Macromolecular Sequence, Structure, and Function

fuNTRp: Identifying protein positions for variation driven functional tuning

Miller M[1,*], Vitale D[2] , Rost B[3,4] and Bromberg Y[1,4,5]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08901, USA [2]Columbian College of Arts and Sciences Data Science Program Corcoran Hall, 725 21st Street NW, Washington DC, 20052, USA [3]Department for Bioinformatics and Computational Biology, Technische Universität München, Boltzmannstr. 3, 85748 Garching/Munich, Germany [4]Institute for Advanced Study at Technische Universität München (TUM-IAS), Lichtenbergstraße 2a 85748, Garching/Munich, Germany [5]Department of Genetics, Rutgers University, Human Genetics Institute, Life Sciences Building, 145 Bevier Road, Piscataway, NJ 08854, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Evaluating the impact of non-synonymous genetic variants is essential for uncovering disease associations. Understanding the corresponding changes in protein sequences can also help with synthetic protein design and stability assessments. Even though hundreds of computational approaches addressing this task exist, and more are being developed, there has been little improvement in their performance in the recent years. One of the likely reasons for this lack of progress might be that most approaches use similar sets of gene/protein features for model development, with great emphasis being placed on sequence conservation. While high levels of conservation clearly highlight residues essential for protein activity, much of the *in vivo* observable variation is arguably weaker in its impact and, thus, requires evaluation of a higher level of resolution.

Results: Here we describe ***fu**nction **N**eutral/**T**oggle/**R**heostat **p**redictor* (***funtrp***), a novel computational method that classifies protein positions by type based on the expected range of mutational impacts at that position: Neutral (most mutations have no or weak effects), Rheostat (range of effects; i.e. functional tuning), or Toggle (mostly strong effects). Three conclusions of our work are most salient. We show that our position types do not correlate strongly with the familiar protein features such as conservation or protein disorder. Moreover, we find that position type distribution varies across different enzyme classes. Finally, we demonstrate that position types reflect experimentally derived functional effects, improving performance of existing variant effect predictors and suggesting a way forward for the development of new ones.

Availability: https://services.bromberglab.org/funtrp; Git: https://bitbucket.org/bromberglab/funtrp/

Contact: mmiller@bromberglab.org

Supplementary information: Supplementary data are available online.

1 Introduction

The recent decades have seen significant advances in high-throughput experimentation and growing sophistication in the analyses of the results. Unfortunately, our ability to perform these experimental analyses cannot keep up with the current pace of sequencing for research and medical purposes (Bruse, et al., 2016; Ellinghaus, et al., 2013; Turner, et al., 2016). On the other hand, advanced computational techniques are enabled by, and crucial for, dealing with this onslaught of data.

Consider experimental techniques like Deep Mutational Scanning (DMS) (Fowler, et al., 2010). DMS allows for simultaneous assessment of the effects of hundreds of thousands of genetic variants. It combines high throughput sequencing with the ability to create large protein

libraries, *i.e.* uniting high throughput selection methods with high throughput sequencing methods. Still, large-scale mutant library generation is limited by a number of factors, such as bias in sequencing preparation and time requirements / difficulties of design of meaningful screening and selection methods. Experimental limitations also include sequencing read length, severely limiting the evaluation of co-acting effects between distant residues (Araya and Fowler, 2011). Thus, it is infeasible to experimentally assess, for example, the effects of all non-synonymous Single Nucleotide Polymorphisms (nsSNPs) of a given individual, much less a population. However, the large-scale mutational fitness landscapes resulting from DMS analyses are an exciting resource for the development of new accurate variant effect prediction approaches (Gray, et al., 2018).

Identifying disease-association of the roughly 10,000 protein sequence changing genetic variants of every individual (Bromberg, 2013) is like looking for the needle in a haystack. Finding variants that alter protein function may help, but variant effects are not black and white, having a range of outcomes (Swint-Kruse, 2016). While some variants may only marginally alter ligand affinity, others can induce drastic changes (Walker, et al., 2010). Moreover, while subtle molecular modifications are difficult to detect, in concert with other mutation-driven changes they can cause phenotypic changes (Kowarsch, et al., 2010; Zabalza, et al., 2014).

Single amino acid substitutions caused by nsSNPs are often associated with specific traits (Box, et al., 1997; Duffy, et al., 2007; Shastry, 2009), diseases (de Ligt, et al., 2013; Kumar, et al., 2017), and pharmacological responses (Halushka, et al., 2003). Moreover, targeted mutagenesis of specific protein sites is an essential tool in the synthetic biology toolkit (Sun, et al., 2015). Given the broad range of their possible applications, it is not surprising that many computational algorithms for the prediction of single amino acid substitution effects have been developed (>200; as of January 2018). The different approaches range in algorithm complexity (*e.g.* random forests (Ioannidis, et al., 2016) or meta-servers (Capriotti, et al., 2013), training/development data sets (*e.g.* cancer (Douville, et al., 2013) or stability changes (Capriotti, et al., 2005), and gene/protein features used (*e.g.* conservation or protein structure (Adzhubei, et al., 2010; Bromberg and Rost, 2007; Ng and Henikoff, 2003). However, they still have room for improvement (Dong, et al., 2015; Mahmood, et al., 2017) and despite their increasing number and complexity, there has, arguably, not been a significant improvement in prediction accuracy over the last decade.

Recently, our collaborators (Meinhardt, et al., 2013) had established a new classification of protein (sequence) position types - *Toggle* and *Rheostat* – where mutations in *Toggle* positions were mostly severely disruptive of protein function, while mutations in *Rheostatic* positions had a complete range of effects. We further demonstrated (Miller, et al., 2017) that existing computational predictors fall short of accurately differentiating between neutral and non-neutral mutations in the two position types. Thus, for example, *Toggle* position mutation experimentally shown to have no-effect on protein function, were still deemed as having an effect by most of the evaluated predictors. We concluded from this work that knowledge of position type could improve prediction accuracy.

Until now, *Toggles* and *Rheostats* were characterized on the basis of the distribution of experimentally validated variant effects per protein sequence position (Hodges, et al., 2018). However, experimental evaluation of variant effects is still very limited in comparison to the number of available protein sequences (*e.g.* UniProtKB (The UniProt, 2017)). Moreover, once the variant effect is experimentally determined, its prediction becomes irrelevant. In other words, having to experimentally establish the position type precludes using it as a feature in a variant effect predictor.

Here, we developed a new machine learning approach, ***fu**nction **N**eutral/**T**oggle/**R**heostat **p**redictor* (***fuNTRp***), to predict position types using a curated set of sequence-based features. *funtrp* classifies protein positions by type based on the expected range of mutational impacts possible at each position; *i.e.* at *Neutral* positions most variation will have no or weak effect, at *Rheostat* positions – a range of effects is possible, *i.e.* functional tuning, and at *Toggle* positions mostly strong effects are expected. We found that protein active/functional regions are enriched in *Rheostats* and *Toggles*, with the latter dominating crucial residues (*e.g.* catalytic sites). While these findings are in line with the conservation landscape, we observed lower than expected correlation between conservation and position types, particularly for *Rheostats*. Curiously, we also found that distribution of position types varied across protein classes, slightly differentiating enzymes from non-enzymes and significantly varying between enzyme functional classes. Notably, we showed that position types correlate with experimental effect annotations; *i.e.* we were able to fairly accurately predict mutation effects simply by considering the position type. Combining *funtrp* annotation with outputs of the existing variant effect predictors further improved prediction accuracy.

These findings suggest that knowledge of position types is critical for evaluating functional effects of variants. Thus, *funtrp* predictions could aid the development of improved variant effect prediction methods.

2 Methods

The *funtrp* training/development process is detailed in Fig. 1. The training datasets are summarized in Supplementary Table S1.

2.1 Training datasets and feature extraction

We extracted quantitative deep mutational scanning (DMS) (Araya and Fowler, 2011; Pitt and Ferre-D'Amare, 2010) amino acid substitution effect data for five proteins (Table 1) (Firnberg, et al., 2014; Melamed, et al., 2013; Starita, et al., 2013; Starita, et al., 2015; Wu, et al., 2016). The DMS approach generates a large set of mutations and estimates of their

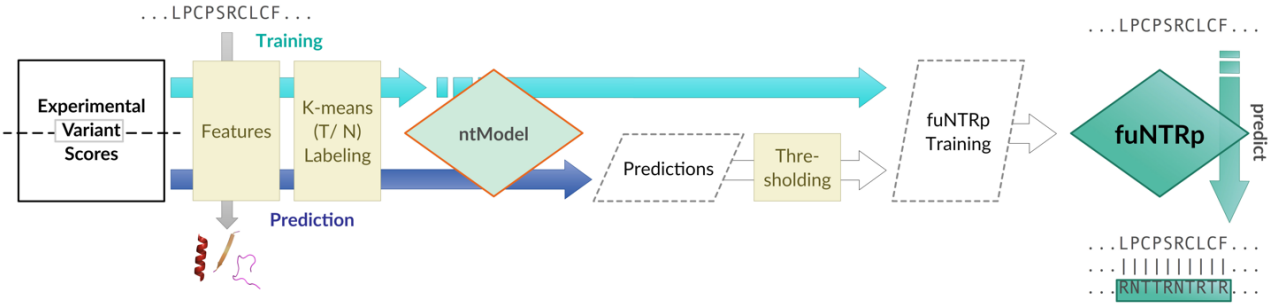

Fig. 1. *funtrp* pipeline. Schematic overview of the *funtrp* pipeline. In training, experimentally measured variant effect scores are extracted for all residues present in selected Deep Mutational Scanning (DMS) datasets. These scores are used in the k-means cluster labeling step to initially label a subset of all (residue) positions as either *Toggle* or *Neutral*. Annotated with a computed set of sequence-based features the subset of cluster-labeled positions is then used to train the *ntModel* to predict the not yet labeled positions from the DMS datasets as either *Toggle* or *Neutral*. After filtering, those are combined with the initially (cluster-labeled) positions and the same set of sequence-based features to train the final *fuNTRp* model.

impacts for every evaluated protein-coding gene. The effects evaluated in this study include impact on E3 ligase activity (Sets 1 and 3), ampicillin resistance (Set 2 and 4), and relative binding affinity of the human Immunoglobulin G F(c) fragment (IgG-FC) (Set 5). Note that from all DMS datasets extracted from the literature only five met our stringent requirements for inclusion into the study, namely: having at least 50 mutated positions, ≥6 variants per position for at least 40% of positions, at least one third of Single Nucleotide Polymorphisms (SNPs) among the variants, wildtype (*wt*) and knockout (*ko*) measurements available, and, notably, available raw datasets in parseable format (data in PDF format and/or not retrievable from contact with the study authors was excluded).

Table 1. Deep mutational scanning datasets used in model training

Gene	Sub-region	Organism	|Variants|	Measured Activity
BRCA1	RING domain	H. sapiens	3080	E3 ligase activity
PAB1	RRM domain	S. cerevisiae	1188	Ampicillin resistance
UBE4B	U-box domain	H. sapiens	926	E3 ligase activity
TEM-1	-	E. coli	5469	Ampicillin resistance
SPG1	GB1	Strepto. sp	467	Binding affinity to IgG

For each protein, effects (scores) of each substitution were standardized using the *wt* measurements reported in the corresponding publication as reference. All scores (including the *wt* and *ko* variant scores) were thus transformed to reflect their absolute distance to *wt*, without differentiating beneficial and deleterious mutations (Eqn. 1).

$$\mathrm{mut_{score}} = |\mathrm{mut_{score}} - wt_{\mathrm{score}}| \quad (1)$$

We further computed ten sequence-based features (Table 2) for each protein. These features included basic amino acid properties, as well as structural properties generated using a Dockerized (Docker, 2018) version of PredictProtein (default parameters) (Yachdav, et al., 2014). Features were chosen based on biological relevance to reflect a broad range of properties associated with protein function.

2.1.1 Filtering sequence positions

In total, our five proteins comprised 822 amino acids (residues) and 11,130 substitutions with measured effect scores. We removed the two unknown amino acids (labeled X in sequence), leaving 820 residues. Note that the number of available experimental scores per residue varied between and within datasets. Also note that only half of the available variants (5,423 of 11,130) satisfied the SNP-possible criteria, *i.e.* the observed amino acid substitutions required no more than one nucleotide change with respect to the wildtype amino acid. Note, we did NOT go back to the gene sequence to find the affected codon, but rather designated as SNP-possible any single nucleotide codon to codon changes representing the wt and substituting amino acids. As SNPs are more common than multi-nucleotide changes, using only the SNP-possible variants more closely mirrored natural selection acting on genes/proteins. This approach also allowed us to avoid compounding effects of the later mutagenesis round mutations, which may have impacted activity more severely.

We removed from any further consideration the 57 positions with fewer than three variant scores as we could not reliably validate any predictions for these positions (7% of 820). With a total of six variants, *Tryptophan* (W) was the amino acid with the least (six) SNP-possible substitutions. Thus, selecting for the first round of training only the positions with at least six SNP-possible variants enabled us to include all *wt* residues, as well as to retain positions with a sufficient number of variants to ensure accurate classification. Thus, we set aside 172 positions (three to five variants; *FewVariants* set) and retained 591 positions (72% of 820) with at least six SNP-possible variants in our dataset – *Clustering* set.

Table 2. Set of sequence-based features used by prediction model

id	Feature	Source	ReliefF**	Rank
1	Solvent Accessibility	PROF (*)	0.18	3
2	Secondary Structure	PROF (*)	0.12	6
3	Residue Flexibility	PROFbval (*)	0.15	4
4	Protein Disorder	MD (*)	0.22	2
5	Amino Acid	-	5e-5	8
6	Residue Size	-	0	10
7	Residue Charge	-	1e-7	9
8	SNP possible	-	7e-4	7
9	Conservation	ConSurf (*)	0.34	1
10	MSA Ratio	-	0.14	5

(*) tools in the PredictProtein pipeline (Yachdav, et al., 2014). (**) Features ranked by importance to funtrp position typing using ReliefF (Kononenko, et al., 1996); weights were rounded. Secondary structure weights were summarized across helix, sheet, and loop motifs (pH, pE, and pL). Feature descriptions and default parameters in Supplementary Table S2.

2.1.2 Toggle and Neutral cluster labeling

We further subdivided the sequence positions in the *Clustering* set into *Neutral* and *Toggle* classes. Note that we previously defined *Toggles* (Miller, et al., 2017) as positions intolerant of any change, while *Neutrals* were new to this work, indicating positions that can tolerate almost all substitutions with no-effect on function. Each of the proteins in our set was evaluated separately and only the *Clustering* set variants and positions were considered. To each protein's set of experimental variant scores, the protein specific *wt* and *ko* scores were added. K-means (Lloyd, 1982) clustering (with k=3) was used to partition each protein position set into three clusters. Variants assigned to the same cluster as the *ko* score were labeled *severe*. Those assigned to the cluster containing the *wt* score were labeled *no-effect*. All variants in the remaining cluster were labeled *intermediate*.

Each sequence position *x* was classified (Eqn. 2) into one of two distinct position types (*Toggle* or *Neutral*) on the basis of the distribution of its variant scores among the three clusters. If the most variants at *x* were assigned to the *no-effect* cluster and no more than one to any other cluster, we labeled this *x Neutral* (N; 153 positions). If most were assigned to the severe cluster and no more than two to any other cluster, we labeled *x* a *Toggle* (T; 66 positions). If none of these two conditions held true, *x* was deemed unknown (372 positions; *Unknown* set).

$$\mathrm{type}(\mathrm{pos_x}) = \begin{cases} \boldsymbol{N}, if\ (\ |variants_x| - |variants_{x\ in\ wt\ cluster}|\ ) \leq 1 \\ \boldsymbol{T}, if\ (\ |variants_x| - |variants_{x\ in\ wt\ cluster}|\ ) \geq 2 \\ unknown, otherwise \end{cases} \quad (2)$$

We excluded all unknown positions from *Clustering* and manually refined the remaining positions on the basis of distributions of experimental scores (Supplementary Fig. S3). Overall, we removed six *Toggle* and six *Neutral* positions with noticeably higher variance and/or different medians of scores as compared to other instances within the same class. We, thus, retained a conservative training set of labeled *Toggle* and *Neutral*

positions with comparable variance and medians of experimental scores (*ntTraining* set; 207 instances: 60 *Toggles*, 147 *Neutrals*).

2.1.3 ntModel and Neutral/Toggle scoring

Using the labeled *ntTraining* set we trained a Random Forest (RF) (Breiman, 2001) classifier (*ntModel*) to predict *Toggle* vs. *Neutral* position types on the basis of the ten features extracted as described above (Table 2). To account for the bias towards the *Neutral* class in the training set, we used over-sampling and trained our model on a balanced input set comprising 414 instances (200% of the unbalanced input). We evaluated the model performance using Leave-One-Out-Cross-Validation (LOO-CV). The model prediction scores were in the [0, 1] range, such that the sum of all type scores was =1. The LOO-CV predictions were used to determine prediction score type thresholds, limiting the number of false positive *Toggle* or *Neutral* predictions to ≤3% (Fig. 2). Based on this required error rate limitation, thresholds were consecutively set at score ≤0.1 for *Neutral* and score ≥0.8 for *Toggle* predictions.

2.1.4 Defining Rheostats

We assessed the predictions close to the middle (0.5) of our RF classifier prediction range. Here the model exhibited the highest uncertainty in deciding whether the position is a *Neutral* or a *Toggle*. We concluded that positions with prediction scores in that range were *Rheostats* – positions in which mutations can result in a whole range of functionality changes. The *Rheostat* score range was set at [0.35, 0.7] – a range containing 50% of all incorrect predictions of our *ntModel*.

2.1.5 *funtrp* and residue labeling

The *FewVariants* and the *Unknown* sets comprised 544 (66% of 820) yet-unlabeled positions. We ran the *ntModel* and used score thresholds, as defined above, to assign final *N, R, T* predictions per position.

New *Toggle* and *Neutral* position variant score distributions were compared to those of the cluster-based (Step 2, above) positions. We retained only those *ntModel-Neutral* positions from this set whose experimental score medians were less than or equal to the highest median score of the clustering-*Neutral* positions from the *ntTraining* set. Similarly, *ntModel-Toggle*s were retained only if their experimental score medians were more than or equal to the lowest median score of the clustering-*Toggle*s. We retained only those *Rheostat*s whose medians were in-between highest clustering-*Neutral* and lowest clustering-*Toggle* median scores were retained. Thus-labeled positions (72 *Neutrals,* 20 *Toggles*, 104 *Rheostats)* were added to the *ntTraining* set to form the *funtrpTraining* set (403 positions: 219 *Neutrals,* 80 *Toggles,* 104 *Rheostats*).

The *funtrpTraining* set was used to train a second RF model, *i.e.* the final *funtrp* model, using the same ten features, over-sampling -based class balancing (806 instances; 200% of the unbalanced input set), and LOO-CV evaluation as in the *ntModel*.

For each position, the *funtrpModel* prediction score for each type (N, R, T) was in the [0,1] range, such that the sum total of all type scores was =1. By default, the position was assigned the highest scoring type. Performance for both models in this work was reported as accuracy, precision, and recall (Eqn. 3; for each position type, *Y*, at every score cutoff, true positives, TP, are positions correctly predicted as *Y*; false positives, FP, are non-*Y* positions predicted as *Y*; false negatives, FN, are *Y* positions predicted as non-*Y*).

$$\text{precision} = \frac{TP}{(TP + FP)}$$
$$\text{recall} = \frac{TP}{(TP + FN)}$$
$$\text{accuracy} = \frac{TP + TN}{(TP + FP + TN + FN)} \quad (3)$$

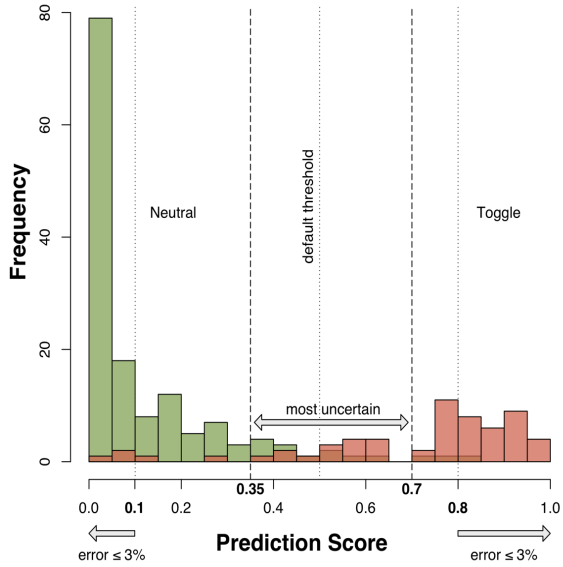

Fig. 2. Determination of *ntModel* thresholds. LOO-CV predictions of the *ntModel* were used to determine prediction score type thresholds. Thresholds were set at score ≤0.1 = *Neutral* and score ≥0.8 = *Toggle*, limiting the number of false positive *Toggle* or *Neutral* predictions to ≤3%. The model exhibits the highest uncertainty in deciding whether a position is a *Neutral* or a *Toggle* at predictions close to the middle (0.5). Positions with prediction scores in the range [0.35, 0.7] (containing 50% of all incorrect predictions of the *ntModel*) were defined as *Rheostats*.

2.2 Predicting position types in protein sets

Neutral, *Rheostat,* and *Toggle* position types were predicted for various curated sets of protein sequences (Supplementary Table S4). Human proteins were extracted from the UniProt Knowledgebase (UniProtKB release 2018_09) (The UniProt, 2017). We predicted position types for all 20,410 manually curated (Swiss-Prot) sequences; for 5% of these (909; 32 enzymes and 877 non-enzymes), no predictions could be made due to errors in extracting the required set of input features. In total 19,501 sequences were processed using *clubber* (Miller, et al., 2017) to distribute computation among multiple High-Performance Cluster (HPC) environments. The subsets of the data were as follows:

(1) The *EXPV* set included 1,250 Swiss-Prot enzymes with experimentally validated, unique, unambiguous E.C. (Enzyme Commission) numbers, compiled as in (Mahlich, et al., 2018).

(2) We extracted all human enzymes with catalytic site annotations from the M-CSA database (Ribeiro, et al., 2018) and retained those which also contained binding site annotations in UniProt (94 proteins; 419 catalytic und 214 binding sites).

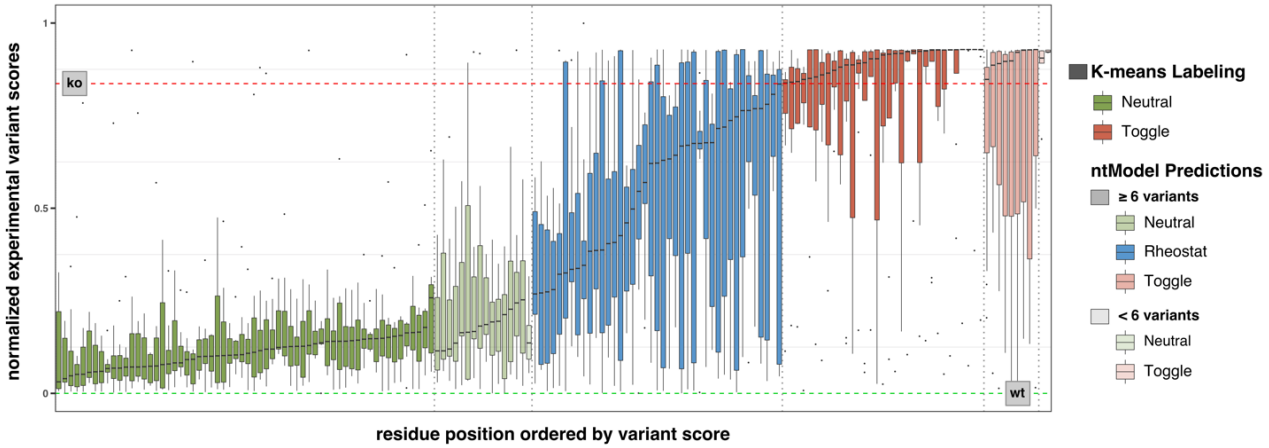

Fig. 3. Distributions of experimental effect scores for TEM-1 (E. coli) positions colored by position type. Positions are colored by assigned/predicted position type (green =*Neutral*, *red* =*Rheostat*, blue =*Toggle*) and ordered by the median of the associated variant score distribution. Positions classified by K-means cluster labeling are shown in bold colors. Those predicted of the *ntModel* are shown in more opaque coloring based on the number of experimental variants at the respective position. The dashed horizontal lines represent data set specific *ko* (red) and *wt* (green) scores. Positions removed during the manual and automatic refinement steps are not shown. Details for the remaining four proteins are available in Supplementary Fig. S3.

(3) We extracted a set of transition metal binding proteins from the PDB as described in (Senn, et al., 2014), (Bromberg Y., 2019) resulting in a set of structural *sahle* spheres. A *sahle* sphere is defined as all residues within a 15Å radius sphere centered on the geometric center of the metal ligand. 231 PDB structures of human proteins containing *sahle* spheres were mapped to UniProt. *fuNTRp* predictions were available for 230 of these.

(4) Swiss-Prot proteins were labeled as disordered (6,309) or ordered (13,192) if at least 50% of their residues were predicted disordered by the MetaDisorder predictor (MD score threshold of ≥ 0.5) (Schlessinger, et al., 2009).

(5) We extracted the Protein Mutant Database (PMD) experimental annotations and SNAP (Bromberg and Rost, 2007), SIFT (Ng and Henikoff, 2003) and PolyPhen-2 (Adzhubei, et al., 2010) predictions of effects of 10,559 variants in 733 proteins from SNPdbe (Schaefer, et al., 2012). For this set we labeled variants as either experimentally benign (SNPdbe score=10), effect (SNPdbe score =3,6,9 or =11,13,16) or knockout (SNPdbe score=0). 728 of these proteins could be mapped unambiguously to UniProtKB. For the remaining five we used the SNPdbe sequences.

To compare position type predictions between different subsets, we calculated the standard error individually for all three position types as follows: for each subset, we randomly resampled 50% of the included residues (without replacement) for 100 times and computed standard error of the mean.

2.3 *funtrp* pipeline implementation

We used a Java based implementation of Random Forest Classification (Breiman, 2001; Smith and Frank, 2016). We used R (R Core Team, 2015) for K-Means Clustering, performance evaluations, and visualizations. Protein features were computed using the Dockerized version of the PredictProtein (Yachdav, et al., 2014) pipeline; available at https://bitbucket.org/bromberglab/predictprotein (manuscript in preparation).

The *funtrp* prediction pipeline was implemented in Python (Version 3.6 or later) and is publicly available via Git repository (https://bitbucket.org/bromberglab/funtrp). The *funtrp* predictor is available as standalone Docker container (bromberglab/funtrp) and as webservice (https://services.bromberglab.org/funtrp).

3 Results

3.1 *funtrp* accurately recognizes position classes

Both RF classifier models were evaluated using LOO-CV (Supplementary Table S5 A,B). *ntModel* achieved an overall accuracy of 92.3% (*Neutrals* = 0.94/0.95 and *Toggles* = 0.88/0.85 precision/recall, respectively, at default cutoff; Eqn. 3). *funtrp* overall accuracy was 85.1% (*Neutrals* = 0.90/0.91, *Toggles* = 0.88/0.80, *Rheostats* = 0.73/0.77 precision, respectively, at default cutoff; Fig. 4; Eqn. 3). Note that that the higher prediction

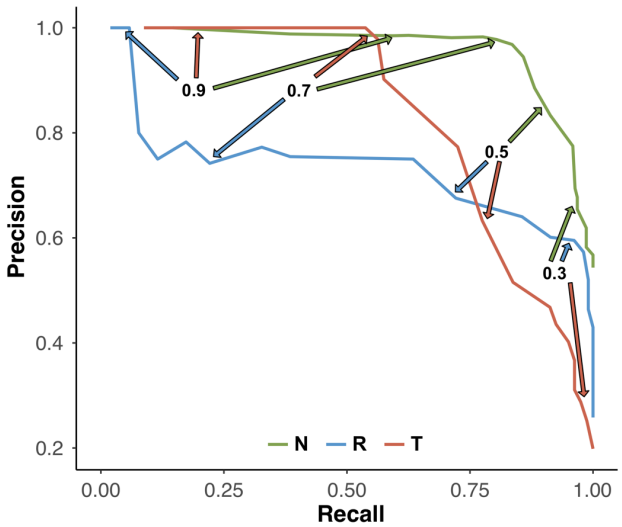

Fig. 4. *funtrp* type classification performance. Precision-Recall curves for LOO-CV predictions of *Neutral*, *Toggle* and *Rheostat* positions for the *funtrpModel*. The performance for all three position types is indicated for different cutoffs. Performance measures were calculated for the single *True* class vs. both remaining classes combined as *Other*, for *Neutral*, *Toggle* and *Rheostat* respectively.

scores of the *funtrp* model correlated with higher precision, albeit lower recall of the predictions.

The slightly lower performance of the *funtrpModel* (vs. the *ntModel*) in differentiating *Toggles* and *Neutrals* is easily attributable to the increase set size and less obvious labels of the added positions. The *funtrp* performance discrepancy among classes was expected. After all, while *Toggles* and *Neutral* are explicitly defined types, *Rheostats* are a collection of different position types. As such, they encompass a much larger range/variability in residue properties. For example, in our training set, a position containing three *intermediate* variants would be as much a *Rheostat* if it additionally contained three *no-effect* variants or three *severe* ones.

Additionally, note that truly benign, *no-effect*, mutations are often subjective and always less obvious and more difficult to identify, experimentally or computationally, then *severe* ones. Thus, the differentiation between *Rheostat* and *Neutral* positions is arguably more complex even with experimental data available. For *funtrp*, the majority (80%) of the incorrectly predicted *Rheostats* were labeled *Neutral*; more than half of these predictions were also unreliable (scores in the [0.4, 0.49] range). Coincidentally, of the incorrectly predicted *Neutral* positions 80% were also labeled as *Rheostats*.

3.2 Individual sequence-based features are not sufficient to describe position types

Using the ReliefF (Kononenko, et al., 1996) feature selection algorithm we ranked the importance of *funtrp* features for labeling sequence positions in Swiss-Prot (Table 2). As expected, evolutionary conservation was ranked most important. However, the assigned weight was only slightly higher than other important features: protein disorder, solvent accessibility, and residue flexibility. These results suggest that none of those features alone could explain the predicted position types.

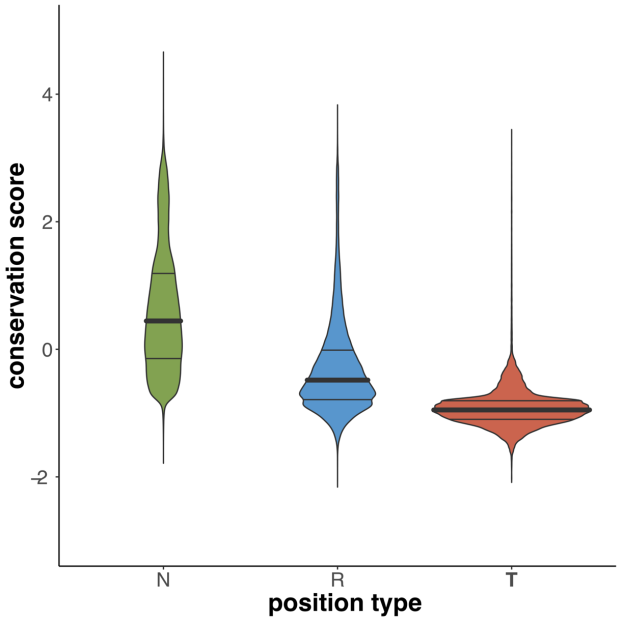

Fig. 6. Conservation of position types. Density distributions of evolutionary conservation (ConSurf) compared between position types. ConSurf predictions scores are by default normalized such as 0 depicts the average score over the entire protein and standard deviation is |1|). Distribution medians are highlighted in bold.

Conservation is widely used as an approximation for residue importance (Capra and Singh, 2007; Shakhnovich, et al., 1996); *i.e.* the more conserved a residue is, the higher the likelihood that its substitution by another amino acid will result in a function disruption. We compared conservation scores (defined by ConSurf (Ashkenazy, et al., 2016) for all positions of experimentally verified enzymes (EXPV). As expected, these were significantly different between the three position types (Fig. 5; medians in bold). ConSurf scores are normalized by default, so that the average score over all residues of one protein is zero, and the standard deviation is one; here, lower scores indicate more conserved residues. *Toggle* positions were predominantly conserved while *Neutral* positions were for mostly non-conserved. *Rheostats*, however, were in-between the other position types and often showed similarly high conservation as the *Toggles*.

To further establish how well a predictor for position types could perform using conservation alone, we computed the number of positions in Swiss-Prot proteins that could be correctly identified as a *funtrp Rheostat, Toggle,* or *Neutral* at a fixed cutoff. The lowest cutoff for *Neutrals* was selected by taking the mean of the distribution medians of *Neutral* and *Rheostat* conservation scores. Similarly, the highest cutoff for *Toggles* was at the mean of *Rheostat* and *Toggle* conservation score medians. *Rheostats* were assigned all other conservation scores. The overall accuracy for this thresholding was 61% (*Neutrals* = 0.80/0.70, *Toggles* = 0.45/0.80, *Rheostats* = 0.44/0.39 precision/recall, respectively; Supplementary Table S6);

Thus, evolutionary conservation - despite being the highest-ranking feature - was not representative of position types. Further, none of the remaining features was likely to perform better then conservation indicated by their consistently lower ReliefF rankings (Table 2). Moreover, arguably, for a given position in a given protein establishing the conservation thresholds for each of the three classes would be infeasible. Note, that we observed the same trends for the training dataset (*funtrpTraining*) for *funtrp* (Supplementary Fig. S7).

3.3 Position type profiles differ across protein classes

Swiss-Prot (Fig. 6A) enzymes had proportionately more *Toggle* and fewer *Neutral* positions than non-enzymes. However, the difference in the number of *Rheostats* between enzymes and non-enzymes was minimal. As *Rheostats* allow for functional flexibility while adapting to different environments, the latter result is expected. On the other hand, we did not expect *Toggle* positions in enzymes, *i.e.* those critical for defining protein activities: active sites, ligand specificity, *etc.,* to represent a larger share of all residues than in non-enzymes. Our results, however, suggest that functionally critical sites are more common in enzymes than expected.

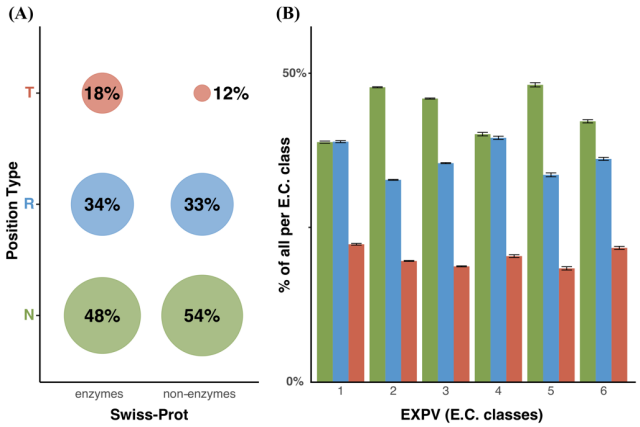

Fig. 5. Distribution of position types per protein class. Distributions are based on entire Swiss-Prot (A) and *EXPV* set (B). Colors are according to position type (green =*Neutral, red =Rheostat*, blue =*Toggle*). Percentages in (A) are rounded and thus do not add up to 100%. Error bars in (B) are computed based on 100 iterations of random subsampling (Methods),

We further compared distributions of position types between the six main enzyme classes (with corresponding E.C.s): Oxidoreductases (1), Transferases (2), Hydrolases (3), Lyases (4), Isomerases (5) and Ligases (6) (Fig. 6B). For proteins with experimental annotations of enzymatic functionality (EXPV set), *Neutral* positions were significantly more frequent for four of the six enzyme classes. For the other two (Oxidoreductases and Lyases) fractions of *Neutrals* and *Rheostats* were similar. Fewer than 25% of all enzymes class positions were *Toggles*. We observed similar trends for all of Swiss-Prot enzyme classes except Oxidoreductases (Supplementary Fig. S8). There were slightly more *Neutrals* in the Swiss-Prot set of EC1 proteins, an observation that can be explained by sequence redundancy of multiple proteins from the same family. The EXPV protein functions are experimentally derived and, thus, the data set tends to be less redundant (98% of the sequences <90% sequence similar). On the other hand, most Swiss-Prot EC annotations are annotated via function transfer by homology – a process (and some error in it (Mahlich, et al., 2018; Schnoes, et al., 2009) that ensure overrepresentation of position types of large families.

3.4 Distribution of position types varies by residue function

We compared the distribution of position types for catalytic sites, binding sites, and *other residues* in Swiss-Prot enzymes (Fig. 7A). Note, that here we included only the 47 proteins containing both binding and catalytic sites, which were non-overlapping, *i.e.* annotated in different positions of the protein.

As expected, the majority of catalytic sites were *Toggles* and only 1% were *Neutral.* Binding sites were less frequently *Toggles* than catalytic sites, but much more frequently so than the *other residues* in the respective proteins, which were predominantly *Neutral*. Curiously, the fraction of *Rheostat* positions did not vary as drastically across the residues sets.

Notably the catalytic site primary actors – the charged amino acids (D, E, R, K, H; Supplementary Fig, S9) (Bartlett, et al., 2002) were unexpectedly low in *Toggles* and *Rheostats* in *other residues*. This finding is particularly interesting in the light of the generic assumptions made about irreplaceability of charged residues. Outside the enzymatic functional sites, the more commonly structure-relevant large hydrophobic amino acids (C, W, Y, M, F) were most often *Toggles*, while the smaller (A, I, L, V) were drastically enriched in *Rheostats* (Supplementary Fig. S9).

3.5 Distribution of position types varies by metal-ligand binding proximity

We evaluated the composition of position types of residues located in the proximity of metal-containing ligands (*sahle* 3D-structure spheres, Methods) for Swiss-Prot proteins. As for functional sites above, we defined three sets of residues: those annotated in Swiss-Prot as metal binding, *sahle* sphere residues within 15A of the ligand center, and *other residues* (Fig. 7B). Note that we excluded from consideration any residues annotated as metal binding and not located within a *sahle* sphere.

Metal binding residues showed a similar distribution of position types as catalytic sites (80% *Toggle,* 5% *Neutral*). Notably, *sahle* spheres were more enriched in *Rheostats* (38%) than were the binding sites described above (26%). However, the latter were more frequently Toggles (59%) than the former (44%). This result suggests that binding sites are critical features of function, while *sahle* spheres encompass residues relevant to functional flexibility. Moreover, outside of *sahle* spheres *Toggles* were the least abundant and more than half of the residues were *Neutral*, suggesting

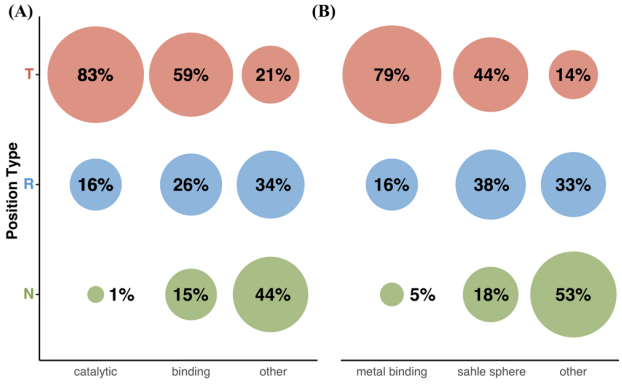

Fig. 7. Distribution of position types across various protein sites. Colors are according to position type (green =*Neutral, red =Rheostat*, blue =*Toggle*). Percentages are rounded and thus do not add up to 100%.

that most of the other residues are significantly less involved in protein function (including stability effects).

Preferred residues for metal binding are C, H, D, and E (Cao, et al., 2017), which is also confirmed by our data (Supplementary Fig. S10). Interestingly, for all of these except glutamate (E) *Toggles* were the dominant position type; for glutamate *Neutrals* and *Rheostats* were strongly enriched.

3.6 Position type profiles enable identification of disordered proteins

Based on MetaDisorder predictions (Methods) we labeled 6,309 Swiss-Prot proteins as disordered and 13,192 as ordered and compared the ratios of position types between these sets. The two classes of proteins were clearly separable by distribution of position types (Supplementary Fig. S11).

Ordered proteins contained more than twice as many *Toggle*s as disordered proteins (19% vs. 8%), while disordered proteins were preferentially *Neutral* (68% vs. 46%). Of the 668 proteins, where *Neutrals* made up over 80% of all residues, 94% (650) were disordered. This result is, to a certain extent, expected due to frequent modulation of function, *i.e. Rheostatic* activity, achieved via structural changes; *e.g.* changes in residue solvent accessibility or secondary structure may, and often do, modulate functionality (Studer, et al., 2013). However, this finding may also indicate that disordered proteins are poorly predicted by *funtrp*, as our method relies on structural features. Another hypothesis based on this observation may be that our definition of position types is not directly applicable to disordered proteins, where changes in functionality may be harder to objectively measure and evaluate.

3.7 Position types can improve variant effect prediction

We evaluated the relationship of position types with experimental annotations of variant effects extracted from the literature (as reported in PMD) and with the predicted variant effect scores (from SNAP, SIFT and PolyPhen-2). Based on PMD effect annotations, sequence positions could be categorized into three main variant impact groups: *no-effect*, *ranged effect* and *knockout* (Supplementary Fig. S12). We compared the composition of predicted position types for each of the effect groups. As expected, the majority (52%) of all 3,223 variants in the *no-effect* group were in *Neutral* positions. However, 20% of *no-effect* variant positions were *Toggles*. On the other hand, the most extreme impact group of *knockout* (2,271)

variants, was comprised of 53% *Toggle* positions 18% *Neutrals*. Note that in this set every sequence position had only one annotated variant. Thus, finding some *no-effect* variants in *Toggle* positions and some *knockout* variants in *Neutral* positions is not unexpected as per our position type definitions. However, as *funtrp* has never been trained to recognize variant effects, the dominant trend of finding variants of expected impact in the right places highlights our method's ability to recognize functionally relevant protein positions.

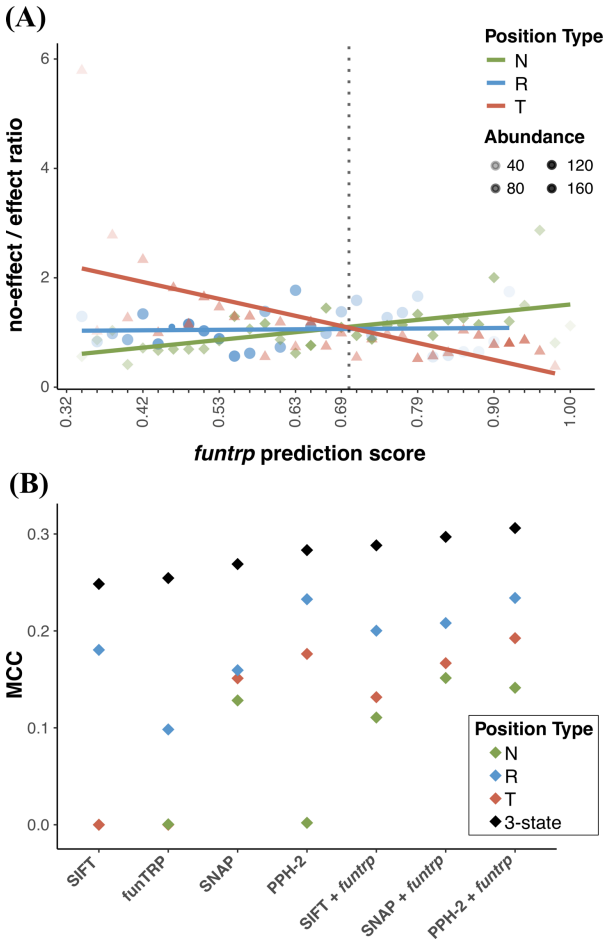

Fig. 8. improving variant effect prediction with position types. (A) Ratio of *no-effect vs. effect* of individual position typea at respective *funtrp* prediction scores for the PMD dataset. Effect annotations are based on PMD annotations. The abundance of position types at a certain score is represented by opacity. (B) Performance comparison of logistic regression models based on predictions of traditional variant effect prediction methods alone and in combination with *funtrp* predictions.

Finally, the variants in the *ranged effect* group were nearly evenly distributed (33%/32%/35% *Neutrals*/*Rheostats*/*Toggles*) across all position types. This is not unexpected, as the *ranged effect* group contains variants with PMD annotations ranging from mild to severe. Interestingly, the fraction of *Rheostat* positions was consistent across all three impact groups, although slightly less for *knockout* and *no-effect* groups (29%). This finding is consistent with our definition of *Rheostats*, which may contain both severe/knockout effect and no-effect variants in addition to everything in-between.

To further highlight the relationship between predicted position types and annotated variant effects in PMD, we calculated the *no-effect vs. effect* (including *ranged effect* and *knockout* variants) ratios individually for every type (Fig. 8A) based on the extracted PMD dataset (Methods). In line with the above results, we found that reliably predicted *Toggle* positions were more likely to have a lower ratio (more *effect* variants), while reliably predicted *Neutrals* had a higher ratio (more *no-effect* variants). Thus, we suggest that variant effect predictors could improve significantly if trained/developed separately with sample data specific to different position types. Specifically, we expect most improvement for *Rheostats*, where increased resolution can be expected once the, arguably, easier *Toggle* and *Neutral* -specific variants are no longer considered.

To compare *funtrp* with common variant effect prediction tools (SNAP, SIFT and PolyPhen-2) we converted predicted position types into approximated variant effect predictions (*Toggle* or *Rheostat* position = *effect* and *Neutral* = *no-effect*). We computed the performance for all four methods on the *no-effect vs. effect* groups extracted from PMD (described above). Note, that performance reported here (Supplementary Table S13) was averaged over 100 iterations, each based on a subsampled dataset (without replacement and balanced regarding the class with fewer instances) from PMD . Note that all methods are expected to perform better on the original unbalanced set of variants, which include significantly more non-neutral effects. This is due to the earlier mentioned difficulty (computational and experimental) of correctly recognizing neutral effects (Bromberg, et al., 2013). All four predictors attained nearly the same accuracy of 62% (+/- 1%), though they did not perform similar within classes (*e.g.* SNAP = 0.59/0.78 and SIFT = 0.65/0.57 precision/recall for *effect* variants, respectively). On the other hand, as mentioned previously *funtrp* is NOT a variant effect prediction method but it reached a performance similar to those of specialized methods.

To quantify the contribution that knowledge of position types can make to prediction method performance, we trained logistic regression models based on prediction scores of traditional variant effect predictors as well as in combination with the information gained by predicted *funtrp* position types (Fig. 8B). This approach consistently improved variant effect predictions. These findings strongly suggest that incorporating position type predictions as features into the more sophisticated variant effect evaluation approaches will improve prediction performance.

Additionally, our new definition of position types will likely contribute to the understanding of biophysics of protein folding and related epistatic mutation effects, as well as highlight prime candidates for directed evolutionary pathways.

Acknowledgements

We would like to thank Dr. Liskin Swint-Kruse (University of Kansas) for all help and comments that made this work possible. We are also grateful to Dr. Predrag Radivojac (Northeastern), Dr. Jay A. Tischfield, Dr. Gary Heiman, Dr. Chengsheng Zhu, Yannick Mahlich, Yanran Wang, and Zishuo Zeng (all Rutgers) for all discussions and to Dr. Sonakshi Bhattacharjee (Columbia) for help with the manuscript. We would also like to express gratitude to all people who deposit their data into publicly available databases and to those who maintain them.

Funding

Y.B. and M.M. were supported by the NIH U01 GM115486 grant. M.M. was also supported by the NIH R01 MH115958 01 grant.

Conflict of Interest: none declared.

References

Adzhubei, I.A., *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* 2010;7(4):248-249.

Araya, C.L. and Fowler, D.M. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol* 2011;29(9):435-442.

Ashkenazy, H., *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* 2016;44(W1):W344-350.

Bartlett, G.J., *et al.* Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 2002;324(1):105-121.

Box, N.F., *et al.* Characterization of melanocyte stimulating hormone receptor variant alleles in twins with red hair. *Human Molecular Genetics* 1997;6(11):1891-1897.

Breiman, L. Random forests. *Mach Learn* 2001;45(1):5-32.

Bromberg, Y. Building a genome analysis pipeline to predict disease risk and prevent disease. *J Mol Biol* 2013;425(21):3993-4005.

Bromberg, Y., Kahn, P.C. and Rost, B. Neutral and weakly nonneutral sequence variants may define individuality. *Proc Natl Acad Sci U S A* 2013;110(35):14255-14260.

Bromberg, Y. and Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* 2007;35(11):3823-3835.

Bromberg Y., M.K., Senn S., Cook L., Nanda V., Falkowski P. Structural Relationships of Metal Binding Sites Suggest Origins of Biological Electron Transfer In. (in preparation); 2019.

Bruse, S., *et al.* Whole exome sequencing identifies novel candidate genes that modify chronic obstructive pulmonary disease susceptibility. *Hum Genomics* 2016;10:1.

Cao, X., *et al.* Identification of metal ion binding sites based on amino acid sequences. *PLoS One* 2017;12(8):e0183756.

Capra, J.A. and Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* 2007;23(15):1875-1882.

Capriotti, E., Altman, R.B. and Bromberg, Y. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics* 2013;14 Suppl 3:S2.

Capriotti, E., Fariselli, P. and Casadio, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 2005;33(Web Server issue):W306-310.

de Ligt, J., Veltman, J.A. and Vissers, L.E. Point mutations as a source of de novo genetic disease. *Curr Opin Genet Dev* 2013;23(3):257-263.

Docker, I. Docker Open Source. In.; 2018.

Dong, C., *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 2015;24(8):2125-2137.

Douville, C., *et al.* CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* 2013;29(5):647-648.

Duffy, D.L., *et al.* A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *American Journal of Human Genetics* 2007;80(2):241-252.

Ellinghaus, D., *et al.* Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology* 2013;145(2):339-347.

Firnberg, E., *et al.* A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol* 2014;31(6):1581-1592.

Fowler, D.M., *et al.* High-resolution mapping of protein sequence-function relationships. *Nat Methods* 2010;7(9):741-746.

Gray, V.E., *et al.* Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst* 2018;6(1):116-124 e113.

Halushka, M.K., Walker, L.P. and Halushka, P.V. Genetic variation in cyclooxygenase 1: effects on response to aspirin. *Clin Pharmacol Ther* 2003;73(1):122-130.

Hodges, A.M., *et al.* RheoScale: A tool to aggregate and quantify experimentally determined substitution outcomes for multiple variants at individual protein positions. *Hum Mutat* 2018;39(12):1814-1826.

Ioannidis, N.M., *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* 2016;99(4):877-885.

Kononenko, I., RobnikSikonja, M. and Pompe, U. ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. *Fr Art Int* 1996;35:31-40.

Kowarsch, A., *et al.* Correlated mutations: a hallmark of phenotypic amino acid substitutions. *PLoS Comput Biol* 2010;6(9).

Kumar, R., *et al.* Disease-causing point-mutations in metal-binding domains of Wilson disease protein decrease stability and increase structural dynamics. *Biometals* 2017;30(1):27-35.

Lloyd, S.P. Least-Squares Quantization in Pcm. *Ieee T Inform Theory* 1982;28(2):129-137.

Mahlich, Y., *et al.* HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics* 2018;34(13):i304-i312.

Mahmood, K., *et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum Genomics* 2017;11(1):10.

Meinhardt, S., *et al.* Rheostats and toggle switches for modulating protein function. *PLoS One* 2013;8(12):e83502.

Melamed, D., *et al.* Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* 2013;19(11):1537-1551.

Miller, M., Bromberg, Y. and Swint-Kruse, L. Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci Rep* 2017;7:41329.

Miller, M., Zhu, C. and Bromberg, Y. clubber: removing the bioinformatics bottleneck in big data analyses. *J Integr Bioinform* 2017;14(2).

Ng, P.C. and Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* 2003;31(13):3812-3814.

Pitt, J.N. and Ferre-D'Amare, A.R. Rapid construction of empirical RNA fitness landscapes. *Science* 2010;330(6002):376-379.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Ribeiro, A.J.M., *et al.* Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res* 2018;46(D1):D618-D623.

Schaefer, C., *et al.* SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics* 2012;28(4):601-602.

Schlessinger, A., *et al.* Improved disorder prediction by combination of orthogonal approaches. *PLoS One* 2009;4(2):e4433.

Schnoes, A.M., *et al.* Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 2009;5(12):e1000605.

Senn, S., *et al.* Function-based assessment of structural similarity measurements using metal co-factor orientation. *Proteins* 2014;82(4):648-656.

Shakhnovich, E., Abkevich, V. and Ptitsyn, O. Conserved residues and the mechanism of protein folding. *Nature* 1996;379(6560):96-98.

Shastry, B.S. SNPs: impact on gene function and phenotype. *Methods Mol Biol* 2009;578:3-22.

Smith, T.C. and Frank, E. Introducing Machine Learning Concepts with WEKA. *Methods Mol Biol* 2016;1418:353-378.

Starita, L.M., *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc Natl Acad Sci U S A* 2013;110(14):E1263-1272.

Starita, L.M., *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* 2015;200(2):413-422.

Studer, R.A., Dessailly, B.H. and Orengo, C.A. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem J* 2013;449:581-594.

Sun, X.J., *et al.* Targeted mutagenesis in soybean using the CRISPR-Cas9 system. *Sci Rep-Uk* 2015;5.

Swint-Kruse, L. Using Evolution to Guide Protein Engineering: The Devil IS in the Details. *Biophys J* 2016;111(1):10-18.

The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;45(D1):D158-D169.

Turner, T.N., *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* 2016;98(1):58-74.

Walker, I.H., Hsieh, P.C. and Riggs, P.D. Mutations in maltose-binding protein that alter affinity and solubility properties. *Appl Microbiol Biotechnol* 2010;88(1):187-197.

Wu, N.C., Olson, C.A. and Sun, R. High-throughput identification of protein mutant stability computed from a double mutant fitness landscape. *Protein Sci* 2016;25(2):530-539.

Yachdav, G., *et al.* PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res* 2014;42(Web Server issue):W337-343.

Zabalza, R., *et al.* Co-occurrence of four nucleotide changes associated with an adult mitochondrial ataxia phenotype. *BMC Res Notes* 2014;7:883.