Fakultät für Informatik

Assessment of variant effect prediction

Jonas Reeb

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

Vorsitzender: Prof. Dr. Julien Gagneur

Prüfende der Dissertation:

1. Prof. Dr. Burkhard Rost

2. Prof. Dr. Dominik Grimm

Die Dissertation wurde am 26.06.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 02.10.2020 angenommen.

# Zusammenfassung

Diese Arbeit befasst sich mit den Problemen und Herausforderungen für Machine-learning-basierte Methoden, die den Effekt von Proteinsequenzvarianten vorhersagen. Gegenwärtig ist es üblich, entweder alte, etablierte Methoden anzuwenden oder neue Ansätze zu entwickeln, obwohl ein tieferes, allgemeines Verständnis für diese Vorhersagemethoden fehlt.

Um bessere Kenntnisse über den Stand der Technik zu erlangen, haben wir neue Datensätze erstellt und damit unabhängige Evaluierungen auf bisher nicht genutzten, experimentellen Daten durchgeführt. Die erzeugten Datensätze bestehen aus Krankheitsvarianten in Tieren sowie High-Throughput-Messungen der Proteinfunktion mit Deep-Mutational-Scanning-Experimenten.

Die Analysen zeigen, dass Vorhersagemethoden dazu neigen, primär Varianten mit starkem, schädlichem Effekt zu detektieren. Diese weisen oft eine hohe Sequenzkonservierung auf. Des Weiteren werden Varianten mit positivem Effekt generell vernachlässigt, was sich in schlechten Vorhersagen derselben widerspiegelt. Allgemein lässt sich festhalten, dass keine einzelne Methode unter allen Umständen die beste Leistung zeigt. Zukünftige Anstrengungen werden daher nötig sein, um die Defizite zu beseitigen. Dies wird es ermöglichen, die Methoden auch für klinische Anwendungen einzubeziehen und damit die umfassende Einführung von Präzisionsmedizin zu unterstützen. Diesbezüglich liegt der Beitrag dieser Arbeit im Bereitstellen des nötigen Wissens, um sowohl die umsichtige Anwendung von aktuellen Methoden als auch die gezielte Neuentwicklung von Vorhersagemethoden voranzutreiben.

# Abstract

This thesis discusses the issues with and challenges faced by machine learning-based methods which predict the effect of protein sequence variants. Current practice favors the usage of either long-established methods or development of novel approaches, despite lacking a deeper understanding of effect predictors in general.

To gain insights into the state of the art, we assembled new datasets and performed independent evaluations on previously untapped experimental data. The generated datasets consist of disease variants in animals, and high-throughput measurements of protein function by deep mutational scanning assays.

The analyses illustrate that prediction methods are heavily biased towards deleterious, high effect variants which exhibit strong sequence conservation. We further show that variants with beneficial effect are generally neglected, resulting in poor performance of their prediction. Overall, no single prediction method performs best under all circumstances. Thus, future endeavors are necessary to remedy these deficiencies. Accounting for them will allow those methods to support clinical action and drive the large-scale adoption of precision medicine. In this regard, the significance of this study lies in providing the knowledge required for both the sensible application of current prediction methods as well as the targeted development of new predictors.

*To Schnuggi* ♡

# Acknowledgements

First of all, my thanks goes to Burkhard, for inviting me with open arms into his lab, for nurturing, teaching and encouraging me throughout my Bachelor's and Master's program, and eventually supervising my PhD thesis. Thank you for all the trips to international conferences, for opportunities created and encouragement provided. The more years go by, the more I realize the luck I had in happening upon a supervisor as understanding and supporting.

Special thanks also to Edda Kloppmann for supervising my Bachelor's and Master's thesis, and thus introducing me to science. I could not have imagined a better person at my side during those formative years and still rely on the foundations built then.

Thank you to Prof. Dominik Grimm and Prof. Julien Gagneur for dedication their valuable time as members of the thesis committee.

Thank you as well to all other Rostlab members for creating a pleasantly stimulating research environment and supporting me on this path. In particular thanks to Inga and Lothar for dealing and helping with bureaucracy, and to Tim for all hard- and software support, nurturing an increasing coffee obsession and being an all-around nice guy to talk to. Finally, thanks to Michael for sharing an office with me for many years—few people are as genuinely kind.

I am grateful to my parents that most had to deal with my idiosyncrasies, yet have always supported me in all my decisions. I increasingly realize the value in your enduring backing that cannot be taken for granted. Furthermore, Sophie, for broadening my horizon and living an example of boldness and determination.

Thanks to all my friends that supported me along the way. Thanks in particular to Jochen without whom I might have never set foot into the mountains that remained unnoticed at my doorstep for so many years. I can hardly imagine a life without them nowadays. In the same vein, thank you Tim. In you I found unwavering friendship and the partner to define myself in the outdoors. We succeed and failed together, and grew stronger out of it. Nothing could have been a better distraction from the occasional setbacks of research.

Finally, Diana. I will not imagine the things we could have experienced had fate brought us together earlier, yet I relish every day since and look forward to decades of adventures together. So much of who I am today is owed to meeting you. Without you, I might not be writing these lines. Thank you for everything.

# Contents

# List of Figures



# List of Tables

# Abbreviations

3D three-dimensional, page 12

CAGI Critical Assessment of Genome Interpretation, page 25

CNV Copy number variation, page 5

DMS Deep mutational scanning, page 10

HTS High-throughput sequencing, page 2

MSA Multiple sequence alignment, page 15

SAV Single amino acid variant, page 5

SIFT Sorting intolerant from tolerant, page 15

SNAP Screening for non-acceptable polymorphisms, page 17

SNV Single nucleotide variant, page 5

VEP Variant effect prediction method, page 14

WES whole-exome sequencing, page 3

WGS Whole-genome sequencing, page 2

wt Wild-type, page 11

# Publications

This work constitutes a cumulative dissertation based on the following peer-reviewed publications which are summarized with individual contributions in Appendix B:

> **Reeb, J.**, Hecht, M., Mahlich, Y., Bromberg, Y., and Rost, B. (2016). Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLOS Computational Biology*, 12(8), e1005047
>
> **Reeb, J.**, Wirth, T., and Rost, B. (2020). Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics*, 21(1), 107

While working on the dissertation I (co)-authored the following publications relevant to the topic:

> Mahlich, Y., **Reeb, J.**, Hecht, M., Schelling, M., De Beer, T. A. P., Bromberg, Y., and Rost, B. (2017). Common sequence variants affect molecular function more than rare variants? *Scientific Reports*, 7(1), 1608
>
> **Reeb, J.**, Goldberg, T., Ofran, Y., Rost, B. (2020). Predictive Methods Using Protein Sequences. In *Bioinformatics* (4th edition, pp. 185-225). Wiley

I further worked on the following publications and other literature:

> Bernhofer, M., Kloppmann, E., **Reeb, J.**, and Rost, B. (2016). TMSEG: novel prediction of transmembrane helices. *Proteins*, 84(11), 1706–1716
>
> **Reeb, J.** and Rost, B. (2019). Secondary Structure Prediction. *Encyclopedia of Bioinformatics and Computational Biology*, 2, 488–496
>
> Kloppmann, E., **Reeb, J.**, Hoenigschmid, P., and Rost, B. (2019). Protein Secondary Structure Prediction in 2018. In *Encyclopedia of Biophysics*. European Biophysical Societies' Association
>
> Punta, M., Kloppmann, E., and **Reeb, J.** (2019). Membrane Protein Structure. In *Encyclopedia of Biophysics*. European Biophysical Societies' Association

Zhou, N., Jiang, Y., [...] **Reeb, J.**, [...] Mooney, S. D., Greene, C. S., Radivojac, P., Friedberg, I. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1), 244

*It's always further than it looks*
*It's always taller than it looks*
*And it's always harder than it looks*

Chapter 1

# Introduction

## 1.1 Genetic variation

Genomes are the manuals of life which contain the information required for the development of an organism. However, not every organism of the same species shares the exact same genome. For humans in particular, genetic variation is what defines us as individuals. It determines the way we look, our behavior, preferences and predispositions. As such, understanding whether a genetic change will have an effect and what it will be is inherently interesting. Both to us as individuals as well as to every industry that aims to offer a product tailored to their customers, such as personalized medicine.

In 2003, the human genome project concluded by presenting "the" human genome after working over a decade on its determination. In truth, it was a mosaic of genetic material from a number of European individuals combined (Collins *et al.*, 2003b). Nevertheless, having the genome sequence did not immediately elucidate the inner workings of our complex organism. Despite knowing the alphabet, grammar and now also the full text of this "book of life", interpreting it still provides a challenge to this day. Researchers alike were, in retrospect, overly optimistic about the opportunities the human genome would offer, in particular regarding the outlook of personalized medicine (Collins *et al.*, 2003a; Shendure *et al.*, 2019). Even 15 years later the promise of drugs developed to every individual's needs is far from being a reality. These days, precision medicine seems more likely, i.e., treatments tailored towards groups of individuals for example with the same genetic makeup (Ashley, 2016; Rost *et al.*, 2016; Morganti *et al.*, 2020; Claussnitzer *et al.*, 2020). However, this is far from trivial as 54 % of protein-coding variants were unique to just one in 60,000 humans (Lek *et al.*, 2016). Furthermore, with around 1,150,000 protein-coding positions in the human genome and 19 possible changes for each, more than 200 million changes are possible. Yet, the outcome is known for only around 85,000 , i.e., less than 0.05 percent (The Uniprot Consortium, 2019; Landrum *et al.*, 2016). Therefore,

one of the crucial impediments for attaining precision medicine is the interpretation of genetic variation (Daneshjou *et al.*, 2017; J. Shendure and J. M. Akey, 2015; Manolio *et al.*, 2017).

To assist in closing the divide between known sequence variants and their outcome, dedicated variant effect prediction methods (VEPs) began to be developed concurrently with the human genome project's conclusion (Ng and Steven, 2001; Ramensky *et al.*, 2002). Their goal is to predict the effect of a sequence variant *in silico* while using only information about sequences and structures from public databases as input and without the need for time-consuming and costly wet-lab experiments. As the speed at which genome sequences and with that genetic variation can be determined rapidly increases, VEPs are struggling to keep up. In the remainder of this chapter, more recent advances in sequencing, the types and effects of genetic variations, as well as popular VEPs and their issues will be discussed.

### 1.1.1 High-throughput sequencing

The human genome project still relied on Sanger Sequencing which remains a viable method for small-scale projects. For larger efforts such as whole-genome sequencing (WGS), it has been replaced by a set of high-throughput sequencing (HTS) methods, also referred to as next-generation sequencing (Goodwin *et al.*, 2016; Levy and Myers, 2016; van Dijk *et al.*, 2018). Common to all of them is an increase in sequencing speed at reduced cost (Figure 1.1). In the context of HTS, a read refers to a stretch of nucleotide sequence that can be detected within one operation from a single molecule. Depending on the particular HTS method, the maximum length of sequencing reads might be shorter, or the error rate higher than for Sanger sequencing (Pfeiffer *et al.*, 2018; Bowden *et al.*, 2019). However, some of the initial shortcomings have been resolved algorithmically or by improvements to the methods themselves. Statistical underpinnings of interpreting the data have also matured to be more robust (Carss *et al.*, 2019).

HTS has enabled a multitude of new experimental venues or spurred research in existing fields. For example, in cancer research, increasing sequencing capabilities have led to the discovery of mutational signatures specific to certain cancer types which are important for clinical action (Alexandrov and Stratton, 2014; Aravanis *et al.*, 2017; Cieslik and Chinnaiyan, 2020). HTS also enabled single-cell sequencing which determines the genetic makeup of just one individual cell and thus elucidates the heterogeneity of cells in, e.g., developmental processes or within healthy as well as cancer tissue (Kolodziejczyk *et al.*, 2015; Gawad *et al.*, 2016). Deep mutational scanning (DMS) describes a framework for evaluating the functional outcome of every possible variant in a protein of interest

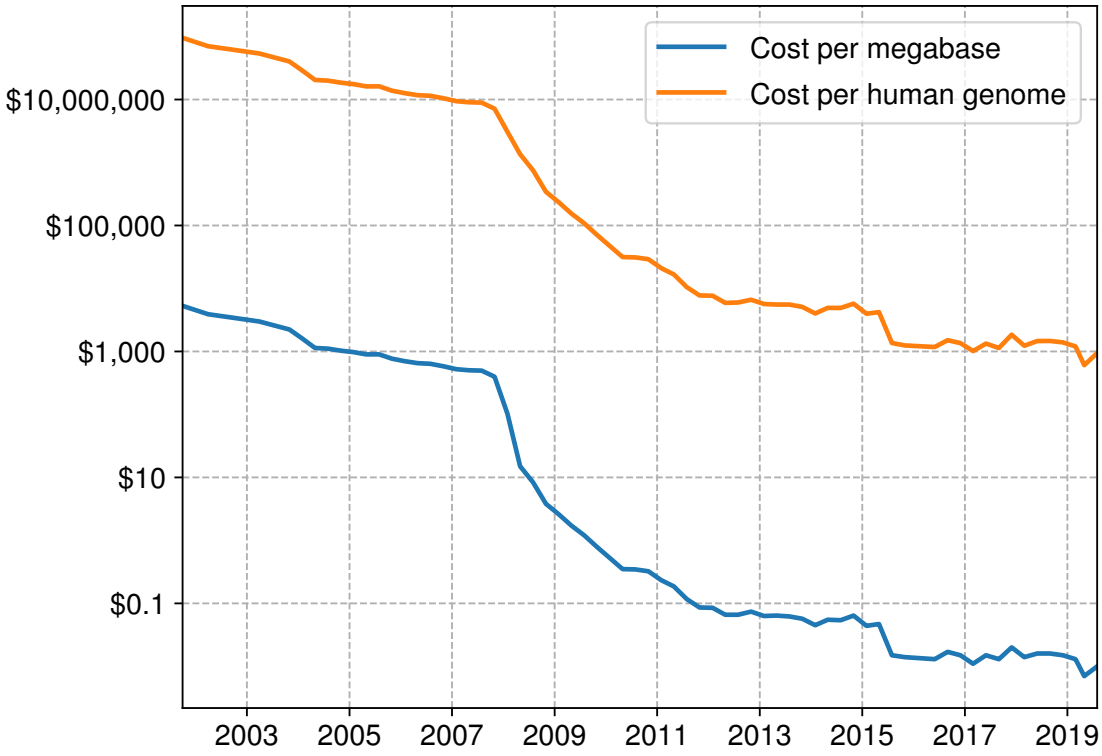



**Figure 1.1. Decline in sequencing costs.** Since the inception of high-throughput sequencing technologies around 2008, the associated costs have decreased exponentially. Denoted amounts include sequencing machines and materials but not associated expenditures such as quality control, management, development and processing of raw sequencing output. Costs are denoted in US$ and the data underlying these plots are provided by the NIH (Wetterstrand, 2019).

and thereby tries to close the gap between sequence data and interpretation (Fowler and Fields, 2014). Perhaps most importantly, whole-exome sequencing (WES), which is limited to protein-coding sequences, as well as WGS have benefited greatly from HTS. This lead to increasingly larger efforts both to determine additional reference genomes as well as to study the genetic variation within (Martinez and Nelson, 2010). Naturally, the study of the human genome landscape has been a particular focal point. For example, just five years after the determination of the human genome, the 1000 genomes project started. At its conclusion in 2015 stood 2,504 human genomes from populations all over the world (The 1000 Genomes Project Consortium, 2010, 2012, 2015). In parallel, another large effort studied the human microbiome and its effects on health (The Human Microbiome Project Consortium, 2012; The Integrative HMP (iHMP) Research Network Consortium, 2014). More recently an initiative limited to the United Kingdom successfully sequenced 100,000 human genomes (Genomics England, 2017) and several other countries are working on similar projects.

Human WES and WGS is now being offered by various companies around the often cited $1,000 threshold and costs are certainly below $10,000 even for single laboratories (Schwarze *et al.*, 2018, 2019). This sparks ethical concerns as well as technological challenges regarding data processing and storage. Moreover, the tremendous amount of data resulting from HTS also highlights the deficits still present in its interpretation (J. Shendure and J. M. Akey, 2015; Manolio *et al.*, 2017).

### 1.1.2 Types of genetic variation

The advances in HTS outlined above form the basis for pervasive detection of genetic variation (Pavlopoulos *et al.*, 2013). However, to determine a change one first has to establish a baseline to deviate from. For a set of common organisms including *Homo sapiens*, the Genome Reference Consortium maintains such a reference (Schneider *et al.*, 2017). With the baseline, currently GRCh38 for human, any deviation from it can be regarded as a manifestation of genetic variation. However, a single reference is necessarily always biased and not an accurate representation of the complete human population. This lead to recent arguments for using a collection of sequences referred to as pan-genome instead (Sherman and Salzberg, 2020). Either way, such differences are not uncommon: Any two unrelated individuals differ in their genomes at four to five million sites, i.e., about 0.1 % of the total 3 billion nucleotides (The 1000 Genomes Project Consortium, 2015). Given the human genome's size, the number of possible changes is in the order of $10^{10}$. These changes can be caused by genetic recombination or DNA repair mechanisms and errors therein, as well as at random. Depending on whether the variants occur in somatic or germline cells, they can be inherited by offspring. Since humans are diploid organisms, we carry two copies of every genetic locus. Each copy is referred to as an allele. The minor allele frequency describes how often the second most common allele is found in a given population. Variants are often considered rare if they have a minor allele frequency of for example less than 1 % or less than 0.1 %, and common otherwise. Since the terms such as mutation or polymorphism carry negative connotations to many readers and are therefore often interpreted wrong, all types of genetic changes will be referred to as variants throughout this thesis (Richards *et al.*, 2015). Epigenetic changes, i.e., changes outside of the genomic sequence are not discussed here. The different types of variation which can occur are outlined in the following two subsections.

#### Structural variants

Alongside many other findings, efforts such as the 1000 genomes project also revealed that the majority of changed nucleotides in the human genome are caused by so-called

structural variants, although they make up only 0.1 % of all variants in the population. It is also noteworthy that current, reference-based HTS approaches as described in Subsection 1.1.1 are less sensitive to many types of structural variation meaning that the real occurrences may be higher (Chaisson *et al.*, 2015). The 2019 gnomAD-SV analysis based on WGS data from almost 15,000 individuals also supports this (Collins *et al.*, 2019).

Typically, structural variants denote large changes that affect at least around 1,000 bases (Feuk *et al.*, 2006). Among them, copy number variations (CNVs), such as insertions, deletions and duplications describe the addition or removal of nucleotide stretches. The length of the duplicated elements can vary from di-nucleotides to whole genes. Furthermore, specific duplicated subsequences show a different number of repeats between individuals. Besides CNVs, inversions describe cases in which part of the sequence is reversed while the overall length remains unaffected. Several of the described variation types can also occur simultaneously leading to more complex changes. Furthermore, these changes can also occur on the level of chromosomes, therefore affecting even larger portions of the genome.

#### Single position variants

Before HTS methods were able to detect the pervasiveness of structural variants, smaller, local changes were thought to be the main source of human genetic variation. Indeed, single nucleotide variants (SNVs) account for around 85 % of the variation every individual carries compared to the human reference genome (The 1000 Genomes Project Consortium, 2015). When including short insertions or deletions, the portion rises to more than 99.9 %. SNVs are often distinguished from single nucleotide polymorphisms, i.e., SNVs that occur in the population with a certain frequency. Within this thesis, no such distinction is made.

Most SNVs are found in non-coding regions of the genome—i.e., outside of genes which are translated into proteins—and are considered rare since they are present in less than 1 % of the population (Lek *et al.*, 2016). SNVs in coding regions can be either synonymous or non-synonymous. Synonymous SNVs, also referred to as silent variants, alter a nucleotide but not the encoded amino acid as a result of the genetic code's redundancy. Non-synonymous SNVs change the encoded amino acid and therefore translate to an altered protein sequence. As a special case, nonsense variants change a tri-nucleotide encoding for an amino acid to a STOP-codon which leads to premature termination of translation.

On the level of proteins, changes of single residues are also referred to as single amino acid variants (SAVs). SAVs are typically denoted in a short-hand such as MET1ILE, mean-

ing that methionine at position 1 is changed to isoleucine. Throughout this work, SAVs will be denoted this way but using the amino acid one letter code, i.e., M1I for the previous example. It is important to note that a SAV is not necessarily the change of just one nucleotide in the gene sequence but could also denote a complete exchange of the underlying codon. To further differentiate SAVs, a change that could be achieved with just one nucleotide alteration is referred to as SNV-possible (Bromberg *et al.*, 2013). However, this definition does not take into account the actual codon found at the position of interest but rather all theoretically possible changes. For example, given a protein with arginine at position 8, the list of SNV-possible variants includes R8W and R8H, although R8W can only be achieved with a single nucleotide change if the underlying codon for arginine is CGG or AGG, while R8H requires one of the codons CGU and CGC. From here on out, the standard genetic code is assumed for the definition of SNV-possible. Opposed to SNV-possible stands the set of 19-non-native variants, i.e., the change of a given amino acid to all other nineteen standard proteinogenic amino acids.

## 1.2 Effect of genetic variation

The outcome of most variants as outlined in Subsection 1.1.2 is unknown, i.e., they are variants of unknown significance. Their effect can range from none at all to the most extreme cases of disease development. For this distinction it is crucial to consider the level at which effects are measured. If a variant does not lead to a disease or any other discernible phenotypic change on the organism level it could be considered to not have an effect, that is, to be neutral or wildtype-like. Nevertheless the variant in question might have an effect on the protein function on a smaller scale, such as reduced enzymatic activity. Knowledge of this effect is still relevant for understanding the changes to molecular pathways and how several co-occurring variants might lead to larger scale effects. Even if the effect is strictly limited to the protein in question, whether one or both copies of the gene are changed, i.e., whether the variant is hetero- or homozygous, can affect its consequences. Finally, variants might indeed not have any measurable effect and just be the manifestation of evolution.

Another factor still highly underrepresented in experimental data is that a variant's effect can be beneficial as well. This is also often referred to as a "gain-of-function". While the typical assumption is that the variant has a deleterious effect, it may be increased as well. An increase in enzymatic activity constitutes a beneficial variant effect on the level of protein function, however this doesn't necessarily imply a positive outcome on the level of the organism. The increased function might throw a pathway out of balance leading

to undesired effects. As another example, increased virulence is a beneficial effect for the virus but not the host, same as increased antibiotic resistance in pathogenic bacteria.

Finally, it is important to realize that many traits do not behave according to straightforward Mendelian inheritance where a single gene is responsible for all phenotypic changes. In so-called complex or quantitative traits, variants at multiple genomic loci all contribute with typically small effect to the overall organismal phenotype (Mackay *et al.*, 2009). In the same way that, e.g., adult human height is influenced by hundreds of known variants at hundreds of loci, diseases can also be attributed to a set of changes which all contribute to factors such as onset or severity (Marouli *et al.*, 2017; Plomin *et al.*, 2009; Manolio *et al.*, 2009). Such polygenic diseases are among the most common human diseases including diabetes, various heart conditions, and cancer types (Khera *et al.*, 2018). Furthermore, much of the phenotypic heritability in humans appears to be driven not by the most common variants but rather extremely rare ones that have been observed only once in the population (Hernandez *et al.*, 2019). At the same time it is hotly debated whether common variants have more functional impact than rare ones (Mahlich *et al.*, 2017; Alhuzimi *et al.*, 2018; Laddach *et al.*, 2019).

In the following, Subsection 1.2.1 will give an overview of the type of effects variants can display. Subsection 1.2.2 highlights different approaches for determining these effects in wet lab experiments, with particular focus on a recent method referred to as deep mutational scanning (Subsection 1.2.3 on page 10). Finally, Subsection 1.2.4 on page 12 outlines various resources that catalog variants for which effects have been determined.

### 1.2.1 Examples of variation effects

A well-known example of structural variation is how often the tri-nucleotide CAG coding for glutamine repeats near the N-terminus of the human protein huntingtin. The number of repeats is highly correlated with the development and onset of Huntington's disease (Myers, 2004; Saudou and Humbert, 2016). Another example is the 32 nucleotide deletion in *CCR5*, commonly termed CCR5-Δ32 which leads to a frameshift and premature termination of the chemokine receptor's translation. While recent research also suggests detrimental effects of this variant, homozygous carriers of CCR5-Δ32 have long been known to be effectively immune to HIV infection (Dean *et al.*, 1996; Barmania and Pepper, 2013). Finally, the duplication of a gene can intuitively affect its expression and prevalence in the cell which might in turn lead to changes in molecule concentrations or pathway activity - an effect referred to as gene dosing. Such CNVs have been implicated with various diseases including Alzheimer's, autism and HIV susceptibility (Feuk *et al.*,

2006; Freeman, 2006). Larger scale structural variants are well known through diseases such as Down syndrome caused by the (partial) duplication of chromosome 21.

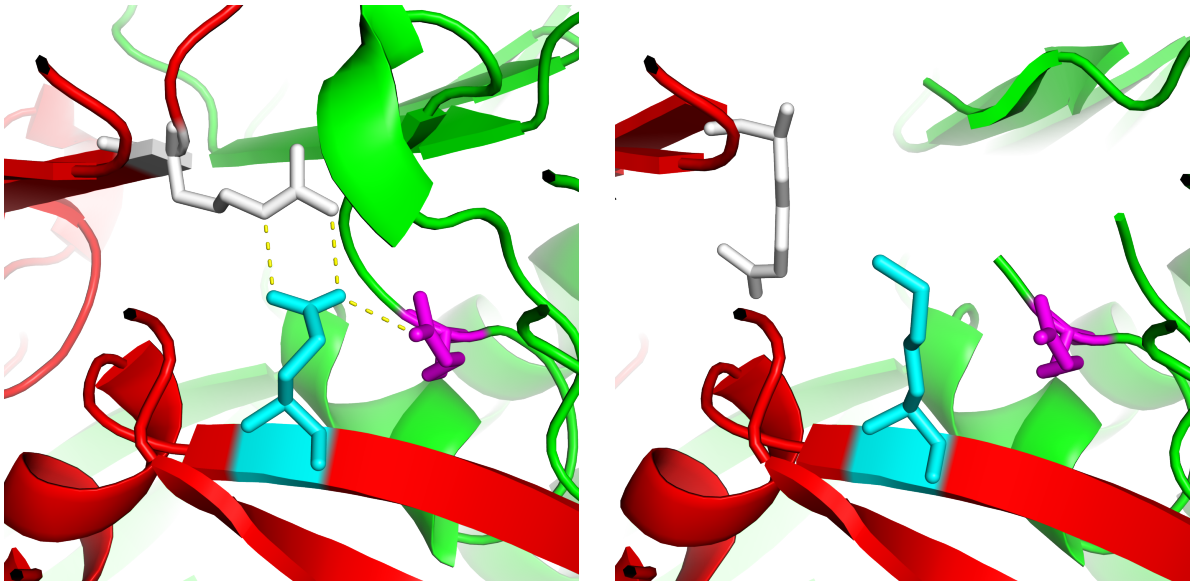
SAVs have traditionally been the focus of research. To give just some examples of their possible effects, OMIM (Subsection 1.2.4, Amberger *et al.*, 2019) currently catalogs some 5,500 phenotypes with a known cause, around 11,000 of which are SAVs (Sherry *et al.*, 2001). Sensitivity to alcohol is one of those phenotypes (OMIM identifier 100650) and common in East Asian populations due to variant E504K in aldehyde dehydrogenase which performs a crucial step in alcohol metabolism (Figure 1.2). Another popular example is variant E7V in hemoglobin which causes sickle-cell anaemia when homozygous, but confers increased resistance to malaria when only one copy of the gene carries the variant. Depending on the context, it can therefore also be seen as an example of a variant with beneficial effect. Similar so-called protective variants have been identified for diseases including type 1 and 2 diabetes, various cancers or inflammatory bowel disease (Butler *et al.*, 2017).

In human, the number of SAVs found in every individual is comparable to the number of synonymous variants - around 10,000 each (The 1000 Genomes Project Consortium, 2015). However, effects are much more studied for the SAVs. Intuitively one might not expect synonymous variants to have an effect on protein function. After all, the resulting amino acid chain remains unaffected. Nonetheless, their common designation as "silent" variants is misleading since they can, for example, influence factors such as translation fidelity by changes to less frequent codons which can in turn affect co-translational protein folding (Sauna and Kimchi-Sarfaty, 2011). Codon changes may further influence regulation mechanisms, post-translational modifications as well as RNA structure and stability. For example, Crouzon syndrome can be caused by the synonymous variants A344A and P361P which affect splice site usage (Fenwick *et al.*, 2014).

### 1.2.2 Experimentally determining variant effect

As outlined above, knowing about the existence of a variant has become increasingly easy while determining its effect largely remains a challenge. In the following, some of the traditional and more recent experimental approaches that tackle this problem will be outlined. Knowing how the effect of a variant was determined is important both for sensible training of prediction methods and clinical action. Since this is the variant type most relevant to the rest of this thesis, the outlined experiments focus on determining the effect of SAVs.

Traditionally, datasets of experimentally verified SAV effect have originated from *in vitro* studies of single variants of interest for example because they were expected to crucially



**(a)** Wild-type **(b)** E504K variant

**Figure 1.2. Effect of a single amino acid variant on enzymatic activity.** Shown is the exemplary effect of a SAV on the function of human aldehyde dehydrogenase. *ALDH2* catalyzes the reaction of acetaldehyde to acetic acid, a crucial step in the degradation of alcohol. The enzyme forms a homotetramer, i.e., four copies of the gene associate as subunits to build the active form of the protein. Visualized here is the interface between two copies colored in red and green, respectively. Both chains are visualized in a cartoon view with α-helices as ribbons and β-sheets as arrows. Specific residues which are described below are further visualized as sticks. **(a)** shows wild-type *ALDH2* where glutamic acid (cyan) at position 504 forms hydrogen bonds (yellow dashed lines) with two arginines at residues 281 (white) and 492 (magenta) in the dimer interface. **(b)** shows the E504K variant (cyan) common in East Asian populations. The SAV is not directly affecting the active site that catalyzes the reaction. However, the missing hydrogen bonds destabilize the overall structure. While the tetramer still folds successfully, introduction of a disordered region indirectly deforms the active site which significantly decreases its activity (Larson *et al.*, 2005). The defective enzyme can have consequences going beyond the immediate discomfort of alcohol sensitivity since drinking despite being a carrier of the variant has been implicated with increased risk for esophageal cancer (Brooks *et al.*, 2009; Ding *et al.*, 2010). 3D structures were rendered in PyMol (Schrödinger LLC, 2019) using PDB IDs 1o05 and 1zum (Larson *et al.*, 2005; Hurley *et al.*, 2001; Berman *et al.*, 2000; Burley *et al.*, 2019)

affect protein functions. A further source for variants that warrant more detailed studies are genome-wide association studies which identify statistically significant differences in SAV frequencies within a large cohort of individuals with and without a trait of interest. However, these analyses are not well suited for the identification of large effect SAVs since these will not be common in the general population (Manolio *et al.*, 2009; Marjoram *et al.*, 2014; Carss *et al.*, 2019). Generally, any type of targeted analysis is limited in scope as it requires prior knowledge of which variants could be interesting. With the advent of HTS

methods, WES or even WGS has become a viable, less targeted approach for studying the variants underlying in particular complex diseases (Bamshad *et al.*, 2011; Kiezun *et al.*, 2012). Compared to association studies these sequencing efforts can detect both common and rare variants. However, this approach still requires a sufficiently large and specific sample of individuals, such as a disease cohort, to confidently associate identified variants with the disease in question (Carss *et al.*, 2019).

Alanine scanning, the systematic replacement of every amino acid by alanine, represents an early approach for breaking out of targeted SAV effect analysis (Cunningham and Wells, 1989). Nonetheless, every variant's effect still had to be measured individually which limits the scale of these studies. Furthermore, alanine scanning elucidates positions which are sensitive to changes but probably not the effect size that can be expected of those SAVs (Gray *et al.*, 2017).

Generally, variants that have an effect are more studied than neutral ones. This is caused by effect variants being more clinically relevant and actionable, but also the difficulty in proving that a certain variant has no effect under all possible conditions. Nonetheless, knowledge of neutral variants is crucial for example when developing prediction methods as they provide the signal which machine learning models should differentiate effect variants from (see Section 1.3). Since few experimentally determined neutral variants are available, authors of prediction method often have to create synthetic datasets of proxy-neutral variants. For example, SNAP used sequence changes between enzymes that perform the same function as likely neutral SAVs (Bromberg and Rost, 2007). CADD, assumes all variants which are fixed in the human population since the last common ancestor to be effectively neutral as they have survived purifying selection (Rentzsch *et al.*, 2019; Sundaram *et al.*, 2018).

### 1.2.3 Deep mutational scanning

More recently, another experimental approach, so-called multiplexed assays for variant effect, has gained popularity (Starita *et al.*, 2017). These assays essentially extend the idea of alanine scanning while using the advances in HTS. As such, their key advantage over previous methods is establishing a link between the variant in question and a specific functional assay which can be measured on a high-throughput scale. Multiple approaches have been developed based on this idea, including deep mutational scanning (DMS) which evaluates the functional consequence of all SAVs in a protein of interest (Araya and Fowler, 2011; Fowler and Fields, 2014).

In DMS, a sequence library with all possible variants is created through for example error-prone PCR or a variety of more specialized protocols such as EMPIRIC, PALS, PFunkel or

Nicking mutagenesis (Hietpas *et al.*, 2012; Fowler *et al.*, 2014; Wrenbeck *et al.*, 2017a). Next, the variant library is introduced into a selection system, i.e., an assay that specifically links the variant sequence to the function of the studied protein. Ideally, the choice of functional assay and intensity of functional pressure are optimized together to capture a wide range of effect as opposed to just a small set of high effect variants. Possible assay-types include display methods such as phage display, cell-based assays which link protein function to cell fitness or cell-sorting, e.g., using fluorescence-activated cell sorting.

After variants have been introduced in the selection system, some sort of functional pressure is imposed. For example, the assay might be constructed such that function of the protein of interest is mandatory for cell survival. HTS is then used to determine variant frequency before and after applying functional pressure. The application of HTS in this step is the major improvement of DMS over earlier methods as it allows these measurements to be efficiently determined simultaneously, and thus greatly increases the possible throughput. To account for high error rates in some HTS methods, many protocols employ paired-end reads, i.e., sequencing the DNA from both ends. Another option is tagging each variant with a unique barcode (Mavor *et al.*, 2016). Barcoding has the additional advantage that it increases the sequence length which can be studied beyond the maximum read length of the HTS method employed.

Finally, variant frequencies from sequencing are transformed into a functional score. In the most simple case, the functional score is the fraction of the variant's frequency after and before applying selective pressure. This way, lower scores indicate variants which have reduced function. Due to experimental design and noise, the wild-type does not always result in a score of 1. Therefore, it is common to normalize the functional score by the wildtype (wt) frequencies as shown in Equation 1.1 where 0 denotes the time point before and 1 after applying selective pressure (Rubin *et al.*, 2017).

$$\forall \text{ variants } v, ER = \frac{ratio_v}{ratio_{wt}} = \frac{\frac{count_{v,1}}{count_{v,0}}}{\frac{count_{wt,1}}{count_{wt,0}}} \tag{1.1}$$

The score may further be logarithmized. Any version of this scoring is commonly referred to as the (log-scaled) enrichment ratio. For more sophisticated analysis like the calculation of confidence measures, toolkits such as Enrich2, dms_tools, PACT or the variant effect map imputation webservice have been developed (Rubin *et al.*, 2017; Bloom, 2015; Klesmith and Hackel, 2019; Wu *et al.*, 2019).

Besides the increase in efficiency and ability to scale to thousands of variants, DMS also has the added benefit of detecting variants with beneficial effect, whereas targeted analyses are typically focused only on deleterious variation. The same goes for the detection of variants that show no effect, i.e., are neutral under the given conditions of the assay. One drawback is that each assay should be specific to the protein's function, hence limiting the scalability of the method. Furthermore, a single protein might lend itself to multiple functional assays or have multiple functions which necessitates a more general or several assays. Recently, a protocol termed VAMP-seq has been introduced as a more widely applicable functional assay based on intracellular protein abundance (Matreyek *et al.*, 2018).

Apart from measurement of a protein's natural function, DMS has further been applied to screen proteins for example for improved drug binding, antibody affinity, using non-native chemical stresses, or non-proteinogenic amino acids, and is also amenable to synthetic proteins (Forsyth *et al.*, 2013; Mavor *et al.*, 2016, 2018; Tinberg *et al.*, 2013; Procko *et al.*, 2013; Whitehead *et al.*, 2012; Fujino *et al.*, 2012; Rogers *et al.*, 2018). DMS studies share some aspects with directed evolution experiments which assay a wide range of variants for engineering a protein of interest towards a specific purpose. In fact, protein engineering has been performed based on the results from DMS studies for example by efficiently screening for mutants that improve ligand binding (Wrenbeck *et al.*, 2017b). Experiments where the assay evaluates which variants can rescue the function of a mutant protein have also been implemented with DMS and further highlight the wide applicability of the method (Wu *et al.*, 2013; Wagenaar *et al.*, 2014).

Finally, in 2019 DMS data was used for inferring protein three-dimensional (3D) structure (Chiasson and Fowler, 2019). Two groups independently presented similar approaches for using the effect of several 100,000 double mutants obtained through DMS. From these they determined co-evolutionary constraints which allowed the inference of 3D protein structure at a resolution rivaling experimental structure determination methods (Schmiedel and Lehner, 2019; Rollins *et al.*, 2019). While the protein domain in question was just 56 amino acids long and the DMS dataset unique in its high number of double mutants, these approaches provide a promising outlook for future usage of DMS data in one of the most challenging tasks of computational biology.

### 1.2.4 Resources of variant effect

The variety of variant effect types and approaches for their determination is reflected in the number of resources cataloging them. This subsection will give a short overview of

some commonly used databases. They serve as resources not only to researchers interested in particular variants but also developers of prediction methods who must carefully assemble their training sets (cf. Section 1.3). The data in these resources might be anything between a mere list of variants imputed from raw (sequencing) data or manually curated information on the specific type and strength of effect caused by the variant.

One of the largest general databases for small variants such as SNVs, insertions and deletions is dbSNP containing almost 700,000,000 entries in the most recent build (August 2019, Sherry *et al.*, 2001). dbSNP reference identifiers are pervasive in the community due to their stability and are used to uniquely identify a specific locus for a type of variation. For example, identifier rs334 describes the locus in human hemoglobin where the SAV E7V causes sickle-cell anemia. As a complement to dbSNP, dbVar similarly catalogs large, structural variants (Lappalainen *et al.*, 2013). Since 2017, both databases are focusing on human variants and do not accept new submission for other organisms. Handling of non-human variation has since been taken over by the European Variation Archive which has a comparably broad focus and is in regular exchange with dbSNP and dbVar (EVA, 2019).

Many databases focus on variants implicated in (human) diseases. Arguably, the most common one is OMIM which contains around 5,500 human phenotypes for which the molecular basis, e.g., the SAV, is known (Amberger *et al.*, 2019). Many of these diseases are monogenic which leads to datasets built from OMIM being highly biased towards variants with strong effect (see also Subsection 1.3.2). OMIA follows the same approach as OMIM for animal phenotypes, in particular model organisms or those relevant for production and breeding (Lenffer *et al.*, 2006; OMIA, 2019). Mouse and rat variants are excluded since they are contained in their own specialized databases (Eppig *et al.*, 2015; Shimoyama *et al.*, 2015). HGMD is similar to OMIM, and a manually curated database focusing exclusively on human germline disease variants (Stenson *et al.*, 2017). It currently contains almost 270,000 variants of which only 170,000 are accessible through the non-professional version. ClinVar is yet another database of human disease-associated variants but less restrictive than OMIM and HGMD. It currently contains around 560,000 unqiue variants (October 2019, Landrum *et al.*, 2016). Finally, UniProtKB/Swiss-Prot offers humsavar, an automatically created file of all around 80,000 human SAVs found in their database. Of these, about 31,000 are associated with a disease (Release 2019_09, The Uniprot Consortium, 2019).

Next to the wealth of resources focusing on disease variants, there are many other databases with entries for a special type of variant. For example, IARC TP53 contains variants of the human *TP53* gene which codes for a transcription factor involved most prominently in tumor progression (Bouaoun *et al.*, 2016). MutHTP aggregates vari-

ants in human transmembrane proteins from sources including humsavar and ClinVar (Kulandaisamy *et al.*, 2018). Variants obtained from multiplexed assays for variant effect, including DMS studies can be found in MAVEDb (Esposito *et al.*, 2019). Other databases are primarily an aggregation of datasets: VariBench is a collection of variant effect datasets which are categorized by specific effect types, diseases or proteins of interest (Sasidharan Nair and Vihinen, 2013; Sarkar *et al.*, 2019). Sets can further be redundancy reduced among each other, thus building a convenient source of training and testing datasets for the development of VEPs. Finally, dbNSFP contains predictions by 26 methods for all possible human SNVs together with cross-references of effect annotations from sources such as dbSNP and ClinVar (Liu *et al.*, 2016).

As mentioned before (see Subsection 1.2.2), experimentally determining neutral variants is difficult and several approaches to tackle this problem exist. Many of the previously mentioned resources also contain variants annotated as neutral. However, for approaches which define neutral variants based on their occurrence in sequencing data the gnomAD database should be pointed out (Karczewski *et al.*, 2019). It contains data from 125,000 human WES and 15,000 human WGS studies and as such presents an invaluable resource of high-quality HTS data. Initially, its focus was solely on exome data, made available by the Exome Aggregation Consoritum (ExAC). Hence, data which is now found in gnomAD is often still attributed to ExAC in publications.

Overlaps between the above-mentioned resources are common and often intended with explicit cross-references. This way, every resource can play to its specific strength and users choose the database best suited to their needs. Which resource to use for a given project is therefore usually a matter of what exactly one wants to achieve and might also be determined by factors such as a the programmatic accessibility. Especially when using these resources for the creation of training and testing datasets of VEPs, the most important point is to be aware of which type of data is being used since it will determine the behavior and potential biases of the resulting method.

## 1.3 Prediction of variant effect

Variant effect prediction methods (VEPs) assess the outcome of a given variant *in silico* avoiding costly and time-consuming experiments. As simple as this definition is, the reality of the field presents itself to be far more complex. With which types of variants (see Subsection 1.1.2 on page 4), what effects (see Section 1.2 on page 6) and based on which data (see Subsection 1.2.4 on page 12) training is performed, are just three major points differentiating VEPs. Probably the most comprehensive overview of VEPs, as well

as some resources, is currently provided by VIPdb (Hu *et al.*, 2019). It catalogs more than 150 VEPs and denotes information about reference, access and in particular which variant types are supported as input. Below, Subsection 1.3.1 describes some exemplary VEPs focusing on the prediction of SAV effects—the most commonly predicted type of variant. Subsection 1.3.2 on page 20 then discusses various challenges of which are important to recognize when using their results.

### 1.3.1 Overview of variant effect predictors

#### Approaches focused on exploiting conservation patterns

The first VEPs harnessed conservation patterns from homology information to determine the evolutionary tolerance of a variant. The assumption behind this approach is that sites are conserved for a reason. They are under more evolutionary pressure not to change in order to maintain function. To exploit this idea, a VEP might first build a multiple sequence alignment (MSA) which represents a family of proteins. SIFT, short for sorting intolerant from tolerant, is an early prediction method based solely on this concept and remains popular to this day (Kumar *et al.*, 2009; Ng and Henikoff, 2003). Given an input sequence of interest, an MSA is created from the results of performing a homology search with PSI-BLAST (Altschul, 1997). From this alignment, SIFT extracts the occurrence of every amino acid at every position, normalized by the frequency of the most common amino acid. If the resulting value for a SAV is below an empirically defined threshold, the variant is predicted to have an effect. In addition, SIFT provides a confidence value for every position which is based on the number of amino acids that are observed at that position in the alignment. The basic concept—SAVs which alter conserved positions are likely to have an effect—still lies at the basis of most current VEPs. PROVEAN is a direct improvement on SIFT using a very similar approach but adding support for small insertions and deletions. Furthermore, it introduces a more sensitive treatment of PSI-BLAST results which makes the method more powerful for large protein families that contain many similar sequences (Choi *et al.*, 2012).

MutationAssessor is another VEP that heavily relies on conservation (Reva *et al.*, 2011). It predicts the functional impact of variants based on two complementary evolutionary scores: In an MSA created by MUSCLE (Edgar, 2004), and using homologs from a BLAST search against UniProtKB, the conservation score measures the change of entropy in every column of the alignment when a variant is introduced. This score is position-specific and depends on the original and mutated amino acid but it is constant for all sequences in the alignment with the same conditions. The second specificity score follows the same idea

of entropy changes introduced by variants. It is based on further clustering all sequences in the alignment into subfamilies which differ at a subset of positions determinant for their functions. Both scores are combined by averaging to form the final prediction score.

### Application of machine learning to the variant effect prediction problem

Machine learning generally describes the act of training a program on performing a specific task without explicitly writing the rules to achieve this but rather by providing a general framework and examples to learn from (Larrañaga *et al.*, 2006; Chicco, 2017). This is important for tasks where such a set of rules cannot be realistically determined or would be too complex to program. In computational biology some of the most common frameworks, also referred to as models, are neural networks, support vector machines, hidden Markov models or random forests (Jensen and Bateman, 2011). In the most common case of so-called supervised learning, the model is trained with a set of inputs for which the intended output is known. The type of inputs, typically referred to as features, can be anything in a representation suitable for the program. For example, a protein sequence can be represented using one-hot encoding, or by its conservation information as obtained from an MSA and formalized in a position-specific scoring matrix (PSSM, Table 1.1). A variant in the sequence could be expressed simply by using the same input representation twice. Once with the original sequence and once with the sequence containing the variant. Given the inputs, the output to be learned has to be defined as well. For a given SAV this could be either 'neutral' or 'effect'—a classification task—, or the strength of effect over a range of values—a regression task. Throughout the training process a set of parameters depending on the specific machine learning model is optimized to yield the desired output, given the inputs. At the same time, care must be taken to train a model that generalizes well and is not just a perfect representation of the training set. This can be ensured by applying the final trained model on an additional test dataset which contains inputs that have not been used for training and are not similar to those. When every step is performed with care, the performance on this test set is an indication of what the model will achieve when applied to unknown data points.

PolyPhen-2 builds upon the previously discussed conservation concept but applies a machine learning approach (Adzhubei *et al.*, 2010, 2013). A naïve Bayes classifier is trained to predict SAV effect using eleven input features of which eight are related to sequence conservation. The other three features represent structural properties of the protein and are related to its accessible surface area and stability. The structural features can only be included in the prediction when a protein with solved 3D structure related to the prediction query can be retrieved. Using this framework, two models of PolyPhen-2 were

**Table 1.1. Numeric representations of an amino acid sequence.** To use a protein sequence as input for most machine learning models, the characters need to be transformed into a series of numbers or replaced by a set of features that represent the information in the amino acid sequence. There are several ways this can be achieved. Two representations are shown here as an example with the first five residues MDLSA of protein *BRCA1*. In every row, that is, at every position in the protein, 20 possible inputs exist. **(a)** shows a so-called one-hot encoding where 20 values represent these 20 possible inputs and only the one actually encountered is set to 1 while all others are set to 0. **(b)** shows one possibility of incorporating information about conserved residues in the input. Based on a homology search for similar proteins, e.g., from related organisms, an MSA can be created. Given the MSA one can assign each possible substitution a value representing how often it was observed in the data. For example, the MSA might show that the methionine at position 1 is often conserved in related proteins and therefore a large value of 9 is assigned (highlighted in green). Other substitutions such as M1P are rarely observed and therefore assigned a small value (highlighted in red). This format is referred to as a PSSM and a common output of homology search tools such as PSI-BLAST.

**(a)** One-hot encoding

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| L | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| A | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**(b)** Evolutionary conservation

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M | −3 | −4 | −4 | −5 | −3 | −3 | −4 | −5 | −4 | 2 | 1 | −4 | 9 | −2 | −5 | −4 | −3 | −4 | −3 | 0 |
| D | 2 | −3 | −1 | 6 | −4 | −2 | 1 | −3 | −3 | −4 | −4 | −2 | −4 | −1 | −3 | 1 | 0 | −5 | −3 | −3 |
| L | −1 | −1 | −3 | −3 | −1 | −1 | 0 | −2 | 0 | 0 | 3 | −1 | 4 | 0 | −3 | −1 | 1 | 1 | −2 | 0 |
| S | 1 | −2 | −2 | −1 | −2 | 1 | 0 | −1 | −1 | −2 | −2 | 1 | 1 | −4 | 2 | 4 | 0 | −4 | −3 | −3 |
| A | 3 | 1 | −1 | −2 | −2 | −1 | −1 | −1 | −1 | −2 | −1 | −1 | 2 | −2 | −3 | 2 | 2 | 1 | −2 | 0 |

trained which both focus on human disease-causing SAVs. One of them tries to adjust for SAVs with less strong effects as they may appear in complex polygenic diseases, the other is trained for SAVs with strong effects as they are common in Mendelian diseases.

SNAP (Screening for non-acceptable polymorphisms, Bromberg and Rost, 2007) and its successor SNAP2 (Hecht *et al.*, 2015) are VEPs similar to PolyPhen-2 in that they apply machine learning to the problem and supplement conservation based input features with structural ones. While SNAP2 is based on larger set of features and uses a neural network

as machine learning model, the first major difference is that the structural features are not extracted from homologs with determined 3D structures. Instead, features such as the secondary structure of every residue are predicted from sequence by other integrated methods—in this case ReProf (Rost and Sander, 1994; Yachdav *et al.*, 2014). The second major difference lies in the training data. While PolyPhen-2 focuses on disease variants, SNAP2 was mostly trained on functional effect SAVs and hence predicts a different type of effect. This fact is also reflected in the raw output values which are almost binary in the case of PolyPhen-2 and more continuous for SNAP2.

MutPred combines tens of different prediction methods, each targeting particular aspects of structure and function (Li *et al.*, 2009). The final model distinguishes between disease-causing variants and polymorphic SAVs from UniProtKB/Swiss-Prot which are considered neutral. Ultimately, the aim of this approach is to highlight the molecular mechanism of a disease as part of the prediction. Through the ensemble of methods, MutPred provides not only the classification of a SAV as effect or neutral, but also a suggestion of why an effect was predicted, e.g., because the SAV was in an important protein-DNA binding site. The concept has been improved and updated to current training datasets with MutPred2 (Pejaver *et al.*, 2017) and three related VEPs have been developed for predicting the effect of variants other than SAVs (Mort *et al.*, 2014; Pagel *et al.*, 2017, 2019).

#### New possibilities from HTS data

Improvements in HTS and methods for high-throughput variant effect determination (see Subsections 1.1.1 and 1.2.2) lead to increasingly large amounts of potential training data. Generally, this is to the benefit of traditional machine learning approaches. However, it further opens the possibility for applying deep learning methods which typically require several magnitudes larger training data (Angermueller *et al.*, 2016; Jurtz *et al.*, 2017; Eraslan *et al.*, 2019). This approach which effectively consists of using neural networks with more than one hidden layer, has recently seen a renaissance in machine learning. As datasets grow, applying deep learning techniques might be necessary to extract the best features from the high-dimensional data, potentially unsupervised, i.e., without any knowledge about the structure and labels of the data (Cao *et al.*, 2019; Riesselman *et al.*, 2017). In particular, recent advances using methodology from the field of natural language processing for the extraction of high quality embeddings, that is, representations of biological sequences, have shown promising results and will likely lead to a new class of VEPs in the near future (Heinzinger *et al.*, 2019; Rives *et al.*, 2019; Shamsi *et al.*, 2020).

CADD also benefits from the increase in HTS data and is one of the fewer methods which use genomic instead of protein sequences as input (Kircher *et al.*, 2014; Rentzsch *et al.*,

2019). CADD was trained on a dataset of variants that arise when comparing the current human genome to the last common ape ancestor's. These variants are fixed and therefore assumed to be effectively neutral as purifying selection would have otherwise removed them from the population. On the other hand, effect variants created *in silico* are not under evolutionary pressure and are expected to have primarily deleterious effects. With this training set, CADD extracts hundreds of features and trains a logicistic regression model for the prediction of variant effect.

With the exponential increase in sequencing data, execution speed has also come into focus as a relevant factor. The VAAST variant prioritizer aims at not necessarily predicting a perfect approximation of every variant's score, but rather providing a rough—and highly scalable—filter for prioritizing variants of high interest that warrant further investigation (Flygare *et al.*, 2018). This is achieved without machine learning using only a likelihood-ratio test based on conservation and whether a variant is homo- or heterozygous. The approach makes the VEP particularly well suited to help interpret the results of WES or WGS studies. While it is this methods' expressed goal, the fact is that every VEP can really only provide a filter for finding variants of interest as results are never accurate enough to directly act on them (see Subsection 1.3.2).

Finally, the increased prevalence of DMS has lead to the development of a VEP based on this new kind of data (Gray *et al.*, 2018). Envision is an ensemble of decision trees trained with a relatively standard feature set including evolutionary and structure information. However, the unique training set consists only of variants from DMS experiments and—unlike most VEPs—the model was trained as a regression predictor, i.e., it aims to predict the strength of effect. Envision thereby extends variant effect prediction to new avenues and further holds the potential for significant improvements since it is based on a type of data that will likely see large increases in the near future.

#### Meta-predictors

dbNSFP categorizes VEPs as conservation scores, functional prediction scores, general prediction scores, and ensemble scores (Liu *et al.*, 2016). An example of the last class is the consensus predictor REVEL which applies a random forest machine learning model (Ioannidis *et al.*, 2016). However, the only input features are predictions of variant effect from a set of 13 VEPs, including the previously described SIFT, PROVEAN, MutationAssessor, PolyPhen-2 and MutPred. Similar approaches are used by VEPs such as Condel, Logit or MetaSVM and MetaLR (González-Pérez and López-Bigas, 2011; Li *et al.*, 2013; Dong *et al.*, 2015).

### 1.3.2 Challenges

#### Reliance on sequence conservation

The majority of VEPs uses sequence conservation as one of or even their only input feature (Subsection 1.3.1). Typically, this information is obtained in one way or another by building an MSA from sequences homologous to the query. This is intuitively a promising approach since positions that show high conservation are likely to be functionally important. A variant in such a position is therefore likely to have an effect and the approach is generally powerful (Andersen *et al.*, 2017; Stein *et al.*, 2019; Zhang *et al.*, 2019). However, training VEPs on this data also brings issues with it. Firstly, not all residues are conserved because of their functional importance and secondly, positions which show low conservation can still be crucial for function (Sun and Yu, 2019). For example, a variant may be species-specific and important for function but will appear as non-conserved when building an MSA from related species' homologous sequences.

The problem of bias towards conserved residues is further compounded by the machine learning applied to train VEPs. It is a highly complex and difficult task to ensure that the training recognizes conservation as an important feature, yet maintains the generalized understanding that not every variant at a conserved position should automatically be predicted as having an effect. Rather, the hope of training a predictor with multiple input features is that the VEP recognizes patterns beyond this simple inference. Unfortunately, evaluations of VEPs throughout the years consistently highlight that this issue is far from solved (Fowler and Fields, 2014; Miller *et al.*, 2017, 2019b; Cline *et al.*, 2019). For example, in a recent study by Sun and Yu (2019) eleven of the twelve evaluated VEPs—including most of those mentioned in Subsection 1.3.1—showed significantly increased false positive predictions of variants as having an effect at conserved positions. Furthermore, all VEPs performed significantly worse in recognizing disease causing variants at residues with low conservation.

As a final point, Miller *et al.* (2017) introduced the distinction of residues in two categories: Toggle positions where a variant either has a strong effect or not and rheostat positions where multiple variants lead to a range of effect strengths. Toggle positions are usually conserved leading to effect variants well recognizable by VEPs (Miller *et al.*, 2017). Furthermore, positions where no variant has an effect are usually non-conserved. However, rheostat positions show varying levels of conservation which motivates their integration in the training process of future VEPs (Miller *et al.*, 2019a).

### Bias towards deleterious effect variants

The training sets of many VEPs are biased towards deleterious effect variants and high effect SAVs as they are common for example in Mendelian disease variants from OMIM (see Subsection 1.2.4). Along with this comes a bias in predictions to typically overestimate effects (Figure 1.3, Richards *et al.*, 2015; Anderson and Lassmann, 2018; Andersen *et al.*, 2017; Niroula and Vihinen, 2019). This issue is worsened by the fact that pathogenic variants are often located at conserved positions (Sun and Yu, 2019). VEPs are therefore often biased towards high effect variants and conservation which both reinforce each other.

Furthermore, neutral or beneficial effect variants are typically not recognized well. For example, the five most commonly used algorithms in dbNSFP agreed for 79 % of pathogenic variants, but only 33 % of neutral ones (Kim *et al.*, 2019). A similar bias was observed when evaluating predictions of ClinVar variants. Here, false positives, i.e., neutral variants predicted as having an effect, were much higher than false negatives. On a set of neutral ExAC variants, the performance of ten VEPs differed widely and even the best method still predicted every twelfth variant as having an effect (Niroula and Vihinen, 2019). Nowadays negligence of neutral variants in training sets is a recognized problem and one common approach to tackle it is using variants that became fixed in the human population. However, this is not an ideal solution either since fixed variants could also have a beneficial effect (Subsection 1.2.4, Rentzsch *et al.*, 2019). Variants with experimentally determined beneficial effect on the other hand should become more common with the rise of DMS studies and can thus be better accounted for during the development of future VEPs.

### Quality and heterogeneity of training data

Given the multitude of resources outlined in Subsection 1.2.4, it may seem that collection of training data is not a challenge in itself. However, as previously discussed, method developers have to exercise great care when assembling their data sets to avoid the introduction of biases to the VEP. The underlying training data is also not perfect and contains inconsistent annotations between different laboratories or large-scale sequencing efforts (Kim *et al.*, 2019)—despite efforts to standardize what is considered an effect at different levels (cf. following Subsection, Richards *et al.*, 2015). Another point that must be recognized is that proteins can have multiple functions. Therefore, some variants might have an effect in one condition or on one function but not the other. For example, this has been extensively studied for Ubiquitin (Mavor *et al.*, 2016, 2018, 2019; Roscoe *et al.*,

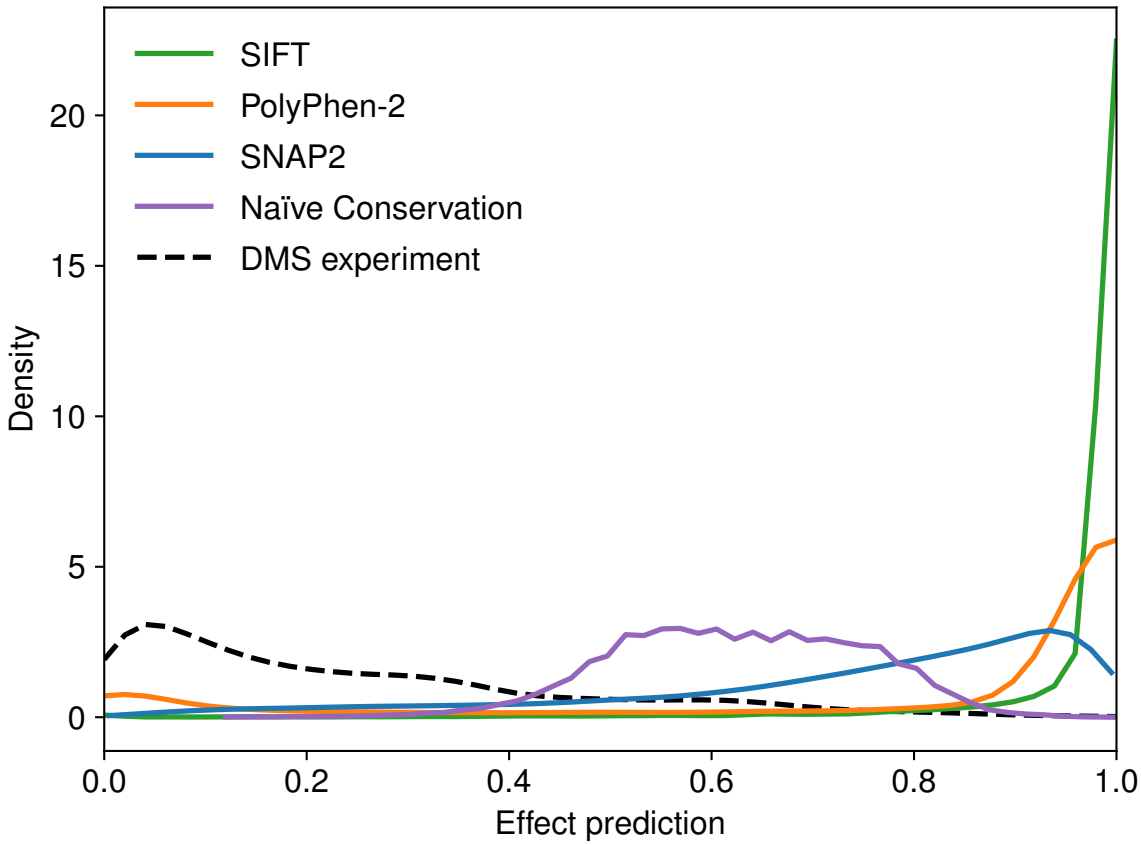



**Figure 1.3. Variant predictions are often biased towards effect.** Shown are predictions of variant effect by the VEPs SIFT, PolyPhen-2 and SNAP2 which were performed on a set of 17,781 SAVs from 22 DMS experiments (see Chapter 3 on page 47). Naïve Conservation represents a simple prediction based on a homology search by PSI-BLAST. The x-axis denotes the prediction score scaled to lie within the interval $[0, 1]$. For all predictions, higher values denote increased affirmation by the VEP that the SAV has an effect. Therefore, a value of 0 denotes a prediction of no effect, i.e., a neutral SAV, while a value of 1 denotes a SAV predicted to very likely have an effect. For the experimentally determined effect (dashed black line), values denote increasing strength of effect as measured by the underlying DMS assays. The y-axis represents the density function from the Gaussian kernel density estimation. Intuitively it shows the number of data points with the respective effect score. All predictions methods show a clear shift of values towards effect when compared to the experimental measurements. Naïve Conservation shows that this effect cannot be attributed solely to reliance on conserved residues for prediction.

2013; Roscoe and Bolon, 2014) and BRCA1 (Starita *et al.*, 2015; Findlay *et al.*, 2018). In such cases one can expect that a VEP either only correctly detects variants which show a signal in all cases or only those specific to one condition or function. The decision between these options lies in how and on what data training was performed. Given that, one could argue that it makes sense to have a specialized VEP for every class of proteins.

However, such a predictor would have a very limited range of application. Furthermore, such measures are not generally necessary. For example, integral membrane proteins are a notoriously difficult to handle and a highly specific protein class (Reeb *et al.*, 2014; Punta *et al.*, 2019). Nonetheless, they are seemingly not posing a particularly hard challenge to VEPs. Although few variants in typical training sets are of membrane proteins, prediction performance was found to be comparable to that on soluble proteins for a set of VEPs including most of those mentioned in Subsection 1.3.1 (Orioli and Vihinen, 2019).

Another major point to consider is that although most resources supply only high quality data, it is far from trivial to arrive at this point and not everything deposited can automatically be considered a gold standard (Arthur *et al.*, 2015; Harrison and Rehm, 2019). Issues arise already at the level of (high-throughput) sequencing where different methods achieve varying fidelity. The problem is further compounded by the choice among numerous read mappers and variant callers which then have to be applied to the raw data, yielding the variants that can finally be used for training a prediction method. On this level, the Genome in a Bottle Consortium provides benchmark human genomes and corresponding variant calling together with tools for automated assessments of HTS pipelines (Zook *et al.*, 2014, 2019). PrecisionFDA is another initiative benchmarking various aspects of HTS pipelines with the goal of making their results more clinically actionable (precisionFDA, 2019). A final aspect is that many variants, in particular SAVs, come from WES since it is cheaper to perform than WGS. However, WGS provides more and higher quality SAVs (Belkadi *et al.*, 2015).

#### Different interpretations of variant effect

Arguably, the largest issue in understanding and comparing the results from different VEPs lies in the fact that it is a matter of definition what exactly constitutes a variant's effect (see Section 1.2). For example, J. Shendure and J. M. Akey (2015) suggested to consider the views of (i) fitness, i.e., how a variant effects the organism's reproductive ability, (ii) pathogenicity, i.e., whether the variant leads to disease development, and (iii) molecular function, which describes the effect on the respective protein's function, e.g., enzymatic activity. The last two in particular are popular choices when developing a VEP but describe two highly different cases (Gray *et al.*, 2018). This heterogeneity of effect is reflected in the goals and thereby training sets, features and prediction behaviors of different VEPs (Sun and Yu, 2019). A common approach to tackle non-consistent prediction results is to apply multiple VEPs and evaluate their consensus (Richards *et al.*, 2015) or explicitly developing a meta-predictor such as REVEL (see Subsection 1.3.1). In theory, such meta-predictors could harness the complementing strengths of differently

trained VEPs. However, there are conflicting results on the success of such approaches for variant effect prediction (Grimm *et al.*, 2015; Daneshjou *et al.*, 2017; Anderson and Lassmann, 2018; Sun and Yu, 2019). Often a meta-analysis may further obfuscate the signal and an expert user would benefit from seeing the raw prediction results while knowing about the potential focus and bias of each method (Vihinen, 2020).

A related point is that not all tools predict variant effect on a continuous scale. This works for toggle positions, but for rheostats a more fine-grained distinction than neutral or effect is necessary. Envision is one of the few VEPs trained as a regression-based prediction method and therefore intrinsically accounts for this. For SNAP2 it has been shown that the output scores correlate with effect strength, although the method was not explicitly trained for this (Bromberg and Rost, 2007; Bromberg *et al.*, 2013; Mahlich *et al.*, 2017). Other methods like PolyPhen-2 or SIFT provide almost binary prediction outputs making such analyses challenging. On a higher level, complex diseases also pose a similar continuous effect challenge as they are elicited by numerous small effect variants (Wray *et al.*, 2013). Precisely because complex diseases are so poorly captured by current VEPs, the increased use of endophenotypes has been suggested (Masica and Karchin, 2016). Endophenotypes are quantifiable traits such as enzymatic activity with clearer genetic cause than for example the complex diseases that develop as a phenotype from multiple underlying endophenotypes. It is reasonable to assume that endophenotypes will be easier to predict by VEPs than disease outcome. Current approaches might only need small adjustments together with new training data to capture them.

Finally, VEPs do not consider the conditions in which the variant appears, including whether a variant appears in the germline or somatic cells (Carss *et al.*, 2019; Gray *et al.*, 2018). VEPs also do not account for additional variants which may occur together with the one being predicted. This is a major flaw since correlated variants can have varying outcomes ranging from reinforcing the effect to a complete rescue (Göbel *et al.*, 1994; Kowarsch *et al.*, 2010). For example, one can imagine that the substitution of a small amino acid like alanine by a large one such as tyrosine can severely affect the stability of the protein, especially when occurring in the protein interior. However, if at the same time a large amino acid which lies opposite in 3D space is substituted by a much smaller one, the variant may be tolerated and not have an effect. Such cases can be one source of incomplete penetrance, i.e., a variant leading to a phenotype in one person but not another. Seminal work by Hopf *et al.* (2017) has lead to the development of EVmutation, a VEP which exploits information about sequence co-evolution and can be applied to predict the effects of multiple variants occurring together. While there are several challenges to this approach, in particular lack of data for the large number of parameters that need to be tuned, EVmutation shows how future VEPs might account for the issue of co-occurring

variants.

### Evaluation and clinical use of VEPs

In the same way that every tool is biased by its training data and therefore predicts a very specific aspect of the effect spectrum, evaluations can fall into the same trap when assembling test datasets. This is intrinsically true for the evaluation every method developer performs as part of the machine learning process while building the tool. However, it is also key for independent evaluations performed by researchers assessing published methods. If a VEP performs well in an evaluation setting, the narrower the test data, the likelier that the achieved performance is just a sign of bias being confirmed in testing and results will not generalize well. Therefore, evaluations should ideally aim to assess as widely as possible but must in addition always outline in detail exactly what data was used to obtain the results since none will ever be pervasive enough to capture all aspects. Ideally, experimental datasets essentially tailored as test for specific VEPs' goals would be created alongside (Miller *et al.*, 2019b). This, however, is currently an unrealistic scenario given the number of avenues that would have to be covered and implicated cost.

In terms of evaluating VEPs, the Critical Assessment of Genome Interpretation (CAGI) should be highlighted (Andreoletti *et al.*, 2019). Strongly related to similar efforts from other fields such as structure (CASP) or function (CAFA) prediction, the idea of CAGI is to use previously unpublished variant effect data and collect blind predictions by VEP developers on this data. Once the submission deadline has closed, the predictions will be evaluated by independent assessors. The need of unpublished variant effect data somewhat limits the pervasiveness that CAGI can achieve. Therefore, it usually consists of several challenges which each focus on, e.g., a particular protein or disease phenotype. In 2018, the fifth iteration of CAGI concluded with the evaluation of 14 challenges assessing a range of variants from regulatory and non-synonymous over those affecting splicing to disease variants, for example in breast cancer. The design of CAGI enables an independent, recurrent view to the field of variant effect prediction and continues to give valuable insights such as VEPs' reliance on sequence conservation and bias towards effect variants (Cline *et al.*, 2019; Miller *et al.*, 2019b; Zhang *et al.*, 2019; Monzon *et al.*, 2019) or optimization for classification of high effect variants rather than regression (Pejaver *et al.*, 2019).

Grimm *et al.* (2015) identified two major points affecting assessment of VEPs which are pervasive in the field: The first, "type 1 circularity", describes an overlap between the training and testing data. This may seem like a trivial point and every method developer should be well aware that testing cannot be performed on samples that the VEP has already

seen in training without introducing a severe bias to resulting performance measures. However, the problem is more complex for independent evaluations since the testing dataset for such an effort has to be kept disjoint from all evaluated VEPs' training data. The issue is particularly poignant for consensus prediction methods as the training data of all underlying methods must be considered. "Type 2 circularity" describes a more complex problem which arises through biases in experiments and variant effect databases. Many database entries are genes which contain a multitude of effect variants. Once an effect variant has been identified in a specific gene, there is increased interest to experimentally study further variants in the gene increasing the known effect variants. When training a VEP on such datasets it is very likely that it will learn not to distinguish those variants as having an effect but rather learning that every variant in the gene has an effect. This creates a bias towards predicting effect variants in general and furthermore makes the detection of neutral variants in such a gene unlikely. However, identifying neutral variants next to effect variants is arguably one of the most important tasks for a VEP. After all, the simple association of considering every variant in a "disease gene" as having an effect would not require a method which employs advanced machine learning. This issue could also be one possible explanation for the fact that SIFT, PolyPhen-2 and SNAP2 differed much more on variants without experimental information than those with known outcomes (Reeb *et al.*, 2016; Mahlich *et al.*, 2017).

In the end, the goal of studying variant effect is to arrive at clinically actionable knowledge about sequence variation which can be used to improve the quality of human lives. Carss *et al.* (2019) name three steps for the clinical interpretation of a variant: (i) being sure that its detection is not a technical error and hence a false positive, (ii) ensuring the variant has an effect on the gene product, and (iii) being confident that the change in function actually causes a clinical phenotype. VEPs answer the second part but not the first and third. Furthermore, summarizing the previous paragraphs one can say that current VEPs are good at distinguishing effect and neutral variants only on average. This makes them useful tools for the prioritization of variants, yet completely unsuited for direct clinical application. In fact, a recent publication found a small set of VEPs to perform significantly worse on clinically applicable variants when compared to typically used evaluation sets (Gunning *et al.*, 2020). Not only are error rates too high but understanding of their occurrence is too low to ethically use them for patient diagnosis. It seems neither likely nor necessarily desirable that VEPs will ever be reliable enough for direct clinical application. However, as the prospect of personal genomes becoming widely available in the near future is increasingly realistic, the demand for interpretation of the raw data will grow. Therefore, the most reasonable outcome might be a combination of prioritization through VEPs together with more focused experimental approaches such as DMS studies

and expert clinical knowledge (Stein *et al.*, 2019; Anderson and Lassmann, 2018; Gelman *et al.*, 2019; Sruthi and Prakash, 2020; Lauschke and Ingelman-Sundberg, 2020).

### 1.3.3 Conclusions

Given the number of VEPs already published, it appears judicious not to carelessly contribute another contender vying with 150 methods for the attention of users. Even more so since experience shows that VEPs such as SIFT or PolyPhen-2 remain in use longer than they should, given their shortcomings compared to more recent approaches. Furthermore, I argue that when VEPs are only successful on average, understanding when and why they fail is more important for the development of the field than slight improvements in a performance measure. The more people publish new methods with volatile evaluations instead of focusing on the understanding of what is already there, the more confusing the field becomes. This thesis is therefore neither directly addressing all of the challenges highlighted in this Subsection nor presenting a new VEP. Rather, it aims to provide insights about two aspects of current variant effect prediction to help understand the results of already published VEPs and drive their development as well as support their clinical adoption.

だって私、あの人を追いかけてる私が好きなんだもの
千年女優

Chapter 2

# Human and animal disease SAVs

## 2.1 Introduction

Since the 20th century, the use of organisms beside humans for studying biology in general or disease in particular has tremendously accelerated research. For example, experiments in fruit flies, zebrafish or thale cress have shaped our understanding of developmental genetics. Mice have proven an invaluable model, especially of human disease (Müller and Grossniklaus, 2010). These so-called model organism allowed research that would have been deemed too costly, time-consuming or unethical in humans (Davis, 2004; Russell *et al.*, 2017). Recent advances including HTS (Subsection 1.1.1) make earlier uses of model organisms such as the discovery of disease genes seem superfluous (Aitman *et al.*, 2011). However, they are likely to still play a significant role in research for example to study the pathways, function and molecular mechanism underlying those disease genes. All these topics are more accessible and less complex in well-studied model organisms with less constraints on the type of experiments that can be performed.

Due to the extensive studies on model organisms as well as production and companion animals, variants and their effect are also known for those species. However, VEPs have for the longest time not made use of these resources, neither for training nor testing. There are multiple reasons for this. Data is far less accessible, especially programmatically, which means that the assembly of a dataset equates significant manual curation work. Furthermore, VEPs are often designed for assessing the effect of variants primarily in humans. Including variants from other organisms during training might increase the noise and contribute to decreased understanding of the model's behavior.

Recently, two VEPs made use of animal variant data and lead to the development of specifically trained methods for the detection of effect variants in mice and dogs (Groß *et al.*, 2018; Capriotti *et al.*, 2019a). However, the question remains to which degree traditional VEPs can be used to assess the effect of animal variants. In fact, how well can

VEPs such as SNAP2, which was trained to predict molecular effect, assess high effect disease variants in either humans or animals? Furthermore, how do disease variants from animals compare to those in human, and what happens if a variant is transferred between organisms in the same way an *in vitro* experiment might do with a model organism?

To answer these questions, we first assembled a dataset of 5,661 human disease SAVs from OMIM and manually curated another set of 117 animal disease SAVs from OMIA (Figure 2.1). Since SNAP2 contained OMIM variants in its training set, we re-trained a version of SNAP2 without any disease variants. The remaining SAVs in the training set mostly originated from PMD. Thus, the resulting VEP is focused on the prediction of molecular effect. Due to their popularity and difference in focus, SIFT and PolyPhen-2 were also evaluated. All three VEPs predicted at least 75 % of the 5,661 SAVs to have an effect, with PolyPhen-2 being the top performer at 85 %. However, this is not surprising since PolyPhen-2 is trained exclusively on human disease variants which are expected to overlap with our set of OMIM SAVs. Furthermore, a set of neutral variants showed that SNAP2 predicted fewest of them to have an effect (18 %) while PolyPhen-2 had the highest number of false positive predictions (25 % of neutral SAVs predicted as effect). SIFT and PolyPhen-2 therefore achieve their higher effect variant detection only at the cost of predicting too much effect in general.

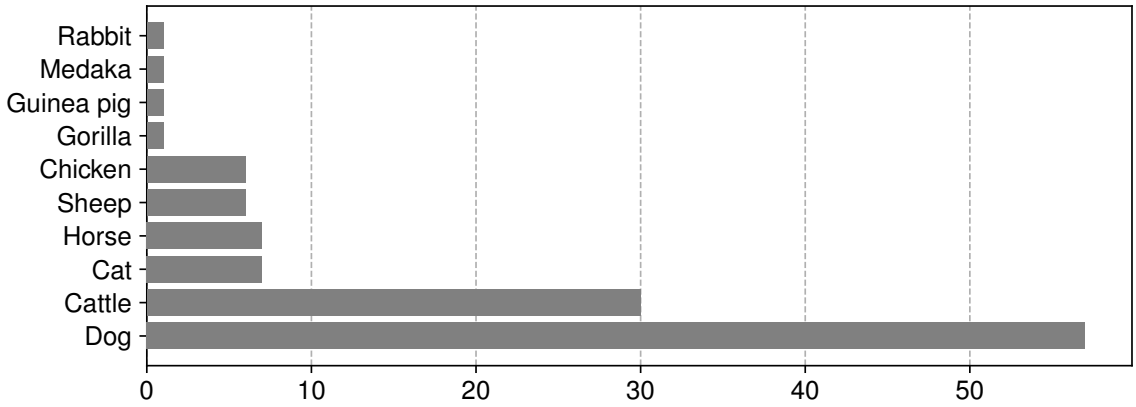



**Figure 2.1. Source organism of 117 animal disease SAVs.** SAVs were manually curated from the OMIA database, release 08/2015 (Lenffer *et al.*, 2006). Taxonomic identifiers of organisms were translated to their common name using the NCBI Taxonomy database (Federhen, 2012).

While it is common practice to observe human disease variants in model organism, VEPs are typically not applied to the same task. Using SNAP2, we evaluated the subset of 4,229 OMIM SAVs which could be mapped to homologous mouse proteins. Effect predicted for these variants was lower in the mouse proteins (78 %) compared to the original human ones (82 %). There are several possible explanations for this finding. For exam-

ple, even with a generally high sequence identity of the mouse orthologs, few changed residues could have an effect on the molecular mechanism which leads to the disease phenotype in humans. If for example the SAV is correlated with another variant, our approach of transferring only one change can lead to a different epistatic environment and thus influence the SAV's effect. It must also be noted that the difference between the two sets are too small for the only 4,229 samples to confidently infer a definite effect. As more monogenic disease SAVs are discovered in humans and animals alike, future analyses based on larger datasets might provide deeper insights into the problem.

Our analysis also provided data points on the issue of VEPs' bias towards sequence conservation (cf. Subsection 1.3.2). We assembled two sets of SAVs which described the same change than each OMIM SAV, i.e., the original and variant amino acid were the same. One set consisted only of SAVs at residues which were at least as conserved as the residue of the original disease SAV. The other set contained SAVs at less conserved positions. Effect predictions for the SAVs at conserved positions were almost as high as those for the original disease SAVs (79 %) while those at less conserved positions were predicted to have significantly less effect (50 %). Furthermore, re-training SNAP2 but excluding features related to sequence conservation, largely attenuated this difference to just 53 % and 50 % predicted effect, respectively. While this is a strong display of the importance conservation has on the predictions of SNAP2, and likely other VEPs as well, we also found that simply using a conservation threshold to perform naïve effect predictions performed worse. Thus, while challenged by conservation bias, VEPs do provide added value to the prediction task.

Overall, this analysis showed that VEPs capture the effect of disease SAVs well. Even when the method was not trained on variants with phenotypic changes but rather effect on protein function, as was the case for SNAP2. Such disease SAVs in humans and animals presented signals even stronger than all effect variants in the SNAP2 training set and were detected even without any disease variants presented during training. This is partially owed to the high sequence conservation of residues where these SAVs are found. VEPs are clearly biased towards predicting effect at such residues, but still outperform simplistic approaches based on sequence conservation thresholds. Future methods must carefully treat their feature sets to reduce the bias of homology information and will at the same time benefit from including additional types of sequence variants in the training set—for example, those from DMS experiments as discussed in Chapter 3.

## 2.2 Journal article

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Predicted Molecular Effects of Sequence Variants Link to System Level of Disease

**Jonas Reeb[1,2]*, Maximilian Hecht[1], Yannick Mahlich[1,3,4], Yana Bromberg[3,4], Burkhard Rost[1,4,5]**

**1** Department of Informatics, Bioinformatics & Computational Biology—i12, Technische Universität München, Garching/Munich, Germany, **2** TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technische Universität München, Garching, Germany, **3** Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey, United States of America, **4** Institute for Advanced Study (TUM-IAS), Garching/Munich, Germany, **5** Institute for Food and Plant Sciences WZW, Technische Universität München, Weihenstephan, Freising, Germany

* reeb@rostlab.org



OPEN ACCESS

**Citation:** Reeb J, Hecht M, Mahlich Y, Bromberg Y, Rost B (2016) Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. PLoS Comput Biol 12(8): e1005047. doi:10.1371/journal.pcbi.1005047

**Editor:** Rachel Karchin, Johns Hopkins University, UNITED STATES

**Received:** January 15, 2016

**Accepted:** July 4, 2016

**Published:** August 18, 2016



**Data Availability Statement:** Data are available from https://rostlab.org/resources/omia

**Funding:** YB was supported in part by an Informatics Research Starter grant from the PhRMA foundation and by NIH/NIGMS grant U01GM115486. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

Developments in experimental and computational biology are advancing our understanding of how protein sequence variation impacts molecular protein function. However, the leap from the micro level of molecular function to the macro level of the whole organism, *e.g.* disease, remains barred. Here, we present new results emphasizing earlier work that suggested some links from molecular function to disease. We focused on non-synonymous single nucleotide variants, also referred to as single amino acid variants (SAVs). Building upon OMIA (Online Mendelian Inheritance in Animals), we introduced a curated set of 117 disease-causing SAVs in animals. Methods optimized to capture effects upon molecular function often correctly predict human (OMIM) and animal (OMIA) Mendelian disease-causing variants. We also predicted effects of human disease-causing variants in the mouse model, *i.e.* we put OMIM SAVs into mouse orthologs. Overall, fewer variants were predicted with effect in the model organism than in the original organism. Our results, along with other recent studies, demonstrate that predictions of molecular effects capture some important aspects of disease. Thus, *in silico* methods focusing on the micro level of molecular function can help to understand the macro system level of disease.

## Author Summary

The variations in the genetic sequence between individuals affect the gene-product, *i.e.* the protein differently. Some variants have no measurable effect (are neutral), while others affect protein function. Some of those effects are so severe they cause so called monogenic Mendelian diseases, *i.e.* diseases triggered by a single letter change. Some *in silico* methods predict the molecular impact of sequence variation. However, both experimental and computational analyses struggle to generalize from the effect upon molecular protein function to the effect upon the organism such as a disease. Here, we confirmed that methods predicting molecular effects correctly capture the type of effects causing Mendelian

diseases in human and introduced a data set for animal diseases that was also captured by predictions methods. Predicted effects were less when *in silico* testing human variants in an animal model (here mouse). This is important to know because "mouse models" are common to study human diseases. Overall, we provided some evidence for a link between the molecular level and some type of disease.

## Introduction

Protein sequences span three orders of magnitude in their lengths (30-30k residues). Aspects of molecular function are often captured by 'sub-units', *e.g.* by domains or domain-like fragments [1,2] that are, on average, about 100 residues long [3,4]. The variation of a single amino acid (SAV) can change the function of a multi-domain protein and many changes in molecular function lead to disease. In fact, OMIM, the database of Online Mendelian Inheritance in Man [5], archives thousands of SAVs that cause Mendelian diseases. On the other hand, databases such as the Protein Mutant Database (PMD) catalogue tens of thousands SAVs altering molecular function; many of those have not been observed to cause a phenotype on the level of the organism. Sequencing everyone on this globe, will we observe almost all possible SAVs? The answer remains subject for speculation. Obvious exceptions include embryonically lethal variants and not all variants will occur in germ lines.

Deep mutational scanning studies that change every residue in a protein to all non-native amino acids suggest a conundrum: for almost every position (each residue) both neutral and effect SAVs exist [6–8], *i.e.* most residue positions are at the same time sensitive and robust to variants. A variety of computational methods predict the effect of SAVs. Although most methods have many goals, we can simplify by distinguishing methods that focus more on predicting the effect of SAVs upon (Mendelian) disease [9–15] and upon molecular function or structure [16–20]. *In silico* methods focusing on molecular function [21,22] correlate more with experimental deep mutational scans than those focusing on disease [8,23].

The "micro" perspective of molecular function is often probed through *in vitro* assays of proteins or cells, while *in vivo* screens often focus on observing the "macro" level through the impact upon the entire organism or system, *e.g.* in form of a disease phenotype. Molecular impact does not directly correspond to system impact, *i.e.* functional effects of variants usually do not directly explain diseases. Relating the two levels of variant effects is of utmost importance, for example to understand diseases and to develop treatments. Successful drugs often mechanistically bridge this gap: the molecular agent (drug) affects the organism/system (disease).

Here, we show a few links that suggest how molecular effect predictions can capture some aspects of diseases. Our findings are largely based on a manually curated set of variants (SAVs) from OMIA (Online Mendelian Inheritance in Animals), a database cataloging expert curated monogenic diseases in animals and their relevant variants [24]. Methods focusing on the molecular impact of variants predict disease-causing variants in animals and human (taken from OMIM [5]). We also addressed the question how prediction methods behave for model systems, *e.g.* by predicting variants in mice to study human diseases. The latter analysis might be particularly relevant in light of a recent discussion about the validity of using mouse models [25,26].

## Results and Discussion

### OMIM variants predicted to have strong effect

SIFT [27] predicts the impact of variants upon molecular protein function by assessing the disruption of conserved residues. SNAP [17] predicts this impact by considering

evolutionary, functional and structural features. Our newer method SNAP2 [16] also trained on disease-causing variants. To avoid the overlap of variant sets used for SNAP2 training and those used in this work, we trained a SNAP2 version, using only variants with impact upon molecular function, *i.e.* leaving out all human disease variants from OMIM or HumVar [28] but keeping the variants from PMD. PolyPhen-2 also uses evolutionary and structural features to predict the effect of disease-causing mutations in human [12]. We predicted the effect of disease-causing SAVs from OMIM through PolyPhen-2, SIFT and the re-trained version of SNAP2 (not using disease variants). All three methods predicted very strong functional effects (Fig 1A). PolyPhen-2 predicted the highest fraction (85%) of the OMIM SAVs to have effects, followed by SNAP2 (78%) and SIFT (76%). Monitoring effect predictions for a set of neutral SAVs (TrNeutral), showed that both PolyPhen-2 and SIFT reached higher effect fractions at the expense of more false positives (TrNeutral bars higher): the differences OMIM-TrNeutral were the same between SNAP2 and PolyPhen-2 (60%). Another crucial difference was that the numbers for SNAP2 were derived without using the data used for training, while the results for PolyPhen-2 overlapped substantially with the training data used for that method. Machine learning methods usually perform better on the training than on the testing data. For instance, the SNAP2 version trained with OMIM reached 80% effect predictions for OMIM as opposed to 78% for the version not trained on OMIM.

Another crucial aspect was that SNAP2 predicted its training set of effect SAVs less well than the OMIM SAVs (Fig 1A: TrEffect 75% vs. OMIM 78%). For us, this was the most outstanding example for a new data set outperforming the training set in 23 years of machine learning in biology [29]. The label "disease" seemingly generates more consistent data than experimental measurements of functional disruption.

Previous analyses showed the strength of the molecular effect to correlate with the SNAP score: higher SNAP scores indicate more reliable predictions and stronger effects [17,30]. This implies that *in silico* predictions can accurately sort thousands of variants relevant for some investigation by their likely molecular impact without the need to provide any additional annotations. Thus, the high amount of SNAP2 effect predictions for OMIM variants (Fig 1A: OMIM higher than for TrEffect) suggested very strong effects upon molecular function. For variants associated with Mendelian disease, this result was expected.

## Manually curated OMIA data set

OMIA [24], the database for Online Mendelian Inheritance in Animals, collects expert annotations for monogenic diseases in animals. Mouse and rat data are excluded, as those variants and annotations are available through the specialized databases RGD [31] and MGD [32]. Unfortunately, none of those resources readily provided the data needed for our analysis. Very few of the, *e.g.* 600 variants with known disease associations in OMIA, which range from large structural variants to single nucleotide variants and SAVs, were in a machine-readable standard format such as "sequence variant XpositionY causes effect". Moreover, the protein sequences referenced by the variants remained obfuscated. Several person-months got us from OMIA to a set of just 117 single disease associated variants with matching sequences (Methods, S2 and S3 Tables). Incidentally, we note that OMIA's value to the genomics, proteomics and health-related research communities might significantly increase if their high-quality manually curated data were readily available to automated analyses across the spectrum of gene- and protein-science. For similar database-related reasons and time constraints, mouse and rat variants could not be included in this analysis. As an additional complication, studies in mouse and rat typically focus on whole gene knockouts rather than on effects of SAVs.

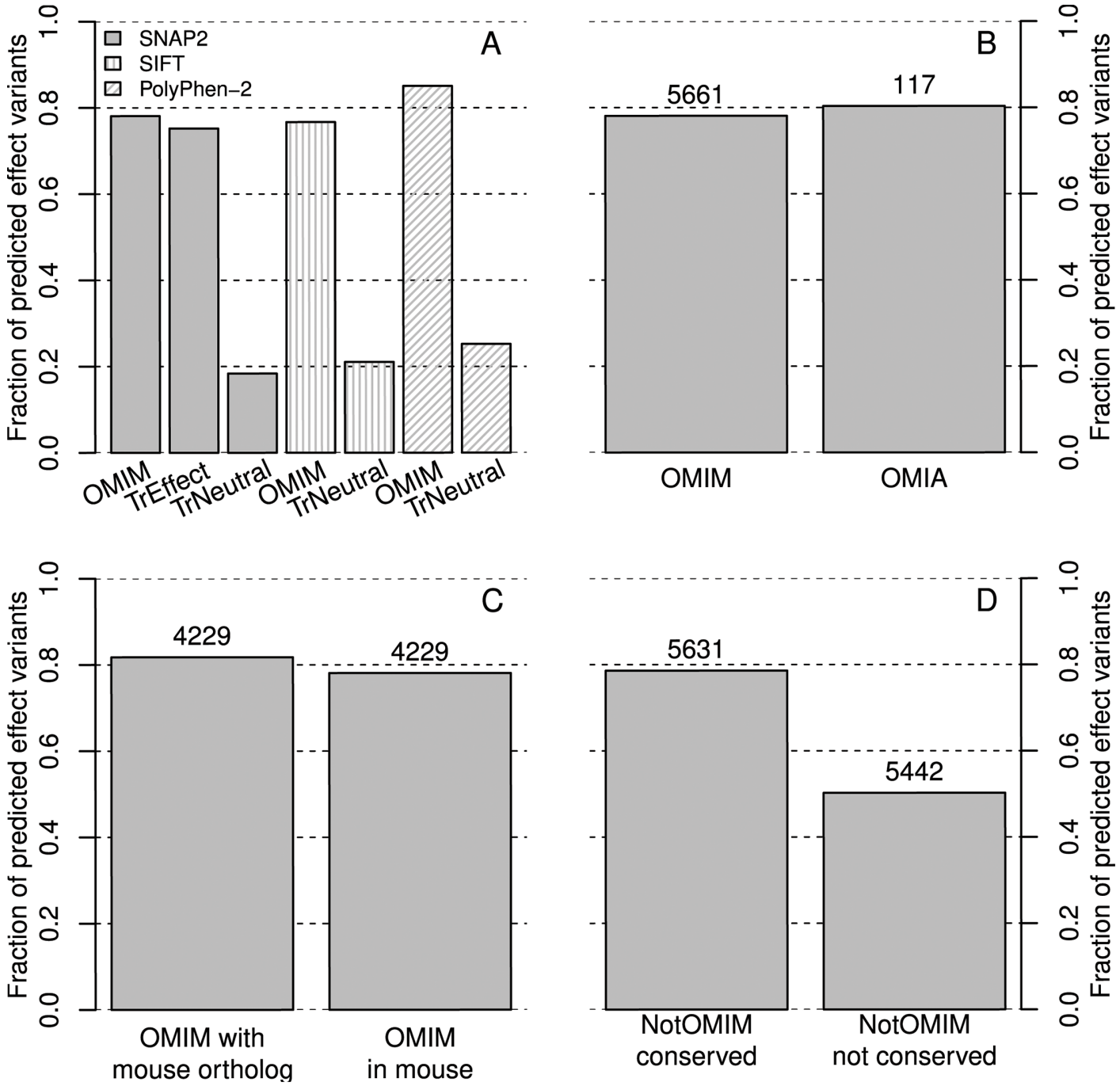



**Fig 1. Predictions of SAV effects upon function and disease across species.** The numbers above bars give the number of SAVs in the set. **A**: Three methods (SNAP2 [16], SIFT [27], PolyPhen-2 [12]) predicted SAV effects upon molecular function (TrEffect/TrNeutral) and upon disease (OMIM). Exclusively for this panel SNAP2 was trained without using disease SAVs from OMIM [5] or HumVar [28]. The SNAP2 version trained exclusively on molecular function clearly captured aspects of OMIM-disease SAVs (leftmost bar OMIM higher than 2nd to the left TrEffect). TrNeutral was the SNAP2 training set of variants without effect. Comparing the bars for TrNeutral and OMIM for each method pointed to differential thresholds: Polyphen-2 correctly predicted more effect in OMIM than SNAP2 but also incorrectly predicted more effect in the neutral data, *i.e*. simply predicted more effect variants. **B:** OMIM is repeated from A. SNAP2 captured disease signals in humans and animals at similar levels. OMIA contained disease SAVs from animals other than mouse and rat (mostly dog and cattle). **C:** SNAP2 predicted OMIM SAVs with less effect in mouse orthologs than in human. Left bar (*OMIM with mouse ortholog*): SNAP2 predictions for the subset of all 4,229 OMIM SAVs for which we found a mouse ortholog. Right bar (*OMIM in mouse*): SNAP2 predictions when putting the human SAV into the mouse sequence. **D:** Disease variants happen in non-random positions. Left bar (*NotOMIM conserved*): in each protein with an OMIM SAV, we predicted the effect of all SAVs with a level of sequence conservation $\geq$ that of the OMIM variant. Right bar (*NotOMIM not conserved*): predictions for SAVs in non-OMIM positions with conservation < that of the OMIM SAV. Obviously, OMIM SAVs were very well conserved.

doi:10.1371/journal.pcbi.1005047.g001

## Slightly more effect for OMIA than for OMIM variants

All methods optimized to predict disease causes, for obvious reasons of data availability and clinical relevance, focus on human variants. In contrast, methods such as SIFT and SNAP2 perform at similar levels for other organisms. Here, we applied SNAP2 to our curated set of OMIA variants (SAVs). Although this data set was small, it was particularly interesting for testing, because those variants had not been available for the training of methods before.

SNAP2 predicted more OMIA variants with effects than in the SNAP2-effect training set (Fig 1A TrEffect 75% vs. Fig 1B OMIA 80%). Additionally, OMIA variants were predicted with slightly higher effect than those from OMIM (Fig 1B: OMIM 78% vs. OMIA 80%). This result suggested Mendelian disease-SAVs to have stronger effect in animals than in human. The simple asymmetry in what is considered a disease in animals and human might explain this observation. For example, non-lethal abnormalities such as variation in hair-growth might be perceived as a human disease, while the equivalent may not be an animal disease worth noting. In fact, the "disease-ness" of hair/fur length differences actually depends on the animal in question; *e.g.* the furs of dogs differ between breeds (an intended result of breeding). OMIA is therefore likely to focus on more lethal variants than OMIM and SNAP2 predictions simply mirror this expectation.

## Disease-variants affect the carrier more than other species

When experimental biology builds an animal model for a human disease, disease-causing human variants are introduced into the animal. Can *in silico* methods achieve the same? We took the mutations (SAVs) from OMIM and predicted the effect of the same variant in the mouse homolog (Fig 1C). The disease-causing SAVs from human were predicted with slightly less effect in the mouse model (Fig 1C: left bar higher than right). We might rationalize this observation by arguing that the OMIM SAV has been observed because it had such a strong effect, slight alterations to the sequence might reduce the signal. Although we have some additional evidence supporting this view (S1 Fig), it remains very speculative. OMIM SAVs are by no means random mutations and in 95% of the cases with OMIM SAVs, the amino acid was the same in human and mouse (not unexpected, given the results presented in the next paragraph). Whatever the cause, this effect should be taken into account when creating animal models for human diseases.

## Position of variant more important than its type

We know that the positions of OMIM variants are not random. *In silico*, we can easily introduce OMIM-like variants elsewhere in the protein. For each OMIM variant (XnY, *i.e.* amino acid X at residue n mutated to amino acid Y), we have to find another position ($m \neq n$) and *in silico* vary XmY. Then we compare the predicted effect XnY to those predicted for XmY. As we suspect that OMIM SAVs tend to be more conserved within the evolution of protein families than randomly chosen positions in the same protein, we can additionally constrain our analysis by postulating that we find positions *m* such that the conservation of $m \geq$ that for *n* (Fig 1D: *NotOMIM conserved*). We can contrast this to a sampling in which we predict the effect for less well-conserved positions (m conserved $<$ n, Fig 1D: *NotOMIM not conserved*). This seemingly simple scheme opens another complication: we could additionally choose variants of the native amino acid against all other 19 non-native ones (19-non native), or we could restrict our variants to the subset of those variants that are reachable by a single nucleotide variation (SNV-possible). For simplicity, we only reported results for the SNV-possible version of randomly chosen variants. We observed that a randomly chosen SNV-possible amino acid variant at each OMIM position was predicted with slightly lower effect than the original OMIM SAV (S1

Fig: *OMIM_rand* vs. *OMIM*). More importantly, our results confirmed the expected importance of residue conservation: SNAP2 predicted almost the same effect for the OMIM variant as for NotOMIM SAVs of similar conservation (Fig 1B *OMIM* vs. Fig 1D *NotOMIM conserved*). Conversely, replacing the disease variant XnY at all positions m with less conservation (XmY) was predicted with substantially lower effect (Fig 1D: *NotOMIM conserved* vs. *NotOMIM not conserved*). Interestingly, random SNV-possible variants at OMIM or NotOMIM conserved positions were predicted with an equal number of effect variants (S1 Fig).

We further applied a version of SNAP2 that did not use conservation (*i.e.* alignments) as input but was otherwise trained as the default version. This alignment-free version predicted the same trend, but with significantly reduced difference between predicted effect at OMIM and NotOMIM positions (S2 and S3 Figs). Repeating the above analyses for the OMIA set produced similar results (S4–S7 Figs).

The strong dependence of results on conservation suggested that predicting disease-causing variants would only require the definition of a single threshold, *i.e.* predict variant as disease if the conservation at its position is above an empirically chosen value. However, we sampled a different conservation threshold for each protein by picking the level of conservation equal to or higher than that observed for each OMIM/OMIA variant. Accordingly, a simple method that predicts every SNV-possible SAV at positions above a single conservation threshold as having an effect, would over-predict effect substantially (S1 Fig, S4 and S6 Figs, S8 Fig).

## Variants with known experimental observations might be biased

SIFT and SNAP2 were optimized on molecular effect variants, PolyPhen-2 [12] on disease variants. Nevertheless, the three agreed on 68% of the variants with known experimental molecular effects [16]. In predicting the effect on molecular function, SNAP2 performed best for difficult variants [16], *i.e.* those that were predicted differently by two methods (as effect by one, as neutral by the other). Most relevant and available experimental results have been used for method development. Do computational methods inherit a bias from the experimental data?

We can address the question about bias in the experimental data through comprehensive *in silico* mutagenesis [33], *i.e.* by predicting the effect of all possible SAVs; such studies are also referred to as the complete *mutability landscape* [21]. There are two approaches for such a complete mutagenesis: 19 non-native SAVs (large-scale *in silico* mutagenesis), or SNV-possible SAVs. The second approach produces a subset of the first with different statistical features [30]. The first solution furthers our understanding of protein function in the context of its mutability landscape; the second simulates the types of changes that can happen in evolution.

Methods differ in their predictions for experimentally annotated SAVs, as well as for *in silico* assays of complete mutagenesis (19-non native SAVs). For instance, SIFT and SNAP2 predictions differ more for all possible SAVs in human than for variants with effect on molecular function from PMD (S1 Table). A similar difference is implied between SIFT and PolyPhen-2 [34]. Although the differences amount to "just" 3–8 percentage points, they imply prediction differences for millions of variants. Why do the predictions of the two methods agree more for experimental annotations than for all possible variants?

Assume that the existing methods converged toward the same solution for known data due to the lack of diversity in the training data, *i.e.* the same data enforces the same lesson. Put differently, the experimental data focuses on some particular type of effect (that might be easier to predict than the types that remain unknown). This assumption would explain our findings but it seems incorrect. Firstly, methods have not used the exact same type of data: some focus on molecular function, others on disease-causing variants. Secondly, prediction agreement between methods is not higher for strong-impact, disease-causing variants from OMIM than

for the neutral and molecular function effect variants from PMD, although stronger variants are predicted better [17,30]. Thirdly, additional recent tests confirm the important differences in predictions for larger data sets, where methods tend to agree more for some observed human variants and less so for others. Thus, the agreement between methods for experimentally annotated data sets is not explained by the assumption that they learned the same from the restricted data.

Could it be that we already have an experimental record for most effect variants? If true, the observed method correlation would be explained. For OMIM, this completeness assumption might not be too far from the truth: It has been argued that through recent advances in deep sequencing the majority of disease-causing variants, in particular in coding regions which are tractable through whole exome sequencing, have already been observed and many are to follow in the near future [35]. However, large-scale *in silico* mutagenesis strongly suggests that many effect variants remain experimentally uncharacterized. If true, the method agreement for experimental annotations would not be explained.

Alternatively, differences between *in silico* mutagenesis predictions and experimental annotations might originate from the bias in the experimental data. Many reasons would explain such a bias. Firstly, the *in vitro* assays may not capture all interactions and constraints under which proteins exist *in vivo*. Secondly, the experimental thresholds for the degree of functional impact (*e.g.* change in ΔΔG of binding) required to report a variant as "effect" or "neutral" are subjective. Computational methods will likely zoom into the most consistent data, *i.e.* the strongest or simplest effects. Bias might also be introduced by the difficulty in relating the molecular to the system level, *e.g.* not every variant that has a high effect on molecular function challenges the organism. Conversely, not every disease is caused by a single SAV. On the contrary, most diseases are likely caused by much more complex mechanisms than single variants. For example, in cancer many variants may affect molecular function; some of these "drive" the cancerous growth, others simply piggyback (passenger mutations). The two have very different biological traits and can be distinguished *in silico* [36]. Nevertheless, the gain from molecular functional effect predictions for describing odds in prognosis is still limited [37].

Finally, the methods' high agreement might originate from the codon usage. While there is no comprehensive explanation that convincingly maps the codon usage to the biophysical features of the encoded amino acids, there are some preferences built into one of the three bases [38]. SNV-possible variants might therefore tend to alter the biophysical features of an amino acid less than other substitutions. Methods such as SNAP2 are trained to consider variants that maintain the biophysical environment of a residue to be more neutral than others. Hence, SNV-possible might be predicted as more neutral than amino acid substitutions that required more than one nucleotide change. However, since most experimental annotations report effect SAVs, the codon usage correlations are unlikely to help explain the agreement.

## Capturing phenotype effects through molecular function predictions?

In order to bridge the gap from effect upon single protein to effect upon organism, we clearly also have to consider the interaction context of a protein. For instance, predicted effects upon molecular function are much more likely to imply effects upon the organism if the protein is a key player in a crucial pathway than if the protein is "just" a structural protein. Indeed, OMIM SAVs may be so damaging because they preferentially hit crucial proteins. OMIM SAVs constitute one link between molecular effect and disease, albeit possibly an exceptional one. PolyPhen-2 and SNAP2 trained on such disease-effects. The fact that they predict those very well, therefore, is not very meaningful. However, when we retrained a version of SNAP2 without

any disease- or system-level related SAVs, we could still predict OMIM SAVs very well (Fig 1). Thus, we established one link between molecular and organism effect.

How could we bridge the gap from the molecular level to that of the organism more efficiently for a larger set of SAVs? As already mentioned: we might succeed by including more relevant knowledge related to interactions. However, success toward this end remains incomplete for the time being. Alternatively, we might consider the integration of gene prioritization tools. These integrate additional orthogonal data such as expression patterns, subcellular localization, information from literature or otherwise manually curated annotations [39,40]. For example, recent work has seen the development of a model to distinguish loss-of-function genes in human, based on conservation and protein interaction data [41]. This however is based on variants that lead to a complete loss of the transcript and therefore not comparable to the SAV effect prediction by SNAP2.

Another idea is to move from the level of SAVs to that of correlated variants [8,23]. This remains challenging: no method can yet predict the effect for all possible pairs of SAVs in all human proteins. However, even for the proteins for which some methods can achieve this: such a refinement might contribute much toward increasing the agreement between computational and experimental deep mutagenesis studies. However, it might contribute little for better bridging the micro and macro level.

### Conclusion

We have presented evidence that methods optimized for predicting the effects of SAVs upon molecular function, such as SNAP2, capture the type of strong effect that leads to monogenic diseases. This was sustained even when excluding disease-causing SAVs from training. Possibly, OMIM-like means "effect upon molecular function strong enough to not have to consider anything else". We also showed that Mendelian disease-causing SAVs in animals from OMIA (mostly dog and cattle) were predicted even more successfully than those from OMIM. Both these results (OMIM higher than training data although not used, OMIA even higher) imply that methods not focused on phenotype level effects, can capture the strong underlying functional effect signal. OMIM-like SAVs often hit the most conserved position, but a trivial prediction solely based on this conservation fell much behind the level of performance reached by methods such as SNAP2 or PolyPhen-2. Generally, computational and experimental analyses of molecular effects of SAVs cannot explain the effects upon the organism. The integration of gene prioritization and the incorporation of additional data from interactions might contribute to bridging this gap.

## Materials and Methods

### Collecting OMIA variants

We annotated sequence variants in animals using the SQL dump of OMIA (release 08/2015) [24]. Gene symbols and the text from the section *Molecular basis* were extracted for all diseases (i) considered as defect by OMIA and (ii) with the causal variant known. We then read the text and publications to extract variant annotations in the standard format of, *e.g.* A11W: native alanine (A) at residue position 11 mutated to tryptophan (W). OMIA already contained 82 variants in this format possibly enabling automated extraction through a regular expression. However, at least one of the 82 was outdated; this fact was mentioned in the description, but would have been missed by automation. Our effort yielded another 96 variants. Thus, we could use 178 OMIA variants in total. Next, we retrieved the protein sequences of the OMIA variants by querying UniProtKB (release 2015_08) with the gene symbol and NCBI taxonomy identifier extracted from OMIA. When we had multiple matches, we chose the top match. Among the

178 variants, three synonymous variants were excluded. Of the remaining 175, 12 had to be excluded because the above protocol did not yield a sequence. In 46 cases a sequence could be retrieved but the amino acid found at the position denoted by OMIA was not the one found in the sequence at that position, *e.g.* for OMIA variant A11W, the amino acid at position 11 in the sequence was not alanine (A). In 110 cases the amino acid was found as expected and in seven additional cases shifting the position by +1 yielded the expected sequence. The "+1" accounts for sequences stored without the initiator methionine. Our final data set of 117 variants from 99 sequences (S2 Table) is available at https://rostlab.org/resources/omia. The attrition rate leading to the 117 mutations is summarized again in S3 Table. Most of the variants in the final dataset were from dogs (39%) and cattle (21%). These ratios were comparable to those for original 178 variants (44% and 21%). We annotated another 12 positions with single amino acid deletions and 48 variants leading to premature stop codons. However, since SNAP2 only predicts effect for changes of amino acids not their removal or premature stop of the amino acid sequence, these were not used in the further analysis.

## OMIM, SNPdbe, and PMD

We extracted 5,661 OMIM [5] variants with sequences from SNPdbe [42]. SNAP2 [16] was trained on SAVs from PMD, the Protein Mutation Database [43] as well as human disease variants from OMIM and HumVar [5,28]. For the sets shown in Fig 1A, we trained a version of SNAP2 on only molecular effect variants, *i.e.* without variants from OMIM or HumVar, and show cross-validation results for that (TrEffect and TrNeutral). In all other cases, the training set of SNAP2 also included disease variants [16].

## Ortholog mapping for OMIM variants to mouse

Human homologs of the animal genes from OMIM were retrieved using the Biomart interface [44] of Ensembl Genes 82 (release 09/2015) [45]. 271 sequences from the OMIM mutation set were removed because they were not found in the Ensembl set. The remaining 1,293 sequence pairs were aligned using the global alignment implemented in BioPython's *globalds* with BLOSUM62 as substitution matrix, gap open -10 and gap extend -0.5 [46]. Variants at positions with insertions (aligned against a gap) were removed. After transferring the variants from the human to the mouse sequence, some variants implied no change because for the human X2Y variant, the mouse had Y as its native amino acid, *i.e.* the "variant" in mouse would have been a synonymous Y2Y. Removing all such cases and their respective variant in human, the final set comprised 4,229 variants (of the original 5,661 OMIM variants) in both human and the mouse homologs, *i.e.* the "*in silico* humanized mouse model" (denoted as "OMIM in mouse" in Fig 1).

## Prediction methods

For all variants, effects were predicted by SIFT [27,47], PolyPhen-2 [12] and SNAP2 [16]. We used SNAP2 with the parameter *tolerate*, that performs predictions even if underlying methods fail, to obtain results for all variants. For some analyses (S2 and S3 Figs, S5 and S7 Figs), we used SNAP2 without alignments as input, by using the *skip* parameter. SIFT predictions were obtained locally with version 4.0.3b [47]. PolyPhen-2 predictions were obtained locally using version 2.2.2 [12]. All three methods used a BLAST database created by merging PDB and UniProtKB (release 2015_08), followed by a redundancy reduction at 80% sequence identity with CD-HIT [48,49]. We used the default cutoffs of each method to obtain binary predictions into either effect or neutral for every variant.

## Statistics

The background effects for the OMIM data (Figs 1D and S1 and S8) were estimated as follows: At every disease variant position, we mutated to either (i) the amino acid denoted in the disease SAV (*OMIM*, Figs 1D and S8) or (ii) considered one randomly out of the SNV-possible variants, *i.e.* mutations to amino acids that could occur by a single nucleotide change (*OMIM_rand*, S1 Fig). This simplification was imposed by the incompleteness in the knowledge of the underlying DNA sequences. We assume that our hack approximation to "all SNV-possible" provides a sufficiently accurate approximation.

For the non-disease positions, we sampled a random set of positions without known disease variants from the same proteins (*NotOMIM*). Non-disease positions were never sampled from the first and last 10 residues of a sequence, since SNAP2 uses an input window size of 21. The predicted effect at the NotOMIM positions was evaluated as before. (i) Either given an OMIM mutation such as I10L, we randomly picked a non-disease position with isoleucine and mutated it to leucine (*NotOMIM*, Figs 1D and S8). (ii) Alternatively, we chose a random SNV-possible variant from non-disease positions (*NotOMIM_rand*, S1 Fig).

For the conserved non-disease positions (*NotOMIM conserved*) we considered only non-disease positions that were at least as conserved as the known disease position. For instance, assume a protein P contains two disease variants X25Y and A100B. Randomly choose one out of all positions other than 25 and 100 in P that is at least as conserved as position 25. Then do the same for position 100 and all other variants in other proteins. Skip, if the disease position is the one most conserved in that protein and there is no other position with an equally high conservation. For the not conserved positions, we accordingly used all positions with conservation lower than that of the OMIM SAV. Conservation was measured through the *information per position* value from PSI-BLAST PSSMs created by querying the OMIM sequences against the 80% redundancy reduced database of UniProtKB and PDB mentioned in the previous section. At each NotOMIM conserved or not conserved position, effects were predicted as outlined above for cases i (*NotOMIM (not) conserved*, Figs 1D and S8) and ii (*NotOMIM_rand (not) conserved*, S1 Fig).

The same was repeated using SNAP2 without alignments as input (S2 and S3 Figs). We also show results for the full set of variants, *e.g.* "all @ NotOMIM_rand not conserved" are all SNV-possible mutations at all non-disease positions that are less conserved than the position of the original OMIM SAV. "all @ NOT-OMIM conserved" are all OMIM SAVs at all eligible non-disease positions (S1 and S8 Figs). All analyses were also performed on the OMIA set (S4–S7 Figs).

## Supporting Information

**S1 Fig. SNAP2 predictions towards random SNV-possible variants at different positions in the OMIM set.** Analogous to Fig 1D of the main paper but mutating positions to random SNV-possible variants instead of using the OMIM SAV. "OMIM" is repeated from Fig 1A as reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in the respective set, instead of a random sample.
(TIF)

**S2 Fig. SNAP2 predictions without alignment input at different positions in the OMIM set.** Analogous to Fig 1D of the main paper but using SNAP2 without alignments input. "OMIM using alignments" is repeated from Fig 1A as a reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in

the respective set, instead of a random sample.
(TIF)

**S3 Fig. SNAP2 predictions without alignment input and towards random SNV-possible variants at different positions in the OMIM set.** Analogous to Fig 1D of the main paper but mutating positions to random SNV-possible amino acids instead of using the OMIM SAV. Additionally, SNAP2 is used without alignment input. "OMIM using alignments" is repeated from Fig 1A as a reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in the respective set, instead of a random sample.
(TIF)

**S4 Fig. SNAP2 predictions at different positions in the OMIA set.** Analogous to Fig 1D of the main paper but on the OMIA set. "OMIA" is repeated from Fig 1B as a reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in the respective set, instead of a random sample.
(TIF)

**S5 Fig. SNAP2 predictions without alignment input at different positions in the OMIA set.** Analogous to Fig 1D of the main paper but using SNAP2 without alignments input and on the OMIA set. "OMIA using alignments" is repeated from Fig 1B as a reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in the respective set, instead of a random sample.
(TIF)

**S6 Fig. SNAP2 predictions towards random SNV-possible variants at different positions in the OMIA set.** Analogous to Fig 1D of the main paper but using OMIA and mutating positions to random SNV-possible variants instead of using the OMIA SAV. "OMIA" is repeated from Fig 1B as reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in the respective set, instead of a random sample.
(TIF)

**S7 Fig. SNAP2 predictions without alignment input and towards random SNV-possible variants at different positions in the OMIA set.** Analogous to Fig 1D of the main paper but using OMIA and mutating positions to random SNV-possible amino acids instead of using the OMIA SAV. Additionally, SNAP2 is used without alignment input. "OMIA using alignments" is repeated from Fig 1B as a reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in the respective set, instead of a random sample.
(TIF)

**S8 Fig. SNAP2 predictions at different positions in the OMIM set.** Analogous to Fig 1D of the main paper. "OMIM" is repeated from Fig 1A as a reference. The numbers above bars give the number of SAVs in the set. Sets prefixed with "all @" contain all possible mutations in the respective set, instead of a random sample.
(TIF)

**S1 Table. Pairwise agreement of effect prediction.** Shown is the percentage of entries in the respective dataset for which the two given methods agree in binary prediction, *i.e.* both predict a neutral or effect variation.
(DOC)

**S2 Table. The set of 117 OMIA mutations.** The 117 mutation extracted by manual review from the OMIA database. Shown are only entries for which a sequence could be found and the mutation mapped onto the sequence (*cf.* S3 Table). All diseases are considered a defect by OMIA annotation. Organism shows the NCBI taxonomy id. Variants marked with *, are those where the position was shifted one forward (Methods, S3 Table). The full set including the sequences is also available at rostlab.org/resources/omia.
(DOC)

**S3 Table. Attrition rate of OMIA annotations.** AA deletion describes cases where a single amino acid is deleted without affecting the reading frame. Nonsense are mutations to a premature stop codon. These two cases were extracted from OMIA but not used in the analysis. For the amino acid substitution set *No seq*. describes that no sequence was found for the given combination of taxonomy id and gene id (Methods). *No match* describes that a sequence was found but the amino acid at the position given by OMIA was not the one expected from the annotated mutation. *Match* are all cases where this was the case, and *Match+1* were the amino acid fit after shifting one position to the right. Highlighted in green are the cases forming the final set of 117 mutations used for the analysis.
(DOC)

## Acknowledgments

Thanks to Tim Karl, Laszlo Kajan, and Guy Yachdav (TUM) for invaluable help with hardware and software; to Inga Weise (TUM) for support with many other aspects of this work; to Janet Kelso (MPI Leipzig) for helpful comments. We are also grateful to the three anonymous reviewers for their important help. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

## Author Contributions

Conceived and designed the experiments: JR MH YB BR. Performed the experiments: JR MH. Analyzed the data: JR MH YB. Contributed reagents/materials/analysis tools: MH YM BR. Wrote the paper: JR MH YB BR.

## References

1. Lees JG, Ranea JA, Orengo CA (2015) Identifying and characterising key alternative splicing events in Drosophila development. BMC Genomics 16: 608. doi: 10.1186/s12864-015-1674-2 PMID: 26275604
2. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, et al. (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Research 43: D376–381. doi: 10.1093/nar/gku947 PMID: 25348408
3. Liu J, Rost B (2004) CHOP proteins into structural domain-like fragments. Proteins: Structure, Function, and Bioinformatics 55: 678–688. doi: 10.1002/prot.20095 PMID: 15103630
4. Liu J, Rost B (2003) Domains, motifs, and clusters in the protein universe. Current Opinion in Chemical Biology 7: 5–11. PMID: 12547420
5. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 33: D514–517. doi: 10.1093/nar/gki033 PMID: 15608251
6. Fowler DM, Stephany JJ, Fields S (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning. Nature Protocols 9: 2267–2284. doi: 10.1038/nprot.2014.153 PMID: 25167058
7. Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. Nature Methods 11: 801–807. doi: 10.1038/nmeth.3027 PMID: 25075907
8. Hopf TA, Ingraham JB, Poelwijk FJ, Springer M, Sander C, et al. (2015) Quantification of the effect of mutations using a global probability model of natural sequence variation. ArXiv e-prints.

9. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics 46: 310–315. doi: 10.1038/ng.2892 PMID: 24487276

10. Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M (2012) PON-P: Integrated predictor for pathogenicity of missense variants. Human Mutation 33: 1166–1174. doi: 10.1002/humu.22102 PMID: 22505138

11. Thusberg J, Vihinen M (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Human Mutation 30: 703–714. doi: 10.1002/humu.20938 PMID: 19267389

12. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nature Methods 7: 248–249. doi: 10.1038/nmeth0410-248 PMID: 20354512

13. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Research 39: e118. doi: 10.1093/nar/gkr407 PMID: 21727090

14. Venselaar H, Camilli F, Gholizadeh S, Snelleman M, Brunner HG, et al. (2013) Status quo of annotation of human disease variants. BMC Bioinformatics 14. doi: 10.1186/1471-2105-14-352

15. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013) Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3: S3–S3. doi: 10.1186/1471-2164-14-S3-S3 PMID: 23819870

16. Hecht M, Bromberg Y, Rost B (2015) Better prediction of functional effects for sequence variants. BMC Genomics 16 Suppl 8: S1. doi: 10.1186/1471-2164-16-S8-S1 PMID: 26110438

17. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research 35: 3823–3835. doi: 10.1093/nar/gkm238 PMID: 17526529

18. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics 12: 151. doi: 10.1186/1471-2105-12-151 PMID: 21569468

19. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research 33: W306–310. doi: 10.1093/nar/gki375 PMID: 15980478

20. Topham CM, Srinivasan N, Blundell TL (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Engineering, Design & Selection 10: 7–21.

21. Hecht M, Bromberg Y, Rost B (2013) News from the protein mutability landscape. Journal of Molecular Biology 425: 3937–3948. doi: 10.1016/j.jmb.2013.07.028 PMID: 23896297

22. Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Engineering, Design & Selection 22: 553–560. doi: 10.1093/protein/gzp030 PMID: 19561092

23. Hopf TA (2015) Phenotype prediction from evolutionary sequence covariation. Munich: TUM.

24. Lenffer J, Nicholas FW, Castle K, Rao A, Gregory S, et al. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. Nucleic Acids Research 34: D599–601. doi: 10.1093/nar/gkj152 PMID: 16381939

25. Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, et al. (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. Proceedings of the National Academy of Sciences 110: 3507–3512. doi: 10.1073/pnas.1222878110 PMID: 23401516

26. Takao K, Miyakawa T (2014) Genomic responses in mouse models greatly mimic human inflammatory diseases. Proceedings of the National Academy of Sciences 112: 1401965111. doi: 10.1073/pnas.1401965111 PMID: 25092317

27. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research 31: 3812–3814. PMID: 12824425

28. Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22: 2729–2734. doi: 10.1093/bioinformatics/btl423 PMID: 16895930

29. Rost B, Sander C (1992) Jury returns on structure prediction. Nature 360: 540. doi: 10.1038/360540b0 PMID: 1281284

30. Bromberg Y, Kahn PC, Rost B (2013) Neutral and weakly nonneutral sequence variants may define individuality. Proceedings of the National Academy of Sciences 110: 14255–14260. doi: 10.1073/pnas.1216613110 PMID: 23940345

31. Shimoyama M, De Pons J, Hayman GT, Laulederkind SJF, Liu W, et al. (2015) The Rat Genome Database 2015: Genomic, phenotypic and environmental variations and disease. Nucleic Acids Research 43: D743–D750. doi: 10.1093/nar/gku1026 PMID: 25355511

32. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, et al. (2015) The Mouse Genome Database (MGD): Facilitating mouse as a model for human biology and disease. Nucleic Acids Research 43: D726–D736. doi: 10.1093/nar/gku967 PMID: 25348401

33. Bromberg Y, Rost B (2008) Comprehensive in silico mutagenesis highlights functionally important residues in proteins. Bioinformatics 24: i207–i212. doi: 10.1093/bioinformatics/btn268 PMID: 18689826

34. Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Human Mutation 34: E2393–2402. doi: 10.1002/humu.22376 PMID: 23843252

35. Boycott KM, Vanstone MR, Bulman DE, Mackenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature Reviews Genetics 14: 681–691. doi: 10.1038/nrg3555 PMID: 23999272

36. Carter H, Karchin R (2014) Predicting the Functional Consequences of Somatic Missense Mutations Found in Tumors. In: Ochs FM, editor. Gene Function Analysis. Totowa, NJ: Humana Press. pp. 135–159. doi: 10.1007/978-1-62703-721-1_8 PMID: 24233781

37. Masica DL, Li S, Douville C, Manola J, Ferris RL, et al. (2015) Predicting survival in head and neck squamous cell carcinoma from TP53 mutation. Human genetics 134: 497–507. doi: 10.1007/s00439-014-1470-0 PMID: 25108461

38. Tolstrup N, Toftgård J, Engelbrecht J, Brunak S (1994) Neural Network Model of the Genetic Code is Strongly Correlated to the GES Scale of Amino Acid Transfer Free Energies. Journal of Molecular Biology 243: 816–820. doi: 10.1006/jmbi.1994.1683 PMID: 7966302

39. Moreau Y, Tranchevent L-C (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics 13: 1–14. doi: 10.1038/nrg3253 PMID: 22751426

40. Bromberg Y (2013) Chapter 15: Disease Gene Prioritization. PLoS Computational Biology 9. doi: 10.1371/journal.pcbi.1002902 PMID: 23633938

41. Macarthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science. doi: 10.1126/science.1215040 PMID: 22344438

42. Schaefer C, Meier A, Rost B, Bromberg Y (2012) SNPdbe: constructing an nsSNP functional impacts database. Bioinformatics 28: 601–602. doi: 10.1093/bioinformatics/btr705 PMID: 22210871

43. Kawabata T, Ota M, Nishikawa K (1999) The Protein Mutant Database. Nucleic Acids Research 27: 355–357. PMID: 9847227

44. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, et al. (2011) Ensembl BioMarts: A hub for data retrieval across taxonomic space. Database 2011: 1–9. doi: 10.1093/database/bar030 PMID: 21785142

45. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, et al. (2015) Ensembl 2015. Nucleic Acids Research 43: D662–D669. doi: 10.1093/nar/gku1010 PMID: 25352552

46. Cock PJa, Antao T, Chang JT, Chapman Ba, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423. doi: 10.1093/bioinformatics/btp163 PMID: 19304878

47. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4: 1073–1081. doi: 10.1038/nprot.2009.86 PMID: 19561590

48. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659. doi: 10.1093/bioinformatics/btl158 PMID: 16731699

49. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25. PMID: 9254694

*And into the mountains I go, to lose my mind and find my soul*

Adapted from John Muir

Chapter 3

# SAVs from DMS studies

## 3.1 Introduction

Traditional datasets used for training and assessing VEPs are often biased by deleterious, high effect disease variants (Section 1.3), e.g., from the database OMIM which contains SAVs that cause disease phenotypes in human. DMS presents a novel experimental approach that has the potential to drive a significant shift in the field of VEPs and beyond by providing less biased high-throughput measurements of deleterious as well as beneficial variant effect on the level of protein function (Subsection 1.2.3).

With more DMS data becoming available, the VEP Envision has recently been trained exclusively on such SAVs (Gray *et al.*, 2018). Furthermore, the method is trained as a regression predictor, i.e., it estimates the degree of effect a SAV has on protein function. We analyzed how well this novel method maintains its performance on a larger dataset of DMS studies which have become available since its development. We also assessed how the focus on regression affects the ability to distinguish between SAVs in just two classes, neutral and effect. Furthermore, the majority of traditional classification VEPs such as SIFT, Polyphen-2 or SNPA2 have not seen this type of data or the specific variants during their training (Kumar *et al.*, 2009; Adzhubei *et al.*, 2013; Hecht *et al.*, 2015). This creates the opportunity to perform an unbiased assessment. For example, we evaluated how those methods deal with the task of predicting the precise degree of effect or what their behavior is towards beneficial effect SAVs. This chapter contains additional data and analyses that are not part of the publication and also extends the discussion to include more recent developments in the fast-growing field of DMS.

## 3.2 Methods

This section introduces methodology which extends the one in the manuscript or is important for understanding the additional findings presented here (Reeb *et al.*, 2020, see Section 3.5).

To build a new dataset for the assessment of VEPs we searched the literature for DMS studies, excluding those with less than 100 variants reported in total. We specifically did not exclude sets with only segments of proteins analyzed, since our goal is not to train on this data but only to analyze agreement. For that, fragments do not pose an issue. Functional effect scores were either retrieved from the Supplemental Information of papers or requested from the authors. The heterogeneity of functional assays employed seemed to be reflected in the various data formats used. Since publication of this work, a new resource for multiplexed variant effect data, MaveDB, has been published (Esposito *et al.*, 2019). This should make future endeavors in a similar direction significantly easier as it provides a central resource with a well-defined data format. The resulting dataset contains SAVs from 22 different DMS studies (Table 3.1). To allow a fair comparison between methods, assessments are always performed on a set of SAVs that every VEP provided a prediction for. SetCommon is the largest common subset of 17,781 deleterious effect SAVs from all DMS studies for which that is the case.

Functional scores from every DMS assay were normalized to lie between 0, denoting no or wt-like effect, to 1, the highest effect SAVs. The resulting score distributions differ significantly between experiments. However, applying further score normalization such as in Gray *et al.* (2018) would not make the distributions more comparable, rather in that analysis, scores were distributed similar to begin with. Since our goal is primarily to analyze the relationship between experimental measurements and predicted effect, not to train on those scores, no additional normalization is required and might even obfuscate trends in the data. Deleterious and beneficial effect SAVs were always treated separately since experimental assays cannot be assumed to behave evenly for both. Furthermore, not splitting by effect type would entail that the highest deleterious effect is treated equally to the highest beneficial effect which does not appear justifiable in any circumstance.

**Table 3.1. Overview of the evaluated deep mutational scanning datasets.** 22 datasets were collected from literature. Dataset identifier denotes the short name used to refer to the set throughout this work. Many analyses were limited to the largest common subset of experiments to which every prediction method could be applied. This subset of SAVs is referred to as SetCommon and consists of the datasets highlighted in bold font. FACS is short for fluorescence-activated cell sorting, DBMS stands for droplet-based microfluidic screening. For assay types, dpl. is short for display.

| Dataset identifier | Protein | Assay type | Protein source organism | Reference |
|---|---|---|---|---|
| ccdB | Toxin CcdB | Growth | *E. coli* | Adkar *et al.* (2012) |
| **YAP1** | Transcriptional coactivator YAP1 | Phage dpl. | *H. sapiens* | Araya *et al.* (2012) |
| **MAPK1** | Mitogen-activated protein kinase 1 | Growth | *H. sapiens* | Brenan *et al.* (2016) |
| **BRCA1** | Breast cancer type 1 susceptibility protein | Growth | *H. sapiens* | Findlay *et al.* (2018) |
| **CCR5** | C-C chemokine receptor type 5 | FACS | *H. sapiens* | Heredia *et al.* (2018) |
| **CXCR4** | C-X-C chemokine receptor type 4 | FACS | *H. sapiens* | Heredia *et al.* (2018) |
| HSP82_2011 | ATP-dependent molecular chaperone HSP82 | Growth | *S. cerevisiae* | Hietpas *et al.* (2011) |
| HSP82_2013 | ATP-dependent molecular chaperone HSP82 | Growth | *S. cerevisiae* | Hietpas *et al.* (2013) |
| HSP82_2013_Exp | ATP-dependent molecular chaperone HSP82 | Growth | *S. cerevisiae* | Jiang *et al.* (2013) |
| **GAL4** | Regulatory protein GAL4 | Growth | *S. cerevisiae* | Kitzman *et al.* (2014) |
| LGK | Levoglucosan kinase | Growth | *L. starkeyi* | Klesmith *et al.* (2015) |
| **PPARG** | Peroxisome proliferator-activated receptor γ | FACS | *H. sapiens* | Majithia *et al.* (2016) |
| **PTEN** | Phosphatase and tensin homolog | FACS | *H. sapiens* | Matreyek *et al.* (2018) |
| **TPMT** | Thiopurine S-methyltransferase | FACS | *H. sapiens* | Matreyek *et al.* (2018) |
| haeIIIM | Modification methylase HaeIII | Growth | *H. aegyptius* | Rockah-Shmuel *et al.* (2015) |
| bgl3 | β-glucosidase | DBMS | *Streptomyces* | Romero *et al.* (2015) |
| GFP | Green fluorescent protein | FACS | *Ae. victoria* | Sarkisyan *et al.* (2016) |
| **Ube4b** | Ubiquitin conjugation factor E4 B | Phage dpl. | *M. musculus* | Starita *et al.* (2013) |
| BRCA1_2015_Y2H | Breast cancer type 1 susceptibility protein | Growth | *H. sapiens* | Starita *et al.* (2015) |
| BRCA1_2015_E3 | Breast cancer type 1 susceptibility protein | Phage dpl. | *H. sapiens* | Starita *et al.* (2015) |
| bla | β-lactamase TEM | Growth | *E. coli* | Stiffler *et al.* (2015) |
| IgG1 | Immunoglobulin gamma-1 heavy chain | Yeast dpl. | *H. sapiens* | Traxlmayr *et al.* (2012) |

The VEPs used for comparison to the functional effect scores include classification methods, PolyPhen-2, SIFT, and SNAP2, as well as the regression predictor Envision which was trained on DMS data (Subsection 1.3.1). Since sequence conservation is such an important feature for the prediction of variant effect, an additional baseline method was created based on PSI-BLAST: The homology search tool was run with three iterations against UniProtKB and estimations of variant effect calculated based on the resulting PSSM. Frequent substitutions, i.e., positive values, were treated as having no or low effect, while uncommon substitutions with negative values were considered as effect predictions. This predictor is referred to as Naïve Conservation in the following.

Two measures were used to assess the agreement of VEPs' predictions with the experimentally determined effect of SAVs. Correlation performance was evaluated with Spearman's ρ (Equation 3.1, Virtanen *et al.*, 2020). As the ranked variable version of the more common correlation measure Pearson's *R*, ρ is more suitable for data that is not normally distributed or contains outliers —both of which cases apply to the DMS datasets analyzed here (Wilcox, 2016). Furthermore, *R* requires a linear relationship between the predicted and measured effects. While this would be desirable, it seems overly strict to require this level of agreement from VEPs, in particular given the heterogeneity of DMS assays. ρ, on the other hand, only measures a monotonic relationship, i.e., if experimentally measured effect increases so should the predicted effect strength. Arguably, this is already sufficiently useful and details of the exact nature of the relationship between predictions and experiments could be determined in future analyses when performance has reached a level where this appears warranted.

$$\text{Spearman's } \rho \ (\rho) = \frac{n\sum_{i=1}^{n} rx_i ry_i - \sum_{i=1}^{n} rx_i \sum_{i=1}^{n} ry_i}{\sqrt{n\sum_{i=1}^{n} rx_i{}^2 - \left(\sum_{i=1}^{n} rx_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} ry_i{}^2 - \left(\sum_{i=1}^{n} ry_i\right)^2}}$$

$$\text{Mean squared error (MSE)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

where

$n$ = Number of SAVs

$x_i$, $y_i$ = Experimentally measured/predicted effect score for SAV $i$

$rx_i$, $ry_i$ = Rank of the $i$-th experimental measure/prediction

(3.1)

The second measure used for evaluation is an error measure, the Mean squared error (MSE, Equation 3.1, Pedregosa *et al.*, 2011). MSE complements the qualitative ρ by pro-

viding an absolute measure which assesses how close every predicted effect score is to the experimentally determined effect score of the respective SAV. This is an important addition since the degree of effect can be important, e.g., in precision medicine applications, but might be obscured in $\rho$ with its limitation to monotonic relationships. A baseline for both $\rho$ and MSE is provided by the Naïve Conservation predictor. 95 % confidence intervals (CIs) for $\rho$ and MSE were estimated using percentile bootstrapping with 1,000 samples and replacement (Bishara and Hittner, 2017).

For classification performance analyses, SAVs were assigned to classes "effect" or "neutral" by two orthogonal methods: (i) In dataset SetCommonSyn95 SAVs are defined as being neutral when their functional score in the DMS assay is within the middle 95 % range of synonymous variants' scores on the same protein. (ii) SetCommonAuthor is constructed with SAVs for which authors of the original publication provided class assignments (Table A.1). Both sets represent the largest common subset of SAVs that each of the five VEPs can perform a prediction for. Given these sets, the prediction performance was assessed using receiver operating characteristic (ROC) curves and the area under that curve (AUC). This choice eschews picking a single threshold at which to assess VEPs and thus allows a more detailed look at every method's performance over the whole range of predicted values. Furthermore, it allows the assessment of Envision where choosing a threshold is not intended in general and every single choice would be arbitrary and potentially biased. ROCs, AUCs, as well as their 95 % CIs were calculated in R using package pROC which implements the DeLong method instead of bootstrapping (`ci.se`, `ci.auc`, R Core Team, 2020; Turck *et al.*, 2011; DeLong *et al.*, 1988). In addition to ROC curves, precision-recall curves show the trade-off between the two measures defined in Equation 3.2 (Pedregosa *et al.*, 2011).

$$
\begin{aligned}
\text{True positive (TP)} &= \text{Effect SAV predicted as having an effect} \\
\text{False positive (FP)} &= \text{Neutral SAV predicted as having an effect} \\
\text{True negative (TN)} &= \text{Neutral SAV predicted as not having an effect} \\
\text{False negative (FN)} &= \text{Effect SAV predicted as not having an effect} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}
\end{aligned}
\tag{3.2}
$$

## 3.3 Additional results and discussion

### 3.3.1 Correlation performance for SIFT and PolyPhen-2

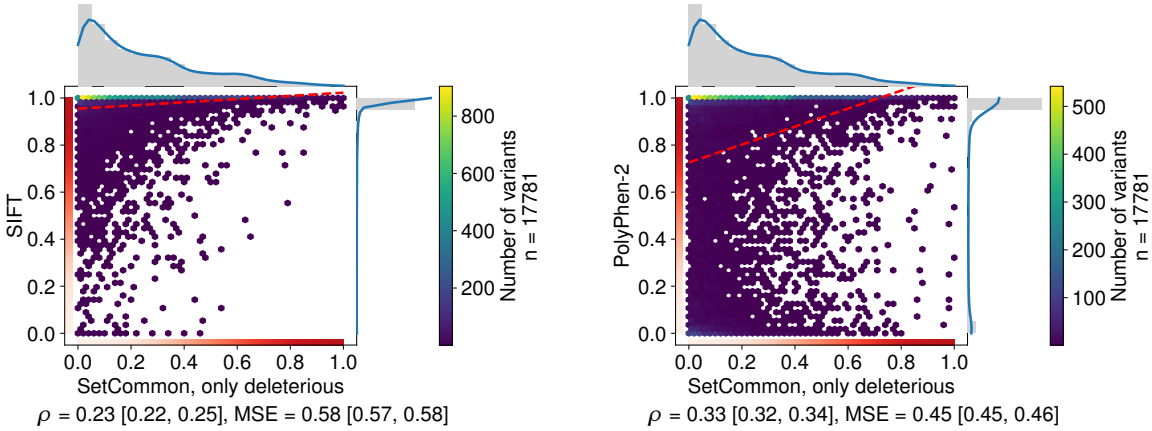
SIFT and PolyPhen-2 belong to a category of VEPs which were specifically designed to predict SAVs only in two classes. Therefore, one could argue that their prediction scores should not be treated as continuous measures. For this reason, they were excluded from the analysis in Reeb *et al.* (2020). However, given the popularity of these tools, users may still be inclined to apply the VEPs in this way—despite the developers' intentions. We provide an overview of the performance to be expected in this case.

On deleterious effect SAVs of SetCommon, both tools show almost binary prediction scores and are heavily skewed towards effect (Figure 3.1). The resulting correlation performance is better than random but both VEPs are outperformed by SNAP2 ($\rho = 0.41$, 95 % CI = $[0.4, 0.42]$) and only PolyPhen-2 achieves better correlation than the baseline represented by Naïve Conservation ($\rho = 0.29$ $[0.28, 0.3]$). Among all five VEPs, MSE is worst for SIFT, followed by Polyphen-2. This is in line with expectations, given their intended application and training which focused on discriminating between two classes and primarily high effect SAVs. Overall, the two VEPs are therefore least suited for estimating effect strength and should indeed not be applied for this purpose.

### 3.3.2 Low correlation for Envision

Envision was trained on data from DMS assays, distinct from the test set used here, and employs a regression prediction as opposed to the classification approach of traditional VEPs. As such, it is poised to perform well in our evaluation. However, the achieved correlation performance is low ($\rho = 0.1$ $[0.08, 0.11]$). Given the significantly higher cross-validation performance of $\rho = 0.5$ reported during training of the method and confirmed by the authors on an independent test set, the weak performance comes as a surprise (Gray *et al.*, 2018). Nonetheless, our findings corroborate another recent study on a set of three DMS assays which also found Envision's performance around the level we observed on our larger set ($\rho \in \{-0.44, 0.09, 0.27\}$, Sruthi and Prakash, 2020). This doesn't invalidate the approach itself and a regression predictor for variant effect is in fact highly desirable. However, while larger SAV datasets from DMS studies enable such endeavors, the volatile performance also indicates that much remains to be learned.

One relevant factor for this result could lie in the completeness of Envision's input features. When available, the VEP makes use of features from protein 3D structure, with the B factor and solvent accessibility being the two most important features overall (Gray *et al.*,




**(a)** SIFT **(b)** PolyPhen-2

**Figure 3.1. Agreement between experimentally determined and predicted SAV effect.** Hexbin plots show the effect of 17,781 deleterious effect SAVs in SetCommon against predicted effect by classification VEPs **(a)** SIFT and **(b)** PolyPhen-2. Marginals denote distributions of the respective scores together with a kernel density estimate (blue). Normalized scores range from no effect (0) to the highest effect observed or possible to predict (1). Footers denote Spearman's ρ and the mean squared error with their respective 95 % confidence intervals determined by a percentile bootstrap. A dashed red line represents the linear least squares regression.

2018). However, 3D structures are not available for all query sequences and the authors highlight that performance on their training set decreases by 39 % without structural features. Indeed, structural features were missing for around 80 % of all residues in the ten proteins from SetCommon (Figure 3.2). On the other hand, there is no clear relationship between the lack of structural input features and the correlation performance achieved. Evolutionary features extracted from MSAs were also important and their exclusion lead to an 18 % decreased performance during training. While all but one of those features are available for more than 96 % of residues in our set, this highlights the impact those inputs have. How exactly the alignments were created between the publication and the automated webserver could therefore lead to differences in performance, as confirmed in personal communication with the authors.

Finally, DMS data is still new, in particular to the field of VEPs. How exactly scores from DMS experiments should be treated and normalized is an active discussion in the field. Some of the differences in performance are therefore likely caused by contrasting approaches in the treatment of the underlying scores. This leads to differently oriented training and evaluation sets which creates challenges similar to those in evaluating traditional VEPs (cf. Subsection 1.3.2).

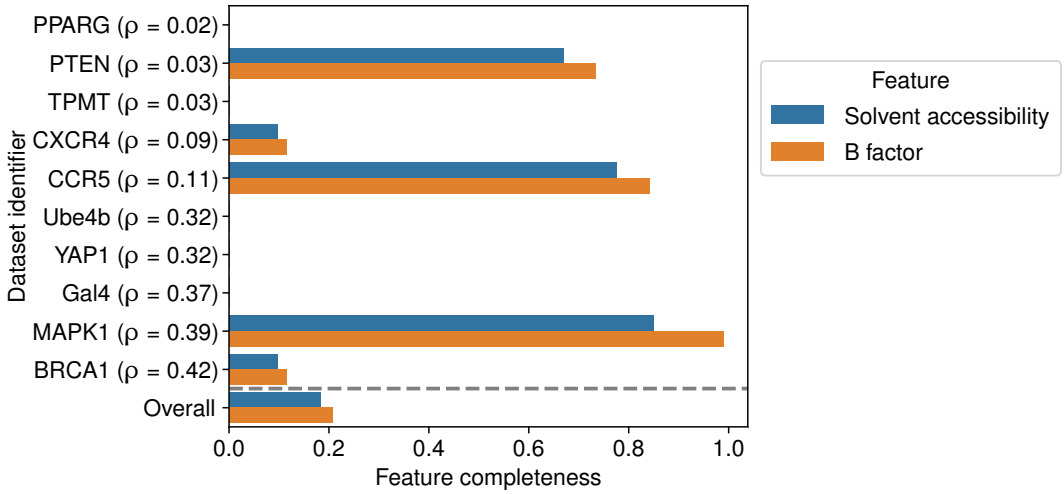


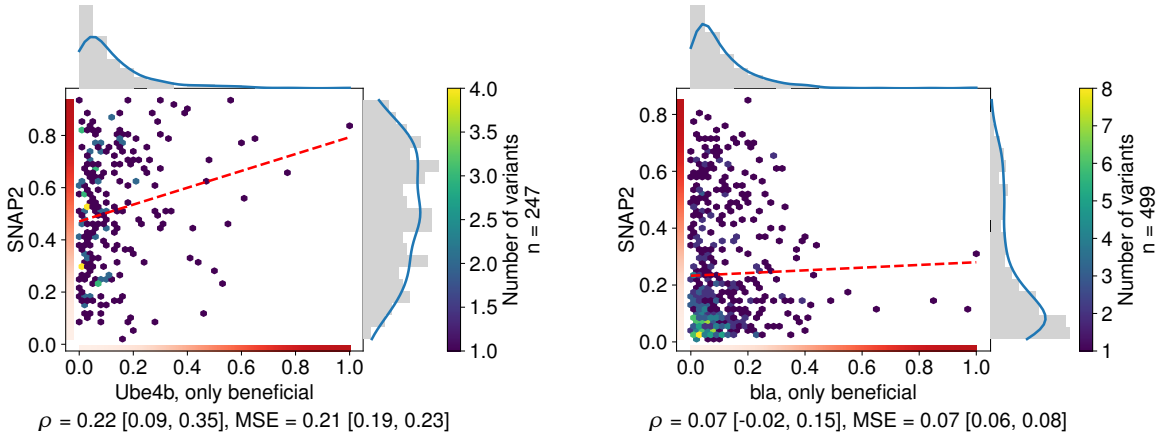

**Figure 3.2. Completeness of the two most important Envision input features.** Bars denote the fraction of residues in the respective proteins where the structural input features "Solvent accessibility" and "B factor" exist. "Overall" refers to all residues in all ten proteins. Datasets are sorted by ascending correlation performance on the subset of SAVs for which Envision and all other VEPs in our analysis provided a prediction.

### 3.3.3 Correlation with beneficial effect variants

Normalizing all scores to lie within the interval $[0, 1]$ is required to perform comparisons on the complete set of SAVs. However, this approach necessarily disregards the underlying score distributions of every DMS experiment. For example, a low MSE is easier to reach when most of the scores are clustered in a small interval with some outliers. Furthermore, the highest effect score measured by one assay might not be directly comparable to that from another due to experimental setups. Therefore, it is worthwhile to investigate agreement of predictions on the level of single DMS experiments as well.

For beneficial effect SAVs, the highest correlation can be observed on variants from Ube4b (Figure 3.3a). While the set contains just 247 SAVs and CIs are accordingly large, the trend appears clear. This is interesting, given that the distribution of experimental scores is clearly shifted towards low effect. One could hypothesize that the training set composition of VEPs such as SNAP2 creates a bias towards only detecting low beneficial effect variants. However, the experimental distribution on β-lactamase TEM (bla) is similarly skewed towards low effect and correlation performance is significantly lower on this set (Figure 3.3b). Furthermore, SNAP2 prediction scores are heavily skewed, in particular when compared to those on deleterious effect SAVs. This indicates that traditional VEPs are likely to mistake beneficial effect variants as having no or low effect. One possible explanation for this is that evolutionary information appears to be a less useful signal for beneficial effect SAVs, as evidenced by the drop in performance of Naïve Conservation

($\rho = -0.08$ $[-0.09, -0.06]$). Intuitively, this makes sense since beneficial effects may often not be detrimental to the organism and are thus under less evolutionary pressure. Given the previously discussed issues in the training of VEPs such as SNAP2 and their bias towards sequence conservation, these results are consistent (cf. Subsection 1.3.2).




**(a)** Ube4b **(b)** bla

**Figure 3.3. SNAP2 predictions of beneficial effect.** Hexbin plots show the effect of beneficial effect SAVs from DMS studies **(a)** Ube4b and **(b)** bla against predictions of the best VEP on these sets, SNAP2. Performance measures, CIs and marginals are as described in Figure 3.1.

### 3.3.4 Influence of DMS functional assay type

The effect values determined by a DMS study are affected by the particular experimental setup employed. As pointed out in Subsection 1.2.3, the functional assays used to impose selective pressure have to be specific for the protein function of interest and can measure activity in various ways. A common choice are growth- or fitness-based assays that evaluate the SAV's effect by coupling protein function to cell growth or survival. Other approaches, such as protein display or fluorescence-activated cell sorting, could be considered a more direct, or at least different, measure of protein function. This distinction can be important since reduced fitness does not necessarily mandate impaired function of the protein in question. Furthermore, lower protein function might show only little effect on fitness (Capriotti *et al.*, 2019b). For example, in a highly conserved region of Hsp82, variants that lead to a 79% reduced protein function resulted in only a 5% fitness decline (Jiang *et al.*, 2013). Variants with less effect on function were indistinguishable from the wild-type due to the shape of the elasticity function between the protein's function and organism fitness.

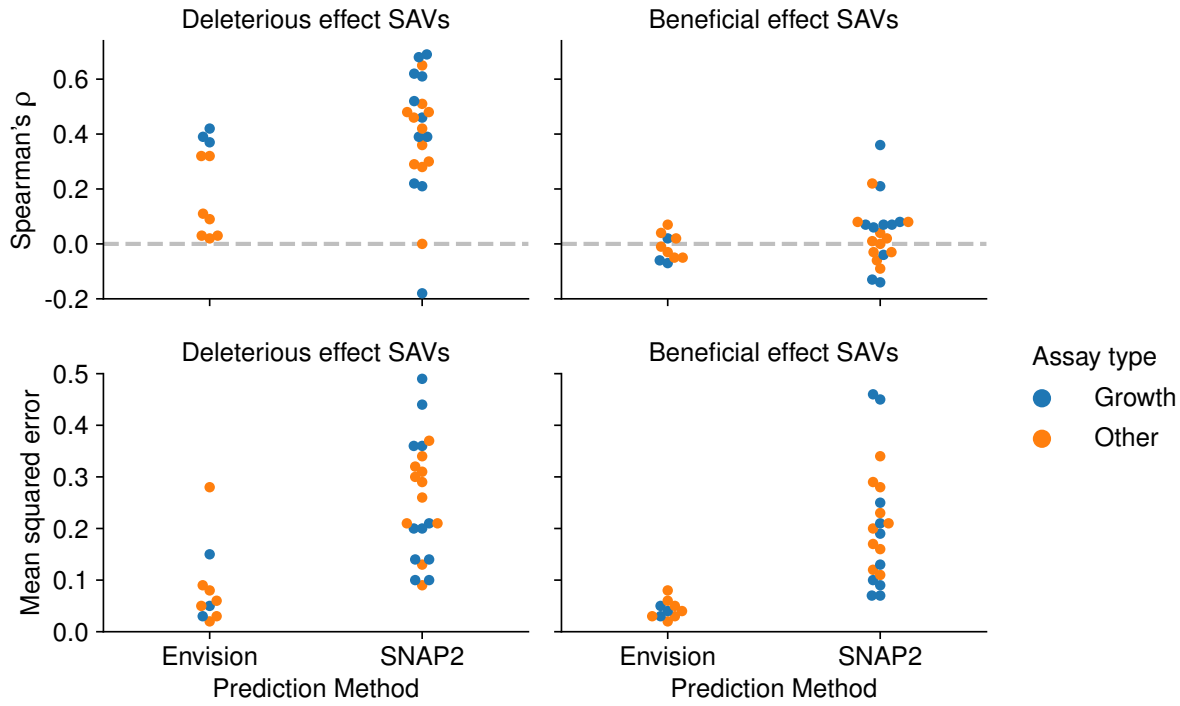



**Figure 3.4. Performance of effect prediction between different DMS assay types.** Scatterplots show the performance of VEPs Envision and SNAP2 in predicting the effect determined through DMS experiments. SAVs are split between either deleterious or beneficial effect. Points are colored by whether the respective dataset's selection assay is growth-based or using some other selection (Table 3.1). Note that Envision predictions exist for only a subset of all DMS datasets and plots thus contain fewer points compared to SNAP2.

Among the 22 primary DMS measurements used in this study, 11 use assays that can be classified as growth-based (Table 3.1). For both deleterious and beneficial effect SAVs no major differences in prediction performance between SAVs from the two assay types can be observed (Figure 3.4). The only clear difference is found for Envision's ρ on deleterious effect. However, with just three growth-based assays the sample size is too small for any reliable inference. The method's training set was also balanced with four out of nine assays being growth-based. If a bias of Envision towards growth-based selections were observed on a larger dataset, this would offer strong support that selection type must be accounted for during training. However, the data presented here indicates no impact of assay type on effect prediction performance.

A recent study evaluated the usage of variant effect determined by DMS studies for identifying human disease variants from ClinVar (Livesey and Marsh, 2019). For this purpose, SAV effects from growth-based assays showed significantly better classification performance. This suggests that VEPs such as PolyPhen-2—which focus on the prediction of variant effect in the context of human disease—could particularly benefit from using

growth-based DMS data in their training.

### 3.3.5 Classification of SAVs based on definitions by authors of DMS studies

The analyses of classification performance in Reeb *et al.* (2020) are based on distinguishing neutral and effect SAVs by the effect of synonymous variants which yields the set referred to as SetCommonSyn95. While this is a sensible approach that generalizes well, an alternative view presents itself for the specific subset of DMS studies analyzed here. For some sets, authors of the studies already provided definitions into classes themselves. One could argue that these are valuable in that they take into account specifics of the protein in question and might thus be more accurate. Such definitions were available for six of the 22 DMS studies and were clustered into two classes. The only beneficial effect variants in this set were 33 SAVs from haeIIIM, thus no distinction between deleterious and beneficial effect was made. In total, the set contained 8,944 SAVs. Of those, only 6,410 SAVs from studies BRCA1, PTEN and TPMT had predictions from all VEPs. This largest common subset is referred to as SetCommonAuthor in the following.

Comparing the ROC curves and resulting AUCs to those from deleterious effect SAVs in SetCommonSyn95 shows slightly better prediction performance by all methods on SetCommonAuthor (cf. Figures A.1a and 3.5). The improved performance could stem from the specific subset of just three DMS studies. However, AUCs for all VEPs on BRCA1, PTEN and TPMT were similar or lower when using class definitions based on synonymous variants' effects. Thus, the increased performance is indeed caused by the difference in assigning classes and indicates that classifications performed by authors of DMS studies capture a different aspect. However, the ranking between methods is only partly affected: Envision performs marginally better but still clearly worse than all VEPs specialized for classification. SIFT also remains the second-worst method separated by a large margin from the top three. Among those, SNAP2 shows the highest performance (AUC = 0.78), closely followed by PolyPhen-2 (0.77) and Naïve Conservation (0.76). Unlike on SetCommonSyn95, these differences are not statistically significant and 95 % confidence intervals of all three VEPs overlap.

Besides the higher overall performance, the shape of the precision recall curves is highly similar between the two classification schemes for all VEPs (Figure A.1b). PolyPhen-2 and SIFT which have almost binary prediction outputs show little room for achieving a trade-off between precision and recall. In fact, the two curves do not overlap until the maximal threshold and a recall of 0 is reached. Even the naïve approach based on PSI-BLAST profiles offers this to some degree, although its prediction output is by design limited to only around 20 distinct values. Envision shows generally poor performance

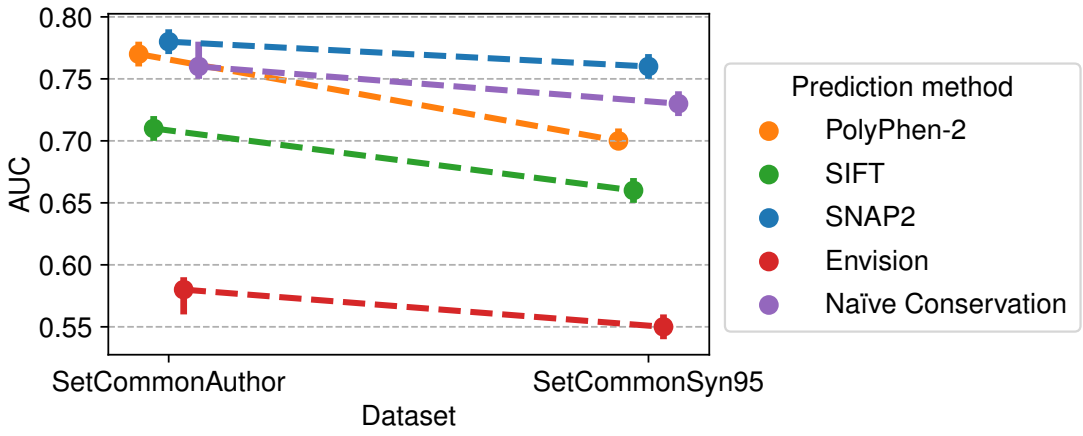



**Figure 3.5. Binary prediction performance between class definitions.** Classification performance of VEPs is summarized by the area under the receiver operating characteristic curve (AUC). SetCommonAuthor consists of 6,410 SAVs classified as either having an effect or being neutral by the authors of the respective DMS studies. SetCommonSyn95 contains 13,796 SAVs with classes defined by the effect exhibited by synonymous variants in each DMS study (Reeb *et al.*, 2020). Error bars denote 95 % confidence intervals.

in line with the achieved AUC. However, more interesting is the sharp drop in precision at a high threshold by more than 0.3 with the following rise to a precision of 1. Closer analysis reveals that performance is relatively stable on BRCA1, while the drop is caused by SAVs from TPMT and the subsequent rise by those from PTEN. Although there is no clear explanation for this behavior, the fact that it only appears for Envision indicates a potential bias and further underlines the VEP to be less well suited for classification. Arguably, the smoothest transition between precision and recall as well as overall stability is achieved by SNAP2. This can be regarded as manifestation of its careful training and also forms the foundation for SNAP2 scores to correlate with effect strength.

## 3.4 Conclusions and outlook

Increasingly large variant effect data from the advances in next-generation sequencing have lead to new discoveries, opportunities, and challenges in the field of variant effect prediction. DMS is one technique based on these developments and has seen tremendous growth in recent years, contributing large amounts of SAVs with known effect on protein function. Our analyses highlight that traditional VEPs trained as classification methods on non-DMS SAVs capture some aspects of effect in this new data. On the other hand, a VEP specifically developed on DMS SAVs offers complementary strengths but also highlights that more improvements are necessary. Generally, the findings of this study give important pointers for the development of better VEPs in the future. For example, no single

score could express the overall performance of methods in all possible applications. Thus, optimization of VEPs during training must not focus on a single value either. Furthermore, deleterious effect SAVs are predicted significantly better than beneficial effect variants. Finally, no method performed well for both classification as well as regression tasks. While this is expected, the benefit VEPs provide towards precision medicine will increase by finding a better trade-off. This could be accomplished by a singular new method or astute combination of complementary approaches.

Several issues regarding the underlying data remain to be solved as well. For example, comparisons of multiple experimental measurements on the same protein showed differing agreements, highlighting the heterogeneity of assays. Naturally, prediction methods can only ever hope to capture one of such facets of effect or, alternatively, detect the set of variants that exhibit an effect on function in every type of functional assay. Another aspect to consider for future endeavors is that methods trained on data from two different assays for the same protein might learn and then predict different types of effect. This can create a situation akin to that of current VEPs trained on various subsets of variants which are thus both difficult to fairly evaluate as well as to apply without expert knowledge of every VEP's specific intricacies.

Regarding experimental values, we evaluated the correlation between predicted effect and functional effect scores as they were published by the authors of DMS studies employing as little normalization as possible. Many authors provided the scores as the logarithm of read counts before and after applying functional pressure, potentially normalized to the wild-type. Others have used more sophisticated approaches to better reflect a variant's impact on function. While all are sensible, it makes scores even more heterogeneous. For a different approach, one might want to retrieve the raw sequencing data for every study and process scores in a common pipeline, including error correction measures. This is possible with methods such as Enrich2, dms_tools or PACT (Rubin *et al.*, 2017; Bloom, 2015; Klesmith and Hackel, 2019). Since noise is a common problem in DMS studies (Starita *et al.*, 2017), this might yield more comparable scores. However, heterogeneity of the employed functional assays and how direct their measure of protein function is will remain.

Future evaluations might concentrate on additional DMS datasets which are not analyzed here. For example, a set of DMS studies has focused on viral proteins and how their variants affect virus proliferation. While the efficiency of invasion and replication within host cells is indeed those proteins' function, none of the prediction methods evaluated here have been trained to recognize the effect of SAVs in viruses. Viral proteins form a particularly interesting case since one of the main features in common variant prediction methods is evolutionary information and viruses are under markedly different

evolutionary pressure (Krupovic and Bamford, 2011). As so patently shown by the 2020 COVID-19 pandemic caused by SARS coronavirus 2, the impact these proteins can have on a global scale is arguably unmatched. Thus, anything helping in understanding their mechanisms and supporting respective research will provide a benefit to population and economic health alike. Recently, a new study found performance in predicting the effect of variants on viral proteins significantly lower when compared to SAVs in human, yeast or bacterial proteins (Livesey and Marsh, 2019). Indeed, one of the main drivers for this effect appeared to be low sequence diversity in alignments of viral proteins which current VEPs cannot account for. Even more so, the best performing VEP overall was among the worst on viral proteins. Therefore, much remains to be discovered in this area and there is a large potential for improvement of VEPs.

Finally, the study mentioned above also found that an unsupervised probabilistic method showed the best performance in predicting SAV effects as determined by DMS studies (Livesey and Marsh, 2019). This should be a clear sign for future VEPs that, despite DMS providing a wealth of potentially valuable training data, improvements in the field may not come from falling into the same traps again and developing incrementally better methods using increasingly complex models. Rather, the evolutionary information in the sequence appears to already provide crucial signals which should first be understood and harnessed in as direct as possible ways while avoiding the currently present biases.

## 3.5 Journal article

Reeb et al. BMC Bioinformatics (2020) 21:107
https://doi.org/10.1186/s12859-020-3439-4

BMC Bioinformatics

RESEARCH ARTICLE Open Access

# Variant effect predictions capture some aspects of deep mutational scanning experiments

Jonas Reeb[1*], Theresa Wirth[1] and Burkhard Rost[1,2,3,4]

## Abstract

**Background:** 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; also referred to as missense mutations, or non-synonymous Single Nucleotide Variants – missense SNVs or nsSNVs) for particular proteins. We assembled SAV annotations from 22 different DMS experiments and normalized the effect scores to evaluate variant effect prediction methods. Three trained on traditional variant effect data (PolyPhen-2, SIFT, SNAP2), a regression method optimized on DMS data (Envision), and a naïve prediction using conservation information from homologs.

**Results:** On a set of 32,981 SAVs, all methods captured some aspects of the experimental effect scores, albeit not the same. Traditional methods such as SNAP2 correlated slightly more with measurements and better classified binary states (effect or neutral). Envision appeared to better estimate the precise degree of effect. Most surprising was that the simple naïve conservation approach using PSI-BLAST in many cases outperformed other methods. All methods captured beneficial effects (gain-of-function) significantly worse than deleterious (loss-of-function). For the few proteins with multiple independent experimental measurements, experiments differed substantially, but agreed more with each other than with predictions.

**Conclusions:** DMS provides a new powerful experimental means of understanding the dynamics of the protein sequence space. As always, promising new beginnings have to overcome challenges. While our results demonstrated that DMS will be crucial to improve variant effect prediction methods, data diversity hindered simplification and generalization.

**Keywords:** Sequence variation, Variant effect prediction, Deep mutational scanning, Non-synonymous sequence variant, Missense variant, Single nucleotide variant

## Background

Recent human sequencing projects conclude that we all carry about 10,000 single amino acid variants (**SAVs**; also referred to as missense mutations, or non-synonymous Single Nucleotide Variants: **nsSNVs**) with respect to the "reference genome" and by 20,000 for every pair of unrelated individuals [1, 2]. Many of these SAVs are assumed to be neutral, while others might change protein function, contributing to complex phenotypes and causing diseases. 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 may be incorrect disease associations [6]. Without improving the ability to interpret SAV effects, both on the level of the

* Correspondence: reeb@rostlab.org; assistant@rostlab.org
[1]Department of Informatics, Bioinformatics & Computational Biology - i12, TUM (Technical University of Munich), Boltzmannstr 3, 85748 Garching/Munich, Germany
Full list of author information is available at the end of the article

organism and the protein, the promise of precision medicine will remain, importantly 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, i.e. all possible amino acid mutations. In such deep mutational scanning (**DMS**) studies [13, 14], a sequence library with all possible variants is subjected to selection. 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 and deleterious effect on protein function are discovered together with a quantification of how much effect. Thus, DMS aims at measuring the landscape of functional fitness for select proteins [18].

DMS also screens proteins for improved drug binding, antibody affinity, using non-native chemical stresses, or non-proteinogenic amino acids, and on synthetic proteins [19–26]. Finally, DMS share objectives with directed evolution, benefiting protein engineering [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, for the Ubiquitin-60S ribosomal protein L40 variant effects have been assessed through their direct impact on yeast growth and through the impaired activation by the E1 enzyme [27, 28]. Similarly, BRCA1 has been assayed through E3 ubiquitin ligase activity and 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]. On top, variants might affect molecular function as assayed by DMS although being clinically benign, i.e. not causing disease.

Long before experimental DMS, prediction methods had addressed the same task in silico [33–41]. These methods were developed on very limited data; many focused on disease-causing SAVs from OMIM [42], others used databases such as PMD [43] cataloguing variants by effect upon protein function or structure. CADD solved the problems of data limitation and bias 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 dataset 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 have been optimized on relatively small, unrepresentative subsets: fewer than 85,000 of all possible 217 million human SAVs (< 0.04%) have some experimental annotations [44, 45]. Methods agree much more with each other for SAVs with than for those without annotations [46].

DMS datasets constitute a uniquely valuable resource for the evaluation of current SAV effect prediction methods [17, 47, 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

### DMS studies not complete yet

Our Deep Mutational Scanning (DMS) analyses began with 22 separate experimental datasets from 18 unique proteins, since some experiments were performed on the same protein (Supplementary Online Material (SOM), Fig. S1a, Table S1) [29, 30, 32, 50–65]. In total the set contained 68,447 variants (Fig. S1); 2358 (3%) of these were synonymous, the other 97% constituted SAVs (or missense mutations).

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 had SAVs with both deleterious and beneficial effects (Table S2; those two could be seen as "disruptive variants" arching over gain- and loss-of-function). Most SAVs were beneficial for only 3 of 22 studies (14%), for the other 19 studies deleterious outnumbered beneficial SAVs by factors of 1.5–22.5 (Fig. S1b). Due to asymmetries in numbers and experimental fidelity, deleterious and beneficial SAVs were analyzed separately.

### Some correlation achieved by all methods

*SetCommon* constituted a subset of all 22 datasets with 32,981 effect SAVs (17,781 deleterious) for which we had predictions from each method (Table 1). Although all predictions differed from the experiments, all correlated slightly positively for deleterious SAVs (Spearman $\rho \geq 0.1$, Fig. 1a-c, Tables 2, S3). The 95% confidence intervals (CIs) of methods did not overlap, and their differences were statistically significant (Table S4).

Both SIFT [39] and PolyPhen-2 [37] are optimized for capturing binary effects, not correlations, as confirmed by recent studies [47, 49]. Consequently, analysis for these was confined to binary predictions. SNAP2 [38] and Envision [49] scores appeared, overall, less binary (Figs. 1a-b). SNAP2 distributions were skewed toward high effect, while Envision also succeeded in detecting

**Table 1** Number of SAVs in aggregated datasets[a]

| | Number of SAVs | | | |
|---|---|---|---|---|
| | Total | Neutral | Deleterious | Beneficial |
| SetAll | 66,089 | 818 [b] | 45,382 | 19,889 |
| SetCommon | 32,981 | 0 | 17,781 | 15,200 |
| SetCommonSyn90 | 15,621 | 8926 | 4545 | 2150 |
| SetCommonSyn95 | 15,621 | 10,587 | 3209 | 1825 |
| SetCommonSyn99 | 15,621 | 13,506 | 1548 | 567 |

[a]*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

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

SAVs with less pronounced effects (Fig. 1a-b). Predictions by Naïve Conservation, based on PSI-BLAST profiles, correlated more with the DMS experiments than Envision (Fig. 1c).

### Envision might approximate experimental values best

When evaluating methods by the numerical difference between experimental and predicted variant effect scores (mean squared error, MSE), Envision appeared best, followed at considerable distance by Naïve Conservation and SNAP2 (Fig. 1, Table 2). However, its low MSE partially originated from predicting no SAV with strong effect (the highest Envision score was 61% of the possible maximum – 0.61). This resembled the experimental distribution skewed towards low effect (Fig. 1b, gray distributions next to x- and y-axes). Indeed, shuffling the prediction scores yielded the same MSE (Fig. S2a). Predicting a normal distribution around the experimental mean, performed slightly worse but still better than all other prediction methods (Fig. S2b). When considering each DMS measurement separately, Envision also appeared to perform best except for the transcriptional coactivator YAP1 (YAP1) with the most uniform distribution of effect scores (similar number of lowest, medium, and strongest effects observed; Fig. S3b, Table S5).

### All classification methods detect increasing effect strength

Do methods work better for SAVs with stronger observed effect? Toward this end, the experimental scores were sorted into 20 bins of increasing effect strength, and the effect predictions in each bin (here referred to as recall) were monitored for all prediction methods. All classification methods tended to reach higher recall levels for SAVs with stronger effects (Fig. 2a, higher values toward the right). Furthermore, all methods also show an increase without a clear saturation point showing that the range of

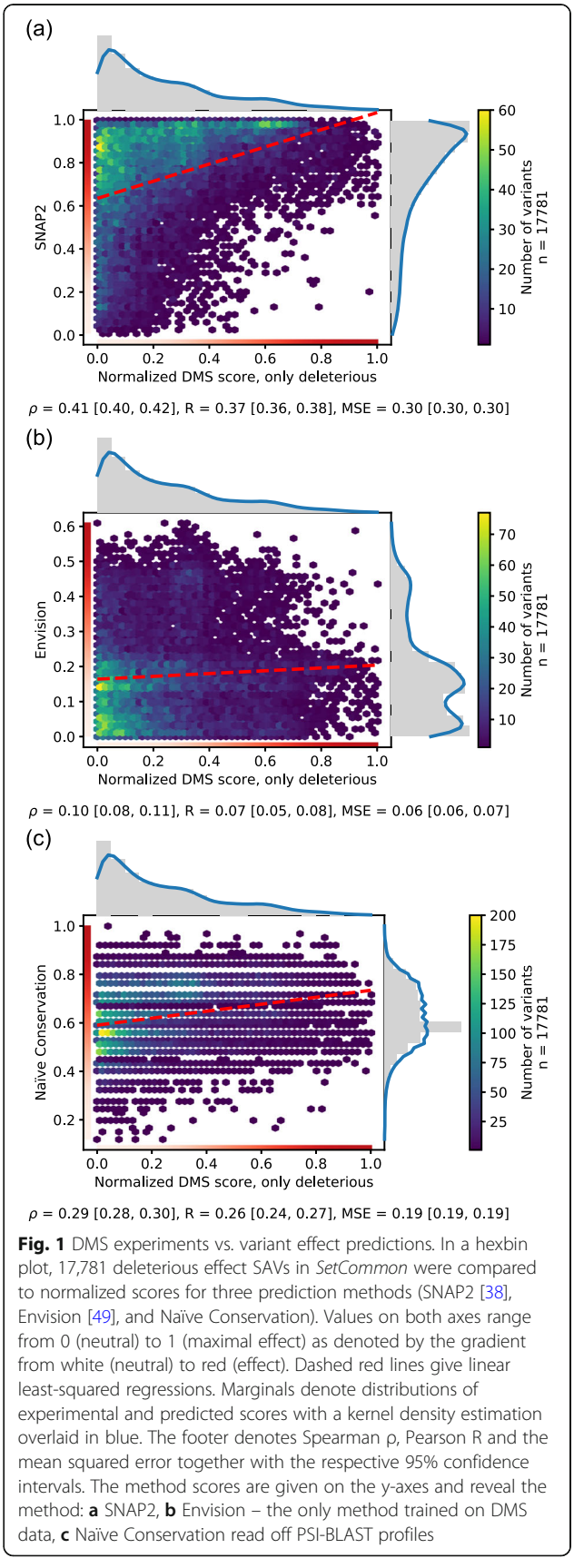



**Fig. 1** DMS experiments vs. variant effect predictions. In a hexbin plot, 17,781 deleterious effect SAVs in *SetCommon* were compared to normalized scores for three prediction methods (SNAP2 [38], Envision [49], and Naïve Conservation). Values on both axes range from 0 (neutral) to 1 (maximal effect) as denoted by the gradient from white (neutral) to red (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** SNAP2, **b** Envision – the only method trained on DMS data, **c** Naïve Conservation read off PSI-BLAST profiles

**Table 2** Pearson ρ and mean squared error (MSE) for methods on *SetCommon*[a]

| | deleterious SAVs ($n = 17{,}781$) | | beneficial SAVs ($n = 15{,}200$) | |
|---|---|---|---|---|
| | ρ | MSE | ρ | MSE |
| 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] |
| Naïve 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] |

[a]*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_Note3). Values in brackets are 95% confidence intervals

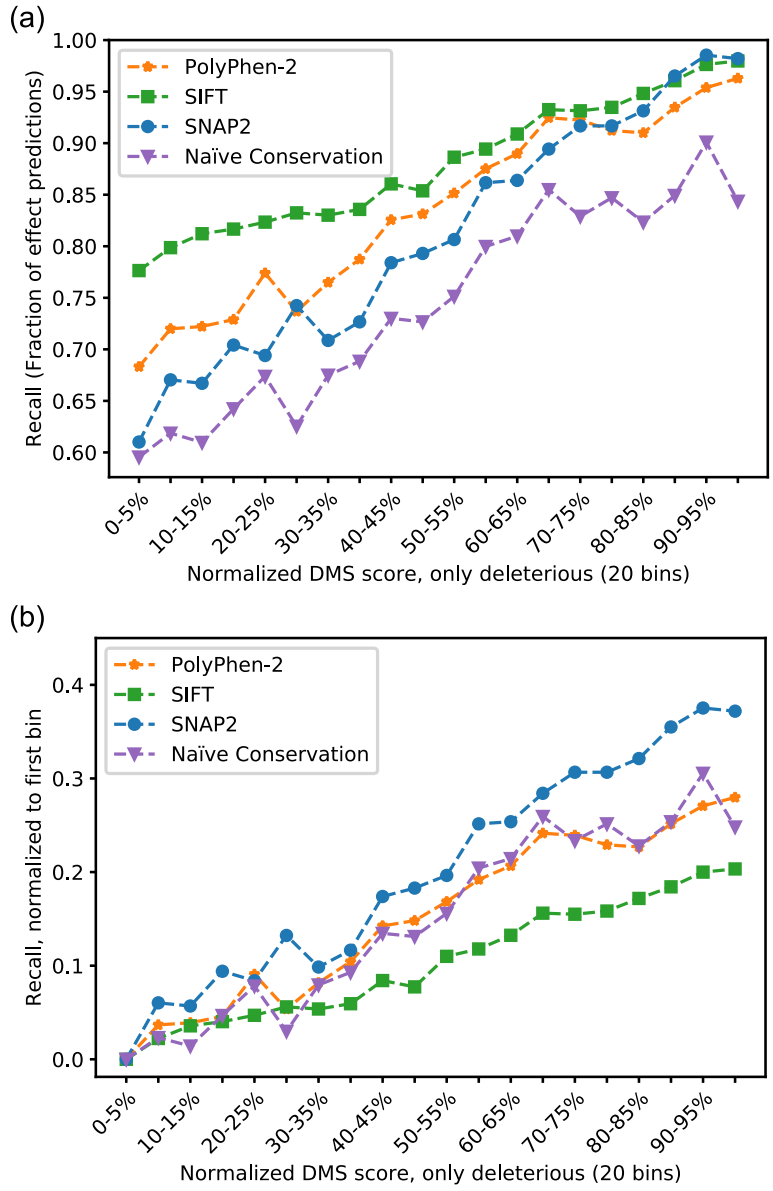



**Fig. 2** Recall proportional to deleterious DMS effect scores. The continuous normalized DMS scores with deleterious effect in *SetCommon* were split into 20 bins of equal size. **a** In each bin the fraction of SAVs predicted as having an effect by the binary classification methods (PolyPhen-2 [37], SIFT [39] and SNAP2 [38]) was shown. Naïve Conservation read off PSI-BLAST profiles was treated as an effect prediction when scores were above 0. For all other methods the default score thresholds were applied. **b** shows the values adjusted for the amount of effect predicted in the first bin

increasing effect strength is detected. For some methods the difference between the least- and most-effect bins was higher than for others, i.e. their predictions distinguished more between high and low experimental scores (Fig. 2b).

### 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, S6, Fig. S4). Furthermore, most methods essentially predicted similar numbers or lower numbers of effect variants irrespective of the observed effect strength with the exception of SNAP2 that detected some high effect SAVs (Fig. S5). The conservation-based prediction also decreased substantially from a Spearman ρ of 0.29 for deleterious to −0.08 for beneficial SAVs (Table 2, Fig. S4c). SNAP2 scores were shifted more toward lower effect than for deleterious SAVs (Fig. 1a and Fig. S4a, gray distributions). In contrast to Spearman ρ, the MSE for beneficial effect SAVs was similar to that for deleterious SAVs. Envision again was by far best (MSE = 0.05, Tables 2, S7, Fig. S6). However, although Envision used 25% beneficial effect SAVs for development (SOM_Note1), the correlation was much lower for beneficial than for deleterious SAVs ($\rho = -0.14$ versus 0.1).

### 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. Such data were available for 11 measurements on 4 proteins (Table S8, Fig. S7); unfortunately, Envision predictions were available for only one of those proteins (BRCA1). For deleterious SAVs, the lowest correlation was that between two measurements on breast cancer type 1 susceptibility protein, BRCA1 and BRCA1_2015_E3 ($\rho = 0.21$, Fig. S7b). Rather than experimental noise, the low correlation might also originate from different experimental setups employed for multifunctional 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. S7h) provided a single case in strong support of such an explanation. To compare prediction methods and experiments, we assessed the difference in ρ and MSE for each combination of the 11 measurements (Fig. 3). Experiments clearly agreed more with each other than with SNAP2 and Naïve Conservation on the same datasets (Fig. 3: all values negative).

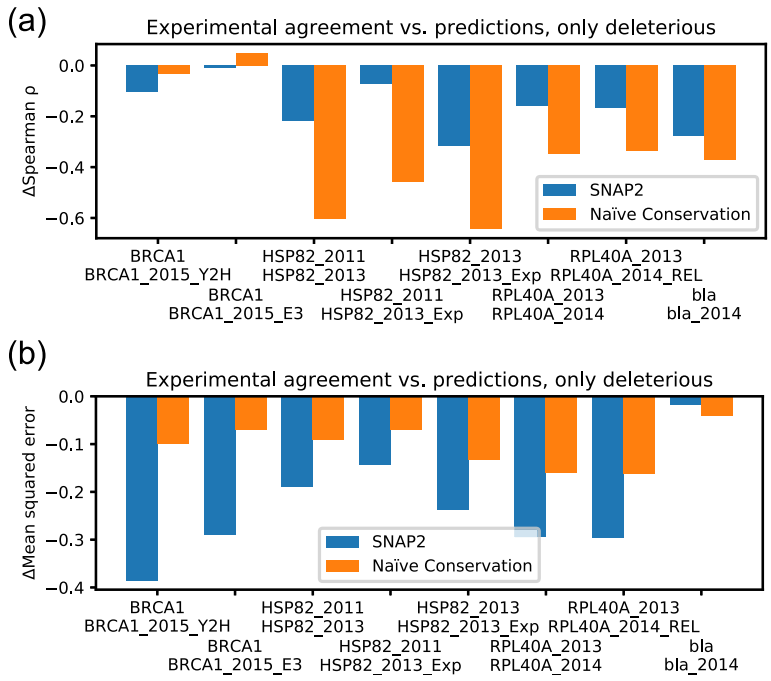



**Fig. 3** Experimental agreement vs. predictions. For every pair of experimental measurements on the same protein (Table S1), the agreement between two experiments and that between each experiment and the predictions of SNAP2 and Naïve Conservation are compared. **a** $\Delta\rho = 0.5*(\rho(\times 1,p1) + \rho(\times 2,p2)) - \rho(\times 1,\times 2)$, (b) $\Delta MSE = MSE(\times 1,\times 2) - 0.5*(MSE(\times 1,p1) + MSE(\times 2,p2))$. Where × 1/× 2 are the experiments and p1/p2 the predictions on the two experiments, all of which are calculated based on the largest possible set of SAVs. Negative values on the y-axes thus imply that the agreement between experiments is higher than that between experiment and prediction, positive values that predictions agree more

Experiments did not correlate at all with each other for beneficial effect (mean ρ = 0.03) although the MSE remained low (mean MSE = 0.05, Table S8, Fig. S8). The major issue for this comparison was the small number of only 572 SAVs.

### Assessment of binary classification (neutral/effect) similar to regression

Scores from binary classification methods (neutral or effect) are often assessed through receiver operating characteristic (ROC) curves avoiding to choose particular thresholds to distinguish neutral and effect. Toward this end, we assigned classes to SAVs through normalization by experimental measurements of synonymous variants [60] (Methods). Other solutions are feasible, each with their own ad hoc parameter choices and flaws implying that the following results provide one snapshot instead of a sustained method ranking.

On the 3209 deleterious effect SAVs of *SetCommonSyn95* (10,587 neutral, Table 1, Fig. S9), SNAP2 achieved the highest area under the curve (AUC, 0.76, 95% CI [0.75, 0.77]). It was the only method statistically significantly better than Naïve Conservation (0.73 [0.72, 0.74], Figs. 4, S10 Table S9). Precision-recall curves also highlighted the smooth transition of SNAP2 scores opposed to those for Naïve Conservation although the peak performance was similar for both (Fig. S11). Envision - not developed for this task - performed better than random, but clearly worse than the classification methods (AUC = 0.55 [0.54, 0.56]). However, the four proteins considered here (BRCA1, PPARG, PTEN and TPMT), also correlated above average for SNAP2, PolyPhen-2 and SIFT (Table S3). Using different thresholds in severity to classify SAVs did not qualitatively change these major findings (*SetCommonSyn90*, *SetCommonSyn99*, Fig. S12a-b).

At their default thresholds SIFT, PolyPhen-2, and SNAP2 consider over two thirds of the neutral variants to have an effect. Interestingly, the behavior of Envision trained on DMS data was the reverse as previously illustrated by the maximal scores reaching only up to 61% of the possible maximal values (and thereby contributing to a seemingly low MSE).

Beneficial SAVs were also difficult to classify: PolyPhen-2 and SNAP2 performed best with AUC = 0.62, followed by SIFT, while Envision predictions were not better than random (Fig. S13, Fig. S12c-d, Table S9). Naïve Conservation also performed significantly worse at a level of random predictions.

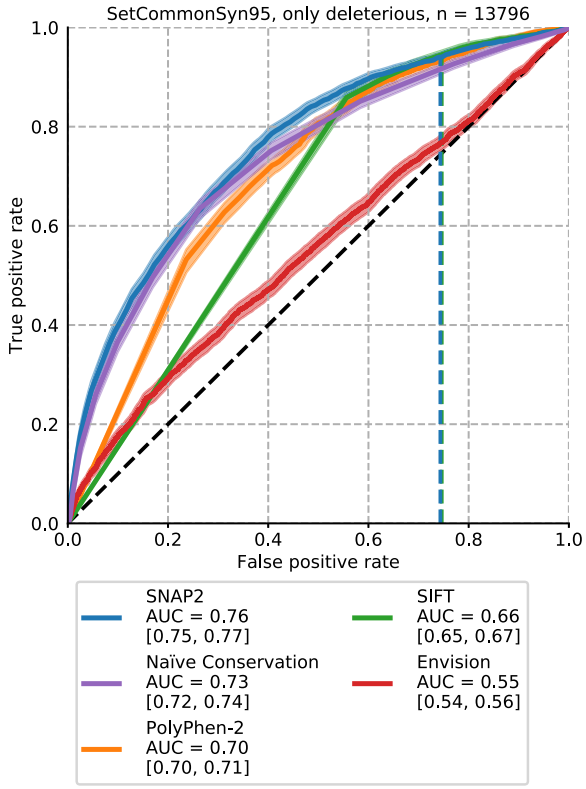



**Fig. 4** Classification performance of all prediction methods. Shown are ROC curves for 13,796 deleterious effect SAVs which were classified into either neutral, defined by the middle 95% of the scores from synonymous variants, or effect (*SetCommonSyn95*). Shaded areas around lines denote 95% confidence intervals. The legend denotes the AUC for each of the five prediction methods, along with the 95% confidence intervals. Horizontal dashed lines denote the default score threshold used by SNAP2 (blue) and SIFT (green)

## Discussion

### No clear winner in predicting effect variants

We compared the predictions of five methods with SAV effects determined by DMS experiments. SNAP2 was trained on binary classification data (effect or neutral). Nevertheless, predictions have been shown to correlate with effect strength [5, 66, 67]. To a degree, the Deep Mutational Scanning (DMS) data replicated this finding, highlighting that even methods trained for classification capture aspects of effect strength.

Sorting DMS scores into 20 bins and including classification methods SIFT and PolyPhen-2 in the analysis, all methods indicated better recognition of high effect SAVs. This finding might be attributed to the bias of classifications methods towards high effect variants, a common criticism in the field [68–71]. We observed the same trend for Naïve Conservation exclusively using PSI-BLAST profiles to predict SAV effects. This emphasized the importance of this signal but to some extend also explained the traditional classification methods' bias since they all rely on this input.

The significantly better performance of Envision in estimating the precise degree of effect especially suggested value in this approach. However, the low MSE was

largely explained by that Envision correctly predicted the overall distribution of experimental scores. Thus, the definite distinction between 'good prediction' and 'advantageous bias' remained elusive.

When treating DMS effect scores as binary assignments (neutral or effect), ROC curves highlighted the high false positive rates of the evaluated classification methods. A similar perspective on over-prediction has recently been observed for ClinVar data [69]. Over-prediction might be encouraged by the way many users of prediction methods mistakenly chose their tools, namely by testing a small set of SAVs they know have an effect and valuing methods highest when they predict effects for more of those.

### Family conservation carries most important signal

Most surprising was the overall good performance of Naïve Conservation. Disease causing SAVs from OMIM typically affect the most conserved residues [46], and machine-learning based predictions have been criticized to largely capture conservation [17, 70, 72–74]. Furthermore, simple conservation patterns can capture aspects of variant effects [75]. Our findings partially validated this for DMS experiments, although the effect distributions observed by DMS and predicted by Naïve Conservation differed substantially (Fig. 1c, gray distributions). Another recent analysis also found a method heavily relying on evolutionary information as one of the best performers on DMS data, although more sophisticated than our naïve approach [48, 76].

### Beneficial effects neither correctly predicted, nor consistent between experiments

The bad correlation and classification performance of beneficial effect SAVs by all methods suggested those to have distinctly different signatures than deleterious SAVs, missed by current approaches. Generally, SAVs with neutral or beneficial effects are often not recognized well [69, 77]. In part, this is attributable to the lack of respective experimentally verified data useable for training sets. For beneficial effect variants, the rise of DMS studies could help to alleviate this problem and lead to the development of less biased methods.

Agreement between experimental studies was particularly low for beneficial effect SAVs. Maybe DMS assays are still biased towards measuring deleterious effects. These results put the seemingly poor predictions of beneficial SAVs into perspective. Generally, the wide variation of correlation between experiments for different datasets/proteins has also been observed in another recent DMS analysis [48].

## Conclusions

Deep mutational scanning (DMS) studies set out to explore the relation between protein sequence and molecular function. We collected 22 DMS experiments and focused on single amino acid variants (SAVs, also referred to as missense mutations or non-synonymous SNVs). Most studies probe only a small subset of all possible variants (for a protein with N residues, there are 19*N non-native SAVs). Two experiments probing the same protein tended to agree more with each other than with predictions for deleterious effect (Fig. 3). Nonetheless, experiments also disagreed significantly (Table S8). No single measure captured all aspects of the comparison between experiments and predictions, e.g. the ranking of methods changed crucially depending on the measure used to compare (Table 2, SOM_Note2).

We analyzed five variant effect prediction methods: *Envision* was trained on DMS data, *PolyPhen-2*, *SIFT* and *SNAP2* were methods developed to classify into effect/neutral, and *Naïve Conservation* (essentially using PSI-Blast conservation to predict effect/neutral) was added to gauge the importance of evolutionary conservation for the prediction. For deleterious SAVs, all methods reached slightly positive Spearman ρ correlations with the DMS experiments (Fig. 1). The classification method SNAP2 correlated most with effect strength, although most of the correlation was explained by simple conservation. The lowest mean squared error (MSE) was achieved by Envision. Its MSE was as low as that between experiments, although most of the low MSE could be explained by correctly predicting the distribution of scores (Fig. 1, Fig. S2a). All methods performed better on SAVs with deleterious (akin to loss-of-function) than with beneficial (gain-of-function) effect. However, experimental agreement was also almost non-existing for beneficial effects.

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. 4). Notably, Naïve Conservation captured effect better than some more advanced tools. Methods performed better for SAVs with stronger experimental effect scores (Fig. 2: higher toward right), although most classifiers tended to substantially over-predict at their default scores (Fig. 4). Overall, our analyses confirm some of the trends from other reviews of DMS data [48, 49].

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, [78]). This is a magnitude matched by a single large DMS experiment. The generality of a single SAV might not be comparable between the sets, yet DMS opens up variant effect prediction to new methodologies, possibly even to deep learning approaches [79, 80]. The enriched data might also allow methods to distinguish between toggle and rheostat positions [73]. Furthermore,

DMS studies contain many beneficial effect SAVs that have, so far, been underrepresented. Finally, DMS focuses on molecular function, i.e. some of the disruptive SAVs (deleterious or beneficial) might correspond to clinically benign SAVs. Nevertheless, DMS will likely give rise to new methods better predicting SAV effects upon molecular protein function and upon organisms. In fact, growth-based DMS assays have been shown to be predictive of human disease SAVs in a recent study [48]. Therefore, a combination of experimental data with new prediction methods might be what is needed to attain the goals of precision medicine.

## Methods

### Dataset collection

Figure 5 sketches the basic workflow of this analysis. 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 S10). 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 S11) [44, 81]. Six of the 22 experiments contained up to five substitutions (pairwise sequence identity ≥98%); those were maintained for prediction. We refer to the combined data as *SetAll* (66,089 SAVs) supplemented by *SetCommon* with 32,981 SAVs for which we had a prediction from every method tested (Table 1). *SetCommon* contained SAVs from ten of the 22 experiments: YAP1, MAPK1, BRCA1, CCR5, CXCR4, GAL4, PPARG, PTEN, TPMT, and Ube4b (Table S1). During completion of this manuscript, MaveDB, a centralized resource of multiplexed assays of variant effect has been published [8, 82]. MaveDB identifiers exist for ten of our 22 datasets (November 2019, Table S10).

*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, 83]. 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.

### Processing functional effect scores

Several DMS studies provide multiple effect scores for the same protein of which we decided on only one per set (Table S12). 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 S13) became 0, and larger values denoted more effect. Next, scores were interpolated, separately for each of the 22 DMS measurements, to lie between 0 and 1 (highest effect). This interpolation did not affect Spearman ρ or the mean squared error within each dataset. 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

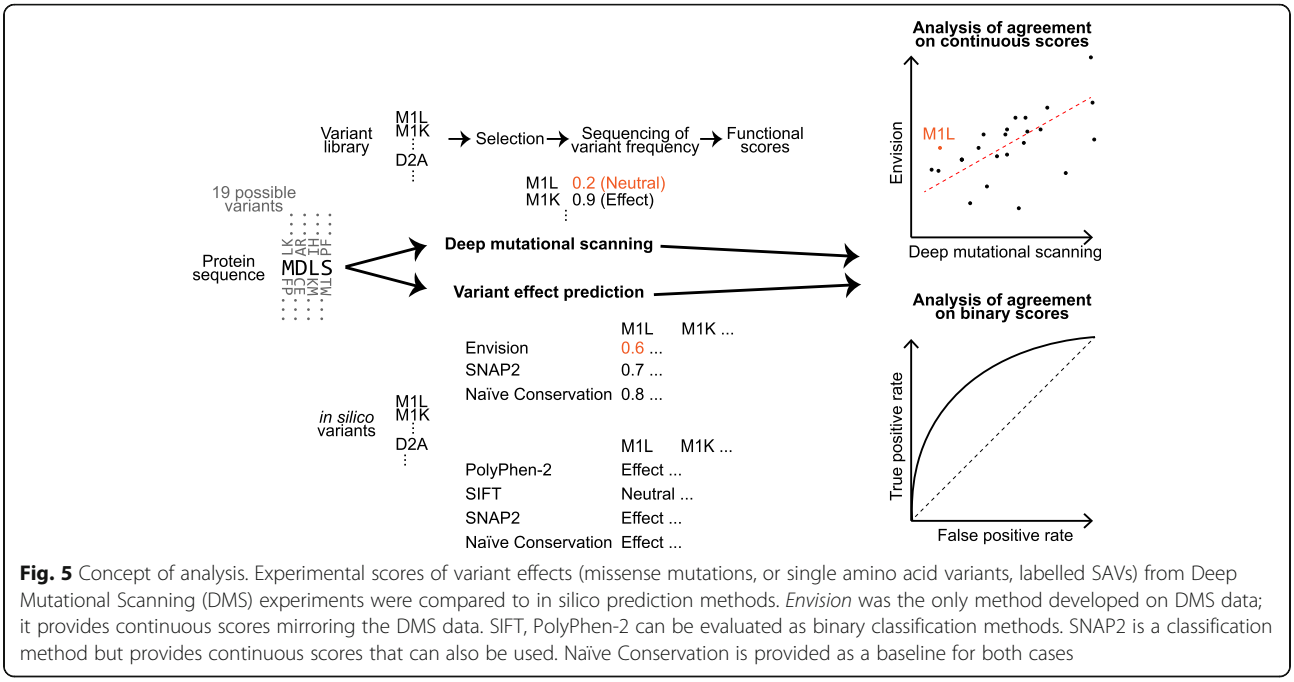



**Fig. 5** Concept of analysis. Experimental scores of variant effects (missense mutations, or single amino acid variants, labelled SAVs) from Deep Mutational Scanning (DMS) experiments were compared to in silico prediction methods. *Envision* was the only method developed on DMS data; it provides continuous scores mirroring the DMS data. SIFT, PolyPhen-2 can be evaluated as binary classification methods. SNAP2 is a classification method but provides continuous scores that can also be used. Naïve Conservation is provided as a baseline for both cases

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 *SetCommon-Syn90|95|99*. Again, deleterious and beneficial effect SAVs were analyzed separately.

### Performance measures

Experiments and predictions were compared through three measures (SOM_Note3, SOM_Note2): (1) **mean squared error** (**MSE**) calculated with the scikit-learn metrics module [84]; (2) **Pearson R** (pearsonr) and (3) **Spearman ρ** (spearmanr) both calculated with the SciPy stats module [85]. 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 [86]). 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) [87, 88]. Additionally, precision-recall curves were created using scikit-learn (precision-recall-curve). These are defined with TP as true positives (predicted and observed as effect), FP as false positives (predicted as effect, observed as neutral), and FN as false negatives (predicted neutral, observed effect): Precision = TP/(TP + FP), Recall = True Positive Rate = TP/(TP + FN) and False Positive Rate = FP/(FP + TN).

### 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 with classification model humdiv on genome assembly GRCh37/hg19 and default parameters [89]. Predictions failed for all relevant residues of the three DMS studies on Hsp82. *Envision* [49] predictions were retrieved online which requires UniProtKB identifiers as input [90]. Therefore, Envision predictions could be analyzed only for ten proteins (Table S14). While SNAP2 and SIFT predicted all SAVs, PolyPhen-2 and Envision failed for some residues, shrinking the size of the datasets. 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*. The prediction was not intended to be the most accurate conservation score possible but rather to represent a suitable baseline since (PSI-)BLAST results are used in some way as input feature by all methods analyzed here.

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_Note4). 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.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1186/s12859-020-3439-4.

**Additional file 1.** Supporting Online Material (SOM) containing additional figures, tables and notes.

### Abbreviations

AUC: Area under the ROC curve; CI: Confidence interval; DMS: Deep mutational scanning; MSE: Mean squared error; ROC: Receiver operating characteristic; SAV: Single amino acid variant

### 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 Littmann (both TUM) for helpful discussions, to Inga Weise (TUM) for administrative support and to Tim Karl (TUM) for help with hard- and software. Particular thanks to the anonymous reviewers who helped importantly to clarify the path through the data plethora.

**Authors' contributions**
TW performed initial analyses and data curation on a smaller set of datasets. JR collected additional datasets, improved on the methodology and expanded the analyses. JR and BR wrote the manuscript. BR conceptualized and supervised the work. All authors read and approved the final manuscript.

**Funding**
This work was supported by the Deutsche Forschungsgemeinschaft (DFG) – project number 640508. This work was also supported by the DFG and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program. Further funding was provided by the Bavarian Ministry for Education through funding to the TUM paying for the positions of the authors. The funding bodies did not play any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Availability of data and materials**
The datasets generated and code for their analysis are available on Mendeley Data (https://doi.org/10.17632/2rwrkp7mfk.1).

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]Department of Informatics, Bioinformatics & Computational Biology - i12, TUM (Technical University of Munich), Boltzmannstr 3, 85748 Garching/Munich, Germany. [2]Institute for Advanced Study (TUM-IAS), Lichtenbergstr 2a, 85748 Garching/Munich, Germany. [3]TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany. [4]Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West, 168th Street, New York, NY 10032, USA.

Received: 12 December 2019 Accepted: 3 March 2020
Published online: 17 March 2020

**References**

1. Tennessen JA, Bigham AW, Connor TDO, Fu W, Kenny EE, Gravel S, Mcgee S, Do R, Liu X, Jun G, et al. Evolution and functional impact of rare coding variation from deep sequencing of human Exomes. Science. 2012;337:64–70.
2. The 1000 Genomes Project Consortium. 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, Worthey E, Chisholm RL, Green ED, Jacob HJ, 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 Comput Biol. 2013;9.
5. Mahlich Y, Reeb J, Hecht M, Schelling M, De Beer TAP, Bromberg Y, Rost B. Common sequence variants affect molecular function more than rare variants? Sci Rep. 2017;7:1608.
6. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, 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 Lett. 2016;590:2327–41.
8. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. Variant interpretation: functional assays to the rescue. Am J Hum Genet. 2017;101:315–25.
9. Capriotti E, Ozturk K, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2019;11(3):e1443.
10. Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, Lena PD, Casadio R, Edwards M, Gifford D, et al. Working toward precision medicine: predicting phenotypes from exomes in the critical assessment of genome interpretation (CAGI) challenges. Hum Mutat. 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. J Mol Biol. 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. Nat Protoc. 2012;7:1382–96.
14. Wrenbeck EE, Faber MS, Whitehead TA. Deep sequencing methods for protein engineering and design. Curr Opin Struct Biol. 2017;45:36–44.
15. Araya CL, Fowler DM. Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. 2011;29:435–442.
16. Fowler DM, Stephany JJ, Fields S. Measuring the activity of protein variants on a large scale using deep mutational scanning. Nat Protoc. 2014;9:2267–84.
17. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11:801–7.
18. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8:610–8.
19. Forsyth CM, Juan V, Akamatsu Y, DuBridge RB, Doan M, Ivanov AV, Zhiyuan M, Polakoff D, Razo J, Wilson K, 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, Gaskins G, Liu Z, Deming L, Axen SD, 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, Green EM, Kenner LR, Mensa B, Morinishi LS, 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, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, 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, Aramini J, Kornhaber G, Hunt JF, Tong L, et al. Computational design of a protein-based enzyme inhibitor. J Mol Biol. 2013;425:3563–75.
24. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers CA, Kamisetty H, Blair P, Wilson IA, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol. 2012;30:543–8.
25. Fujino Y, Fujita R, Wada K, Fujishige K, Kanamori T, Hunt L, Shimizu Y, Ueda T. Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. Biochem Biophys Res Commun. 2012;428:395–400.
26. Rogers JM, Passioura T, Suga H. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. Proc Natl Acad Sci. 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. J Mol Biol. 2013;425:1363–77.
28. Roscoe BP, Bolon DNA. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. J Mol Biol. 2014;18:1199–216.
29. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S. 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, Janizek JD, Huang X, Starita LM, Shendure J. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562:217–22.
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 Res. 2014;42:1–8.
32. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet. 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 Mutat. 2009;30:1237–44.

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;14(Suppl 3).
35. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–D894.
36. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426: Elsevier B.V.:2692–701.
37. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat 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 Res. 2012;40:452–7.
40. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99: American Society of Human Genetics: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;14(Suppl 3):S3.
42. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47: D1038–43.
43. Kawabata T, Ota M, Nishikawa K. The protein mutant database. Nucleic Acids Res. 1999;27:355–7.
44. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15.
45. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.
46. Reeb J, Hecht M, Mahlich Y, Bromberg Y, Rost B. Predicted molecular effects of sequence variants link to system level of disease. PLoS Comput Biol. 2016;12:e1005047.
47. Pejaver V, Babbi G, Casadio R, Folkman L, Katsonis P, Kundu K, Lichtarge O, Martelli PL, Miller M, Moult J, et al. Assessment of methods for predicting the effects of PTEN and TPMT protein variants. Human Mutat. 2019;40:1495–506.
48. Livesay B, Marsh JA. Using deep mutational scanning data to benchmark computational phenotype predictors and identify pathogenic missense mutations. *bioRxiv*. 2019.
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 e113.
50. Adkar BV, Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, Swarnkar MK, Gokhale RS, Varadarajan R. 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. Proc Natl Acad Sci. 2012;109:16858–63.
52. Brenan L, Andreev A, Cohen O, Pantel S, Kamburov A, Cacchiarelli D, Persky NS, Zhu C, Bagul M, Goetz EM, et al. Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. Cell Rep. 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. J Immunol. 2018;200:3825–39.
54. Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. Proc Natl Acad Sci. 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 Genet. 2013;9.
57. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single amino acid mutagenesis. Nat 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 Synth Biol. 2015;4:1235–43.
59. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, Patel KA, Zhang X, Broekema MF, Patterson N, et al. Prospective functional classification of all possible missense variants in PPARG. Nat Genet. 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 Comput Biol. 2015;11:1–28.
61. Romero PA, Tran TM, Abate AR. Dissecting enzyme function with microfluidic-based deep mutational scanning. Proc Natl Acad Sci. 2015;112: 7159–64.
62. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, 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, Shendure J, Brzovic PS, Fields S, Klevit RE. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proc Natl Acad Sci. 2013;110:E1263–72.
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, Grillari J, Rüker F, Obinger C. Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. J Mol Biol. 2012;423:397–412.
66. Bromberg Y, Kahn PC, Rost B. Neutral and weakly nonneutral sequence variants may define individuality. Proc Natl Acad Sci U S A. 2013;110:14255–60.
67. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35:3823–35.
68. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–23.
69. Niroula A, Vihinen M. How good are pathogenicity predictors in detecting benign variants ? PLoS Comput Biol. 2019;15:1–17.
70. Andersen LL, Terczyńska-Dyla E, Mørk N, Scavenius C, Enghild JJ, Höning K, Hornung V, Christiansen M, Mogensen TH, Hartmann R. Frequently used bioinformatics tools overestimate the damaging effect of allelic variants. Genes Immun. 2017;20:10–22.
71. Anderson D, Lassmann T. A phenotype centric benchmark of variant prioritisation tools. Genomic Medicine. 2018;3.
72. Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;36:513–23.
73. Miller M, Bromberg Y, Swint-Kruse L. Computational predictors fail to identify amino acid substitution effects at rheostat positions. Sci Rep. 2017;7:41329.
74. Cline MS, Babbi G, Bonache S, Cao Y, Casadio R, Cruz X, Díez O, Gutiérrez-Enríquez S, Katsonis P, Lai C, et al. Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. Human Mutat. 2019;40: 1546–56.
75. Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and mechanistic models for disease-causing protein variants. Trends Biochem Sci. 2019;44:575–88.
76. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nature Methods. 2018;15: Springer US:816–22.
77. Kim, Y., Ki, C., & Jang, M. (2019). Challenges and Considerations in Sequence Variant Interpretation for Mendelian Disorders. Annals of Laboratory Medicine, 39(5), 421. https://doi.org/10.3343/alm.2019.39.5.421.
78. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.
79. Rives A, Goyal S, Meier J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv preprint. 2019.
80. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics. 2019;20(1):723. https://doi.org/10.1186/s12859-019-3220-8.
81. Pundir S, Martin MJ, O'Donovan C. UniProt tools. Curr Protoc Bioinformatics. 2016;53:1.29.21–21.29.15.

82. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. 2019;20(1):223. https://doi.org/10.1186/s13059-019-1845-6.
83. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map of a Gene's fitness landscape. Mol Biol Evol. 2014;31:1581–92.
84. Fabian P, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Pedregosa F, Varoquaux G, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
85. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
86. Wilcox RR. Comparing dependent robust correlations. Br J Math Stat Psychol. 2016;69:215–24.
87. Turck N, Vutskits L, Sanchez-Pena P, Robin X, Hainard A, Gex-Fabry M, Fouda C, Bassem H, Mueller M, Lisacek F, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;8:12–77.
88. R Core Team. R: a language and environment for statistical computing. In: R Foundation for Statistical Computing; 2018.
89. PolyPhen2 Webserver. http://genetics.bwh.harvard.edu/pph2/bgi.shtml Accessed: 15 Apr 2019.
90. Envision webserver. https://envision.gs.washington.edu/shiny/envision_new/ Accessed: 15 Apr 2019.

## Publisher's Note

*So go ahead, break stuff.*
*Break yourself on the once-hard edges of yourself.*
*And recycle the debris into the foundation of your future.*
Mark Twight - Kiss Or Kill: Confessions of a Serial Climber

Chapter 4

# Conclusion

The field of variant effect prediction is in a state of contradiction. In a health care system with access to ever cheaper sequencing capacities, VEPs play an important part in providing actionable interpretation. Hundreds of such tools are available. Yet, our understanding of performance and behavior is at odds with their increasing significance and the role they are considered for in precision medicine. By analyzing the most popular VEPs currently used, this thesis has highlighted shortcomings and provides directions for future enhancements.

We assembled novel datasets of monogenic disease-causing SAVs in human as well as mouse, demonstrating that these variants exhibit a high effect signal. The popularity of these SAVs for training of VEPs identifies a major source of bias, further compounded by the variants' high sequence conservation. A second assessment provided further support for this as well as other known issues, and farther elucidated original problems. Using experimental data from DMS assays, we proved that no VEP currently captures all aspects of variant effect on protein function in a continuous form. Nor does a single commonly used score suffice for portraying the results of such an assessment. These points persist even when evaluating prediction performance in the more established approach of classifying SAVs as either having an effect or not. Additionally, variants with beneficial effect on protein function poignantly emphasized how they are neglected by current VEPs and thus present one of the major challenges for future methods. At the same time, they provided a strong point for using DMS data in prospective analyses as well as VEP development.

Overall, there is still plenty of room for improvements in a field within computational biology that was established decades ago. While this thesis presents the challenges for VEPs in 2020, it does not offer the gratification of an immediate solution. Integrating the knowledge presented by this thesis in new methods will be the task of future research. Likely, improvements will emerge from both better understanding of problems and careful application of new method development techniques, as well as a further increase in

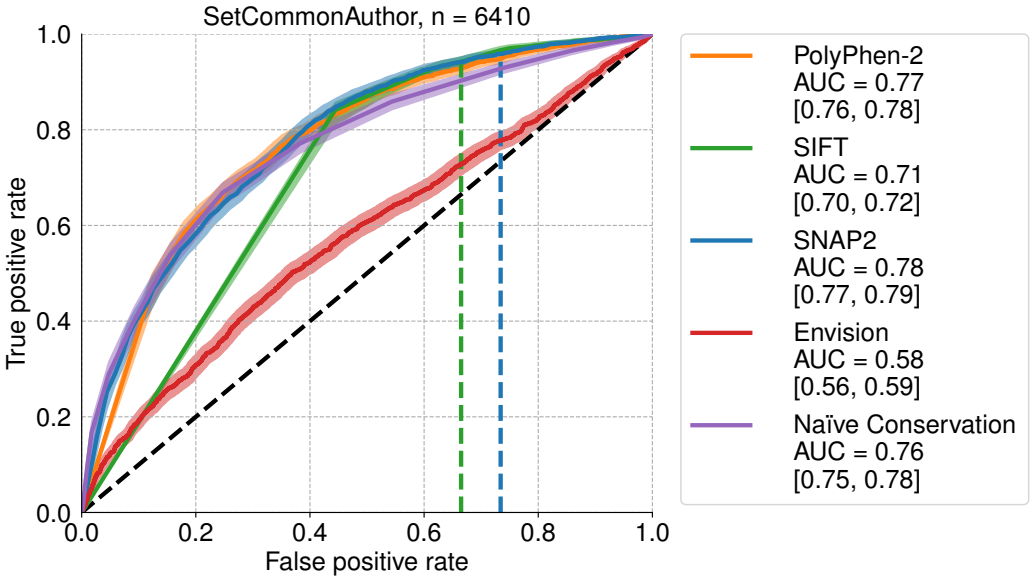
experimental data. It is reasonable to assume that these changes will not be wholly completed within a single generation of novel VEPs. However, if the field seeks to continue supporting human healthcare, these tasks must be tackled sooner rather than later. Finding the most suitable combination of predictions and experimental measurements will be the determinant for enabling true precision medicine.

Appendix A

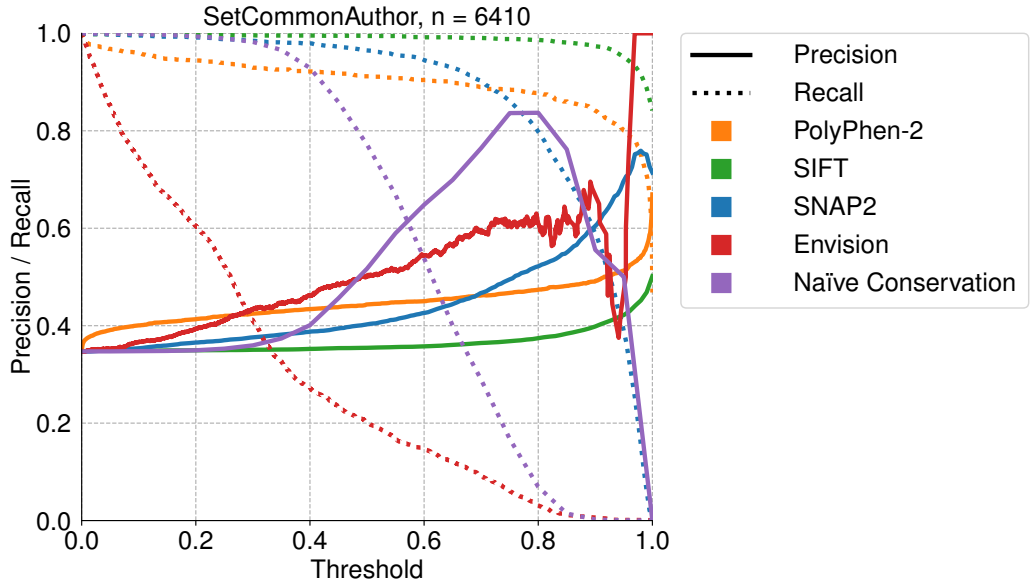
# Additional data

**Table A.1. SAV classifications provided by DMS study authors.** Detailed information on how classification of SAVs into either effect or neutral were performed for our analysis. References to columns or worksheets are indexed at 1 and refer to the files denoted in Table S10 of Reeb *et al.*, 2020

| Dataset identifier | Classification scheme |
|---|---|
| BRCA1 | Column 17 provides a classification into 3 bins: functional (FUNC), non-functional (LOF) and intermediate (INT). The average binned value for each position is the highest effect class encountered, i.e. INT if at least one INT, LOF if at least one LOF, otherwise FUNC. For our classification into neutral and effect, INT is discarded, FUNC is assigned as neutral, LOF as effect. |
| HSP82_2011 | The manuscript implies a classification of $< -0.4$ (deleterious), $< 0.1$ (neutral), $> 0.1$ (beneficial). The values of every codon are averaged. No beneficial effect variants exist. |
| HSP82_2013 | Author-provided binning from the manuscript based on the normalized selection coefficient $s$ is: beneficial ($s > 0.01$), wt-like ($-0.01 < s < 0.01$) deleterious ($-0.5 < s < -0.01$), strongly deleterious ($s < -0.5$). wt-like is mapped to neutral, the rest to effect. This score uses the same measurement as the one for correlation, i.e. the selection coefficient at 30°. No beneficial effect variants exist for this condition. |
| PTEN | Column 6 contains a classification into four states and is parsed into two binary scores. 'possibly low' and 'possibly wt' entries are ignored, 'wt' is assigned as neutral, 'low' as effect. |
| TPMT | Same as PTEN. |
| haeIIIM | The manuscript contains cutoffs for classification into deleterious ($\leq 0.6$), neutral $(]0.6, 1.1])$ and beneficial effects ($> 1.1$). Deleterious and beneficial are both assigned as effect. We use the strongest selection after 17 rounds. |




**(a)** Receiver operating characteristic curves




**(b)** Precision-recall curves

**Figure A.1. Classification performance on SAVs with class definitions by authors of DMS studies.** Analogously to Figure 4 in Reeb *et al.* (2020), the performance of five VEPs is assessed on set of SAVs that were categorized as either neutral or effect by authors of the respective DMS studies (see Table A.1). **(a)** shows ROC curves with shaded areas and numbers in parentheses denoting 95 % confidence intervals. **(b)** plots precision and recall against increasingly strict thresholds for considering a prediction as having an effect. Thresholds for every method were scaled to the interval $[0, 1]$ for plotting.

Appendix B

# Publication summaries with individual contributions

## B.1 Predicted Molecular Effects of Sequence Variants Link to System Level of Disease

This publication primarily investigates two aspects of variant effect prediction: (i) The behavior of VEPs regarding the prediction of disease variants in humans as well as animals, and (ii) the relationship of those and other variants to sequence conservation.

Disease variants in humans as found in databases like OMIM traditionally constituted a major part of VEPs' training data (Amberger *et al.*, 2019). Here, we first evaluated how a set of 5,661 SAVs from OMIM is predicted by the commonly used VEPs SNAP2, PolyPhen-2 and SIFT (Hecht *et al.*, 2015; Adzhubei *et al.*, 2013; Kumar *et al.*, 2009). All three showed high sensitivity, predicting at least 75 % of the SAVs to have an effect. However, an additional set of neutral variants highlighted that SIFT and PolyPhen-2 achieved their higher effect variant detection only at the cost of predicting too much effect in general. Interestingly, SNAP2 was re-trained for this analysis without disease SAVs and still predicted more effect for the set of OMIM SAVs than for its own training set of effect variants from other sources. An additional set of 117 disease SAVs from animals was manually curated from the OMIA database (Lenffer *et al.*, 2006; OMIA, 2019). While this sample size is too small to draw definite conclusions on larger sets, even more animal disease SAVs were predicted to have an effect than those from humans.

In experimental biology it is common to study variants outside of their natural host. For example, human diseases and their underlying SAVs are commonly studied in mouse models. We transferred the concept to an *in silico* evaluation by introducing OMIM SAVs into the respective mouse homologs. In this case, less variants in the animal model were predicted to have an effect compared to human. Reasons for this behavior could for example be slight changes between the human and homologous sequences which influence the SAVs' effects through a different epistatic environment.

The final analysis aimed to shed light on the importance and bias of VEPs regarding sequence conservation. We showed that OMIM SAVs when transferred to other positions in the sequence which were at least as conserved as the original position, were predicted with similar levels of effect. More so, random variants at other, at least equally conserved positions, were also predicted similarly. This behavior mostly perished when removing features related to sequence conservation from the SNAP2 prediction input. Together, these findings indicate that conservation is a primary driver of effect prediction. While the feature is clearly relevant, VEPs are easily biased to consider every change at conserved positions to have an effect.

Jonas Reeb (JR), Maximilian Hecht (MH), Yannick Mahlich and Burkhard Rost (BR) conceptualized the work. JR performed data curation, analyses, methodology, and visualization. MH provided help with SNAP2 and contributed to additional analyses and data curation. BR and Yana Bromberg provided supervision. BR provided funding. JR wrote the initial manuscript draft with BR. All authors reviewed and approved of the final manuscript.

## B.2 Variant effect predictions capture some aspects of deep mutational scanning experiments

DMS represents a novel experimental approach for the high-throughput measurement of SAV effects, utilizing the advances in next-generation sequencing. Due to the data's novelty only one VEP, Envision, has so far been trained on DMS SAVs. We harnessed this unique opportunity to provide an unbiased assessment of the established VEPs SNAP2, SIFT, and PolyPhen-2 together with Envision on a set of 17,781 SAVs with deleterious effect.

The effect signal of SAVs from DMS experiments was captured by VEPs only to a degree. Classification methods such as SNAP2 performed best when assessing the correlation between experimentally measured and predicted effect on a continuous scale (Spearman's $\rho = 0.41$, 95 % confidence interval $= [0.4, 0.42]$). Generally, classification VEPs performed better for SAVs with stronger observed effect and detected the increase in effect well. On the other hand, Envision which was trained as a regression predictor, appeared to best approximate the precise degree of effect (Mean squared error $= 0.09$ $[0.06, 0.07]$). However, this seemingly good performance is at least partially owed to not predicting any SAVs as having a high effect. Interestingly, a naïve prediction method based solely on PSI-BLAST sequence conservation scores exhibited the second best performance after SNAP2 ($\rho = 0.29$ $[0.28, 0.3]$).

Assessing VEPs on a classification task, i.e., predicting whether a SAV is neutral or has an effect, showed SNAP2 as the best performing method (AUC = 0.76 [0.75, 0.77]). Yet, this was the only VEP statistically significantly better than the naïve prediction approach (AUC = 0.73 [0.72, 0.74]). Envision, not trained on this task, performed better than random but clearly worse than all other VEPs. Additional analyses of 15,200 SAVs with beneficial effect showed severely reduced prediction performance by all methods, effectively on the level of random guessing.

Overall, this analysis provided unique insights regarding the performance of established VEPs on this new type of experimental dataset. Traditional classification methods and a novel regression approach showed complementary strengths. However, the agreement between independent experimental measurements was always higher than between predictions and experiments. Furthermore, a simplistic approach using only sequence conservation information already provided much of the signal captured. Finally, beneficial effect SAVs also highlighted that much remains to be done for future VEPs.

Jonas Reeb (JR) and Burkhard Rost (BR) conceptualized the initial idea for the work. Theresa Wirth performed preliminary analyses and data curation. JR expanded the concept with BR. JR established novel resources and performed data curation for new and final analyses. JR developed the methodology and visualization. BR provided supervision and funding. JR wrote the original manuscript draft with BR. All authors reviewed and approved of the final manuscript.

# References

Author lists have been abbreviated when longer than five items.


Adkar, B. V., Tripathi, A., Sahoo, A., *et al.* (2012). Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure*, **20**(2), 371–381. 49

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, **76**(1), 7.20.1–7.20.41. 16, 47, 77

Adzhubei, I. A., Schmidt, S., Peshkin, L., *et al.* (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, **7**(4), 248–249. 16

Aitman, T. J., Boone, C., Churchill, G. A., *et al.* (2011). The future of model organisms in human disease research. *Nature Reviews Genetics*, **12**(8), 575–582. 29

Alexandrov, L. B. and Stratton, M. R. (2014). Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics and Development*, **24**(1), 52–60. 2

Alhuzimi, E., Leal, L. G., Sternberg, M. J., and David, A. (2018). Properties of human genes guided by their enrichment in rare and common variants. *Human Mutation*, **39**(3), 365–370. 7

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402. 15

Amberger, J. S., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, **47**(D1), D1038–D1043. 8, 13, 77

Andersen, L. L., Terczyńska-Dyla, E., Mørk, N., *et al.* (2017). Frequently used bioinformatics tools overestimate the damaging effect of allelic variants. *Genes & Immunity*, **20**(1), 10–22. 20, 21

Anderson, D. and Lassmann, T. (2018). A phenotype centric benchmark of variant prioritisation tools. *npj Genomic Medicine*, **3**(1), 5. 21, 24, 27

Andreoletti, G., Pal, L. R., Moult, J., and Brenner, S. E. (2019). Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation*, **40**(9), 1197–1201. 25

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, **12**(7), 878. 18

Aravanis, A. M., Lee, M., and Klausner, R. D. (2017). Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell*, **168**(4), 571–574. 2

Araya, C. L. and Fowler, D. M. (2011). Deep mutational scanning: assessing protein function on a massive scale. *Trends in Biotechnology*, **29**(9), 435–442. 10

Araya, C. L., Fowler, D. M., Chen, W., *et al.* (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, **109**(42), 16858–16863. 49

Arthur, J. W., Cheung, F. S., and Reichardt, J. K. (2015). Single nucleotide differences (SNDs) continue to contaminate the dbSNP database with consequences for human genomics and health. *Human Mutation*, **36**(2), 196–199. 23

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, **17**, 507–522. 1

Bamshad, M. J., Ng, S. B., Bigham, A. W., *et al.* (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, **12**(11), 745–755. 10

Barmania, F. and Pepper, M. S. (2013). C-C chemokine receptor type five (CCR5): An emerging target for the control of HIV infection. *Applied and Translational Genomics*, **2**(1), 3–16. 7

Belkadi, A., Bolze, A., Itan, Y., *et al.* (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(17), 5473–5478. 23

Berman, H. M., Westbrook, J., Feng, Z., *et al.* (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**(1), 235–42. 9

Bishara, A. J. and Hittner, J. B. (2017). Confidence intervals for correlations when data are not normal. *Behavior Research Methods*, **49**(1), 294–309. 51

Bloom, J. D. (2015). Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, **16**(1), 1–13. 11, 59

Bouaoun, L., Sonkin, D., Ardin, M., *et al.* (2016). TP53 variations in human cancers: new lessons from the IARC TP53 Database and genomics data. *Human Mutation*, **5**(1), 7–20. 13

Bowden, R., Davies, R. W., Heger, A., *et al.* (2019). Sequencing of human genomes with nanopore technology. *Nature Communications*, **10**(1), 1869. 2

Brenan, L., Andreev, A., Cohen, O., *et al.* (2016). Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Reports*, **17**(4), 1171–1183. 49

Bromberg, Y. and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, **35**(11), 3823–35. 10, 17, 24

Bromberg, Y., Kahn, P. C., and Rost, B. (2013). Neutral and weakly nonneutral sequence variants may define individuality. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(35), 14255–14260. 6, 24

Brooks, P. J., Enoch, M. A., Goldman, D., Li, T. K., and Yokoyama, A. (2009). The alcohol flushing response: An unrecognized risk factor for esophageal cancer from alcohol consumption. *PLoS Medicine*, **6**(3), 0258–0263. 9

Burley, S. K., Berman, H. M., Bhikadiya, C., *et al.* (2019). RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, **47**(D1), D464–D474. 9

Butler, J. M., Hall, N., Narendran, N., Yang, Y. C., and Paraoan, L. (2017). Identification of candidate protective variants for common diseases and evaluation of their protective potential. *BMC Genomics*, **18**(1), 1–11. 8

Cao, Y., Sun, Y., Karimi, M., *et al.* (2019). Predicting Pathogenicity of Missense Variants with Weakly Supervised Regression. *bioRxiv*. 18

Capriotti, E., Montanucci, L., Profiti, G., *et al.* (2019a). Fido-SNP: the first webserver for scoring the impact of single nucleotide variants in the dog genome. *Nucleic Acids Research*, pages 1–6. 29

Capriotti, E., Ozturk, K., and Carter, H. (2019b). Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **11**(3), e1443. 55

Carss, K., Goldstein, D., Aggarwal, V., and Petrovski, S. (2019). Variant Interpretation and Genomic Medicine. In D. J. Balding, I. Moltke, and J. Marioni, editors, *Handbook of Statistical Genomics*, volume 2, chapter 27, pages 761–798. JohnWiley & Sons Ltd., 4th edition. 2, 9, 10, 24, 26

Chaisson, M. J., Wilson, R. K., and Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, **16**(11), 627–640. 5

Chiasson, M. and Fowler, D. M. (2019). Mutagenesis-based protein structure determination. *Nature Genetics*, **51**(July). 12

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, **10**(1), 1–17. 16

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, **7**(10). 15

Cieslik, M. and Chinnaiyan, A. M. (2020). Global genomics project unravels cancer's complexity at unprecedented scale. *Nature*, **578**(7793), 39–40. 2

Claussnitzer, M., Cho, J. H., Collins, R., *et al.* (2020). A brief history of human disease genetics. *Nature*, **577**(7789), 179–189. 1

Cline, M. S., Babbi, G., Bonache, S., *et al.* (2019). Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. *Human Mutation*, **40**(9), 1546–1556. 20, 25

Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003a). A vision for the future of genomics research. *Nature*, **431**(April), 835–847. 1

Collins, F. S., Morgan, M., and Patrinos, A. (2003b). The Human Genome Project: Lessons from large-scale biology. *Science*, **300**(5617), 286–290. 1

Collins, R. L., Brand, H., Karczewski, K. J., *et al.* (2019). An open resource of structural variation for medical and population genetics. *bioRxiv*. 5

Cunningham, B. and Wells, J. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, **244**(4908), 1081–1085. 10

Daneshjou, R., Wang, Y., Bromberg, Y., *et al.* (2017). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*, **38**(9), 1182–1192. 2, 24

Davis, R. H. (2004). The age of model organisms. *Nature Reviews Genetics*, **5**(1), 69–76. 29

Dean, M., Carrington, M., Winkler, C., *et al.* (1996). Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the CKR5 Structural Gene. *Science*, **273**(5283), 1856–1862. 7

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach. *Biometrics*, **44**(3), 837–845. 51

Ding, J. H., Li, S. P., Cao, H. X., *et al.* (2010). Alcohol dehydrogenase-2 and aldehyde dehydrogenase-2 genotypes, alcohol drinking and the risk for esophageal cancer in a Chinese population. *Journal of Human Genetics*, **55**(2), 97–102. 9

Dong, C., Wei, P., Jian, X., *et al.* (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, **24**(8), 2125–2137. 19

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113. 15

Eppig, J. T., Blake, J. A., Bult, C. J., *et al.* (2015). The Mouse Genome Database (MGD): Facilitating mouse as a model for human biology and disease. *Nucleic Acids Research*, **43**(D1), D726–D736. 13

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*. 18

Esposito, D., Weile, J., Shendure, J., *et al.* (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, **20**(1), 223. 14, 48

EVA (2019). `https://www.ebi.ac.uk/eva/`. Accessed 2019/10/31. 13

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, **40**(Database issue), D136–43.

30

Fenwick, A. L., Goos, J. A., Rankin, J., *et al.* (2014). Apparently synonymous substitutions in FGFR2 affect splicing and result in mild Crouzon syndrome. *BMC Medical Genetics*, **15**(1), 1–6. 8

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, **7**(2), 85–97. 5, 7

Findlay, G. M., Daza, R. M., Martin, B., *et al.* (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, **562**(7726), 217–222. 22, 49

Flygare, S., Hernandez, E. J., Phan, L., *et al.* (2018). The VAAST Variant Prioritizer (VVP): Ultrafast, easy to use whole genome variant prioritization tool. *BMC Bioinformatics*, **19**(1), 1–13. 19

Forsyth, C. M., Juan, V., Akamatsu, Y., *et al.* (2013). Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *mAbs*, **5**(4), 523–532. 12

Fowler, D. M. and Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nature Methods*, **11**(8), 801–807. 3, 10, 20

Fowler, D. M., Stephany, J. J., and Fields, S. (2014). Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature Protocols*, **9**(9), 2267–2284. 11

Freeman, J. L. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, **16**(8), 949–961. 8

Fujino, Y., Fujita, R., Wada, K., *et al.* (2012). Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochemical and Biophysical Research Communications*, **428**(3), 395–400. 12

Gawad, C., Koh, W., and Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics*, **17**(3), 175–188. 2

Gelman, H., Dines, J. N., Berg, J., *et al.* (2019). Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Medicine*, **11**(1), 1–11. 27

Genomics England (2017). The 100,000 Genomes Project Protocol v3. `https://doi.org/10.6084/m9.figshare.4530893.v2`. Accessed 2020/06/19. 3

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics*, **18**(4), 309–317. 24

González-Pérez, A. and López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *The American Journal of Human Genetics*, **88**(4), 440–449. 19

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, **17**(6), 333–351. 2

Gray, V. E., Hause, R. J., and Fowler, D. M. (2017). Analysis of large-scale mutagenesis data to assess the impact of single amino acid substitutions. *Genetics*, **207**(1), 53–61. 10

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. (2018). Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Systems*, **6**(1), 116–124.e3. 19, 23, 24, 47, 48, 52

Grimm, D. G., Azencott, C. A., Aicheler, F., *et al.* (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, **36**(5), 513–523. 24, 25

Groß, C., de Ridder, D., and Reinders, M. (2018). Predicting variant deleteriousness in non-human species: Applying the CADD approach in mouse. *BMC Bioinformatics*, **19**(1), 1–10. 29

Gunning, A. C., Fryer, V., Fasham, J., *et al.* (2020). Assessing performance of pathogenicity predictors using clinically-relevant variant datasets. *bioRxiv*, pages 1–12. 26

Harrison, S. M. and Rehm, H. L. (2019). Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. *Genome Medicine*, **11**(1), 72. 23

Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, **16**(Suppl 8), S1. 17, 47, 77

Heinzinger, M., Elnaggar, A., Wang, Y., *et al.* (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, **20**(1), 723. 18

Heredia, J. D., Park, J., Brubaker, R. J., *et al.* (2018). Mapping Interaction Sites on Human Chemokine Receptors by Deep Mutational Scanning. *The Journal of Immunology*, **200**(11), 3825–3839. 49

Hernandez, R. D., Uricchio, L. H., Hartman, K., *et al.* (2019). Ultrarare variants drive substantial cis heritability of human gene expression. *Nature Genetics*, **51**(9), 1349–1355. 7

Hietpas, R., Roscoe, B., Jiang, L., and Bolon, D. N. (2012). Fitness analyses of all possible point mutations for regions of genes in yeast. *Nature Protocols*, **7**(7), 1382–1396. 11

Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. (2011). Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*, **108**(19), 7896–7901. 49

Hietpas, R. T., Bank, C., Jensen, J. D., and Bolon, D. N. A. (2013). Shifting fitness landscapes in response to altered environments. *Evolution*, **67**(12), 3512–3522. 49

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., *et al.* (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, **35**(2), 128–135. 24

Hu, Z., Yu, C., Furutsuki, M., *et al.* (2019). VIPdb, a genetic Variant Impact Predictor Database. *Human Mutation*, **40**(9). 15

Hurley, T. D., Perez-Miller, S., and Breen, H. (2001). Order and disorder in mitochondrial aldehyde dehydrogenase. *Chemico-Biological Interactions*, **130-132**, 3–14. 9

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., *et al.* (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics*, **99**(4), 877–885. 19

J. Shendure and J. M. Akey (2015). The origins, determinants, and consequences of human mutations. *Science*, **349**(6255), 1478–1483. 2, 4, 23

Jensen, L. J. and Bateman, A. (2011). The rise and fall of supervised machine learning techniques. *Bioinformatics*, **27**(24), 3331–3332. 16

Jiang, L., Mishra, P., Hietpas, R. T., Zeldovich, K. B., and Bolon, D. N. A. (2013). Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLoS Genetics*, **9**(6), e1003600. 49, 55

Jurtz, V. I., Johansen, A. R., Nielsen, M., *et al.* (2017). An introduction to deep learning on biological sequence data: Examples and solutions. *Bioinformatics*, **33**(22), 3685–3690. 18

Karczewski, K. J., Francioli, L. C., Tiao, G., *et al.* (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. 14

Khera, A. V., Chaffin, M., Aragam, K. G., *et al.* (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, **50**(9), 1219–1224. 7

Kiezun, A., Garimella, K., Do, R., *et al.* (2012). Exome sequencing and the genetic basis of complex traits. *Nature Genetics*, **44**(6), 623–630. 10

Kim, Y.-e., Ki, C.-s., and Jang, M.-a. (2019). Challenges and Considerations in Sequence Variant Interpretation for Mendelian Disorders. *Annals of Laboratory Medicine*, **39**(5), 421. 21

Kircher, M., Witten, D. M., Jain, P., *et al.* (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, **46**(3), 310–5. 18

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. (2014). Massively Parallel Single Amino

Acid Mutagenesis. *Nature Methods*, **44**(12), 3516–3521. 49

Klesmith, J. R. and Hackel, B. J. (2019). Improved mutant function prediction via PACT: Protein Analysis and Classifier Toolkit. *Bioinformatics*, pages 1–6. 11, 59

Klesmith, J. R., Bacik, J. P., Michalczyk, R., and Whitehead, T. A. (2015). Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. *ACS Synthetic Biology*, **4**(11), 1235–1243. 49

Kolodziejczyk, A., Kim, J. K., Svensson, V., Marioni, J., and Teichmann, S. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, **58**(4), 610–620. 2

Kowarsch, A., Fuchs, A., Frishman, D., and Pagel, P. (2010). Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions. *PLoS Computational Biology*, **6**(9), e1000923. 24

Krupovic, M. and Bamford, D. H. (2011). *Protein Conservation in Virus Evolution*. John Wiley & Sons, Ltd, Chichester, UK. 60

Kulandaisamy, A., Binny Priya, S., Sakthivel, R., *et al.* (2018). MutHTP: Mutations in human transmembrane proteins. *Bioinformatics*, **34**(13), 2325–2326. 14

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, **4**(8), 1073–1081. 15, 47, 77

Laddach, A., Ng, J. C.-F., and Fraternali, F. (2019). Missense variants in health and disease target distinct functional pathways and proteomics features. *bioRxiv*. 7

Landrum, M. J., Lee, J. M., Benson, M., *et al.* (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, **44**(D1), D862–D868. 1, 13

Lappalainen, I., Lopez, J., Skipper, L., *et al.* (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Research*, **41**(D1), 936–941. 13

Larrañaga, P., Calvo, B., Santana, R., *et al.* (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, **7**(1), 86–112. 16

Larson, H. N., Weiner, H., and Hurley, T. D. (2005). Disruption of the coenzyme binding site and dimer interface revealed in the crystal structure of mitochondrial aldehyde dehydrogenase "Asian" variant. *Journal of Biological Chemistry*, **280**(34), 30550–30556. 9

Lauschke, V. M. and Ingelman-Sundberg, M. (2020). Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. *npj Genomic Medicine*, **5**(1), 9. 27

Lek, M., Karczewski, K. J., Minikel, E. V., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), 285–291. 1, 5

Lenffer, J., Nicholas, F. W., Castle, K., *et al.* (2006). OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Research*, **34**(Database issue), D599–601. 13, 30, 77

Levy, S. E. and Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*, **17**(1), 95–115. 2

Li, B., Krishnan, V. G., Mort, M. E., *et al.* (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, **25**(21), 2744–2750. 18

Li, M. J., Wang, L. Y., Xia, Z., Sham, P. C., and Wang, J. (2013). GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Research*, **41**(Web Server issue), 150–158. 19

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, **37**(3), 235–241. 14, 19

Livesey, B. J. and Marsh, J. A. (2019). Using deep mutational scanning data to benchmark computational phenotype predictors and identify pathogenic missense mutations. *bioRxiv*. 56, 60

Mackay, T. F. C., Stone, E. a., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, **10**(8), 565–77. 7

Mahlich, Y., Reeb, J., Hecht, M., *et al.* (2017). Common sequence variants affect molecular function more than rare variants? *Scientific Reports*, **7**(1), 1608. 7, 24, 26

Majithia, A. R., Tsuda, B., Agostini, M., *et al.* (2016). Prospective functional classification of all possible missense variants in PPARG. *Nature Genetics*, **48**(12), 1570–1575. 49

Manolio, T. a., Collins, F. S., Cox, N. J., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–53. 7, 9

Manolio, T. A., Fowler, D. M., Starita, L. M., *et al.* (2017). Bedside Back to Bench: Building Bridges between Basic and Clinical Genomic Research. *Cell*, **169**(1), 6–12. 2, 4

Marjoram, P., Zubair, A., and Nuzhdin, S. V. (2014). Post-GWAS: where next? More samples, more SNPs or more biology? *Heredity*, **112**(1), 79–88. 9

Marouli, E., Graff, M., Medina-Gomez, C., *et al.* (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, **542**(7640), 186–190. 7

Martinez, D. a. and Nelson, M. A. (2010). The next generation becomes the now generation. *PLoS genetics*, **6**(4), e1000906. 3

Masica, D. L. and Karchin, R. (2016). Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. *PLOS Computational Biology*, **12**(5), e1004725. 24

Matreyek, K. A., Starita, L. M., Stephany, J. J., *et al.* (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics*, **50**(6), 874–882. 12, 49

Mavor, D., Barlow, K., Thompson, S., *et al.* (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*, **5**(MAY2016), 1–23. 11, 12, 21

Mavor, D., Barlow, K. A., Asarnow, D., *et al.* (2018). Extending chemical perturbations of the ubiquitin fitness landscape in a classroom setting reveals new constraints on sequence tolerance. *Biology Open*, **7**(7), bio036103. 12, 21

Mavor, D., Bolon, D. N. A., and Mishra, P. (2019). Mapping the growth effect of previously hidden ubiquitin alleles using an overexpression based mutational scan. *bioRxiv*. 21

Miller, M., Bromberg, Y., and Swint-Kruse, L. (2017). Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Scientific Reports*, **7**(December 2016), 41329. 20

Miller, M., Vitale, D., Kahn, P. C., Rost, B., and Bromberg, Y. (2019a). Funtrp: Identifying Protein Positions for Variation Driven Functional Tuning. *Nucleic Acids Research*. 20

Miller, M., Wang, Y., and Bromberg, Y. (2019b). What went wrong with variant effect predictor performance for the PCM1 challenge. *Human Mutation*, **40**(9), 1486–1494. 20, 25

Monzon, A. M., Carraro, M., Chiricosta, L., *et al.* (2019). Performance of computational methods for the evaluation of pericentriolar material 1 missense variants in CAGI-5. *Human Mutation*, **40**(9), 1474–1485. 25

Morganti, S., Tarantino, P., Ferraro, E., *et al.* (2020). Role of Next-Generation Sequencing Technologies in Personalized Medicine. In G. Pravettoni and S. Triberti, editors, *P5 eHealth: An Agenda for the Health Technologies of the Future*, chapter 8, pages 125–154. Springer International Publishing, Cham. 1

Mort, M., Sterne-Weiler, T., Li, B., *et al.* (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biology*, **15**(1), R19. 18

Müller, B. and Grossniklaus, U. (2010). Model organisms - A historical perspective. *Journal of Proteomics*, **73**(11), 2054–2063. 29

Myers, R. H. (2004). Huntington's Disease Genetics. *NeuroRx*, **1**(2), 255–262. 7

Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic*

*Acids Research*, **31**(13), 3812–3814. 15

Ng, P. C. and Steven, H. (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Research*, **11**(5), 863–874. 2

Niroula, A. and Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? *PLoS Computational Biology*, **15**(2), 1–17. 21

OMIA (2019). `https://omia.org/`. Accessed 2019/10/31. 13, 77

Orioli, T. and Vihinen, M. (2019). Benchmarking subcellular localization and variant tolerance predictors on membrane proteins. *BMC Genomics*, **20**(S8), 1–15. 23

Pagel, K. A., Pejaver, V., Lin, G. N., *et al.* (2017). When loss-of-function is loss of function : assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics*, **33**. 18

Pagel, K. A., Antaki, D., Lian, A., *et al.* (2019). Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLOS Computational Biology*, **15**(6), e1007112. 18

Pavlopoulos, G. A., Oulas, A., Iacucci, E., *et al.* (2013). Unraveling genomic variation from next generation sequencing data. *Biodata Mining*, **6**. 4

Pedregosa, F., Michel, V., Grisel, O., *et al.* (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830. 50, 51

Pejaver, V., Urresti, J., Lugo-Martinez, J., *et al.* (2017). MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv*, pages 1–28. 18

Pejaver, V., Babbi, G., Casadio, R., *et al.* (2019). Assessment of methods for predicting the effects of PTEN and TPMT protein variants. *Human Mutation*, **40**(9), 1495–1506. 25

Pfeiffer, F., Gröber, C., Blank, M., *et al.* (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, **8**(1), 1–14. 2

Plomin, R., Haworth, C. M. a., and Davis, O. S. P. (2009). Common disorders are quantitative traits. *Nature Reviews Genetics*, **10**(12), 872–8. 7

precisionFDA (2019). `https://precision.fda.gov/`. Accessed 2019/10/30. 23

Procko, E., Hedman, R., Hamilton, K., *et al.* (2013). Computational design of a protein-based enzyme inhibitor. *Journal of Molecular Biology*, **425**(18), 3563–3575. 12

Punta, M., Kloppmann, E., and Reeb, J. (2019). Membrane Protein Structure. In *Encyclopedia of Biophysics*. Springer, Berlin. 23

R Core Team (2020). R: A language and environment for statistical computing. http://www.r-project.org. 51

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, **30**(17), 3894–3900. 2

Reeb, J., Kloppmann, E., Bernhofer, M., and Rost, B. (2014). Evaluation of transmembrane helix predictions in 2014. *Proteins: Structure, Function, and Bioinformatics*, **83**(3), 473–484. 23

Reeb, J., Hecht, M., Mahlich, Y., Bromberg, Y., and Rost, B. (2016). Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLOS Computational Biology*, **12**(8), e1005047. 26

Reeb, J., Wirth, T., and Rost, B. (2020). Variant effect predictions capture some aspects of deep mutational scanning experiments. *BMC Bioinformatics*, **21**(1), 107. 48, 52, 57, 58, 75, 76

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, **47**(D1), D886–D894. 10, 18, 21

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, **39**(17), 37–43. 15

Richards, S., Aziz, N., Bale, S., *et al.* (2015). Standards and guidelines for the interpretation of sequence

variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, **17**(5), 405–423. 4, 21, 23

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. (2017). Deep generative models of genetic variation capture mutation effects. *arXiv*, pages 1–25. 18

Rives, A., Goyal, S., Meier, J., *et al.* (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*. 18

Rockah-Shmuel, L., Tóth-Petróczy, Á., and Tawfik, D. S. (2015). Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Computational Biology*, **11**(8), 1–28. 49

Rogers, J. M., Passioura, T., and Suga, H. (2018). Nonproteinogenic deep mutational scanning of linear and cyclic peptides. *Proceedings of the National Academy of Sciences*, **115**(43), 201809901. 12

Rollins, N. J., Brock, K. P., Poelwijk, F. J., *et al.* (2019). Inferring protein 3D structure from deep mutation scans. *Nature Genetics*, **51**(July). 12

Romero, P. A., Tran, T. M., and Abate, A. R. (2015). Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, **112**(23), 7159–7164. 49

Roscoe, B. P. and Bolon, D. N. A. (2014). Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *Journal of Molecular Biology*, **18**(9), 1199–1216. 22

Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. A. (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *Journal of Molecular Biology*, **425**(8), 1363–1377. 21

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Bioinformatics*, **19**(1), 55–72. 18

Rost, B., Radivojac, P., and Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. *FEBS Letters*, **590**, 2327–2341. 1

Rubin, A. F., Gelman, H., Lucas, N., *et al.* (2017). A statistical framework for analyzing deep mutational scanning data. *Genome Biology*, **18**(1), 1–15. 11, 59

Russell, J. J., Theriot, J. A., Sood, P., *et al.* (2017). Non-model model organisms. *BMC Biology*, **15**(1), 55. 29

Sarkar, A., Yang, Y., and Vihinen, M. (2019). Variation Benchmark Datasets: Update, Criteria, Quality and Applications. *bioRxiv*, page 634766. 14

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., *et al.* (2016). Local fitness landscape of the green fluorescent protein. *Nature*, **533**(7603), 397–401. 49

Sasidharan Nair, P. and Vihinen, M. (2013). VariBench: A Benchmark Database for Variations. *Human Mutation*, **34**(1), 42–49. 14

Saudou, F. and Humbert, S. (2016). The Biology of Huntingtin. *Neuron*, **89**(5), 910–926. 7

Sauna, Z. E. and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, **12**(10), 683–91. 8

Schmiedel, J. M. and Lehner, B. (2019). Determining protein structures using deep mutagenesis. *Nature Genetics*, **51**(July). 12

Schneider, V. A., Graves-Lindsay, T., Howe, K., *et al.* (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, **27**(5), 849–864. 4

Schrödinger LLC (2019). The PyMOL Molecular Graphics System, Version 2.2 Schrödinger, LLC. 9

Schwarze, K., Buchanan, J., Taylor, J. C., and Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine*, **20**(10),

1122–1130. 4

Schwarze, K., Buchanan, J., Fermont, J. M., *et al.* (2019). The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, **0**(0), 1–10. 4

Shamsi, Z., Chan, M., and Shukla, D. (2020). TLmutation : predicting the effects of mutations using transfer learning . *bioRxiv*. 18

Shendure, J., Findlay, G. M., and Snyder, M. W. (2019). Genomic Medicine–Progress, Pitfalls, and Promise. *Cell*, **177**(1), 45–57. 1

Sherman, R. M. and Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews Genetics*, pages 1–12. 4

Sherry, S. T., Ward, M.-H., Kholodov, M., *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**(1), 308–311. 8, 13

Shimoyama, M., De Pons, J., Hayman, G. T., *et al.* (2015). The Rat Genome Database 2015: Genomic, phenotypic and environmental variations and disease. *Nucleic Acids Research*, **43**(D1), D743–D750. 13

Sruthi, C. K. and Prakash, M. (2020). Deep2Full: Evaluating strategies for selecting the minimal mutational experiments for optimal computational predictions of deep mutational scan outcomes. *PLoS ONE*, **15**(1), e0227621. 27, 52

Starita, L. M., Pruneda, J. N., Lo, R. S., *et al.* (2013). Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, **110**(14), E1263–E1272. 49

Starita, L. M., Young, D. L., Islam, M., *et al.* (2015). Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*, **200**(2), 413–422. 22, 49

Starita, L. M., Ahituv, N., Dunham, M. J., *et al.* (2017). Variant Interpretation: Functional Assays to the Rescue. *American Journal of Human Genetics*, **101**(3), 315–325. 10, 59

Stein, A., Fowler, D. M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends in Biochemical Sciences*, **44**(7), 575–588. 20, 27

Stenson, P. D., Mort, M., Ball, E. V., *et al.* (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, **136**(6), 665–677. 13

Stiffler, M., Hekstra, D., and Ranganathan, R. (2015). Evolvability as a Function of Purifying Selection in TEM-1 $\beta$-Lactamase. *Cell*, **160**(5), 882–892. 49

Sun, H. and Yu, G. (2019). New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Scientific Reports*, **9**(1), 1–11. 20, 21, 23, 24

Sundaram, L., Gao, H., Padigepati, S. R., *et al.* (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, **50**(8), 1161–1170. 10

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–73. 3

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65. 3

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74. 3, 4, 5, 8

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**(7402), 207–214. 3

The Integrative HMP (iHMP) Research Network Consortium (2014). The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease.

*Cell Host & Microbe*, **16**(3), 276–289. 3

The Uniprot Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**(D1), D506–D515. 1, 13

Tinberg, C. E., Khare, S. D., Dou, J., *et al.* (2013). Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, **501**(7466), 212–216. 12

Traxlmayr, M. W., Hasenhindl, C., Hackl, M., *et al.* (2012). Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. *Journal of Molecular Biology*, **423**(3), 397–412. 49

Turck, N., Vutskits, L., Sanchez-Pena, P., *et al.* (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **8**, 12–77. 51

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, **34**(9), 666–681. 2

Vihinen, M. (2020). Problems in variation interpretation guidelines and in their implementation in computational tools. *Molecular Genetics & Genomic Medicine*, (February), 1–10. 24

Virtanen, P., Gommers, R., Oliphant, T. E., *et al.* (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, **17**(3), 261–272. 50

Wagenaar, T. R., Ma, L., Roscoe, B., *et al.* (2014). Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell and Melanoma Research*, **27**(1), 124–133. 12

Wetterstrand (2019). `https://genome.gov/sequencingcostsdata`. Accessed 2019/11/19. 3

Whitehead, T. A., Chevalier, A., Song, Y., *et al.* (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnology*, **30**(6), 543–548. 12

Wilcox, R. R. (2016). Comparing dependent robust correlations. *British Journal of Mathematical and Statistical Psychology*, **69**(3), 215–224. 50

Wray, N. R., Yang, J., Hayes, B. J., *et al.* (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, **14**(7), 507–515. 24

Wrenbeck, E. E., Faber, M. S., and Whitehead, T. A. (2017a). Deep sequencing methods for protein engineering and design. *Current Opinion in Structural Biology*, **45**, 36–44. 11

Wrenbeck, E. E., Azouz, L. R., and Whitehead, T. A. (2017b). Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature Communications*, **8**, 1–10. 12

Wu, N. C., Young, A. P., Dandekar, S., *et al.* (2013). Systematic Identification of H274Y Compensatory Mutations in Influenza A Virus Neuraminidase by High-Throughput Screening. *Journal of Virology*, **87**(2), 1193–1199. 12

Wu, Y., Weile, J., Cote, A. G., *et al.* (2019). A web application and service for imputing and visualizing missense variant effect maps. *Bioinformatics*, **35**(January), 3191–3193. 11

Yachdav, G., Kloppmann, E., Kajan, L., *et al.* (2014). PredictProtein-an open resource for online prediction of protein structural and functional features. *Nucleic Acids Research*, **42**(Web Server issue), W337–43. 18

Zhang, J., Kinch, L. N., Cong, Q., *et al.* (2019). Assessing predictions on fitness effects of missense variants in calmodulin. *Human Mutation*, **40**(9), 1463–1473. 20, 25

Zook, J. M., Chapman, B., Wang, J., *et al.* (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, **32**(3), 246–251. 23

Zook, J. M., McDaniel, J., Olson, N. D., *et al.* (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, **37**(5), 561–566. 23