Assessment of predicted enzymatic activity of alpha-N-acetylglucosaminidase (NAGLU) variants of unknown significance for CAGI 2016

Wyatt Travis Clark
Laura Kasak
Constantina Bakolitsa
Zhiqiang Hu
Gaia Andreoletti
Aneeta Uppal
Lukas Folkman
Colby T. Ford
Yaoqi Zhou
Castrense Savojardo
Pier Luigi Martelli
Rita Casadio
Giulia Babbi
Yizhou Yin
Conor Nodzak
Predrag Radivojac
Xinghua Shi
Yana Bromberg
Panagiotis Katsonis
Olivier Lichtarge

Kunal Kundu
Lipika R. Pal
John Moult
Qifang Xu
Roland Dunbrack
Meng Wang
Liping Wei
David Jones
Vikas Pejaver
Sean D. Mooney
G. Karen Yu
Steven E. Brenner
Jonathan H. LeBowitz
Correspondence: Wyatt Travis Clark

*Present address:

Contract grant sponsor: NIH (U41 HG007346, R13 HG006650, R01 MH105524, R01 LM009722, GM079656 and GM066099), National Institute of Aging (R01-AG061105), National Health and Medical Research Council of Australia (1059775, 1083450, 1121629), NIGMS grant (1 U01 GM115486 01).

ORCID IDs:

Abstract

Keywords: alpha-N-acetylglucosaminidase, variants of unknown significance, machine learning, CAGI, enzymatic activity, critical assessment

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 (ref). Although the sheer number of VUS may be high, having *a-priori* knowledge of their likely disease relevance can facilitate the pre-screening 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evaluate such algorithms, the Critical Assessment of Genome Interpretation (CAGI) creates challenges using unpublished experimental data to evaluate the performance of blinded phenotype prediction algorithms (Hoskins et al., 2017).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 (OMIM# 252920), one of the more commonly diagnosed types, are located in the gene encoding the lysosomal hydrolase, α-N-acetylglucosaminidase (*NAGLU*, OMIM# 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, Aldamiz-Echevarria, 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, Voko, & Siffel, 2018). To date, no effective treatment for Sanfilippo syndrome exists although several promising approaches are being developed, including gene therapy, bone marrow stem cell transplantation and small molecules (Gaffke, Pierzynowska, Piotrowska, & Wegrzyn, 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. 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. This challenge attracted 17 submissions from ten groups. Most of the models utilized sequence information (n=15), one third of the methods also added structure-based features in addition to sequence (n=6). To the best of our knowledge, this is the largest assessment of predicted enzyme activity for rare population missense variants in CAGI.

Results

Participation in the NAGLU CAGI Challenge

There were 17 submitted sets of predictions from 10 individual teams for the CAGI NAGLU Challenge (Table 1:). 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. We estimated that a large fraction of participating models, nine in total, 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, Henikoff, & Ng, 2009). All but one model utilized information from multiple sequence alignments (MSAs) or position specific scoring matrices (PSSMs). Three models utilized HGMD as a source of training data.

Table 1:

A list of participating teams and submitted predictive models. PSSM, position specific scoring matrices; MSA, multiple sequence alignments; ML, machine learning; PMD, Protein Mutation Database; HGMD, Human Gene Mutation Database.

PI	Model Name	PubMed	PolyPhen/SIFT/Provean Based Features	Structure Based Features	PSSM/MSA Based Features	ML Method	Training Database
Bromberg	SNAP-1	17526529, 18757876	Yes	Yes	Yes	Neural Network	PMD
Bromberg	SNAP-2	26110438	Yes	Yes	Yes	Neural Network	PMD
Moult	Moult Consensus	28544272	Yes	Yes	Yes	Support Vector Regression	
Lichtarge	Evolutionary Action	25217195	No	No	Yes	None	
Wei	iFish	27527004	Yes	Yes	Yes	SVM	
Mooney	MutPred	19734154	Yes	No	Yes	Random Forest	HGMD

Mooney	MutPred2 w/o homology	bioRxiv 134981	Yes	No	Yes	Neural Network Ensemble	HGMD 2013
Mooney	MutPred2 w homology	bioRxiv 134981	Yes	No	Yes	Neural Network Ensemble	HGMD 2013
Jones	HHblits w/ real contacts	22198341?	No	Yes	Yes	Logistic regression	
Jones	HHblits w/ predicted contacts	22198341?	No	No	Yes	Logistic regression	
Jones	HHblits w/o contacts	22198341?	No	No	No	Logistic regression	
Jones	PAM250 PSSM		No	No	Yes	Logistic regression	
Ford	PolyPhen2 Random Forest	bioRxiv 598870	Yes	No	Yes	Random Forest	1000 Genomes, NCBI, wANNOVAR
Casadio	INPS3D	27153629	No	Yes	Yes	SVM	
Casadio	SNPs&GO	23819482	No	Yes	Yes	SVM	
Zhou	EASE-MM	26804571	No	No	Yes	Support Vector Regression	ProTherm
Dunbrack	Dunbrack-SVM	22965855	Yes	Yes	Yes	SVM	

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. Percent wild-type (***%wt***) activity values were converted to fraction wild-type (***fwt***) activity values. When calculating AUC we utilized both 0.10 and 0.15 ***fwt*** activity thresholds at which variants were designated as either neutral or disease-causing but considered 0.15 as a threshold 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). Although AUC values were generated by sampling mutations, ROC 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 models are represented in Supplementary Information.

Figure: Evaluation Metrics

Resulting evaluation metrics for each model (A: Pearson's r, B: Spearman's rho, 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.

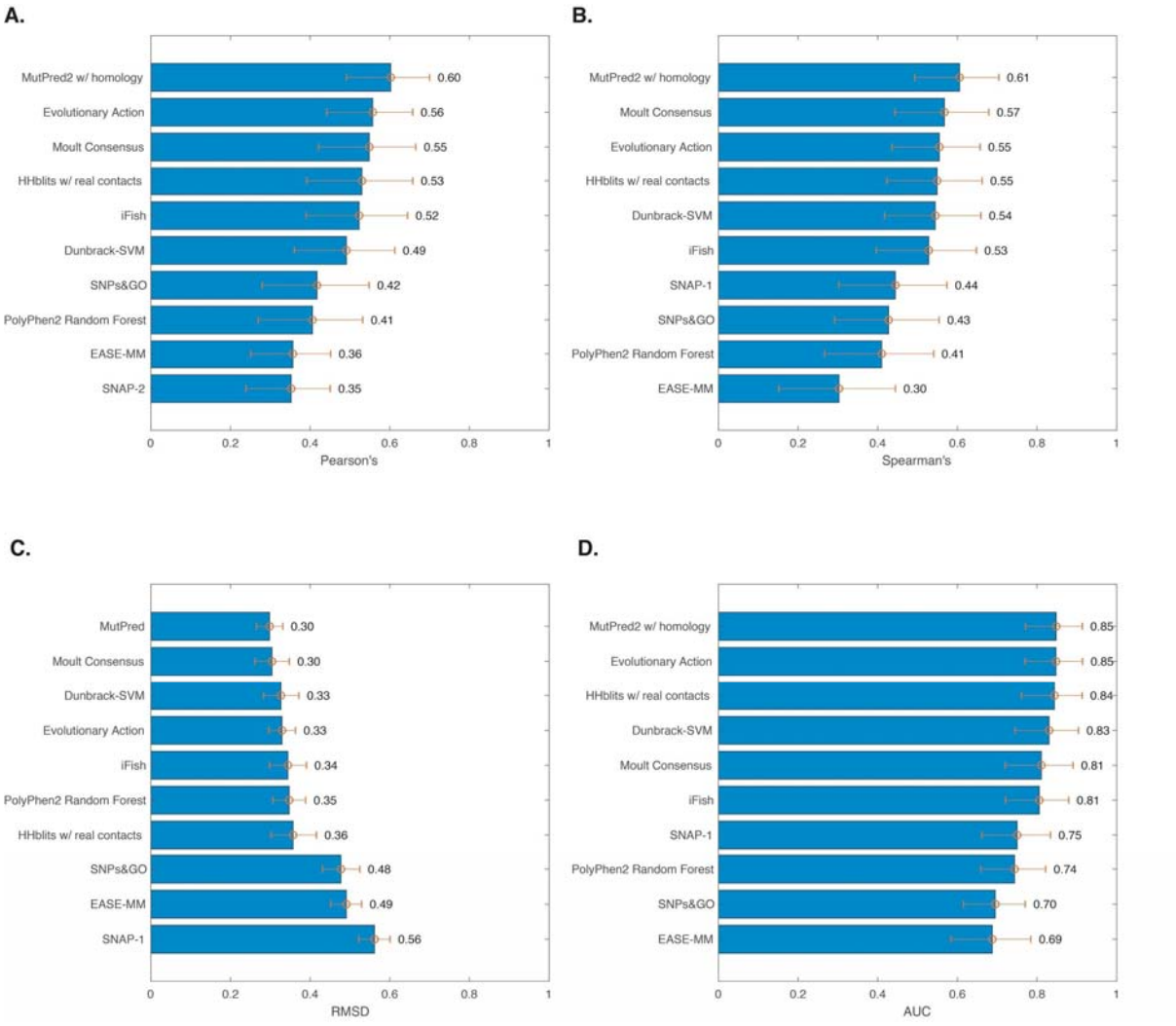


Figure 1.

Figure: ROC Curves

ROC curves for each team's top performing model when using 0.15 ***fwt*** activity as the threshold at which pathogenic and benign mutations are distinguished. 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.

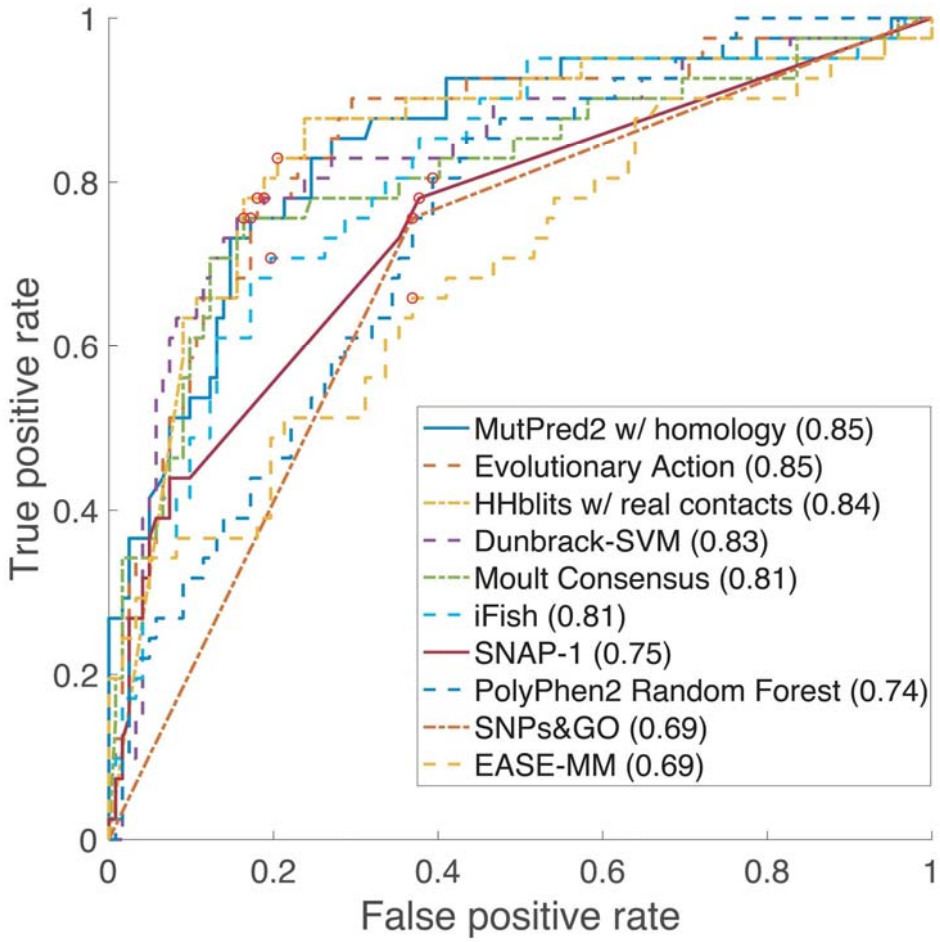


Figure 2.

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=0.60), followed by the Evolutionary Action model (r=0.56), and Moult Consensus (r=0.55) respectively. The same three teams performed the best in terms of Spearman's ρ as well (MutPred w/ homology (ρ=0.61), Moult Consensus (ρ=0.57), and Evolutionary Action (ρ=0.55)). 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, whereasFigure 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 measure 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. We observed that the SNAP-1 and SNAP-2 models had rather different ranking when comparing Spearman's and Pearson's correlation coefficients. Both models had similar values for each metric. The SNAP models had the 7th highest Spearman's correlation coefficient, but did not perform as well in terms of Pearson's correlation, ranking last amongst all teams. This suggests that while the SNAP models performed reasonably well in terms of producing a proper ranking of mutations according to ***fwt*** activity, the predicted scores generated by these models might not have been properly scaled.

Easy and difficult to predict mutations

We determined whether any mutations were easy or difficult to predict. This was done by measuring the average RMSE 1.) across all predictors, and 2.) 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 non-conservative substitutions of buried residues (Error Simple Table). 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 be pathogenic by most predictors but was shown to actually increase activity. Both this variant and p.Gly596Cys, another benign variant predicted to be pathogenic, involve non-conservative substitutions and are buried in the NAGLU structure.

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 ***fwt*** values (Figure 3). Furthermore, we observed that the 6 top performing models as measured by Pearson's correlation coefficients were all more correlated with each other than with ***fwt*** activity values, and that these correlations were found to be statistically significant through bootstrapping simulation (Methods).

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 of 0.82. For none of the 10^4 bootstrap samples generated did we observe that two models were more correlated with ***fwt*** activity values than they were with each other. This suggests that these models perform reasonably well at predicting ***fwt*** 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 ***fwt*** activity values. In order 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, and the proportion of total variance explained was 5% higher than the best supplemental method (REVEL, Suppl data).

Figure: Pairwise Correlations

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 Methods Determining statistical significance of correlation coefficients.

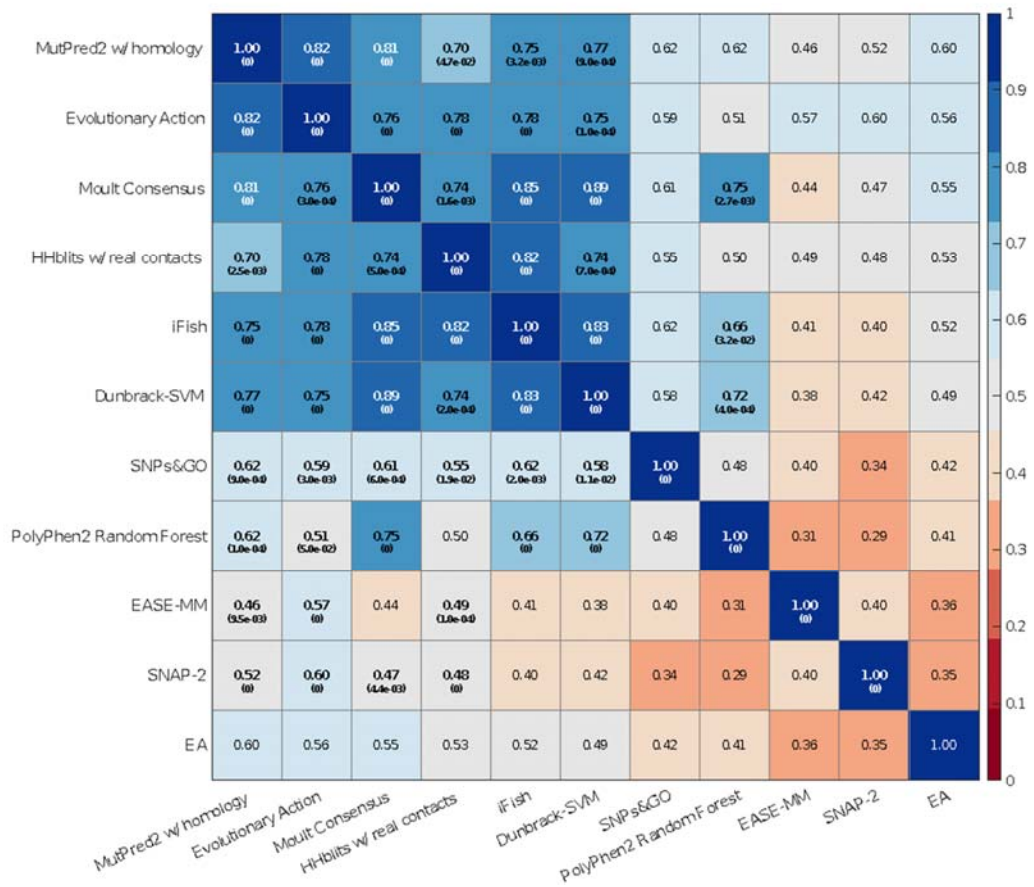


Figure 3.

Comparison to Supplemental Models

We also selected several commonly used supplemental models with which we could compare submitted models, including REVEL, Polyphen, SIFT and CADD (Methods Additional predictions used for evaluation). REVEL performed the best out of all off the shelf methods (Pearson's r = 0.56) although it was not as correlated with observed ***fwt*** activity values as MutPred 2 w/ homology (Pearson's r = 0.60) and the two models were highly correlated with each other (Pearson's r = 0.89) (Supplemental Table X). We observed that for both models, prediction scores were statistically significantly more correlated with each other than they were with observed ***fwt*** activity values; for none of 10^4 bootstrap simulations did we observe higher correlation between either model and fwt activity value than with each model (Methods Determining statistical significance of correlation coefficients). 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 (Suppl Fig X A). This indicates that REVEL itself can be a good representative of all commonly used models that were selected.

Conclusions/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 more difficult than merely predicting whether a mutation is pathogenic because the target variable is continuous. 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 more difficult and nuanced task than simply predicting pathogenicity, we found that participants in the 2016 NAGLU CAGI challenge performed surprisingly well. This performance is even more startling in light of the fact that many models were not explicitly trained for the more difficult task of predicting enzymatic activity, instead being designed for the simpler 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 training

methods for each model suggest that a common feature type is the primary driver behind the high level of correlation between these methods. Of all the feature types employed by participating models, sequence conservation was the most common. Limitations in the use of homology in non-synonymous single nucleotide variation impact prediction, include the effect of number of homologous sequences, phylogenetic depth and alignment choice on method accuracy (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, there is 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 severalindependent transfections and was consistently found to have 3% wild-type activity. This particular mutation has a Non-Finnish European allele frequency in gnomad of 2.53×10^{-4}, whereas the most common known disease causing mutation, p.Ser612Gln, has an allele frequency of 2.21×10^{-4}. Unless p.Arg464Gln is in severe linkage disequilibrium with all other pathogenic mutations within the European population, 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 read-outs in similar over-expression based cell line systems.

Methods

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). In order 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 Human Gene Mutation Database (HGMD) (Stenson et al., 2003).

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 et al., 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 (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 monomer of PDB 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.A16V 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 Supplemental Table.

Predictor performance evaluation

We calculated a number of metrics in order to give a robust view of the performance of each team's submissions. 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 as well as RMSE with observed enzymatic activity values for each set of predictions. 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 rho 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.

Some metrics, such as sensitivity, specificity, and AUC, assume a binary target variable. In these cases we designated pathogenic variables as positives, and benign as negatives. We used 0.15% wt 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 inClark et al., 2018. 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. Optimal positions on the ROC curve, designated by red dots in Figure: ROC Curves, 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.

Determining statistical significance of correlation coefficients

For any two models, X_i andX_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. In order 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 $r(X_i, EA)_k$ respectively. We then calculated the percentage of times $r(X_i, X_j)_k$ was greater than $r(X_i, EA)_k$; 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 less than 5%.

Determining the uniqueness of predictions using a linear regression model

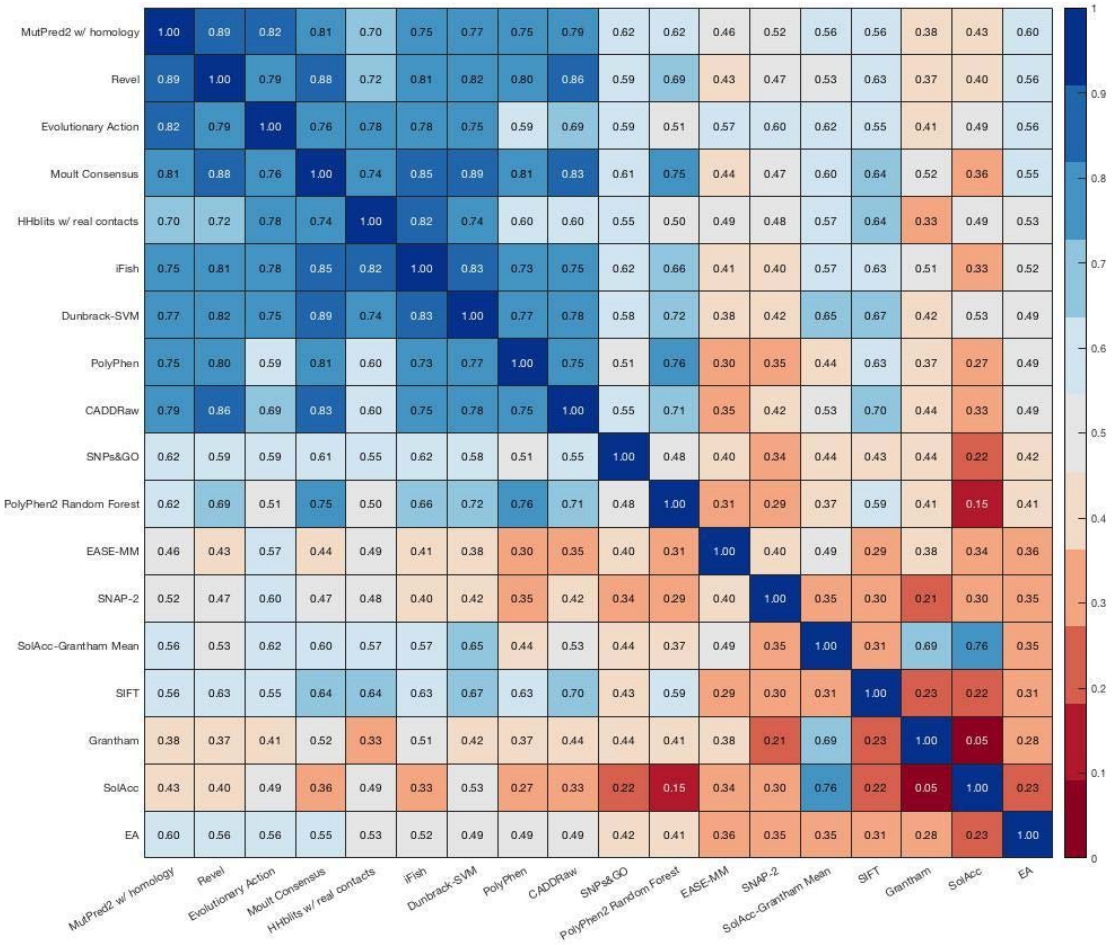
To estimate the specific contribution of each prediction model to the variance with experimental results (R^2), a multiple linear regression model was applied. First, a linear regression model was built for every single model. The top model from each group was chosen based on the highest adjusted R^2 values. 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.

Supplemental

Supplemental Figures

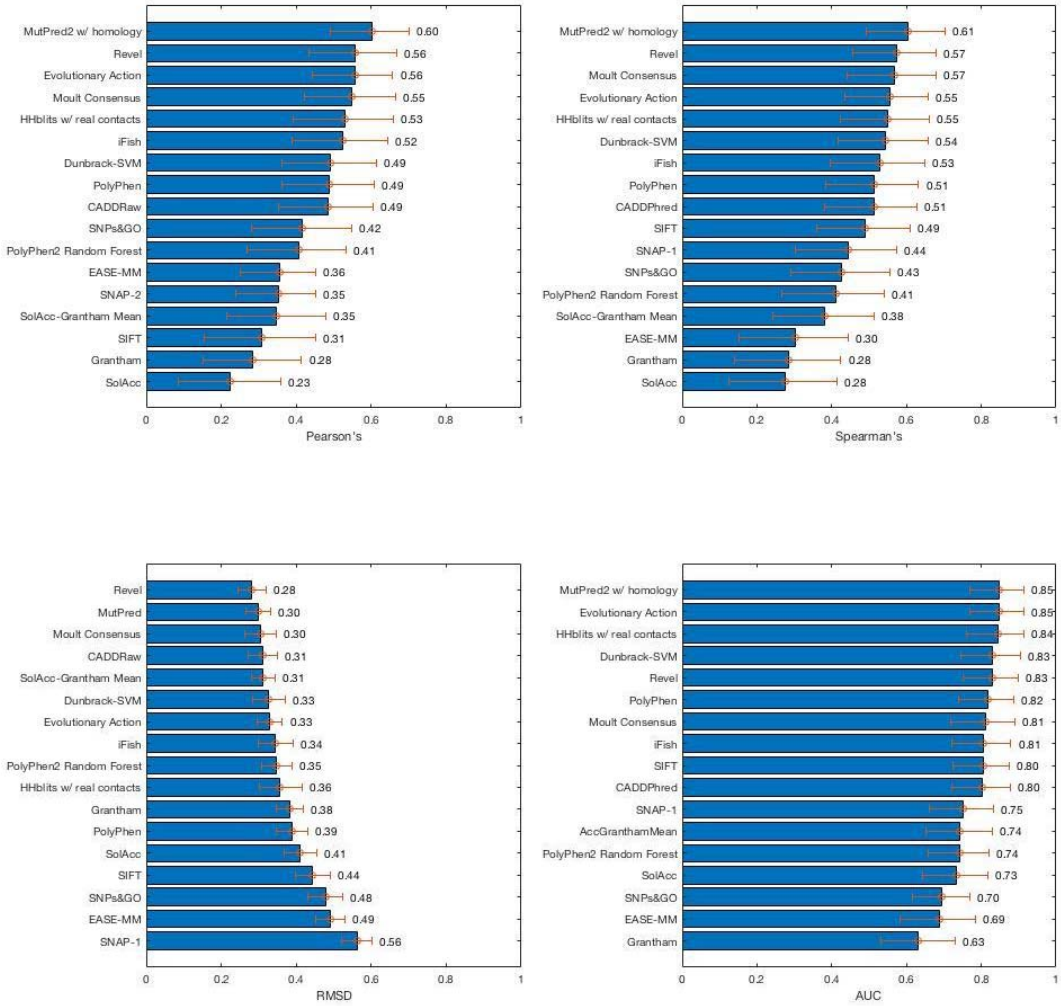
Supplemental Figure: All vs All heatmap.

Without significance.



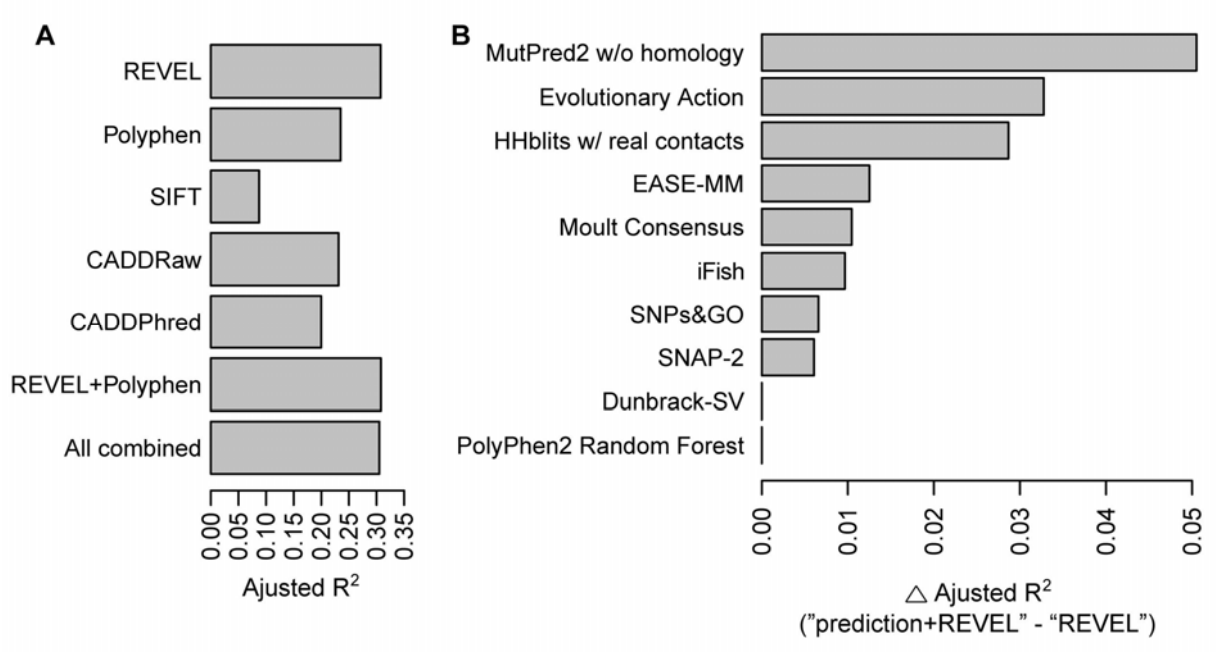
Supplemental Figure: Evaluation Metrics

Various statistics with supplemental models.



Supplemental Figure: Linear Mixed Model

Percentage of variance in predictions explained by each model



Supplemental Data

This spreadsheet has participant information and some results

Supplemental Table

This spreadsheet has supplemental results.

Supplemental Table

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. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Andrade, F., Aldamiz-Echevarria, L., Llarena, M., & Couce, M. L. (2015). Sanfilippo syndrome: Overall review. *Pediatr Int, 57*(3), 331-338. doi:10.1111/ped.12636

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

Carvalho, M. A., Couch, F. J., & Monteiro, A. N. (2007). Functional assays for BRCA1 and BRCA2. *Int J Biochem Cell Biol, 39*(2), 298-310. doi: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. doi: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. doi:10.1371/journal.pone.0200008
Coutinho, M. F., Lacerda, L., & Alves, S. (2012). Glycosaminoglycan storage disorders: a review. *Biochem Res Int, 2012*, 471325. doi:10.1155/2012/471325
Gaffke, L., Pierzynowska, K., Piotrowska, E., & Wegrzyn, G. (2018). How close are we to therapies for Sanfilippo disease? *Metab Brain Dis, 33*(1), 1-10. doi: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. *Hum Mutat, 38*(5), 569-580. doi: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. *Hum Mutat, 38*(9), 1039-1041. doi: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. *Am J Hum Genet, 99*(4), 877-885. doi:10.1016/j.ajhg.2016.08.016
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 Sci, 23*(12), 1650-1666. doi:10.1002/pro.2552
Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi:10.1038/nprot.2009.86
Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. doi:10.1038/nature19057
Meijer, O. L. M., Te Brinke, H., Ofman, R., L, I. J., Wijburg, F. A., & van Vlies, N. (2017). Processing of mutant N-acetyl-alpha-glucosaminidase in mucopolysaccharidosis type IIIB fibroblasts cultured at low temperature. *Mol Genet Metab, 122*(1-2), 100-106. doi:10.1016/j.ymgme.2017.07.005
Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res, 47*(D1), D886-D894. doi:10.1093/nar/gky1016
Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. L. Q. A. (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. *Genet Med, 17*(5), 405-424. doi:10.1038/gim.2015.30
Rost, B., & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins, 20*(3), 216-226. doi:10.1002/prot.340200303
Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., . . . Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat, 21*(6), 577-581. doi:10.1002/humu.10212

Zelei, T., Csetneki, K., Voko, Z., & Siffel, C. (2018). Epidemiology of Sanfilippo syndrome: results of a systematic literature review. *Orphanet J Rare Dis, 13*(1), 53. doi:10.1186/s13023-018-0796-4

Tang, H., & Thomas, P. D. (2016). Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics, 203*(2), 635-647. doi:10.1534/genetics.116.190033