# Simple Topological Properties Predict Functional Misannotations in a Metabolic Network.

Rodrigo Liberal, John W. Pinney*.
Centre for Integrative Systems Biology and Bioinformatics, Imperial College London, London SW7 2AZ, UK.
* E-mail: j.pinney@imperial.ac.uk

## Supplementary Figures

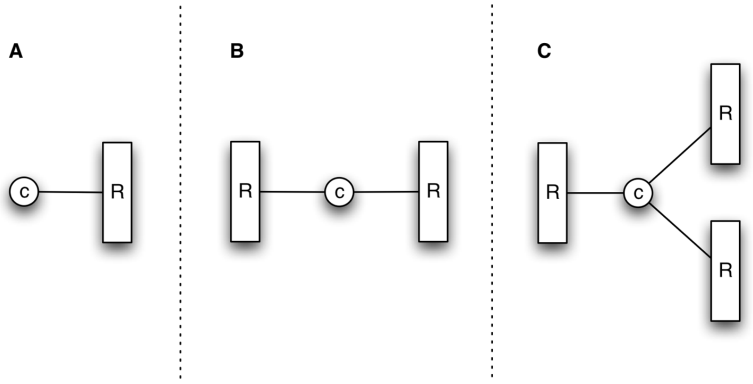


**Figure S1. Classification of compounds.** Each compound C involved in a reaction R belongs to one of these three classes: A - connected to just one reaction (an unpaired compound); B - connected to exactly two reactions (a chokepoint compound); C - connected to more than two reactions.

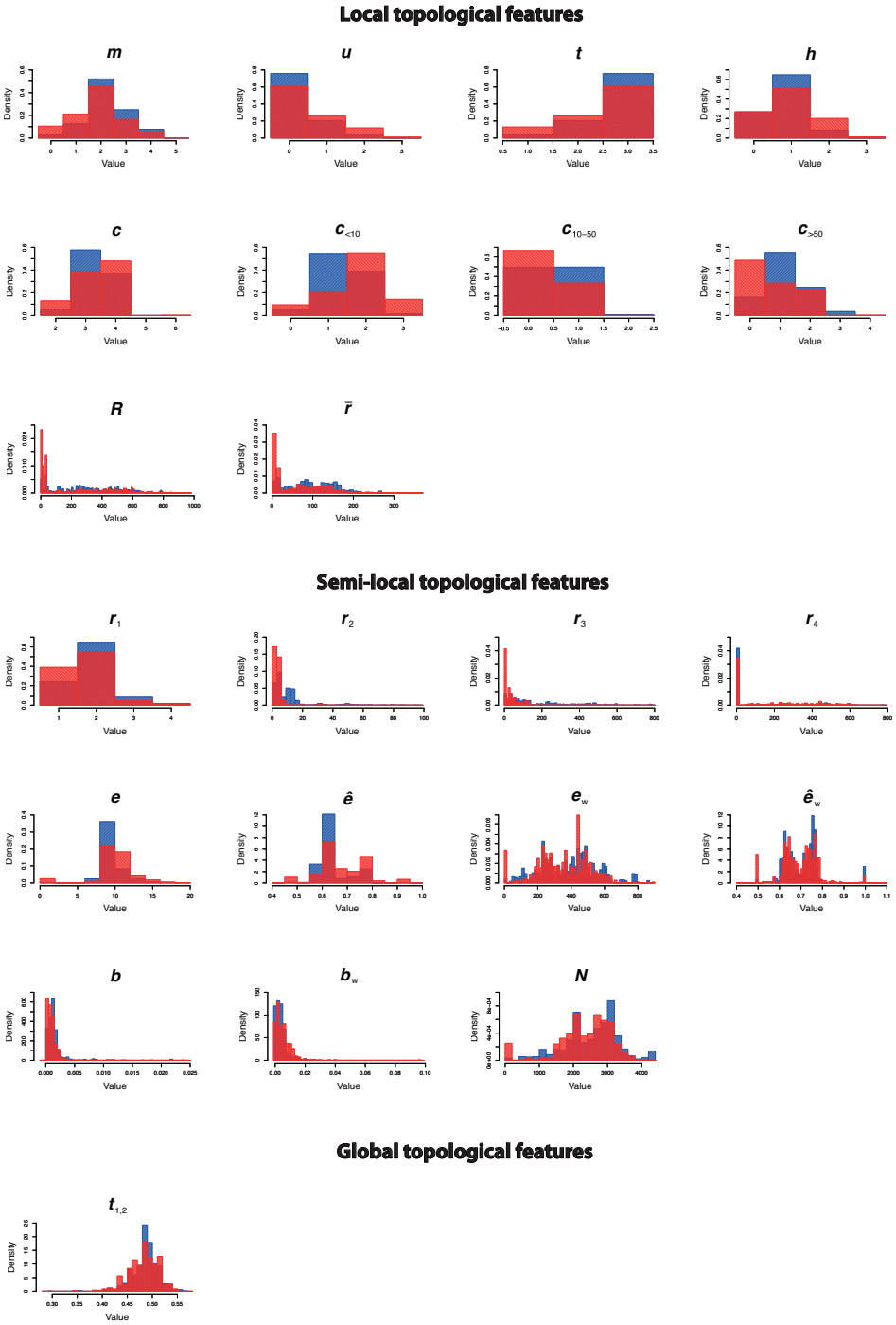


**Figure S2. Feature histograms.** Visualisation of the potential value of each attribute in distinguishing the correct functional assignments from the incorrect ones (red - incorrect annotations; blue - correct annotations). See Table 1 in main article for feature definitions.

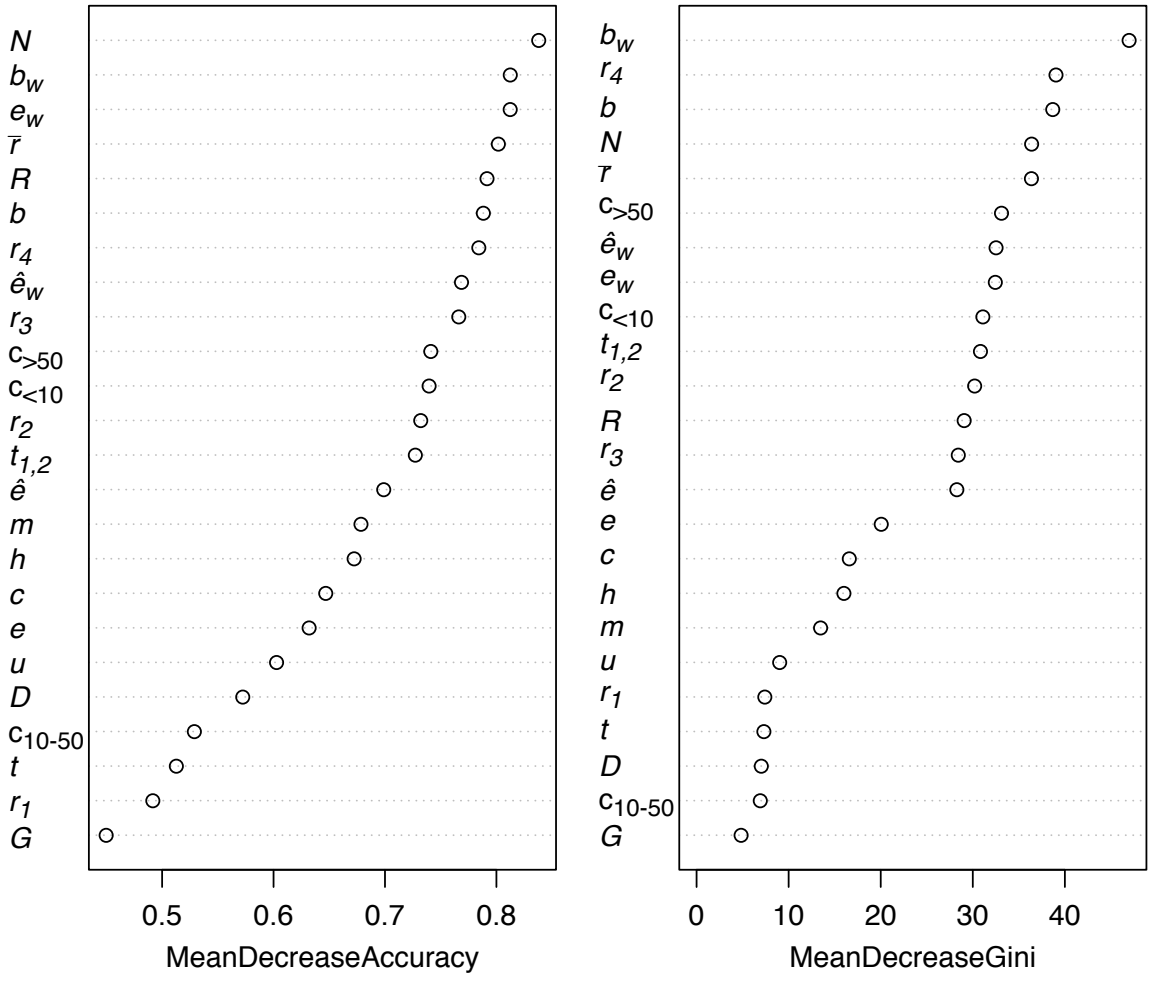


**Figure S3. Feature predictiveness.** These scores, obtained from the `importance` function of the randomForest R package, are used to assess the relative contribution of each feature to the performance of the predictor. left: average accuracy decrease when each feature is removed. right: average entropy decrease for each feature.

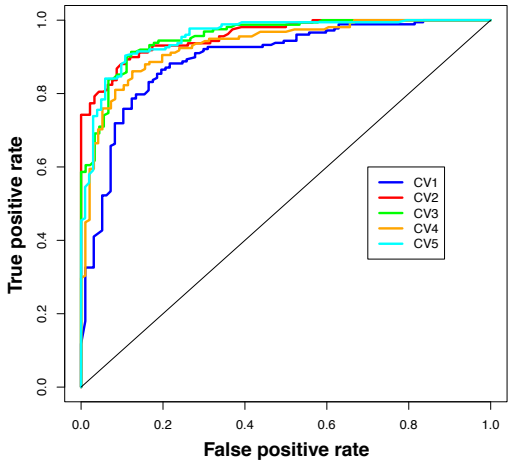


**Figure S4. fivefold cross validation ROC curves.**

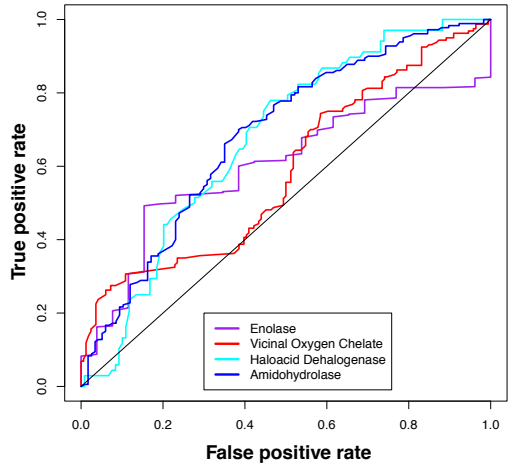


**Figure S5. superfamily cross validation ROC curves.**

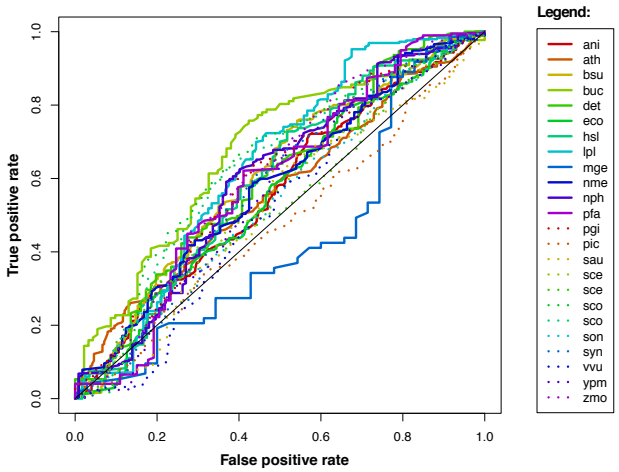


**Figure S6. Classifier performance in curated models ROC curves.**

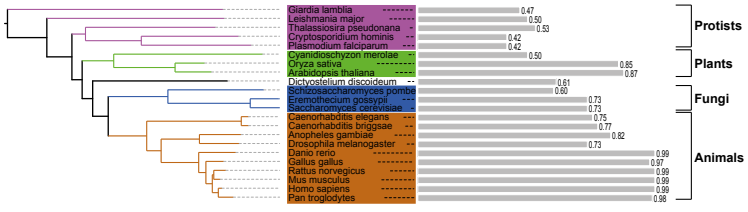


**Figure S7. Predicted quality of draft metabolic networks across a eukaryote phylogeny.** The classifier was applied to all eukaryote species present in iTOL. To the left is the eukaryote phylogenetic tree. The quality values are represented by bars next to the species names.

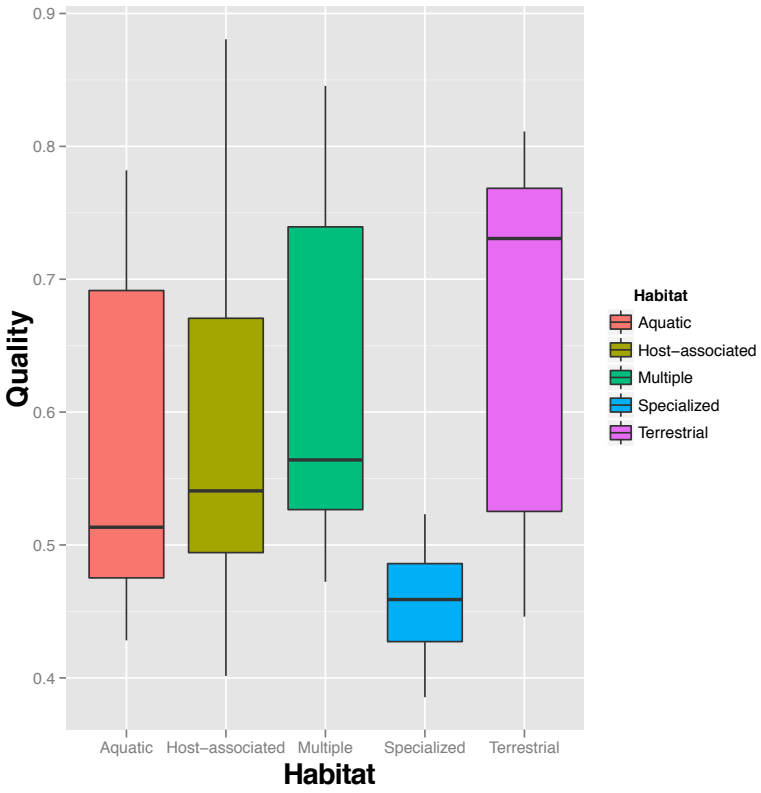


**Figure S8. Bacterial habitats.** The box plots show the predicted annotation quality (percentage of correctly annotated reactions according to the classifier) of the bacterial species, grouped by habitat.

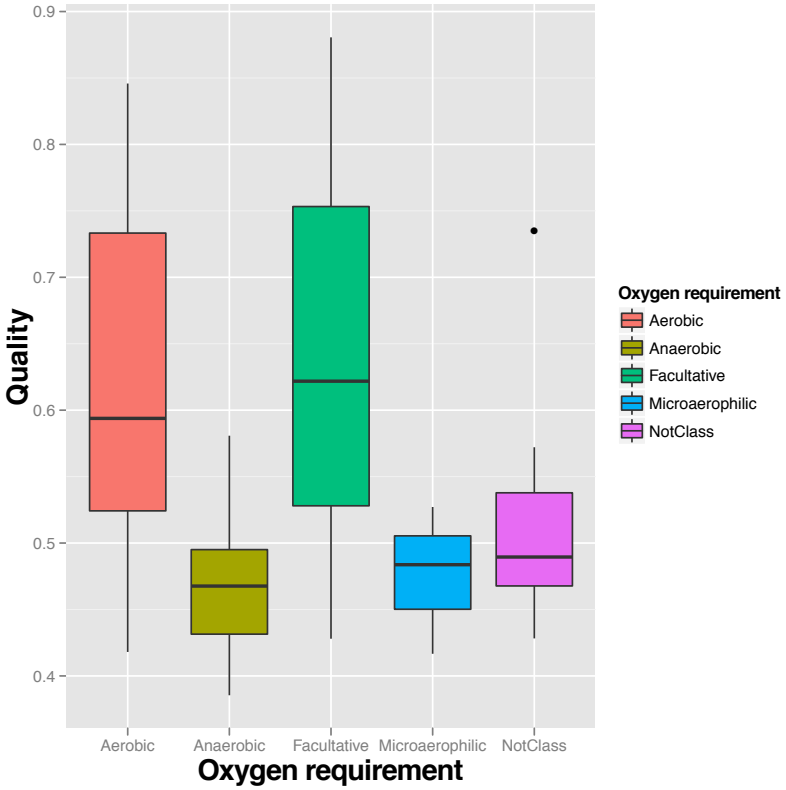


**Figure S9. Bacterial oxygen requirement.** The box plots show the predicted annotation quality (percentage of correctly annotated reactions according to the classifier) of the bacterial species, grouped by oxygen requirement.

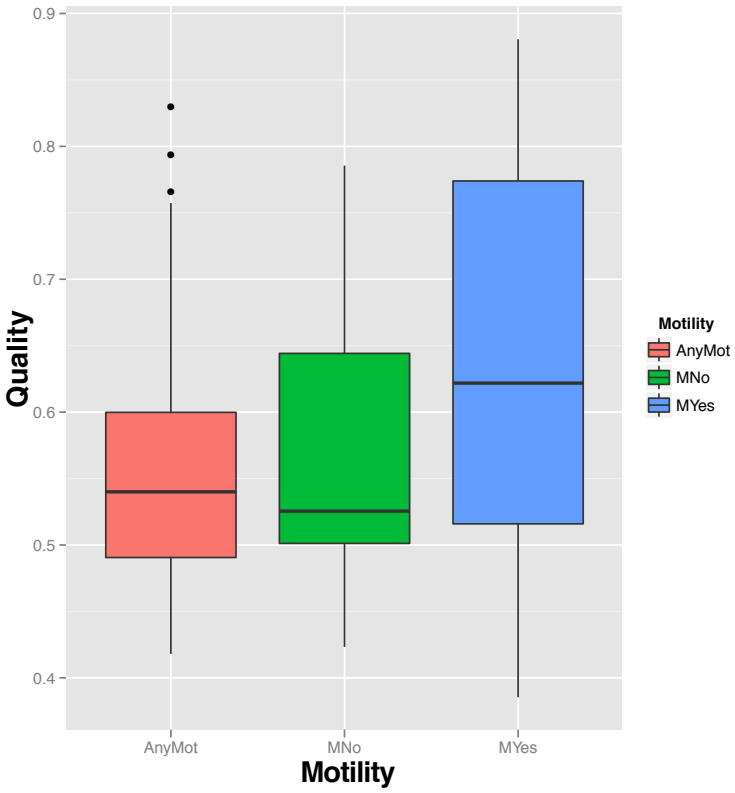


**Figure S10. Bacterial motility.** The box plots show the predicted annotation quality (percentage of correctly annotated reactions according to the classifier) of the bacterial species, grouped by motility.

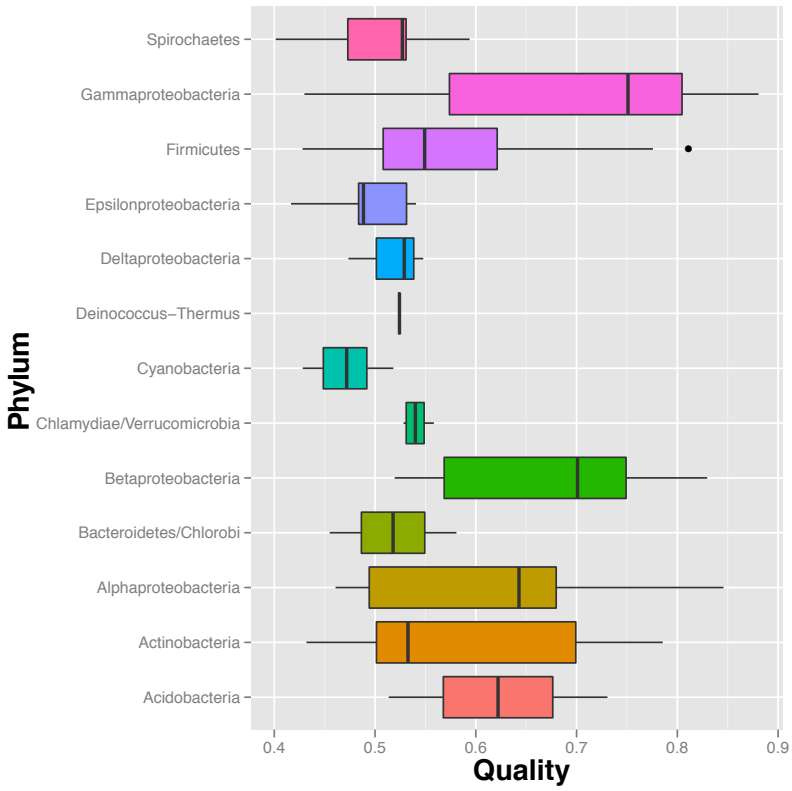


**Figure S11. Bacterial phyla.**The box plots show the predicted annotation quality (percentage of correctly annotated reactions according to the classifier) of the bacterial species grouped by phylum.

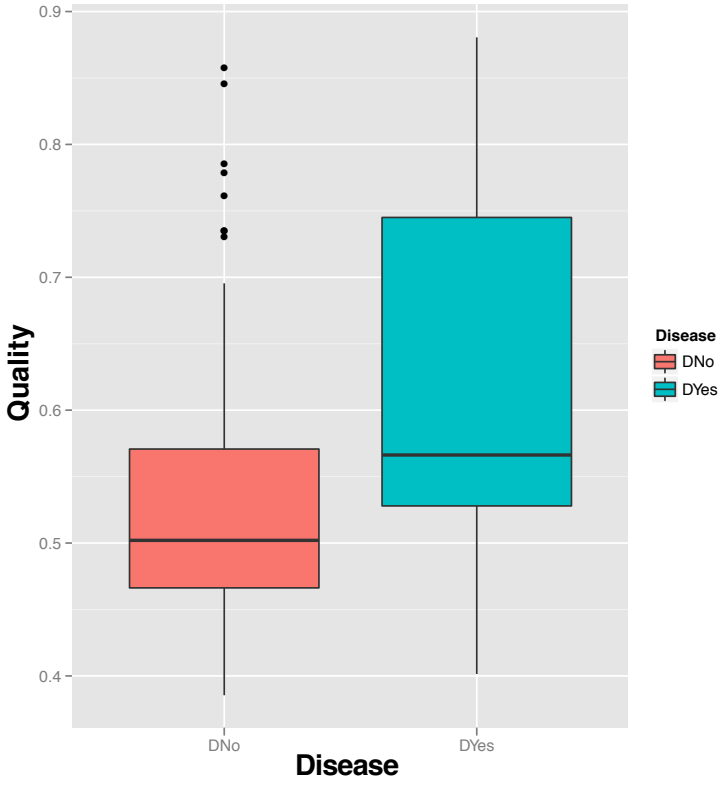


**Figure S12. Bacterial disease association.**The box plots show the predicted annotation quality (percentage of correctly annotated reactions according to the classifier) of the species grouped by disease association.

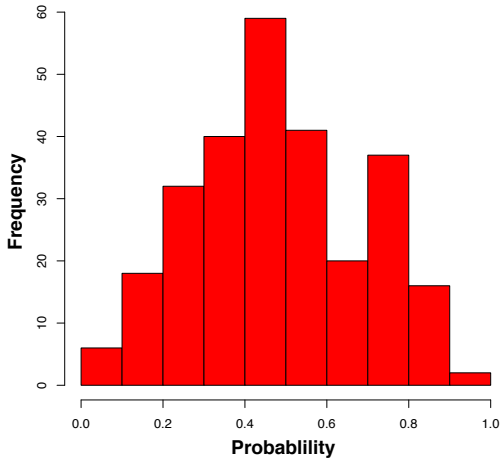


**Figure S13.** ***Plasmodium falciparum*** **enzymatic function quality histogram.**

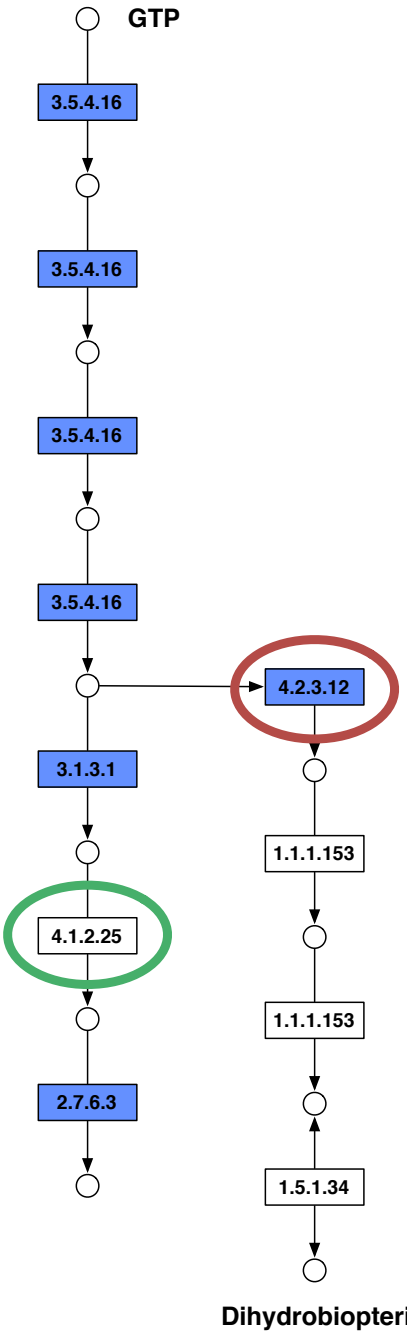


**Figure S14.** Part of the *P. falciparum* folate biosynthesis pathway taken from KEGG. All the reactions that do not have an enzyme assigned have a white background. Marked by a green circle is the missing link referred by Dittrich *et al.*. With a red circle is the reaction to which the enzyme PTPS was originally assigned.

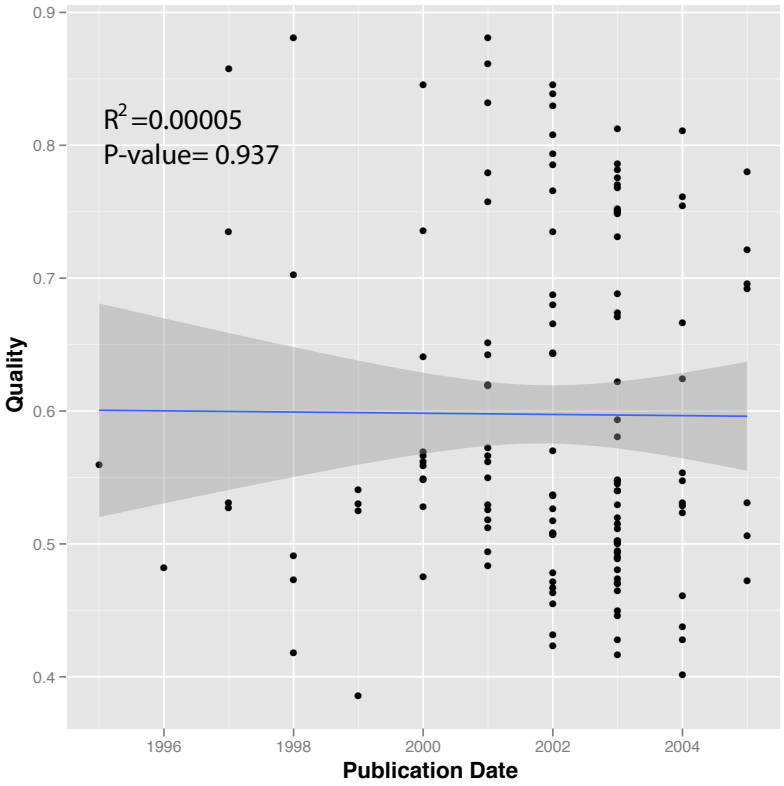


**Figure S15. Variation of predicted annotation quality with publication date in bacteria.**

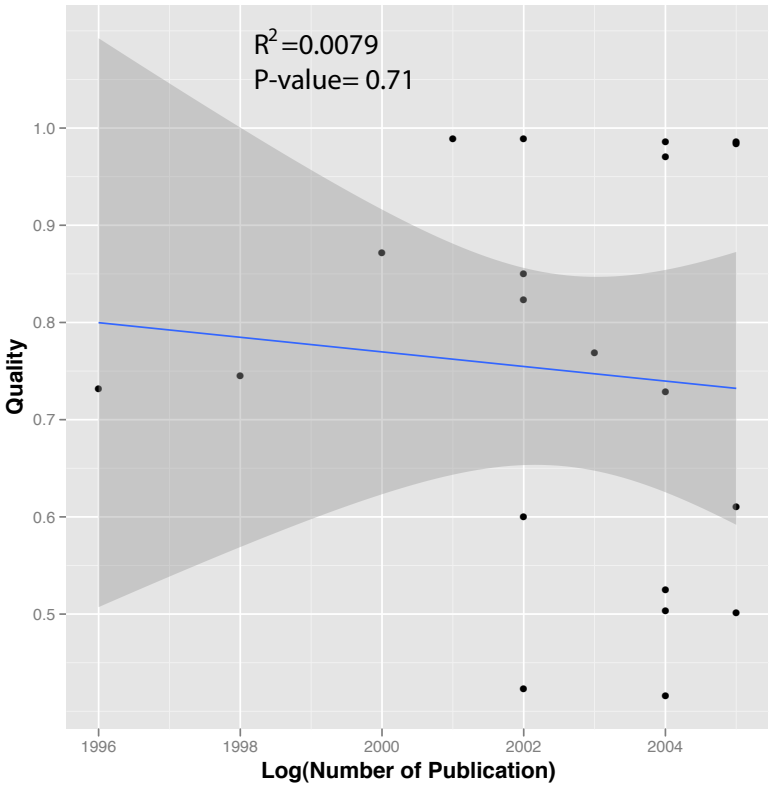


**Figure S16. Variation of predicted annotation quality with publication date in eukaryotes.**

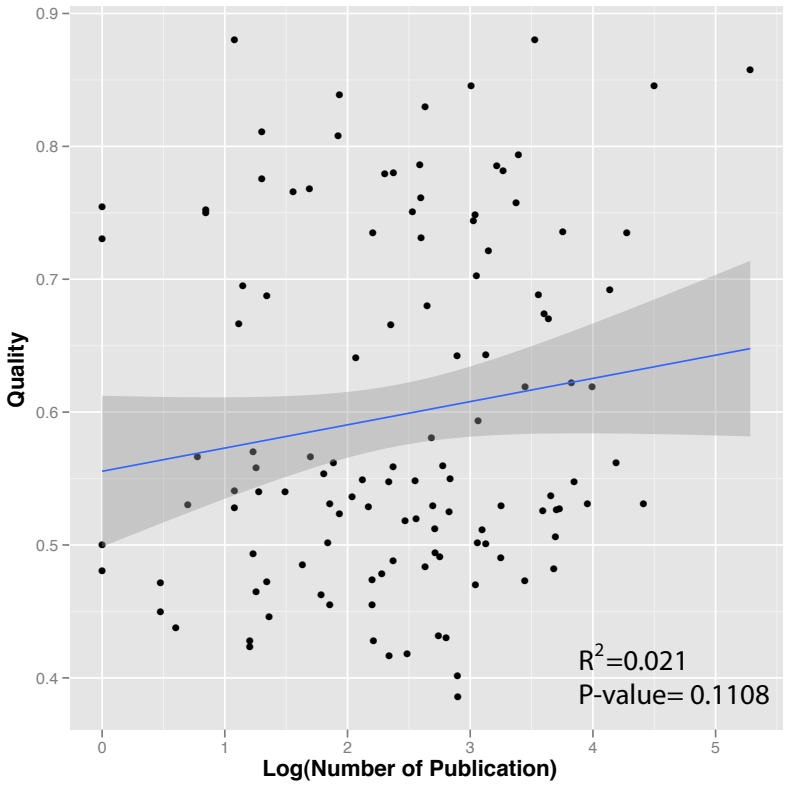


**Figure S17. Variation of predicted annotation quality in bacteria with number of publications.**