Received: 16 April 2019 | Revised: 2 June 2019 | Accepted: 9 June 2019

DOI: 10.1002/humu.23843

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

SPECIAL ARTICLE

Evaluating the predictions of the protein stability change upon single amino acid substitutions for the FXN CAGI5 challenge

Castrense Savojardo[1]* | Maria Petrosino[2]* | Giulia Babbi[1] | Samuele Bovo[1] | Carles Corbi-Verge[3] | Rita Casadio[1,4] | Piero Fariselli[5] | Lukas Folkman[6] | Aditi Garg[7] | Mostafa Karimi[8] | Panagiotis Katsonis[9] | Philip M. Kim[3,10,11] | Olivier Lichtarge[9,12,13,14] | Pier Luigi Martelli[1] | Alessandra Pasquo[15] | Debnath Pal[7] | Yang Shen[8] | Alexey V. Strokach[11] | Paola Turina[16] | Yaoqi Zhou[6,17] | Gaia Andreoletti[18] | Steven E. Brenner[18] | Roberta Chiaraluce[2] | Valerio Consalvi[2] | Emidio Capriotti[16]

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

[2]Department of Biochemical Sciences "A. Rossi Fanelli", Sapienza University of Roma, Roma, Italy

[3]Donnelly Center for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada

[4]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), Bari, Italy

[5]Department of Medical Sciences, University of Torino, Torino, Italy

[6]School of Information and Communication Technology, Griffith University, Southport, Queensland, Australia

[7]Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, India

[8]Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas

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

[10]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

[11]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

[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]Diagnostics and Metrology Laboratory, FSN-TECFIS-DIM, ENEA CR Frascati, Frascati, Italy

[16]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

[17]Institute for Glycomics, Griffith University, Southport, Queensland, Australia

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

Correspondence
Roberta Chiaraluce and Valerio Consalvi, Department of Biochemical Sciences "A. Rossi Fanelli", Sapienza University of Roma, Piazzale Aldo Moro 5, 00185 Roma, Italy.
Email: roberta.chiaraluce@uniroma1.it (R. C.) and valerio.consalvi@uniroma1.it (V. C.).

Emidio Capriotti, Department of Pharmacy and Biotechnology, University of Bologna, Via Francesco Selmi 3, 40126 Bologna, Italy.
Email: emidio.capriotti@unibo.it

Abstract

Frataxin (FXN) is a highly conserved protein found in prokaryotes and eukaryotes that is required for efficient regulation of cellular iron homeostasis. Experimental evidence associates amino acid substitutions of the FXN to Friedreich Ataxia, a neurodegenerative disorder. Recently, new thermodynamic experiments have been performed to study the impact of somatic variations identified in cancer tissues on protein stability. The Critical Assessment of Genome Interpretation (CAGI) data provider at the University of Rome measured the unfolding free energy of a set of variants (FXN challenge data set) with far-UV circular dichroism and intrinsic

*Castrense Savojardo and Maria Petrosino are the co-first authors.

Funding information

NIH, Grant/Award Numbers: U41 HG007346, R13 HG006650; NSERC Discovery Grant, Grant/Award Number: 386671; National Institute of General Medical Sciences, Grant/Award Number: R35GM124952; Australia Research Council, Grant/Award Number: DP180102060; National Health and Medical Research Council, Grant/Award Number: 1121629

fluorescence spectra. These values have been used to calculate the change in unfolding free energy between the variant and wild-type proteins at zero concentration of denaturant ($\Delta\Delta G^{H_2O}$). The FXN challenge data set, composed of eight amino acid substitutions, was used to evaluate the performance of the current computational methods for predicting the $\Delta\Delta G^{H_2O}$ value associated with the variants and to classify them as destabilizing and not destabilizing. For the fifth edition of CAGI, six independent research groups from Asia, Australia, Europe, and North America submitted 12 sets of predictions from different approaches. In this paper, we report the results of our assessment and discuss the limitations of the tested algorithms.

KEYWORDS

machine learning, protein folding, protein stability, single amino acid variant, free energy change

1 | INTRODUCTION

The human frataxin (FXN) is a protein localized in the mitochondria and cytoplasm of the cells that promotes the heme biosynthesis, the assembly, and repair of iron-sulfur clusters by delivering Fe^{2+} to proteins involved in these pathways. Frataxin may play a role in the protection against iron-catalyzed oxidative stress (Lupoli, Vannocci, Longo, Niccolai, & Pastore, 2018).

FXN single-nucleotide variants have been associated with Friedreich Ataxia (MIM# 229300), a degenerative disorder primarily affecting the nervous system (Pandolfo, 2008). Moreover, FXN might play a role in cancer as previous studies have shown that it protects tumor cells against oxidative stress and apoptosis but also acts as a tumor suppressor (Guccini et al., 2011; Schulz et al., 2006). The Catalog of Somatic Mutations in Cancer (COSMIC) database (Tate et al., 2019) collects a set of *FXN* somatic variations identified in cancer tissues. To investigate the possible thermodynamic effect of those variations on protein stability a subset of eight variants were expressed as a soluble recombinant protein in *Escherichia coli* (Petrosino et al., 2019). For this data set of amino acid substitutions, the stability of the variant proteins is experimentally measured with circular dichroism and fluorescence and compared with wild type. These measures have been used for the FXN challenge of the fifth edition of the Critical Assessment of Genome Interpretation (CAGI5). For the FXN challenge participants were asked to predict the variation of free energy change at zero concentration of denaturant $\Delta\Delta G^{H_2O}$ upon single-point protein variation. During the last decades, several methods have been developed to predict the impact of amino acid variants on protein stability (Compiani & Capriotti, 2013). These available algorithms are mainly based on energy functions designed to assess the stability free energy of the protein and its variants and/or machine-learning-based methods trained to predict the stability changes upon variation. In this manuscript, we scored the performance of six research groups in predicting the measured $\Delta\Delta G^{H_2O}$ value (regression task) and its class (classification task) for eight FXN single amino acid variants. The performances of all the groups are compared with those achieved by state-of-the-art methods (Capriotti, Fariselli, & Casadio, 2005; Guerois, Nielsen, & Serrano, 2002) to estimate the possible improvement with respect to previously developed algorithms. For the calibration of the predictions, previous experimental thermodynamic data on a different set of variants (Faraj, Gonzalez-Lebrero, Roman, & Santos, 2016) were used as a reference.

2 | MATERIAL AND METHODS

2.1 | Data set and classification

The CAGI5 FXN challenge data set consists of eight coding variants of the *FXN* gene. These variants encode for single amino acid substitutions reported in the COSMIC database. A representation of the variation sites in the three-dimensional structure of FXN (PDB: 1EKG) is provided in Figure 1.

For each protein variant, the unfolding free energy change (ΔG_u) at different denaturant concentrations was experimentally determined with circular dichroism and fluorescence. These measures were used to calculate the unfolding free energy at zero concentration of denaturant (ΔG^{H_2O}). Finally, the change of ΔG^{H_2O} of the variant protein $\Delta\Delta G^{H_2O}$ was calculated using the following equation:

$$\Delta\Delta G^{H_2O} = \Delta G^{H_2O}_{mut} - \Delta G^{H_2O}_{wt} \qquad (1)$$

The average experimental values of the $\Delta\Delta G^{H_2O}$ obtained by circular dichroism and fluorescence (Table S1) were used for the challenge. An exception to this procedure is the case of the variant p.W173C which does not fold. In this case, we assumed its unfolding free energy equal to 0 kcal/mol and the $\Delta\Delta G^{H_2O}$ equal to the negative of the ΔG^{H_2O} of the wild-type protein.

We also assessed the quality of the predictions by calculating the performance of the methods in classification mode. For this task we selected a threshold of −1.0 kcal/mol to discriminate between destabilizing ($\Delta\Delta G^{H_2O} < -1.0$ kcal/mol) and not destabilizing variants ($\Delta\Delta G^{H_2O} \geq -1.0$ kcal/mol). With this assumption, five variations in the data set were classified as destabilizing and the remaining three as

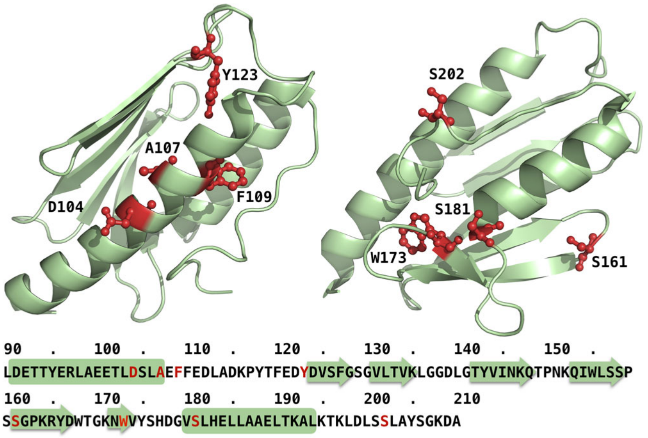
Y123
S202
A107
F109
D104
S181
W173
S161
90 . 100 . 110 . 120 . 130 . 140 . 150 .
LDETTYERLAEETLDSLAEFFEDLADKPYTFEDYDVSFGSGVLTVKLGGDLGTYVINKQTPNKQIWLSSP
160 . 170 . 180 . 190 . 200 . 210
SSGPKRYDWTGKNWVYSHDGVSLHELLAAELTKALKTKLDLSSLAYSGKDA

FIGURE 1 Mapping of the eight variation sites of the frataxin challenge data set on the three-dimensional structure of the protein (PDB: 1EKG)

not destabilizing. A visual representation of the similarity between the $\Delta\Delta G^{H_2O}$ obtained by different experimental techniques (circular dichroism and fluorescence) and the classification of the variations are shown in Figure S1.

The final set of eight variations with the relative average $\Delta\Delta G^{H_2O}$ and their experimental errors are reported in Table 1.

2.2 | Experimental measures

Human *FXN* variants were obtained with specific mutagenesis primers with polymerase chain reaction, using wild type as a template. Wild type and variants were then expressed in *E. coli* and purified. The structural conformation of the variants was compared to that of the wild type by monitoring the near and far-UV circular dichroism and intrinsic fluorescence spectra. The thermodynamic stability was measured at different concentrations of denaturant (Urea) by monitoring the spectral changes (far-UV circular dichroism and intrinsic fluorescence emission) induced by urea. The spectral changes were extrapolated to zero denaturant concentration ($\Delta\Delta G^{H_2O}$). For equilibrium transition studies, FXN wild type and variants were incubated at 20°C at increasing concentrations of urea (0–9M). After 10 min, equilibrium was reached and both intrinsic fluorescence emission and far-UV CD spectra were recorded in parallel at 20°C. To test the reversibility of the unfolding, FXN wild type and variants were unfolded at 20°C in 9.0M urea. After 10 min, refolding was started by 10-fold dilution of the unfolding mixture at 20°C into solutions of the same buffer used for unfolding containing decreasing urea concentrations. After 24 hr, intrinsic fluorescence emission and far-UV CD spectra were recorded at 20°C. All denaturation experiments were performed in triplicate. For thermal denaturation studies, FXN wild type and variants were heated from 20 to 95°C and then cooled from 95 to 20°C. The dichroic activity at 222 nm was continuously monitored every 0.5°C. Melting temperature (T_m) values were calculated by taking the first derivative of the ellipticity at 222 nm with respect to temperature. All denaturation experiments were performed in triplicate. More details about the procedure for the calculation of the $\Delta\Delta G^{H_2O}$ and the analysis of the thermodynamic data are described in supplementary materials.

2.3 | Challenge participants and prediction methods

Six groups participated in the CAGI5 FXN challenge by submitting a total or 12 sets of predictions using different procedures. The Lichtarge Lab at the Baylor College of Medicine, referred as Group 1, submitted one set of predictions (G1-1) using Evolutionary Action method (Katsonis & Lichtarge, 2014). The output of the program was normalized to return $\Delta\Delta G^{H_2O}$ values between 0 and −3 kcal/mol. The Biocomputing Group (Group 2) from the University of Bologna provided one batch of predictions (G2-1) using INPS-3D (Savojardo, Fariselli, Martelli, & Casadio, 2016). For this challenge, the 1EKG structure from the Protein Data Bank was considered as wild type. The Zhou Lab at the Griffith

TABLE 1 Frataxin challenge data set of amino acid substitutions

DNA (hg38)	mRNA (NM_000144.4)	Protein (NP_000135.2)	$\Delta\Delta G^{H_2O}$kcal/mol	Destabilizing
chr9:g.69053187A>G	c.311A>G	p.D104G	0.4±0.4	No
chr9:g.69053196C>T	c.320C>T	p.A107V	0.0±0.6	No
chr9:g.69053201T>C	c.325T>C	p.F109L	−2.8±0.4	Yes
chr9:g.69053244A>C	c.368A>C	p.Y123S	−5.1±0.3	Yes
chr9:g.69065035G>T	c.482G>T	p.S161I	−3.1±0.4	Yes
chr9:g.69072648G>T	c.519G>T	p.W173C	−9.5±0.3[a]	Yes
chr9:g.69072671C>T	c.542C>T	p.S181F	−3.0±0.4	Yes
chr9:g.69072734C>T	c.605C>T	p.S202F	−0.2±0.4	No

Note: The mean variation of unfolding free energy change at zero solvent concentration ($\Delta\Delta G^{H_2O}$) is calculated as the mean $\Delta\Delta G^{H_2O}$ values of fluorescence and circular dichroism experiments (see Table S1). The standard deviation (σ) is obtained summing the errors associated with both types of measures. Destabilizing are the variants with $\Delta\Delta G^{H_2O} < -1.0$ kcal/mol.

[a] The variant p.W173C does not fold into a three-dimensional. Thus, for calculating the $\Delta\Delta G^{H_2O}$ of p.W173C we assumed that its $\Delta G^{H_2O} = 0$ kcal/mol. It follows that $\Delta\Delta G^{H_2O}$ is equal to $-$ ΔG^{H_2O} of the wild type, which is −9.50 kcal/mol.

University, labeled as Group 3, submitted three sets of predictions (G3-1, G3-2, and G3-3) using Evolutionary, Amino acid, and Structural Encodings with Multiple Models (EASE-MM) algorithm (Folkman, Stantic, Sattar, & Zhou, 2016). For the assessment we considered only one set of predictions (G3-1) because the three batches of predictions returned the same $\Delta\Delta G^{H_2O}$ values. The Shen Lab at the Texas A&M University (Group 4) submitted two groups of predictions (G4-1, G4-2) using Interconnected Cost Function Network (iCFN; Karimi & Shen, 2018). This method was modified to fit the experimental $\Delta\Delta G^{H_2O}$ values for FXN variants from a previous work (Correia, Pastore, Adinolfi, Pastore, & Gomes, 2008). The Pal Lab at the Indian Institute of Science in Bangalore, labeled as Group 5, submitted two batches of unscaled predictions (G5-1, G5-2) using GROMACS (Van Der Spoel et al., 2005). This approach uses molecular dynamics simulations to estimate the stability of unfolded and native conformations for the wild type and variants. The Kim Lab at the University of Toronto (Group 6) submitted three batches of predictions (G6-1, G6-2, and G6-3) using the ELAPSIC algorithm (Berliner, Teyra, Colak, Garcia Lopez, & Kim, 2014; Witvliet et al., 2016). ELAPSIC is a meta-predictor that combines predictions from other methods with sequence and structure-based features using a gradient boosting algorithm. During the assessment, we observed that predictions submitted by Group 6 showed a strong negative correlation with the experimental data. This is due to the difference between the challenge's request of predicting the variation of unfolding free energy change ($\Delta\Delta G_u$) and the predictions of folding free energy change ($\Delta\Delta G_f$) submitted by Kim's Lab. For this reason, we also scored the inverse of the three sets of Group 6 predictions (G6-R1, G6-R2, and G6-R3).

Finally, to estimate the improvement of the performance between more recent algorithms and state-of-the-art methods, we included in our assessment the performance of FoldX (Guerois et al., 2002) and I-Mutant2.0 (Capriotti et al., 2005).

In the supplementary materials, we described more in detail the methods and procedures used by each group to perform their predictions. A summary of all the submissions is reported in Table S2.

2.4 | Prediction assessment

For the evaluation of the predictions we considered eight measures of performance for the regression and classification tasks defined in supplementary materials (Section S3). Comparing the predicted and experimental values of $\Delta\Delta G^{H_2O}$ of each protein variant, we calculated three types of correlations (Person, Spearman, and Kendall-Tau) and two types of errors (root mean square error [RMSE] and the mean absolute error [MAE]). Furthermore, we considered a threshold of −1.0 kcal/mol for classifying variants in destabilizing ($\Delta\Delta G^{H_2O}$ < -1.0 kcal/mol) and not destabilizing variants ($\Delta\Delta G^{H_2O}$ ≥ -1.0 kcal/mol). Using this threshold for the binary classification task, we scored the predictions calculating the balanced accuracy (BQ_2), the Matthews correlation coefficient (MCC) and the area under the receiving operator characteristic curve (AUC). Finally, we ranked all the submissions considering each one of the eight measures of performance and by calculating the average value of the ranks, which is used to select the best predictions. In the second part of the assessment, we determined the significance of the differences between the performance of two methods with the Kolmogorov–Smirnov (KS) test. The KS test was used to compare the distribution of the ranks for each measure of performance.

Another important issue in the evaluation of the most reliable predictions is the presence of outliers in the experimental data set. With outlier, we refer to an experimental measure that, for different reasons, is considered to be less accurate or reliable than others. In general, it is expected that most of the methods will fail in the prediction of the outliers. According to this assumption, in our assessment, we also scored the performance of the algorithms removing the outliers from the initial FXN challenge data set. In particular, for this calculation we removed from the data set the variant p.W173C for which the ΔG^{H_2O} was set to 0 kcal/mol because it was not folding properly.

The definitions of the eight measures of performance considered for this assessment are reported in Supporting Information Materials.

3 | RESULTS

3.1 | Assessment and performance evaluation

In our assessment, we first evaluated the success of the participants in predicting the value of $\Delta\Delta G^{H_2O}$. For this task, we calculated five performance measures, three of which score the correlations between experimental and predicted data ($r_{P,}$ r_S, and r_{KT}) and two the prediction errors (RMSE and MAE). The performance in the regression task for the best predictions of each group are reported in Figure S2. According to the calculated scores, Group 3 resulted in the best predictions reaching the highest Pearson correlation coefficient (r_P = 0.84) and lowest root-mean-square-deviation (RMSE = 2.94 kcal/mol). Our analysis also showed that Group 6 resulted in negative values of the Pearson correlation coefficient close to −1 (r_P = −0.89). Assuming that Group 6 predicted the variation of the $\Delta\Delta G^{H_2O}$ of folding instead of the unfolding, we decided to include in our assessment the opposite of the predictions submitted by Group 6. The performances of participants were compared with those achieved by state-of-the-art methods by including in our assessment the predictions returned by FoldX and I-Mutant2.0. Furthermore, we combined the regression measures with three classification scores (BQ_2, MCC, and AUC) obtained using a threshold of −1.0 kcal/mol for discriminating between destabilizing and not destabilizing variants. The assessment, including eight scores of performance sorted by the average of the rank orders of each method, is summarized in Table 2.

The results showed that the opposite predictions of Submission 1 from Kim Lab (G6-R1) achieved the top average rank calculated over the eight measures of performance. It is worth noting that FoldX scored second in the ranking achieving on average lower performance on the binary classification task and better results in the prediction of the $\Delta\Delta G^{H_2O}$ value with respect to the Kim's Lab R1 submission. Additional details about the comparison between Kim's Lab R1 submission and the prediction of the state-of-the-art methods are shown in Figure 2.

TABLE 2 Assessment of the predictions of the six groups and the state-of-the-art methods (FoldX and I-Mutant2.0)

Group	Submission	r_P	r_S	r_{KT}	RMSE	MAE	BQ_2	MCC	AUC	Rank
Kim Lab	G6-R1[a]	0.82	0.69	0.50	2.4	1.7	0.80	0.60	0.93	1.75
FoldX	–	0.84	0.64	0.57	2.2	1.7	0.73	0.47	0.87	2.00
Zhou Lab	G3-1	0.85	0.64	0.64	3.0	2.3	0.70	0.45	0.80	2.88
Kim Lab	G6-R2[a]	0.71	0.57	0.43	2.7	2.0	0.63	0.26	0.80	4.13
Biocomp	G2-1	0.74	0.52	0.36	3.2	2.3	0.80	0.60	0.80	4.25
Lichtarge Lab	G1-1	0.46	0.60	0.50	3.1	2.2	0.63	0.26	0.87	4.38
I-Mutant2.0	–	0.75	0.55	0.43	3.3	2.5	0.70	0.45	0.73	4.75
Kim Lab	G6-R3[a]	0.89	0.57	0.50	3.9	3.7	0.50	0.00	0.80	5.25
Shen Lab	G4-2	−0.02	0.12	0.07	4.1	2.6	0.70	0.45	0.60	7.00
Shen Lab	G4-1	−0.09	0.17	0.07	3.9	2.7	0.60	0.29	0.60	7.25
Pal Lab	G5-1	0.57	0.43	0.29	41	36	0.63	0.26	0.67	7.88
Kim Lab	G6-2	−0.71	−0.57	−0.43	6.2	4.4	0.50	0.00	0.20	9.13
Kim Lab	G6-1	−0.89	−0.57	−0.50	10.9	9.5	0.50	0.00	0.20	10.00
Kim Lab	G6-3	−0.82	−0.69	−0.50	6.4	4.5	0.50	0.00	0.07	10.00
Pal Lab	G5-2	−0.42	−0.64	−0.50	1,441	1,378	0.50	0.00	0.27	10.13

Note: The eight measures of performance are defined in supplementary materials. Zhou Lab submitted three sets of predictions with the same $\Delta\Delta G^{H_2O}$ values. For this reason, we reported only the measure of performance for Submission 1.

[a] The submissions of Kim's Lab that were reversed. Confusion matrices for the binary classification are reported in Table S3.

3.2 | Data set outlier

The analysis of all the submitted predictions revealed that on average all the groups failed in the prediction of the $\Delta\Delta G^{H_2O}$ for the variant p.W173C. Excluding Group 5, for this variant, the difference between the average predicted and experimental $\Delta\Delta G^{H_2O}$ is approximately 6 kcal/mol (see Figure 3). A possible motivation of the strong discrepancy between predicted and experimental $\Delta\Delta G^{H_2O}$ values is the partial indetermination of the ΔG^{H_2O} of the unfolding of the p.W173C variant. Indeed, this protein variant did not fold into a three-dimensional structure. For this reason, we arbitrarily assigned to the p.W173C variant a ΔG^{H_2O} equal to 0 kcal/mol, which implies an equal fraction of folded and unfolded protein at equilibrium. According to this observation, the protein variant p.W173C was considered an outlier and we performed a second assessment of the predictions removing it from the FXN challenge data set. Sorting all the predictions, according to the average ranking based on the eight measures of performance, we observed that the G6-R1 from Kim's Lab and FoldX predictions scored in the first and second position, respectively. The difference with respect to the previous assessment including all the FXN variants is the third position in the ranking achieved by the Biocomputing Group. As expected for all the submissions the RMSE and MAE values decreased. Thus, removing the variant p.W173C from the data set, the average RMSE for the top four ranking submissions was approximately 1.7 kcal/mol while it was approximately 2.6 kcal/mol for all the variants.

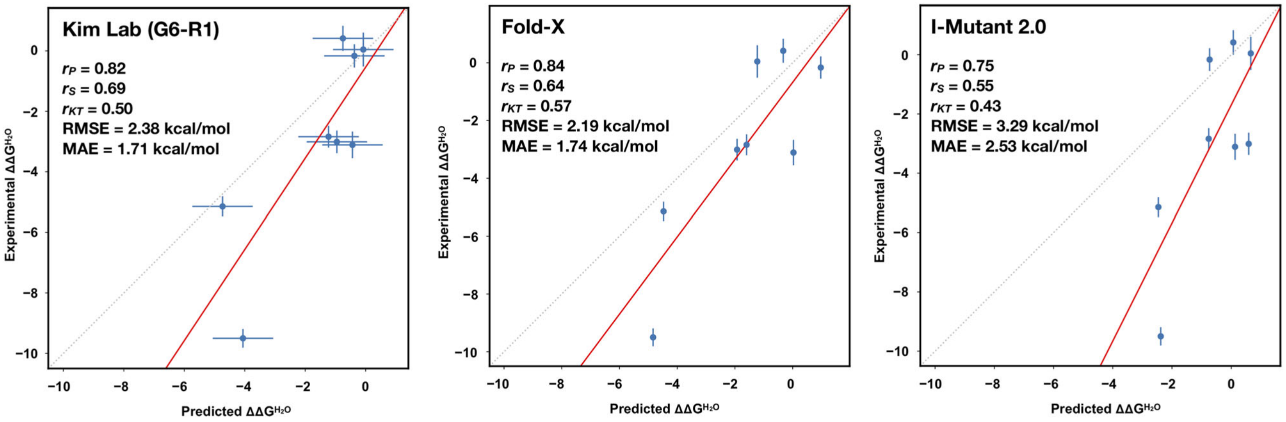

FIGURE 2 Comparison between the performance achieved in the regression task by the top ranking submission from Kim Lab (G6-R1), FoldX, and I-Mutant2.0. r_P, r_S, r_{KT}, RMSE, and MAE are defined in Supporting Information Materials. MAE, mean absolute error; RMSE, root mean square error

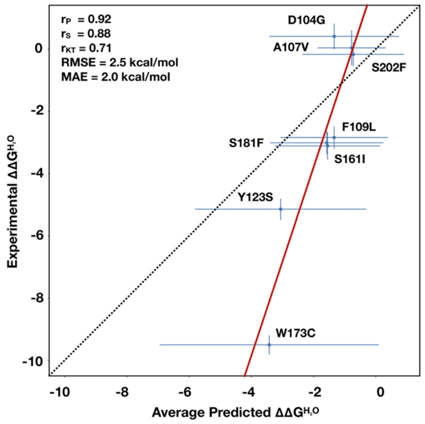

FIGURE 3 Linear regression between the average predicted and experimental $\Delta\Delta G^{H_2O}$. The average predictions are calculated excluding the prediction from Group 5 and considering only one submission from Group 3. r_P, r_S, r_{KT}, RMSE, and MAE are defined in Supporting Information Materials. MAE, mean absolute error; RMSE, root mean square error

3.3 | Methods and predictions similarity

In the last part of our assessment, we compared all the submissions to calculate the level of similarity among the predictions. For the comparison, we assigned to each submission a ranking vector based on the eight measures of performance defined in Supporting Information Materials. The statistical difference among such vectors was calculated with the KS test. In Figure 4 we summarized our analysis assigning a blue color to the submissions that had significantly different ranking distributions ($p < .05$). Contrarily, red spots are assigned to the pairs of submissions which were statistically indistinguishable. The results showed that R1 (the reverse of Submission 1 from Kim's Lab) is not statistically different from FoldX. This observation is consistent with the fact that ELAPSIC algorithm, used by Kim's group, includes the calculation of $\Delta\Delta G$ values with FoldX. Our analysis also revealed that after Kim's lab and FoldX predictions, the submissions from the Zhou Lab, Biocomputing Group, and Lichtarge Lab were statistically indistinguishable. The performances of methods from the previous groups are comparable with those achieved by I-Mutant2.0. These observations, which are valid for the whole FXN data set (Figure 4a), are partially confirmed after removing the p.W173C variant. In this case, the ranking of the predictions from Kim's Lab is statistically different from FoldX (Figure 4b) while the second group of submissions (Biocomputing Group, Zhou Lab, and Lichtarge Lab) remains statistically indistinguishable.

4 | DISCUSSION

The assessment of the FXN challenge of the CAGI5 experiment provided an opportunity to evaluate the performance of the available variant annotation methods for predicting the impact of single amino acid variations on protein stability. In detail, we scored each submission by considering the performance of the corresponding method in predicting the $\Delta\Delta G^{H_2O}$ values (regression task) and by correctly classifying the variants in destabilizing and not destabilizing (classification task). The results showed that, in the regression task, the best methods achieved a Pearson correlation coefficient >0.8 and a RMSE <2.4 kcal/mol (see Table 2). After removing from the data set p.W173C, which represents an outlier with respect to all the other

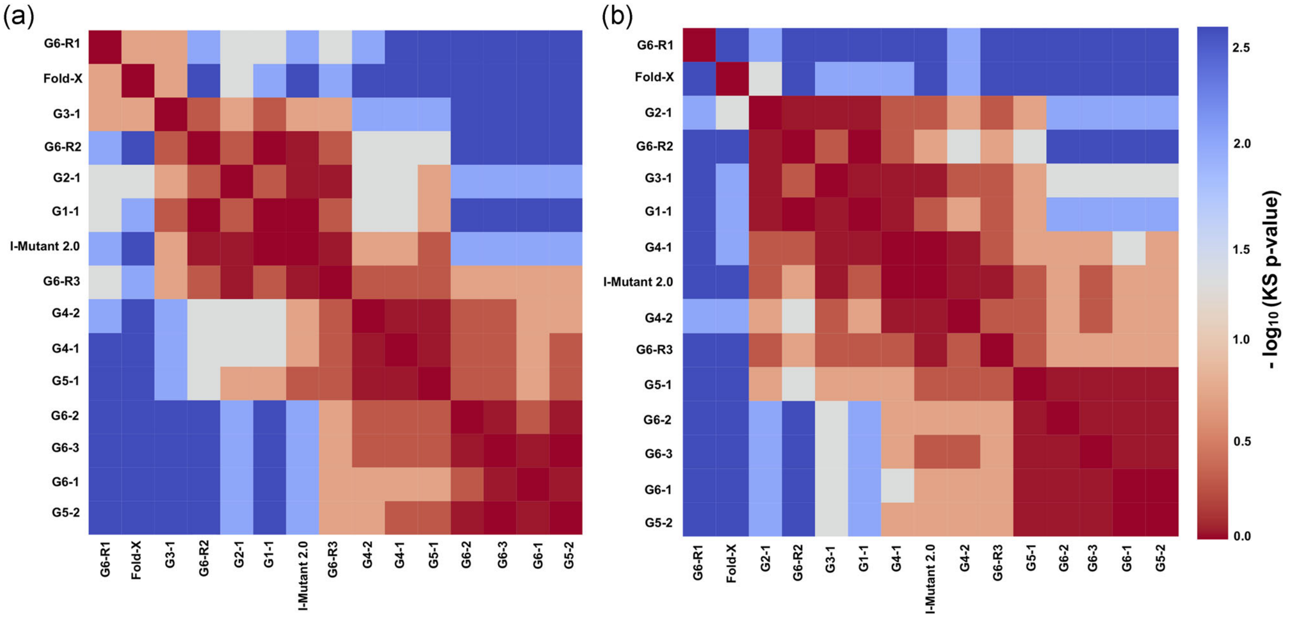

FIGURE 4 Similarity between the predictions based on the Kolmogorov–Smirnov test among the ranking vectors from the eight measures of performance. The color of each cell is proportional to the −log10 of the Kolmogorov–Smirnov p value. Similarities calculated considering the whole frataxin data set and excluding the variant p.W173C are plotted in panels (a) and (b), respectively

TABLE 3 Assessment of the predictions submitted by the 6 groups and returned by state-of-the-art methods (FoldX and I-Mutant2.0) excluding the p.W173C variant

Group	Submission	r_P	r_S	r_{KT}	RMSE	MAE	BQ_2	MCC	AUC	Rank
Kim Lab	G6-R1[a]	0.75	0.57	0.43	1.5	1.2	0.75	0.55	0.92	1.13
FoldX	–	0.73	0.46	0.43	1.5	1.3	0.71	0.42	0.83	2.13
Biocomp	G2-1	0.72	0.32	0.24	1.9	1.6	0.75	0.55	0.75	3.50
Kim Lab	G6-R2[a]	0.65	0.39	0.33	1.8	1.4	0.58	0.17	0.75	3.75
Zhou Lab	G3-1	0.57	0.46	0.52	2.1	1.8	0.63	0.35	0.75	4.25
Lichtarge Lab	G1-1	0.20	0.46	0.43	2.1	1.6	0.58	0.17	0.83	4.38
Shen Lab	G4-1	0.50	0.50	0.33	2.2	1.8	0.63	0.35	0.67	4.88
I-Mutant2.0	–	0.58	0.36	0.33	2.3	1.9	0.63	0.35	0.67	5.25
Shen Lab	G4-2	0.22	0.07	0.05	2.5	1.6	0.75	0.55	0.58	5.63
Kim Lab	G6-R3[a]	0.66	0.36	0.33	4.1	3.8	0.50	0.00	0.75	6.00
Pal Lab	G5-1	0.09	0.14	0.05	38	33	0.58	0.17	0.58	8.13
Kim Lab	G6-2	−0.65	−0.39	−0.33	4.5	3.1	0.50	0.00	0.25	8.50
Kim Lab	G6-3	−0.66	−0.36	−0.33	8.4	7.8	0.50	0.00	0.25	9.13
Kim Lab	G6-1	−0.75	−0.57	−0.43	4.5	3.2	0.50	0.00	0.08	9.50
Pal Lab	G5-2	−0.51	−0.54	−0.43	1,472	1,404	0.50	0.00	0.33	9.63

Note: The eight measures of performance are defined in Supporting Information Materials. Zhou Lab submitted three sets of predictions with the same $\Delta\Delta G^{H_2O}$ values. For this reason, we reported only the measure of performance for Submission 1.

[a]The submissions of Kim's Lab that were reversed. Confusion matrices for the binary classification are reported in Table S4.

variants, the RMSE values of the best submissions decrease below 1.5 kcal/mol (see Table 3). For the classification task, we select a $\Delta\Delta G^{H_2O}$ threshold of −1.0 kcal/mol for discriminating between destabilizing ($\Delta\Delta G^{H_2O}$ < −1.0 kcal/mol) and not destabilizing variants ($\Delta\Delta G^{H_2O}$ ≥ −1.0 kcal/mol). Using such threshold, the best predictions (reversed Submission 1 from Kim's Lab) achieved remarkable BQ_2, MCC, and AUC scoring 0.80, 0.60, and 0.93, respectively (see Table 2). Slightly lower performance was obtained when the p.W173C variant was removed from the FXN data set. The evaluation of the similarities among the submissions showed that although the reverse Submission 1 (R1) from ELAPSIC scores better than FoldX for the classification task, the difference between the ranking distributions of the two methods is not significant (KS p = .19). A significant difference between the ranking distribution of G6-R1 Kim's Lab and FoldX predictions is found when the p.W173C variant is removed from the data set. In this case, Kim's Lab R1 submission ranks in the first position for seven over eight measures of performance considered in our assessment. Comparing the ranking distribution of the second block of groups we found that the predictions from Zhou Lab, Biocomputing Group, and Lichtarge Lab are statistically indistinguishable. Finally, the analysis of the predictions from Group 5, which adopted a molecular dynamics-based approach, shows the largest $\Delta\Delta G^{H_2O}$ resulting in the highest RMSE values. As suggested by the Group 5 submitters, their predictions could have been improved by normalizing the energies obtained from the simulations.

In conclusion, the assessment of the predictions submitted for the FXN challenge confirmed that the methods for predicting the protein stability change upon variation achieved a good level of performance, especially in the classification task. For the prediction of the $\Delta\Delta G^{H_2O}$ values, the best methods achieved good performance in terms of correlation coefficient but the error is still high (RMSE ~2.0 kcal/mol). Finally, we observed that all the algorithms fail to predict the $\Delta\Delta G^{H_2O}$ of p.W173C. variant which has a high impact on protein stability. Our hypothesis is that the high error level is due to the low number of experimental data for highly destabilizing variants in the training set. This hypothesis is consistent with the observation that machine-learning-based methods such as INPS-3D and EASE-MM resulted in higher RMSE than FoldX which implements an energy-functions-based approach.

Although the selection of a single protein and the limited number of variants in FXN challenge data set do not allow to generalize the results of our assessment, nevertheless it is noteworthy that the most accurate methods achieved good performance in terms of correlation coefficient and RMSE.

ACKNOWLEDGMENTS

P. M. K. would acknowledge support from a NSERC Discovery Grant (#386671). This project is in part supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM124952 to Y. S.). This work was supported in part by Australia Research Council DP180102060 and by National Health and Medical Research Council (1121629) of Australia to Y. Z. We acknowledge John Moult for his support in the organization of the CAGI5 FXN challenge. The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding authors.

ORCID

Castrense Savojardo http://orcid.org/0000-0002-7359-0633
Lukas Folkman http://orcid.org/0000-0002-5811-8875
Panagiotis Katsonis http://orcid.org/0000-0002-7172-1644
Yang Shen http://orcid.org/0000-0002-1703-7796
Gaia Andreoletti http://orcid.org/0000-0002-0452-0009
Steven E. Brenner http://orcid.org/0000-0001-7559-6185
Roberta Chiaraluce http://orcid.org/0000-0001-7748-2237
Emidio Capriotti http://orcid.org/0000-0002-2323-0963

REFERENCES

Berliner, N., Teyra, J., Colak, R., Garcia Lopez, S., & Kim, P. M. (2014). Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One*, *9*(9), e107353. https://doi.org/10.1371/journal.pone.0107353

Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, *33*(Web Server issue), W306–W310. https://doi.org/10.1093/nar/gki375

Compiani, M., & Capriotti, E. (2013). Computational and theoretical methods for protein folding. *Biochemistry*, *52*(48), 8601–8624. https://doi.org/10.1021/bi4001529

Correia, A. R., Pastore, C., Adinolfi, S., Pastore, A., & Gomes, C. M. (2008). Dynamics, stability and iron-binding activity of frataxin clinical mutants. *The FEBS Journal*, *275*(14), 3680–3690. https://doi.org/10.1111/j.1742-4658.2008.06512.x

Faraj, S. E., Gonzalez-Lebrero, R. M., Roman, E. A., & Santos, J. (2016). Human frataxin folds via an intermediate state. Role of the C-terminal region. *Scientific Reports*, *6*, 20782. https://doi.org/10.1038/srep20782

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

Guccini, I., Serio, D., Condo, I., Rufini, A., Tomassini, B., Mangiola, A., … Malisan, F. (2011). Frataxin participates to the hypoxia-induced response in tumors. *Cell Death & Disease*, *2*, e123. https://doi.org/10.1038/cddis.2011.5

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology*, *320*(2), 369–387. https://doi.org/10.1016/S0022-2836(02)00442-4

Karimi, M., & Shen, Y. (2018). iCFN: An efficient exact algorithm for multistate protein design. *Bioinformatics*, *34*(17), i811–i820. https://doi.org/10.1093/bioinformatics/bty564

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

Lupoli, F., Vannocci, T., Longo, G., Niccolai, N., & Pastore, A. (2018). The role of oxidative stress in Friedreich's ataxia. *FEBS Letters*, *592*(5), 718–727. https://doi.org/10.1002/1873-3468.12928

Pandolfo, M. (2008). Friedreich ataxia. *Archives of Neurology*, *65*(10), 1296–1303. https://doi.org/10.1001/archneur.65.10.1296

Petrosino, M., Pasquo, A., Novak, L., Toto, A., Gianni, S., Mantuano, E., … Consalvi, V. (2019). Characterization of human frataxin missense variants in cancer tissues. *Human Mutation*, https://doi.org/10.1002/humu.23789

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

Schulz, T. J., Thierbach, R., Voigt, A., Drewes, G., Mietzner, B., Steinberg, P., … Ristow, M. (2006). Induction of oxidative metabolism by mitochondrial frataxin inhibits cancer growth: Otto Warburg revisited. *The Journal of Biological Chemistry*, *281*(2), 977–981. https://doi.org/10.1074/jbc.M511064200

van derSpoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, *26*(16), 1701–1718. https://doi.org/10.1002/jcc.20291

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., … Forbes, S. A. (2019). COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Witvliet, D. K., Strokach, A., Giraldo-Forero, A. F., Teyra, J., Colak, R., & Kim, P. M. (2016). ELASPIC web-server: Proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. *Bioinformatics*, *32*(10), 1589–1591. https://doi.org/10.1093/bioinformatics/btw031

SUPPORTING INFORMATION

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

How to cite this article: Savojardo C, Petrosino M, Babbi G, et al. Evaluating the predictions of the protein stability change upon single amino acid substitutions for the FXN CAGI5 challenge. *Human Mutation*. 2019;40:1392–1399. https://doi.org/10.1002/humu.23843