A computational method for predicting the most likely evolutionary trajectories in the stepwise accumulation of resistance mutations

R. Charlotte Eccleston[a,1], Emilia Manko[a], Susana Campino[a], Taane G. Clark[a,b], Nicholas Furnham[a]

[a]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom; [b]Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

[1]R. Charlotte Eccleston

Email: charlotte.eccleston@lshtm.ac.uk

Author Contributions: R.C.E., S.C., T.G.C. and N.F. conceived and designed the study. E.M. performed the bioinformatic processing of the raw sequencing data. R.C.E performed the research. R.C.E. analyzed the data. R.C.E. wrote the first draft of paper. All authors commented on and edited the manuscript and approved the final version. R.C.E. compiled the final manuscript.

Competing Interest Statement: None.

Classification: Major: Biological Sciences, Minor: Biophysics and Computational Biology, Evolution

Keywords: Antimicrobial resistance, epistasis, pathogen evolution

This PDF file includes:

Main Text

Figures 1 to 4

Table 1

Abstract

Pathogen evolution of drug resistance often occurs in a stepwise manner via accumulation of multiple mutations that in combination have a non-additive impact on fitness, a phenomenon known as epistasis. Epistasis complicates the sequence-structure-function relationship and undermines our ability to predict evolution. Here we test how well Rosetta Flex ddG is able to estimate epistasis between resistance mutations and observe good agreement with experimental data. By using these predictions to parameterize an existing mechanistic model of molecular evolution, we also investigate how well such a model is able to capture evolution to resistance where the order of fixation is influenced by epistasis. We apply this method to predict evolutionary trajectories to known quadruple mutations associated with resistant phenotypes for malaria. Resistance to the combination drug sulfadoxine-pyrimethamine (SP) in malaria-causing species *Plasmodium falciparum (Pf)* and *Plasmodium vivax (Pv)* has arisen via the accumulation of multiple point mutations in the DHFR and DHPS genes. The evolution of resistance in the DHFR genes of *Pf* and *Pv* has been studied extensively. Four known *Pf*DHFR pyrimethamine resistance mutations are highly prevalent in field-isolates and multiple studies have shown epistatic interactions between these mutations determine the accessible evolutionary trajectories to the highly resistant quadruple mutation N51I,C59R,S108N,I164L. We simulated the possible evolutionary trajectories to this quadruple *Pf*DHFR mutation as well as the homologous *Pv*DHFR mutations. In both cases, the most probable pathways agreed well with those determined experimentally. This method can be applied to any drug-target system where the drug acts by binding to the target.

57 **Significance Statement**

58 Antimicrobial resistance (AMR) is a major public health threat, resulting from the overuse and
59 misuse of antimicrobial drugs. Our ability to monitor emerging resistance and direct the most
60 appropriate treatment strategy (stewardship) would be strengthened by the development of
61 methods to accurately predict the mutational pathways to resistance. Here, we present a
62 computational method to predict the most likely pathways to resistance that result from
63 combinations of mutations that alter drug binding affinity. We utilized the Rosetta Flex ddG
64 protocol to parameterize a thermodynamic evolutionary model to predict the order of fixation of
65 each mutation in the pathway. This approach accurately captures known resistance trajectories
66 and can be applied to any system where resistance arises via drug binding affinity changes.

67

68 **Main Text**

69 **Add in arguments from presentation**

70 **Introduction**

71 The development of new antimicrobial therapeutics and the design of more successful drug
72 deployment strategies to reduce the prevalence of resistance, requires an understanding of the
73 underlying molecular evolution. Antimicrobial resistance (AMR) poses a huge global health threat
74 through a wide range of mechanisms [1-3]. One of the major routes to resistance, and the focus
75 of the work presented here, is genomic variation within protein coding regions. Of particular
76 significance are single-nucleotide polymorphisms (SNPs) in the antimicrobial target gene that
77 alter the protein structure and prevent efficient binding of the antimicrobial drug. Provided these
78 SNPs do not prevent the target from carrying out its function, the resistant strains will proliferate
79 within the population [4].

80

81 The evolution of resistance is affected by the interplay between selection for resistance, selection
82 for protein function, drug concentration and mutational bias, and it is also influenced by a
83 phenomenon known as epistasis [5-7].

Epistasis between mutations within the same protein arises due to energetic interactions between the amino acids, where the impact of a mutation depends upon the protein sequence [8]. When epistasis occurs between two or more mutations, their combined impact on protein fitness or a physical trait such as stability or binding affinity, does not equal the sum of their independent impacts. Epistasis been observed in the evolution of many pathogens [9-11], including the evolution of resistance in *Plasmodium falciparum* [6, 12] and *Plasmodium vivax* [7] . Epistasis can influence the evolution of resistance because it determines the order of fixation of mutations and the accessibility of evolutionary trajectories to resistance phenotypes [5, 13]. It may also have important consequences for the success of AMR management strategies that aim to reduce resistance via the cessation of use of a particular drug. In theory, stopping the use of a drug should result in reversion of resistance mutations, because most resistance mutations incur a large fitness cost in the absence of drug [14, 15]. However, the success of this strategy has been mixed, and in some cases bacterial populations remained resistant [16-18], which may be the result of compensatory mutations (a type of epistasis), which mitigate the deleterious impact of resistance mutations, allowing them to remain in a population and thus retain resistance even in the absence of drug selection pressures [19].

Fragment-based drug discovery (FBDD) and AMR surveillance strategies require methods to predict evolutionary trajectories to resistance. For example, by identifying mutations involved in resistance trajectories that might impact the effectiveness of an antimicrobial drug, specific regions of a target molecule can be exploited or avoided, thus creating ‘evolution proof’ drugs. Therefore, understanding how epistasis arises and predicting which mutations will interact, is important for anticipating future mutations, designing new drugs and developing strategies to minimize resistance.

Evolution towards drug-resistant phenotypes in malaria species *P. falciparum* and *P. vivax* has been shown to occur in a stepwise manner, due to epistatic interactions between mutations, and

112 the most likely trajectories to resistance phenotypes have been predicted using experimental
113 measures of resistance [6, 7, 12].
114
115 *P. falciparum* and *P. vivax* parasites cause the majority of malaria infections and have evolved
116 strong resistance to many antimalarial drugs, including pyrimethamine [12] and sulfadoxine [20].
117 There were an estimated 241 million new cases of malaria world-wide in 2020, resulting in
118 approximately 627,000 deaths predominately among children under 5 years of age [21]. *P.*
119 *falciparum* malaria has been treated with the combination drug sulfadoxine-pyrimethamine (SP)
120 since 1970s, which targets the folate metabolic pathway. Numerous resistance mutations have
121 arisen within its genome as a result of SNPs in *P. falciparum* dihydrofolate reductase (*Pf*DHFR)
122 and dihydropteroate synthase (*Pf*DHPS) genes, which are the targets of pyrimethamine and
123 sulfadoxine respectively [20, 22]. Although SP is not usually used to treat *P. vivax*, co-infections
124 with *P. falciparum* have meant SP resistance mutations have also arisen in the *P. vivax* genome
125 [23]. The enzymes of the folate pathway are largely conserved across *Plasmodium* species, and
126 so polymorphisms in equivalent positions have been observed in *P. vivax* DHFR (*Pv*DHFR) and
127 DHPS (*Pv*DHPS) and are thought to confer resistance to SP [24, 25].
128
129 The DHFR gene encodes an enzyme that uses NADPH to synthesize tetrahydrofolate, a co-
130 factor in the synthesis of amino acids [26], and pyrimethamine acts to disrupt this process,
131 thereby blocking DNA synthesis and slowing down growth. Stepwise acquisition of multiple
132 mutations leading to resistance to pyrimethamine has been observed in both *Pf*DHFR [6, 12] and
133 *Pv*DHFR [7].
134
135 Resistance In *Pf*DHFR has been studied extensively and a combination of four mutations – Asn-
136 51 to Ile (N51I), Cys-59 to Arg (C59R), Ser-108 to Asn (S108N) and Ile-164 to Leu (I164L) – has
137 been reported to result in resistance to pyrimethamine [27] by altering the binding pocket and
138 reducing the affinity for the drug [28]. Epistasis in both pyrimethamine binding free energy and the
139 concentration required to inhibit cell growth by 50% (IC50) has been observed experimentally for

140 combinations of these four mutations [6, 12]. This means that mutations which on their own are
141 not associated with a resistance phenotype, can be when in combination with other mutations.
142 Epistasis between these mutations has been shown to determine the evolutionary trajectories to
143 the quadruple mutation N51I,C59R,S108N,I164L, which is strongly associated with
144 pyrimethamine resistance [6].
145
146 A similar investigation was conducted into the homologous set of *Pv*DHFR mutations – Asn-50 to
147 Ile (N50I), Ser-59 to Arg (S58R), Ser-117 to Asn (S117N) and Ile-173 to Leu (I173L) - and the
148 accessible evolutionary trajectories to the quadruple mutation [7], some combinations of which
149 have been observed to result in pyrimethamine resistance both *in vivo* and *in vitro* [25, 29].
150 Evolutionary simulations accounting for growth rates, IC50 measurements for increasing
151 concentrations of pyrimethamine and nucleotide bias predicted the most likely pathways to the
152 quadruple mutation for different drug concentrations. The observed trajectories at each
153 concentration were influenced by epistasis between the mutations and the adaptive conflict
154 between endogenous function and acquisition of drug resistance. These studies, along with other
155 investigations [5, 30] have highlighted the prevalence of epistasis among resistance mutations
156 and the importance of considering epistatic interactions between mutations when predicting
157 evolutionary trajectories to drug resistance.
158
159 The predictability of evolution is a central topic in biology of interest to experimentalists and
160 theorists alike [31-34] (for a review of the topic see [35]). By using experimentally measured
161 values to characterize the empirical fitness landscapes and simulate evolutionary trajectories, the
162 work in [6] and [7] is determining the predictability of evolution in these landscapes by assessing
163 which trajectories are accessible and the level of determinism associated with the evolution.
164 Whilst such experimental methods have been successful in capturing epistasis, characterizing
165 evolutionary landscapes and predicting evolutionary trajectories, they are expensive and time
166 consuming.
167

168 The development of computational methods to predict resistance trajectories would enable fast
169 and efficient predictions and would be more widely accessible than lab-based methods.
170 Computational tools could help narrow down the pool of mutations to be studied experimentally
171 and would also be applicable to difficult to study targets. Some target-specific computational tools
172 to predict individual resistance mutations have been developed [36, 37]. However, such tools are
173 target specific and so not generalizable. Furthermore, they only consider independent mutations
174 on a single structure and so ignore epistasis between resistance mutations. Therefore, they are
175 not suitable for predicting evolutionary trajectories to resistance.
176
177 To determine a generalizable computational method to predict evolutionary trajectories to
178 resistance, we need to consider the main determinants of resistance. Rodrigues et al. (2016) [38]
179 investigated three mutations in *Escherichia coli* DHFR associated with trimethoprim resistance
180 and considered activity, binding affinity, fold stability, and intracellular abundance. They found
181 that whilst resistance is a trade-off between these factors, binding affinity is the single most
182 predictive trait of resistance, especially at later points in evolution. Therefore, we decided to
183 investigate if predictions of binding affinity change can be used to predict the order of fixation of
184 resistance mutations involved in evolutionary trajectories to resistance.
185
186 Rosetta Flex ddG [39] is the current state-of-the art method for predicting changes in protein-
187 protein and protein-ligand binding free energy. Rosetta is a software suite for macromolecular
188 modelling and design that uses all-atom mixed physics- and knowledge-based potentials, and
189 provides a diverse set of protocols to perform specific tasks, such as structure prediction,
190 molecular docking and homology modelling [40]. The Flex ddG protocol has been found to
191 perform better than machine learning methods and comparably to molecular dynamics methods
192 when tested on a large dataset of ligand binding free energy changes upon protein mutation [41,
193 42]. However, its ability to capture epistasis has not yet been tested. Therefore, we investigated
194 how well Flex ddG can capture epistasis between resistance mutations in *Pf*DHFR and observed
195 a good agreement with experimental data.

196

197 Next, we used the Flex ddG predictions to parameterize a fitness function and applied this fitness
198 function in an existing model of molecular evolution. We used this method to predict evolutionary
199 trajectories to known resistant quadruple mutants in both *Pf*DHFR and *Pv*DHFR, where the
200 evolutionary trajectories have been studied experimentally [6, 7]. Good agreement was observed
201 between the most likely trajectories to the quadruple mutations predicted by our model and those
202 predicted experimentally.

203

204 The main advantage of this approach is that it does not require access to an experimental 'wet'
205 lab and can be carried out by anyone with access to a high-performance computer (to run the
206 Flex ddG predictions). It is generalizable to any antimicrobial drug that acts by binding to its target
207 and can be easily applied to any drug-target complex for which there is an available structure.
208 Therefore, it can be used to study complexes and systems that might be problematic
209 experimentally. It enables accurate assessment of the predictability of the evolutionary
210 landscape, in other words it can predict whether we would expect to see constrained evolutionary
211 trajectories on a fitness landscape as a result of epistatic interactions in drug binding free energy.

212

213 In addition, we analyzed if evolutionary pathways can be inferred from the frequency of mutations
214 found in isolate data. We determined the frequency of mutations in *Pf*DHFR and *Pv*DHFR, and
215 inferred the most likely evolutionary pathways under the assumption that the most likely mutation
216 at each step corresponds to the most frequent mutation. We carried out this analysis first upon a
217 combined set of global isolates from all regions and then upon isolates from each region
218 individually. In general, the most likely pathways inferred from the global isolate data agreed
219 remarkably well with both the experimentally determined pathways and the pathways predicted
220 by our computational method. When considering geographical regions separately, the inferred
221 pathways from several regions agreed well with the experimental pathways and our predicted
222 pathways, however the most likely pathways inferred in some regions differed from the main
223 pathways, highlighting the importance of considering the evolution in different regions separately.

Results

Rosetta Flex-ddG accurately predicts binding free energy changes and captures epistasis

We investigated if Flex ddG predictions agree with experimentally measured binding free energy and if these predictions can be used to calculate the non-additivity in binding free energy (interaction energy), which for a double mutant defines the epistasis between the two mutations and, for a triple mutant or higher, captures the level of epistatic interactions. We calculated the interaction energy by finding the difference between the predicted change in binding free energy of a multiple mutation and the sum of the predictions of their independent binding free energy changes. A positive value of the interaction energy indicates the sum of the independent impacts is more destabilizing than the impact of the multiple mutation and a negative value indicates the sum is less destabilizing than the combined impact.

The change in binding free energy was predicted using Rosetta Flex ddG for the combinatorically complete set of the four *Pf*DHFR pyrimethamine resistance mutations N51I, C59R, S108N and I164L. (Note on notation: lists of single mutations are written X, Y, Z, multiple mutations are written X,Y,Z (i.e. no space between the commas and the mutations) and pathways are written X/Y/Z to denote the order of fixation).

We compared the predictions to the data from [12] in which they present the binding free energy changes for a subset of the possible combinations of mutations, the sum of the independent mutations (calculated for multiple mutants to compare to the experimentally determined binding free energy changes of multiple mutants) and the interaction energy of the multiple mutants (Table 1). In Table 1, a positive $\Delta\Delta G$ value indicates a destabilising mutation and a negative $\Delta\Delta G$ value indicates a stabilising mutation (Note: Rosetta Flex ddG calculates the change in binding free energy as $\Delta\Delta G = \Delta G_{mut} - \Delta G_{WT}$, whereas [12] calculated the change as the reverse, $\Delta\Delta G = \Delta G_{WT} - \Delta G_{mut}$, where *WT* indicates the wild-type free energy and *mut* indicates the mutant free energy. Therefore, in [12], a mutation that destabilized the binding corresponded to a negative

252 $\Delta\Delta G$, whilst here we have reversed the signs of their data to enable comparison with our
253 predictions).
254
255 The authors of the Flex ddG protocol suggest conducting a minimum of 35 runs and to take the
256 average of the distribution as the prediction for that mutation [39]. To determine the optimum
257 number of runs per mutation, we ran predictions for 35 (Figure S1, Table S1) and 150 (Figure S2,
258 Table 1) runs and a single prediction at 300 runs (Figure S3), and concluded 150 runs provides
259 the best trade-off between accuracy and efficiency (see Supplementary text for detailed
260 discussion). Therefore, we will be discussing the predictions for n=150 going forward
261
262 Examining the distributions of the predicted change in binding free energy for each of the
263 mutations considered in [12] (Figure S2) we can see that the distributions are not Gaussian but
264 have a defined peak, and so the average of the distribution may not be the best summary metric
265 to use to compare to the data. We therefore chose the peak of the distribution as the summary
266 metric of each prediction. We compared the predictions for 150 runs and the data from [12]
267 (Table 1) and observed a correlation of 0.635 between for the binding free energy data, 0.435 for
268 the sum of the independent predictions for multiple mutants and 0.765 for the interaction energy.
269 We found 8/9 binding free energy predictions were correctly classified, 4/5 of the sum of the
270 independent predictions were correctly classified and 4/5 of the interaction energies were
271 correctly classified.
272
273 Mutation S108N was the only single mutation to destabilize pyrimethamine binding in both the
274 experimental data and the Flex ddG predictions. However, in the experimental data the double
275 mutation N51I,S108N is more destabilizing to binding than single mutation S108N, but the peak
276 of the Flex ddG distribution was stabilizing. The triple mutation C59R,S108N,I164L was found
277 experimentally to be the most destabilizing of the triple mutations, however the peak of the Flex
278 ddG distribution predicted it to be only mildly destabilizing and the least destabilizing of the triple
279 mutations. Furthermore, the quadruple mutation was found experimentally to have the most

destabilizing impact out of all combinations of single and multiple mutations, however, Flex ddG predicted it to be less destabilizing than the double mutation C59R,S108N.

Considering the interaction energy, the incorrectly classified mutation was again N51I,S108N, where the interaction energy was predicted to be positive, but found experimentally to be negative, because the sum of the individual predictions was destabilizing but the double mutation itself was predicted to have a stabilizing impact. Both the experimental data and our predictions found that the quadruple mutation had the largest magnitude interaction energy reflecting the greatest difference between the stabilizing impact of the sum of the individual mutations and the destabilizing impact of the quadruple mutation itself.

The triple mutation C59R,S108N,I164L was found to have a large positive interaction energy in both the experimental data and our predictions, where the triple mutation was found to be destabilizing whilst the sum of the individual impacts was stabilizing to pyrimethamine binding. We also observed large negative interaction energy between S108N and C59R, where C59R is stabilizing in the wildtype background but destabilizing in the background of S108N, an example of sign epistasis and in agreement with the observations of both [12] and [6]. However, whilst the interaction energy of the triple mutation N51I,C59R,S108N was positive for both the experimental data and predictions, in our predictions its magnitude was much smaller compared to the data. Both single mutations N51I and C59R were predicted to be only marginally stabilizing – almost neutral - to pyrimethamine binding, whilst in the experimental data both mutations have a large stabilizing impact. Furthermore, the triple mutation was predicted to be only marginally more destabilizing than single mutation S108N, resulting in the small negative interaction energy.

We use these predictions in the following sections to parameterize an evolutionary model to determine if the predicted evolutionary trajectories agree with experimentally determined evolutionary trajectories.

308 **A thermodynamic evolutionary model predicts the most likely evolutionary trajectories to**
309 **quadruple mutations in both *Pf*DHFR and *Pv*DHFR**
310 We simulated the evolutionary trajectories to the quadruple mutants described above for the
311 genes *Pf*DHFR and *Pv*DHFR using an evolutionary model, adapted from previous studies [43-
312 45]. In this model, selection acts to reduce the binding affinity between target protein and the
313 antimalarial drug with which the mutations have been associated with resistance. Briefly, starting
314 from the wild-type protein, we randomly sample a value from the Flex ddG distributions for each
315 of the four single mutations and calculate the fitness of the mutated protein (Eq. 1), and the
316 fixation probability (Eq. 2). A mutation is then chosen with a probability proportional to the fixation
317 probability and this is repeated until the quadruple mutation is reached. If the set of sampled
318 mutations at a step all have a fixation probability of zero, the algorithm terminates at that point in
319 the pathway and begins the next run at the single mutation step. Therefore, it is not guaranteed
320 that a run will reach the quadruple mutation. We carried out 10,000 runs and determined i) the
321 number of runs that reached a single, double, triple or the quadruple mutation before the run
322 ended (files ending '_endpoint_numbers.csv'), ii) the frequency of the observed trajectories up to
323 the quadruple mutation, including trajectories that terminated before the quadruple mutation (files
324 ending '_pathway_endpoints.csv'), iii) the frequency at which a mutational step was chosen in all
325 runs (files ending '_total_pathway_probabilities.csv) and iv) the most likely trajectories to the
326 quadruple mutation predicted by our simulations (see files ending '_quadruple_pathways.csv').
327
328 To determine how well our simulations reflect the evolutionary process to the *Pf*DHFR quadruple
329 mutation N51I,C59R,S108N,I164L, we compared our results to experimentally determined
330 evolutionary trajectories as presented by [6]. In our simulations, the quadruple mutation was
331 reached in approximately 12% of runs (see file 'PfDHFR_endpoint_numbers.csv' in the
332 Supplementary data). The majority of the runs (75%) terminated at a double mutation, with
333 S108N/C59R the most likely trajectory over all. The algorithm was often unable to proceed
334 passed S108N/C59R in many runs because the Flex ddG distribution for C59R,S108N is
335 concentrated around large destabilizing values (Figure S2f) whilst the distributions of the two

336 possible next steps, N51I,C59R,S108N and C59R,S108N,I164L, are concentrated around lower
337 destabilizing values (Figures S2g and h). Therefore, in many instances, the change in binding
338 free energy caused by the next step in the pathway were predicted to be stabilizing, and thus
339 were not be chosen by the algorithm. This demonstrates the dependence of the method upon the
340 accuracy of the Flex ddG. In contrast, the majority of runs in the simulations based on IC50
341 measurements presented in [6] reached the quadruple mutation (see Figure 2 in [6]).
342
343 However, since we are interested in how epistasis influences the order of fixation of mutations in
344 an evolutionary trajectory to a high-resistance quadruple mutation, we compared the most likely
345 trajectories to the quadruple mutation predicted by our simulations to the most likely trajectories
346 to the quadruple mutation predicted in [6] and observed remarkable agreement. The top two most
347 likely trajectories predicted by our model to the *Pf*DHFR quadruple mutation were
348 S108N/C59R/N51I/I164L, S108N/C59R/I164L/N51I, respectively (Figure 1a) and these pathways
349 correspond to the top two most likely pathways to the quadruple determined in [6]. The third most
350 likely trajectory to the quadruple mutation in [6] was predicted to be S108N/N51I/C59R/I164L,
351 however this pathway was predicted to be very unlikely in our simulations, due to the fact that the
352 distribution of Flex ddG predictions for double mutation N51I,S108N was mostly stabilizing to
353 pyrimethamine binding (Figure S2e), whereas all of the S108N distribution was destabilizing to
354 pyrimethamine binding, so this step was unlikely to be chosen by the evolutionary algorithm.
355
356 Considering the frequency at which the single mutations were chosen as the first step in all
357 simulated pathways (see file 'PfDHFR_total_pathway_probabilities.csv'), S108N was the most
358 likely single mutation and C59R was the second most likely single mutation, in agreement with
359 the two most likely first steps in the pathways predicted in [6]. In our simulations, N51I was the
360 third most likely single mutation and I164L has zero probability of being chosen as the first
361 mutation, due to the fact that all 150 predicted binding affinity changes were stabilizing to
362 pyrimethamine binding. The most likely pathway to a double mutation realized in all trajectories in

363 both our simulations and the simulations in [6] is S108N/C59R. Similarly, the most likely pathway
364 to a triple mutation realized in all our simulations and in [6] was S108N/C59R/N51I.
365
366 To simulate the evolutionary pathways for *Pv*DHFR, we also carried out predictions of binding
367 free energy changes for the homologous set of four mutations in *Pv*DHFR, (N50I, S58R, S117N
368 and I173L). As described in the previous section, we carried out 150 Flex ddG runs per mutation.
369 Unfortunately, there is no binding affinity data available for the mutations in *Pv*DHFR to compare
370 to the Flex ddG predictions. However, Jiang et al. (2013) [7] predicted pathways to the *Pv*DHFR
371 quadruple mutation for four pyrimethamine concentrations using simulations informed by both
372 drug resistance and catalytic activity, to which we can compare our simulations. In their
373 simulations, the quadruple mutation fixed in 99.8% of runs for the highest pyrimethamine
374 concentration, but it did not fix for the three lower concentrations. In our simulations, the
375 quadruple mutation was reached in 47% of runs (see file 'PvDHFR_endpoint_numbers.csv'),
376 whilst 48% of runs terminated at a double mutation. The most likely pathway endpoint overall in
377 our simulations was S58R/I173L, which occurred in 29% of runs, however this path was not a
378 frequent trajectory observed in the simulations in [7]. The double mutation S58R,I173L was
379 predicted by Flex ddG to have a large destabilizing impact on pyrimethamine binding, with all
380 runs predicted to be destabilizing and many predicted to have a medium to large impact.
381 However, the triple mutation S58R,S117N,I173L was predicted to have a smaller destabilizing
382 impact than S58R,S117N, making pathway S58R/I173L/N50I unlikely, whilst N50I,S58R,I173L
383 was predicted to be stabilizing to pyrimethamine in all Flex ddG runs and therefore pathway
384 S58R/I173L/N50I had a zero probability of occurring in the simulations. This resulted in many
385 runs terminating at step S58R/I173L.
386
387 Considering the order of fixation of mutations in the quadruple mutation, if we compare the most
388 likely evolutionary trajectories to the quadruple mutation predicted by our simulations (see file
389 'PvDHFR_quadruple_pathways.csv') to the most likely evolutionary trajectories to the quadruple
390 mutation presented in [7], we observed good agreement for the largest of the four concentrations

391 they considered. The most likely pathway to the quadruple mutation predicted by our simulations
392 was S58R/S117N/I173L/N50I (Figure 1b) which corresponds to the second most likely pathway to
393 the quadruple mutation predicted in [7] for the highest pyrimethamine concentration. Our second
394 most likely pathway to the quadruple mutation (S58R/S117N/N50I/I173L) corresponds to the first
395 most likely pathway predicted in [7] for the highest pyrimethamine concentration. There were two
396 other possible pathways to the quadruple at the highest concentration, S117N/N50I/S58R/I173L
397 and N50I/S117N/S58R/I173L, which correspond to the sixth and seventh most likely pathways to
398 the quadruple in our simulations.
399
400 The first step in the evolutionary trajectories determined in [7] for the highest concentration was
401 S58R whereas for the three lower concentrations it was S117N. The most likely first step in all
402 pathways predicted by our simulations was S58R (Figure 1b), whilst S117N was the second most
403 likely first step (see 'PvDHFR_total_pathway_probabilities.csv'). Analyzing the Flex ddG
404 distributions, the S58R predictions are mostly destabilizing to pyrimethamine and it is the only
405 single mutation to reduce the binding affinity when considering both the average and the peak of
406 the distribution (Figure S4). A large proportion of the Flex ddG predictions for single mutation
407 S117N are mildly stabilizing to pyrimethamine and whilst the average is mildly destabilizing, the
408 peak of the distribution is a mild stabilizing effect (Figure S5).
409
410 To quantify the predictability of an evolutionary landscape, previous studies have calculated the
411 Gibbs-Shannon entropy distribution of the path weights [34, 35], namely $S = -\sum P_i \, lnP_i$, where P_i
412 is the probability of the ith pathway and the value of S ranges from 0 to $\ln(n)$ for n equally likely
413 pathways. The lower the value of S the higher the predictability of the evolution i.e. most of the
414 probability is concentrated around a small number of pathways, which suggests epistasis is
415 influential in constraining the accessible trajectories. The value of S when considering the
416 probability distribution of all realized evolutionary in the simulations was 1.29 for *Pf*DHFR and
417 2.56 for *Pv*DHFR (both simulations have an equal number of possible pathways because they
418 have an equal number of mutations, so the values of S are comparable and the maximum value

419 of S for both simulations is 4.16). This means the evolutionary trajectories were more constrained
420 in the *Pf*DHFR simulations than in the *Pv*DHFR simulations and suggests that epistasis between
421 the mutations plays a greater role in constraining the trajectories in the evolution of *Pf*DHFR
422 resistance. Unfortunately, the probabilities of all possible pathways determined in [6] and [7] are
423 not made available (the data is represented in pathway diagrams, the probabilities of a step are
424 indicated by line thickness and only the probabilities of the most likely pathways annotated),
425 therefore we cannot calculate the corresponding values of S for these distributions for
426 comparison.
427
428 **The frequency of mutations in isolate data can be used to infer evolutionary trajectories to**
429 **multiple resistance mutations**
430 It was noted in [6] that their most likely pathways to the *Pf*DHFR quadruple mutation were
431 consistent with combinations of these four mutations observed in high frequencies in worldwide
432 surveys of *P. falciparum* polymorphisms. To expand on this idea, we analyzed the frequency of
433 the possible combinations the mutations in *Pf*DHFR and *Pv*DHFR found in our isolate data to
434 identify if there is agreement between these frequencies, the experimentally determined
435 trajectories and our predicted trajectories and if, therefore, isolate frequency data may be used to
436 infer evolutionary trajectories. We inferred evolutionary trajectories from the frequency data by
437 assuming if a specific mutation was found in high frequency (and is part of the combinatorically
438 complete set of four mutations found in the four genes) then it is likely to be part of the
439 evolutionary trajectory towards the quadruple mutation. To infer the first step in the most likely
440 trajectory, we considered the frequency of single mutations of the set of four mutations
441 considered for each gene and selected the most frequent mutation. To infer subsequent steps in
442 the trajectory, we considered the frequency of only those mutations that contain the previous
443 mutation and another of the set of four mutations in some combination and chose the most
444 frequent mutation at each step. We also inferred alternative pathways which from the frequency
445 data are less likely than the main pathway, but still a possibility due to the occurrence of
446 intermediate mutations in the isolate data. To do this, we considered each step in the most likely

trajectory and identified any other high frequency mutations that would enable alternative pathways from the double mutation onwards. If there were no alternative pathways, we began the process again but chose the second most frequent single mutation (if applicable) and built the pathway from there. In the event of multiple alternative pathways, we are unable to determine their relative likelihoods, only that they are less likely than the most likely pathway.

Considering the total frequency of each mutation in the set of four *Pf*DHFR mutations (N51I, C59R, S108N and I164L) in the isolate data (Figure 2a), S108N was the most frequent single mutation (72/6762 isolates), C59R,S108N the most frequent double mutation (496/6762 isolates) and N51I,C59R,S108N the most frequent triple mutation (3714/6762 isolates). The quadruple mutation N51I,C59R,S108N,I164L was found in 1439/6762 isolates. This suggests the pathway proceeds in the order S108N/C59R/N51I/I164L, in agreement with the most likely pathway to the quadruple mutation from both our evolutionary simulations and those using experimental data [6].

Triple mutation C59R,S108N,I164L was found in 101/6762 isolates, suggesting that the second most likely pathway to the quadruple from our simulations and experimental data, S108N/C59R/I164L/N51I, is a possible less likely trajectory to the quadruple mutation. Double mutation N51I,S108N was the second most frequent double mutation in the isolate data (198/6762 isolates), allowing for a third most likely pathway S108N/N51I/C59R/I164L. This agrees with the third most likely pathway presented by [6], however it was the fifth most likely pathway to the quadruple in our simulations.

Single mutations C59R and N51I were the second and third most prevalent single mutations in our isolate data, found in 5/6762 and 1/6762 of isolates, respectively. They were also the second and third most likely first step in our pathway predictions (see file 'PfDHFR_total_pathway_probabilities.csv'). Single mutation I164L was absent from the isolate data and had a zero probability of being selected as the first step of our evolutionary trajectories.

475 Considering the set of four *Pv*DHFR mutations (N50I, S58R, S117N and I173L), in our isolate
476 data, the mutations S58R and S117N are fixed at these locations and the wild-type alleles are
477 now considered to have an Arginine at codon 58 and Asparagine at codon 117. However in [7]
478 they consider the wild-type allele to have a Serine at codons 58 and 117 and therefore we have
479 changed our definition of the wild-type allele to agree with [7] for ease of comparison with their
480 evolutionary pathways and our own.
481
482 The most frequent single mutation was S117N (70/847 isolates), the most frequent double
483 mutation was S58R,S117N (237/847) and the only observed triple mutation was
484 S58R,S117N,I173L (2/847) (Figure 2b). The quadruple mutation N50I,S58R,S117N,I173L was
485 not observed in our isolate data, and has not been reported in the literature either. By considering
486 the frequency of the possible combinations of mutations, we inferred the evolution towards triple
487 mutation S58R,S117N,I173L most likely occurs via pathway S117N/S58R/I173L. This
488 corresponds to the third most likely pathway to a triple mutation when considering all pathways
489 observed in our simulations, however this pathway is not observed in any of the most frequent
490 pathways at any of the four concentrations studied in [7].
491
492 Single mutation S58R was the second most frequent single mutation in our isolate data (20/847).
493 This supports the predicted first evolutionary steps in [7] which were predicted to be S58R for the
494 highest pyrimethamine concentration and S117N for three lower concentrations of
495 pyrimethamine, which suggests both single mutations are possible, but S117N is more likely for a
496 lower pyrimethamine concentrations. An alternative pathway to the triple mutation could therefore
497 be S58R/S117N/I173L, which may be more likely under higher pyrimethamine concentrations.
498 This corresponds to the most likely pathway to a triple mutation in our simulations and is part of
499 the second most likely pathway to the quadruple mutation at the highest pyrimethamine
500 concentration considered in [7].
501

Analysis of geographical distribution of mutations found in our isolate data reveals alternative pathways to resistance

The analysis in the previous section inferred evolutionary trajectories using the frequency data of mutations found in worldwide isolate data. We next considered the evolutionary trajectories by geographical location to determine if there are any differences in the inferred trajectories compared to the global trajectories, and which areas agree with the trajectories predicted by our simulations. The *P. falciparum* isolates were grouped in to seven geographical regions, as defined by the United National Statistics Division: South America (Brazil, Colombia and Peru), West Africa (Benin, Burkina Faso, Cameroon, Cape Verde, Cote d'Ivoire, Gabon, Gambia, Ghana, Guinea, Mali, Mauritania, Nigeria and Senegal), Central Africa (Congo [DRC]), East Africa (Eritrea, Ethiopia, Kenya, Madagascar, Malawi, Tanzania, Uganda), South Asia (Bangladesh), Southeast Asia (Cambodia, Indonesia, Laos, Myanmar, Thailand and Vietnam) and Melanesia (Papua New Guinea). The *P. vivax* isolates were grouped into seven broad geographical regions, as defined by the United Nations Statistics Division: Central America (Mexico), South America (Brazil, Colombia, Guyana, Panama, Peru), East Africa (Ethiopia, Eritrea, Madagascar, Sudan, Uganda), South Asia (Afghanistan, Bangladesh, India, Pakistan, Sri Lanka), Southeast Asia (Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Thailand and Vietnam), East Asia (China) and Melanesia (Papua New Guinea). The mutation frequency data discussed in this section for each country can be found in Supplementary data folders 'PfDHFR/IsolateMutationFrequency' and 'PvDHFR/IsolateMutationFrequency'. For an analysis of the frequencies and inferred pathways per country per region, as well as information on additional mutations found in the data, see Supplementary text.

As in the previous section, we inferred the most likely pathway by assuming the most frequent mutation at each step corresponds to the most likely evolutionary trajectory. Considering the frequency of mutations found in *Pf*DHFR isolates from the different geographical regions (Figure 3), the most likely evolutionary pathway to the quadruple mutation N51I,C59R,S108N,I164L was inferred to agree with the main pathway, S108N/C59R/N51I/I164L, predicted by our evolutionary

model and the data presented in [6], as well as the most likely pathway inferred by considering the frequency of all worldwide *Pf*DHFR isolate data in West Africa (S108N: 31/2594; C59R,S108N: 211/2594; N51I,C59R,S108N: 1739/2594; N51I,C59R,S108N,I164L: 1/2594), South Asia (S108N: 2/86; C59R,S108N: 39/86; N51I,C59R,S108N: 15/86; N51I,C59R,S108N,I164L: 17/86) and Southeast Asia (S108N: 3/2650; C59R,S108N: 186/2650; N51I,C59R,S108N: 920/2650; N51I,C59R,S108N,I164L: 1419/2650).

This main pathway to the quadruple was also inferred to be a possible alternative pathway in East Africa (S108N: 3/904; C59R,S108N: 31/904; N51I,C59R,S108N: 782/904, N51I,C59R,S108N,I164L: 2/904). Additionally, S108N/C59R/I164L/N51I was inferred to be an alternative in South Asia (C59R,S108N,I164L: 11/86) and Southeast Asia (C59R,S108N,I164L: 90/2650), corresponding to the second most likely pathway to the quadruple predicted by our simulations, the data in [6] and total frequency data.

However, in East Africa S108N/N51I/C59R/I164L was the most likely pathway to the quadruple mutation (S108N,N51I: 67/904), which corresponds to the fifth most likely pathway to the quadruple in our simulations and the third most likely pathway presented in [6] and inferred from the total frequency data. This was also an alternative pathway in West Africa (S108N,N51I: 32/2594) .

Furthermore, the quadruple mutation was not observed in the isolate data from Central Africa, South America, and Melanesia. In Central Africa, evolution up to the triple mutation N51I,C59R,S108N was observed and the most likely inferred pathway to this mutation was S108N/N51I/C59R (S108N: 6/359; N51I,S108N: 86/359; N51I,C59R,S108N: 258/359), with an alternative less likely pathway of S108N/C59R/N51I (C59R,S108N: 6/359), corresponding to the fifth and first most likely trajectories to a triple mutation in our simulations, respectively. In South America, evolution up to the triple mutation N51I,S108N,I164L was observed in the isolate data and was inferred to follow the pathway S108N/N51I/I164L (S108N: 26/50, N51I,S108N: 12/50;

558 N51I,S108N,I164L: 4/50), however this pathway was very unlikely in our simulations. In
559 Melanesia, evolution up to the double mutation C59R,S108N was observed in the isolate data
560 and was inferred using the frequency data to have followed the pathway S108N/C59R (S108N:
561 1/119; C59R,S108N: 23/119), which corresponds to the most likely pathway of all our
562 evolutionary runs.
563
564 Next, we considered the frequency of combinations of the four *Pv*DHFR mutations N50I, S58R,
565 S117N and I173L in different geographical regions and used these frequencies to infer
566 evolutionary trajectories (Figure 4). As mentioned previously, the quadruple mutation is not
567 observed in the isolate data and so we will infer trajectories up to triple mutant combinations of
568 the four mutations where possible, and compare to the most likely pathways to triple mutations in
569 our simulations and in the experimental data. The triple mutation S58R,S117N,I173L was the only
570 triple mutant combination observed in the isolate data and was only found in South America.
571 Analyzing the frequency of the constituent mutations (S117N: 34/257; S58R,S117N: 74/257;
572 S59R,S117N,I173L: 2/257), the most likely inferred pathway to this mutation is
573 S117N/S58R/I173L. This was the third most likely triple mutation pathway in our simulations,
574 however it was not an observed pathway in [7]. There are two other possible pathways inferred
575 from the frequency data to this triple mutation: S117N/I173L/S58R (S117N,I173L: 1/257), and
576 S58R/S117N/I173L (S58R: 9/257). Pathway S117N/I173L/S58R was the fourth most likely
577 pathway to a triple mutation in our simulations when considering all possible pathways but was
578 not observed in the simulations in [7] and S58R/S117N/I173L the most likely pathway to a triple
579 mutation in our simulations and was observed as part of the second most likely pathway to the
580 quadruple mutation at the highest pyrimethamine concentration considered in [7].
581
582 Evolution only up to double mutation S58R,S117N was found in Southeast Asia, East Africa and
583 Melanesia and S117N/S58R was the most fixation order in East Africa, whilst S58R/S117N was
584 the most likely fixation order in Southeast Asia and Melanesia. Pathways S58R/S117N and
585 S117N/S58R were the second and third most frequently observed double mutation pathways

586 observed in our simulations when considering all possible pathways observed. In South Asia
587 evolution up to double mutations S58R,S117N and N50I,S117N was observed, following
588 pathways S117N/S58R and S117N/N50I, respectively. In East Asia, evolution only up to single
589 mutation S58R was observed and in Central America no steps in the evolutionary pathway
590 including combinations of these four mutations were found.
591

592 **Discussion**
593
594 We have presented a method for predicting the most likely evolutionary trajectories to multiple
595 mutants by parameterizing thermodynamic evolutionary model using Flex ddG predictions. The
596 most likely pathways predicted by our model to the pyrimethamine-resistant quadruple *Pf*DHFR
597 mutant correspond well to those predicted in [6], generated using experimentally determined IC50
598 values of *Pf*DHFR pyrimethamine binding. The two most likely pathways based on experimental
599 IC50 values were found in the top two most likely pathways to the quadruple mutation based on
600 our simulations using predictions of binding free energy. Whilst our simulations disagreed with the
601 simulations in [6] in terms of which were the most frequently realized pathways out of the total
602 number of runs, where a realized pathway does not necessarily have to reach the quadruple
603 mutation, our model is able to capture the most likely order of fixation of mutations leading to a
604 particular multiple mutant in general agreement with the simulations in [6].
605

606 We also simulated the most likely evolutionary trajectories to the *Pv*DHFR quadruple mutation
607 N50I,S58R,S117N,I173L and compared our results to those predicted in [7]. They considered the
608 relative growth rates of the different alleles at different drug concentrations when simulating
609 evolutionary trajectories, which incorporate both change in pyrimethamine binding affinity (K_i) and
610 catalytic activity (k_{cat}). Our top two most likely pathways to the quadruple correspond to their top
611 two most likely pathways for the highest pyrimethamine concentration they consider, albeit in
612 reverse order. At high pyrimethamine concentrations, it is likely mutations that significantly reduce
613 binding affinity will be selectively favoured even if there is a slight reduction in catalytic activity.
614 Indeed, Rodrigues et al. (2016) [38] observed a clear tradeoff between catalytic activity and

binding affinity for increased drug resistance and found that whilst many molecular features affect drug resistance, drug binding affinity was the key determinant of drug resistance in later stages of evolution. This may be why our predictions agree well their predictions for high pyrimethamine concentration, but not for low-to-middle pyrimethamine concentrations, because even though ligand concentration is included in our equation for protein fitness (Eq. 1), our model cannot account for adaptive conflict between K_i and k_{cat}. This highlights a limitation of our method as it only accounts for changes in binding affinity and does not account for changes in protein function.

As previously mentioned, DHFR catalyzes the reduction of substrate DHF via oxidation of cofactor NADPH. Therefore, in the case of the DHFR enzyme, a future iteration of the model could include the impact resistance mutations have on binding of these two ligands, as a proxy for changes to enzyme function. However, this would require a much more complex model of protein fitness and would be much more computationally expensive.

A further limitation of the evolutionary model used here is that it operates in the weak mutation regime, in which the mutation rate is so low that mutations appear and fix in isolation. However, this assumption breaks down when considering large microbial populations where clonal interference means that mutations can arise simultaneously and compete for fixation [46]. This can lead to a process known as ‘greedy adaptation’, in which the mutation of larger beneficial effect is fixed with certainty [47]. Clonal interference has been shown to emerge rapidly in laboratory cultivated *P. falciparum,* where the parasite cycles through only asexual stages, suggesting it may influence the dynamics of the emergence of resistance [48]. The evolutionary model used here may therefore overestimate the fixation probability of mutations with milder beneficial effects and underestimate the fixation probability of mutations with larger beneficial effects [35]. Future iterations of this work could be improved by making using a fixation probability that models of clonal interference such as the work of Gerrish and Lenski (1998) [46] and Campos et al. (2004) [49]. However, such models are more difficult to implement as they can require species-specific derivations for certain functions.

643

644 Mutations occurring at a drug-binding site may also reduce the protein's thermodynamic stability
645 [50] and therefore may not be selected for, even if they improve the resistance phenotype.
646 Therefore, our model may also be improved by including selection for mutations that do not
647 reduce thermodynamic stability relative to the wild-type enzyme. However, it must also be noted
648 that most proteins are marginally stable [51, 52], a property which may have evolved either as an
649 evolutionary spandrel [53, 54] (a characteristic that arises as a result of non-adaptive processes
650 which is then used for adaptive purposes [55]) or due to selection for increased flexibility to
651 improve certain functionalities [56, 57]. Therefore, the model would also have to account for the
652 fact that a resistance mutation that increases protein stability relative to the wild-type stability may
653 also result in a reduction in fitness.

654

655 Despite the limitations of our computational method to predict evolutionary trajectories by only
656 considering the impact on drug binding, it is able to accurately predict the most likely order of
657 fixation of mutations in a trajectory in general agreement with trajectories determined using
658 experimental values such as IC50 or more complex fitness landscapes informed by multiple
659 parameters including drug concentration and growth rates. This supports the findings in [38], that
660 drug binding is a major determinant of resistance, especially at later stages of evolution. It also
661 suggests evolution in such landscapes is more predictable than might be expected, since
662 trajectories can be predicted considering only the impact on binding affinity.

663

664 We also inferred evolutionary pathways from the total frequency in worldwide clinical isolate data
665 as well as from different geographical regions. This analysis suggests evolutionary pathways may
666 be inferred from the frequency of mutations found in isolate data, however it requires a large
667 number of isolates to properly sample the mutations in the population. Furthermore, this method
668 can only be used to predict trajectories once resistance has emerged in a population, whereas
669 the computational method presented here can predict evolutionary trajectories before of a new
670 drug.

671

672 This analysis also suggested that different regions often follow different evolutionary trajectories
673 and that the most likely evolutionary trajectories predicted by our model, and experimental
674 trajectories, are not always the most prevalent. Geographical differences in the distribution of
675 resistant alleles may be the result of drug regimens and gene flow in parasite populations.
676 Combination drug SP was first used in 1967 to treat *P. falciparum* in Southeast Asia, and
677 resistance was first noted that same year on the Thai-Cambodia and Thai-Myanmar borders [58].
678 In Africa, SP was first used in the 1980s, with resistance occurring later that decade. However,
679 analysis of *Pf*DHFR genotypes and microsatellite haplotypes surrounding the DHFR gene in
680 Southeast Asia and Africa suggest a single resistant lineage that appeared in Southeast Asia
681 accumulated multiple mutations, including the triple N51I,C59R,S108N [59, 60], migrated to Africa
682 and spread throughout the continent [61-63]. Variation in the frequency of *Pf*DHFR mutants
683 across Africa occurs because of differences in the timing of chloroquine withdrawal and
684 introduction of SP, as well as continued use of SP for intermittent preventive treatment (IPTp) in
685 pregnant women residing in areas of moderate to high malaria transmission intensity [64, 65].
686

687 Pyrimethamine resistance increased in West Papua in the early 1960s following the introduction
688 of mass drug administration [66]. Microsatellite haplotype analysis suggests C59R,S108N in
689 Melanesia has two lineages, one of which originated in Southeast Asia whilst the other evolved
690 indigenously [60].
691

692 Pyrimethamine resistance in South America looks surprisingly different from the distributions in
693 Africa and Southeast Asia. SP was introduced in South America and low-level resistance was first
694 noted in Colombia in 1981 [67]. Microsatellite haplotype analysis suggests pyrimethamine
695 resistance evolved indigenously in South America, with at least two distinct lineages detected. A
696 triple mutant lineage (C50I,N51I,S108N) was identified in Venezuela that possibly evolved from
697 double mutant N51I,S108N [62]. A second triple mutant lineage (N51I,S108N,I164L) was

698 identified in Peru and Bolivia which also possibly evolved from a distinct double mutant
699 (N51I,S108N) lineage [68].
700
701 In general, the *Pv*DHFR gene is much more polymorphic than *Pf*DHFR gene, with over 20 alleles
702 observed in a limited geographical sampling [29], whereas fewer *Pf*DHFR alleles have been
703 observed despite much more extensive surveillance with non-synonymous changes and
704 insertions/deletions occurring rarely [69]. It also appears that the origin of *Pv*DHFR
705 pyrimethamine resistance mutation is much more diverse than *Pf*DHFR. Hawkins et al. (2008)
706 [70] investigated isolates from Colombia, India, Indonesia, Papua New Guinea, Sri Lanka,
707 Thailand and Vanuatu. They found multiple origins of the double *Pv*DHFR mutant 58R,117N and
708 three independent origins of triple mutant 58R,61M,117T and quadruple mutants
709 57I,61M,117T,173F and 57L,58R,61M,117T in Thailand, Indonesia and Papua New
710 Guinea/Vanuatu. Shaukat et al. (2021) [71] assessed the evolutionary origin of *Pv*DHFR
711 pyrimethamine resistance mutations in Punjab, Pakistan and found multiple origins of single
712 mutation S117N and a common origin of double mutant 58R,S117N and triple mutants
713 57L,58R,117N, 58R,61M,117N and 58R,118N,I173L. This is in contrast to the evolutionary origin
714 of pyrimethamine resistance in *Pf*DHFR, where mutations in Africa shared a common origin with
715 a resistance lineage from Asia.
716
717 This highlights the need to distinguish between geographical regions and account for existing
718 resistance alleles within that region and trace their lineages when attempting to predict the next
719 step in evolutionary trajectories to highly resistant multiple mutants. Given the current dominant
720 resistance allele from a specific region, our method could be used to predict the most likely next
721 steps from a subset of likely mutations.
722
723 We have presented a computational method for predicting the most likely evolutionary trajectories
724 that has demonstrated good agreement with trajectories predicted experimentally and has the
725 advantage of being much quicker and more cost-effective than laboratory-based methods. This

726 method can be applied to any system in which a drug binds to a target molecule, provided a
727 structure of the complex exists or can be produced via structural modelling. Given the threat
728 antimicrobial resistance poses, methods to accurately and efficiently predict future trajectories are
729 vital and can inform treatment strategies and aid drug development.
730
731
732 **Materials and Methods**
733
734 **Homology modelling**
735 Homology modelling was carried out in Modeller [72] to produce complete structures of the target
736 proteins bound to their drug molecules. Several crystal structures of *Pf*DHFR exist in the Protein
737 Data Bank (PDB). The entry 3QGT provides the crystal structure of wild-type *Pf*DHFR complexed
738 with NADPH, dUMP and pyrimethamine, however residues in the ranges 86-95 and 232-282 are
739 missing from the structural model. Homology modelling was used to complete the structure using
740 a second wild-type *Pf*DHFR structure PDB entry 1J3I along with a wild-type *Pv*DHFR structure
741 PDB entry 2BLB.
742
743 To produce a complete structural model of *Pv*DHFR, PDB entry 2BLB was used as a template,
744 which provides the X-ray crystal structure of wild-type *P. vivax* DHFR in complex with
745 pyrimethamine. This structure was only missing a loop section between residues 87-105 and so
746 Modeller was used to build this missing loop.
747
748 **Flex ddG binding free energy predictions**
749 The Rosetta Flex ddG protocol was used to estimate the change in binding free energy upon
750 mutation, $\Delta\Delta G = \Delta G_{mut} - \Delta G_{WT}$, for each step in all possible mutational trajectories for a set of
751 stepwise resistance mutations (see supplementary data 'Flex_ddG' folder for examples of a
752 Rosetta script, resfile and command line. The protein-ligand structure files and ligand parameter
753 files can be found in the folders named for the specific targets). To predict the change in binding
754 free energy for a single or multiple mutation, we used the structure of the target protein with the
755 drug molecule bound as input to Flex ddG and ran the protocol for 150 times per mutation to

produce a distribution of predictions of the change in the free energy of binding. We then found the peak of the distribution to produce a single estimate of the change in the binding free energy for the mutation, denoted $\Delta\Delta G_X^*$ for mutation *X*.

To predict the stepwise evolutionary trajectories, we must consider the interactions between the mutations in the pathway. The interaction energy (or epistasis) in the binding free energy between two mutations *X* and *Y,* can be written $\varepsilon_{XY} = \Delta\Delta G_{X,Y} - (\Delta\Delta G_X + \Delta\Delta G_Y)$. This quantifies by how much the change in binding free energy of the double mutant *X,Y* deviates from additivity of the single mutants, where each are calculated with respect to the wild-type. Therefore, the change in binding free energy when mutation *Y* occurs in the background of mutation *X* can be written $\Delta\Delta G_{X/Y} = \Delta\Delta G_{X,Y} - \Delta\Delta G_X$, where $\Delta\Delta G_{X/Y} = \Delta\Delta G_Y + \varepsilon_{XY}$.

For a third mutation, *Z*, occurring in the background of double mutation *X,Y*, the interaction energy between *Z* and *X,Y* is $\varepsilon_{XY,Z} = \Delta\Delta G_{X,Y,Z} - (\Delta\Delta G_{X,Y} + \Delta\Delta G_Z)$. The quantity $\varepsilon_{XY,Z}$ is not the same as the third order epistasis between mutations *X*, *Y*, and *Z,* or the interaction energy $\varepsilon_{XYZ} = \Delta\Delta G_{X,Y,Z} - (\Delta\Delta G_X + \Delta\Delta G_Y + \Delta\Delta G_Z)$, as it does not account for the interaction between *X* and *Y,* rather it only quantifies the interaction between *Z* and the two mutations *X* and *Y*..
Therefore, the change in binding free energy when mutation *Z* occurs in the background of double mutant *X,Y* can be calculated as $\Delta\Delta G_{X,Y/Z} = \Delta\Delta G_{X,Y,Z} - \Delta\Delta G_{X,Y}$, where $\Delta\Delta G_{X,Y/Z} = \Delta\Delta G_Z + \varepsilon_{XY,Z}$.

To estimate the change in binding free energy when mutation *Y* occurs in the background of mutation *X*, $\Delta\Delta G_{X/Y}$ for stepwise pathway *X/Y*, we subtracted the predictions $\Delta\Delta G_X^i$ for the first mutation *X,* from the predictions for the double mutation *X,Y,* $\Delta\Delta G_{X,Y}^i$, to create a set of 150 'predictions' for the change in binding free energy when *Y* occurs in the background of *X,* $\Delta\Delta G_{X/Y}^i$ i.e. $\Delta\Delta G_{X/Y}^i = \Delta\Delta G_{X,Y}^i - \Delta\Delta G_X^i$ for $i = \{1, \dots, 150\}$. To estimate the change in binding free energy when mutation *Z* occurs in the background of mutations *X* and *Y* we calculated. $\Delta\Delta G_{X,Y/Z}^i =$

783 $\Delta\Delta G^{i}_{X,Y,Z} - \Delta\Delta G^{i}_{X,Y}$. We applied a similar method for the quadruple mutations, so that we had a set
784 of 'predictions' for each step in the possible evolutionary trajectories.
785
786 **Simulating Evolutionary Trajectories**
787 The Rosetta energy function is a mix of a combination of physic-based and statistics-based
788 potentials and so raw predictions using this function don't up match up with physical energy units
789 (e.g. kcal/mol or kJ/mol). However, the authors of Flex ddG applied a generalized additive model
790 (GAM)-like approach to the Rosetta energy function to reweight its terms and to fit experimentally
791 known values (in kcal/mol). The resulting nonlinear reweighting model reduced the absolute error
792 between the predictions and experimental values and so improved the agreement with
793 experimentally determined interface $\Delta\Delta G$ values. They found that by doing this the Flex ddG
794 predictions of binding free energy changes were in a similar range as experimental binding free
795 energy changes and observed improved correlation and classification of mutations as stabilising
796 or destabilising (Barlow et al. 2018 [39]). Therefore, we assume Flex ddG can provide
797 approximate predictions of binding free energy changes comparable to experimental changes in
798 kcal/mol and can therefore be used to parameterize a thermodynamic model.
799
800 To predict the most likely evolutionary trajectories to reach a quadruple mutant we used a model
801 based in thermodynamics and statistical mechanics where the fitness of a protein is determined
802 by the probability it would not be bound to a ligand, $P_{unbound}$. We consider a two-state system in
803 which the protein can either be bound or unbound and do not explicitly account for if the protein is
804 folded or unfolded in either the bound or unbound state. For ligand concentration $[L]$ it can be
805 shown that the probability a protein is unbound is

806 $$P_{unbound} = \frac{1}{\frac{[L]}{K_d} + 1} \quad (1)$$

807 where K_d is the protein-ligand dissociation constant and can be calculated as $c_0 e^{\Delta G/kT}$ where c_0 is
808 a reference ligand concentration (set here arbitrarily to 1M), ΔG is the protein-ligand binding free
809 energy, k is the Boltzmann constant and T is the temperature in Kelvin.

Starting from the wild-type protein, with binding free energy ΔG_{WT} and fitness $P_{unbound}^{WT}$, we extract one sample *i* from the 150 values of the predicted binding affinity changes for the single mutations to determine the binding free energy after mutation *X*, $\Delta G_X^i = \Delta G_{WT} + \Delta\Delta G_X^i$, and calculate the fitness of each single mutant protein $P_{unbound}^{X(i)}$. We can calculate the probability the mutation will fix in the population using the Kimura fixation probability for a haploid organism

$$p_{fix} = \frac{1 - e^{-2s}}{1 - e^{-2sN_e}} \tag{2}$$

where N_e is the effective population size (set to 10^6 as previous models in [43, 44]) and *s* is the selection coefficient $s = (P_{unbound}^{X,i} - P_{unbound}^{WT})/P_{unbound}^{WT}$. We also took in to account the mutational bias of *Plasmodium falciparum* using the nucleotide mutation matrix calculated in [6]. The probabilities of fixation for each mutation were normalised by the sum of the probabilities of fixation for all possible mutations at that step in the trajectory. A mutation is then chosen with a probability proportional to this normalised probability of fixation.

Once a single mutation is chosen, the binding free energy is set to ΔG_X^i of the chosen mutation, and a value is sampled from the distribution of each of the possible next steps, *X/Y* in the trajectory i.e. $\Delta\Delta G_{X/Y}^i$. This continues until the end of the trajectory is reached. If the fixation probabilities of all mutations sampled at a step are effectively zero, no mutation is chosen at that step and the algorithm begins again by choosing a single mutation. Therefore, not all of the runs produce a complete trajectory and some will terminate before reaching the quadruple mutation. The algorithm was written in R.

We calculate the probabilities, P_i, of each realized pathway (even those which don't reach the quadruple mutation) by dividing the total number of times that specific pathway occurs by the number of runs. We calculate the probability of a particular step by dividing the number of times that step occurs in all realized pathways by the total number of runs.

837 **SNP data**

838 *P. falciparum* and *P. vivax* data was obtained from publicly available raw sequence data from
839 European Nucleotide Archive. These data include Illumina raw sequences from the MalariaGEN
840 Community Project for *P. falciparum [73]* and *P. vivax [74]. P. vivax* data additionally includes the
841 Public Health England Malaria Reference Laboratory isolates from returning travelers to UK from
842 regions where malaria is endemic (study accession number ERP128476) [75]. Data was filtered
843 and processed to SNP data with the methodology described in the recent publications [64] and
844 [75] respectively. In this study, we analysed genotype data for 6,762 high-quality isolates from 32
845 countries across regions of Africa, South East Asia, Oceania and South America to identify the
846 genetic diversity in the *PfDHFR* gene. A similar analysis was carried out on 847 *P. vivax* isolates
847 spanning 25 countries across East Africa, South Asia, South East Asia, Southern South East Asia
848 and South America to identify genetic diversity in *PvDHFR* gene. SNPs occurring in non-unique,
849 low quality or low coverage regions were discarded, and those with a missense effect in the
850 candidate genes were analysed. Functional annotation was done with *SnpEff* (version 5.0) [76]
851 with the following options: *-no-downstream -no-upstream*.

852
853
854 **Acknowledgments**

855 R.C.E and N.F. are funded by the Medical Research Council UK (Grant no. MR/T000171/1).
856 T.G.C is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1,
857 MR/R025576/1, MR/R020973/1 and MR/T000171/1) and BBSRC (Grant no. BB/R013063/1). S.C
858 is funded by Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and
859 MR/R020973/1) and Bloomsbury SET. E.M is funded by a Newton Institutional Links Grant
860 (British Council, no. 261868591. The funders had no role in study design, data collection and
861 analysis, decision to publish, or preparation of the manuscript. R.C.E would like to thank
862 Tanushree Tunstall (London School of Hygiene and Tropical Medicine) for the helpful
863 discussions. R.C.E would also like to thank Professor Richard Goldstein (University College
864 London) for introducing her to the evolutionary algorithm used in this paper.

865
866 **Data Availability**

867
868 Supplementary data can be found at 10.5281/zenodo.6670770
869
870 **References**
871

872 1. Sun, D., et al., *Editorial: Horizontal Gene Transfer Mediated Bacterial Antibiotic*
873 *Resistance.* Frontiers in microbiology, 2019. **10**: p. 1933-1933.
874 2. Davies, J. and D. Davies, *Origins and evolution of antibiotic resistance.* Microbiology and
875 molecular biology reviews : MMBR, 2010. **74**(3): p. 417-433.
876 3. Levy, S.B. and B. Marshall, *Antibacterial resistance worldwide: causes, challenges and*
877 *responses.* Nature Medicine, 2004. **10**(12): p. S122-S129.
878 4. Blair, J.M.A., et al., *Molecular mechanisms of antibiotic resistance.* Nature Reviews
879 Microbiology, 2015. **13**(1): p. 42-51.
880 5. Weinreich, D.M., et al., *Darwinian Evolution Can Follow Only Very Few Mutational Paths*
881 *to Fitter Proteins.* Science, 2006. **312**(5770): p. 111.
882 6. Lozovsky, E.R., et al., *Stepwise acquisition of pyrimethamine resistance in the malaria*
883 *parasite.* Proc Natl Acad Sci, U. S. A., 2009. **106**(29): p. 12025-12030.
884 7. Jiang, P.-P., et al., *Accessible Mutational Trajectories for the Evolution of Pyrimethamine*
885 *Resistance in the Malaria Parasite Plasmodium vivax.* Journal of Molecular Evolution,
886 2013. **77**(3): p. 81-91.
887 8. Starr, T.N. and J.W. Thornton, *Epistasis in protein evolution.* Protein science : a
888 publication of the Protein Society, 2016. **25**(7): p. 1204-1218.
889 9. Khan, A.I., et al., *Negative epistasis between beneficial mutations in an evolving*
890 *bacterial population.* Science, 2011. **332**(6034): p. 1193-1196.
891 10. Gong, L.I., M.A. Suchard, and J.D. Bloom, *Stability-mediated epistasis constrains the*
892 *evolution of an influenza protein.* eLife, 2013. **2**: p. e00631.
893 11. Sanjuán, R., et al., *Epistasis and the adaptability of an RNA virus.* Genetics, 2005. **170**: p.
894 1001-1008.
895 12. Sirawaraporn, W., et al., *Antifolate-resistant mutants of Plasmodium falciparum*
896 *dihydrofolate reductase.* Proc Natl Acad Sci U S A, 1997. **94**: p. 1124-1129.
897 13. Weinreich, D.M., R.A. Watson, and L. Chao, *Perspective: Sign epistasis and genetic*
898 *constraint on evolutionary trajectories.* Evolution, 2005. **59**(6): p. 1165-74.
899 14. Melnyk, A.H., A. Wong, and R. Kassen, *The fitness costs of antibiotic resistance*
900 *mutations.* Evol, Appl, 2015. **8**(3): p. 273-283.
901 15. Vogwill, T. and R.C. MacLean, *The genetic basis of the fitness costs of antimicrobial*
902 *resistance: a meta-analysis approach.* Evolutionary applications, 2015. **8**(3): p. 284-295.
903 16. Costelloe, C., et al., *Effect of antibiotic prescribing in primary care on antimicrobial*
904 *resistance in individual patients: systematic review and meta-analysis.* BMJ, 2010. **340**:
905 p. c2096.
906 17. Enne, V.I., *Reducing antimicrobial resistance in the community by restricting prescribing:*
907 *can it be done?* Journal of Antimicrobial Chemotherapy, 2010. **65**(2): p. 179-182.
908 18. Sundqvist, M., et al., *Little evidence for reversibility of trimethoprim resistance after a*
909 *drastic reduction in trimethoprim use.* J. Antimicrob Chemother, 2010. **65**: p. 350-360.
910 19. Andersson, D.I., *Persistence of antibiotic resistant bacteria.* Curr Opin Microbiol., 2003.
911 **6**: p. 452-456.
912 20. Wang, P., et al., *Sulfadoxine resistance in the human malaria parasite Plasmodium*
913 *falciparum is determined by mutations in dihydropteroate synthetase and an additional*
914 *factor associated with folate utilization.* Mol Microbiol., 1997. **23**(5): p. 979-986.
915 21. *World Malaria Report: 2021*. 2021, World Health Organization, Geneva, Switzerland.

22. Brooks, D.R., et al., *Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, Plasmodium falciparum, with differing resistance to sulfadoxine.* Eur, J. Biochem, 1994. **224**(2): p. 397-405.
23. Snounou, G. and N.J. White, *The co-existence of Plasmodium: sidelights from falciparum and vivax malaria in Thailand.* Trends in Parasitology, 2004. **20**(7): p. 333-339.
24. Korsinczky, M., et al., *Sulfadoxine resistance in Plasmodium vivax is associated with a specific amino acid in dihydropteroate synthase at the putative sulfadoxine-binding site.* Antimicrob Agents, Chemother, 2004. **48**(6): p. 2214-2222.
25. Hastings, M.D., et al., *Dihydrofolate Reductase Mutations in Plasmodium vivax from Indonesia and Therapeutic Response to Sulfadoxine plus Pyrimethamine.* The Journal of Infectious Diseases, 2004. **189**(4): p. 744-750.
26. Kompis, I.M., R.L. Islam K Fau - Then, and R.L. Then, *DNA and RNA synthesis: antifolates.* Chem, Rev, 2005. **105**(2): p. 593-620.
27. Ferlan, J.T., et al., *Mutagenesis of dihydrofolate reductase from Plasmodium falciparum: analysis in Saccharomyces cerevisiae of triple mutant alleles resistant to pyrimethamine or WR99210.* Mol Biochem, Parasitol, 2001. **113**(0166-6851 (Print)): p. 139-150.
28. Yuthavong, Y., et al., *Malarial (Plasmodium falciparum) dihydrofolate reductase-thymidylate synthase: structural basis for antifolate resistance and development of effective inhibitors.* Parasitology,, 2005. **130**(0031-1820 (Print)): p. 249-259.
29. Hawkins, V.N., et al., *Antifolates can have a role in the treatment of Plasmodium vivax.* Trends, Parasitol, 2007. **25**(5): p. 213-222.
30. Tamer, Y.T., et al., *High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a Rugged Fitness Landscape in the Presence of Trimethoprim Selection.* Molecular Biology and Evolution, 2019. **36**(7): p. 1533-1550.
31. Achaz, G., et al., *The reproducibility of adaptation in the light of experimental evolution with whole genome sequencing.* Adv Exp Med Biol, 2014. **781**: p. 211-31.
32. Lobkovsky, A. and E. Koonin, *Replaying the Tape of Life: Quantification of the Predictability of Evolution.* Frontiers in Genetics, 2012. **3**.
33. Palmer, M.E., A. Moudgil, and M.W. Feldman, *Long-term evolution is surprisingly predictable in lattice proteins.* Journal of The Royal Society Interface, 2013. **10**(82): p. 20130026.
34. Szendro Ivan, G., et al., *Predictability of evolution depends nonmonotonically on population size.* Proceedings of the National Academy of Sciences, 2013. **110**(2): p. 571-576.
35. de Visser, J.A.G.M. and J. Krug, *Empirical fitness landscapes and the predictability of evolution.* Nature Reviews Genetics, 2014. **15**(7): p. 480-490.
36. Karmakar, M., et al., *Structure guided prediction of Pyrazinamide resistance mutations in pncA.* Scientific Reports, 2020. **10**(1): p. 1875.
37. Portelli, S., et al., *Prediction of rifampicin resistance beyond the RRDR using structure-based machine learning approaches.* Scientific Reports, 2020. **10**(1): p. 18120.
38. Rodrigues, J.V., et al., *Biophysical principles predict fitness landscapes of drug resistance.* Proceedings of the National Academy of Sciences, 2016. **113**(11): p. E1470.
39. Barlow, K.A., et al., *Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation.* The Journal of Physical Chemistry B, 2018. **122**(21): p. 5389-5399.
40. Alford, R.A.-O., et al., *The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J. Chem Theory Comput, 2017. **13**: p. 3031-3048.

41. Aldeghi, M., V. Gapsys, and B.L. de Groot, *Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation.* ACS central science, 2018. **4**(12): p. 1708-1718.
42. Aldeghi, M., V. Gapsys, and B.L. de Groot, *Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches.* ACS Central Science, 2019. **5**(8): p. 1468-1474.
43. Eccleston, R.C., D.D. Pollock, and R.A. Goldstein, *Selection for cooperativity causes epistasis predominately between native contacts and enables epistasis-based structure reconstruction.* Proceedings of the National Academy of Sciences, 2021. **118**(16): p. e2010057118.
44. Pollock, D.D., G. Thiltgen, and R.A. Goldstein, *Amino acid coevolution induces an evolutionary Stokes shift.* Proceedings of the National Academy of Sciences, 2012. **109**(21): p. E1352.
45. Pollock, D.D., et al., *Mechanistic Models of Protein Evolution*, in *Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts*, P. Pontarotti, Editor. 2017, Springer International Publishing: Cham. p. 277-296.
46. Gerrish, P.J. and R.E. Lenski, *The fate of competing beneficial mutations in an asexual population.* Genetica, 1998. **102**(0): p. 127.
47. Jain, K., J. Krug, and S.-C. Park, *EVOLUTIONARY ADVANTAGE OF SMALL POPULATIONS ON COMPLEX FITNESS LANDSCAPES.* Evolution, 2011. **65**(7): p. 1945-1955.
48. Jett, C., A. Dia, and I.H. Cheeseman, *Rapid emergence of clonal interference during malaria parasite cultivation.* bioRxiv, 2020: p. 2020.03.04.977165.
49. Campos, P.R.A., C. Adami, and C.O. Wilke, *Modelling Stochastic Clonal Interference*, in *Modelling in Molecular Biology*, G. Ciobanu and G. Rozenberg, Editors. 2004, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 21-38.
50. Wang, X., G. Minasov, and B.K. Shoichet, *Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs.* J Mol Biol, 2002. **320**(1): p. 85-95.
51. Vogl, T., et al., *Thermodynamic Stability of Annexin V E17G: Equilibrium Parameters from an Irreversible Unfolding Reaction.* Biochemistry, 1997. **36**(7): p. 1657-1668.
52. Ruvinov, S., et al., *Engineering the Independent Folding of the Subtilisin BPN' Prodomain: Analysis of Two-State Folding versus Protein Stability.* Biochemistry, 1997. **36**(34): p. 10414-10421.
53. Taverna, D.M. and R.A. Goldstein, *Why are proteins marginally stable?* Proteins: Structure, Function, and Bioinformatics, 2002. **46**(1): p. 105-109.
54. Goldstein, R.A., *The evolution and evolutionary consequences of marginal thermostability in proteins.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(5): p. 1396-1407.
55. Gould, S.J., et al., *The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme.* Proceedings of the Royal Society of London. Series B. Biological Sciences, 1979. **205**(1161): p. 581-598.
56. Závodszky, P., et al., *Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins.* Proceedings of the National Academy of Sciences, 1998. **95**(13): p. 7406.
57. Tsou, C.L., *Active site flexibility in enzyme catalysis.* Ann N Y Acad Sci, 1998. **864**: p. 1-8.
58. Björkman, A. and P.A. Phillips-Howard, *The epidemiology of drug-resistant malaria.* Transactions of The Royal Society of Tropical Medicine and Hygiene, 1990. **84**(2): p. 177-180.
59. Roper, C., et al., *Intercontinental spread of pyrimethamine-resistant malaria.* Science, 2004. **305**(5687): p. 1124.

1011 60. Mita, T., et al., *Independent Evolution of Pyrimethamine Resistance in Plasmodium*
1012 *falciparum Isolates in Melanesia.* Antimicrobial Agents and Chemotherapy, 2007. **51**(3):
1013 p. 1071-1077.
1014 61. Maïga, O., et al., *A Shared Asian Origin of the Triple-Mutant dhfr Allele in Plasmodium*
1015 *falciparum from Sites across Africa.* The Journal of Infectious Diseases, 2007. **196**(1): p.
1016 165-172.
1017 62. McCollum Andrea, M., et al., *Common Origin and Fixation of Plasmodium falciparum*
1018 *dhfr and dhps Mutations Associated with Sulfadoxine-Pyrimethamine Resistance in a*
1019 *Low-Transmission Area in South America.* Antimicrobial Agents and Chemotherapy,
1020 2007. **51**(6): p. 2085-2091.
1021 63. McCollum Andrea, M., et al., *Hitchhiking and Selective Sweeps of Plasmodium*
1022 *falciparum Sulfadoxine and Pyrimethamine Resistance Alleles in a Population from*
1023 *Central Africa.* Antimicrobial Agents and Chemotherapy, 2008. **52**(11): p. 4089-4097.
1024 64. Turkiewicz, A., et al., *Genetic diversity of the Plasmodium falciparum GTP-cyclohydrolase*
1025 *1, dihydrofolate reductase and dihydropteroate synthetase genes reveals new insights*
1026 *into sulfadoxine-pyrimethamine antimalarial drug resistance.* PLoS Genet, 2020. **16**(12):
1027 p. e1009268.
1028 65. Ravenhall, M., et al., *Characterizing the impact of sustained sulfadoxine/pyrimethamine*
1029 *use upon the Plasmodium falciparum population in Malawi.* Malar J, 2016. **15**(1): p. 575.
1030 66. Verdrager, J., *Epidemiology of the emergence and spread of drug-resistant falciparum*
1031 *malaria in South-East Asia and Australasia.* J Trop Med Hyg, 1986. **89**(6): p. 277-89.
1032 67. Espinal, T.C.A., et al., *Sensitivity of Plasmodium Falciparum to Antimalarial Drugs in*
1033 *Colombia.* The American Journal of Tropical Medicine and Hygiene, 1985. **34**(4): p. 675-
1034 680.
1035 68. Zhou, Z., et al., *Decline in sulfadoxine-pyrimethamine-resistant alleles after change in*
1036 *drug policy in the Amazon region of Peru.* Antimicrob Agents Chemother, 2008. **52**(2): p.
1037 739-41.
1038 69. Gregson, A. and C.V. Plowe, *Mechanisms of Resistance of Malaria Parasites to*
1039 *Antifolates.* Pharmacological Reviews, 2005. **57**(1): p. 117.
1040 70. Hawkins, V.N., et al., *Multiple origins of resistance-conferring mutations in Plasmodium*
1041 *vivax dihydrofolate reductase.* Malaria Journal, 2008. **7**(1): p. 72.
1042 71. Shaukat, A., et al., *Phylogenetic analysis suggests single and multiple origins of*
1043 *dihydrofolate reductase mutations in Plasmodium vivax.* Acta Trop, 2021. **215**: p.
1044 105821.
1045 72. Webb, B. and A. Sali, *Comparative Protein Structure Modeling Using MODELLER.* Curr
1046 Protoc Bioinformatics, 2016. **54**: p. 5.6.1-5.6.37.
1047 73. Ahouidi, A., et al., *An open dataset of Plasmodium falciparum genome variation in 7,000*
1048 *worldwide samples.* Wellcome Open Res, 2021. **6**: p. 42.
1049 74. Adam, I., et al., *An open dataset of Plasmodium vivax genome variation in 1,895*
1050 *worldwide samples.* Wellcome Open Res, 2022. **7**: p. 136.
1051 75. Benavente, E.D., et al., *Distinctive genetic structure and selection patterns in*
1052 *Plasmodium vivax from South Asia and East Africa.* Nat Commun, 2021. **12**(1): p. 3160.
1053 76. Cingolani, P., et al., *A program for annotating and predicting the effects of single*
1054 *nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster*
1055 *strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.

1056
1057 **Figures and Tables**
1058
1059 **Table 1. Correlation between Flex-ddG predictions for 150 runs and experimental data for**
1060 ***Pf*DHFR pyrimethamine resistance mutations**

Mutation	$\Delta\Delta G_{exp}$*(kcal/mol)	Exp. Sum**	Exp. I.E.***	$\Delta\Delta G_{Flex\ ddG}$† (kcal/mol)	Sum‡	I.E.§
N51I	-0.783			-0.065		
C59R	-0.184			-0.055		
S108N	1.297			0.275		
I164L	-0.351			-0.65		
N51I,S108N	1.89	0.514	1.376	-0.13	0.21	-0.34
C59R,S108N	2.29	1.113	1.177	0.575	0.22	0.355
N51I,C59R,S108N	2.595	0.33	2.265	0.225	0.155	0.07
C59R,S108N,I164L	3.283	0.762	2.521	0.025	-0.43	0.455
N51I,C59R,S108N,I164L	3.761	-0.021	3.782	0.425	-0.495	0.92
Pearson correlation				0.635	0.435	0.765
Correctly classified				8/9	4/5	4/5

1061
1062 *Experimentally measured *Pf*DHFR-pyrimethamine binding free energy change data from [12]
1063 **Sum of experimental values of binding free energy change for independent mutations
1064 ***Interaction energy calculated as the difference between experimentally measured values of
1065 binding free energy change of multiple mutant compared to the sum of the independent mutations
1066 involved.
1067 † Change in *Pf*DHFR-pyrimethamine binding free energy predicted by Flex ddG calculated as the
1068 average of the distribution of runs. Free energy predictions from Rosetta are in Rosetta Energy
1069 Units, however the authors of Flex ddG applied a generalized additive model to reweight the
1070 predictions and make the output more comparable to units of kcal/mol [39].
1071 ‡Sum of Flex ddG predictions for independent mutations
1072 §Interaction energy calculated as the difference between Flex ddG predicted binding free energy
1073 change of multiple mutant compared to the sum of the independent mutations.
1074
1075

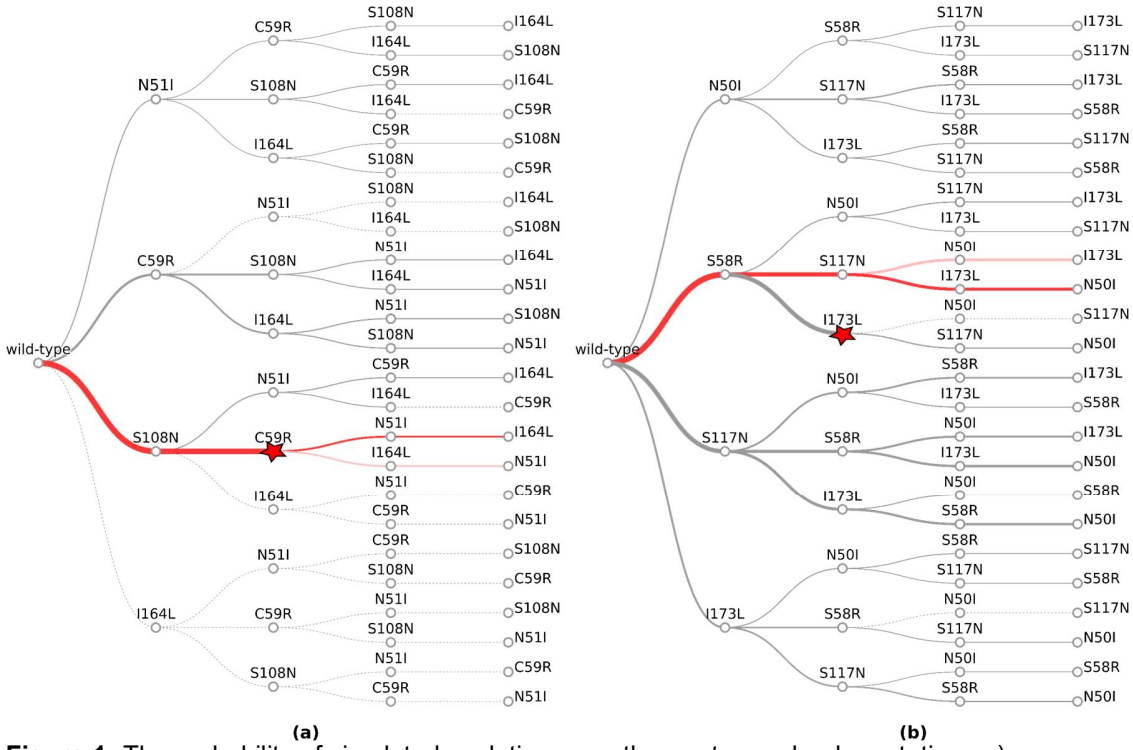
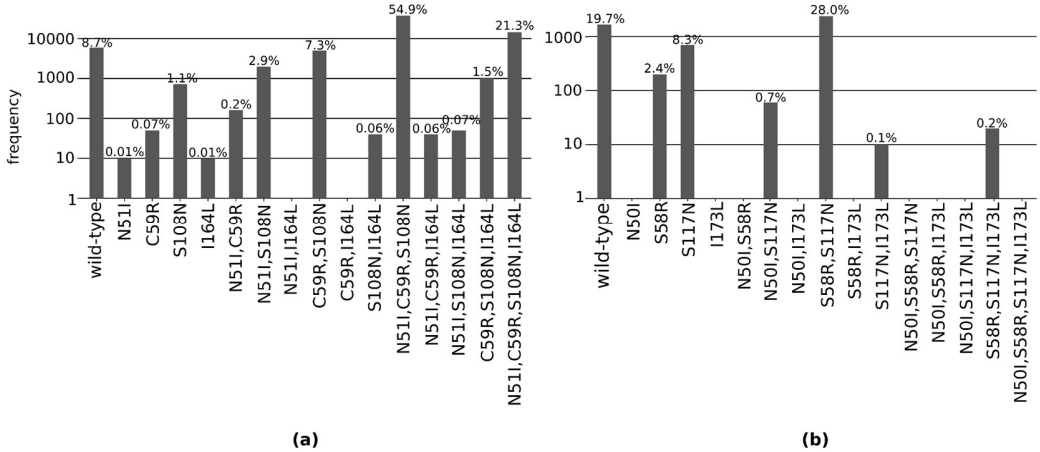


Figure 1: The probability of simulated evolutionary pathways to quadruple mutations a) N51I,C59R,S108N,I164L in *Pf*DHFR and b) N50I,S58R,S117N,I173L in *Pv*DHFR. Line thickness indicates the total probability of a mutation when considering all pathways it can occur in, determined from the frequency of that step in all realized mutational pathways from all runs. Dotted lines indicate zero probability of a mutation at that step. The most likely pathway in total is denoted by a red star. The most likely pathway to the quadruple mutation is highlighted in dark red and the second most likely pathway to the quadruple mutation is highlighted in light red. The probabilities corresponding to these plots can be found in Supplementary files 'PfDHFR_total_pathway_probabilities.csv' and 'PvDHFR_total_pathway_probabilities.csv' for a) and b), respectively.



(a) (b)

Figure 2: The total frequency of the combinations of mutations found in our isolate data for sets of four mutations a) N51I, C59R, S108N and I164L in *Pf*DHFR, and b) N50I, S58R, S117N and I173L in *Pv*DHFR. The frequencies are also given as the percentage of the total number of isolates, which for *Pf*DHFR is 6762 and *Pv*DHFR is 847.

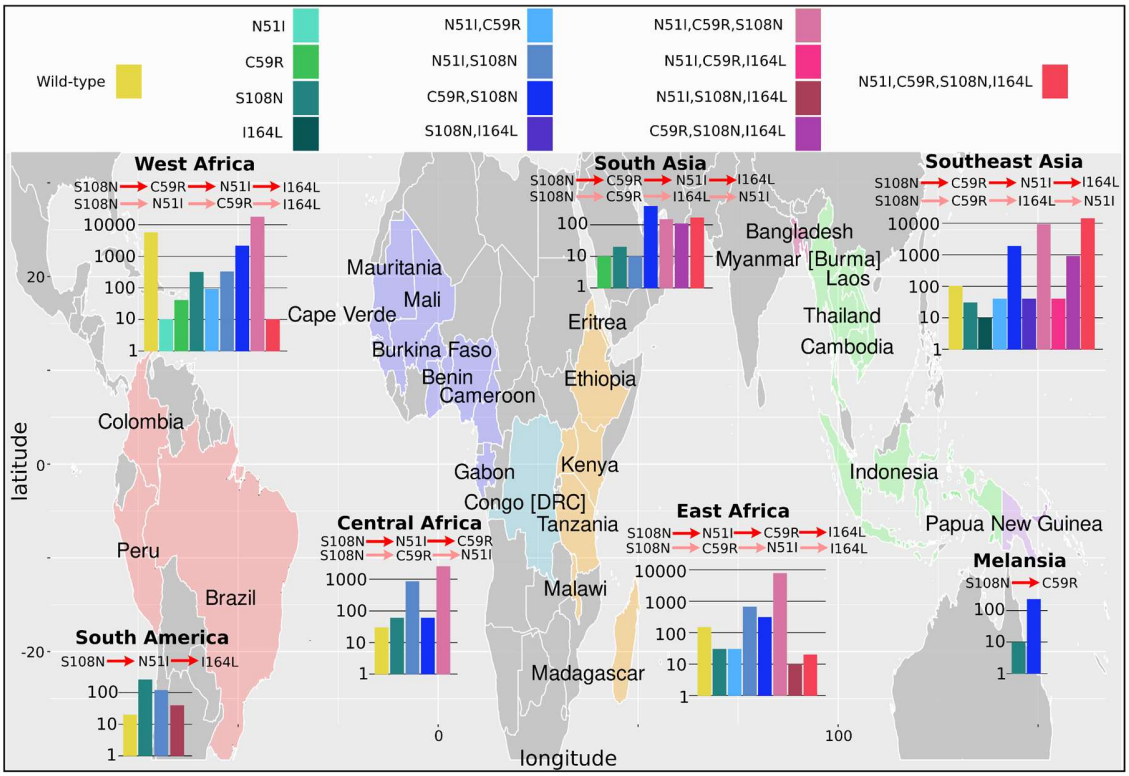


Figure 3. The *Pf*DHFR isolate data was grouped into seven geographical areas: South America, West Africa, Central Africa, East Africa, South Asia, Southeast Asia and Melansia. The bar charts display the frequency (log scale) of the combinations of the four mutations N51I, C59R, S108N and I164L. The frequency data has been multiplied by a factor of 10 to enable clear identification of those mutations occurring in one isolate only. The most likely evolutionary trajectory inferred from the frequency of combinations are included above the corresponding frequency chart from which the pathways were inferred indicated by mutations separated by dark red arrows. Alternative pathways are indicated by mutations separated by light red arrows. Where only single mutations are present a pathway is not inferred. (See Supplementary Text for further explanation and see Supplementary data folder 'PfDHFR/IsolateMutationFrequency' for the frequency of all mutations found in the isolate data from these regions).

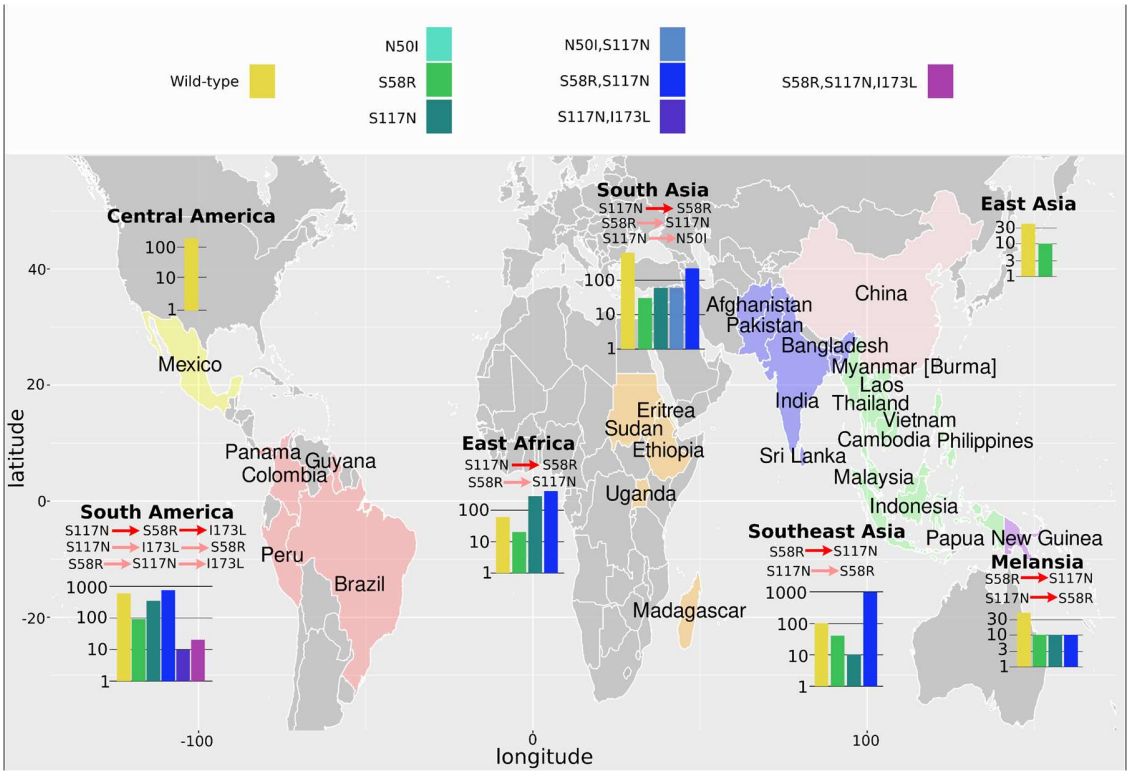


Figure 4. The *Pv*DHFR isolate data was grouped into seven geographical areas: Central America, South America, East Africa, South Asia, East Asia, Southeast Asia and Melansia. The bar charts display the frequency (log scale) of the combinations of the four mutations N50I, S58R, S117N and I173L. The frequency data has been multiplied by a factor of 10 to enable clear identification of those mutations occurring in one isolate only. The most likely evolutionary trajectory inferred from the frequency of combinations are included above the corresponding frequency chart from which the pathways were inferred indicated by mutations separated by dark red arrows. Alternative pathways are indicated by mutations separated by light red arrows. Where only single mutations are present a pathway is not inferred. (See Supplementary Text for further explanation and see Supplementary data folder 'PvDHFR/IsolateMutationFrequency' for the frequency of all mutations found in the isolate data from these regions).