models [17]. For these reasons, we used the Adjusted Geometric Mean (AGM) well-suited for bioinformatics imbalance problems [16]. It has been shown both analytically and empirically to perform better than F-score. It has no parameters (like beta in F-score). It is given by Eq. 1 where GM is the geometric mean, SP is specificity, Nn is the proportion of negative samples, and SE is sensitivity.

\[
AGM = \frac{GM + SP \times N_n}{1 + N_n}, \quad SE > 0 \quad (1)
\]

\[
AGM = 0, \quad SE = 0
\]

AGM's purpose is to increase the sensitivity while keeping the reduction of specificity to a minimum. Also, the higher the degree of imbalance, the higher reaction to changes in specificity. It returns values between 0 (the worst prediction) and 1 (a perfect prediction).
We calculate AGM for each structure separately. To assess the overall quality, we use macro-averaged F1 and AGM scores. That is, we take an average of overall scores for each structure. This way we do not favor more frequent classes.

Significance testing and effect size

Null hypothesis significance testing (nhst) is a commonly used statistical method for comparing classifier performances [26,27] although the authors mention their caveats. In the case where the test datasets are not random (like the benchmark datasets used in the evaluation of SS prediction), using classical nhst is problematic [26]. Random permutation tests based on the Fisher-Pitman model of inference [28] are an alternative that is strongly recommended by us in that case. In our experiments, we used a one-sided paired sample permutation test for difference in mean classifier performances (perm.paired.loc function from wPerm R package).

The tests are performed at the sequence level. Tests for separate structures are performed only on the subsets of sequences for which it was possible to calculate a given metric (e.g., if the structure is present in the ground truth or prediction).

Here (to our knowledge, for the first time), we propose a new methodology to compare the significance of classifier performance differences. Significance testing as well as permutation tests alone do not resolve the problem of inferential interpretation. Statistical significance shows only that an effect exists, practical significance - the effect size - shows that the effect is large enough to be meaningful in the real world. Statistical significance alone can be misleading because it’s influenced by the sample size. Increasing the sample size always makes it more likely to find a statistically significant effect, no matter how small the effect is in the real world. Effect sizes are independent of the sample size and are an essential component when evaluating the strength of a statistical claim. Some authors [29] proposed to use confidence intervals for estimation of effect size, but they require a random sample to enable inference. Cohen's effect size $d$ [30] that we propose to use in our study for a paired-samples can be calculated by dividing the mean difference by the standard deviation of the differences. Whether an effect size should be interpreted as negligible ($d < 0.01$), very small ($d < 0.2$), small ($d < 0.5$), medium ($d < 0.8$), or large ($d < 1.2$) depends on the context (application) and its operational definition [31]. Thus, we propose to report statistical significance (denoted by p-values) together with practical significance represented by effect sizes (here, Cohen's effect size $d$ for a paired-samples).

ProteinUnet2 architecture

U-Net architectures have proven to be extremely effective in image segmentation tasks [32, 33]. The U-shaped architecture of ProteinUnet2 is based on the idea from our previous ProteinUnet for secondary structure prediction [15] (for which the results are presented in Supplementary Table S1). The new architecture was adjusted to handle multiple inputs by using multiple contractive paths, one for each input (Fig. 6). After each down-block, the features of all inputs are concatenated together and passed to the up-block via a skip connection. There are two output layers with softmax activations connected to the last up-block, separately for SS3 and SS8. In ProteinUnet2, we limited the maximum supported sequence length from 1024 to 704 to further improve training and inference times without losing accuracy. Anyway, SPOT-1D and SAINT were not trained with proteins longer than 700, and there are no proteins longer than 704 in our datasets. The input features and the number of filters were selected experimentally as described in the next section.

To mitigate the problem of the increased number of inputs and parameters of the network, in the final ProteinUnet2 architecture (Figure 6), we modified the architecture to be similar to the Attention U-Net [34]. That is, we decreased the number of convolutions in each down-block from 3 to 2, added dropouts with 0.1 rate between convolutions in all blocks, and applied attention gates right before the concatenation operation. ProteinUnet2 was implemented in the environment containing Python 3.8 with TensorFlow 2.4 accelerated by CUDA 11.0 and cuDNN 8.0. The code and trained models are available on the CodeOcean platform (https://codeocean.com/capsule/0425426) ensuring high reproducibility of the results.

Feature representation and selection

ProteinUnet2 takes a sequence of feature vectors $X = (x_1, x_2, x_3, ..., x_N)$ as input, where $x_i$ is the feature vector corresponding to the $i$th residue, and it returns two sequences of structure probabilities vectors $Y = (y_1, y_2, y_3, ..., y_N)$ as output, where $y_i$ is the vector of 3 or 8 probabilities of $i$th residue being in one of SS3 or SS8 states. The 8 states are specified by the secondary structure assignment program Define Secondary Structure of Proteins (DSSP) [35]. There are three helix states: 310-helix (G), alpha-helix (H), and pi-helix (I); three strand states: beta-bridge (B) and beta-strand (E); and three coil types: high curvature loop (S), beta-tum (T), and pi-helix (I); three coil types: high curvature loop (S), beta-turn (T), and coil (C).
and coil (C). These 8 classes are converted into 3-class problem by grouping the states: G, H, and I into H; B and E into E; and S, T, and C into C.

Similar to SPOT-1D, our final model contains 20 features from PSSM [10], and 30 features from HHM profiles [11]. The features were standardized to ensure 0 mean and SD of 1 in the training data. Additionally, we use contact maps generated by SPOT-Contact [36]. We use the same windowing scheme as described in SPOT-1D, but we do not standardize the contact maps as they are already in the acceptable range < 0, 1>. The window size of 50 was selected experimentally based on the results from Supplementary Table S1 that shows F1 scores and accuracies on the largest TEST2016 set for a single ProteinUnet trained with different input features on TR10029 and validated on VAL983. Supplementary Table S1 suggests that SPOT-Contact features gave better results of SS8 prediction than any other input alone. The worst results are reported for 7 physicochemical properties [37]. Thus, we did not investigate them further in ProteinUnet2.

Supplementary Table S2 shows the F1 scores and accuracies on TEST2016 for our proposed ProteinUnet2 trained with different combinations of input features and a different number of filters in down-blocks. It reveals that SPOT-Contact features alone outperformed combined PSSM and HHblits. However, the combination of all these 3 features (keeping the same number of filters) increased F1 scores for all SS8 structures when comparing to any other feature alone. Most of our results are better for the higher number of filters, but we did not test numbers higher than 64 to avoid overfitting and to keep the number of filters in all blocks the same as in the original ProteinUnet. Thus, we decided to investigate further only the combination PHSA 64 attention from Supplementary Table S2. The architecture for this combination is presented in Fig. 6.

Training procedures and ensembling

For the initial experiments presented in Supplementary Table S1 and Supplementary Table S2 the single models were trained on TR10029 dataset and validated on VAL983. However, in the final model, datasets TR10029 and VAL983 were combined and then divided into 10 stratified folds to ensure a similar ratio of each SS8 structure in each fold. There were nine factors of stratification: the sequence length - shorter/longer than mean sequence length, and one factor for each of 8 structures occurrence - fewer/more occurrences than a mean number of occurrences per chain. We trained 10 models, each time using a different fold as a validation set and the rest as the training set. The models were trained to optimize the categorical cross-entropy loss using Adam optimizer [38] with batch size 8 and an initial learning rate 0.001. The learning rate was reduced by a factor of 0.1 when there was no improvement in the validation loss for 4 epochs. The training was running until the validation loss was not improving for 7 epochs. Finally, the ensemble was created from the models with the lowest validation loss for each fold by taking the average of their softmax outputs, forming the final ProteinUnet2 prediction.

Abbreviations

AGM
Adjusted geometric mean
BRNN
Bidirectional recurrent neural network
CASP
The Critical Assessment of protein Structure Prediction
LSTM
Long short-term memory
PSSM
Position-Specific Scoring Matrix
SS3
3-class secondary structure
SS8
8-class secondary structure

Declarations

Ethics approval and consent to participate
Consent for publication
Not applicable.

Availability of data and materials
The prediction code and trained models are available on CodeOcean platform ensuring high reproducibility of the results: https://codeocean.com/capsule/0425426. The data were shared by authors of SPOT-1D and SAINT.

Competing interests
None declared.

Funding
This work was supported by Statutory Research funds of Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland (02/100/BK_21/0008).

Authors' contributions
KS proposed metrics, designed the validation and statistical testing procedures, and commented on the results. KK designed the architecture of Pro-teinUnet2, implemented the code, generated results, prepared figures, and edited the manuscript. TS prepared the literature review, generated SPOT-1D and SAINT predictions, measured the running times, and edited the manuscript. IR provided substantive support, reviewed the manuscript, and analysed the incorrect predictions. All authors read and approved the final manuscript. KK and KS should be regarded as Joint First Authors.

Acknowledgements
We would like to thank the authors of SPOT-1D and SAINT for providing the PSSM, HHblits, and contact maps of the proteins in the training and test sets.

Authors' information
1Department of Applied Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
2Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College, Medyczna 7, 30-688 Krakow, Poland

References
1. Anfinsen CB. Principles that Govern the Folding of Protein Chains. Science. 1973;181:223–230. doi: 10.1126/science.181.4096.223.
2. Kendrew J, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958;181:662–6.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–242. doi: 10.1093/nar/28.1.235.
4. Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K, Zhou Y. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform. 2018;19:482–494. doi: 10.1093/bib/bbw129.

5. Smolarczyk T, Roterman-Konieczna I, Stapor K. Protein Secondary Structure Prediction: A Review of Progress and Directions. Curr Bioinforma. 2020;15:90–107.

6. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974;13:222–45.

7. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol. 1978;120:97–120.

8. Lim VI. Algorithms for prediction of α-helical and β-structural regions in globular proteins. J Mol Biol. 1974;88:873–94.

9. Jiang Q, Jin X, Lee S-J, Yao S. Protein secondary structure prediction: A survey of the state of the art. J Mol Graph Model. 2017;76:379–402. 10.1016/j.jmgm.2017.07.015.

10. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci U S A. 1993;90:7558–62.

11. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012;9:173–5.

12. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292:195–202.

13. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Valencia A, editor. Bioinformatics. 2019 Jul;35(14):2403–10.

14. Uddin MR, Mahbub S, Rahman MS, Bayzid MS. SAINT: self-attention augmented inception-inside-inception network improves protein secondary structure prediction. Bioinformatics. 2020 Nov 1;36(17):4599–608.

15. Kotowski K, Smolarczyk T, Roterman-Konieczna I, Stapor K. ProteinUnet—An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. J Comput Chem. 2021;42(1):50–9.

16. Batuwiwa R, Palade V. Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning. J Bioinform Comput Biol. 2012 Aug 1;10(04):1250003.

17. Japkowicz N. Assessment Metrics for Imbalanced Learning. In: Imbalanced Learning: Foundations, Algorithms, and Applications. The Institute of Electrical and Electronics Engineers, Inc.; 2013. p. 187–206.

18. Zhang J, Wang Q, Barz B, He Z, Kosztin I, Xu D. MUFOLD: A new solution for protein 3D structure prediction. Proteins. 2010 Apr;78(5):1137–52.

19. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, Sommer MOA, Winther O, Nielsen M, Petersen B, Marcantili P. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins Struct Funct Bioinforma. 2019;87(6):520–7.

20. Laskowski RA, Jablonska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci. 2018;27(1):129–34.

21. Humphrey W, Dalke A, Schulten K. VMD – Visual Molecular Dynamics. J Mol Graph. 1996;14:33–8.

22. Banach M, Fabian P, Stapor K, Konieczny L, Roterman I. Structure of the Hydrophobic Core Determines the 3D Protein Structure—Verification by Single Mutation Proteins. Biomolecules. 2020 May 14;10(5):767.

23. Ling CX, Sheng VS. Class Imbalance Problem. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning [Internet]. Boston, MA: Springer US; 2010 [cited 2021 Jun 29]. p. 171–171. Available from: https://doi.org/10.1007/978-0-387-30164-8_110

24. Wang S, Peng J, Ma J, Xu J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Sci Rep. 2016;6:doi: 10.1038/srep18962.

25. Weiss GM. Mining with rarity: a unifying framework. ACM SIGKDD Explor Newsl. 2004;6:7–19.

26. Stapor K, Kseniewicz P, García S, Woźniak M. How to design the fair experimental classifier evaluation. Appl Soft Comput. 2021 Jun 1;104:107219.

27. Japkowicz N, Shah M. Evaluating Learning Algorithms: A Classification Perspective [Internet]. Cambridge: Cambridge University Press; 2011. [cited 2021 Jul 25]. Available from: https://www.cambridge.org/core/books/evaluating-learning-

Loading [MathJax]/jax/output/CommonHTML/jax.js
algorithms/3CB22D16AB609D1770C24CA2CB5A11BF

28. Good PI. Permutation, Parametric, and Bootstrap Tests of Hypotheses. 3rd Edition. New York: Springer-Verlag New York; 2005.
29. Berrar D, Lozano JA. Significance tests or confidence intervals: which are preferable for the comparison of classifiers? J Exp Theor Artif Intell. 2013;25:189–206.
30. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. New York: Routledge; 1988.
31. Sawilowsky SS. New Effect Size Rules of Thumb. J Mod Appl Stat Methods. 2009;8:597–9.
32. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb;18(2):203–11.
33. Kotowski K, Nalepa J, Dudzik W. Detection and Segmentation of Brain Tumors from MRI Using U-Nets. In: Crimi A, Bakas S, editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer International Publishing; 2020. p. 179–90. (Lecture Notes in Computer Science).
34. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocke B, Rueckert D. Attention U-Net: Learning Where to Look for the Pancreas. ArXiv180403999 Cs [Internet]. 2018 May 20 [cited 2021 Mar 26]; Available from: http://arxiv.org/abs/1804.03999
35. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
36. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics. 2018 Dec 1;34(23):4039–45.
37. Fauchère J, Charton M, Kier LB, Verloop A, Pliska V. Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res. 1988;32:269–78.
38. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv. 2014;cs.LG.

Figures
Figure 1

Boxplots of macro-averaged F1 (top row), macro-averaged AGM (middle row), and Q8 (bottom row) at the sequence level for TEST2016 (left column), TEST2018 (middle column), and CASP12 (right column). The red triangles denote mean scores. The exact values of means and standard deviations are given in Supplementary Table S5.
Figure 2

Macro-averaged AGM for predicted SS8 as a function of the number of effective homologous sequences for the TEST2016 set. The circle/square over the point means that ProteinUnet2 has statistically significantly larger mean than SPOT-1D/SAINT for that Neff value (one-sided paired permutation test at $p < 0.01$).

Figure 3

The primary and secondary structure of chain B in 5d6h protein from PDB.

Figure 4

3D visualization of chain B in 5d6h protein with missing amino acids
Figure 5

Frequencies of secondary structures. The frequencies of 8 secondary structures in TEST2016 dataset.

Figure 6

The schematic architecture of ProteinUnet2. AA is a one-hot encoded sequence of amino acids. FC and FE are the numbers of features in contractive and expanding paths.

Supplementary Files
