Received: 22 April 2019 | Revised: 27 June 2019 | Accepted: 15 July 2019

DOI: 10.1002/humu.23875

SPECIAL ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Assessment of predicted enzymatic activity of α-*N*-acetylglucosaminidase variants of unknown significance for CAGI 2016

Wyatt T. Clark[1] | Laura Kasak[2,3] | Constantina Bakolitsa[2] | Zhiqiang Hu[2] | Gaia Andreoletti[2] | Giulia Babbi[4] | Yana Bromberg[5] | Rita Casadio[4] | Roland Dunbrack[6] | Lukas Folkman[7] | Colby T. Ford[8] | David Jones[9] | Panagiotis Katsonis[10] | Kunal Kundu[11,20] | Olivier Lichtarge[10,12,13,14] | Pier L. Martelli[3] | Sean D. Mooney[15] | Conor Nodzak[8] | Lipika R. Pal[11] | Predrag Radivojac[16] | Castrense Savojardo[4] | Xinghua Shi[8] | Yaoqi Zhou[17] | Aneeta Uppal[8] | Qifang Xu[6] | Yizhou Yin[11,20] | Vikas Pejaver[16,18] | Meng Wang[19] | Liping Wei[19] | John Moult[11,21] | Guoying Karen Yu[1] | Steven E. Brenner[2] | Jonathan H. LeBowitz[1]

[1]Human Genetics, BioMarin Pharmaceutical, San Rafael, California

[2]Department of Plant and Microbial Biology, University of California, Berkeley, California

[3]Department of Biomedicine, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia

[4]Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

[5]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey

[6]Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania

[7]Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH, Zurich, Switzerland

[8]Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, Charlotte, North Carolina

[9]Bioinformatics Group, Department of Computer Science, University College London, London, UK

[10]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

[11]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland

[12]Department of Biochemistry & Molecular Biology, Baylor College of Medicine, Houston, Texas

[13]Department of Pharmacology, Baylor College of Medicine, Houston, Texas

[14]Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, Texas

[15]Buck Institute for Research on Aging, Novato, California

[16]Department of Computer Science, Indiana University, Bloomington, Indiana

[17]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana

[18]Department of Informatics, Indiana University, Bloomington, Indiana

[19]Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, P.R. China

[20]Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, Maryland

[21]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland

Correspondence
Wyatt T. Clark, BioMarin Pharmaceutical, 105 Digital Dr., Novato, CA 94949.
Email: wyatt.clark@bmrn.com

Abstract

The NAGLU challenge of the fourth edition of the Critical Assessment of Genome Interpretation experiment (CAGI4) in 2016, invited participants to predict the impact

© 2019 Wiley Periodicals, Inc.

Present address:
Gaia Andreoletti, Institute for Computational Health Sciences, University of California, San Francisco, CA; Predrag Radivojac, Khoury College of Computer Sciences, Northeastern University, Boston, MA; Yizhou Yin, Visterra, Inc., Waltham, MA; Sean Mooney, Department of Biomedical Informatics and Medical Education, University of Washington, Seattle WA.

Funding information
National Institute of Health, Grant/Award Numbers: U41 HG007346, R01 GM079656, R01 GM066099, U01 GM115486, R01 MH105524, R01 LM009722, R01 AG061105, R13 HG006650; National Health and Medical Research Council of Australia, Grant/Award Numbers: 1121629, 1059775, 1083450; Eesti Teadusagentuur, Grant/Award Number: IUT34-12

of variants of unknown significance (VUS) on the enzymatic activity of the lysosomal hydrolase α-*N*-acetylglucosaminidase (NAGLU). Deficiencies in NAGLU activity lead to a rare, monogenic, recessive lysosomal storage disorder, Sanfilippo syndrome type B (MPS type IIIB). This challenge attracted 17 submissions from 10 groups. We observed that top models were able to predict the impact of missense mutations on enzymatic activity with Pearson's correlation coefficients of up to .61. We also observed that top methods were significantly more correlated with each other than they were with observed enzymatic activity values, which we believe speaks to the importance of sequence conservation across the different methods. Improved functional predictions on the VUS will help population-scale analysis of disease epidemiology and rare variant association analysis.

KEYWORDS
CAGI, critical assessment, enzymatic activity, machine learning, Sanfilippo syndrome, variants of unknown significance, α-*N*-acetylglucosaminidase, NAGLU

1 | INTRODUCTION

The exponential increase in genetic data over the past decade has confronted researchers with an unprecedented number of rare variants of unknown disease significance (VUS) detected in the human population. Such data present both a challenge and an opportunity. In the context of newborn screening, a clinician might be asked to interpret only a handful of mutations in a specific gene of relevance, but that given gene might have hundreds of missense VUS in databases such as gnomAD (Karczewski et al., 2019). Although the sheer number of VUS may be high, having a priori knowledge of their likely disease relevance can facilitate the prescreening of such mutations. Next to observing a mutation in a confidently diagnosed patient, experimental characterization remains a valuable method for validating the impact of detected variants. However, this process can be time-consuming, costly, and impractical. In an attempt to bridge this gap, many computational methods have been developed to predict the impact of missense variants on protein function (Gallion et al., 2017; Tang & Thomas, 2016). As part of the effort to test and independently evaluate such algorithms, the Critical Assessment of Genome Interpretation (CAGI) creates challenges using unpublished experimental data to evaluate the performance of blinded phenotype prediction algorithms (Hoskins et al., 2017).

Sanfilippo syndrome, also known as mucopolysaccharidosis type III (MPS III), is a rare autosomal recessive inherited metabolic disease caused by a deficiency in one of four lysosomal enzymes catalyzing distinct steps in the sequential degradation of heparan sulfate (Coutinho, Lacerda, & Alves, 2012). Each enzyme deficiency defines a separate subtype: IIIA, IIIB, IIIC, IIID, although symptoms and disease progression are largely indistinguishable between types. The resultant accumulation of heparan sulfate within lysosomes, particularly in the brain and liver, leads to a severe neurological phenotype and death in the second decade. Mutations leading to type IIIB (MIM# 252920), one of the more commonly diagnosed types, are located in the gene encoding the lysosomal hydrolase, α-*N*-acetylglucosaminidase (NAGLU; MIM# 609701).

The accumulation of heparan sulfate due to partial or complete loss of NAGLU enzyme activity occurs in various tissues and cells; however, the clinical signs are mostly associated with the central nervous system (Birrane et al., 2019), causing severe cognitive disabilities, behavioral problems and developmental regression, leading to death in adolescence or early adulthood. The age of onset of Sanfilippo Type B is 1–4 years (Andrade, Aldámiz-Echevarría, Llarena, & Couce, 2015) and the estimate for lifetime risk at birth (number of patients per 100,000 live births) varies substantially in European populations from 0.05 in Sweden to 0.78 in Greece (Zelei, Csetneki, Vokó, & Siffel, 2018). To date, no effective treatment for Sanfilippo syndrome exists although several promising approaches are being developed, including enzyme replacement therapy, gene therapy, bone marrow stem cell transplantation and small molecules (Aoyagi-Scharber et al., 2017; Gaffke, Pierzynowska, Piotrowska, & Węgrzyn, 2018). Because newborns are asymptomatic at birth, early diagnosis is critical for improved management and outcome of therapeutic trials. The development of algorithms capable of reliably distinguishing between pathogenic and benign *NAGLU* alleles is an important step in this direction.

For the NAGLU challenge of the fourth edition of the CAGI experiment (CAGI4) in 2016, participants were asked to predict the impact of VUS on the enzymatic activity of NAGLU. Variants were selected for testing based on being present in the ExAC version v0.3 and not being present in Human Gene Mutation Database (HGMD; Lek et al., 2016). The enzymatic activity of these missense mutations in NAGLU had been previously measured in transfected cell lysates (Clark, Yu, Aoyagi-Scharber, & LeBowitz, 2018). Of the 163 VUS tested, 41 (25%) decreased the activity of NAGLU to levels consistent with known Sanfilippo Type B pathogenic alleles. Previous analysis of variants that were found to decrease activity to levels consistent with disease found that they were more likely to be buried and close to the active site of the protein.

TABLE 1 A list of participating teams and submitted predictive models

PI	Model name	PubMed	PolyPhen/SIFT/ Provean-based features	Structure-based features	PSSM/MSA based features	ML method	Training database
Bromberg	SNAP-1	Bromberg and Rost (2007)	Yes	Yes	Yes	Neural network	PMD
Bromberg	SNAP-2	Bromberg and Rost (2007)	Yes	Yes	Yes	Neural network	PMD
Moult	Moult Consensus	Yin, Kundu, Pal, and Moult (2017)	Yes	Yes	Yes	Support vector regression	
Lichtarge	Evolutionary Action	Katsonis and Lichtarge (2014)	No	No	Yes	None	
Wei	iFish	Wang and Wei (2016)	Yes	Yes	Yes	SVM	
Mooney	MutPred	Li et al. (2009)	Yes	No	Yes	Random forest	HGMD
Mooney	MutPred2 w/o homology	Pejaver et al. (2017)	Yes	No	Yes	Neural network ensemble	HGMD 2013
Mooney	MutPred2 w/ homology	Pejaver et al. (2017)	Yes	No	Yes	Neural network ensemble	HGMD 2013
Jones	HHblits w/ real contacts	Remmert, Biegert, Hauser, and Söding (2011)	No	Yes	Yes	Logistic regression	
Jones	HHblits w/ predicted contacts	Remmert et al. (2011)	No	No	Yes	Logistic regression	
Jones	HHblits w/o contacts	Remmert et al. (2011)	No	No	No	Logistic regression	
Jones	PAM250 PSSM		No	No	Yes	Logistic regression	
Ford	PolyPhen2 Random Forest	Ford, Uppal, Nodzak, and Shi (2019)	Yes	No	Yes	Random forest	1000 Genomes, NCBI, wANNOVAR
Casadio	INPS3D	Savojardo, Fariselli, Martelli, and Casadio (2016)	No	Yes	Yes	SVM	
Casadio	SNPs&GO	Capriotti et al. (2013)	No	Yes	Yes	SVM	
Zhou	EASE-MM	Folkman, Stantic, Sattar, and Zhou (2016)	No	No	Yes	Support vector regression	ProTherm
Dunbrack	Dunbrack-SVM	Wei, Xu, and Dunbrack (2013)	Yes	Yes	Yes	SVM	

Note: Detailed model descriptions provided by participants are available as Supporting Information Data.

Abbreviations: HGMD, Human Gene Mutation Database; ML, machine learning; MSA, multiple sequence alignments; PMD, Protein Mutation Database; PSSM, position-specific scoring matrices; SVM, support vector machine.

This challenge attracted 17 submissions from 10 groups (Table 1). Most of the models utilized sequence information (*n* = 16), one-third of the methods also added structure-based features in addition to sequence (*n* = 8). To the best of our knowledge, this is the largest assessment of predicted enzyme activity for rare population missense variants in CAGI.

2 | METHODS

2.1 | Selection and testing of NAGLU variants

For the CAGI challenge, we attempted to select missense variants that were both observed in the population and which were of unknown disease significance (Clark et al., 2018). To do this, we relied on the v0.3 release of the Exome Aggregation Consortium's (ExAC) collection of exome sequencing data comprising 60,706 individuals as a source for observed missense mutations (Lek et al., 2016). As a source of disease-associated variants, we relied on the 2016 v1 version of the HGMD (Stenson et al., 2003). All missense mutations are reported for Ensembl protein ENSP00000225927.1 for NAGLU. Figure S1 shows a schematic of the observed activity of tested variants and their amino acid positions.

2.2 | Predictor performance evaluation

We calculated a number of metrics to give a robust view of the performance of each team's submissions. For analysis, percent

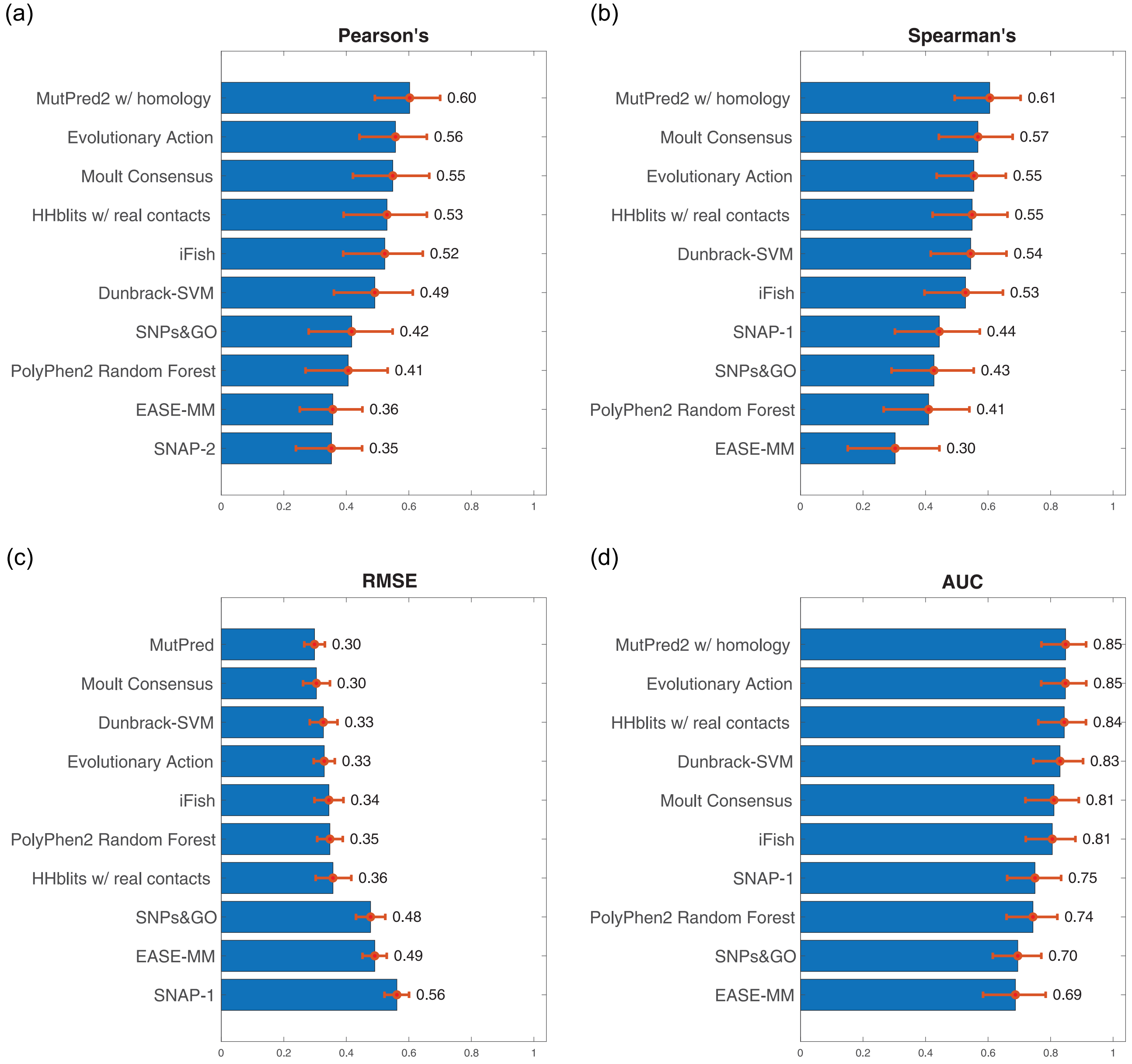


FIGURE 1 Resulting evaluation metrics for the top model for each team. (a) Pearson's *r*. (b) Spearman's *ρ*. (c) RMSE. (d) *AUC*. Error bars represent empirical 95% confidence intervals for each metric generated through 10^4 iterations of bootstrap sampling of data points. Only the top-performing model for each team is shown for each metric. *AUC* was calculated by designating missense mutations with <0.15 observed *fwt* activity as positives (Section 2.2). Resulting metrics for all models are available in Table S1 and Figure S2. *AUC*, area under the receiver operating characteristic curve; *fwt*, fraction wild-type; RMSE, root-mean-squared error

wild-type (*%WT*) activity values were converted to fraction wild-type (*fwt*) activity values. Our analysis treated the experiment both as a binary classification problem and as one with a continuous-valued target variable.

We calculated the Pearson and Spearman correlation coefficients, root-mean-squared error (RMSE) with observed enzymatic activity values and area under the receiver operating characteristic curve (*AUC*) for each set of predictions (Figure 1, Figure S2, Table S1). Predictor values submitted through the CAGI challenge were not normalized. Although linear transformations of predicted values, like *z*-score normalization, will not impact Pearson's *r* or Spearman's *ρ* they would impact RMSE values. RMSE represents the most stringent metric that we used to evaluate predictions as it requires predictions to be properly scaled. As a supplement, precision and recall curves were generated (Figure S3).

Some metrics, such as sensitivity, specificity, and *AUC*, assume a binary target variable. In these cases, we designated pathogenic variables as true positives, and benign as true negatives. We used 0.15 *fwt* activity as a threshold with which we distinguished pathogenic from benign variables. This level of *fwt* activity is consistent with what we observed from previously identified pathogenic mutations as described in Clark et al. (2018). We also calculated *AUC* and *F*-max values for thresholds ranging from 0.05 to 0.95 in increments of 0.05 (Table S2, Table S3). For each predictor, a sliding decision threshold was varied from the highest predictor score to its lowest. Because, in this instance, low predictor scores designate positives, each predicted mutation with a score below the threshold was chosen as a predicted positive. All others were designated as predicted negative, or benign, data points. A simple way to achieve the same impact would be to multiply predictions of each model by −1 and proceed as one normally would when calculating binary metrics. We also generated receiver operating characteristic (ROC) curves (Figure 2). Optimal positions on the ROC curve, designated by red dots in Figure 2 were determined as the point with the lowest square root of the sum of the square of the false positive rate and false negative rate.

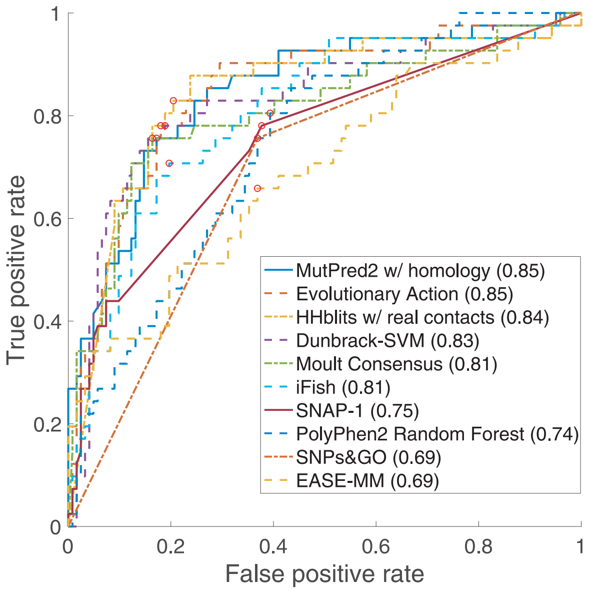


FIGURE 2 ROC curves for each team's top-performing model when using 0.15 observed *fwt* activity as the threshold at which pathogenic and benign mutations are distinguished (Section 2.2). Red dots represent the positions in each ROC curve closest to the upper left-hand side of the plot and were calculated by finding the point with the lowest square root of the sum of the square of the false positive rate and false negative rate. *fwt*, fraction wild-type; ROC, receiver operating characteristic

2.3 | Determining statistical significance of correlation coefficients

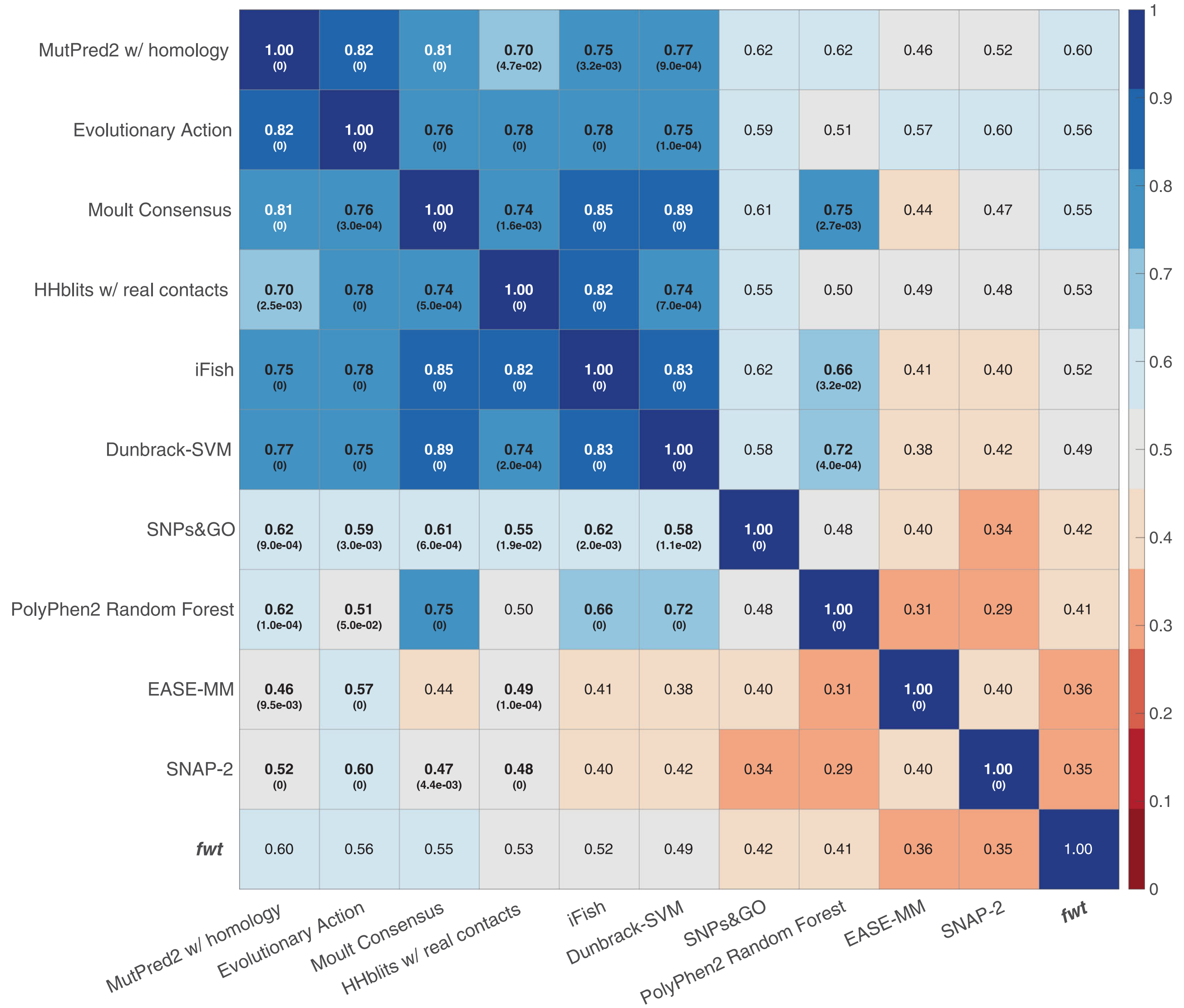
We calculated pairwise correlation coefficients for all models (Figure 3, Table S4, Table S5). For any two models, X_i and X_j, we calculated whether the correlation of model X_i with the experimentally observed enzymatic activity values, EA, was statistically significantly different than its correlation with model X_j using bootstrap simulation (Table S6). To do this 10,000 random samples of 163 data points were generated using sampling with replacement of the original data. For each pair of predictors, each sample, s_k, was used to calculate the correlation coefficient of X_i with and X_j and EA; $r(X_i, X_j)_k$ and $r(X_i, \mathrm{EA})_k$, respectively. We then calculated the percentage of times $r(X_i, X_j)_k$ was greater than $r(X_i, \mathrm{EA})_k$; and deemed the correlation of model X_i with EA to be significantly greater than the correlation with model X_j if this value was <5%. Figure S4 shows a heatmap of pairwise correlation coefficients for all models, including off-the-shelf methods.

2.4 | Determining the uniqueness of predictions using a linear regression model

To estimate the specific contribution of each prediction model to the variance with experimental results (R^2), a multiple linear regression model was applied (Figure 4). First, a linear regression model was built for every single model. The top model from each group was chosen based on the highest adjusted R^2 values (Table S7). Next, models were combined with the best performing model, and the linear regression equation was recalculated to evaluate the contributions of each model to the variance (Table S8).

2.5 | Additional predictions used for evaluation

We compared the predictions submitted to the NAGLU challenge to several off-the-shelf methods. As a simple method, we considered Grantham scores (Grantham, 1974). Quantitative scores for PolyPhen and SIFT were obtained from the ExAC VCF file and were generated using VEP v81 (Adzhubei et al., 2010; Kumar, Henikoff, & Ng, 2009). We previously analyzed categorical predictions produced by SIFT and PolyPhen. Here we only considered quantitative scores for both predictors. CADD annotations were obtained from CADD v1.4 (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019). REVEL predictions were taken from the June 3, 2016 release of predictions



	MutPred2 w/ homology	Evolutionary Action	Moult Consensus	HHblits w/ real contacts	iFish	Dunbrack-SVM	SNPs&GO	PolyPhen2 Random Forest	EASE-MM	SNAP-2	*fwt*
MutPred2 w/ homology	**1.00** (0)	**0.82** (0)	**0.81** (0)	**0.70** (4.7e-02)	**0.75** (3.2e-03)	**0.77** (9.0e-04)	0.62	0.62	0.46	0.52	0.60
Evolutionary Action	**0.82** (0)	**1.00** (0)	**0.76** (0)	**0.78** (0)	**0.78** (0)	**0.75** (1.0e-04)	0.59	0.51	0.57	0.60	0.56
Moult Consensus	**0.81** (0)	**0.76** (3.0e-04)	**1.00** (0)	**0.74** (1.6e-03)	**0.85** (0)	**0.89** (0)	0.61	**0.75** (2.7e-03)	0.44	0.47	0.55
HHblits w/ real contacts	**0.70** (2.5e-03)	**0.78** (0)	**0.74** (5.0e-04)	**1.00** (0)	**0.82** (0)	**0.74** (7.0e-04)	0.55	0.50	0.49	0.48	0.53
iFish	**0.75** (0)	**0.78** (0)	**0.85** (0)	**0.82** (0)	**1.00** (0)	**0.83** (0)	0.62	**0.66** (3.2e-02)	0.41	0.40	0.52
Dunbrack-SVM	**0.77** (0)	**0.75** (0)	**0.89** (0)	**0.74** (2.0e-04)	**0.83** (0)	**1.00** (0)	0.58	**0.72** (4.0e-04)	0.38	0.42	0.49
SNPs&GO	**0.62** (9.0e-04)	**0.59** (3.0e-03)	**0.61** (6.0e-04)	**0.55** (1.9e-02)	**0.62** (2.0e-03)	**0.58** (1.1e-02)	**1.00** (0)	0.48	0.40	0.34	0.42
PolyPhen2 Random Forest	**0.62** (1.0e-04)	**0.51** (5.0e-02)	**0.75** (0)	0.50	**0.66** (0)	**0.72** (0)	0.48	**1.00** (0)	0.31	0.29	0.41
EASE-MM	**0.46** (9.5e-03)	**0.57** (0)	0.44	**0.49** (1.0e-04)	0.41	0.38	0.40	0.31	**1.00** (0)	0.40	0.36
SNAP-2	**0.52** (0)	**0.60** (0)	**0.47** (4.4e-03)	**0.48** (0)	0.40	0.42	0.34	0.29	0.40	**1.00** (0)	0.35
fwt	0.60	0.56	0.55	0.53	0.52	0.49	0.42	0.41	0.36	0.35	1.00

FIGURE 3 Pairwise Pearson's correlation coefficients between each team's top model as well as observed enzymatic activity values. A bold/starred value for row *i*, column *j* indicates that method *i* is statistically significantly more correlated with model *j* than with *fwt* values at the 0.05 level. Statistical significance was calculated as described in Section 2. Pairwise Pearson's correlation coefficients, standard deviations and *p* values for all methods, including supplemental methods are shown in Table S4–S6 and Figure S4. *fwt*, fraction wild-type

(Ioannidis et al., 2016). Because quantitative scores produced by Grantham, PolyPhen, REVEL, and CADD are negatively correlated with enzymatic activity (a higher score indicated a higher likelihood of being pathogenic) scores were inverted by subtracting them from 1. This is a linear transformation that will only impact the sign of correlation values but will allow a fair comparison of RMSE value produced by these predictors to other models. In the case of CADD raw and Phred scores, this was done after normalizing those scores to the range (0–1) by subtracting the minimum value raw or Phred score from a prediction then dividing by the maximum value minus the minimum. Again, this is only a linear transformation that will not impact correlation or *AUC* values but will facilitate a fair comparison of RMSE values between models. Grantham scores were normalized in a similar fashion as CADD scores, but a minimum value of 0 was assumed.

Relative solvent accessibility was calculated by first calculating the solvent accessibility of each amino acid in the monomer of the Protein Data Bank structure 4XWH using DSSP. Raw solvent accessibility values were then normalized by dividing by maximum solvent accessibility values (Rost & Sander, 1994). Because residues 1 through 23 are a signaling peptide and are proteolytically cleaved, they are not present in the PDB structure for NAGLU. This means there is no solvent accessibility value for the p.Arg16Val missense mutation. We replaced this missing value with the average relative solvent accessibility value for the remaining 162 amino acids in the evaluation set. A final, ad hoc model was generated by taking the average of normalized Grantham scores and relative solvent accessibility.

All results for such methods are included along with all submitted models in Table S9.

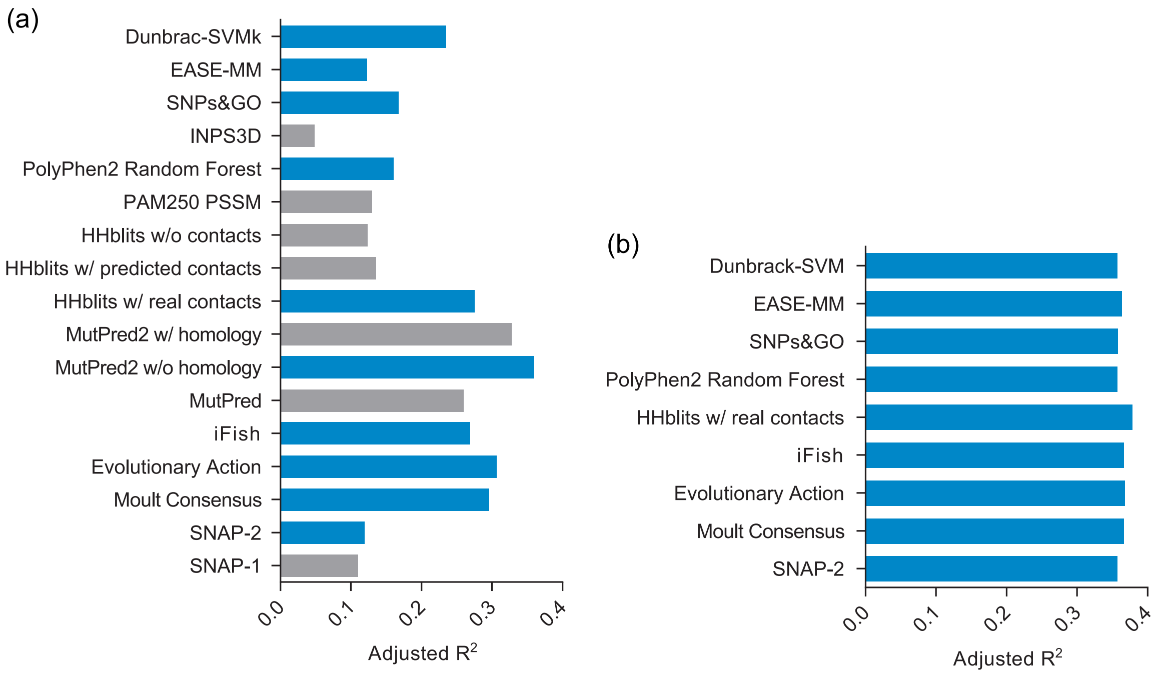


FIGURE 4 Percentage of variance in predictions explained by each model. Adjusted R^2 values from the linear model with (a) a single method and (b) a combination of the best performing method (MutPred2 w/o Homology) and any other tool

3 | RESULTS

3.1 | Participation in the NAGLU CAGI challenge

There were 17 submitted sets of predictions from 10 individual teams for the CAGI NAGLU Challenge (Table 1, model descriptions available in Supporting Information Data). Of these 17 submissions, six models (Moult Consensus, iFish, HHBlits w/ real contacts, INPS3D, SNP&GP, and Dunbrack-SVM) utilized one of the NAGLU protein structures. Nine models utilized the output of commonly used predictors of variant functional effect such as PolyPhen, SIFT, or PROVEAN (Adzhubei et al., 2010; Choi & Chan, 2015; Kumar et al., 2009). All but one model ($n = 16$) utilized information from multiple sequence alignments or position-specific scoring matrices. Three models utilized HGMD as a source of training data.

3.2 | Analysis of predicted enzymatic activities

We utilized several metrics when assessing performance in the CAGI4 challenge. Averages and standard deviations for each metric were obtained by randomly sampling the 163 variants with replacement 10^4 times. When calculating (*AUC*) we utilized 0.15 *fwt* activity thresholds at which variants were designated as either neutral or disease-causing for our primary analysis. This threshold was based on the upper limit of *fwt* activity measured in previously observed pathogenic mutations (Clark et al., 2018). We also calculated *AUC* and *F*-max for thresholds ranging from 0.05 to 0.95 in increments of 0.05 (Table S2, Table S3). Although *AUC* and *F*-max values were generated by sampling mutations, ROC and precision-recall curves were generated from unsampled data. In cases where more than one model for a team ranked highly according to a particular metric, we only mention the top-performing model, although all results for all models are shown in Table S1. Precision and recall curves were generated as a supplemental figure (Figure S3).

Figure 1 shows several metrics (Pearson's *r*, Spearman's *ρ*, *AUC*, RMSE) used to evaluate the performance of each predictor. We found that the MutPred w/ homology model performed the best in terms of Pearson's *r* ($r = .60$), followed by the Evolutionary Action model ($r = .56$), and Moult Consensus ($r = .55$) respectively. The same three teams performed the best in terms of Spearman's *ρ* as well (MutPred w/ homology [$\rho = .61$], Moult Consensus [$\rho = .57$], and Evolutionary Action [$\rho = .55$]). It should be noted that Spearman and Pearson correlation coefficients were very correlated for all models. RMSE represents the most stringent measure of model performance that we utilized. MutPred obtained the lowest RMSE (0.30), followed by Moult Consensus (0.30) and Dunbrack-SVM (0.32). Figure 2 shows ROC curves for the top 10 performing submissions according to *AUC*, and Figure 1 shows the obtained *AUC* values for these models. We found that MutPred2 w/ homology performed the best in terms of *AUC* (*AUC* = 0.85), followed by Evolutionary Action (*AUC* = 0.85) and HHblits w/ real contacts (*AUC* = 0.84).

Although each of these metrics measures a different aspect of a predictor's performance, we found a large amount of agreement between metrics in the overall ranking of models. For example, MutPred2 w/ homology performed the best according to Pearson's *r*, Spearman's *ρ*, and *AUC*. In terms of RMSE, the MutPred2 w/ homology only slightly underperformed compared to the MutPred model.

3.3 | Easy and difficult to predict mutations

We determined whether any mutations were easy or difficult to predict (Table S10, Figure S5). This was done by measuring the average RMSE: (a) across all predictors, and (b) for the top 5 models (Mutpred, MutPred2 w/o homology, Moult consensus, Dunbrack-SVM, and Evolutionary Action). We found several mutations for which experimentally observed activities were both easy and difficult for models to predict.

The majority of easiest to predict deleterious mutations involved nonconservative substitutions of buried residues. Within the NAGLU structure, these variants are predicted to affect protein stability via disruption of aromatic clusters or stacking, salt bridges and hydrogen bonding networks, as well as through proximity to the active site or interference with the binding site pocket.

The majority of hardest to predict mutations involved moderate or conservative substitutions of partially or fully solvent-exposed residues. Interpretation of their effects within the NAGLU structure was not immediately obvious. One possibility involves an effect on protein solubility, especially in the context of the enzyme's trimerization. The hardest to predict mutation (p.Pro283Leu) was predicted to have low activity by most predictors but was shown to actually increase activity. Both this variant and p.Gly596Cys, another benign variant predicted to have low activity, involve nonconservative substitutions and are buried in the NAGLU structure.

3.4 | Correlation between predictive models

While it may be easy to focus on which model performed the best in terms of a particular measure, we observed that top models from each team were significantly more correlated with at least one other model from another team than they were with *f*wt values (Figure 3). Furthermore, we observed that the six top-performing models as measured by Pearson's correlation coefficients were all more correlated with each other than with *f*wt activity values and that these correlations were found to be statistically significant through bootstrapping simulation (Section 2).

For example, although the MutPred2 w/ homology and Evolutionary Action models were correlated with observed activity values with coefficients of 0.60 and 0.56 respectively, they were correlated with each other with a Pearson's $r = .82$. For none of the 10^4 bootstrap samples generated did we observe that the two models were more correlated with *f*wt activity values than they were with each other. This suggests that these models perform reasonably well at predicting *f*wt activities for NAGLU, they are better at recapitulating each other's behavior although they are presumably based on distinct, and rather different, methodology.

In light of the high correlation between models, we did not observe that combining the best performing model with any other tool improved correlation with *f*wt activity values. To determine this, all models were fit to a linear regression model and the best tool out of all submissions from the same group was chosen based on the adjusted R^2 values (shown in black in Figure 4). As R^2 values, in this case, should be equivalent to squared Pearson's correlation coefficients, MutPred2 w/o homology was found to explain the highest proportion of variance (36%). The best combination (MutPred2 w/o homology and HHblits w/ real contacts) increased the adjusted R^2 value only by 0.02%. This implies that MutPred2 w/o homology itself can represent all of the other tools.

3.5 | Comparison to supplemental models

We also evaluated the performance of several commonly used off-the-shelf tools as supplemental models including REVEL, Polyphen, SIFT and CADD (Figure S2, Section 2, Table S9). REVEL performed the best out of all off-the-shelf methods (Pearson's $r = .56$) although it was not as correlated with observed *f*wt activity values as MutPred2 w/ homology (Pearson's $r = .60$) and the two models were highly correlated with each other (Pearson's $r = .89$) (Table S4). We observed that for both models, prediction scores were statistically significantly more correlated with each other than they were with observed *f*wt activity values; for none of 10^4 bootstrap simulations did we observe a higher correlation between either model and *f*wt activity value than with each model (Section 2). It is important to point out that REVEL uses predictions from the previous version of MutPred as features. Furthermore, MutPred2 w/ homology was trained using an older version of HGMD (June 2013) than REVEL (2015.2 version of HGMD).

While we observed that top methods performed better than PolyPhen scores, in the case of MutPred w/ homology, we did not find the difference to be significant in terms of Pearson's correlation coefficients. Solvent accessibility and Grantham scores performed the poorest in terms of Pearson's and Spearman's correlation coefficients, but Grantham scores had lower RMSE than PolyPhen scores.

Out of all supplemental models, the proportion of variance explained based on a linear regression model was the highest for REVEL (adjusted $R^2 = 30.1\%$), and there was no improvement when additional models were added (Figure S6). This indicates that REVEL itself can be a good representative of all commonly used models that were selected.

4 | CONCLUSIONS AND DISCUSSION

For the CAGI NAGLU challenge, we asked participants to predict the impact of missense mutations on the enzymatic activity of NAGLU. This task is different than predicting whether a model could perform poorly if it is not able to distinguish between a benign mutation that has 60% wild-type activity and one that has 90%, or a pathogenic mutation with 0% activity and one with 10%. Although this is a different task than predicting pathogenicity, we found that participants in the 2016 NAGLU CAGI challenge performed well. This performance was obtained contrary to the fact that many models were not explicitly trained for the task of predicting enzymatic activity, instead being designed for the slightly different task of distinguishing pathogenic from benign variants.

Although models performed well, we did observe that the top methods were significantly more correlated with each other than they were with observed activity values. In many cases, such as with MutPred2 and Evolutionary Action, methods were highly correlated in spite of having relatively distinctive methodologies; one being a supervised machine learning model, the second being one based on a calculus of evolutionary variations. The starkly different methodologies of these two models suggest that a common feature type is the primary driver behind the high level of correlation between methods. Of all the feature types employed by participating models, sequence conservation was the most common. While we observed that sequence conservation was a unifying feature across almost all methods, we were unable to observe any relationship between the training data used for supervised methods and their performance. In fact, one of the top methods, Evolutionary Action, was a method that did not use a training data set of known pathogenic mutations at all, instead focused on evolutionary conservation. There are technical details, such as the phylogenetic depth at which one should measure conservation and the choice of alignment algorithm, that must be considered when using homology to infer the impact of mutations (Katsonis et al., 2014).

The high level of correlation between in silico models also has implications for the interpretation of variants in a clinical setting. As noted by the ACMG guidelines for variant interpretation, many tools rely on the same underlying data to make predictions, and single predictors should not be counted individually as evidence that a variant is pathogenic (Richards et al., 2015). Considering predictions from multiple tools will not necessarily add additional information regarding a particular mutation.

Our current in vitro enzyme activity assay is limited to testing missense coding variants, and as shown by Clark et al. (2018), there is a very good agreement with observed activity and pathogenicity of known and well-annotated disease variants. However, the in vitro assay may not always correlate with enzyme activities tested directly from patient samples. Mutations were introduced into a vector containing the NAGLU cDNA. Splicing, promoter/enhancer, and epigenetic mutations will be missed. Also, protein is being expressed at super-physiological levels. Some mutations may result in protein aggregation at high concentrations, but not at endogenous levels. Furthermore, proteins being expressed in cell lines from different organisms, or even different tissues, can exhibit variability in activity (Meijer et al., 2017). We can point to at least one mutation, p.Arg464Gln, whose activity we were surprised by. p.Arg464Gln was generated in multiple, sequence confirmed independent constructs, and its activity was checked by several independent transfections and was consistently found to have 3% wild-type activity. This particular mutation has a non-Finnish-European allele frequency in gnomAD v2.1 of 1.74×10^{-4}, whereas the most common known disease-causing mutation, p.Ser612Gly, has an allele frequency of 1.23×10^{-4}. Given the high frequency of this variant compared to known pathogenic mutations, it is surprising it has not yet appeared in a patient.

Functional data from genes with clear functional readouts are important. While genes such as BRCA1 are important in the context of cancer, determining the impact of a missense mutation on its function is not a simple task (Carvalho, Couch, & Monteiro, 2007). Functional screening data on more genes like *NAGLU* can help train better models, which, in turn, can produce better predictions on genes that are more difficult to assay. More data will also allow researchers to determine trends amongst mutations that are easy and difficult to predict, as well as those that might not produce accurate activity readouts in similar overexpression based cell line systems.

DATA ACCESSIBILITY

The data that support the findings of this study are openly available as supplementary files. Submitted predictions for each model are available to registered users from the CAGI web site at: https://genomeinterpretation.org/content/4-NAGLU

ACKNOWLEDGEMENT

The authors are grateful to the CAGI organizers, data providers and participants. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650.

ORCID

Wyatt T. Clark http://orcid.org/0000-0002-5041-3669
Laura Kasak http://orcid.org/0000-0003-4182-2396
Constantina Bakolitsa http://orcid.org/0000-0002-6980-9831
Zhiqiang Hu http://orcid.org/0000-0001-8854-3410
Gaia Andreoletti http://orcid.org/0000-0002-0452-0009
Giulia Babbi http://orcid.org/0000-0002-9816-4737
Yana Bromberg http://orcid.org/0000-0002-8351-0844
Rita Casadio http://orcid.org/0000-0002-7462-7039
Roland Dunbrack http://orcid.org/0000-0001-7674-6667
Lukas Folkman http://orcid.org/0000-0002-5811-8875
Colby T. Ford http://orcid.org/0000-0002-7859-3622
Panagiotis Katsonis http://orcid.org/0000-0002-7172-1644
Kunal Kundu http://orcid.org/0000-0002-4452-4290
Olivier Lichtarge http://orcid.org/0000-0003-4057-7122
Pier L. Martelli http://orcid.org/0000-0002-0274-5669
Sean D. Mooney http://orcid.org/0000-0003-2654-0833
Lipika R. Pal http://orcid.org/0000-0002-3390-110X
Predrag Radivojac http://orcid.org/0000-0002-6769-0793
Castrense Savojardo http://orcid.org/0000-0002-7359-0633
Yaoqi Zhou http://orcid.org/0000-0002-9958-5699
Yizhou Yin http://orcid.org/0000-0002-5365-2294
Vikas Pejaver http://orcid.org/0000-0002-1943-0284
Liping Wei http://orcid.org/0000-0002-1795-8755
John Moult http://orcid.org/0000-0002-3012-2282
Steven E. Brenner http://orcid.org/0000-0001-7559-6185
Jonathan H. LeBowitz http://orcid.org/0000-0001-9323-2422

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Andrade, F., Aldámiz-Echevarría, L., Llarena, M., & Couce, M. L. (2015). Sanfilippo syndrome: Overall review. *Pediatria Internazionale*, *57*(3), 331–338. https://doi.org/10.1111/ped.12636

Aoyagi-Scharber, M., Crippen-Harmon, D., Lawrence, R., Vincelette, J., Yogalingam, G., Prill, H., … Bunting, S. (2017). Clearance of heparan sulfate and attenuation of CNS pathology by intracerebroventricular BMN 250 in Sanfilippo Type B mice. *Molecular Therapy. Methods & Clinical Development*, *6*, 43–53. https://doi.org/10.1016/j.omtm.2017.05.009

Birrane, G., Dassier, A. L., Romashko, A., Lundberg, D., Holmes, K., Cottle, T., … Meiyappan, M. (2019). Structural characterization of the α-*N*-acetylglucosaminidase, a key enzyme in the pathogenesis of Sanfilippo syndrome B. *Journal of Structural Biology*, *205*(3), 65–71. https://doi.org/10.1016/j.jsb.2019.02.005

Bromberg, Y., & Rost, B. (2007). SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, *35*(11), 3823–3835. https://doi.org/10.1093/nar/gkm238

Capriotti, E., Calabrese, R., Fariselli, P., Martelli, P., Altman, R. B., & Casadio, R. (2013). WS-SNPs&GO: A web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics*, *14*(Suppl 3), S6. https://doi.org/10.1186/1471-2164-14-S3-S6

Carvalho, M. A., Couch, F. J., & Monteiro, A. N. A. (2007). Functional assays for BRCA1 and BRCA2. *The International Journal of Biochemistry & Cell Biology*, *39*(2), 298–310. https://doi.org/10.1016/j.biocel.2006.08.002

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, *31*(16), 2745–2747. https://doi.org/10.1093/bioinformatics/btv195

Clark, W. T., Yu, G. K., Aoyagi-Scharber, M., & LeBowitz, J. H. (2018). Utilizing ExAC to assess the hidden contribution of variants of unknown significance to Sanfilippo Type B incidence. *PLoS One*, *13*(7):e0200008. https://doi.org/10.1371/journal.pone.0200008

Coutinho, M. F., Lacerda, L., & Alves, S. (2012). Glycosaminoglycan storage disorders: A review. *Biochemistry Research International*, *2012*, 1–16. https://doi.org/10.1155/2012/471325

Folkman, L., Stantic, B., Sattar, A., & Zhou, Y. (2016). EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *Journal of Molecular Biology*, *428*(6), 1394–1405. https://doi.org/10.1016/j.jmb.2016.01.012

Ford, C. T., Uppal, A., Nodzak, C. M., & Shi, X. (2019). Prediction of the effect of naturally occurring missense mutations on cellular N-Acetyl-glucosaminidase enzymatic activity. *bioRxiv*. Retrieved from https://doi.org/10.1101/598870

Gaffke, L., Pierzynowska, K., Piotrowska, E., & Węgrzyn, G. (2018). How close are we to therapies for Sanfilippo disease? *Metabolic Brain Disease*, *33*(1), 1–10. https://doi.org/10.1007/s11011-017-0111-4

Gallion, J., Koire, A., Katsonis, P., Schoenegge, A. M., Bouvier, M., & Lichtarge, O. (2017). Predicting phenotype from genotype: Improving accuracy through more robust experimental and computational modeling. *Human Mutation*, *38*(5), 569–580. https://doi.org/10.1002/humu.23193

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, *185*(4154), 862–864.

Hoskins, R. A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., & Brenner, S. E. (2017). Reports from CAGI: The critical assessment of genome interpretation. *Human Mutation*, *38*(9), 1039–1041. https://doi.org/10.1002/humu.23290

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., & Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., … MacArthur, D. G. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. Retrieved from https://doi.org/10.1101/531210

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Science*, *23*(12), 1650–1666. https://doi.org/10.1002/pro.2552

Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Research*, *24*(12), 2050–2058. https://doi.org/10.1101/gr.176214.114

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., … Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, *25*(21), 2744–2750. https://doi.org/10.1093/bioinformatics/btp528

Meijer, O. L. M., Te Brinke, H., Ofman, R., Ijlst, L., Wijburg, F. A., & van Vlies, N. (2017). Processing of mutant *N*-acetyl-α-glucosaminidase in mucopolysaccharidosis type IIIB fibroblasts cultured at low temperature. *Molecular Genetics and Metabolism*, *122*(1-2), 100–106. https://doi.org/10.1016/j.ymgme.2017.07.005

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K. A., Lin, G. N., Nam, H. -J., & Radivojac, P. (2017). MutPred2: Inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv*. Retrieved from https://doi.org/10.1101/134981

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–423. https://doi.org/10.1038/gim.2015.30

Rost, B., & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, *20*(3), 216–226. https://doi.org/10.1002/prot.340200303

Savojardo, C., Fariselli, P., Martelli, P. L., & Casadio, R. (2016). INPS-MD: A web server to predict stability of protein variants from sequence and structure. *Bioinformatics*, *32*(16), 2542–2544. https://doi.org/10.1093/bioinformatics/btw192

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S. T., … Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Human Mutation*, *21*(6), 577–581. https://doi.org/10.1002/humu.10212

Tang, H., & Thomas, P. D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, *203*(2), 635–647. https://doi.org/10.1534/genetics.116.190033

Wang, M., & Wei, L. (2016). iFish: Predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. *Scientific Reports*, *6*, 31321. https://doi.org/10.1038/srep31321

Wei, Q., Xu, Q., & Dunbrack, R. L., Jr. (2013). Prediction of phenotypes of missense mutations in human proteins from biological assemblies. *Proteins*, *81*(2), 199–213. https://doi.org/10.1002/prot.24176

Yin, Y., Kundu, K., Pal, L. R., & Moult, J. (2017). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4NAGLU (Human *N*-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation*, *38*(9), 1109–1122. https://doi.org/10.1002/humu.23267

Zelei, T., Csetneki, K., Vokó, Z., & Siffel, C. (2018). Epidemiology of Sanfilippo syndrome: Results of a systematic literature review. *Orphanet Journal of Rare Diseases*, *13*(1), 53. https://doi.org/10.1186/s13023-018-0796-4

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Clark WT, Kasak L, Bakolitsa C, et al. Assessment of predicted enzymatic activity of α-*N*-acetylglucosaminidase variants of unknown significance for CAGI 2016. *Human Mutation*. 2019;40:1519–1529. https://doi.org/10.1002/humu.23875