doi: 10.1093/nargab/lqaa014

Family-specific analysis of variant pathogenicity prediction tools

Jan Zaucha[1], Michael Heinzinger[2], Svetlana Tarnovskaya[3], Burkhard Rost[2] and Dmitrij Frishman[1,*]

[1]Department of Bioinformatics, Technical University of Munich, 85354 Freising, Germany, [2]Department of Informatics, Bioinformatics & Computational Biology—i12, Technical University of Munich, 85748 Garching, Germany and [3]Almazov National Medical Research Centre, St. Petersburg 197341, Russia

Received December 19, 2019; Revised February 12, 2020; Editorial Decision February 15, 2020; Accepted February 25, 2020

ABSTRACT

Using the presently available datasets of annotated missense variants, we ran a protein family-specific benchmarking of tools for predicting the pathogenicity of single amino acid variants. We find that despite the high overall accuracy of all tested methods, each tool has its Achilles heel, i.e. protein families in which its predictions prove unreliable (expected accuracy does not exceed 51% in any method). As a proof of principle, we show that choosing the optimal tool and pathogenicity threshold at a protein family-individual level allows obtaining reliable predictions in all Pfam domains (accuracy no less than 68%). A functional analysis of the sets of protein domains annotated exclusively by neutral or pathogenic mutations indicates that specific protein functions can be associated with a high or low sensitivity to mutations, respectively. The highly sensitive sets of protein domains are involved in the regulation of transcription and DNA sequence-specific transcription factor binding, while the domains that do not result in disease when mutated are responsible for mediating immune and stress responses. These results suggest that future predictors of pathogenicity and especially variant prioritization tools may benefit from considering functional annotation.

INTRODUCTION

The quest for automated annotation of single amino acid variants (SAVs) has led to the development of numerous methods for predicting the deleteriousness of missense mutations. All of these methods rely on supervised machine learning models, trained on a collection of manually annotated variants to provide a probability of pathogenicity for each queried mutant protein sequence. Internally, the algorithms employ a combination of different modalities of proteins, i.e. sequence, structure or meta-based features. Typically, the most informative feature is a metric of evolutionary conservation with more conserved sites being more sensitive to mutations. Although their accuracy is constantly increasing, with state-of-the-art models already achieving area under the curve values (AUC) of over 0.9, many of the tools exhibit low-to-moderate agreement (Pearson correlation <0.7) between the binary predictions they yield (1). What is more, it has recently become apparent that some methods greatly overpredict the pathogenicity of variants, assigning a disease-causing effect to even up to a third of benign variants (2).

MetaSVM (1), which combines the output of multiple individual methods to arrive at an aggregate prediction, leveraged this to yield particularly high performance in benchmarks (3). However, our recent analysis revealed that in select proteins its predictions still prove rather unreliable. For example, in the cardiac sodium channel, MetaSVM greatly exaggerates the pathogenicity of mutations, predicting a deleterious effect in 75% of SAVs that are, in fact, annotated as neutral (4). One line of reasoning that could explain such results is that this protein, although conserved in evolution (as captured by most models predicting a deleterious effect for the majority of mutations), does not always result in disease upon mutation—the observed evolutionary conservation may have resulted from the reduced fitness that the nowadays-benign mutations could have been bringing to affected hosts when competing for survival in the wild. In fact, the first tool that attempted to account for the particular context of the mutation in question was the weighted version of the FATHHM method, which was tuned to yield the best accuracy in a species-specific manner (5). Such context-specific prediction methods can be taken a step further in order to account for all the characteristics of each individual protein family. This can be accomplished, provided a considerable number of curated SAVs are already available. In the case when a method clearly over- or underestimates the predicted effects of mutations in a certain protein family, the most straightforward approach for obtaining more

[*]To whom correspondence should be addressed. Tel: +49 8161 712134; Fax: +49 8161 712186; Email: d.frishman@wzw.tum.de

reliable results is shifting the pathogenicity threshold. Nevertheless, in the case of the cardiac sodium channel, even after increasing the pathogenicity threshold of MetaSVM, the expected accuracy of predicting the effects of mutations in this protein remains low—at roughly 55%.

Therefore, we hypothesized that different prediction methods may differ in their reliability, depending on the family of the queried protein. One trivial reason for this is that each method is biased towards the dataset used for training. Another reason is that different input features capture the various functional effects, which a mutation can have to a different degree. For example, since FATHMM (5) relies on amino acid residue transition probabilities encoded within hidden Markov model representations of each protein domain family, it should be expected to be particularly reliable for predicting cases of missense variants that disrupt the structure (and thus function) of any protein belonging to that family. However, it may be less accurate in cases where a mutation affects a protein from a family that comprises many domains specializing in specific tasks, each one requiring a different residue side chain at the same position in order to fulfil its particular role. In each individual protein by itself, strict conservation of that crucial residue is to be expected. However, at the family level, all of the different residues are equally probable within the corresponding HMM match state. In the latter example, methods capturing the evolutionary conservation at specific genomic positions [such as CADD (6)] should be expected to yield more accurate predictions.

Altogether, this reasoning points to the idea that predictions of pathogenicity could be tuned with respect to the specific characteristics of each protein. The major caveat, for the time being, is that such endeavour would require a large number of manually annotated SAVs already available for each protein. Nevertheless, owing to the recurrence of protein domains within the genome, annotated mutations occurring within the same protein family can be pooled together to determine which tools and parameters are optimal for predicting the pathogenicity of other SAVs in those families. As proof of concept, we apply this idea on a subset of Pfam domain families, for which a sufficient number of annotated SAVs are available. We show that such fine-tuning could, in principle, allow obtaining significantly more reliable predictions of pathogenicity than standard off-the-shelf methods. Although predicting which method and threshold of pathogenicity should be used in families lacking enough curated SAVs proved difficult, based on a functional enrichment analysis of families annotated exclusively by benign or pathogenic SAVs, we provide indications suggesting that future predictors of pathogenicity might benefit from utilizing annotations of protein function. Finally, a further discussion of the results provokes reconsidering the definition of a neutral mutation.

MATERIALS AND METHODS

Annotation of mutations

Mutations with precalculated predictions of deleteriousness from popular methods published to date were extracted from the dbNSFP database (1); the predictions from the SNAP2 method were calculated on our server. The pathogenicity scores selected for analysis were chosen to be the raw outputs of the respective methods (indicated as ‘raw/score’ within dbNSFP). Where available, annotations specialized for predicting the effects of protein coding variants were selected (indicated as ‘coding’ within dbNSFP). Methods for which precalculated predictions were unavailable for all variants were excluded from the analysis. In the case when multiple versions of a method were published as multiple separate scores, each one was included in the analysis. Altogether, 26 different scores developed across 23 studies were considered in the benchmarking (Supplementary File 4).

dbNSFP provides annotations of pathogenicity primarily from ClinVar, accessed on 22 July 2019 (7); we additionally added variant annotations from SwissVar (8) and UniProt (9). Since there is a surplus of variants annotated to be pathogenic, in order to increase the breadth of the analysis, we included an additional list of mutations that are most likely neutral. Following Bendl *et al*. (10), we added mutations from the VariSNP (11) database. Briefly, neutral variants extracted from dbSNP (12) were filtered to exclude all pathogenic mutations found in ClinVar, SwissProt and PhenCode (13) as well as any SNPs that occurred in cancer [COSMIC (14)] or the NHGRI GWAS catalogue (15).

The Pfam database (16) was used to map mutations in proteins to their respective domain families (mapping was achieved via UniProt accession and position of the mutation in the respective protein). There are 6512 distinct Pfam families within the human proteome, covering 71% of all protein sequences at 45% residue coverage (16). A total of 88 687 labelled mutations were mapped to 3422 Pfam domains (at least one mutation), 63 398 of which were annotated as pathogenic (or ‘likely pathogenic’, as annotated in UniProt), while the remaining 25 289 variants were regarded as likely benign. In order to ensure a high quality of the comparisons, families with <10 curated mutations in the two classes, i.e. ‘pathogenic’ or ‘neutral’, were excluded from the benchmarking steps. Three hundred fifty-two families have at least 10 annotated mutations in each class, giving a total of 14 916 neutral and 42 161 pathogenic SAVs. Forty-nine families have been annotated exclusively with neutral mutations, while 106 families feature only deleterious variants (Supplementary Files 1 and 2). In order to achieve higher data coverage, SAVs could be mapped to the clan level of the Pfam hierarchy. However, since some clans are very prominent, the specificity of each protein family would be lost.

Annotation of Pfam domain families

For each protein family, the corresponding alignment based on representative proteomes (clustered at sequence identity of 75%) was extracted from the Pfam database. The alignment depth provides a measure of prevalence across reference proteomes. The number of effective sequences was obtained by redundancy clustering at 80% sequence identity using CD-HIT (17). JPred4 (18) was used to obtain known (with the ‘pdb’ option set to true) or predicted secondary structure. A mapping of Pfam domains to PDB (19) structures was obtained from the Pfam database (available for 294 of the 352 Pfam families included in the benchmarking analysis) (20). For predicting, which prediction tool will offer the highest accuracy in the given protein family, the following characteristic features were extracted from the multiple sequence alignment corresponding to each Pfam: alignment depth (number of proteins in the alignment), domain length and mean weighted Shannon entropy (calculated using Mstat-X).

Apart from the sequence-based characteristics, we considered structure-specific data, including fraction of sequence annotated with secondary structure elements, fraction of residues forming helices and strands, and features related to specific inter-residue contacts extracted from the PDB files. First, contact density was taken as the total number of contacts normalized by domain length. Second, the maximum inter-residue contact connectivity was calculated from the residue connectivity graph [the igraph library was used for this task (21)]. This feature describes the minimum number of vertices that have to be removed from the graph to eliminate all paths between a pair of nodes; it was shown to correlate with designability (22). Third, relative contact order is a measure of the locality of the inter-amino acid contacts in the protein's native state tertiary structure; it describes the protein's compactness and is predictive of folding rates (23).

Relative contact order is calculated as

$$\mathrm{CO} = \frac{1}{L \cdot N} \sum^{N} \Delta S_{ij},$$

where L is the protein length, N is the total number of contacts and ΔS_{ij} is the separation (in sequence space) between residues i and j, which are in contact. For the calculation of the above features, a pair of residues was regarded to be in contact if the physical distance between the corresponding alpha carbon atoms of the residues was no more than 6 Å; local contacts between residues up to five positions apart in sequence space were disregarded. The final feature considered was the network degree (number of interacting partners) within the domain–domain interaction network [obtained from the 3did database (24)]. The distributions of features grouped according to top-performing prediction methods (limited to the top 8 methods, each of which performed the best in at least 15 families) were tested for statistical significance using the non-parametric Kruskal–Wallis *H*-test. The distributions of features attributed to families annotated exclusively with neutral or exclusively pathogenic mutations were tested to detect statistically significant differences using the non-parametric Mann–Whitney *U*-test.

Benchmarking the performance of methods

For each Pfam and each prediction tool, a univariate logistic regression model with balanced class weights (samples were weighted inversely proportional to class frequencies in the input data) was fitted. The decision boundaries (corresponding to the threshold of pathogenicity) were extracted directly from the parameters of the fitted model (negative ratio of the fitted *y*-intercept and the corresponding regression coefficient). The accuracy (sum of true positive and true negative predictions divided by the total number of predictions) of each prediction tool on the given Pfam was taken as the mean accuracy obtained over 100 randomly chosen balanced (equal number of pathogenic and neutral mutations) subsamples; the standard deviation serves as a confidence measure for the reported value. Spearman's correlation coefficients were calculated between the accuracy achieved and each family's characteristics.

RESULTS

The top-performing methods differ across families

In protein families [Pfam domains (16)] that were annotated with at least 10 mutations of each class (neutral or pathogenic), we investigated which methods provide the highest accuracy of predictions. The performance of tools across Pfam families is summarized in Table 1 and Figure 1; detailed results for all families analysed independently are available for inspection in Supplementary File 3.

We found that different methods achieve different levels of accuracy in predicting the deleteriousness of mutations, depending on the particular protein family that is analysed. Most importantly, however, no single method can be regarded as reliable in all individual families. CADD (6) proved to be the top method in most individual Pfam families (77 out of 352, 22%) but on average, its accuracy (84%) was only the second best. Moreover, its performance was not satisfactory across all families analysed; for example, it achieved only 34% accuracy on the RING (Really Interesting New Gene) finger, a type of zinc finger domain, which is among the most prevalent domain families across reference proteomes. The tool that provided the best accuracy overall (85%) is REVEL (25); moreover, most likely owing to the fact that it is an ensemble method leveraging strengths of 13 other tools, it yielded one of the highest accuracies in the worst case (50% in guanylate kinases) and it was never the worst performing prediction tool across all families analysed. A marginally better performance in the worst case (51% in the ThiF family) was achieved by Polyphen2-HVAR (26), but overall its average accuracy was considerably lower (76%). As a consequence, our results suggest that REVEL should be the tool of choice in families for which the best performing method is unknown. Nevertheless, in such cases, the accuracy cannot be guaranteed to exceed 50%, which still demonstrates room for improvement. On the other hand, methods that have been found to yield the worst performance in the highest number of individual families are GenoCanyon (27), LRT (28), MutationTaster (29), FATHMM-XF (coding) (30) and FATHMM (5). It should be noted that most of the aforementioned methods are among the tools that were developed the earliest. In this regard, we detected a moderate correlation between prediction accuracy and the tool's publication date (Spearman's $\rho = 0.49$, $P = 0.01$) suggesting that incremental improvements are constantly being made (Figure 2). However, it must be noted that the recently developed methods have the advantage of being trained on larger and newer datasets, which undoubtedly feature a greater overlap with the benchmark set used in this study. Therefore, their superior performance may be partly attributed to this bias.

The key observation that we would like to point out is that across all methods the worst accuracy was achieved in different Pfam families (with the exception of guanylate kinases, which proved to be the hardest case for three tools).

Table 1. Performance of methods across families

	Number of Pfams				
Method	With best accuracy	With worst accuracy	Mean accuracy	Worst accuracy	Hardest Pfam
CADD	**77**	2	0.84	0.34	PF00097
REVEL	66	**0**	**0.85**	0.50	PF00625
VEST4	54	1	0.83	0.48	PF18199
PrimateAI	29	7	0.78	0.35	PF04732
M-CAP	19	2	0.79	0.34	PF12031
Eigen-PC	15	5	0.78	0.41	PF00625
MetaLR	15	1	0.81	0.45	PF01044
MetaSVM	15	1	0.82	0.46	PF04558
SNAP2	13	2	0.76	0.45	PF04814
FATHMM-XF	8	44	0.66	0.30	PF00396
MVP	8	1	0.77	0.47	PF00531
DEOGEN2	7	3	0.76	0.46	PF01624
Eigen	6	1	0.80	0.43	PF00625
PROVEAN	5	5	0.75	0.41	PF00230
DANN	3	14	0.67	0.40	PF00858
FATHMM-MKL	2	11	0.68	0.40	PF04814
FATHMM	2	30	0.67	0.39	PF00364
MutationAssessor	2	**0**	0.77	0.50	PF01344
LRT_Omega	2	16	0.67	0.27	PF08645
Polyphen2_HVAR	1	**0**	0.76	**0.51**	PF00899
Polyphen2_HDIV	1	**0**	0.74	0.50	PF00622
SIFT4G	1	3	0.72	0.42	PF11577
SIFT	1	2	0.70	0.40	PF00003
GenoCanyon	0	79	0.59	0.39	PF04757
MutationTaster	0	58	0.59	0.43	PF15156
LRT	0	64	0.60	0.33	PF00616

The columns include the method name (from dbNSFP), number of families in which the method achieved the best accuracy, number of families in which the method achieved the worst accuracy, mean accuracy across all families, worst accuracy in a specific Pfam and the hardest Pfam for this method (corresponding to the lowest accuracy achieved).

In fact, when using the best performing method in each family, the accuracy of predictions is universally high—the mean accuracy is 92% and the lowest accuracy score in any family is 68% (for the glycoside hydrolase family 22). Figure 3 shows the performance gains achieved by using the best method in each family, compared with the scores achieved when sticking to the single tool (CADD) that achieved the best performance in the highest number of families. Applying the best performing tool for the given family allows predicting the effects of mutations with at least 85% accuracy in 75% of families.

Functional enrichment of families annotated exclusively with benign or pathogenic variants

In an attempt to gain insight into the characteristics of protein families that are particularly sensitive, or conversely, robust to mutations, we performed a functional enrichment of the Pfam domains annotated exclusively by pathogenic and neutral mutations, respectively. A domain-centric functional enrichment (31) allows transferring functional annotation from the gene level [as provided by the Gene Ontology (32,33)] to protein domains.

The results obtained indicate that protein families, which are particularly sensitive to mutations, are involved in regulation of transcription, specifically DNA sequence-specific transcription factor binding and regulation of RNA polymerase II activity (Supplementary File 1). This suggests that enhancing biological complexity through fine regulation of transcription by DNA sequence-specific binding comes at the cost of becoming more prone to replication errors. On the other hand, the relatively robust Pfam domains are mainly responsible for immune or stress responses (Supplementary File 2). One explanation for this is that these proteins are very adaptable due to the constantly changing pressures exerted by external stimuli (e.g. pathogens). However, it should be noted that the majority of these functions are necessary for survival only in specific conditions (presence of external pathogens or stress stimuli) and may not need to be activated otherwise. Therefore, although mutations of these protein families do not immediately cause disease, it is questionable whether they do not reduce the host's fitness by impeding its ability to withstand environmental pressures in the event of their occurrence. Therefore, the available annotation may not always be a complete representation of the mutation's effect.

This points to the need for a better definition of what is meant by a neutral mutation. Two methods with unique approaches to the problem of predicting which variants are neutral, based on their definitions of neutrality, are PrimateAI (34) (aims to predict which variants are evolutionarily neutral by leveraging additional data on common variants from primates) and SNAP2 (35) (trained to predict whether a variant causes any functional effect, rather than disease itself). Based on the most commonly used definition of deleteriousness, in which a variant is deemed pathogenic only if it is directly implicated in causing disease, the latter method may be underperforming in benchmarks.

On top of this, we observed mild but statistically significant signals indicating that particularly sensitive protein

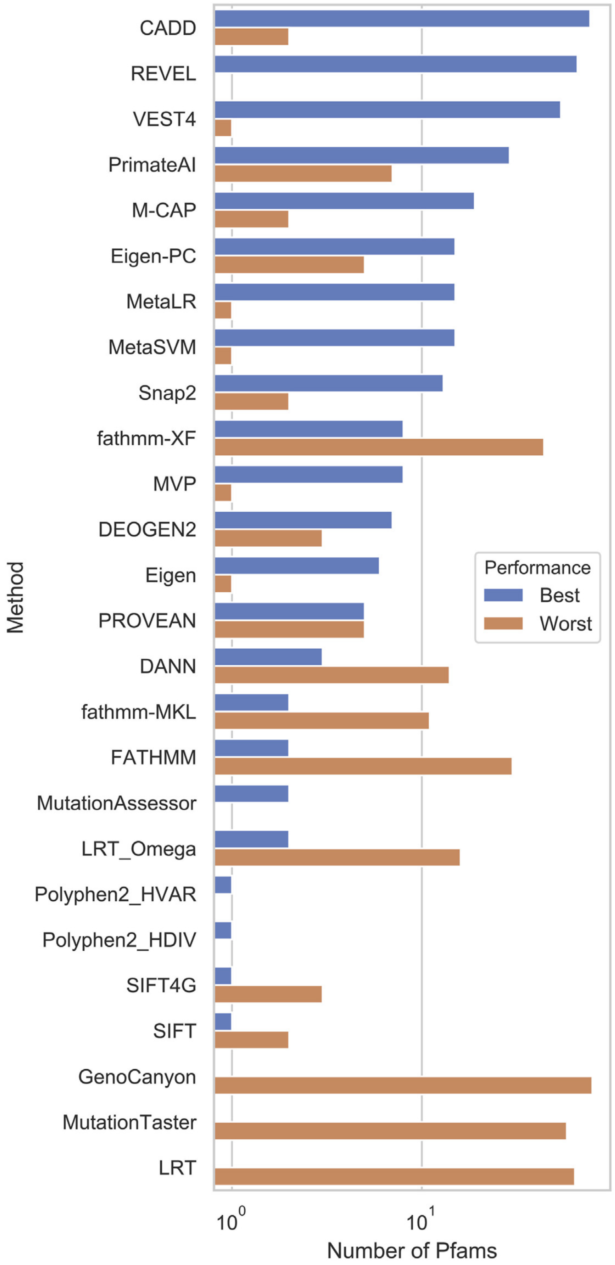

Figure 1. Overview of methods with regard to how many individual families they are optimal for (blue; optimal = yielding highest accuracy for all pooled mutations in a given Pfam) or suboptimal for (orange; suboptimal = yielding lowest accuracy) when predicting the pathogenicity of variants.

families have a higher degree in the domain–domain interaction network (Figure 4), while the mutation-insensitive families are more prevalent across reference proteomes (Figure 5). The former observation agrees with previous results indicating that deletions of proteins with a high protein–protein interaction network degree are more likely to be lethal (36). This finding is in line with the observation that mutation-sensitive families are enriched for transcription factor binding functions, which are known to have a high network degree (37)). The latter observation, on the other hand, may be a reflection of paralogue compensation (38) (copies of the domain encoded by other genes may compensate for the loss of a specific instance).

Can we predict which method is optimal for analysing mutations in a specific protein family?

Since the majority of families have not yet been annotated with curated variants, it is unknown which method is the most suitable for predicting the pathogenicity of their mutations. We sought to find universal features of protein families that may be informative for identifying the most accurate prediction tool. We investigated the following features: prevalence across reference proteomes (measured in terms of the total alignment depth), number of effective sequences in the alignment, domain length, fraction of secondary structure, fraction of helix, fraction of extended strands, mean entropy, maximum inter-residue contact connectivity [shown to correlate with designability (22)], contact order (predictive of folding rates) (23), and contact density. Unsatisfactorily, none of the features tested carry enough information that would allow predicting the best method for analysing the pathogenicity of mutations in a family-specific manner (Supplementary Figures S1–S11). Additionally, we tested a recently reported abstract embedding of protein sequences in the form of 1024-dimensional feature vectors (39). For this analysis, the embeddings for each Pfam were compressed to 16 dimensions by using an autoencoder, which was trained in a previous step compressing all protein sequences in UniRef50 (40). However, the unsupervised t-SNE clustering (41) did not show any indication of adding value towards our goal (Supplementary Figure S12). For the time being, our attempts at predicting the top-performing method in a family-specific manner remain futile; however, as more annotations of pathogenicity become available, it will be possible to analyse more Pfam domain families explicitly.

CADD tends to be more accurate in less prevalent protein families or ones that have fewer interaction partners

We also investigated whether any protein family features correlate with the achieved prediction accuracies across methods. In most cases, Spearman's correlation coefficient was close to zero, but we found two features exhibiting a weak negative correlation with the prediction accuracy of CADD: the tool yields a higher accuracy in protein families that are less prevalent across reference proteomes (indicated by a negative correlation between prediction accuracy and alignment depth: Spearman's $\rho = -0.2$; $P = 0.0006$; Supplementary Figure S13) and in those families that have fewer interaction partners (indicated by a negative correlation between prediction accuracy and number of domain interaction partners in the domain–domain interaction network: Spearman's $\rho = -0.22$; $P = 0.0001$; Supplementary Figure S14). Considering that CADD only relies on the characteristics of the specific genomic locus where the mutation occurred, this implies that its prediction of pathogenicity could potentially be improved by adding information extracted from other members of each protein's family.

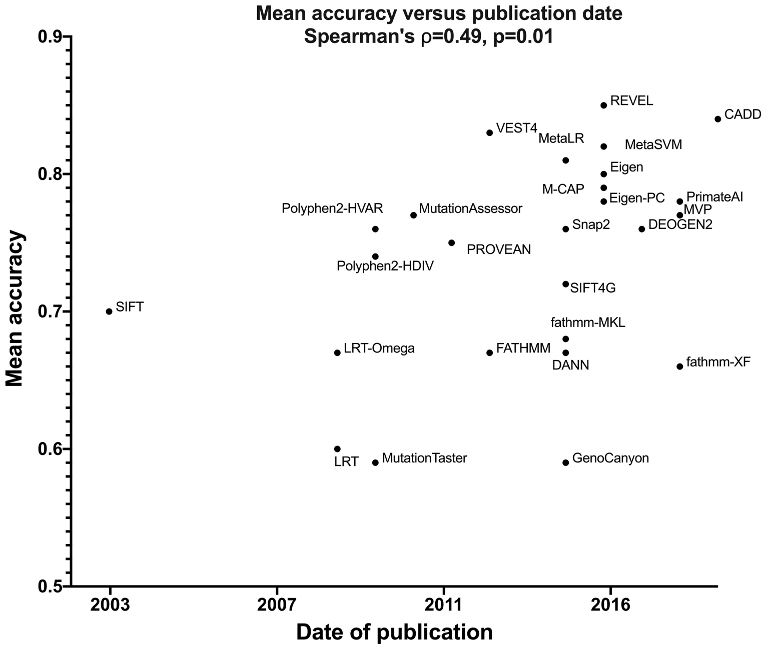

Figure 2. Mean prediction accuracy across families analysed versus publication date of the tool.

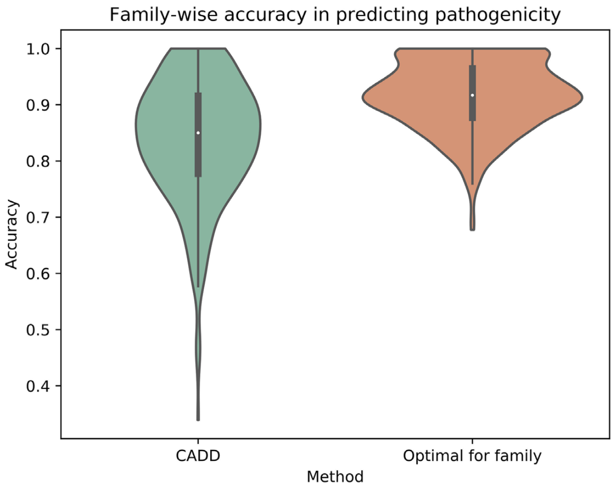

Figure 3. Accuracy for classifying neutral and pathogenic variants measured for each Pfam independently: as achieved by the overall top-performing method CADD (green) or the method achieving the highest accuracy for the given family (orange).

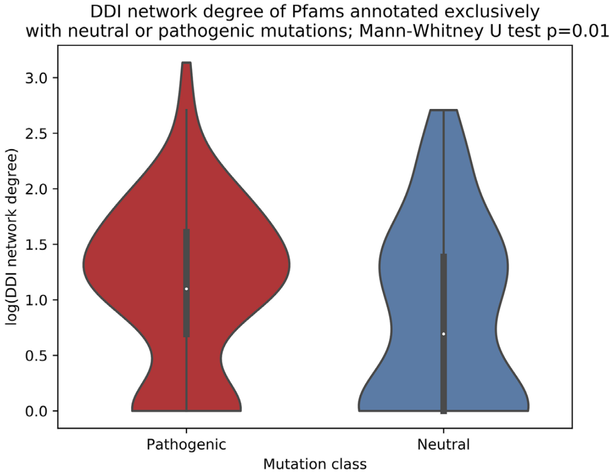

Figure 4. Distribution of domain–domain interaction network degrees in families annotated exclusively with pathogenic (red) or neutral (blue) mutations.

DISCUSSION

Altogether, we have shown that despite the very high overall accuracy of most tools for predicting the pathogenicity of mutations, virtually each method has its Achilles heel, i.e. the protein families for which its predictions are unreliable. We have also shown that thanks to the heterogeneity of the approaches, for all families analysed, there is at least one prediction tool that provides good levels of reliability. Nevertheless, predicting which method is optimal in each particular case proves difficult. This is additionally apparent from the fact that ensemble methods (included in the analysis) do not alleviate the problem of inaccurate predictions in specific families, for which their predictions are also unreliable. For the time being, we are unable to suggest a method for predicting which tool is optimal for predicting the pathogenicity of mutations in the families lacking enough already curated data. On average, the safest approach is to use REVEL, which was shown to achieve an accuracy of no less than 50% in any Pfam analysed.

The functional enrichment of families annotated exclusively by benign or pathogenic mutations indicates that functional annotation may be of value in predicting whether a mutation is implicated in disease. However, doing so will require caution since functional annotation is highly biased by the current state of knowledge on the disease. Notably, worth considering is the definition of neutrality used in the

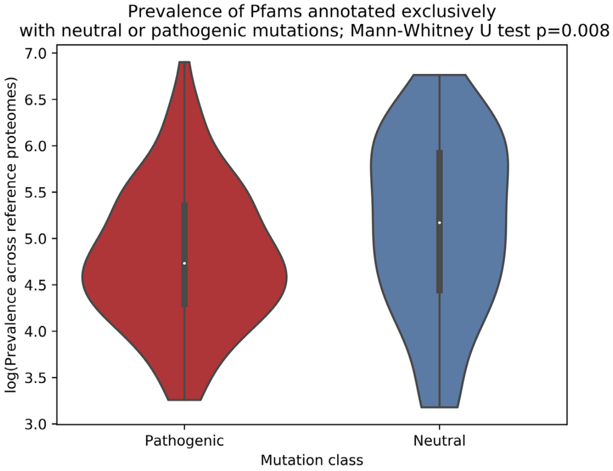

Figure 5. Distribution of prevalence across reference proteomes of families annotated exclusively with pathogenic (red) or neutral (blue) mutations.

annotation of variants, since mutations that do not immediately cause a disease may not necessarily lack any molecular effects (42).

The final remark regards the apparent superiority of the more recently developed methods. In part, it may be attributed to their higher levels of sophistication and design ingenuity. However, another important factor at play is the bias in the distribution of variants used in the training of each individual method (43). Due to the constantly expanding and evolving annotation of SAVs, the training sets used in the development of the newer methods should be expected to be more similar to the benchmark dataset used in this study. It would be interesting to see how much better (if at all) would the older methods fare following an update of their parameters to reflect the characteristics of the most current training set.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

Deutsche Forschungsgemeinschaft [FR 1411/14-1 to D.F., RO 1320/4-1 to B.R.]; Russian Science Foundation [17-15-01292 to S.T.].

Conflict of interest statement. None declared.

REFERENCES

1. Liu,X., Wu,C., Li,C. and Boerwinkle,E. (2016) dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, **37**, 235–241.
2. Niroula,A. and Vihinen,M. (2019) How good are pathogenicity predictors in detecting benign variants? *PLoS Comput. Biol.*, **15**, e1006481.
3. Anderson,D. and Lassmann,T. (2018) A phenotype centric benchmark of variant prioritisation tools. *npj Genomic Med.*, **3**, 5.
4. Tarnovskaya,S., Korkosh,V., Zhorov,B.S. and Frishman,D. (2019) Predicting variant pathogenicity in the cardiac sodium channel using paralogue annotation. *Biophys. J.*, **116**, 391a.
5. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L., Edwards,K.J., Day,I.N.M. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57–65.
6. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
7. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
8. Mottaz,A., David,F.P.A., Veuthey,A.-L. and Yip,Y.L. (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, **26**, 851–852.
9. Famiglietti,M.L., Estreicher,A., Gos,A., Bolleman,J., Géhant,S., Breuza,L., Bridge,A., Poux,S., Redaschi,N., Bougueleret,L. *et al.* (2014) Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.*, **35**, 927–935.
10. Bendl,J., Stourac,J., Salanda,O., Pavelka,A., Wieben,E.D., Zendulka,J., Brezovsky,J. and Damborsky,J. (2014) PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.*, **10**, e1003440.
11. Schaafsma,G.C.P. and Vihinen,M. (2015) VariSNP, a benchmark database for variations from dbSNP. *Hum. Mutat.*, **36**, 161–166.
12. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
13. Giardine,B., Riemer,C., Hefferon,T., Thomas,D., Hsu,F., Zielenski,J., Sang,Y., Elnitski,L., Cutting,G., Trumbower,H. *et al.* (2007) PhenCode: connecting ENCODE data with mutations and phenotype. *Hum. Mutat.*, **28**, 554–562.
14. Forbes,S.A., Beare,D., Bindal,N., Bamford,S., Ward,S., Cole,C.G., Jia,M., Kok,C., Boutselakis,H., De,T. *et al.* (2016) COSMIC: high-resolution cancer genetics using the catalogue of somatic mutations in cancer. In: *Current Protocols in Human Genetics*. Wiley, Hoboken, NJ, USA, pp. 10.11.1–10.11.37.
15. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.*, **42**, 1001–1006.
16. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
17. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
18. Drozdetskiy,A., Cole,C., Procter,J. and Barton,G.J. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.*, **43**, W389–W394.
19. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S. *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.
20. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
21. Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJ. Complex Syst.*, **1965**, 1–9.
22. Leelananda,S.P., Jernigan,R.L. and Kloczkowski,A. (2016) Predicting designability of small proteins from graph features of contact maps. *J. Comput. Biol.*, **23**, 400–411.
23. Plaxco,K.W., Simons,K.T. and Baker,D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.
24. Mosca,R., Céol,A., Stein,A., Olivella,R. and Aloy,P. (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **42**, D374–D379.

25. Ioannidis,N.M., Rothstein,J.H., Pejaver,V., Middha,S., McDonnell,S.K., Baheti,S., Musolf,A., Li,Q., Holzinger,E., Karyadi,D. *et al.* (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.*, **99**, 877–885.
26. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
27. Lu,Q., Hu,Y., Sun,J., Cheng,Y., Cheung,K.-H. and Zhao,H. (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.*, **5**, 10576.
28. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
29. Schwarz,J.M., Rödelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
30. Rogers,M.F., Shihab,H.A., Mort,M., Cooper,D.N., Gaunt,T.R. and Campbell,C. (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, **34**, 511–513.
31. Fang,H. and Gough,J. (2013) DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res.*, **41**, D536–D544.
32. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
33. The Gene Ontology Consortium (2019) The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
34. Sundaram,L., Gao,H., Padigepati,S.R., McRae,J.F., Li,Y., Kosmicki,J.A., Fritzilas,N., Hakenberg,J., Dutta,A., Shon,J. *et al.* (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.*, **50**, 1161–1170.
35. Hecht,M., Bromberg,Y. and Rost,B. (2015) Better prediction of functional effects for sequence variants. *BMC Genomics*, **16**, S1.
36. Jeong,H., Mason,S.P., Barabási,A.-L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
37. Padi,M. and Quackenbush,J. (2015) Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators. *BMC Syst. Biol.*, **9**, 80.
38. Diss,G., Ascencio,D., DeLuna,A. and Landry,C.R. (2014) Molecular mechanisms of paralogous compensation and the robustness of cellular networks. *J. Exp. Zool. Part B: Mol. Dev. Evol.*, **322**, 488–499.
39. Heinzinger,M., Elnaggar,A., Wang,Y., Dallago,C., Nachaev,D., Matthes,F. and Rost,B. (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, **20**, 723.
40. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
41. van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
42. Miller,M., Vitale,D., Kahn,P.C., Rost,B. and Bromberg,Y. (2019) funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res.*, **47**, e142.
43. Grimm,D.G., Azencott,C.-A., Aicheler,F., Gieraths,U., MacArthur,D.G., Samocha,K.E., Cooper,D.N., Stenson,P.D., Daly,M.J., Smoller,J.W. *et al.* (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, **36**, 513–523.