The moderately efficient enzyme: evolutionary and physico-chemical trends shaping enzyme parameters

Supporting Information

Index

A. Data set Properties

Obtaining and filtering the kinetic parameters

Kinetic parameters of more than 5000 different enzymes were extracted and filtered from the BRENDA database (http://www.brenda-enzymes.info/, last downloaded in July 2010) (*1*). We used the turnover numbers and the K_M values of each enzyme in every organism where the data is available. We removed all data explicitly referring to mutated enzymes. We downloaded and parsed the database and created a local one using SQLite (http://www.sqlite.org/), containing around 90,000 entries for K_M and 35,000 entries for k_{cat} (Table 1).

A central problem in utilizing the BRENDA database is that many of the substrates for which the kinetic data were measured are not the natural substrates. We therefore used the KEGG database (http://kegg.jp/, last downloaded in May 2010) (*2*), which contains a comprehensive list of naturally occurring metabolic reactions. For each enzyme in BRENDA, we identified the set of natural reactants by comparing the reactant names with those associated with this enzyme in the KEGG database. Table 1 shows the number of kinetic parameters obtain by this procedure.

Finally, in all the analyses performed, the cofactor substrates were omitted to avoid a bias from compounds that are expected to be subject to a different kind of evolutionary pressure. Furthermore, these cofactors frequently participate in more than one reaction catalyzed by the same enzyme (for example, NADH participates in many reactions catalyzed by alcohol dehydrogenase), and hence assigning their kinetic parameters to a single reaction ID (in KEGG) is virtually impossible in many cases.

Dependencies between the kinetic parameters

We tested for dependencies between kinetic parameters measured for the same enzyme. We found that k_{cat} associated with forward and backward reactions of the same enzyme are correlated (R^2=0.4, p-value<10^{-6}, Methods, Table S1). Since the transition-state of a reaction is the same for both directions, this correlation might indicate that the activation energy barriers are similar with respect to the enzyme-substrate complex and the enzyme-product complex.

Based on theoretical considerations, a correlation between k_{cat} and K_M is expected for a specific reaction carried out by different enzymes at different organisms or tissues (*3*). This was supported by various experiments, for example the carbon-fixing enzyme Rubisco (*4, 5*). We tested for a global dependence between these parameters and found only a weak (though statically significant) positive correlation between k_{cat} and K_M when analyzing all enzymes together (Fig. S1, R^2=0.09, p-value<10^{-6}). The correlation varies when considering different individual enzyme classes (Fig. S1). Most EC classes display a wide range of both k_{cat} and K_M, where some show no correlation between the two, while others present correlations up to R^2=0.3 (Fig. S1).

Why is the data set noisy?

We analyzed the consistency of the values described throughout the literature and found that our data set is considerably noisy. For example, k_{cat} and K_M values measured in different studies for the same reaction, substrate and enzyme vary, on average, by a factor of 2.5 and 2.9, respectively (R^2=0.37, Table S1; in all cases, for R^2 calculations, k_{cat} and K_M were logarithmically transformed). By taking the median across all publications referring to the same kinetic parameter, we were able to decrease this noise. Still, even after such a procedure, the correlation between the k_{cat} values for different substrates of the same multi-substrate reaction (expected to have identical values) was found to be only R^2=0.63 (Table S1). Further analysis of the dispersion of values and of the global dependencies between kinetic parameters, such as the correlation between k_{cat} and K_M, is given in the SI and in Figure S1.

There are several possible reasons for the low correlation values:

1. The different kinetic parameters were measured under varying conditions such as pH, temperature, ionic strength and metal-ion and co-factor concentrations. As a consequence, while

some enzymes were measured close to their optimal conditions, others were tested far from optimum. Nevertheless, we do not think there are systematic differences in these conditions that can explain the trends observed.

2. We found up to 20% of the values in the Brenda database do not correspond to the values reported in the corresponding reference papers. These discrepancies are the result of erroneous copying and unit mismatch. In several cases, the values in the database were orders of magnitude higher/lower than those given in the original works.

3. In the global analysis we perform, many sub-groups might not be expected to display the identified correlations. For example, different EC classes that are characterized by a complex mechanism usually do not present any significant correlation between MW and K_M or k_{cat} and K_M. In such cases the global inclusion of these groups lowers the observed correlations.

4. There are systematic differences between the measured *in-vitro* parameters and the *in-vivo* ones (*6*). These differences can be as high as three orders of magnitude in either way, and are different for each enzyme (*6*).

B. Context Affects Enzyme Kinetics

Metabolic classification affects enzyme kinetics

The KEGG database provided us an association of enzyme-substrate pairs with metabolic modules, where each pair was assigned to one or more of the >300 metabolic modules. We manually assigned each metabolic module to one of four primary groups: Central-CE (carbohydrate-energy) metabolism; Central-AFN (amino-acids, fatty-acids and nucleotide) metabolism; intermediate metabolism and secondary metabolism. The exact assignment of the modules is given in Table S2. Each enzyme-substrate pair could therefore be assigned to one of the four primary groups. If a pair was associated with several modules, assigned to different primary groups, we assigned the pair to the "more central" group, according to the order above.

In the main text we show that enzymes assigned to central metabolism have significantly higher k_{cat} and k_{cat}/K_M values as compared to secondary metabolism enzymes. We validate this by testing specific groups of enzymes, categorized functionally, structurally and by organism affiliation:

1. We repeated the analysis of different metabolic groups using only one EC class at a time thus testing for possible functional biases. We observe the same overall trends in most EC classes, as shown in Figure S4A for the EC class for which we have the largest kinetic dataset, 1.1.1.X (oxidoreductases acting on the CH-OH groups as a donor with NAD^+ or $NADP^+$ as acceptor).

2. We use CATH, a manually curated classification of protein domain structures (*7*), to categorize enzymes according to their general structure (http://chemistry.st-and.ac.uk/staff/jbom/group/EC_PDB_CATH.html). We divide enzyme-substrate pairs in different CATH classes into the four groups discussed above. We observe the same overall trends in most CATH classes, as shown in Figure S4B-D for the CATH classes for which we have the largest kinetic dataset, 1.10.XXX (mainly alpha, orthogonal bundle), 3.10.XXX (mixed alpha-beta, role) and 3.30.XXX (mixed alpha-beta, 2-layer sandwich).

3. We repeated the analysis of different metabolic groups using only enzymes form a single organism, testing for possible host biases. We observe the same overall trends in all organisms from which we had enough kinetic data: *E. coli*, *S. cerevisiae* and human.

The effect of host of an enzyme on its kinetics

We expect that if metabolic context affect enzyme kinetics the host organism may also play a role in shaping the evolutionary pressure imposed on an enzyme. For example, enzymes operating in unicellular organisms might be expected to be under stronger selective pressure to increase rate as compared to enzymes of multi-cellular organisms. While in a multi-cellular organism the increase in flux within a given cell or tissue probably has a marginal effect on the fitness of the whole organism, a higher flux in a unicellular organism is expected to translate more directly to a higher growth rate and fitness advantage.

Unfortunately, we were able to find only marginal reinforcements to the above hypothesis. As shown in Figure 1, there are no significant differences between the overall distributions of the kinetic parameters of prokaryotic and eukaryotic enzymes. To refine our analysis we considered only enzymes for which we can compared the kinetic parameters in different hosts. We analyzed human, *E. coli* and *S. cerevisiae*, which are the only organisms for which the datasets are large for pair-wise comparisons. We found that enzymes operating in *E. coli* and *S. cerevisiae* show similar average k_{cat} values (Figs. S3) and both exhibit ~ 2-fold higher average k_{cat} values than their human counterparts (Figs. S3). While this difference is quite small and the comparison is problematic due to the small number of pairs in the data set and the different preparation techniques (e.g. human enzymes expressed in *E. coli*), the differences seem consistent along the entire range of catalytic efficiencies, and statistically significant ($p\text{-value}<0.05$). We further detected a significant difference for tRNA synthetases (EC 6.1.1.X), where we observed a median k_{cat} for prokaryotic enzymes which is more than double that of eukaryotic ones (Fig. S3). Notably, the activity of these enzymes is closely associated with cellular growth (*8-10*). We verified that those differences are not due to different measurement temperatures and pH values.

C. Substrate Physico-Chemical Properties Affect K_M

Calculating and estimating the physico-chemical properties of substrates

The molecular mass, number of hydrogen bond acceptors and donors, number of charged atoms and number of rotatable bonds were calculated using Pybel, the Python wrapper for OpenBabel (http://openbabel.org/wiki/Main_Page) (*11*). Using the same software package we corrected all compounds to be in the protonation level most abundant at pH 7. The total number of hydrogen bonds, representing the total hydrogen bond inventory of the molecule (*3*), was taken as hydrogen bond donors + hydrogen bond acceptors.

The molecular 3D-structure, essential for determining the surface area of the molecules, was also estimated using OpenBabel. We then used *asa.py* (http://boscoh.com/protein/asapy) (*12*) to calculate the surface area of each atom in every compound and hence the *total* surface area of the 3D-structure. We used the solvent-excluded surface area, representing the “cavity” the molecule creates in bulk solvent (*13*). We also computed the *polar* surface area, i.e. the area of the polar atoms only (oxygen, nitrogen and the hydrogen atoms attached to them). The non-polar surface area is the difference between total surface area and polar surface area.

We used XLOGP3 (*14*) and ALogPS (*15*) to predict LogP. Both programs gave similar results.

The effect of the physico-chemical properties on K_M

We found that only MW and LogP significantly affect K_M, with R^2=0.13 and R^2=0.24, respectively (Table S3). These two substrate properties are uncorrelated (Table S3) and their contribution to K_M is independent. Indeed, performing linear regression on K_M, using both MW and LogP as predictors, results is an almost additive R^2=0.31.

The surface area, polar surface area, charge, absolute formal charge, number of charged atoms, number of hydrogen bond donors, hydrogen bond acceptors, total hydrogen bonds (acceptor + donors) and number of rotatable bonds has only a minor effect on K_M (Table S3).

We find a correlation between the non-polar surface area (NPSA) and K_M. However, we suspect that this stems from the positive correlation between NPSA and both MW and LogP. Supporting this suggestion, we find that adding NPSA to MW and LogP as the explanatory parameters for the regression of K_M did not result in any increase in R^2. Also, discarding MW or LogP and using NPSA instead for the regression of K_M, resulted in significantly lower R^2.

The energetic contribution of each non-hydorgen atom to binding

A previous study indicates that the energetic contribution of each non-hydrogen atom to the binding of small ligands to receptors is ~1.5 kcal/mol (*16*). Yet, a later study suggested that a typical contribution of a non-hydrogen atom is only about 0.3 kcal/mol (*17*). Since we get a slope of ~-0.006 Da^{-1} between $\log_{10}(K_M)$ and MW, and assuming that K_M ~ K_D (dissociation constant of the substrate-enzyme complex) and that the mass of a non-hydrogen atom is commonly ~12g/mol, we find the energetic contribution of a non-hydrogen atom to be $RT\cdot12\cdot0.006\cdot\ln(10)$ ~ 0.1 kcal/mol. We note that the lower energetic contribution we observe is sensible considering the noise in our data set and that K_M is not identical to K_D for numerous enzymes: complex catalytic mechanisms together with the effect of k_{cat}, both serve to increase K_M, as compared to K_D.

Substitution with large modifiers might serve to decrease K_M

Apart from their conventional function in compound activation, substitution with Coenzyme A (CoA), NDPs (UDP, CDP, etc.) and phosphate decreases K_M systematically, suggesting that a further possible role of these compounds is to serve as affinity enhancers for small molecules.

One might claim that since the substituted compounds must be produced from un-substituted compounds the affinity problem seems to be just shifted to upstream enzymes catalyzing the substitution reaction. However, we argue that it is easier to optimize only one enzyme to accept an un-substituted compound than to optimize all the downstream enzymes utilizing that metabolite.

Focusing on CoA, we note that from all the enzymes that can accept the un-substituted forms of acetate, propionate and butanoate, we found that the ones which produce the CoA-substituted counterparts (ligase and transferase) are characterized by the highest affinity towards the small molecules.

Recognizing small substrates by the attached CoA involves a major drawback: enzymes might accept other substituted small substrates rather than those they intended to, indicating a tradeoff between increasing the affinity and retaining specificity. Indeed, many of the enzymes that accept the CoA substituted compounds are non-specific. One of the best studied examples of this is the enzyme acetyl-CoA carboxylase, which is known to accept acetyl-CoA, propionyl-CoA and butryl-CoA (*18*).

D. Statistical Analysis & Figures Supplementary Information

We calculated p-values for observed R^2 using a Monte Carlo permutation test (also known as approximate permutation test or random permutation test) (*19-21*). The p-value corresponds to the null assumption that $R^2=0$ and was calculated by shuffling the Y-values, randomly assigning them to X-values, and calculating the resulting R^2. This process was repeated 10^6 times. Using the resulting distribution of R^2 values, the p-value was calculated as the fraction of times where the randomized R^2 values were equal or higher than the original R^2.

The cyan area in Figure 3 was drawn as follows. For each data point, we examined the 150 point interval centered around that point (75 points on each side). We calculated the mean X-value and K_M for the window as well as the standard error around the mean K_M. This process was repeated for each data point. Plotted are the standard errors around the mean K_M as a function of the mean X-value calculated this way.

For analyzing the effect of metabolic context and enzyme host on the kinetic parameters we referred to combinations of enzyme plus substrate. If multiple values exist in the database for a given enzyme-substrate combination, corresponding to different organisms and publications, we used their median. For analyzing the effect of physico-chemical properties on K_M we referred to substrates only. If multiple values exist in the database for a given substrate, corresponding to different enzymes, organisms and publications, we used their median.

The p-value for the hypothesis that the median of the ratio of k_{cat} values between different organisms is equal to 1 (no effect of the organism) was calculated using Wilcoxon signed rank test. Similar approach was taken for the ratio of K_M values between substituted and un-substituted (by modifiers) substrates.

E. References

1. Pharkya, P., Nikolaev, E. V., and Maranas, C. D. (2003) Review of the BRENDA Database, *Metab Eng 5*, 71-73.

2. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment, *Nucleic Acids Res 36*, D480-484.

3. Fersht, A. (1998) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W. H. Freeman, New York.

4. Tcherkez, G. G., Farquhar, G. D., and Andrews, T. J. (2006) Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized, *Proc Natl Acad Sci U S A 103*, 7246-7251.

5. Savir, Y., Noor, E., Milo, R., and Tlusty, T. (2010) Cross-species analysis traces adaptation of Rubisco towards optimality in a low dimensional landscape, *Proc Natl Acad Sci U S A Accepted*.

6. Wright, B. E., Butler, M. H., and Albe, K. R. (1992) Systems analysis of the tricarboxylic acid cycle in Dictyostelium discoideum. I. The basis for model construction, *J Biol Chem 267*, 3101-3105.

7. Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M., and Orengo, C. A. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution, *Nucleic Acids Res 35*, D291-297.

8. Putzer, H., Laalami, S., Brakhage, A. A., Condon, C., and Grunberg-Manago, M. (1995) Aminoacyl-tRNA synthetase gene regulation in Bacillus subtilis: induction, repression and growth-rate regulation, *Mol Microbiol 16*, 709-718.

9. Kyriacou, S. V., and Deutscher, M. P. (2008) An important role for the multienzyme aminoacyl-tRNA synthetase complex in mammalian translation and cell growth, *Mol Cell 29*, 419-427.

10. Comer, M. M., Dondon, J., Graffe, M., Yarchuk, O., and Springer, M. (1996) Growth rate-dependent control, feedback regulation and steady-state mRNA levels of the threonyl-tRNA synthetase gene of Escherichia coli, *J Mol Biol 261*, 108-124.

11. O'Boyle, N. M., Morley, C., and Hutchison, G. R. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit, *Chem Cent J 2*, 5.

12. Shrake, A., and Rupley, J. A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin, *J Mol Biol 79*, 351-371.

13. Richmond, T. J. (1984) Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect, *J Mol Biol 178*, 63-89.

14. Cheng, T., Zhao, Y., Li, X., Lin, F., Xu, Y., Zhang, X., Li, Y., Wang, R., and Lai, L. (2007) Computation of octanol-water partition coefficients by guiding an additive model with knowledge, *J Chem Inf Model 47*, 2140-2148.

15. Tetko, I. V., and Tanchuk, V. Y. (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program, *J Chem Inf Comput Sci 42*, 1136-1145.

16. Kuntz, I. D., Chen, K., Sharp, K. A., and Kollman, P. A. (1999) The maximal affinity of ligands, *Proc Natl Acad Sci U S A 96*, 9997-10002.

17. Hopkins, A. L., Groom, C. R., and Alex, A. (2004) Ligand efficiency: a useful metric for lead selection, *Drug Discov Today 9*, 430-431.

18. Bettey, M., Ireland, R. J., and Smith, A. M. (1992) Purification and characterization of acetyl CoA carboxylase from developing pea embryos, *J. Plant Physiol. 140*, 513-520.

19. Dwass, M. (1957) Modified randomization tests for nonparametric hypotheses, *The Annals of Mathematical Statistics*, 181.

20. Vadiveloo, J. (1983) On the theory of modified randomization tests for nonparametric hypotheses, *Communications in Statistics - Theory and Methods 12*, 1581-1596.

21. Nichols, T. E., and Holmes, A. P. (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples, *Hum Brain Mapp 15*, 1-25.

Figure S1

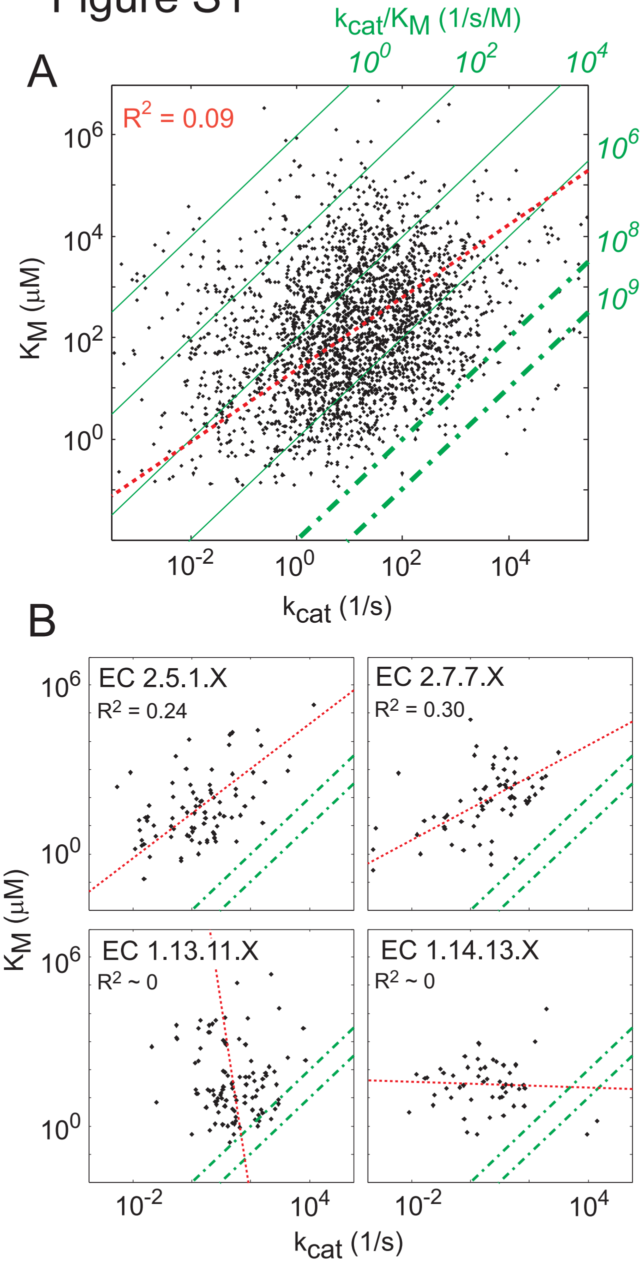

Figure S1

The correlation between k_{cat} and K_M (A) across all reactions (R^2=0.09) and (B) for reactions associated with specific EC classes. Green corresponds to k_{cat}/K_M isovalue lines. Bold, dashed lines represent the diffusion limit for small metabolites interacting with proteins ($k_{cat}/K_M < 10^8 - 10^9$). Red corresponds to the trend lines, calculated by orthogonal least-square fitting. Each dot represents a combination of an enzyme, a substrate and an organism. EC classes: 2.5.1.X (R^2=0.24): Transferase enzymes, transferring alkyl or aryl groups, other than methyl groups; 2.7.7.X (R^2=0.30): Nucleotidyltransferases; 1.13.11.X (R^2~0): Oxidase enzymes, acting on a single donor and incorporating two atoms of oxygen and 1.14.13.X (R^2~0): Oxidase enzymes, acting on two donors, where NADH or NADPH is one donor,

Figure S2

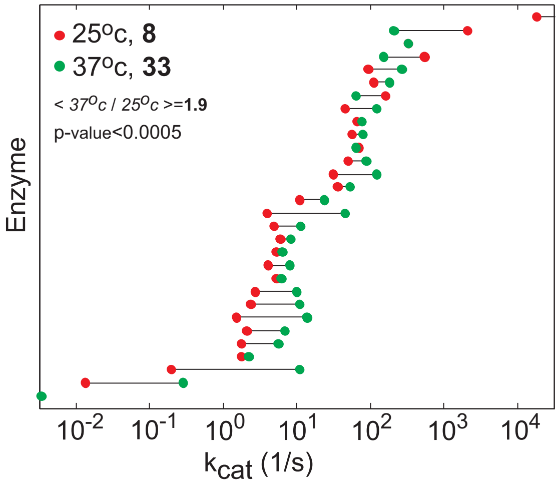

Figure S2

Enzymes operating at different temperatures have significantly different k_{cat} values. Each line corresponds to a single enzyme-substrate pair. Bold numbers correspond to the median of each distribution.

Figure S3

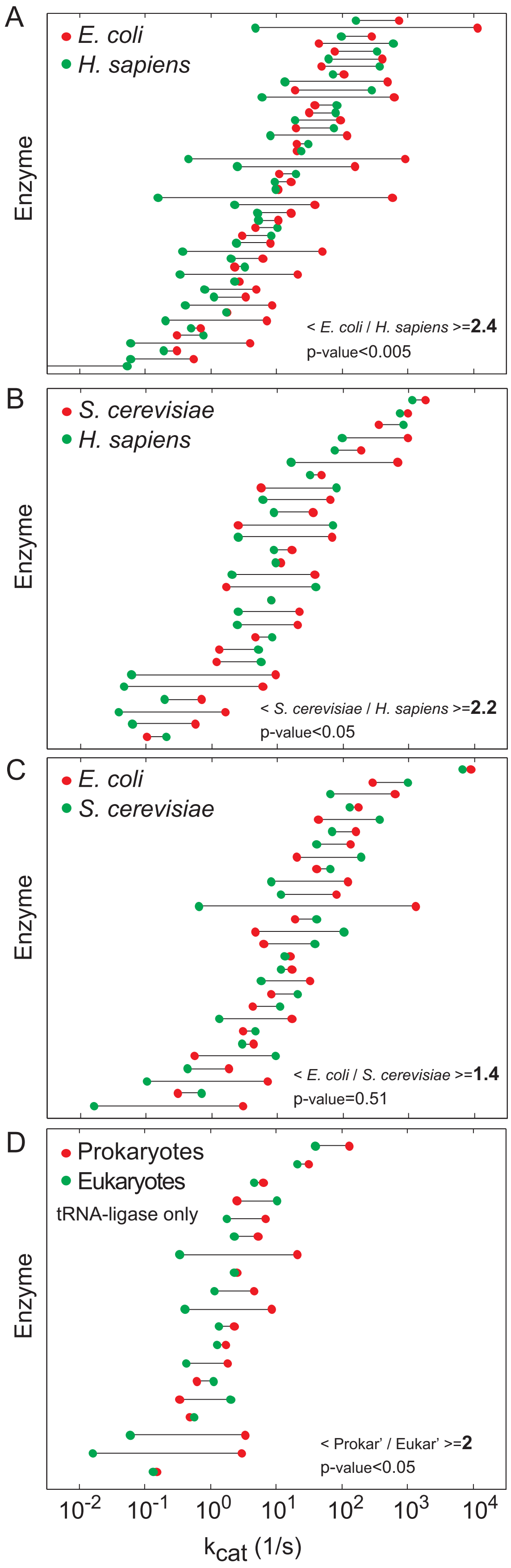

Figure S3

Enzymes operating in different hosts have significantly different k_{cat} values. Each line corresponds to a single enzyme-substrate pair.

Figure S4

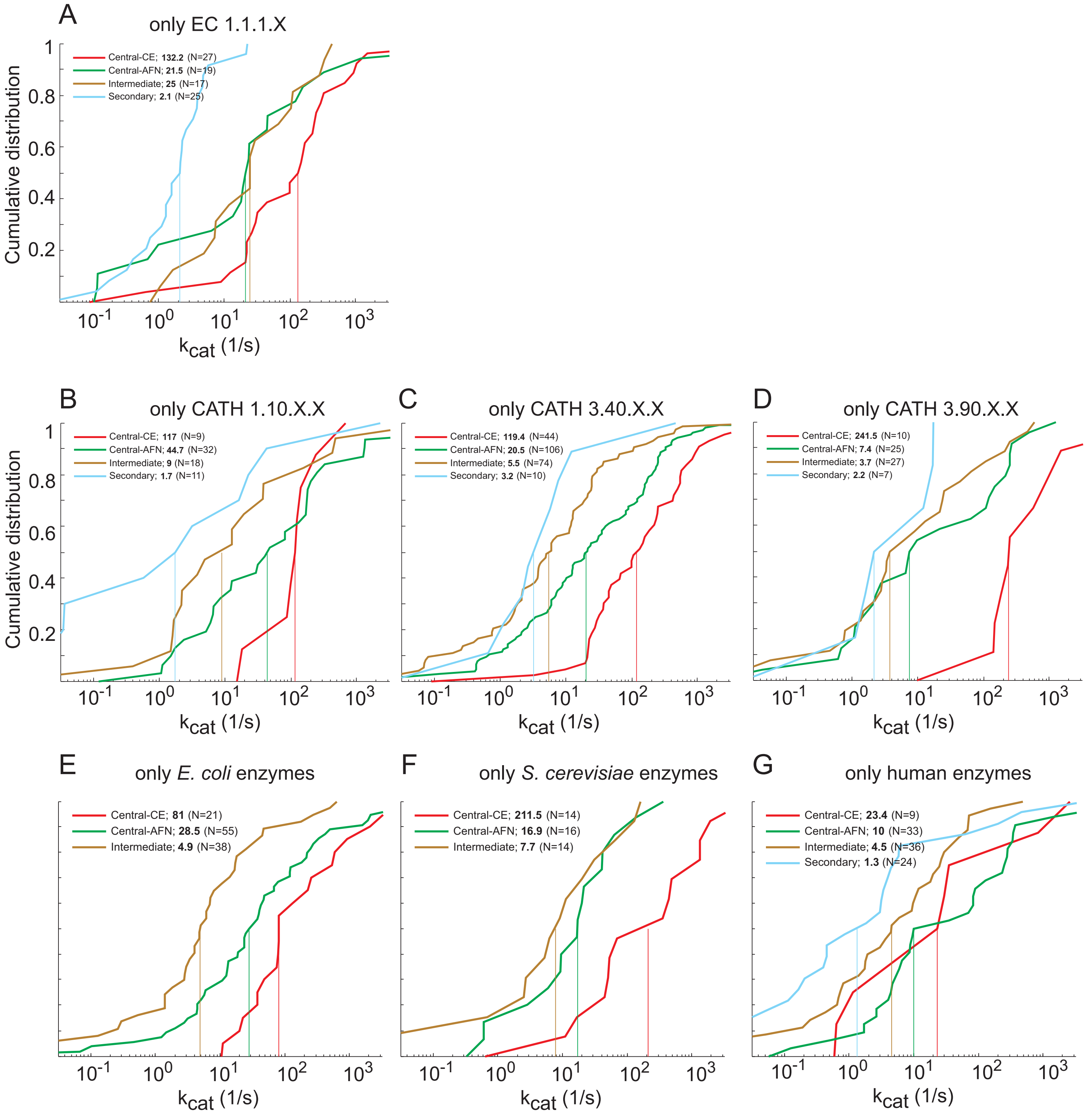

Figure S4

Enzymes operating within different metabolic groups have significantly different k_{cat} values. (A) enzymes of the EC class 1.1.1.X (oxidoreductases acting on the CH-OH groups as a donor with NAD^+ or $NADP^+$ as acceptor). (B) enzymes of the CATH class 1.10.XXX (mainly alpha, orthogonal bundle). (C) enzymes of the CATH class 3.10.XXX (mixed alpha-beta, role) and (D) enzymes of the CATH class 3.30.XXX (mixed alpha-beta, 2-layer sandwich). (E) only *E. coli* enzymes. (F) only *S. cerevisiae* enzymes. (G) only human enzymes.

Figure S5

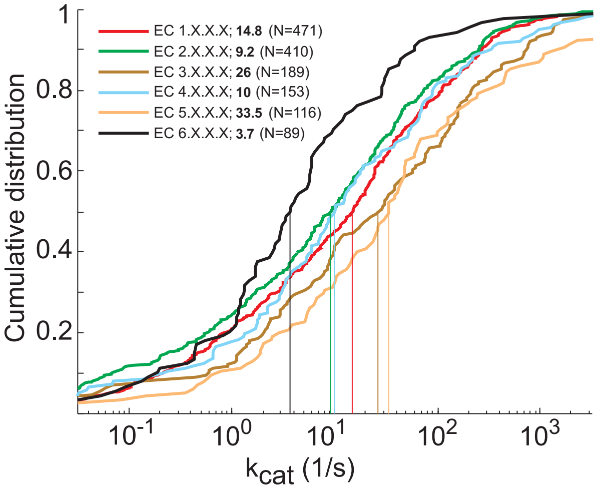

Figure S5

Enzymes of different classes have significantly different k_{cat} values. Isomerase enzymes (EC 5.X.X.X) are significantly faster as compared to any other EC classes with p-value<0.005, except hydrolases (EC 3.X.X.X). Ligase enzymes (EC 6.X.X.X) are significantly slower as compared to any other EC classes with p-value<0.05. Bold numbers in the legend correspond to the median of each distribution, while numbers in parentheses represent the number of measured values in each distribution.

Figure S6

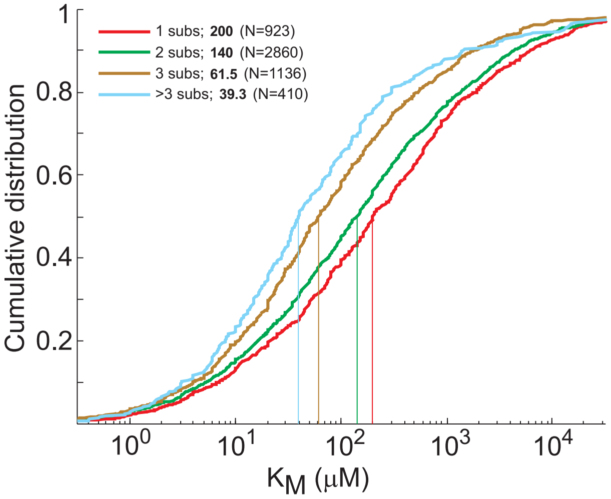

Figure S6

Reactions with a higher number of substrates are characterized by lower K_M values. The difference between any two groups is significant (p-value<0.01). Bold numbers in the legend correspond to the median of each distribution; in parentheses is the number of measured values in each distribution.

Figure S7

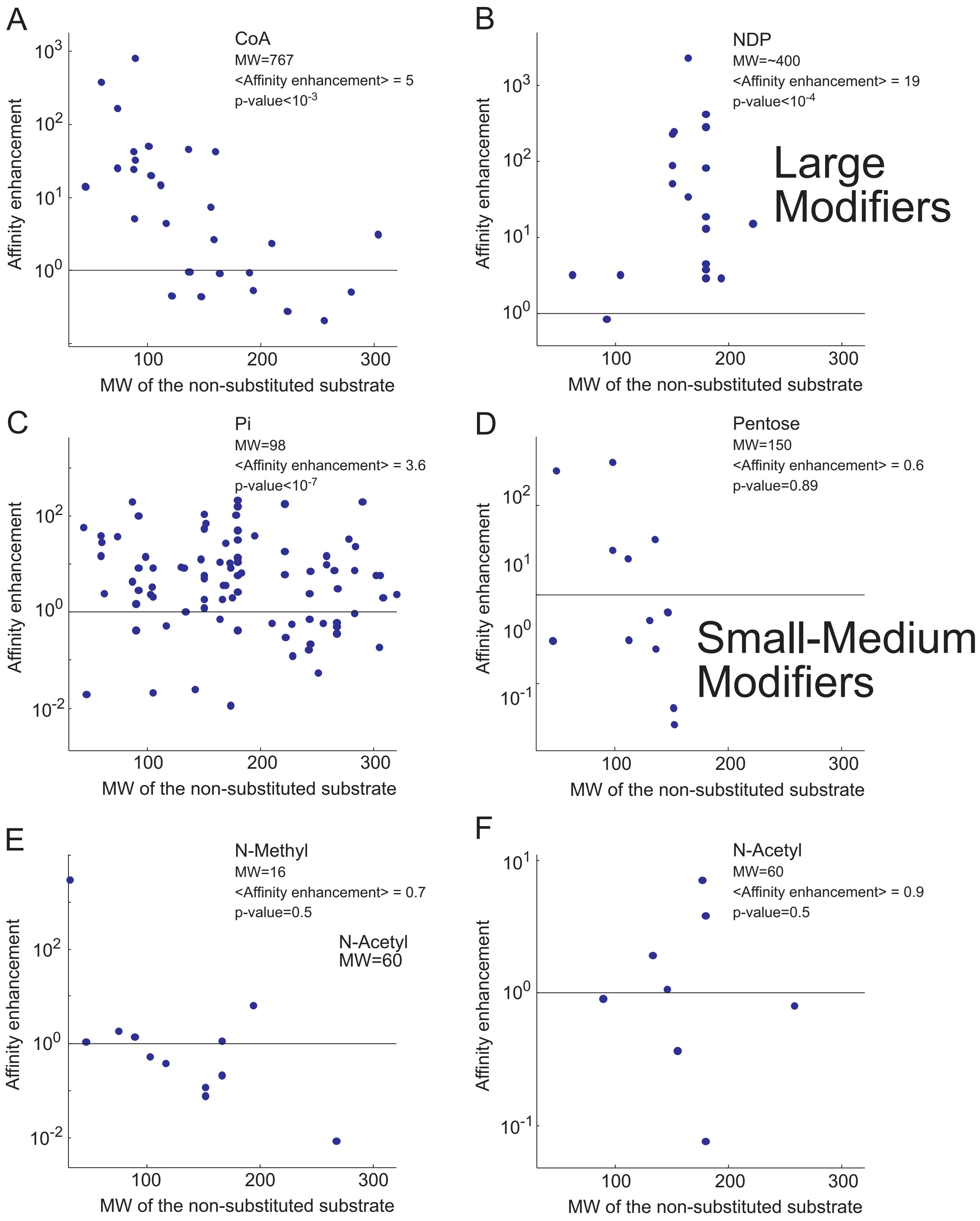

Figure S7

The correlation between the substrate molecular mass and the KM enhancement upon substitution with various modifiers. The X-axis corresponds to the molecular mass of the non substituted substrate, while the Y-axis corresponds to the ratio between the KM of the non substituted compound and its substituted counterpart. Black dashed line represents a ratio of 1, indicating no change in the kinetic parameter.