Computational Saturation Screen Reveals the Landscape of Mutations in Human Fumarate Hydratase

David Shorthouse[1], Michael W J Hall[2,3], Benjamin A Hall[1*]

1 Department of Medical Physics and Biomedical Engineering, UCL, London, WC1E 6BT

2 MRC Cancer Unit, University of Cambridge, Cambridge, CB2 0XZ, UK

3 Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK

Protein engineering, mutation, delta-delta-g, high-throughput, molecular dynamics

ABSTRACT: Single amino acid substitutions within protein structures often manifest with clinical conditions in humans. The mutation of a single amino can significantly alter protein folding and stability, or change protein dynamics to influence function. The chemical engineering field has developed a large toolset for predicting the influence of point mutations with the aim of guiding the design of improved and more stable proteins. Here we reverse this general protocol, and adapt these tools for prediction of damaging mutations within proteins. Mutations to Fumarate Hydratase (FH), an enzyme of the citric acid cycle, can lead to human disease. Inactivation of FH by mutation causes leiomyomas and renal cell carcinoma by subsequent fumarate buildup and reduction in available malate. We present a scheme for accurately predicting the clinical effects of every possible mutation in FH by adaptation to a database of characterized damaging and benign mutations. Using energy prediction tools Rosetta and FoldX coupled with molecular dynamics simulations, we accurately predict individual mutations as well as mutational hotspots with high disruptive capability in FH. Furthermore, through dynamic analysis we find that hinge regions of the protein can be stabilized or destabilized by mutations, with mechanistic implications for the functional ability of the enzyme. Finally we categorize all potential mutations in FH into functional groups, predicting which known mutations in the human population are Loss of Function (LOF), therefore having clinical implications, and validate our findings through metabolomics data of characterized human cell lines.

INTRODUCTION

The influences of individual point mutations on a protein structure are many and varied. Mutations may influence the structure in numerous ways – leading to unfolding of the protein, inactivation of enzymatic activity, and an alteration of the dynamics of the structure. Many mutations to proteins in humans are associated with pathogenic disease, including all cancers, many inherited syndromes, and prion diseases[1]. Understanding the potential effects of mutations is paramount to the diagnosis and treatment of disease, and as such, prediction of the effects of a mutation is of large importance. Many mutations can be disruptive simply through alteration of a binding site, but for mutations that are not near to a binding site, the energic change to the structure, ΔΔG, is important as these mutations can lead to unfolding if the energy change is sufficiently high.

Numerous methods for predicting the ΔΔG of mutations have been derived, based on chemical analysis of a protein structure. Tools such as FoldX[2,3], which uses an empirical force field to predict the alterations in a protein induced by mutation, and methods included as part of the Rosetta suite[4,5], which uses Monte-Carlo based dynamics to predict energetic effects of mutations are generally employed for the prediction of stabilizing mutations in enzymatic optimization and design. Here we reverse this workflow to predict mutations that destabilize protein structures from analysis of every potential mutation that can occur. We thus provide a prediction derived from the enzyme chemistry and biophysics on whether any possible mutation will result in damage to protein function, when coupled with other heuristic knowledge such as binding site locations.

We apply our approach to the tetrameric enzyme Fumarate Hydratase. Fumarate hydratase (FH) is a member of the tricarboxylic acid cycle occurring in the mitochondria. FH activity in the cell is responsible for the reversible conversion of the metabolite fumarate into malate, and the knockout or mutational inactivation of FH in kidneys is linked to a pathogenic buildup of fumarate[6,7]. As a result FH is associated with numerous kidney pathologies – including cancer. Fumarate has been described as part of a novel classification of molecules named "oncometabolites", metabolites responsible for cancer transformation. Precisely how the buildup of fumarate can be oncogenic is unclear, but FH loss is associated with renal cell carcinoma, and recent work points towards suppression of DNA repair responses, epithelial-to-mesenchymal transformation, and promotion of mitosis[8–10]. Understanding the effects of mutations on the

activity and assembly of FH is of importance for the understanding and stratification of germline mutations, which can predispose patients to hereditary leiomyomatosis and renal cell cancer (HLRCC), a dominant negative condition caused by single allelic mutation of FH[11,12]. Previous work has identified mutants linked with inherited and de-novo FH-related conditions, including cancer[13] – most notably, the FH mutation database represents a comprehensive list of mutations and their effects, if known, on FH activity[14].

Here we present a workflow for understanding how current pathogenic mutations in FH inactivate the enzyme, and predicting novel loss of function (LOF) mutations. We find LOF mutations fall into three categories – those that disrupt the binding site, those that influence the assembly/folding of the protein, and those that alter the dynamics of hinge domains in the protein. Guided by the previously categorized mutations, we assess and classify every potential mutation in the available fumarate hydratase structure to study the landscape of potential mutations. We consider the structural and biological implications of each mutation, and thus can predict mechanistic effects of every potential mutant. We validate our workflow on historic mutational screens of bacterial T4 lysozyme[15]. Overall we predict that 66% of all mutations to FH influence activity or assembly. We further validate our predictions through studying the Cancer Cell Line Encyclopaedia (CCLE)[16,17] and show that previously unstudied mutations that we predict to be damaging to the function of FH result in altered metabolite levels expected from disruption to the activity of FH. We define a workflow for the chemical analysis of every potential single amino acid substitution in any soluble protein for which structural data is available.

MATERIALS AND METHODS

All data used in this study, including the code used in generating all figures from raw data, and the raw data to generate them is available publicly at: https://github.com/shorthouse-mrc/Fumarate_Hydratase

FH mutation database. The FH mutation database was downloaded from the Leiden Open Variation Database[14]. Missense mutations were manually curated into categories (Loss of Function, Benign, and Unknown) based on their implied clinical classification, and variant remarks, which contained information regarding FH enzymatic activity.

Mutational Clustering. Mutational clustering was performed with the non-random mutational clustering (NMC) algorithm, which attempts to discern the likelihood of a mutation spectrum occurring by random chance. NMC returns clusters of mutations that are statistically significant. We chose to run the NMC algorithm using the R library iPAC[18], using an alpha cutoff value of 0.05, and the Bonferroni multiple test correction method.

Gaussian Network Modelling (GNM). GNM was implemented using the Prody package in python[19]. A Kirchhoff matrix was constructed using the gnm.buildKirchhoff command with the parameters cutoff = 10.0 and gamma = 1.0. Normal modes were then calculated using the gnm.calcModes() command. Predicted hinges were assessed using the gnm.getHinges() command.

Molecular Dynamics Simulations. Molecular dynamics was performed using Gromacs version 2018.1[20]. We chose to simulate proteins using the GROMOS 54a7 forcefield[21]. The protein structures were first repaired using FoldX[2] "RepairPDB" with the following command:

```
$foldx --command=RepairPDB --pdb=5upp.pdb --ionStrength=0.05
--pH=7 --vdwDesign=2
```

The protein was then placed in a cubic box size 15 x 15 x 15 nm and solvated with roughly 90,000 spc water molecules. Counterions were introduced to a neutral charge, and to a concentration of 0.05 mol/litre. The system was energy minimized using the steepest descents algorithm until the maximum force, Fmax, of the system reached below 1000 kJ/mol/nm. Equilibration was performed using the NVT, followed by the NPT ensembles for 100 ps each. We chose to use the verlet cutoff scheme and PME electrostatics, and utilized periodic boundary conditions in the x, y, and z planes. Molecular dynamics was performed for 200 ns retaining velocities from the NPT equilibration. We used the V-rescale temperature coupling scheme, and Parrinello-Rahman isotropic pressure coupling.

FoldX ΔΔG Calculations. FoldX predicted ΔΔG was calculated using the PositionScan command within FoldX4. Positionscan was run on each residue in the protein structure sequentially using the following command:

```
$foldx --command=PositionScan --pdb=5upp_repaired.pdb
--ionStrength=0.05 --pH=7 --vdwDesign=2 --pdbHydrogens=false
--positions=49
```

for positionscan on the 49th residue. This command was repeated for each residue in the protein to generate an energy for every possible mutation. In the case of FH, we mutated a single subunit within the full tetrametric complex.

Rosetta ΔΔG Calculations. Rosetta predicted ΔΔG was calculated using the cartesian_ddg method as described in Kellogg et al[22]:

```
$path/to/source/bin/cartesian_ddg.static.linuxgccrelease -in:file:s
5upp_repaired.pdb -in::file::fullatom -database /path/to/database/ -
ignore_unrecognized_res true -ignore_zero_occupancy false -
fa_max_dis 9.0 -ddgccartesian -ddg::mut_file mutfile.txt -ddg::iterations
3 -ddg::dump_pdbs true -ddg::suppress_checkpointing true -ddg::mean
true -ddg::min true -ddg:output_silent true -bbnbr 1 -beta_nov16_cart >
logfile.log
```

ΔΔG was calculated by averaging the energy of 3 models of each mutation and comparing it to the WT calculation. In the case of FH we mutated a single subunit with the tetrameric protein complex.

Umap. We used Umap[23] based on the github repository at www.github.com/lmcinnes/unmap - using default parameters.

Cancer Cell Line Encyclopedia Data. Cancer Cell Line Encyclopedia (CCLE) mutation data was downloaded from the Broad Institute at: https://portals.broadinstitute.org/ccle/data . Metabolomics data was obtained from the supplementary data of Li et al[17].

Data Analysis. MDanalysis[24] and Biopython[25] were used for analysis of structural data. Data analysis workflows, and code for generation of all figures from primary data is available in jupyter notebooks available at https://github.com/shorthouse-mrc/Fumarate_Hydratase.

RESULTS AND DISCUSSION

Evidence of mutational clustering in FH. To start our study of mutations within FH we looked to pre-existing, characterized mutations associated with human disease. Human FH is formed as a homotetramer of subunits generated from the FH gene. Each subunit contains 3 domains, Domain 1, Domain 2, and Domain 3 (D1, D2, and D3 respectively) (Figure 1a). D1 is formed from residues in the range 49-188, D2 is formed from residues in the range 189-439, and D3 from residues in the range 440-510. The full functional protein is an assembly of 4 subunits and contains 4 identical binding pockets made of interactions between 3 subunits (Figure 1b). There are two proposed regions of importance for catalysis of the fumarate/malate conversion; Site A, the known active site (hereafter referred to as the binding site), and Site B, a region of proposed but unknown functional importance[26,27]. For this study we chose to only include the known catalytic site, Site A, defined as residues HIS176, ASN182, SER186, SER187, ASN188, THR234, HIS235, LYS371, VAL372, ASN373, and GLU378 (Figure 1c). We do not consider Site B due to the unknown and conflicting evidence surrounding its importance. For this study we chose to focus on the crystal structure 5upp[28], which covers residues 49-510 of the 510 residue protein assembled into a homotetramer.

To study mutations known or suspected to have roles in human disease, we investigated the Fumarate Hydratase Mutation Database[14], which contains 378 mutations, including 113 that are distinct missense, at the time of this study. The Fumarate Hydratase Mutation Database attempts to pool all observed mutations in FH, including those that are benign, and a large number of mutations have no clinical or functional annotation. Mutations that are not known to be benign (i.e those either labelled as loss-of-function, or those which are uncharacterised) are shown in Figure 1d.

We calculated 1D clusters of mutations across the entire FH sequence. We chose to include the top 5 predicted clusters, ranked by significance, and with a size less than 50 residues long (all calculated clusters are available in Table S1). We find the most significant clusters are all within the region of the more prevalent mutations in residues 230 and 233, indicating that this region is statistically highly over mutated, and potentially a mutationally vulnerable site.

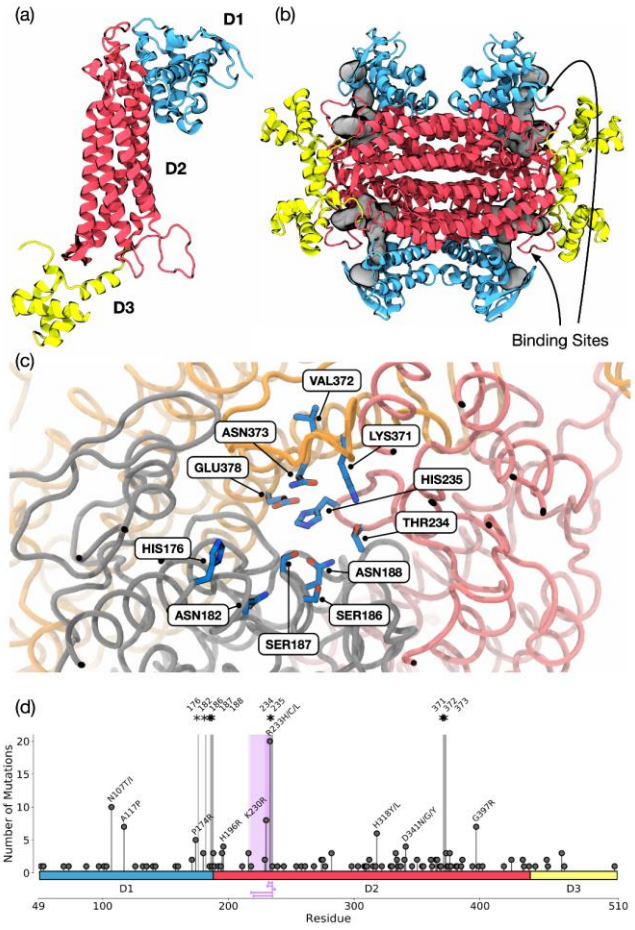


Figure 1. Structure and observed mutations in Fumarate Hydratase. (a) Structure of a single subunit of FH showing the D1, D2, and D3 regions. (b) Structure of an assembled homotetramer of FH. Binding sites are highlighted and made up on an interface between 3 subunits. (c) Close up of the binding site of FH showing the residues involved in catalytic activity. (d) Mutational spectrum of non-benign single amino acid substitutions in FH. D1, D2, and D3 regions are highlighted in blue, red, and yellow respectively. Stars (*) indicate residues involved in catalytic activity that make up the binding site of FH. Purple highlight and lines represent the top 5 mutational clusters as calculated by the NMC algorithm.

Classification of mutations by proximity to the binding site and protein hinges. Residues of the catalytic site in FH have been previously identified as essential for the conversion of fumarate to malate. We define binding site-associated residues as those with alpha-carbons (CA) within 6 Å of the CA of any binding site residue. Mutations this close to the binding site are likely to disrupt the assembly of the binding site, and we predict them to be Loss of Function (LOF). A significant number of known mutations are binding-site adjacent (Figure 2a) and significantly mutated.

We also surmised that regions involved in the "hinging" of FH domains may influence the binding site assembly due to the proximity and reliance of the quaternary structure of multiple domains to make up the binding pocket. We used Gaussian Network Modelling (GNM) within Prody[19,29] to predict hinge residues between the subdomains of the protein (Figure 2b). We used the second normal mode for calculation of hinge residues, as the first normal mode reflects a flexing of the quaternary structure (confirmed with anisotropic network modelling - ANM). Calculating the hinge residues results in residues 196, 198, 232, 242, 270, 317, 401, 411, and 448 being the most likely "hinge points" in the structure, these residues are shown on a single subunit of FH, coloured by eigenvector direction in Figure 2c. To assess how mutations to this region disrupt the quaternary structure of FH, we chose to simulate the known R233H mutant, and the wild type (WT) tetrameric assemblies for 200ns each using molecular dynamics simulations. Measuring the angles between CA atoms of two residues in the centre of the D2 and D3 domains with respect to the hinge reveals that the R233H mutant reduces the angle of the domains by an average of 8 degrees, and so leads to a partial occlusion of the catalytic site of FH (Figure 2d). From this evidence we conclude that disruption of these hinges are likely to alter the binding site and assembly of FH – and are likely loss-of-function. We chose to treat all mutations with CA atoms within 6 Å of any hinge residue as potentially LOF through disruption of the protein quaternary structure.

Overall we infer that mutations near to a binding site, or a hinge region of the protein are likely loss-of-function. A significant proportion of mutations from our database can be classed as either binding site-associated, or hinge-associated, including a number of known loss-of-function variants. Whilst 42 residues in the 462 amino acid protein structure (9%) are classified as being "binding site-associated", we find that 11 of the 30 (36%) known LOF mutations are within these residues, showing a clear statistical enrichment towards binding site-associated mutations vs mutations occurring randomly across the structure (Chi squared p value < 0.001). Similarly, 55 of the 462 (12%) amino acids in the protein structure are classified as "hinge-associated", and we find 7 of the 30 (23%) within the FH mutation database fulfil this classification, showing a lesser, but still large occurrence bias and enrichment (Chi squared p value < 0.001). Distance calculations for all potential mutations are included in Table S2.

High-Throughput mutational stability screen of FH in silico. To study how mutations that are not near the binding site or hinge regions may have effects on the structure of the protein, we sought to generate predicted mutational energy changes (ΔΔG) for every potential amino acid substitution in the FH structure. To validate our methods we chose to run a mutation saturation screen on the 1lyd structure of the bacterial T4 lysozyme[30] (Figure S1), which has been experimentally screened previously[31]. We used two methods to calculate ΔΔG; the FoldX empirical forcefield method for saturation screening, and the Rosetta cartesian_ddg method, which utilizes monte-carlo dynamics to explore the mutant conformation and influence on the protein energetics. Calculating ΔΔG for every potential substitution in T4 lysozyme results in a high rank correlation between the Foldx and Rosetta methods (spearman rank = 0.68, p <0.001, pearson r = 0.58) (Figure S2, Table S3). The methods generally agree on overall destabilizing and stabilizing mutants with disagreements between the methods generally being the magnitude of extreme destabilizing (> 10 Kcal/mol) mutations, and mutations with ΔΔG between 1 and -1 Kcal/mol.

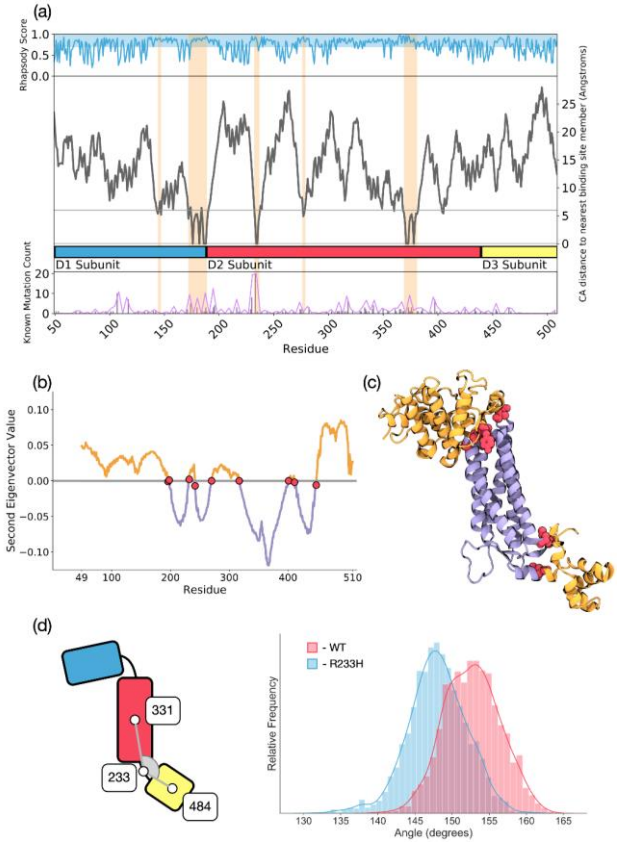


Figure 2. Mutations can be categorised on proximity to functional regions of FH. (a) Alpha carbon (CA) distance from a binding site residue. Shown is: Top: average rhapsody score for each residue, Middle: distance of each residue from a binding site residue by CA distance, Bottom: mutational frequency for each residue. Orange highlights show some regions have high rhapsody scores, low binding site distance, and high mutational frequency. (b) Second normal mode eigenvectors per residue for a single subunit of FH. Residues with an eigenvector above the line are moving generally opposed to those with an eigenvector below the line. Predicted hinge residues are shown in red. (c) Single subunit of FH coloured according to eigenvector direction (positive as orange and negative as purple). Hinge residues are highlighted as red. (d) Molecular Dynamics simulations of hinge mutations shows altered hinge flexibility. Left: Schematic of the angle measured in each simulation, Right: Angle of WT (red), and R233H mutant FH (blue) over a 200 ns equilibrium molecular dynamics simulation.

In the historic study, mutations were classified according to the growth of plaques, with mutants classified into normal growth (mutations that have no effect on the protein), mutations that induce small and large growth impairment, and mutations that result in no growth, and are therefore not tolerated. We averaged the ΔΔG from both applied methods, and binned each mutation into their respective category as determined by Rennell et al. Average ΔΔG correlates significantly with experimental growth rate (Figure 3a). Mutations that don't influence growth experimentally have a significantly lower average ΔΔG (median value 1.65 Kcal/mol) compared to those that showed no growth at all (median value 9.99 Kcal/mol), confirming the validity of the ΔΔG calculations. ROC analysis (Figure 3b) identifies the best cutoff for discriminating between small growth impairment and benign mutations is in the range of 2.5 – 3 Kcal/mol (lowest distance to corner for cutoff 2.65 Kcal/mol). We chose to apply a conservative 2.5 Kcal/mol definition for LOF, as for patient screening a larger number of false positives is favourable over false negatives, and this cutoff has been used previously in studies of mutational energy[32,33]. ROC analysis for all mutations identifies a slightly higher cutoff of 3 Kcal/mol as optimal, but agrees that 2.5 Kcal/mol is suitable for good discrimination between LOF and benign mutations (Figure S3). We further confirm the validity of the method by calculating the number of mutants of each residue that are predicted to be disruptive through both the experimental data, and the predicted ΔΔG – applying our cutoff of 2.5 Kcal/mol, Figure 3c. We find a strong correlation between the predicted and experimental disruption rate (pearson r = 0.68, p<0.001), in particular both reproducing peaks at residue 25 and residue 100. Predicted ΔΔGs are included in Table S3.

Having validated our workflow, we next applied the calculation to the human fumarate hydratase structure 5upp. Because HLRCC is caused by single allelic mutation of FH we chose to mutate a single subunit within a tetrameric structure, to represent cases where a single mutant allele combines with 3 wild type (WT) copies of FH. We again find a good agreement between the FoldX and Rosetta methods (spearman r = 0.67, p <0.0001, pearson r = 0.56) for all mutational energies (Figure 3d). This correlation is especially good given previous reports showing an overlap between Foldx and Rosetta-ddg of only 12%-25% when considering stabilizing mutations[34]. Notably however, both methods appear to agree on predictions of mutations with extremely high energy (though not necessarily on the magnitude), but there is a significant portion of the distribution that shows a reasonably poor correlation, particularly mutations that have a predicted ΔΔG between 1 and -1 Kcal/mol. As further validation that our cutoff of 2.5 Kcal/mol is appropriate for defining destabilizing mutations, we find a good separation between non-binding site adjacent mutations that are annotated as benign or damaging (Figure 3e). Due to the exposed loops within the FH structure, we additionally chose to apply a surface area-based cutoff to exclude mutations that are of high energy but within the solvent exposed loops, as these mutations will not lead to structural unfolding. We count mutations as damaging only if they have a predicted average ΔΔG >= 2.5 Kcal/mol, and a relative solvent accessibility (RSA) below 20%, in keeping with previous studies[35,36], indicating that it is buried within the structure. Across all potential mutations we find that ~45% (3968 out of 8778) meet this criterion (Figure S4). This fits well with the mutational screen of, T4 lysozyme, which found that 45% of mutational sites assessed lead to structural inactivation of enzymatic function.

We find that regions of higher overall mutational disruption are those packed within the centre of D1, and on the interface of D1 and D2, suggesting disruption to the D1/D2 interface that will alter the binding site conformation of the protein. (Figure 4)

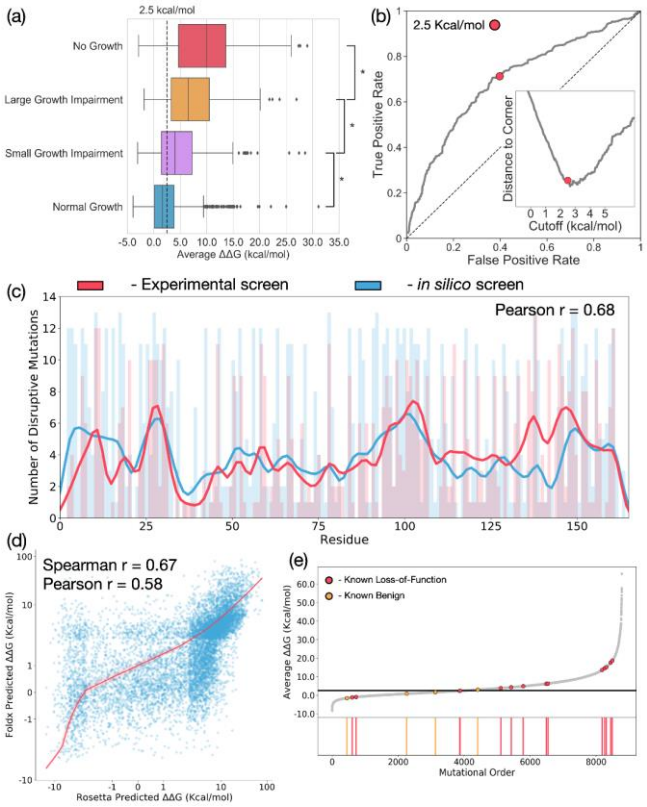


Figure 3. High-Throughput mutational stability screen of T4 lysozyme and FH in silico. a-c) T4 lysozyme. d-e) FH. (a) Average ΔΔG for each category of mutant in 1lyd. Categories derived from Rennel et al. * indicates student t-test p <0.05. (b) ROC curve for discrimination between mutations causing normal growth and small growth impairment as classified by Rennel et al. Inset: distance to corner. (c) Number of disruptive mutations per residue, red line represents experimental findings from Rennel et al., blue represents in silico mutations with a ΔΔG > 2.5 Kcal/mol. (d) Correlation between Foldx and Rosetta methods for all mutations in Human Fumarate Hydratase. (e) Position of known loss-of-function (red) and known benign (orange) mutations on the ΔΔG spectrum, from low to high. Black line represents 2.5 Kcal/mol cutoff.

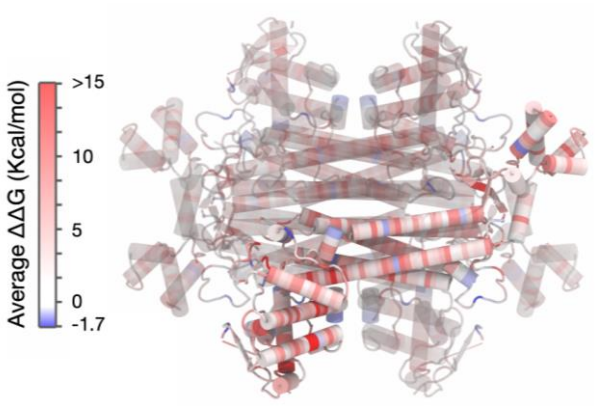


Figure 4. Crystal structure of Human Fumarate Hydratase 5upp, coloured by average residue ΔΔG.

Existing mutations are accurately categorised based on known phenotypic effects. Overall, by combining the previously defined metrics, we define a scheme for classify mutations as predicted LOF, or predicted benign, as well as sub-categorising LOF mutations into binding-site associated, hinge-site associated, and destabilizing. (Figure 5a). The initial structure is relaxed using FoldX, before the binding site and hinge regions are calculated and classified, mutations that are potentially destabilizing are defined based on average energy from the Rosetta and FoldX mutation methods, plus filtered for buried mutations through calculating the RSA. This results in a categorisation for every mutation, where is it classified as predicted silent, binding site, hinge site, or destabilizing (including combinations of disruptive mutation types) (Table S2). We classify 5811 out of 8778 (66%) mutations as predicted LOF, similar to a study of mutational effects on TP53, which found that roughly 50-60% of all possible mutations were functionally disruptive[37]. We compared our predictions with known mutations in the FH mutations database. We predicted a classification for all mutations within the database and compared this their known effects: loss of function (LOF), benign, or unknown. In total 34 mutations had a known (or implied) functional effect, either LOF or benign, whilst 74 were classified as unknown (Table S4). We find that 24 out of 30 (80%) mutations are correctly classified as LOF using our classification scheme, and 3 out of 4 (75%) are correctly classed as benign (Figure 5b). Of the mutations incorrectly classified as benign when they are known to be LOF, two mutations involve cysteine (C434Y, Y465C), which is known to be modelled poorly by Rosetta cartesian_ddg[38], The single mutation our methodology classified as LOF when it is listed as benign within the FH mutation database is R268G. We predict the R268G mutation to be both destabilizing (ΔΔG > 2.5 Kcal/mol, RSA < 0.2) and hinge-associated. Whilst the mutation is listed as benign, no experimental information is cited, and PolyPhen-2[39], and Rhapsody also classify this particular mutation as damaging, indicating that the benign classification for this mutation may be questionable. We ran a molecular dynamics simulation of the R268G mutant. Simulations predict that mutant R268G reduces the hinge angle of the D1/D2 domains by ~5 degrees (Figure S5), and supports previous evidence from analysis of the pathogenic R233H mutant, that hinges within he protein can affect binding site assembly. Of the 74 unknown mutations, we predict that 28 are functionally benign, and 46 are potential LOF mutations.

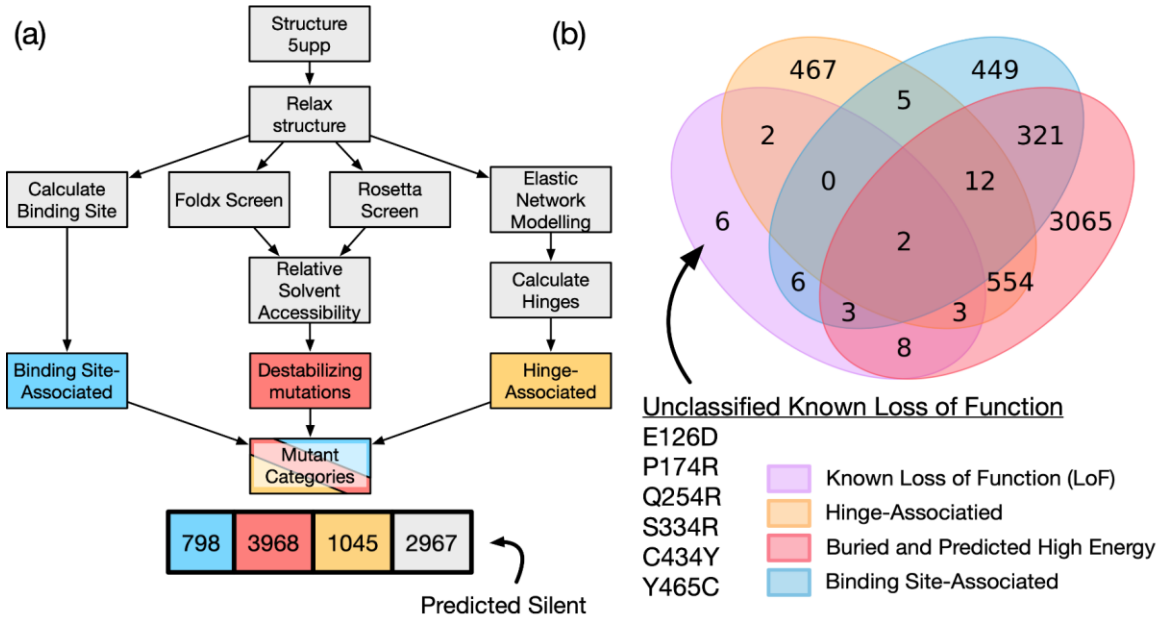


Figure 5. Classification of functionally characterized mutations. (a) schema for categorization of mutations in human fumarate hydratase. (b) Overlap between Hinge-associated (orange), destabilizing (red), and binding site associated mutations (blue). 30 known Loss-of-function mutations are included (purple). 24 mutations are correctly identified as LoF, whilst 6 are incorrectly classified as benign.

Mutations with unknown properties can be accurately predicted to be functional or neutral. To visualise all potential mutations in FH we chose to cluster all mutations using umap[23]. Umap clusters items by similarity, in a manner similar to principal component analysis, or TSNE[40]. We ran umap on all mutations using the 4 major axes involved in the classification – Minimum distance to a binding site residue, minimum distance to a hinge residue, average ΔΔG of mutation, and RSA for each residue (Figure 6a). We find that distinct regions of the plot cluster into functionally different mutations when coloured by classification. There is a region specifically for hinge-associated mutations, binding site-associated, and unknown (not predicted damaging) mutations. In particular, the region of "unknown" (not classified as damaging) mutations overlaps significantly with a number of predicted destabilizing mutations, indicating that discrimination between these mutations is difficult, and perhaps not accurate with currently available data. We find that most of the benign mutations, aside from R268G are found clearly within the regions of predicted benign mutations. R268G clusters with the hinge mutation region as expected from our previous classification. For the known LOF mutations, we find they mostly cluster within the well defined regions for binding site, hinge, and destabilizing mutations. There are some mutations, particularly those which were misclassified, that fall within ambiguous regions of state space in the mutational landscape, and so are hard to classify using our defined criterion.

To test the predictive power of our classification scheme we used the Cancer Cell Line Encyclopedia to look for changes in metabolite levels associated with mutations in FH[17,41]. We find 42 mutations (35 unique) in FH within 34 individual cell lines (Table S5). Selecting only for missense mutations yielded 25 mutations (20 unique) within 23 unique cell lines. We classified the mutations according to our criterion as either predicted LOF, or predicted benign. We find that by analysis of metabolomics data included in the CCLE database, mutations that we predict to be LOF have a higher average level of fumarate/mateate/alpha-ketoisovalerate detected in media than cells with predicted benign mutations ($p = 0.035$) – indicating that these cell lines may have an accumulation of fumarate as a result of inactive levels of FH (Figure 6b, c).

CONCLUSIONS

We have shown, using a comprehensive combination of techniques, that we can categorise with a high degree of confidence the functional effects of any potential missense mutation in FH. Beyond FH, we present an integrated series of methods that can be adapted for mutationally screening any protein for functionally relevant mutations in a reasonably small amount of computational time. Our workflow predicts the functional effects of all mutations that can be compared to existing methods based on machine-learning principles such as rhapsody and polyphen, at significantly lower time and effort expenditure than experimental characterization. Whilst some other methods incorporate some manner of structural analysis in their predictions, ours demonstrates a new perspective, as it explicitly models every potential mutation in a structure, allowing it to interface directly with other computational techniques in the field such as molecular dynamics simulations to further study mutations of interest.

Biologically we propose three ways in which mutations can potentially disrupt the catalytic activity of FH, in particular we find that addition of hinge altering mutations are necessary for classification of many known LOF mutations, indicating that there is a biological relevance, and hinting at a mechanism for mutations that change the flexibility and stiffness of protein hinges in this case. Additionally, we chose to exclude site B from our analysis of mutation disruption and find that we are able to classify almost all known mutations without its inclusion. This implies that mutations in site B may not have functional or disease-related relevance, despite some evidence that site B can alter catalytic activity of the enzyme[42]. This is reinforced by the fact that 27 of the 462 residues within the protein structure are classified as near site B (6%), and only 3 of 30 residues in the FH mutation database (10%) are near to site B, showing a poor to negligible enrichment of mutations in site B (Chi squared $p = 0.187$).

Fumarate hydratase represents a good first-use case for a high-throughput mutational screen due to the need to understand mutations in their functional context, but as mutational detection techniques, and high-throughput mutational studies increase the need to be able to classify mutations confidently as benign and LOF is more important. Here we show that our method classifies known LOF and benign mutations with a high level of confidence, and predict which mutations discovered in the human population are likely to have functional relevance, and therefore predispose patients to particular metabolic diseases.

Whilst the accuracy of our method with the current data is high, there are clear regions where the analysis is not able to discriminate between mutations on the borderline between destabilizing and benign, this results from the lack of accuracy in the mutational ΔΔG calculations, despite using a consensus of the best available methods at time of study[5]. As better methods become available it will be of interest to improve upon this work to attempt a more accurate classification. It should also be noted that whilst a ΔΔG cutoff of 2.5 Kcal/mol to discriminate benign vs destabilizing mutants is a good starting point for any given structure, ROC analysis should be performed if the number of "ground truth" mutational effects is high enough to confirm this is appropriate for any particular protein structure.

Finally, whilst the work here focusses on a single molecule within the TCA cycle, FH, structural data has existed for a large number of enzymes within the cycle for some time[43–45] and it would be of great interest to look into mutations across entire metabolic pathways. With this study laying the groundwork, it will be of future interest to model all mutations in all enzymes, and attempt to further link these with genomic and metabolomic data that is already available.

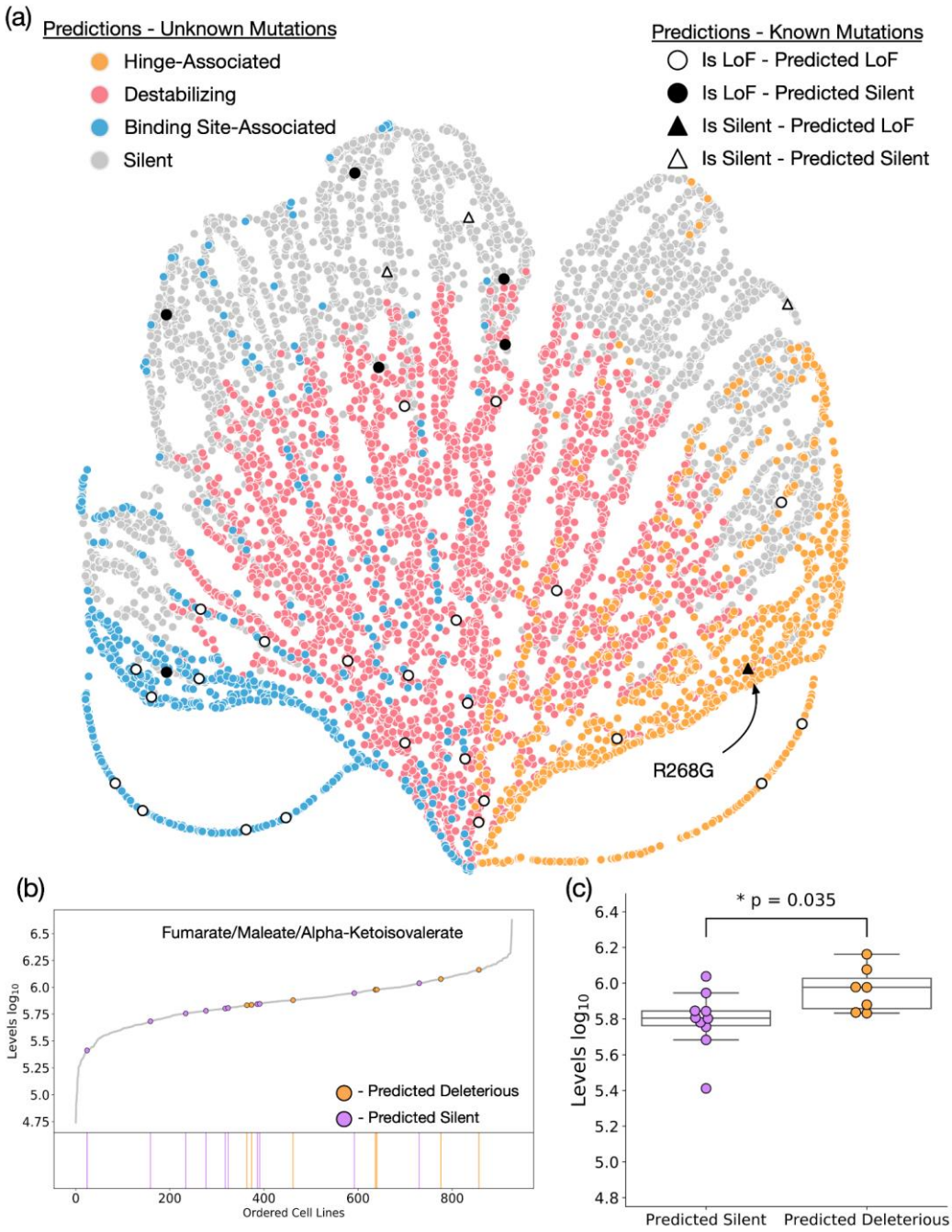


Figure 6: Mutational Landscape of Fumarate Hydratase. A) Umap for all mutations in FH. Mutations are coloured by classification. Hinge-associated (orange), Destabilizing (red), and Binding site-associated (blue) are shown clustered into groups. Predicted silent mutations (grey) are also shown. Overlayed are our predictions for characterized mutations in the FH mutation database. Mutations that are known Loss-of-Function (LOF) are circular and coloured according to whether we predict them to be LOF (black) or silent (white). Known benign mutations are in triangles, and also coloured according to whether we predict them to be LOF (black) or silent (white). The questionable known benign mutation R268G is labelled B) Mutations in the Cancer Cell Line Encyclopedia (CCLE) metabolomics data. All cell lines are ranked according to their detected levels of Fumarate/Maleate/Alpha-Ketoisovalerate. Coloured are cell lines with mutations in FH that we predict to be LOF (orange), or silent (purple). C) Swarmplot for levels of Fumarate/Maleate/Alpha-Ketoisovalerate in mutant FH cell lines. Mutations predicted to be silent are significantly lower than mutations predicted to be LOF (p value represents independent T test). Error bars represent 1.5 * interquartile range (IQR).

ASSOCIATED CONTENT

Supporting Information. Additional figures are contained in supplementary information.

DATA AND SOFTWARE AVAILABILITY

All data generated for this manuscript is available in supplementary information. Supplementary tables 1 and 2 contain all predicted values for every mutation studied. Code required to generated every figure within the manuscript is available at: https://github.com/shorthouse-mrc/Fumarate_Hydratase

AUTHOR INFORMATION

Corresponding Author – Benjamin A Hall, b.hall@ucl.ac.uk

Present Addresses – Department of Medical Physics and Biomedical Engineering, UCL, London, WC1E 6BT

Author Contributions

DS and BAH conceived the study and wrote the manuscript. DS generated all data and performed all analysis. MWJH contributed to development and testing of the pipeline. All authors were responsible for editing of the manuscript.

Funding Sources

This work was supported by the Medical Research Council (grant no. MR/S000216/1). M.W.J.H. acknowledges support from the Harrison Watson Fund at Clare College, Cambridge. B.A.H. acknowledges support from the Royal Society (grant no. UF130039).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the Frezza group, in particular Christian Frezza for support and constructive feedback during the generation of this manuscript.

ABBREVIATIONS

LOF; Loss-of-function; GOF, Gain-of-function, FH, Fumarate Hydratase; WT, Wild type; RSA, Relative Solvent Accessibility; NMC, nonrandom mutational clustering.

REFERENCES

(1) Shendure, J.; Akey, J. M. The Origins, Determinants, and Consequences of Human Mutations. *Science* **2015**, *349* (6255), 1478–1483.

(2) Delgado, J.; Radusky, L. G.; Cianferoni, D.; Serrano, L. FoldX 5.0: Working with RNA, Small Molecules and a New Graphical Interface. *Bioinformatics* **2019**, *35* (20), 4168–4169.

(3) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33* (Web Server), W382–W388.

(4) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031–3048.

(5) Strokach, A.; Corbi-Verge, C.; Kim, P. M. Predicting Changes in Protein Stability Caused by Mutation Using Sequence-and Structure-based Methods in a CAGI5 Blind Challenge. *Hum. Mutat.* **2019**, *40* (9), 1414–1423.

(6) Coman, D.; Kranc, K. R.; Christodoulou, J. *Fumarate Hydratase Deficiency*; University of Washington, Seattle, 1993.

(7) Yang, M.; Soga, T.; Pollard, P. J.; Adam, J. The Emerging Role of Fumarate as an Oncometabolite. *Front. Oncol.* **2012**, *2*, 85.

(8) Leshets, M.; Ramamurthy, D.; Lisby, M.; Lehming, N.; Pines, O. Fumarase Is Involved in DNA Double-Strand Break Resection through a Functional Interaction with Sae2. *Curr. Genet.* **2018**, *64* (3), 697–712.

(9) Yogev, O.; Yogev, O.; Singer, E.; Shaulian, E.; Goldberg, M.; Fox, T. D.; Pines, O. Fumarase: A Mitochondrial Metabolic Enzyme and a Cytosolic/Nuclear Component of the DNA Damage Response. *PLoS Biol.* **2010**, *8* (3), e1000328.

(10) Schmidt, C.; Sciacovelli, M.; Frezza, C. Fumarate Hydratase in Cancer: A Multifaceted Tumour Suppressor. *Semin. Cell Dev. Biol.* **2020**, *98*, 15–25.

(11) Skala, S. L.; Dhanasekaran, S. M.; Mehra, R. Hereditary Leiomyomatosis and Renal Cell Carcinoma Syndrome (HLRCC): A Contemporary Review and Practical Discussion of the Differential Diagnosis for HLRCC-Associated Renal Cell Carcinoma. *Arch. Pathol. Lab. Med.* **2018**, *142* (10), 1202–1215.

(12) Alam, N. A.; Olpin, S.; Leigh, I. M. Fumarate Hydratase Mutations and Predisposition to Cutaneous Leiomyomas, Uterine Leiomyomas and Renal Cancer. *Br. J. Dermatol.* **2005**, *153* (1), 11–17.

(13) Clark, G. R.; Sciacovelli, M.; Gaude, E.; Walsh, D. M.; Kirby, G.; Simpson, M. A.; Trembath, R. C.; Berg, J. N.; Woodward, E. R.; Kinning, E.; Morrison, P. J.; Frezza, C.; Maher, E. R. Germline *FH* Mutations Presenting With Pheochromocytoma. *J. Clin. Endocrinol. Metab.* **2014**, *99* (10), E2046–E2050.

(14) Bayley, J.-P.; Launonen, V.; Tomlinson, I. P. The FH Mutation Database: An Online Database of Fumarate Hydratase Mutations Involved in the MCUL (HLRCC) Tumor Syndrome and Congenital Fumarase Deficiency. *BMC Med. Genet.* **2008**, *9* (1), 20.

(15) Rennell, D.; Bouvier, S. E.; Hardy, L. W.; Poteete, A. R. Systematic Mutation of Bacteriophage T4 Lysozyme. *J. Mol. Biol.* **1991**, *222* (1), 67–88.

(16) Ghandi, M.; Huang, F. W.; Jané-Valbuena, J.; Kryukov, G. V.; Lo, C. C.; McDonald, E. R.; Barretina, J.; Gelfand, E. T.; Bielski, C. M.; Li, H.; Hu, K.; Andreev-Drakhlin, A. Y.; Kim, J.; Hess, J. M.; Haas, B. J.; Aguet, F.; Weir, B. A.; Rothberg, M. V.; Paolella, B. R.; Lawrence, M. S.; Akbani, R.; Lu, Y.; Tiv, H. L.; Gokhale, P. C.; de Weck, A.; Mansour, A. A.; Oh, C.; Shih, J.; Hadi, K.; Rosen, Y.; Bistline, J.; Venkatesan, K.; Reddy, A.; Sonkin, D.; Liu, M.; Lehar, J.; Korn, J. M.; Porter, D. A.; Jones, M. D.; Golji, J.; Caponigro, G.; Taylor, J. E.; Dunning, C. M.; Creech, A. L.; Warren, A. C.; McFarland, J. M.; Zamanighomi, M.; Kauffmann, A.; Stransky, N.; Imielinski, M.; Maruvka, Y. E.; Cherniack, A. D.; Tsherniak, A.; Vazquez, F.; Jaffe, J. D.; Lane, A. A.;

Weinstock, D. M.; Johannessen, C. M.; Morrissey, M. P.; Stegmeier, F.; Schlegel, R.; Hahn, W. C.; Getz, G.; Mills, G. B.; Boehm, J. S.; Golub, T. R.; Garraway, L. A.; Sellers, W. R. Next-Generation Characterization of the Cancer Cell Line Encyclopedia. *Nature* **2019**, *569* (7757), 503–508.

(17) Li, H.; Ning, S.; Ghandi, M.; Kryukov, G. V.; Gopal, S.; Deik, A.; Souza, A.; Pierce, K.; Keskula, P.; Hernandez, D.; Ann, J.; Shkoza, D.; Apfel, V.; Zou, Y.; Vazquez, F.; Barretina, J.; Pagliarini, R. A.; Galli, G. G.; Root, D. E.; Hahn, W. C.; Tsherniak, A.; Giannakis, M.; Schreiber, S. L.; Clish, C. B.; Garraway, L. A.; Sellers, W. R. The Landscape of Cancer Cell Line Metabolism. *Nat. Med.* **2019**, *25* (5), 850–860.

(18) Ye, J.; Pavlicek, A.; Lunney, E. A.; Rejto, P. A.; Teng, C.-H. Statistical Method on Nonrandom Clustering with Application to Somatic Mutations in Cancer. *BMC Bioinformatics* **2010**, *11* (1), 11.

(19) Bakan, A.; Meireles, L. M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27* (11), 1575–1577.

(20) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

(21) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40* (7), 843–856.

(22) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (3), 830–838.

(23) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3* (29), 861.

(24) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32* (10), 2319–2327.

(25) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423.

(26) Picaud, S.; Kavanagh, K. L.; Yue, W. W.; Lee, W. H.; Muller-Knapp, S.; Gileadi, O.; Sacchettini, J.; Oppermann, U. Structural Basis of Fumarate Hydratase Deficiency. *J. Inherit. Metab. Dis.* **2011**, *34* (3), 671–676.

(27) Rose, I. A.; Weaver, T. M. The Role of the Allosteric B Site in the Fumarase Reaction. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (10), 3393.

(28) Ajalla Aleixo, M. A.; Rangel, V. L.; Rustiguel, J. K.; de Pádua, R. A. P.; Nonato, M. C. Structural, Biochemical and Biophysical Characterization of Recombinant Human Fumarate Hydratase. *FEBS J.* **2019**, *286* (10), 1925–1940.

(29) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian Dynamics of Folded Proteins. *Phys. Rev. Lett.* **1997**, *79* (16), 3090–3093.

(30) Rose, D. R.; Phipps, J.; Michniewicz, J.; Birnbaum, G. I.; Ahmed, F. R.; Muir, A.; Anderson, W. F.; Narang, S. Crystal Structure of T4-Lysozyme Generated from Synthetic Coding DNA Expressed in Escherichia Coli. *Protein Eng.* **1988**, *2* (4), 277–282.

(31) Rennell, D.; Bouvier, S. E.; Hardy, L. W.; Poteete, A. R. Systematic Mutation of Bacteriophage T4 Lysozyme. *J. Mol. Biol.* **1991**, *222* (1), 67–88.

(32) Abildgaard, A. B.; Stein, A.; Nielsen, S. V.; Schultz-Knudsen, K.; Papaleo, E.; Shrikhande, A.; Hoffmann, E. R.; Bernstein, I.; Gerdes, A.-M.; Takahashi, M.; Ishioka, C.; Lindorff-Larsen, K.; Hartmann-Petersen, R. Computational and Cellular Studies Reveal Structural Destabilization and Degradation of MLH1 Variants in Lynch Syndrome. *eLife* **2019**, *8*.

(33) Bromberg, Y.; Rost, B. Correlating Protein Function and Stability through the Analysis of Single Amino Acid Substitutions. *BMC Bioinformatics* **2009**, *10* (S8), S8.

(34) Buß, O.; Rudat, J.; Ochsenreither, K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Comput. Struct. Biotechnol. J.* **2018**, *16*, 25–33.

(35) Chen, H. Prediction of Solvent Accessibility and Sites of Deleterious Mutations from Protein Sequence. *Nucleic Acids Res.* **2005**, *33* (10), 3193–3199.

(36) Li, X.; Pan, X.-M. New Method for Accurate Prediction of Solvent Accessibility from Protein Sequence. *Proteins Struct. Funct. Genet.* **2001**, *42* (1), 1–5.

(37) Kotler, E.; Shani, O.; Goldfeld, G.; Lotan-Pompan, M.; Tarcic, O.; Gershoni, A.; Hopf, T. A.; Marks, D. S.; Oren, M.; Segal, E. A Systematic P53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol. Cell* **2018**, *71* (1), 178-190.e8.

(38) Frenz, B.; Lewis, S. M.; King, I.; DiMaio, F.; Park, H.; Song, Y. Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy. *Front. Bioeng. Biotechnol.* **2020**, *8*..

(39) Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R. A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* **2010**, *7* (4), 248–249.

(40) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (Nov), 2579–2605.

(41) Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V.; Sonkin, D.; Reddy, A.; Liu, M.; Murray, L.; Berger, M. F.; Monahan, J. E.; Morais, P.; Meltzer, J.; Korejwa, A.; Jané-Valbuena, J.; Mapa, F. A.; Thibault, J.; Bric-Furlong, E.; Raman, P.; Shipway, A.; Engels, I. H.; Cheng, J.; Yu, G. K.; Yu, J.; Aspesi, P.; de Silva, M.; Jagtap, K.; Jones, M. D.; Wang, L.; Hatton, C.; Palescandolo, E.; Gupta, S.; Mahan, S.; Sougnez, C.; Onofrio, R. C.; Liefeld, T.; MacConaill, L.; Winckler, W.; Reich, M.; Li, N.; Mesirov, J. P.; Gabriel, S. B.; Getz, G.; Ardlie, K.; Chan, V.; Myer, V. E.; Weber, B. L.; Porter, J.; Warmuth, M.; Finan, P.; Harris, J. L.; Meyerson, M.; Golub, T. R.; Morrissey, M. P.; Sellers, W. R.; Schlegel, R.; Garraway, L. A. The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* **2012**, *483* (7391), 603–607.

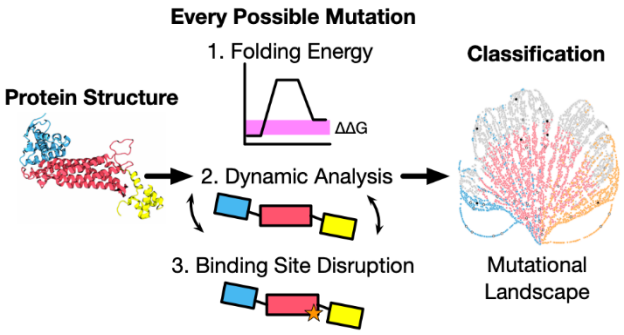
(42) Alam, N. A.; Olpin, S.; Rowan, A.; Kelsell, D.; Leigh, I. M.; Tomlinson, I. P. M.; Weaver, T. Missense Mutations in Fumarate Hydratase in Multiple Cutaneous and

Uterine Leiomyomatosis and Renal Cell Cancer. *J. Mol. Diagn.* **2005**, *7* (4), 437–443.

(43) Chapman, A. D. M.; Cortés, A.; Dafforn, T. R.; Clarke, A. R.; Brady, R. L. Structural Basis of Substrate Specificity in Malate Dehydrogenases: Crystal Structure of a Ternary Complex of Porcine Cytoplasmic Malate Dehydrogenase, α-Ketomalonate and TetrahydoNAD 1 1Edited by R. Huber. *J. Mol. Biol.* **1999**, *285* (2), 703–712.

(44) Lauble, H.; Kennedy, M. C.; Beinert, H.; Stout, C. D. Crystal Structures of Aconitase with Isocitrate and Nitroisocitrate Bound. *Biochemistry* **1992**, *31* (10), 2735–2748.

(45) Remington, S.; Wiegand, G.; Huber, R. Crystallographic Refinement and Atomic Models of Two Different Forms of Citrate Synthase at 2·7 and 1·7 Å Resolution. *J. Mol. Biol.* **1982**, *158* (1), 111–152.


Every Possible Mutation
1. Folding Energy
ΔΔG
Protein Structure
2. Dynamic Analysis
Classification
3. Binding Site Disruption
Mutational
Landscape