The Landscape of Mutations in Fumarate Hydratase

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

*Correspondence to be addressed to B.A.H (bh418@mrc-cu.cam.ac.uk)

[1] MRC Cancer Unit, University of Cambridge, Cambridge, CB2 0XZ, UK

[2] Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK

ABSTRACT

Fumarate Hydratase (FH) is an enzyme of the citric acid (TCA) cycle that is responsible for reversibly catalysing the conversion between fumarate and malate. FH loss and subsequent buildup of the oncometabolite fumarate causes hereditary leiomyomatosis and renal cell carcinoma.

We sought to explore the mutational landscape of FH in silico, to predict the functional effects of many detected mutations, and categorise detected but un-characterised mutations in human populations. Using mutational energy predicting tools such as Rosetta and FoldX we can accurately predict mutations and mutational hotspots with high disruptive capability. Furthermore, through performing molecular dynamics simulations we show that hinge regions of the protein can be stabilized or destabilized by mutations, with new mechanistic implications of the consequences on the binding affinity of the enzyme for its substrates.

We can additionally categorise a large majority of mutations and potential mutations into functional groups. This allows us to predict which detected mutations in the human population are likely to be loss-of-function, and therefore predispose patients to papillary renal carcinoma through considering only mutations to the protein binding site, hinges, and those that are buried deep within the protein. We additionally link mutation data to publicly available metabolomics data, and show that we can accurately predict which mutations in cancer cell lines are functionally relevant.

INTRODUCTION

Fumarate hydratase (FH) is a member of the tricarboxylic acid (TCA) cycle occurring in the mitochondria, and enzymatically metabolises fumarate within the cytosol. 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 an oncogenically-associated buildup of fumarate[1,2]. As a result the enzyme FH has been described as a tumor suppressor, and fumarate as part of a novel classification of molecules named “oncometabolites”. Precisely how the buildup of fumarate can be oncogenic is unknown, but recent work points towards suppression of DNA repair responses, EMT, and promotion of mitotic entry upon fumarate buildup[3–5].

Understanding the effects of mutations on the activity and assembly of FH is of importance for the understanding and stratification of germline mutations in FH, which can predispose patients with a single mutated or deleted allele to hereditary leiomyomatosis and renal cell cancer (HLRCC) upon mutational inactivation of their remaining wild-type copy[6,7]. Previous work has identified mutants linked with inherited and de-novo FH-related conditions, including cancer[8] – most notably, the FH mutation database represents a comprehensive list of mutations and their effects, if known, on FH activity[9].

In recent years numerous methods have been developed for estimating the effects of single point mutations (SNPs) on the stability of a protein structure in silico. Notable methods include FoldX[10,11], 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[12,13], which uses Monte-Carlo based dynamics to predict energetic effects of mutations. Additionally, molecular dynamics can be used to more comprehensively investigate mutant protein structure, though at significantly higher computational cost. With the advent of high-throughput methods such as CRISPR screening, and larger projects being undertaken to screen populations for mutations and disease, coupled with large-scale disease-focussed data generating projects such as The Cancer Genome Atlas (TCGA)[14] and the International Cancer Genome Consortium (ICGC)[15], the number and diversity of mutations being implicated in disease is rapidly expanding. Whilst methods to attempt to sift functionally relevant

mutations from synonymous to detect highly mutated genes exist in the form of statistical tests such as DN/DS[16], mutsig[17], and oncodrive[18], including some methods that take into account structure of the protein such as Rhapsody[19], there is scope for detailed, structure-informed, chemically aware methods to classify mutations, including those not yet observed, into Loss-of-Function (LOF) and benign categories.

Here we computationally screen 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 details of every potential mutant. We confirm that our method predicts known functionally relevant mutations, and predict from existing databases of mutations which have an unknown effect, which of them will be damaging to the activity of FH. 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)[20,21] 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.

RESULTS

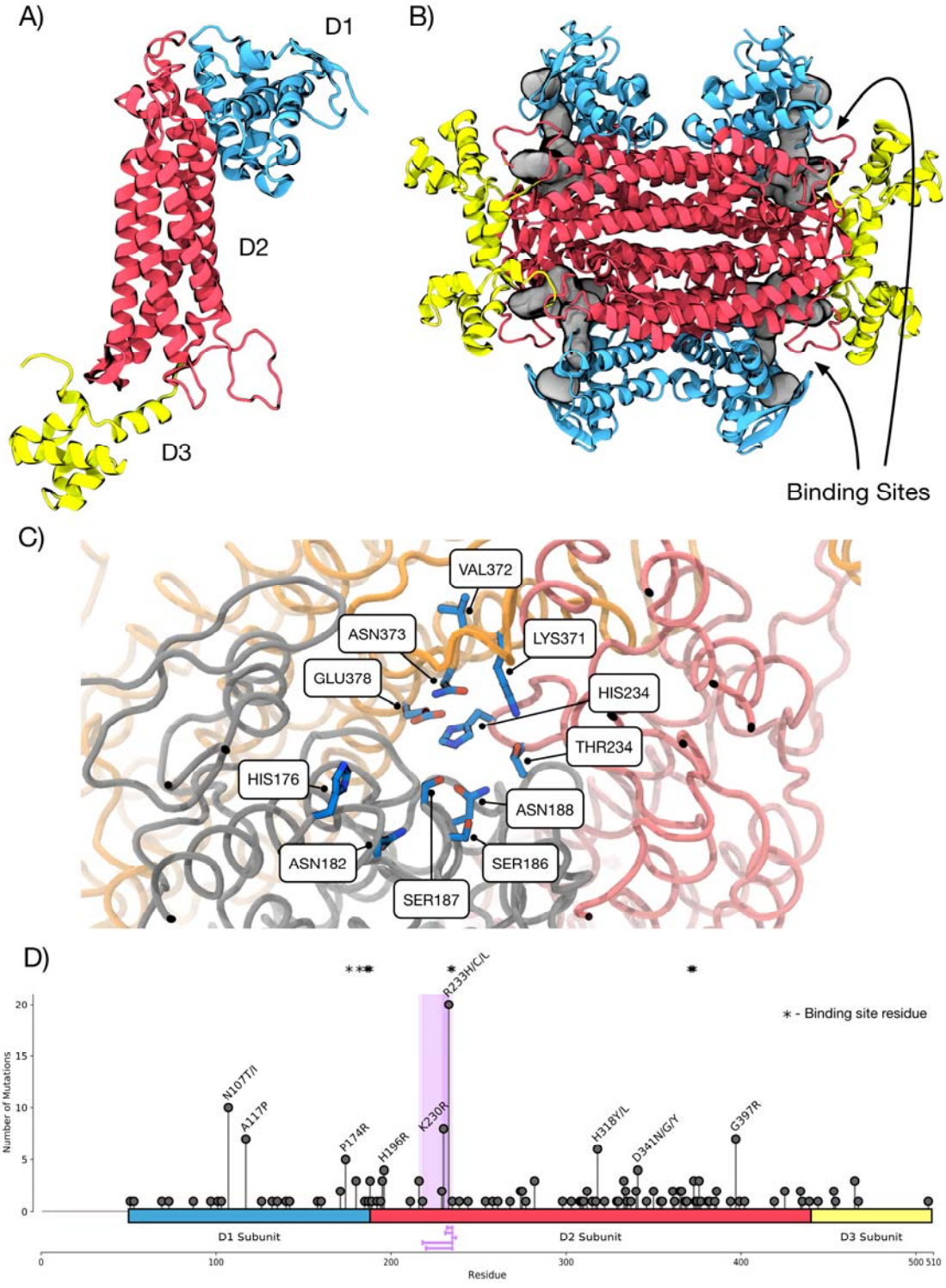
Evidence of mutational clustering in FH

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 1 A**). 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 1 B**). 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[22,23]. 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 1 C**). 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[24], 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[9], 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 1 D**. In particular, mutations at amino acids 107, 117, 230, and 233 are reported at a higher frequency than other mutations and may indicate regions of mutational vulnerability in the structure.

We applied the NMC clustering method to look for clustering of mutations across the 1D sequence of the protein[25]. We chose to include the top 5 predicted clusters, ranked by significance, and with a size less than 50 residues long. We find the most significant clusters are all within the region of the more prevalent mutations in

147 residues 230 and 233, indicating that this region is statistically highly over mutated,
148 and potentially a mutationally vulnerable site.



149
150 **Figure 1**: Structure and current mutations in Fumarate Hydratase. A) Structure of a
151 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 SNPs 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 chose to define binding site-associated residues as those with alpha-carbons (CA) within 6 Å of the CA of any binding site residue. Plotting the resultant distances for each residue in the FH structure shows that specific clusters of residues in the vicinity of the binding site are also significantly mutated (**Figure 2 A**). In particular, there is a high frequency of mutations between residues 172-189, 232-237, 277-278, and 369-381 that correspond to mutations likely to alter the binding site via proximity by our definition. Additionally, generating the Rhapsody scores[19] for each residue results in regions of predicted pathogenicity that also align with the binding site regions – reinforcing that mutations neighbouring binding sites are likely to be pathogenic purely via proximity and disruption of the precise conformation of sidechains necessary for catalysis. Whilst Rhapsody represents a potentially useful single metric for assessing mutational disruption, incorporates evidence from sequence and structure alone, without biological context.

Due to the three-domain structure of FH we surmised that regions involved in the "hinging" of these 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. To calculate predicted hinges within the structure we used Gaussian Network Modelling (GNM) within prody[26,27] to calculate the major normal modes for an individual subunit of FH. We find that the second mode best represents the hinging mode expected around the three domains of the protein. Calculating the hinge residues from the second normal mode results in residues 196, 198, 232, 242, 270, 317, 401, 411, and 448 being the most likely "hinge points" in the structure

(**Figure 2 B**), these residues are shown on a single subunit of FH, coloured by eigenvector direction in **Figure 2 C**. In order to assess whether mutation of these domains was sufficient to disrupt the quaternary structure of the protein and thus the binding site we chose to simulate a known mutation within a hinge region that is found at a high frequency in the FH mutation database using molecular dynamics simulations. We chose the simulate the R233H mutant, and the wild type (WT) tetrameric assemblies for 200ns each. Measuring the angles between CA atoms of two residues in the centre of the D2 and D3 regions 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 2 D**). From this evidence we conclude that disruption of these hinges are likely to alter the binding site and assembly of FH – and are likely pathogenic. 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 find that mutations near to either the binding site, or a hinge region of the protein are likely to disrupt or alter the protein function. We find that, from the FH mutation database, a significant proportion of mutations can be classed as either binding site-associated, or hinge-associated, including a number of known loss-of-function (LOF) variants. Whilst 42 residues in the 461 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 bias towards binding site-associated mutations. Similarly, 55 of the 461 (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. Distance calculations for all potential mutations are included in **Supplementary Table 1**.

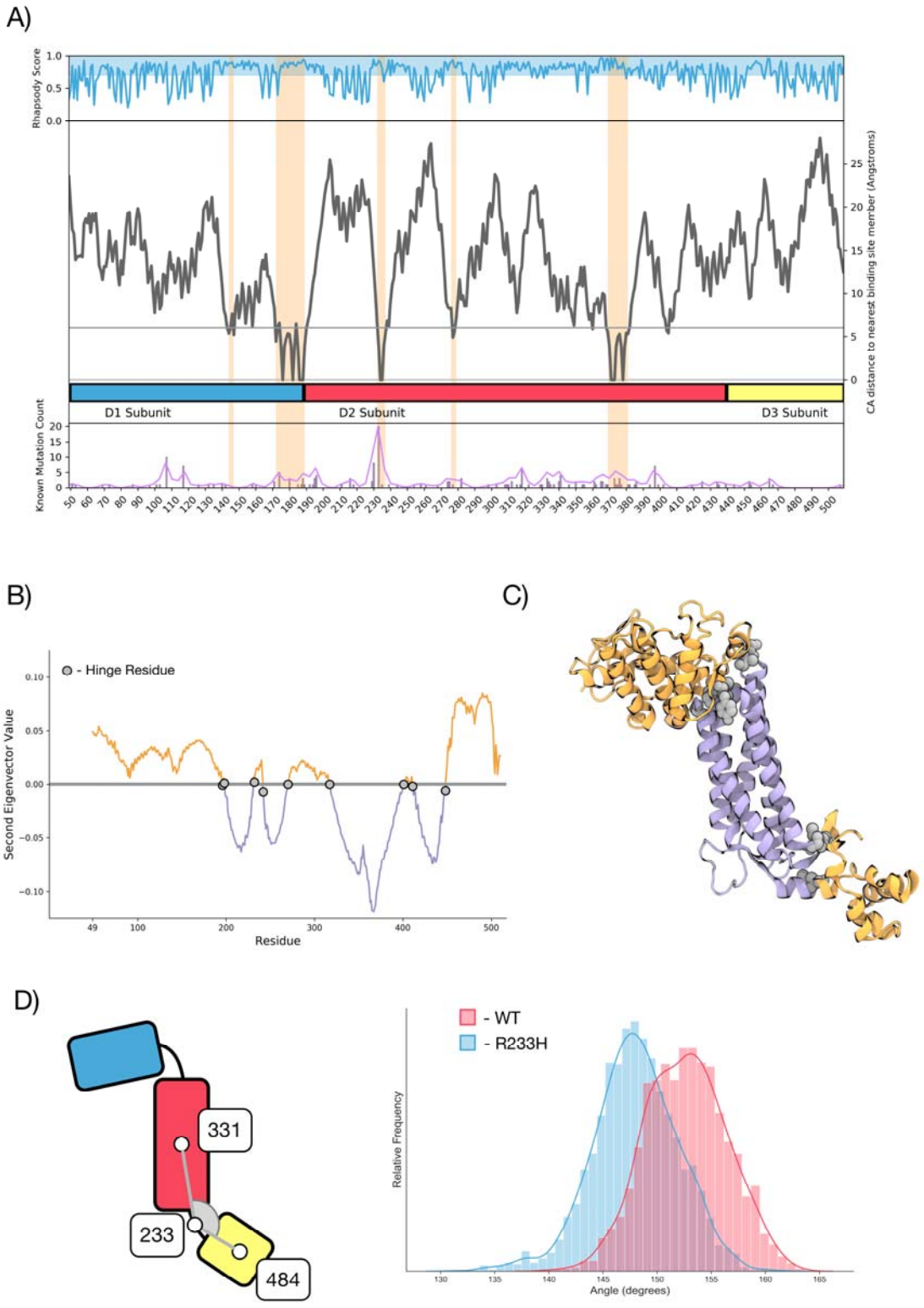


Figure 2: Mutations can be categorised on proximity to functional regions of FH. A) Alpha carbon (CA) distance from a binding site residue. Shown is: Top: average

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 grey. C) Single subunit of FH coloured according to eigenvector direction (positive as organge and negative as purple). Hinge residues are highlighted as grey. 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.

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. We chose to use two conceptually different methods and use an average between the two methods to study each potential mutant. We chose to utilize the FoldX method[10,11], and the Rosetta cartesian_ddg method[12,13], (hereafter described as the Rosetta method) to perform mutational energy calculations. Both methods have been shown to generate accurate predictions on the CAGI5 blind challenge datasets, but overlaps between the two methods on the same dataset indicate that they generally predict different mutations correctly, the combined overlap between the two being a good indicator of mutational ΔΔG[13].

To perform mutant calculations, the pdb structure 5upp was first relaxed using the FoldX RelaxPDB method, before each mutation and its resultant ΔΔG was calculated. We additionally calculate the Relative Solvent Accessible Surface Area (RSA) for each wild-type (WT) residue. Mutations on the surface of the protein are unlikely to dramatically alter the folding of the protein, so we chose to only consider a mutation potentially destabilizing if it is buried, defined as having an RSA <= 0.2.

We find a good agreement between the FoldX and Rosetta methods, with an r of 0.67 ($p < 0.0001$) for all mutational energies (**Figure 3 A**). Notably however, both methods appear to agree on predictions of mutations with extremely high energy, 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. We chose to study the average predicted energy of each mutation by taking the average ΔΔG from the two methods. Ranking the average ΔΔG over all ~9000 mutations results in a distribution of all mutational energies across the mutational landscape (**Figure 3 B**). We find that mutations known to be LOF, and that are not within 6 Å of either the binding site or hinge regions tend to cluster near the upper end of the distribution, indicating that they affect the stability of the protein. Mutations that are known benign tend to fall near the lower end of the distribution. We chose a cutoff of 2.5 kcal/mol for classifying mutations as potentially destabilizing, and any mutation over this value for average energy, and with a RSA < 0.2 was classified as destabilizing. Across all potential mutations we find that ~45% (3968 out of 8778) meet this criterion (**Figure 3 C**). This fits roughly with historical data of mutational stability in T4 lysozyme, which found that 45% of mutational sites lead to structural inactivation of enzymatic function[28].

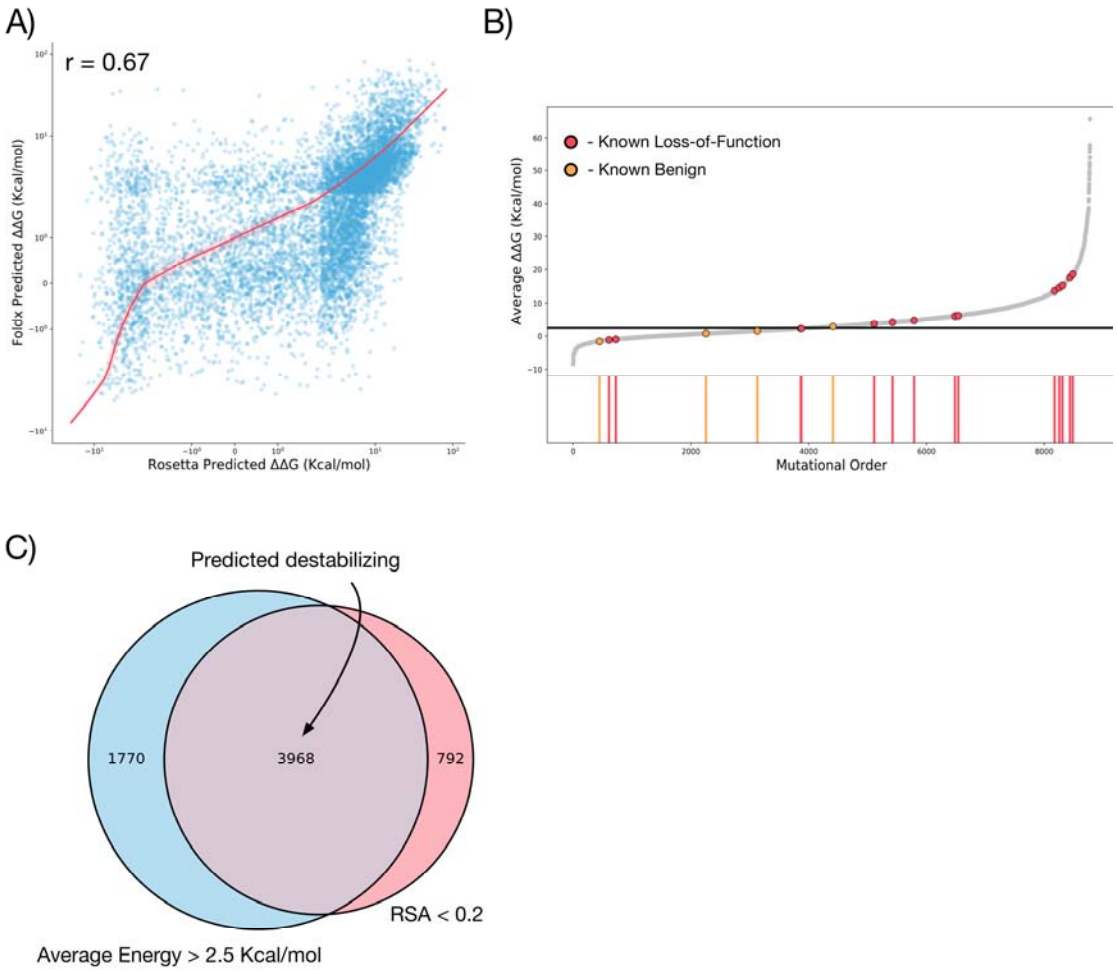


Figure 3: Prediction of Destabilizing Mutations. A) Comparison of ΔΔG calculations from FoldX and Rosetta. Correlation r is spearmans rank. B) Position of known loss-of-function (red) and benign mutations (orange) on the ΔΔG spectrum. Mutations are ordered in acsending mutational ΔΔG. Black line represents 2.5 Kcal/mol cutoff. C) Overlap between residues with a high predicted mutational energy (Defined as those with average ΔΔG > 2.5 Kcal/mol) and buried residues (RSA < 0.2). In total 3968 mutations are classified as destabilizing by taking the overlap between these two criterion.

Plotting mutational frequency for both methods, and their average for each residue (**Figure 4 A**) reveals that the most destabilizing mutations predicted by either method are in regions with a large number of buried amino acids, as expected. When plotting these mutations on the structure of the protein (**Figure 4 B**), we find the most significantly destabilizing mutations are those packed within the centre of D1, and on the interface between D1 and D2. This location suggests mutational disruption will alter the position of the D1/D2 interface, and thus will affect the binding site conformation, whereas mutations within the core D2 region are likely to influence the stability of the fully assembled tetramer.

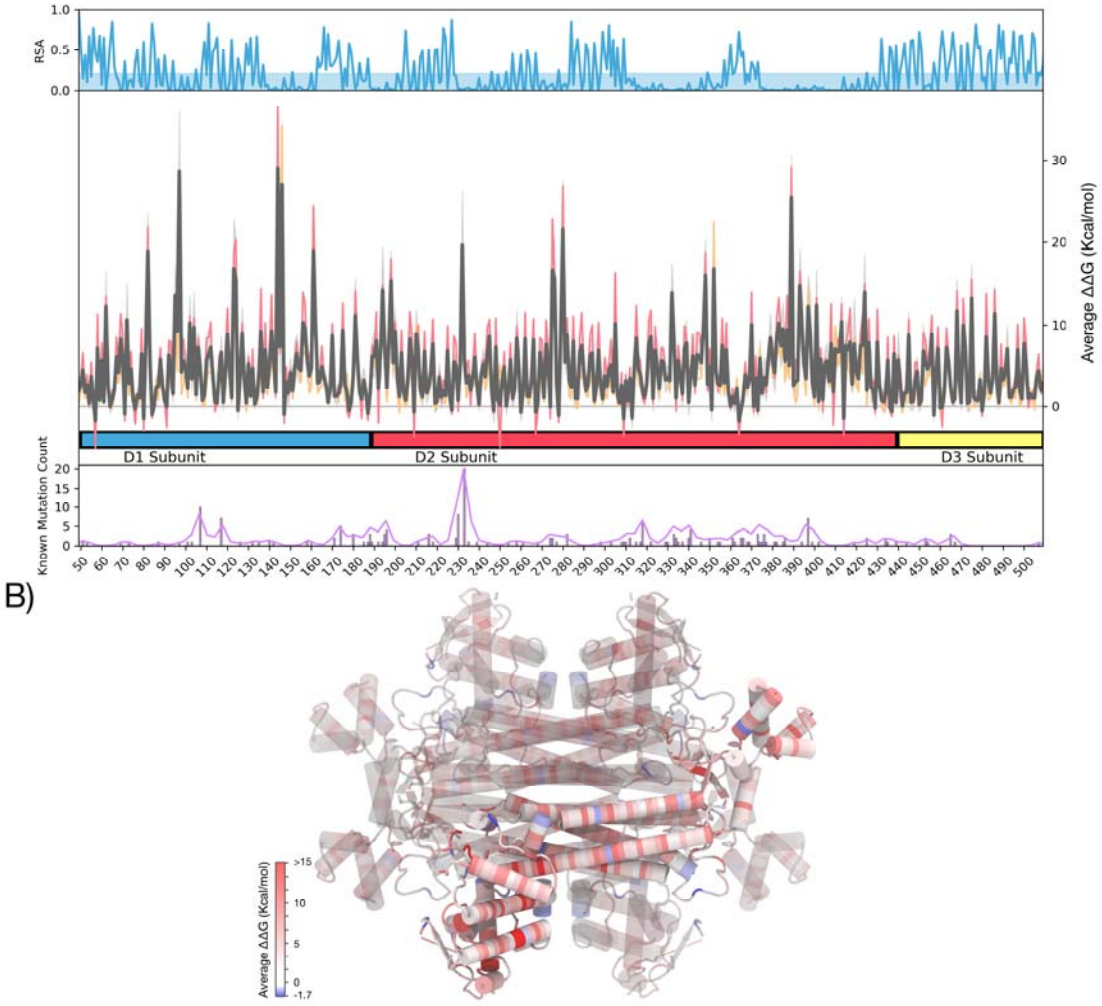


Figure 4: Predicted ΔΔG for every mutation in Fumarate Hydratase. A) Average mutational energy per residue in the FH structure. Top: Relative solvent accessible surface area (RSA) for every residue. Blue highlight indicates RSA < 0.2, classified

as buried. Middle: average ΔΔG for each residue (Grey). Red and Orange lines represent average Rosetta and FoldX calculations respectively. Bottom: Mutational frequency from the FH mutation database for each residue in FH. B) Mutational ΔΔG applied to the structure 5upp of FH. Red indicates high average ΔΔG, and so represents areas where mutations are likely to disrupt the strucutre. Blue represents regions of generally stabilizing mutations.

Existing mutations are accurately categorised based on known phenotypic effects

Overall, we define a scheme for classifying mutations into different categories of potentially disrupting, predicted LOF substitutions (**Figure 5 A**). The initial structure is relaxed using FoldX, before the binding site and hinge regions are calculated and classified, additionally mutations that are potentially destabilizing are defined based on average energy from the Rosetta and FoldX mutation methods, plus screened for buried mutations through calculating the RSA. This results in a categorisation for every mutation, where each is classified as predicted silent, binding site, hinge site, or destabilizing (including combinations of disruptive mutation types) (**Supplementary Table 1**). Overall we classify 5811 out of 8778 (66%) mutations as potentially functionally disruptive, similar to a study of mutational effects on TP53, which found that roughly 50-60% of all possible mutations were functionally disruptive[29].

To study the accuracy of our classification we chose to interrogate all mutations within the FH mutation database. We classed all mutations within the database as either loss of function (LOF), benign, or of unknown functional effect. In total 34 mutations had a known (or implied) functional effect, whilst 73 were classified as unknown (**Supplementary Table 2**). We sought to validate our functional classifications (binding site associated, hinge associated, or destabilizing) against the mutations that are known to be LOF (**Figure 5 B**). We find that 24 out of 30 (80%) mutations are correctly classified as LOF, and 3 out of 4 (75%) are correctly classed as benign. 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, and in fact, is classified as a stabilizing mutations by Rosetta, with a predicted ΔΔG of -5.2 kcal/mol, though FoldX classifies it as destabilizing. The mutation incorrectly classified as deleterious 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[30], and Rhapsody also classify this particular mutation as damaging, indicating that the benign classification for this mutation may be questionable. To further explore this mutation 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 (**Supplementary Figure 1**), and supports previous evidence from the R233H mutant, that hinges within the protein can effect binding site assembly. Of the 73 unknown mutations, we predict that 28 are functionally benign, and 45 are potential LOF mutations.

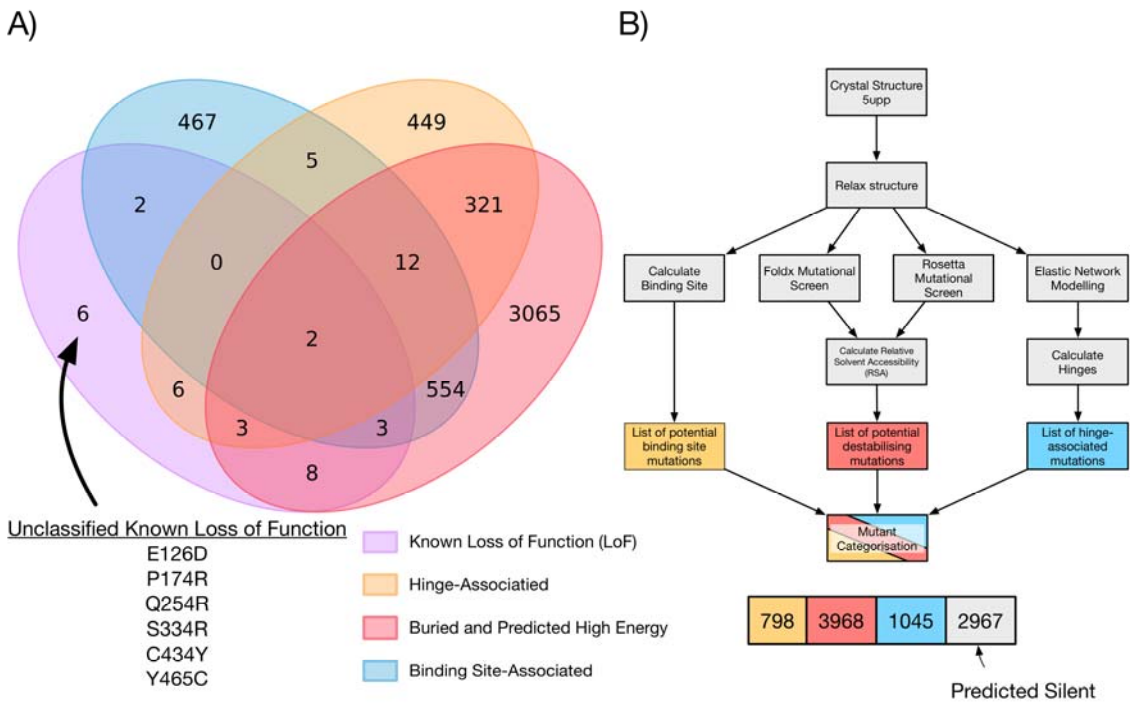


Figure 5: Prediction of known Loss-of-Function (LOF) mutations. Venn diagram showing overlap between Hinge-associated (orange), destabilizing (red) and Binding site-associated (blue) mutations. 30 known LOF mutations are included (purple) 24

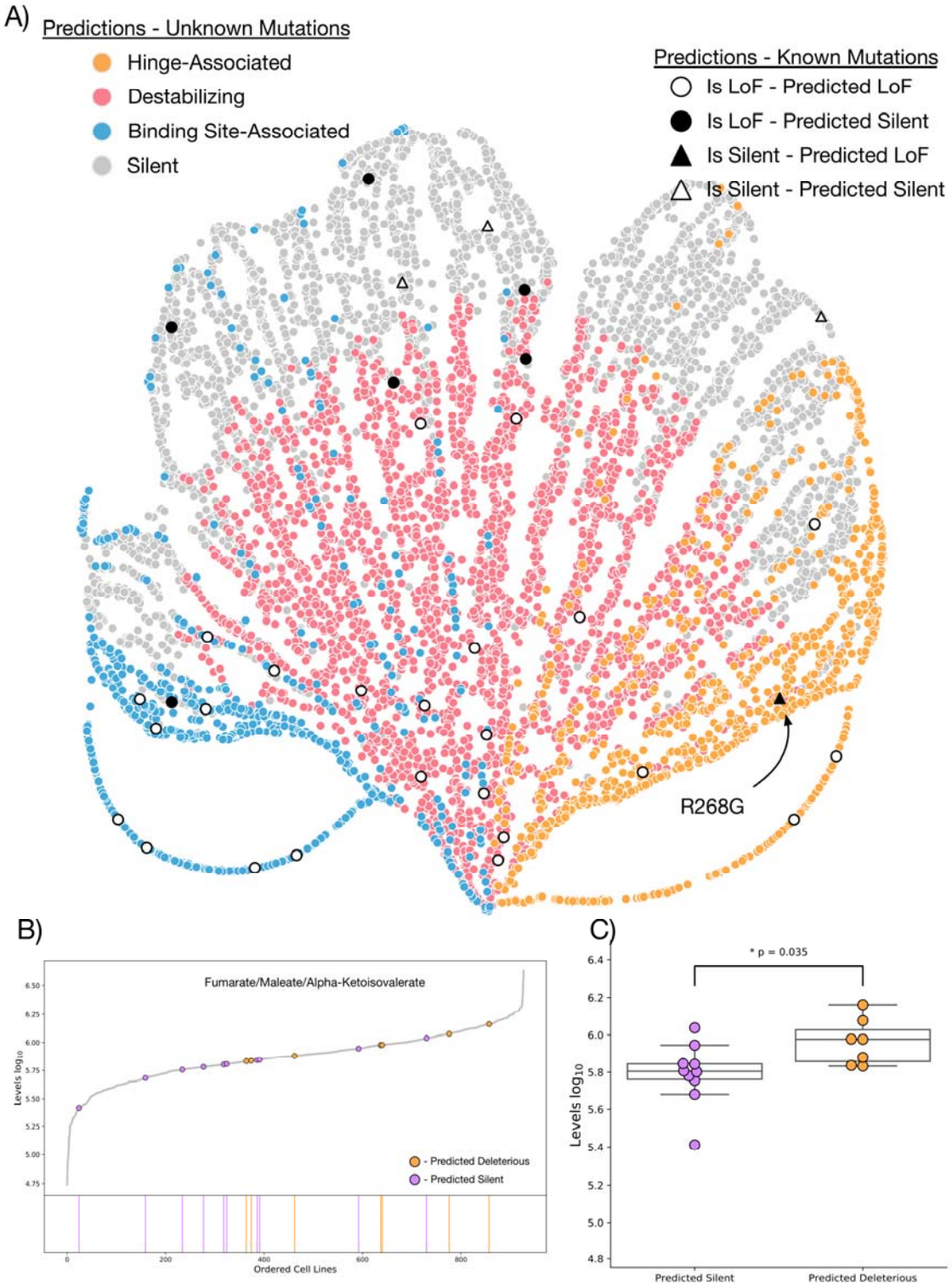
mutations are correctly categorized as LOF, whilst 6 are incorrectly categorized as benign mutations. B) Schema for categorization of mutations in FH. The structure is initially relaxed using FoldX RelaxPDB, residues within 6 Å of the binding site are calculated resulting in a list of 798 binding site associated mutations (orange). FoldX and Rosetta are used to calculate the ΔΔG for every mutation and this is subset by the relative solvent acessible surface area resulting in 3968 potentially destabilizing mutations (red). Elastic network modelling is performed to generate hinge regions of the protein, and residues within 6 Å of hinges are calculated, resulting in 1045 hinge associated mutations (blue). 2967 mutations are predicted to be silent.

Mutations with unknown properties can be accurately predicted to be functional or neutral

To study all potential mutations in FH we chose to plot all mutations using umap[31]. We ran umap on the 4 major axis involved in the classification in this study for every mutations – minimum distance to a binding site residue, minimum distance to a hinge residue, average ΔΔG of mutation, and RSA for each residue (**Figure 6 A**). 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. Also shown are the known benign and loss of function mutations. 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. Overall, we find that mutations broadly separate as expected using umap, and that unclassified mutations can be plotted on the resultant distribution –

confidence of the classification of any individual mutation can be inferred from where it fits within the landscape.

To test the predictive power of our classification we used the Cancer Cell Line Encyclopedia to look for changes in metabolite levels associated with mutations in FH[21,32]. We find 42 mutations (35 unique) in FH within 34 individual cell lines (**Supplementary Table 3**). 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 6 B,C**).



410

411 **Figure 6**: Mutational Landscape of Fumarate Hydratase. A) Umap for all mutations
412 in FH. Mutations are coloured by classification. Hinge-associated (orange),
413 Destabilizing (red), and Binding site-associated (blue) are shown clustered into
414 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 bo 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).

DISCUSSION

In conclusion, we have shown, using a comprehensive combination of techniques, that we can categorise accurately 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[33]. This is reinforced by the fact that 27 of the 461 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 when compared to similar calculations for site A.

Fumarate hydratase represents a good first-use case for 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 accurately classifies known LOF and benign mutations with a high degree of accuracy, 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 the best available methods at time of study[13]. As better methods become available it will be of interest to improve upon this work to attempt a more accurate classification.

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[34–37], 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.

AUTHOR CONTRIBUTIONS

DS and BAH conceived the study and wrote the manuscript. DS generated all data and performed all analysis. All authors were responsible for editing of the manuscript.

ACKNOWLEDGEMENTS

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

COMPETING FINANCIAL INTEREST

The authors declare no competing financial interest.

DATA AVAILABILITY

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

METHODS

FH mutation database

The FH mutation database was downloaded from the Leiden Open Variation Database[9] (https://databases.lovd.nl/shared/variants/FH/unique). 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 NMC clustering 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[25], using an alpha cutoff value of 0.05, and the Bonferroni multiple test correction method (see supplementary code).

Gaussian Network Modelling

GNM was implemented using the Prody package in python[38].

Molecular Dynamics Simulations

Molecular dynamics was performed using Gromacs version 2018.1[39]. We chose to simulate proteins using the GROMOS 54a7 forcefield[40].

The protein structure was first repaired using FoldX[10] “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 and solvated with spc water. 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, F_{max}, 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.pdb --ionStrength=0.05 --pH=7 --vdwDesign=2 --pdbHydrogens=false --positions=49
```

For positionscan on the 49th residue.

Rosetta ΔΔG Calculations

Rosetta predicted ΔΔG was calculated using the cartesian_ddg method as described in Kellogg et al:

```
$path/to/source/bin/cartesian_ddg.static.linuxgccrelease -in:file:s 5upp.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.

Umap

We used Umap[31] based on the github repository at www.github.com/lmcinnes/unmap

Cancer Cell Line Encyclopedia Data

Cancer Cell Line Encylopedia (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[21].

Data Analysis

Both MDanalysis[41] and Biopython[42] were used for analysis of structural data.

BIBLIOGRAPHY

1. Ewbank, C., Kerrigan, J. F. & Aleck, K. *Fumarate Hydratase Deficiency*. *GeneReviews*® (University of Washington, Seattle, 1993).

2. Yang, M., Soga, T., Pollard, P. J. & Adam, J. The emerging role of fumarate as an oncometabolite. *Front. Oncol.* **2**, 85 (2012).

3. 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.* **64**, 697–712 (2018).

4. Yogev, O. *et al.* Fumarase: A Mitochondrial Metabolic Enzyme and a Cytosolic/Nuclear Component of the DNA Damage Response. *PLoS Biol.* **8**, e1000328 (2010).

5. Schmidt, C., Sciacovelli, M. & Frezza, C. Fumarate hydratase in cancer: A multifaceted tumour suppressor. *Semin. Cell Dev. Biol.* (2019).

6. 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.* **142**, 1202–1215 (2018).

7. Alam, N. A., Olpin, S. & Leigh, I. M. Fumarate hydratase mutations and predisposition to cutaneous leiomyomas, uterine leiomyomas and renal cancer. *Br. J. Dermatol.* **153**, 11–17 (2005).

8. Clark, G. R. *et al.* Germline *FH* Mutations Presenting With Pheochromocytoma. *J. Clin. Endocrinol. Metab.* **99**, E2046–E2050 (2014).

9. 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.* **9**, 20 (2008).

10. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**,

4168–4169 (2019).

11. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).

12. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

13. 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.* **40**, 1414–1423 (2019).

14. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Poznan, Poland)* **19**, A68-77 (2015).

15. International Cancer Genome Consortium, T. I. C. G. *et al.* International network of cancer genome projects. *Nature* **464**, 993–8 (2010).

16. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

17. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

18. Tamborero, D., Lopez-Bigas, N. & Gonzalez-Perez, A. Oncodrive-CIS: A Method to Reveal Likely Driver Genes Based on the Impact of Their Copy Number Changes on Expression. *PLoS One* **8**, e55489 (2013).

19. Ponzoni, L., Oltvai, Z. N. & Bahar, I. Rhapsody: Pathogenicity prediction of human missense variants based on protein sequence, structure and dynamics. *bioRxiv* 737429 (2019).

20. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

21. Li, H. *et al.* The landscape of cancer cell line metabolism. *Nat. Med.* **25**, 850–860 (2019).

22. Picaud, S. *et al.* Structural basis of fumarate hydratase deficiency. *J. Inherit. Metab. Dis.* **34**, 671–6 (2011).

23. Rose, I. A. & Weaver, T. M. The role of the allosteric B site in the fumarase reaction. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3393 (2004).

24. 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.* **286**, 1925–1940 (2019).

25. 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).

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

27. Bahar, I., Atilgan, A. R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2**, 173–181 (1997).

28. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–88 (1991).

29. Kotler, E. *et al.* A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol. Cell* **71**, 178-190.e8 (2018).

30. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

31. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

32. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

33. Alam, N. A. *et al.* Missense Mutations in Fumarate Hydratase in Multiple

674 Cutaneous and Uterine Leiomyomatosis and Renal Cell Cancer. *J. Mol.*
675 *Diagnostics* **7**, 437–443 (2005).

676 34. Remington, S., Wiegand, G. & Huber, R. Crystallographic refinement and
677 atomic models of two different forms of citrate synthase at 2·7 and 1·7 Å
678 resolution. *J. Mol. Biol.* **158**, 111–152 (1982).

679 35. Lauble, H., Kennedy, M. C., Beinert, H. & Stout, C. D. Crystal structures of
680 aconitase with isocitrate and nitroisocitrate bound. *Biochemistry* **31**, 2735–
681 2748 (1992).

682 36. Chapman, A. D. ., Cortés, A., Dafforn, T. ., Clarke, A. . & Brady, R. . Structural
683 basis of substrate specificity in malate dehydrogenases: crystal structure of a
684 ternary complex of porcine cytoplasmic malate dehydrogenase, α-
685 Ketomalonate and TetrahydoNAD 1 1Edited by R. Huber. *J. Mol. Biol.* **285**,
686 703–712 (1999).

687 37. Yankovskaya, V. *et al.* Architecture of Succinate Dehydrogenase and Reactive
688 Oxygen Species Generation. *Science (80-.).* **299**, 700–704 (2003).

689 38. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein Dynamics Inferred from
690 Theory and Experiments. *Bioinformatics* **27**, 1575–1577 (2011).

691 39. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations
692 through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**–
693 **2**, 19–25 (2015).

694 40. Schmid, N. *et al.* Definition and testing of the GROMOS force-field versions
695 54A7 and 54B7. *Eur. Biophys. J.* **40**, 843–856 (2011).

696 41. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis:
697 A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*
698 **32**, 2319–2327 (2011).

699 42. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational
700 molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

701