Variant effect predictions capture some aspects of deep mutational scanning experiments

Jonas Reeb^{1,*}, Theresa Wirth¹, & Burkhard Rost^{1,2}

¹ TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany

² Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany & Columbia University, Department of Biochemistry and Molecular Biophysics, 701 West, 168th Street, New York, NY 10032, USA

* Corresponding author: assistant@rostlab.org, <http://www.rostlab.org/>
Tel: +49-289-17-811 (email rost:assistant@rostlab.org)

Running Title: Variant effect prediction on deep mutational scanning

Document statistics: Abstr:173 words, Text: 4200 words, 86 references, 1 figure; 2 tables; 25 pages

Journal:

Submitted: 28th November 2019

Abstract

Deep mutational scanning (DMS) studies exploit the mutational landscape of sequence variation by systematically and comprehensively assaying the effect of single amino acid variants (SAVs) for particular proteins. Different experimental protocols proxy effect through a diversity of measures. We evaluated three early prediction methods trained on traditional variant effect data (PolyPhen-2, SIFT, SNAP2) along with a regression method optimized on DMS data (Envision). On a common subset of 32,981 SAVs, all methods capture some aspects of variant effects, albeit not the same. Early effect prediction methods correlated slightly more with measurements and better classified binary states (effect or neutral), while Envision predicted better the precise degree of effect. Most surprising was that a simple approach predicting residues conserved in families (found and aligned by PSI-BLAST) in many cases outperformed other methods. All methods predicted beneficial effects (gain-of-function) significantly worse than deleterious (loss-of-function). For the few proteins with several DMS measurements, experiments agreed more with each other than predictions with experiments. Our findings highlight challenges and opportunities of DMS for improving variant effect predictions.

1

2 **Key words:** sequence variation, variant effect prediction, deep mutational scanning, non-
3 synonymous sequence variant, deep mutational scanning

4

5 **Abbreviations used:** **DMS**, deep mutational scanning; **SAV**, single amino acid
6 variant; **ROC**, receiver operating characteristic; **AUC**, area under the ROC curve; **CI**,
7 confidence interval; Mean squared error; **MSE**

Introduction

Recent human sequencing projects conclude that we all carry about 10,000 single amino acid variants (**SAVs**) with respect to the “reference genome” [1, 2]. Many of these SAVs are assumed to be neutral, while others might change protein function, contributing to complex phenotypes and causing diseases through a single change. Unfortunately, the gap between SAVs with and without experimental characterization continues to widen [3]: for only one in 10,000 of the known SAVs some experimental information is available [4, 5]. On top, many of those for which something is known are disease associations that may be wrong [6]. As long as the ability to interpret SAV effect does not improve, both on the level of the organism and the protein, the promise of precision medicine remains unmet [7-10].

Through the increased efficiency of sequencing, a procedure formerly used primarily *in silico* [11, 12] has become feasible for experiments, namely assessing the effect of all possible SAVs in a protein. In such deep mutational scanning (**DMS**) studies, a sequence library with all possible variants is created [13, 14] and exposed to a selection system. In the simplest case, the (logarithmic) difference between sequence frequencies with and without selection pressure yield an effect score for individual or combinations of variants [8, 15-17]. Variants with beneficial effect on protein function are discovered alongside the more commonly reported deleterious effect variants. DMS therefore aims at measuring the landscape of functional fitness for select proteins [18].

DMS has also been applied to screen proteins for improved drug binding, antibody affinity, using non-native chemical stresses, or non-proteinogenic amino acids, and on synthetic proteins [19-26]. DMS share objectives with directed evolution experiments which assay a wide range of mutations to engineer proteins with specific purposes. In fact, protein engineering has benefited from DMS studies [14].

One major challenge for DMS is the development of an assay to measure effect. Evaluating proteins with multiple functions requires multiple assays [8]. For instance, the effect of variants on Ubiquitin-60S ribosomal protein L40 has been assessed through their direct impact on yeast growth, as well as, through the impaired activation by the E1 enzyme [27, 28]. BRCA1 has been assayed through E3 ubiquitin ligase activity, through BARD1 binding and transcript abundance [29, 30]. Even for the same assay, specific experimental conditions might influence measurements [31]. Recently, a protocol for measuring protein abundance has been suggested as a proxy for function and applicable to many proteins [32]. The conclusions from DMS studies are limited by the validity of their functional assays; inferences of more complex effect relationships such as disease risk or clinically actionable pathogenicity often remain too speculative [8, 17].

Long before experimental DMS, prediction methods had addressed the same task *in silico* [33, 34] [35] [36-41]. These methods were developed on very limited datasets of effect and neutral variants. Many methods focused on disease-causing SAVs, e.g. from OMIM [42]. Others used databases such as the heterogeneous dbSNP [43] or PMD [44] which catalogue variants according to their direct effect upon protein function or structure. CADD solved the problem of limited data sets and some bias in databases by considering all mutations that have become fixed in the human population as neutral and a simulated set of all other variants as having an effect [35]. The training data set determines the type of effect methods can learn. Consequently, methods differ and work only on the type of SAV used for development. Given the limitations in today's data, all methods are optimized on relatively small, unrepresentative subsets: fewer than 85,000 of all possible 217 million human SAVs have some experimental annotations [45, 46]. Methods agree much more with one another for the SAVs with known annotation than for SAVs without annotations [47].

DMS data sets constitute a uniquely valuable resource for the evaluation of current SAV effect prediction methods [17, 48], not the least, because most have not used those

data. The Fowler lab has, recently, published an excellent analysis of prediction methods on DMS datasets and developed a new regression-based prediction method, Envision, trained only on DMS data [49]. Here, we focus on the analysis of a larger set of DMS studies and present trends in their correlation with SAV effects predicted by four variant effect prediction methods.

Results and Discussion

DMS experiments focus on subsets of all possible variants. The base dataset for our analyses of DMS studies consisted of 22 functional effect measures from 18 proteins, nine of those from human (Supplementary Online Material (SOM), Fig. S1a, Table S1) [29, 30, 32, 50-65]. The number of variants per study differed from hundreds to 10,000. In total the set contained 68,447 variants; 2,358 (3%) of these were synonymous, the other 97% constituted SAVs.

DMS still does not imply that all possible variants are tested [66]. On the contrary, only ten of the 22 sets (45%) scored some variants for at least 98% of the residues (Table S1). Four DMS studies provided functional scores for over 90% of all possible 19 non-native SAVs. On average, 66% of the residues in the 22 DMS sets were found to have SAVs with both deleterious and beneficial effects (Table S2). The majority of SAVs were beneficial for 3 of 22 studies (14%), namely for C-C chemokine receptor type 5 (CCR5), Mitogen-activated protein kinase 1 (MAPK1) and Peroxisome proliferator-activated receptor γ (PPARG). For the other 19 studies deleterious outnumbered beneficial SAVs by factors of 1.5-22.5 (Fig. S1b). Due the asymmetry in numbers and experimental fidelity, deleterious and beneficial effect SAVs were analyzed separately.

Higher correlation for binary classification methods. SetCommon comprised 32,981 effect SAVs (17,781 deleterious) with predictions from each method (Table 1). All prediction methods reached a positive correlation on the set of deleterious SAVs (Spearman $\rho \geq 0.1$, Fig. 1a-d, Table 2). Highest was SNAP2 ($\rho = 0.41$), followed by PolyPhen-2, SIFT and Envision [37-39, 49]. 95% confidence intervals (CIs) never overlapped, and these differences were statistically significant (Table S3).

Both SIFT and PolyPhen-2, by design, have almost binary prediction scores (either neutral or effect), and both score distributions are highly skewed toward effect. Both are optimized for capturing binary effects, not correlations, which was confirmed by two recent studies [48, 49]. The relatively high Spearman ρ values for these were, thus, somehow unexpected. On the other hand, SNAP2 and Envision scores appeared, overall, less binary (Figs. 1c-d, S3). For Envision this was expected, given that it was trained as a regression predictor on non-binary data. In contrast, SNAP2 was trained on binary classification data (effect or neutral). Nevertheless, predictions have been shown to correlate with effect strength [5, 67, 68]. SNAP2 distributions were skewed toward high effect, while Envision also succeeded in detecting SAVs with less pronounced effects (Fig. 1c-d).

Envision approximates experimental values best. When evaluating methods by the numerical difference between experimental and predicted variant effect scores (mean squared error, MSE), Envision was best, followed at considerable distance by SNAP2, PolyPhen-2 and SIFT (Fig. 1a-d, Table 2). Envision was found to best distinguish low from high effect variants [49]. The training as regression method might have given it an advantage over classification methods. Envision's good performance on this metric, partially originated from predicting no SAV with strong effect (the highest Envision score was 0.61, with a possible maximum of 1). Such a distribution resembled the overall experimental distribution that was skewed towards lower effect (Fig. 1d, gray distributions next to x- and y-axes). Predicting most SAVs somewhere in the middle constraints the maximally possible MSE. Indeed, shuffling the prediction scores achieved the same MSE (Fig. S2a). Furthermore, predicting a normal distribution around the mean of the experimental values performed slightly worse but still better than all other prediction methods (Fig. S2b, Table 2). On the other hand, MSE did not increase when limiting the evaluation to just the 25% strongest effect SAVs, although the benefit of predicting no strong effect SAVs should be

reduced here (Table S4). When considering each DMS measurement separately, Envision also appeared to perform best except for transcriptional coactivator YAP1 (YAP1) which indeed had the most uniform distribution of effect scores (similar number of lowest, medium, and strongest effects observed; Fig. S3b, Table S5).

Conservation captured deleterious effects well. Most surprising was the good performance of a naïve method exclusively using conservation (from PSI-BLAST profiles) to predict SAV effects. This naïve prediction appeared to correlate more with the DMS experiments than the specialized methods SIFT and Envision (Fig. 1e). Furthermore, the MSE was lower for Conservation than for all specialists except Envision (Fig. 1, Table 2). Disease causing SAVs from OMIM typically affect the most conserved residues [47], and it has been argued that machine-learning based predictions of variant effects are of limited applicability because they often largely capture conservation [17, 69-72]. Furthermore, non-specialized methods such as analyses of conservation patterns also capture some aspects of variant effect on function [73]. To some extent, our findings confirm this to be valid for DMS experiments, as well. Nonetheless, the distributions of effect predicted by Conservation and observed by DMS differed substantially (Fig. 1e, gray distributions).

High effect SAVs from DMS not generally predicted better. The DMS sets for immunoglobulin gamma-1 heavy chain (IgG1) and modification methylase HaeIII (haeIIIM) were most skewed toward effect (Fig. S3o,v). SNAP2, PolyPhen-2 and SIFT had lower than average MSE on these two sets and also better than average correlation for haeIIIM (Tables S5, S6). Bias in the data used for developing those methods might explain this observation [5, 70, 71, 74]. However, the same trend was observed for the conservation-based prediction suggesting that for those two DMS experiments evolutionary information might capture variant effects best.

To investigate whether methods predict the strongest effect SAVs better, we analyzed subsets with the 75%, 50%, or 25% strongest effect SAVs (strongest deleterious: SetDelStrongest75|50|25, beneficial: SetBenStrongest75|50|25). MSE was higher for stronger deleterious effect SAVs for PolyPhen-2, SNAP2 and Conservation, much less so for SIFT, and not for Envision (Table S4). Therefore, the accuracy in predicting exactly how much effect a SAV has was not increased for strong effect SAVs. Furthermore, the Spearman ρ correlation decreased for all methods for SetDelStrongest75, 50 and 25. This again suggested most methods to succeed only at distinguishing lower from high effect SAVs, on average. These results were surprising, because, for instance, SNAP2 prediction scores correlated with effect strengths, i.e. stronger effect SAVs tend to be predicted better [47, 68, 75]. This correlation might originate from so-called toggle positions, i.e. residues for which SAVs have either a strong effect or none [71, 76].

Beneficial effects difficult to predict. Unlike for deleterious SAVs, no method correlated, on average, with beneficial effect SAVs ($-0.14 \leq \rho \leq 0.02$, Tables 2, S7, Fig. S4) and performance was only slightly increased when focusing on high effect subsets (Table S4). This suggested beneficial effect SAVs to have distinctly different signatures from their deleterious counterparts, too distinct for current methods to capture. The conservation-based prediction also decreased substantially from Spearman ρ of 0.29 for deleterious to -0.08 for beneficial SAVs (Table 2, Fig. S4e). Furthermore, SNAP2 and PolyPhen-2 show a clear shift of prediction scores towards lower effect compared to predictions of deleterious SAVs (cf. Fig 1a,c and Fig. S4a,c, gray distributions). This was particularly pronounced for levoglucosan kinase (LGK), haeIIM, β -glucosidase (bgl3) and β -lactamase TEM (bla) which suggests that beneficial effect SAVs were often mistaken as having no or low effect (Fig. S5j, n, o, t). Possibly, because most SAVs in the development sets of these methods were deleterious.

Unlike Spearman ρ , the MSE for beneficial effect SAVs was similar to that for deleterious variants. Envision again performed best by far (MSE=0.05, Tables 2, S3), also when treating each set separately (Table S8, Fig. S5). Envision was trained with 25% beneficial effect variants (SOM_Note1). However, the decreased correlation performance ($\rho = -0.14$ versus 0.1), highlighted that just including these variants in training did not suffice to learn their signatures.

Experimental agreement sets the benchmark for prediction methods. The above comparisons of experimental and predicted SAV effects raise the question of what agreement can realistically be obtained. One proxy for an answer is the comparison of different DMS studies conducted on the same protein.

For deleterious SAVs, the correlation between experiments was high (Spearman $\rho = 0.65$ vs. 0.41 for the top predictions from SNAP2, Table S9, Fig. S6). The lowest correlation was that between two measurements on breast cancer type 1 susceptibility protein, BRCA1 and BRCA1_2015_E3 ($\rho=0.21$, Fig. S6b). This was the only set for which prediction methods outperformed experimental agreement. Rather than experimental noise, the low correlation might also originate from different experimental setups employed for multi-functional proteins such as BRCA1. The strong correlation ($\rho=0.93$) between two experiments that measured the same condition for bla (beta-lactamase TEM precursor; bla and bla_2014, Fig. S6h) provided a single case in strong support of such an explanation. Possibly, prediction methods only detect one of multiple aspects of effect, or the effects measured by all functional assays. The MSE between experiments was not lower than that between experiments and Envision (mean MSE=0.07, Table S9, Table S5).

In contrast to deleterious SAVs, experiments did not correlate for beneficial effect (mean $\rho=0.03$) although the MSE remained low (mean MSE=0.05, Table S9, Fig. S7). The major issue for this comparison were the small numbers (572 SAVs from five assays).

Maybe experimental assays are biased towards measuring deleterious effects. This also put the poor predictions of beneficial SAVs effects into perspective.

Classification into neutral and effect provides similar picture as regression analyses.

Scores from binary classification methods (neutral or effect) such as SNAP2, PolyPhen-2 and SIFT primarily measure the method's confidence in the predicted class. Classification performance is often assessed through receiver operating characteristic (ROC) curves avoiding the choice of a particular single threshold in making the decision between neutral and effect. Toward this end, we assigned classes to the SAVs in SetCommon through normalization by experimental measurements of synonymous variants [60] (Methods).

On the 3,209 deleterious effect SAVs of SetCommonSyn95 (10,587 neutral, Table 1, Fig. S8), SNAP2 achieved the highest area under the curve (AUC, 0.76, 95% CI [0.75, 0.77]). It was the only method better (with a statistically significant difference) than the naïve conservation-based method (0.73 [0.72, 0.74], Fig. 1f, Table S10). Precision-recall curves also highlight the smooth transition of SNAP2 opposed to the naïve approach of Conservation although both methods share similar peak performance (Fig S9). For this task, for which it had not been developed, Envision performed better than random, but clearly worse than the specialized classification methods (AUC = 0.55 [0.54, 0.56]). On the BRCA1 set, Envision performed on par with the best methods (Fig. S10a). The reason remained unclear, but for this set correlation was also highest (Table S6). Overall, the simple assumption that classification methods perform better on the task they were trained for was not supported by our results: The four proteins considered here (BRCA1, PPARG, PTEN and TPMT), also correlated above average for SNAP2, PolyPhen-2 and SIFT (Table S6). Using less or more stringent definitions for classifying experimental SAVs (SetCommonSyn90 and SetCommonSyn99, see Methods, Fig. S11a-b) did not change these results.

1 Beneficial SAVs were also difficult to classify: PolyPhen-2 and SNAP2 performed
2 best with AUC=0.62, followed by SIFT, while Envision predictions were not better than
3 random (Fig. S4f, Fig. S11c-d, Table S10). Naïve Conservation also performed significantly
4 worse at a level of random predictions.

Conclusions

Deep mutational scanning (DMS) studies set out to explore the relation between protein sequence and function by systematically probing sequence variants. We collected 22 DMS experiments from the literature and focused on single amino acid variants (SAVs). Most studies probe only a small subset of all possible variants (for a protein with N residues, there are $19 \cdot N$ non-native sequence variants). We set out to evaluate whether DMS studies agree with each other and with predictions. One result was that two experiments probing the same protein tend to agree more with each other than either of those agrees with predictions. Nonetheless, different experiments use different proxies for function and disagree substantially, and all experiments differ significantly from all prediction methods. On top of the problem that different experiments probe different proxies for function, we failed to identify a single measure for how similar experiments and predictions are (see SOM_Note2). Lacking a convincing solution, we discussed different aspects and scoring schemes, showing as much as possible the raw data.

We analyzed five variant effect prediction methods: *Envision*, *naïve Conservation* (essentially using PSI-Blast conservation to predict effect/neutral), *PolyPhen-2*, *SIFT* and *SNAP2*. Looking at the distributions for experimental and predicted SAV effects (Fig. 1), we noted that for deleterious SAVs all methods reached positive Spearman ρ correlations with the DMS experiments. Traditional classification methods actually achieved highest correlations with effect strength, although much of their performance can be achieved by evolutionary information, i.e. read off PSI-BLAST profiles. On the other hand, the lowest mean squared error (MSE) was achieved by *Envision*, the only tool currently available which has been trained and optimized for regression on DMS data. In fact, the *Envision* MSE was as low as that between experiments.

Sorting performance by Spearman ρ and by MSE, put different methods on top showing that methods captured diverse aspects measured by the DMS studies. All methods performed better on SAVs with deleterious than with beneficial effect. This highlighted the common negligence and lack of data for beneficial effect SAVs. However, experiments also agreed much more with each other for deleterious than for beneficial effect SAVs which demonstrates the challenge in experimentally describing beneficial SAVs comprehensively and correctly.

Although binary classification methods, surprisingly, captured aspects of non-binary measurements, they performed much better for the binary classification task (projecting DMS results onto neutral vs. effect; Fig. 1f). Notably, the naïve Conservation prediction captured effect better than some more advanced tools.

The challenge for the next generation of prediction methods will be to learn from the diversity of DMS. To give just one example: OMIM, a popular source of training data, contained ~11,000 SAVs referenced in dbSNP (02/2019). This is a magnitude matched by a single large DMS experiment. The generality of a singular SAV in both sets might not be comparable, yet DMS opens up variant effect prediction to new methodologies. For instance, increasingly popular deep learning approaches which typically require training sets magnitudes larger than what has traditionally been used for variant effect [77, 78]. Furthermore, the increase in amount and diversity of training data could help to improve the applicability of methods by being less reliant on conservation. It might also allow methods to account for different SAV types such as toggle and rheostat positions [71]. Finally, DMS studies contain many beneficial effect SAVs that have, so far, been underrepresented in databases. Clearly, DMS should play a major role for the future of variant effect prediction methods and give rise to new methods which are better estimators of SAV effects both on the level of protein function as well as organism-level disease-causing variants. Only then will we come closer to attaining the goals of precision medicine.

Methods

Dataset collection. We retrieved all DMS datasets available by June 2019 that report over 100 SAVs available from the literature. Functional effect scores were taken directly from the supplemental material published or requested from the authors (Table S11). The data were formatted in a variety of formats including Excel, and tab- or comma-separated files. Scores were manually mapped either to the UniProtKB identifier given in the publication or to its closest BLAST match (Table S12) [45, 79]. Six of the 22 experiments contained up to five substitutions (pairwise sequence identity $\geq 98\%$); those were maintained for prediction. We refer to the combined data as *SetAll* (66,089 SAVs) supplemented by *SetCommon* with 32,981 SAVs from ten of the 22 experiments for which we had a prediction from every method tested (Table 1). During completion of this manuscript, MaveDB, a centralized resource of multiplexed assays of variant effect has been published [8, 80]. MaveDB identifiers exist for ten of our 22 datasets (November 2019, Table S11).

SetAll contained several proteins with multiple independent experimental measurements. Inclusion of additional sets analyzed previously [49], yielded a total of three measurements for Hsp82 and BRCA1 and two for both beta-lactamase and ubiquitin (Table S1) [27, 28, 81]. Performance measures were calculated only on SAVs and not between DMS measurements from the same publication. For analysis of beneficial effect SAVs, all studies on Hsp82 had to be excluded since the sets contain only three of those SAVs each.

Three subsets of *SetAll* were created containing the SAVs with the 75%, 50%, or 25% highest deleterious or beneficial effect for each DMS experiment (*SetBenStrongest75/50/25* and *SetDelStrongest75/50/25*); the YAP1 data set was removed for *SetBenStrongest25*, because it contained fewer than 50 SAVs after filtering.

Processing functional effect scores. Several DMS studies provide multiple effect scores for the same protein. We generally chose the scores with the highest evolutionary pressure (Table S13). In the following processing, effect scores were left as provided by the authors as much as possible but adjusted such that the wild-type score for each measurement (Table S14) became 0, and larger values denoted more effect up to the maximum at 1 (SOM_Note3). Beneficial and deleterious effects had to be analyzed separately because experimental assays were not symmetrical and further normalization might over- or underrepresent effects. The resulting score distributions differed significantly between experiments (e.g. in contrast to the more homogeneous subset used previously [49]).

We also created sets with binary classifications (effect vs. neutral) from all DMS studies with synonymous variants: The middle 95% of effect score values from synonymous variants was used to define which SAVs were considered neutral. All SAVs outside this range were considered as effect. We applied the same procedure using 90% or 99% of synonymous variants' values and refer to the thresholding schemes as *syn90*, *syn95*, and *syn99*. Applying these schemes to the four experiments in SetCommon which have synonymous variants (BRCA1, PTEN, TPMT, PPARG) yields *SetCommonSyn90/95/99*. Again, deleterious and beneficial effect SAVs were analyzed separately.

Performance measures. Experiments and predictions were compared through three measures (SOM_Note4, SOM_Note2): (1) mean squared error (**MSE**) calculated with the scikit-learn metrics module [82]; (2) Pearson R (pearsonr) and (3) Spearman ρ (spearmanr) both calculated with the SciPy stats module [83]. For convenience linear least-squares regression lines (linregress) were added to the correlation plots. Pearson R was added for ease of comparison to others but not discussed as it is not robust and most datasets violated both its validity assumptions (normal distribution & absence of significant

outliers [84]). We further found no evidence to supplement MSE by a measure more robust to outliers (SOM_Note2). 95% confidence intervals (**CIs**) for R, ρ and MSE were estimated using a percentile bootstrap with 1000 random samples with replacement.

The performance of binary predictions (effect vs. neutral) was measured through receiver operating characteristic (**ROC**) curves and the area under those curves (**AUC**) calculated through the pROC package in R, which was also used to calculate 95% confidence intervals of ROC (*ci.se*) and AUC (*ci.auc*) [85, 86]. Additionally, precision-recall curves were created using scikit-learn (*precision-recall-curve*).

Prediction methods. The sequences determined during dataset collection were used as input to a set of commonly used variant effect prediction methods. Each method was run to predict the effect of all 19 non-native amino acids at every position in the protein. *SNAP2* [38] was run locally using default parameters on UniProtKB (Release 2018_09). *SIFT* version 6.2.1 [39] was run locally (UniProtKB/TrEMBL Release 2018_10). *PolyPhen-2* [37] predictions were retrieved from the webserver in batch mode (<http://genetics.bwh.harvard.edu/pph2/bgi.shtml>) with classification model humdiv on genome assembly GRCh37/hg19 and default parameters. Predictions failed for all relevant residues of the three DMS studies on Hsp82. *Envision* [49] predictions were retrieved online (https://envision.gs.washington.edu/shiny/envision_new/) which requires UniProtKB identifiers as input. Therefore, Envision predictions could be analyzed only for ten proteins (Table S15). While SNAP2 and SIFT predicted all SAVs, PolyPhen-2 and Envision failed for some residues, shrinking the size of the data sets. We always report performance on the largest common subset of SAVs per dataset.

As a baseline, predictions were also created by running PSI-BLAST with three iterations on UniProtKB (Release 2018_09). Scores from the resulting profile (position-

specific scoring matrix) had their signs flipped and were then directly used as a measure of effect, i.e. less frequent substitutions have a higher effect than conserved ones. We refer to this method as naïve *Conservation*.

For SIFT, scores were reversed such that higher values implied higher effect. The same was done for Envision predictions of deleterious effect. Envision predictions of beneficial effect were treated separately and mapped to the range of [0,0.2]. This yielded the same performance than scaling between [0,1] or no scaling (SOM_Note5). Finally, prediction scores of all methods were adjusted to lie between 0 (no effect) and 1 (highest effect) using the theoretical maximum and minimum prediction value of every method.

Acknowledgements

The authors wish to thank all groups that work on DMS and readily provided their data, either as part of their manuscript or swiftly upon personal contact. Thanks also to Michael Bernhofer and Maria Schelling for helpful discussions, Inga Weise for administrative support and Tim Karl for help with hard- and software. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) – project number 5091000. Further funding was provided by the Bavarian Ministry for Education through funding to the TUM paying for the positions of the authors.

References

- [1] Tennesen JA, Bigham AW, Connor TDO, Fu W, Kenny EE, Gravel S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*. 2012;64-70.
- [2] Consortium TGP. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
- [3] Manolio TA, Fowler DM, Starita LM, Haendel MA, MacArthur DG, Biesecker LG, et al. Bedside Back to Bench: Building Bridges between Basic and Clinical Genomic Research. *Cell*. 2017;169:6-12.
- [4] de Beer TAP, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM. Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset. *PLoS Computational Biology*. 2013;9.
- [5] Mahlich Y, Reeb J, Hecht M, Schelling M, De Beer TAP, Bromberg Y, et al. Common sequence variants affect molecular function more than rare variants? *Scientific Reports*. 2017;7:1608.
- [6] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285-91.
- [7] Rost B, Radivojac P, Bromberg Y. Protein function in precision medicine: deep understanding with machine learning. *FEBS Letters*. 2016;590:2327-41.
- [8] Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, et al. Variant Interpretation: Functional Assays to the Rescue. *American Journal of Human Genetics*. 2017;101:315-25.
- [9] Capriotti E, Ozturk K, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*. 2018:1-20.
- [10] Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*. 2017;38:1182-92.
- [11] Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. *Bioinformatics (Oxford, England)*. 2008;24:2397-8.
- [12] Hecht M, Bromberg Y, Rost B. News from the protein mutability landscape. *Journal of molecular biology*. 2013;425:3937-48.
- [13] Hietpas R, Roscoe B, Jiang L, Bolon DNA. Fitness analyses of all possible point mutations for regions of genes in yeast. *Nature protocols*. 2012;7:1382-96.
- [14] Wrenbeck EE, Faber MS, Whitehead TA. Deep sequencing methods for protein engineering and design. *Current Opinion in Structural Biology*. 2017;45:36-44.
- [15] Araya CL, Fowler DM. Deep mutational scanning: assessing protein function on a massive scale. *Trends in Biotechnology*. 2011;29.
- [16] Fowler DM, Stephany JJ, Fields S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature Protocols*. 2014;9:2267-84.
- [17] Fowler DM, Fields S. Deep mutational scanning: A new style of protein science. *Nature Methods*. 2014;11:801-7.
- [18] Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*. 2007;8:610-8.
- [19] Forsyth CM, Juan V, Akamatsu Y, DuBridge RB, Doan M, Ivanov AV, et al. Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *mAbs*. 2013;5:523-32.
- [20] Mavor D, Barlow K, Thompson S, Barad BA, Bonny AR, Cario CL, et al. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*. 2016;5:1-23.

- [21] Mavor D, Barlow KA, Asarnow D, Birman Y, Britain D, Chen W, et al. Extending chemical perturbations of the ubiquitin fitness landscape in a classroom setting reveals new constraints on sequence tolerance. *Biology Open*. 2018;7:bio036103.
- [22] Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, et al. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. 2013;501:212-6.
- [23] Procko E, Hedman R, Hamilton K, Seetharaman J, Fleishman SJ, Su M, et al. Computational design of a protein-based enzyme inhibitor. *Journal of Molecular Biology*. 2013;425:3563-75.
- [24] Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnology*. 2012;30:543-8.
- [25] Fujino Y, Fujita R, Wada K, Fujishige K, Kanamori T, Hunt L, et al. Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochemical and Biophysical Research Communications*. 2012;428:395-400.
- [26] Rogers JM, Passioura T, Suga H. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. *Proceedings of the National Academy of Sciences*. 2018;115:201809901.
- [27] Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DNA. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of Molecular Biology*. 2013;425:1363-77.
- [28] Roscoe BP, Bolon DNA. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *Journal of molecular biology*. 2014;18:1199-216.
- [29] Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*. 2015;200:413-22.
- [30] Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*. 2018.
- [31] Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Research*. 2014;42:1-8.
- [32] Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*. 2018;50:874-82.
- [33] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation* 2009. p. 1237-44.
- [34] Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC genomics* 2013.
- [35] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*. 2018:1-9.
- [36] Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of Molecular Biology: Elsevier B.V.*; 2014. p. 2692-701.
- [37] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nature Methods*. 2010;7:248-9.

- [38] Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. *BMC Genomics*. 2015;16:S1.
- [39] Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*. 2012;40:452-7.
- [40] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics: American Society of Human Genetics*; 2016. p. 877-85.
- [41] Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC genomics*2013. p. S3.
- [42] Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*. 2019;47:D1038-D43.
- [43] Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*. 2001;29:308-11.
- [44] Kawabata T, Ota M, Nishikawa K. The protein mutant database. *Nucleic Acids Research*. 1999;27:355-7.
- [45] Consortium TU. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*. 2019;47:D506-D15.
- [46] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*. 2016;44:D862-D8.
- [47] Reeb J, Hecht M, Mahlich Y, Bromberg Y, Rost B. Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLOS Computational Biology*. 2016;12:e1005047.
- [48] Pejaver V, Babbi G, Casadio R, Folkman L, Katsonis P, Kundu K, et al. Assessment of methods for predicting the effects of PTEN and TPMT protein variants. *Human Mutation*2019. p. 1495-506.
- [49] Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Systems*. 2018;6:116-24.e3.
- [50] Adkar BV, Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, et al. Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure*. 2012;20:371-81.
- [51] Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*. 2012;109:16858-63.
- [52] Brenan L, Andreev A, Cohen O, Pantel S, Kamburov A, Cacchiarelli D, et al. Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Reports*. 2016;17:1171-83.
- [53] Heredia JD, Park J, Brubaker RJ, Szymanski SK, Gill KS, Procko E. Mapping Interaction Sites on Human Chemokine Receptors by Deep Mutational Scanning. *The Journal of Immunology*. 2018;200:3825-39.
- [54] Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*. 2011;108:7896-901.
- [55] Hietpas RT, Bank C, Jensen JD, Bolon DNA. Shifting fitness landscapes in response to altered environments. *Evolution*. 2013;67:3512-22.
- [56] Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DNA. Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLoS Genetics*. 2013;9.
- [57] Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively Parallel Single Amino Acid Mutagenesis. *Nature Methods*. 2014;44:3516-21.

- [58] Klesmith JR, Bacik JP, Michalczyk R, Whitehead TA. Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in *E. coli*. *ACS Synthetic Biology*. 2015;4:1235-43.
- [59] Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, et al. Prospective functional classification of all possible missense variants in PPARG. *Nature Genetics*. 2016;48:1570-5.
- [60] Rockah-Shmuel L, Tóth-Petróczy Á, Tawfik DS. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Computational Biology*. 2015;11:1-28.
- [61] Romero PA, Tran TM, Abate AR. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*. 2015;112:7159-64.
- [62] Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, et al. Local fitness landscape of the green fluorescent protein. *Nature*. 2016;533:397-401.
- [63] Starita LM, Pruneda JN, Lo RS, Fowler DM, Kim HJ, Hiatt JB, et al. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*. 2013;110:E1263-E72.
- [64] Stiffler Michael A, Hekstra Doeke R, Ranganathan R. Evolvability as a Function of Purifying Selection in TEM-1 β -Lactamase. *Cell*. 2015;160:882-92.
- [65] Traxlmayr MW, Hasenhindl C, Hackl M, Stadlmayr G, Rybka JD, Borth N, et al. Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. *Journal of Molecular Biology*. 2012;423:397-412.
- [66] Wu Y, Weile J, Cote AG, Sun S, Knapp J, Verby M, et al. A web application and service for imputing and visualizing missense variant effect maps. *Bioinformatics*. 2019;35:3191-3.
- [67] Bromberg Y, Kahn PC, Rost B. Neutral and weakly nonneutral sequence variants may define individuality. *Proceedings of the National Academy of Sciences of the United States of America*. 2013;110:14255-60.
- [68] Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic acids research*. 2007;35:3823-35.
- [69] Andersen LL, Terczyńska-Dyla E, Mørk N, Scavenius C, Enghild JJ, Höning K, et al. Frequently used bioinformatics tools overestimate the damaging effect of allelic variants. *Genes and Immunity*. 2017;20:10-22.
- [70] Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*. 2015;36:513-23.
- [71] Miller M, Bromberg Y, Swint-Kruse L. Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Scientific Reports*. 2017;7:41329.
- [72] Cline MS, Babbi G, Bonache S, Cao Y, Casadio R, Cruz X, et al. Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. *Human Mutation* 2019. p. 1546-56.
- [73] Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends in Biochemical Sciences*. 2019:1-14.
- [74] Niroula A, Vihinen M. How good are pathogenicity predictors in detecting benign variants ? *PLoS computational biology*. 2019;15:1-17.
- [75] Bromberg Y, Rost B. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics*. 2008;24:207-12.
- [76] Miller M, Vitale D, Kahn PC, Rost B, Bromberg Y. Funtrp: Identifying Protein Positions for Variation Driven Functional Tuning. *Nucleic Acids Research* 2019.

- [77] Rives A, Goyal S, Meier J, Guo D, Ott M, Zitnick CL, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv preprint. 2019.
- [78] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling the language of life - deep learning protein sequences. bioRxiv preprint. 2019.
- [79] Pundir S, Martin MJ, O'Donovan C. UniProt Tools. Current protocols in bioinformatics. 2016;53:1.29.1-1..15.
- [80] Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al. An open-source platform to distribute and interpret data from multiplexed assays of variant effect. bioRxiv. 2019:555797.
- [81] Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. Molecular Biology and Evolution. 2014;31:1581-92.
- [82] Fabian P, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.
- [83] Jones E, Oliphant T, Peterson P, others a. SciPy: Open Source Scientific Tools for Python. 2001-.
- [84] Wilcox RR. Comparing dependent robust correlations. British Journal of Mathematical and Statistical Psychology. 2016;69:215-24.
- [85] Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;8:12-77.
- [86] Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2018.

Figure captions

Figure 1: Agreements between DMS experiments and variant effect predictions. Panels (a)-(e) show the performance for five prediction methods (PolyPhen-2 [37], SIFT [39], SNAP2 [38], Envision [49], and naïve Conservation) through hexbin plots of 17,781 deleterious effect SAVs in SetCommon (see Methods, Processing). Values on axes range from 0 (neutral) to 1 (maximal effect). Dashed red lines give linear least-squared regressions. Marginals denote distributions of experimental and predicted scores with a kernel density estimation overlaid in blue. The footer denotes Spearman ρ , Pearson R and the mean squared error together with the respective 95% confidence intervals. The method scores are given on the y-axes and reveal the method: (a) PolyPhen-2, (b) SIFT, (c) SNAP2, (d) Envision – the only method trained on DMS data, (e) naïve Conservation read off PSI-BLAST profiles. Panel (f) gives ROC curves for 13,796 deleterious effect SAVs which were classified into either neutral, defined by the middle 95% of the scores from synonymous variants, or effect (SetCommonSyn95, see Methods). Shaded areas around lines denote 95% confidence intervals. The legend denotes the AUC for each method along with the 95% confidence intervals. Horizontal dashed lines denote the default score threshold used by SNAP2 (blue) and SIFT (green).

Tables

Table 1: Number of SAVs in aggregated datasets ^Δ.

	Number of SAVs			
	Total	Neutral	Deleterious	Beneficial
SetAll	66089	818 *	45382	19889
SetCommon	32981	0	17781	15200
SetCommonSyn90	15621	8926	4545	2150
SetCommonSyn95	15621	10587	3209	1825
SetCommonSyn99	15621	13506	1548	567

^Δ SetAll depicts the total number of SAVs collected, while SetCommon contains only SAVs with predictions from every analyzed method. SetCommonSyn contains all SAVs with predictions where a thresholding scheme could be applied to yield classification of SAVs into neutral and effect (see Methods). The number of SAVs in every single DMS experiment are depicted in Fig. S1 and Table S1.

* The ccdB set classifies variant effect in categories and contains 818 non-synonymous variants which fall in the same category as the wt. Hence these SAVs could be considered neutral.

Table 2: Performance measures for five prediction methods on SetCommon ^Δ.

	deleterious SAVs (n = 17781)		beneficial SAVs (n = 15200)	
	ρ	MSE	ρ	MSE
PolyPhen-2	0.33 [0.32, 0.34]	0.45 [0.45, 0.46]	-0.09 [-0.10, -0.07]	0.4 [0.40, 0.41]
SIFT	0.23 [0.22, 0.25]	0.58 [0.57, 0.58]	-0.14 [-0.15, -0.12]	0.64 [0.63, 0.64]
SNAP2	0.41 [0.40, 0.42]	0.3 [0.30, 0.30]	0.02 [0.01, 0.04]	0.23 [0.23, 0.24]
Envision	0.1 [0.08, 0.11]	0.06 [0.06, 0.07]	-0.14 [-0.16, -0.13]	0.05 [0.04, 0.05]
Conservation	0.29 [0.27, 0.30]	0.19 [0.19, 0.19]	-0.08 [-0.09, -0.06]	0.19 [0.19, 0.20]

^Δ SetCommon denotes the set of SAVs with predictions from every method (see Methods). ρ denotes Spearman ρ (higher is better), MSE the mean squared error (lower is better, Methods, SOM_Note4). Normal Dist. denotes a random prediction of scores normally distributed around the mean of the experimental values and is given as a baseline. Values in brackets are 95% confidence intervals.