The catalytic residues of an enzyme comprise the amino acids located in the active center responsible for accelerating the enzyme-catalyzed reaction. These residues lower the activation energy of reactions by performing several catalytic functions. Decades of enzymology research has established general themes regarding the roles of specific residues in these catalytic reactions, but it has been more difficult to explore these roles in a more systematic way. Here, we review the data on the catalytic residues of 648 enzymes, as annotated in the Mechanism and Catalytic Site Atlas (M-CSA), and compare our results with those in previous studies. We structured this analysis around three key properties of the catalytic residues: amino acid type, catalytic function, and sequence conservation in homologous proteins. As expected, we observed that catalysis is mostly accomplished by a small set of residues performing a limited number of catalytic functions. Catalytic residues are typically highly conserved, but to a smaller degree in homologues that perform different reactions or are nonenzymes (pseudoenzymes). Cross-analysis yielded further insights revealing which residues perform particular functions and how often. We obtained more detailed specificity rules for certain functions by identifying the chemical group upon which the residue acts. Finally, we show the mutation tolerance of the catalytic residues based on their roles. The characterization of the catalytic residues, their functions, and conservation, as presented here, is key to understanding the impact of mutations in evolution, disease, and enzyme design. The tools developed for this analysis are available at the M-CSA website and allow for user specific analysis of the same data.

Enzymes have evolved to catalyze biological reactions that are either too slow or too promiscuous to happen in the cytoplasm of the cell without help. Enzymes make these reactions faster by stabilizing the transition state structures or by providing an altogether different chemical path not available in solution. Either way, this results in a lower transition state energy (ΔG°), which translates to a faster turnover number, or catalytic rate (kcat) (1). Additionally, enzymes confine highly energetic species, which are commonly formed in the intermediate steps of these reactions, to the active center, limiting side reactions and driving activity toward the desired products. Our discussion concentrates on catalysis by proteins.

The subject of enzyme catalysis and particularly enzyme evolution can be approached from different angles. Much initial work in this field focused on single enzymes or, as examples and technical capabilities grew, enzyme subfamilies, but overlaying the rich informatic data available can greatly accelerate these investigations. The classification of extant enzymes in families according to their function, structure, and sequence reveals evolutionary links and suggests which features of the enzyme are associated with function conservation or the appearance of new functions. This is the general idea behind the classification of proteins as done by PFAM (2), CATH (3), or SCOP (4) or similar resources specifically tailored for enzymes, such as FunTree (5), SFLD (6), or EzCatDB (7). A complementary approach is to reconstruct the ancient sequences, so that the evolutionary mutation history can be explicitly analyzed (8). When it comes to the process by which enzymes evolve new functions, promiscuity is usually regarded as key. Either at the reaction or substrate level, promiscuity provides a pool of latent, but currently unused, functions. These functions can become useful through mutations that increase the activity of the enzyme toward a substrate or reaction or simply by exposing the enzyme to a new environment where an adequate substrate is present (9, 10). From a practical perspective, knowledge about enzymatic catalysis and evolution is the foundation for enzyme design, be it with methods of directed evolution (11, 12), rational design (13), or de novo protein design (14).

Here, we provide a systematic overview of the current knowledge of enzymatic catalysis from the point of view of the catalytic residues. We use the Mechanism and Catalytic Site Atlas (M-CSA)3 (15) database as a manually curated data set of catalytic residues and other mechanistic information. We start by considering the frequency of the 20 amino acid types as catalytic residues and whether they act through their side chain or backbone. This analysis highlights the most common catalytic amino acids and their general chemical properties (e.g. charge, size, hydrophobicity, or aromaticity), and, using catalytic propensity, we infer how selection pressure is different in catalytic

---

3 The abbreviations used are: M-CSA, Mechanism and Catalytic Site Atlas; MACiE, Mechanism, Annotation, and Classification in Enzymes; CSA, Mechanism and Catalytic Site Atlas; EMO, enzyme mechanism ontology; PLP, pyridoxal phosphate.
residues when compared with the rest of the sequence. The second part of the analysis focuses on the catalytic functions and on which amino acids are responsible for those functions, revealing the specificity of each amino acid for each catalytic function. For reactant roles (see Fig. 1), another layer of analysis is possible based on the chemical group with which the amino acid reacts. Finally, we focus on the conservation of the catalytic residues in Swiss-Prot homologous sequences. In this last section, we integrate all of the previous analysis, by checking how conservation is affected by each amino acid type and function. In particular, we investigate factors that determine conservation and how conservation changes if we look at homologues that have the same function, the same catalytic function but on a different substrate, different enzymatic functions, or nonenzymatic proteins (i.e. pseudoenzymes (17)).

Approximately half of catalytic reactions require one or more cofactors for successful completion. Cofactors actively assist the catalytic amino acids in enzyme catalysis, performing roles that are complementary to amino acids (18). However, it is not possible to directly compare them with the roles of amino acids, as the modes of analysis required for amino acids and cofactors are distinct. For example, the conservation and propensity scores, as calculated here, are not applicable for cofactors. Additionally, cofactors are much more diverse than amino acid
acids, necessitating the development of suitable systems for classification of these entities. For that reason, they will not be considered further here.

Together with increasing our knowledge of biocatalysis, mechanistic and catalytic residue information can be important for the identification of function in nonannotated structures and sequences and for the prediction of the effect of mutations of the catalytic residues. Currently, function prediction for uncharacterized protein sequences hinges on the identification of close homologues with known function. However, the attribution of functional annotation based solely on homology can be problematic, especially because we are interested in homologues with different functions (19, 20). For these challenging cases, a systematic knowledge of the possible roles of the catalytic residues and their mutation profile, as presented here, can be valuable. Specifically, this knowledge can go beyond expectations based on general themes of amino acid function derived from individual studies by providing more contextual information and pointing toward potential roles that might be unexpected. Additionally, by suggesting the effect of mutations in protein function, the data are also useful to understand how these variations in sequence shaped evolutionary past and current genetic variation and what mutations should be explored in enzyme design projects. In this review, we therefore endeavor to explore known trends and themes with a global lens and to demonstrate how a carefully curated database such as the M-CSA can provide a valuable starting point for a variety of scientific exploits.

**Mechanism and Catalytic Site Atlas**

We have previously developed and currently maintain a manually curated database of catalytic residues and enzyme mechanisms called the Mechanism and Catalytic Site Atlas (M-CSA) (15), which can be accessed at [www.ebi.ac.uk/thornton-srv/m-csa/](http://www.ebi.ac.uk/thornton-srv/m-csa/). M-CSA was built upon the CatRes data set (21) and two previous databases: Mechanism, Annotation, and Classification in Enzymes (MACiE), a database of enzyme mechanisms (22, 23), and Catalytic Site Atlas (CSA) (22, 23), a database of catalytic sites (24, 25). M-CSA contains two types of entries with two different levels of annotation. All of the 964 entries list the catalytic residues of the enzyme and a broad description of their role in the mechanism. For a subset of 684 entries, the database also contains detailed information about the mechanistic steps of the reaction, including a pictorial curly arrow representation and the role of the residues in each step (Figs. S1 and S2 show how the two levels of annotation are presented on the website). Since the last published update of M-CSA (15), we have extended the number of entries with a detailed mechanism description from 423 to 684 (and from 280 in the previous analysis paper (26)). In this review, we look at these 684 entries with curated mechanistic information and functional description. Furthermore, we ignore mechanism proposals that we judged to be less well-validated (i.e. labeled in the database with less than three stars). This brings the total number of analyzed mechanisms to 648.

The annotation of the residue roles in M-CSA is done using an ontology specifically developed for annotating enzyme mechanisms, the enzyme mechanism ontology (EMO) (25). Fig. 1 (top) shows all of the catalytic functions that we capture and how these are organized in EMO. Catalytic roles related to the breaking and formation of bonds are grouped under “Reactant Roles,” whereas other catalytic roles are grouped under “Spectator Roles.” “Interaction Roles” describe how the catalytic residues interact with other molecular species in the active site, although the interaction roles themselves do not necessarily describe a catalytic function. For the 648 mechanisms analyzed in this review, we annotate the function of 3,351 residues across 3,373 mechanism steps, which gives an average of 5.2 residues/ enzyme as well as 5.2 steps/mechanism. There are a total of 19,586 residue-role associations in these steps, 8,041 if each association is only counted once for each mechanism. Additional details about the annotation and data processing are available in the website documentation and references given therein.

All of the data presented in this analysis about the individual catalytic residues and their functions were manually extracted from the over 3,000 published papers. Each enzyme page in M-CSA contains links to the relevant papers used to curate that entry.

While preparing this analysis, we have developed new tools in M-CSA relating to the catalytic site annotation and devised ways to represent these on the website. (a) There is now code that detects all of the reactant functions directly from the curly arrow schemes, according to the schematics in Fig. 1 (bottom), which will make future curation easier and more consistent. (b) Mechanism files are now parsed and interpreted from a chemical point of view, using SMARTS patterns, as generated by RD-Kit (25). SMARTS is a language to describe chemical groups or molecules in a consistent and compact manner, which facilitates computational handling of these data. (c) We have added a new feature on the website that generates plots using up-to-date M-CSA data and where custom filters can be applied (available at [www.ebi.ac.uk/thornton-srv/m-csa/stats/custom-plots](http://www.ebi.ac.uk/thornton-srv/m-csa/stats/custom-plots)). We have used this feature to generate most of the plots shown in this review. (d) A newly developed public API is now the preferred way to access up-to-date M-CSA data in bulk, as opposed to flat files. We have created a Polymer web component ([www.polymer-project.org](http://www.polymer-project.org)) that can be added to any website to show the catalytic mechanism of specific enzymes. Information about the API and the web component can be found at [www.ebi.ac.uk/thornton-srv/m-csa/download](http://www.ebi.ac.uk/thornton-srv/m-csa/download).

**Previous data sets and applications**

The data sets derived from CatRes, MACiE, and CSA, from which M-CSA evolved, have been used in the past to study enzymes from different perspectives. Bartlett et al. (27) showed how enzymes evolved new functions by keeping most of the mechanism and catalytic machinery intact, while changing the individual steps of the reaction. The integration of MACiE data with EC classification (28), the CATH structural classification (3), and reaction similarity analysis with EC-Blast (29) was also used to explain the evolution of ligases (30). In another exam-

---

4 Please note that the JBC is not responsible for the long-term archiving and maintenance of this site or any other third party hosted site.
ple, the chemical components of mechanisms, as annotated in MACIE (and now in M-CSA), were used to reconstruct the evolutionary history of enzymatic catalysis (31).

CSA has been used as a curated source of catalytic residues for diverse purposes. Some examples include using these residues as a reference set to benchmark a method that predicts functionally important residues based on sequence (32); as both training and test sets of a machine learning method to identify catalytic residues based on sequence and structural data (33); or simply as a list of catalytic residues to annotate the results of a structure prediction tool (34). In a recent example using the M-CSA data set (among other data sources), the authors created a method to predict the effect of mutations in specific residue positions (35).

**Previous analyses of catalytic residues**

Analyses of catalytic residues from a broad set of enzyme families are surprisingly rare in the literature, and they are mostly linked to the M-CSA parent databases. An exception to this is an early analysis of 37 catalytic residues belonging to 17 enzymes (37), in which His, Asp, and Glu were found to be the most common catalytic residues, followed by Lys and Arg. For this data set and a list of under 100 homologues, all of the catalytic residues were perfectly conserved. Despite the low number of examples, the results obtained have been validated using the larger data sets below.

Other studies on the functions of catalytic residues have been performed using the CatRes, MACIE, and CSA data sets. The CatRes study (21) included 178 enzymes with 615 catalytic residues. The reported frequency and catalytic propensity of the catalytic residues are remarkably similar to those shown here, despite the smaller data set. Catalytic residues were found more commonly in loop regions than expected, and with less frequency in α-helices. They were also less exposed to the solvent compared with other protein residues and are less flexible, as measured by the β-factors. The overall conservation was shown to be higher for catalytic residues and the surrounding residues. With respect to the catalytic functions, “acid base,” “transition state stabilizer,” and “activation roles” were found to be the most common. Two more recent analyses using the MACIE data set (26, 36) (at the time containing 223 and 280 mechanisms, respectively) used a more detailed controlled vocabulary for the annotation of the residue roles and explicit description of the reaction steps. These improvements revealed a matrix of residue-role associations more similar to the current one. The slightly larger data set also allowed for the separate analysis of the EC classes, which showed differences in the frequency and the role associations of the catalytic residues. Nevertheless, the identification and interpretation of the chemical constraints that influence residue specificity proved harder to systematize.

The data presented in this review complement previous results by means of additional data and modes of analysis. The main improvements are summarized in the following points: (a) with respect to previous versions of M-CSA (or MACIE and CSA), we have updated every entry in the database to include new literature and mechanism proposals; (b) the number of annotated mechanisms has more than doubled since the last similar analysis using MACIE (from 280 to 684), which makes the residue-roles associations more robust and reveals some associations not previously detected; (c) instead of using the EC classification as a proxy for different types of reactions, we now use an explicit chemical description of the reactant groups, using SMILES annotation, allowing for a better identification of residue specificity rules, such as in Fig. 3; (d) this is the first time the catalytic functions have been analyzed following the adoption of the EMO (see Fig. 1), which harmonizes CSA and MACIE annotation and provides a more detailed specification of residue roles compared with previous studies; (e) exploration of the associations between conservation and function via conservation analysis was done for the first time.

**Amino acid frequency and catalytic propensity**

We start by surveying the overall frequency by which the 20 standard amino acids appear as catalytic residues. Fig. 2 shows how frequently each amino acid type has a catalytic role while acting through either its side chain (top left) or main chain (bottom left). Most residues perform their catalytic function through the side chain (89% of the residues act through the side chain). Histidine, followed by the other negatively and positively charged residues, is the most common catalytic residue. These are followed by amino acids with polar and aromatic groups and finally by hydrophobic amino acids. The functional analysis we present below clarifies why some amino acids are more commonly catalytic over others.

Although every residue has a backbone (with the same chemistry), not every residue main chain is equally likely to be catalytic. This ranking is led by glycine, followed by the other three smallest amino acids, Ser, Ala, and Cys, a hint that accessibility and flexibility are important for the catalytic roles performed by the main chain. It is interesting to note that in the case of Cys, Ser, and Thr, where the main-chain atoms have catalytic roles, a significant number of the corresponding side chains are also involved in the reaction (26 of 29, 21 of 38, and 7 of 21, respectively), an indication that the side chain rather than the backbone of these amino acids is probably driving selection.

Another way to look at residue frequency is to use catalytic propensity. This value is the ratio between the residue’s frequency among all catalytic residues and its frequency in the overall protein. The functions that catalytic residues need to perform are distinct from the functions of other amino acids. Therefore, residues that can perform these catalytic functions but are not as common elsewhere in the protein have high catalytic propensities. Fig. 2 (top right) shows that His appears in the catalytic site 8.2 times more often than expected when compared with its baseline frequency, whereas Cys is 4.6 times more common. Other charged residues are commonly catalytic, but because they are more frequent in the sequence, they have lower catalytic propensities (from 1.7 to 3.0). Hydrophobic residues have very low catalytic propensities because they are rarely catalytic, and some of them, especially the smaller ones, are very common in the protein sequence.

The catalytic propensities of residues that act through their main chain are shown in Fig. 2 (bottom right). Of the four small residues identified in the overall frequency plot, only three (Gly, Cys, and Ser) have catalytic propensities significantly higher than 1: 3.5, 5.8, and 1.8, respectively. Ala is as common in the
The catalytic site as in the sequence, so its catalytic propensity is close to 1.

**The catalytic roles and residue-role associations**

In the same way catalysis is dominated by a handful of amino acids, some catalytic functions are also much more common than others. Fig. 3 shows the frequency of residue roles in catalysis in our data set divided across the three main function groups, reactant (Fig. 3A), spectator (Fig. 3B), and interaction (Fig. 3C), as performed by residues that act through the side chain. There are 12 roles that occur more than 100 times in the data set. Among reactant functions (see Fig. 2), the three proton transfer roles together with nucleophile and nucleofuge (a leaving group that keeps the lone pair of electrons) account for more than 90% of the annotations (2,269 of a total of 2,482), 93% if we include the remaining heterolytic roles electrophile and electrofuge (a leaving group that does not keep the lone pair of electrons). Single and electron pair roles are much less frequent, representing 6% of the data set, and radical chemistry involving catalytic residues accounts for less than 1% of all of the annotations. There are only two cases of hydride transfers to catalytic residues, both associated with the cleavage of a disulfide bond.

For spectator roles, a single role, electrostatic stabilization, accounts for 60% of the cases (1,316 of a total of 2,199), whereas the combination of activator roles amounts to 29% and steric interactions to 8%. Finally, interaction roles, for which there are 2,559 annotations, are dominated by hydrogen bond interactions and metal binding.
Figure 3. A–C, number of occurrences of the catalytic functions in M-CSA grouped by type. Functions are only counted once for each enzyme and catalytic residue combination. D, amino acid-role association matrix for “reactant” catalytic roles. Totals for each amino acid and function are shown in green. Other squares are colored according to frequency, from very common (dark red) to less common (light yellow) and not observed (white). E, number of residues in M-CSA that have the nucleophile/nucleofuge roles. F, nucleophilic amino acids and their electrophilic chemical group counterparts. Electrophilic groups observed fewer than five times are grouped under Other. The first atom in the SMARTS pattern is the reaction center. A pictorial key of these electrophilic groups is given in Fig. S5. Only functions performed through the side chains are considered in all plots.
There is an important caveat regarding the annotation of these roles, which makes direct comparisons between reactant functions and the other two groups difficult. Each reaction step is defined by the bonds that are being formed or broken. For each curly arrow in the mechanism files, the respective reactant functions can be unambiguously identified. This is done both by a process of automatic annotation extraction and manual checking by a curator. Most spectator and interaction roles are different because the attribution of functions is more subjective. For example, some researchers may consider an electrostatic stabilization catalytic role for a residue that is ignored by others (a consistent way to do this would be to calculate the electrostatic contribution of every residue to catalysis, but that is currently impractical). For this reason, we only annotate a function like electrostatic stabilization if it is explicitly mentioned in the literature.

Having established which catalytic roles are more frequent, we now turn to identifying catalytic residues that are responsible for each role and establishing how specific these residue-role associations are. Understanding the specificity of each role is necessary if we want to predict the impact of mutations on the function of the protein. Fig. 3D shows which residues catalyze each “reactant” role and how often they occur. Please note that this plot does not include “spectator” or “interaction” roles, so only amino acids that catalyze reactant roles are included. For brevity, similar heat maps for spectator and interaction functions are not discussed, but they are shown in the supporting material (Figs. S3 and S4) and on the website. Because proton acceptor and proton donor are the most common roles, the most common catalytic residues are also the ones that perform these two functions. His, Glu, and Asp, in particular, are responsible for more than 60% of the proton transfers that involve catalytic residues. Whereas the eight nucleophilic residues are all associated with proton transfers, the relative frequency of occurrence is distinct. Cys, Lys, and Ser are collectively the nucleophiles in 78% of all of the cases in the data set (see the “Residue specificity to interacting chemical groups” for specificity among nucleophiles). Lys is the only amino acid annotated as an electron pair acceptor/donor. This role is commonly associated with the formation of a double bond between Lys and the pyridoxal phosphate (PLP) cofactor. This is the only case where two electrons are transferred from the residue to another chemical species. We do not explicitly annotate the movement of pairs of electrons if it happens within the same molecule. Roles associated with single-electron transfers can be performed by 13 amino acids. Most of these annotations happen in single-electron transport chains, where the electron is transferred from one part of the protein (usually from a metal) to the substrate. Residues in these electron chains are annotated with the three single-electron roles (acceptor, donor, and relay), and they do not hold the single electron in a stable configuration, which is a more specific function. The only residues that do become stable radicals are Cys, Tyr, and Trp. The remaining roles can only be performed by a limited set of residues; Cys, Glu, and Asp can be electrophiles and electronductors, whereas Cys, Tyr, and Gly react with hydrogen radicals. Only Cys is observed to perform radical reactions with heavy atoms and hydride reactions.

**Residue specificity to interacting chemical groups**

The previous plot can answer some questions regarding the specificity of the catalytic roles. In particular, it suggests that a catalytic residue mutation responsible for a specific function that lands on an empty square in the heat map is deleterious to the enzyme function. The next question is what happens when the mutation results in a residue that can do the same function. For example, what is the effect on enzyme function if a nucleophile is mutated to another nucleophilic residue? To answer this question, we must look at the chemical species that is the target of that catalytic function. For the nucleophilic residues, which will be the electrophile receiving the shared electrons; for a proton acceptor, it will be the proton donor, for example. Every reactant role has a corresponding target role, as can be seen in Fig. 1 (bottom). In this section, we have chosen the nucleophile role for a more detailed analysis, but the same plots can be recreated on the website for other reactant roles.

Fig. 3E shows how frequently each amino acid has nucleophilic activity in our data set. Only 8 amino acids are observed to be nucleophiles, with Cys, Ser, and Lys being responsible for almost 80% of the annotations. The distribution is markedly different from the overall amino acid count as shown in Fig. 2. To better understand why evolution favored this distribution and how specific each amino acid is, we analyzed the data for all of the nucleophilic attacks to identify their electrophilic counterpart. The results are shown in Fig. 3F.

When looking at the target chemical groups, some specificity rules become obvious. Cys and Ser are overwhelmingly common as the nucleophiles that attack amide bonds, Lys is very specific to the different PLP chemical environments that arise when PLP binds to other molecules covalently. Cys is the only amino acid that can form bonds with other sulfur compounds. Phosphates are a very promiscuous electrophile, being the targets of six amino acids. Phosphate is also the only chemical group that is nucleophilically attacked by His. Asp and Glu seem to be well-equipped to attack acetals, as seen often in glycosidic bonds. Cys, which was the most promiscuous amino acid in terms of the possible roles it performs, is also able to nucleophilically attack the most diverse set of chemical groups. Unfortunately, for most other chemical groups, the data are too sparse to offer any conclusions. It is impossible to say for sure whether the lack of examples reflects lack of data, lack of annotation, low biological abundance for this chemical reaction, or real lack of activity. This comment also applies to other less frequent reactant roles.

**Conservation of catalytic residues in Swiss-Prot homologues**

Homologues (defined as evolutionary relatives) were found using sequence similarity searches of Swiss-Prot. Multiple-sequence alignments of each set of homologues, one for each enzyme in M-CSA, were used to check the conservation of the catalytic residues. We also consider a more restrictive subset of Swiss-Prot during the analysis that includes only the entries for which there is functional experimental evidence. The conser-
vation values for each catalytic residue in a family correspond to the percentage of the homologues of each enzyme where the residue is conserved. The conservation for groups of residues (belonging to several enzyme families) is the average of each catalytic residue conservation, where all of the residues have the same weight. More details about the homology search and the conservation calculation can be found in the supporting material.

Homologues are divided into four different groups, with respect to the reference enzyme in M-CSA. Below, we give examples of proteins homologous to *Lactobacillus delbrueckii* D-lactate dehydrogenase (UniProtKB P26297, EC 1.1.1.28). The four groups are as follows.

(a) Homologues that share at least one complete EC number are grouped under homologues that catalyze the “same reaction.” These are usually the same enzyme in related species (orthologous enzymes), such as D-lactate dehydrogenase in *Escherichia coli* (UniProtKB P52643, EC 1.1.1.28).

(b) Homologues that share one EC subclass (third level EC), but not a complete EC, are considered to be homologues that perform the same reaction on “other substrates,” such as glyceraldehyde dehydrogenase (UniProtKB P36234, EC 1.1.1.29).

(c) Homologues that are enzymes but do not share any EC sub-class are considered to be catalyzing “other reactions,” as is the case of phosphonate dehydrogenase (UniProtKB O69054, EC 1.20.1.1).

(d) Homologues without an EC number are considered to be nonenzymes/pseudoenzymes. For example, C terminus–binding protein 1 (UniProtKB Q20595) binds DNA and regulates gene expression, but it does not require dehydrogenase activity (38). Other well-known examples of pseudoenzymes are reviewed elsewhere (39).

The overall conservation for catalytic residues across all homologues in our data set is 84.2% (83.2% for the experimental evidence data set). Conservation of catalytic residues is highest among homologues that catalyze the same reaction, 93.7% (93.4%), followed by enzymes that perform the same reaction on different substrates, 80.4% (80.5%), and then enzymes that catalyze a different reaction, 64.7% (62.8%). Finally, conservation is lowest in non-enzymatic homologues/pseudoenzymes, where on average only 50.3% (49.2%) of catalytic residues are conserved.

These results are intuitive; larger differences in function require more changes in the catalytic machinery. We note that even when the function is identical (i.e. the same chemical reaction on the same substrate), the catalytic machinery is not always conserved.

Conservation is still relatively high for the second group: enzymes that catalyze the same reaction on a different substrate. In theory, only binding residues should change in these cases, and catalytic residues should be strictly conserved. In practice, many binding residues are also catalytic, and small changes to the catalytic residues may be needed to accommodate a different substrate, so the slightly lower conservation score is expected.

For enzymes that catalyze a different reaction, more mutations are required, but almost two-thirds of the catalytic residues are still conserved. This suggests that the evolution of new enzyme reactions is done with only slight changes to the catalytic residues and the underlying catalytic functions. In fact, previous studies have shown that new enzyme functions are commonly evolved by changing some of the chemical steps the enzyme catalyzes along the reaction or by rearranging them in a different order (26). For this to be possible, residues involved in the intact steps need to be conserved.

The interpretation of the results for pseudoenzymes is more nuanced. On the one hand, loss of catalytic function can be achieved by means other than changes in the catalytic residues, such as mutations in the binding pocket or in allosteric regions. This means some pseudoenzymes can have a perfectly conserved catalytic machinery, despite their catalytic inactivity. On the other hand, when loss of function occurs, the evolutive pressure to keep the catalytic residues in place disappears, so mutations can accumulate freely. Finally, for pseudoenzymes to be retained, they must keep or acquire a noncatalytic function, which is commonly associated with the ability to bind the original substrate (40) and thus requires the conservation of the catalytic residues that are essential for substrate binding. The observed conservation values, which, at 50%, are lower than any of the other enzyme homologue groups, indicate that the effect of loss of selective pressure for catalytic residues is significant. When looking at active sites as a whole, we find that only 24.7% of the pseudoenzymes with 2 or more residues have all of the catalytic residues conserved.

**Conservation by amino acid type and role**

Fig. 4 (top) shows the conservation of catalytic residues across homologues that catalyze the same function. The data are divided by amino acid type and residue role, to show how these affect conservation. The figure only includes the six most common catalytic functions, but the same plot can be generated for all functions and homologue sets using the website. Figures with the conservation values for the same six functions and different sets of homologues are included in the supporting information (Fig. S6). Conservation in homologues that catalyze the same EC number is high throughout all functions and amino acids, but some patterns can be observed. The better-conserved residues have higher catalytic propensities, in particular for Cys and all of the charged residues. Figs. S7–S10 illustrate the correlation between conservation and catalytic propensity for different sets of homologues (the correlation is clearly lost for pseudoenzymes). The distribution of conservation along catalytic functions is suggestive. The most conserved functions are the ones performed by a smaller subset of residues and also the ones that perform more specific chemical roles (such as nucleophile, metal ligand, and activator as opposed to proton acceptor, hydrogen bond donor, and electrostatic stabilizer). Conservation for combinations of functions and amino acids follows the same general trend.

Fig. 4 (bottom) shows how conservation changes for the same six functions when looking at different sets of homologues. Homologues with the same EC number are well-conserved. For enzymes with a different fourth-level EC, there is a generalized small fall in conservation, but changes are not uniform across functions. Nucleophilic residues remain very well-conserved, whereas there is a bigger drop for all of the other roles, suggesting that to catalyze the same reaction in a different substrate, some mutations are tolerated except for very specific roles. For
enzymes that catalyze an altogether different reaction, some reaction steps may have been lost, together with residues involved in these steps, and hence the observed big drop in the conservation of nucleophiles, proton acceptors, and activators. The nonenzyme case shows very clearly the loss of evolutionary pressure toward keeping the reactant functions in place. Nucleophilic residues are the least conserved in pseudoenzymes, among the six roles shown. The relatively higher conservation of metal-binding residues can be attributed to binding roles that these proteins may still be performing.

Conclusions

The data presented here and available on the M-CSA website demonstrate how enzymes can catalyze thousands of reactions with a limited set of amino acids performing a small set of functions. Part of the challenge of understanding how this chemical diversity is possible is to create tools that summarize the data without ignoring its complexity. We have tried to tackle this problem by creating a series of plots that cover different levels of detail, from general statistics, such as the frequency and propensity of the amino acids (Fig. 2), to more specific data, such as the chemical groups of the substrate that interact with nucleophiles (Fig. 3F). Although it is not possible to cover here all of the catalytic functions at that level of detail, we hope to have shown that the same analysis and reasoning can be extended to them and easily performed using the M-CSA website.

The results under “Conservation by amino acid type and role” show the potential to use mutations in catalytic residues to identify evolutionary events of change and loss of function. The ability to do this could bring significant improvement to the functional annotation of uncharacterized proteins in UniProtKB by adding a discriminatory power that methods based solely on global homology cannot have. In particular, we envisage the creation of a machine learning method to identify change or loss-of-function events based on these data.

The sample of enzyme mechanisms available in M-CSA provides good but not exhaustive coverage of the EC space, as can be appreciated in Table 1 for the seven EC classes. Coverage is calculated both at the fourth (complete EC number) and third (sub-subclass) levels of the EC classification. EC sub-subclasses can be roughly understood as types of chemical reactions, with the fourth level distinguishing between different substrates undergoing the same reaction. For this reason, enzymes within the same sub-subclass commonly share the same mechanism, and the coverage at this level of the EC classification is more informative than the coverage at the complete EC number. Overall, M-CSA covers more than 60% of all sub-subclasses. Coverage is particularly lacking for EC 1 and EC 7 (45.6 and 30% coverage, respectively, although there are only 10 sub-subclasses belonging to EC 7). This representation bias partially reflects our curation backlog but also reflects the fact that some of these mechanisms have not been solved yet. The last two columns of Table 1 show the total numbers and coverage for EC numbers that can be associated with a PDB structure. This filter excludes reactions for which the mechanism is probably not known in the literature, because a structure is typically necessary to propose one. By this measure, the M-CSA coverage of EC space increases to almost 80%. We aim to update M-CSA in the coming year with mechanisms recently described in the literature.

We do not expect that the prevalence of common residues and roles will change significantly as more mechanisms are
Table 1

Coverage of EC space by M-CSA entries with a detailed mechanism

The subset of EC numbers associated with at least one PDB structure is also considered. The M-CSA database contains an additional 280 entries with catalytic site annotation but not detailed mechanism description. These entries are not considered throughout this review. The figures for EC coverage including these entries can be seen on the website.

| Top level code | ECs in M-CSA | Number of ECs | M-CSA coverage of ECs | ECs with PDB structure | M-CSA coverage of ECs with PDB structure |
|----------------|--------------|---------------|-----------------------|------------------------|----------------------------------------|
|                |              |               |                       |                        |                                        |
| **Third-level EC** |              |               |                       |                        |                                        |
| EC 1. Oxidoreductases | 67           | 147           | 45.6                  | 101                    | 66.3                                   |
| EC 2. Transferases | 30           | 39            | 76.9                  | 34                     | 88.2                                   |
| EC 3. Hydrolases | 47           | 66            | 71.2                  | 51                     | 92.2                                   |
| EC 4. Lyases | 17           | 17            | 100.0                 | 17                     | 100.0                                  |
| EC 5. Isomerases | 17           | 19            | 89.5                  | 19                     | 89.5                                   |
| EC 6. Ligases | 10           | 12            | 83.3                  | 11                     | 90.9                                   |
| EC 7. Translocases | 3            | 10            | 30.0                  | 7                      | 42.9                                   |
| **Total** | 191          | 310           | 61.6                  | 240                    | 79.6                                   |
| **Fourth-level EC** |              |               |                       |                        |                                        |
| EC 1. Oxidoreductases | 129          | 1888          | 6.8                   | 700                    | 18.4                                   |
| EC 2. Transferases | 123          | 1896          | 6.5                   | 826                    | 14.9                                   |
| EC 3. Hydrolases | 161          | 1305          | 12.3                  | 752                    | 21.4                                   |
| EC 4. Lyases | 91           | 703           | 12.9                  | 311                    | 29.3                                   |
| EC 5. Isomerases | 57           | 302           | 18.9                  | 156                    | 36.5                                   |
| EC 6. Ligases | 27           | 216           | 12.5                  | 109                    | 24.8                                   |
| EC 7. Translocases | 3            | 83            | 3.6                   | 26                     | 11.5                                   |
| **Total** | 591          | 6,393         | 9.2                   | 2,880                  | 20.5                                   |

added to the database, but it is likely that parts of our analysis relative to less represented roles, and especially those associated with the oxidoreductases, such as radical or electron transfer chemistry, will improve with better coverage.

How enzymes have evolved and how they are able to catalyze such a broad set of chemical reactions remain two of the most important questions in the field of protein science. By collecting data about catalytic mechanisms in M-CSA, we hope to provide open and machine-readable information to be used by other researchers or in studies like the one presented here. Improved understanding will help in the design of enzymes with novel functions, using evolutionary paradigms to improve activity.

The current analysis can be improved in a number of ways. The first would be the inclusion of structural information. At the moment, our analysis disregards any structural considerations, such as possible compensatory mutations that happen close to a mutated residue. Also, we are limited here by the reductionist approach of considering each residue as an independent unit. Enzymatic reactions require the participation of several residues, substrates, and cofactors interacting across multiple chemical steps (in three-dimensional space). Because selection acts at this reaction level (reactions can be understood as the required chemical phenotype, where it does not matter how they are catalyzed, as long as they happen), an integrated analysis needs to be implemented to effect a more complete understanding of the catalytic process. An essential component of such an integrated view would be the characterization of the mutations of catalytic residues in evolutionary protein phylogenetic trees, such that the effect of mutations can be associated with changes of function.

Acknowledgments—We thank Amelia Brasnett, Morwenna Hall, Charity Hornby, and James Willey for help in curating the last version of M-CSA and Roman Laskowski for helpful discussions. We thank the scientific editor of JBC, Catherine Goodman, for extensive suggestions on improving the clarity of the review.

References

1. Nelson, D. L., and Cox, M. M. (2017) Lehninger Principles of Biochemistry, 7th Ed., pp. 246–269, W. H. Freeman, New York
2. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., et al. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432 CrossRef Medline
3. Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., Tolulope, A., Scholes, H. M., Senatorov, I., Bujan, A., Ceballos Rodriguez-Conde, F., Dowling, B., Thornton, J., and Orenco, C. A. (2019) CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.* 47, D280–D284 CrossRef Medline
4. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 42, D310–D314 CrossRef Medline
5. Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Laskowski, R. A., Orenco, C. A., and Thornton, J. M. (2012) Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Comput. Biol.* 8, e1002403 CrossRef Medline
6. Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Norris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., et al. (2013) The structure-function linkage database. *Nucleic Acids Res.* 42, D521–D530 Medline
7. Nagano, N., Nakayama, N., Ikeda, K., Fukuie, M., Yokota, K., Doi, T., Kato, T., and Tomii, K. (2015) ExCatDB: the enzyme reaction database, 2015 update. *Nucleic Acids Res.* 43, D453–D458 CrossRef Medline
8. Hochberg, G. K. A., and Thornton, J. W. (2017) Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* 46, 247–269 CrossRef Medline
9. Copley, S. D. (2015) An evolutionary biochemist’s perspective on promiscuity. *Trends Biochem. Sci.* 40, 72–78 CrossRef Medline
10. Khersonsky, O., and Tawfik, D. S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* 79, 471–505 CrossRef Medline
11. Renata, H., Wang, Z. J., and Arnold, F. H. (2015) Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution. *Angew. Chem. Int. Ed. Engl.* 54, 3351–3367 CrossRef Medline
12. Zeymer, C., and Hilvert, D. (2018) Directed evolution of protein catalysts. *Annu. Rev. Biochem.* 87, 131–157 CrossRef Medline
**JBC REVIEWS: Function and conservation of catalytic residues**

13. Ebert, M. C., and Pelletier, J. N. (2017) Computational tools for enzyme improvement: why everyone can—and should—use them. *Curr. Opin. Chem. Biol.* 37, 89–96 CrossRef Medline

14. Huang, P.-S., Boyken, S. E., and Baker, D. (2016) The coming of age of de novo protein design. *Nature* 537, 320–327 CrossRef Medline

15. Ribeiro, A. J. M., Holliday, G. L., Furnham, N., Tyzack, J. D., Ferris, K., and Thornton, J. M. (2018) Mechanism and catalytic site atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* 46, D618–D623 CrossRef Medline

16. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D618–D623 CrossRef Medline

17. Murphy, J. M., Mace, P. D., and Eyers, P. A. (2017) Live and let die: insights into pseudoenzyme mechanisms from structure. *Curr. Opin. Struct. Biol.* 47, 95–104 CrossRef Medline

18. Fischer, J. D., Holliday, G. L., Rahman, S. A., and Thornton, J. M. (2010) The structures and physicochemical properties of organic cofactors in biocatalysis. *J. Mol. Biol.* 403, 803–824 CrossRef Medline

19. Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* 5, e1000605 CrossRef Medline

20. Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., et al. (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227 CrossRef Medline

21. Bartlett, G. J., Porter, C. T., Borkakoty, N., and Thornton, J. M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* 324, 105–121 CrossRef Medline

22. Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O’Boyle, N. M., Murray-Rust, P., Thornton, J. M., and Mitchell, J. B. O. (2005) MACiE: a database of enzyme reaction mechanisms. *Bioinformatics* 21, 4315–4316 CrossRef Medline

23. Holliday, G. L., Andreini, C., Fischer, J. D., Rahman, S. A., Almonacid, D. E., Williams, S. T., and Pearson, W. R. (2012) MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.* 40, D783–D789 CrossRef Medline

24. Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* 32, D129–D133 CrossRef Medline

25. Furnham, N., Holliday, G. L., de Beer, T. A., Jacobsen, J. O. B., Pearson, W. R., and Thornton, J. M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* 42, D485–D489 CrossRef Medline

26. Holliday, G. L., Fischer, J. D., Mitchell, J. B. O., and Thornton, J. M. (2011) Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *FEBS J.* 278, 3835–3845 CrossRef Medline

27. Bartlett, G. J., Borkakoty, N., and Thornton, J. M. (2003) Catalysing new reactions during evolution: economy of residues and mechanism. *J. Mol. Biol.* 331, 829–860 CrossRef Medline

28. McDonald, A. G., and Tipton, K. F. (2014) Fifty-five years of enzyme classification: advances and difficulties. *FEBS J.* 281, 583–592 CrossRef Medline

29. Rahman, S. A., Cuesta, S. M., Furnham, N., Holliday, G. L., and Thornton, J. M. (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods* 11, 171–174 CrossRef Medline

30. Holliday, G. L., Rahman, S. A., Furnham, N., and Thornton, J. M. (2014) Exploring the biological and chemical complexity of the ligases. *J. Mol. Biol.* 426, 2096–2111 CrossRef Medline

31. Nath, N., Mitchell, J. B. O., and Caetano-Anollés, G. (2014) The natural history of biocatalytic mechanisms. *PLoS Comput. Biol.* 10, e1003642 CrossRef Medline

32. Capra, J. A., and Singh, M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics* 23, 1875–1882 CrossRef Medline

33. Miller, M., Vitale, D., Kahn, P. C., Rost, B., and Bromberg, Y. (2019) funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res.* 47, e142 CrossRef Medline

34. Song, J., Li, F., Takemoto, K., Haffari, G., Akutsu, T., Chou, K.-C., and Webb, G. I. (2018) PReval, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *J. Theor. Biol.* 443, 125–137 CrossRef Medline

35. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858 CrossRef Medline

36. Miller, M., Vitale, D., Kahn, P. C., Rost, B., and Bromberg, Y. (2019) funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res.* 47, e142 CrossRef Medline

37. Holliday, G. L., Mitchell, J. B. O., and Thornton, J. M. (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.* 390, 560–577 CrossRef Medline

38. Zvelebil, M. J. J. M., and Sternberg, M. J. E. (1988) Analysis and prediction of the location of catalytic residues in enzymes. *Protein Eng.* 2, 127–138 CrossRef Medline

39. Chen, S., Whetstone, R. J., Ghosh, S., Hanover, J. A., Gali, R. R., Grosu, P., and Shi, Y. (2009) The conserved NAD(H)-dependent coressor CTBP-1 regulates Caenorhabditis elegans life span. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1496–1501 CrossRef Medline

40. Ribeiro, A. J. M., Das, S., Dawson, N., Zaru, R., Orchard, S., Thornton, J. M., Orengo, C., Zeqiraj, E., Murphy, J. M., and Eyers, P. A. (2019) Emerging concepts in pseudoenzyme classification, evolution, and signaling. *Sci. Signal.* 12, eaat979 CrossRef Medline

41. King, J. E., and Jura, N. (2016) Structural basis for the non-catalytic functions of protein kinases. *Structure* 24, 7–24 CrossRef Medline