GigaScience

Integrating deep mutational scanning and lowthroughput mutagenesis data to predict the impact of amino acid variants

--Manuscript Draft--

Manuscript Number:	GIGA-D-23-00040	
Full Title:	Integrating deep mutational scanning and lowthroughput mutagenesis data to predict the impact of amino acid variants	
Article Type:	Research	
Funding Information:	Melbourne Research Scholarship	Mr Yunfan Fu
	National Health and Medical Research Council (116955)	Professor Anthony Troy Papenfuss
	Lorenzo and Pamela Galli Medical Research Trust	Professor Anthony Troy Papenfuss
	Stafford Fox Medical Research Foundation	Professor Anthony Troy Papenfuss
	National Human Genome Research Institute (RM1HG010461)	Dr Alan F. Rubin
	National Human Genome Research Institute (UM1HG011969)	Dr Alan F. Rubin
Abstract:	Background: Evaluating the impact of amino acid variants has been a critical challenge for studying protein function and interpreting genomic data. High-throughput experimental methods like deep mutational scanning (DMS) can measure the effect of large numbers of variants in a target protein, but because DMS studies have not been performed on all proteins, researchers also model DMS data computationally to estimate variant impacts by predictors. Results: In this study, we extended a linear regression-based predictor to explore whether incorporating data from alanine scanning (AS), a widely-used low-throughput mutagenesis method, would improve prediction results. To evaluate our model, we collected 146 AS datasets, mapping to 54 DMS datasets across 22 distinct proteins. Conclusions: We show that improved model performance depends on the compatibility of the DMS and AS assays, and the scale of improvement is closely related to the correlation between DMS and AS results.	
Corresponding Author:	Alan F Rubin, PhD Walter and Eliza Hall Institute of Medical Research Parkville, VIC AUSTRALIA	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Walter and Eliza Hall Institute of Medical Research	
Corresponding Author's Secondary Institution:		
First Author:	Yunfan Fu	
First Author Secondary Information:		
Order of Authors:	Yunfan Fu	
	Justin Bedő	
	Anthony Troy Papenfuss, BSc (Hons) PhD	
	Alan F. Rubin	
Order of Authors Secondary Information:		
Additional Information:		

Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.	Yes

Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

Integrating deep mutational scanning and low-throughput mutagenesis data to predict the impact of amino acid variants

Authors:

Yunfan Fu[1,2], Justin Bedő[1,3,*], Anthony T. Papenfuss[1,2,4,*,**], Alan F. Rubin[1,2,*,**]

Affiliations:

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, 3052, Victoria, Australia.

[2]Department of Medical Biology, The University of Melbourne, Melbourne, VIC 3010, Australia.

[3]School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC 3010, Australia.

[4]Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia.

* Contributed equally

** To whom correspondence should be addressed (papenfuss@wehi.edu.au & alan.rubin@wehi.edu.au)

Abstract

Background: Evaluating the impact of amino acid variants has been a critical challenge for studying protein function and interpreting genomic data. High-throughput experimental methods like deep mutational scanning (DMS) can measure the effect of large numbers of variants in a target protein, but because DMS studies have not been performed on all proteins, researchers also model DMS data computationally to estimate variant impacts by predictors.

Results: In this study, we extended a linear regression-based predictor to explore whether incorporating data from alanine scanning (AS), a widely-used low-throughput mutagenesis method, would improve prediction results. To evaluate our model, we collected 146 AS datasets, mapping to 54 DMS datasets across 22 distinct proteins.

Conclusions: We show that improved model performance depends on the compatibility of the DMS and AS assays, and the scale of improvement is closely related to the correlation between DMS and AS results.

Keywords: deep mutational scanning, alanine scanning, machine learning, predictor

1 Introduction

Deep mutational scanning (DMS) is a functional genomics method that can experimentally measure the impact of many thousands of protein variants by combining high-throughput sequencing with a functional assay [1]. In a typical DMS, a cDNA library of genetic variants of a target gene is generated, containing all possible single amino acid substitutions. This variant library is then expressed in a functional assay system where the variants can be selected based on their properties. The change in variant frequency in the pre- and post-selection populations is determined by high-throughput sequencing which is then used to calculate a multiplexed functional score that captures the variant’s impact [2–4]. The versatility of DMS assays makes it possible to measure variant impact on a wide range of protein properties, including protein binding [5,6], protein abundance [7–9], catalytic activity [10,11] and cell growth rate [12–14].

Computational studies have used DMS data to build predictive models of variant impact. These predictors use supervised or semi-supervised learning models trained on experimental DMS data and various protein features to make predictions [15–21]. Envision is one such method that used protein structural, physicochemical, and evolutionary features to predict variant effect scores and was trained on DMS data from 8 proteins using gradient boosting [15]. Another method, DeMaSk, predicted DMS scores by combining two evolutionary features (protein positional conservation and variant homologous frequency) with a DMS substitution matrix and was trained on data from 17 proteins using a linear model [17]. Deep learning algorithms have also been applied to build protein fitness predictors [16,18], which are usually based only on variant sequences.

Low-throughput mutagenesis experiments that measure tens of variants at a time have also been used extensively to study diverse protein properties, including substrate binding affinity [22,23], protein stability [24,25], and protein activity [26,27]. Alanine scanning (AS) is a widely-used low-throughput mutagenesis method [28,29], and AS data are available for many proteins. In this method, each targeted protein residue is substituted with alanine, and the impacts of these variants are measured by a functional assay [30]. AS experiments are typically used to identify functional hot spots or critical residues in the target protein [31,32] and have been used as a source of independent validation for DMS studies [27,33–35].

In this study, we explore whether a predictive model can be improved by incorporating low-throughput mutagenesis data (Fig 1). We find that AS data can increase prediction accuracy and that the improvement is related to the similarity of the functional assays and the correlation of DMS and AS results.

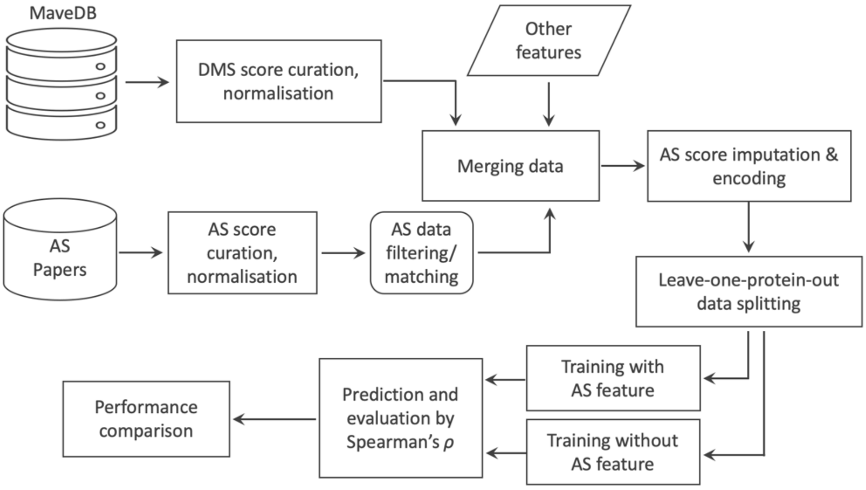
MaveDB
DMS score curation, normalisation
Other features
Merging data
AS score imputation & encoding
AS Papers
AS score curation, normalisation
AS data filtering/ matching
Leave-one-protein-out data splitting
Training with AS feature
Training without AS feature
Prediction and evaluation by Spearman's ρ
Performance comparison

Fig 1. Workflow for model training and testing. DMS and AS datasets are collected from online resources and are normalized. DMS and AS datasets targeting the same protein are then matched, filtered and merged. Two predictors are constructed and tested: the first uses DMS data, AS data and other protein features, and the second uses only DMS data and the same other protein features.

2 Results

2.1 Overview of DMS and alanine scanning (AS) data

To build the predictive model, 130 DMS datasets were collected from MaveDB [36,37] (Supplementary table 1). We searched the literature and found 146 AS datasets targeting the same proteins as 54 of the DMS datasets. In total, we obtained both DMS and AS data for 22 different proteins: 17 human proteins, three yeast proteins, and two bacterial proteins. Most DMS experiments were highly complete, with a mean coverage of 95.0% of all possible single amino acid substitutions assayed in the target region, comprising 373,219 total protein variant measurements. AS data were only available on a small number of protein residues (Fig 2), and we were able to curate 1,480 alanine substitution scores from the 146 studies. Variant scores from collected DMS and AS studies were linearly normalized to a common scale (see Methods) to make them comparable across datasets (Fig S1).

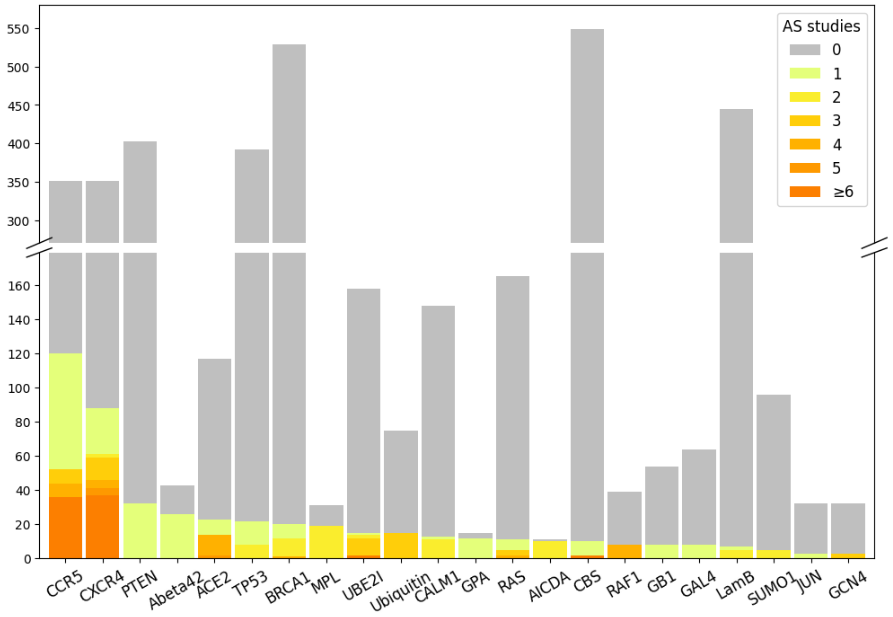

Fig 2. DMS data generally cover more protein residues than AS data. Each bar shows the number of residues assayed by DMS studies on given target proteins. Colour indicates the number of AS studies available for the DMS-tested residues.

2.2 The correlation of DMS and AS scores is related to assay compatibility

To evaluate the similarity of AS and DMS scores, we calculated Spearman's correlation (ρ) between the AS scores and DMS scores for the same alanine substitutions. Since each protein may have results from several AS and DMS experiments, we calculated ρ between each possible pair. The median ρ over DMS and AS data (DMS/AS) pairs was 0.2, indicating that the experimental scores were poorly correlated overall (Fig 3).

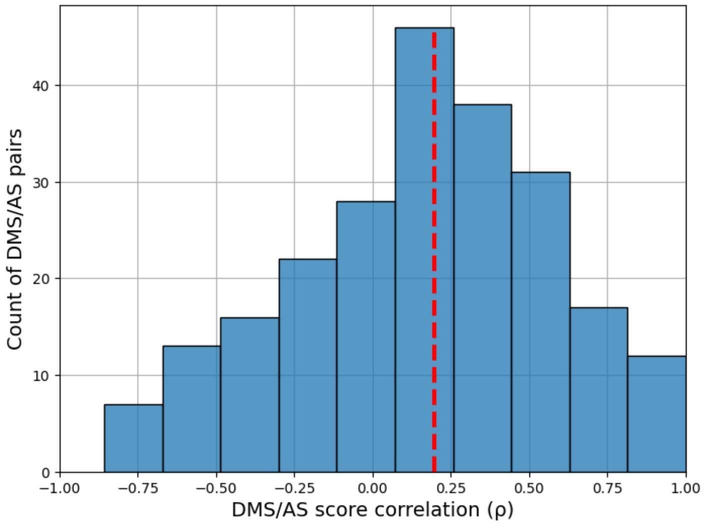

Fig 3. Correlation between DMS and AS data shows substantial variation. We calculated Spearman's ρ between alanine substitution scores in each pair of AS and DMS data. The results for pairs with less than three alanine substitutions are removed. The red dashed line shows the median ρ.

We then considered if differences between AS and DMS assay designs might contribute to this low agreement between scores. To explore this, we developed a decision tree (Fig S2) to classify whether DMS/AS pairs had low, medium, or high assay compatibility, which we defined as a similarity measurement of the functional assays performed. For example, the DMS assay measuring the binding affinity of a cell surface protein, CXCR4, to its natural ligand [38] has high compatibility with the AS experiment also measuring this ligand binding but has low compatibility with the study on CXCR4's ability to facilitate virus infection [39]. A full assay compatibility table can be found in Supplementary Table 1 with the compatibility classifications and justification for each pair. We then compared DMS and AS score correlation for each compatibility class and found that score correlations were closely related to assay compatibility. Data from low compatibility assays had a median correlation of 0.15, rising to 0.19 for medium compatibility assays and 0.40 for high compatibility assays (Fig 4). This link between assay compatibility and score correlation indicates that our decision tree approach was able to capture the similarity between assay systems.

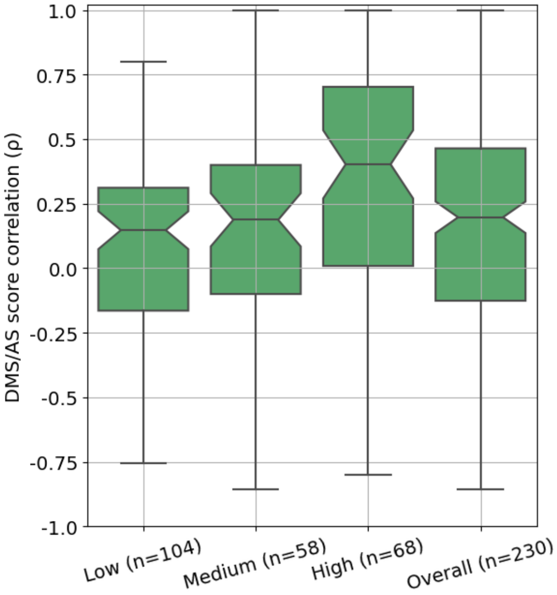

Fig 4. DMS and AS data pairs with high assay compatibility show a higher score correlation. Each box represents Spearman's ρ between DMS and AS data pairs of classified assay compatibility or the overall result. The correlation coefficients are calculated between alanine substitution scores in each pair of AS and DMS data. Results for data pairs with less than three alanine substitutions are removed.

2.3 Compatible AS data improve DMS score prediction accuracy

To test if incorporating AS data into DMS score models would improve prediction accuracy, we decided to build a new model based on DeMaSk [17]. We chose DeMaSk because it showed better performance compared to similar methods and was straightforward to modify. The published DeMaSk model predicts DMS scores using protein positional conservation, variant homologous frequency, and substitution score matrix, and we incorporated AS data as an additional feature. Our new predictor was modelled with all 130 DMS we collected and we applied a leave-one-protein-out cross-validation approach to training and testing [15]. Prediction performance was evaluated using the Spearman's correlation (ρ) between the experimentally-derived DMS scores and the predicted scores for each pair of DMS and AS studies. The performance of our DMS/AS model was compared with a model trained only on DMS data, equivalent to retrained DeMaSk (Fig S3), by calculating the change of prediction ρ (see Methods).

We trained our model with either all or a subset of AS data we collected (Fig 5, Table S1). We first integrated all 146 AS data collected for training and evaluation but observed only a modest improvement of prediction ρ (Fig 5 left box, and Fig S4). We then retrained and evaluated our model on filtered AS data with only high compatibility assays, and observed a median increase in prediction Spearman's ρ of 0.1 compared to the results with no AS data (Fig 5 middle box, and Fig S4). However, training with both high and medium compatibility pairs reduced the performance improvement (Fig S5). These results indicate that medium and low compatibility pairs might provide inconsistent training data, degrading model performance. We also evaluated the impact of including high compatibility AS data in an alternative model based on Envison [15], and found similar results (Fig S6). To differentiate between high assay compatibility and high DMS/AS score correlation, we trained the model using the most highly correlated AS result for each DMS dataset (see Methods). Although the upper quartile was high, the median performance change of this predictor was lower than the high assay compatibility model, suggesting that matching with the highest score correlation alone is insufficient (Fig 5 right box).

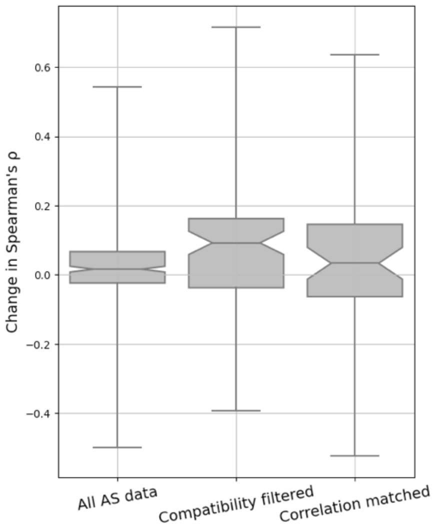

Fig 5. Performance of variant impact prediction is improved using AS data with high assay compatibility. The change of prediction ρ for each DMS and AS data pair is shown as box plots. A higher value represents higher prediction accuracy achieved for using AS data. Different approaches to filtering/matching the data are shown on the x-axis: "All AS data" used all available data; "Compatibility filtered" used only data of high assay compatibility; "Correlation matched" used only data with the highest regularised correlation for each DMS dataset.

To further explore the higher performance of compatibility-filtered predictor, we examined the relationship between prediction ρ change and score correlation for each high compatibility DMS/AS pair (Fig 6). For most pairs, prediction performance was improved by using AS data, and the scale of improvement was also related to the score correlation. This relationship could also be observed for multiple DMS/AS pairs from an individual protein, such as CXCR4 and CCR5. We saw the same trend in the predictor trained with all DMS/AS pairs but noted that the performance even of highly correlated pairs was worse, likely due to the influence of low compatibility training data on the model (Fig S7).

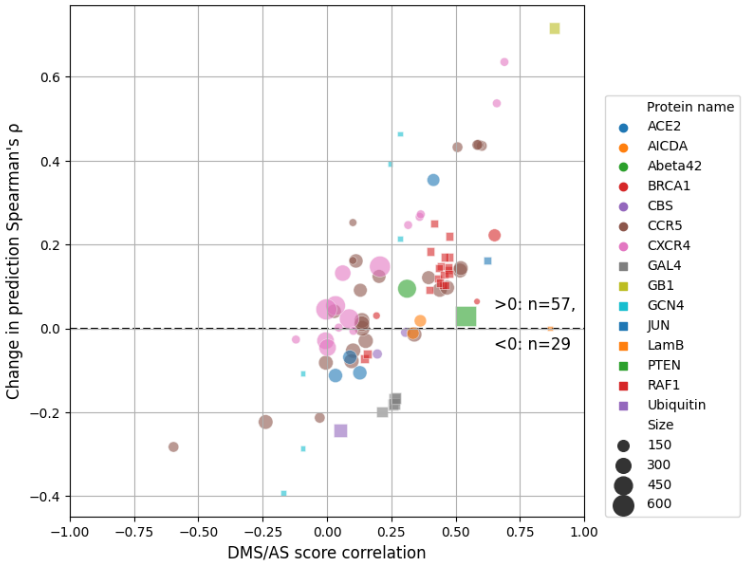

Fig 6. Prediction performance change is related to DMS and AS score correlation. Each dot represents a filtered DMS/AS data pair of high assay compatibility. The vertical axis shows the change of prediction *ρ* by using AS data (larger means higher performance achieved by using AS data). The horizontal axis shows the DMS/AS score correlation for *all* variants on the matched residues rather than just alanine substitutions. The colours and shapes of the dots correspond to the target protein, and size indicates the number of variants in each data pair.

We also explored the consequences of the sparsity of AS data on our model in two ways: by using a boosting approach that focuses only on residues with AS data (Fig S8) and by using complete alanine substitution information from DMS as the AS feature (Fig S9). Both of these approaches performed very similarly to the primary model constructed using high-compatibility DMS/AS data and simple mean score imputation.

To test the influence of amino acids on our predictor, we grouped the prediction results by either wild-type or variant amino acid and calculated the prediction improvement when AS data were included (Fig 7). We found that 14 of 19 wild-type amino acids performed better with the addition of AS data, with cysteine showing the largest improvement and performing worst in the model lacking AS data. 18 of 20 variant amino acids benefited from the inclusion of AS data, with marginal performance decrease on lysine and aspartic acid (|Δρ|<0.01) (Fig 7).

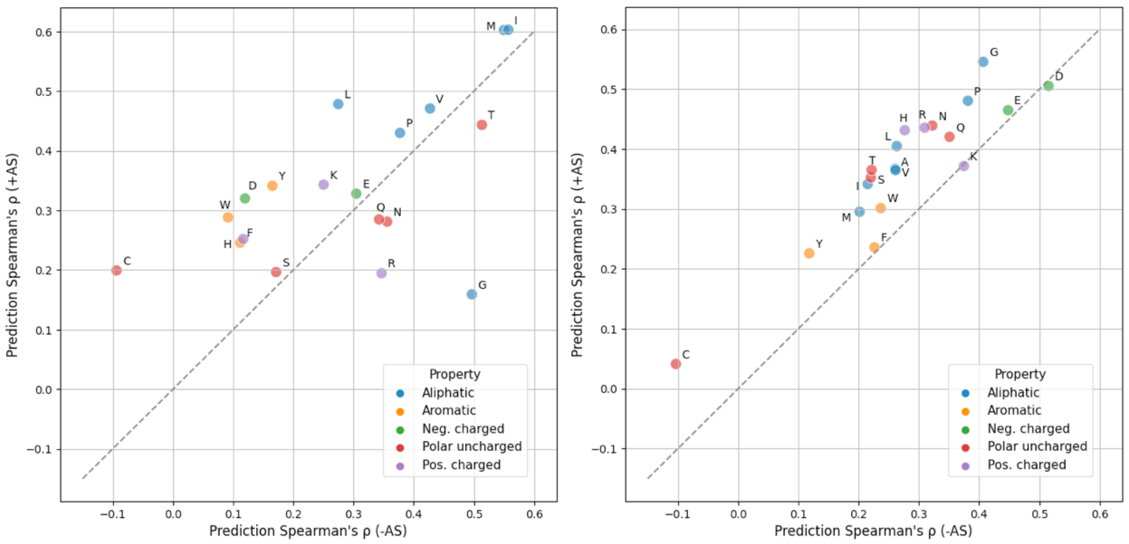

Fig 7. Model perfomance is generally improved for each wild-type and variant amino acid. Prediction Spearman's ρ when using (y-axis) or not using (x-axis) AS data on each wild-type (left) or variant (right) amino acid is shown in the scatter plots. The results are coloured according to the property of each amino acid type. Alanine (A) result is not applicable in the first figure since alanine scanning data are always missing when the wildtype is alanine itself. Absolute count for each amino acid can be found in Fig S10. (Neg.: negatively, Pos.: positively)

3 Discussion

In this study, we integrated alanine scanning (AS) data into deep mutational scanning (DMS) score prediction, leading to modest improvements in the accuracy of variant score prediction. We also explored the impact of the diversity of protein properties measured by DMS and AS. Filtering DMS and AS data based on our manual classification of assay type compatibility led to improved prediction performance.

A potential shortcoming of our current approach is that AS data were available for only a small proportion of the DMS data. Although most recent DMS studies can analyze variants of the whole protein, most AS experiments only cover a handful of residues in the target protein, leaving missing AS scores for the vast majority of residues. We explored this here and found that alternative methods for addressing the sparsity of AS data did not improve or degrade performance, but we anticipate further improved prediction accuracy if the low completeness and unevenness of AS data are appropriately handled before modelling, such as by advanced imputation methods [48,49].

In this study, we identified the importance of DMS/AS assay compatibility as a crucial factor for improving prediction accuracy. An issue with using this concept is that it further shrinks already sparse data. It also fails to take advantage of the fact that even for low compatible assays some fundamental information like protein stability can still be mutually captured. Instead of hard filtering, proper implementation of this underlying information may facilitate variant impact prediction in the future. Nonetheless, filtering on assay compatibility still leads to performance improvement. We also briefly explored whether the consistency of DMS and AS scores can be considered more directly by matching the best correlated AS data for each DMS dataset. Consistency is partially driven by assay compatibility but also reflects other features of the data, such as bias and noise. While we picked the most correlated pair for each DMS, we did not threshold the correlation, potentially including data pairs that were poor matches.

The concepts of compatibility and data quality are also relevant to training any DMS-based predictors. DMS assays have been developed to measure variant impacts to distinct protein properties, and a variant can behave similarly to wildtype when measured by one assay yet show altered protein properties in other assay results, which are frequently found in regions with specific biochemical functions [50–55]. With more experimental assays to be applied, the diverse measurements may impede the progress of future DMS-based predictors unless this assay effect is properly addressed, for example, by building assay specific predictors. Measurement error is another source of DMS data heterogeneity that potentially affects the model performance. In our current study, DMS scores of protein variants are weighted equally while training. Adjustable weighting can be applied in future studies to adapt the distinct experimental error between individual variants and datasets, reducing the influence of low-confident data.

In summary, we conclude that the careful inclusion of low-throughput mutagenesis data improves the prediction of DMS scores, and the approaches described here can potentially be applied to other prediction methods.

4 Availability of supporting source code and requirements

Project name: DMS_with_Alanine_scan

Project home page: https://github.com/PapenfussLab/DMS_with_Alanine_scan

Operating system: Platform independent

Programming language: Python

Other requirements: Python 3.10.6

Licence: MIT Licence

5 List of abbreviations

DMS: deep mutational scanning

AS: alanine scanning

6 Supporting information

Supplementary Table 1: All candidate DMS and alanine scanning data with detailed dataset information.

Supplementary Table 2: Normalized DMS dataset with protein property features.

Supplementary Table 3: Normalized alanine scanning dataset.

7 Author contributions

YF developed the software and wrote the initial draft of the manuscript. AFR conceived the study. JB, AFR, and ATP oversaw the project. All authors reviewed, contributed to, and approved the manuscript.

8 Funding

YF is supported by Melbourne Research Scholarship. ATP was supported by an Australian National Health and Medical Research Council (NHMRC) Senior Research Fellowship (1116955). JB, AFR and ATP were supported by the Lorenzo and Pamela Galli Medical Research Trust. JB and ATP were supported by the Stafford Fox Medical Research Foundation. AFR was supported by the National Human Genome Research Institute of the NIH under award numbers RM1HG010461 and UM1HG011969. The research benefitted from support from the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support.

9 Methods

9.1 DMS data collection

DMS data were downloaded from MaveDB [36,37] which were then filtered and curated. DMS experiments targeting antibody and virus proteins were removed because of their potentially unique functionality. We retrieved the UniProt accession ID of target proteins by searching the protein names or sequences in UniProt [56], and proteins lacking available UniProt ID were also excluded. Datasets that are computationally processed or their wildtype-like and nonsense-like scores (see Normalization) cannot be identified were also filtered out (Supplementary Table 1). All missense variants with only a single amino acid substitution were curated from the DMS studies for our analysis. A total of 130 DMS experiments from 53 studies [5,6,9–14,27,33–35,38,57–94] were collected for our analysis.

9.2 Collection of AS data and other features

The following process was used to search for candidate AS studies. Papers were identified by searching on PubMed and Google Scholar for the “alanine scan” or “alanine scanning” together with the name of candidate proteins. While searching in Google Scholar, we included the protein’s UniProt ID rather than molecule name as the search term to reduce false positives. Appropriate AS data were collected from the search results. Western blot results were transformed to values by ImageJ if it was the only experimental data available in the study. A total 146 AS experiments were collected from 45 distinct studies [22–24,26,27,39–42,44,45,84,95–127].

Protein features of Shannon entropy and the logarithm of variant amino acid frequency were downloaded from the DeMaSk online toolkit [17]. The substitution score matrix feature was calculated from the mean of training DMS scores for each of the 380 possible amino acid substitutions before each iteration of cross-validation.

9.3 Normalization

DMS and AS datasets were normalized to a common scale using the following approach adapted from previous studies [15,43]. Let D denotes a protein study measuring scores s_i^D for a single variant i, s_{wt}^D denotes the scores for wildtype and s_{non}^D represents the score for nonsense-like variants. The normalized scores $s_i'^D$ are given by:

$$s_i'^D := \frac{s_i^D - s_{wt}^D}{s_{wt}^D - s_{non}^D} + 1$$

Wild-type scores were directly identified from the paper or the median score of synonymous variants. For DMS data, since not all DMS studies report score of nonsense variants, we defined the nonsense-like scores as the median DMS scores for the 1% missense variants with the strongest loss of function for each dataset. For AS data, nonsense-like scores were either defined according to the paper or using the extreme values (Supplementary Table 1).

9.4 AS data filtering and matching

AS data subsets were filtered/matched according to either assay compatibility or score correlation. For assay compatibility filtering, DMS and AS assay pairs were first classified into three levels of compatibility (Fig S2). For each DMS dataset, we first tried to use only AS data with high assay compatibility for further modelling, removing AS data of medium and low assay compatibility. We then also tried to model with AS data of both high and medium assay compatibility.

For score correlation matching, Spearman's correlation (ρ) is calculated between alanine substitution scores in each pair of AS and DMS data. To avoid influence from the size of AS datasets, we regularised the ρ value by empirical copula [128]:

$$\rho_r := \rho \times \frac{n-1}{n+1}$$

where ρ_r is the regularised correlation coefficient, and n is the number of alanine substitutions used for correlation calculation. For each DMS dataset, AS result with the highest ρ_r was picked for modelling.

9.5 AS data pre-processing

AS data were pre-processed prior to modelling. For variants without available (filtered/matched) AS data, their AS scores were imputed with the mean value of all available AS scores. Then the AS data were encoded by the wild-type and variant amino acid type with one-hot-encoding. For each variant, the AS feature is expanded with two one-hot vectors. Each of the vectors has 19 zeros and one non-zero value which was the AS score, with the location of the non-zero value indicating the wild-type or variant amino acid type.

9.6 Training and evaluation of DMS score predictor

To build the predictors, we performed linear regression using the function `sklearn.linear_model.LinearRegression` from scikit-learn [129]. Training and validation data were separated with leave-one-protein-out cross-validation. In this process, data from one protein were withheld for subsequent validation, and the rest were used for training. This process was iterated over all proteins in the data. Variants were inversely weighted during the training process by the number of measurements available, thus compensating for some regions having greater coverage with DMS and AS assays. Predictors were trained on protein features, DMS data and (optionally) AS data using four different filtering or matching strategies: i) all DMS/AS data, ii) compatibility-filtered DMS/AS data, iii) correlation-matched DMS/AS data, and iv) a control, constructed using DMS data only.

In the evaluation process, let *V* be protein variants assayed by both DMS study *D* and AS study *A*. Variant scores are predicted by the previously mentioned predictors either using AS data

($\hat{s}_V^A$) or not ($\hat{s}_V$). Spearman's correlation (ρ) was calculated between the DMS scores s_V^D and each set of predicted scores. The difference of ρ was used to evaluate the performance change ($\Delta\rho_V$).

$$\rho_V^A = \text{Spearman's correlation}(\hat{s}_V^A, s_V^D)$$

$$\rho_V = \text{Spearman's correlation}(\hat{s}_V, s_V^D)$$

$$\Delta\rho_V = \rho_V^A - \rho_V$$

To evaluate, we iterated over variants from each pair of DMS/AS studies. Results were dropped for variants *V* with only one protein residue available during analysis and visualization.

10 References

1. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nature Methods*. 2014; doi: 10.1038/nmeth.3027.

2. Findlay GM. Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. *Human Molecular Genetics*. 2021; doi: 10.1093/hmg/ddab219.

3. Geck RC, Boyle G, Amorosi CJ, Fowler DM, Dunham MJ. Measuring Pharmacogene Variant Function at Scale Using Multiplexed Assays. *Annual Review of Pharmacology and Toxicology*. 2022; doi: 10.1146/annurev-pharmtox-032221-085807.

4. Weile J, Roth FP. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Hum Genet*. 2018; doi: 10.1007/s00439-018-1916-x.

5. Diss G, Lehner B. The genetic landscape of a physical interaction. *eLife*. 2018; doi: 10.7554/eLife.32472.

6. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al.. High-resolution mapping of protein sequence-function relationships. *Nature Methods*. 2010; doi: 10.1038/nmeth.1492.

7. Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, et al.. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *The American Journal of Human Genetics*. 2021; doi: 10.1016/j.ajhg.2021.07.001.

8. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*. 2022; doi: 10.1038/s41586-022-04586-4.

9. 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; doi: 10.1038/s41588-018-0122-z.

10. Mighell TL, Evans-Dutson S, O'Roak BJ. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *The American Journal of Human Genetics*. 2018; doi: 10.1016/j.ajhg.2018.03.018.

11. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell*. 2015; doi: 10.1016/j.cell.2015.01.035.

12. Ahler E, Register AC, Chakraborty S, Fang L, Dieter EM, Sitko KA, et al.. A Combined Approach Reveals a Regulatory Mechanism Coupling Src's Kinase Activity, Localization, and Phosphotransferase-Independent Functions. *Molecular Cell*. 2019; doi: 10.1016/j.molcel.2019.02.003.

13. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, et al.. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nature Genetics*. Nature Publishing Group; 2018; doi: 10.1038/s41588-018-0204-y.

14. 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; doi: 10.1016/j.jmb.2013.01.032.

15. Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Systems*. 2018; doi: 10.1016/j.cels.2017.11.003.

16. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods*. 2019; doi: 10.1038/s41592-019-0598-1.

17. Munro D, Singh M. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. Xu J, editor. *Bioinformatics*. 2020; doi: 10.1093/bioinformatics/btaa1030.

18. Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low- N protein engineering with data-efficient deep learning. *Nature Methods*. Nature Publishing Group; 2021; doi: 10.1038/s41592-021-01100-y.

19. Høie MH, Cagiada M, Beck Frederiksen AH, Stein A, Lindorff-Larsen K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Reports*. 2022; doi: 10.1016/j.celrep.2021.110207.

20. Wu Y, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. *The American Journal of Human Genetics*. 2021; doi: 10.1016/j.ajhg.2021.08.012.

21. Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol*. 2022; doi: 10.1038/s41587-021-01146-5.

22. Block C, Janknecht R, Herrmann C, Nassar N, Wittinghofer A. Quantitative structure-activity analysis correlating Ras/Raf interaction in vitro to Raf activation in vivo. *Nature Structural Biology*. Nature Publishing Group; 1996; doi: 10.1038/nsb0396-244.

23. Sloan DJ, Hellinga HW. Dissection of the protein G B1 domain binding site for human IgG Fc fragment. *Protein Science*. 1999; doi: 10.1110/ps.8.8.1643.

24. Fleming KG, Engelman DM. Specificity in transmembrane helix–helix interactions can define a hierarchy of stability for sequence variants. *PNAS*. National Academy of Sciences; 2001; doi: 10.1073/pnas.251367498.

25. Shibata Y, White JF, Serrano-Vega MJ, Magnani F, Aloia AL, Grisshammer R, et al.. Thermostabilization of the Neurotensin Receptor NTS1. *Journal of Molecular Biology*. 2009; doi: 10.1016/j.jmb.2009.04.068.

26. Brzovic PS, Heikaus CC, Kisselev L, Vernon R, Herbig E, Pacheco D, et al.. The Acidic Transcription Activator Gcn4 Binds the Mediator Subunit Gal11/Med15 Using a Simple Protein Interface Forming a Fuzzy Complex. *Molecular Cell*. 2011; doi: 10.1016/j.molcel.2011.11.008.

27. Gajula KS, Huwe PJ, Mo CY, Crawford DJ, Stivers JT, Radhakrishnan R, et al.. High-throughput mutagenesis reveals functional determinants for DNA targeting by activation-induced deaminase. *Nucleic Acids Research*. 2014; doi: 10.1093/nar/gku689.

28. Kortemme T, Kim DE, Baker D. Computational Alanine Scanning of Protein-Protein Interfaces. *Science's STKE*. American Association for the Advancement of Science; 2004; doi: 10.1126/stke.2192004pl2.

29. Morrison KL, Weiss GA. Combinatorial alanine-scanning. *Current Opinion in Chemical Biology*. 2001; doi: 10.1016/S1367-5931(00)00206-4.

30. Cunningham BC, Wells JA. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*. American Association for the Advancement of Science; 1989; doi: 10.1126/science.2471267.

31. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. *Current Opinion in Structural Biology*. 2002; doi: 10.1016/S0959-440X(02)00283-X.

32. Eustache S, Leprince J, Tufféry P. Progress with peptide scanning to study structure-activity relationships: the implications for drug discovery. *Expert Opinion on Drug Discovery*. 2016; doi: 10.1080/17460441.2016.1201058.

33. Olson CA, Wu NC, Sun R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*. 2014; doi: 10.1016/j.cub.2014.09.072.

34. Staller MV, Holehouse AS, Swain-Lenz D, Das RK, Pappu RV, Cohen BA. A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Systems*. 2018; doi: 10.1016/j.cels.2018.01.015.

35. Gray VE, Sitko K, Kameni FZN, Williamson M, Stephany JJ, Hasle N, et al.. Elucidating the Molecular Determinants of Aβ Aggregation with Deep Mutational Scanning. *G3 (Bethesda)*. 2019; doi: 10.1534/g3.119.400535.

36. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al.. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol*. 2019; doi: 10.1186/s13059-019-1845-6.

37. Rubin AF, Min JK, Rollins NJ, Da EY, Esposito D, Harrington M, et al.. MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. bioRxiv;

38. 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*. American Association of Immunologists; 2018; doi: 10.4049/jimmunol.1800343.

39. Tian S, Choi W-T, Liu D, Pesavento J, Wang Y, An J, et al.. Distinct Functional Sites for Human Immunodeficiency Virus Type 1 and Stromal Cell-Derived Factor 1α on CXCR4 Transmembrane Helical Domains. *JVI*. 2005; doi: 10.1128/JVI.79.20.12667-12673.2005.

40. Chabot DJ, Zhang P-F, Quinnan GV, Broder CC. Mutagenesis of CXCR4 Identifies Important Domains for Human Immunodeficiency Virus Type 1 X4 Isolate Envelope-Mediated Membrane Fusion and Virus Entry and Reveals Cryptic Coreceptor Activity for R5 Isolates. *J Virol*. 1999; doi: 10.1128/JVI.73.8.6598-6609.1999.

41. Han DP, Penn-Nicholson A, Cho MW. Identification of critical determinants on ACE2 for SARS-CoV entry and development of a potent entry inhibitor. *Virology*. 2006; doi: 10.1016/j.virol.2006.01.029.

42. Fujita–Yoshigaki J, Shirouzu M, Ito Y, Hattori S, Furuyama S, Nishimura S, et al.. A Constitutive Effector Region on the C-terminal Side of Switch I of the Ras Protein. *J Biol Chem*. American Society for Biochemistry and Molecular Biology; 1995; doi: 10.1074/jbc.270.9.4661.

43. Gray VE, Hause RJ, Fowler DM. Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics*. 2017; doi: 10.1534/genetics.117.300064.

44. Hidalgo P, Ansari AZ, Schmidt P, Hare B, Simkovich N, Farrell S, et al.. Recruitment of the transcriptional machinery through GAL11P: structure and interactions of the GAL4 dimerization domain. *Genes Dev*. 2001; doi: 10.1101/gad.873901.

45. Rodríguez-Escudero I, Oliver MD, Andrés-Pons A, Molina M, Cid VJ, Pulido R. A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. *Human Molecular Genetics*. 2011; doi: 10.1093/hmg/ddr337.

46. Schröter C, Günther R, Rhiel L, Becker S, Toleikis L, Doerner A, et al.. A generic approach to engineer antibody pH-switches using combinatorial histidine scanning libraries and yeast display. *mAbs*. 2015; doi: 10.4161/19420862.2014.985993.

47. Starace DM, Bezanilla F. Histidine Scanning Mutagenesis of Basic Residues of the S4 Segment of the Shaker K+ Channel. *J Gen Physiol*. 117:469–902001;

48. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012; doi: 10.1093/bioinformatics/btr597.

49. 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. Schwartz R, editor. *Bioinformatics*. 2019; doi: 10.1093/bioinformatics/btz012.

50. Cagiada M, Johansson KE, Valanciute A, Nielsen SV, Hartmann-Petersen R, Yang JJ, et al.. Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. Ozkan B, editor. *Molecular Biology and Evolution*. 2021; doi: 10.1093/molbev/msab095.

51. Cagiada M, Bottaro S, Lindemose S, Schenstrøm SM, Stein A, Hartmann-Petersen R, et al.. Discovering functionally important sites in proteins. bioRxiv;

52. Jepsen MM, Fowler DM, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Chapter 5 - Classifying disease-associated variants using measures of protein activity and stability. In: Pey AL, editor. *Protein Homeostasis Diseases*. Academic Press;

53. Matreyek KA, Stephany JJ, Ahler E, Fowler DM. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Med*. 2021; doi: 10.1186/s13073-021-00984-x.

54. Mighell TL, Thacker S, Fombonne E, Eng C, O’Roak BJ. An Integrated Deep-Mutational-Scanning Approach Provides Clinical Insights on PTEN Genotype-Phenotype Relationships. *The American Journal of Human Genetics*. 2020; doi: 10.1016/j.ajhg.2020.04.014.

55. Nielsen SV, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Multiplexed assays reveal effects of missense variants in MSH2 and cancer predisposition. *PLOS Genetics*. Public Library of Science; 2021; doi: 10.1371/journal.pgen.1009496.

56. The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, et al.. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. 2021; doi: 10.1093/nar/gkaa1100.

57. Andrews B, Fields S. Distinct patterns of mutational sensitivity for λ resistance and maltodextrin transport in Escherichia coli LamB. *Microb Genom*. 2020; doi: 10.1099/mgen.0.000364.

58. Bandaru P, Shah NH, Bhattacharyya M, Barton JP, Kondo Y, Cofsky JC, et al.. Deconstruction of the Ras switching cycle through saturation mutagenesis. *eLife*. 2017; doi: 10.7554/eLife.27810.

59. Bolognesi B, Faure AJ, Seuma M, Schmiedel JM, Tartaglia GG, Lehner B. The mutational landscape of a prion-like domain. *Nat Commun*. 2019; doi: 10.1038/s41467-019-12101-z.

60. Bridgford JL, Lee SM, Lee CMM, Guglielmelli P, Rumi E, Pietra D, et al.. Novel drivers and modifiers of MPL-dependent oncogenic transformation identified by deep mutational scanning. *Blood*. American Society of Hematology; 2020; doi: 10.1182/blood.2019002561.

61. Chan KK, Dorosky D, Sharma P, Abbasi SA, Dye JM, Kranz DM, et al.. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science*. American Association for the Advancement of Science; 2020; doi: 10.1126/science.abc0870.

62. Chiasson MA, Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, Verby M, et al.. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *Elife*. 2020; doi: 10.7554/eLife.58026.

63. Elazar A, Weinstein J, Biran I, Fridman Y, Bibi E, Fleishman SJ. Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. Shan Y, editor. *eLife*. eLife Sciences Publications, Ltd; 2016; doi: 10.7554/eLife.12125.

64. 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; doi: 10.1038/s41586-018-0461-z.

65. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A Comprehensive, High-Resolution Map of a Gene’s Fitness Landscape. *Mol Biol Evol*. 2014; doi: 10.1093/molbev/msu081.

66. Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*. 2011; doi: 10.1073/pnas.1016024108.

67. Hietpas RT, Bank C, Jensen JD, Bolon DNA. Shifting fitness landscapes in response to altered environments. *Evolution*. 2013; doi: 10.1111/evo.12207.

68. Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DNA. Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLOS Genetics*. Public Library of Science; 2013; doi: 10.1371/journal.pgen.1003600.

69. Jiang RJ. Exhaustive Mapping of Missense Variation in Coronary Heart Disease-related Genes [Thesis]. University of Toronto;

70. Keskin A, Akdoğan E, Dunn CD. Evidence for Amino Acid Snorkeling from a High-Resolution, *In Vivo* Analysis of Fis1 Tail-Anchor Insertion at the Mitochondrial Outer Membrane. *Genetics*. 2017; doi: 10.1534/genetics.116.196428.

71. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. *Nat Methods*. 2015; doi: 10.1038/nmeth.3223.

72. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, et al.. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation

Pattern and Evolutionary Conservation. *Molecular Cell*. Elsevier; 2018; doi: 10.1016/j.molcel.2018.06.012.

73. Kowalsky CA, Whitehead TA. Determination of binding affinity upon mutation for type I dockerin–cohesin complexes from Clostridium thermocellum and Clostridium cellulolyticum using deep sequencing. *Proteins: Structure, Function, and Bioinformatics*. 2016; doi: 10.1002/prot.25175.

74. McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. *Nature*. 2012; doi: 10.1038/nature11500.

75. Melamed D, Young DL, Gamble CE, Miller CR, Fields S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA*. 2013; doi: 10.1261/rna.040709.113.

76. Mishra P, Flynn JM, Starr TN, Bolon DNA. Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. *Cell Reports*. 2016; doi: 10.1016/j.celrep.2016.03.046.

77. Nedrud D, Coyote-Maestas W, Schmidt D. A large-scale survey of pairwise epistasis reveals a mechanism for evolutionary expansion and specialization of PDZ domains. *Proteins: Structure, Function, and Bioinformatics*. 2021; doi: 10.1002/prot.26067.

78. Newberry RW, Arhar T, Costello J, Hartoularos GC, Maxwell AM, Naing ZZC, et al.. Robust Sequence Determinants of α-Synuclein Toxicity in Yeast Implicate Membrane Binding. *ACS Chem Biol*. 2020; doi: 10.1021/acschembio.0c00339.

79. Newberry RW, Leong JT, Chow ED, Kampmann M, DeGrado WF. Deep mutational scanning reveals the structural basis for α-synuclein activity. *Nat Chem Biol*. 2020; doi: 10.1038/s41589-020-0480-6.

80. Roscoe BP, Bolon DNA. Systematic Exploration of Ubiquitin Sequence, E1 Activation Efficiency, and Experimental Fitness in Yeast. *Journal of Molecular Biology*. 2014; doi: 10.1016/j.jmb.2014.05.019.

81. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, et al.. Local fitness landscape of the green fluorescent protein. *Nature*. Nature Publishing Group; 2016; doi: 10.1038/nature17995.

82. Silverstein RA, Sun S, Verby M, Weile J, Wu Y, Roth FP. A systematic genotype-phenotype map for missense variants in the human intellectual disability-associated gene GDI1. bioRxiv;

83. 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. *PNAS*. 2013; doi: 10.1073/pnas.1303309110.

84. 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; doi: 10.1534/genetics.115.175802.

85. Starita LM, Islam MM, Banerjee T, Adamovich AI, Gullingsrud J, Fields S, et al.. A Multiplex Homology-Directed DNA Repair Assay Reveals the Impact of More Than 1,000 BRCA1 Missense Substitution Variants on Protein Function. *The American Journal of Human Genetics*. 2018; doi: 10.1016/j.ajhg.2018.07.016.

86. Suiter CC, Moriyama T, Matreyek KA, Yang W, Scaletti ER, Nishii R, et al.. Massively parallel variant characterization identifies *NUDT15* alleles associated with thiopurine toxicity. *Proc Natl Acad Sci USA*. 2020; doi: 10.1073/pnas.1915680117.

87. Sun S, Weile J, Verby M, Wu Y, Wang Y, Cote AG, et al.. A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. *Genome Med*. 2020; doi: 10.1186/s13073-020-0711-1.

88. Thompson S, Zhang Y, Ingle C, Reynolds KA, Kortemme T. Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme. *eLife*. 2020; doi: 10.7554/eLife.53476.

89. Trenker R, Wu X, Nguyen JV, Wilcox S, Rubin AF, Call ME, et al.. Human and viral membrane–associated E3 ubiquitin ligases MARCH1 and MIR2 recognize different features of CD86 to downregulate surface expression. *Journal of Biological Chemistry*. Elsevier; 2021; doi: 10.1016/j.jbc.2021.100900.

90. Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, et al.. A framework for exhaustively mapping functional missense variants. *Mol Syst Biol*. 2017; doi: 10.15252/msb.20177908.

91. Weile J, Kishore N, Sun S, Maaieh R, Verby M, Li R, et al.. Shifting landscapes of human MTHFR missense-variant effects. *The American Journal of Human Genetics*. Elsevier; 2021; doi: 10.1016/j.ajhg.2021.05.009.

92. Wrenbeck EE, Bedewitz MA, Klesmith JR, Noshin S, Barry CS, Whitehead TA. An Automated Data-Driven Pipeline for Improving Heterologous Enzyme Expression. *ACS Synth Biol*. American Chemical Society; 2019; doi: 10.1021/acssynbio.8b00486.

93. Zhang L, Sarangi V, Moon I, Yu J, Liu D, Devarajan S, et al.. CYP2C9 and CYP2C19: Deep Mutational Scanning and Functional Characterization of Genomic Missense Variants. *Clinical and Translational Science*. 2020; doi: https://doi.org/10.1111/cts.12758.

94. Zinkus-Boltz J, DeValk C, Dickinson BC. A Phage-Assisted Continuous Selection Approach for Deep Mutational Scanning of Protein–Protein Interactions. *ACS Chem Biol*. American Chemical Society; 2019; doi: 10.1021/acschembio.9b00669.

95. Bernier-Villamor V, Sampson DA, Matunis MJ, Lima CD. Structural Basis for E2-Mediated SUMO Conjugation Revealed by a Complex between Ubiquitin-Conjugating Enzyme Ubc9 and RanGAP. *Cell*. 108:122002;

96. Blanpain C, Doranz BJ, Vakili J, Rucker J, Govaerts C, Baik SSW, et al.. Multiple Charged and Aromatic Residues in CCR5 Amino-terminal Domain Are Involved in High Affinity Binding of Both Chemokines and HIV-1 Env Protein. *J Biol Chem*. 1999; doi: 10.1074/jbc.274.49.34719.

97. Brzovic PS, Keeffe JR, Nishikawa H, Miyamoto K, Fox D, Fukuda M, et al.. Binding and recognition in the assembly of an active BRCA1/BARD1 ubiquitin-ligase complex. *Proceedings of the National Academy of Sciences*. 2003; doi: 10.1073/pnas.0836054100.

98. Chen S, Wu J, Zhong S, Li Y, Zhang P, Ma J, et al.. iASPP mediates p53 selectivity through a modular mechanism fine-tuning DNA recognition. *Proc Natl Acad Sci USA*. 2019; doi: 10.1073/pnas.1909393116.

99. Chupreta S, Holmstrom S, Subramanian L, Iñiguez-Lluhí JA. A Small Conserved Surface in SUMO Is the Critical Structural Determinant of Its Transcriptional Inhibitory Properties. *MCB*. 2005; doi: 10.1128/MCB.25.10.4272-4282.2005.

100. Cobb JA, Roberts DM. Structural Requirements for N-Trimethylation of Lysine 115 of Calmodulin. *Journal of Biological Chemistry*. 2000; doi: 10.1074/jbc.M002332200.

101. Coyne RS, McDonald HB, Edgemon K, Brody LC. Functional Characterization of BRCA1 Sequence Variants using a Yeast Small Colony Phenotype Assay. *Cancer Biology & Therapy*. 2004; doi: 10.4161/cbt.3.5.809.

102. Denker K, Orlik F, Schiffler B, Benz R. Site-directed Mutagenesis of the Greasy Slide Aromatic Residues Within the LamB (Maltoporin) Channel of Escherichia coli: Effect on Ion and Maltopentaose Transport. *Journal of Molecular Biology*. 2005; doi: 10.1016/j.jmb.2005.07.025.

103. Dragic T, Trkola A, Lin SW, Nagashima KA, Kajumo F, Zhao L, et al.. Amino-Terminal Substitutions in the CCR5 Coreceptor Impair gp120 Binding and Human Immunodeficiency Virus Type 1 Entry. *J Virol*. 1998; doi: 10.1128/JVI.72.1.279-285.1998.

104. Dragic T, Trkola A, Thompson DAD, Cormier EG, Kajumo FA, Maxwell E, et al.. A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembrane helices of CCR5. *Proceedings of the National Academy of Sciences*. 2000; doi: 10.1073/pnas.090576697.

105. Ecsédi P, Gógl G, Hóf H, Kiss B, Harmat V, Nyitray L. Structure Determination of the Transactivation Domain of p53 in Complex with S100A4 Using Annexin A2 as a Crystallization Chaperone. *Structure*. 2020; doi: 10.1016/j.str.2020.05.001.

106. Kopecká J, Krijt J, Raková K, Kožich V. Restoring assembly and activity of cystathionine β-synthase mutants by ligands and chemical chaperones. *Journal of Inherited Metabolic Disease*. 2011; doi: 10.1007/s10545-010-9087-5.

107. Kožich V, Sokolová J, Klatovská V, Krijt J, Janošík M, Jelínek K, et al.. Cystathionine β-synthase mutations: effect of mutation topology on folding and activity. *Hum Mutat*. 2010; doi: 10.1002/humu.21273.

108. Kruger W d., Wang L, Jhee K h., Singh R h., Elsas II L j.. Cystathionine β-synthase deficiency in Georgia (USA): Correlation of clinical and biochemical phenotype with genotype. *Human Mutation*. 2003; doi: 10.1002/humu.10290.

109. Lee SY, Pullen L, Virgil DJ, Castañeda CA, Abeykoon D, Bolon DNA, et al.. Alanine Scan of Core Positions in Ubiquitin Reveals Links between Dynamics, Stability, and Function. *Journal of Molecular Biology*. 2014; doi: 10.1016/j.jmb.2013.10.042.

110. Li W, Zhang C, Sui J, Kuhn JH, Moore MJ, Luo S, et al.. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J*. 2005; doi: 10.1038/sj.emboj.7600640.

111. Lin G, Baribaud F, Romano J, Doms RW, Hoxie JA. Identification of gp120 Binding Sites on CXCR4 by Using CD4-Independent Human Immunodeficiency Virus Type 2 Env Proteins. *JVI*. 2003; doi: 10.1128/JVI.77.2.931-942.2003.

112. Mascle XH, Lussier-Price M, Cappadocia L, Estephan P, Raiola L, Omichinski JG, et al.. Identification of a Non-covalent Ternary Complex Formed by PIAS1, SUMO1, and UBC9 Proteins Involved in Transcriptional Regulation. *Journal of Biological Chemistry*. 2013; doi: 10.1074/jbc.M113.486845.

113. Matthews EE, Thévenin D, Rogers JM, Gotow L, Lira PD, Reiter LA, et al.. Thrombopoietin receptor activation: transmembrane helix dimerization, rotation, and allosteric modulation. *The FASEB Journal*. 2011; doi: https://doi.org/10.1096/fj.10-178673.

114. Mayfield JA, Davies MW, Dimster-Denk D, Pleskac N, McCarthy S, Boydston EA, et al.. Surrogate Genetics and Metabolic Profiling for Characterization of Human Disease Alleles. *Genetics*. 2012; doi: 10.1534/genetics.111.137471.

115. Navenot J-M, Wang Z, Trent JO, Murray JL, Hu Q, DeLeeuw L, et al.. Molecular anatomy of CCR5 engagement by physiologic and viral chemokines and HIV-1 envelope glycoproteins: differences in primary structural requirements for RANTES, MIP-1α, and vMIP-II binding11Edited by P. E. Wright. *Journal of Molecular Biology*. 2001; doi: 10.1006/jmbi.2001.5086.

116. Peng L, Damschroder MM, Cook KE, Wu H, Dall'Acqua WF. Molecular basis for the antagonistic activity of an anti-CXCR4 antibody. *mAbs*. 2016; doi: 10.1080/19420862.2015.1113359.

117. Peterson BR, Sun LJ, Verdine GL. A critical arginine residue mediates cooperativity in the contact interface between transcription factors NFAT and AP-1. *Proceedings of the National Academy of Sciences*. 1996; doi: 10.1073/pnas.93.24.13671.

118. Rabut GEE, Konner JA, Kajumo F, Moore JP, Dragic T. Alanine Substitutions of Polar and Nonpolar Residues in the Amino-Terminal Domain of CCR5 Differently Impair Entry of Macrophage- and Dualtropic Isolates of Human Immunodeficiency Virus Type 1. *J Virol*. 1998; doi: 10.1128/JVI.72.4.3464-3468.1998.

119. Ransburgh DJR, Chiba N, Ishioka C, Toland AE, Parvin JD. Identification of Breast Tumor Mutations in *BRCA1* That Abolish Its Function in Homologous DNA Recombination. *Cancer Res*. 2010; doi: 10.1158/0008-5472.CAN-09-2850.

120. Tan Y, Tong P, Wang J, Zhao L, Li J, Yu Y, et al.. The Membrane-Proximal Region of C–C Chemokine Receptor Type 5 Participates in the Infection of HIV-1. *Front Immunol*. 2017; doi: 10.3389/fimmu.2017.00478.

121. Towler WI, Zhang J, Ransburgh DJR, Toland AE, Ishioka C, Chiba N, et al.. Analysis of BRCA1 Variants in Double-Strand Break Repair by Homologous Recombination and Single-Strand Annealing. *Human Mutation*. 2013; doi: 10.1002/humu.22251.

122. Trent JO, Wang Z, Murray JL, Shao W, Tamamura H, Fujii N, et al.. Lipid Bilayer Simulations of CXCR4 with Inverse Agonists and Weak Partial Agonists. *J Biol Chem*. 2003; doi: 10.1074/jbc.M307850200.

123. Van Gelder P, Dumas F, Bartoldus I, Saint N, Prilipov A, Winterhalter M, et al.. Sugar Transport through Maltoporin of *Escherichia coli* : Role of the Greasy Slide. *J Bacteriol*. 2002; doi: 10.1128/JB.184.11.2994-2999.2002.

124. VanBerkum MF, Means AR. Three amino acid substitutions in domain I of calmodulin prevent the activation of chicken smooth muscle myosin light chain kinase. *J Biol Chem*. American Society for Biochemistry and Molecular Biology; 266:21488–951991;

125. Wei Q, Wang L, Wang Q, Kruger WD, Dunbrack RL. Testing computational prediction of missense mutation phenotypes: Functional characterization of 204 mutations of human cystathionine beta synthase. *Proteins: Structure, Function, and Bioinformatics*. 2010; doi: 10.1002/prot.22722.

126. Williams AD, Shivaprasad S, Wetzel R. Alanine Scanning Mutagenesis of Aβ(1-40) Amyloid Fibril Stability. *Journal of Molecular Biology*. 2006; doi: 10.1016/j.jmb.2006.01.041.

127. Zhang J, Rao E, Dioszegi M, Kondru R, DeRosier A, Chan E, et al.. The Second Extracellular Loop of CCR5 Contains the Dominant Epitopes for Highly Potent Anti-Human Immunodeficiency Virus Monoclonal Antibodies. *AAC*. 2007; doi: 10.1128/AAC.01302-06.

128. Nelsen RB. An introduction to copulas. 2nd ed. New York: Springer;

129. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al.. Scikit-learn: Machine Learning in Python. *Journal of machine Learning research*. :2825–30 2011;

130. González J, Dai Z, Hennig P, Lawrence ND. Batch Bayesian Optimization via Local Penalization. arXiv;

Supplementary material

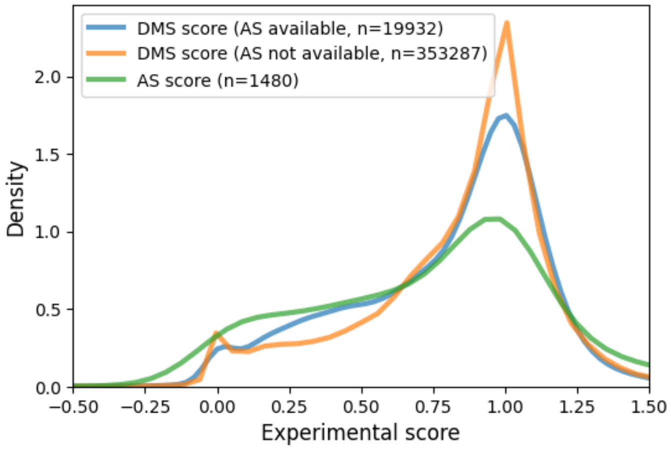

Fig S1. DMS and AS score distribution. The figure shows the kernel estimated density of normalized AS scores and DMS scores for variants with or without available AS data.

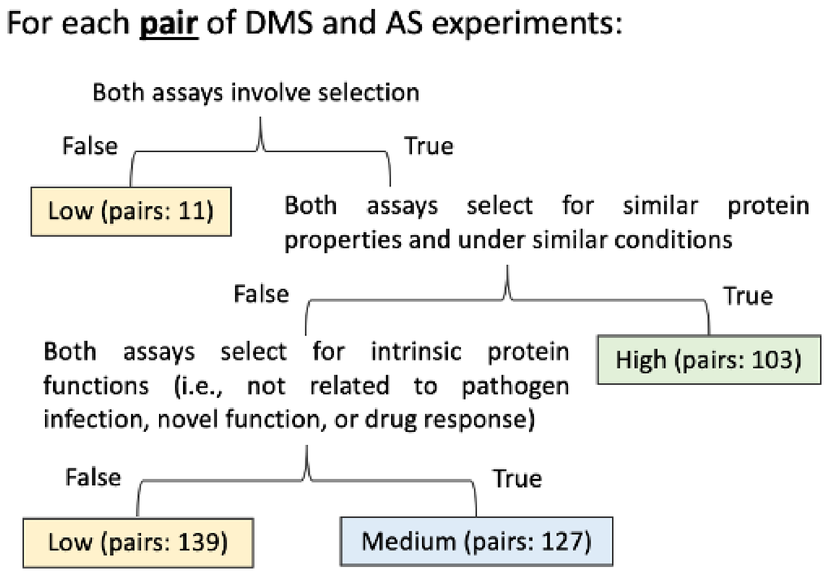
For each pair of DMS and AS experiments:
Both assays involve selection
False
True
Low (pairs: 11)
Both assays select for similar protein properties and under similar conditions
False
True
High (pairs: 103)
Both assays select for intrinsic protein functions (i.e., not related to pathogen infection, novel function, or drug response)
False
True
Low (pairs: 139)
Medium (pairs: 127)

Fig S2. Decision tree for classifying the DMS and AS assay compatibility. The end-nodes show the classified assay compatibility. The number indicates the count of assay pairs for each compatibility level (low, medium, high).

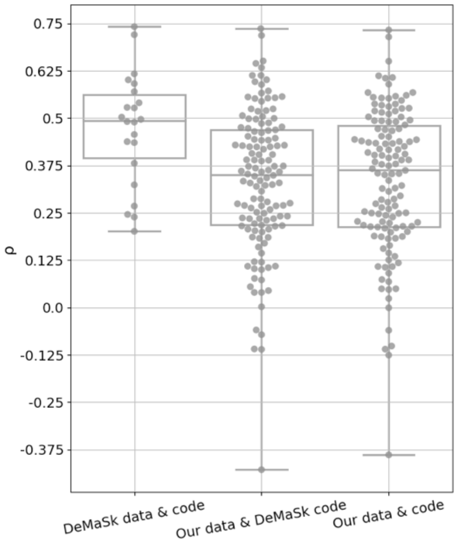

Fig S3. Comparison between published and re-implemented predictors. The plot shows leave-one-protein-out cross-validation performance on predictors built from the published DeMaSk code or our code. The predictors were trained and evaluated on DMS data either provided by the DeMaSk study or curated by our own. The "DeMaSk data & code" result is similar to the published result. For the "Our data & DeMaSk code" result, we used our own data and published code which shows a median performance around 0.35. This is probably because

many more DMS results are included in our data. The similarity of results achieved using “Our data & code” demonstrates the correctness of our re-implementation.

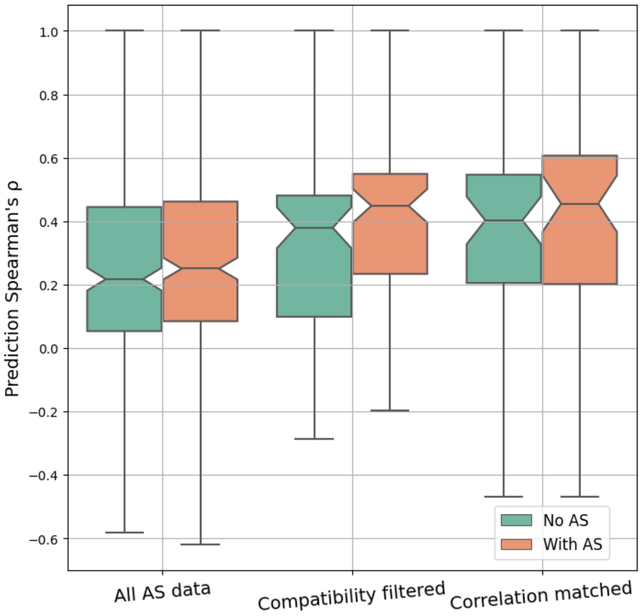

Fig S4. Performance comparison between predictors using AS data or not. The Spearman’s ρ between experiment DMS scores and predicted scores for each DMS and AS data pair are shown as box plots. Different approaches to filtering/matching the data are shown on the x-axis: “All AS data” used all available data; “Compatibility filtered” used only data of high assay compatibility; “Correlation matched” used only data with the highest regularised correlation for each DMS dataset. The figure does not include data without available (filtered/matched) AS scores. This means that the different results are not directly comparable since they are visualized on different subsets of DMS/AS data pairs (for example, “All AS data” contains all DMS/AS data pairs, but “Compatibility filtered” contains only data pairs of high assay compatibility). Control results are shown as green boxes for predicting without AS data as a feature. The underlying ρ for each data pair in the control results is the same, but the boxes are shifted due to data filtering/matching.

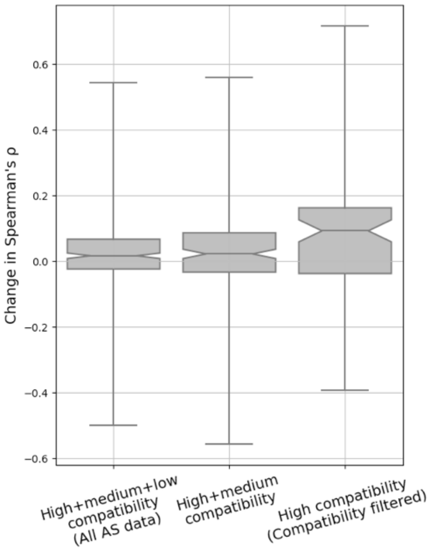

Fig S5. The performance of variant impact prediction for using data of different assay compatibility levels. The change of prediction Spearman's ρ for each DMS and AS data pair is shown as box plots. A higher value represents higher prediction accuracy achieved for using AS data. Different data filtering methods are shown on the x-axis.

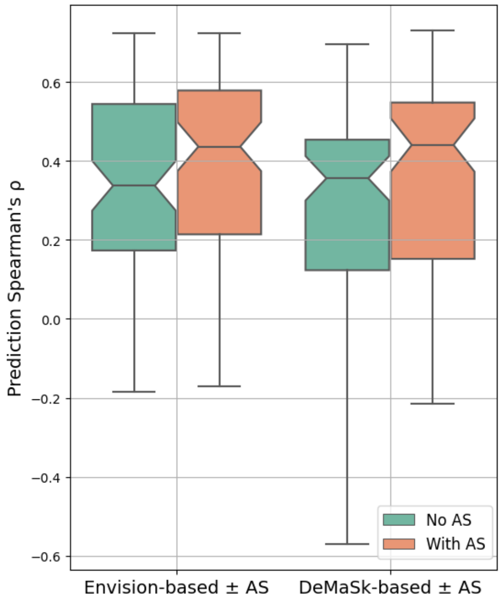

Fig S6. Prediction performance is improved while incorporating high compatibility AS data into the Envision model. The Spearman's ρ between experiment DMS scores and predicted scores for each high compatible DMS/AS assay pair are shown as box plots. The x-axis shows the predictor used, either Envision or DeMaSk. Control results are shown as green boxes for predicting without AS data as a feature.

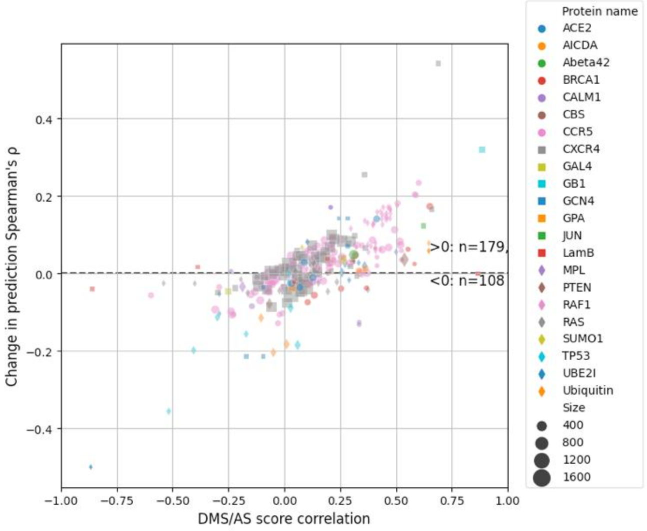

Fig S7. Prediction performance change for using all AS data. Each dot represents a DMS/AS data pair. The vertical axis shows the change of prediction ρ by using AS data (larger means higher performance achieved by using AS data). The horizontal axis shows the DMS/AS score correlation for *all* variants on the matched residues rather than just alanine substitutions. The colours and shapes of the dots correspond to the target protein, and size indicates the number of variants in each data pair.

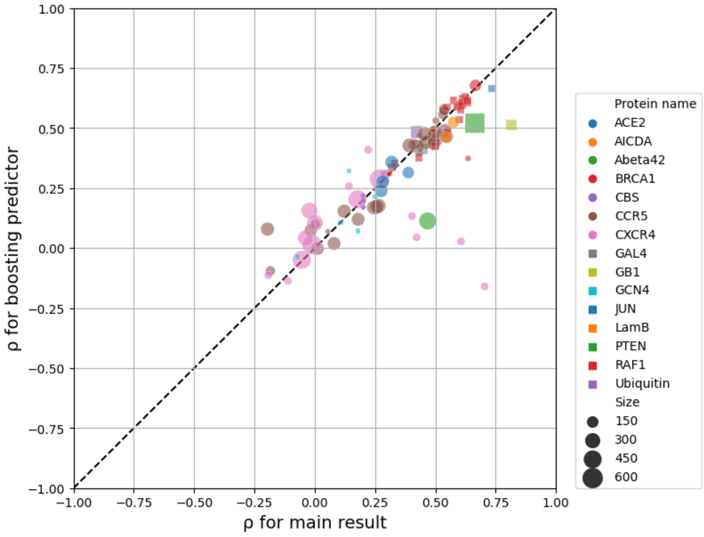

Fig S8. Boosting setup shows similar performance as the main result. Each dot represents a filtered DMS/AS data pair of high assay compatibility. The vertical and horizontal axes show the prediction Spearman's ρ for either modelled with boosting or the one-step (main result) setup. The colours and shapes of the dots correspond to the target protein, and size indicates the number of variants in each data pair.

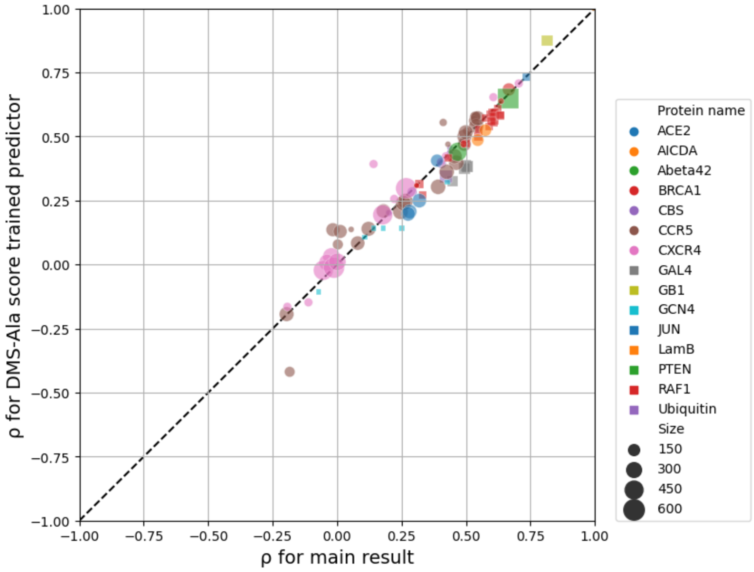

Fig S9. Training with DMS scores of alanine substitutions shows similar performance as the main result. The vertical and horizontal axes show the prediction Spearman's ρ for predictors either trained with DMS score of alanine substitutions (DMS-Ala) or AS data of high assay compatibility (main result), yet all evaluated on high compatibility AS data. The colours and shapes of the dots correspond to the target protein, and size indicates the number of variants in each data pair.

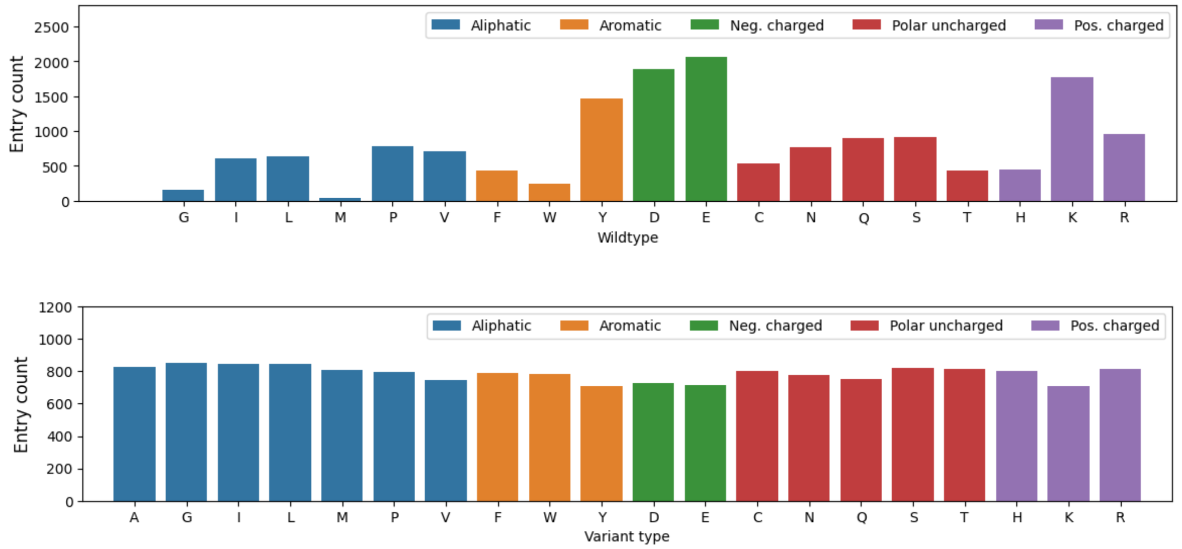

Fig S10. Count of variant entries for each wild-type or variant amino acid of high assay compatibility data. (Neg.: negatively, Pos.: positively)

Table S1. Amount of data with AS scores available

Data composition	Protein	DMS dataset	AS dataset[1]	Variant entries[2]
All AS	22	54	146	70446
Compatibility filtered	15	35	60	15739
High+medium assay compatibility	21	51	105	28380
Correlation matched	22	54	32	7940

1. This column shows how many unique AS datasets are included.
2. Include duplicated variants caused by multiple experiments targeting the same protein variant.

Supplementary information

Applying AS data to Envision method

We re-implemented a predictor based on Envision [15] to incorporate AS data. Features used in Envision were downloaded from its online toolkit. All Envision features are used for modelling except for substitution type (wt_mut) which has low importance according to the published result and our pilot studies yet is computationally expensive in our setup. Protein data were excluded if their features were not available online. DMS and AS data pairs with high assay compatibility were used for modelling. Missing feature values were imputed by the mean values for numerical features or the most frequent values for categorical features. Categorical features are encoded with the one-hot encoder. We used `sklearn.ensemble.GradientBoostingRegressor` from scikit-learn package [129] to build the predictor, and hyperparameters were tuned by Bayesian Optimization [130] with Group K-Fold (protein-30-fold) cross-validation. The training and evaluation process were similar to that previously described. For comparison, we repeated the DeMaSk-based analysis on the same subset of data.

Boosting with AS data

To deal with the sparsity of AS data, we tested a variant impact predictor based on boosting. A first linear regression predictor was trained with all training DMS data using the three DeMaSk features without AS data, which was the same as the control predictor mentioned previously. We then calculated the prediction error by subtracting the predicted scores from DMS scores, and a second linear regression predictor was trained to predict the error. The second predictor was trained only on DMS/AS data of high assay compatibility and used both protein features and the encoded AS scores. The final prediction result was the sum of the outputs from these two predictors.

Replacing AS data with DMS scores of alanine substitutions

We investigated another potential approach to overcome the sparsity of AS data by replacing the AS feature with the DMS scores of alanine substitutions (DMS-Ala). For all DMS datasets we collected, their AS feature values, regardless of availability, were replaced by the DMS-Ala scores on the same residue. Missing scores were imputed by the mean value of all DMS-Ala scores. A regression model was trained and evaluated as previously described, using the three DeMaSk features as well as the DMS-Ala scores. The AS data of high assay compatibility are still used for the testing process.

Supplementary Table 1

Supplementary Table 2

Supplementary Table 3

WEHI
brighter together

The Walter and Eliza Hall Institute of Medical Research
ABN 12 004 251 423
1G Royal Parade Parkville Victoria 3052 Australia
T +61 3 9345 2555 F +61 3 9347 0852
www.wehi.edu.au

Dr Scott Edmunds
Editor-in-Chief
Gigascience

Dear Dr Edmunds,

Please find our enclosed manuscript entitled "**Integrating deep mutational scanning and low-throughput mutagenesis data to predict the impact of amino acid variants**" for your consideration for publication in *Gigascience*.

The key contributions of our work are:

- We developed the first predictor of protein variant impact integrating high-throughput and low-throughput mutagenesis data, in our case, deep mutational scanning and alanine scan data.
- We demonstrate that integrative variant impact predictors improve model performance only when the high and low throughput data are generated by related assay types.

In this work, we collected high-throughput deep mutational scanning (DMS) data from an online database, with 370,000 protein variants and low throughput alanine scanning data of matched proteins from published papers. We defined a decision tree to classify low- and high-throughput assays to distinct levels of similarity across multiple categories, which we call assay compatibility. We then explored models of variant impact trained on these data. Our results showed the connection between experiment assay compatibility and the predictor's performance built from these data.

This is an original research article, and we have no conflicts of interest to disclose. All authors have participated in the preparation of this manuscript and approved the submission of it. We confirm that this work has not been published nor is currently under consideration for publication elsewhere.

Thank you for your consideration of this manuscript.

Yours sincerely,

Yunfan Fu

Mr Yunfan Fu

A.P

Prof Anthony T. Papenfuss

Dr Alan F. Rubin