Research article

The 3D modules of enzyme catalysis: deconstructing active sites into distinct functional entities

Ioannis G. Riziotis[1*], António J. M. Ribeiro[1], Neera Borkakoti[1], Janet M. Thornton[1]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, CB10 1SD, Cambridge, UK

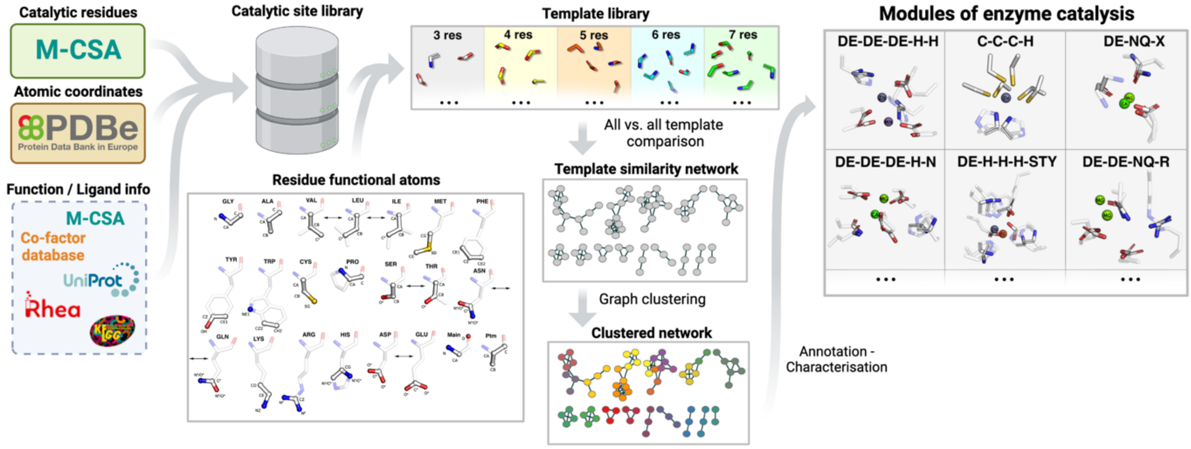
[*]*To whom correspondence should be addressed: riziotis@ebi.ac.uk, tel.: 0044 77 490 32067*

Running title:

The 3D modules of enzyme catalysis

Abstract

Enzyme catalysis is governed by a limited toolkit of residues and organic or inorganic co-factors. Therefore, it is expected that recurring residue arrangements will be found across the enzyme space, which perform a defined catalytic function, are structurally similar and occur in unrelated enzymes. Leveraging the integrated information in the Mechanism and Catalytic Site Atlas (M-CSA) (enzyme structure, sequence, catalytic residue annotations, catalysed reaction, detailed mechanism description), 3D templates were derived to represent compact groups of catalytic residues. A fuzzy template-template search, allowed us to identify those recurring motifs, which are conserved or convergent, that we define as the "modules of enzyme catalysis". We show that a large fraction of these modules facilitate binding of metal ions, co-factors and substrates, and are frequently the result of convergent evolution. A smaller number of convergent modules perform a well-defined catalytic role, such as the variants of the catalytic triad (i.e. Ser-His-Asp/Cys-His-Asp) and the saccharide-cleaving Asp/Glu triad. It is also shown that enzymes whose functions have diverged during evolution preserve regions of their active site unaltered, as shown by modules performing similar or identical steps of the catalytic mechanism. We have compiled a comprehensive library of catalytic modules, that characterise a broad spectrum of enzymes. These modules can be used as templates in enzyme design and for better understanding catalysis in 3D.



Keywords

catalysis, enzyme, catalytic residues, structure, modules, active site, convergence

Introduction

Complex catalytic processes in biological systems are often performed by one or more discrete functional components, comprised of a small number of chemical groups in a defined conformation. Those groups are contributed by basic units, the catalytic residues, that have evolved to be arranged in a defined geometry and are brought into close proximity by the folding of the polypeptide chain. Over the years, our group has been collecting and curating information on the structure and function of enzyme catalytic sites, cataloguing over 1000 enzyme families in the Mechanism and Catalytic Site Atlas (M-CSA)[1,2]. In this database, we have empirically observed that several enzymes in nature adhere to the "discrete functional entities" paradigm, with a simple example shown in Fig. 1. Quinoprotein Glucose Dehydrogenase (M-CSA 104, E.C. 1.1.5.2) is a bacterial oxidoreductase catalysing the oxidation of glucose to gluconolactone with the concurrent reduction of ubiquinone to ubiquinol[3,4]. The enzyme requires pyrroloquinoline quinone (PQQ) and a Ca^{2+} ion as co-factors participating in the oxidation-reduction reaction as well as seven catalytic residues. Even a brief inspection of the catalytic site 3D structure reveals two discrete, compact residue clusters, of three and four residues. The four-residue cluster (Ala293, Tyr295, Asp297, Glu333) binds a Ca^{2+} ion, with the conformation of the residues being driven by organometallic interactions, while the three residue cluster is responsible for the core catalytic process, involving the glucopyranose substrate and the PQQ co-factor. Specifically, a catalytic His168 has the role of the zwitterionic proton donor and acceptor (general acid/base), providing the starting point of a hydride transfer from the C1-OH of the glucose by deprotonation, to the C5 of the PQQ co-factor. The adjacent Asp187 and Arg252 act as activator of His444 and electrostatic stabiliser of negative charge on the PQQ cofactor, respectively. These two entities are simultaneously independent in the 3D space and co-dependent through the course of the reaction (the presence of Ca^{2+} is critical for the activity of the enzyme). Such metal binding sites, which are present in proteins with different folds and functions, are often a result of convergent evolution, with the residue geometry being driven by the coordination sphere of the bound metal. Interestingly in this example, two residues of the metal binding site (Ala293 and Tyr295) ligate the metal through the oxygen atoms of the backbone, and those are the most variable residue positions in the multiple sequence alignment of active sites of the homologous family sequences. This clearly explains the reduced evolutionary pressure in those positions, and the "generic" nature of such motifs within varying structural contexts. Similarly, oxyanion holes in serine proteinases (and other enzymes involving deprotonation of oxygen moieties, usually polymer-cleaving enzymes[5–7]) are formed by two backbone amide groups creating a positively charged pocket that stabilises a negative charge. This is another example of an evolutionarily "generic" motif that has evolved independently in multiple different enzymes. On the other hand, there are residue arrangements that define the specific function of an enzyme and are better conserved and less frequently a result of convergent evolution. Those are usually catalytic sites of enzymes performing complex reactions with several mechanistic steps, involving multiple different functional groups, like in the case of Class A Beta-lactamases (EC 3.5.2.6). Other Beta-lactamases have also converged at the reaction level, with some of which to share similarities in mechanism. A few, for example, use Ser as a nucleophile although surrounding residues are different.

Such observations led us to define the concept of the catalytic "module": a self-contained, compact arrangement of residues, performing one or more tasks in the enzyme catalysis landscape. Each module is named after its residue composition (one letter-code, also accounting for fuzziness – see Methods). Modules can be catalytically active, such as the Arg-His-Asp triad in Quinoprotein Glucose Dehydrogenase, or might be more generic, such as the metal binding site in the same enzyme. Catalytically active modules, however, are not necessarily invariable in terms of 3D conformation and function. For instance, the Arg-His-Asp catalytic triad (the respective module would be named "DE-H-R", to indicate the presence an acid residue, Asp or Glu) is also present in phospholipase C and performs a different function (phosphodiester bond cleavage)[8] with the 3D arrangement of residues differing too. Furthermore, structural convergence can also apply to catalytically active modules, like the Ser-His-Asp catalytic triad, the most definitive example of active site convergent evolution in enzymology, which is present in multiple analogues of serine proteinases, all independently evolved[9–12].

Here we aim to collect generic and function specific modules that recur in enzymes. These can be derived from enzymes with different folds (convergent) or related enzymes with the same fold but different function (divergent). In addition, we elaborate on the concept of "partial" templates to decompose active sites into compact residue arrangements, identifying recurring instances of those arrangements observed to date, to catalogue, classify and functionally characterise most generic modules known to date governing enzyme catalysis.

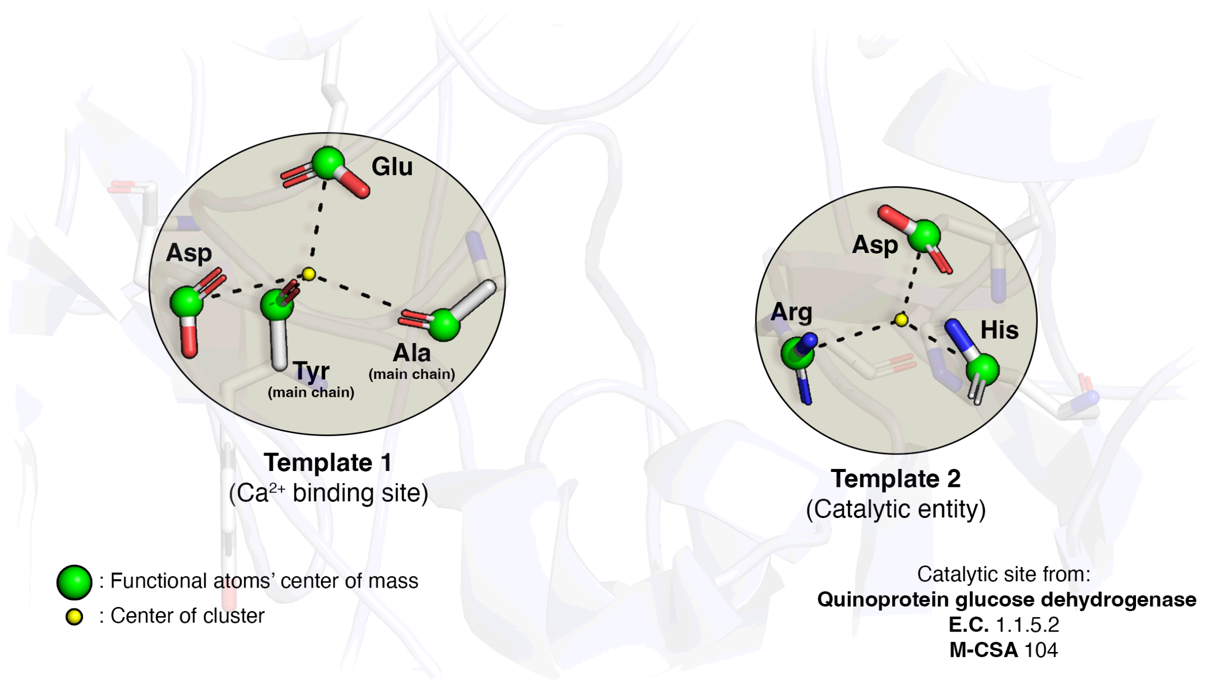


Fig. 1: The concept of active site modularity and template fragmentation method. Functional atoms of catalytic residues are shown in sticks (white: C, red: O, blue: N) with their centres of geometry shown as green spheres. Residue clusters ("partial" templates), identified by the k-means algorithm in 3D space, are shaded in grey circles and centres of each cluster are shown as yellow spheres. A partial template of size s consists of the first s residues closest to the cluster centre (dashed lines indicate distances).

Results

Definitions

First, it is important to distinguish the definitions for the terms "catalytic site", "template" and "module" used throughout this paper:

- **Catalytic site**: A collection of three or more catalytically important and conserved residues, manually annotated on enzyme structures in the Mechanism and Catalytic Site Atlas.
- **Template**: An abstract representation of an active site that includes only three functional atoms per catalytic residue. Templates are used as 3D probes to identify similar motifs in a structure, and also to account for matching fuzziness in the atom or residue type. Templates can be "complete" if they represent all catalytic residues of a site, or "partial" if they represent a subset of densely grouped catalytic residues.
- **Module**: A set of recurring partial templates (identical or chemically similar), with defined residue composition and similar geometry, found in at least two evolutionary

related (divergent) or unrelated proteins (convergent). A module might contain a mixture of divergent or convergent templates.

Template library

Using the methods described further below, we derived templates from M-CSA, which contains 1003 enzymes at the time of this study, with 835 functions, as defined by the 4th level of EC nomenclature[1]. The M-CSA was initially constructed to include examples of enzymes that are unique in at least one of the following characteristics: mechanism, overall reaction, or catalytic machinery[13]. With respect to mechanisms, divergent enzymes most often evolve to process different substrates (reflected in a different 4th EC level). Only occasionally do they evolve to perform completely different mechanisms[14]. However, a new entry for a divergent enzyme would not be added in M-CSA if this only differs in substrate specificity, while catalytic residues and mechanism is conserved. The template library was derived to explore evolution of enzyme active sites and mechanisms.

Basic statistics of derived templates

Fig. 2a presents some statistics for the template library. In total, we extracted 3953 partial templates (to reflect 813 active sites), providing the principal dataset for all analyses. 3-residue templates are the most abundant (1474), with the number of larger templates decreasing with size (4-residue: 1042, 5-residue: 640, 6-residue: 417, 7-residue: 234, 8-residue: 146). Templates cover all seven EC classes (1: Oxidoreductases, 2: Transferases, 3: Hydrolases, 4: Lyases, 5: Isomerases, 6: Ligases, 7: Translocases), with hydrolases being clearly more numerous, and translocases (only recently added to E.C. nomenclature) occupying only a small fraction. This distribution, reflecting the enzyme functional space covered by UniProt and PDBe. Structural space is also diverse, covering 377 unique superfamilies, as defined at the H level of the CATH domain fold classification[15] and 259 unique topologies (T level). However, two topologies are overrepresented, the Rossmann Fold (3.40.50) and the TIM barrel (3.20.20), highlighting the dense network of evolutionary relationships in enzymes and the relatively limited but functionally diverse structural space[16].

Geometric analysis

To explore the landscape of enzyme active sites, we perfomed an all-by-all comparison of the coordinate templates, using RMSD as the comparison metric. The results are presented in Fig. 2b, where we explore the dependence of RMSD against the template size and the evolutionary relationship of the proteins from which the templates derive. We define a pair of templates as "divergent" if their parent proteins belong to the same CATH superfamily (H level), while in "convergent" templates, parent proteins are unrelated and belong to different CATH superfamilies. Out of 1572 pairs of matched templates, 1419 were convergent and 153 were divergent. It should be commented that those numbers reflect the construction of the M-CSA (see previous paragraph) and not the diversity in nature, where divergence dominates protein evolution (and the emergence of new functions) compared to convergence. Additionally, ~37% of the template dataset consists of 3-residue templates for which, due to the presence of fewer matching constraints, it is easier to find matches from unrelated proteins. In contrast, larger templates requiring more geometrical constraints, are frequently found in active sites of related proteins. Divergent and convergent templates behave differently when their geometry is compared. In both cases, there is a broad spread of RMSD values. RMSD is sensitive to fluctuations when small coordinate sets (like templates) are compared. It would be expected for RMSD to increase when the template size increases, and this is the case in convergent template pairs, which overall exhibit higher RMSD values than divergent pairs. RMSD is also more variable in 3,4 and 5-residue template pairs, with significantly lower mean values (<1Å) than in convergent pairs, indicating that these functional residues tend to retain a stable geometry during evolution. In 6 and 7-residue divergent templates, RMSD tends to be much lower and inversely proportional with template size, in contrast to convergent templates in which RMSD and size are proportional. This can be explained since the likelihood of matching two templates by chance, within the set pairwise distance cutoff (3.5Å), dramatically decreases with the increase of the template size. Furthermore, RMSD for 4 and 5-residue pairs is distributed bimodally; values are distributed in a region around the mean value (~0.5Å) and spread up to >2.0Å. The low RMSD region corresponds to nearly identical templates, while higher RMSD could indicate the presence of one or more flexible residues. It has been shown that residues of auxiliary function in catalysis, such as electrostatic stabilisers or ones with a steric role, tend to have more variable geometry compared to those involved directly in the formation/breakage of bonds)[17], therefore, their presence might influence template structural variability.

Templates are extensively used as tools to identify functional sites in protein structures and for functional annotation[12,18–20]. In Fig. 2c we benchmark the ability of our template library to identify catalytic residues and examine the effect of template size and template/query evolutionary relationship on RMSD. 6405 non-redundant chains from the PDB[21], for which catalytic residues are annotated in M-CSA, were scanned by our library of partial templates using the constraint based structural template matching program *Jess*[22]. Searching complete PDB structures means that we find template matches within the active sites, but also distributed around the protein, since some geometrical arrangements of three residues occur quite widely in structures but are not necessarily associated with enzyme mechanism. Such hits are called 'spurious' below.

Out of all matches, 4768 (3-residue: 2413, 4-residue: 1246, 5-residue: 563, 6-residue: 315, 7-residue: 155, 8-residue: 76) were divergent (template and target structure from the same superfamily) and 7223 (3-residue: 6477, 4-residue: 524, 5-residue: 133, 6-residue: 51, 7-residue: 30, 8-residue: 8) were convergent (different superfamilies). Unsurprisingly, matches where template parent protein and query protein are homologous (divergent) exhibit much lower RMSD values than convergent matches. This is useful when assessing the quality of template-template matches, indicating that convergent template geometries will be expected to be less similar than divergent ones. Also, template size is shown to be a definitive factor affecting template structural similarity. While in divergent matches, the shape of the RMSD distributions is retained when the overall number of hits increases with decrease in template size, in convergent examples, 3-residue matches are much more wide-spread, in a bimodal distribution, reflecting spurious hits. The global maximum of the histogram lies at ~1Å, therefore, 3-residue convergent matches, either between templates and queries or between two templates, will be expected to exhibit higher RMSD values, with a higher likelihood of being spurious compared to matches from larger templates.

However, the chance of identifying a hit by chance drops as the query size is smaller; in the case of module identification discussed below, the queries are templates themselves, therefore the chance of two templates being structurally similar by chance (and not being functionally relevant) is relatively low.

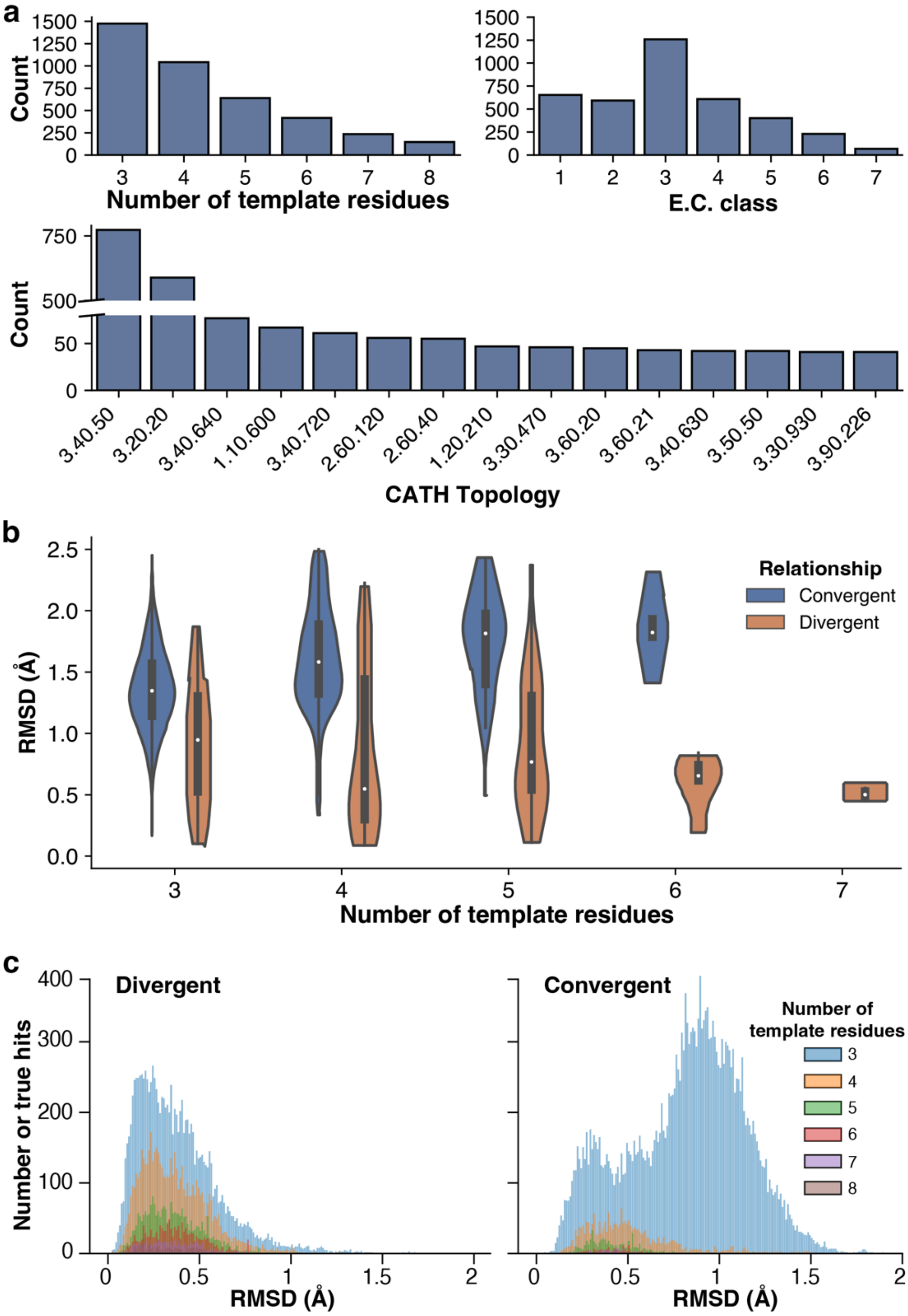


Fig. 2: Analyses of the template library (total templates: 3953). a: Basic statistics. The upper left bar plot shows the distribution of template sizes while the upper right plot shows the functional coverage of templates in terms of their EC class (1: Oxidoreductases, 2: Transferases, 3: Hydrolases, 4: Lyases, 5: Isomerases, 6: Ligases, 7: Translocases). The bottom plot shows the structural coverage of the domains from which the templates derive, in terms of the Topology level in the CATH hierarchy. b: Pairwise template superposition to quantify dissimilarity expressed as the RMSD over functional atoms. Violin plots illustrate the distribution of RMSD values for template pairs of different sizes (3,4,5,6 and 7 residues). Template pairs were split according to whether they derive from domains within the same CATH superfamily (divergent – coloured blue) or

different families (convergent – coloured orange). c: Benchmarking of catalytic template search. 6405 non-redundant PDB chains for which catalytic residues are explicitly annotated in M-CSA were scanned by the catalytic template library. Histograms show the distribution of RMSD (Å) on hits containing at least 3 catalytic residues. The left panel corresponds to cases where the query protein is homologous with the protein from which the template has derived (common ancestor, same or divergent function), while the right panel shows cases of hits where template parent protein and query protein are non-homologous (unrelated lineage, convergence in parts of the active site and potentially in function). Histograms are further resolved by the number of residues in the template, shown in different colours.

The modules of catalysis

Overview

We used the template cross comparison analysis to capture templates that recurred in more than one protein superfamily and to construct enzyme modules. We found the following matches, as illustrated in Fig. 3. 706 out of 3953 partial templates (3-residue: 579, 4-residue: 215, 5-residue: 69, 6-residue: 17, 7-residue: 4, 8-residue: 0) were grouped into 161 modules. Out of those, 87 modules were 3-residue, 44 were 4-residue, 22 were 5-residue, 6 were 6-residue and 2 were 7-residue. Fig. 3 represents the modules as a network of template-template relationships, where each node represents one or more templates from the same CATH homologous superfamily. The connections represent similarities with RMSD ≤ 2.5Å. The number of nodes in one module, represents the number of distinct superfamilies in which that module is found. As observed in the case of the Arg-His-Asp triad, the geometry of templates can vary within a module and multiple conformers could form structural clusters. Graph clustering allowed us to identify those clusters, using as a metric the connectivity (modularity) of each putative cluster. As clearly seen in Fig. 3, members of the same cluster tend to be interconnected (e.g. the two large clusters of the DE-DE-H module), but not necessarily fully connected (e.g. DE-H-NQ module).

To visualise and illustrate the landscape of the modules of catalysis. Fig. 4 shows a hierarchical pie chart with five levels of classification: 1) Module size (number of catalytic residues), 2) fuzzy residue composition (the primary “module” level), 3) structural cluster 4) CATH homologous superfamily and 5) EC number of hosting enzyme. Structural variability within structural clusters is expressed by the average RMSD of the templates to a cluster representative.

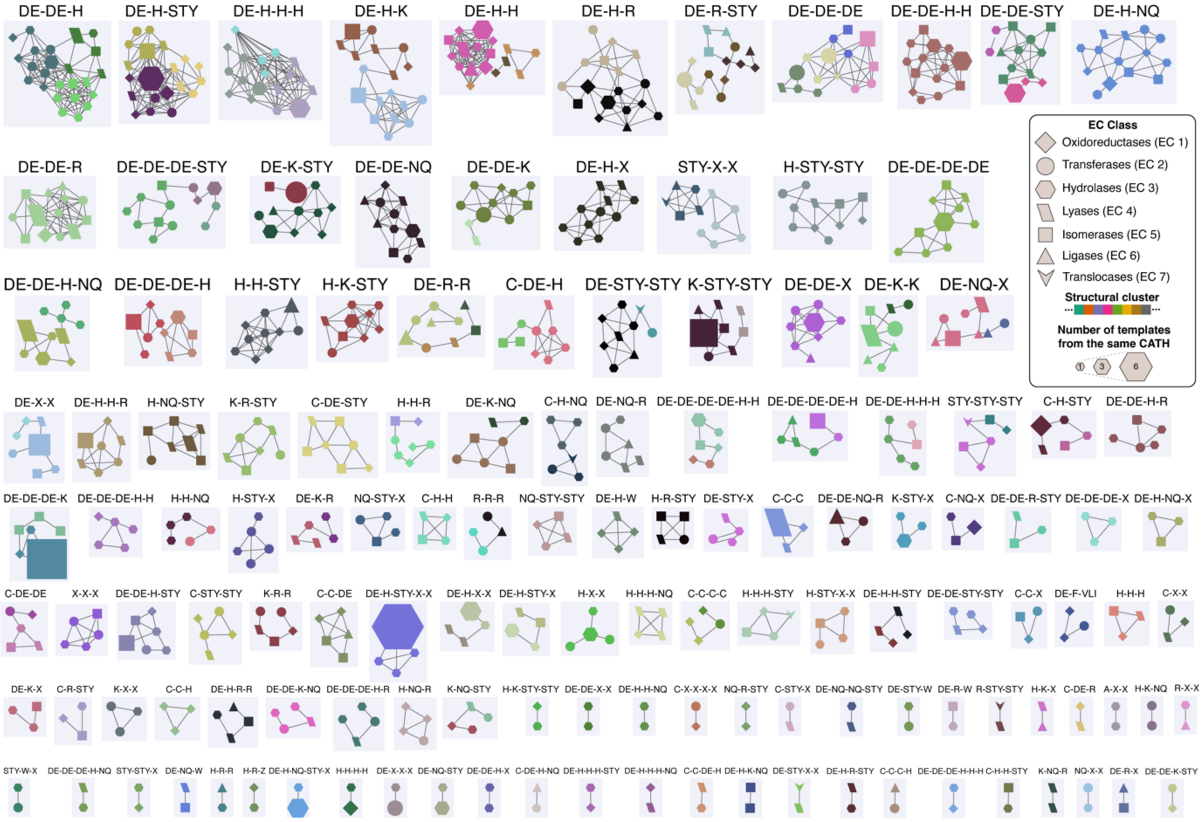


Fig. 3: Graph representation of the modules landscape. Each node represents a template or a group of templates from the same CATH domain (size proportional to the number of merged nodes). Each subgraph represents a "module" named after its residue composition, where similar templates (RMSD ≤ 2.5Å) are connected by an edge. Different structural clusters within a module are indicated by a different arbitrary colour. Different node shapes correspond to different EC classes. For better visualisation, only modules with two or more templates from a different CATH domain are shown. For better visualisation and to reduce clutter, a sample of 300 random templates within the modules set was collected to draw the figure. Graph construction and clustering (Leiden algorithm) were performed in Cytoscape[23].

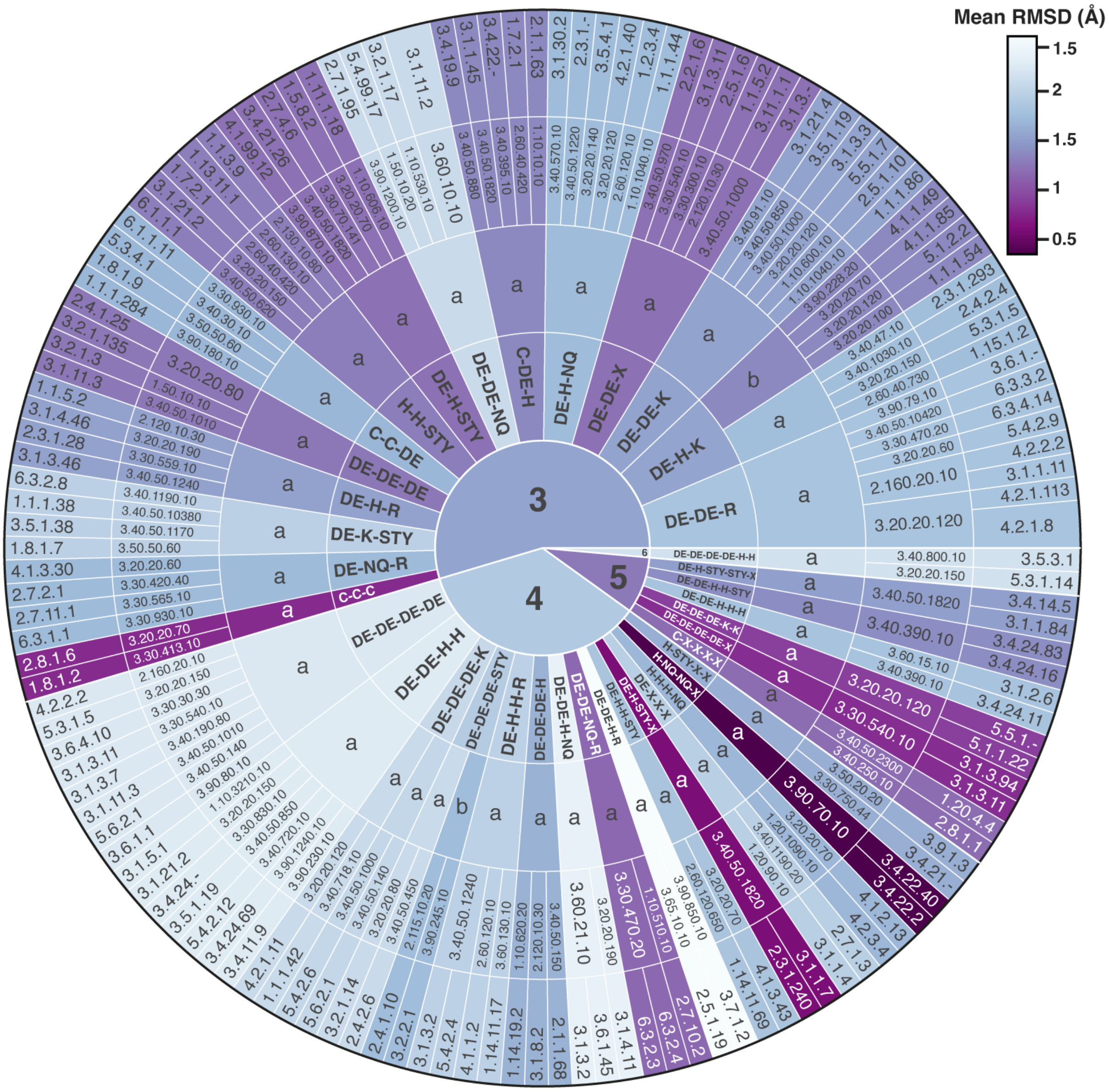


Fig. 4: Classification of the modules. Levels of classification are represented as concentric circles, from the centre outwards, as follows: 1) Number of residues 2) Residue composition 3) Structural cluster 4) CATH superfamily 5) EC code. Each classification group is coloured according to the mean RMSD of its corresponding conformational clusters, quantifying the structural variability within. Lower RMSD values (darker colours) indicate relatively invariable motifs in 3D, while higher RMSD values (lighter colours) indicate variable motifs.

Divergent and convergent evolution in modules

It is important to discuss the structural and functional characteristics of modules in the context of divergent and convergent evolution. We find that most modules include both divergent and convergent templates, with their ratio (# divergent/# convergent) increasing with the number of residues. This is visible in Fig. 4, where four out of seven 5-residue modules in the figure derive from divergent proteins (same CATH superfamily). Divergence and convergence are also reflected in the graph landscape in Fig. 3. Several examples (DE-H-STY-X-X, DE-DE-

DE-K, C-C-C, K-STY-STY, DE-DE-H-NQ) contain several divergent templates (larger nodes), as well as convergent ones. The question is whether the functional role of these modules is retained when enzymes diverge.

Table S2 is a collection of the divergent subset of templates in the module library (43 modules), grouped by CATH superfamily (52 groups). Most of them contain only a single superfamily group, where templates are likely to perform a similar role. For instance, the two templates from divergent enzymes in the C-NQ-X module (CATH 3.60.20.10), correspond to a triad of a catalytic nucleophile (Cys) and two oxyanion hole residues (Asn and Gly (backbone)). The two enzymes containing this module, Glutamate Synthase (EC: 1.4.7.1) and Amidophosphoribosyltransferase (EC: 2.4.2.14), catalyse different overall reactions. However, both retain part of the mechanism conserved (hydrolysis of glutamine to glutamate) using the same conserved triad, with the latter having an extra, coupled functionality facilitated by a separate group of catalytic residues[24,25]. This is likely a gain-of-function evolutionary scenario, where part of the active site remains conserved for maintaining a particular part of the mechanism, and an emerging new function utilises this site as a component.

Some modules in Table S2 contain multiple CATH superfamily groups. In those cases, residue function might be similar both within and across superfamilies, or similar within one superfamily but different across superfamilies. The first case is exemplified by the DE-DE-DE-H module, found in enzymes from two superfamilies (3.20.20.120 and 3.30.420.10). This is an acid residue triad binding Mg^{2+} and Mn^{2+} ions accompanied by a zwitterionic His acting as a proton shuttle[26–30]. The second scenario is represented by the DE-K-STY module. Here, five enzymes are grouped into two superfamilies, 3.40.50.300 and 3.40.640.10. In the first group, the module is a group of transition state electrostatic stabilisers[31,32], while in the second, it interacts with a PLP co-factor[33–35]. Many papers have been published discussing divergent evolution[36,37] and this dominates the emergence of new functions.

For the modules suggesting convergent evolution (131 or 83% of a total of 161 modules) we analysed whether structural convergence is accompanied by functional convergence or whether the change of EC class observed in proteins containing these modules is coincidental. This is the ultimate question addressed in this paper, with mostly convergent templates being under the spotlight.

Structural characterisation

We explored structural and functional diversity of the modules, by looking at them individually. Fig. 5a shows the number of unique CATH superfamilies (H level, left panel) and unique EC numbers (to the 4th EC level, right panel) associated with each of the first 100 most populated modules. Most modules are highly diverse both structurally and functionally, with the numbers of unique domains and ECs being correlated (i.e. structure-function correlation). However, several do not to follow this trend; for instance, the second most abundant module (DE-H-STY, otherwise known as the Ser-His-Asp catalytic triad of serine proteinases), is seen in ~30 different enzymes (21 unique EC subsubclasses), whose structural space covers ~20 structural folds. This underlines the high functional diversity of this module as a catalytic entity[9,38–40]. Functional diversity is also visible when looking at the extended module DE-H-STY-X-X which includes the two residues forming the oxyanion hole using their backbone atoms. Similarly, generic modules that bind metal ions, for instance DE-DE-DE-K and K-STY-STY, have also remained structurally unaltered while enzyme functions diverged. Those two perfectly exemplify the evolutionary constraints placed by co-factor binding on catalytic residue conformation. The former module is almost always seen to bind Mg^{2+} or Mn^{2+} ions; three acidic residues (Asp or Glu) in any combination are directly coordinated by the metal, while an amide group contributed by an adjacent Lys residue, acts as a stabiliser of negative charges during this protein/ligand interaction. The need to bind organic co-factors might also lead to convergence during evolution, like in the K-STY-STY module that is observed in crystal structures to interact with NAD and its analogues, making it one of the few examples in our dataset where a consistent residue geometry is coordinated by a flexible organic co-factor.

Fig. 5b and c demonstrate the relationship between template conformational variation within modules (average RMSD) and diversity of the proteins from which the constituent templates derive from, in terms of structural folds (CATH domain diversity) and function (EC number diversity) respectively. In both cases, when no structure/function diversity is present, template conformational variation tends to be low. This is evident among enzymes of identical function (zero EC diversity), where average RMSD is always below 1Å. In cases of enzymes belonging to the same CATH superfamily (zero fold diversity), the average RMSD is slightly more variable, while most modules stay below 1Å. This behaviour is explained by the fact that proteins with an evolutionary link, sharing a similar overall fold, have diverged into enzymes of discrete function, where some key structural elements of their active site have been conserved in 3D, either perfectly (RMSD < 1Å) or partially (1Å < RMSD < 2Å). When diversity increases, in both structure and function, a subtle linear trend with RMSD is observed (Pearson correlation coefficient: 0.37 and 0.27 respectively). This highlights that templates with a defined function are either well conserved or have converged to a similar conformation (the geometry of functional atoms needs to be preserved in order for the chemistry to happen), and their structural evolution is relatively independent from the evolution of the overall protein fold.

A caveat should be stated here: By definition, modules only contain templates sharing the same fuzzy residue composition. This captures functional changes for an arrangement of conserved or conservatively mutated residues. It does not capture changes caused by non-conservative mutations (e.g. mutation of a charged catalytic residue to a hydrophobic one). These non-conservative mutations of active site residues often lead to change[41] or even loss of function[42,43].

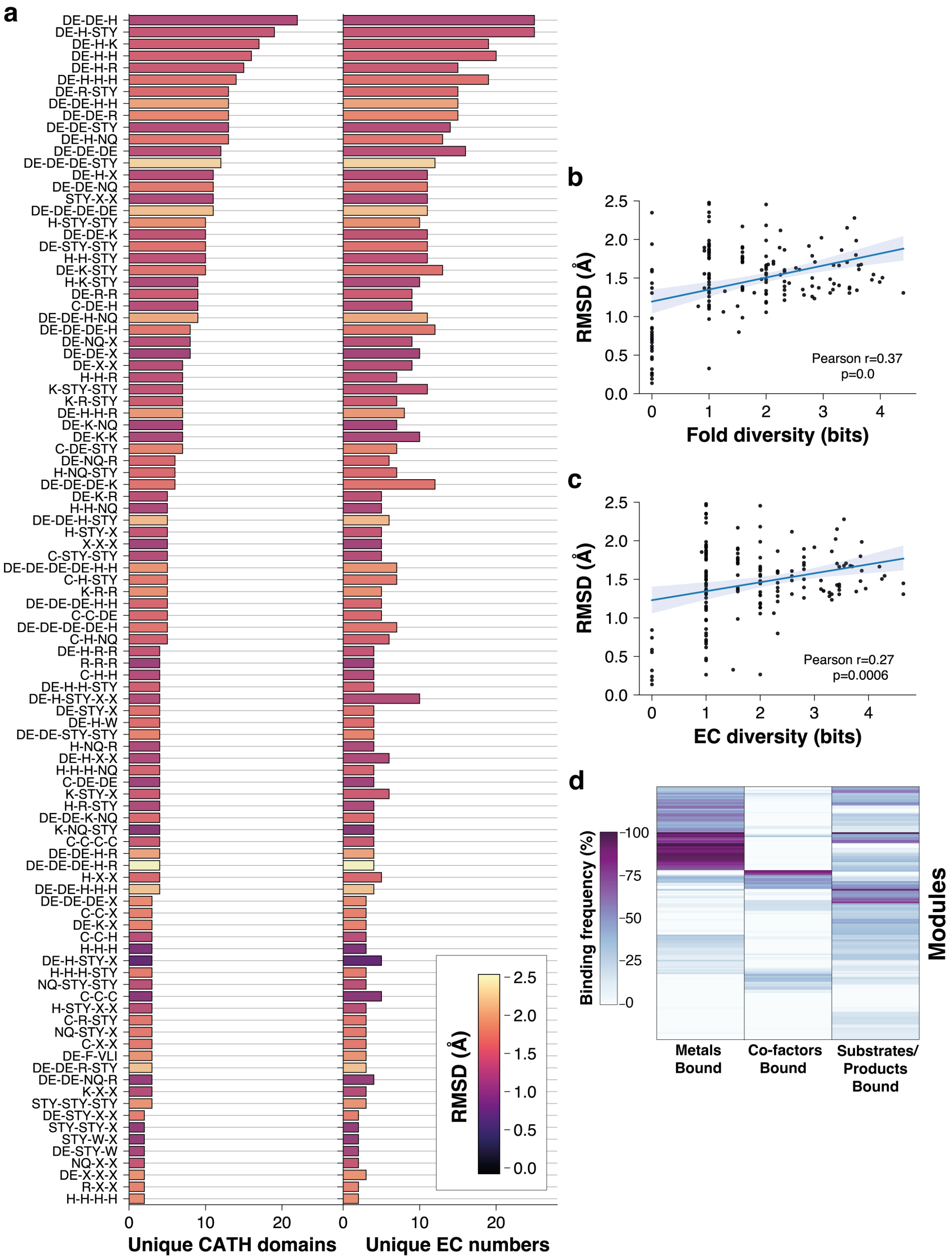


Fig. 5: Structural characterisation of catalytic modules. a: Structural and functional diversity of the proteins accommodating the templates within each module. Diversity is expressed as the unique number of CATH domains and EC numbers respectively. To reduce clutter, only modules of ≥3 templates are shown. Average RMSD of all templates against a reference template is shown in colour scale. b and c: Average RMSD against the reference template for each module, plotted against the structural diversity (b) and functional diversity (c), expressed as the Shannon entropy of CATH and EC numbers respectively. Each point corresponds to a template and linear regression is drawn as a blue line with the respective confidence interval shaded. d: Heatmap of the bound ligand binding frequency scores (x-axis) for each module (y-axis).

Functional characterisation

Templates and modules have been identified with the exclusive use of geometrical and knowledge based criteria, so that further scrutiny is required to establish any potential relationship between module geometry (structure) and mechanism of action (function). Here, three approaches to achieve this are discussed (details in Methods section). The first is based on investigating which type of ligands lie in proximity to each template. The second is the enrichment of each residue position with functional role annotations and their frequencies, taken from the M-CSA entry for that enzyme. The third is the identification of mechanistic steps in which template residues participate, to search for common instances within a module. Exemplary 3D models of convergent modules are shown as superimposed templates in Fig. 6, while comprehensive module descriptions for the whole library are presented in Table S3. This is populated with examples of enzymes and folds represented by a module, the structural variation (average RMSD) between members of the module, annotations of ligands bound in proximity and catalytic residue functional roles, as discussed below.

Modules can be characterised by bound ligand preferences

The first approach for functional characterisation explores the idea that a conserved/convergent arrangement of catalytic residues will be associated with a set of ligands with which it interacts (as taken from PDB structures for members of the module). For this, each module was assigned with a binding frequency score for each ligand category (substrate/product[44], co-factor[45], metal ion[46,47]) that approximates the propensity to bind a ligand falling in those categories. Those scores are presented as a heatmap in Fig. 5d. Almost all modules documented in Table S3 appear to interact with a ligand (given the distance cutoff we set). Of these, the majority involve metal binding, fewer bind cofactors and even fewer bind substrates.

At a glance, ~1/3 of the modules bind a metal with 25% frequency or more. On the other hand, only a few modules seem to interact with organic co-factors, with respective propensities not exceeding 50% in most cases. This is surprising, since “generic” compounds like organic co-factors, are found across a plethora of domains and functions and would be expected to interact with a limited variety of local 3D motifs. However, this is not the case, as shown in the examples of Fig. 6. Not only can a particular co-factor bind and interact with different modules, but also a single module can accommodate one or more co-factors and their analogues, in multiple different poses.

On the contrary, metal ions and metal clusters[48] are precicely define the geometry of binding residues as shown in multiple examples in Fig. 6. This reflects the organometallic interactions driven by the ion coordination sphere, restricting the degree of side chain flexibility. For instance, in several modules consisting of His and Asp/Glu residues (Fig. 6), the ion-interacting functional atoms are fixed, and superpose well even in unrelated enzymes. A general conclusion drawn here is that metal binding motifs are frequently convergent in 3D, while motifs binding organic co-factors are significantly more variable. This is because large co-factors are flexible, therefore can bind in various ways, and some of them might contribute different chemical groups in different catalytic mechanisms.

With respect to the “substrate/product” ligand category, module annotations are not as clear as those in the “metal” and “co-factor” categories. Usually, a cognate ligand analogue will lie close to the catalytic residues, overlapping with residues involved in metal and co-factor binding. This is due to the relatively loose criteria we have set to identify neighbouring ligands (distance cut-off and the requirement of at least one ligand atom to lie close to the module residues) and the fact that usually enzymes are co-crystallised with a cognate ligand analogue that often interacts with co-factors. This highlights the multi-functional nature of several modules, as presented in multiple cases in Table S3. Based on these observations, it can be deduced that “substrate/product” annotation is typically of low confidence (substrate/product analogues of high similarity to the cognate ones are rarely seen in crystal structures[44]). For this reason, complementary approaches are necessary to characterise the modules, as discussed in the following paragraphs.

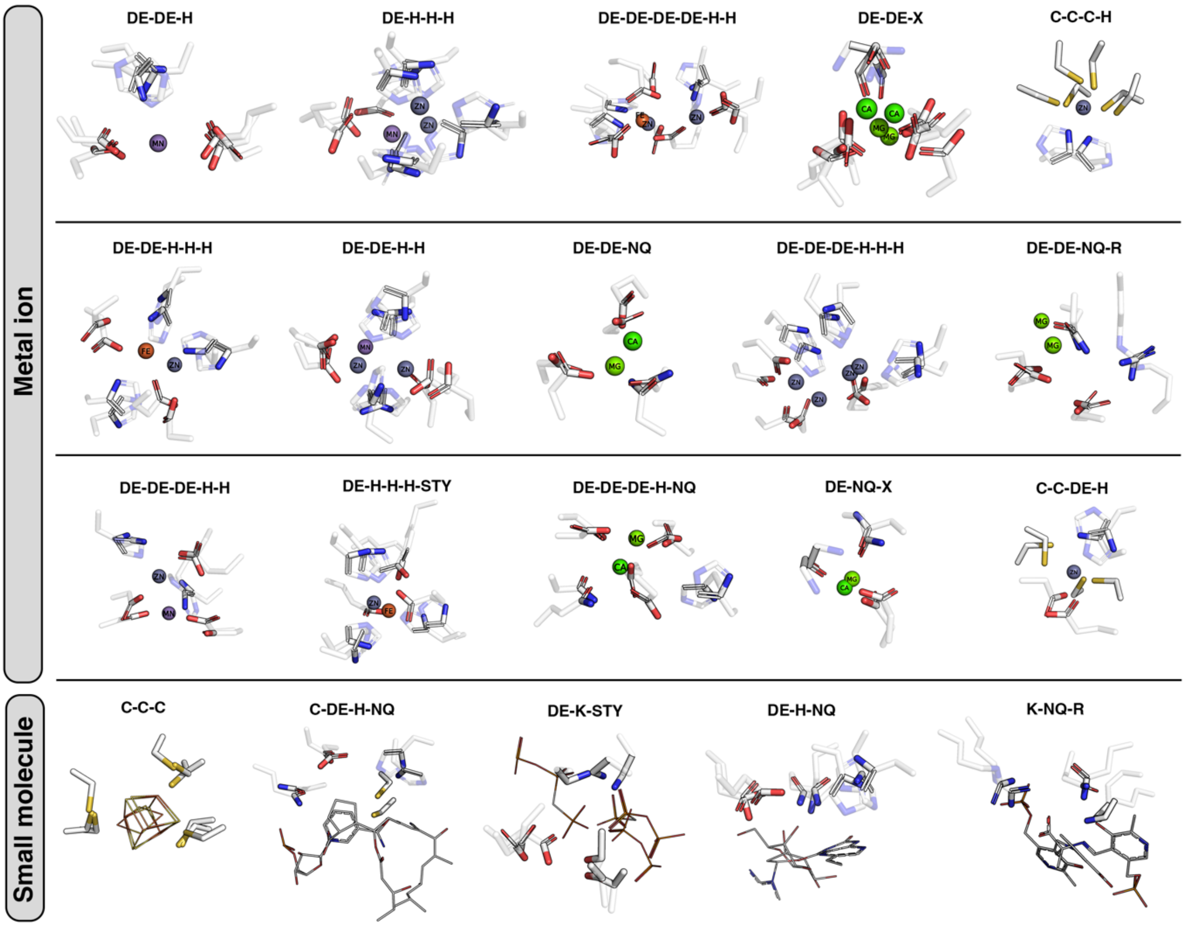


Fig. 6: A collection of example modules, shown as superimposed convergent templates. Catalytic residues are shown as opaque and transparent sticks, indicating functional and non-functional atoms. For residues contributing side-chain atoms, backbone is removed, whereas for backbone contributing residues, side chain is removed. Ligands are shown as spheres (metals) and lines (small molecules). Atoms are coloured according to their type (white: C, red: O, blue: N, yellow: S, orange: P). Bulky small molecules were truncated to reduce clutter. 3D models were prepared in PyMol[49].

Module enrichment with catalytic residue functional roles

The second functional characterisation approach was the enrichment of each aligned residue position in a module with mechanistic role annotations from M-CSA. In M-CSA each residue is assigned a catalytic function in the mechanism – Broadly these can be classified as one of three roles - catalytic (directly involved in catalytic rearrangements), ligand binding or stabiliser/spectator role. In Fig. 7, we present the functional preferences of modules consisting of three or more convergent templates. For clarity, we decided to include only one instance of each CATH domain in each module, to reduce bias from over-represented folds (e.g. DE-DE-DE-K module, Fig. 3). Pie charts in Fig. 7 show the relative occupancies of role categories (for role classification see Table S1) on each residue position, and the occupancies for the entire module (leftmost pie charts).

Through visual inspection of functional occupancies, we defined four coarse categories that classify modules according to function: 1) Mainly metal binding – where at least 3 catalytic residues are annotated as “metal ligand”; 2) Mixed – where a mixture of functions including “metal ligand” are observed; 3) Mixed - Catalytic/Spectator – where a combination of non-radical covalent catalysis and interaction/spectator roles dominates; 4) Mainly Spectator – where Interaction/Spectator occupy ≥85% of roles.

Metal ligand binding modules are straightforward to characterise. This role is rarely ambiguous and most modules in the library are strongly correlated with co-crystallised metal ions in vicinity (see previous paragraph). However, conflation with roles of other categories, as well as the presence of multiple 3D conformers of the same residue composition, often introduce ambiguity. For instance, the DE-DE-H module is so generic that “proton shuttle”, “metal ligand” and “interaction/spectator” roles are observed in almost equal occupancies. In fact, this module serves multiple functions, one of which is Mn^{2+} and Mg^{2+} binding. The next two categories are of mixed function. Characterisation is not as trivial in those cases, and each module needs to be examined individually to elucidate its precise functions. However, a few broad observations can be made. For instance, His-Asp/Glu containing catalytic triads are captured and are, not unexpectedly, highly recurring. Modules corresponding to other triads (e.g. DE-H-R or DE-H-NQ) are also highly represented in this category. This reinforces the hypothesis of functional convergent evolution. Indeed, the acid-base pair (DE-H) is preserved and plays an activating role, while function is determined by the third residue, usually a catalytic nucleophile. We also defined a special category for modules of high interaction/spectator role content. Four modules are in this category (K-R-STY, X-X-X, R-R-R, DE-F-VLI) and all have an auxiliary role, such as charge stabilisation (X-X-X gamma-turn hole), stabilisation of radicals (DE-F-VLI) and interaction with co-factors (R-R-R and K-R-STY). A special comment should be made on the presence of oxyanion holes in the library. Usually those are formed by the backbone of two low-conserved residues. Since our method does not capture two-residue modules, all instances of oxyanion holes are found as two aspecific residue pairs (X-X) and a residue from the adjacent catalytic module. In almost all example, the X-X pair has interaction/spectator roles in high occupancies.

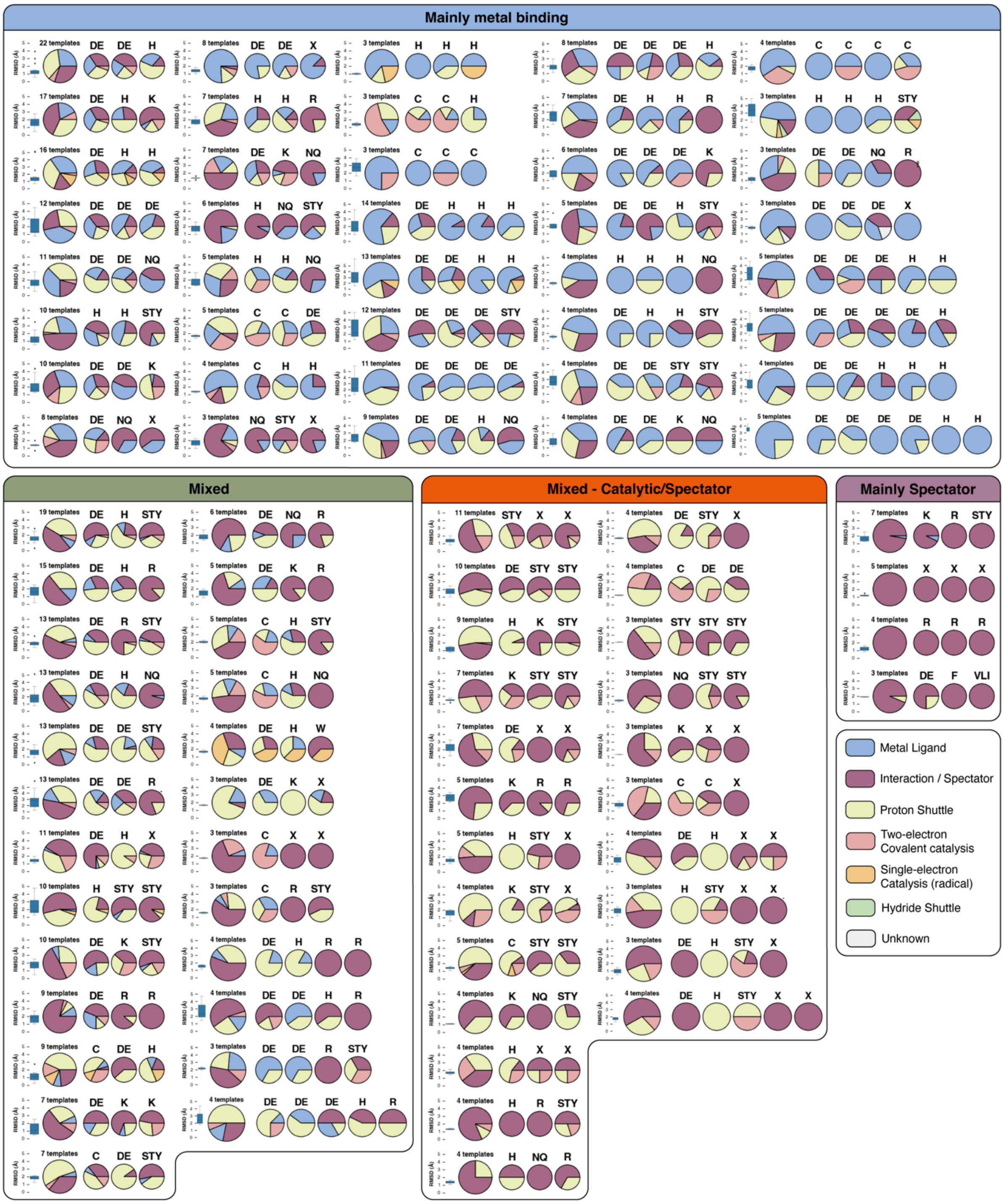


Fig. 7: Functional characterisation of catalytic modules. For each module, the occupancies of functional role categories at each residue position are shown as pie charts. Residues at each position are named by the amino-acid one letter code, with multiple letters indicating different, chemically equivalent residues (the letter X is used for residues of unspecific type contributing backbone atoms). The leftmost pie chart shows the sum of occupancies over all residues. Conformational variation among constituent templates is shown as a boxplot of RMSD values over a representative template. To reduce bias from over-represented structural folds, one template from each unique CATH domain is used to collect annotations for this analysis. Modules with <3 residues are not shown in this figure. Categorisation was performed by manual inspection, considering the relative role frequencies.

Modules as mechanistic components

The third functional characterisation approach based on the concept of the "rules of enzyme catalysis"[50]. Catalytic mechanisms can be broken down in a series of reactions involving catalytic residue functional atoms and substrate/co-factor atoms: the reaction centres of the mechanism. We applied a semi-automated method to identify recurring mechanistic steps in modules, as illustrated in the examples below.

The first example is the DE-DE-DE module. This functionally heterogenous module facilitates metal binding (see Table S3) as well as various catalytic and supporting functions (Fig. 7). We found that this triangular arrangement of acidic residues is a common motif in saccharide-cleaving enzymes. In Fig. 8a, it is shown that this module with this particular functionality is found in at least three functionally convergent and structurally diverse enzymes: Glucan 1,4-alpha-maltohydrolase (EC 3.2.1.133), Glucansucrase (E.C. 2.4.1.125) and Xyloglucan endotransglycosylase (EC 2.4.1.207). The latter two share three very similar rules, directly involving two Asp/Glu residues in the reaction centres, while the third acts as an electrostatic stabiliser. Co-crystallised substrate analogues (glucan disaccharides) bind close to the two catalytic Asp/Glu residues, though at slightly different poses, as seen in 3D in Fig. 8a. In Fig. 8b, the first step of the catalytic mechanism, taken from M-CSA entry 1002 (Glucantransferases), which is shared through the respective rule (rule ID: 799170) is illustrated in 3D (left panel). In this reference mechanism, elucidated by QM/MM simulations[51], the glucosyl group anomeric carbon of the sucrose incurs a nucleophilic attack from Asp477 (covalent catalysis role). Subsequently, the oxygen of the glycosidic bond is protonated by Glu515, causing cleavage, and leaving of the fructosyl group.

The second example (Fig. 9), is well studied. The Cys-His-Asp catalytic triad is one of the most profound examples of convergent evolution, found in several enzymes of the hydrolase (E.C. 3.-.-.-) and transferase (E.C. 2.-.-.-) classes. In Fig. 9a, we present three enzymes of those classes: Arylamine Acetyltransferase (E.C. 2.3.1.118), Ulp1 Peptidase (E.C. 3.4.22.68) and Glutamyl Hydrolase (E.C. 3.4.19.9). All have unrelated folds, and contain the Cys-His-Asp triad (C-DE-H) in their active sites. These three templates are mapped to several almost identical mechanistic steps, with the templates from Ulp1 Peptidase and Glutamyl Hydrolase sharing four consecutive and identical rules. As an example in Fig. 9b, we show the third step of the reference mechanism (Ulp1 Peptidase, M-CSA entry 820). The respective rule (rule ID: 828513) represents three electron transfers in a series, starting from the δ nitrogen atom of the His imidazole ring towards the catalytic Cys which is covalently attached to the peptide bond, and through a water molecule. The Asp/Glu serves consistently as pKa regulator to stabilise the His activator. A comment here regards the template from Glutamyl Hydrolase. As seen in the superimposed models in Fig. 9a, the side-chain conformation of the catalytic Cys is considerably different from the other two examples. However, it is important to note that the position of the endpoint functional sulfur atom is similar across all three analogues, regardless of the orientation of the side chain. This is a good example of how evolution might “invent” the same functional configuration multiple times: what is required, is the catalytically crucial atoms to be in a precise position, regardless of the geometry of the hosting residues and the overall structural context. This can also be reflected in the RMSD of fitted templates, exhibiting increased values in cases like this, although the function is the same.

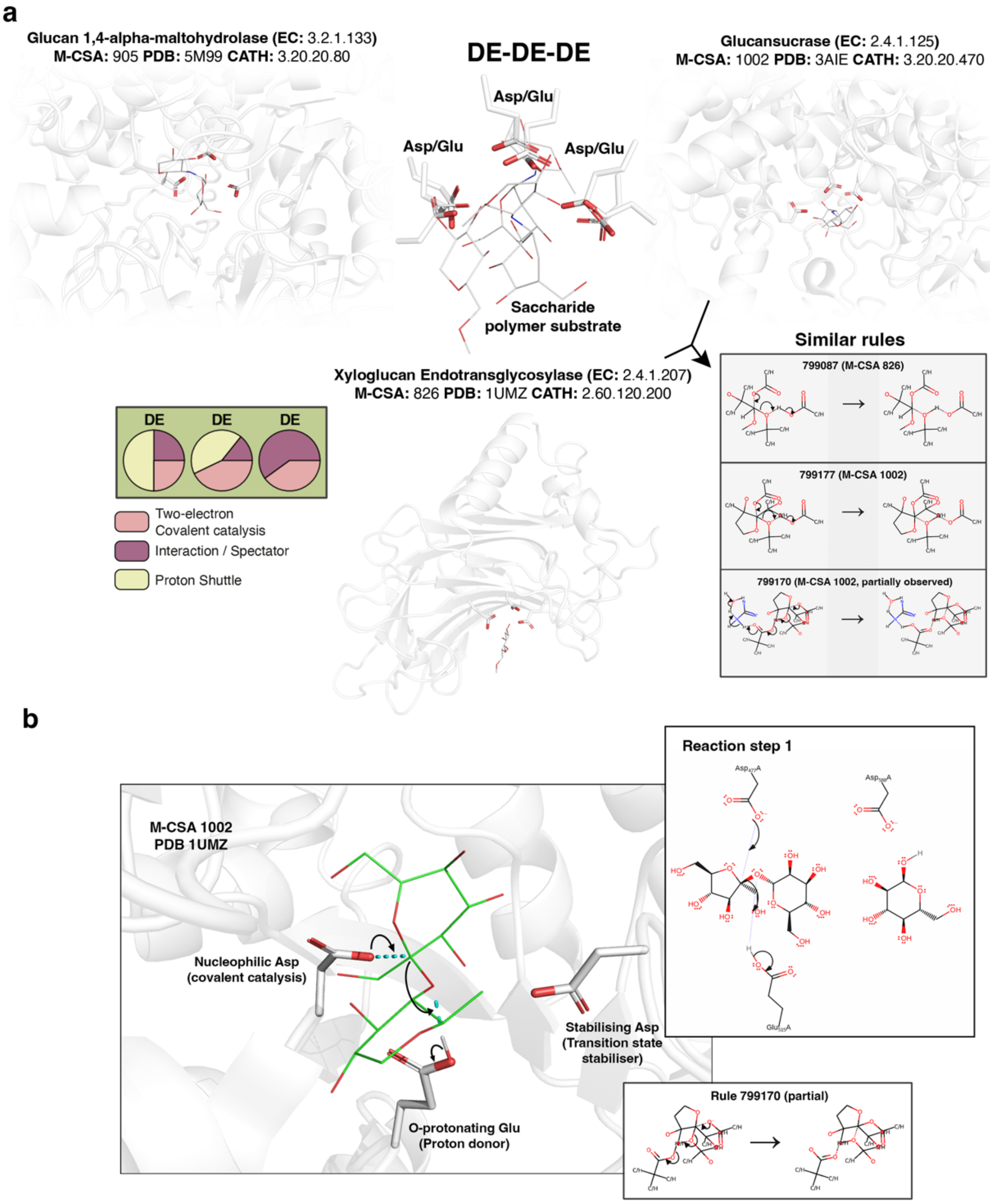

a
Glucan 1,4-alpha-maltohydrolase (EC: 3.2.1.133)
M-CSA: 905 PDB: 5M99 CATH: 3.20.20.80
DE-DE-DE
Glucansucrase (EC: 2.4.1.125)
M-CSA: 1002 PDB: 3AIE CATH: 3.20.20.470
Asp/Glu
Asp/Glu
Asp/Glu
Saccharide polymer substrate
Xyloglucan Endotransglycosylase (EC: 2.4.1.207)
M-CSA: 826 PDB: 1UMZ CATH: 2.60.120.200
Similar rules
799087 (M-CSA 826)
799177 (M-CSA 1002)
799170 (M-CSA 1002, partially observed)
DE
DE
DE
Two-electron Covalent catalysis
Interaction / Spectator
Proton Shuttle
b
M-CSA 1002
PDB 1UMZ
Nucleophilic Asp (covalent catalysis)
Stabilising Asp (Transition state stabiliser)
O-protonating Glu (Proton donor)
Reaction step 1
Rule 799170 (partial)

Fig. 8: *Mechanistic characterisation of the DE-DE-DE module. a: Superposition of three templates of the module over their functional atoms (center). Catalytic residue side chains, and bound ligands are shown as thick and light sticks respectively. Functional atoms are distinguished by increased opacity. Templates along with their structural context are shown individually around the central panel (context is shown in semi-transparent cartoon representation). Functional roles of each residue position are shown in the form of a pie chart in the lower left panel. Shared rules of highly chemical similarity are shown in the lower right panel, in the form of chemical equations. Here, since no identical rules are shared, the name of the M-CSA entry is shown next to each rule ID. b: Example of a catalytic step (initial step of saccharide bond cleavage), shared between two templates of the module. The left panel and upper right panels show the step in 3D and 2D representation, where curly arrows indicate electron transfers towards covalent bonds (dashed sticks/lines). For the 3D representation, the bound ligand has been slightly modified by substituting some atoms in order for its chemical structure to match the cognate ligand of the reaction. The lower right panel shows the respective rule of the step in the form of a chemical equation. Atoms throughout the figure are coloured according to their type (white: C, red: O, blue: N, yellow: S, orange: P). All 3D models were prepared in PyMol*[49].

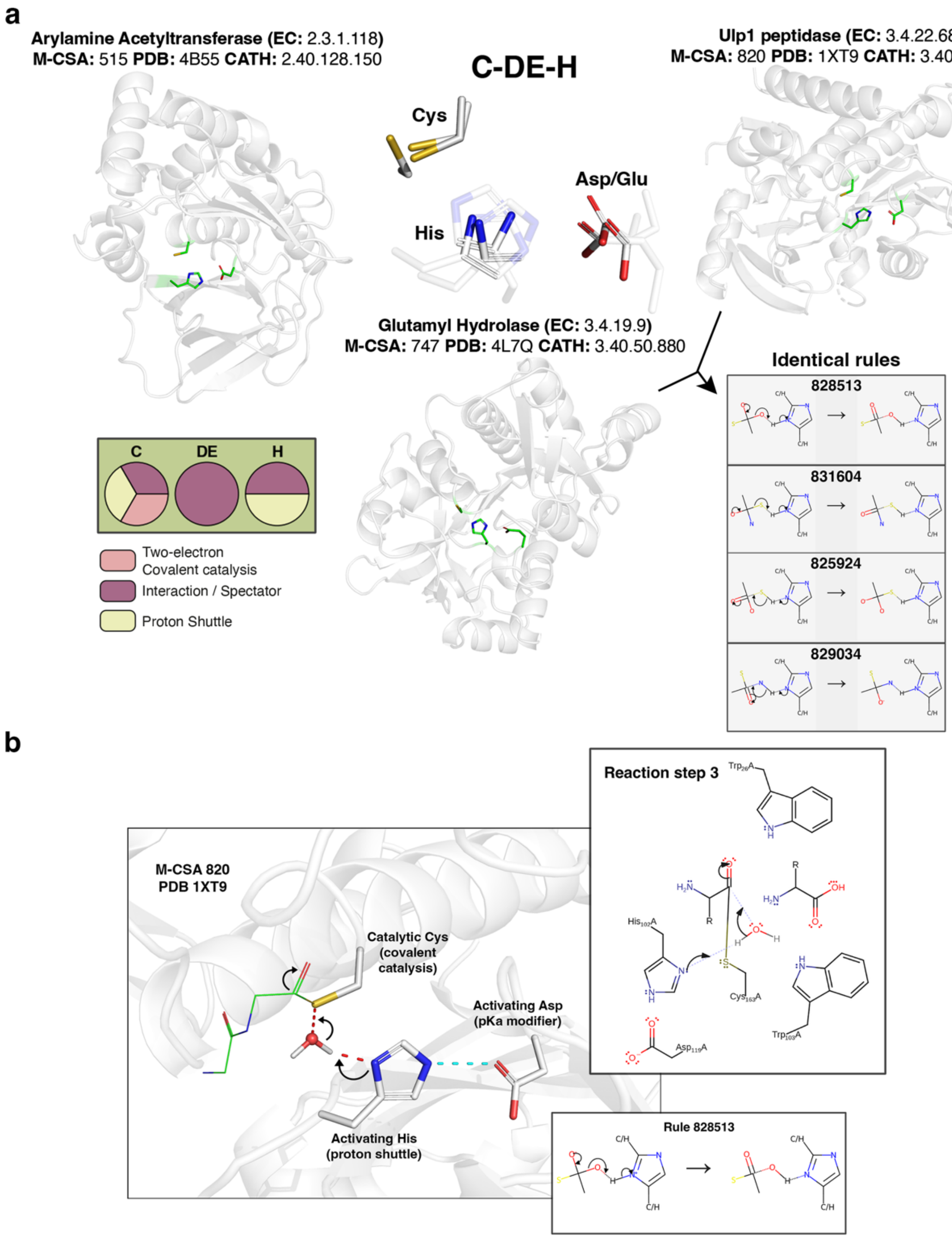

a
Arylamine Acetyltransferase (EC: 2.3.1.118)
M-CSA: 515 PDB: 4B55 CATH: 2.40.128.150
C-DE-H
Ulp1 peptidase (EC: 3.4.22.68)
M-CSA: 820 PDB: 1XT9 CATH: 3.40.395.10
Cys
Asp/Glu
His
Glutamyl Hydrolase (EC: 3.4.19.9)
M-CSA: 747 PDB: 4L7Q CATH: 3.40.50.880
Identical rules
828513
831604
825924
829034
C
DE
H
Two-electron Covalent catalysis
Interaction / Spectator
Proton Shuttle
b
M-CSA 820
PDB 1XT9
Catalytic Cys (covalent catalysis)
Activating Asp (pKa modifier)
Activating His (proton shuttle)
Reaction step 3
Rule 828513

Fig. 9: Mechanistic characterisation of the C-DE-H module. a: Superposition of three templates of the module over their functional atoms (center). Catalytic residue side chains are shown as sticks. Functional atoms are distinguished by increased opacity. Templates along with their structural context are shown individually around the central panel (context is shown in semi-transparent cartoon representation). Functional roles of each residue position are shown in the form of a pie chart in the lower left panel. Four identical and consecutive shared rules of are shown in the lower right panel, in the form of chemical equations. b: Example of a catalytic step (third step, involving electron transfers from the His to the covalently attached substrate to the catalytic Cys), shared between two templates of the module. The left panel and upper right panels show the step in 3D and 2D representation, where curly arrows indicate electron transfers. For the 3D representation, the bound ligand has been slightly modified by substituting some atoms in order for its chemical structure to match the cognate ligand of the reaction. The lower right panel shows the respective rule of the step in the form of a chemical equation. Atoms throughout the figure are coloured according to their type (white: C, red: O, blue: N, yellow: S, orange: P). All 3D models were prepared in PyMol[49].

Discussion

Collecting and archiving information on the elements at the heart of biological catalysis has allowed the definition and identification of currently known "3D modules of enzyme catalysis". By definition, a catalytic module is found in several enzymes, which are diverse in function and/or structure. Through a novel methodology involving 3D templates, we identified and catalogued 161 modules of varying abundance, and conformational/functional diversity.

The ultimate challenge of this analysis, was functional characterisation. A fully characterised and well-defined module would require each residue position to correspond to a precise functional role and/or a mechanistic step. Ideally, since catalytic modules participate in a series of reactions during the course of the catalytic transformation (steps), similarities in the detailed mode of action of each enzyme needs to be investigated .We show that a large number of catalytic modules facilitate binding of metal ions that are naturally not involved directly in electron/proton exchanges, making their characterisation relatively straightforward (annotation by ligand type or by role). Characterisation in mixed-role and non-metal binding modules however, as discussed , is more complex. Residue functional annotation, primarily for stabilising residues relies in manual curation and the available information of the literature, thus inconsistencies might be present. As can be summarized from Fig. 7, the striking majority of modules (apart from metal binders) are of mixed function, complicating automated characterisation. However, this also reflects the complexity of biological catalysis, the multitude of residue roles and also conformational variability. Another issue regards the sensitivity of template-template matching: Side chain conformations are rather variable and even in divergent enzymes[17] they often change their orientations and would not be picked up by template searching.

Looking at bound ligands in template neighbourhood provided good insight on module functionality. We characterised each module according to its residue-ligand interaction propensity and showed that ~1/3 of modules in the library facilitate metal binding with high propensity and in several cases, ligands of more than one type can interact with a module. Several examples of convergent modules in Fig. 6 show that coordination by metals is a very important factor applying evolutionary pressure to the conformation of binding residues, resulting in finding instances of such modules in multiple, unlinked by evolution, enzymes.

Functional atoms in these modules differ only slightly in geometry, something reflected in low RMSD values, despite the fact that residues contributing them might lie in completely different positions in the sequence and may not even share the same sequence order. Conversely, we found surprising that convergent patterns facilitating interaction with organic co-factors were not identified in this analysis. The explanation here is that large, flexible co-factors assisting multiple enzymatic functions (especially in pleiotropic enzymes[52]), might bind in several different ways and sometimes contribute different atoms according to what the mechanistic procedure requires. From our observations, two types of structural variation are distinguished: a flexible co-factor adopts several poses and can interact with different residue arrangements (for instance, several modules are found to interact with NAD, according to Table S3), and a residue arrangement can bind more than one co-factors. Finally, substrate binding is observed to be the least straightforward annotation, since it is highly likely that a substrate, especially a bulky one, will be located close to all catalytic residues and can co-exist with metals and co-factors.

Overall, our analysis shows that structurally similar modules can be involved in more than one enzymatic function, often being radically different in function themselves, and may have emerged more than once during evolution. A good example is the DE-DE-DE module, which is found to bind metals as well as being involved directly in catalysis, even though the residue geometry is relatively similar in both functions. Therefore, it is crucial to assess all three approaches of characterisation in order to assign the function of a module, taking also into account the structural diversity within. A limitation of this work is that this library is populated only by data stored in M-CSA. This automatically biases the analyses towards well studied enzymes that include detailed mechanistic information. However, as shown in Fig. 2a, the database covers a large functional landscape in terms of EC classes, therefore the amount and diversity of active site annotations in the current version provides a good start in examining the abundance of modules. There are two ways of expanding this coverage: 1) Use catalytic residue annotations from other resources like UniProt[53]. This though, comes with a compromise in reliability, since the advantage of M-CSA is the human-hand annotation of functionally important residues, compared to the automated ways used in general purpose resources. 2) Expansion of structural coverage and identification of further convergent evolution examples by scanning experimental and/or predicted[54–56] protein structures with the templates forming the modules. We are currently developing a protocol to perform exactly this, with the biggest challenge lying in distinguishing true, functional template matches from spurious ones, especially when functional information on the target proteins is limited (e.g. in uncharacterized predicted models from trEMBL[53]).

Beyond the academic interest that led us to investigate the modularity of active sites, these modules can be used as tools in computational enzymology. A library of functionally characterised, recurring small motifs can be used to identify functional sites in experimental or predicted protein structures. For each module, one or more templates can serve as probes to identify instances of the module in a protein structure[12,57–59], using template matching software like *Jess*[22]. Template matches can then benefit from the annotations provided in this study (see Table S3). This can be particularly useful, especially when investigating evolutionary relationships between enzymes and events like residue substitution and conservation of particular segments of the active site. Another application is the development of enzymes of novel function. We showed that active sites can be partially conserved in 3D, especially within residue groups coordinated by a ligand or serving a very precise catalytic function. This knowledge, in the form of the module library, can be used when developing a novel enzyme, incorporating *de novo* and directed evolution techniques[60,61]. A relevant concept, the concept of "theozymes"[62], can also use modules as templates for theoretical moulding of the active site, to perform a novel catalytic mechanism.

Although all data used here are available upon request, we are planning to release the template library in a detailed and entry-based structure within M-CSA in the close future. The immediate plan is to map residue-level mechanistic annotation in all modules. This will make modules an even more robust tool, to provide mechanistic insight in functional annotation and investigation of evolutionary relationships.

Methods

Template library methods

Catalytic site library

A library of enzyme catalytic site structures was generated by integrating data primarily from the Mechanism and Catalytic Site Atlas (M-CSA)[1] and the Protein Data Bank (PDBe)[63]. A detailed description of the method for building a catalytic site library is provided in the methods section of our recent work on catalytic site 3D variation[17]. In brief, curated catalytic residue annotations for 1003 enzyme families were collected from M-CSA, 734 of which included a detailed description of the catalytic mechanism in a computer readable format[50]. These annotations were used to extract atomic coordinates of catalytic residues from experimentally resolved protein structures (biological assemblies) and a clustering algorithm was applied to reconstruct all catalytic sites. The use of a clustering algorithm is important in this step: catalytic residues in identified homologues are annotated at the sequence level, and in cases of active sites consisting of residues from multiple identical chains, all residues sharing the same identifier will be annotated on the structure. Therefore, geometrical heuristics needed to be applied to identify the correct combination of catalytic residues forming an active site. Further information was added to the catalytic site library, such as co-factor and catalytic reaction annotations, cognate/bound ligand similarity scores[44] and mapping to protein sequences from SwissProt[53]. Lastly, active sites extracted from an M-CSA family, were overlayed onto their respective reference active site. Each ensemble of superimposed active sites effectively represents the different conformers captured in experimentally resolved structures.

Extraction of "complete" templates

In our recent review[12], we showed that each alternative conformational state of an active site can be represented by a consensus template –an abstract active site signature ("complete" template) containing only a subset of three functional atoms per catalytic residue. The selection of the atoms depends on the chemical nature of the residue, as well as on its function location (backbone or side chain atoms) within the context of the catalytic reaction, as annotated in M-CSA. Since templates are primarily used to probe protein structures for similar atom arrangements through a template matching algorithm[19,20,22,64,65], it is important to account for matching fuzziness in the residue and atom level. For instance, chemically equivalent amino acids should be able to be matched interchangeably (Asp ↔ Glu, Asn ↔ Gln, Ser ↔ Thr ↔ Tyr, Val ↔ Leu ↔ Ile), as well as carbonyl O atoms in Asp/Glu residues or $N_{\delta1}$ and $C_{\delta2}$ atoms in His residues, to account for side chain flipping and ambiguity in crystal structures[66].

Fig. S2 schematically describes the process for deriving "complete" templates for 813 M-CSA enzyme families. The steps are the following: Active sites from M-CSA families with ≥3 annotated catalytic residues were first constructed. Sites within a family, were then superimposed over their functional atoms in an all-vs-all fashion. For the purposes of this study, only active sites having all catalytic residues conserved (compared to the curated reference catalytic residues) or only containing chemically equivalent mutations were included. If aligned residues are known to function via their main chain or known to be post-translationally modified, any mutation is allowed. Rigid body superposition was RMSD-based and performed in an iterative manner (by updating the average structure at each cycle) using a modified, gaussian-weighted *Kabsch* algorithm[67,68] as introduced by Damm et al.[68]. Three functional atoms per catalytic residue are used for superposition with the exact definitions shown in Fig. S1. The next step was the calculation of an average active site structure and the selection of a representative template active site having the closest similarity to the average after superposition. After excluding those structures for which a template could not be generated due to insufficient electron density (some residues might be missing from the structure), we ended up with 813 "complete" templates. Those are formatted using a modified PDB nomenclature[64] and are compatible with the template matching software *Jess*[22,64].

Fragmentation of complete templates into "partial" templates

Complete templates were further fragmented into smaller, "partial" templates of 3,4,5,6,7 and 8 residues, lying in close proximity. The idea behind this is that residue groups contributing to a particular function, will be typically expected to form tight clusters. Thus, for a complete template consisting of S residues, we need to identify those partial templates which optimally cluster S residues into k partial templates. The fragmentation algorithm is purely geometrical and includes the following steps (Fig. 1): 1) Calculation of the centre of mass of the three functional atoms of each catalytic residue, 2) calculation of the number of partial templates (residue clusters) $k \approx S/s$, where S is the size of the complete template (number of residues) and *s* is the desired size of the partial template. It should be noted that potential partial templates of all possible sizes were generated for each complete template. 3) Application of k-means algorithm in 3D cartesian space to define *k* clusters of catalytic residues' centres of mass and 4) calculation of the centre of each cluster. 5) Finally, the first *s* catalytic residues closest to each centre are selected to define the partial template. Applying this method to all 813 complete templates led to a total of 3953 partial templates, with sizes spanning from 3 to 8 residues. Statistics on the template library (sizes, E.C. and fold distributions) are shown in Fig. 1c and are discussed further in the Results section.

Module library methods

Template-template search

Templates of the same size were cross-compared to capture groups of similar templates. This template-template search was performed in an all-vs-all fashion using the software *Jess*[22]. The standard functionality of Jess involves a template and a query structure; in our case the query structures are replaced by templates themselves, permitting the identification of template pairs that share structural similarity up to a defined atom displacement (functional atoms' RMSD). Template matching is fuzzy both at the residue and atom level as mentioned in the previous paragraph, allowing matching of chemically similar residues and atoms (Fig S1). Here again, for residues known to function via their main chain or those which are post-translationally modified, three backbone atoms of any residue can be matched. *Jess* requires two user-defined parameters to perform the search: a maximum template-query pairwise atom distance (set to 2.5Å) and a maximum RMSD after optimal superposition of the template to the query (set to 2.0Å). These cutoffs were set empirically, after performing multiple template-template search runs to determine a good balance between true positive and spurious matches. Using these parameters, *Jess* will try to identify a set of atoms satisfying the constraints coded within the template (atom and residue types) allowing a maximum variation equal to the atom distance cutoff, and then perform rigid body superposition, outputting matches with RMSD lower or equal than the RMSD cutoff. The final output of these template-template matching runs consisted of 15207 template pairs (*p*-*q* pairs). These hits were reduced to 1481 after removing redundant pairs (*p*-*q*/*q*-*p* permutations), and instances where at least one template in a pair is a subset of a larger one existing in the set. Furthermore, hits where both *p* and *q* derive from the same PDB structure were removed.

Graph representation and clustering

The 1481 template relationships constitute a network that can be drawn in the form of a graph (Fig. 3), consisting of a set of smaller subgraphs, within which all templates share the same residue composition (accounting for residue type fuzziness as explained previously). Subgraphs of the same residue composition define a set of "modules". Structural clusters within each module were identified by graph clustering, using the Leiden algorithm and RMSD as similarity metric. Finally, a total number of 708 templates were assigned to 123 modules and 213 structural clusters.

Alignment of residues within modules and clusters

To facilitate precise residue level annotation of the templates within each module (e.g. to derive functional role statistics in each residue position), it was important for all residues to be aligned consistently. The approach is as follows: a reference template is found within each structural cluster of a module; in the corresponding subgraph, this is the node with the highest centrality (defined as the number of edges connected to a node). The residue order of the reference template is used to find the correct residue order of as many templates within the cluster as possible, from the raw output of the template-template search (before cleaning-up). Often, subgraphs are not fully connected (not all templates match all templates under the defined matching parameters), therefore for templates for which an alignment could not be found in the similarities set, this was found by freshly performing pairwise structural alignment to the reference, through applying the same gaussian-weighted superposition algorithm described further above in active site superposition. Last, to achieve residue alignment within the module, aligned residues within the clusters were reordered by finding the correct residue correspondence of each cluster reference to the reference of the largest cluster. A point of care here is the following: since structural similarity of functional atoms of chemically similar residues is the only criterion to find the ideal aligned positions, ambiguity is often unavoidable, especially in templates of convergent geometry containing more than one instance of a residue type (or equivalent residues such as D/E, S/T/Y or V/L/I).

Module annotation

After assignment to a module and structural cluster, and subsequent residue reordering, each template was annotated with structural and functional information. Basic information includes template size, mean inter-residue distance and RMSD to the corresponding reference template. The name of a module corresponds to its fuzzy residue composition (e.g. DE-DE-NQ-X), where each residue is written in one-letter code. Alternative residues in a position are written as multiple letters, with X and Z used for unspecific residue types contributing backbone atoms and post-translationally modified residues, respectively. Further annotations were collected from the M-CSA public API, such as EC number[69], CATH superfamily[15] and functional roles[70,71] for each residue position. The latter were grouped into empirical categories (metal ligand, proton shuttle, hydride shuttle, two-electron covalent catalysis, single-electron catalysis (radical), interaction/spectator) with the exact definitions[70,71] presented in Table S1. It should be noted that catalytic residue roles annotations rely on the underlying literature and discretion of the curator when describing the mechanism. Roles, especially those belonging to the Spectator/Interaction class, are often considered auxiliary and are subject to ambiguity. Also, residues might perform different roles in different steps during the course of the catalytic mechanism. For these reasons, “Metal Ligand” and “Reactant” classes are prioritised over “Spectator/Interaction” ones, when they co-exist at a residue, since the former are annotated more accurately using experimental/computational evidence.

In order to identify ligands interacting with module residues, we utilised the CSA-3D framework of our previous work[17], with the exact methodology being described thoroughly therein. For each template in a module, all ligands located 6Å or closer to all catalytic residue (minimum interatomic distance between ligand and residue in 3D space) are extracted and classified in four categories: Metal ions, co-factors (organic or inorganic co-factors composed of more than one atom, such as NAD and Fe-S metal clusters respectively), substrates (small molecule or polymer) and crystallographic artefacts (such as glycerol or ions irrelevant to the catalysed reaction). Classification involves an initial step of ion, co-factor and artefact identification by looking up their PDB three-letter codes in respective curated databases[45,47]. Any ligands not falling within these categories are classified as putative substrates/products. Subsequently, each ligand is compared as per its chemical similarity

(PARITY score[44]) with all cognate ligands of the reaction that the respective enzyme is known to catalyse (reaction information is collected from KEGG[72,73], Rhea[73] or M-CSA). For non-polymer substrates/products, in order for a ligand to be classified as a substrate, it needs to resemble one of the cognate substrates/products at ≥70% chemical similarity. Crystallographic artefacts are ignored since they are usually biologically irrelevant. Since a template represents a cluster of similar, homologous and conserved active sites, ligand information is collected from all those sites. Frequency of occurrence for each ligand in each template is calculated, and frequencies are aggregated over all templates of a module. Additionally, a score is calculated for each of the three ligand categories in each template: if a site within a template contains at least one ligand of a given category, the corresponding score in the template is incremented by 1. An average score is then calculated at the module level.

Last, we added mechanistic information associated with the catalytic site from which each template derives; this information is coded in the form of "the rules of enzyme catalysis", a concept introduced in our recent work[50]. Each rule represents a complete or partial step of the catalytic mechanism and is coded in a SMARTS string, representing a reaction involving the reaction centres of the catalytic step plus all atoms within two bonds of the reaction centre atoms. We compiled a database of these rules as part of M-CSA, that can be used to search an active site in 3D for possible mechanisms. We have also shown that rules can be a tool to investigate evolutionary relationships between enzymes at the mechanism level[14,74], by pairwise and stepwise rule comparison[75]. This is the basis of the characterisation method here: all rules mapped to respective M-CSA entries of the templates within a module are compared. However, as of yet, M-CSA rules are mapped on the active site level and not on the residue level, complicating the automation of the module annotation. Therefore, each case needs to be examined individually to check whether the rules shared between templates actually involve catalytic residues from these particular templates. A systematic analysis with precise residue-level rule mapping is currently in progress, and we plan to release a detailed, fully annotated library, where modules can be used as tools themselves for identifying enzyme mechanistic relationships in the context of convergent and divergent evolution. In the current analysis we used the version 1.0 rule set; in order to derive a set of generic rules able to capture mechanistic steps that are similar but not identical, rules were clustered according to three similarity features: reactants and products chemical similarity, bond changes and reaction centres similarity. The average of these scores was used as a similarity metric to perform hierarchical clustering and group together rules being ≥70% similar. If two rules share a chemical similarity of ≥70%[76], templates are considered to share a common mechanistic step, adding further evidence of functional convergence.

References

[1] A.J.M. Ribeiro, G.L. Holliday, N. Furnham, J.D. Tyzack, K. Ferris, J.M. Thornton, Mechanism and Catalytic Site Atlas (M-CSA): A database of enzyme reaction mechanisms and active sites, Nucleic Acids Res. 46 (2018) D618–D623. https://doi.org/10.1093/nar/gkx1012.

[2] N. Furnham, G.L. Holliday, T.A.P. de Beer, J.O.B. Jacobsen, W.R. Pearson, J.M. Thornton, The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes, Nucleic Acids Res. 42 (2014) D485–D489. https://doi.org/10.1093/NAR/GKT1243.

[3] A.R. Dewanti, J.A. Duine, Ca2+-assisted, direct hydride transfer, and rate-determining tautomerization of C5-reduced PQQ to PQQH2, in the oxidation of β-D-glucose by soluble, quinoprotein glucose dehydrogenase, Biochemistry. 39 (2000). https://doi.org/10.1021/bi992810x.

[4] M. Leopoldini, N. Russo, M. Toscano, The preferred reaction path for the oxidation of methanol by PQQ-containing methanol dehydrogenase: Addition-elimination versus hydride-transfer mechanism, Chemistry - A European Journal. 13 (2007). https://doi.org/10.1002/chem.200601123.

[5] L. Simón, J.M. Goodman, Enzyme catalysis by hydrogen bonds: The balance between transition State binding and substrate binding in oxyanion holes, Journal of Organic Chemistry. 75 (2010). https://doi.org/10.1021/jo901503d.

[6] R. Ménard, A.C. Storer, Oxyanion Hole Interactions in Serine and Cysteine Proteases, Biol Chem Hoppe Seyler. 373 (1992). https://doi.org/10.1515/bchm3.1992.373.2.393.

[7] C. Martinez, A. Nicolas, H. van Tilbeurgh, M.P. Egloff, C. Cambillau, C. Cudrey, R. Verger, Cutinase, a Lipolytic Enzyme with a Preformed Oxyanion Hole, Biochemistry. 33 (1994). https://doi.org/10.1021/bi00167a011.

[8] R.J. Kubiak, X. Yue, R.J. Hondal, C. Mihai, M.D. Tsai, K.S. Bruzik, Involvement of the Arg - Asp - His catalytic triad in enzymatic cleavage of the phosphodiester bond, Biochemistry. 40 (2001). https://doi.org/10.1021/bi002371y.

[9] L. Polgár, The catalytic triad of serine peptidases, Cellular and Molecular Life Sciences. 62 (2005). https://doi.org/10.1007/s00018-005-5160-x.

[10] Ö.D. Ekici, M. Paetzel, R.E. Dalbey, Unconventional serine proteases: Variations on the catalytic Ser/His/Asp triad configuration, Protein Sci. 17 (2008) 2023. https://doi.org/10.1110/PS.035436.108.

[11] A.R. Buller, C.A. Townsend, Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad, Proceedings of the National Academy of Sciences. 110 (2013) E653–E661. https://doi.org/10.1073/PNAS.1221050110.

[12] I.G. Riziotis, J.M. Thornton, Capturing the geometry, function, and evolution of enzymes with 3D templates, Protein Science. 31 (2022) e4363. https://doi.org/https://doi.org/10.1002/pro.4363.

[13] G.L. Holliday, C. Andreini, J.D. Fischer, S.A. Rahman, D.E. Almonacid, S.T. Williams, W.R. Pearson, MACiE: Exploring the diversity of biochemical reactions, Nucleic Acids Res. 40 (2012). https://doi.org/10.1093/nar/gkr799.

[14] N. Furnham, N.L. Dawson, S.A. Rahman, J.M. Thornton, C.A. Orengo, Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies, J Mol Biol. 428 (2016) 253–267. https://doi.org/10.1016/j.jmb.2015.11.010.

[15] I. Sillitoe, N. Dawson, T.E. Lewis, S. Das, J.G. Lees, P. Ashford, A. Tolulope, H.M. Scholes, I. Senatorov, A. Bujan, F. Ceballos Rodriguez-Conde, B. Dowling, J. Thornton, C.A. Orengo, CATH: Expanding the horizons of structure-based functional annotations for genome sequences, Nucleic Acids Res. 47 (2019) D280–D284. https://doi.org/10.1093/nar/gky1097.

[16] S. Romero-Romero, S. Kordes, F. Michel, B. Höcker, Evolution, folding, and design of TIM barrels and related proteins, Curr Opin Struct Biol. 68 (2021). https://doi.org/10.1016/j.sbi.2020.12.007.

[17] I.G. Riziotis, A.J.M. Ribeiro, N. Borkakoti, J.M. Thornton, Conformational Variation in Enzyme Catalysis: A Structural Study on Catalytic Residues, J Mol Biol. 434 (2022). https://doi.org/10.1016/j.jmb.2022.167517.

[18] R.A. Laskowski, J.D. Watson, J.M. Thornton, ProFunc: A server for predicting protein function from 3D structure, Nucleic Acids Res. 33 (2005). https://doi.org/10.1093/nar/gki414.

[19] A. Stark, R.B. Russell, Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures, Nucleic Acids Res. 31 (2003). https://doi.org/10.1093/nar/gkg506.

[20] G.J. Kleywegt, Recognition of spatial motifs in protein structures, J Mol Biol. 285 (1999) 1887–1897. https://doi.org/10.1006/jmbi.1998.2393.

[21] G. Wang, R.L. Dunbrack, PISCES: A protein sequence culling server, Bioinformatics. 19 (2003). https://doi.org/10.1093/bioinformatics/btg224.

[22] J.A. Barker, J.M. Thornton, An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis, Bioinformatics. 19 (2003) 1644–1649. https://doi.org/10.1093/bioinformatics/btg226.

[23] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software Environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003). https://doi.org/10.1101/gr.1239303.

[24] R.H.H. Van Den Heuvel, D.I. Svergun, M. V. Petoukhov, A. Coda, B. Curti, S. Ravasio, M.A. Vanoni, A. Mattevi, The active conformation of glutamate synthase and its binding to ferredoxin, J Mol Biol. 330 (2003). https://doi.org/10.1016/S0022-2836(03)00522-9.

[25] J.L. Smith, Glutamine PRPP amidotransferase: Snapshots of an enzyme in action, Curr Opin Struct Biol. 8 (1998). https://doi.org/10.1016/S0959-440X(98)80087-0.

[26] Y. Cheng, D.J. Patel, Crystallographic structure of the nuclease domain of 3′hExo, a DEDDh family member, bound to rAMP, J Mol Biol. 343 (2004). https://doi.org/10.1016/j.jmb.2004.08.055.

[27] G. Andrés Cisneros, L. Perera, R.M. Sehaaper, L.C. Pedersen, R.E. London, L.G. Pedersen, T.A. Darden, Reaction mechanism of the ε subunit of E. coli DNA polymerase III: Insights into active site metal coordination and catalytically significant residues, J Am Chem Soc. 131 (2009). https://doi.org/10.1021/ja8082818.

[28] J.F. Rakus, A.A. Fedorov, E. V. Fedorov, M.E. Glasner, B.K. Hubbard, J.D. Delli, P.C. Babbitt, S.C. Almo, J.A. Gerlt, Evolution of enzymatic activities in the enolase

superfamily: L-rhamnonate dehydratase, Biochemistry. 47 (2008). https://doi.org/10.1021/bi800914r.

[29] M. Nagar, B.N. Wyatt, M. St. Maurice, S.L. Bearne, Inactivation of mandelate racemase by 3-hydroxypyruvate reveals a potential mechanistic link between enzyme superfamilies, Biochemistry. 54 (2015). https://doi.org/10.1021/acs.biochem.5b00221.

[30] S.Y. Wen, A.A. Fedorov, E. V. Fedorov, S.C. Almo, J.A. Gerlt, Evolution of enzymatic activities in the enolase superfamily: L-talarate/galactarate dehydratase from Salmonella typhimurium LT2, Biochemistry. 46 (2007). https://doi.org/10.1021/bi7008882.

[31] K. Yamada, N. Kunishima, K. Mayanagi, T. Ohnishi, T. Nishino, H. Iwasaki, H. Shinagawa, K. Morikawa, Crystal structure of the Holliday junction migration motor protein RuvB from Thermus thermophilus HB8, Proc Natl Acad Sci U S A. 98 (2001). https://doi.org/10.1073/pnas.98.4.1442.

[32] G. Yang, T. Sandalova, K. Lohman, Y. Lindqvist, A.R. Rendina, Active site mutants of Escherichia coli dethiobiotin synthetase: Effects of mutations on enzyme catalytic and structural properties, Biochemistry. 36 (1997). https://doi.org/10.1021/bi9631677.

[33] A.C. Eliot, J. Sandmark, G. Schneider, J.F. Kirsch, The dual-specific active site of 7,8-diaminopelargonic acid synthase and the effect of the R391A mutation, Biochemistry. 41 (2002). https://doi.org/10.1021/bi026339a.

[34] C.E. Pugh, J.L. Harwood, R.A. John, Mechanism of glutamate semialdehyde aminotransferase: Roles of diamino- and dioxo-intermediates in the synthesis of aminolevulinate, Journal of Biological Chemistry. 267 (1992). https://doi.org/10.1016/s0021-9258(18)45985-6.

[35] D.M.E. Szebenyi, F.N. Musayev, M.L. Di Salvo, M.K. Safo, V. Schirch, Serine hydroxymethyltransferase: Role of Glu75 and evidence that serine is cleaved by a retroaldol mechanism, Biochemistry. 43 (2004). https://doi.org/10.1021/bi049791y.

[36] J.A. Gerlt, P.C. Babbitt, Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies, Annu Rev Biochem. 70 (2001). https://doi.org/10.1146/annurev.biochem.70.1.209.

[37] F. Baier, J.N. Copp, N. Tokuriki, Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence-Function Relationships, Biochemistry. 55 (2016). https://doi.org/10.1021/acs.biochem.6b00723.

[38] A.C. Wallace, R.A. Laskowski, J.M. Thornton, Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases, Protein Science. 5 (1996). https://doi.org/10.1002/pro.5560050603.

[39] R.D. Gupta, Recent advances in enzyme promiscuity, Sustainable Chemical Processes. 4 (2016). https://doi.org/10.1186/s40508-016-0046-9.

[40] T. Devamani, A.M. Rauwerdink, M. Lunzer, B.J. Jones, J.L. Mooney, M.A.O. Tan, Z.J. Zhang, J.H. Xu, A.M. Dean, R.J. Kazlauskas, Catalytic Promiscuity of Ancestral Esterases and Hydroxynitrile Lyases, J Am Chem Soc. 138 (2016). https://doi.org/10.1021/jacs.5b12209.

[41] I. Sillitoe, N. Furnham, FunTree: Advances in a resource for exploring and contextualising protein function evolution, Nucleic Acids Res. 44 (2016) D317–D323. https://doi.org/10.1093/nar/gkv1274.

[42] J.M. Murphy, H. Farhan, P.A. Eyers, Bio-Zombie: The rise of pseudoenzymes in biology, Biochem Soc Trans. 45 (2017) 537–544. https://doi.org/10.1042/BST20160400.

[43] A.J.M. Ribeiro, J.D. Tyzack, N. Borkakoti, J.M. Thornton, Identifying pseudoenzymes using functional annotation: pitfalls of common practice, FEBS Journal. 287 (2020). https://doi.org/10.1111/febs.15142.

[44] J.D. Tyzack, L. Fernando, A.J.M. Ribeiro, N. Borkakoti, J.M. Thornton, Ranking Enzyme Structures in the PDB by Bound Ligand Similarity to Biological Substrates, Structure. 26 (2018) 565-571.e3. https://doi.org/10.1016/j.str.2018.02.009.

[45] J.D. Fischer, G.L. Holliday, J.M. Thornton, The CoFactor database: Organic cofactors in enzyme catalysis, Bioinformatics. 26 (2010) 2496–2497. https://doi.org/10.1093/bioinformatics/btq442.

[46] C. Andreini, I. Bertini, G. Cavallaro, G.L. Holliday, J.M. Thornton, Metal ions in biological catalysis: From enzyme databases to general principles, Journal of Biological Inorganic Chemistry. 13 (2008). https://doi.org/10.1007/s00775-008-0404-5.

[47] C. Andreini, I. Bertini, G. Cavallaro, G.L. Holliday, J.M. Thornton, Metal-MACiE: A database of metals involved in biological catalysis, Bioinformatics. 25 (2009). https://doi.org/10.1093/bioinformatics/btp256.

[48] Y.W. Lin, Rational design of artificial metalloproteins and metalloenzymes with metal clusters, Molecules. 24 (2019). https://doi.org/10.3390/molecules24152743.

[49] W.L. DeLano, The PyMOL Molecular Graphics System, Version 2.3, Schrödinger LLC. (2020).

[50] A.J.M. Ribeiro, I.G. Riziotis, J.D. Tyzack, N. Borkakoti, J.M. Thornton, EzMechanism: An Automated Tool to Propose Catalytic Mechanisms of Enzyme Reactions, BioRxiv. (2022).

[51] G.A. Jaña, F. Mendoza, M.I. Osorio, J.B. Alderete, P.A. Fernandes, M.J. Ramos, V.A. Jiménez, A QM/MM approach on the structural and stereoelectronic factors governing glycosylation by GTF-SI from: Streptococcus mutans, Org Biomol Chem. 16 (2018). https://doi.org/10.1039/c8ob00284c.

[52] J.D. Dimitrov, T.L. Vassilev, Cofactor-mediated protein promiscuity, Nat Biotechnol. 27 (2009). https://doi.org/10.1038/nbt1009-892a.

[53] C. UniProt, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Res. 51 (2023) D523–D531. https://doi.org/10.1093/nar/gkac1052.

[54] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature. 596 (2021). https://doi.org/10.1038/s41586-021-03819-2.

[55] J.M. Thornton, R.A. Laskowski, N. Borkakoti, AlphaFold heralds a data-driven revolution in biology and medicine, Nat Med. 27 (2021). https://doi.org/10.1038/s41591-021-01533-0.

[56] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Res. 50 (2022). https://doi.org/10.1093/nar/gkab1061.

[57] R.A. Laskowski, J.D. Watson, J.M. Thornton, Protein function prediction using local 3D templates, J Mol Biol. 351 (2005). https://doi.org/10.1016/j.jmb.2005.05.067.

[58] E.C. Meng, B.J. Polacco, P.C. Babbitt, Superfamily active site templates, Proteins: Structure, Function and Genetics. 55 (2004) 962–976. https://doi.org/10.1002/prot.20099.

[59] S. Jones, J.A. Barker, I. Nobeli, J.M. Thornton, Using structural motif templates to identify proteins with DNA binding function, Nucleic Acids Res. 31 (2003). https://doi.org/10.1093/nar/gkg386.

[60] E.T. Farinas, T. Bulter, F.H. Arnold, Directed enzyme evolution, Curr Opin Biotechnol. 12 (2001) 545–551. https://doi.org/10.1016/S0958-1669(01)00261-0.

[61] M. Goldsmith, D.S. Tawfik, Directed enzyme evolution: Beyond the low-hanging fruit, Curr Opin Struct Biol. 22 (2012). https://doi.org/10.1016/j.sbi.2012.03.010.

[62] F. Richter, A. Leaver-Fay, S.D. Khare, S. Bjelic, D. Baker, De novo enzyme design using Rosetta3, PLoS One. 6 (2011). https://doi.org/10.1371/journal.pone.0019230.

[63] S. Velankar, C. Best, B. Beuth, C.H. Boutselakis, N. Cobley, A.W. Sousa da Silva, D. Dimitropoulos, A. Golovin, M. Hirshberg, M. John, E.B. Krissinel, R. Newman, T. Oldfield, A. Pajon, C.J. Penkett, J. Pineda-Castillo, G. Sahni, S. Sen, R. Slowley, A. Suarez-Uruena, J. Swaminathan, G. van Ginkel, W.F. Vranken, K. Henrick, G.J. Kleywegt, PDBe: Protein Data Bank in Europe, Nucleic Acids Res. 38 (2009). https://doi.org/10.1093/nar/gkp916.

[64] A.C. Wallace, N. Borkakoti, J.M. Thornton, Tess: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites, Protein Science. 6 (1997) 2308–2323. https://doi.org/10.1002/pro.5560061104.

[65] A. Stark, S. Sunyaev, R.B. Russell, A model for statistical significance of local similarities in structure, J Mol Biol. 326 (2003). https://doi.org/10.1016/S0022-2836(03)00045-7.

[66] I.K. Mcdonald, J.M. Thornton, The application of hydrogen bonding analysis in x-ray crystallography to help orientate asparagine, glutamine and histidine side chains, Protein Engineering, Design and Selection. 8 (1995) 217–224. https://doi.org/10.1093/protein/8.3.217.

[67] W. Kabsch, A solution for the best rotation to relate two sets of vectors, Acta Crystallographica Section A. 32 (1976) 922–923. https://doi.org/10.1107/S0567739476001873.

[68] K.L. Damm, H.A. Carlson, Gaussian-weighted RMSD superposition of proteins: A structural comparison for flexible proteins and predicted protein structures, Biophys J. 90 (2006) 4558–4573. https://doi.org/10.1529/biophysj.105.066654.

[69] K.F. Tipton, Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, Academic Press. 6 (1992).

[70] G.L. Holliday, J.B.O. Mitchell, J.M. Thornton, Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis, J Mol Biol. 390 (2009) 560–577. https://doi.org/10.1016/J.JMB.2009.05.015.

[71] A.J.M. Ribeiro, J.D. Tyzack, N. Borkakoti, G.L. Holliday, J.M. Thornton, A global analysis of function and conservation of catalytic residues in enzymes, J Biol Chem. 295 (2020) 314. https://doi.org/10.1074/JBC.REV119.006289.

[72] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (2017) D353–D361. https://doi.org/10.1093/nar/gkw1092.

[73] R. Alcántara, K.B. Axelsen, A. Morgat, E. Belda, E. Coudert, A. Bridge, H. Cao, P. De Matos, M. Ennis, S. Turner, G. Owen, L. Bougueleret, I. Xenarios, C. Steinbeck, Rhea - A manually curated resource of biochemical reactions, Nucleic Acids Res. 40 (2012) D754–D760. https://doi.org/10.1093/nar/gkr1126.

[74] G.J. Bartlett, N. Borkakoti, J.M. Thornton, Catalysing new reactions during evolution: Economy of residues and mechanism, J Mol Biol. 331 (2003) 829–860. https://doi.org/10.1016/S0022-2836(03)00734-4.

[75] A.J.M. Ribeiro, I.G. Riziotis, J.D. Tyzack, N. Borkakoti, J.M. Thornton, Using mechanism similarity to understand enzyme evolution, Biophys Rev. 14 (2022) 1273–1280. https://doi.org/10.1007/s12551-022-01022-9.

[76] S.A. Rahman, S.M. Cuesta, N. Furnham, G.L. Holliday, J.M. Thornton, EC-BLAST: A tool to automatically search and compare enzyme reactions, Nat Methods. 11 (2014). https://doi.org/10.1038/nmeth.2803.

Data availability

All data is available upon request.

Conflict of interest

Authors declare no conflict of interest.

Author contributions

IGR: Conceptualisation, Software, Methodology, Validation, Formal analysis, Investigation, Visualisation, Writing – Original Draft preparation, Writing – Review & Editing

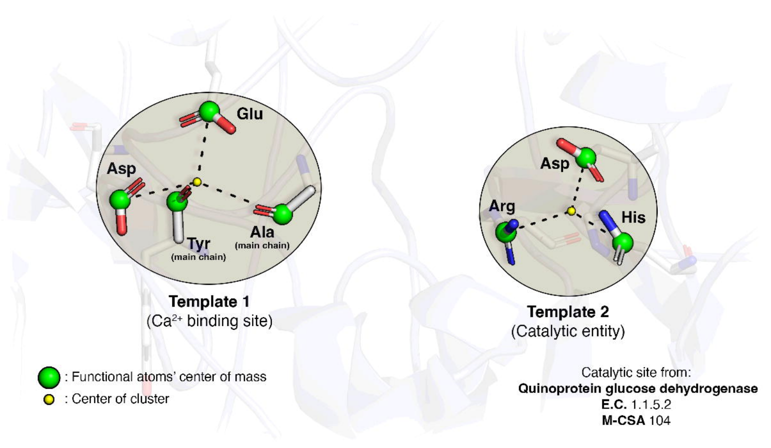
AJMR: Conceptualisation, Data Curation, Writing – Review & Editing

NB: Conceptualisation, Writing – Review & Editing

JMT: Conceptualisation, Supervision, Resources, Funding Acquisition, Project Administration, Writing – Review & Editing

Acknowledgements

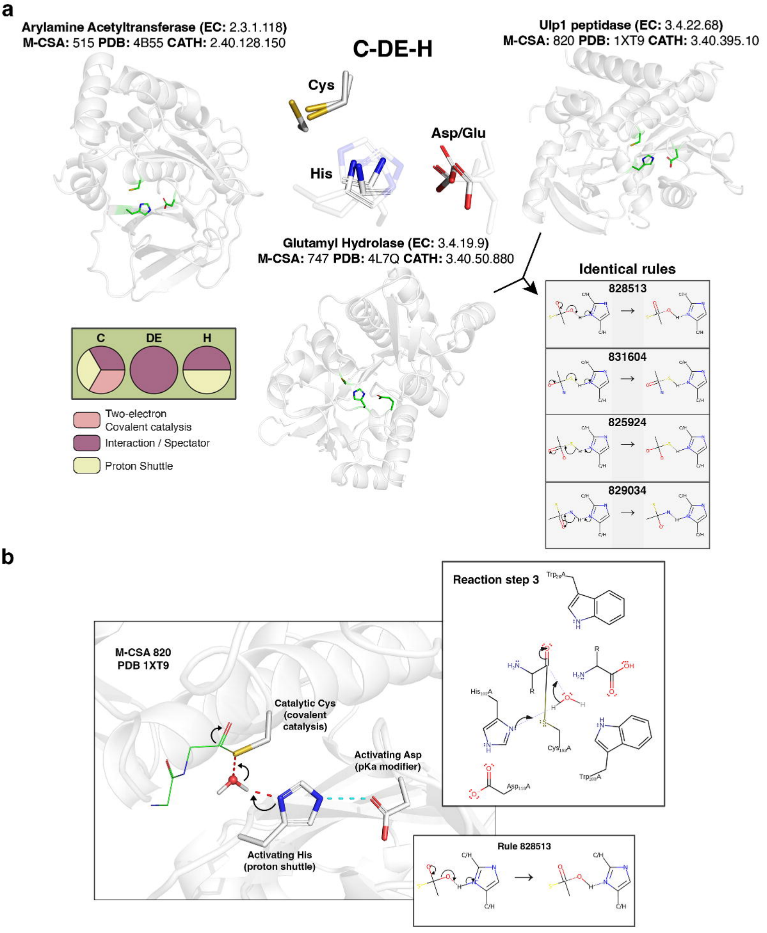
The work was supported by: the EMBL International PhD Programme (IGR) and the European Molecular Biology Laboratory (AJMR, NB, JMT)

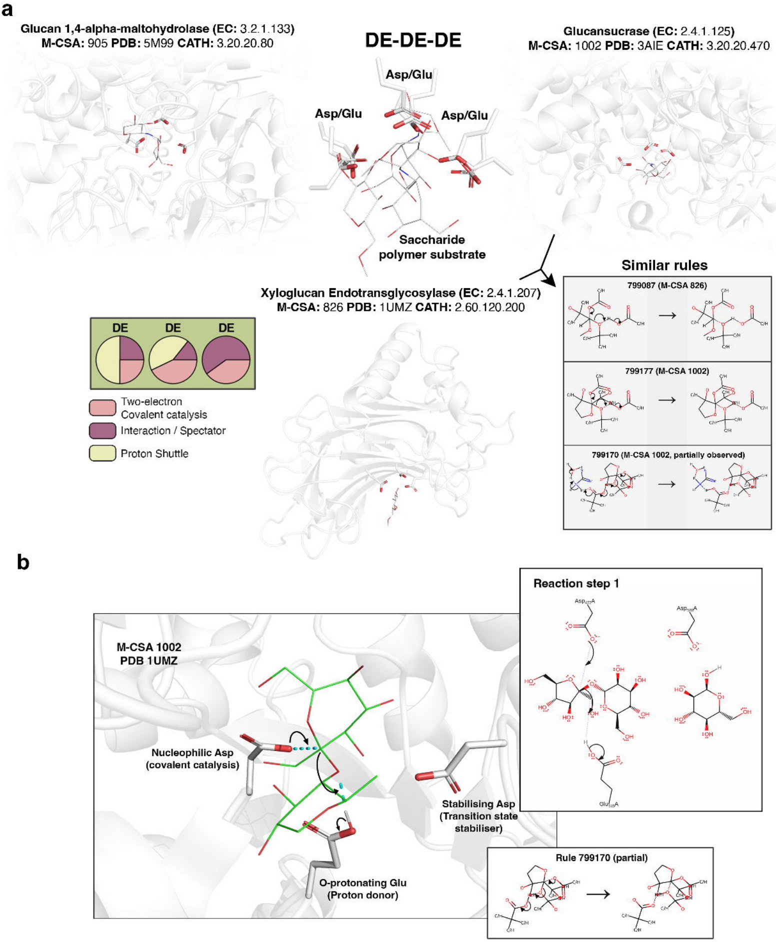


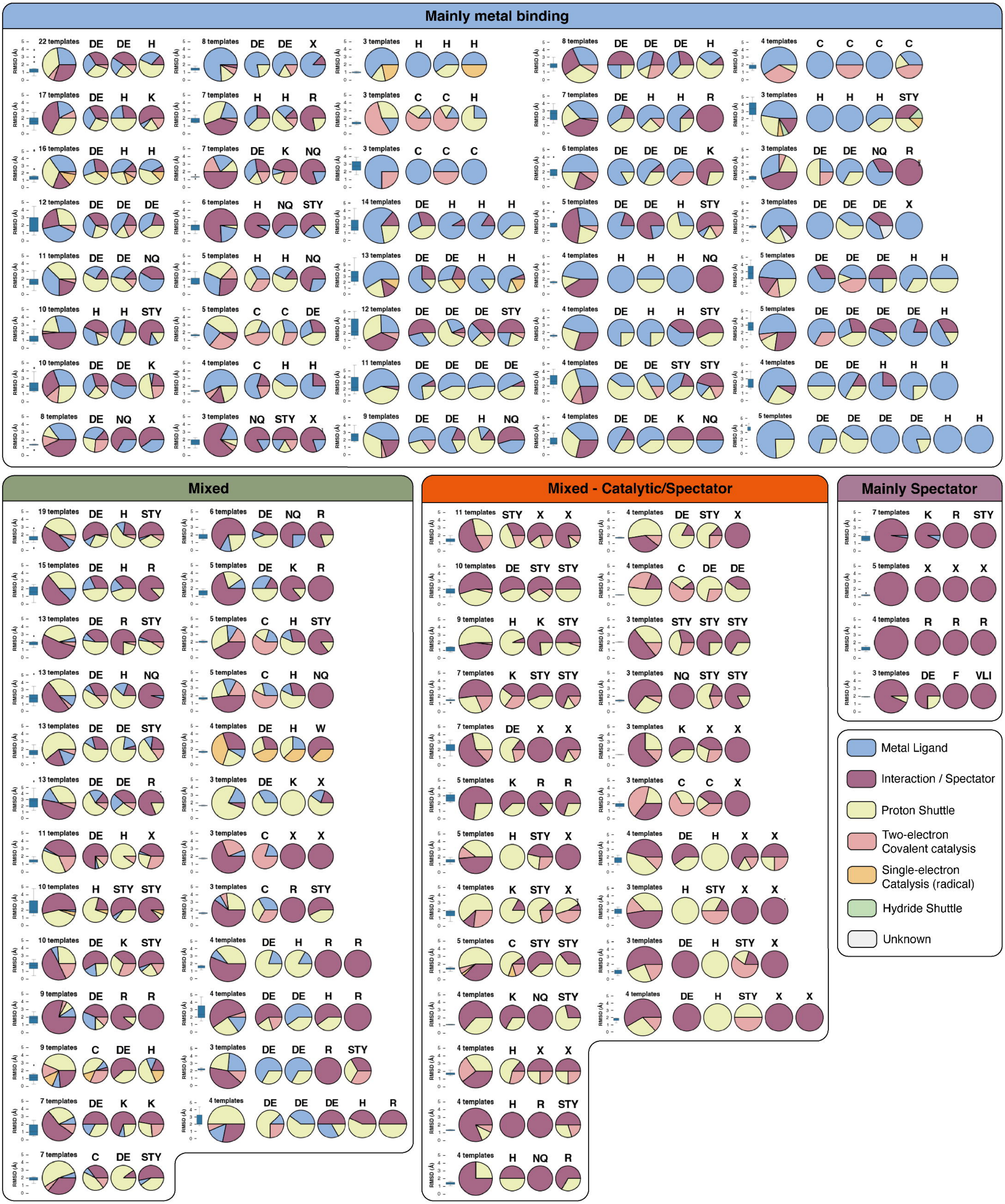
Template 1
(Ca^{2+} binding site)

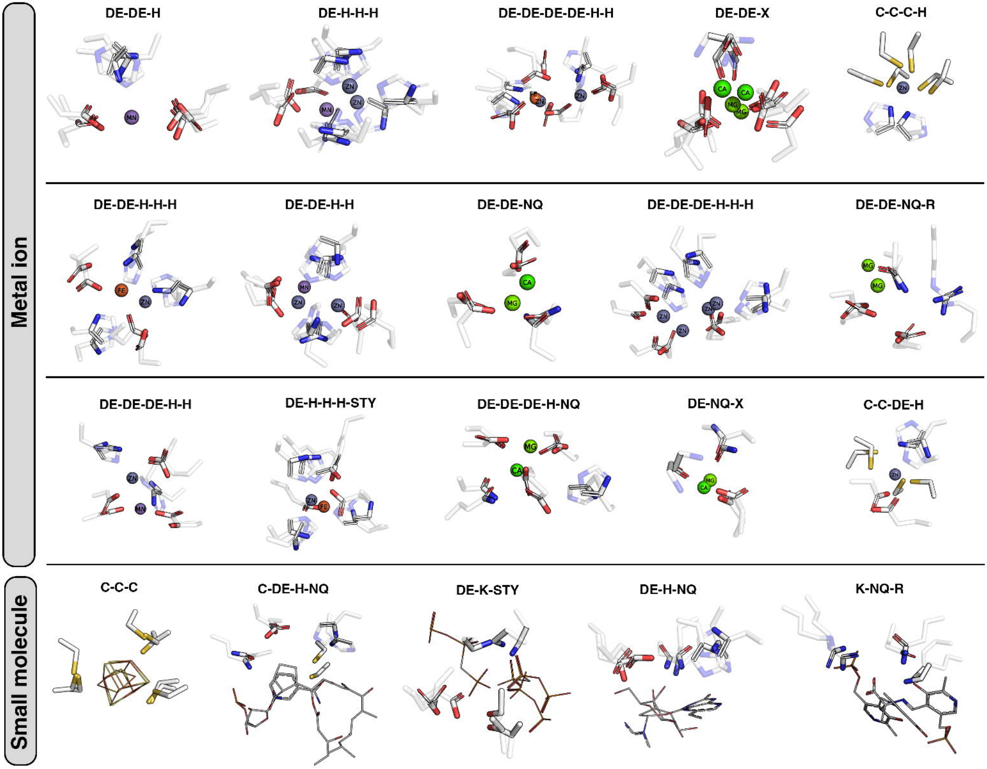
Template 2
(Catalytic entity)

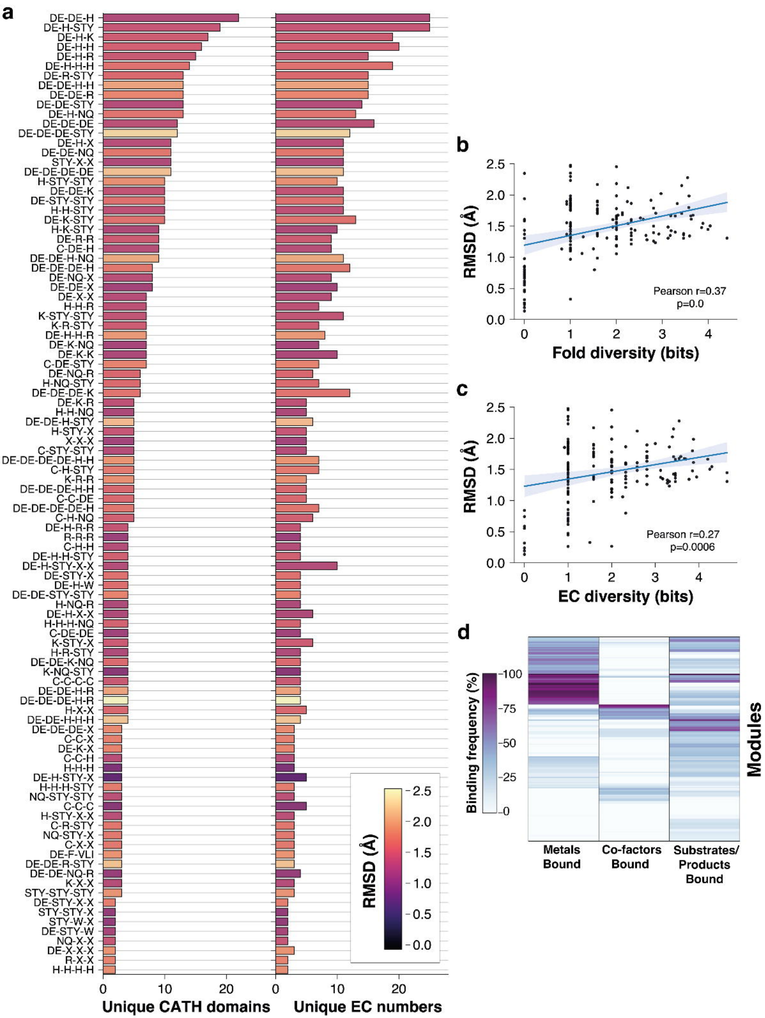
Catalytic site from:
Quinoprotein glucose dehydrogenase
E.C. 1.1.5.2
M-CSA 104

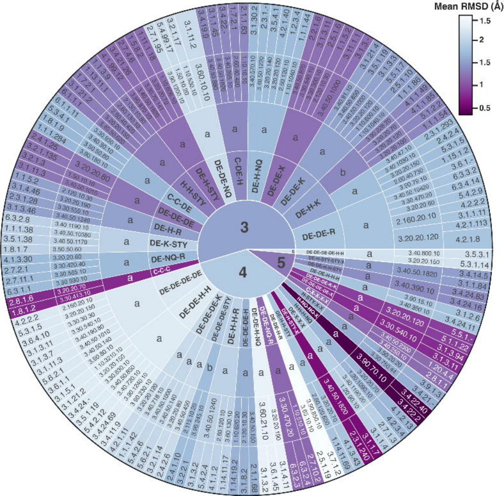

a
Arylamine Acetyltransferase (EC: 2.3.1.118)
M-CSA: 515 PDB: 4B55 CATH: 2.40.128.150
C-DE-H
Cys
His
Asp/Glu
Ulp1 peptidase (EC: 3.4.22.68)
M-CSA: 820 PDB: 1XT9 CATH: 3.40.395.10
Glutamyl Hydrolase (EC: 3.4.19.9)
M-CSA: 747 PDB: 4L7Q CATH: 3.40.50.880
C
DE
H
Two-electron Covalent catalysis
Interaction / Spectator
Proton Shuttle
Identical rules
828513
831604
825924
829034
C/H
b
M-CSA 820
PDB 1XT9
Catalytic Cys (covalent catalysis)
Activating Asp (pKa modifier)
Activating His (proton shuttle)
Reaction step 3
Trp28A
His102A
Cys163A
Asp119A
Trp203A
R
Rule 828513

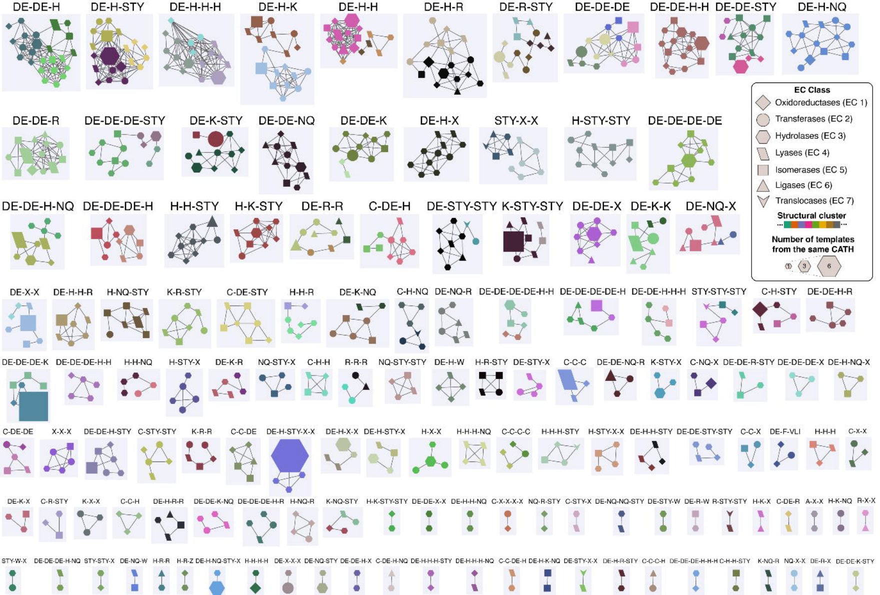

a
Glucan 1,4-alpha-maltohydrolase (EC: 3.2.1.133)
M-CSA: 905 PDB: 5M99 CATH: 3.20.20.80
DE-DE-DE
Asp/Glu
Asp/Glu
Asp/Glu
Saccharide polymer substrate
Glucansucrase (EC: 2.4.1.125)
M-CSA: 1002 PDB: 3AIE CATH: 3.20.20.470
Xyloglucan Endotransglycosylase (EC: 2.4.1.207)
M-CSA: 826 PDB: 1UMZ CATH: 2.60.120.200
DE
DE
DE
Two-electron Covalent catalysis
Interaction / Spectator
Proton Shuttle
Similar rules
799087 (M-CSA 826)
799177 (M-CSA 1002)
799170 (M-CSA 1002, partially observed)
b
M-CSA 1002
PDB 1UMZ
Nucleophilic Asp
(covalent catalysis)
Stabilising Asp
(Transition state stabiliser)
O-protonating Glu
(Proton donor)
Reaction step 1
Asp477A
Asp588A
Glu515A
Rule 799170 (partial)

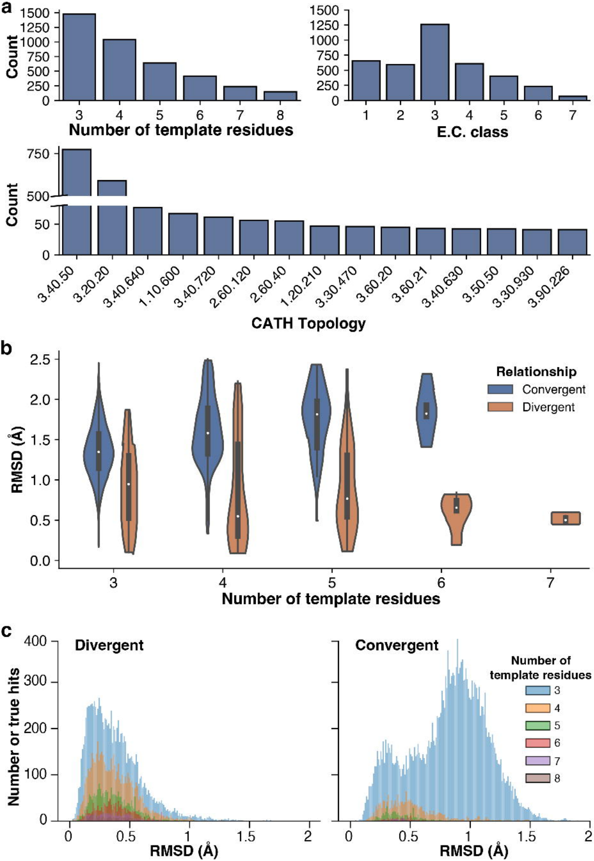

Mainly metal binding
22 templates DE DE H
8 templates DE DE X
3 templates H H H
8 templates DE DE DE H
4 templates C C C C
17 templates DE H K
7 templates H H R
3 templates C C H
7 templates DE H H R
3 templates H H H STY
16 templates DE H H
7 templates DE K NQ
3 templates C C C
6 templates DE DE DE K
3 templates DE DE NQ R
12 templates DE DE DE
6 templates H NQ STY
14 templates DE H H H
5 templates DE DE H STY
3 templates DE DE DE X
11 templates DE DE NQ
5 templates H H NQ
13 templates DE DE H H
4 templates H H H NQ
5 templates DE DE DE H H
10 templates H H STY
5 templates C C DE
12 templates DE DE DE STY
4 templates DE H H STY
5 templates DE DE DE DE H
10 templates DE DE K
4 templates C H H
11 templates DE DE DE DE
4 templates DE DE STY STY
4 templates DE DE H H H
8 templates DE NQ X
3 templates NQ STY X
9 templates DE DE H NQ
4 templates DE DE K NQ
5 templates DE DE DE DE H H
Mixed
19 templates DE H STY
15 templates DE H R
13 templates DE R STY
13 templates DE H NQ
13 templates DE DE STY
13 templates DE DE R
11 templates DE H X
10 templates H STY STY
10 templates DE K STY
9 templates DE R R
9 templates C DE H
7 templates DE K K
7 templates C DE STY
6 templates DE NQ R
5 templates DE K R
5 templates C H STY
5 templates C H NQ
4 templates DE H W
3 templates DE K X
3 templates C X X
3 templates C R STY
4 templates DE H R R
4 templates DE DE H R
3 templates DE DE R STY
4 templates DE DE DE H R
Mixed - Catalytic/Spectator
11 templates STY X X
10 templates DE STY STY
9 templates H K STY
7 templates K STY STY
7 templates DE X X
5 templates K R R
5 templates H STY X
4 templates K STY X
5 templates C STY STY
4 templates K NQ STY
4 templates H X X
4 templates H R STY
4 templates H NQ R
4 templates DE STY X
4 templates C DE DE
3 templates STY STY STY
3 templates NQ STY STY
3 templates K X X
3 templates C C X
4 templates DE H X X
3 templates H STY X X
3 templates DE H STY X
4 templates DE H STY X X
Mainly Spectator
7 templates K R STY
5 templates X X X
4 templates R R R
3 templates DE F VLI
RMSD (Å)
0 1 2 3 4 5
Metal Ligand
Interaction / Spectator
Proton Shuttle
Two-electron Covalent catalysis
Single-electron Catalysis (radical)
Hydride Shuttle
Unknown

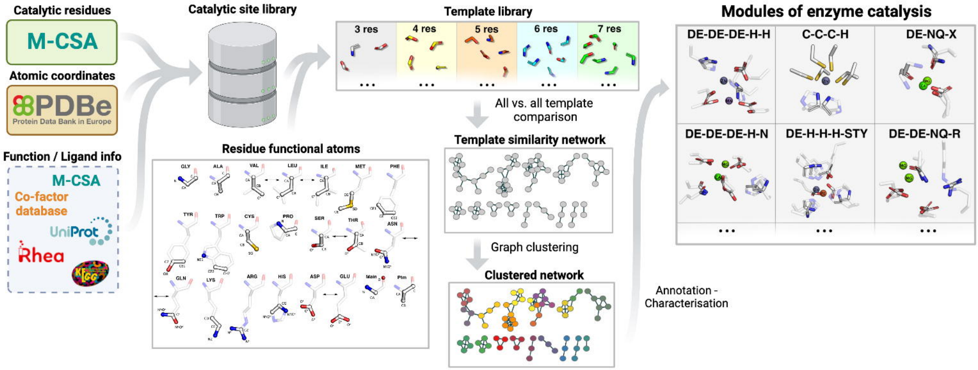

Metal ion
DE-DE-H
DE-H-H-H
DE-DE-DE-DE-H-H
DE-DE-X
C-C-C-H
DE-DE-H-H-H
DE-DE-H-H
DE-DE-NQ
DE-DE-DE-H-H-H
DE-DE-NQ-R
DE-DE-DE-H-H
DE-H-H-H-STY
DE-DE-DE-H-NQ
DE-NQ-X
C-C-DE-H
Small molecule
C-C-C
C-DE-H-NQ
DE-K-STY
DE-H-NQ
K-NQ-R
MN
ZN
FE
CA
MG


a
DE-DE-H
DE-H-STY
DE-H-K
DE-H-H
DE-H-R
DE-H-H-H
DE-R-STY
DE-DE-H-H
DE-DE-R
DE-DE-STY
DE-H-NQ
DE-DE-DE
DE-DE-DE-STY
DE-H-X
DE-DE-NQ
STY-X-X
DE-DE-DE-DE
H-STY-STY
DE-DE-K
DE-STY-STY
H-H-STY
DE-K-STY
H-K-STY
DE-R-R
C-DE-H
DE-DE-H-NQ
DE-DE-DE-H
DE-NQ-X
DE-DE-X
DE-X-X
H-H-R
K-STY-STY
K-R-STY
DE-H-H-R
DE-K-NQ
DE-K-K
C-DE-STY
DE-NQ-R
H-NQ-STY
DE-DE-DE-K
DE-K-R
H-H-NQ
DE-DE-H-STY
H-STY-X
X-X-X
C-STY-STY
DE-DE-DE-DE-H-H
C-H-STY
K-R-R
DE-DE-DE-H-H
C-C-DE
DE-DE-DE-DE-H
C-H-NQ
DE-H-R-R
R-R-R
C-H-H
DE-H-H-STY
DE-H-STY-X-X
DE-STY-X
DE-H-W
DE-DE-STY-STY
H-NQ-R
DE-H-X-X
H-H-H-NQ
C-DE-DE
K-STY-X
H-R-STY
DE-DE-K-NQ
K-NQ-STY
C-C-C-C
DE-DE-H-R
DE-DE-DE-H-R
H-X-X
DE-DE-H-H-H
DE-DE-DE-X
C-C-X
DE-K-X
C-C-H
H-H-H
DE-H-STY-X
H-H-H-STY
NQ-STY-STY
C-C-C
H-STY-X-X
C-R-STY
NQ-STY-X
C-X-X
DE-F-VLI
DE-DE-R-STY
DE-DE-NQ-R
K-X-X
STY-STY-STY
DE-STY-X-X
STY-STY-X
STY-W-X
DE-STY-W
NQ-X-X
DE-X-X-X
R-X-X
H-H-H-H
0
10
20
Unique CATH domains
0
10
20
Unique EC numbers
RMSD (Å)
2.5
2.0
1.5
1.0
0.5
0.0
b
RMSD (Å)
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
Fold diversity (bits)
Pearson r=0.37
p=0.0
c
RMSD (Å)
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
EC diversity (bits)
Pearson r=0.27
p=0.0006
d
Binding frequency (%)
100
75
50
25
0
Modules
Metals Bound
Co-factors Bound
Substrates/ Products Bound


Mean RMSD (Å)
1.5
2
1.5
1
0.5
3
4
5
6
DE-DE-NQ
C-DE-H
DE-H-NQ
DE-DE-X
DE-DE-K
DE-H-K
DE-DE-R
DE-H-STY
H-H-STY
C-C-DE
DE-DE-DE
DE-H-R
DE-K-STY
DE-NQ-R
C-C-C
DE-DE-DE-DE
DE-DE-H-H
DE-DE-DE-K
DE-DE-DE-STY
DE-H-H-R
DE-DE-DE-H
DE-DE-H-NQ
DE-DE-NQ-R
DE-DE-H-R
DE-H-H-STY
DE-H-STY-X
DE-X-X-X
H-H-H-NQ
H-NQ-NQ-X
H-STY-X-X
C-X-X-X-X
DE-DE-DE-DE-X
DE-DE-DE-K-K
DE-DE-H-H-H
DE-DE-H-H-STY
DE-H-STY-STY-X
DE-DE-DE-DE-H-H
a
b
3.60.10.10
1.10.606.10
3.20.20.70
3.30.70.141
3.40.50.1820
3.90.870.10
2.130.10.80
2.60.130.10
2.60.40.420
3.20.20.150
3.40.50.620
3.30.930.10
3.40.30.10
3.50.50.60
3.90.180.10
3.20.20.80
1.50.10.10
3.40.50.1010
2.120.10.30
3.20.20.190
3.30.559.10
3.40.50.1240
3.40.1190.10
3.40.50.10380
3.40.50.1170
3.50.50.60
3.20.20.60
3.30.420.40
3.30.565.10
3.30.930.10
3.20.20.70
3.30.413.10
1.10.530.10
1.50.10.20
3.90.1200.10
3.40.50.880
3.40.50.1820
3.40.365.10
2.60.40.420
1.10.10.10
3.40.570.10
3.40.50.1220
3.20.20.140
3.20.20.120
2.80.120.10
1.10.1040.10
3.40.50.970
3.30.540.10
3.30.300.10
2.120.10.30
3.40.50.1000
3.40.91.10
3.40.50.850
3.40.50.1000
3.20.20.120
1.10.600.10
1.10.1040.10
3.90.228.20
3.20.20.70
3.20.20.120
3.20.20.100
3.40.47.10
3.40.1030.10
3.20.20.150
2.80.40.730
3.90.79.10
3.40.50.10420
3.30.470.20
3.20.20.60
2.160.20.10
3.20.20.120
3.40.800.10
3.20.20.150
3.40.50.1820
3.40.390.10
3.60.15.10
3.40.390.10
3.20.20.120
3.30.540.10
3.40.50.2300
3.40.250.10
3.50.20.20
3.30.750.44
3.90.70.10
3.20.20.70
1.20.1090.10
3.40.1190.20
1.20.90.10
3.40.50.1820
2.60.120.650
3.20.20.70
3.90.850.10
3.65.10.10
1.10.510.10
3.30.470.20
3.20.20.190
3.60.21.10
3.40.50.150
2.120.10.30
1.10.620.20
3.60.130.10
2.50.120.10
3.40.50.1240
3.90.245.10
2.115.10.20
3.40.50.450
3.20.20.80
3.40.50.140
3.40.50.1000
3.40.718.10
3.20.20.120
3.90.230.10
3.90.1240.10
3.40.720.10
3.40.50.850
3.30.830.10
3.20.20.150
1.10.3210.10
3.90.80.10
3.40.50.140
3.40.50.1010
3.40.190.80
3.30.540.10
3.30.30.30
3.20.20.150
2.160.20.10
3.4.19.9
3.1.1.45
3.4.22.-
1.7.2.1
2.1.1.63
3.1.11.2
3.2.1.17
5.4.99.17
2.7.1.95
1.11.1.18
1.5.8.2
2.7.4.6
3.4.21.26
4.1.99.12
1.1.3.9
1.13.11.1
1.7.2.1
3.1.21.2
6.1.1.1
6.1.1.11
5.3.4.1
1.8.1.9
1.1.1.284
2.4.1.25
3.2.1.135
3.2.1.3
3.1.11.3
1.1.5.2
3.1.4.46
2.3.1.28
3.1.3.46
6.3.2.8
1.1.1.38
3.5.1.38
1.8.1.7
4.1.3.30
2.7.2.1
2.7.11.1
6.3.1.1
2.8.1.6
1.8.1.2
3.1.30.2
2.3.1.-
3.5.4.1
4.2.1.40
1.2.3.4
1.1.1.44
2.2.1.6
3.1.3.11
2.5.1.6
1.1.5.2
3.11.1.1
3.1.3.-
3.1.21.4
3.5.1.19
3.1.3.3
5.5.1.7
2.5.1.10
1.1.1.86
4.1.1.49
4.1.1.85
5.1.2.2
1.1.1.54
2.3.1.293
2.4.2.4
5.3.1.5
1.15.1.2
3.6.1.-
6.3.3.2
6.3.4.14
5.4.2.9
4.2.2.2
3.1.1.11
4.2.1.113
4.2.1.8
3.5.3.1
5.3.1.14
3.4.14.5
3.1.1.84
3.4.24.83
3.4.24.16
3.1.2.6
3.4.24.11
5.5.1.-
5.1.1.22
3.1.3.94
3.1.3.11
1.20.4.4
2.8.1.1
3.9.1.3
3.4.21.-
3.4.22.40
3.4.22.2
4.1.2.13
4.2.3.4
2.7.1.3
3.1.1.4
2.3.1.240
3.1.1.7
4.1.3.43
1.14.11.69
3.7.1.2
2.5.1.19
2.7.10.2
6.3.2.4
6.3.2.3
3.1.4.11
3.6.1.45
3.1.3.2
2.1.1.68
3.1.8.2
1.14.19.2
1.14.11.17
4.1.1.2
5.4.2.4
3.1.3.2
3.2.2.1
2.4.1.10
2.4.2.6
3.2.1.14
5.6.2.1
5.4.2.6
1.1.1.42
4.2.1.11
3.4.11.9
3.4.24.69
5.4.2.12
3.5.1.19
3.4.24.-
3.1.21.2
3.1.5.1
3.6.1.1
5.6.2.1
3.1.11.3
3.1.3.7
3.1.3.11
3.6.4.10
5.3.1.5
4.2.2.2


DE-DE-H
DE-H-STY
DE-H-H-H
DE-H-K
DE-H-H
DE-H-R
DE-R-STY
DE-DE-DE
DE-DE-H-H
DE-DE-STY
DE-H-NQ
EC Class
Oxidoreductases (EC 1)
Transferases (EC 2)
Hydrolases (EC 3)
Lyases (EC 4)
Isomerases (EC 5)
Ligases (EC 6)
Translocases (EC 7)
Structural cluster
Number of templates from the same CATH
1
3
6
DE-DE-R
DE-DE-DE-STY
DE-K-STY
DE-DE-NQ
DE-DE-K
DE-H-X
STY-X-X
H-STY-STY
DE-DE-DE-DE
DE-DE-H-NQ
DE-DE-DE-H
H-H-STY
H-K-STY
DE-R-R
C-DE-H
DE-STY-STY
K-STY-STY
DE-DE-X
DE-K-K
DE-NQ-X
DE-X-X
DE-H-H-R
H-NQ-STY
K-R-STY
C-DE-STY
H-H-R
DE-K-NQ
C-H-NQ
DE-NQ-R
DE-DE-DE-DE-H-H
DE-DE-DE-DE-H
DE-DE-H-H-H
STY-STY-STY
C-H-STY
DE-DE-H-R
DE-DE-DE-K
DE-DE-DE-H-H
H-H-NQ
H-STY-X
DE-K-R
NQ-STY-X
C-H-H
R-R-R
NQ-STY-STY
DE-H-W
H-R-STY
DE-STY-X
C-C-C
DE-DE-NQ-R
K-STY-X
C-NQ-X
DE-DE-R-STY
DE-DE-DE-X
DE-H-NQ-X
C-DE-DE
X-X-X
DE-DE-H-STY
C-STY-STY
K-R-R
C-C-DE
DE-H-STY-X-X
DE-H-X-X
DE-H-STY-X
H-X-X
H-H-H-NQ
C-C-C-C
H-H-H-STY
H-STY-X-X
DE-H-H-STY
DE-DE-STY-STY
C-C-X
DE-F-VLI
H-H-H
C-X-X
DE-K-X
C-R-STY
K-X-X
C-C-H
DE-H-R-R
DE-DE-K-NQ
DE-DE-DE-H-R
H-NQ-R
K-NQ-STY
H-K-STY-STY
DE-DE-X-X
DE-H-H-NQ
C-X-X-X-X
NQ-R-STY
C-STY-X
DE-NQ-NQ-STY
DE-STY-W
DE-R-W
R-STY-STY
H-K-X
C-DE-R
A-X-X
H-K-NQ
R-X-X
STY-W-X
DE-DE-DE-H-NQ
STY-STY-X
DE-NQ-W
H-R-R
H-R-Z
DE-H-NQ-STY-X
H-H-H-H
DE-X-X-X
DE-NQ-STY
DE-DE-H-X
C-DE-H-NQ
DE-H-H-H-STY
DE-H-H-H-NQ
C-C-DE-H
DE-H-K-NQ
DE-STY-X-X
DE-H-R-STY
C-C-C-H
DE-DE-DE-H-H-H
C-H-H-STY
K-NQ-R
NQ-X-X
DE-R-X
DE-DE-K-STY


a
Count
1500
1250
1000
750
500
250
0
3
4
5
6
7
8
Number of template residues
1
2
3
4
5
6
7
E.C. class
750
500
50
0
3.40.50
3.20.20
3.40.640
1.10.600
3.40.720
2.60.120
2.60.40
1.20.210
3.30.470
3.60.20
3.60.21
3.40.630
3.50.50
3.30.930
3.90.226
CATH Topology
b
RMSD (Å)
2.5
2.0
1.5
1.0
0.5
0.0
Relationship
Convergent
Divergent
3
4
5
6
7
Number of template residues
c
Divergent
Convergent
Number or true hits
400
300
200
100
0
0
0.5
1
1.5
2
RMSD (Å)
Number of template residues
3
4
5
6
7
8


Catalytic residues
M-CSA
Atomic coordinates
PDBe
Protein Data Bank in Europe
Function / Ligand info
M-CSA
Co-factor database
UniProt
Rhea
KEGG
Catalytic site library
Template library
3 res
4 res
5 res
6 res
7 res
• • •
All vs. all template comparison
Template similarity network
Graph clustering
Clustered network
Residue functional atoms
GLY
ALA
VAL
LEU
ILE
MET
PHE
TYR
TRP
CYS
PRO
SER
THR
ASN
GLN
LYS
ARG
HIS
ASP
GLU
Main
Ptm
Annotation - Characterisation
Modules of enzyme catalysis
DE-DE-DE-H-H
C-C-C-H
DE-NQ-X
DE-DE-DE-H-N
DE-H-H-H-STY
DE-DE-NQ-R
• • •