# Identifying genotype-phenotype correlations via integrative mutation analysis

Edward Airey1,2,3, Stephanie Portelli1,2,3, Joicymara S. Xavier4, YooChan Myung1,2,3, Michael Silk1,2,3, Malancha Karmakar1,2,3, João P. L. Velloso4, Carlos H. M. Rodrigues1,2,3, Hardik H. Parate1,2,3, Anjali Garg1,2,3, Raghad Al-Jarf1,2,3, Lucy Barr3, Juliana A. Geraldo4, Pâmela M. Rezende4, Douglas E.V. Pires1,3,5,\*, David B. Ascher1,2,3,6,\*

1Structural Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Australia;

2ACRF Facility for Innovative Cancer Drug Discovery, Bio21 Institute, University of Melbourne, Australia;

3Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Australia.

4Instituto René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, 30190-002, Brazil;

5School of Computing and Information Systems, University of Melbourne, Australia;

6Department of Biochemistry, Cambridge University, UK

\*To whom correspondence should be addressed D.B.A. Tel: +61 90354794; Email: david.ascher@unimelb.edu.au or da382@cam.ac.uk. Correspondence may also be addressed to D.E.V.P. douglas.pires@unimelb.edu.au.

ORCID ID

Douglas E.V. Pires: 0000-0002-3004-2119

David B. Ascher: 0000-0003-2948-2413

## Running Head

Genotype-phenotype correlations via integrative mutation analysis

## Abstract

Mutations in protein coding regions can lead to large biological changes and are associated with genetic conditions, including cancers and Mendelian diseases, as well as drug resistance. Although whole genome and exome sequencing has been helping to elucidate potential genotype-phenotype correlations, there is a large gap between the identification of new variants and deciphering their molecular consequences. A comprehensive understanding of these mechanistic consequences is crucial to better understand and treat diseases in a more personalised and effective way. This is particularly relevant considering estimates that over 80% of mutations associated with a disease are incorrectly assumed to be causative. A thorough analysis of potential effects of mutations is required to correctly identify the molecular mechanisms of disease and enable the distinction between disease-causing and non-disease-causing variation within a gene. Here we present an overview of our integrative mutation analysis platform, which focuses on refining the current genotype-phenotype correlation methods by using the wealth of protein structural information.

## Keywords

Genotype-phenotype correlations, graph-based signatures, mCSM, mutation, protein structure, protein interactions.

## Introduction

Proteins are versatile molecules, responsible for orchestrating a wide range of biological processes. They are comprised of a single polypeptide chain of amino acids, which fold in 3D space into dynamic structures. How a protein folds is important for determining its functions, including activities and interactions with other molecules. These structures are highly coordinated and conserved across evolution, and small perturbations in the amino acid sequence can disrupt these shapes, functions and interactions[1,2]. While missense mutations, causing a change to a single amino acid, are generally less structurally disruptive than nonsense mutations, their effects are highly variable and can be wide-ranging, making their molecular consequences harder to determine. Despite their subtle effects, missense substitutions are related with many different genetic conditions, including cancer, Mendelian diseases as well as the emergence of drug resistance.

The introduction of a missense mutation can have many molecular effects, including altering how the protein folds, its dynamics, post-translational modifications, half-life, localisation, activity and molecular interactions[3]. When analysing a new mutation, an integrative approach is therefore important to consider the effects it might have on all of these aspects. This enables the identification of specific functional, and structural changes imparted by the mutations, which is essential for a molecular understanding. This can also explain why mutations in the same protein might lead to different diseases, why mutations might cluster in 3D space and how those genetic changes present phenotypically.

Although many assume that an unfavourable phenotype (*e.g.*, pathogenic, drug-resistant) is the result of large, overall destabilizing mutations, mutations with milder effects are often more prevalent in a population, as they are generally under less selective pressure [4,5]. For example, when we analysed mutations in three different tuberculosis proteins that lead to resistance, we showed that the most frequent resistant mutations were more likely to be associated with overall mild functional effects, and associated reduced fitness cost, allowing for increased prevalence within the bacterial population [4].

Experimentally elucidating the biophysical effects of mutations is an expensive and time-consuming task, usually limited to a small number of variants in proteins with amenable assays. Over the years, the accumulation of information of experimentally characterised mutations has enabled the development and improvement of computational mutational analysis tools[6]. These computational platforms have shown to be invaluable assets to decipher genotype-phenotype correlations in cancer[7-18], Mendelian diseases[19-25] and detection of antimicrobial resistance[26-28,14,29-31,4,32-34], helping to guide clinical decisions and drive further research. Here, we introduce a general computational pipeline that uses *in silico* biophysical predictions and machine learning approaches to best harness the wealth of available biological and protein structural information to give insights into genotype-phenotype correlation for clinical use[10].

The mutation cutoff scanning matrix (mCSM) platform is the only comprehensive collection of *in silico* tools for quantitatively predicting the effects of missense mutations on protein folding, structure, dynamics and interactions. This platform includes tools which calculate all possible molecular interactions (Arpeggio[35]), account for changes in protein stability (mCSM-Stability[36], SDM[37], DUET[38]), dynamics (DynaMut[39]), protein interactions with other proteins (mCSM-PPI[36], mCSM-PPI2[40], mCSM-AB[41]), nucleic acids (mCSM-DNA[36], mCSM-NA[42]), and small molecule ligands (mCSM-lig[43], CSM-lig[44]).

These tools were built using the concept of graph-based signatures, which represent the geometry and physicochemical properties of the wildtype protein structure environment as a network or graph, composed of a series of nodes, describing the local mutation environment, and edges, describing the distances between interacting ‘layers’ of surrounding residues. Information on the mutation is captured using the pharmacophore change between the wild-type and the mutant residue, including whether charges or hydrogen donors/acceptors have been gained or lost[36].

This platform allows for accurate biophysical predictions, which when complemented with other protein analytical tools, can provide a detailed landscape on the specific mutational effects on a protein. We have implemented these within an analytical and supervised machine learning predictive pipeline (Figure 1), to enable easy and fast characterisation of novel mutations and their likely clinical phenotypes. This approach has been shown to have big implications in diagnostic and personalised medicine in the post-genomic era.

## Materials

## Data curation

2.1.1. Mutation curation

The foremost requirement for training a machine learning model is the quality and availability of appropriate experimental/clinical data, with suitable representation of the classes under comparison. With respect to human disease, a wealth of freely accessible collections of curated data exist. Previously reported mutations through publications and functional studies are available from dbSNP[45], the largest freely available repository of genetic variation. Variants with evidence of pathogenicity can be viewed from the Human Gene Mutation Database (HGMD)[46] and ClinVar[47], and from disease-specific datasets such as the Catalogue of Somatic Mutations in Cancer (COSMIC). Standing variation is available from genomic sequencing efforts of healthy populations, including over 140,000 healthy humans in gnomAD[48] and 50,000 whole exomes currently available in UK Biobank[49].

When combining data from multiple sources, it is important that every datapoint is comparable to the others. If using genetic coordinates, it is important to ensure that they are found on the same assembly of the genome (GRCh38 vs GRCh37, for example). The mutations themselves (whether reported as genetic or amino acid changes) must be reported on the same transcript, as most genes have multiple reported coding sequences.

2.1.2. Protein structure curation

The sequence and functional information for the specific protein of interest can be obtained from Uniprot (<https://www.uniprot.org/>)[50]. To run the mCSM tools we need either crystallographic structures, which can be downloaded from the protein data bank (PDB;<http://www.rcsb.org/>)[51] or generated via homology modelling or molecular docking (to run mCSM-PPI, mCSM-Lig or mCSM-NA). Once we have the variant information collected from the resources in 2.1.1., we map these variants on to the identified protein structures to help us visualize the spread and also identify potential hotspots. This step can easily be done using visualisation software such as PyMol, as it enables selection of residues being mutated in a 3D manner.

## An overview of computational tools to analyse missense mutations

Over the past two decades there has been an unprecedented growth in both computational power and the amount of biological data available. This has facilitated the development of numerous sequence (Table 1) and structural (Table 2) based computational tools to guide mutation characterisation.

The mCSM platform is the only available approach to consider all possible molecular effects, and has therefore formed the central component of our mutational analysis pipeline. All mCSM Platform tools are available freely as websites compatible with most web-browsers, but Google Chrome is recommended. A summary of these methods and each link to access them is described in Table 3.

## Methods

## Predicting and Analysing Structural and Biophysical Effects of Mutations using the mCSM Platform

The mCSM methods can be divided based on their specific purpose. As shown in Figure 1, methods are chosen depending on interactions made, and what structural information is available. Below we discuss how each type of predictor can be used and interpreted.

* The user should choose the appropriate tools based on what information is available on their protein of interest (Figure 1).
* In general, each mCSM tool requires a wild-type protein file, in the PDB format, and the single-point mutation or a list of mutations. Some tools may require additional specific information; Table 4 shows the inputs required for each tool. Notes 1 and 2 highlight some common issues with the submission inputs.

## mCSM Platform Output

3.2.1. Arpeggio

The results of Arpeggio are shown in Figure 2.

* After submitting a job, an overview of the type and number of atomic interactions within the protein are shown (Figure 2A). Arpeggio calculates all types of molecular interactions (Table 5), which are displayed and downloadable along with a visual representation of the atomic contacts overlaid on the protein structure (Figure 2B).
* The number of each interaction/contact and PyMOL session files can be downloaded for a more detailed analysis.

3.2.2. MTR-Viewer

Gene viewer

* The MTR gene viewer[5] results page (Figure 3) shows predicted MTR scores in an interactive line graph with a control panel which allows users to adjust the window size (31 codons is selected by default), and the ethnicity (all populations shown by default) for MTR estimates. The line graph (Figure 3A) displays the regions that have high variation, low-MTR scored; thosein red are most likely to be pathogenic. If the user selects a specific ethnicity, from Latino, Non-Finnish European and South Asian, the ethnicity-specific MTR scores are shown in blue on the line graph.
* The first lollipop plot (Figure 3B) shows observed missense (yellow) and synonymous (green) variations based on gnomeAD.
* If the gene of interest is a ClinVar pathogenic gene, their pathogenic (red) and benign (blue) missense variants are also displayed under the gnomeAD lollipop plot (Figure 3C).
* Users can browse results of alternative-transcript (Figure 3D) of the given query if available.

Variant query

* The variant query result page (Figure 4) shows MTR scores for each user-supplied missense variant, providing the estimated regional intolerance. Low MTR scores indicate stronger purifying selection within the population. Users can also press 'view' next to a variant to show its position within its gene transcript.

3.2.3. mCSM-Stability/PPI/DNA

The impact of mutations on protein stability, protein-protein binding affinity and protein-DNA affinity can be predicted by mCSM-Stability, mCSM-PPI, mCSM-DNA with three types of prediction; single, multiple and systematic mutation.

Single mutation

* If the single mutation option is selected in one of the tools within the mCSM platform, it will be shown on a results page after processing. This information includes the predicted value changes (protein stability, protein-protein interaction, protein-DNA interaction) as measured by the change in Gibbs Free Energy ΔΔG kcal/mol (Figure 5), which are then classified as highly destabilising (ΔΔG ≤ -2 kcal/mol), destabilising (-2 kcal/mol < ΔΔG < 0 kcal/mol), stabilising (0 kcal/mol ≤ ΔΔG < 2 kcal/mol) , or highly stabilising (ΔΔG ≥ 2 kcal/mol).
* If the structure of a complex is submitted to mCSM-Stability, it will calculate the predicted change in stability of the entire complex. It is therefore often advisable to also run predictions on a PDB file containing the protomer chain alone.
* For mCSM-PPI and mCSM-DNA, for mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there are fewer mutations located further away than 12 Å in the datasets used to train the methods.
* Also shown is an interactive 3D visual representation of the uploaded PDB file (Figure 5 A, right).

Multiple or systematic

* If the option for inputting a list of mutations or systematic was used to analyse the PDB file, then after processing, the results will be shown in a tabulated form (Figure 5B), including mutation specific information such as the residue solvent accessibility (RSA), as well as the predicted ΔΔG.
* Each result is also classified, using the predicted ΔΔG value, as highly destabilising, destabilising, stabilising, or highly stabilising.
* Users can search the result table or download the entire results into a tab-separated text file.

3.2.4. SDM

SDM uses environment-specific amino acid substitution tables[37] and structural features including residue depth[14] and packing density to predict the impact of mutations on protein stability. The result page of single and list mutation is as follows.

Single mutation

* The single mutation result page (Figure 6A) provides predicted protein stability changes (ΔΔG), in addition to structural information implemented in SDM including secondary structure, RSA, residue depth and residue occluded packing density (OSP), sidechain-sidechain hydrogen bond (HBOND\_SS), sidechain-main chain amide hydrogen bond (HBOND\_SN), sidechain-main chain carbonyl hydrogen bond (HBOND\_SO). The integrated 3D viewer also shows the provided structure and its wild-type amino acids in ribbon and stick representation, respectively (Figure 6A).
* The predicted stability changes (ΔΔG) will be shown with a negative sign if the mutation is predicted as a destabilising mutation (ΔΔG < 0.0 kcal/mol : destabilising, ΔΔG < -2.5 kcal/mol : highly destabilising) and displayed in red; if a stabilising mutation (ΔΔG ≥ 0.0 kcal/mol : stabilising, ΔΔG > 2.5 kcal/mol : highly stabilising) is predicted, then the ΔΔG will be positive and displayed in blue.

Multiple mutations

* The predicted SDM ΔΔG for a given mutation list is displayed in a tabulated format (Figure 6B) with their structural features. Users can download all mutant PDB structures and their predicted values in individual files.

3.2.5 DUET

Single mutation

* The DUET result page (Figure 7A) provides the predicted stability changes (ΔΔG) with integrated features such as secondary structure and stability changes from mCSM and SDM. The outcome of DUET is classified into destabilising (ΔΔG ≤ 0 kcal/mol) and stabilising (ΔΔG > 0 kcal/mol). While DUET refers to both mCSM and SDM scores, the prediction result can vary from the two methods.
* In the structure viewer (Figure 7A right panel), the wild-type amino acid is shown in stick representation and users can download the corresponding mutant structure file in PDB format.

Systematic mutations

* With the systematic prediction (Figure 7B), users can examine the predicted changes in protein stability using DUET, mCSM, and SDM for all nineteen possible mutations at a given residue position.
* The predictions and the structural information used to calculate the DUET scores are displayed in a downloadable table.

3.2.6. DynaMut

Users can use DynaMut to assess the impact of mutations on protein dynamics and stability with single and list mutation prediction.

Single mutation

* The results of mutational effects on protein dynamics and stability are shown in three different tabs (Figure 8A): ΔΔG predictions, interatomic interactions, deformation and fluctuation analysis.
* The ΔΔG prediction page provides predicted values from normal mode analysis (NMA)-based prediction (ΔΔG ENCoM), vibrational entropy energy changes (ΔΔSVib ENCoM) and other structure-based stability predictions (ΔΔG mCSM, ΔΔG SDM, ΔΔG DUET). Users can visually assess mutational effects on protein flexibility which is coloured on the protein structure by vibrational entropy (Figure 8B) for the region gaining (red) or losing (blue) their flexibility. This 3D visual representation can be also downloaded into a Pymol session, high resolution image and CSV file.
* Through the interatomic interactions tab (Figure 8A), users can compare the differences of molecular interactions between wild-type and mutant structures. The PDB structure with interatomic interactions can be retrieved as a Pymol session file.
* The mutational effects on protein dynamics are shown in the deformation and fluctuation tab (Figure 8A). Users can evaluate the changes of the amount of local flexibility and atomic fluctuation upon mutation in 3D visual representation, and the results are downloadable as a CSV file and a Pymol session file.

Multiple mutations

* For a given mutation list, DynaMut gives all predicted values, including ΔΔGStabilityENCoM, ΔΔSVib ENCoM and ΔΔGStability DynaMut, in table format (Figure 8C). Users can conduct a more detailed analysis accessing the single prediction page of each mutation by clicking on the ‘Detail’ button.

3.2.7. mCSM-PPI2

mCSM-PPI2 supports two types of protein-protein affinity prediction: mutation prediction and binding analysis. The mutation prediction gives predicted protein-protein affinity changes based on a given protein-protein complex and the mutation information. On the other hand, the binding analysis considers interface residues within 5 Å from different chains in the complex structure for alanine scanning and saturation mutagenesis.

Single mutation

* mCSM-PPI2 displays predicted binding affinity changes (ΔΔG) upon mutation in two classes, destabilising (ΔΔG < 0 kcal/mol) and stabilising (ΔΔG > 0 kcal/mol). Mutation details such as the distance to the interface from the given mutation position are also shown (Figure 9).
* For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
* Users can assess the mutational impact in atomic/residue level through a 3D interactive viewer and a 2D graph (Figure 9). The molecular viewer provides Arpeggio inter/intra interactions for wild-type and mutant structures and the interaction changes between wild-type and mutant allows for investigation of the relationship between non-bonded interaction and protein-protein affinity. For the residue-level analysis, the 2D graph can be used to study the inter-residue interactions of wild-type and mutant in a simple and user-friendly representation.

List mutation

* For multiple mutation analysis, the result page shows predicted ΔΔG with mutation details in a tabulated format. Users can access more detailed results of each mutation through its single mutation result page and download all entries on the table as a CSV file.

Alanine scanning

* To identify residues with a greater contribution to the energy of binding (hot-spot) at the interface of interaction, alanine scanning can be used by predicting protein-protein binding affinity changes upon mutations to alanine across all identified interface residues. The predicted ΔΔG values are displayed in table, bar chart and 3D viewer (Figure 10A).
* Users can assess the effects of alanine mutation on the interface residues through a bar graph and 3D viewer coloured in red and blue for destabilising and stabilising mutations, respectively.

Saturation mutagenesis

* The saturation mutagenesis provides the most exhaustive prediction, showing predicted ΔΔG for all identified interface residues when they are changed into nineteen different amino acids. The results are shown in table, heatmap and 3D molecule viewer and the interface residues of the 3D viewer are coloured by the average ΔΔG of all mutations for each residue (Figure 10B).

3.2.8. mCSM-NA

Single prediction

* The predicted protein-nucleic acid affinity changes (destabilising : ΔΔG < 0 Kcal/mol, stabilising : ΔΔG > 0 Kcal/mol) on a given structure are shown (Figure 11A) with other properties such as the type of nucleic acid, solvent accessibility of wild-type protein, and predicted mutational effects from mCSM-Stability.
* For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
* The molecule visualisation panel (Figure 11B) shows the protein-nucleic acid complex with the wild-type amino acid, and the mutation as a stick representation. mCSM-NA allows users to further investigate inter/intra residue interactions by downloading Pymol session file.

List mutation

* mCSN-NA provides predicted protein-nucleic acid affinity changes, wild-type RSA, and mutation information for a given list of mutations in a table (Figure 11C) and the table is also downloadable as a TSV format.

3.2.9. mCSM-lig

* mCSM-lig predicts affinity changes (log affinity fold) between a given protein and its ligand upon mutation (Figure 12A) using additional information such as the closest distance between wild-type residue and ligand and the protein stability change (Kcal/mol) from DUET. The stabilising and destabilising mutations are shown in positive and negative values respectively.
* For mutations further than 12 Å from the interaction, the mCSM predictions are not considered, and are set to 0, as the graph-based signatures capture a smaller radius of environmental data, and there were fewer mutations located further away than 12 Å in the datasets used to train the methods.
* The wild-type amino acid and ligand are shown in stick and sphere representations in 3D molecule viewer, respectively (Figure 12B).

## Identification of driving molecular consequences

The outputs of the predictive tools described in Section 3.2 provide the basis for an initial heuristic examination. When trying to interpret the molecular consequences of a specific variant, it is important to remember that phenotypic outcomes are often the result of the combination of multiple molecular changes. For coding mutations, we initially ask ourselves two questions:

1. Is the mutation located within 5 Å of an interface? If so, is the mutation is more likely to disrupt the interaction (ΔΔG < ±0.5 kcal/mol) based on the corresponding mCSM output above (e.g. mCSM-PPI, mCSM-DNA, mCSM-NA, mCSM-Lig)? If the mutation is further than 12 Å away from an interface, it is less likely to disrupt the interaction directly, and the mCSM predictions are less reliable.
2. Is the mutation likely to disrupt protein folding and stability? mCSM-Stability, SDM, DUET and DynaMut provide insight into this, with mutations leading to ΔΔG < ±0.5 kcal/mol more likely to have a significant biological effect. Mutations at buried residues are more likely to have a larger effect on protein stability.
3. Is the mutation a special case that is more likely to lead to disruption of the protein due to the unique geometry restraints of the residues (see Notes 3 and 4)

To more exhaustively explore how mutations in a protein are leading to a phenotype, and to identify those molecular features that best capture the driving of the molecular mechanisms an investigation into the performance of each inputted feature should be conducted in order to construct the highest performing predictive model.

A more robust method for selecting which features are most informative can be performed using feature selection in R, a statistical programming language. While R is powerful enough to create classification models in itself, we can also use it to measure which features from our predictive tools’ output are most effective in stratifying our mutations. Two effective approaches are:

1. A random forest classification algorithm to measure feature importance using a set of mutations with known class labels (e.g. pathogenic / non-pathogenic, deleterious / non-deleterious).
2. The Boruta Algorithm performs permutations of the data to statistically compare each feature’s importance with that attainable at random, and uses this to eliminate uninformative features. The package in R provides a graphical output using boxplots.

Features that score highly provide evidence that the molecular consequence that they measure is relevant to how mutations are leading to the phenotype of interest. The algorithm can also highlight high levels of correlation between features. When two or more features are highly correlated and are likely measuring the same information, only one should be used in subsequent predictive model development to remove redundancy, minimise noise and avoid bias from heavily weighting a model in favour of a particular attribute. The model should also have the fewest possible features that perform best. Using too many features risks generating a model that only performs accurately on the training data but cannot be generalised to real-world data.

## Machine Learning Phenotypes: Building a predictive classifier

An initial understanding of molecular mechanisms imparted by disease-causing mutations is a crucial step towards establishing a genotype-phenotype correlation. However, manual analysis of different results can often miss underlying, statistically significant relationships amongst different mutational measurements, which can help relate them to the phenotype. Machine learning, and in particular supervised learning, addresses this issue by providing a set of tools for the efficient analysis of labelled data (*e.g.*, experimentally characterised mutations) in order to derive a model that describes a phenomenon, aiming for its generalisation (applying it to data never seen before). The identification of patterns and associations within the data being analysed will further help the predictive model to establish a distinction between mutations within the same gene leading to different phenotypes, and hence the development of an effective predictive tool that can be used to interpret novel clinical variants.

Here, our goal is to build a machine learning classifier to distinguish between pathogenic vs. non-pathogenic mutations in a given gene. Multiple steps are required to use machine learning effectively in order to obtain a non-biased, accurate predictor:

1. Dataset curation: Machine learning algorithms require a well-curated dataset which can be presented in the form of a comma separated value (.csv) file. In a supervised machine learning approach, all data labels (here, pathogenic or non-pathogenic for each mutation) must be known in order to enable correlations to be assessed between labels (*e.g.*, phenotypes) and features/properties used as evidence to represent each data point (*e.g.*, mutations). The source of the data is an essential aspect of building a classifier, as the quality of the resultant predictive model directly depends on the quality of the data used to build it. In our case, accurate clinical sources are required to justify labelling mutations as either pathogenic or non-pathogenic. In this case, generally, non-pathogenic variants can be curated from population variant databases such as GnomAD, usually taking into account frequent mutations. Even common variants, however, may still be linked to a disease, especially if it is a weakly penetrative mutation or recessive condition, which would add noise to the data set and consequently make the task of building a general predictive model more challenging. In situations where other biologically relevant information is present, such as cellular fitness cost, it is essential that this type of information is present for every mutation in your dataset, as a supervised algorithm will not be able to handle missing labels in your input. The initial dataset should contain a representative set of mutations within all the phenotype classes (pathogenic and non-pathogenic), and ideally, present a balanced number of instances between classes, to help prevent any biases towards overrepresented classes in the resultant model. More details on metrics used to evaluate the performance of predictive models on an imbalanced dataset are discussed below.
2. Feature generation: The feature generation stage is crucial as it is intended to provide descriptive information about each mutation, which will be used by the learning algorithm to finally classify the phenotype of a mutation. As described above, features can encompass a diverse range of mutational information:
	1. Protein stability and dynamics (mCSM-Stability, DUET, SDM, Dynamut).
	2. Protein functional changes such as changes in affinity for other proteins (mCSM-PPI2), nucleic acids (mCSM-NA) and ligands (mCSM-lig).
	3. At the residue level, changes in protein pharmacophore and local residue environment such as changes in interatomic interactions (Arpeggio) are also important, as some mutations at the same locus can have different phenotypes.
	4. Sequence-level predictors (SIFT, Polyphen, SNAP2).
	5. Evolutionary-based predictors (ConSurf), population based mutational tolerance (MTR-Viewer), as well as amino acid substitution matrices (e.g. PAM30, BLOSUM62, PSSM) offer added information on the likelihood of one mutation to change into another.

Feature generation is directly dependent on the wild-type biological functions of the protein, which is why an understanding of the biological relevance is important at the very beginning of this process.

1. Training and Testing sets: The data collected has to be divided into training and testing sets. This is important to assess the generalization power of a classifier, *i.e.* its ability to correctly predict on new data, and to ensure that it has not been over- or under-trained. One pitfall to avoid is that the data used to train the model should be different, non-redundant, from the data used to test the model. It is a common practice to divide the original dataset into the Training and Test sets in the beginning of the machine learning experimentation process, depending on the overall size of the dataset. For smaller datasets, a larger proportion of the data may need to be segregated into the Test set to provide sufficient data points to accurately measure performance of the trained model. This process can be done in a bootstrapping procedure or through cross-validation, when the original data set is divided into k-folds and each is taken iteratively as the test set while the remaining are used as training (k-fold cross-validation).
2. Feature selection: The features selected for training the model can have a large effect on its accuracy. Thus, it is important to select only informative features, and eliminate irrelevant or non-discriminative ones, which are a common source of noise. Feature selection can also help reduce overfitting, improve accuracy and reduce training time considerably, as it aims to generate simpler, more concise models. Feature selection methods provided in the Python machine learning library, Scikit-Learn[52], include univariate selection, feature importance, correlation matrix, and recursive feature elimination or addition. Alternatively, forward stepwise selection can also be performed as a greedy heuristics in which features are included iteratively, one at a time, based on their individual performance contributions.
3. Machine learning platforms: Different tools have been developed for implementing machine learning. Some offer a graphical user interface (GUI), such as Weka[53], while some run as python packages through the command line, such as Scikit-Learn. Different packages for different programming languages offer similar algorithms and options to adjust the algorithm parameters according to specific tasks. The major classification algorithms we test are: Naive Bayes, Decision Trees, K-Nearest Neighbour, Support Vector Machines and Ensemble Classifiers. It is good practice to try and compare representative algorithms of all these classes, provided that the algorithm is compatible with the dataset type. Within weka, this can be done automatically using the auto-weka function. In cases where the training set is unbalanced, oversampling or undersampling of the training data can be used to achieve a better representation of classes within the classification model-building stage, and prevent model bias in always detecting the predominant class and achieving a false high performance.
4. Model validation: Within the development stage, the primary tool in the validation of a model is the use of a non-redundant independent test set, also called blind test.

Validation can be furthered by the use of internal data testing such as *k*-fold cross validation. When using *k*-fold cross validation, the overall dataset is divided into *k* number of subsets. One of these subsets is used as a test set, while the remaining (*k*-1) subsets are used to train a model. The process is repeated *k* times, until all the data have been used in both training and test sets. The final model performance is calculated as the average between the performances of all the *k* iterations. Since this is an average value, it gives an overall performance metric of model evaluation. We will often vary *k* based on the size of the dataset. In cases where the training set is small (*e.g.*, ~200 data points), we will often try leave-one-out validation, which is where *k* is equal to the size of the dataset. The final performance is the average of all the training iterations. One important aspect while selecting adequate predictive models is consistency in performance between the training and test sets. This usually indicates a robust model, within which discrepancies might indicate overfitting (*e.g*., a considerably higher performance on training in comparison with the test set might indicate overfitting).

1. Model evaluation: Several different evaluation metrics may be used for classification tasks, which are generally calculated on values obtained from a confusion matrix. This matrix is a summary of the total number of data points, their actual phenotype, and their predicted phenotype (Table 6). From the distributions of data points across the compartments of the matrix, descriptive metrics can be calculated:

i. accuracy (number of correct predictions: [(TP+TN)/TOTAL]),

ii. precision (rate of correctly predicted positive instances from all assigned as positives: [TP/(TP+FP)],

iii. recall (rate of correctly predicted positive instances from all real positive instances: [TP/(TP+FN)],

iv. f-score (a weighted average of recall and precision), and

v. Matthews correlation coefficient (MCC; a balanced measure between true positives and true negatives

[(TP X TN) - (FP X FN)/

√(TP+FP)(TP+FN)(TN+FP)(TN+FN)] )

Where:

TP = True positive

TN = True negative

FP = False positive

FN = False negative

Classifier performance can also be described graphically using a Receiver Operating Characteristic curve, which compares the True Positive Rate and True Negative Rate. The closer the area under the curve is to 1, the better the performance of the classifier.

These metrics should be used in a combinatorial fashion across all elements of training, test, and cross validation stages for comparison of model performance during different stages of classifier optimisation. In cases where the dataset is imbalanced, balanced measures such as MCC should be prioritized, as other measures might bias for an overtrained model on the dominant dataset.

## Notes

1. Often following curation, the distribution of number of pathogenic and benign mutations is unbalanced, which can affect subsequent efforts to build predictive tools using machine learning. Two approaches that can help include oversampling of the underrepresented class, or undersampling of the overrepresented class. Evaluation metrics that are less biased towards unbalanced classes, such as the Matthew’s correlation coefficient, precision-recall curves and Kendall correlations, should also be preferentially used.
2. The chain ID for the provided PDB file is a mandatory field for all the structure-based methods and blank characters are not allowed. It is possible that homology modelling tools might not automatically add a chain ID. If this is the case, the user will need to modify the PDB file prior to submission to the servers. There are several tools available to perform this task (e.g. <http://www.canoz.com/sdh/renamepdbchain.pl> )
3. Special cases: Mutations to and from prolines. Prolines are the only amino acid whose amino group is connected to the sidechain, which in the context of the peptide bond, greatly limits torsional angles. The nature of this residue, therefore, needs to be taken into account while analysing mutation effects. For instance, (i) mutations to prolines in the middle of alpha-helices can introduce kinks, affecting local structure and (ii) since prolines are commonly found in turns and loops, their substitution might interfere with the formation of secondary structures such as hairpins.
4. Special cases: mutations of positive-phi glycines. Similarly to prolines, positive phi glycines, while rare in experimental structures, should also be given special consideration due to their torsional angles. Glycines are the only residues capable of adopting positive-phi angles. These glycines are usually conserved across evolution, meaning that mutations on positive-phi glycines, especially on loops and hairpins, tend to be destabilising.

## Acknowledgements

This work was supported by Australian Government Research Training Program Scholarships [to S.P., M.K., Y.M., C.H.M.R.]; the Jack Brockhoff Foundation [JBF 4186, 2016 to D.B.A.]; a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1 to D.B.A. and D.E.V.P.]; the National Health and Medical Research Council of Australia [APP1072476 to D.B.A.].

## References

1. Jatana N, Ascher DB, Pires DEV, Gokhale RS, Thukral L (2019) Human LC3 and GABARAP subfamily members achieve functional specificity via specific structural modulations. Autophagy:1-17. doi:10.1080/15548627.2019.1606636

2. Abayakoon P, Jin Y, Lingford JP, Petricevic M, John A, Ryan E, Wai-Ying Mui J, Pires DEV, Ascher DB, Davies GJ, Goddard-Borger ED, Williams SJ (2018) Structural and Biochemical Insights into the Function and Evolution of Sulfoquinovosidases. ACS Cent Sci 4 (9):1266-1273. doi:10.1021/acscentsci.8b00453

3. Ascher DB, Cromer BA, Morton CJ, Volitakis I, Cherny RA, Albiston AL, Chai SY, Parker MW (2011) Regulation of insulin-regulated membrane aminopeptidase activity by its C-terminal domain. Biochemistry 50 (13):2611-2622. doi:10.1021/bi101893w

4. Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N (2018) Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 8 (1):15356. doi:10.1038/s41598-018-33370-6

5. Silk M, Petrovski S, Ascher DB (2019) MTR-Viewer: identifying regions within genes under purifying selection. Nucleic Acids Res 47 (W1):W121-W126. doi:10.1093/nar/gkz457

6. Pires DE, Blundell TL, Ascher DB (2015) Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 43 (Database issue):D387-391. doi:10.1093/nar/gku966

7. Blaszczyk M, Harmer NJ, Chirgadze DY, Ascher DB, Blundell TL (2015) Achieving high signal-to-noise in cell regulatory systems: Spatial organization of multiprotein transmembrane assemblies of FGFR and MET receptors. Prog Biophys Mol Biol 118 (3):103-111. doi:10.1016/j.pbiomolbio.2015.04.007

8. Jafri M, Wake NC, Ascher DB, Pires DE, Gentle D, Morris MR, Rattenberry E, Simpson MA, Trembath RC, Weber A, Woodward ER, Donaldson A, Blundell TL, Latif F, Maher ER (2015) Germline Mutations in the CDKN2B Tumor Suppressor Gene Predispose to Renal Cell Carcinoma. Cancer Discov 5 (7):723-729. doi:10.1158/2159-8290.CD-14-1096

9. Pacitto A, Ascher DB, Wong LH, Blaszczyk BK, Nookala RK, Zhang N, Dokudovskaya S, Levine TP, Blundell TL (2015) Lst4, the yeast Fnip1/2 orthologue, is a DENN-family protein. Open Biol 5 (12):150174. doi:10.1098/rsob.150174

10. Pires DE, Chen J, Blundell TL, Ascher DB (2016) In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 6:19848. doi:10.1038/srep19848

11. Albanaz ATS, Rodrigues CHM, Pires DEV, Ascher DB (2017) Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 12 (6):553-563. doi:10.1080/17460441.2017.1322579

12. Casey RT, Ascher DB, Rattenberry E, Izatt L, Andrews KA, Simpson HL, Challis B, Park SM, Bulusu VR, Lalloo F, Pires DEV, West H, Clark GR, Smith PS, Whitworth J, Papathomas TG, Taniere P, Savisaar R, Hurst LD, Woodward ER, Maher ER (2017) SDHA related tumorigenesis: a new case series and literature review for variant interpretation and pathogenicity. Mol Genet Genomic Med 5 (3):237-250. doi:10.1002/mgg3.279

13. Jubb HC, Pandurangan AP, Turner MA, Ochoa-Montano B, Blundell TL, Ascher DB (2017) Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Prog Biophys Mol Biol 128:3-13. doi:10.1016/j.pbiomolbio.2016.10.002

14. Pandurangan AP, Ascher DB, Thomas SE, Blundell TL (2017) Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. Biochem Soc Trans 45 (2):303-311. doi:10.1042/BST20160422

15. Sibanda BL, Chirgadze DY, Ascher DB, Blundell TL (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science 355 (6324):520-524. doi:10.1126/science.aak9654

16. Rodrigues CH, Ascher DB, Pires DE (2018) Kinact: a computational approach for predicting activating missense mutations in protein kinases. Nucleic Acids Res 46 (W1):W127-W132. doi:10.1093/nar/gky375

17. Hnizda A, Fabry M, Moriyama T, Pachl P, Kugler M, Brinsa V, Ascher DB, Carroll WL, Novak P, Zaliova M, Trka J, Rezacova P, Yang JJ, Veverka V (2018) Relapsed acute lymphoblastic leukemia-specific mutations in NT5C2 cluster into hotspots driving intersubunit stimulation. Leukemia 32 (6):1393-1403. doi:10.1038/s41375-018-0073-5

18. Andrews KA, Ascher DB, Pires DEV, Barnes DR, Vialard L, Casey RT, Bradshaw N, Adlard J, Aylwin S, Brennan P, Brewer C, Cole T, Cook JA, Davidson R, Donaldson A, Fryer A, Greenhalgh L, Hodgson SV, Irving R, Lalloo F, McConachie M, McConnell VPM, Morrison PJ, Murday V, Park SM, Simpson HL, Snape K, Stewart S, Tomkins SE, Wallis Y, Izatt L, Goudie D, Lindsay RS, Perry CG, Woodward ER, Antoniou AC, Maher ER (2018) Tumour risks and genotype-phenotype correlations associated with germline variants in succinate dehydrogenase subunit genes SDHB, SDHC and SDHD. J Med Genet 55 (6):384-394. doi:10.1136/jmedgenet-2017-105127

19. Usher JL, Ascher DB, Pires DE, Milan AM, Blundell TL, Ranganath LR (2015) Analysis of HGD Gene Mutations in Patients with Alkaptonuria from the United Kingdom: Identification of Novel Mutations. JIMD Rep 24:3-11. doi:10.1007/8904\_2014\_380

20. Nemethova M, Radvanszky J, Kadasi L, Ascher DB, Pires DE, Blundell TL, Porfirio B, Mannoni A, Santucci A, Milucci L, Sestini S, Biolcati G, Sorge F, Aurizi C, Aquaron R, Alsbou M, Lourenco CM, Ramadevi K, Ranganath LR, Gallagher JA, van Kan C, Hall AK, Olsson B, Sireau N, Ayoob H, Timmis OG, Sang KH, Genovese F, Imrich R, Rovensky J, Srinivasaraghavan R, Bharadwaj SK, Spiegel R, Zatkova A (2016) Twelve novel HGD gene variants identified in 99 alkaptonuria patients: focus on 'black bone disease' in Italy. Eur J Hum Genet 24 (1):66-72. doi:10.1038/ejhg.2015.60

21. Ramdzan YM, Trubetskov MM, Ormsby AR, Newcombe EA, Sui X, Tobin MJ, Bongiovanni MN, Gras SL, Dewson G, Miller JML, Finkbeiner S, Moily NS, Niclis J, Parish CL, Purcell AW, Baker MJ, Wilce JA, Waris S, Stojanovski D, Bocking T, Ang CS, Ascher DB, Reid GE, Hatters DM (2017) Huntingtin Inclusions Trigger Cellular Quiescence, Deactivate Apoptosis, and Lead to Delayed Necrosis. Cell Rep 19 (5):919-927. doi:10.1016/j.celrep.2017.04.029

22. Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S (2017) Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res 27 (10):1715-1729. doi:10.1101/gr.226589.117

23. Trezza A, Bernini A, Langella A, Ascher DB, Pires DEV, Sodi A, Passerini I, Pelo E, Rizzo S, Niccolai N, Spiga O (2017) A Computational Approach From Gene to Structure Analysis of the Human ABCA4 Transporter Involved in Genetic Retinal Diseases. Invest Ophthalmol Vis Sci 58 (12):5320-5328. doi:10.1167/iovs.17-22158

24. Ascher DB, Spiga O, Sekelska M, Pires DEV, Bernini A, Tiezzi M, Kralovicova J, Borovska I, Soltysova A, Olsson B, Galderisi S, Cicaloni V, Ranganath L, Santucci A, Zatkova A (2019) Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype-phenotype correlations in the largest cohort of patients with AKU. Eur J Hum Genet 27 (6):888-902. doi:10.1038/s41431-019-0354-0

25. Soardi FC, Machado-Silva A, Linhares ND, Zheng G, Qu Q, Pena HB, Martins TMM, Vieira HGS, Pereira NB, Melo-Minardi RC, Gomes CC, Gomez RS, Gomes DA, Pires DEV, Ascher DB, Yu H, Pena SDJ (2017) Familial STAG2 germline mutation defines a new human cohesinopathy. NPJ Genom Med 2:7. doi:10.1038/s41525-017-0009-4

26. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, Coeck N, Hill-Cawthorne GA, Nair MB, Mallard K, Ramsay A, Campino S, Hibberd ML, Pain A, Rigouts L, Clark TG (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 14:31. doi:10.1186/s12916-016-0575-9

27. Silvino AC, Costa GL, Araujo FC, Ascher DB, Pires DE, Fontes CJ, Carvalho LH, Brito CF, Sousa TN (2016) Variation in Human Cytochrome P-450 Drug-Metabolism Genes: A Gateway to the Understanding of Plasmodium vivax Relapses. PLoS One 11 (7):e0160172. doi:10.1371/journal.pone.0160172

28. White RR, Ponsford AH, Weekes MP, Rodrigues RB, Ascher DB, Mol M, Selkirk ME, Gygi SP, Sanderson CM, Artavanis-Tsakonas K (2016) Ubiquitin-Dependent Modification of Skeletal Muscle by the Parasitic Nematode, Trichinella spiralis. PLoS Pathog 12 (11):e1005977. doi:10.1371/journal.ppat.1005977

29. Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, Spelman DW, Padiglione A, Peleg AY, Holt KE (2018) Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genom 4 (3). doi:10.1099/mgen.0.000165

30. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN, Thwaites G, Edwards DJ, Nath AP, Pham K, Ascher DB, Farrar J, Khor CC, Teo YY, Inouye M, Caws M, Dunstan SJ (2018) Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 50 (6):849-856. doi:10.1038/s41588-018-0117-9

31. Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, Denholm JT, Ascher DB (2018) Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy. Am J Respir Crit Care Med 198 (4):541-544. doi:10.1164/rccm.201712-2572LE

32. Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, Arumugam S, Rajan L, Ebenezer M, Ascher DB, Arnold E, Blundell TL (2018) Structural Implications of Mutations Conferring Rifampin Resistance in Mycobacterium leprae. Sci Rep 8 (1):5016. doi:10.1038/s41598-018-23423-1

33. Karmakar M, Rodrigues CHM, Holt KE, Dunstan SJ, Denholm J, Ascher DB (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS One 14 (5):e0217169. doi:10.1371/journal.pone.0217169

34. Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW (2014) Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 4:4765. doi:10.1038/srep04765

35. Jubb HC, Higueruelo AP, Ochoa-Montano B, Pitt WR, Ascher DB, Blundell TL (2017) Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J Mol Biol 429 (3):365-371. doi:10.1016/j.jmb.2016.12.004

36. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30 (3):335-342. doi:10.1093/bioinformatics/btt691

37. Pandurangan AP, Ochoa-Montano B, Ascher DB, Blundell TL (2017) SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res 45 (W1):W229-W235. doi:10.1093/nar/gkx439

38. Pires DE, Ascher DB, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 42 (Web Server issue):W314-319. doi:10.1093/nar/gku411

39. Rodrigues CH, Pires DE, Ascher DB (2018) DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 46 (W1):W350-W355. doi:10.1093/nar/gky300

40. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. Nucleic Acids Res 47 (W1):W338-W344. doi:10.1093/nar/gkz383

41. Pires DE, Ascher DB (2016) mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. Nucleic Acids Res 44 (W1):W469-473. doi:10.1093/nar/gkw458

42. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45 (W1):W241-W246. doi:10.1093/nar/gkx236

43. Pires DE, Blundell TL, Ascher DB (2016) mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 6:29575. doi:10.1038/srep29575

44. Pires DE, Ascher DB (2016) CSM-lig: a web server for assessing and comparing protein-small molecule affinities. Nucleic Acids Res 44 (W1):W557-561. doi:10.1093/nar/gkw390

45. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29 (1):308-311. doi:10.1093/nar/29.1.308

46. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136 (6):665-677. doi:10.1007/s00439-017-1779-6

47. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46 (D1):D1062-D1067. doi:10.1093/nar/gkx1153

48. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Vallabh Minikel E, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv:531210. doi:10.1101/531210

49. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12 (3):e1001779. doi:10.1371/journal.pmed.1001779

50. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res 46 (5):2699. doi:10.1093/nar/gky092

51. Rose PW, Prlic A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao YP, Valasatava Y, Voigt M, Westbrook JD, Woo J, Yang H, Young JY, Zardecki C, Berman HM, Burley SK (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res 45 (D1):D271-D281. doi:10.1093/nar/gkw1000

52. Pedregosa F, Ga, #235, Varoquaux l, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, #201, Duchesnay d (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825-2830

53. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc.,

## Figure Captions

**Figure 1.** An overview of the mechanistic characterisation of mutations to understand their biological consequences, and to guide the development of tools to predict phenotypic outcomes.

**Figure 2.** **The output of the Arpeggio tool.** A) showing an overview of the output for the inputted protein including the different types of interactions. B) Visualisation of the interactions shown on a protein structure.

**Figure 3. The MTR Gene Viewer result page.** (A) The line graph shows MTR scores in red for variations distant from neutrality across the transcript according to selected window size (codons) and subpopulation option. (B) The lollipop plot shows observed gnomeAD variation in yellow and green for missense and synonymous variation (C) The second lollipop plot displays pathogenic (red) and benign (blue) missense variants based on ClinVar annotation. (D) The alternate transcripts can be shown in a table with RefSeq ID.

**Figure 4. MTR Variant Queries result page.** The calculated results and information for the given input variants (or a CSV) are shown in a tabulated format. User can check the details through MTR Gene Viewer by clicking on the view button.

**Figure 5. The result pages for mCSM-Stability, mCSM-PPI and mCSM-DNA.** (A) mCSM-Stability (single mutation) and (B) mCSM-PPI (multiple/systematic mutation). (A) The single prediction for example mCSM-Stability page supports 3D interactive viewer for structural analysis. (B) The results and information from multiple/systematic prediction for example mCSM-PPI are shown in a table.

**Figure 6. SDM prediction results for single and list prediction**. (A) The single prediction displays the predicted ΔΔG with information used on the left panel and 3D structure in a ribbon (protein) and a stick (wild-type amino acid) representation. (B) The list prediction gives detailed structural information and predicted ΔΔG in a tabulated form highlighted according to stabilising (blue) and destabilising (red) mutation.

**Figure 7. DUET result pages for single and systematic prediction.** (A) The single prediction result of DUET shows predicted ΔΔG across SDM and mCSM-Stability with mutation details. (B) The systematic prediction results are shown in table including ΔΔG from DUET, SDM and mCSM-Stability and relative solvent accessible area of wild-type structure.

**Figure 8. DynaMut result pages.** The single prediction shows predicted DynaMut ΔΔG (A, left) and predicted protein stability (ΔΔG) from mCSM-Stability, SDM and DUET and flexibility changes (ΔΔG ENCoM). Users can check vibrational energy changes upon mutation in the panel B. For a multiple mutation, (B) list prediction result page shows predicted DynaMut ΔΔG and links to access the corresponding single prediction in table.

**Figure 9. mCSM-PPI2 single prediction result page.** The predicted ΔΔG is shown along with two interaction viewers: 3D interactive molecule viewer for atomic interaction analysis and 2D diagram for residue-level interaction analysis.

**Figure 10. mCSM-PPI2 interface scanning result pages.** The result pages of (A) alanine scanning and (B) saturation mutagenesis provide a bar chart and a heatmap coloured by predicted ΔΔG and average predicted ΔΔG from the nineteen possible mutations, respectively.

**Figure 11. mCSM-NA result pages for single and list mutation prediction.** In the single prediction result page, predicted protein-DNA affinity changes and mutation information are displayed in the prediction details (A) and the 3D viewer shows protein-DNA complex and wild-type amino acid in a ribbon and stick representation (B). The results of list prediction are shown in a tabulated form (C) and users can save the results in a TSV format.

**Figure 12. mCSM-lig result page.** (A)The predicted affinity change between protein-ligand upon mutation is shown in logarithm scale. (B) The protein and ligand are displayed in 3D viewer with a ribbon (for protein), a stick (for wild-type amino acid), and a sphere (for ligand) representation.

## Table Captions

**Table 1.** Comparison of available sequence-based predictive tools for mutation analysis.

**Table 2.** Comparison of available structure-based predictive tools for mutation analysis.

**Table 3.** Computational tools available in the mCSM platform.

**Table 4.** Information required to run each mCSM program.

**Table 5.** Atomic interactions calculated by Arpeggio.

**Table 6:** Description of a confusion matrix.

## Tables

**Table 1.**

|  |
| --- |
| **Protein stability & Dynamics** |
| **Method** | **Corr.a** |
| I-Mutant 2.0 | 0.62 |
| Auto-Mute | 0.64\* |
| MUpro | 0.75 |
| DynaMine | 0.63 |
| DDGun | 0.49 |
| INPS-MD/3D | 0.58 |
| iStable | 0.56**b** |
| iPTREEE - STAB | 0.70 |
| ProMaya | 0.79 |

**a** Pearson’s correlation **b**MCC

**Table 2.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Protein stability & Dynamics** | **Protein-protein affinity** | **Protein-nucleic acid affinity** | **Protein-small molecule affinity** |
| **Method** | **Corr.a** | **Method** | **Corr.b** | **Method** | **Corr.c** | **Method** | **Corr.d** |
| mCSM-Stability | 0.69 | mCSM-PPI | 0.16 | mCSM-NA | 0.70 | mCSM-lig | 0.63 |
| DUET | 0.68  | mCSM-PPI2 | 0.42 |  |  |  |  |
| DynaMut | 0.70 | BeAtMuSiC | 0.28 |  |  |  |  |
| SDM2 | 0.61 | MutaBind | 0.41 |  |  |  |  |
| STRUM | 0.79 | FoldX | 0.12 |  |  |  |  |
| PopMuSiC 2.1 | 0.63 | MMPBSA | 0.19 |  |  |  |  |
| CUPSAT | 0.78 |  |  |  |  |  |  |
| Eris | 0.75 |  |  |  |  |  |  |
| INPS-MD/3D | 0.72 |  |  |  |  |  |  |

**a** Pearson’s correlation when evaluated on blind-test sets derived from the ProTherm database.

**b** Kendall rank correlation coefficient on 1007 single-point mutations from CAPRI (T55)

**c** Pearson’s correlation on 331 single-point mutations from 38 protein-nucleic acid complexes

**d** Pearson’s correlation on 763 single-point mutations from 200 protein-ligand complexes

**Table 3.**

|  |  |  |
| --- | --- | --- |
| **mCSM Tool** | **Type** | **Function** |
| **Arpeggioa** | Protein Interaction | Calculates 13 different types of interactions between atoms including hydrogen bonds, halogen bonds, carbonyl interactions and others. |
| **MTR-Viewerb** | Missense Tolerance | A measure of a gene’s regional tolerance to missense variation.  |
| **mCSM-Stabilityc** | Stability | Predict the effects of a mutation on the overall protein stability |
| **SDM2d** | Stability | Predicts the change in protein stability due to a single mutation using conformationally constrained environment-dependent amino acid substitution tables. |
| **DUETe** | Stability | Uses mCSM-Stability and SDM2 in order to create a consensus prediction the effects of a mutation on protein stability |
| **DynaMutf** | Flexibility | Looks to predict the effects of a mutation on protein stability, flexibility, and dynamics |
| **mCSM-PPIg** | Protein Interaction | Predicts the effects of a mutation within a specified protein on its impact with overall protein-protein interactions. |
| **mCSM-PPI2h** | Protein Interaction | Creates a similar prediction to PPI but incorporates the effects of mutations on inter-residue non-covalent interaction network using graph kernels, evolutionary information, complex network metrics and energetic terms. |
| **mCSM-DNAi** | Protein Interaction | Predicts the impact of mutations on the proteins interaction with DNA. |
| **mCSM-NAj** | Protein Interaction | Predicts the impact of mutations on the proteins interaction with nucleic acids, and uses pharmacophore and information about nucleic acid properties. |
| **mCSM-Ligk** | Protein Interaction | Predicts the effects of single-point mutations on the stability of a protein-ligand complex. |

a[**http://biosig.unimelb.edu.au/arpeggioweb/**](http://biosig.unimelb.edu.au/arpeggioweb/)

b[**http://biosig.unimelb.edu.au/mtr-viewer/**](http://biosig.unimelb.edu.au/mtr-viewer/)

c[**http://biosig.unimelb.edu.au/mcsm/stability**](http://biosig.unimelb.edu.au/mcsm/stability)

d[**http://marid.bioc.cam.ac.uk/sdm2**](http://marid.bioc.cam.ac.uk/sdm2)

e[**http://biosig.unimelb.edu.au/duet/**](http://biosig.unimelb.edu.au/duet/)

f[**http://biosig.unimelb.edu.au/dynamut/**](http://biosig.unimelb.edu.au/dynamut/)

g[**http://biosig.unimelb.edu.au/mcsm/protein\_protein**](http://biosig.unimelb.edu.au/mcsm/protein_protein)

h[**http://biosig.unimelb.edu.au/mcsm\_ppi2/**](http://biosig.unimelb.edu.au/mcsm_ppi2/)

i[**http://biosig.unimelb.edu.au/mcsm/protein\_dna**](http://biosig.unimelb.edu.au/mcsm/protein_dna)

j[**http://biosig.unimelb.edu.au/mcsm\_na/**](http://biosig.unimelb.edu.au/mcsm_na/)

k[**http://biosig.unimelb.edu.au/mcsm\_lig/**](http://biosig.unimelb.edu.au/mcsm_lig/)

**Table 4.**

|  |  |  |
| --- | --- | --- |
| **mCSM Tool** | **Task** | **Inputs** |
| **Step 1** | **Step 2** |
| **Arpeggio** | Calculate | Molecule in PDB format or PDB accession code. | Select what you would like to calculate interactions to. You can select any (including multiple) part of the PDB file using the syntax:/1/2/3Where: 1. Chain ID. 2. Residue number. 3. Atom name. |
| **MTR-Viewer** | Gene Viewer | Gene, ensembl ID or Refseq ID | Select window size and overlay sub-population |
| Variant Queries | Variants as GrCh37 genomic coordinates. |  |
| **mCSM-Stability, mCSM-PPI, mCSM-DNA**  | Prediction | Wild-type protein file in a PDB format. For mCSM-PPI and mCSM-DNA, the structure of the complex in PDB format is required. | Single mutation (code and mutation chain), file with a list of mutations and its respective chains or code of residue and the mutation chain.  |
| **SDM2** | Prediction | Wild-type protein structure in a PDB format or PDB accession code. | Single mutation (code and mutation chain) or residue/position code and the mutation chain.  |
| **DUET** | Prediction | Wild-type protein structure in a PDB format or PDB accession code. | Single mutation (code and mutation chain) |
| **DynaMut** | Analysis | Wild-type protein structure in a PDB format or PDB accession code. | The selection of a Force Field and email (optional field). |
| Prediction | Wild-type protein structure in a PDB format or PDB accession code. | Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field). |
| **mCSM-PPI2** | Prediction | The structure of the complex in PDB format or corresponding PDB accession code. | Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and email (optional field). |
| Analysis | The structure of the complex in PDB format or corresponding PDB accession code. | Mutation details (Alanine scanning or saturation mutagenesis) and email (optional field). |
| **mCSM-NA** | Prediction | The structure of the complex in PDB format or corresponding PDB accession code. | Single mutation (code and mutation chain) or file with a list of mutations and its respective chains, and the selection of the Nucleic Acid Type.  |
| **mCSM-Lig** | Prediction | The structure of the complex in PDB format or corresponding PDB accession code. | Single mutation (code and mutation chain) and ligand information (3-letter ligand ID and estimated wild-type affinity). |

**Table 5.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Atomic Interaction** | **Description** | **Arpeggio Class** | **Bond Energy (kJ/mol)** |
| Van der Waals (dipole) | Permanent, induces and instantaneous dipoles | VWD | 1-9 |
| Hydrophobic | Between aliphatic and aromatic atoms | Hydrophobic | 4-12 |
| Hydrogen bond | Between carboxyl, amide, imidazole, guanidine, amino, hydroxyl and phenolic groups | Hydrogen bonds, Weak hydrogen bond, Polar contacts, Halogen bonds, Carbonyl interactions | 8-40 |
| Pi Interactions | From/to rings | Aromatic contacts | 6-70 |
| Electrostatic | Between carboxyl and amino groups | Ionic interactions, Metal complex | 42-84 |

**Table 6:**

|  |  |
| --- | --- |
|  | Actual value |
| **Positive** | **Negative** |
| Predicted value | **Positive** | True positive | False positive |
| **Negative** | False negative | True negative |