Rapid, qualitative prediction of antimicrobial resistance by alchemical free energy methods

Philip W Fowler* [1,2]

[1]Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Headley Way, Oxford, OX3 9DU, UK

[2]National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK

Abstract

The emergence of antimicrobial resistance (AMR) threatens modern medicine and necessitates more personalised treatment of bacterial infections. Sequencing the whole genome of the pathogen(s) in a clinical sample offers one way to improve clinical microbiology diagnostic services, and has already been adopted for tuberculosis in some countries. A key weakness of a genetics clinical microbiology is it cannot return a result for rare or novel genetic variants and therefore predictive methods are required. Non-synonymous mutations in the *S. aureus dfrB* gene can be successfully classified as either conferring resistance (or not) by calculating their effect on the binding free energy of the antibiotic, trimethoprim. The underlying approach, alchemical free energy methods, requires large amounts of molecular dynamics simulations to be run.

We show that a large number (N=15) of binding free energies calculated from a series of very short (50 ps) molecular dynamics simulations are able to satisfactorily classify all seven mutations in our clinically-derived testset. A result for a single mutation could therefore be returned in less than an hour, thereby demonstrating that this or similar methods are now sufficiently fast (and reproducible) for clinical use, which is a necessary pre-condition for starting the certification process.

*To whom correspondence should be addressed: philip.fowler@ndm.ox.ac.uk, @philipwfowler

Introduction

Much of modern medicine relies on being able to prevent and treat bacterial infections. The effectiveness of antibiotics is diminishing since resistance is evolving faster than the rate at which new antibiotics are being developed and brought to market. This rise of antibiotic resistance (AMR) is now accepted as posing a threat to modern medicine requiring urgent and concerted action [1–3]. Clearly activity is required on all fronts, including improving infection control and encouraging the development of new antibiotics. An important part of any solution will be helping clinicians make appropriate treatment decisions by improving the coverage, portability, speed, accuracy and cost of species identification and antibiotic susceptibility testing. A particularly promising approach is to sequence the genome of any infecting pathogen(s) found in a clinical sample and, by looking up genetic variants found in genes known to confer resistance to the action of antibiotics, return a prediction of the effectiveness, or otherwise, of a panel of antibiotics to the clinician [4–8].

Genetic clinical microbiology has been shown to be cheaper, faster and probably more accurate than traditional culture-based clinical microbiology for the drug susceptibility testing of tuberculosis [9] and, in addition, facilitates the rapid identification of epidemiological clusters, allowing outbreaks to be rapidly identified. Public Health England adopted whole-genome sequencing for species identification and antibiotic susceptibility testing of tuberculosis in 2017 [3, 10] and other pathogens are likely to follow suit. Although catalogues relating genetic variants to phenotype have been carefully and extensively developed, they all share a common weakness: such an approach is fundamentally inferential and so cannot make a prediction when it encounters a genetic variant not present in the catalogue, such as is the case for rare genetic mutations. *Predictive* methods are therefore needed to give the clinician some information about the likely effectiveness of a drug in treating an infection whilst, at the very least, they wait for the clinical sample to be cultured and tested [11].

Trimethoprim (TMP) is a competitive inhibitor of *S. aureus* dihydrofolate reductase (DHFR, Fig. 1a), an enzyme in the essential folic acid pathway encoded by the chromosomal gene *dfrB*. It is usually administered in combination with sulfamethoxazole, which also inhibits the bacterial folic acid pathway, and is used to treat urinary tract and soft tissue infections. Predicting resistance to trimethoprim is a good test of a novel method since there exists a large amount of structural, biophysical and clinical data and the most common mutation that confers resistance to trimethoprim is F99Y (Fig. 1b),

which is a comparatively small mutation and therefore is a stringent test of any predictive method. Since DHFR is essential, our hypothesis is that non-synonymous protein mutations can confer resistance by reducing how well the antibiotic, but not the natural substrate (dihydrofolic acid, DHA), binds. This reduces the problem to calculating how the binding free energy of the drug ($\Delta\Delta G_{tmp}$) changes upon introducing the protein mutation. If the mutation reduces the binding free energy below a certain threshold, then one can predict the mutation confers resistance. It has been shown that alchemical free energy methods, a simulation method derived from classical statistical mechanics, can be successfully employed to predict the effect of individual amino acid mutations on the action of trimethoprim [12]. By applying some simple kinetic theory to clinically observed minimum inhibitory concentrations of trimethoprim for resistant and susceptible samples that work was also able to establish that for a mutation to confer resistance, $\Delta\Delta G_{tmp} \geq 0.8$ kcal/mol.

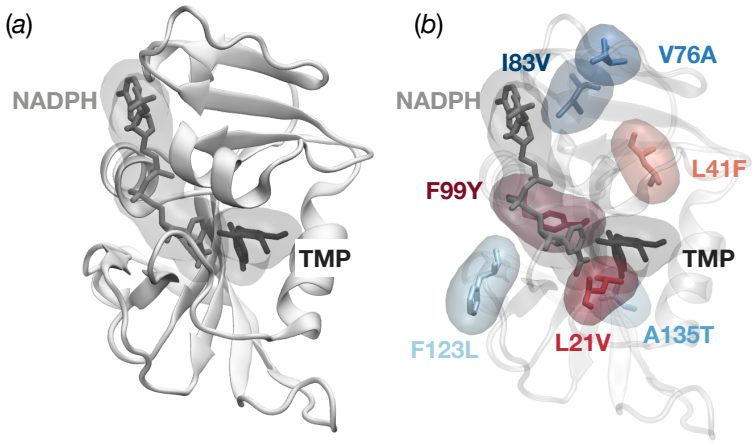

Figure 1: The structure of *S. aureus* DHFR [14] showing (*a*) the overall topology and the trimethoprim (TMP) binding site and (*b*) the location of the seven mutations studied. The three mutations that confer resistance are coloured in different shades of red, whilst the four mutations that have no clinical effect on the action of trimethoprim are coloured in different shades of blue.

For such a method to be deployed clinically it must be both fast and consume as little computational resource as possible. Whilst broadly successful, the previous study required 32,344 molecular dynamics simulations to be run, yielding a total of 8.1 μs. At the time of writing, one can simulate about 10 ns per day of DHFR using 4 computer cores slaved to a single consumer-grade graphics processing unit (GPU). The calculations underlying a single prediction therefore would require 9,720 CPU hours and 2,430

GPU hours which, although feasible, is still too large for routine use.

Any method must also meet the thresholds for accuracy and reproducibility as laid out by the existing international standards for new drug susceptibility testing methods [13]. The relevant criteria are the very major discrepancy (VMD) and major discrepancy (MD) rates. The former is defined as the number of samples that classified as susceptible by the method under test which the reference method determined as being resistant as a proportion of the total number of resistant samples and, to pass, VMD $\leq$ 3%. The definition of the major discrepancy rate is similar but inverted, i.e. the number of samples incorrectly interpreted as resistant that are susceptible. Again, to pass, MD $\leq$ 3%.

In this paper we shall examine how varying the computational resource allocated to the calculations affects the qualitative prediction and its reproducibility and thereby answer the question: *"just how quickly can we reliably predict the effect of a mutation in* dfrB *on the action of trimethoprim?"*. The answer to this question will guide whether it is yet feasible to deploy this kind of approach clinically.

Results

Datasets

A previous study calculated 32 independent values of $\Delta\Delta G_{tmp}$ and $\Delta\Delta G_{fol}$ for each of seven mutations [12]. Three of the mutations (F99Y, F99Y/L21V, L41F) are known to confer resistance (Fig. 1b) whilst the remaining four (F123L, A135T, V76A, I83V) have no clinical effect on the action of TMP. Each pair of free energies ($\Delta\Delta G_{tmp}$ & $\Delta\Delta G_{fol}$) required 13 separate alchemical free energies to be calculated, each of which in turn used between 8 and 16 molecular dynamics (MD) simulations at different values of the progress parameter, λ. By the standards of the field, all of the MD simulations were short, at just 250 ps in duration. We call this collection of simulations Set2 (Table 1). So that we may assess the impact of simulation duration on the accuracy and precision of the free energy calculations, and thence the sensitivity and specificity of predicting antimicrobial resistance, we extended the simulations underlying ten of the 32 free energy pairs by an order of magnitude, i.e. to 2.5 ns (Set1, Table 1). We shall not consider any further the effect of the mutations on the binding free energy of the natural substrate (DHA).

Name	Number of values of $\Delta\Delta G_{tmp}$ per mutation	Simulation duration (ps)
Set1	10	2,500
Set2	32	250

Table 1: We took a published set of thirty two values of $\Delta\Delta G_{tmp}$ per mutation (Set2) and extended the simulations underlying ten of these values by an order of magnitude creating a second set (Set1). Note therefore that there is overlap between the two sets since the first 250 ps of all the simulations in Set1 appear also in Set2.

Varying the duration of the simulations affects the estimated precision of $\Delta\Delta G_{tmp}$ for F99Y.

First let us consider how extending the duration of all the molecular dynamics simulations necessary for the alchemical free energy calculations affects the accuracy and precision of how the F99Y mutant affects the binding free energy of trimethoprim. The mean value of $\Delta\Delta G_{tmp}$ and its associated error (95% confidence) was calculated as a function of the simulation duration (t) for the F99Y mutation in both Set1 and Set2 using bootstrapping ($n = 100$). Comparing $\Delta\Delta G_{tmp}(t)$ to published thermodynamic data for this mutant measured using isothermal titration calorimetry (ITC) [14–18] shows how although increasing the duration of the simulations beyond 250 ps brings the predicted value into agreement with the experimental data (Fig. 2a), this is mainly through a gradual but sustained increase in the estimated uncertainty, which is clearly unsatisfactory. The lack of precision is such that once more than $\sim$1 ns of data are included the qualitative prediction would also not be definitive and on occasion will be incorrect. In contrast, as the duration of the simulation shrinks (< 50 ps) the magnitude of the estimated error increases, until, again, the lack of precision is such that the calculated value agrees with experiment and the method no longer reliably predicts this mutation to confer resistance.

Increasing the duration of the simulations leads to an apparent decrease in precision

There is no experimental binding free energy data for the other six mutations, so let us now investigate how varying the simulation duration affects the estimated error of $\Delta\Delta G_{tmp}$ along with the sensitivity and specificity of the resistant/susceptible prediction and the very major discrepancy (VMD) and major discrepancy (MD) rates. Using Set1 we calculated how the free energy of trimethoprim binding ($N = 10$) is affected by each of the seven protein mutations (Fig. 3a & b). As the simulation duration is increased, the estimated errors also increase for all mutations, as was seen for F99Y (Fig. 2a).

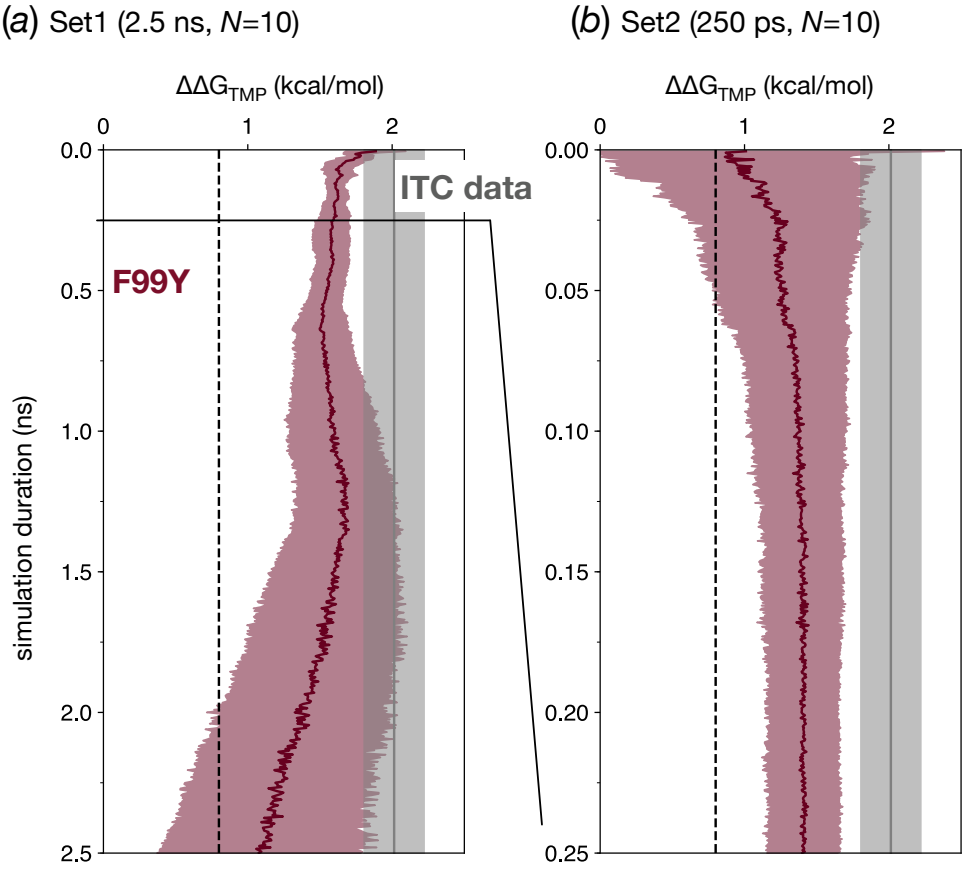

Figure 2: The estimated error for the $\Delta\Delta G_{tmp}$ for the F99Y mutation initially falls as the simulation duration increases, but then rises again. (*a*) The calculated value of $\Delta\Delta G_{tmp}$ for F99Y for $N = 10$ using Set1 becomes less precise as the simulations are extended, resulting in agreement with the published isothermal titration calorimetry (ITC) data [14–18]. Ultimately the lack of precision leads to the method no longer reliably predicting this mutation to confer resistance. (*b*) To examine the behaviour as the simulation duration is shortened, the same analysis was repeated but this time using Set2. The precision rapidly increases in the first 50-100 ps, after which the method reliably and correctly predicts this mutation to confer resistance.

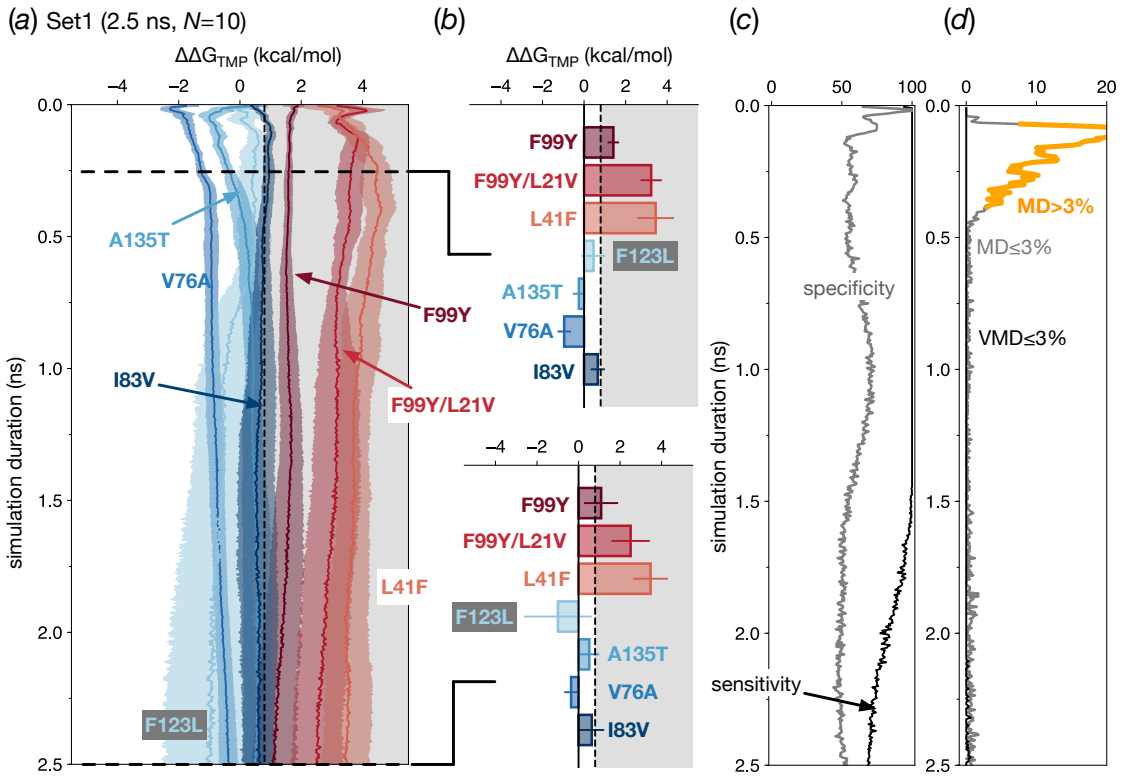

Figure 3: Keeping the number of calculations constant ($N = 10$) and increasing the simulation duration by an order of magnitude leads to an apparent decrease in precision. (*a*) How the calculated binding free energy of trimethoprim ($\Delta\Delta G_{tmp}$) varies by mutation as the simulation duration is increased. At each duration, ten values of $\Delta\Delta G_{tmp}$ for the mutation are drawn from the dataset with replacement and their mean and standard deviation are calculated. This is repeated 100 times to provide an estimate of mean and standard error. The latter is at 95% confidence and was calculated using the appropriate t-statistic. The region where $\Delta\Delta G_{tmp} > 0.8$ kcal/mol, and therefore resistance can be inferred, is shaded grey. The mutations are coloured using the same scheme as in Fig. 1. (*b*) Bar charts for each of the seven mutations at $t = 250$ ps and $t = 2500$ps. The former corresponds to the duration used in a previous study [12]. The longer duration leads to larger estimated errors. (*c*) Converting the values of $\Delta\Delta G_{tmp}$ into a ternary Resistant/Uncertain/Susceptible classification (Fig. S1) allows us to calculate the sensitivity and specificity of the method. (*d*) The very major discrepancy (VMD) rate remains below the required 3% threshold throughout whilst the major discrepancy (MD) rate reaches a maximum of 24% before falling below the threshold after $t = 0.38ns$

Each bootstrapped value of $\Delta\Delta G_{tmp}$ ($n = 100$) is then converted into a ternary prediction. If calculated value of $\Delta\Delta G_{tmp}$ for the mutation in question is indistinguishable from the threshold of 0.8 kcal/mol [12], then it is classified as Uncertain (U), otherwise it is classified as Resistant (R) or Susceptible (S) depending whether it lies entirely above or below the threshold. This is repeated for all mutations as a function of the simulation duration (Fig. S1). This analysis makes clear what one can observe in Fig. 3: since the values of $\Delta\Delta G_{tmp}(t)$ for the three resistant mutations are all ≥ 0.8 kcal/mol, except at larger simulation durations, they are uniformly predicted to be Resistant, except in the case of F99Y which has an increasing probability of being predicted Uncertain once $t > 1.5$ ns. Since the values of three of the four susceptible mutations lie close to the threshold for at least some values of t, the picture here is more complex. V76A is always predicted to be Susceptible, whilst F123L and A135T are predicted to be Susceptible or Uncertain with varying proportions. I83V has a high initial probability of being incorrectly classified as Resistant (Fig. S1) and is then mostly predicted to have an Uncertain phenotype (with a small chance of being correctly predicted susceptible) when $t > 0.5$ ns.

The resulting sensitivity is high as one would expect from the performance of the three resistant mutations (mean 93.4 %, maximum 100 %, minimum 68.3 %), whilst the specificity is mostly between 50-60 % (mean 57.6 %, maximum 100 %, minimum 44.0 %). Formally since the definitions of sensitivity and specificity assume a binary not a ternary phenotype the implication of a low specificity is that the method is incorrectly classifying Susceptible samples as Resistant, which is not the case since they are, for the most part, being classified as Uncertain. We are therefore perhaps being unduly conservative by including the cases where an Uncertain phenotype has been predicted.

What is curious is that whilst the values of $\Delta\Delta G_{tmp}$ do not yet converged and the precision is likely underestimated, they appear sufficiently accurate to allow a reasonable prediction of whether a mutation confers resistance to trimethoprim or not across a wide range of simulation durations.

Reducing the simulation duration increases error

Let us, therefore, use Set2 to examine how the method performs when *very* short durations are used to calculate $\Delta\Delta G_{tmp}$. Since Set2 has a larger number of simulations we shall now calculate the mean and uncertainty in $\Delta\Delta G_{tmp}$ for each mutation using $N = 30$, again with bootstrapping. Now the opposite trend is observed: initially the estimated errors are large but as the simulation duration is increased, their magnitudes decrease.

We have therefore recapitulated for all mutations the trends observed for F99Y (Fig. 2).

Bearing in mind that we are now drawing more values of $\Delta\Delta G_{tmp}$ from a larger dataset, we observe similar trends: the values of $\Delta\Delta G_{tmp}$ for all three resistant mutations tend to lie above the resistance threshold, however we are now probing *very* short simulation durations and can see (Fig. S2) how there is an appreciable probability that these mutations are classified as Uncertain, rather than Resistant, when $t < 50$ ps, leading to an initial rise in the sensitivity (Fig. 4c, mean 96.3 %, maximum 100 %, minimum 45 %) whilst the VMD rate remains below 3% throughout (Fig. 4d). This time, the values of $\Delta\Delta G_{tmp}$ for A135T and V76A remain firmly below the resistance threshold, resulting these mutations being consistently classified as Susceptible, with the exception of $t < 20$ps for A135T (Fig. S2). The values of $\Delta\Delta G_{tmp}$ for F123L and I83V, are consistently predicted to be close to the resistance threshold with the result that they are predicted to be Uncertain or Susceptible in varying proportions, except at very short durations ($t < 20$ ps) for F123L where there is a small probability that it is predicted Resistant (Fig. S3). The specificity is again lower, as expected (mean 71.8 %, maximum 83.4 %, minimum 39.8 %) and, apart from an initial peak in the major discrepancy rate at small values of t, the MD rate also remains below the required 3% threshold.

Using a fixed amount of computational resource

Keeping the number of values of $\Delta\Delta G_{tmp}$ contributing to the average constant ensures that the amount of computer resource *increases* as the length of the simulations increase. Perhaps a more helpful question to answer is, if one has a fixed amount of computational resource, should one calculate a large number of values of $\Delta\Delta G_{tmp}$ using very short simulations, or should instead one calculate relatively few values of $\Delta\Delta G_{tmp}$ using longer simulations?

First, let us investigate how one might best use the amount of resource required to calculate three independent values of $\Delta\Delta G_{tmp}$ using simulations of duration 2500 ps (Fig. 5). The estimated errors when the mean is calculated using 3 values of $\Delta\Delta G_{tmp}$ derived from simulations 2.5 ns long are much larger than those than when e.g. the mean is calculated from 30 values of $\Delta\Delta G_{tmp}$ derived from simulations 0.25 ns long (Fig. 5b). There is therefore a much greater probability that a value of $\Delta\Delta G_{tmp}$ for any given mutation is predicted as having an Uncertain phenotype, or even an incorrect phenotype (Fig. S3). Only V76A is consistently predicted to have the correct phenotype, with the all other six mutations having an increasing chance of being predicted Uncertain as t is

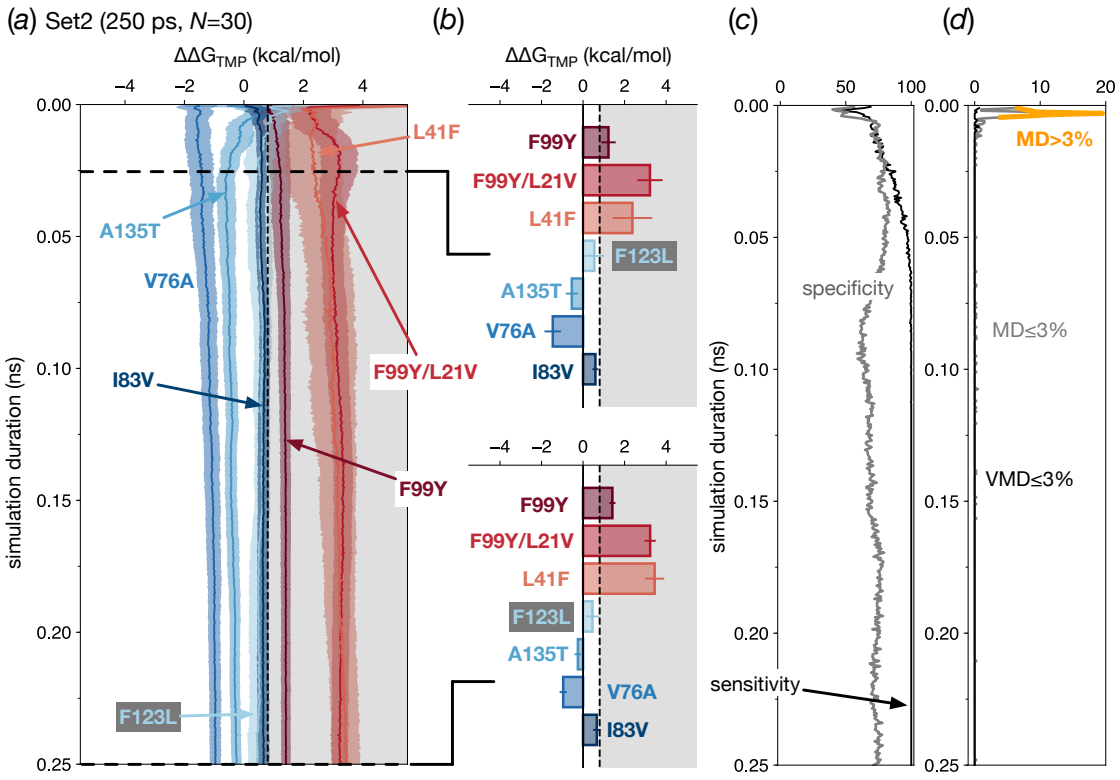

Figure 4: Keeping the number of calculations constant ($N = 30$) and decreasing the simulation duration by an order of magnitude also leads to an apparent decrease in precision. (*a*) How the calculated binding free energy of trimethoprim ($\Delta\Delta G_{tmp}$) varies by mutation as the simulation duration is increased. The process followed is the same as in Fig. 3 except that thirty values of $\Delta\Delta G_{tmp}$ are drawn from Set2, rather than ten from Set1. (*b*) Bar charts for each of the seven mutations at $t = 25$ ps and $t = 250$ps. The latter corresponds to the duration used in a previous study [12]. (*c*) Converting the values of $\Delta\Delta G_{tmp}$ into a ternary Resistant/Uncertain/Susceptible classification (Fig. S2) allows us to calculate the sensitivity and specificity of the method. (*d*) For this set, the very major discrepancy (VMD) rate remains below the required 3% threshold throughout whilst the major discrepancy (MD) rate reaches a maximum of 20% before falling below the threshold after t = 0.005ns

increased (and therefore N decreases) with F123L, A135T and I83V also having a small probability of being predicted Resistant. As was observed before, I83V also has a high probability of being predicted Resistant at low values of t, even given the corresponding high values of N. The values of the sensitivity and specificity are correspondingly more modest, as one would expect given ≤ 30 % of the computer resource is being used compared to Fig. 3. The very major discrepancy and major discrepancy rates remain below and above, respectively, the 3 % threshold throughout.

Since the magnitude of the estimated error increases with simulation duration, leading to a worse classification performance, we conclude that the additional statistical power introduced by having more values of $\Delta\Delta G_{tmp}$ contributing to the mean is dominating the improved exploration of phase space made possible by longer individual trajectories. We note, however, that since Set1 only contains ten independent values of $\Delta\Delta G_{tmp}$ for each mutation, the bootstrapping process will be creating datasets with large numbers of repeated values as the simulation duration is reduced and the number of values increases which may be introducing artefacts into our analysis particularly in the region $t < 0.8$ ns.

Let us therefore switch to using Set2, since this has 32 independent values of $\Delta\Delta G_{tmp}$ for each mutation available so that we may investigate just how short the simulations can be (Fig. 6). Apart from probably spurious behaviour at very short times, the magnitude of estimated error remains more constant as t increases (and N decreases, Fig. 6a), suggesting that the statistical and phase space effects are more balanced in this regime. Only one of the three resistant mutations (F99Y/L21V) is classified as Resistant throughout, whilst the other two have a roughly constant probability of being predicted Uncertain. Two (V76A & A135T) of the susceptible mutations are consistently classified as Susceptible, with both having a small chance of being classified Uncertain at large values of t. The other two are most likely to be predicted as having an Uncertain phenotype, with both having a very small chance of being incorrectly predicted as Resistant and F123L having a high probability of being incorrectly classified as Resistant if $t < 20$ ps. Aggregating these effects leads to approximately constant values of the sensitivity (mean 80.2 %) and specificity (mean 58.8 %), although we emphasise that by including the cases predicted Uncertain in the denominator of the sensitivity and specificity calculations, we are probably being over-conservative and excluding these cases would result in much higher values for the sensitivity and specificity.

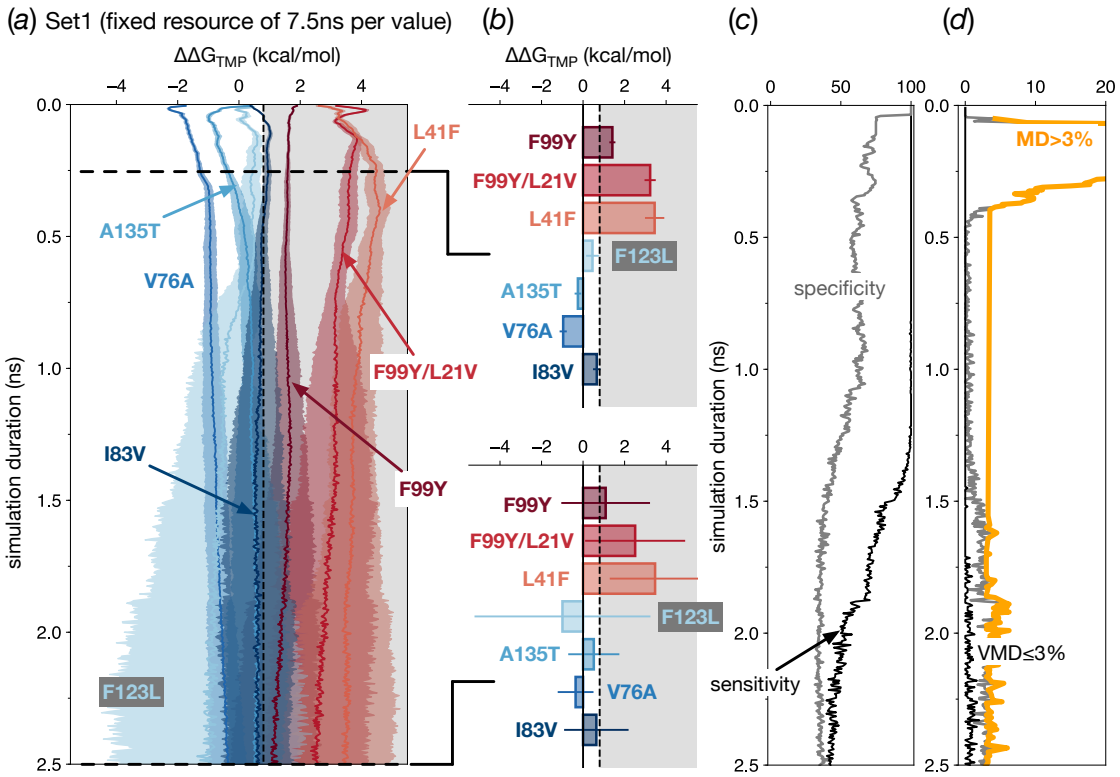

Figure 5: Given a computational resource equivalent to the amount necessary to calculate 3 values of $\Delta\Delta G_{tmp}$ using simulations 2.5 ns long, then running more, shorter simulations results in a better classification method. (*a*) How the calculated binding free energy of trimethoprim ($\Delta\Delta G_{tmp}$) varies by mutation as the simulation duration is increased (and the corresponding number of values of $\Delta\Delta G_{tmp}$ drawn from Set1 decreases). (*b*) Bar charts for each of the seven mutations at $t = 250$ ps (n=30) and $t = 2500$ps (n=3). (*c*) Converting the values of $\Delta\Delta G_{tmp}$ into a ternary Resistant/Uncertain/Susceptible classification (Fig. S3) allows us to calculate the sensitivity and specificity of the method. (*d*) For this set, the very major discrepancy (VMD) rate remains below the required 3% threshold throughout whilst the major discrepancy (MD) remains above the 3% threshold throughout, and reaches a maximum of 25%.

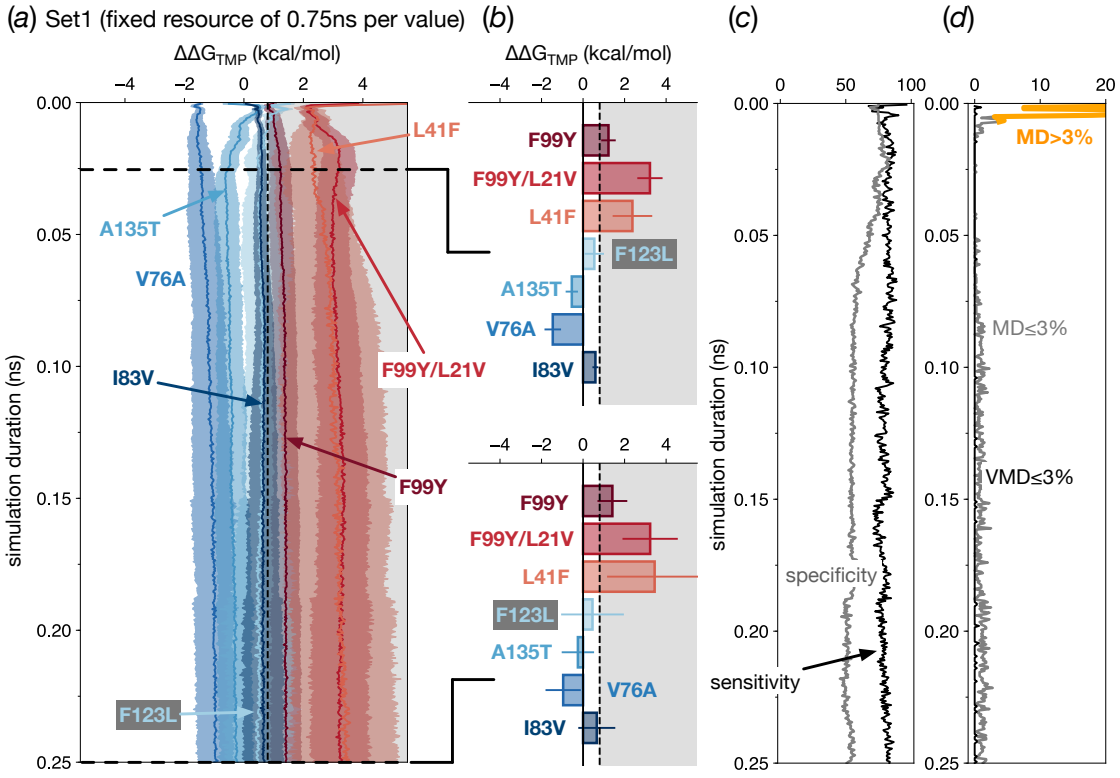

Figure 6: Given a computational resource equivalent to the amount necessary to calculate 3 values of $\Delta\Delta G_{tmp}$ using simulations 0.25 ns long, then running more, shorter simulations results in a better classification method. This dataset uses a tenth of the computational resource applied to Fig. 5. (*a*) How the calculated binding free energy of trimethoprim ($\Delta\Delta G_{tmp}$) varies by mutation as the simulation duration is increased (and the corresponding number of values of $\Delta\Delta G_{tmp}$ drawn from Set1 decreases). (*b*) Bar charts for each of the seven mutations at $t = 250$ ps (n=30) and $t = 2500$ps (n=3). (*c*) Converting the values of $\Delta\Delta G_{tmp}$ into a ternary Resistant/Uncertain/Susceptible classification (Fig. S3) allows us to calculate the sensitivity and specificity of the method. (*d*) For this set, the very major discrepancy (VMD) rate remains below the required 3% threshold throughout whilst the major discrepancy (MD) remains above the 3% threshold throughout, and reaches a maximum of 25%.

How fast could we predict whether a mutation confers resistance to trimethoprim?

Given the detailed analysis in Fig. 3-6 we are able to optimise our choice of (N,t), however one does not *a priori* usually have this amount of information and, in any case, it is possible that as yet unseen mutations in *dfrB* could have effects not congruent with our test set of seven mutations. For example, our test set does not contain any mutation that is marginally resistant and hence the magnitude of the sensitivity is always greater than the specificity, which may be true in general. This analysis therefore should be taken as indicative. That said, it is clearly interesting to consider just how rapidly one could predict whether a mutation confers resistance or not.

Let us choose $N = 15$ and $t = 50$ ps (Fig. 6) since the trajectories are long enough to avoid the observed transients in classification behaviour and thence major discrepancy rate but are still short enough to run quickly using a consumer grade GPU. We estimate that a single 50 ps trajectory will take ~ 7.5 minutes to complete, assuming GPU is large enough (or there is more than one on the motherboard) so that all the replicas can run on the same node so they can undergo Hamiltonian replica exchange. We have restricted ourselves in this paper to only calculating $\Delta\Delta G_{tmp}$ and therefore each value requires 8 alchemical free energies to be calculated, each in turn requiring 8-16 coupled molecular dynamics simulations, making a total of 960-1,920 simulations, which is still daunting. If we arbitrarily decide that a prediction must be complete within one hour, then we would require twenty nodes, each with one or more GPUs and 8-16 CPU cores, which is moderately large and could easily be provided by a commercial cloud platform. That would leave 15 minutes for setup and analysis. The paradigm of preparing all simulation input files on a single machine would not be sufficiently fast, and hence one would have to distribute all the setup tasks using the same high performance computer that the simulations would run on. We conclude that, whilst there are obvious challenges, it is now feasible to predict whether individual mutations confer resistance to an antibiotic using free energy methods and that this can be done fast enough to be clinically useful.

Prevalence of the studied mutations in the European Nucleotide Archive

The seven different non-synonymous *dfrB* mutations studied here were selected from a relatively small dataset of 501 unrelated *S. aureus* isolates collected from patients in the UK [12, 19]. Our analyses and conclusion depend on our testset of seven mutations being representative of the likely mutations one might encounter clinically in *dfrB*. To estimate how prevalent these mutations are globally, we searched an index of the European Nucleotide Archive [20]. Due to how the index is created, only results for amino

acids 11-148 incl. were returned, and samples containing multiple amino acid mutations fewer than ten positions apart are unlikely to have been detected. That said, this is as comprehensive a scan of all deposited *S. aureus* genomes as is currently possible and approximately 19,200 *S. aureus* genomes were searched (Table S1). The three *dfrB* mutations implicated in samples that were resistant to trimethoprim were detected. Although F99Y was found in 13.8 % (69) of the clinical isolates, the prevalence in the ENA was only 0.7 % (137), suggesting the clinical dataset was substantially enriched for trimethoprim resistance. Both L41F and L21V F99Y were only detected once in the original clinical dataset (0.2 %). Whilst the double L21V F99Y was not present in the much larger ENA dataset, 421 genomes (2.3 %) containing the L21V mutation without F99Y were detected, however the effect of L21V on its own on the effectiveness of trimethoprim is unknown. The L41F mutation was found in the ENA, but at a very low prevalence (0.02 %, n=3). Of the four mutations, A135T (n=162, 32.3 %), V76A (89, 17.8 %), I83V (8, 1.6 %) and F123L (5, 1.0 %), identified as susceptible in the clinical study [19], only two were found in the ENA: A135T (n=6,578, 34.3 %) and V76A (1,412, 7.4 %). Further examination of the results returned for *k*-mers used to probe variation at Val76 and Phe123 showed that comparatively few sets of short-reads in the ENA were identified around these positions, suggesting that some samples were missed since they contained multiple amino acid mutations within the width of the *k*-mer (21 amino acids) or that similar *k*-mers are found in other species and therefore our ability to detect variation at these sites using this method is probably limited.

Discussion

We conclude that it is now possible to rapidly and reproducibly predict whether individual non-synonymous mutations confer resistance (or not) to trimethoprim, an antibiotic, and we have shown that, for this system at least, it is theoretically possible to make a prediction in less than one hour. This relies on our observation that calculating a large number of values of $\Delta\Delta G_{tmp}$ using very short alchemical molecular dynamics simulations allows the seven mutations in our test set to be adequately classified with reasonable sensitivities and specificities and low very major and major discrepancy rates, thereby using an order of magnitude less computational resource than in a previous study [12].

The pattern we observe of the magnitude of the estimated error reducing as the simulation duration increases before reaching a minimum and then increasing again was commented on over fifteen years ago [21]. We infer that the initial high variance is due to each set of alchemical molecular dynamics simulations are starting from a different structure seeded from one or more longer equilibrium trajectories and therefore starting in a different part of phase space. As the simulations progress, and explore phase space, a process sped up by the use of Hamiltonian replica exchange, they begin to converge, reducing the observed variation. Then a point is reached where simulations stochastically start to access new parts of phase space, that were perhaps not explored by the equilibrium simulation due to an energetic barrier, and the variance starts to increase again again. It is therefore likely that our simulations, and hence our values for $\Delta\Delta G_{tmp}$, are not converged. Only by running far longer simulations would it be possible to infer that convergence. Curiously, since our objective is producing a ternary classification, success is only notably dependent on quantitative accuracy (and hence presumably convergence) if the mutation under study has a value of $\Delta\Delta G_{tmp}$ that is close to the resistance threshold.

Our conclusions are reliant on the seven mutations that form the testset and we have investigated whether these are representative of the genetic variation one might expect to observe clinically. Since several of mutations were detected at either very low prevalences or not at all in the European Nucleotide Archive, we conclude that our testset is not truly representative and therefore our conclusions may not transfer into the clinic. Clearly the method needs to be tested on additional *dfrB* mutations as well as other antibiotic/protein target combinations in a wide range of pathogenic bacterial before we can make more definitive statements about its applicability.

Although we were able to show that calculating 15 values of $\Delta\Delta G_{tmp}$ from alchemical simulations only 50 ps long led to acceptable classification behaviour, this is almost certainly a form of overfitting since we had a free choice of a wide range of combinations and may have simply chosen one that works best for our testset. It will only be possible to gain confidence through applying such parameter combinations 'blind' to other mutations. We have also restricted ourselves here to only considering the effect of the mutation on the binding of the antibiotic; previous work has shown that taking the effect on the natural substrate, DHA, into account changes and may improve the prediction [12]. Also, all of the mutations studied (with the possible exception of the double mutant, F99Y/L21V) are tractable by alchemical methods: it remains to be seen how successful this approach will be for mutations involving a change in electrical charge, that involve a proline, or simply require a large number of atoms to be perturbed. We have assumed here that all the eight alchemical free energies that are required to calculate a single value of $\Delta\Delta G_{tmp}$ all converge and behave similarly, which is almost certainly not true. Further work will be needed to assess if different types of alchemical transition (e.g. removing the electrical charges from the alchemical atoms) require more or less simulation time and/or numbers of molecular dynamics simulations. It is possible up to another order of magnitude saving is available through careful dynamic control (i.e. 'steering' [22]) of the makeup and number of alchemical free energies run. This will necessarily complicate the calculation of errors which was done here at the level of $\Delta\Delta G_{tmp}$ and in future will likely have to be done at the level of each alchemical free energy with errors then added in quadrature.

The translation of genetics into clinical microbiology shows no sign of abating, with the most progress being made using whole-genome sequencing for antibiotic susceptibility testing (AST) of tuberculosis where catalogues of observed genetic variants and their associated effects on different drugs are most advanced [23]. Since this approach is purely inferential, there remains a need to develop predictive methods [11]. Even a low-quality prediction for a single antibiotic may prove useful clinically, since it will be viewed alongside the results for other drugs which often allows nonsensical predictions to be discounted. Such a Bayesian approach has already been shown to improve AST of tuberculosis [11]. In any case, if the prediction affects the clinical decision, it is likely that the sample would still be sent for culturing and testing.

Although we have focussed primarily on calculating the effect of the mutation on the binding free energy of the antibiotic, there are a range of other methods that could

be brought to bear. Machine learning methods, using genetic, structural and/or chemical features, are likely to sufficiently accurate and also fast [11, 24, 25]. Such methods could be used to screen out mutations that have no effect on an antibiotic, leaving only the marginal cases for computationally more intensive approaches such as we are proposing. One key advantage of our approach we have not discussed is that it rarely makes an incorrect classification. This potentially enables a guided process whereby simulations are run until a definite (R/S) prediction is returned, saving further computational resource and time. More work needs to be done if this, or other, methods are to be formally certified for use in AST. For example, to pass the relevant international standard [13], not only must the very major and major discrepancy rates be ≤ 3 %, but also there must be a high level of categorial agreement with a reference method. This will necessitate carefully designed and thorough studies in collaboration with clinical microbiology laboratories using standard methods.

Clinical microbiology is built upon a binary paradigm of a sample being Resistant or Susceptible. Converting a quantitative measurement which has a confidence limit into a binary result necessitates a third Uncertain category for those cases that cannot be definitely classified as R or S. This is a subtle point (and one can probably trace its origins back to the Law of the Excluded Middle) but it is nonsensical to deny predictive and experimental methods the option of returning an Uncertain result. The Clinical Laboratory and Standards Institute, which provides clinical microbiology standards mainly to the U.S.A, has adopted a ternary system, however, the European Committee on Antimicrobial Susceptibility Testing have only recently begun to introduce such a category into some antibiotic/pathogen combinations [26]. Many of the tools rely on the binary paradigm and new tools and language need developing, as we have seen here when calculating the sensitivity and specificity of our method.

As computational resource becomes faster and more widespread, the broader field of molecular simulation is gradually moving away from running single simulations towards running large number of replicas [27, 28]. This potentially exposes underlying problems with the molecular dynamics codes and how we, as computational scientists, typically work. For example, when setting up and running thousands of molecular dynamics simulations the time taken to setup a simulation becomes appreciable. Likewise, one must use some kind of object store or file hierarchy to archive all the simulation files. Without progress in these areas, it is possible that the time taken to setup, copy files, queue simulations, retrieve and analysis files could become the limiting factor in

speeding up and therefore applying these methods clinically.

In the field of alchemical free energy calculations much attention has understandably been focussed on demonstrating that free energies can be calculated that agree with experimental data to a high degree of precision. A high degree of precision and accuracy is spurious in this application and one might speculate that applying alchemical methods in this way is a sign that the field is maturing. Finally, whilst thermodynamic integration is usually described as an equilibrium method, it is not obvious if this remains true when the duration of an alchemical simulation is only 50 ps. Despite this, it is both illuminating and encouraging to look back over the last thirty years on the progress made by the field of alchemical free energy calculations [29] and infer what might be possible in just a few more years.

Methods

The GROMACS molecular dynamics [30] simulations underlying the alchemical free energy calculations were setup and run as described previously [12] and followed best practice [31], including using `pmx` [32] to mutate the wild-type structure of DHFR with trimethoprim bound [14]. The alchemical simulations underlying ten of the thirty-two values of $\Delta\Delta G_{tmp}$ for each of the seven mutations were extended by an order of magnitude (from 0.25 ns to 2.5 ns) to enable this study. To cope with the very large numbers of MD simulations, all data was stored in a file hierarchy and tagged using the datreant Python module [33]. All simulation data was then parsed and alchemical free energies were calculated as a function of simulation time t using a purpose-written Python class. In each case, only the second half of each dataset was used and then a single alchemical free energy was calculated using thermodynamic integration (via the trapezium rule). This yielded a large table of alchemical free energies for both Set1 and Set2 that was stored as a Pandas dataframe [34]. The resulting values of $\Delta\Delta G_{tmp}$ were then calculated and stored. These dataframes can be found in the Supplemental Information. Bespoke Python code then read these tables and applied the bootstrapping process described in the main body of the paper to produce the Fig. 3-6. Standard errors were calculated in the usual way and converted to a 95% confidence interval using the appropriate t-statistic for the number of values. All graphs were plotted using Matplotlib [35] and all protein images were rendered using VMD [36]. The BIGSI search index for microbial genomes [20] is interrogated using a k-mer where $k \geq 61$ bases. Since searching the index for amino acid mutations involves permuting the bases in a triplet, leading to 80 different variations, we wrote a Python module, called `pygsi`, that automatically interrogates the index using a 63-mer constructed with 20 base pairs flanking the codon of interest [37]. Once all the permutations have been tried, the code moves on to the next codon.

Data Accessibility Statement

The two tables containing all the values of $\Delta\Delta G_{tmp}$ for the different mutations used to construct the figures in this paper are provided as Supplemental Files in the CSV format and are described in the Supplemental Information. The results of searching the BIGSI bacterial genomic index are provided in Table S1.

Acknowledgements

The research was funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Center (BRC). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. We are grateful to the Science and Technology Facilities Research Council and Amazon Web Services for providing additional computer time. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. O'Neill J (2016) Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. Technical report.

2. Davies SC (2013) Annual Report of the Chief Medical Officer - Vol 2. Technical report, Department of Health, UK Government.

3. Public Health England (2018) Tuberculosis in England: 2018. Technical report.

4. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW (2012) *Nat Rev Genetics* 13:601–12.

5. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J, Peacock SJ (2012) *PLoS pathogens* 8:e1002824.

6. Ellington M, Ekelund O, Aarestrup F, Canton R, Doumith M, Giske C, Grundman H, Hasman H, Holden M, Hopkins K, Iredell J, Kahlmeter G, Köser C, MacGowan A, Mevius D, Mulvey M, Naas T, Peto T, Rolain JM, Samuelsen Ø, Woodford N (2017) *Clinical Microbiology and Infection* 23:2–22.

7. Tagini F, Greub G (2017) *Eur J Clin Micro Infect Dis* 36:2007–2020.

8. Balloux F, Brønstad Brynildsrud O, van Dorp L, Shaw LP, Chen H, Harris KA, Wang H, Eldholm V (2018) *Trends in Microbiology* 26:1035–1048.

9. Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordsworth S, Wyllie D, Xu L, Crook DW (2016) *Lancet Resp Med* 4:49–58.

10. Walker TM, Cruz ALG, Peto TE, Smith EG, Esmail H, Crook DW (2017) *Lancet Infec Disease* 17:359–361.

11. Brankin AE, Fowler PW (2019) *ACS Central Science* 5:1312–1314.

12. Fowler PW, Cole K, Gordon NC, Kearns AM, Llewelyn MJ, Peto TE, Crook DW, Walker AS (2018) *Cell Chemical Biology* 25:339–349.e4.

13. International Organization for Standardization (2007) ISO 20776-2: Clinical laboratory testing and in vitro diagnostic test systems. Technical report, International Standards Organization.

14. Oefner C, Bandera M, Haldimann A, Laue H, Schulz H, Mukhija S, Parisi S, Weiss L, Lociuro S, Dale GE (2009) *J Antimicrobial Chem* 63:687–698.

15. Pires DEV, Blundell TL, Ascher DB (2015) *Nuc Acid Res* 43:D387–D391.

16. Dale GE, Broger C, D' Arcy A, Hartman PG, DeHoogt R, Jolidon S, Kompis I, Labhardt AM, Langen H, Locher H, Page MG, Stüber D, Then RL, Wipf B, Oefner C (1997) *J Mol Biol* 266:23–30.

17. Frey KM, Georgiev I, Donald BR, Anderson AC (2010) *Proceedings of the National Academy of Sciences* 107:13707–13712.

18. Frey KM, Viswanathan K, Wright DL, Anderson AC (2012) *Antimicrob Agent Chemo* 56:3556–3562.

19. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TEA, Crook DW, Walker AS, Golubchik T (2014) *J Clin Microbiol* 52:1182–91.

20. Bradley P, den Bakker HC, Rocha EPC, McVean G, Iqbal Z (2019) *Nature Biotechnology* 37:152–159.

21. Chipot C, Pearlman D (2002) *Mol Sim* 28:1–12.

22. Fowler PW, Geroult S, Jha S, Waksman G, Coveney PV (2007) *J Chem Theory Comput* 3:1193–1202.

23. The CRyPTIC Consortium, 100000 Genomes Project (2018) *New Eng J Med* 379:1403–1415.

24. Carter JJ, Walker TM, Walker AS, Whitfield MG, Morlock GP, Peto TE, Posey JE, Crook DW, Fowler PW (2019) *bioRxiv doi:101101/518142* .

25. Aldeghi M, Gapsys V, de Groot BL (2019) *ACS Central Science* 5:1468–1474.

26. Davies TJ, Stoesser N, Sheppard AE, Abuoun M, Fowler P, Quan TP, Griffiths D, Vaughan A, Morgan M, Phan HTT, Jeffery KJ, Andersson M, Ellington MJ, Ekelund

O, Mathers AJ, Robert A, Woodford N, Crook DW, Peto TEA, Anjum MF, Sarah A (2019) *biorXiv doi:101101/511402* .

27. Knapp B, Ospina L, Deane CM (2018) *Journal of Chemical Theory and Computation* 14:6127–6138.

28. Bhati AP, Wan S, Hu Y, Sherborne B, Coveney PV (2018) *Journal of Chemical Theory and Computation* 14:2867–2880.

29. Bash P, Singh U, Brown F, Langridge R, Kollman P (1987) *Science* 235:574.

30. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) *SoftwareX* 1-2:19–25.

31. Klimovich PV, Shirts MR, Mobley DL (2015) *J Comp Aided Mol Des* 29:397–411.

32. Gapsys V, Michielssens S, Seeliger D, de Groot BL (2015) *J Comp Chem* 36:348–54.

33. Dotson DL, Seyler SL, Linke M, Gowers RJ, Beckstein O (2016) In Proc 15th Python Sci Conf, edited by S Benthall, S Rostrup, 51–56.

34. McKinney W (2010) In Proceedings of the 9th Python in Science Conference, edited by SvdW Millman, Jarrod, 51–56.

35. Hunter JD (2007) *Computing in Science & Engineering* 9:90–95.

36. Humphrey W, Dalke A, Schulten K (1996) *J Mol Graph* 14:33–38.

37. Fowler PW (2017). pygsi: a Python class to interrogate BIGISI. doi: 10.5281/zenodo.1407085.