Predicting antifolate resistance in the unculturable fungal pathogen *Pneumocystis jirovecii*

Authors:
Francois D. Rouleau[1,2,3,4,*], Alexandre K Dubé[1,2,3,4], Alicia Pageau[1,2,3,4], Lyne Désautels[5], Philippe J. Dufresne[5] and Christian R Landry[1,2,3,4]

Affiliations:

1) Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, QC, Canada
2) Département de Biochimie, Microbiologie et de Bioinformatique, Faculté des Sciences et de Génie, Université Laval, Québec, QC, Canada
3) PROTEO - Regroupement Québécois de Recherche sur la Fonction, l'Ingénierie et les Applications des Protéines, QC, Canada
4) Centre de Recherche en Données Massives (CRDM_UL), Québec, QC, Canada
5) Laboratoire de santé publique du Québec (LSPQ), Institut national de santé publique du Québec (INSPQ), Sainte-Anne-de-Bellevue, QC, Canada

*Author for correspondence:
Francois D. Rouleau: Francois.rouleau.2@ulaval.ca

Author contributions:

Contributor Role	Role Definition
Conceptualization	Francois D. Rouleau, Alexandre K Dubé, Christian R. Landry
Data Curation	Francois D. Rouleau
Formal Analysis	Francois D. Rouleau, Alicia Pageau
Funding Acquisition	Christian R Landry, Philippe J. Dufresne, Francois D. Rouleau
Investigation	Francois D. Rouleau, Lyne Désautels, Alexandre K Dubé, Christian R Landry
Methodology	Francois D. Rouleau, Alexandre K Dubé, Christian R Landry
Project Administration	Francois D. Rouleau, Philippe J. Dufresne, Alexandre K Dubé, Christian R Landry
Resources	Philippe Dufresne, Christian R Landry
Software	Francois D. Rouleau, Alicia Pageau
Supervision	Alexandre K Dubé, Philippe J. Dufresne, Christian R Landry
Validation	Francois D. Rouleau, Alexandre K Dubé, Christian R Landry
Visualization	Francois D. Rouleau
Writing – Original Draft Preparation	Francois D. Rouleau

Writing – Review and Editing	Francois D. Rouleau, Alexandre K Dubé, Alicia Pageau, Lyne Desautels, Philippe J. Dufresne, Christian R Landry

Abstract (293/300 words)

Pneumocystis jirovecii is a fungal pathogen causing Pneumocystis pneumonia in humans, mainly in immunocompromised individuals. Infections by *P. jirovecii* are treated using the antifolate combination drug trimethoprim-sulfamethoxazole (TMP-SMX), targeting the dihydrofolate reductase (DHFR) and the dihydropteroate synthase (DHPS). In recent years, there has been an increase of treatment failure, with no mutations observed in the DHPS, implying the potential evolution of resistance through this pathogen's DHFR (PjDHFR). Experimental methods to study this pathogen are limited, as it cannot be grown *in vitro*. Model fungi are insensitive to TMP-SMX due to unknown mechanisms, preventing the use of functional complementation to study mutations causing resistance to this specific drug combination. In a previous study, we conducted deep mutational scanning (DMS) on PjDHFR to identify resistance mutations to methotrexate (MTX), another antifolate drug. Here, by leveraging this data, as well as computational data modeling aspects of protein function and stability in the PjDHFR-MTX complex, we train a machine learning model to predict the effect of mutations on MTX resistance. We find that the model can predict the effect of mutations outside of its training dataset (balanced accuracy on training set: 98.3%, and 88.3% on testing set). We also find that the best predictors of resistance, such as distance to ligand and effect on region flexibility, are coherent with previously established models, and that experimental data about the effect of mutations on protein function is critical to optimize model performance. Using this model on computational data generated using the PjDHFR-TMP complex, we predict the effect of mutations on resistance to TMP. We predict TMP resistance mutations in PjDHFR that did not confer resistance to MTX, one of which had been characterized *in vitro* as reducing affinity to TMP by 100-folds. We compare the predictions from this model to PjDHFR sequences from previously and newly sequenced clinical samples. Our results offer a resource to interpret the impact of amino acid variants in PjDHFR on TMP resistance, as well as methods to predict resistance in hard-to-study organisms.

Short title: Predicting antifolate resistance in an unculturable fungus

Author summary (200/200):

Pneumocystis jirovecii is a fungal pathogen causing pneumonia in immunocompromised humans. Infections by *P. jirovecii* are treated using drugs that prevent this pathogen from making folate, an essential component of many cellular mechanisms. In recent years, this treatment has been failing in an increasing number of cases, implying the evolution of resistance to this treatment. As *P. jirovecii* does not grow in the lab, the investigation of this resistance has been difficult, and common lab models do not respond to the drugs used to treat it. To overcome these limitations, we use a combination of experimental data and computer modeling to train a machine learning model to predict how genetic changes in one of the drug targets might cause drug resistance in this pathogen. The presented model predicts mutations in the drug target that may make this pathogen resistant to treatment, including mutations that have been previously characterized *in vitro* as drastically reducing drug binding. To investigate if our model predicted mutations that accrued in nature, we also sequenced the largest number of this pathogen's drug target to date. Our study provides new tools to predict drug resistance in hard-to-study pathogens, helping to understand and potentially respond to treatment failure.

Introduction

The number of human pathogens is estimated to be around 1400 species, with approximately 300 being fungi [1,2]. One such pathogen is *Pneumocystis jirovecii*, a human-specific obligate parasitic commensal, and opportunistic pathogen that causes Pneumocystis pneumonia (PCP) in vulnerable patients [3]. In healthy patients, *P. jirovecii* usually does not cause PCP, and rather leads to asymptomatic infection and transient colonization. Environmental exposure to *P. jirovecii* is frequent, with reports from some countries indicating it can be present in up to 40% of individuals in a population [4–6]. This pathogenic fungus has also been known to cause PCP in HIV-negative infants [7]. Historically, PCP has been most prevalent in HIV-positive individuals, but with the advent of efficient antiretroviral therapies, its occurrence has been decreasing [8]. However, infection rates have been increasing in immunocompromised, non-HIV populations [9,10]. *P. jirovecii* has recently been placed on the WHO fungal pathogen priority list [11].

Research on *P. jirovecii* has been historically difficult, as this pathogen has not been reproducibly culturable *in vitro* [12]. *Pneumocystis* species are not known to have an ecological niche outside of mammal lung epithelial tissues, and all known species are host-specific, with no known reports of transmission of a given *Pneumocystis sp.* outside of its associated niche [13,14]. *P. jirovecii* has the peculiarity of being insensitive to most classical antifungals, as they have not been found to efficiently treat PCP [15]. As such, the antifolate combination drug Trimethoprim-Sulfamethoxazole (TMP-SMX, Bactrim), is used as the mainstay drug for both prophylaxis and treatment against PCP. These drugs target the dihydrofolate reductase (DHFR) and the dihydropteroate synthase (DHPS), respectively [16,17]. In the last two decades, instances of TMP-SMX treatment failure have been increasing in PCP cases, implying the evolution of resistance in this fungus [18–23]. There is therefore a need to investigate mechanisms of resistance, and to gain insight into the mode of action of this drug combination in fungal infections.

Historically, treatment failure has been associated with mutations in the DHPS enzyme, the target of SMX. However, an increasing number of studies have also reported mutations in the DHFR when failure of TMP-SMX treatment is observed [24–27]. Furthermore, reports from eastern Canada, where 265 DHPS from clinical samples have been sequenced, only found DHPS mutations associated with treatment failure in three samples (~1.5%), with those mutations very rarely observed in patient with TMP-SMX treatment failure as well, suggesting a greater potential contribution from the DHFR to resistance than previously estimated, or the presence of a different mechanism of resistance (data available at Laboratoire de santé publique du Québec (LSPQ)).

As there are currently no susceptibility assays available for *Pneumocystis jirovecii*, our ability to study the evolution of resistance in this pathogen is limited to biochemical assays *in vitro* and *in silico* approaches [12]. An early study focused on testing *in vitro* biochemical constants, such as the inhibitory constant (K_i), the Michaelis constant (K_m)

and the catalytic constant (k_{cat}) of protein variants that had been identified in clinical studies in patient samples where TMP-SMX treatment had failed. Their experiments identified that these variants had decreased affinity for TMP. However, the effect of mutations on the protein canonical function *in vivo* could not be validated, limiting the broader conclusions of this study [24]. Another study focused on using computational tools to identify the effects of mutations on resistance. Using molecular dynamics to model the effects of mutants characterized in [24], the authors used a machine learning model to classify the mutations in sensitive/resistant classes. The resulting model accurately classified variants in ~65% of cases and relied on the two most predictive features identified through principal component analysis. However, their conclusions were limited by the small amount of *in vivo* information available, as well as computational constraints, since extensive molecular dynamics were necessary for each variant tested. This only allowed them to focus on 19 PjDHFR mutants, and again without knowledge on the effect of mutations on protein function, limiting the broader conclusions of this study [28].

Given the current limitations, there is a need to find ways to test mutational effects, ideally in a context where the impact of mutations on DHFR function can also be assayed. We recently used functional complementation of DHFR in *S. cerevisiae* to systematically investigate the effect of mutations on resistance to antifolates through mutations in *P. jirovecii*'s DHFR (PjDHFR) [29]. Because *S. cerevisiae* is not sensitive to TMP, we could not investigate the effects of mutations in PjDHFR on TMP resistance *in vivo*. However, we used deep mutational scanning (DMS) of PjDHFR to identify mutations that caused resistance to another similar competitively inhibiting antifolate molecule, methotrexate (MTX). We reasoned that if we could use computational tools to predict MTX resistance accurately using this experimental data as ground truth, we could use the same tools to predict resistance to TMP. Indeed, as both MTX and TMP are antifolate competitive inhibitors of the same protein, and share molecular properties and binding site, we expect that resistance mechanisms should share characteristics as well [30–32].

Here, we leverage our previous DMS data on PjDHFR's resistance to MTX to train a machine learning model to predict resistance from the protein structure and properties. We use a suite of molecular features in the model, including an *in silico* analysis of the biophysical impacts of mutations on proteins [33–35]. One important piece of data for making these predictions are the effects of mutations on the canonical function of PjDHFR as mutations that appear to be susceptible to MTX or that could be predicted to decrease interaction affinity between the protein and MTX could in fact be inactivating or destabilizing the enzyme. We therefore performed an additional set of experiments to estimate the impact of all amino acid substitutions in PjDHFR on its ability to complement the *S. cerevisiae* DHFR. Then, using the model trained to predict resistance to MTX and our experimental data as a reference data set, we use the model to predict resistance to TMP. Our computational and experimental approaches are outlined in Figure 1 [33–35].

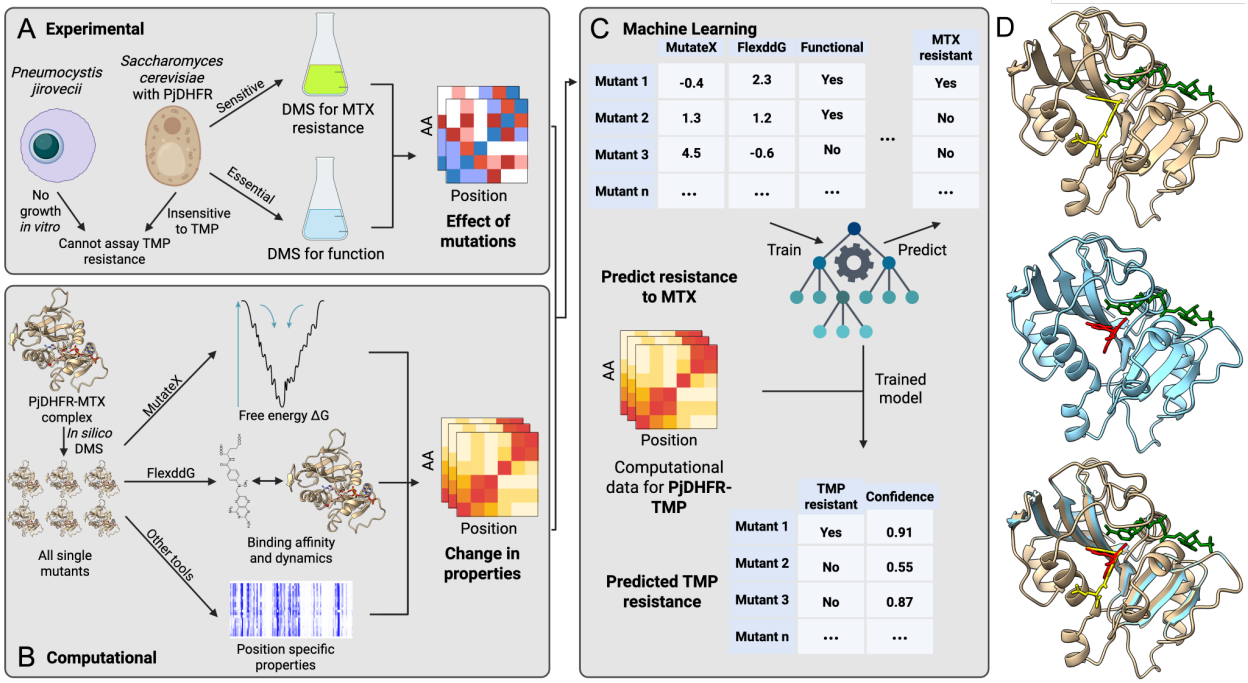


Figure 1: General workflow of the approach to predict drug resistance to TMP in *Pneumocystis jirovecii*. A) Using *S. cerevisiae* strains where the DHFR deletion was functionally complemented with the *P. jirovecii* ortholog (PjDHFR), we measured the impact of all amino acid substitutions on both function and MTX resistance. This is used downstream as ground truth when training the model. B) Using PjDHFR-MTX 3D structure from [36] (AlphaFold model UniProt A0EPZ9), we used modeling suites to conduct *in silico* DMS to estimate the effect of mutations in different ligand-dependent and independent effects. These data were used to train a machine learning model to predict MTX resistance based on the different mutation properties. C) The model was used on similar data generated using the PjDHFR-TMP complex, and predictions were made on the effect of mutations in PjDHFR resistance to TMP, which cannot be tested experimentally. D) Structures for the PjDHFR-MTX complex (top, PjDHFR in beige, MTX in yellow and NADPH in green, PBD: 3CD2), PjDHFR-TMP complex (middle, PjDHFR in blue, TMP in red and NADPH in green, PDB: 1DYR), and superimposition of both structures (bottom). Panel A, B and C were created in BioRender. 1, U. (2025) https://BioRender.com/2ja2o1o

Results

DMS of PjDHFR reveals functional constraints and conserved LOF hotspots

To measure the effect of mutations on protein function (Fig 1A), we screened every possible single amino acid mutant in PjDHFR in a temperature sensitive (TS) DHFR constructions (TS-DHFR), in four independent replicates (S1 fig.) We transformed the mutant library built on plasmids in each replicate and grew each for approximately 6 generations. The screen was conducted in two conditions: 20 °C was set as the permissive condition, and 37 °C as the selection condition for the TS-DHFR strains, where the genomic DHFR should be fully inactivated. The 20 °C condition was selected to minimize the effect of heat shock and incubation on TS-DHFR strain growth, and we expected small effects from the PjDHFR library in this condition.

Mutation effects were first measured on a per-codon basis, and correlation between replicates and fragment overlaps was evaluated by codon, then normalized by the fitness of the wild-type sequence (S2A, S2B fig., Eq. 1). We found that selection coefficients were correlated between fragment overlaps in both temperature conditions (Pearson's $0.91 < r < 0.98$, mean of 0.94). Once we observed that the replicates were correlated in the selection condition (S2 fig., Pearson's $0.84 < r < 0.94$, mean of 0.88), codons were grouped by amino acid at each position across replicates, with the median score per amino acid substitution used as selection coefficient for subsequent analyses (Fig 2A). The resulting score, now standardized between fragments and replicates, was used as the selection coefficient for subsequent analyses. A heatmap of the selection coefficient for the screen at the permissive temperature (20 °C) for TS-DHFR strains is available in S3 fig. At 37 °C, we observed that most mutations had either a neutral or deleterious effect on protein function (selection coefficient of <= 0.0), and that most positions where several mutations led to deleterious effects were in contact with one or both DHFR's canonical substrates, folate (DHF) and nicotinamide adenine dinucleotide phosphate (NADPH) (Fig 2A).

We observed a correlation between the selection coefficient from the MTX resistance screen, and selection coefficients for function measured at 20 °C (Spearman's $\rho = 0.44$, $p < 1e\text{-}100$, Fig 2B) and 37 °C (Spearman's $\rho = 0.52$, $p < 1e\text{-}100$, Fig 2C). Mutants behaved similarly on both temperature conditions (Spearman's $\rho = 0.78$, $p < 1e\text{-}100$, Fig 2D), implying that even the low temperature transformation protocol and incubation at 20 °C caused fitness defects in the TS-DHFR strains. However, as all biological, codon and fragment overlap replicates are well correlated, we expect that this effect had minimal consequences on selection coefficient in selection conditions (S2 fig.) After grouping the data from both assays, we covered 3758/3914 (96.0%) of all possible amino acid substitutions with selection coefficients for both MTX resistance and functionality.

Most mutations had little to no detectable effect on protein function. Indeed, 3140/3758 amino acid substitutions (83.5% of covered mutations) were grouped as being wild-type-like in the gaussian mixture model (selection coefficient >= -0.124 at 37 °C, S4A fig.), and therefore having a neutral effect on protein function at this expression level and experimental condition. All tested positions had at least one substitution with a neutral effect, except position G125, where all substitutions were deleterious. Across 204 amino acid positions, there were 71 positions where all substitutions had a selection coefficient >= -0.124, indicating that all mutations at that position had a neutral effect and that these positions are highly tolerant to amino acid substitutions. Overall, the median number of neutral substitutions per position was 18/19, indicating that most positions are tolerant to amino acid substitutions.

We observed 583 amino acid substitutions as having deleterious effects with a high degree of confidence (14.2% of substitutions, selection coefficient < -0.185, p-value < 0.05, Benjamini-Hochberg (BH) corrected, S4 fig.), causing large growth defects when grown at 37 °C in the TS-DHFR backgrounds. We define these amino acid substitutions as loss-of-function (LOF) mutations (Fig 2A-C, mutants below dotted black lines). Most positions with such strong deleterious effects often had many mutations leading to LOF, with very few positions where only one mutation caused strongly deleterious effects, except in the case of substitutions towards proline. Interestingly, 39.9% LOF mutations are found within a radius of 9Å of each other (S5 fig.) This structured region, composed of a β-sheet and an α helix from positions 54 to 64 (β-strand into α-helix), 75 to 81 (β-strand) and 121 to 126 (β-strand), is in close proximity to both the folate and NADPH binding pockets.

To support these experimental observations, we computed GEMME scores, which is a score used to predict mutational effects based on sequence conservation across orthologs, where a lower score represents a substitution that is more likely to be deleterious [37]. This region has statistically significantly more negative GEMME scores than the rest of the protein (Mann-Whitney U rank test, $p < 0.05$), validating it's a high degree of importance for protein function (S5 fig.) In PjDHFR in selection conditions, only 58 mutations (1.4% of amino acid mutants, -0.124 > selection coefficient > -0.185, p-value < 0.05, BH corrected) were identified as having statistically significant intermediate effects, with all other mutations being non-significantly different from either wild-type residues or stop codons. Interestingly, for most positions with low tolerance to mutations, the conservation of amino acid properties appeared to be key to retain protein function, as opposed to the presence of a specific amino acid, as shown in previous studies [38,39]. The 14.2% of mutations causing LOF in PjDHFR (206aa) is comparable to *folA*, the DHFR ortholog in *Escherichia coli* (160aa), where 12% of missense mutations lead to LOF [40], consistent with the high degree of structural conservation in orthologous DHFRs [41]. These regions are also known to be critical for function in *folA* [42]. Evolutionarily conserved positions with low tolerance to mutations, such as positions G20, R75 and

G125 in PjDHFR (G15, R57 and G96 in *folA*) are also sensitive to most amino acid substitutions in *folA* [40].

For a mutation to lead to resistance, some level of protein function must be maintained. Indeed, even if a mutation would prevent binding between PjDHFR and MTX, if that mutation also prevents protein function, it will not lead to resistance, as it would not allow for cell survival or proliferation. Here, there appears to be few mutations where a tradeoff between function and resistance is observed, as most mutations that affect function led to LOF (Spearman's ρ = 0.52 between function and resistance at 37 °C, $p < 1e\text{-}100$). However, as our thresholds for calling functional/non-functional (Function selection coefficient >= 0.185) and resistant/sensitive (MTX resistance selection coefficient >= 0.550) mutants in our experiments were set using gaussian mixture models (S4A, S4B fig.), some mutants close to either threshold could be classified as non-functional and resistant, which would imply a tradeoff. These include mutants I10D, G124C, G124D, F199D and F199G. Across the entire protein, only 11 mutations (0.2% of amino acid mutants, DMS score -0.124 >= at 37 °C and >= 0.220 in MTX condition, p-value < 0.05) resulted in a partially functional protein that showed resistance to MTX (Fig 2B-C, bottom right of graphs), with only two mutation (I10D and F199D) being identified as highly resistant when setting thresholds using gaussian mixture models. Otherwise, mutations that lead to MTX resistance appear to have little to no effect on canonical protein function in the assay based on functional complementation of *S. cerevisiae'*s DHFR. This is similar to recent results obtained when studying *ERG11*, another generally essential gene targeted by antifungal drugs, where no tradeoff was observed between function and resistance [43].

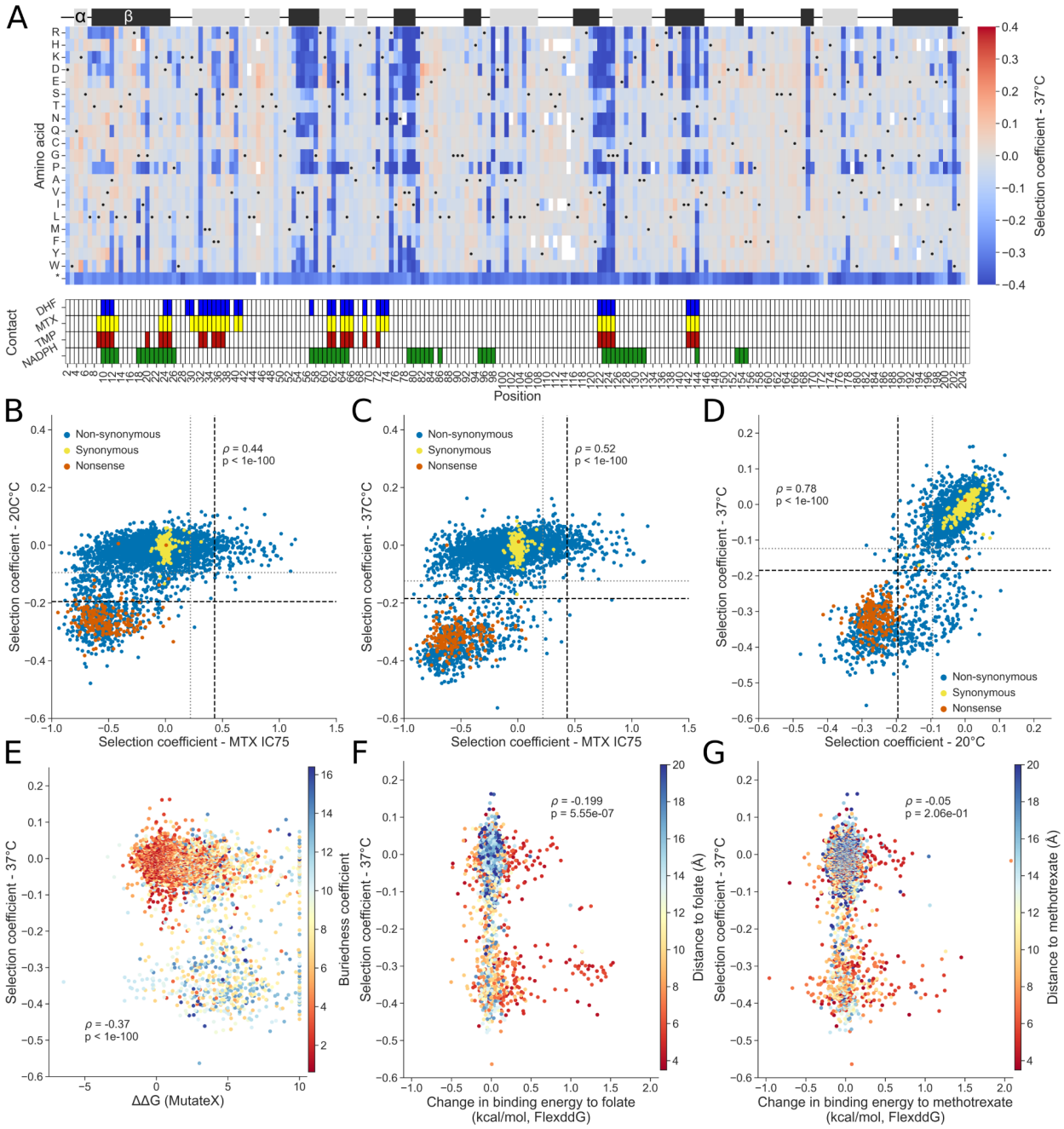


Figure 2: Effects of amino acid substitutions on PjDHFR function as assayed in *S. cerevisiae* TS-DHFR strains. A) Heatmap of selection coefficients at 37 °C, where TS-DHFR alleles are non-functional. Selection coefficient is measured in relation to the median of silent mutations. Contact was defined as amino acids with an α-carbon located less than 8 Å from MTX, DHF, TMP or NADPH (PDB: 3CD2 (MTX and NADPH) and 4CD2 (DHF)). Black dots represent the wild-type sequence. The track at the top of the figure represents secondary structures along the sequence, with α-helices in light grey and beta strands in dark grey. B) Comparison between screening done at 20 °C and MTX screening at IC75 from [36] (Spearman's ρ = 0.44, p < 1e-100). Horizontal lines are thresholds for function established from the mixture models (grey line, partially functional <= -0.09, non-functional, black line, <= -0.195) and vertical lines are thresholds from [36]. C) Comparison between screening done at 37 °C and MTX screening at IC75 (Spearman's ρ = 0.52, p < 1e-100, grey line, partially functional <= -0.124, non-functional, black line, <= -

0.185). D) Comparison between screening done at 20 °C and 37 °C (Spearman's ρ = 0.78, p < 1e-100). Grey and black lines are as in panels B and C. E) Comparison between screening done at 37 °C and computed MutateX changes in free energy for all mutants, colored by residue buriedness, (Spearman's ρ = -0.45, p < 1e-100). F) Comparison between computed FlexddG changes in interaction energy between binding energy to DHF and screening done at 37 °C, colored by proximity of α-carbon to the closest heavy atom on the ligand for residues in contact with the ligand (<8 Å, Spearman's ρ = -0.199, p = 5.55e-07)). G) Comparison between computed FlexddG changes in interaction energy between binding energy to MTX and screening done at 37 °C, colored by proximity of α-carbon to the closest heavy atom on the ligand.

Complementary modeling approaches help explain functional effects of amino acid substitutions

No single modeling method can yet recapitulate all aspects of protein dynamics and function. We can therefore expect that different modeling methods will each explain part of the experimental effects we observed during DMS screening. We mainly focused on two modeling tools which allow us to predict effects caused by mutations. The first tool, MutateX (v0.8) [32], is based on FoldX (suite 5) [43], and allows to measure the effect of a mutation on a protein's folding free energy, where mutations with positive ΔΔG are being predicted as being destabilizing, and mutations with a negative score as stabilizing. By using FoldX empirical force-fields to calculate the energy of both the wild-type and mutant structures for each amino acid mutant across the protein, as well as FoldX repair function to minimize clashes within the structure before computation, this software estimates the ΔΔG caused by mutations based on five independent simulations.

The second tool, FlexddG [33], allows to model the effect of mutations on the interaction free energy between two molecules and computes the changes caused by amino acid substitutions. It was used to compute changes in binding free energy between PjDHFR and its different substrates and inhibitors (Fig1B.) FlexddG uses a Rosetta ensemble-based method to estimate changes of binding free energy (ΔΔG between protein and ligand) by sampling structural variations caused by mutations through a "backrub" method, minimization steps and side-chain repacking [34]. For each mutation, 35 000 backrub steps are conducted on 100 individual instances to generate a statistically robust distribution of ΔΔG scores, which are then averaged to yield the final estimate of interaction free energy changes.

We observed that what correlated best with protein function was the changes in folding free energy from MutateX (Fig 2E, Spearman's ρ = -0.45, p-value < 1e-100.) Highly buried residues often had higher change in folding free energy than less buried residues and are predicted to destabilize the protein to a greater extent, as expected. Using FlexddG, we also observed a correlation between changes in binding affinity between PjDHFR and DHF, and protein function, highlighting this method's ability to capture the effects of mutations on free energy of binding. When considering only residues in contact with the ligand, we found a weak but statistically significant negative correlation where mutations with a greater destabilizing effect are more deleterious (Fig 2F, Spearman's ρ = -0.199, p-value = 5.55e-07.) This correlation was weaker, and not

statistically significant, when comparing changes in binding energy between PjDHFR and MTX with function across the protein, implying that this method can accurately capture small differences between ligand molecules (Fig 2G, Spearman's ρ = -0.05, p-value = 2.06e-1), such as a single heavy atom between DHF and MTX.

Random forest model highlights importance of ligand interactions and protein functionality in resistance prediction

From various experiments and computational analyses, we believed that we had gathered comprehensive data on different aspects of protein function, resistance, and mechanisms. If this data could be used to predict resistance to MTX using machine learning (ML), we hypothesized that the resulting model could be used to predict resistance to TMP, as they share biophysical properties and mechanisms of action (Fig 1C). Based on the gaussian mixture model from [36], mutations were classified as being either resistant (selection coefficient >= 0.550), or sensitive to MTX (selection coefficient < 0.550), and this was used as the ground truth for classification to train the ML model.

To predict resistance to MTX, we used experimental and computational features. The computational features can be split into two categories, ligand-dependent and ligand-independent. Details on how features were computed and gathered are available in the materials and methods section, and details about each feature is available in S1 Table. Ligand-dependent features are affected by the presence of a different ligand bound to the protein (DHF, MTX or TMP), and are not intrinsic properties of the protein. They include the change in binding free energy to the ligand caused by the amino acid change, the effect of mutation on region flexibility, and distance between α-carbon of a residue and the closest heavy atom on the bound ligand. Ligand-independent features are intrinsic properties of the protein or rely solely on protein sequence and include change in free folding energy (computed in the absence of ligands), residue side chain length in the wild-type protein, residue buriedness (compound measurement of solvent accessibility and distance to protein outer residues), and evolutionary constraints. Experimental features include the effect of the amino acid changes on function (selection coefficient measured in TS-DHFR strains in the absence of drugs), whether a position is involved in catalysis, and allostery, and are ligand-independent features [44]. The dataset, composed of 3758 amino acid substitutions for which data was available for each feature, was split into training and testing sets in a 80%:20% ratio with a fixed seed, and z-score normalized to put all data on a similar scale. Using the training set, different models were tested and optimized to find the best performing one (see materials and methods for details on model selection). To assess model performance, we used several evaluation tools, such as confusion matrices, receiver operating characteristic, precision-recall display (Fig 3A). We also examined the balanced accuracy score, defined as the sum of recall for all classes which allows for better estimate of model performance for unbalanced classes.

368 Overall, the best performing model was a Random Forest (RF), with balanced
369 accuracy scores of 98.3% (2907/3005 substitutions accurately classified) on the training
370 set and 88.3% (716/752) on the testing set (Fig 3A, see Model code on GitHub for detailed
371 parameters). Mutants with probability of >0.5 were predicted as being resistant (class 1),
372 and mutants with probability <= 0.5 were predicted as sensitive (class 0). Among the 111
373 amino acid changes experimentally identified as leading to MTX resistance, only 4 were
374 misclassified as sensitive by the model (false negatives, V11E, V11M, S37P and G124P).
375 Using the SHapley Additive exPlanations explainer function, we identified the relative
376 importance of features used for prediction in the model (Fig 3A) and investigated the
377 classification decision for the misclassified mutations. Mutation V11E (resistance
378 probability in MTX of 0.33, selection coefficient in MTX of 0.857, selection coefficient at
379 37 °C of 0.065) and V11M (resistance probability in MTX of 0.36, selection coefficient in
380 MTX of 0.745, selection coefficient at 37 °C of 0.015) were predicted as being
381 destabilizing, explaining most of their classification as non-resistant (S6A fig.). Mutation
382 S37P (resistance probability in MTX of 0.36, selection coefficient in MTX of 0.760,
383 selection coefficient a 37 °C of 0.039) was mischaracterized mainly because of its
384 allosteric effect score, as it was scored as having an allosteric effect with high confidence
385 in [45]. According to the model explainer, this allosteric score can have a strong effect on
386 prediction in both directions (class 0 and class 1), depending on other position-specific
387 factors. As for mutation G124P (resistance probability in MTX of 0.25, selection coefficient
388 in MTX of 0.813, selection coefficient at 37 °C of -0.325), it was identified as being non-
389 functional, which is the major factor explaining its misclassification by the model (S6A
390 fig.). Despite good overall performance, some effects could not be captured by the model.
391 The feature which contributed most to successful overall classification of mutants
392 was distance to ligand, which is coherent with results from [36], in which all but two
393 positions where multiple substitutions lead to resistance were in contact with the antifolate
394 inhibitor (Fig 2A, Fig 3A).
395 Mutant functionality was the second most important feature in prediction of
396 resistance. Across all mutations identified as being resistant to MTX, only 5 out of 111
397 were identified as being non-functional, with all 5 being close to the identified threshold
398 between gaussian distributions (Fig 3B-C). Therefore, except in few cases near the
399 margins, non-functionality is mutually exclusive to resistance. The misclassification of
400 G124P supports that the model can capture this effect in most cases, classifying non-
401 functional mutants as sensitive. When removing the experimental measures of LOFs as
402 a model feature, we observed reduced model performance, showing an increase of ~50%
403 in false positives across both sets (Balanced accuracy score of 96.9% on training set and
404 83.6% on testing set, S7A fig.). While this model was able to accurately classify most
405 mutations that did not lead to MTX resistance (true negatives), it performed strictly worse
406 overall than the model that includes experimental measurements of function. Assessing

407 the effect of mutations on function is therefore a key feature to accurately predict the
408 effect of mutations on resistance to antifolates.
409 Another important contributor of model decision was residue buriedness, where
410 most residues with several resistance mutations fell in an intermediate range of
411 buriedness, being neither deeply buried nor very exposed to solvent. Change in folding
412 free energy, where mutations with a higher absolute change in folding free energy were
413 often predicted as being sensitive was also a highly ranked feature. This is coherent with
414 the negative correlation observed with function (Fig 2E). Interquartile range (IQR) of the
415 change in distance between residue α-carbon and ligand heavy atom caused by
416 mutations, a metric computed from FlexddG output structures, was also amongst the top
417 predictive features. IQR of the change in distance is a measure of how much a given
418 mutation affects the surrounding region's flexibility, where mutations that reduced
419 flexibility were less likely to be predicted as resistance-causing. Interestingly, IQR of the
420 change in distance also had a statistically significant negative correlation with the effect
421 of mutations on function, meaning that mutations which reduce flexibility when bound to
422 inhibitors more often have a deleterious effect on function (Spearman's ρ = -0.24, p-value
423 < 1e-100, S8 fig.).
424 While binary classification of MTX resistance based on gaussian mixture models
425 was used for classification, the measurements of MTX resistance were on a continuous
426 scale. We therefore also used regression-based models to predict resistance score for
427 each mutation (S9 fig.) and can compare the decision made by the best-performing
428 classification model, and the selection coefficient predicted by regression models. Using
429 the experimentally determined threshold from the screen for MTX resistance, we can
430 consider mutants with a predicted score >0.550 as being predicted as resistant by the
431 regression model.
432 When using the simple linear regression model, no mutant is predicted as having
433 a score of >0.550, which implies that no mutation can lead to MTX resistance, according
434 to this model (R^2 = 0.185, RMSE = 0.236, Spearman's ρ = 0.326, p-value = 1.213e-93,
435 S9A fig.). When using the support vector regression (SVR) method (R^2 = 0.323, RMSE =
436 0.215, Spearman's ρ = 0.618, p-value < 1e-100, S9B fig.), predicted scores are skewed
437 towards more negative values comparatively to ground truth (35/111 predicted as having
438 a score of >0.550), indicating poor model fitting despite improvement on the linear
439 regression model. Based on the best performance of the random forest classification
440 method, we expected the RF regression to outperform other regression-based methods,
441 which is what we observed (R^2 = 0.541, RMSE = 0.177, Spearman's ρ = 0.864, p-value
442 < 1e-100, S9C fig.). When compared to the RF classification model, the regression-based
443 model predicted less false positives, but more false negatives, with most predictions being
444 close to the established classification threshold. As we believe that in the case of
445 predicting the effects of mutations on antifolate resistance, minimizing false negatives
446 should be prioritized, we focused on the RF classification model as the best performing

model. The complete dataset used to train and test the ML models on PjDHFR's resistance to MTX, as well as the classification scores for all mutants using the best performing model, is available as S1 File.

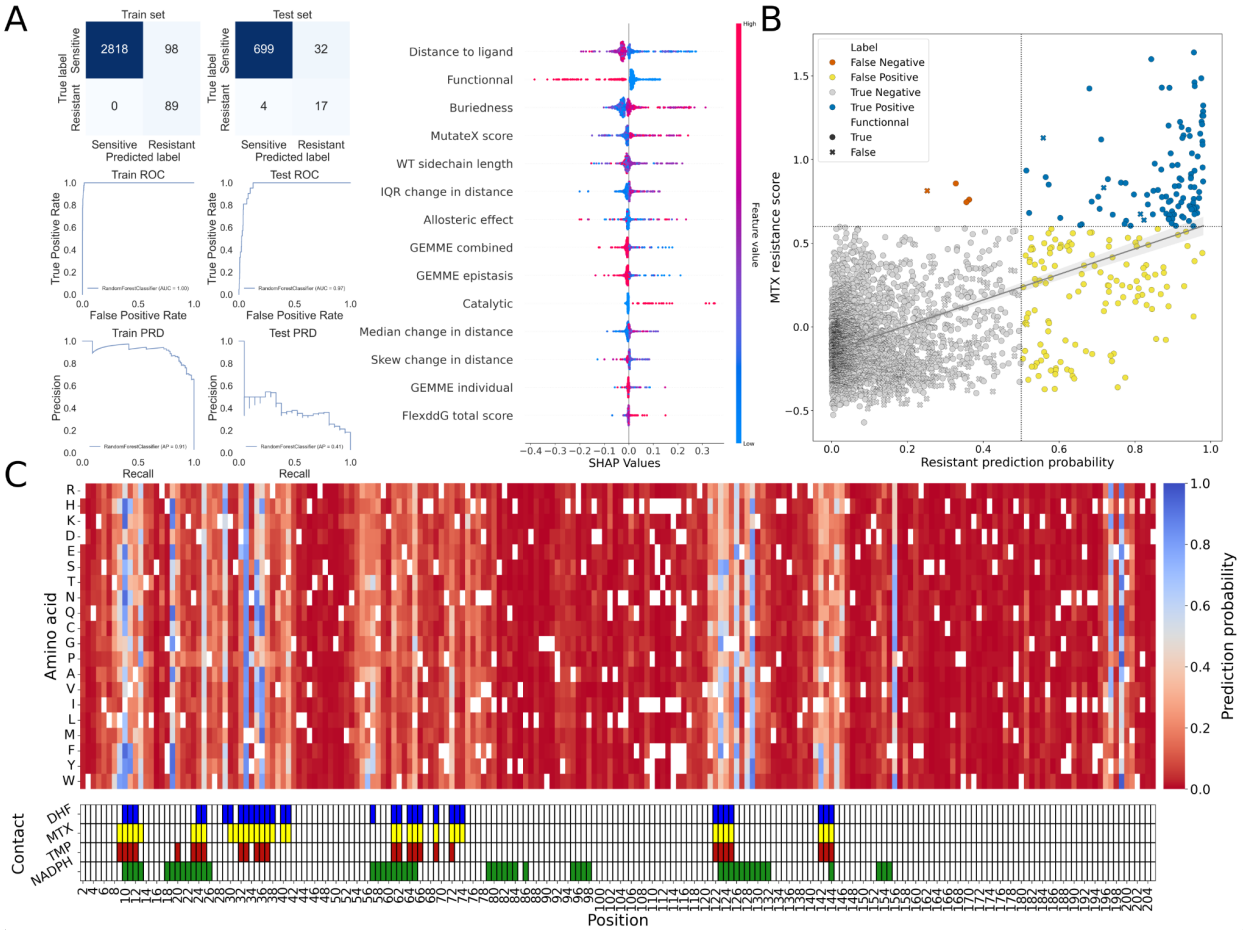


Figure 3: Visualization of model performance. A) Confusion matrices and ROC/PRD curves for train and testing set splits, with SHAP value explainer plot, with relative impact of features on model decision. B) Scatterplot comparing ground truth from MTX DMS screening vs model's predicted probability of being resistant. True negatives in grey, true positives in blue, false positives in yellow and false negatives in orange. Mutations that were identified as functional in the function DMS screening have round markers and mutations identified as non-functional have X-shaped markers. C) Heatmap with prediction probability of being resistant to MTX. Mutations with probability >0.5 are identified as resistant, and mutations <= 0.5 are classified as sensitive.

Ligand-specific features enable TMP resistance prediction from MTX-trained model

Given the adequate performance of our model for MTX resistance, we applied it to predict TMP resistance (Fig 1C, Fig 4) but this time using computational features modeled using the PjDHFR-TMP structure for predictions. As expected from feature importance during model training, mutations that are predicted to lead to resistance with

high confidence are often mutations in close proximity to TMP heavy atoms, as well as positions where several mutations lead to MTX resistance. The model predicted 213/3758 (5.67%) mutations as leading to TMP resistance, with 74/3758 (1.97%) having a prediction probability of >0.8. Of all mutations predicted as leading to TMP resistance, 89/213 (41.78%) were experimentally identified as leading to MTX resistance, and 186 (87.32%) were predicted by the original model as leading to MTX resistance. We expected to find an overlap between mutations predicted to lead to MTX and to TMP resistance, as both antifolates share structural and mechanistic properties. The model also accurately captured the effect of functionality on resistance. Only three mutants were predicted as being TMP resistant while being identified as non-functional. All three mutants, G124C (Resistance probability in TMP of 0.55, selection coefficient in MTX of 0.831, selection coefficient at 37 °C of -0.233), F199D (probability of 0.61, MTX of 0.673, 37 °C of -0.242) and F199G (probability of 0.61, MTX of 0.638, 37 °C of -0.335) are part of the 11 mutations where a tradeoff between MTX resistance and function was observed.

The model also predicted several mutations that did not lead to MTX resistance. All mutations predicted as leading to resistance are in the right range of distance to ligand and buriedness. These mutations, I10K (probability of 0.94), F36C (0.91), F121H (0.52), I123G (0.82), F156V (0.75) and Y197Q (0.81) highlight the model's ability to make novel predictions based on ligand-dependent features, which vary between the PjDHFR-MTX and PjDHFR-TMP structures. All predictions were in the right range of distance to ligand and buriedness, and subsequently driven by other parameters (S6B fig.). Of these, F36C has been observed in the clinic and biochemically characterized as reducing affinity between PjDHFR and TMP. Despite the importance of ligand-independent variables, such as buriedness and wild-type side chain length, the presented model can differentiate and make predictions based on changes in bound molecules, and this without explicit structural information. All data used to make predictions on the effect of mutations on resistance to TMP is available in S2 file.

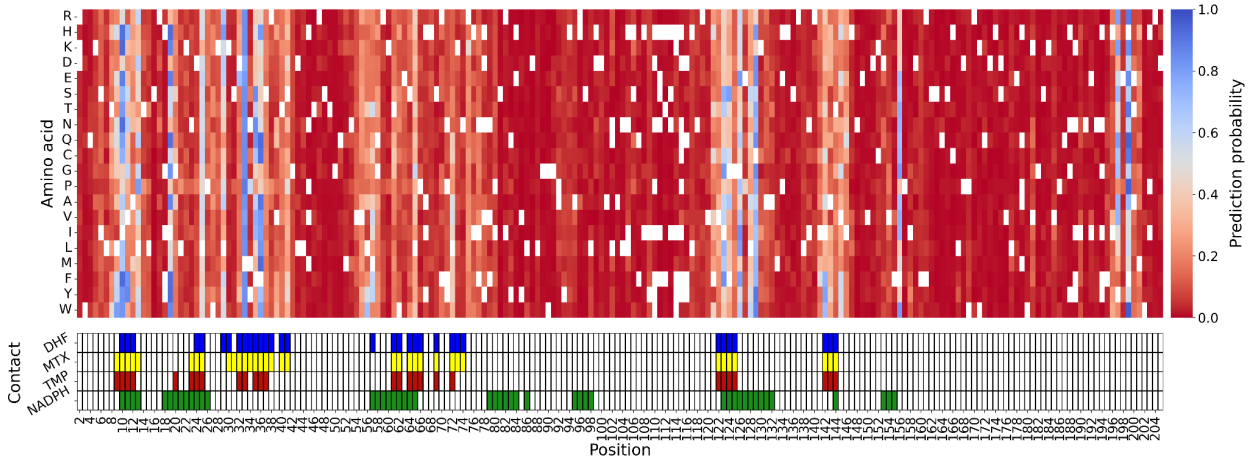


Figure 4: Confidence score for mutations predicted to lead to TMP resistance. Using the model trained on resistance to MTX, we predicted the effect of mutations in TMP resistance using data from the different

498 modeling strategies used on the PjDHFR-TMP complex structure. Mutations with probability >0.5 are
499 identified as resistant, and mutations <= 0.5 are classified as sensitive.
500
501 Next, we examined if our predictions captured mutations previously shown or
502 suspected to cause resistance, for instance ones that have been identified in clinical
503 studies and/or biochemically characterized. Few are available, and fewer yet have been
504 linked to failure of treatment, biochemically characterized, or identified in more than one
505 study (S2 table) [22,24,25,46–48]. Amongst the mutations identified in patient samples
506 that have been characterized *in vitro*, three had been identified as being most likely
507 resistance-causing, based on their impact on K_i, K_m and k_{cat} [22,24]. These amino acid
508 changes, F36C, L65P, and A67V, caused 100-fold, 20-fold and 40-fold increases in TMP
509 K_i in PjDHFR, respectively. The classification model predicted mutation F36C as leading
510 to TMP resistance with a confidence of 0.91 but did not predict mutations L65P
511 (Resistance probability 0.33) and A67V (0.03) as being likely to lead to resistance (S6C
512 fig for details on model decision). While all three mutations have effects on biochemical
513 constants that could be linked to resistance, their effect on function had not been
514 investigated in a biological system prior to this study. We found that several mutations at
515 position L65, including L65P, caused LOF, implying that this mutation cannot lead to TMP
516 resistance *in vivo*. However, while L65P was not identified as being likely to lead to TMP
517 resistance, mutation L65G was, with a prediction of 0.58, which is coherent with the large
518 effect on K_i caused by another mutation at this position. Neither F36C nor L65G were
519 experimentally identified as leading to MTX resistance in our previous study, meaning
520 that the model can make novel predictions based on data it had not encountered
521 previously.
522
523 **PjDHFRs from clinical samples show no mutations associated with increased**
524 **organismal load in patients or predicted as leading to TMP resistance**
525 Despite increased interest on the effect of mutations in PjDHFR on treatment
526 efficiency and patient outcome, the size of the available dataset on sequence
527 polymorphism is limited. To try and expand this dataset, and screen for potential
528 informative resistance mutations from patient samples, we sequenced PjDHFR in
529 *Pneumocystis jirovecii*-positive samples from the Laboratoire de Santé Publique du
530 Québec (LSPQ), a large public laboratory in Eastern Canada. We used multiplexed
531 amplicon sequencing specifically targeting PjDHFR in these samples. Using data from a
532 standard qPCR assay able to detect intervals between 3 000 and 3 000 000 genomic
533 copies per mL of bronchoalveolar wash, we investigate if DHFR mutations correlate with
534 larger organismal load in patients.
535 Between 2019 to 2024, 636 respiratory samples were found positive for *P. jirovecii*
536 as part of the *Pneumocystis* qPCR detection and quantification program (>5000 genome
537 copies/mL) from across the province of Quebec. From these, we successfully amplified
538 and sequenced the coding sequence of 531 PjDHFRs. As samples were sequenced with

539 a minimum of 20x coverage across all bases, we identified the presence of different
540 alleles within individual samples, when polymorphisms were identified in at least 30% of
541 reads at a given position. We identified 104 samples (19.6%) containing substitutions
542 relative to the *P. jirovecii* reference genome (RU7), with 94 harboring synonymous
543 substitutions (17.7%). The mutations found most often were the previously reported
544 G90G (T270C in the spliced sequence) silent mutation [25–27,49], as well as novel
545 synonymous ones (S9A fig.). Ten samples had amino acid substitutions, including the
546 previously reported G18W, A67V and C166Y, and 2 nonsense mutations [22,25,26].
547 A67V was one of two amino acid changes identified in more than one sample, being found
548 in four samples isolated at different hospitals between 2020 and 2024, and C166Y, found
549 in two samples. A subset of eight samples was validated by Sanger sequencing, and all
550 matched the amplicon sequencing results. We identified 44 samples (8.2% of all samples,
551 42.3% of samples with at least one polymorphism) that likely had two different PjDHFR
552 alleles (S9B fig.). When found, mutation A67V was always accompanied by a wild-type
553 allele. As *P. jirovecii* is suspected to produce a diploid zygote as part of its mainly
554 homothallic sexual reproduction, we cannot confidently say that the presence of two
555 alleles is due to coinfection [50,51]. However, as the transmissible form of *P. jirovecii* is
556 suspected to be haploid, coupled with the relatively high read threshold we set for both
557 assembly and polymorphisms, we believe that co-infection by two different strains is a
558 likely cause of the presence of different alleles within single samples, as shown by others
559 [14,51,52]. Most importantly, of all the mutations identified from clinical samples in this
560 study, none were predicted by the model to lead to TMP resistance in PjDHFR.
561 Samples were classified as being either wild-type, containing synonymous, non-
562 synonymous or nonsense mutations. We then compared each class with the qPCR count
563 of *P. jirovecii* genome copies within samples, and copy counts were standardized to
564 account for the method's detection limit of between 3 000 and 3 000 000 copies. We
565 found that the presence of non-synonymous mutations did not lead to a statistically
566 significantly higher qPCR count than in samples with no sequence changes relative to the
567 reference (S9C fig.). For substitution A67V, three out of four samples had counts of less
568 than 20 000, less than half of the median across all samples (~43 000 copies). The fourth
569 sample containing this mutation had a qPCR count of 1 794 000 copies, being above the
570 85th percentile of the distribution. From this distribution of counts, we cannot confidently
571 associate A67V with a higher *P. jirovecii* load in patients, despite biochemical evidence
572 that this mutation reduces affinity between TMP and PjDHFR. These results, where
573 mutations in PjDHFR are not associated with higher organism load, are coherent with the
574 presented model not predicting the identified mutations as being resistance-causing.
575 Furthermore, across all 531 sequenced samples, 265 also had their DHPS sequenced
576 during multi-locus sequence typing at the LSPQ, and mutations were only identified in
577 three samples, with only two being at position 55, associated in DHPS with SMX
578 resistance. We find no overlap between model predictions of TMP resistance and

579 mutations found in PjDHFR from patient samples. In addition, we find no correlation
580 between the presence of mutations and organismal load in patients' samples, and
581 therefore can make no associations between mutations in PjDHFR and outcome.
582
583 **Discussion**
584
585 **A combined experimental and computational framework for making predictions on**
586 **antifolate resistance in *P. jirovecii's* DHFR**
587 The study of how mutations in *P. jirovecii* affect its sensitivity to antifolate drugs
588 has always been difficult. As TMP and MTX share structural and mechanistic properties,
589 we hypothesized that data obtained on the effects of mutations in PjDHFR on resistance
590 to MTX could help predict the effect of mutations on resistance to TMP (Fig 1). Here, by
591 using experimental data from DMS screens on resistance to MTX and on protein function,
592 and computational data from various modeling strategies, we trained a machine learning
593 model to predict the effects of mutations in PjDHFR on resistance to TMP (Fi 2, Fi 3). The
594 resulting random forest model accurately captures the mutually exclusive character of
595 LOFs mutations and resistance, as well as the effects of modeled features such as
596 protein/ligand contacts, changes in protein free folding energy, and residue conservation.
597 While there was a significant overlap between the predictions made on the effects of
598 amino acid substitutions on resistance to MTX and TMP, the model predicts substitutions
599 that were not classified as MTX resistant would lead to TMP resistance, highlighting its
600 ability to make novel predictions based on new data (Fig 4). No single modeling method
601 can capture all effects of amino acid substitutions in PjDHFR, as each focus on individual
602 aspects of protein properties. Different positions have different levels of stringency for
603 specific amino acid properties. Therefore, prediction of mutational effects should be
604 conducted using several tools focusing on different features, and proteins must be
605 investigated in a per-residue manner to identify tolerated permutations at given positions.
606 Other properties, such as which residues are involved in catalysis and which residues are
607 allosterically linked, often still need to be experimentally determined [44,45,53]. While
608 computational methods can allow the modeling of granular properties (FlexddG, MutateX)
609 or integrative scores (GEMME), certain experimental methods cannot yet be fully
610 replaced, and are still necessary to assess biological functions.
611 Of the three mutations of interest identified in clinical samples that had been
612 biochemically characterized in [24], the model predicted that one would lead to TMP
613 resistance (F36C, not identified as leading to MTX resistance), one was predicted as not
614 leading to TMP resistance, as it was experimentally identified as non-functional (L65P),
615 and one was predicted as not leading to TMP resistance because of other factors (A67V).
616 To further compare model predictions to clinical data, we sequenced PjDHFRs from 531
617 patient samples from across the province of Québec and identified both new and
618 previously reported PjDHFR polymorphisms. No mutations identified from these patient
619 samples were predicted as leading to TMP resistance, and none was consistently

associated with high organismal load. This sequencing experiment also highlighted endemic sequence variation in PjDHFR sequence, as well as potential dynamics of coinfection by different *P. jirovecii* strains within patients.

However, as there yet exists no experimental methods to validate the predictions made by the presented model, or to test mutants identified from clinical samples for resistance to TMP, the presented results remain predictive. Despite this, our study offers novel resources to help understand mechanisms of antifolate resistance in PjDHFR, as well as a general workflow to predict the effect of mutations on resistance in other hard-to-study organisms. By identifying which features correlate best with MTX resistance and building a predictive model which we could validate using DMS data on MTX resistance, we believe we were able to make informed predictions about the effects of those same mutations on TMP resistance. With a balanced accuracy score of 88.3% in the testing set for MTX resistance, we expect that predictions regarding the effect of mutations on TMP resistance should have a similar overall balanced accuracy.

We believe that this framework could be applied to other protein/ligand combinations for which DMS data is available, and potentially help orient drug development. Indeed, by identifying position-dependent resistance-causing features, therapeutic molecules could be tailored to maximise binding to mutants that cause resistance to other therapeutic agents. Furthermore, this framework could help direct experimental testing efforts to specific positions of interest across the protein. These positions could be investigated individually, without the need for costly full-protein DMS, where it is expected that most mutations will have a neutral or deleterious effect. With the increase in available DMS data for many clinically-relevant proteins/ligand pairs, such frameworks are critical to make the most of the large dataset generated by these experiments.

Functional data from DMS screens reveals widespread lack of tradeoff in antifolate resistance and enhances TMP resistance model accuracy

When conducting a DMS screen on the effects of PjDHFR mutations on protein function, we observed that 14.2% of mutations lead to an LOF-like phenotype, which is consistent with reports in other essential fungal proteins and in bacterial DHFR [40,43,54]. Such mutations were concentrated at positions with very low tolerance, where most mutations lead to LOF, such as the 9Å diameter region of PjDHFR with a median selection coefficient of -0.363 (S4 fig.). Often, when only a subset of mutations led to LOF at a given position, these mutations shared functional characteristics such as side-chain length, polarity, or charge. An exception were substitutions towards proline, which were consistently more deleterious across the protein, consistent with previous reports [38,39,55]. Very few mutations were measured as having an intermediate effect on function (58/3758, 1.54%). This could be explained by the relatively high strength promoter used in the system, as high expression can mask the effects of slightly

destabilizing mutations [56]. Consistent with the small number of intermediate-effect mutations, even fewer mutations (11/3758, 0.29%) exhibited a tradeoff between function and MTX resistance (Fig 2C).

Experimental measurement of mutant functionality was critical to model performance. This feature ranked as the second most important predictive one and helped minimize both false positives and false negatives in the RF classification model (Fig 3, S7 fig.). In most cases, resistance and LOF were mutually exclusive, enabling accurate classification of non-functional mutants as non-resistant. Furthermore, several top predictive features were at least partially correlated with functionality. Having direct experimental data on this phenotype proved essential for training reliable classification models. Furthermore, having experimental data on the effect of mutations on protein function was key to minimize false positives, when compared to other models [28]. Overall, the best-performing features included ligand-independent features such as mutant functionality, residue buriedness, wild-type side chain length, and predicted changes in protein free energy (MutateX). Critical ligand-dependent features include proximity to ligand heavy atoms and changes in region flexibility, modeled via FlexddG outputs (Fig 3A, S1 Table).

Class imbalance negatively affect model performance when predicting antifungal resistance

When working with antifungal resistance, a minority of all possible single mutations will lead to resistance, unless LOF is enough by itself to cause resistance [54]. While it is advantageous from a clinical standpoint, it can cause ML datasets classes (sensitive/resistant) to be highly unbalanced, with a large majority of mutations having little to no effect on resistance and can therefore lead to poor model fitting. We had previously established a method that uses a gaussian mixture model to dissect complex distributions, which allowed us to establish thresholds for MTX resistance and function (S4 fig.). Only 111/3758 (2.8%/97.2%) mutations in PjDHFR were identified as leading to MTX resistance, which leads to imbalanced class distribution in the dataset, and therefore a potentially limited predictive power. Despite this, when training predictive models such as this one, we believe that it is critical to have data on as many amino acid substitutions as possible, hence the relevance of both experimental and *in silico* DMS. Having an interpretable model also allows us to rationalize the mechanisms of resistance, which then allows us to use critical judgment on the predictions that lie near the threshold used for binary classification based on the nature of the mutations involved.

While the model accurately classified the majority of mutations, this imbalance in classes is reflected in some evaluation metrics. For instance, the area under the curve (AUC) of the receiver operator characteristic curve (ROC), not affected by dataset imbalance, being high (0.97 on the test set) is explained by the very high amount of true negatives and indicates a well performing model. However, the AUC of the precision-

recall display curve (PRD), which is much more sensitive to class imbalance, being lower (0.41 on the test set) is caused by the relatively high number of false positives and low number of true positives, which negatively impacts precision (Fig 3A). True negatives, while influencing the ROC curve, do not affect the PRD curve. It is key to use several evaluation metrics when training and testing classification models, and be critical of model performance, based on desired outcomes. Furthermore, when working with datasets with such imbalance, bootstrapping, oversampling and SMOTE sampling often lead to overfitting, minimizing predictive power outside of the training dataset, and should therefore be avoided when possible [57]. This was also reflected in the balanced accuracy scores obtained for training (98.3%) and testing (88.3%). With 21 possible true positives in the test set (2.79% of observations), misclassification of mutants as either false negative or false positive has a large impact on balanced accuracy. Class imbalance in relatively small datasets such as this one (< 5000 observations) has been known to affect prediction accuracy.

While the ML-based methodology presented in this study uses experimental measurements for model training and validation, other variant effects predictor software use purely computational features for ML-less modeling and prediction[58,59]. One such tool is Resistor. Based on the protein design software Osprey, this algorithm uses predicted changes in binding affinity to ligands caused by mutations as well as their synergistic effects and makes ranked predictions based on mutation likelihood. Such algorithms are not affected by dataset class balance as they do not require *ab initio* ground truths like ML-based models. They also focus on fewer, empirical features, and therefore offer a complimentary approach to ML-based methods such as the one presented here [58,59].

Nevertheless, the model successfully predicted that certain mutations, not measured as conferring MTX resistance, would lead to TMP resistance in PjDHFR. This included mutations found in literature and biochemically identified as reducing affinity to TMP, with one of the main predictors, region flexibility (here IQR change in distance) being identified as a good predictor of TMP resistance in a previous study [28].

Expanding the clinical genomics landscape of *P. jirovecii* DHFR highlight endemic genetic variation

To illustrate how our model could be put to use, we sequenced the largest dataset of PjDHFR from clinical isolates presented in literature to date and compounded the results of many previous studies. Despite sequencing PjDHFRs from 531 patient samples spread across five years, different geographical areas, with differences in organismal loads, and finding both new and previously reported mutations, no mutations from this patient cohort could be confidently linked to TMP-SMX resistance. However, we believe that this high-throughput sequencing of PjDHFRs sheds light on endemically circulating genotypes of PjDHFRs in eastern Canada, such as G90G, A67V and C166Y, as well as

740 co-infection by different strains of *P. jirovecii*. While the DHPS has been sequenced for
741 many clinical samples when PCP was diagnosed, the DHFR locus had not been as
742 thoroughly investigated in this organism before this study. The lack of genetic information
743 on the endemic and clinical genetic variation in PjDHFR compounds the difficulty to study
744 resistance on this system. There is a need to sequence further genomes of *P. jirovecii*,
745 or at least of its drug targets, even in the event of treatment success. To further validate
746 these results, it is necessary to conduct such high-throughput sequencing studies across
747 other geographical regions to uncover more information on the evolution of TMP-SMX
748 resistance/treatment failure in *P. jirovecii,* including in other antifolate-resistance
749 associated genes with robust clinical outcome data. This would be especially relevant in
750 regions of the world where failure of prophylaxis and treatment is high, and where second-
751 line treatments are less available [60]. As the mechanism of action of TMP-SMX in this
752 fungi is still poorly understood, with antifungal activity often solely attributed to SMX, such
753 studies could help better understand how this drug combination contributes to treatment
754 outcome [61,62].
755
756 **Conclusion**:
757 Recent advances in experimental technologies enable an unprecedented level of
758 granularity in studying the effects of mutations on proteins. However, applying these high-
759 throughput methods to hard-to-study organisms, such as unculturable organisms,
760 remains challenging. Computational methods and machine learning are key tools to
761 address these limitations but are not yet accurate enough to replace experimental
762 validation, particularly when addressing complex phenomena like the evolution of
763 resistance. Nevertheless, we believe that high-throughput computational methods,
764 combined with the pattern-recognition capabilities of machine learning, are critical for
765 guiding experimental design and identification of key determinants of drug resistance,
766 especially in hard-to-study organisms.

767 **Materials & Methods**

768 **General methods**

769 All yeast transformations were conducted using the standard lithium acetate
770 transformation protocol with heat-shock temperature adjusted to 25 °C when using
771 temperature sensitive strains [63]. All PCR reactions were conducted using KAPA HiFi
772 Hotstart DNA polymerase (Sigma-Aldrich, catalog #BK1000) unless stated otherwise. All
773 mutations identified from the literature and from clinical samples in this study are provided
774 in S2 table. Details for all strains are available in S3 Table, plasmids in S4 Table,
775 oligonucleotides in S5 Table, media in S6 Table and PCR reactions in S7 Table.
776

777 **Temperature sensitive DHFR strains**

778 Yeast strains used in this study were taken from the Yeast Temperature-sensitive
779 (TS) collection [64]. TS-DHFR replicates were taken from individual wells from the
780 collection as biological replicates, and spread on YPD + 200µg/mL Geneticin
781 (YPD+G418) agar plates (Plate 4A position B8, plate 4B position B7, plate 4C position
782 A8, plate 4D position A7 in the TS collection [64]). Four isolated colonies were harvested
783 from each biological replicate and grown in YPD+G418 liquid media. Each culture was
784 diluted to 1.0 OD_{600} and serial-dilution spot assays were conducted on YPD+G418 plates
785 and grown at 20 °C and 37 °C for 72h with pictures taken every 24h (S1 fig.). A colony
786 from each biological replicate that exhibited growth at 20 °C but not at 37 °C (Colony #1,
787 S1 fig.) was selected to ensure the presence of the temperature-sensitive phenotype.
788 Complementation by PjDHFR was tested by transforming pAG416-GPD-PjDHFR into all
789 four strains and observing growth at 37 °C in synthetic-complete solid media without uracil
790 and with monosodium glutamate (SC-URA+MSG) media.

791 All subsequent experiments were conducted in quadruplicate using each individual
792 TS-DHFR strain. All subsequent experiments with these strains were conducted at 20 °C
793 unless stated otherwise.
794

795 **PjDHFR DMS library transformation and pooled screens**

796 The PjDHFR DMS library from [36] was pooled in equivalent molar concentrations
797 by position in four fragments, with an 11 codon overlap between each adjacent fragment.
798 Fragment 1 comprised positions 2 to 63, fragment 2 from 52 to 115, fragment 3 from 104
799 to 167 and fragment 4 from 156 to 206. Each fragment pool was individually transformed
800 into competent cells from each of the four TS-DHFR strains (biological replicates), with
801 16 technical replicates per TS-DHFR strain, and incubated at 20 °C for seven days in SC-
802 URA+MSG media to minimize the selection bias against slow-growing PjDHFR mutants.
803 Using SC-URA+MSG liquid media, the colonies on plates were harvested in sterile falcon
804 tubes for each fragment and biological replicates, with technical replicates pooled at this
805 step. These were diluted to 0.1 OD_{600} in fresh liquid SC-URA+MSG and grown at both 20
806 °C and 37 °C with 250 RPM orbital agitation for 24h, then passed to 0.1 OD_{600} in fresh

807 liquid SC-URA+MSG for another 24h. At each timepoint, 5.0 OD_{600} units were spun down
808 at 224 g for 3 minutes, and pellets were harvested for downstream DNA extraction and
809 frozen at -80 °C. Glycerol stocks were made in triplicate.
810 Plasmid DNA extractions were conducted using Zymoprep Yeast Plasmid
811 Miniprep II (Zymo Research, cat #D2004), with an added pellet freezing step after initial
812 pelleting, and with 3h incubation with zymolyase. Extracted plasmids were amplified using
813 the primers listed in S5 Table and purified using using magnetic beads (AMPureXP from
814 Beckman Coulter Life Sciences, catalog #A63882). Sequencing was conducted using the
815 Element AVITI with paired-end 150bp, yielding ~100M total reads, resulting in 1800 reads
816 per variant at each timepoint and condition. Parsed and analysed mutation calling is
817 available in S4 File for the pooled assay at 20 °C (permissive condition) and S5 File for
818 the pooled assay at 37 °C (selective condition). Raw reads for the function screen (both
819 temperatures) are available at BioProject PRJNA1243435.
820
821 **Read processing, mutations calling and statistical analysis**
822 Reads were processed and analyzed as described in [36]. Briefly, R1 and R2 reads
823 were merged using Pandaseq (v2.11) using default PHRED score threshold [65], and
824 aggregated and counted using VSEARCH (v2.22.1) [66]. Changes in read frequency
825 were calculated between T_0 and T_{Final} for all variants. Frequencies were calculated as the
826 proportion of reads containing a given codon relative to the total number of reads in that
827 sample. Selection coefficients for all variants were computed using the following equation,
828 with the median score of silent mutations relative to the WT sequence as the
829 normalization baseline (wild-type score):
830
831 Eq 1: $$Selection\ coefficient = \frac{Log2(\frac{Freq\ mutant\ T_{final}}{Median\ freq\ silent\ T_{final}}) - Log2(\frac{Freq\ mutant\ T_0}{Median\ freq\ silent\ T_0})}{g}$$
832
833 Where g, the number of generations, was estimated based on optical density measured
834 at the start and end of the different timepoints. Mutations in codons that lead to the same
835 amino acid substitutions were grouped after computing the selection coefficient, and the
836 median score was used for subsequent analysis of the effect of amino acid substitution.
837 Statistical analysis, thresholding and control for false-discovery rate were conducted in
838 an identical manner as described in [36]. Using the scikit-learn (sklearn, v1.5.1)
839 GaussianMixture algorithm, the distributions of fitness scores were decomposed in their
840 underlying gaussian distributions, reflecting the statistical distribution of selection
841 coefficients, and thresholds were set accordingly. We controlled for false discovery rate
842 using the distribution of silent mutations as reference for a Welch's one-sided *t*-test
843 followed by a Benjamini-Hochberg (BH) FDR control with a threshold of 0.05 (S4 Fig.).
844

845 **Structural analysis modeling of changes in protein free energy and in binding**
846 **affinity to ligands**
847 The structure of PjDHFR was recovered from AlphaFoldDB (UniProt: A0EPZ9).
848 The crystal structures of PjDHFR, in complex with MTX/NADPH (PDB: 3CD2)
849 DHF/NADPH (PDB: 4CD2), and TMP/NADPH (PDB: 1DYR) were recovered from the
850 Protein Data Bank (PDB). Using the ChimeraX (version 1.6) [67] matchmaker algorithm,
851 the structures were aligned with each other and ligands from crystal structures were
852 extracted and added to the PjDHFR AlphaFold structure.
853 The resulting ligand-protein complex structures were then relaxed ten times using
854 the FoldX (suite 5) optimize command to remove van der Waals clashes [68]. We used
855 MutateX v0.8 [33] protocol from FoldX suite 5 [68] on the entire protein, with 10
856 incorporated minimization steps following mutations to reduce atom clashes sometimes
857 caused by longer sidechains, before ΔΔG measurements [69]. Apart from additional
858 minimization steps, analysis was conducted as described in [36].
859 Using RosettaDDGPrediction (Rosetta v3.12 and 2022.11), adapted with custom
860 scripts to allow high-throughput computing on AllianceCan's servers, the FlexddG
861 protocol with saturation mutagenesis was run across the entire length of the PjDHFR-
862 ligand structures [34,35,70]. The effects of all mutations were measured on the changes
863 in Gibbs free energy (ΔΔG) between PjDHFR and the ligand of interest, either MTX, TMP
864 or NADPH. This was done 100 times with 35 000 backrub steps for each mutant and all
865 individual structures were analyzed with a modified version of the basic FlexddG analysis
866 script, available on GitHub [34]. As final scores, the median ΔΔG of each mutation was
867 used for downstream analysis.
868 We measured the changes in distance between the α-carbon of the mutated and
869 wild-type residue and the closest heavy atom on the ligand. FlexddG generates, for each
870 mutation of interest, a structure for both the wild-type and the mutated sequence before
871 back-rubbing, and measures ΔΔG by comparing the wild-type structure's ΔG with the
872 mutant's ΔG. By using each structure, we could measure the effect of each mutation on
873 the protein's flexibility by comparing the position of the α-carbon in the wild-type and
874 mutant structures, which yielded the root-mean-square fluctuation of each structure, or
875 the change in distance to ligand between the wild-type structure and the mutant. This
876 allowed us to get a distribution of the effective structural variation caused by the mutation.
877 From this distribution, we extracted three values: 1) the median value of changes in
878 distance, representing the overall change in residue position caused by the mutation, 2)
879 the skewness of the distribution, which represents the asymmetry of the distribution, or
880 the region's propensity to be in a region of space more than in another, either generally
881 closer or further from the ligand, and 3) the interquartile range of the distribution, IQR,
882 representing the "flexibility" of the protein region caused by the mutation,. A wide
883 distribution (high IQR) means the mutated residue increases region flexibility, as the
884 mutated region can occupy a broader range of positions further from its wild-type position.

885
886 **Sequence conservation and epistatic interaction between positions**
887 Using the MPI Bioinformatics Toolkit portal, we ran HHblits on the reference RU7
888 PjDHFR amino acid sequence using default settings [49,71,72]. The resulting 1040
889 sequences with < 70% sequence homology were aligned, and gaps were removed (S5
890 File, raw alignment in S6 File). This alignment was used as the input for GEMME to
891 compute epistasis and conservation effects scores [37]. All three scores, Combined,
892 Epistasis and Individual, were compiled for all possible single mutants based on
893 sequence conservation and co-occurrence across lineages. All data for GEMME analysis,
894 and its output, is available on GitHub.
895
896 **Other amino acid and position properties**
897 Amino acid properties for both wild-type and mutant sequences were defined using
898 Expasy amino acid properties [73]. All available properties were correlated with each
899 other, and property with the highest median correlation within each group was kept to
900 minimize the dimensionality of the dataset. Variance inflation factor (VIF) was also used
901 to minimize correlations within the dataset. Buriedness was computed using a script
902 available at https://github.com/rodogi/buriedness.git, which is a metric similar to
903 accessible surface area (ASA), but with no zero values, allowing comparison between
904 buried and very buried residues, being more informative about relative residue depth than
905 strict ASA.
906 The same script used to measure changes in position in FlexddG structures was
907 used to measure distances between ligands and α-carbon in the wild-type protein as the
908 distance to ligand feature. We used DSSP to assign letter scores to all positions to identify
909 if they are part of secondary structures [74,75].
910
911 **Machine learning model selection, training, and testing**
912 To predict the effect of mutations on PjDHFR's resistance to TMP, the PjDHFR-
913 TMP dataset (S2 file) was used. The features used to train the model were FlexddG score,
914 MutateX score, a binary classification of the mutant's functionality, α- carbon distance to
915 the closest ligand's heavy atom, GEMME scores, and effects of the mutation on structure
916 flexibility, as measured from FlexddG output structures, the length of the side chain of the
917 wild-type amino acid, allostery confidence score, importance in catalysis and buriedness
918 were also used (S1 Table) [40,44,76]. Other features present in the raw dataframe were
919 dropped from the dataset before training but are kept in the raw file to help visualise model
920 performance after training and testing.
921 The model was created using scikit-learn (scikit-learn, v1.5.1) [77]. Using a fixed
922 seed (999), the dataset was split into training and test sets, with 20% reserved for testing,
923 using train_test_split(). Data was scaled using the standard scaler provided by the
924 package, StandardScaler(), which performs z-score standardisation, helpful to make all

925 data on similar scales [77]. Both classifier and regression methods were tested, including
926 but not limited to logistic regression, random forests, decision trees, support vector
927 machines and XGboost. A randomized hyperparameter search was used for broad model
928 selection, estimating performance using the median across five cross-validation. Across
929 all models, the random forest method consistently performed best. Classification
930 performance was assessed using precision, recall, F1-score, and log loss (binary cross-
931 entropy), while regression models were evaluated using RMSE, R^2, and rank correlation.
932 Once the best performing method was selected, an exhaustive grid search was
933 performed, with parameters granularity defined according to the computation time
934 required by the given model. We used SHapley Additive exPlanations (SHAP, v0.42.1)
935 values to identify which model features could best explain the effect of mutations on
936 resistance to MTX [78].
937
938 **Predictions for the effect of mutations on TMP resistance in PjDHFR**
939 In parallel with generating data on the PjDHFR-MTX complex, we performed the
940 same analyses on the PjDHFR-TMP complex and compiled a dataset using the same
941 features as those used to train the ML model. This dataset was then used to predict the
942 impact of mutations on TMP resistance using the trained model. Using the predict(), and
943 predict_proba() methods from the sklearn package, predictions on TMP resistance-
944 causing mutations were made from the dataset generated with the PjDHFR-TMP
945 complex. All code used to train the model on MTX resistance, and to make predictions on
946 TMP resistance, is available on GitHub.
947
948 **Sequencing PjDHFR from clinical samples**
949 Primary samples were acquired by the Laboratoire de Santé Publique du Québec, in
950 Montréal, as part of a province-wide *Pneumocystis jirovecii* monitoring program. In total,
951 636 samples were obtained from December 2018 to May 2024 from 41 hospitals across
952 the province. DNA was extracted using Biomerieux's EMAG® standard protocol. qPCR
953 was conducted using Altona Diagnostics' RealStar® *Pneumocystis jirovecii* PCR Kit 1.0
954 (Cat #551013) to quantify *Pneumocystis jirovecii* genomic copy numbers from
955 standardized primary sample volumes. For this specific kit, no housekeeping gene is used
956 as a copy number normalization control, but samples are systematically tested for the
957 presence of PCR inhibitors to minimize copy count bias.
958 Primers specifically targeting PjDHFR were then used on the extracted DNA to
959 obtain a full length PjDHFR amplicon [79]. Using the Row-Column-Plate PCR method
960 described in [80], with further steps to minimize inter-sample contamination, Briefly, full
961 length PjDHFRs amplicons were split into 3 fragments using fragment-specific primers
962 (F1 - 586bp, F2 - 579bp and F3 - 533bp, 113bp overlap between adjacent fragments),
963 and barcoded in a sample-wise manner, where fragments from the same sample shared
964 barcodes. In total, 531 PjDHFR amplicons were prepared for pooled amplicon sequencing

using Illumina MiSeq for ~700 000 PE300 reads, or ~1300 reads per PjDHFR sample (Oligos in S4 table).

Using custom scripts, sequencing data was parsed and assembled into full length sequences (bwa-mem2 v2.2.1, samtools v1.21, cutadapt v2.6 [81–83].) To ensure high-quality assembly, samples with fewer than 125 total reads or fewer than 25 reads per fragment were excluded. After filtering, 531 high-quality PjDHFR sequences were successfully assembled from clinical samples. Using custom scripts, SNPs were called by comparing assembled amplicons to the RU7 reference sequence. With a minimum of 25x coverage at every position, we were able to identify intra-sample polymorphisms. Polymorphisms were called when a non-consensus base was present in over 30% of reads mapping to a given position. All raw reads for amplicon sequencing of clinical samples are available at BioProject SAMN47937957.

Acknowledgements

We would like to thank the Landry lab for their helpful comments throughout the project. We also thank Mélanie Côté for her work in acquiring and sorting the clinical samples.

Funding

The project was funded by a Genome Canada and Genome Quebec grant (6569, https://genomecanada.ca/). CRL holds the Canada Research Chair in Cellular Systems and Synthetic Biology (https://www.chairs-chaires.gc.ca/home-accueil-eng.aspx). FDR was supported by fellowships from FRQNT, PROTEO, EvoFunPath NSERC CREATE program and the Vanier Canada Graduate Scholarship agency. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

Raw read sequencing from function assays is available at BioProject accession PRJNA1243435**.** Raw read sequencing from PjDHFR sequencing from patient samples is available at BioProject accession SAMN47937957**.** All scripts for figures and analysis, as well as processed data, is available on GitHub at https://github.com/Landrylab/Rouleau_et_al_2025.git. Detailed files and structures for FlexddG analysis are available upon request, as the complete output is around 4.8Tb of data (~4.2M structures + accompanying files). FlexddG input and output files are available at Zenodo DOI: 10.5281/zenodo.15269626

1002 **References**

1003 1. Stop neglecting fungi. Nat Microbiol. 2017;2: 17120.

1004 2. Microbiology by numbers. Nat Rev Microbiol. 2011;9: 628.

1005 3. Ricciardi A, Gentilotti E, Coppola L, Maffongelli G, Cerva C, Malagnino V, et al.
1006 Infectious disease ward admission positively influences P. jiroveci pneumonia (PjP)
1007 outcome: A retrospective analysis of 116 HIV-positive and HIV-negative
1008 immunocompromised patients. PLoS One. 2017;12: e0176881.

1009 4. Morris A, Norris KA. Colonization by pneumocystis jirovecii and its role in disease.
1010 Clinical Microbiology Reviews. 2012. pp. 297–317. doi:10.1128/CMR.00013-12

1011 5. Medrano FJ, Montes-Cano M, Conde M, de la Horra C, Respaldiza N, Gasch A, et
1012 al. Pneumocystis jirovecii in general population. Emerg Infect Dis. 2005;11: 245–
1013 250.

1014 6. Ponce CA, Gallo M, Bustamante R, Vargas SL. Pneumocystis colonization is highly
1015 prevalent in the autopsied lungs of the general population. Clin Infect Dis. 2010;50:
1016 347–353.

1017 7. O'Brien KL, Baggett HC, Brooks WA, Feikin DR, Hammitt LL, Higdon MM, et al.
1018 Causes of severe pneumonia requiring hospital admission in children without HIV
1019 infection from Africa and Asia: the PERCH multi-country case-control study. Lancet.
1020 2019;394: 757–779.

1021 8. Ibrahim A, Chattaraj A, Iqbal Q, Anjum A, Rehman MEU, Aijaz Z, et al.
1022 Pneumocystis jiroveci Pneumonia: A Review of Management in Human
1023 Immunodeficiency Virus (HIV) and Non-HIV Immunocompromised Patients.
1024 Avicenna J Med. 2023;13: 23–34.

1025 9. Roembke F, Heinzow HS, Gosseling T, Heinecke A, Domagk D, Domschke W, et
1026 al. Clinical outcome and predictors of survival in patients with pneumocystis jirovecii
1027 pneumonia--results of a tertiary referral centre. Clin Respir J. 2014;8: 86–92.

1028 10. Nasr M, Mohammad A, Hor M, Baradeiya AM, Qasim H. Exploring the Differences
1029 in Pneumocystis Pneumonia Infection Between HIV and Non-HIV Patients. Cureus.
1030 2022;14: e27727.

1031 11. Casalini G, Giacomelli A, Antinori S. The WHO fungal priority pathogens list: a
1032 crucial reappraisal to review the prioritisation. Lancet Microbe. 2024;5: 717–724.

1033 12. Riebold D, Mahnkopf M, Wicht K, Zubiria-Barrera C, Heise J, Frank M, et al. Axenic
1034 long-term cultivation of Pneumocystis jirovecii. J Fungi (Basel). 2023;9: 903.

1035 13. Cissé OH, Ma L, Kovacs JA. Retracing the evolution of Pneumocystis species, with
1036 a focus on the human pathogen Pneumocystis jirovecii. Heitman J, editor. Microbiol
1037 Mol Biol Rev. 2024;88. doi:10.1128/mmbr.00202-22

1038 14. Ma L, Cissé OH, Kovacs JA. A Molecular Window into the Biology and
1039 Epidemiology of Pneumocystis spp. Clin Microbiol Rev. 2018;31.
1040 doi:10.1128/CMR.00009-18

1041 15. White PL, Price JS, Backx M. Therapy and Management of Pneumocystis jirovecii
1042 Infection. Journal of Fungi. 2018;4: 127–127.

1043 16. Tomblyn M, Chiller T, Einsele H, Gress R, Sepkowitz K, Storek J, et al. Guidelines
1044 for preventing infectious complications among hematopoietic cell transplantation
1045 recipients: a global perspective. Biol Blood Marrow Transplant. 2009;15: 1143–
1046 1238.

1047 17. Lee SM, Cho YK, Sung YM, Chung DH, Jeong SH, Park J-W, et al. A Case of
1048 Pneumonia Caused by Pneumocystis jirovecii Resistant to Trimethoprim-
1049 Sulfamethoxazole. Korean J Parasitol. 2015;53: 321–327.

1050 18. Liu A, Sun R, Cao G, Liu X, Zhu H, Yang J. Prognostic factors and clinical efficacy
1051 of second-line treatments of Pneumocystis jirovecii pneumonia for non-HIV patients
1052 after first-line treatment failure. BMC Infect Dis. 2022;22: 546.

1053 19. Huang Y-S, Yang J-J, Lee N-Y, Chen G-J, Ko W-C, Sun H-Y, et al. Treatment of
1054 Pneumocystis jirovecii pneumonia in HIV-infected patients: a review. Expert Rev
1055 Anti Infect Ther. 2017;15: 873–892.

1056 20. Singh Y, Mirdha BR, Guleria R, Kabra SK, Mohan A, Chaudhry R, et al. Genetic
1057 polymorphisms associated with treatment failure and mortality in pediatric
1058 Pneumocystosis. Sci Rep. 2019;9: 1192.

1059 21. Nahimana A, Rabodonirina M, Francioli P, Bille J, Hauser PM. Pneumocystis
1060 jirovecii dihydrofolate reductase polymorphisms associated with failure of
1061 prophylaxis. J Eukaryot Microbiol. 2003;50 Suppl: 656–657.

1062 22. Nahimana A, Rabodonirina M, Bille J, Francioli P, Hauser PM. Mutations of
1063 Pneumocystis jirovecii Dihydrofolate Reductase Associated with Failure of
1064 Prophylaxis. Antimicrob Agents Chemother. 2004;48: 4301–4301.

1065 23. Helweg-Larsen J, Benfield T, Atzori C, Miller RF. Clinical efficacy of first- and
1066 second-line treatments for HIV-associated Pneumocystis jirovecii pneumonia: a tri-
1067 centre cohort study. J Antimicrob Chemother. 2009;64: 1282–1290.

1068 24. Queener SF, Cody V, Pace J, Torkelson P, Gangjee A. Trimethoprim resistance of
1069 dihydrofolate reductase variants from clinical isolates of Pneumocystis jirovecii.
1070 Antimicrob Agents Chemother. 2013;57: 4990–4998.

1071 25. Wu Y, Shi H, Li W, An Y, Shao Y, Rao X, et al. Analysis of mutations in
1072 Pneumocystis jirovecii dihydropteroate synthase and dihydropteroate reductase
1073 genes among non-HIV patients in China. Infect Drug Resist. 2024;17: 5619–5627.

1074 26. Costa MC, Esteves F, Antunes F, Matos O. Genetic characterization of the

1075 dihydrofolate reductase gene of Pneumocystis jirovecii isolates from Portugal. J
1076 Antimicrob Chemother. 2006;58: 1246–1249.

1077 27. Singh Y, Mirdha BR, Guleria R, Khalil S, Panda A, Chaudhry R, et al. Molecular
1078 detection of DHFR gene polymorphisms in Pneumocystis jirovecii isolates from
1079 Indian patients. J Infect Dev Ctries. 2015;9: 1250–1256.

1080 28. Leidner F, Yilmaz K, Schiffer CA. Deciphering Antifungal Drug Resistance in
1081 Pneumocystis jirovecii DHFR with Molecular Dynamics and Machine Learning. Cite
1082 This: J Chem Inf Model. 2021;61. doi:10.1021/acs.jcim.1c00403

1083 29. Barclay BJ, Nagel MG, Huang T. Dihydrofolate Reductase is not the Target of
1084 Trimethoprim in Saccharomyces Cerevisiae. In: Ayling JE, Nair MG, Baugh CM,
1085 editors. Chemistry and Biology of Pteridines and Folates. Boston, MA: Springer US;
1086 1993. pp. 551–554.

1087 30. Fernández-Villa D, Aguilar MR, Rojo L. Folic acid antagonists: Antimicrobial and
1088 immunomodulating mechanisms and applications. Int J Mol Sci. 2019;20: 4996.

1089 31. Zhao R, Goldman ID. Resistance to antifolates. Oncogene. 2003;22: 7431–7457.

1090 32. Visentin M, Zhao R, Goldman ID. The Antifolates. Hematol Oncol Clin North Am.
1091 2012;26: 629–629.

1092 33. Tiberti M, Terkelsen T, Degn K, Beltrame L, Cremers TC, da Piedade I, et al.
1093 MutateX: an automated pipeline for in silico saturation mutagenesis of protein
1094 structures and structural ensembles. Brief Bioinform. 2022;23.
1095 doi:10.1093/bib/bbac074

1096 34. Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, et al.
1097 Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein
1098 Binding Affinity upon Mutation. J Phys Chem B. 2018;122: 5389–5399.

1099 35. Sora V, Laspiur AO, Degn K, Arnaudi M, Utichi M, Beltrame L, et al.
1100 RosettaDDGPrediction for high-throughput mutational scans: From stability to
1101 binding. Protein Sci. 2023;32: e4527.

1102 36. Rouleau FD, Dubé AK, Gagnon-Arsenault I, Dibyachintan S, Pageau A, Després
1103 PC, et al. Deep mutational scanning of Pneumocystis jirovecii dihydrofolate
1104 reductase reveals allosteric mechanism of resistance to an antifolate. PLoS Genet.
1105 2024;20: e1011252.

1106 37. Laine E, Karami Y, Carbone A. GEMME: A simple and fast Global Epistatic Model
1107 predicting Mutational Effects. Mol Biol Evol. 2019;36: 2604–2619.

1108 38. Dunham AS, Beltrao P. Exploring amino acid functions in a deep mutational
1109 landscape. Mol Syst Biol. 2021;17: e10305.

1110 39. Schneider A, Cannarozzi GM, Gonnet GH. Empirical codon substitution matrix.

1111 BMC Bioinformatics. 2005;6: 134.

1112 40. Nguyen TN, Ingle C, Thompson S, Reynolds KA. The genetic landscape of a
1113 metabolic interaction. Nat Commun. 2024;15: 3351.

1114 41. Wallace LA, Robert Matthews C. Highly divergent dihydrofolate reductases
1115 conserve complex folding mechanisms. J Mol Biol. 2002;315: 193–211.

1116 42. Kazan IC, Mills JH, Ozkan SB. Allosteric regulatory control in dihydrofolate
1117 reductase is revealed by dynamic asymmetry. Protein Sci. 2023;32: e4700.

1118 43. Bédard C, Gagnon-Arsenault I, Boisvert J, Plante S, Dubé AK, Pageau A, et al.
1119 Most azole resistance mutations in the Candida albicans drug target confer cross-
1120 resistance without intrinsic fitness cost. Nat Microbiol. 2024;9: 3025–3040.

1121 44. Reynolds KA, McLaughlin RN, Ranganathan R. Hot spots for allosteric regulation
1122 on protein surfaces. Cell. 2011;147: 1564–1575.

1123 45. McCormick JW, Russo MA, Thompson S, Blevins A, Reynolds KA. Structurally
1124 distributed surface sites tune allosteric regulation. Elife. 2021;10.
1125 doi:10.7554/eLife.68346

1126 46. Siripattanapipong S, Leelayoova S, Mungthin M, Worapong J, Tan-Ariya P. Study of
1127 DHPS and DHFR genes of Pneumocystis jirovecii in Thai HIV-infected patients.
1128 Med Mycol. 2008;46: 389–392.

1129 47. Zhu M, Ye N, Xu J. Clinical characteristics and prevalence of dihydropteroate
1130 synthase gene mutations in Pneumocystis jirovecii-infected AIDS patients from low
1131 endemic areas of China. PLoS One. 2020;15: e0238184–e0238184.

1132 48. Vera C, Rueda ZV. Transmission and Colonization of Pneumocystis jirovecii.
1133 Journal of Fungi 2021, Vol 7, Page 979. 2021;7: 979–979.

1134 49. Ma L, Chen Z, Huang DW, Kutty G, Ishihara M, Wang H, et al. Genome analysis of
1135 three Pneumocystis species reveals adaptation mechanisms to life exclusively in
1136 mammalian hosts. Nat Commun. 2016;7: 10740.

1137 50. Skalski JH, Kottom TJ, Limper AH. Pathobiology of Pneumocystis pneumonia: life
1138 cycle, cell wall and cell signal transduction. FEMS Yeast Res. 2015;15: fov046.

1139 51. Cissé OH, Ma L, Wei Huang D, Khil PP, Dekker JP, Kutty G, et al. Comparative
1140 population genomics analysis of the mammalian fungal pathogen Pneumocystis.
1141 MBio. 2018;9. doi:10.1128/mBio.00381-18

1142 52. Richard S, Almeida JMGCF, Cissé OH, Luraschi A, Nielsen O, Pagni M, et al.
1143 Functional and expression analyses of the Pneumocystis MAT genes suggest
1144 obligate sexuality through primary homothallism within host lungs. MBio. 2018;9:
1145 e02201-17.

1146 53. Wan Q, Bennett BC, Wilson MA, Kovalevsky A, Langan P, Howell EE, et al. Toward
1147 resolving the catalytic mechanism of dihydrofolate reductase using neutron and
1148 ultrahigh-resolution X-ray crystallography. Proc Natl Acad Sci U S A. 2014;111:
1149 18225–18230.

1150 54. Després PC, Cisneros AF, Alexander EMM, Sonigara R, Gagné-Thivierge C, Dubé
1151 AK, et al. Asymmetrical dose responses shape the evolutionary trade-off between
1152 antifungal resistance and nutrient use. Nat Ecol Evol. 2022;6: 1501–1515.

1153 55. Gray VE, Hause RJ, Fowler DM. Analysis of large-scale Mutagenesis data to
1154 assess the impact of single amino acid substitutions. Genetics. 2017;207: 53–61.

1155 56. Cisneros AF, Gagnon-Arsenault I, Dubé AK, Després PC, Kumar P, Lafontaine K,
1156 et al. Epistasis between promoter activity and coding mutations shapes gene
1157 evolvability. Sci Adv. 2023;9: eadd9109.

1158 57. Meng D, Li Y. An imbalanced learning method by combining SMOTE with Center
1159 Offset Factor. Appl Soft Comput. 2022;120: 108618.

1160 58. Guerin N, Feichtner A, Stefan E, Kaserer T, Donald BR. Resistor: An algorithm for
1161 predicting resistance mutations via Pareto optimization over multistate protein
1162 design and mutational signatures. Cell Syst. 2022;13: 830-843.e3.

1163 59. Guerin N, Kaserer T, Donald BR. Protocol for predicting drug-resistant protein
1164 mutations to an ERK2 inhibitor using RESISTOR. STAR Protoc. 2023;4: 102170.

1165 60. Saah AJ, Hoover DR, Peng Y, Phair JP, Visscher B, Kingsley LA, et al. Predictors
1166 for failure of Pneumocystis carinii pneumonia prophylaxis. Multicenter AIDS Cohort
1167 Study. JAMA. 1995;273: 1197–1202.

1168 61. Gleckman R, Blagg N, Joubert DW. Trimethoprim: mechanisms of action,
1169 antimicrobial activity, bacterial resistance, pharmacokinetics, adverse reactions, and
1170 therapeutic indications. Pharmacotherapy. 1981;1: 14–20.

1171 62. Hida S, Yoshida M, Nakabayashi I, Miura NN, Adachi Y, Ohno N. Anti-fungal
1172 activity of sulfamethoxazole toward Aspergillus species. Biol Pharm Bull. 2005;28:
1173 773–778.

1174 63. Gietz RD, Schiestl RH. High-efficiency yeast transformation using the LiAc/SS
1175 carrier DNA/PEG method. Nat Protoc. 2007;2: 31–34.

1176 64. Li Z, Vizeacoumar FJ, Bahr S, Li J, Warringer J, Vizeacoumar FS, et al. Systematic
1177 exploration of essential yeast gene function with temperature-sensitive mutants. Nat
1178 Biotechnol. 2011;29: 361–367.

1179 65. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq:
1180 paired-end assembler for illumina sequences. BMC Bioinformatics. 2012;13: 31.

1181 66. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open

1182 source tool for metagenomics. PeerJ. 2016;4: e2584.

1183 67. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF
1184 ChimeraX: Structure visualization for researchers, educators, and developers.
1185 Protein Sci. 2021;30: 70–82.

1186 68. Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: working with RNA,
1187 small molecules and a new graphical interface. Bioinformatics. 2019;35: 4168–
1188 4169.

1189 69. Usmanova DR, Bogatyreva NS, Ariño Bernad J, Eremina AA, Gorshkova AA,
1190 Kanevskiy GM, et al. Self-consistency test reveals systematic bias in programs for
1191 prediction change of stability upon mutation. Bioinformatics. 2018;34: 3653–3658.

1192 70. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al.
1193 ROSETTA3: an object-oriented software suite for the simulation and design of
1194 macromolecules. Methods Enzymol. 2011;487: 545–574.

1195 71. Gabler F, Nam S-Z, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein
1196 sequence analysis using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformatics.
1197 2020;72: e108.

1198 72. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A
1199 completely reimplemented MPI Bioinformatics Toolkit with a new HHpred server at
1200 its core. J Mol Biol. 2018;430: 2237–2243.

1201 73. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al.
1202 Protein identification and analysis tools on the ExPASy server. The Proteomics
1203 Protocols Handbook. Totowa, NJ: Humana Press; 2005. pp. 571–607.

1204 74. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et
1205 al. A series of PDB related databases for everyday needs. Nucleic Acids Res.
1206 2011;39: D411-9.

1207 75. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition
1208 of hydrogen-bonded and geometrical features. Biopolymers. 1983;22: 2577–2637.

1209 76. O'Boyle NM, Brewerton SC, Taylor R. Using buriedness to improve discrimination
1210 between actives and inactives in docking. J Chem Inf Model. 2008;48: 1269–1278.

1211 77. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API
1212 design for machine learning software: experiences from the scikit-learn project.
1213 arXiv [cs.LG]. 2013. Available: http://arxiv.org/abs/1309.0238

1214 78. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv
1215 [cs.AI]. 2017. doi:10.5555/3295222.3295230

1216 79. Maitte C, Leterrier M, Le Pape P, Miegeville M, Morio F. Multilocus sequence typing
1217 of Pneumocystis jirovecii from clinical samples: how many and which loci should be

1218 used? J Clin Microbiol. 2013;51: 2843–2849.

1219 80. Yachie N, Petsalaki E, Mellor JC, Weile J, Jacob Y, Verby M, et al. Pooled-matrix
1220 protein interaction screens using Barcode Fusion Genetics. Mol Syst Biol. 2016;12:
1221 863–863.

1222 81. Vasimuddin M, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of
1223 BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed
1224 Processing Symposium (IPDPS). IEEE; 2019. pp. 314–324.

1225 82. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve
1226 years of SAMtools and BCFtools. Gigascience. 2021;10: giab008.

1227 83. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing
1228 reads. EMBnet J. 2011;17: 10.

1229

Supplementary files:

S1_file.csv. MTX training and testing dataset with model predictions.
Contains dataset for training and testing the machine learning model, as well as prediction outputs from the model for PjDHFR-MTX.

S2_file.csv. TMP dataset with model predictions.
Contains dataset for making predictions on PjDHFR-TMP, as well as prediction output for PjDHFR-TMP. Data is not de-scaled.

S3_file.csv. DMS screening data of the function assay in permissive condition.
Parsed read data in long format for screening done at 20 °C for all fragments in all strains, in codon format.

S4_file.csv. DMS screening data of the function assay in selective condition.
Parsed read data in long format for screening done at 37 °C for all fragments in all strains, in codon format.

S5_file.fasta. DHFRs aligned using HHblits with gaps removed.
HHBlits DHFR alignment with gaps removed for GEMME analysis. All DHFRs have < 70% homology to PjDHFR to not skew GEMME output with highly similar sequences.

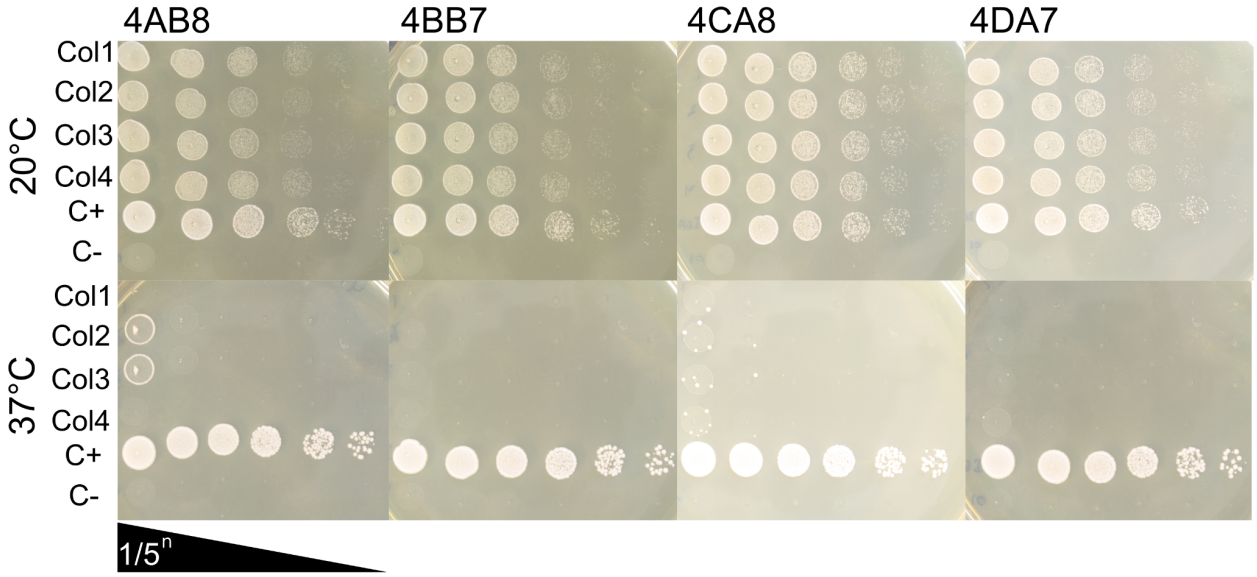
S6_file.fasta. DHFR aligned using HHblits with gaps presents.
HHBlits DHFR alignment with gaps present. All DHFRs have < 70% homology to PjDHFR to not skew GEMME output with highly similar sequences.

Supplementary tables:

S1_table.csv: Description and reference for all features used in model training.
S2_table.csv: Compilation of DHFR mutations leading to drug resistance found in literature, as well as mutations found in this study.
S3_table.csv: Details on strains used in this study.
S4_table.csv: Details on plasmids used in this study.
S5_table.csv: Details on oligonucleotides used in this study.
S6_table.csv: Details on media used in this study.
S7_table.csv: Details on PCR cycles used in this study.

1261 **Supplementary figures:**

1262

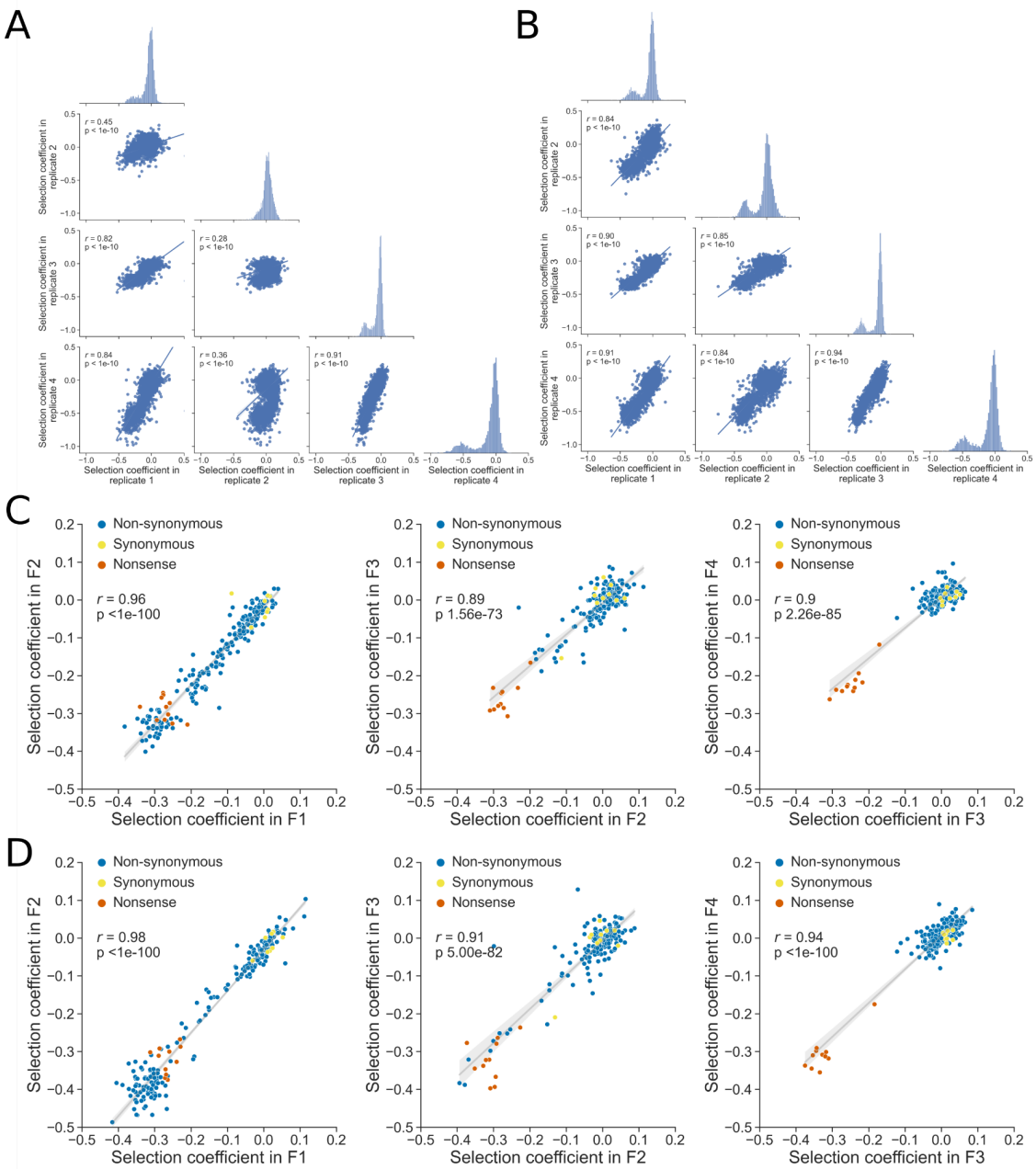


1263

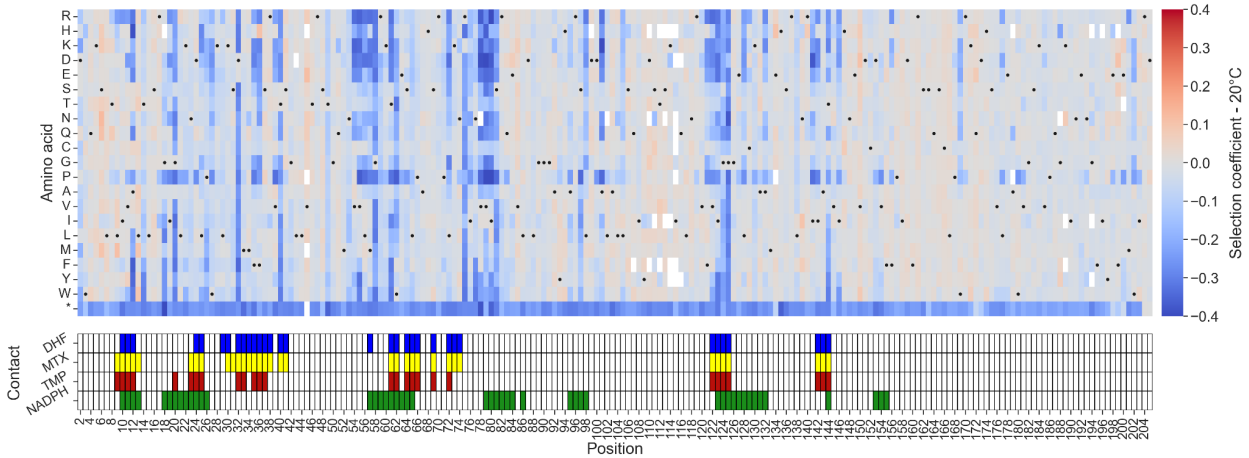
1264 **Supplementary figure 1: Spot-dilution assay to validate temperature-dependent growth of the four**
1265 **TS-DHFR strains.** From each TS-DHFR strain from the temperature sensitive allele collection [64],
1266 samples from the collection glycerol stock were spread on YPD+G418 plates. From each biological
1267 replicate, four isolated colonies were selected and grown overnight in YPD+G418 liquid media. Cultures
1268 were then diluted to 1.0 OD_{600} and spotted on YPD+G418 plates with 5-fold serial dilutions. One plate was
1269 incubated at 20 °C (permissive condition, top) and at 37 °C (selection condition, bottom) for 72h. On each
1270 plate, strain IGA130 (*DHFR*-WT) and strain FDR0001 (*Δdhfr,* inducible DfrB1 to allow for growth) were used
1271 as positive and negative growth controls, respectively [36].

1272

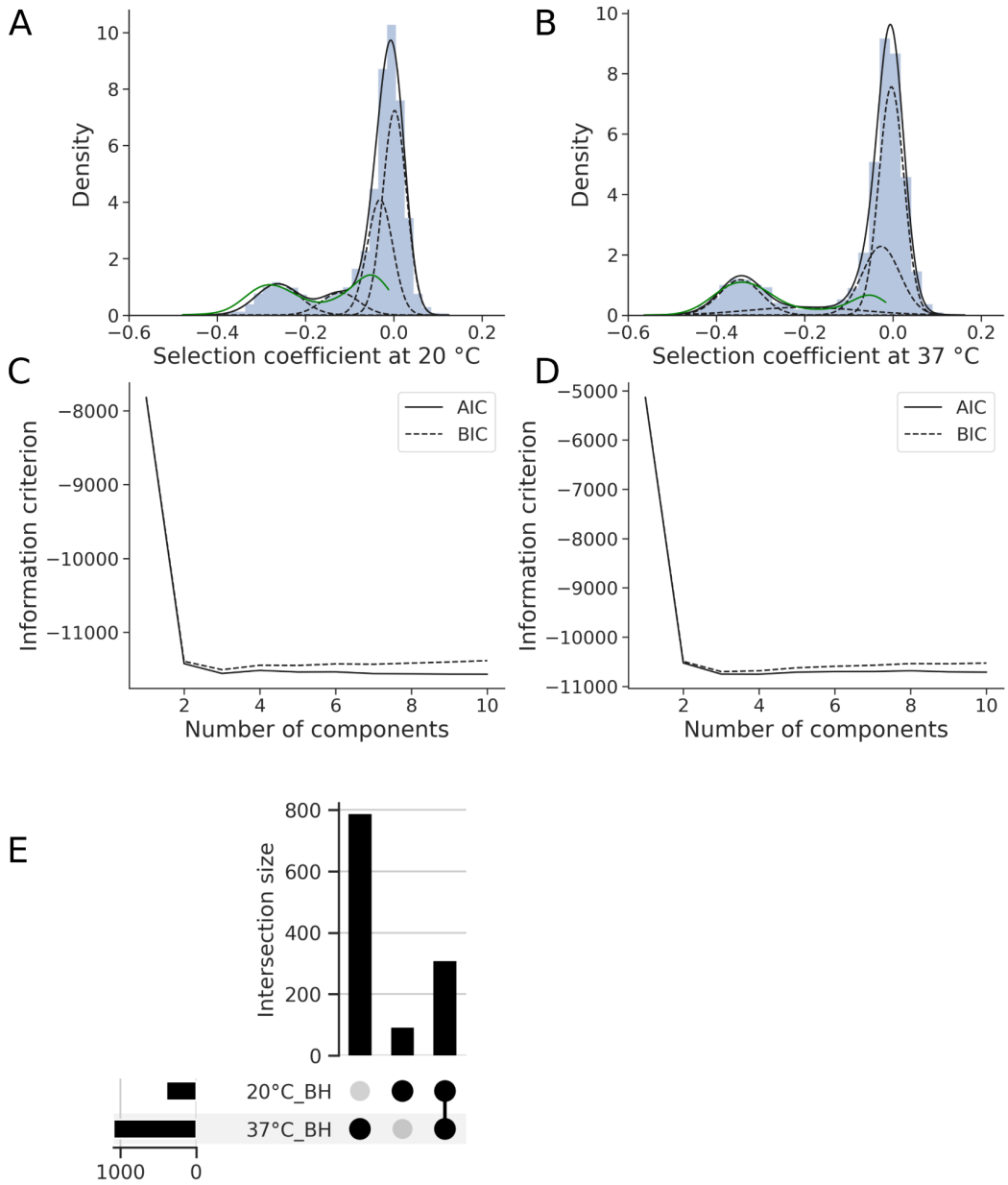
1273



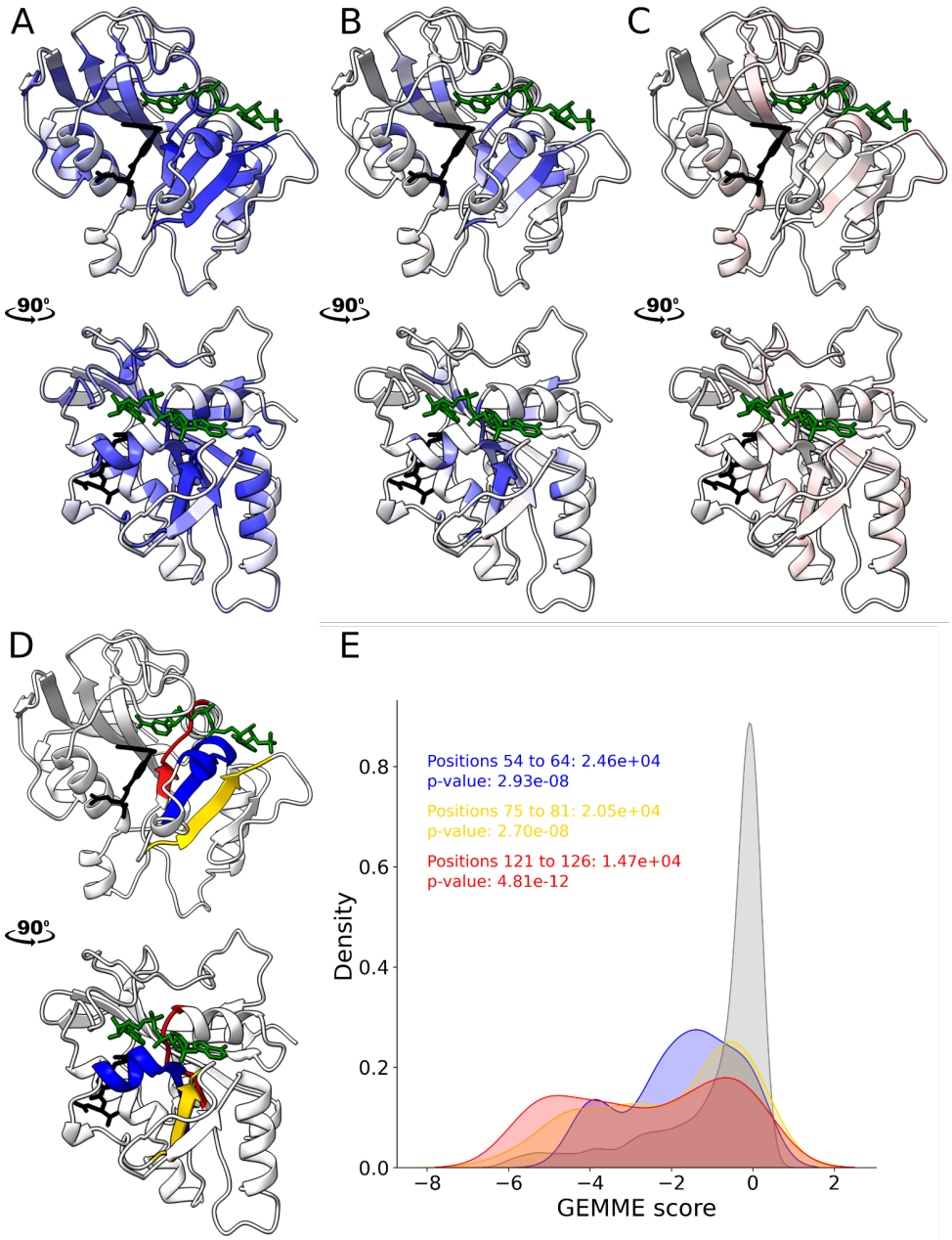
Supplementary figure 2: Correlation among biological replicates of screens at 20 °C and 37 °C. A) Scatterplots of selection coefficients for mutated codons in the different biological replicates at 20 °C. Axes are scaled to allow comparison of the different conditions. Density plots show the distribution of selection coefficients for the replicate on the x-axis. B) Scatterplots of selection coefficients for mutated codons in the different biological replicates at 37 °C. Axes are scaled to allow comparison of the different conditions. C) Scatterplots of selection coefficients for amino acids present in the overlap between the different fragments at 20 °C to ensure good correlation between the different fragments. D) Scatterplots of selection coefficients for amino acids present in the overlap between the different fragments at 37 °C to ensure good correlation between the different fragments. Scatterplots for each fragment overlap show similar patterns at 20 °C and at 37 °C.



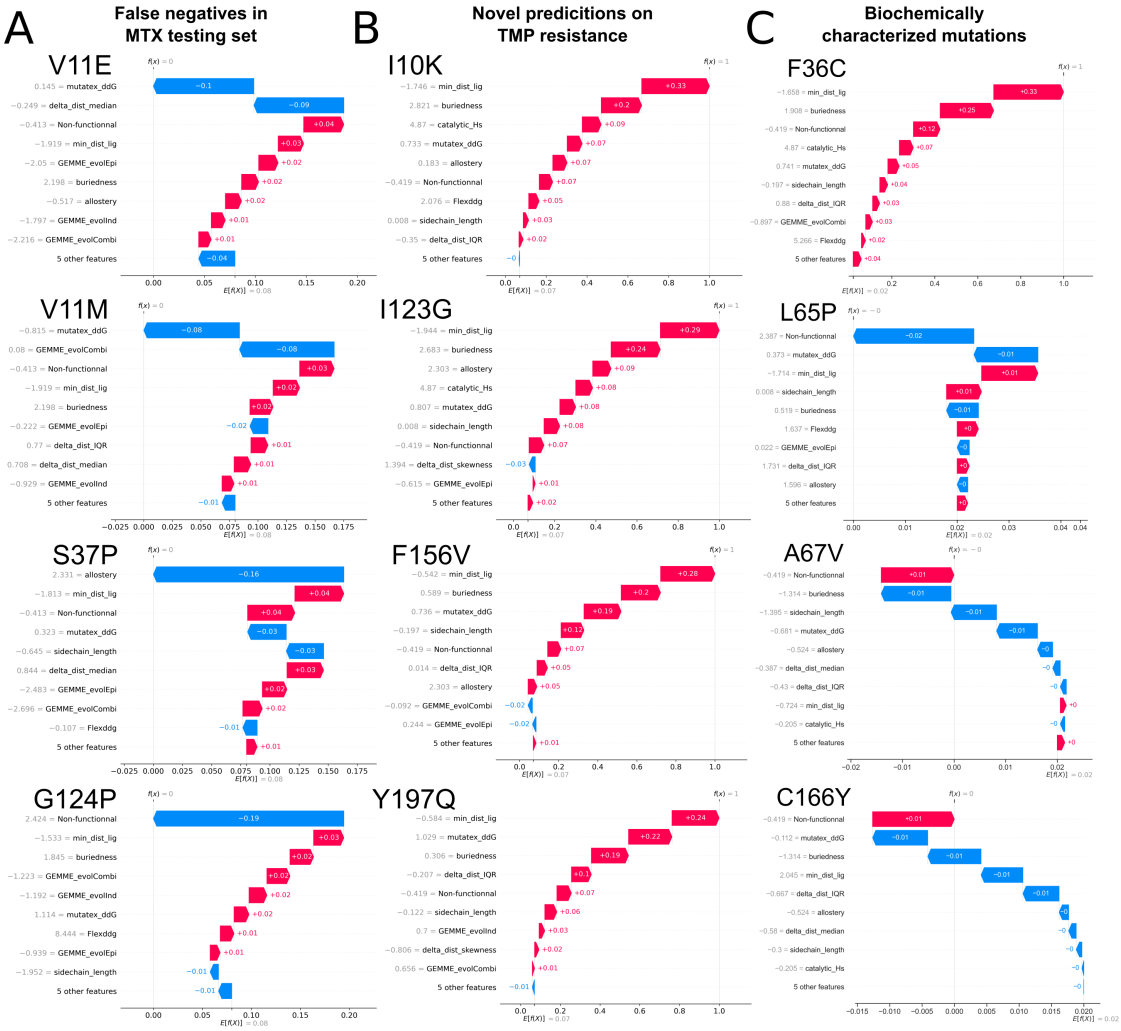
Supplementary figure 3: Heatmap of selection coefficients for screening done at 20 °C. Selection coefficient is measured in relation to the median of silent mutations. Positions of contacting residues along PjDHFR (PDB: 3CD2 (MTX and NADPH), 4CD2 (DHF) and 1DYR (TMP)). Contact was established as amino acids with an α-carbon located less than 8 Å from MTX, DHF, TMP or NADPH. Black dots represent the wild-type sequence



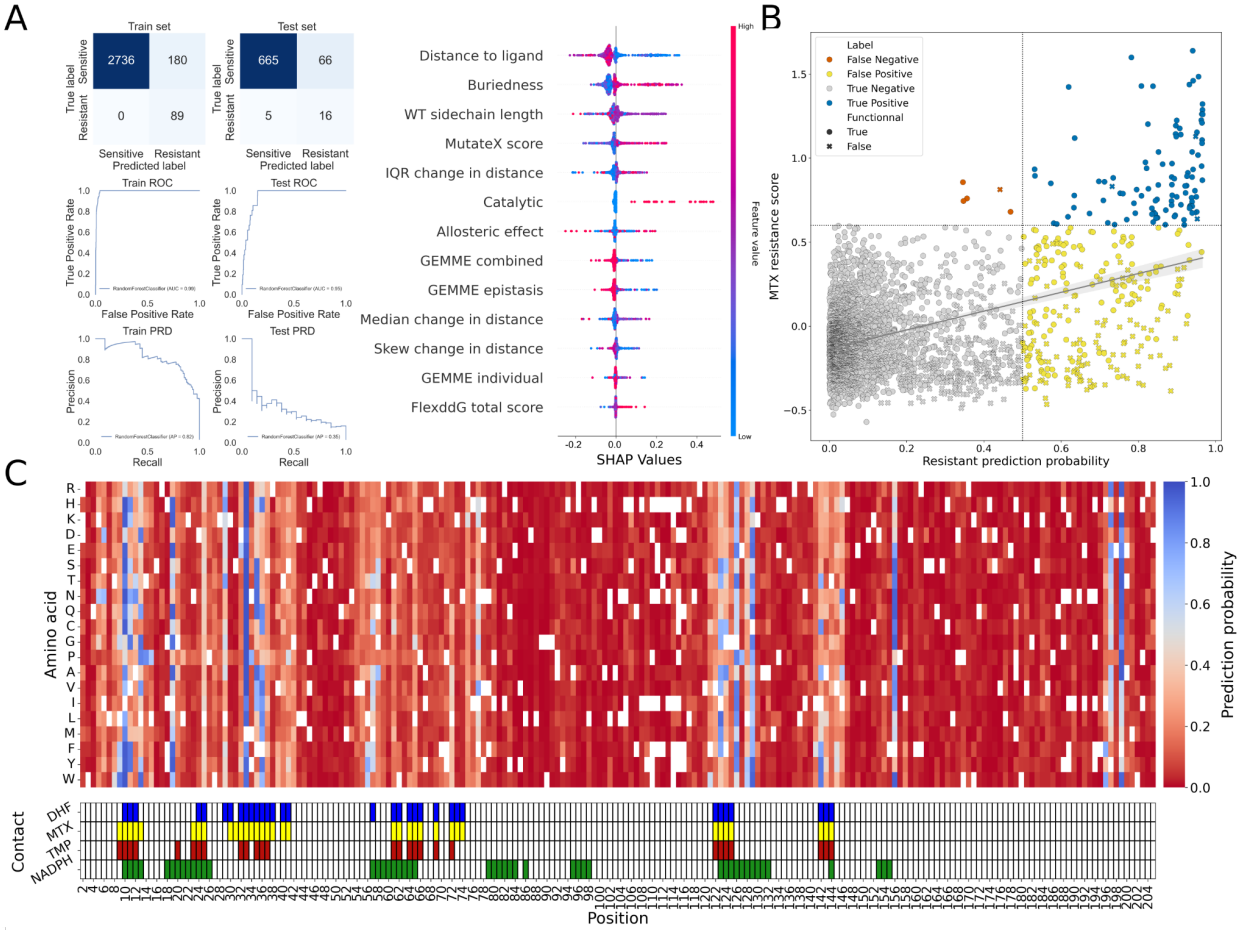
Supplementary figure 4: Gaussian mixture model optimization for the permissive condition (20 °C) and non-permissive condition (37 °C). Optimization of information criteria for A) 20 °C and B) 37 °C. Best Gaussian mixture model (dashed lines represent underlying Gaussians) to recapitulate the underlying distribution (black line/ histogram) of C) 20 °C and D) 37 °C. Density curve of significant mutants for Benjamini-Hoch- berg (green) correction is visible. E) Intersects between statistically significant mutants in 20 °C and 37 °C. Upset plot showing the intersections of the Benjamini Hochberg-FDR corrected groups (control for the false discovery rate of significantly resistant mutants at 95% confidence).



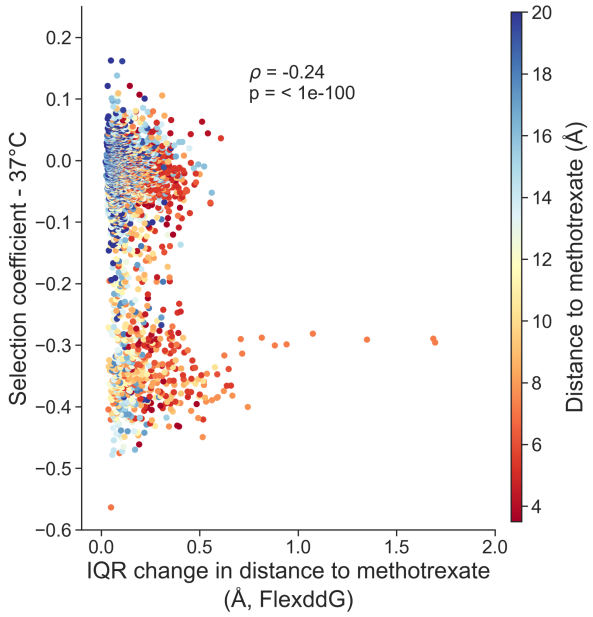
Supplementary figure 5: Mutation effect on function at 37 °C mapped on structure and analysis of region with a proportionally higher rate of LOF. A) Position-wise minimum, colored as in figure 2A. B) Position-wise median score. C) Position-wise maximum score. D) Visualization of the region with a high rate of LOF. Positions 54 to 64 are colored in blue, positions 75 to 81 in yellow and 121 to 126 in red. E) Distribution of GEMMECombined scores for all positions in grey, positions 54 to 64 are colored in blue, positions 75 to 81 in yellow and 121 to 126 in red. One-sided Mann-Whitney U scores and p-values for each distribution are presented. Mutations at these regions are predicted as being more likely to be deleterious than most mutations across the protein. NADPH is colored in green, and folate is colored in black.



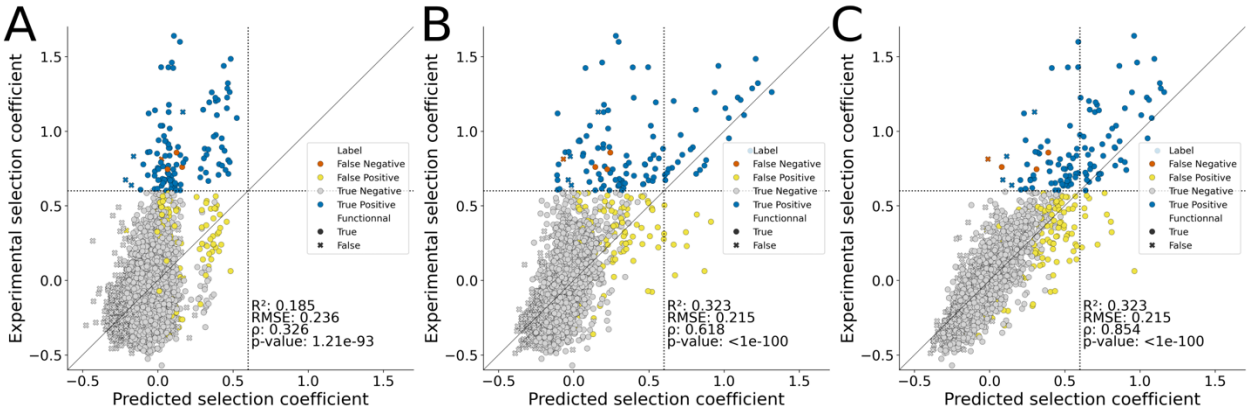
Supplementary figure 6: Waterfall plots obtained from the SHAP explainer values for mutants of interest. A) Waterfall plots detailing classification decisions for mutants that were classified as false negatives in the testing dataset for MTX resistance (leading to resistance experimentally but not being predicted as leading to resistance by the model.) V11E (Prediction probability of leading to MTX resistance of 0.33) and V11M (0.36) are both predicted as being destabilizing by MutateX, S37P (0.36) misclassification is mainly driven by its classification as a high confidence allosteric mutation by [44], and G124P (0.25) was classified as being non-functional. B) Waterfall plots detailing classification decisions for mutants that were predicted by the models as leading to TMP resistance. I10K (probability of 0.94), F121H (0.52), I123G (0.82), and Y197Q (0.81) are all novel predictions made by the model (Not measured/predicted as leading to MTX resistance but predicted as leading to TMP resistance. C) Waterfall plots detailing classification decisions for mutants that were observed either in this or in previous study and have been linked to TMP resistance in PjDHFR. F36C (Prediction probability of leading to TMP resistance of 0.91) was predicted by the model as leading to TMP resistance (and not observed as leading to MTX resistance), but not L65P (0.33), A67V (0.03) or C166Y (0.01). L65P was experimentally identified as being non-functional. A67V was classified as not leading to TMP resistance mainly because of its buriedness level, as well as effect on protein stability. C166Y was classified as not leading to TMP resistance because of its distance to the ligand, buriedness level and effect on protein stability. All SHAP values are calculated on z-score scaled data.



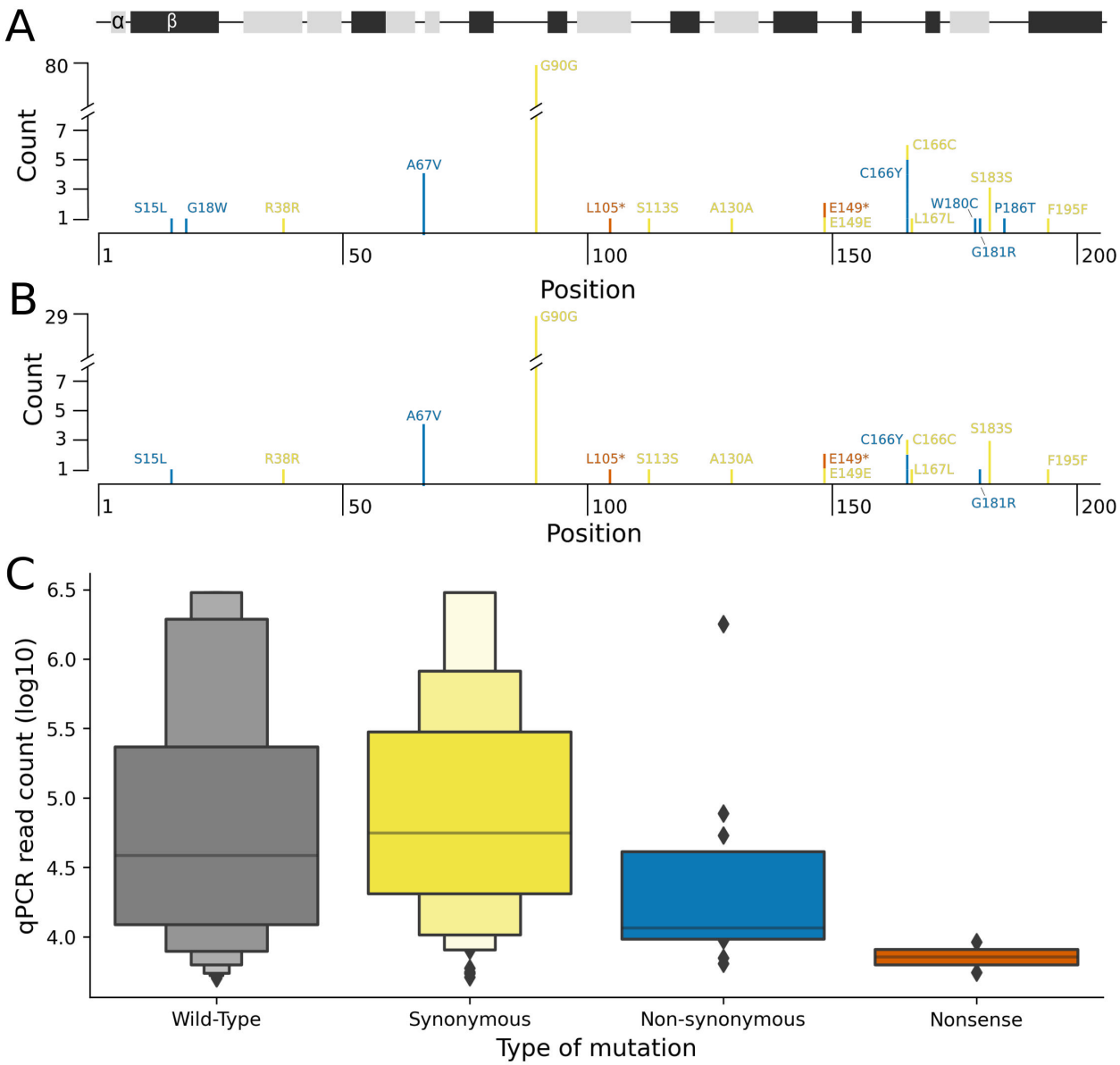
Supplementary figure 7: Visualization of model performance when function-related features are taken out of the model. A) Confusion matrices and ROC/PRD curves for train and testing set splits, with SHAP value explainer plot, with relative impact of features on model decision. True positives and false negatives stayed mostly consistent with the complete model, but much more false positives were predicted, highlighting the importance of functional information B) Scatterplot comparing ground truth from MTX DMS screening vs model's predicted probability. True negatives in grey, true positives in blue, false positives in yellow and false negatives in orange. Mutations that were identified as functional in the function DMS screening have round markers and mutations identified as non-functional have X-shaped markers. X-shaped markers are much more numerous in the false positive quadrant C) Heatmap with prediction probability of being resistant to MTX. Mutations with probability >0.5 are identified as resistant, and mutations <= 0.5 are classified as sensitive.



Supplementary figure 8: Interquartile range of the change in distance between the residue's α-carbon and the closest heavy atom on the ligand. For each structure generated for all mutations during FlexddG saturation mutagenesis, the median position of the α-carbon of the wild-type structures was used as the baseline. This was compared to the position of the corresponding α-carbon of the mutated residue of all 100 structures for each mutant, resulting in a distribution of changes in distance between the wild-type α-carbon and the corresponding mutant's α-carbon. From this distribution, we extracted the interquartile range of the distribution, representing the "flexibility" of the protein region caused by the mutation. A wide distribution (high IQR) means the mutated residue increases region flexibility, as the mutated region can occupy a broader range of positions further from its wild-type position.



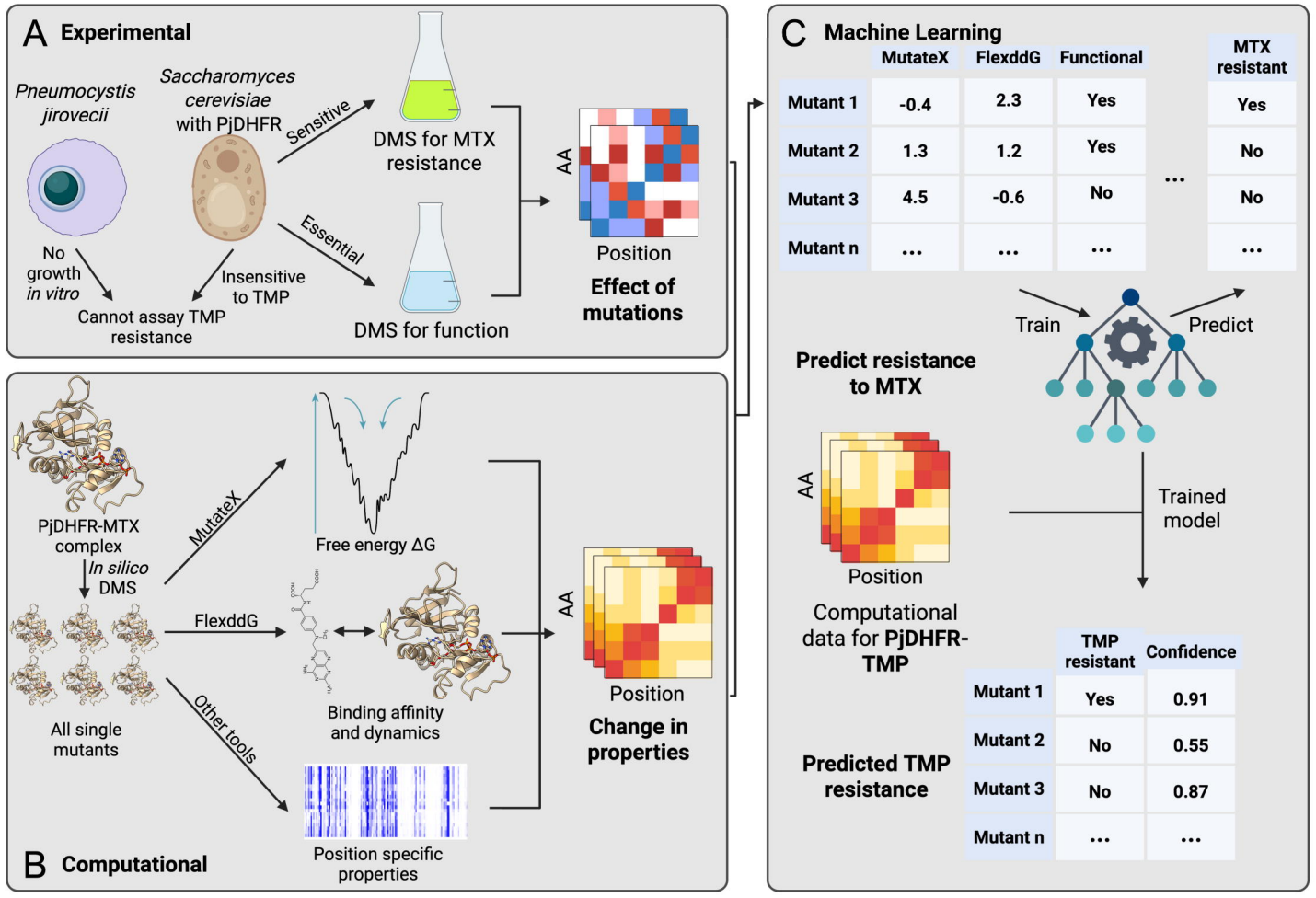
Supplementary figure 9: Summary figures of regression-based models' performance. Regression models trained in parallel to their classification counterparts and compared to the best performing model's classification of mutants. A) For a simple Linear Regression model, B) a Support Vector Regression (SVR) model and C) for a Random Forest Regression (RFR) model. The best performing model was the RFR model, as with classification. The RFR model was more consistently linear, but despite predicting less false positives, also tended to yield more false negative predictions. All correlations are Spearman's *ρ.*

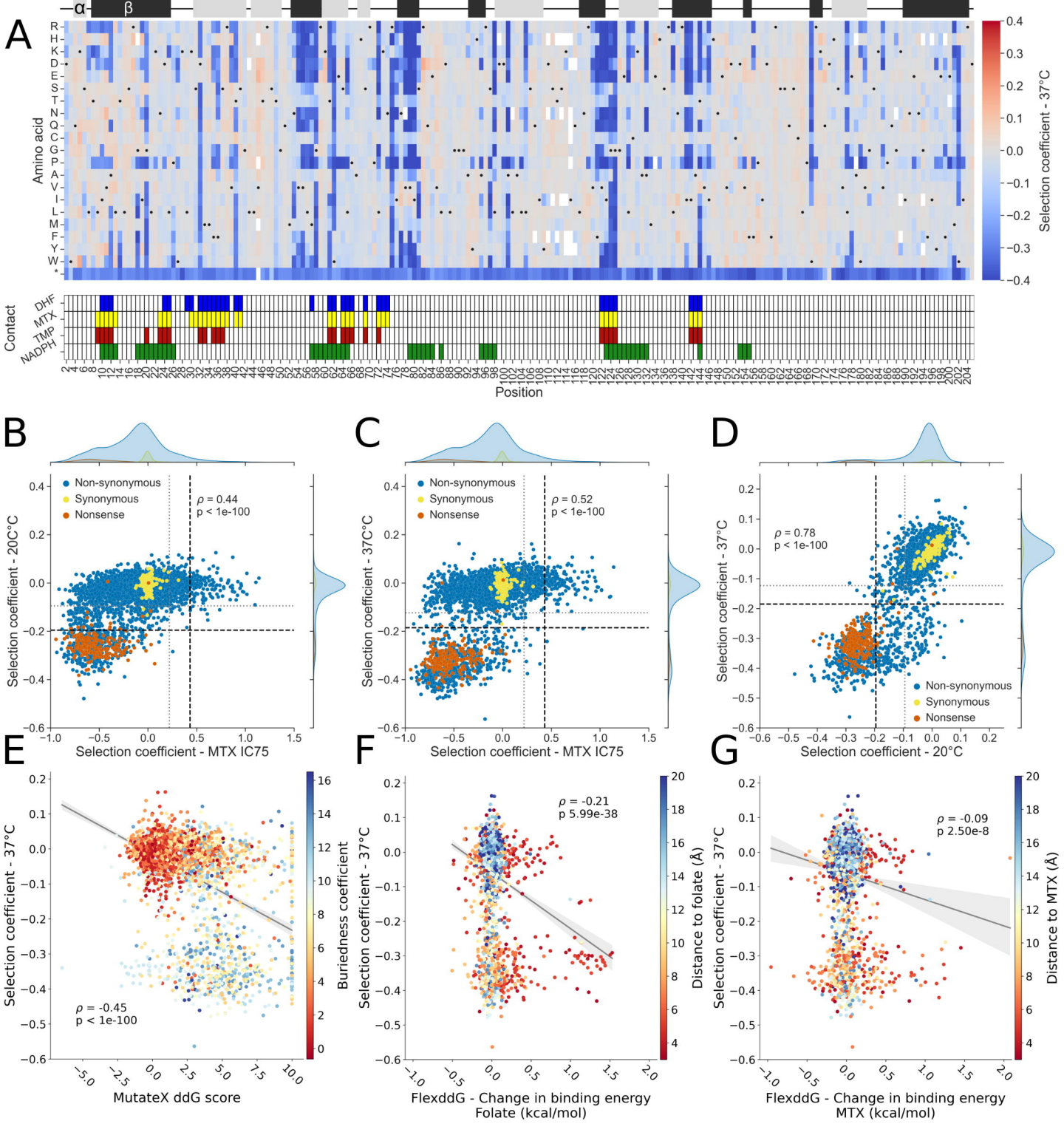


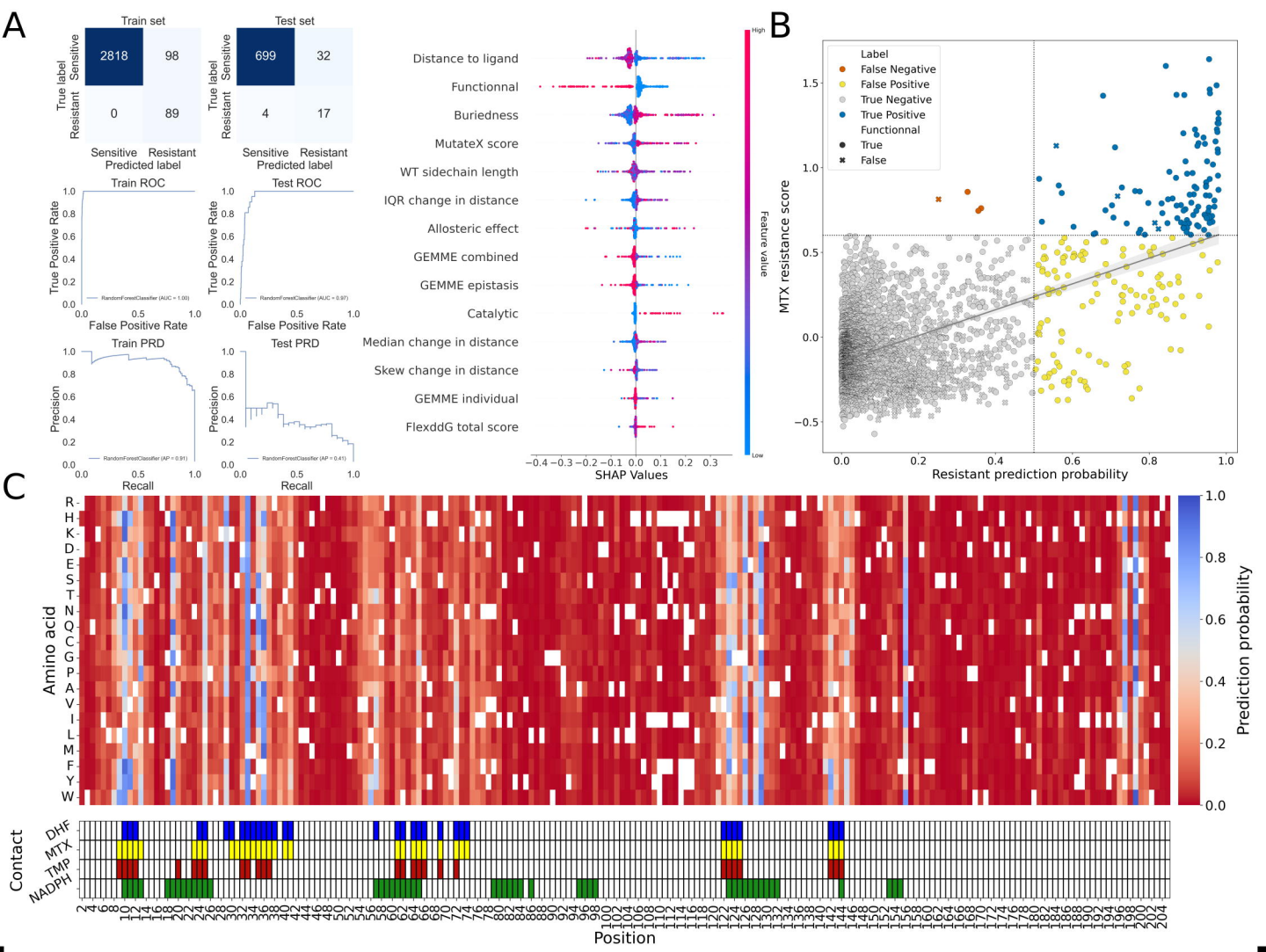
Supplementary figure 10: Nucleotide substitutions regarding the RU7 reference sequence found in PjDHFRs isolated from clinical cases in Eastern Canada. A) Position and count of all mutations found during sequencing of 531 PjDHFRs. The y axis is discontinuous because of the high prevalence of a silent mutation at position 90. The track at the top of the figure represents secondary structures along the sequence, with α-helices in light grey and beta strands in dark grey. Only the coding sequence (spliced sequence) was considered for this representation. B) Positions and counts of all mutations found in samples where more than one PjDHFR allele was observed. Most mutations identified were accompanied by a wild-type allele. C) Boxplot representing the different mutation types observed as well as their accompanying qPCR counts. No mutation type has significantly greater qPCR counts than the distribution counts from wild-type sequences (Welch's one-sided t-test, unequal variance.)

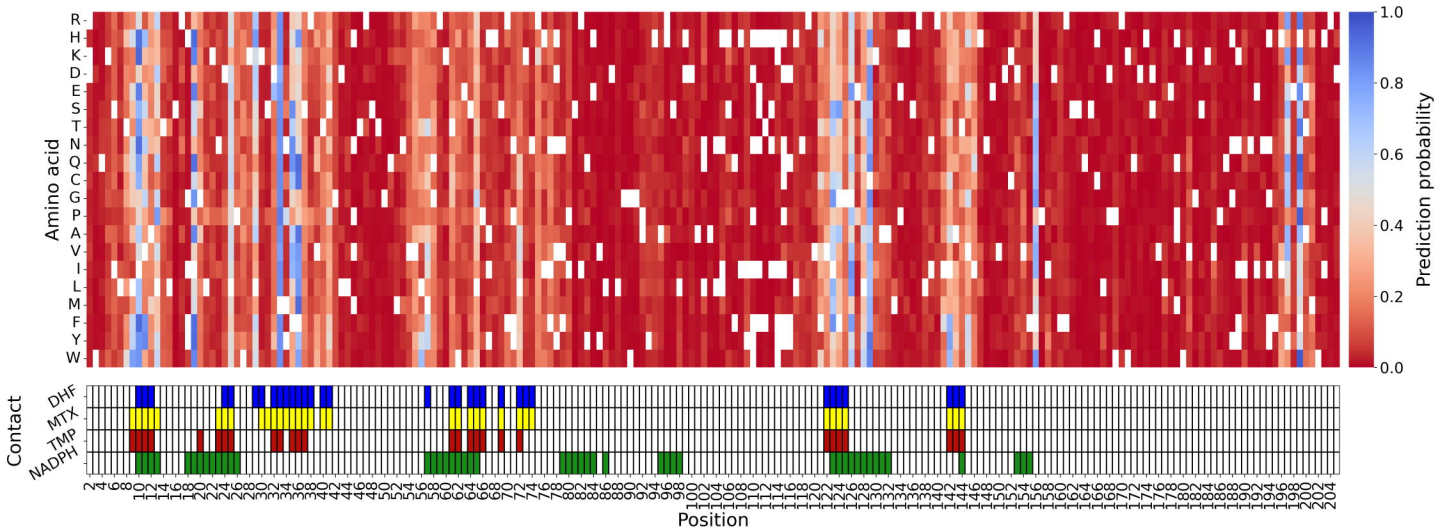
Feature	Type of data	Tool/Method	Ligand dependency	Source
Change in binding free energy between protein and ligand	Computational	FlexddG	Ligand-dependent	[33,34]
Change in protein folding free energy	Computational	MutateX	Ligand-independent	[32]
Distance between α-carbon and closest ligand heavy atom	Computational	Custom script	Ligand-dependent	[35]
Length of wild-type amino acid sidechain	Computational	From protein sequence	Ligand-independent	[70]
Residue buriedness	Computational	Buriedness computation	Ligand-independent	https://github.com/rodogi/buriedness
GEMME individual score	Computational	GEMME	Ligand-independent	[36]
GEMME epistasis score	Computational	GEMME	Ligand-independent	[36]
GEMME combined score	Computational	GEMME	Ligand-independent	[36]
Median change distance between α-carbon and closest ligand heavy atom caused by amino acid substitution	Computational	FlexddG, custom script	Ligand-dependent	This study
Interquartile range of change distance between α-carbon and closest ligand heavy atom caused by amino acid substitution	Computational	FlexddG, custom script	Ligand-dependent	This study
Skewness of change distance between α-carbon and closest ligand heavy atom caused by amino acid substitution	Computational	FlexddG, custom script	Ligand-dependent	This study
Participation in catalysis	Experimental	Several methods	Does not apply	[45]
Allostery confidence score	Experimental	Photosensitive domain	Does not apply	[46]
Selection coefficient in selection condition for function (37 °C)	Experimental	DMS competition assay	Does not apply	This study

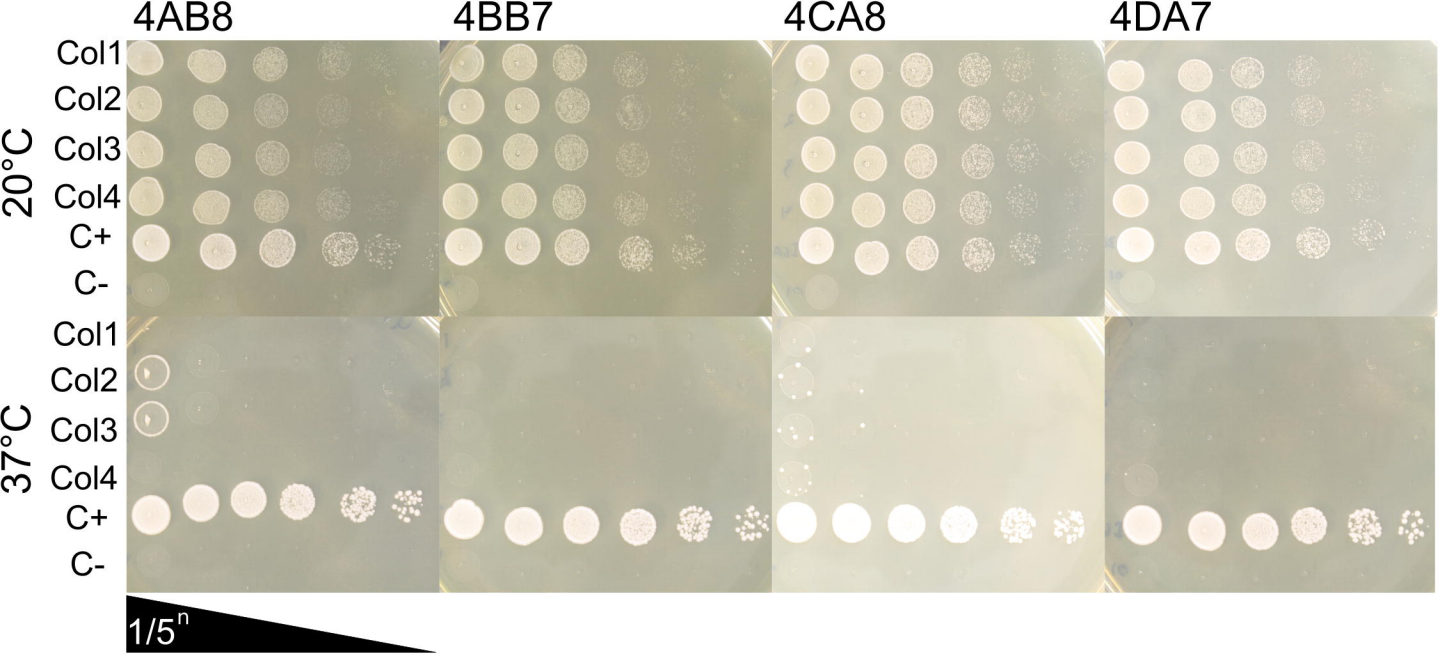
Supplementary table 1: Description of all features used in model training, testing and predictions. Features are grouped by the type of data acquired, the tool or method used to acquire the data and their status of ligand-dependency. The reference for all tools or data source is also available.

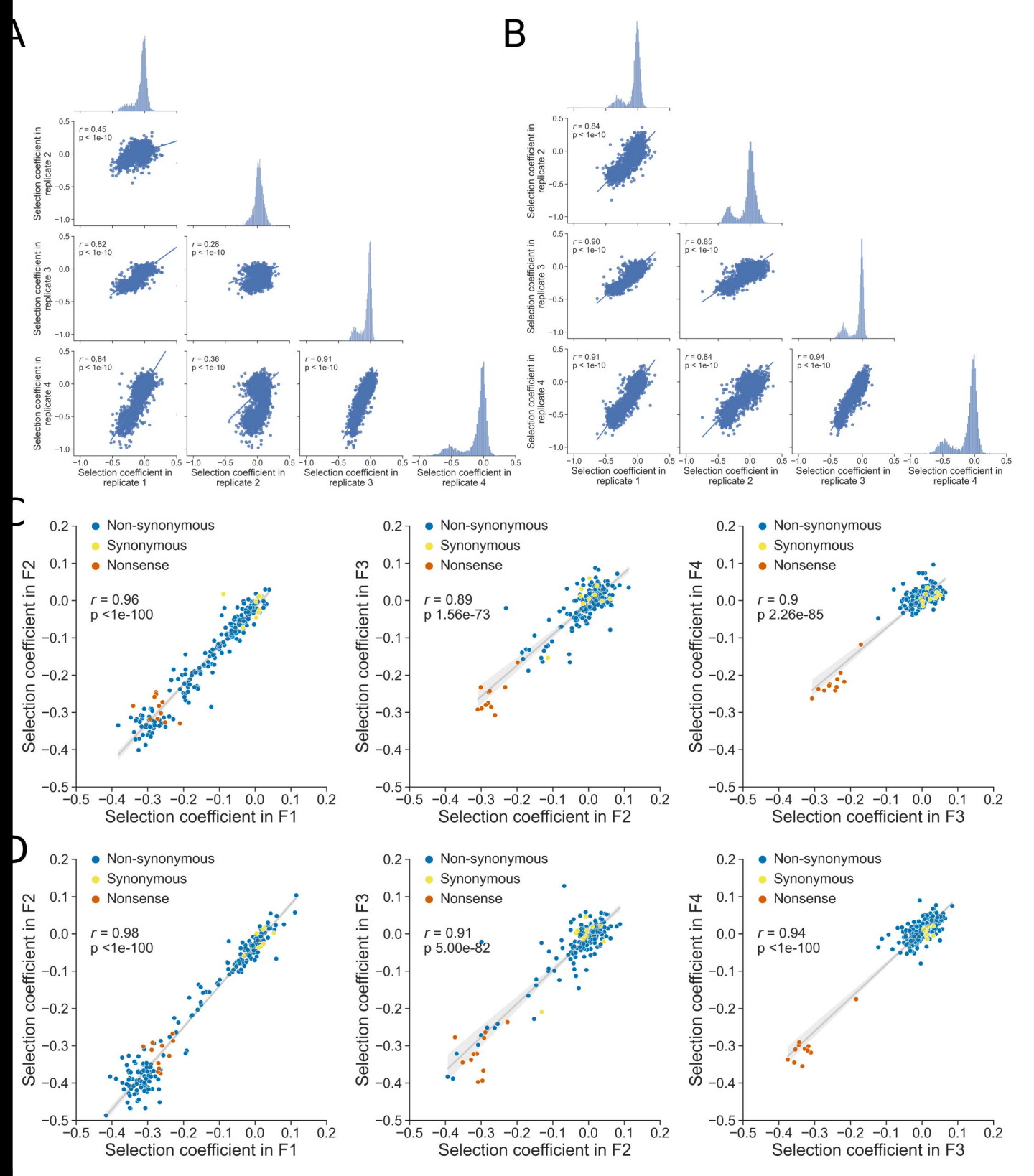

A Experimental
Pneumocystis jirovecii
Saccharomyces cerevisiae with PjDHFR
Sensitive
Essential
No growth in vitro
Insensitive to TMP
Cannot assay TMP resistance
DMS for MTX resistance
DMS for function
AA
Position
Effect of mutations
B Computational
PjDHFR-MTX complex
In silico DMS
All single mutants
MutateX
FlexddG
Other tools
Free energy ΔG
Binding affinity and dynamics
Position specific properties
AA
Position
Change in properties
C Machine Learning
MutateX
FlexddG
Functional
MTX resistant
Mutant 1
-0.4
2.3
Yes
Yes
Mutant 2
1.3
1.2
Yes
No
Mutant 3
4.5
-0.6
No
No
Mutant n
...
Train
Predict
Predict resistance to MTX
AA
Position
Trained model
Computational data for PjDHFR-TMP
TMP resistant
Confidence
Mutant 1
Yes
0.91
Mutant 2
No
0.55
Mutant 3
No
0.87
Mutant n
Predicted TMP resistance

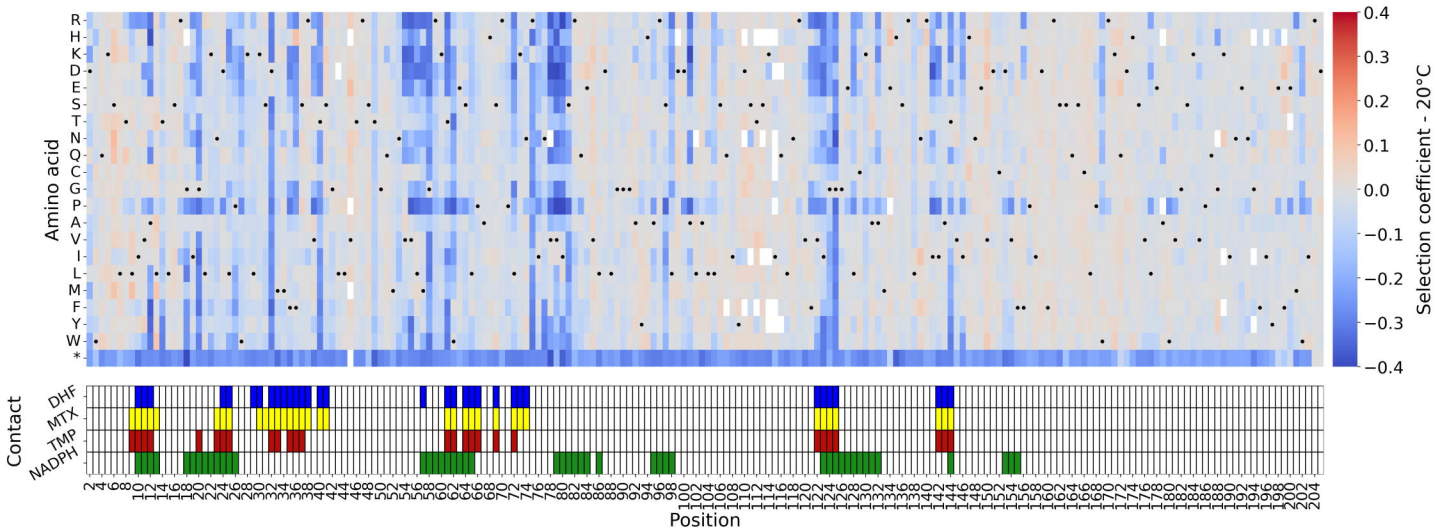

A
α
β
Amino acid
R
H
K
D
E
S
T
N
Q
C
G
P
A
V
I
L
M
F
Y
W
*
Selection coefficient - 37°C
Contact
DHF
MTX
TMP
NADPH
Position
B
Non-synonymous
Synonymous
Nonsense
ρ = 0.44
p < 1e-100
Selection coefficient - 20C°C
Selection coefficient - MTX IC75
C
Non-synonymous
Synonymous
Nonsense
ρ = 0.52
p < 1e-100
Selection coefficient - 37C°C
Selection coefficient - MTX IC75
D
ρ = 0.78
p < 1e-100
Non-synonymous
Synonymous
Nonsense
Selection coefficient - 37°C
Selection coefficient - 20°C
E
ρ = -0.45
p < 1e-100
Selection coefficient - 37°C
MutateX ddG score
Buriedness coefficient
F
ρ = -0.21
p 5.99e-38
Selection coefficient - 37°C
FlexddG - Change in binding energy Folate (kcal/mol)
Distance to folate (Å)
G
ρ = -0.09
p 2.50e-8
Selection coefficient - 37°C
FlexddG - Change in binding energy MTX (kcal/mol)
Distance to MTX (Å)

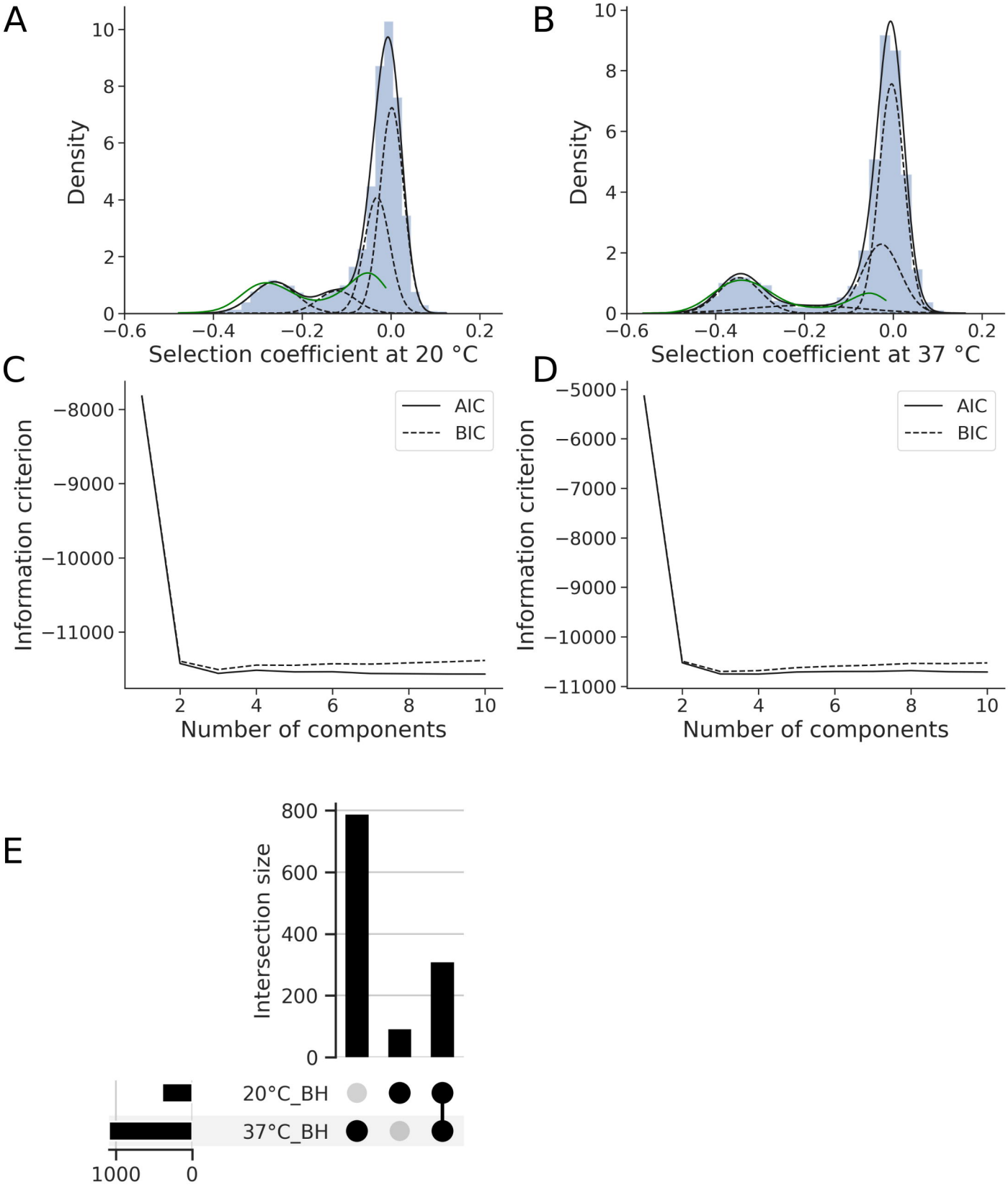

A
Train set
Test set
True label
Sensitive
Resistant
2818
98
0
89
699
32
4
17
Sensitive
Resistant
Predicted label
Train ROC
Test ROC
True Positive Rate
False Positive Rate
RandomForestClassifier (AUC = 1.00)
RandomForestClassifier (AUC = 0.97)
Train PRD
Test PRD
Precision
Recall
RandomForestClassifier (AP = 0.91)
RandomForestClassifier (AP = 0.41)
Distance to ligand
Functionnal
Buriedness
MutateX score
WT sidechain length
IQR change in distance
Allosteric effect
GEMME combined
GEMME epistasis
Catalytic
Median change in distance
Skew change in distance
GEMME individual
FlexddG total score
SHAP Values
High
Low
Feature value
B
Label
False Negative
False Positive
True Negative
True Positive
Functionnal
True
False
MTX resistance score
Resistant prediction probability
C
Amino acid
Prediction probability
Contact
DHF
MTX
TMP
NADPH
Position

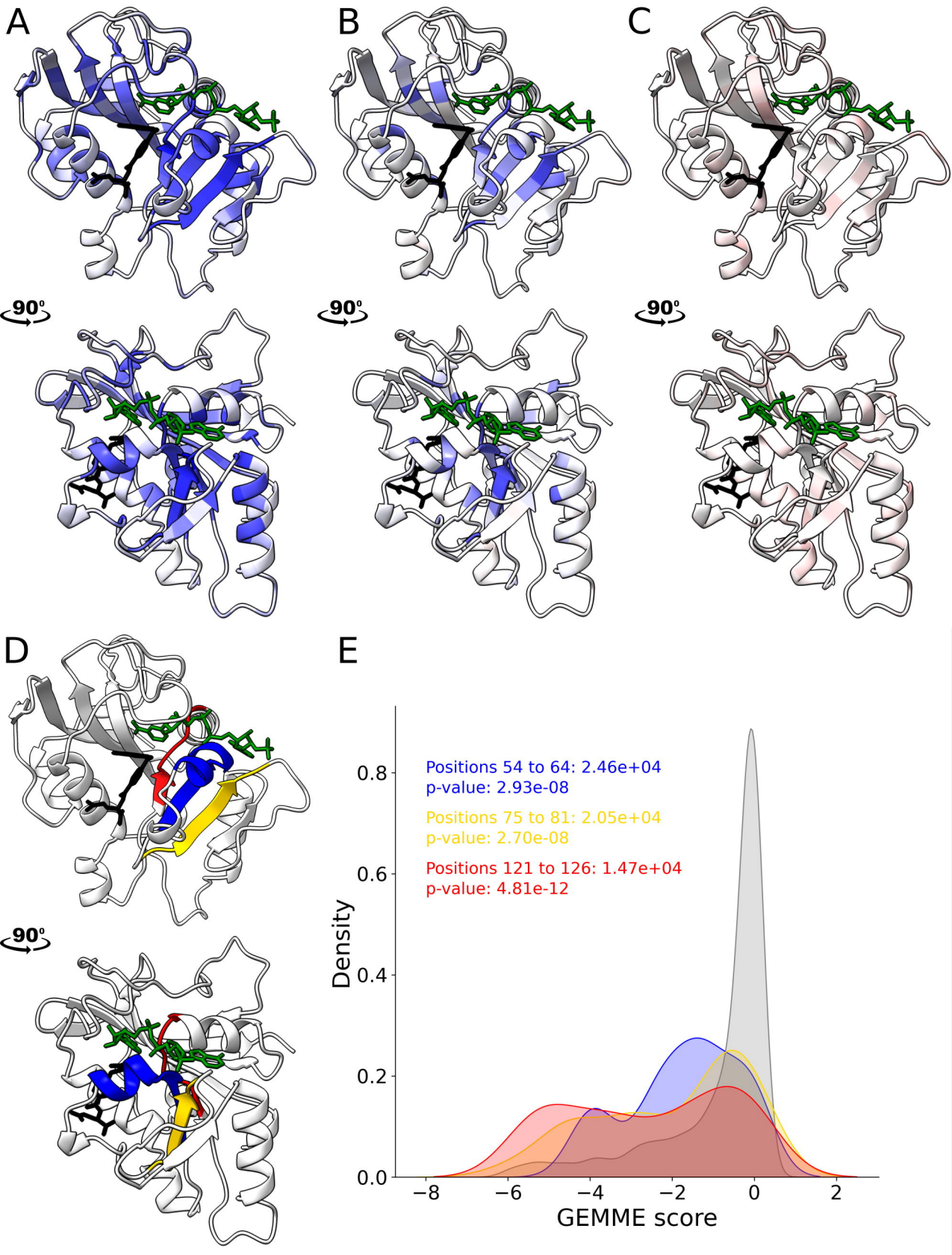

Amino acid
R
H
K
D
E
S
T
N
Q
C
G
P
A
V
I
L
M
F
Y
W
Prediction probability
1.0
0.8
0.6
0.4
0.2
0.0
Contact
DHF
MTX
TMP
NADPH
Position

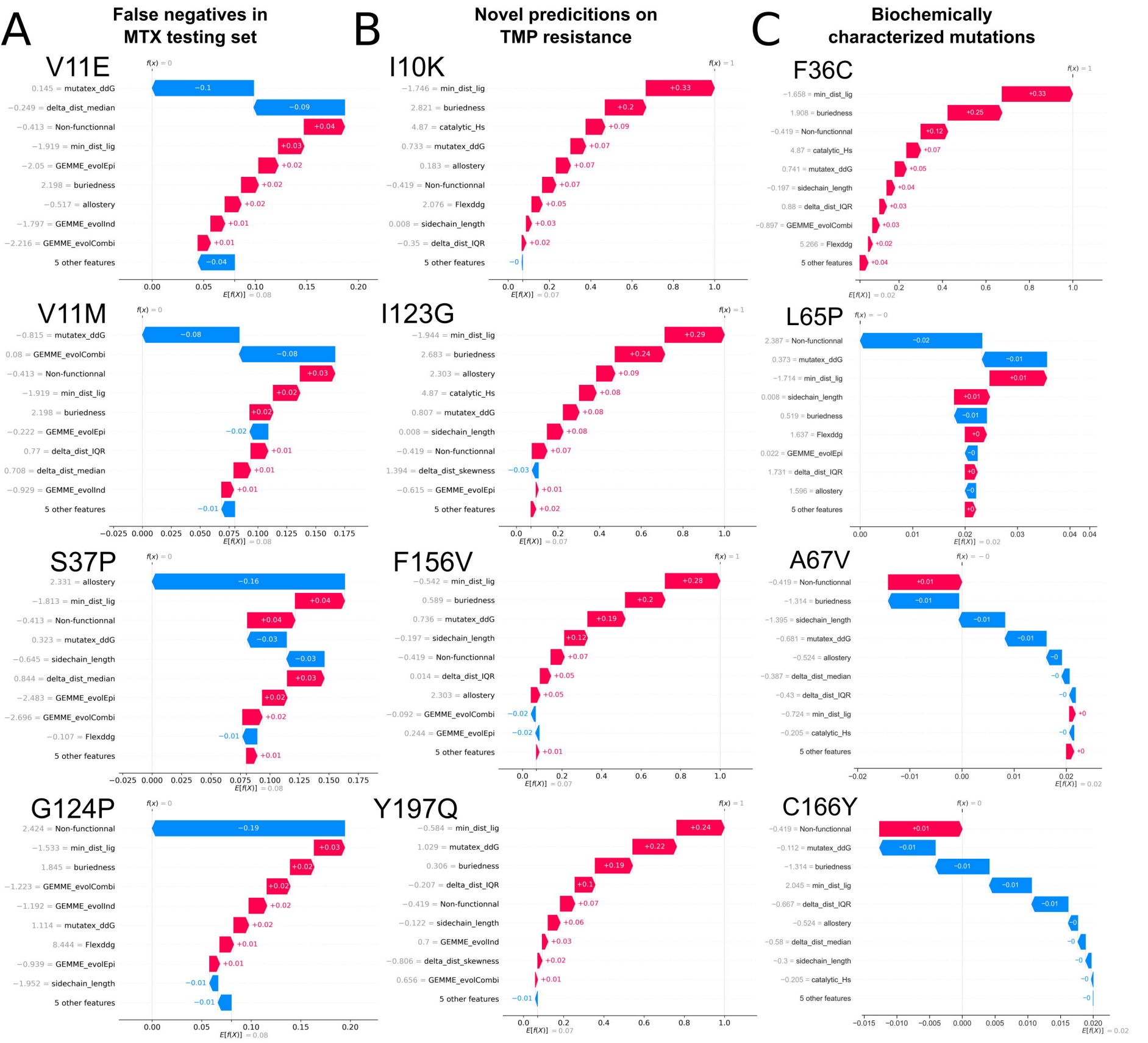

4AB8
4BB7
4CA8
4DA7
20°C
Col1
Col2
Col3
Col4
C+
C-
37°C
Col1
Col2
Col3
Col4
C+
C-
1/5n

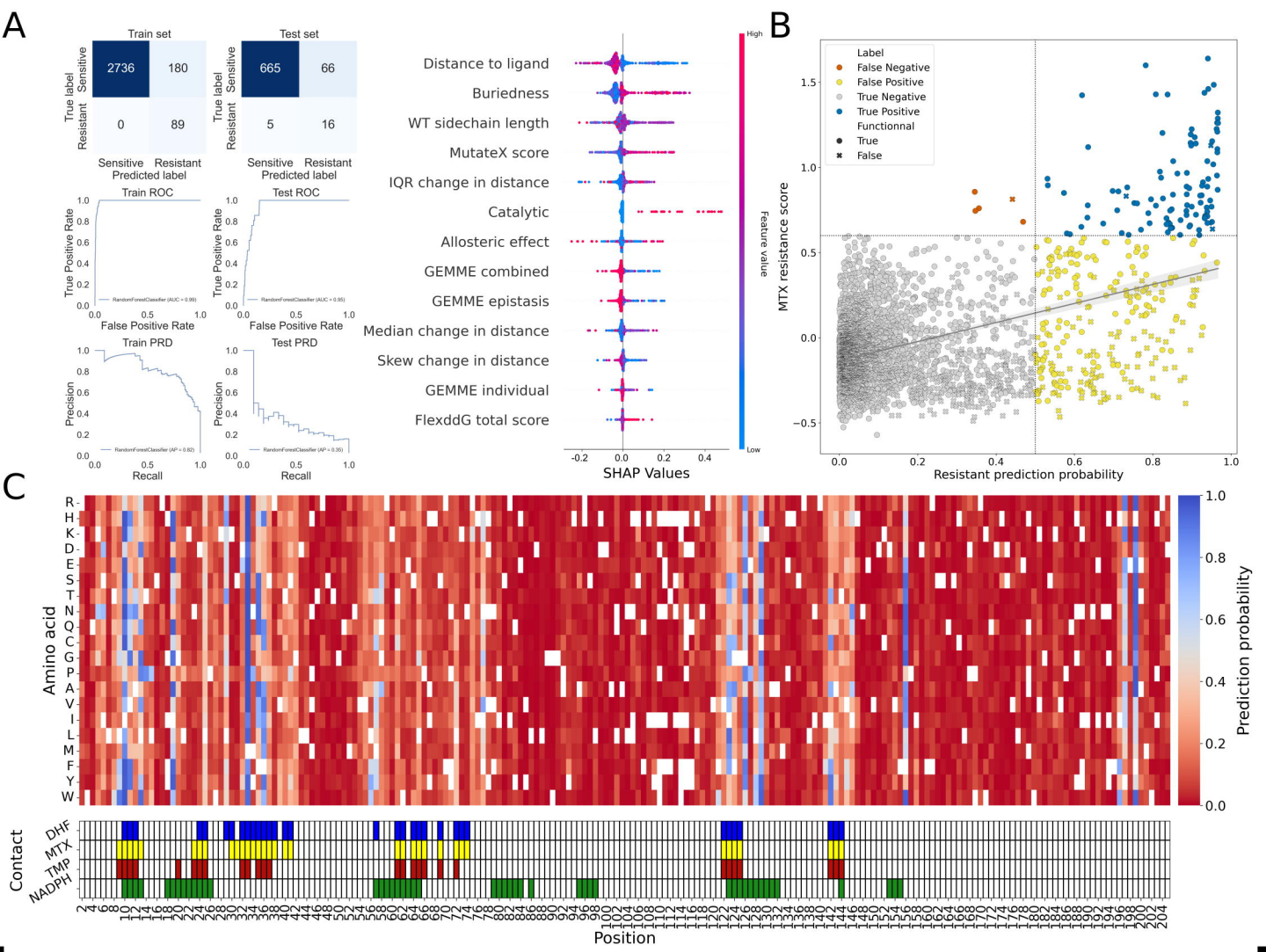

A
B
C
D
Selection coefficient in replicate 1
Selection coefficient in replicate 2
Selection coefficient in replicate 3
Selection coefficient in replicate 4
r = 0.45
p < 1e-10
r = 0.82
p < 1e-10
r = 0.28
p < 1e-10
r = 0.84
p < 1e-10
r = 0.36
p < 1e-10
r = 0.91
p < 1e-10
r = 0.84
p < 1e-10
r = 0.90
p < 1e-10
r = 0.85
p < 1e-10
r = 0.91
p < 1e-10
r = 0.84
p < 1e-10
r = 0.94
p < 1e-10
Non-synonymous
Synonymous
Nonsense
Selection coefficient in F1
Selection coefficient in F2
Selection coefficient in F3
Selection coefficient in F4
r = 0.96
p <1e-100
r = 0.89
p 1.56e-73
r = 0.9
p 2.26e-85
r = 0.98
p <1e-100
r = 0.91
p 5.00e-82
r = 0.94
p <1e-100

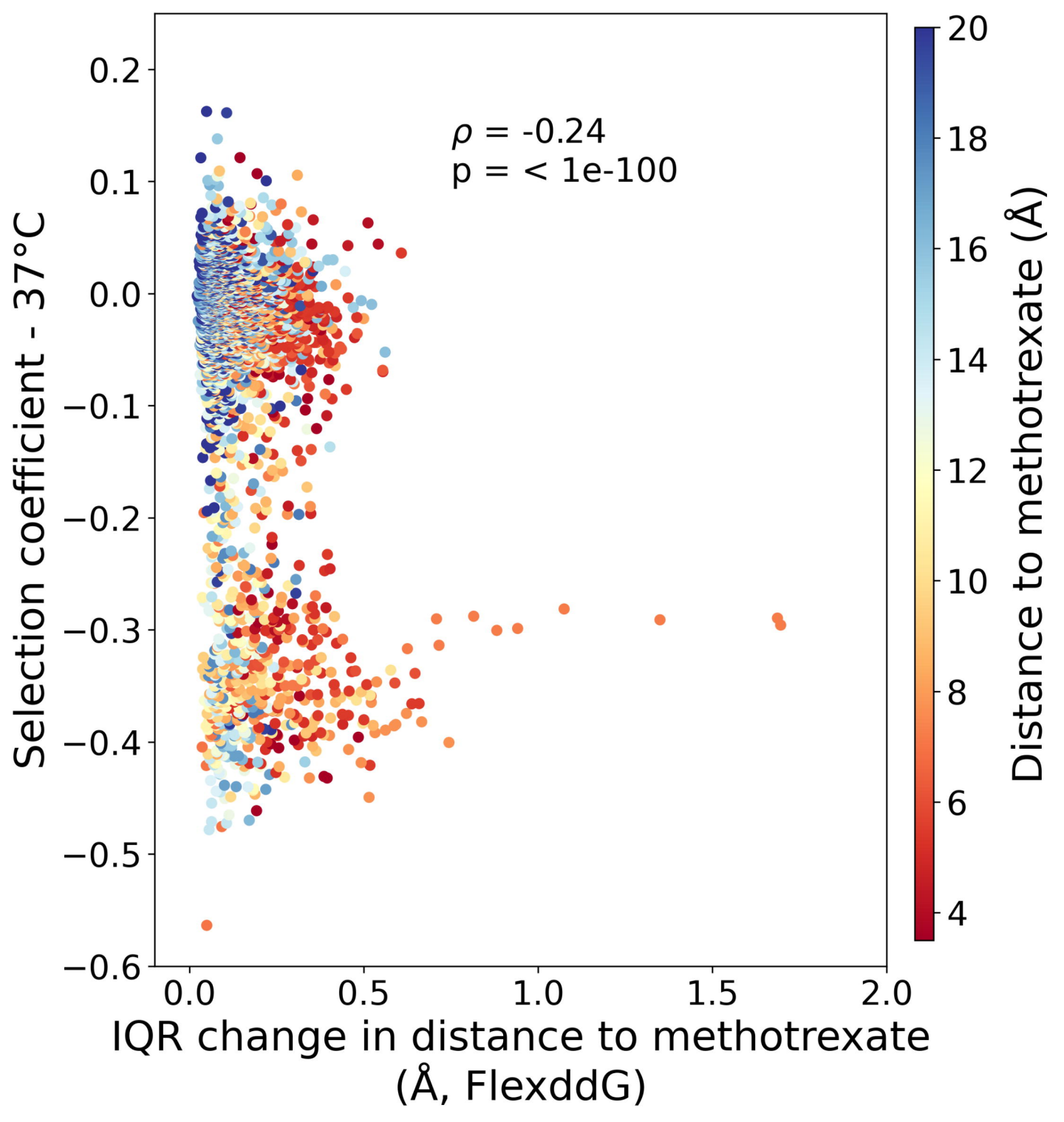

Amino acid
R
H
K
D
E
S
T
N
Q
C
G
P
A
V
I
L
M
F
Y
W
*
Selection coefficient - 20°C
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
Contact
DHF
MTX
TMP
NADPH
Position

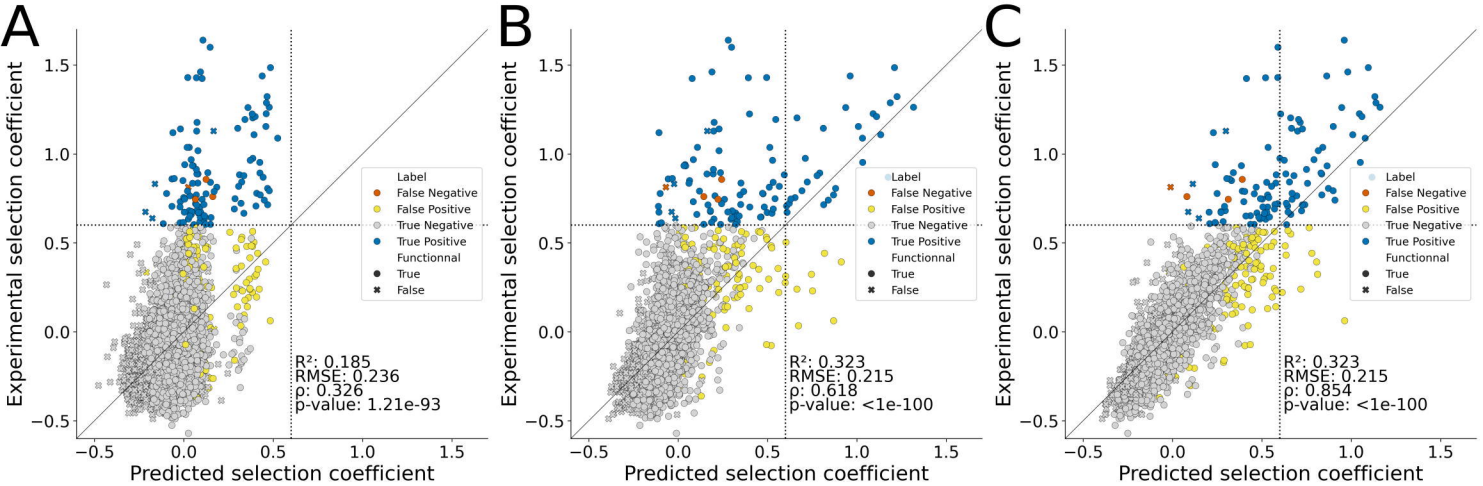

A
Density
Selection coefficient at 20 °C
B
Density
Selection coefficient at 37 °C
C
Information criterion
Number of components
AIC
BIC
D
Information criterion
Number of components
AIC
BIC
E
Intersection size
20°C_BH
37°C_BH
1000
0

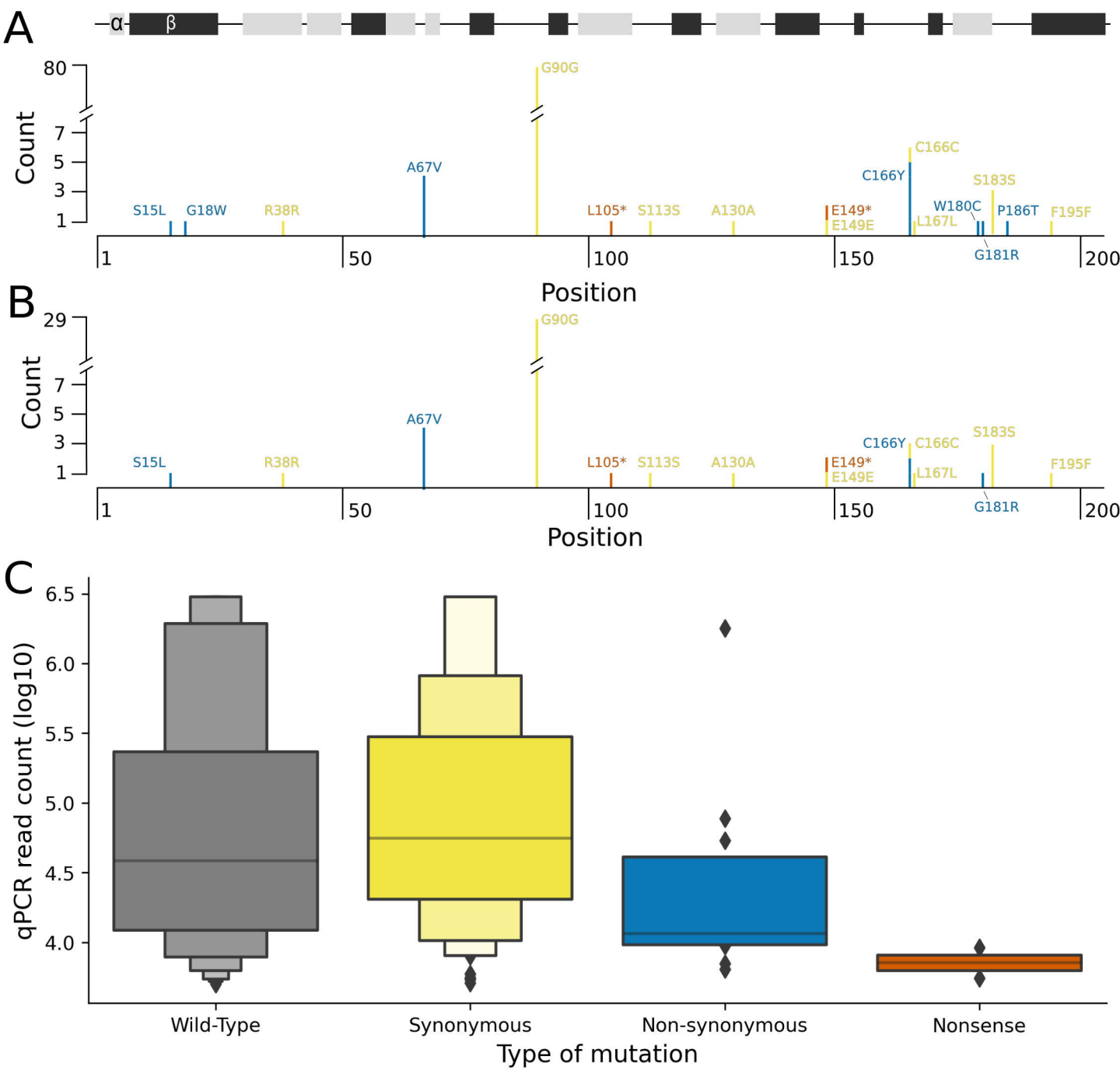

A
B
C
90°
90°
90°
D
90°
E
Positions 54 to 64: 2.46e+04
p-value: 2.93e-08
Positions 75 to 81: 2.05e+04
p-value: 2.70e-08
Positions 121 to 126: 1.47e+04
p-value: 4.81e-12
Density
0.0
0.2
0.4
0.6
0.8
−8
−6
−4
−2
0
2
GEMME score


A
False negatives in MTX testing set
B
Novel predicitions on TMP resistance
C
Biochemically characterized mutations
V11E
f(x) = 0
0.145 = mutatex_ddG −0.1
−0.249 = delta_dist_median −0.09
−0.413 = Non-functionnal +0.04
−1.919 = min_dist_lig +0.03
−2.05 = GEMME_evolEpi +0.02
2.198 = buriedness +0.02
−0.517 = allostery +0.02
−1.797 = GEMME_evolInd +0.01
−2.216 = GEMME_evolCombi +0.01
5 other features −0.04
E[f(X)] = 0.08
I10K
f(x) = 1
−1.746 = min_dist_lig +0.33
2.821 = buriedness +0.2
4.87 = catalytic_Hs +0.09
0.733 = mutatex_ddG +0.07
0.183 = allostery +0.07
−0.419 = Non-functionnal +0.07
2.076 = Flexddg +0.05
0.008 = sidechain_length +0.03
−0.35 = delta_dist_IQR +0.02
5 other features −0
E[f(X)] = 0.07
F36C
f(x) = 1
−1.658 = min_dist_lig +0.33
1.908 = buriedness +0.25
−0.419 = Non-functionnal +0.12
4.87 = catalytic_Hs +0.07
0.741 = mutatex_ddG +0.05
−0.197 = sidechain_length +0.04
0.88 = delta_dist_IQR +0.03
−0.897 = GEMME_evolCombi +0.03
5.266 = Flexddg +0.02
5 other features +0.04
E[f(X)] = 0.02
V11M
f(x) = 0
−0.815 = mutatex_ddG −0.08
0.08 = GEMME_evolCombi −0.08
−0.413 = Non-functionnal +0.03
−1.919 = min_dist_lig +0.02
2.198 = buriedness +0.02
−0.222 = GEMME_evolEpi −0.02
0.77 = delta_dist_IQR +0.01
0.708 = delta_dist_median +0.01
−0.929 = GEMME_evolInd +0.01
5 other features −0.01
E[f(X)] = 0.08
I123G
f(x) = 1
−1.944 = min_dist_lig +0.29
2.683 = buriedness +0.24
2.303 = allostery +0.09
4.87 = catalytic_Hs +0.08
0.807 = mutatex_ddG +0.08
0.008 = sidechain_length +0.08
−0.419 = Non-functionnal +0.07
1.394 = delta_dist_skewness −0.03
−0.615 = GEMME_evolEpi +0.01
5 other features +0.02
E[f(X)] = 0.07
L65P
f(x) = −0
2.387 = Non-functionnal −0.02
0.373 = mutatex_ddG −0.01
−1.714 = min_dist_lig +0.01
0.008 = sidechain_length +0.01
0.519 = buriedness −0.01
1.637 = Flexddg +0
0.022 = GEMME_evolEpi −0
1.731 = delta_dist_IQR +0
1.596 = allostery −0
5 other features +0
E[f(X)] = 0.02
S37P
f(x) = 0
2.331 = allostery −0.16
−1.813 = min_dist_lig +0.04
−0.413 = Non-functionnal +0.04
0.323 = mutatex_ddG −0.03
−0.645 = sidechain_length −0.03
0.844 = delta_dist_median +0.03
−2.483 = GEMME_evolEpi +0.02
−2.696 = GEMME_evolCombi +0.02
−0.107 = Flexddg −0.01
5 other features +0.01
E[f(X)] = 0.08
F156V
f(x) = 1
−0.542 = min_dist_lig +0.28
0.589 = buriedness +0.2
0.736 = mutatex_ddG +0.19
−0.197 = sidechain_length +0.12
−0.419 = Non-functionnal +0.07
0.014 = delta_dist_IQR +0.05
2.303 = allostery +0.05
−0.092 = GEMME_evolCombi −0.02
0.244 = GEMME_evolEpi −0.02
5 other features +0.01
E[f(X)] = 0.07
A67V
f(x) = −0
−0.419 = Non-functionnal +0.01
−1.314 = buriedness −0.01
−1.395 = sidechain_length −0.01
−0.681 = mutatex_ddG −0.01
−0.524 = allostery −0
−0.387 = delta_dist_median −0
−0.43 = delta_dist_IQR −0
−0.724 = min_dist_lig +0
−0.205 = catalytic_Hs −0
5 other features +0
E[f(X)] = 0.02
G124P
f(x) = 0
2.424 = Non-functionnal −0.19
−1.533 = min_dist_lig +0.03
1.845 = buriedness +0.02
−1.223 = GEMME_evolCombi +0.02
−1.192 = GEMME_evolInd +0.02
1.114 = mutatex_ddG +0.02
8.444 = Flexddg +0.01
−0.939 = GEMME_evolEpi +0.01
−1.952 = sidechain_length −0.01
5 other features −0.01
E[f(X)] = 0.08
Y197Q
f(x) = 1
−0.584 = min_dist_lig +0.24
1.029 = mutatex_ddG +0.22
0.306 = buriedness +0.19
−0.207 = delta_dist_IQR +0.1
−0.419 = Non-functionnal +0.07
−0.122 = sidechain_length +0.06
0.7 = GEMME_evolInd +0.03
−0.806 = delta_dist_skewness +0.02
0.656 = GEMME_evolCombi +0.01
5 other features −0.01
E[f(X)] = 0.07
C166Y
f(x) = 0
−0.419 = Non-functionnal +0.01
−0.112 = mutatex_ddG −0.01
−1.314 = buriedness −0.01
2.045 = min_dist_lig −0.01
−0.667 = delta_dist_IQR −0.01
−0.524 = allostery −0
−0.58 = delta_dist_median −0
−0.3 = sidechain_length −0
−0.205 = catalytic_Hs −0
5 other features −0
E[f(X)] = 0.02


A
Train set
Test set
True label
Sensitive
Resistant
Predicted label
2736
180
0
89
665
66
5
16
Train ROC
Test ROC
True Positive Rate
False Positive Rate
Train PRD
Test PRD
Precision
Recall
Distance to ligand
Buriedness
WT sidechain length
MutateX score
IQR change in distance
Catalytic
Allosteric effect
GEMME combined
GEMME epistasis
Median change in distance
Skew change in distance
GEMME individual
FlexddG total score
SHAP Values
Feature value
High
Low
B
Label
False Negative
False Positive
True Negative
True Positive
Functionnal
True
False
MTX resistance score
Resistant prediction probability
C
Amino acid
Prediction probability
Contact
DHF
MTX
TMP
NADPH
Position


ρ = -0.24
p = < 1e-100
Selection coefficient - 37°C
IQR change in distance to methotrexate
(Å, FlexddG)
Distance to methotrexate (Å)
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
0.0
0.5
1.0
1.5
2.0
20
18
16
14
12
10
8
6
4


A
B
C
Experimental selection coefficient
Predicted selection coefficient
Label
False Negative
False Positive
True Negative
True Positive
Functionnal
True
False
R²: 0.185
RMSE: 0.236
ρ: 0.326
p-value: 1.21e-93
R²: 0.323
RMSE: 0.215
ρ: 0.618
p-value: <1e-100
R²: 0.323
RMSE: 0.215
ρ: 0.854
p-value: <1e-100


A
α
β
Count
80
7
5
3
1
G90G
A67V
C166C
C166Y
S183S
S15L
G18W
R38R
L105*
S113S
A130A
E149*
E149E
L167L
W180C
P186T
F195F
G181R
1
50
100
150
200
Position
B
29
Count
7
5
3
1
G90G
A67V
C166Y
C166C
S183S
S15L
R38R
L105*
S113S
A130A
E149*
E149E
L167L
F195F
G181R
1
50
100
150
200
Position
C
qPCR read count (log10)
6.5
6.0
5.5
5.0
4.5
4.0
Wild-Type
Synonymous
Non-synonymous
Nonsense
Type of mutation