*Virus Evolution*, 2021, 7(1): veaa106

doi: 10.1093/ve/veaa106
Research Article

# Tuning intrinsic disorder predictors for virus proteins

Gal Almog,[1,†] Abayomi S. Olabode,[1,*,†,‡] and Art F.Y. Poon[1,2,3,§]

[1]Department of Pathology & Laboratory Medicine, Western University, Dental Sciences Building, Rm. 4044 London, Ontario, Canada, N6A 5C1, [2]Department of Applied Mathematics, Western University, Middlesex College Room 255, 1151 Richmond Street London, Ontario, Canada, N6A 5B7 and [3]Department of Microbiology & Immunology, Western University, 1151 Richmond Street London, Ontario, Canada, N6A 3K

[*]Corresponding author: E-mail: aolabode@uwo.ca

[†]The first two authors contributed equally to the study.

[‡]https://orcid.org/0000-0002-6620-8694

[§]https://orcid.org/0000-0003-3779-154X

## Abstract

Many virus-encoded proteins have intrinsically disordered regions that lack a stable, folded three-dimensional structure. These disordered proteins often play important functional roles in virus replication, such as down-regulating host defense mechanisms. With the widespread availability of next-generation sequencing, the number of new virus genomes with predicted open reading frames is rapidly outpacing our capacity for directly characterizing protein structures through crystallography. Hence, computational methods for structural prediction play an important role. A large number of predictors focus on the problem of classifying residues into ordered and disordered regions, and these methods tend to be validated on a diverse training set of proteins from eukaryotes, prokaryotes, and viruses. In this study, we investigate whether some predictors outperform others in the context of virus proteins and compared our findings with data from non-viral proteins. We evaluate the prediction accuracy of 21 methods, many of which are only available as web applications, on a curated set of 126 proteins encoded by viruses. Furthermore, we apply a random forest classifier to these predictor outputs. Based on cross-validation experiments, this ensemble approach confers a substantial improvement in accuracy, e.g., a mean 36 per cent gain in Matthews correlation coefficient. Lastly, we apply the random forest predictor to severe acute respiratory syndrome coronavirus 2 ORF6, an accessory gene that encodes a short (61 AA) and moderately disordered protein that inhibits the host innate immune response. We show that disorder prediction methods perform differently for viral and non-viral proteins, and that an ensemble approach can yield more robust and accurate predictions.

**Key words:** intrinsically disordered proteins; protein disorder prediction; virus proteins; ensemble classifier; machine learning.

## 1. Introduction

For almost a century, it was assumed that proteins required a properly folded and stable three-dimensional or tertiary structure in order to function (Lichtenthaler 1995; Necci et al. 2018; Uversky 2019). More recently, it has become evident that many proteins and protein regions are disordered, which are referred to as intrinsically disordered proteins (IDPs) and intrinsically disordered protein regions (IDPRs), respectively. Both IDPs and IDPRs can perform important biological functions despite lacking a properly folded and stable tertiary structure (Wright and Dyson 1999; Uversky 2019).

These kinds of proteins are an important area of research because they play major roles in cell regulation, signaling, differentiation, survival, apoptosis, and proliferation (Kozlowski and Bujnicki 2012; Katuwawala, Oldfield, and Kurgan 2019). Some are also postulated to be involved in disease etiology and could represent potential targets for new drugs (Uversky,

Oldfield, and Dunker 2008; Hu et al. 2016). Virus-encoded IDPs facilitate multiple functions such as adaptation to new or dynamic host environments, modulating host gene expression to promote virus replication, or counteracting host-defense mechanisms (Gitlin et al. 2014; Xue et al. 2014; Mishra et al. 2020). IDPRs may be more tolerant of non-synonymous mutations than ordered protein regions (Walter et al. 2019), which may partly explain why virus genomes can tolerate high mutation rates (Tokuriki et al. 2009; Sanjuán et al. 2010). Viruses also have very compact genomes with overlapping reading frames (Cotmore et al. 2005; Holmes 2009), in which mutations may potentially modify multiple proteins. This may confer viruses a greater capacity to acquire novel functions and interactions (Belshaw, Pybus, and Rambaut 2007). Overlapping regions tend to be more structurally disordered when compared to non-overlapping regions (Rancurel et al. 2009).

Several experimental techniques are available to detect IDPs and IDPRs. The most common methods identify either protein regions in crystal structures that have unresolvable coordinates (X-ray crystallography) or regions in nuclear magnetic resonance (NMR) structures that have divergent structural conformations (Ferreon et al. 2010; DeForte and Uversky 2016; Katuwawala, Oldfield, and Kurgan 2019). Other experimental techniques include circular dichroism (CD) spectroscopy and limited proteolysis (LiP) (Necci et al. 2018). The challenge, however, is that these methods are very labor-intensive and difficult to scale up to track the rapidly accumulating number of unique protein sequences in public databases (Ferreon et al. 2010; Katuwawala, Oldfield, and Kurgan 2019). At the time of this writing, over 60 million protein sequences have been deposited in the Uniprot database, yet only 0.02 per cent of these sequences have been annotated for disorder (Katuwawala, Oldfield, and Kurgan 2019). As a result, numerous computational techniques that could potentially predict intrinsic disorder in protein sequences have been developed. These techniques work based on the assumptions that compared to IDPs and IDPRs, ordered proteins have a different amino acid composition as well as levels of sequence conservation (Dunker et al. 2001; Uversky 2002). To date, about sixty predictors for intrinsic disorder in proteins have been developed (Atkins et al. 2015; Necci et al. 2018; Liu, Wang, and Liu 2019), which can be broadly classified into three major categories. The first category, the scoring function-based methods, predict protein disorder solely based on basic statistics of amino acid propensities, physio-chemical properties of amino acids, and residue contacts in folded proteins to detect regions of high energy. A second category is characterized by the use of machine learning classifiers (e.g. regularized regression models or neural networks) to predict protein disorder based on amino acid sequence properties. The third category are meta-predictors that predict disorder from an ensemble of predictive methods from the other two categories (Li et al. 2015; Necci et al. 2018; Katuwawala, Oldfield, and Kurgan 2019).

Different predictors of intrinsic disorder are developed on a variety of methodologies and will inevitably vary with respect to their sensitivities and biases in application to different protein sequences. As a result, it has been relatively difficult to benchmark these methods to identify a single disorder prediction method that can be classified as the most accurate relative to the others (Atkins et al. 2015). The DisProt database is a good resource for obtaining experimental data that has been manually curated for disorder in proteins, and can be used for benchmarking the performance of disorder predictors. As of 27 April 2020, the Disprot protein database contained $n = 3{,}500$ proteins of which 126 were virus-encoded proteins that have been annotated for intrinsic disorder as a presence–absence characteristic at the amino acid level (Piovesan et al. 2017; Hatos et al. 2020). Previously, Tokuriki et al. (2009) reported preliminary evidence that when compared to non-viruses, viral proteins possess many distinct biophysical properties including having shorter disordered regions. We are not aware of a published study that has previously benchmarked predictors of intrinsic disorder specifically for viral proteins. Here, we report results from a comparison of twenty-one disorder predictors on viral proteins from the DisProt database to firstly determine which methods work best for viruses, and secondly to generate inputs for an ensemble predictor that we evaluate alongside the predictors used individually.

## 2. Methods

### 2.1 Data collection

The Database of Protein Disorder (DisProt) (DisProt, 2000) was used to collect virus protein sequences annotated with intrinsically disordered regions, based on experimental data derived from various detection methods; e.g., X-ray crystallography, NMR spectroscopy, CD spectroscopy (both far and near UV), and protease sensitivity. DisProt records include the amino acid sequence and all disordered regions annotated with the respective detection methods as well as specific experimental conditions. At the time of our study, DisProt contained 3,500 author-verified proteins, of which all viral proteins were collected for the present study. A total of 126 virus proteins were obtained, derived from different detection methods. Similarly, a set of 126 non-viral proteins was sampled at random without replacement from the protein database for comparison.

We evaluated a number of disorder prediction programs and web applications. From the methods tested, we selected a subset of predictors favoring those that were developed more recently, are actively maintained, and performed well in previous method comparison studies (Necci et al. 2018; Nielsen and Mulder 2019). Where alternate settings or different versions based on training data were available for a given predictor, we tested all combinations. Our final set of 21 prediction methods tested were: SPOT-Disorder2 (Hanson et al. 2019), PONDR-FIT (Xue et al. 2010), IUPred2 (short and long) (Mészáros, Erdős, and Dosztányi 2018), PONDR (VLXT, XL1-XT, CAN-XT, VL3-BA, and VSL2 variants) (Peng et al. 2006), Disprot (VL2 and variants VL2-V, -C and -S; VL3, VL3H, and VSLB) (Vucetic et al. 2003), CSpritz (short and long) (Walsh et al. 2011), and ESpritz (variants trained on X-ray, NMR, and Disprot data) (Walsh et al. 2012). Although several other predictor models have been released online, the respective web services were unavailable or broken over the course of our data collection.

To obtain disorder predictions from the methods that were only accessible as web applications, i.e., with no source code or compiled binary standalone distribution, we wrote Python scripts to automate the process of submitting protein sequence inputs and parsing HTML outputs. We used Selenium in conjunction with ChromeDriver (v81.0.4044.69) (ChromeDriver: WebDriver for Chrome 2000) to automate the web browsing and form submission processes. For each predictor, we implemented a delay of 90 s between consecutive protein sequence queries to avoid overloading the webservers hosting the respective predictor algorithms with repeated requests. Due to issues with the Disprot webserver, we were only able to obtain predictions for the non-viral protein data set for thirteen of the predictors.

We converted each DisProt record to a binary vector corresponding to ordered/disordered state of residues in the amino acid sequence. To compare results between disorder prediction algorithms, we dichotomized continuous-valued residue predictions, i.e., intrinsic disorder probability, by locating the threshold that maximized the Matthews correlation coefficient (MCC) for each predictor applied to the DisProt training data. This optimal threshold was estimated using Brent's root-finding algorithm as implemented by the *optim* function in the R statistical computing environment (version 3.4.4). In addition, we calculated the accuracy, specificity, and sensitivity for each predictor from the contingency table of DisProt residue labels and dichotomized predictions.

### 2.2 Ensemble classifier training and validation

To assess whether the accuracy of existing predictors could be further improved on the virus-specific data set, we trained an ensemble classifier on the outputs of all predictors as features. Specifically, we used the random forest method implemented in the *scikit-learn* (version 0.23.1) Python module (Pedregosa et al. 2011), which employs a set of de-correlated decision trees and averages their respective outputs to obtain an ensemble prediction (Breiman 2001). To reduce bias, random forests fit the same decision trees to many bootstrap samples of the training data, and each committee of trees 'votes' for a particular classification (Hastie, Tibshirani, and Friedman 2001). By splitting the trees based on different samples of features, random forests reduce the correlation between trees and the overall variance.

We split the viral protein data into random testing and training subsets, with 30 per cent of protein sequences reserved for testing. Due to class imbalance in the data (i.e. only a minority of residues are labeled as disordered), we used stratified random sampling using the 'StratifiedShuffleSplit' function in the *scikit-learn* module. This function stratifies the data by label so that a constant proportion of labels is maintained in the training subset. Continuous-valued outputs from each predictor were normalized to a zero mean and unit variance. Thus, we did not apply the dichotomizing thresholds to these features (predictor outputs) when training the random forest classifier.

We used five-fold cross-validation to tune the four hyperparameters of the random forest classifier; namely: (1) the number of decision trees; (2) the maximum depth of any given decision tree; and the minimum number of samples required to split (3) an internal node or (4) a leaf node. To further minimize the effect of class imbalance in our data, we used over-sampling to balance the data with synthetic cases (Japkowicz 2000). As suggested in Hemmerich, Asilar, and Ecker (2020), we applied an over-sampling procedure at every iteration of the cross-validation analysis to avoid over-optimistic results. We used the Python package *imbalanced-learn* (Lemaître, Nogueira, and Aridas 2017) to over-sample the minority class (residues in intrinsically disordered regions) using the synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002). SMOTE generates new cases by sampling the original data at random with replacement, evaluates each sample's $k$ nearest neighbors in the feature space, and then generates new synthetic samples along the vectors joining the sample to one of the neighboring points. Over-sampling enables decision trees to be more generalizable by amplifying the decision region of the minority class.

Using the optimized tuning parameters, we fit the final model on all of the training data. We applied this final model to generate predictions on the reserved testing data and calculated the MCC, sensitivity, specificity, and accuracy. We repeated this process ten times with randomly generated seeds to split the data into training and testing subsets, and averaged these performance metrics across replicates.

### 2.3 Comparison to non-viral data

To characterize how the performance of individual disorder predictors might vary among proteins from viruses and non-viruses, we computed the root mean square error (RMSE) for all continuous-valued predictions relative to the Disprot label (0, 1). We visualized this error distribution using principal component analysis (PCA). As well, we trained a support vector machine (SVM) on the RMSE values to determine whether the virus/non-virus labels were separable in this space. We used the default radial basis kernel with the C-classification SVM method implemented in the R package *e1071* (Chang and Lin 2011), with hundred training subsets sampled at random without replacement for half of the data, and the remaining half for validation.

### 2.4 Data availability

We have released the data generated in this study and Python scripts for automating queries to the disorder prediction web servers under a permissive free software license at https://github.com/PoonLab/tuning-disorder-virus.

## 3. Results and discussion

### 3.1 Viral and non-viral proteins have similar levels of disorder

We obtained 126 viral and 126 randomly selected non-viral protein sequences from the DisProt database. The sequences were already annotated manually by a panel of experts for the presence or absence of disorder at each amino acid position, based on experimental data (Hatos et al. 2020). Supplementary Tables S1 and S2 summarize the composition of the viral and non-viral protein datasets, respectively. The viral protein data set represents twenty-two virus families and forty-eight species. Not surprisingly, human immunodeficiency virus type 1 was disproportionately represented in these data with sixteen entries corresponding to seven different gene products. Similarly, the non-viral protein data set was predominated by seventy-five human proteins, followed by twenty-three proteins from the yeast *Saccharomyces cerevisiae*. We found no significant difference in amino acid sequence lengths between viruses and all other organisms (Wilcoxon rank-sum test, $P = 0.60$), with median lengths of 355 [interquartile range, IQR: 145–846] and 395 [203–729] amino acids, respectively. Furthermore, the dispersion in sequence lengths was significantly greater among viral proteins relative to the nonviral proteins (Ansari-Bradley test, $P = 0.0028$). There was no significant difference in the proportion of residues in disordered regions between the viral and non-viral data (Wilcoxon $P = 0.97$). The mean proportions were 0.30 (interquartile range, IQR [0.07–0.42]) for viral and 0.30 [0.07–0.47] for non-viral proteins, and similar numbers of proteins exhibited complete disorder (13 and 9, respectively).

### 3.2 Divergent predictions of disorder in viral proteins

Our first objective was to benchmark the performance of different predictors of intrinsic protein disorder to determine which predictor conferred the highest accuracy for viral proteins. These predictors generate continuous-valued outputs that generally correspond to the estimated probability that the residue

is in an intrinsically disordered region. To create a uniform standard for comparison to the binary presence-absence labels, we optimized the disorder prediction thresholds as a tuning parameter for each predictor for the viral and non-viral datasets, respectively (Supplementary Tables S3 and S4). Put simply, residues with values above the threshold were classified as disordered. We used both the MCC (ranging from −1 to + 1 (Boughorbel, Jarray, and El-Anbari 2017)) and area under the receiver-operator characteristic curve (AUC, ranging from 0 to 1) to quantify the performance of each predictor.

These quantities were significantly correlated (Spearman's $\rho = 0.95$, $P = 5.2 \times 10^{-6}$) and identified ESpritz.Disprot, CSpritz.Long, and SPOT.Disorder2 as the most effective predictors for the viral proteins (Fig. 1A). ESpritz.Disprot obtained the highest overall values for both MCC and AUC (0.46 and 0.85, respectively). We note that SPOT.Disorder2 has recently been reported to exhibit a high degree of prediction accuracy for proteins of varying length (Hanson et al. 2017). In contrast, the predictors Disprot-VL2-V, PONDR-XL1, and PONDR-CAN performed very poorly on the viral dataset with MCC < 0.2 and AUC < 0.65. VL2-V is a 'flavour' of the VL2 predictors which were allowed to specialize on different subsets of a partitioned training set; for example, V tended to call higher levels of disorder in proteins of Archaebacteria (Vucetic et al. 2003). Similarly, PONDR-XL1 was optimized to predict longer disordered regions and PONDR-CAN was trained specifically on calcineurins (a protein phosphatase) that is known to perform poorly on other proteins (Romero et al. 2001).

Figure 1B compares the MCC values for non-viral and viral protein data sets. Predictors exhibited substantially less variation in MCC for the non-viral data—put another way, the majority of predictors were more accurate at predicting disorder in viral proteins. The entire set of MCC, AUC, sensitivity, and specificity values for both data sets are summarized in Supplementary Tables S3 and S4. To examine potential differences among predictors in greater detail, we calculated the RMSE for each protein and predictor and used a principal components analysis to visualize the resulting matrix (Fig. 2). The PCA indicated that the different predictors did not exhibit markedly divergent error profiles at the level of entire proteins. However, an SVM classifier trained on a random half of these data obtained, on average, an AUC of 0.75 ($n = 100$, range = 0.65–0.83), indicating that the viral and non-viral protein labels were appreciably separable with respect to these RMSE values.

### 3.3 Ensemble prediction

Ensemble classifiers are expected to perform better than their constituent models because they can reduce overfitting of the data by the latter (Attia 2012). Although multiple predictive models of protein disorder employ an ensemble approach, none of them has been trained specifically on viral protein data. We trained a random forest classifier on the outputs of the predictors used in our study using ten random training subsets of the viral protein data. Next, we validated the performance of this ensemble model in comparison to these individual predictors to determine if training on viral data conferred a significant advantage. We found that the ensemble classifier performed substantially better, with a mean MCC of 0.72 (range 0.62–0.86). This corresponded to a roughly 27 per cent improvement relative to ESpritz.Disorder, the best performing disorder predictor on these data (Fig. 1).

To examine the relative contribution of the different predictors used as inputs for the ensemble method, we evaluated the feature importance of each input (Fig. 3)—roughly the

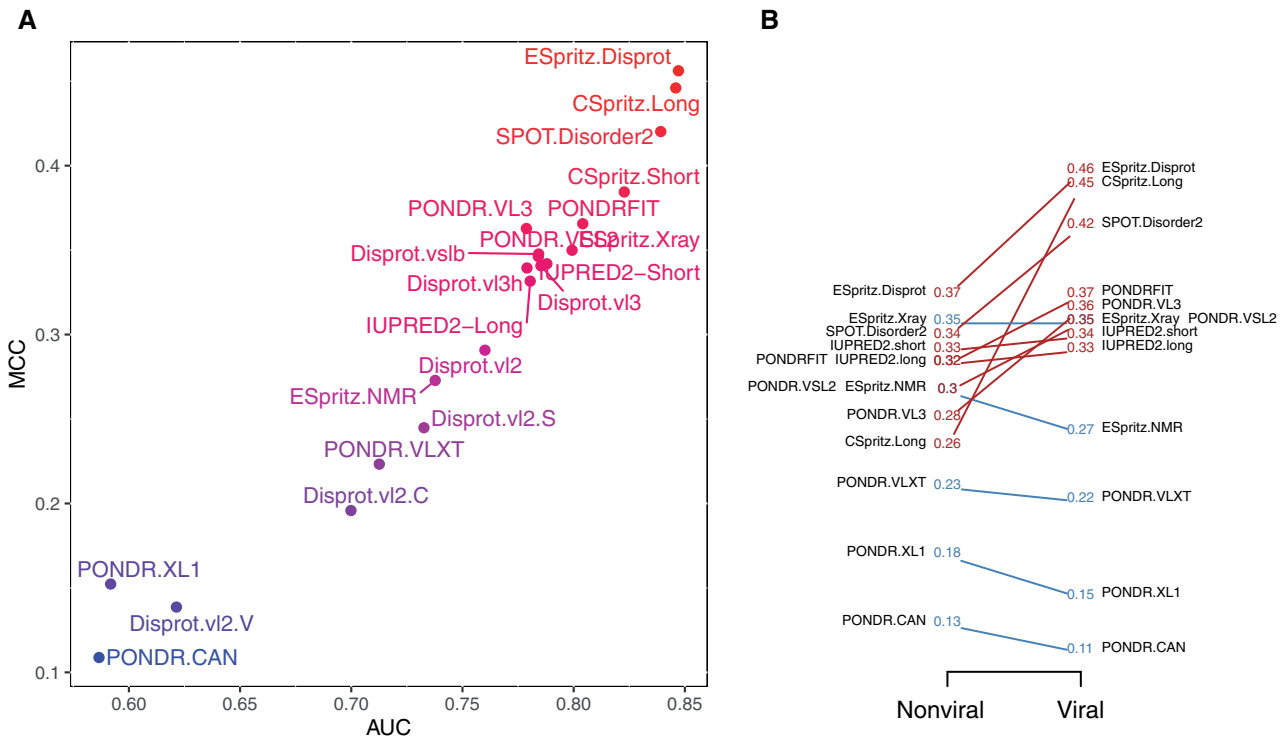


**Figure 1.** Performance of predictors on viral data set. (A) Scatterplot of MCC and AUC values for 21 predictors applied to the viral protein data set. (B) Slopegraph comparing the MCC values for 13 predictors applied to both non-viral and viral data sets. Because the three variants of the ESpritz model obtained identical MCC values, the corresponding labels were merged. Two labels (PONDRFIT, PONDR.VSL2) were displaced to prevent overlaps on the left and right sides, respectively.

prevalence of that feature among the decision trees comprising the random forest. We observed that the individual accuracy of a predictor did not necessarily correspond to its feature importance. Specifically, the best predictors (ESpritz.Disprot, CSpritz.Long, and SPOT-Disorder.2) tended to be assigned higher importance values. On the other hand, both Disprot-VL2.C and Disprot-VL2.V also displayed high importance despite having some of the worst accuracy measures when evaluated individually (Fig. 1).

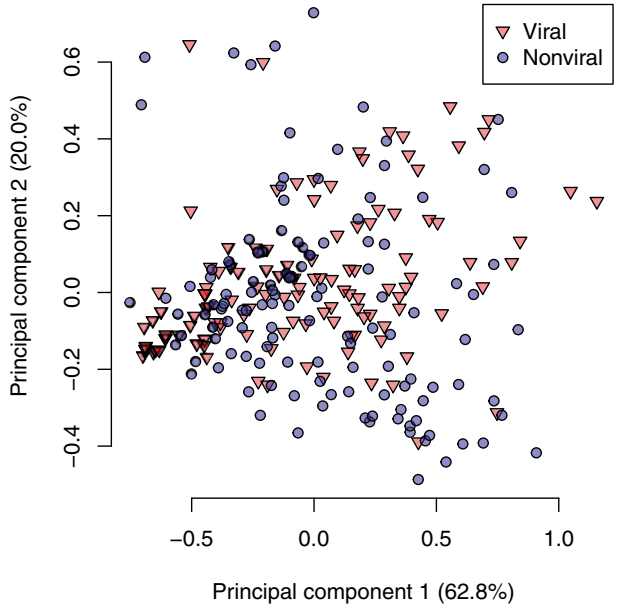


**Figure 2.** Principal components analysis plot of the root mean squared errors (RMSEs) for 13 disorder predictors on viral (red, triangles) and non-viral (blue, circles) protein sequences. The percentages of total variance explained by the first two principal components are indicated in parentheses in the respective axis labels.

### 3.4 Example: SARS-CoV-2 accessory protein 6

To illustrate the use of our ensemble model on a novel protein, we applied this model and the twenty-one individual predictors to the accessory protein encoded by ORF6 in the novel 2019 coronavirus that was first isolated in Wuhan, China (designated severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)). ORF6 is one of the eight accessory genes of this virus. Its protein product is involved in antagonizing interferon activity thereby suppressing host immune response (Yuen et al. 2020). The protein is predicted to be highly disordered, particularly in its C-terminal region that contains short linear motifs involved in numerous biological activities (Giri et al. 2020). We used a heatmap (Fig. 4) to visually summarize results from the ensemble method and individual predictors, mapped to the ORF6 amino acid sequence. Overall, most predictors assigned a higher probability of disorder in the C-terminal region of the protein, with the conspicuous exception of PONDR-XL1 and PONDR-CAN, which did not predict any disordered residues in this region. We also observed considerable variation among predictors around this overall trend. Although the PONDR-XL1 predictor is documented to omit the first and last fifteen residues from disorder predictions, we observed that only fourteen residues were reported this way—this treatment was also obtained for PONDR-CAN, although it was not a documented behavior of that predictor.

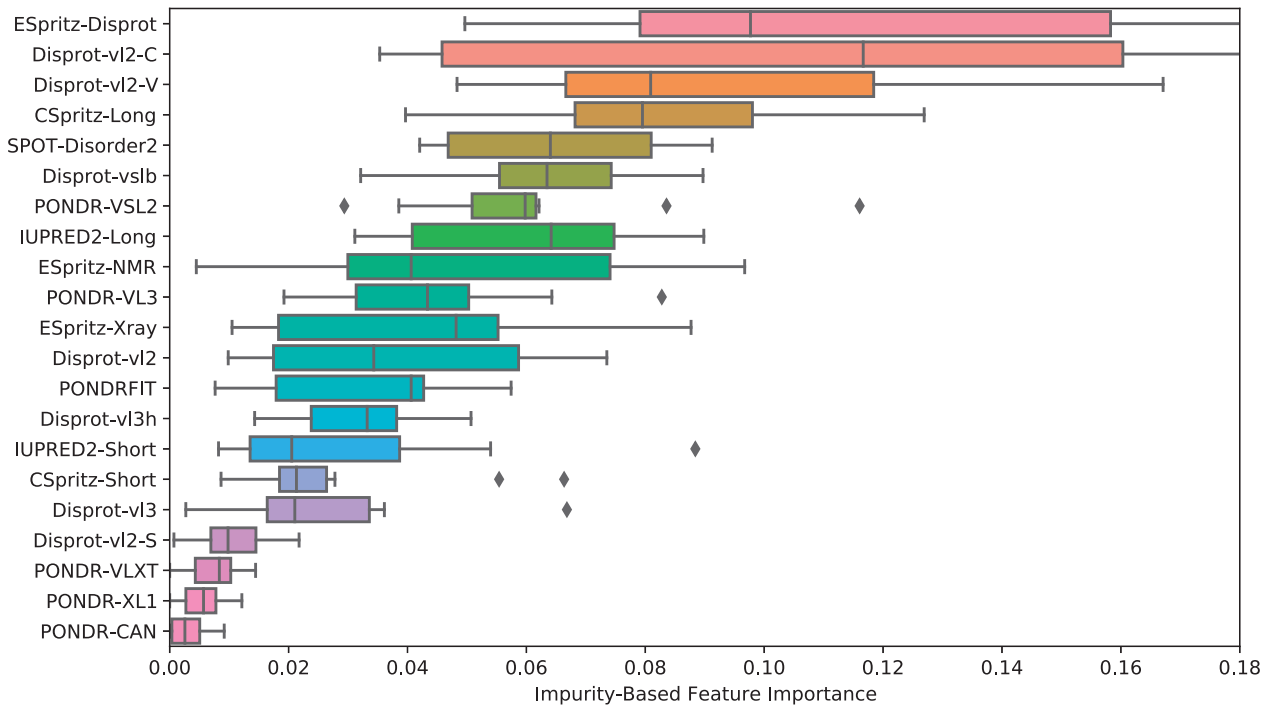


**Figure 3.** Box plot of the average decrease in Gini impurity by each feature in the random forest, for 10 random runs of the random forest model. Vertical line indicates the median, the box is the interquartile range (IQR; range from first to third quartiles). The left whisker extends to the first datum greater that Q1 − 1.5 × IQR and the right whisker extends to the last datum smaller than Q3 + 1.5 × IQR. Individual points are outliers that lie outside this range.

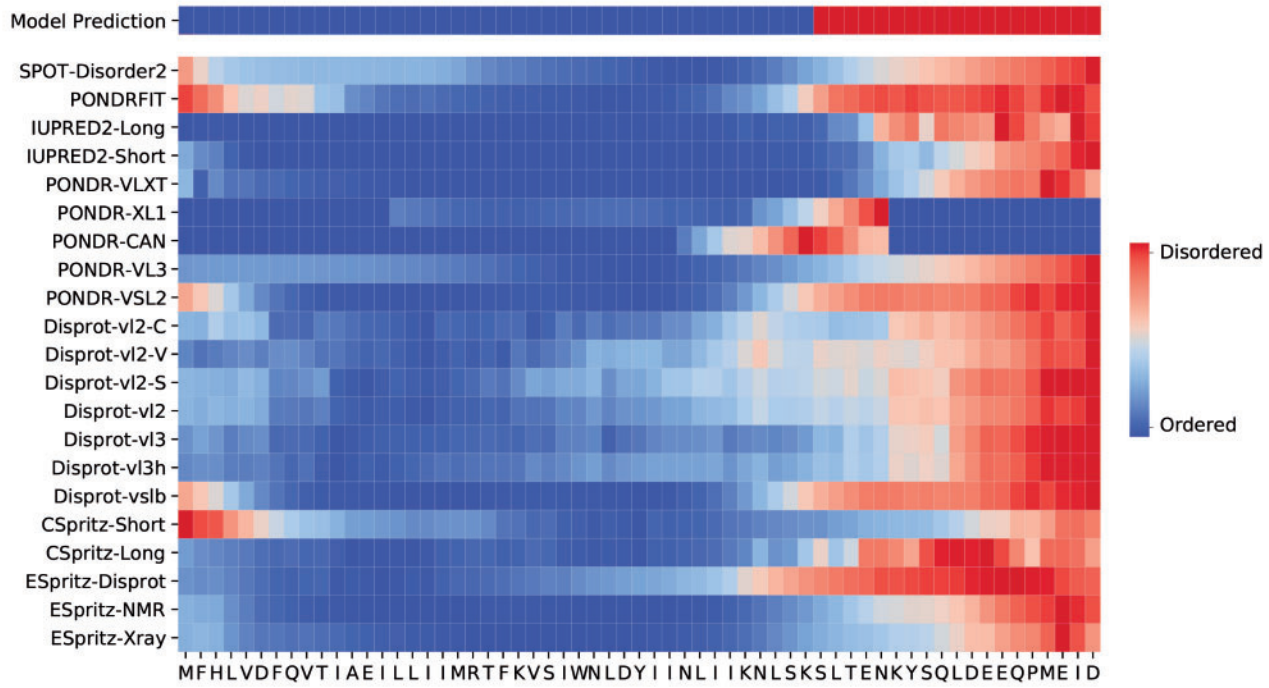


**Figure 4.** Disorder predictions for novel ORF6 in SARS-CoV-2. The first row represents the random forest model predictions, with subsequent rows corresponding to individual predictors. The entire protein length is represented on the x-axis, each grid is an amino acid. Red squares indicate disordered predictions and blue squares indicate ordered predictions.

### 3.5 Concluding remarks

IDPRs play an essential role in many viral functions (Mishra et al. 2020). It is therefore important to predict these regions accurately in order to make biological inferences from sequence variation. In this study, we found that disorder is as prevalent in virus proteins as non-virus proteins, and that predictive models of intrinsic disorder exhibit different biases when evaluated on viral versus non-viral proteins. Moreover, we show that the inherent variation among different predictors can yield discordant results when applied to the same virus protein sequence, and that this variation can be mitigated using an ensemble learning approach.

Though our results suggest that an ensemble method can yield more accurate predictions of intrinsic disorder—or at least, predictions that were more concordant with an expert-curated database of intrinsic protein disorder (Hatos et al. 2020)—we note that many of these predictors could only be accessed through web applications. Requiring access to a number of online resources that are not always available (due, for example, to a local network outage) presents a significant obstacle to the practical utility of an ensemble learning approach. Hence, we encourage researchers in the field of disorder prediction to support open science by releasing their source code or compiled binaries for local execution.

## Data availability

All data and scripts are released under a permissive free license (MIT) at https://github.com/PoonLab/tuning-disorder-virus.

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Funding

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2018-05516) and from the Canadian Institutes of Health Research (CIHR, PJT-155990).

**Conflict of interest:** None declared.

## References

Atkins, J. D. et al. (2015) 'Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies', *International Journal of Molecular Sciences*, 16: 19040–54.

Attia, A. (2012) 'Ensemble Prediction of Intrinsically Disordered Regions in Proteins', *BMC Bioinformatics*, 13: 111.

Belshaw, R., Pybus, O. G., and Rambaut, A. (2007) 'The Evolution of Genome Compression and Genomic Novelty in RNA Viruses', *Genome Research*, 17: 1496–504.

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017) 'Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric', *PLoS One*, 12: e0177678.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45: 5–32.

Chang, C.-C., and Lin, C.-J. (2011) 'Libsvm: A Library for Support Vector Machines', *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2: 1–27.

Chawla, N. V. et al. (2002) 'SMOTE: Synthetic Minority over-Sampling Technique', *Journal of Artificial Intelligence Research*, 16: 321–57.

ChromeDriver: WebDriver for Chrome (2000). https://chromedriver.chromium.org/ accessed 20 Oct 2020.

Cotmore, S. F. et al. (2005). Structure and organization of the viral genome. *Parvoviruses*. London, UK: Hodder Arnold, pp. 73–94.

DeForte, S., and Uversky, V. N. (2016) 'Resolving the Ambiguity: Making Sense of Intrinsic Disorder When PDB Structures Disagree', *Protein Science*, 25: 676–88.

DisProt (2000). https://www.disprot.org/ accessed 27 Apr 2020.

Dunker, A. K. et al. (2001) 'Intrinsically Disordered Protein', *Journal of Molecular Graphics and Modelling*, 19: 26–59.

Ferreon, A. C. M. et al. (2010). 'Chapter 10 - Single-Molecule Fluorescence Studies of Intrinsically Disordered Proteins', in Walter, NG (ed.) *Methods in Enzymology*, Vol. 472, pp. 179–204. Academic Press.

Giri, R. et al. (2020). Dark proteome of newly emerged sars-cov-2 in comparison with human and bat coronaviruses. *bioRxiv*.

Gitlin, L. et al. (2014) 'Rapid Evolution of Virus Sequences in Intrinsically Disordered Protein Regions', *PLoS Pathogens*, 10: e1004529.

Hanson, J. et al. (2019) 'Spot-disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning', *Genomics, Proteomics & Bioinformatics*, 17: 645–56.

—— et al. (2017) 'Improving Protein Disorder Prediction by Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks', *Bioinformatics (Oxford, England)*, 33: 685–92.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.

Hatos, A. et al. (2020) 'Disprot: Intrinsic Protein Disorder Annotation in 2020', *Nucleic Acids Research*, 48: D269–76.

Hemmerich, J., Asilar, E., and Ecker, G. F. (2020) 'Cover: Conformational Oversampling as Data Augmentation for Molecules', *Journal of Cheminformatics*, 12: 1–12.

Holmes, E. C. (2009) 'The Evolutionary Genetics of Emerging Viruses', *Annual Review of Ecology, Evolution, and Systematics*, 40: 353–72.

Hu, G. et al. (2016) 'Untapped Potential of Disordered Proteins in Current Druggable Human Proteome', *Current Drug Targets*, 17: 1198–205.

Japkowicz, N. (2000) 'The Class Imbalance Problem: Significance and Strategies', in *Proc. of the Int'l Conf. on Artificial Intelligence*, Vol. 56.

Katuwawala, A., Oldfield, C., and Kurgan, L. (2019) 'Accuracy of Protein-Level Disorder Predictions', *Briefings in Bioinformatics*, 46: 48.

Kozlowski, L. P., and Bujnicki, J. M. (2012) 'Metadisorder: A Meta-Server for the Prediction of Intrinsic Disorder in Proteins', *BMC Bioinformatics*, 13: 111.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017) 'Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning', *The Journal of Machine Learning Research*, 18: 559–63.

Li, J. et al. (2015) 'An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014', *International Journal of Molecular Sciences*, 16: 23446–62.

Lichtenthaler, F. W. (1995) '100 Years "Schlüssel-Schloss-Prinzip": What Made Emil Fischer Use This Analogy?', *Angewandte Chemie International Edition in English*, 33: 2364–74.

Liu, Y., Wang, X., and Liu, B. (2019) 'A Comprehensive Review and Comparison of Existing Computational Methods for Intrinsically Disordered Protein and Region Prediction', *Briefings in Bioinformatics*, 20: 330–46.

Mészáros, B., Erdős, G., and Dosztányi, Z. (2018) 'Iupred2a: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding', *Nucleic Acids Research*, 46: W329–W337.

Mishra, P. M. et al. (2020) 'Intrinsically Disordered Proteins of Viruses: Involvement in the Mechanism of Cell Regulation and Pathogenesis', *Progress in Molecular Biology and Translational Science*. Dancing Protein Clouds: Intrinsically Disordered Proteins in Health and Disease, Part B, 174: 1–78.

Necci, M. et al. (2018) 'A Comprehensive Assessment of Long Intrinsic Protein Disorder from the Disprot Database', *Bioinformatics*, 34: 445–52.

Nielsen, J. T., and Mulder, F. A. (2019) 'Quality and Bias of Protein Disorder Predictors', *Scientific Reports*, 9: 5137.

Pedregosa, F. et al. (2011) 'Scikit-Learn: Machine Learning in Python', *The Journal of Machine Learning Research*, 12: 2825–30.

Peng, K. et al. (2006) 'Length-Dependent Prediction of Protein Intrinsic Disorder', *BMC Bioinformatics*, 7: 208.

Piovesan, D. et al. (2017) 'Disprot 7.0: A Major Update of the Database of Disordered Proteins', *Nucleic Acids Research*, 45: D219–27.

Rancurel, C. et al. (2009) 'Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into de Novo Protein Creation', *Journal of Virology*, 83: 10719–36.

Romero, P. et al. (2001) 'Sequence Complexity of Disordered Protein', *Proteins: Structure, Function, and Genetics*, 42: 38–48.

Sanjuán, R. et al. (2010) 'Viral Mutation Rates', *Journal of Virology*, 84: 9733–48.

Tokuriki, N. et al. (2009) 'Do Viral Proteins Possess Unique Biophysical Features?', *Trends in Biochemical Sciences*, 34: 53–9.

Uversky, V. N. (2002) 'What Does It Mean to Be Natively Unfolded? ', *European Journal of Biochemistry*, 269: 2–12.

—— (2019) 'Intrinsically Disordered Proteins and Their 'Mysterious' (Meta) Physics', *Frontiers in Physics*, 7: 10.

——, Oldfield, C. J., and Dunker, A. K. (2008) 'Intrinsically Disordered Proteins in Human Diseases: Introducing the d2 Concept', *Annual Review of Biophysics*, 37: 215–46.

Vucetic, S. et al. (2003) 'Flavors of Protein Disorder', *Proteins: Structure, Function, and Bioinformatics*, 52: 573–84.

Walsh, I. et al. (2011) 'Cspritz: Accurate Prediction of Protein Disorder Segments with Annotation for Homology, Secondary Structure and Linear Motifs', *Nucleic Acids Research*, 39: W190–6.

—— et al. (2012) 'Espritz: Accurate and Fast Prediction of Protein Disorder', *Bioinformatics*, 28: 503–9.

Walter, J. et al. (2019) 'Comparative Analysis of Mutational Robustness of the Intrinsically Disordered Viral Protein VPg and of Its Interactor eIF4E', *PloS One*, 14: e0211725.

Wright, P. E., and Dyson, H. J. (1999) 'Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm', *Journal of Molecular Biology*, 293: 321–31.

Xue, B. et al. (2010) 'PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids', *Biochimica et Biophysica Acta*, 1804: 996–1010.

—— et al. (2014) 'Structural Disorder in Viral Proteins', *Chemical Reviews*, 114: 6880–911.

Yuen, C.-K. et al. (2020) 'SARS-COV-2 nsp13, nsp14, nsp15 and orf6 Function as Potent Interferon Antagonists', *Emerging Microbes & Infections*, 9: 1418–29.