

# Structure-based activity prediction for an enzyme of unknown function

Johannes C. Hermann<sup>1</sup>, Ricardo Marti-Arbona<sup>2</sup>, Alexander A. Fedorov<sup>3</sup>, Elena Fedorov<sup>3</sup>, Steven C. Almo<sup>3</sup>, Brian K. Shoichet<sup>1</sup> & Frank M. Raushel<sup>2</sup>

**With many genomes sequenced, a pressing challenge in biology is predicting the function of the proteins that the genes encode. When proteins are unrelated to others of known activity, bioinformatics inference for function becomes problematic. It would thus be useful to interrogate protein structures for function directly. Here, we predict the function of an enzyme of unknown activity, Tm0936 from *Thermotoga maritima*, by docking high-energy intermediate forms of thousands of candidate metabolites. The docking hit list was dominated by adenine analogues, which appeared to undergo C6-deamination. Four of these, including 5-methylthioadenosine and S-adenosylhomocysteine (SAH), were tested as substrates, and three had substantial catalytic rate constants ( $10^5 \text{ M}^{-1} \text{ s}^{-1}$ ). The X-ray crystal structure of the complex between Tm0936 and the product resulting from the deamination of SAH, S-inosylhomocysteine, was determined, and it corresponded closely to the predicted structure. The deaminated products can be further metabolized by *T. maritima* in a previously uncharacterized SAH degradation pathway. Structure-based docking with high-energy forms of potential substrates may be a useful tool to annotate enzymes for function.**

For enzymes of unknown function, substrate prediction based on structural complementarity is, in principle, an alternative to bioinformatics inference of function<sup>1,2</sup>. Structure-based prediction becomes attractive when the target enzyme has little relationship to orthologues of known activity, making inference unreliable<sup>3,4</sup>. Whereas structure-based prediction has been used with some successes for inhibitor design, substrate prediction has proven difficult<sup>5–8</sup>. In addition to the well-known problems of sampling and scoring in docking, substrate prediction confronts several additional challenges. These include the many possible substrates to consider and the many reactions that an enzyme might catalyse<sup>9–11</sup>. Furthermore, enzymes preferably recognize transition states over the ground state structures that are usually represented in docking databases<sup>12–14</sup>.

## Docking metabolites as high-energy intermediates

If, in its most general form, structure-based substrate prediction seems daunting, it may be simplified by several pragmatic choices. If we focus on a single class of reactions, here those catalysed by the amidohydrolase superfamily (AHS), of which Tm0936 is a member, we reduce the number of possible reactions from practically unbounded to a limited set of mechanistically related transformations. Thus, the 6,000 catalogued members of the AHS catalyse ~30 reactions in biosynthetic and catabolic pathways<sup>15–17</sup>. All adopt a common ( $\beta/\alpha$ )<sub>8</sub>-barrel fold and almost all are metallo-enzymes that cleave carbon-hetero-atom bonds. The problem of activity prediction may be further simplified by focusing on a single source of likely substrates, here the KEGG metabolite database<sup>18</sup>. Although substrate identification remains challenging—there are probably hundreds of molecules that are specifically recognized, not all of which are metabolites—it is at least a finite problem.

To address the challenge of transition state recognition, ground state structures were transformed into structures mimicking the

high-energy intermediates that occur along the enzyme reaction coordinate. We will refer to these transition-state-like geometries as high-energy intermediates; this form of the substrate is among those that should best complement steric and electronic features of the enzyme active site<sup>14,19</sup>. All functional groups potentially recognized by AHS enzymes, for each of the 4,207 metabolites that bore them, were converted into high-energy intermediate geometries, with their appropriate charge distributions (Fig. 1). For instance, aromatic amines, which in the ground state are planar, are converted computationally into tetrahedral centres, representing the high-energy intermediate for deamination. Similarly, tetrahedral phosphates are converted into trigonal, bipyramidal forms. Overall, 28 amidohydrolase reactions operating on 19 functional groups were modelled by these high-energy structures, leading to the calculation of about 22,500 different forms of the metabolites. In retrospective calculations, docking these high-energy intermediate structures into seven well-studied amidohydrolases consistently identified the correct substrate from among the thousands of decoy molecules, typically outperforming docking of the ground state forms of the same molecules<sup>20,21</sup>.

These retrospective results encouraged us to prospectively predict the substrates of Tm0936 from *T. maritima*. The X-ray structure of the enzyme had been determined as part of a broad structural genomics effort (PDB codes 1p1m and 1j6p), and it can be assigned to the AHS by fold classification and the identity of certain active site groups. Despite this, its substrate preference is anything but clear. By sequence similarity, Tm0936 most resembles the large chlorohydrolase and cytosine deaminase subgroup, which is often used to annotate amidohydrolases of unknown function<sup>17</sup>. Consistent with the view that this reflects an assignment to a broad subfamily and not a functional annotation, we tested 14 cytosine derivatives as Tm0936 substrates; no turnover was detected for any of them (see Methods). In an effort to find the true substrate, we therefore docked the

<sup>1</sup>Department of Pharmaceutical Chemistry, University of California, San Francisco, MC 2550 1700 4th Street, San Francisco, California 94158-2330, USA. <sup>2</sup>Department of Chemistry, P.O. Box 30012, Texas A&M University, College Station, Texas 77842-3012, USA. <sup>3</sup>Department of Biochemistry, Albert Einstein College of Medicine, Ullmann Building, Room 411, 1300 Morris Park Avenue, Bronx, New York 10461, USA.

database of high-energy intermediates into the structure of Tm0936, sampling thousands of configurations and conformations of each molecule. Each of these was scored by electrostatic and van der Waals complementarity, corrected for ligand desolvation energy, and ranked accordingly (see Methods)<sup>22,23</sup>.

The molecules best-ranked computationally were dominated by adenine and adenosine analogues, which make up 9 of the 10 top-scoring docking hits (Table 1, Supplementary Fig. 1). For all of these, an exocyclic nitrogen has been transformed into a tetrahedral, high energy centre, as would occur in a deamination reaction. The dominance of adenine and adenosine analogues, in this form, is due to nearly ideal interactions with the active site. An example is the docked structure of the high-energy intermediate for the deamination of 5-methylthioadenosine (MTA), the 6th ranked molecule (Fig. 2).

### Experimental testing of the predicted substrates

On the basis of the docking ranks and compound availability, we selected four potential substrates for deamination by Tm0936: MTA, SAH, adenosine and adenosine monophosphate (AMP), all of which scored well (5th, 6th, 14th, 80th out of 4,207 docked metabolites), underwent the same reaction, and chemically resembled one another (Table 2). Although there were other high-ranking molecules in the docking hit list, most were single representatives of a chemotype and lacked the virtue of consistency of the adenines in general and the adenosines in particular. By extension, we also investigated the well-known metabolite *S*-adenosyl-L-methionine (SAM), a close analogue of SAH, even though its docking rank, at 511th, was poor.

Of these five molecules, three had substantial activity as substrates, with MTA and SAH reaching  $k_{cat}/K_m$  values of  $1.4 \times 10^5$  and  $5.8 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$  respectively, and adenosine close to  $10^4 \text{ M}^{-1}\text{s}^{-1}$  (Table 2 and Supplementary Information). The first order rate constant for the spontaneous deamination of adenosine in water is  $1.8 \times 10^{-10} \text{ s}^{-1}$ , making this enzyme proficient for these substrates. Tm0936 is relatively active compared to other adenosine deaminases<sup>24</sup>, especially because the optimal temperature for this thermophilic enzyme is almost certainly higher than the 30 °C at which it was assayed. Consistent with the docking predictions, SAM was not deaminated by Tm0936, despite its close similarity to SAH. Conversely, AMP, which did rank relatively well (80th of 4,207), was also not an enzyme substrate. The inability of the docking programme to fully de-prioritize AMP reflects some of the well-known problems in docking scoring functions, in this case balancing ionic interactions and desolvation penalties for the highly charged phosphate group of AMP.

To investigate the mechanism further, we determined the structure of Tm0936 in complex with the purified product of the SAH deamination reaction, *S*-inosylhomocysteine (SIH), to 2.1 Å resolution by X-ray crystallography (Fig. 3, Methods). The differences between the docked prediction and the crystallographic result are minor, with every key polar and non-polar interaction represented in both structures (except that we docked the tetrahedral intermediate and the X-ray structure is of the ground state product). Indeed, the

**Table 1 | The occurrence of adenine analogues among the top-ranked docking hits**

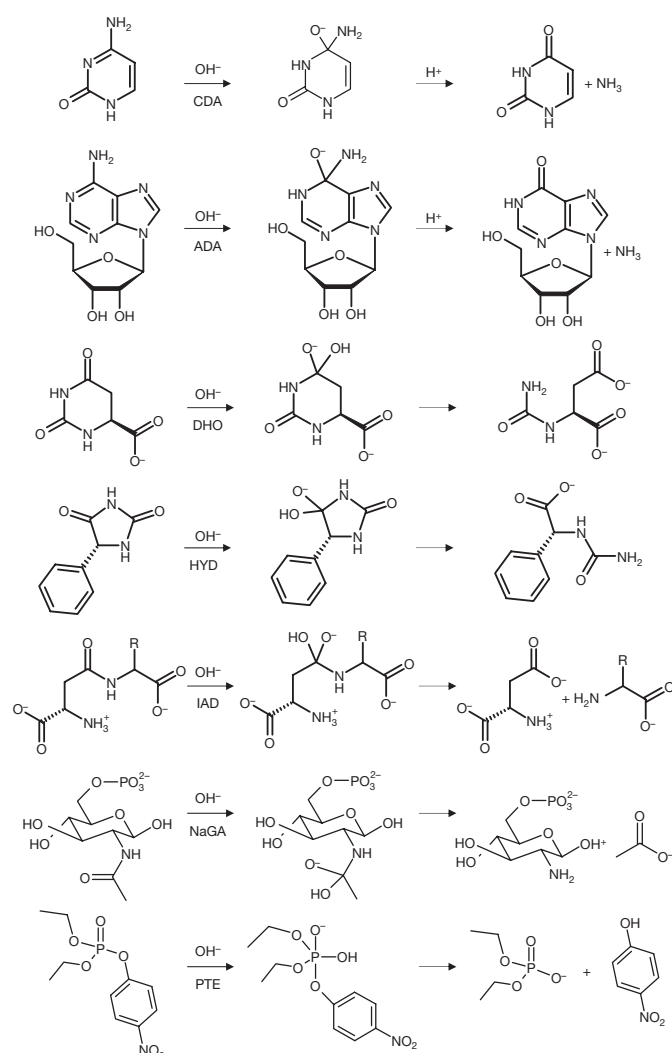
Analogues in docking hit list	Top 10 ranked hits	Top 20 ranked hits	Top 100 ranked hits	Top 300 ranked hits
Adenine analogues	9	17	32	44
Enrichment factor	34	32	12	6

The enrichment factor is measured relative to the abundance of the analogues among the 4,207 potential substrates docked.

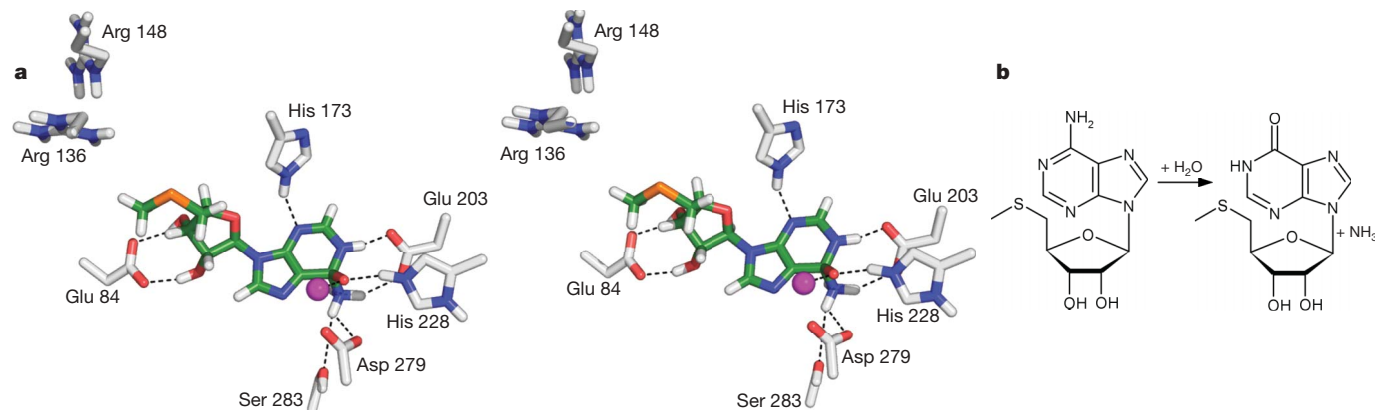
correspondence between the docked and crystallographic structures is closer than one might expect for inhibitor predictions, where docking has been more commonly used<sup>25–28</sup>. This may reflect the advantages of docking substrates in high-energy intermediate geometries, which encode more of the information necessary to specify fit.

### Metabolic pathway of a family of MTA/SAH deaminases

It is tempting to speculate that Tm0936 is not simply an isolated enzyme acting on particular substrates, but is involved in the deamination of metabolites in a previously uncharacterized MTA/SAH pathway. The deamination of adenosine itself is well known in all kingdoms of life, and the deamination of SAH to SIH has been reported in one organism, *Streptomyces flocculus*<sup>29</sup>. Very recently it was shown that MTA is deaminated in *Plasmodium falciparum* in an alternative degradation pathway of adenosine analogues<sup>30</sup>. To investigate whether the products of the deamination reactions, catalysed by Tm0936, SIH and MTI, could be further metabolized by other



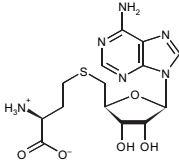
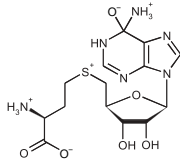
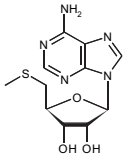
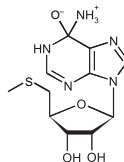
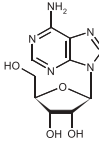
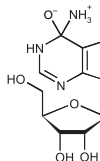
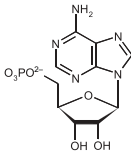
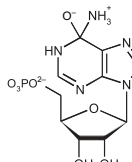
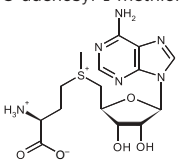
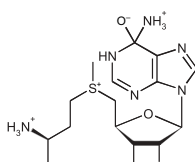
**Figure 1 | Sample transformations of metabolites from their ground state structure into the high-energy intermediate forms that were used for docking.** Transformations were computed according to the conserved reaction mechanism of amidohydrolases, a nucleophilic attack of a hydroxide at an electrophilic centre atom. Every transformable functional group for each molecule was processed independently. If the high-energy structure was chiral, all stereoisomers were calculated. Reactions catalysed by the amidohydrolases cytosine deaminase (CDA), adenosine deaminase (ADA), dihydroorotase (DHO), D-hydantoinase (HYD), isoaspartyl-D-dipeptidase (IAD), *N*-acetyl-D-glucosamine-6-phosphate deacetylase (NaGA) and phosphotriesterase (PTE) are shown.



**Figure 2 | Binding and conversion of MTA by Tm0936.** **a**, Stereoview of MTA in its high-energy intermediate form docked into the active site of Tm0936. Oxygen atoms are coloured red; enzyme carbons, grey; ligand carbons, green; hydrogens, white; nitrogens, blue; sulphur, orange; and the metal ion, purple. The oxyanion, representing the nucleophilic hydroxyl, ion-pairs with the metal ion and His 228. The ammonia leaving group is placed between Glu 203 and Asp 279, at 3.2 Å and 2.9 Å, respectively, also

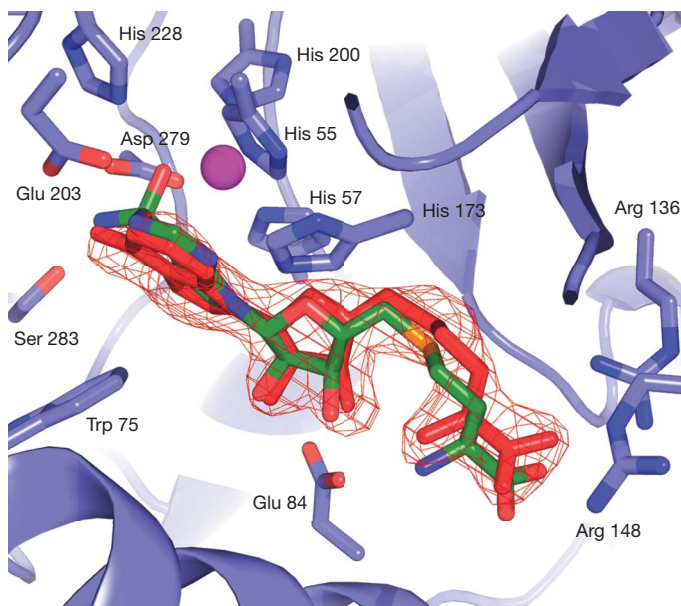
interacting with Ser 283 (3.2 Å). The N1-nitrogen donates a hydrogen bond to Glu 203, whereas N3 accepts one from His 173 (2.5 Å and 2.9 Å). Ribose hydroxyls hydrogen bond to Glu 84 (2.8 Å and 2.9 Å). Adenosines larger than MTA, such as SAH, make additional interactions with more distal residues, such as Arg 136 and Arg 148. All figures were rendered using PyMOL (<http://pymol.sourceforge.net>). **b**, The deamination of MTA to MTI, a reaction catalysed by Tm0936.

**Table 2 | Docking ranks and Tm0936 catalytic constants for five predicted substrates**

Substrate tested	Docked high-energy intermediate form	Dock rank	Relative docking scores (kcal mol <sup>-1</sup> )*	K <sub>m</sub> (μM)	k <sub>cat</sub> (s <sup>-1</sup> )	k <sub>cat</sub> /K <sub>m</sub> (M <sup>-1</sup> s <sup>-1</sup> )
 S-adenosyl-L-homocysteine		5	0	210 ± 40	12.2 ± 0.8	5.8 × 10 <sup>4</sup>
 5-Methyl-thioadenosine		6	4.4	44 ± 4	7.2 ± 0.2	1.4 × 10 <sup>5</sup>
 Adenosine		14	9.5	250 ± 40	2.3 ± 0.2	9.2 × 10 <sup>3</sup>
 Adenosine-5-monophosphate		80	20.2	ND	<10 <sup>-3</sup>	ND
 S-adenosyl-L-methionine		511	35.2	ND	<10 <sup>-3</sup>	ND

Deamination was measured by the production of ammonia. The standard deviations are given.

\* Docking energies relative to the best-ranked compound shown, SAH. Higher energies indicate worse scores. ND, not determined.



**Figure 3 | Comparing the docking prediction and the crystallographic result.** Superposition of the crystal structure of Tm0936 in complex with SIH (red) and the docking predicted structure of the high-energy intermediate of SAH (carbons in green). Enzyme carbons are coloured light blue, SAH and enzyme oxygen atoms are coloured red, nitrogens blue and sulphurs orange. The purple sphere represents the divalent metal ion. An  $F_o - F_c$  omit electron density map for SIH is shown, contoured at  $4.1 \sigma$ . The structure was determined at 2.1 Å resolution.

enzymes in *T. maritima*, we measured the activity of S-adenosyl homocysteinase (Tm0172), which hydrolyses SAH to homocysteine and adenosine, using SIH as a potential substrate. We found that Tm0172 catalyses the formation of homocysteine from either SIH or SAH about equally well (Supplementary Table 1 and Supplementary Information). This is consistent with Tm0172 and Tm0936 participating in a degradation pathway, though it does not confirm it. We cannot exclude the possibility that Tm0936 functions as an adenosine deaminase in *T. maritima*, because no other enzyme in the organism has been identified that serves this role.

What is clear is that Tm0936 has orthologues across multiple species. On the basis of the conservation of characteristic residues that interact with the substrate and product in the docked and X-ray structures, respectively, 78 other previously unannotated AHS enzymes from different species may now be classified as MTA/SAH/adenosine deaminases (Supplementary Fig. 2 and Supplementary Information). In all of these sequences, the metal-ligating residues (His 55, His 57, His 200 and Asp 279, Tm0936 numbering) are conserved, as are the residues recognizing the reactive centre (His 228, Ser 259, Ser 283 and Glu 203). Specificity is conferred by interactions between the substrate and Trp 75, Glu 84 and His 173, all of which are also conserved among the 78 amidohydrolases. Active site residues that vary include Arg 136 and Arg 148, which in Tm0936 interact with the  $\alpha$ -carboxylate of the homocysteine moiety of SAH. These latter interactions are not critical to the activity of the enzyme, because these arginines do not seem to interact with MTA or adenosine, but they may be important for the recognition of SAH.

Many of the Tm0936 orthologues cluster with other genes that can now be associated with the metabolism of SAM, SAH or MTA. For example, in *T. maritima* Tm0936 is closely associated with Tm0938, which is currently annotated as a SAM-dependent methyl transferase. In *Bacillus cereus*, the Tm0936 orthologue is Bc1793, which is also closely associated with a SAM-dependent methyl transferase, Bc1797. In *Pseudomonas aeruginosa*, the Tm0936 orthologue, Pa3170, is adjacent to UbiG-methyltransferase, Pa3171. Other orthologues are adjacent or close to adenosyl homocysteinase, 5'-methylthioadenosine

phosphorylase, MTA/SAH nucleosidase and other SAM-dependent methyl transferases.

### Predicting function from form

This work describes one case of successful function prediction by structure-based docking, and it is appropriate to consider caveats. Our recognition of Tm0936 as an amidohydrolase limited the number of possible reactions to be considered. When even the gross mechanistic details of an enzyme cannot be inferred, this will not be possible. Restricting ourselves to metabolites was also helpful, but this too will not always be appropriate. Finally, we were fortunate that Tm0936 experienced little conformational change between the apo structure and that of the product complex. Enzymes that undergo large conformational changes along their reaction coordinates will be more challenging for docking.

If prudence warns against over-generalization, it is also unlikely that Tm0936 represents an isolated case. Other enzyme structures will be broadly classifiable by mechanism, and whereas conformational change remains a serious challenge, retrospective studies suggest that it is not insurmountable. Indeed, the most important technical innovation adopted here, modelling substrates as high-energy intermediates, was particularly useful when docking to apo structures in those studies (Supplementary Table 2 and Supplementary Information)<sup>20</sup>. Thus, the prediction and determination that Tm0936 acts as an MTA/SAH deaminase illustrates the possibilities of this and related structure-based approaches, at least for a subset of targets. The enzyme has no obvious sequence similarity to any known adenosine deaminase and exploits interactions not previously identified in the active sites of these enzymes. The very pathway in which Tm0936 participates seems novel. Structure-based docking of high-energy intermediates should be a useful tool to decrypt the activity of enzymes of unknown function, and will be especially interesting for those targets where bioinformatics inference is unreliable.

### METHODS

**Molecular docking.** The 1.5 Å X-ray structure of Tm0936 (Protein Data Bank (PDB) code 1P1M) was used in docking calculations. High-energy intermediates of potential substrates were calculated<sup>20</sup> and docked into the enzyme structure using the program DOCK3.5.54. Poses were scored for electrostatic and van der Waals complementarity and penalized for ligand desolvation<sup>31,32</sup>.

**Enzymology.** Tm0936 and Tm0172 from *T. maritima* were cloned, expressed and purified using standard techniques. The deamination reaction was measured by coupling the production of ammonia to the oxidation of NADH catalysed by glutamate dehydrogenase. The decrease in the concentration of NADH was followed spectrophotometrically at 340 nm. The chemical identities of the deaminated products were confirmed by mass spectrometry and specific changes in the ultraviolet absorption (UV) spectra for the deamination of adenosine derivatives. The SAH hydrolase activity by Tm0172 was determined by reaction of the free thiol group of the homocysteine product with dithio-bis(2-nitrobenzoic acid), monitoring the absorbance at 412 nm.

**X-ray crystallography.** Tm0936 was co-crystallized with ZnCl<sub>2</sub> and SIH. X-ray diffraction data were collected at the NSLS X4A beamline (Brookhaven National Laboratory). The structure of the Tm0936-SIH complex was determined by molecular replacement, using apo Tm0936 (PDB code 1J6P) as the search model. The structure has been deposited in the protein data bank (PDB code 2PLM).

**Full Methods** and any associated references are available in the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

Received 24 January; accepted 7 June 2007.

Published online 1 July 2007.

- Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36**, 307–340 (2003).
- Gerlt, J. A. & Babbitt, P. C. Can sequence determine function? *Genome. Biol.* **1**, REVIEWS0005 (2000).
- Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15**, 132–133 (1999).
- Devos, D. & Valencia, A. Intrinsic errors in genome annotation. *Trends Genet.* **17**, 429–431 (2001).
- Schapira, M., Abagyan, R. & Totrov, M. Nuclear hormone receptor targeted virtual screening. *J. Med. Chem.* **46**, 3045–3059 (2003).

6. Rao, M. S. & Olson, A. J. Modelling of factor Xa-inhibitor complexes: a computational flexible docking approach. *Proteins* **34**, 173–183 (1999).
7. Sukuru, S. C. *et al.* Discovering new classes of *Brugia malayi* asparaginyl-tRNA synthetase inhibitors and relating specificity to conformational change. *J. Comput. Aided. Mol. Des.* **20**, 159–178 (2006).
8. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).
9. Macchiarulo, A., Nobeli, I. & Thornton, J. M. Ligand selectivity and competition between enzymes *in silico*. *Nature Biotechnol.* **22**, 1039–1045 (2004).
10. Kalyanaraman, C., Bernacki, K. & Jacobson, M. P. Virtual screening against highly charged active Sites: identifying substrates of  $\alpha$ - $\beta$  barrel enzymes. *Biochemistry* **44**, 2059–2071 (2005).
11. Irwin, J. J. & Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
12. Schramm, V. L. Enzymatic transition states and transition state analogues. *Curr. Opin. Struct. Biol.* **15**, 604–613 (2005).
13. Hermann, J. C., Ridder, L., Holtje, H. D. & Mulholland, A. J. Molecular mechanisms of antibiotic resistance: QM/MM modelling of deacylation in a class A  $\beta$ -lactamase. *Org. Biomol. Chem.* **4**, 206–210 (2006).
14. Warshel, A. & Florian, J. Computer simulations of enzyme catalysis: finding out what has been optimized by evolution. *Proc. Natl Acad. Sci. USA* **95**, 5950–5955 (1998).
15. Holm, L. & Sander, C. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins* **28**, 72–82 (1997).
16. Seibert, C. M. & Raushel, F. M. Structural and catalytic diversity within the amidohydrolase superfamily. *Biochemistry* **44**, 6383–6391 (2005).
17. Pegg, S. C. *et al.* Leveraging enzyme structure–function relationships for functional inference and experimental design: the structure–function linkage database. *Biochemistry* **45**, 2545–2555 (2006).
18. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
19. Tantillo, D. J. & Houk, K. N. Transition state docking: a probe for noncovalent catalysis in biological systems. Application to antibody-catalyzed ester hydrolysis. *J. Comput. Chem.* **23**, 84–95 (2002).
20. Hermann, J. C. *et al.* Predicting substrates by docking high-energy intermediates to enzyme structures. *J. Am. Chem. Soc.* **128**, 15882–15891 (2006).
21. Nowlan, C. *et al.* Resolution of chiral phosphate, phosphonate, and phosphinate esters by an enantioselective enzyme library. *J. Am. Chem. Soc.* **128**, 15892–15902 (2006).
22. Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W. & Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **322**, 339–355 (2002).
23. Lorber, D. M. & Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Curr. Top. Med. Chem.* **5**, 739–749 (2005).
24. Radzicka, A. & Wolfenden, R. A proficient enzyme. *Science* **267**, 90–93 (1995).
25. Mohan, V., Gibbs, A. C., Cummings, M. D., Jaeger, E. P. & DesJarlais, R. L. Docking: successes and challenges. *Curr. Pharm. Des.* **11**, 323–333 (2005).
26. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303**, 1813–1818 (2004).
27. Kairys, V., Fernandes, M. X. & Gilson, M. K. Screening drug-like compounds by docking to homology models: a systematic study. *J. Chem. Inf. Model.* **46**, 365–379 (2006).
28. Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **11**, 580–594 (2006).
29. Speedie, M. K., Zully, J. J. & Brothers, P. S-adenosylhomocysteine metabolism in *Streptomyces flocculus*. *J. Bacteriol.* **170**, 4376–4378 (1988).
30. Tyler, P. C., Taylor, E. A., Fröhlich, R. F. G. & Schramm, V. L. Synthesis of 5'-methylthio coformycins: specific inhibitors for malarial adenosine deaminase. *J. Am. Chem. Soc.* **129**, 6872–6879 (2007).
31. Meng, E. C., Shoichet, B. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comp. Chem.* **13**, 505–524 (1992).
32. Gschwend, D. A. & Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal. *J. Comput. Aided Mol. Des.* **10**, 123–132 (1996).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** This work was supported by grants from the National Institutes of Health, supporting docking analyses (to B.K.S.), large scale structural analysis (to S.C.A.), and function prediction (to F.M.R., B.K.S. and S.C.A.). F.M.R. thanks the Robert A. Welch Foundation for support. J.C.H. thanks the Deutsche Akademie der Naturforscher Leopoldina for a fellowship. We thank J. Irwin, V. Thomas and K. Babaoglu for reading this manuscript. The clone for Tm0172 was kindly supplied by the Joint Center for Structural Genomics.

**Author Contributions** J.C.H. designed the docking database, performed the docking runs, and analysed the docking results. F.M.R. and R.M.-A. performed the enzymatic characterization of Tm0936 and Tm0172, including cloning and purification of the proteins. S.C.A., E.F. and A.A.F. determined the X-ray structure of Tm0936 with S-inosyl-homocysteine. J.C.H. and B.K.S. largely wrote the paper. All authors discussed the results and commented on the manuscript.

**Author Information** The complex structure of Tm0936 with SIH has been deposited in the PDB (accession code 2PLM). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Correspondence and requests for materials related to docking should be addressed to B.K.S. ([shoichet@cgl.ucsf.edu](mailto:shoichet@cgl.ucsf.edu)).

## METHODS

**Molecular docking.** The 1.5 Å X-ray structure of Tm0936 was used in the docking calculations (PDB code 1P1M). The active site metal ion was assigned a charge of +1.4, the remaining charge of 0.6 was distributed among the ligating residues, His 55, His 57, His 200 and Asp 279, to keep the correct net charge and to account for charge distribution effects in metal complexes<sup>20,33</sup>. His 228 was protonated according to its assumed function as the base to activate the catalytic water molecule by abstracting a proton; the hydroxide ion itself was removed from the active site, because it is part of each high-energy intermediate structure that we dock<sup>20</sup>.

The programs CHEMGRID and DISTMAP were used to compute docking grids for van der Waals potentials and excluded volume, respectively<sup>31</sup>. The electrostatic potential grid was calculated with DELPHI using an internal dielectric of 2 and an external dielectric of 78 (ref. 34). A manually curated set of spheres based on a set calculated by the program SPHGEN was used to orient molecules in the binding site<sup>36</sup>.

High-energy intermediates of potential substrates were docked into Tm0936 using the docking program DOCK3.5.54 (ref. 23). Initial ligand orientations were sampled using receptor and ligand bin sizes of 0.5 Å and a ligand and receptor overlap of 0.4 Å. A distance tolerance of 1.5 Å was used for matching receptor and ligand spheres. An average of more than a million poses per molecule were calculated, and those that sterically fit the site were scored for electrostatic and van der Waals complementarity and penalized for ligand desolvation<sup>32</sup>. The best scoring orientation was rigid-body-minimized according to these energies. The details of the preparation of high-energy structures for docking, the molecular docking procedure, and methods for the analysis of the results have been previously described<sup>20</sup>.

**Enzymatic characterization of Tm0936.** All compounds and coupling enzymes were obtained from Sigma or Aldrich, unless otherwise specified. The genomic DNA from *T. maritima* was purchased from the American Type Culture Collection (ATCC). The oligonucleotide synthesis and DNA sequencing reactions were performed by the Gene Technology Laboratory of Texas A&M University. The pET30a(+) expression vector was acquired from Novagen. The T4 DNA ligase and the restriction enzymes, *NdeI* and *EcoRI*, were purchased from New England Biolabs. The Platinum *Pfx* DNA polymerase and the Wizard Plus SV Mini-Prep DNA purification kit were obtained from Invitrogen and Promega, respectively. The glycerol stock of the plasmid encoding Tm0172 was kindly provided by the Joint Center for Structural Genomics.

**Cloning of Tm0936.** The gene encoding Tm0936 from *Thermotoga maritima* was amplified from the genomic DNA by standard PCR methods stipulated in the manufacturer's instructions using oligonucleotide primers with *NdeI* and *EcoRI* restriction sites at either end (Supplementary Table 3). The PCR products were purified, digested with *NdeI* and *EcoRI*, ligated to the expression vector pET30a(+) using T4 DNA ligase, and then transformed into XL1Blue cells. Individual colonies containing the plasmid were selected on LB plates containing 50 µg ml<sup>-1</sup> kanamycin and then used to inoculate 5 ml cultures of LB. The entire coding regions of the plasmids containing the *Tm0936* gene were sequenced to confirm the fidelity of the PCR amplification.

**Purification of Tm0936.** Cells harbouring the plasmid for the expression of Tm0936 were grown overnight and a single colony was used to inoculate 50 ml of LB media containing 50 µM kanamycin, and subsequently used to inoculate 2 l of the same medium. Cell cultures were grown at 37 °C with a rotary shaker until an  $A_{600}$  of ~0.6 was reached. Induction was initiated by the addition of 1.0 mM isopropyl-thiogalactoside (IPTG), and further incubated overnight at 30 °C. The bacterial cells were isolated by centrifugation at 5,200 × *g* for 15 min at 4 °C. The pellet was re-suspended in 50 mM HEPES buffer, pH 7.5 (buffer A), containing 5 µg ml<sup>-1</sup> RNase and 0.1 mg ml<sup>-1</sup> PMSF per gram of wet cells and then disrupted by sonication. The soluble protein was separated from the cell debris by centrifugation at 14,000 × *g* for 15 min and heated at 65 °C for 15 min to precipitate the *Escherichia coli* proteins. The soluble protein was separated from the precipitated protein by centrifugation at 14,000 × *g* for 15 min, loaded onto a 6 ml Resource Q anion ion exchange column (GE Health Care) and eluted with a gradient of NaCl in 20 mM HEPES, pH 8.5 (buffer B). The fractions containing Tm0936 were pooled and re-precipitated by saturation with ammonium sulphate, centrifuged at 14,000 × *g* for 15 min at 4 °C, and resuspended in a minimum amount of buffer A. The final step in the purification was accomplished by chromatography on a High Load 26/60 Superdex 200 prep grade gel filtration column (GE Health Care) and eluted with buffer A. The purity of the protein during the isolation procedure was monitored by SDS-PAGE.

**Purification of Tm0172.** Cells harbouring the plasmid for the expression of Tm0172 were grown overnight and a single colony was used to inoculate 50 ml of LB media containing 100 µM ampicillin and subsequently used to inoculate 2 l of the same medium. Cell cultures were grown at 37 °C with a rotary

shaker until an  $A_{600}$  of ~0.6 was reached. Induction was initiated by the addition of 1.0 mM arabinose, and further incubated overnight at 37 °C. The bacterial cells were isolated by centrifugation at 5,200 × *g* for 15 min at 4 °C. The pellet was re-suspended in 20 mM Tris-Cl buffer, 5 mM imidazole and 500 mM NaCl at pH 7.5 (buffer A), containing 0.1 mg ml<sup>-1</sup> phenylmethylsulphonyl fluoride per gram of wet cells and then disrupted by sonication. The soluble protein was separated from the cell debris by centrifugation at 14,000 × *g* for 15 min and heated at 65 °C for 15 min to precipitate the *E. coli* proteins. The soluble protein was separated from the precipitated protein by centrifugation at 14,000 × *g* for 15 min, loaded onto a Chelating Sepharose Fast Flow column for histidine-tagged fusion protein purification and eluted with a gradient of imidazole in buffer A. Fractions containing the desired protein were pooled by catalytic activity and purity. The purity of the protein during the isolation procedure was monitored by SDS-PAGE.

**Metal analysis and amino acid sequence verification.** The purified Tm0936 was subjected to amino-terminal amino acid sequence analysis by the Protein Chemistry Laboratory at Texas A&M University. The first five amino acids were MIIGN, which agrees with the protein sequence reported for Tm0936. The protein concentration was determined spectrophotometrically at 280 nm using a SPECTRAMax-340 microplate reader (Molecular Devices). An extinction coefficient of 51,020 M<sup>-1</sup>cm<sup>-1</sup> was used for Tm0936 on the basis of the protein sequence. The metal content of the purified protein was determined by inductively coupled plasma emission-mass spectrometry (ICP-MS) and found to contain 1.2 equivalents of Zn per subunit.

**Determination of SAH deaminase activity.** The measurement of the deaminating properties of Tm0936 was conducted by coupling the production of ammonia to the oxidation of NADH with glutamate dehydrogenase. The decrease in the concentration of NADH was followed spectrophotometrically at 340 nm using a SPECTRAMax-340 microplate reader. The standard assay was modified from the report in ref. 36, and contained 100 mM HEPES at pH 8.0, 7.4 mM α-ketoglutarate, 0.4 mM NADH, 6 units of glutamate dehydrogenase, Tm0936 and the appropriate compound in a final volume of 250 µl at 30 °C. Following the initial, purely bioinformatic predictions of cytosine deaminase activity, the following compounds were tested for enzymatic activity at a concentration of 10 mM using this protocol: cytosine, 5-methylcytosine, 5-fluorocytosine, 6-aminouracyl, 4,6-diamino-2-hydroxypyrimidine, 2-deoxycytidine, cytosine-β-D-arabinofuranoside, cytidine, cytidine-5'-diphosphocholine, cytidine-5'-monophosphate, 2'-deoxycytidine-5'-diphosphate, cytidine-5'-diphosphate, cytidine-5'-triphosphate, cytidine-3'-phosphate. Subsequently, we cast a wider net looking for activity on *N*-formimino-L-glutamate, *N*-formimino-L-aspartate, *N*-formimino-L-glycine. It was only with the structure-based docking predictions that we turned to deamination of adenosines, first directly testing the docking predicted metabolites adenosine, adenosine-5'-monophosphate, 5'-methylthioadenosine, *S*-adenosine-5'-homocysteine. Eventually we tested also several other analogues including adenosine-5'-diphosphate, adenosine-5'-triphosphate, *S*-adenosine-5'-methionine, folate, thiamine, pterin, and guanine. Only adenosine, *S*-adenosine-5'-homocysteine, and 5'-methylthioadenosine were found to be substrates.

The products of the reaction catalysed by Tm0936 were confirmed by mass spectroscopy and by changes in the UV spectra. When *S*-adenosine-5'-homocysteine was incubated with Tm0936, the mass spectral signal for SAH at a  $[M+H]$  of 385 *m/z* disappeared and was replaced by a new signal at a  $[M+H]$  of 386 *m/z* that corresponds to the mass expected for *S*-inosyl-5'-homocysteine (SIH). The UV spectrum for SAH has a maximal absorbance at 260 nm and after the addition of Tm0936 the absorbance maximum shifts to 250 nm. These results are consistent with the deamination of the adenine moiety of the substrate and conversion to an inosyl substituent. Similar results were observed for 5'-methylthioadenosine ( $[M+H]$  of 298 *m/z* and  $A_{\max}$  of 260 nm) on conversion to 5'-methylthioinosine ( $[M+H]$  of 299 *m/z* and  $A_{\max}$  of 250 nm) and adenosine ( $[M+H]$  of 268 *m/z* and  $A_{\max}$  of 260 nm) on conversion inosine ( $[M+H]$  of 269 *m/z* and  $A_{\max}$  of 250 nm) with an isobestic point at 251 nm (Supplementary Information and Supplementary Fig. 3).

**Determination of SAH hydrolase activity.** The homocysteinase activity of Tm0172 was determined by reaction of the free thiol group of the homocysteine product with DTNB. The increase in the absorbance at 412 nm was monitored using an extinction coefficient of 13,600 M<sup>-1</sup>cm<sup>-1</sup> (ref. 37). The standard assay contained 100 mM HEPES at pH 8.0, 1.0 mM DTNB, 1.0 mM EDTA, 13 µM Tm0172 and the appropriate substrate in a final volume of 250 µl at 30 °C. The following compounds were tested for catalytic activity at concentrations up to 10 mM: *S*-adenosine-5'-homocysteine, *S*-inosyl-5'-homocysteine, 5'-methylthioadenosine and 5'-methylthioinosine. Activity was obtained only for SAH and SIH.

**Data analysis.** The kinetic parameters,  $k_{\text{cat}}$ , and  $k_{\text{cat}}/K_m$  were determined by fitting the initial velocity data to the equation (1), where  $v$  is the initial velocity,  $E_T$  is the enzyme concentration,  $k_{\text{cat}}$  is the turnover number,  $S$  is the substrate

concentration, and  $K_m$  is the Michaelis constant<sup>38</sup>. In the cases where substrate inhibition was observed an extra parameter ( $K_{is}$ ) was included to calculate the apparent inhibition constant for the substrate, as observed in equation (2) (ref. 39):

$$v/E_T = k_{cat}S/(K_m + S) \quad (1)$$

$$v/E_T = k_{cat}S/[K_m + S + (S^2/K_{is})] \quad (2)$$

**Sequence alignment.** A multiple sequence alignment of Tm0936 with the likely orthologues from other organisms is presented in Supplementary Figure 2 of Supplementary Information.

**X-ray crystallography.** Tm0936 was co-crystallized with S-inosyl-homocysteine (SIH) and  $ZnCl_2$ . The enzyme solution at  $12.9 \text{ mg ml}^{-1}$  in 20 mM HEPES, pH 8.0 was incubated for 60 min at 4 °C with 10 mM SIH and 0.5 mM  $ZnCl_2$ . The ternary complex was crystallized by hanging drop vapour diffusion using 1  $\mu\text{l}$  of the protein–ligand solution and 1  $\mu\text{l}$  of a reservoir solution containing 3.5 M Na formate, 0.5 mM  $ZnCl_2$ , pH 7.0. Crystals appeared in 1–2 days and exhibited diffraction consistent with the space group  $P3_221$  ( $a = 113.28 \text{ \AA}$ ,  $c = 80.30 \text{ \AA}$ , with 1 molecule of the ternary complex per asymmetric unit). X-ray diffraction data to 2.1  $\text{\AA}$  were collected at the NSLS X4A beamline (Brookhaven National Laboratory) on an ADSC CCD detector. Diffraction data were integrated and scaled using the programs DENZO and SCALEPACK<sup>40</sup>. The final 2.1  $\text{\AA}$  data set was 93.2% complete with  $R_{merge} = 0.097$ .

**Structure determination and model refinement.** The structure of the ternary Tm0936·SIH·Zn complex was solved by molecular replacement with the program PHASER<sup>41</sup>, using apo Tm0936 (PDB code 1J6P) as the search model. The solution was subsequently refined with CNS<sup>42</sup>. The bound SIH and Zn were clearly visible in the electron density maps after the first cycle of rigid body refinement of the protein molecule alone. Iterative cycles of manual rebuilding with TOM<sup>43</sup> and refinement with CNS resulted in a model with  $R_{cryst}$  and  $R_{free}$  of 0.209 and 0.238, respectively. The final structure contains 3,210 protein atoms, 1 inhibitor molecule, 1 Zn atom, and 76 water molecules for one monomer of the complex in the asymmetric unit (Supplementary Information and Supplementary Table 4).

33. Irwin, J. J., Raushel, F. M. & Shoichet, B. K. Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry* **44**, 12316–12328 (2005).
34. Gilson, M. K. & Honig, B. H. Calculation of electrostatic potentials in an enzyme active site. *Nature* **330**, 84–86 (1987).
35. Kuntz, I. D. et al. A Geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **161**, 269–288 (1982).
36. Muszbek, L., Polgar, J. & Fesus, L. Kinetic determination of blood coagulation Factor XIII in plasma. *Clin. Chem.* **31**, 35–40 (1985).
37. Ellman, G. L. A colorimetric method for determining low concentrations of mercaptans. *Arch. Biochem. Biophys.* **74**, 443–450 (1958).
38. Cleland, W. W. Statistical analysis of enzyme kinetic data. *Methods Enzymol.* **63**, 103–138 (1979).
39. Cleland, W. W. Substrate inhibition. *Methods Enzymol.* **63**, 500–513 (1979).
40. Otwinowski, Z. & Minor, W. in *Methods in Enzymology* Vol. 276 (eds Carter, C. W. & Sweet, R. M.) 307–326 (Academic Press, New York, 1997).
41. Storoni, L. C., McCoy, A. J. & Read, R. J. Likelihood-enhanced fast rotation functions. *Acta Crystallogr. D* **60**, 432–438 (2004).
42. Brunger, A. T. et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
43. Jones, T. A. Diffraction methods for biological macromolecules. Interactive computer graphics: FRODO. *Methods Enzymol.* **115**, 157–171 (1985).