Emerging concepts in pseudoenzyme classification, evolution and signaling

António J. M. Ribeiro[1], Sayoni Das[2], Natalie Dawson[2], Rossana Zaru[1], Sandra Orchard[1], Janet T. Thornton[1], Christine Orengo[2], Elton Zeqiraj[3], James M. Murphy[4,5] and Patrick A. Eyers[6*]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2] Structural and Molecular Biology, UCL, Gower St, London, WC1E 6BT
[3] Astbury Centre for Structural Molecular Biology, Molecular and Cellular Biology, Faculty of Biological Sciences, Astbury Building, Room 8.109, University of Leeds, Leeds, LS2 9JT, UK
[4] Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia
[5] Department of Medical Biology, University of Melbourne, Parkville, Victoria 3052, Australia
[6] Department of Biochemistry, Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK

* Corresponding Author: patrick.eyers@liverpool.ac.uk

ONE SENTENCE SUMMARY:

Introduction of a framework to support the rapidly expanding field of pseudoenzyme analysis

GLOSS:

Enzymes are protein biological catalysts that speed-up the rate of chemical reactions in all living organisms. Enzymes regulate both metabolism and cell signaling, and bioinformatics has revealed surprising variations within the enzyme catalytic machinery, which converts a significant percentage into pseudoenzymes that lack a catalytic function. In this resource, we review recent progress in the identification, classification and functional annotation of pseudoenzymes, revealing them as key regulators of many fundamental biological processes.

ABSTRACT:

The 21st century is witnessing an explosive surge in our understanding of pseudoenzyme-driven regulatory mechanisms in biology. Pseudoenzymes are proteins that have sequence homology with enzyme families but which are proven or predicted to lack enzyme activity due to mutations in otherwise-conserved catalytic amino acids. The best-studied pseudoenzymes are pseudokinases, although examples from other families are emerging at a rapid rate as experimental approaches "catch-up" with an avalanche of freely-available informatics data. Kingdom-wide analysis in prokaryotes, archaea and eukaryotes reveals that between 5-10% of enzyme families contain pseudoenzymes, with significant expansions and contractions

seemingly associated with specific signaling niches. Pseudoenzymes can allosterically activate canonical enzymes, act as scaffolds to control assembly of signalling complexes and their localization, serve as molecular switches or regulate signalling networks through substrate or enzyme sequestration. Molecular analysis of pseudoenzymes is rapidly advancing knowledge of how they perform their own non-catalytic functions, and permitting the discovery of surprising, and previously unappreciated, functions of their intensively-studied enzyme counterparts. Excitingly, upon further examination, some pseudoenzymes have also proven to possess novel enzyme activities that could not have been predicted a priori. Pseudoenzymes can be targeted and manipulated by small molecules and therefore represent new therapeutic targets (or anti-targets, where intervention should be avoided) in various diseases. In this review, which brings together broad bioinformatics and cell signaling approaches in the field, we highlight a selection of new and recent findings relevant to a contemporary understanding of pseudoenzyme-based biology.

INTRODUCTION:

Genomic sequencing and annotation and the subsequent mining of datasets from varied organisms confirms that most well-characterized enzyme families encode pseudoenzyme homologs, which are predicted to be enzymatically-inactive due to the loss of at least one key catalytic amino acid residue. This basic description, garnered from primary sequencing data, has allowed the bioinformatic identification of pseudoenzyme genes (all of which are transcribed as mRNAs and translated into proteins) in >20 different protein families [1], including well-studied paradigms amongst the pseudokinases, pseudophosphatases and pseudoproteases [2-10]. As detailed in Table 1, subtle changes in catalytic and substrate-binding sites have likely led to the appearance of pseudoenzymes from classical enzyme templates, almost certainly following gene-duplication events. Not surprisingly, therefore, pseudoenzymes share a similar overall protein fold when compared with catalytically-active enzymes from the same family [11, 12]. Of further interest, detailed comparison between pseudoenzymes and related enzyme counterparts can also unearth physiological non-catalytic biological functions of enzymes, driven by the adoption of regulated inactive conformation(s). Such additional functions are also broadly encapsulated in the related concept of protein “moonlighting” [13], where proteins are able to ‘multi-task’, by performing different cellular functions as a consequence of their distinct environmental interactomes. Although an absence of conserved catalytic residues in a pseudoenzyme does not unequivocally prove catalytic deficiency, very high sequence and/or structural conservation suggests that pseudoenzyme sequences have been functionally selected across all the existing branches of life, and have been preserved to regulate specific aspects of cell biology through catalytically-independent mechanisms. Despite the clear presence of pseudoenzymes in a significant percentage of proteomes, we still understand very little about their individual function, especially relative to their active enzyme counterparts. Two recent international pseudoenzyme meetings, which were held in Liverpool, UK (2016) and Sardinia, Italy (2018), served as complementary hubs to bring together expertise from across the breadth of international biological communities, and to discuss and dissect how computational, theoretical and experimental data can be combined to make rapid progress in this new field. Below, we review and evaluate and discuss some of these bioinformatic, structural and biochemical approaches, which are rapidly revealing the extent, evolution and distribution of pseudoenzymes across the kingdoms of life.

Experimentally-established functional classes of pseudoenzymes

Pseudoenzymes are found amongst many metabolic and signaling classes of enzyme superfamilies (see Table 1 for an annotated selection). In terms of signaling outputs, pseudoenzyme actions can be rationalised based on four types of functional mechanisms: (i) regulating catalytic outputs of conventional (canonical) enzymes; (ii) acting as integrators of signaling events and/or toggling between signaling states as molecular switches; (iii) controlling assembly and localization of signaling hubs; or (iv), binding substrates/subunits to control the activity of conventional enzymes [14].

The first of these mechanisms, for which an increasing number of examples are available in the protein kinase, phosphatase and ubiquitin signaling literature, all retain clear "enzyme-like" overall folds. These pseudoenzymes evolved the ability to regulate catalysis of a *bona fide* enzyme-associated partner in order to generate a graded biological output in which the pseudoenzyme and enzyme interaction remains the crucial controlling factor. The kinase-pseudokinase pair is perhaps the most well-known example from this pseudoenzyme class [15]. However, this recurring theme has recently been exemplified in the ubiquitin field, with the analysis of the BRCC36 isopeptidase complex (BRISC) and the BRCA1-A complex, in which the catalytically-active JAMM (JAB1/MPN/Mov34 metalloenzyme) domain Lys_{63}-specific deubiquitinase (DUB) BRCC36 is physically partnered with the pseudo-DUBs Abraxas1 and Abraxas2 [16]. The BRISC complex permits cooperation between a newly discovered "moonlighting" function of the metabolic enzyme serine hydroxymethyltransferase 2 (SHMT2), the pseudo-DUB Abraxas2 and three tandem pseudo-E2 (UEV) domains of BRCC45 (**Fig. 1**). Complex assembly is regulated by SHMT2, which requires pyridoxal-5'-phosphate (PLP) to form a tetramer that catalyzes its canonical role in nucleotide and amino acid metabolism. Catalytically-inactive SHMT2 dimers, but not active, PLP-bound tetramers, inhibit BRISC DUB activity, and intracellular PLP abundance controls this newly discovered moonlighting SHMT2 function, which leads to the co-ordinated regulation of immune signaling in cells [17].

A second class of pseudoenzymes act as binary "switches" by integrating signals in the form of posttranslational modifications (PTMs) or by binding to metabolic ligands (perhaps ancient substrates), triggering inter-conversion between active and inactive conformations. Several of these are presented in more detail in Table 1.

The third category of pseudoenzymes possess distinct biological functions by operating as protein interaction scaffolds, generating (sub)cellular focal points that nucleate assembly of protein complexes, or regulate the localization of a binding partner. Examples are shown in Table 1, with the most profligate example to date being pre-mRNA-processing-splicing factor

8 (PRPF8), which contains four pseudoenzyme domains (Table 1) that act as scaffolds for assembly of the enzymatic spliceosome complex [16, 18]. The further study of protein interactions made by pseudoenzyme domains ('pseudoenzyme interactomes') will provide interesting information about the role of these proteins in the cell. Well-known examples include the interaction of the STYX pseudophosphatase with components of the canonical ERK pathway [19] and independent targeting to FBXW7, a subunit of the SCF Ubiquitin E3 ligase [20]. In a broad sense, molecular interaction data analysis can now be facilitated by using the detailed curation model employed by the International Molecular Exchange (IMEx) Consortium of molecular interaction databases [21], which accurately maps interaction data to specific binding-regions of proteins, and captures the effects of site-directed mutagenesis on potential protein:protein interactions [22].

The final major pseudoenzyme category contains examples in which the protein has repurposed canonical features of a fold that is shared with active enzyme relatives, probably to act as competitors for either substrate binding (as catalytic "traps") or higher order complex assembly. A good example of this fourth category are pseudophosphatases, which usually retain the ability to bind to "substrates", such as phosphorylated peptides, but are no longer able to enzymatically process them, providing a trapping or localization mechanism *in situ* [5, 23]. The same might also be true of pseudoproteases [2], which possess reasonable binding affinity for substrates (and can therefore control their fate), but do not cleave their substrates, performing a catalytically-independent regulatory role. Outside of these notable pseudoenzyme families, relatively few examples of competitor and/or "decoy" pseudoenzymes have been identified to date. However, new data confirm their expansion within plant, fungal and pathogen proteomes [24-27], including, for example, a new viral pseudoenzyme that provides the pathogen with an advantage by directly competing with host cell defence mechanisms [28].

Below, we present the fruits of the findings from the first two international meetings held between members of the pseudoenzyme community, including new state-of-the-art computational approaches that can exploit both genomic and proteomic (and in the future metabolomic) datasets. Together, these are revealing a vast number of pseudoenzymes embedded within genomes, and helping to create the first advanced analytical frameworks to probe and understand how pseudoenzymes function in a multitude of biological niches.

Bioinformatic resources for pseudoenzyme analysis: How computational scientists can identify and curate pseudoenzymes

The UniProt knowledgebase (www.uniprot.org) provides the scientific community with a comprehensive, high quality and freely-accessible resource of protein sequence and

functional information for many species [29]. For enzymes, which represent between 20 and 40% of most proteomes, UniProtKB provides additional information about Enzyme Classification (EC), catalytic activity, cofactors, enzyme regulation, kinetics and pathways, all based on critical assessment of published experimental data. Bioinformatic and structural data are also used to enrich the annotation of the sequence with the identification of active sites and binding sites. Whereas the annotation of enzymes is very well defined, the curation of pseudoenzymes has proven much more challenging. The main issue resides in their identification, especially when the only evidence usually available to define a pseudoenzyme, is the *absence* of critical active site residues. Indeed, to infer a lack of catalytic activity based solely on sequence analysis can be misleading. Experimental evidence can also be difficult to interpret, especially when it contradicts sequence analysis-based prediction and catalysis may be present or absent among orthologs. During the curation process, all the available evidence must be manually assessed to decide whether the protein could be considered a pseudoenzyme or not (**Fig. 2A**). Another important challenge is how to translate this information into consistent, meaningful annotation to facilitate the subsequent retrieval of pseudoenzymes by users. Following the two international pseudoenzyme conferences held in 2016 (Liverpool, UK) and 2018 (Sardinia, Italy), the annotation of pseudoenzymes in UniProt underwent a revision to reflect advances in the pseudoenzyme field. For example, the UniProt curators improved the usage of protein names to directly reflect their inactive status. In addition to providing the enzyme family to which they are related, as well as the position of the "catalytically inactive" domain, UniProt explains the reason why a protein is considered a pseudoenzyme in a "Caution" comment in the enzyme function section (**Fig. 2B**). Importantly, the source of the evidence used to infer the lack of catalytic activity—experimental, sequence analysis or orthology-based—has also been added to make pseudoenzyme definitions as evidence-based as currently practicable (see Table 1).

As an experimental test-bed for the curation and annotation of procedures to create pseudoenzyme databases, UniProtKB have recently reviewed and updated the annotation of the complete kinome [30] and phosphatome of the model worm *C. elegans*, which contain 438 kinases and 237 phosphatases, respectively. Among 208 kinase genes whose function has been experimentally characterized, 41 are annotated as pseudokinases, although supporting evidence is often lacking beyond predicted sequence information. Indeed, for > 90% of worm pseudokinases, a lack of catalytic activity is based only on sequence analysis (Fig. 3A). For genes that lack functional characterization, the sequence analysis predicted lack of kinase activity can be supported by experimental evidence in orthologs. However, this is not always possible, especially for families that contains mainly *C.elegan*s specific genes. This is the case for the two most abundant kinase families, CK1 and TK. For the CK1 family, 11 of the 83 members (13%) and for the TK-KIN-16 group 4 of the 15 members (26%) are predicted

to be inactive based on the loss of the active site. The TK-Fer group is predicted to contain only active members. A similar situation is found in the *C.elegans* phosphatome. Of the 237 phosphatases, 8 are annotated as “inactive” pseudoenzymes, although only one pseudophosphatase possesses documented experimental evidence supporting the facile informatics-based prediction of catalytic sterility (**Fig. 3B**). These two examples illustrate the need to improve experimental characterization of pseudoenzymes, as inferences based only on sequence analysis could result in the incorrect attribution of inactivity, as the well-known example of the with-no-lysine (WNK1; UniProtKB Q9JIH7) family of serine/threonine-protein kinases ably demonstrates [31, 32]. Based on sequence analysis, WNK enzymes would be predicted to be inactive; however, under various experimental conditions tested, they are clearly catalytically active [6]. The WNK members demonstrate the importance of biochemical and structural analysis in understanding enzymes and pseudoenzymes in enzyme superfamilies, such as phosphotransferases; in the case of the WNK “pseudokinases”, this has led to their more accurate re-classification to atypical kinases.

Perhaps the most difficult challenge facing the curation field is the vast amount of protein sequences that continue to be made available through genome sequencing projects. For example, in just one month between UniProt releases in October and November of 2018, over 3 million new protein sequences were added to the database. In order to provide users with information on these proteins (without experimental evidence in nearly all cases), automatic annotation rule-based systems have been devised to enrich the annotation of protein sequences based on protein family membership. These identification systems, based on related manually curated protein entries, provide users with basic predicted functional annotation; an example being UniProtKB H9J2B7, a predicted tyrosine-protein kinase receptor based on automatic sequence detection analysis. Despite the development of rules and tools aimed at identifying differences and similarities in natural variation, such rules may not recognize pseudoenzymes even within a closely related enzyme family.

Evolutionary biology as a driver in the pseudokinase field

Whereas several workflows exist for annotating enzymes, the “rules” for automatic annotation of pseudoenzymes are currently lacking. Focused progress needs to be made on this issue across multiple different enzyme families in order to fully appreciate pseudoenzyme diversity and potential biological ubiquity. Reliable identification of catalytic site residues that can be used to predict a loss of enzymatic activity remains difficult. However, increased experimental characterization of pseudoenzymes in the last decade, and new evolutionary analyses of

226 catalytic sites and catalytic domains, continues to provide invaluable information that improves
227 manual curation and expands the automatic identification/annotation of pseudoenzymes. A
228 case-in-point are the pseudokinases, for which a kinome-wide database has now been
229 assembled across all known eukaryotic, bacterial, and archaeal proteomes [24]. This analysis,
230 published in *Science Signaling*, has revealed that pseudokinases are present across all three
231 domains of life; in total, approximately 30,000 eukaryotic, 1,500 bacterial, and 20 archaeal
232 pseudokinase sequences were classified into 86 distinct pseudokinase families, including ~30
233 well conserved pseudokinase families that were not previously reported [24]. The rich diversity
234 of pseudokinases that occurs across the kingdoms of life exhibits notable family-specific
235 expansions in animals, plants, fungi, and bacteria, where pseudokinases (and
236 pseudoenzymes in general) had previously received cursory attention. Pseudokinase
237 expansions are often accompanied by domain shuffling, which appears to have promoted new
238 roles in plant innate immunity, modulation of plant-fungal interactions, and bacterial signaling.
239 Mechanistically, the ancestral kinase fold, an ideal template for the generation of new
240 functions in pseudoenzymes, has diverged in many distinct ways through the enrichment of
241 unique signature sequence motifs, generating in-turn a slew of new pseudokinase families.
242 The catalytic kinase domain is also repurposed for non-canonical nucleotide binding or
243 atypical catalysis, or to stabilize unique, catalytically-inactive kinase conformations associated
244 with catalytically-independent types of signaling (see below). To conveniently compare these
245 complex datasets, an annotated, searchable collection of all known predicted pseudokinase
246 sequences, and their evolutionary relationships, has been captured in the freely available
247 protein kinase ontology (ProKinO) [24, 33]
248 (http://vulcan.cs.uga.edu/prokino/hierarchy/ProteinPseudokinaseDomain).

249

Structural biology as a key driver in the pseudokinase and pseudoenzyme fields

Structural studies continue to play a pivotal role in advancing our knowledge of functional evolution among enzyme superfamilies, most notably the kinase superfamily, which is composed of a very broad range of small molecule, antibiotic, glycan and protein kinases [34]. Notable examples of these important pseudoenzymes are discussed here briefly, beginning with the remarkable finding that nucleotides can bind in different modes in predicted pseudoenzymes, in some cases leading to unusual (and completely unexpected) types of catalysis. For example, an inverted conformation of ATP was observed in the pseudokinase FAM20A (**Fig. 4A**) [35, 36], which is a catalytically-inactive pseudoenzyme regulator of the conventional protein kinase FAM20C [37]. However, a structure of SelO revealed a similar atypical ATP-binding mode, but rather than being catalytically-inactive, SelO has evolved the capacity to "AMPylate", rather than phosphorylate, a broad spectrum of protein substrates (**Fig. 4B**) [38]. Additionally, several (pseudo)kinases have been confirmed to operate as phosphorylate sugar (as opposed to amino acid) residues, including FAM20B [37, 39] (**Fig. 4C**) and SgK196/Protein *O*-mannose kinase (POMK) [40] (**Fig. 4D**). Intriguingly, the bacterial pseudokinase SidJ has recently been shown to catalyse protein glutamylation of the SidE family of ubiquitin E3 ligases, inhibiting their catalytic output [41, 42].

Although not classically considered to be a specific catalytic motif, the variable glycine-rich loop sequence in kinases (Gly-X-Gly-X-X-Gly) serves as a flap-like structural feature to promote and lock ATP-binding. Interestingly, this motif has been found to be dispensable for catalytic activity in an atypical catalytically-active coccidian kinase, which has been termed WNG1 (with no glycine-1) [43]. A recent study suggests a catalytically active, druggable conformation in Drosophila BubR1 that can directly phosphorylate the motor protein CENP-E [44]. This likely differentiates it from human BubR1, which is reported to be an inactive pseudokinase [45] lacking a Gly-rich loop and that not bind detectably to ATP [46]. Although it is more common that proteins annotated as pseudokinases based on the absence of the traditional catalytic residues (which do not, by definition, include the Gly-rich loop) are indeed catalytically-defective when scrutinized at the appropriate biochemical level, these examples underscore the versatility of the kinase fold to evolve diverse functions, especially when released from the evolutionary pressure of maintaining a catalytically-competent fold [47]. More broadly among pseudoenzymes, such functions likely speak to a non-catalytic substrate binding to an ancestral enzyme, which has evolved to perform a catalytic function, as has been demonstrated during *in vitro* evolution experiments [48, 49] and by using structural features to evolve canonical activity in the pseudokinase CASK, where 4 amino acid substitutions are required to regenerate a Mg-ATP dependent enzyme [50]. In a similar vein, improving the curation of (non-kinase) pseudoenzyme families will provide the scientific

community with valuable information to understand the evolution of these proteins, the aetiology of related diseases and the development and repurposing of pseudoenzyme-targeted drugs, a useful bonus in carefully-conducted pseudoenzyme studies [7, 51, 52].

A new approach for the identification of pseudoenzymes using UniProt

To help take the pseudoenzyme field forwards, we have designed and tested a simple computational pipeline to identify and annotate pseudoenzymes using sequence alignments, UniProt annotations assembled from the primary literature (The UniProt Consortium, 2017), and information from the Mechanism and Catalytic Site Atlas (M-CSA), a database that currently contains defined catalytic residues and mechanistic data for 964 enzymes [53]. We have limited our current analysis to those enzymes for which we have a good or detailed knowledge of their catalytic mechanism (such as those in M-CSA) and therefore know the specific residues involved. To identify pseudoenzymes amongst these data, we start by finding all the SwissProt sequences that are homologous to entries in M-CSA, by using phmmer with an E-value cut-off of 10^{-10} [54]. This value is reasonably stringent and will act as a filter to include only close homologues. More distant relatives are better identified using the structure-based approaches developed by Orengo and colleagues, as described below and in related work [12, 55]. Our procedure yields a broad collection of sequences from all the domains of life, although the sample can be biased by the (current) uneven representation of sequences in SwissProt and enzymes in M-CSA. After their identification, we categorize each homologue as enzyme or non-enzyme according to its annotation in SwissProt. There are at least three types of annotation in SwissProt that we can use to identify enzymes: EC numbers, UniProt keywords, and GO terms. These annotations provide us with three possible rules. A sequence can be categorized as an enzyme if it (i) has at least one EC number; (ii) is annotated with at least one catalytic UniProt keyword (Oxidoreductase, Transferase, Hydrolase, Lyase, Isomerase, Ligase, Translocase, or their hierarchical children); or (iii) is annotated with a catalytic GO term ("catalytic activity" and its child terms). Unfortunately, the three rules are not currently entirely consistent. For example, there are approximately 3,000 sequences with a catalytic keyword but no EC number, and about 35,000 sequences possess a catalytic GO term but have no catalytic keyword. These differences may be due to out-of-date annotation, no available EC number in the EC classification, or GO terms that are automatically transferred based on homology, which will label pseudoenzymes as enzymes. Overall, the rule (ii) based on the UniProt keywords seems to be the most comprehensive without using extended annotation based uniquely on homology, and so we have applied it for the purpose of this review. There are no pseudoenzymes in 669 of the enzyme families curated in M-CSA. However, in 237 of the enzyme families where pseudoenzymes are found, these account for less than 10% of the sequences in the family. The number of families where pseudoenzymes

323 are more common than 10% is 88 (the sum of the last 9 columns in Figure 5). However, the
324 percentage of families identified as including pseudoenzymes is currently a *minimum* number
325 - more can and will be found with deeper searches. Below, we highlight how we can use data
326 from individual enzyme families to try to understand how pseudoenzymes have evolved, an
327 important central question in evolutionary biology.
328

Evolution of pseudoenzymes

Knowing the number of pseudoenzymes associated with each M-CSA enzyme entry is an important starting point for understanding how these proteins came to "lose" (or repurpose) their catalytic function, assuming that they evolved from canonical enzymes in the first place. The reconstruction of their evolutionary trees is necessary to tackle some of the fundamental questions that the field would need to answer. These include (i) how common are the evolutionary 'jumps' that either transform enzymes into pseudoenzymes or pseudoenzymes into enzymes?; (ii) which of these transformations is more common?; and (iii) can we identify the conserved mutation(s) in the catalytic residues that drive these changes? Below, we have attempted to answer the last question by generating and annotating phylogenetic trees for all the enzymes in M-CSA. In the phylogenetic tree for β-amylase (**Fig. 6**), which contains two pseudoenzymes, the homologues belong mostly to plants, except for five that possess a bacterial origin. The red and green circle on the tree represents the point where one pseudoenzyme diverged from an ancestral enzyme. The two existing pseudoenzymes evolved from this pseudoenzyme ancestor, so there is only one "loss-of-function" event represented in the tree. The green circle at the base of the tree means that the last common ancestor for all the homologues in the tree was most likely an enzyme, consistent with the directionality of evolution from enzyme to pseudoenzyme. The pseudoenzymes, which are both classified as "inactive" in UniProtKB and have no associated EC numbers, have the most mutations in their catalytic residues of all the homologues identified. Furthermore, the pseudoenzymes are the only homologues in the tree with substitutions replacing the catalytic residue, Glu_{381}. This acidic side-chain acts as the general base that activates the nucleophilic water molecule to a reactive hydroxide ion, while other residues such as Thr_{343}, Leu_{384}, and Asp_{102} have stabilization roles that are not critical for the reaction, and so can (and are) readily replaced during evolution without affecting enzyme function. It is currently impossible to say if these mutations were the original cause for the loss of function, or if they happened after the loss of function, which may have occurred through other means such as simple accumulation of mutations without corrective selective pressure. After the loss-of-function event, a lack of corrective selection pressure will lead to the accumulation of mutations in the catalytic residues. For this enzyme set, the mutation of catalytic residues alone does not lead necessarily to loss-of-function, as some of the other "active β-amylases have catalytic mutations. Gain-of-function may also be generated through means other than catalytic residue mutations, as illustrated by the β/α-amylase protein (P21543), which has an additional domain with the α-amylase activity (Fig. 6, dark blue).

Use of GO terminology to estimate the global extent of pseudoenzymes

There are currently a large number of entries in UniProt (about 35,000) that do not have a catalytic keyword but do possess a catalytic GO term. This inconsistency is brought about by GO annotation that is automatically extended from enzymes to their homologues, regardless of the presence or absence of actual catalytic activity. Consequently, many of the sequences annotated with a catalytic GO term are themselves pseudoenzymes. To illustrate this contradiction, we also performed an analysis where we removed all the sequences where annotation is contradictory. This led to a smaller total number of pseudoenzymes identified within the M-CSA entries homologues set, as would be expected (fig. S1), and the number of enzyme families with pseudoenzyme members drops from 325 (32.7% of all the families) to 132 (13.3%). Clearly, this type of analysis remains rather subjective unless backed up with supporting data, and represents one of the major experimental challenges in the pseudoenzyme field, where biochemical and cellular analysis are often complex, time-consuming and costly.

Identifying pseudoenzymes in proteomic databases by exploiting the Class, Architecture, Topology, Homology (CATH) resource

Pseudoenzymes have now been identified in many major enzyme families across the tree of life; predictably, this has been predominantly through computational sequence-based analyses [30, 56] (Table 1). Most studies have traditionally compared the sequence of a relative of unknown structure and functional residues against sequences of relatives, which have been structurally characterized and annotated with known catalytic residues that have been confirmed experimentally [14]. However, one study has indicated that homologous proteins that share a common core domain structure can often acquire new enzyme functions by changes both in the nature of their catalytic residues and in the absolute position of these catalytic residues in the protein scaffold [55]. As a result, it can be very difficult to provide confident predictions of pseudoenzymes by only looking at deviations from known catalytic sites, most notably if this is done so in the absence of experimental characterization and/or phenotypic or mutational information. More complex approaches are therefore deemed suitable wherever possible.

As an extension of the approach described above using enzyme terminology, we have also systematically investigated the distribution of pseudoenzymes in the protein universe, using protein families from the CATH-Gene3D resource [57, 58], which links protein domain sequences to structures and experimental functions. CATH-Gene3D classifies ~460,000 domain structures and ~95 million protein domain sequences into ~6,200 evolutionary superfamilies [57]. These can then be sub-classified into Functional Families (FunFams) that share highly similar structures and functions, based on sequence patterns – specificity-

determining positions (SDPs) and other conserved positions (**Fig. 7A**) [59]. Although it is difficult to achieve complete separation in some extremely diverse superfamilies, the accuracy of these functional family classifications has been validated by comparison against experimental data and by endorsement through blind independent assessment in the Critical Assessment of protein Function Annotation (CAFA) algorithm-based evaluation of functional annotations [60]. The number of functional families reflects the functional diversity of a particular superfamily and can be used to explore how protein function is modulated in diverse superfamilies.

We examined 383 enzyme superfamilies in CATH/Gene 3D v.4.2 that contain well-known (experimentally-validated) catalytic domains [53, 55], and identified the proportion of functional families that have enzyme annotations and compared them to those that lack any enzyme annotation. These are highly populated superfamilies accounting for 64% of sequences in all CATH enzyme superfamilies and 60% of all sequences in CATH. A functional family was considered to have enzyme annotations if it has at least one relative that has an EC annotation in UniProtKB [29] and an experimental Gene Ontology (GO) [61] annotation for 'catalytic activity'. For a third (131) of these enzyme superfamilies, all functional families were annotated as enzymes (**Fig. 7B**). However, approximately 252 enzyme superfamilies (two-thirds) had varying proportions of functional families that had no enzyme annotations in the EC classification or GO (**Fig. 7C**), suggesting that these are very likely to be pseudoenzymes. To collate this information in a searchable manner, we have created the first on-line list of putative pseudoenzyme superfamilies (https://uclorengogroup.github.io/cath-pseudoenzymes/index.html). In order to explore some of these pseudoenzyme-containing families, we also developed a protocol that first generates structure-guided multiple sequence alignments of multiple functional families within a superfamily that share very similar structures (such as, wherein relatives superimpose with <RMSD 5Å), and are therefore likely to share catalytic mechanisms. For these structural clusters, grouping structurally similar relatives from different functional families, we next examined whether there was at least one functional family within them with known catalytic residues [53], and at least one family with no enzyme annotation. A functional family lacking enzyme annotation was identified as a putative pseudoenzyme family if we observed a loss or significant change in known catalytic site residues (identified in the enzyme family) in all relatives of the putative pseudoenzyme family, determined using the comprehensive structural cluster alignment. In other words, the putative pseudoenzyme family has conserved residues in the active site that differ significantly in their physicochemical nature or 3-dimensional location to the known catalytic residues of the enzyme family. Clearly, the final confirmation of the sequence as a pseudoenzyme can only come from exhaustive experimental testing, as is now becoming much more commonplace in major signaling superfamilies, such as the phosphatases and pseudophosphatases, kinases

and pseudokinases, and proteases and pseudoproteases [5, 14]. Finally, analysis of some of these structural clusters in superfamilies enabled us to identify families containing previously reported pseudoenzymes, such as the uclear ransfer actor 2 (NTF2) and calsequestrin [62]. For example, the CATH superfamily 3.10.450.50 contains ~38% of functional families with enzyme annotations, whereas 62% of the families do not have any enzyme annotations. The distinct functional family containing NTF2 is structurally very similar (< 4Å RMSD, same structural cluster) to the enzyme families containing scytalone hydratases and steroid delta-isomerases but lacks the catalytic machinery of those enzyme relatives (**Fig. 8**). Another interesting example is N-Myc downstream-regulated gene 2 (NDRG2), which is structurally similar to the enzyme families in the α/β-hydrolase superfamily (CATH 3.40.50.1820) and is thought to be a tumor suppressor [63].

The future of pseudoenzyme research

The evolutionary conservation and prevalence of genes that encode pseudoenzymes is now abundantly clear in the natural world, from well-studied model organisms to newly sequenced genomes from previously uncharacterized species. However, one major barrier to accurate pseudoenzyme cataloguing prior to experimental triage is the broad diversity of terminologies previously used in their literature description, differing definitions of what constitutes a *bona fide* pseudoenzyme, and an historical prejudice that proteins lacking catalytic activity are of unlikely to serve biologically-important functions. Over the last 20 years, various studies have referred to pseudoenzymes as “non-enzymes”, “prozymes”, “dead” or “catalytically-defective/inactive” or, even more confusingly, “atypical” or “non-canonical” enzymes. We therefore propose the broad adoption of “pseudoenzyme” as the essential descriptor that will behave as a critical rate-determining step in allowing the direction of this relatively new field to be strategically plotted and accurately delivered. Formally, we propose that pseudoenzymes be defined as “*the predicted catalytically-defective counterparts of enzymes owing to an absence of one or more catalytic residue*”. It should come as no surprise that in some cases in which this simple description is applied, residual catalytic activities have been reported amongst proteins defined as pseudoenzymes (mainly pseudokinases), including the isolated HER3/EGFR3, JAK2 (JH2), CASK and TRIB2 pseudokinase domains [7, 50, 64, 65]. However, because these very modest catalytic activities are either dispensable for biological function, or not conserved in vitro among paralogous proteins across species (such as HER3/EGFR3[2) [24], this vestigial (or residual) catalytic activity is probably not a defining feature of their biological function. Technically, establishing and quantifying such activities remains enormously challenging, because *ab initio* prediction of substrates and co-factors is not facile and even trace co-purifying/contaminating proteins in recombinant pseudoenzyme preparations (which may lack auxiliary endogenous factors) can lead to erroneous attributions of catalytic functions.

Other cases discussed in this review pose a more serious definitional challenge: diverse proteins predicted to be ‘catalytically-dead’ pseudokinases, like SelO, SidJ, and SgK196/POMK, have instead evolved distinct, and quantifiable, catalytic functions, illustrating a potential weakness of *ab initio* prediction of protein function based on a comparative assessment of sequence. However, these important findings serve to highlight the versatility of protein domains as ancestral folds for the creation of newly evolved activities. Moreover, studying these “odd” folds, and various adaptions at the amino acid level, can potentially lead to paradigm-shifting discoveries, including the three examples listed above. Rather than being defined solely as pseudoenzymes by virtue of the unconventional positions of their catalytic

residues, such catalytic proteins can more accurately be considered as enzymes that exhibit atypical or novel catalytic mechanisms. Historically, this is best exemplified by WNK1, which guided by structural studies was revealed to contain a compensatory ATP-positioning lysine in the β2 rather than the β3 strand found in the vast majority of protein kinases [32]. The discovery of catalytically active WNG1, one of several secreted *Toxoplasma* atypical kinases that together form a clade (termed ROP33, 34 and 35) comes as an additional pleasant surprise [43], especially given that the related pseudokinase Bradyzoite pseudokinase 1 (BPK1), which possesses no detectable nucleotide binding *in vitro* [46], was previously shown to be involved in both the development and infectivity of *Toxoplasma* cysts [66]. The discovery of non Gly-rich loop-containing atypical protein kinases adds an interesting new twist to the WNK1 paradigm, because it has generally been assumed that the **complete** absence of a Gly-rich loop would preclude ATP binding and catalysis amongst phosphotransferases [46, 67-69]. However, it is now abundantly clear from these, and other, examples that a detailed knowledge of protein structure aligned with robust biochemical assays continues to be essential for understanding and validating the novel catalytic or pseudoenzyme-based mechanisms that are at play.

A rich diversity of pseudoenzymes in genomes revealed by bioinformatics

Databases such as UniProtKB, and the biocurators that populate new and existing entries with information extracted from the biomedical literature, play a critical role in cataloguing experimentally-evaluated pseudoenzymes. These characterised proteins, however, represent only a small percentage of the pseudoenzymes believed to be present in even well-studied model genetic organisms. Moreover, their identification is critical in enabling the recognition of pseudoenzymes in all proteomes, and for the development of more sophisticated recognition criteria, which should allow these proteins to be distinguished from catalytically active family members. While bioinformatic studies can readily predict a lack of conventional activities based on founder members of a particular enzyme class or mechanism, they cannot exclude the acquisition of novel enzymatic functions or predict the biological nature of the non-catalytic functions a pseudoenzyme might perform, which is the major reason to study them. Nonetheless, the evolution of non-catalytic functions by pseudoenzymes provides a window into formerly unrecognized regulatory functions mediated by conventional enzymes, such as ATP-binding site occupation in the absence of catalysis in PKC [70], and the stabilization of NMYC by catalytically-inactive Aurora A [71]. While this “moonlighting” allows enzymes to perform additional functions beyond their first-characterized catalytic function(s), it is likely that gene duplications have allowed pseudoenzymes to evolve specific functions, often within the same pathway as the parental enzyme. Duplication allows a parental enzyme to perform an essential function, since it relieves the selective pressures on the duplicated enzyme that

would otherwise constrain active site geometries for catalysis and substrate recognition. Divergence arising from relief of such selective pressure is very well illustrated by structural plasticity among pseudokinase domains. For example, the pseudokinase domains of mouse and human MLKL exhibit divergence in the structures of their “pseudoactive” site clefts and the relative contributions of pseudoactive site residues to ATP binding, which is a relatively common feature of many catalytically-inactive pseudokinases [47, 72]. Additionally, the positions of one of the core elements in active protein kinases, the αC helix, whose Glu positions the ATP-binding lysine in β3 of the N-lobe for catalysis, is highly variable among both kinase and pseudokinase domains. Indeed, in some cases, such as the pseudokinase Sgk223, the αC helix is absent in crystal structures [73] and/or has evolved alternative functions, such as a serving as a regulated platform for peptide-binding [74, 75] amongst the Tribbles pseudokinases [76], which can themselves be targeted with small molecules originally developed as inhibitors of canonical kinases [52, 77].

We believe that bioinformatics approaches, including those under development discussed here, will permit the identification of many new pseudoenzymes amongst the large number of enzyme families in annotated proteomes. Initially, this will be by leveraging our knowledge of protein structure and catalytic mechanisms among active enzyme counterparts in order to predict deficiencies in catalytic residues within their pseudoenzyme cousins. Furthermore, the continued massive expansion of the sequence repositories, particularly with many hundreds of millions of novel enzyme sequences coming from metagenome studies, combined with increasing protein structure data, will enhance the power of those bioinformatic methods that analyse the presence or absence of highly sequence-conserved residue positions in experimentally uncharacterised functional families. Expanding the size and diversity of family membership gives clearer, more accurate sequence patterns. These can then be more easily compared against sequence patterns of known enzyme families in the same superfamily in order to more accurately predict loss of essential catalytic machinery. In addition, they might also identify novel (and experimentally testable) catalytic machinery, spatially located in the active site pocket that is potentially indicative of a shift or change in signaling function. Such approaches will facilitate broad mining of protein sequences to identify candidate pseudoenzymes, but will ultimately rely on a combination of structural and biochemical studies, ideally including the mapping of enzyme <-> pseudoenzyme evolutionary trajectories, to formally evaluate catalytic-deficiency and/or the acquisition of non-canonical catalytic mechanisms.

Improved use of mechanistic data to better understand pseudoenzyme evolution

In this review, we have shown how Uniprot data can be used to identify pseudoenzymes associated with a dataset of enzymes of interest. One such dataset is the M-CSA, which contains information about the catalytic residues and the reaction mechanisms of 964 enzymes. By creating annotated phylogenetic trees, we are able to show where in the evolutionary past the loss of function events occur and which catalytic residues mutations are associated with those events. In the particular example shown, we observe that not all catalytic mutations lead to loss of function, and that the use of this rule to all enzyme families, in general, may be too simplistic. Our future major aim is to analyse the phylogenetic trees of all enzymes in M-CSA that possess pseudoenzyme relatives. By using annotation specific to M-CSA, we will understand if the loss of function events are related with particular catalytic residue functions and the specific chemistry the enzyme catalyses. For example, a mutation in a residue acting as an electrostatic stabilizer may be tolerated, while a mutation in a nucleophile may not be. Furthermore, this tolerance to mutations may be dependent on the specific reaction and the types of chemical groups in substrate(s), which adds another level of complexity.

Overview of different approaches to identify and characterise pseudoenzymes

In this review, three ways are described to identify and characterize pseudoenzymes across protein families. The first is the manual curation by UniProt using the literature or manually checked computational annotation, so that although this method is labour intensive, it is the most accurate of the three. Efforts are presented to characterize all the kinases and phosphatases in *C. elegans*, for which about 20% and 3% of these family members are pseudoenzymes, respectively. The pseudoenzymes identified and annotated in this manner will eventually form the 'gold standard' from which other automated methods will develop. The second method uses sequence homology to identify close relatives and existing SwissProt data to identify pseudoenzymes, which are then annotated with the catalytic residues of the original enzyme, to check their conservation. This sequence-based method identifies pseudoenzymes in about one third of the enzymatic families in M-CSA. The final method uses the CATH structural database as a starting point to identify related proteins, and UniProt and Go annotation to categorize them as either enzyme or non-enzyme. Protein structure is more conserved during evolution than sequence, hence, structural methods can see further into the past to uncover more ancient relationships. Therefore, as expected, this method detects more families with pseudoenzymes than the previous one, identifying about two thirds of a set of enzymatic CATH superfamilies as pseudoenzyme-containing families. The Achilles heel of the second and third methods is their reliance on the lack of catalytic annotation as a useful

criterion for defining non-enzymes. This is a general problem in the pseudoenzyme field, which must use negative evidence, or lack of observed experimental catalytic activity, as a benchmark. The problem is compounded when using databases, where lack of annotation does not mean that catalytic activity was evaluated. More experiments are therefore needed to test more broadly for enzymatic activity and manual curation of the absence of evidence needs to be distinguished from clear experimental proof of a lack of catalysis. Until then, comparative computational approaches (including those described above) are the most powerful method for pseudoenzyme identification that we currently possess.

CONCLUSIONS:

The next decade will be an exciting, and potentially transformative, period of rapid development in the pseudoenzyme field, as experimental and bioinformatic findings rapidly merge, creating new databases that bring together exploitable information for specialists and non-specialists alike. Feeding this information into experimental workflows will rapidly lead to a revolution in our understanding of enzyme and pseudoenzyme evolution and inform fundamental fields ranging from protein folding and enzyme mechanism to cell signalling, metabolism, and drug discovery. Such endeavours will require much broader comparative studies between enzyme families from diverse species, rather than the piecemeal approaches currently favored for studying enzymes and pseudoenzymes in isolation away from their physiological environment. Beyond readily searchable comparative datasets, a major outcome for such studies will also be the creation of benchmarks for studying, and predicting, the effects of evolutionary and disease-associated mutations that take place in enzymes and pseudoenzymes, especially when these changes are conserved in molecular “hot-spots”. More data together with more sophisticated methods will facilitate the development of highly accurate tools that can identify pseudoenzymes computationally. The assembly of these new sets of “rules” (and their allied rule-breakers) will then create truly useful, and biologically informative, outputs. This will help to loosen current phenomenologically complex classifications that can constrain, and often blur, the numerous strands of enzyme-based research taking place. In particular, the fruits of these labors are likely be a complete species-level catalog of pseudoenzymes across hundreds of distinct enzyme families and superfamilies and the prioritization of biochemical, cellular and guided-evolution frameworks to study pseudoenzymes that are of interest across scientific disciplines.

REFERENCES AND NOTES:

1. Murphy, J.M., H. Farhan, and P.A. Eyers, *Bio-zombie: the rise of pseudoenzymes in biology.* Biochem Soc Trans, 2017. **45**(2): p. 95-104.

2. Adrain, C. and M. Freeman, *New lives for old: evolution of pseudoenzyme function illustrated by iRhoms.* Nat Rev Mol Cell Biol, 2012. **13**(8): p. 489-98.
3. Eyers, P.A. and J.M. Murphy, *The evolving world of pseudoenzymes: proteins, prejudice and zombies.* BMC Biol, 2016. **14**(1): p. 98.
4. Jacobsen, A.V. and J.M. Murphy, *The secret life of kinases: insights into non-catalytic signalling functions from pseudokinases.* Biochemical Society Transactions, 2017. **45**(3): p. 665-681.
5. Reiterer, V., P.A. Eyers, and H. Farhan, *Day of the dead: pseudokinases and pseudophosphatases in physiology and disease.* Trends Cell Biol, 2014. **24**(9): p. 489-505.
6. Eyers, P.A. and J.M. Murphy, *Dawn of the dead: protein pseudokinases signal new adventures in cell biology.* Biochemical Society Transactions, 2013. **41**(4): p. 969-74.
7. Kung, J.E. and N. Jura, *Prospects for pharmacological targeting of pseudokinases.* Nature reviews. Drug discovery, 2019.
8. Chen, M.J., J.E. Dixon, and G. Manning, *Genomics and evolution of protein phosphatases.* Science signaling, 2017. **10**(474).
9. Boudeau, J., et al., *Emerging roles of pseudokinases.* Trends in cell biology, 2006. **16**(9): p. 443-52.
10. Zeqiraj, E. and D.M. van Aalten, *Pseudokinases-remnants of evolution or key allosteric regulators?* Current opinion in structural biology, 2010. **20**(6): p. 772-81.
11. Pils, B. and J. Schultz, *Inactive enzyme-homologues find new function in regulatory processes.* J Mol Biol, 2004. **340**(3): p. 399-404.
12. Todd, A.E., C.A. Orengo, and J.M. Thornton, *Sequence and structural differences between enzyme and nonenzyme homologs.* Structure, 2002. **10**(10): p. 1435-51.
13. Jeffery, C.J., *The demise of catalysis, but new functions arise: pseudoenzymes as the phoenixes of the protein world.* Biochemical Society Transactions, 2019. **47**(1): p. 371-379.
14. Murphy, J.M., P.D. Mace, and P.A. Eyers, *Live and let die: insights into pseudoenzyme mechanisms from structure.* Current opinion in structural biology, 2017. **47**: p. 95-104.
15. Zeqiraj, E., et al., *Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation.* Science, 2009. **326**(5960): p. 1707-11.
16. Walden, M., et al., *Pseudo-DUBs as allosteric activators and molecular scaffolds of protein complexes.* Biochemical Society Transactions, 2018. **46**(2): p. 453-466.
17. Walden, *Metabolic control of SHMT2 oligomerization regulates ubiquitin-dependent cytokine signaling.* Nature, 2019. **In Press**.
18. Bertram, K., et al., *Cryo-EM structure of a human spliceosome activated for step 2 of splicing.* Nature, 2017. **542**(7641): p. 318-323.
19. Reiterer, V., et al., *Pseudophosphatase STYX modulates cell-fate decisions and cell migration by spatiotemporal regulation of ERK1/2.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(31): p. E2934-43.
20. Reiterer, V., et al., *The pseudophosphatase STYX targets the F-box of FBXW7 and inhibits SCFFBXW7 function.* The EMBO journal, 2017. **36**(3): p. 260-273.
21. Del-Toro, N., et al., *Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set.* Nature communications, 2019. **10**(1): p. 10.
22. Orchard, S., et al., *Protein interaction data curation: the International Molecular Exchange (IMEx) consortium.* Nature methods, 2012. **9**(4): p. 345-50.
23. Reiterer, V. and H. Farhan, *STYX: a pseudophosphatase that regulates MAPK signaling and ubiquitin ligases* Biochem Soc Trans, 2017. **In press**.
24. Kwon, A., et al., *Tracing the origin and evolution of pseudokinases across the tree of life.* Science signaling, 2019. **12**(578).
25. Wang, J., et al., *Reconstitution and structure of a plant NLR resistosome conferring immunity.* Science, 2019. **364**(6435).
26. Wang, J., et al., *Ligand-triggered allosteric ADP release primes a plant NLR complex.* Science, 2019. **364**(6435).

27. Klymiuk, V., A. Fatiukha, and T. Fahima, *Wheat tandem kinases provide insights on disease-resistance gene flow and host-parasite co-evolution.* The Plant journal : for cell and molecular biology, 2019.
28. Olson, A.T., et al., *A poxvirus pseudokinase represses viral DNA replication via a pathway antagonized by its paralog kinase.* PLoS pathogens, 2019. **15**(2): p. e1007608.
29. Consortium, T.U., *UniProt: a worldwide hub of protein knowledge.* Nucleic acids research, 2019. **47**(D1): p. D506-D515.
30. Zaru, R., M. Magrane, and C. O'Donovan, *From the research laboratory to the database: the Caenorhabditis elegans kinome in UniProtKB.* The Biochemical journal, 2017. **474**(4): p. 493-515.
31. Xu, B., et al., *WNK1, a novel mammalian serine/threonine protein kinase lacking the catalytic lysine in subdomain II.* The Journal of biological chemistry, 2000. **275**(22): p. 16795-801.
32. Min, X., et al., *Crystal structure of the kinase domain of WNK1, a kinase that causes a hereditary form of hypertension.* Structure, 2004. **12**(7): p. 1303-11.
33. Gosal, G., K.J. Kochut, and N. Kannan, *ProKinO: an ontology for integrative analysis of protein kinases in cancer.* PloS one, 2011. **6**(12): p. e28782.
34. Kannan, N., et al., *Structural and functional diversity of the microbial kinome.* PLoS biology, 2007. **5**(3): p. e17.
35. Cui, J., et al., *Structure of Fam20A reveals a pseudokinase featuring a unique disulfide pattern and inverted ATP-binding.* eLife, 2017. **6**.
36. Tagliabracci, V.S., et al., *A Single Kinase Generates the Majority of the Secreted Phosphoproteome.* Cell, 2015. **161**(7): p. 1619-32.
37. Zhang, H., et al., *Structure and evolution of the Fam20 kinases.* Nature communications, 2018. **9**(1): p. 1218.
38. Sreelatha, A., et al., *Protein AMPylation by an Evolutionarily Conserved Pseudokinase.* Cell, 2018. **175**(3): p. 809-821 e19.
39. Wen, J., et al., *Xylose phosphorylation functions as a molecular switch to regulate proteoglycan biosynthesis.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(44): p. 15723-8.
40. Zhu, Q., et al., *Structure of protein O-mannose kinase reveals a unique active site architecture.* eLife, 2016. **5**.
41. Black, M.H., et al., *Bacterial pseudokinase catalyzes protein polyglutamylation to inhibit the SidE-family ubiquitin ligases.* Science, 2019. **364**(6442): p. 787-792.
42. Qiu, J., et al., *A unique deubiquitinase that deconjugates phosphoribosyl-linked protein ubiquitination.* Cell research, 2017. **27**(7): p. 865-881.
43. Beraki, T., et al., *Divergent kinase regulates membrane ultrastructure of the Toxoplasma parasitophorous vacuole.* Proceedings of the National Academy of Sciences of the United States of America, 2019. **116**(13): p. 6361-6370.
44. Huang, Y., et al., *BubR1 phosphorylates CENP-E as a switch enabling the transition from lateral association to end-on capture of spindle microtubules.* Cell research, 2019.
45. Suijkerbuijk, S.J., et al., *The vertebrate mitotic checkpoint protein BUBR1 is an unusual pseudokinase.* Developmental cell, 2012. **22**(6): p. 1321-9.
46. Murphy, J.M., et al., *A robust methodology to subclassify pseudokinases based on their nucleotide-binding properties.* The Biochemical journal, 2014. **457**(2): p. 323-34.
47. Murphy, J.M., et al., *Insights into the evolution of divergent nucleotide-binding mechanisms among pseudokinases revealed by crystal structures of human and mouse MLKL.* The Biochemical journal, 2014. **457**(3): p. 369-77.
48. Clifton, B.E., et al., *Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein.* Nature chemical biology, 2018. **14**(6): p. 542-547.
49. Kaltenbach, M., et al., *Evolution of chalcone isomerase from a noncatalytic ancestor.* Nature chemical biology, 2018. **14**(6): p. 548-555.

50. Mukherjee, K., et al., *Evolution of CASK into a Mg2+-sensitive kinase.* Science signaling, 2010. **3**(119): p. ra33.
51. Byrne, D.P., D.M. Foulkes, and P.A. Eyers, *Pseudokinases: update on their functions and evaluation as new drug targets.* Future medicinal chemistry, 2017. **9**(2): p. 245-265.
52. Foulkes, D.M., et al., *Covalent inhibitors of EGFR family protein kinases induce degradation of human Tribbles 2 (TRIB2) pseudokinase in cancer cells.* Science Signaling, 2018. **11**(549).
53. Ribeiro, A.J.M., et al., *Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites.* Nucleic acids research, 2018. **46**(D1): p. D618-D623.
54. Potter, S.C., et al., *HMMER web server: 2018 update.* Nucleic acids research, 2018. **46**(W1): p. W200-W204.
55. Furnham, N., et al., *Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies.* Journal of molecular biology, 2016. **428**(2 Pt A): p. 253-267.
56. Kwon, *Tracing the origin and evolution of pseudokinases across the tree of life.* Science Signaling, 2019. **In Press**.
57. Sillitoe, I., et al., *CATH: expanding the horizons of structure-based functional annotations for genome sequences.* Nucleic acids research, 2019. **47**(D1): p. D280-D284.
58. Lewis, T.E., et al., *Gene3D: Extensive prediction of globular domains in proteins.* Nucleic acids research, 2018. **46**(D1): p. D1282.
59. Das, S., et al., *Functional classification of CATH superfamilies: a domain-based approach for protein function annotation.* Bioinformatics, 2015. **31**(21): p. 3460-7.
60. Jiang, Y., et al., *An expanded evaluation of protein function prediction methods shows an improvement in accuracy.* Genome biology, 2016. **17**(1): p. 184.
61. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature genetics, 2000. **25**(1): p. 25-9.
62. Sharir-Ivry, A. and Y. Xia, *Nature of Long-Range Evolutionary Constraint in Enzymes: Insights from Comparison to Pseudoenzymes with Similar Structures.* Molecular biology and evolution, 2018. **35**(11): p. 2597-2606.
63. Hwang, J., et al., *Crystal structure of the human N-Myc downstream-regulated gene 2 protein provides insight into its role as a tumor suppressor.* The Journal of biological chemistry, 2011. **286**(14): p. 12450-60.
64. Bailey, F.P., et al., *The Tribbles 2 (TRB2) pseudokinase binds to ATP and autophosphorylates in a metal-independent manner.* The Biochemical journal, 2015. **467**(1): p. 47-62.
65. Ungureanu, D., et al., *The pseudokinase domain of JAK2 is a dual-specificity protein kinase that negatively regulates cytokine signaling.* Nature structural & molecular biology, 2011. **18**(9): p. 971-6.
66. Buchholz, K.R., P.W. Bowyer, and J.C. Boothroyd, *Bradyzoite pseudokinase 1 is crucial for efficient oral infectivity of the Toxoplasma gondii tissue cyst.* Eukaryotic cell, 2013. **12**(3): p. 399-410.
67. Hanks, S.K. and T. Hunter, *Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification.* FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 1995. **9**(8): p. 576-96.
68. Walker, J.E., et al., *Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold.* The EMBO journal, 1982. **1**(8): p. 945-51.
69. Stefely, J.A., et al., *Mitochondrial ADCK3 employs an atypical protein kinase-like fold to enable coenzyme Q biosynthesis.* Molecular cell, 2015. **57**(1): p. 83-94.
70. Cameron, A.J., et al., *PKC maturation is promoted by nucleotide pocket occupation independently of intrinsic kinase activity.* Nature structural & molecular biology, 2009. **16**(6): p. 624-30.

71. Otto, T., et al., *Stabilization of N-Myc is a critical function of Aurora A in human neuroblastoma.* Cancer cell, 2009. **15**(1): p. 67-78.
72. Murphy, J.M., et al., *The pseudokinase MLKL mediates necroptosis via a molecular switch mechanism.* Immunity, 2013. **39**(3): p. 443-53.
73. Patel, O., et al., *Structure of SgK223 pseudokinase reveals novel mechanisms of homotypic and heterotypic association.* Nature communications, 2017. **8**(1): p. 1157.
74. Murphy, J.M., et al., *Molecular Mechanism of CCAAT-Enhancer Binding Protein Recruitment by the TRIB1 Pseudokinase.* Structure, 2015. **23**(11): p. 2111-21.
75. Jamieson, S.A., et al., *Substrate binding allosterically relieves autoinhibition of the pseudokinase TRIB1.* Science Signaling, 2018. **11**(549).
76. Eyers, P.A., K. Keeshan, and N. Kannan, *Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease.* Trends in cell biology, 2017. **27**(4): p. 284-298.
77. Jamieson, S.A., et al., *Substrate binding allosterically relieves autoinhibition of the pseudokinase TRIB1.* Sci Signal, 2018. **11**(549).
78. Zeqiraj, E., et al., *ATP and MO25alpha regulate the conformational state of the STRADalpha pseudokinase and activation of the LKB1 tumour suppressor.* PLoS Biol, 2009. **7**(6): p. e1000126.
79. Littlefield, P., et al., *Structural analysis of the EGFR/HER3 heterodimer reveals the molecular basis for activating HER3 mutations.* Sci Signal, 2014. **7**(354): p. ra114.
80. Lupardus, P.J., et al., *Structure of the pseudokinase-kinase domains from protein kinase TYK2 reveals a mechanism for Janus kinase (JAK) autoinhibition.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. 8025-30.
81. Rostislavleva, K., et al., *Structure and flexibility of the endosomal Vps34 complex reveals the basis of its function on membranes.* Science, 2015. **350**(6257): p. aac7365.
82. Edmund, A.B., et al., *The pseudokinase domains of guanylyl cyclase-A and -B allosterically increase the affinity of their catalytic domains for substrate.* Sci Signal, 2019. **12**(566).
83. Murphy, J.M., et al., *The pseudokinase MLKL mediates necroptosis via a molecular switch mechanism.* Immunity, 2013. **39**(3): p. 443-453.
84. Jacobsen, A.V., et al., *HSP90 activity is required for MLKL oligomerisation and membrane translocation and the induction of necroptotic cell death.* Cell Death Dis, 2016. **7**: p. e2051.
85. Kung, J.E. and N. Jura, *The pseudokinase TRIB1 toggles an intramolecular switch to regulate COP1 nuclear export.* The EMBO journal, 2019. **38**(4).
86. Ha, B.H. and T.J. Boggon, *The crystal structure of pseudokinase PEAK1 (Sugen kinase 269) reveals an unusual catalytic cleft and a novel mode of kinase fold dimerization*. J Biol Chem, 2018. **293**(5): p. 1642-1650.
87. Lecointre, C., et al., *Dimerization of the Pragmin Pseudo-Kinase Regulates Protein Tyrosine Phosphorylation.* Structure, 2018. **26**(4): p. 545-554 e4.
88. Ha, B.H. and T.J. Boggon, *The crystal structure of pseudokinase PEAK1 (Sugen kinase 269) reveals an unusual catalytic cleft and a novel mode of kinase fold dimerization.* The Journal of biological chemistry, 2018. **293**(5): p. 1642-1650.
89. Liu, L., et al., *Homo- and Heterotypic Association Regulates Signaling by the SgK269/PEAK1 and SgK223 Pseudokinases.* The Journal of biological chemistry, 2016. **291**(41): p. 21571-21583.
90. Safari, F., et al., *Mammalian Pragmin regulates Src family kinases via the Glu-Pro-Ile-Tyr-Ala (EPIYA) motif that is exploited by bacterial effectors.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(36): p. 14938-43.
91. Tanaka, H., H. Katoh, and M. Negishi, *Pragmin, a novel effector of Rnd2 GTPase, stimulates RhoA activity.* The Journal of biological chemistry, 2006. **281**(15): p. 10355-64.
92. Stefely, J.A., et al., *Cerebellar Ataxia and Coenzyme Q Deficiency through Loss of Unorthodox Kinase Activity.* Molecular cell, 2016. **63**(4): p. 608-620.

93. Childers, W.S., et al., *Cell fate regulation governed by a repurposed bacterial histidine kinase.* PLoS Biol, 2014. **12**(10): p. e1001979.
94. Ng, A.A., et al., *The CMT4B disease-causing phosphatases Mtmr2 and Mtmr13 localize to the Schwann cell cytoplasm and endomembrane compartments, where they depend upon each other to achieve wild-type levels of protein expression.* Hum Mol Genet, 2013. **22**(8): p. 1493-506.
95. Reiterer, V., et al., *The pseudophosphatase STYX targets the F-box of FBXW7 and inhibits SCFFBXW7 function.* EMBO J, 2017. **36**(3): p. 260-273.
96. Blanchetot, C., L.G. Tertoolen, and J. den Hertog, *Regulation of receptor protein-tyrosine phosphatase alpha by oxidative stress.* The EMBO journal, 2002. **21**(4): p. 493-503.
97. Chang, D.W., et al., *c-FLIP(L) is a dual function regulator for caspase-8 activation and CD95-mediated apoptosis.* EMBO J, 2002. **21**(14): p. 3704-14.
98. Adrain, C., et al., *Tumor necrosis factor signaling requires iRhom2 to promote trafficking and activation of TACE.* Science, 2012. **335**(6065): p. 225-8.
99. Christova, Y., et al., *Mammalian iRhoms have distinct physiological functions including an essential role in TACE regulation.* EMBO Rep, 2013. **14**(10): p. 884-90.
100. Klemencic, M., et al., *Phylogenetic Distribution and Diversity of Bacterial Pseudo-Orthocaspases Underline Their Putative Role in Photosynthesis.* Frontiers in plant science, 2019. **10**: p. 293.
101. Piao, S., et al., *Crystal structure of a clip-domain serine protease and functional roles of the clip domains.* The EMBO journal, 2005. **24**(24): p. 4404-14.
102. Geng, X., et al., *Caenorhabditis elegans caspase homolog CSP-2 inhibits CED-3 autoactivation and apoptosis in germ cells.* Cell death and differentiation, 2009. **16**(10): p. 1385-94.
103. Geng, X., et al., *Inhibition of CED-3 zymogen activation and apoptosis in Caenorhabditis elegans by caspase homolog CSP-3.* Nature structural & molecular biology, 2008. **15**(10): p. 1094-101.
104. Eddins, M.J., et al., *Mms2-Ubc13 covalently bound to ubiquitin reveals the structural basis of linkage-specific polyubiquitin chain formation.* Nat Struct Mol Biol, 2006. **13**(10): p. 915-20.
105. Lingaraju, G.M., et al., *Crystal structure of the human COP9 signalosome.* Nature, 2014. **512**(7513): p. 161-5.
106. Worden, E.J., C. Padovani, and A. Martin, *Structure of the Rpn11-Rpn8 dimer reveals mechanisms of substrate deubiquitination during proteasomal degradation.* Nature structural & molecular biology, 2014. **21**(3): p. 220-7.
107. Pathare, G.R., et al., *Crystal structure of the proteasomal deubiquitylation module Rpn8-Rpn11.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**(8): p. 2984-9.
108. Schafer, I.B., et al., *The structure of the Pan2-Pan3 core complex reveals cross-talk between deadenylase and pseudokinase.* Nature structural & molecular biology, 2014. **21**(7): p. 591-8.
109. Ceccarelli, D.F., et al., *FAM105A/OTULINL Is a Pseudodebuiquitinase of the OTU-Class that Localizes to the ER Membrane.* Structure, 2019.
110. Strickland, M., et al., *Tsg101 chaperone function revealed by HIV-1 assembly inhibitors.* Nat Commun, 2017. **8**(1): p. 1391.
111. Walden, M., et al., *Metabolic control of BRISC-SHMT2 assembly regulates immune signalling.* Nature, 2019. **570**(7760): p. 194-199.
112. Mandel, C.R., et al., *Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease.* Nature, 2006. **444**(7121): p. 953-6.
113. Hillen, H.S., et al., *Mechanism of Transcription Anti-termination in Human Mitochondria.* Cell, 2017. **171**(5): p. 1082-1093 e13.
114. Lazzaretti, D., et al., *The bicoid mRNA localization factor Exuperantia is an RNA-binding pseudonuclease.* Nat Struct Mol Biol, 2016. **23**(8): p. 705-13.

115. Rosenberg, O.S., et al., *Substrates Control Multimerization and Activation of the Multi-Domain ATPase Motor of Type VII Secretion.* Cell, 2015. **161**(3): p. 501-12.
116. Foster, R., et al., *Identification of a novel human Rho protein with unusual properties: GTPase deficiency and in vivo farnesylation.* Mol Cell Biol, 1996. **16**(6): p. 2689-99.
117. Wennerberg, K., et al., *Rnd proteins function as RhoA antagonists by activating p190 RhoGAP.* Curr Biol, 2003. **13**(13): p. 1106-15.
118. Stiegler, A.L. and T.J. Boggon, *PseudoGTPase domains in p190RhoGAP proteins: a mini-review.* Biochem Soc Trans, 2018. **46**(6): p. 1713-1720.
119. Stiegler, A.L. and T.J. Boggon, *The N-Terminal GTPase Domain of p190RhoGAP Proteins Is a PseudoGTPase.* Structure, 2018. **26**(11): p. 1451-1461 e4.
120. Basilico, F., et al., *The pseudo GTPase CENP-M drives human kinetochore assembly.* Elife, 2014. **3**: p. e02978.
121. Schroeder, C.M., et al., *A Ras-like domain in the light intermediate chain bridges the dynein motor to a cargo-binding region.* Elife, 2014. **3**: p. e03351.
122. Ranok, A., et al., *Structural and thermodynamic insights into chitooligosaccharide binding to human cartilage chitinase 3-like protein 2 (CHI3L2 or YKL-39).* J Biol Chem, 2015. **290**(5): p. 2617-29.
123. Schimpl, M., et al., *Human YKL-39 is a pseudo-chitinase with retained chitooligosaccharide-binding properties.* Biochem J, 2012. **446**(1): p. 149-57.
124. Wong, W., et al., *Structure of Plasmodium falciparum Rh5-CyRPA-Ripr invasion complex.* Nature, 2019. **565**(7737): p. 118-121.
125. Kuzina, E.S., et al., *Structures of ligand-occupied beta-Klotho complexes reveal a molecular mechanism underlying endocrine FGF specificity and activity*. Proceedings of the National Academy of Sciences of the United States of America, 2019.
126. Velez, N., C.A. Brautigam, and M.A. Phillips, *Trypanosoma brucei S-adenosylmethionine decarboxylase N terminus is essential for allosteric activation by the regulatory subunit prozyme.* J Biol Chem, 2013. **288**(7): p. 5232-40.
127. Volkov, O.A., et al., *Relief of autoinhibition by conformational switch explains enzyme activation by a catalytically dead paralog.* Elife, 2016. **5**.
128. Willert, E.K., R. Fitzpatrick, and M.A. Phillips, *Allosteric regulation of an essential trypanosome polyamine biosynthetic enzyme by a catalytically dead homolog.* Proc Natl Acad Sci U S A, 2007. **104**(20): p. 8275-80.
129. He, S., et al., *Viral pseudo-enzymes activate RIG-I via deamidation to evade cytokine production.* Mol Cell, 2015. **58**(1): p. 134-46.
130. Nguyen, S., et al., *Allosteric activation of trypanosomatid deoxyhypusine synthase by a catalytically dead paralog.* J Biol Chem, 2013. **288**(21): p. 15256-67.
131. Afanador, G.A., D.R. Tomchick, and M.A. Phillips, *Trypanosomatid Deoxyhypusine Synthase Activity Is Dependent on Shared Active-Site Complementation between Pseudoenzyme Paralogs.* Structure, 2018. **26**(11): p. 1499-1512 e5.
132. Richter, V., et al., *Structural and functional analysis of MiD51, a dynamin receptor required for mitochondrial fission.* The Journal of cell biology, 2014. **204**(4): p. 477-86.
133. Yao, J., et al., *Human pantothenate kinase 4 is a pseudo-pantothenate kinase.* Protein science : a publication of the Protein Society, 2019.
134. Allali-Hassani, A., et al., *Structural and chemical profiling of the human cytosolic sulfotransferases.* PLoS biology, 2007. **5**(5): p. e97.
135. Rao, F.V., et al., *Structure of a bacterial putative acetyltransferase defines the fold of the human O-GlcNAcase C-terminal domain.* Open Biol, 2013. **3**(10): p. 130021.
136. Bozatzi, P. and G.P. Sapkota, *The FAM83 family of proteins: from pseudo-PLDs to anchors for CK1 isoforms.* Biochem Soc Trans, 2018. **46**(3): p. 761-771.
137. Fulcher, L.J., et al., *The DUF1669 domain of FAM83 family proteins anchor casein kinase 1 isoforms.* Sci Signal, 2018. **11**(531).

138. Banumathi, S., et al., *Structure of the neurotoxic complex vipoxin at 1.4 A resolution.* Acta Crystallogr D Biol Crystallogr, 2001. **57**(Pt 11): p. 1552-9.
139. Larson, H.N., H. Weiner, and T.D. Hurley, *Disruption of the coenzyme binding site and dimer interface revealed in the crystal structure of mitochondrial aldehyde dehydrogenase "Asian" variant.* J Biol Chem, 2005. **280**(34): p. 30550-6.
140. Teoh, M.L., P.J. Walasek, and D.H. Evans, *Leporipoxvirus Cu,Zn-superoxide dismutase (SOD) homologs are catalytically inert decoy proteins that bind copper chaperone for SOD.* J Biol Chem, 2003. **278**(35): p. 33175-84.
141. Liu, W., et al., *Mechanism for KRIT1 release of ICAP1-mediated suppression of integrin activation.* Molecular cell, 2013. **49**(4): p. 719-29.
142. Elkhateeb, E., et al., *The role of mouse 2',5'-oligoadenylate synthetase 1 paralogs.* Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 2016. **45**: p. 393-401.
143. Courtney, S.C., et al., *Identification of novel host cell binding partners of Oas1b, the protein conferring resistance to flavivirus-induced disease in mice.* Journal of virology, 2012. **86**(15): p. 7953-63.

ACKNOWLEDGEMENTS:

We thank all the scientists and support staff who organized, attended and contributed to the scientific input at “Pseudoenzymes 2016: from signalling mechanisms to disease”, which took place at the Liverpool Maritime Museum, UK, in September 2016 and “Pseudoenzymes 2018: From molecular mechanisms to cell biology”, which took place in May 2018 in Sardinia, Italy, both of which informed this review.

FUNDING:

This work was initially funded by a Royal Society Research Grant (to PAE) and subsequently by North West Cancer Research grants CR1088 and CR1097 (to PAE). We thank the Biochemical Society and EMBO for awarding dedicated conference and workshop funding to support the development of the pseudoenzyme field.

AUTHOR CONTRIBUTIONS:

AB, SD, ND, RZ, SO, JT and CO performed bioinformatics analyses, and created databases and websites. EZ, JMM and PAE assembled the information in Table 1. All authors co-wrote the manuscript, and all authors approved the final version prior to submission.

COMPETING INTERESTS:

The authors declare that they have no conflicts of interest related to the content of this review.

FIGURE LEGENDS:

Figure 1. Regulated assembly of the macromolecular BRISC-SHMT2 pseudoenzyme-containing complex. Top, schematic of the SHMT2 dimer-tetramer oligomerization transition, which is regulated by pyridoxal-5'-phosphate (PLP), the active co-factor form of vitamin B_6. The SHMT2 dimer, which is inactive as a methyltransferase (left) specifically interacts with, and inhibits, the pseudoenzyme-containing BRISC complex (bottom), revealing a new 'moonlighting' role for SHMT2. The PLP-bound SHMT2 tetramer, which is active as a methyltransferase, is unable to bind or inhibit the BRISC complex. Bottom, schematic of the multimeric BRISC-SHMT2 DUB complex, which contains the active DUB BRCC36 (MPN+) and the inactive Abraxas2 pseudo-DUB (MPN-). BRCC45 contains three pseudo-E2 domains (UEVs), and is discussed further in Table 1 and in the text.

Figure 2. Computational annotation of enzymes and pseudoenzymes in UniProt. (A) Workflow for expert curation of pseudoenzymes. Reported evidence is based on experimental data, ortholog similarity and sequence analysis. **(B)** Worked example of the "inactive" annotation for the *C.elegans* pseudophosphatase egg-4 (UniProt KB O01767), including the protein name, the protein-coding evidence (status) and the caution comment.

Figure 3. **Analyzing the (pseudo)kinome and (pseudo)phosphatome in a model worm. (A)** Calculated percentage of pseudokinases in the *C. elegans* kinome (grey). **(B)** Percentage of phosphatases in the *C. elegans* phosphatome (grey).

Figure 4. Diversity in ATP-binding mode and the acquisition of non-canonical catalytic functions amongst bioinformatically-annotated pseudokinases. (A) The catalytically-inactive pseudokinase, FAM20A, still binds ATP (red sticks), but in an inverted conformation [35] (PDB, 5wrs). Like the related FAM20B, the position of the αC helix (orange) differs from that in canonical protein kinases. **(B)** The highly atypical annotated pseudokinase SelO [38] (PDB, 6eac), can catalyse protein AMPylation via an unusual catalytic mechanism. Like FAM20A **(A)**, SelO binds ATP in an inverted conformation, but in addition, it catalyses AMP transfer to protein substrates. ATP analog AMP-PNP is shown as red sticks; Mg^{2+} and Ca^{2+} are shown as yellow and green spheres, respectively. **(C)** The predicted pseudokinase, FAM20B (PDB, 5xoo), is actually a catalytically-active xylose kinase [37] involved in proteoglycan synthesis. The sugar substrate is shown as green sticks; an adenine (red sticks) is also modelled in the structure. **(D)** Like FAM20B, the predicted pseudokinase, SgK169/protein *O*-mannose kinase (POMK; PDB, 5gza) is actually a sugar kinase [40]. SgK169/POMK closely resembles a typical protein kinase fold with conventional αC helix (orange) position, nucleotide-binding mode (red sticks) and Mg^{2+} cofactors (yellow spheres), and excitingly, this structure captures the protein bound to its sugar substrate (green sticks).

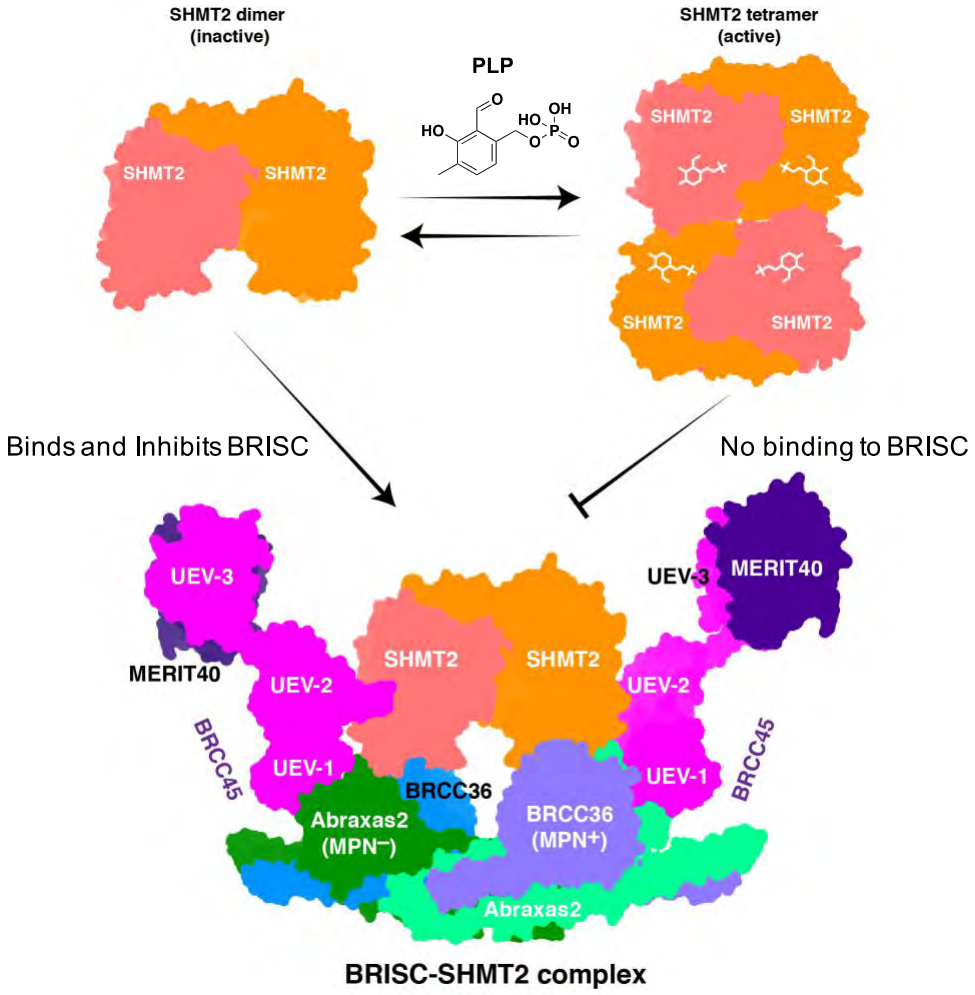
Figure 5. Estimating the ratio of pseudoenzymes in known enzyme families.
A family is defined here as the group of Swiss-Prot sequences that are homologous to one enzyme/entry in M-CSA. Each entry in M-CSA corresponds to one enzyme with a unique enzyme mechanism, so the same EC number can be represented more than once if evolved independently with distinct mechanisms. Sequences are categorised as enzymes if they have a catalytic UniProt keyword and as pseudoenzymes otherwise. The green bar corresponds to enzyme families that contain only enzymes, and the blue bars to enzyme families that contain a variable percentage of pseudoenzymes.

Figure 6. Annotated phylogenetic tree for soybean β-amylase.
Phylogenetic tree for homologous sequences of soybean β-amylase (P10538) annotated with the catalytic residues identified in M-CSA, their EC numbers and PFAM domains. The tree contains 20 homologues from plants, of which 6 are chloroplastic and 5 are bacterial. The two "inactive" pseudoenzymes identified (Q9FM68 and Q8VYW2) belong to Thale Cress (*Arabidopsis thaliana*), and one of these (Q9FM68) is known to have physiological regulatory functions in this organism.

Figure 7. Analysis of the distribution of pseudoenzymes. (A) Sequences in CATH superfamilies are sub-classified into functional families (FunFams) predicted to share similar structures and functions, and which can be used to understand protein function evolution. **(B)** Distribution of the number of enzyme superfamilies (containing catalytic domains) that have varying proportions of functional families with enzyme annotations. **(C)** Showing the number of putative pseudoenzyme families in enzyme superfamilies (containing catalytic domains). These remain to be confirmed by further analysis and experimental testing.

Figure 8: Pseudoenzymes in the NTF2 family. An example comparing the known pseudoenzyme family of NTF2 (shown in blue) and the related enzyme family of scytalone hydratases (shown in orange) in CATH superfamily 3.10.450.50. The established catalytic residues for the Scytalone hydratase family (EC 4.2.1.94) are shown in red. The structural alignment between the structures 1stdA00 and 1ounA00 from the enzyme and pseudoenzyme families was generated using CATH-superpose (https://cath-tools.readthedocs.io/en/latest/tools/cath-superpose/). On the right-hand panel, the highly conserved residues in the two families are shown for structurally equivalent positions lying in the active site of the enzyme family. The height of the characters reflects the degree of conservation and the colours change according to physicochemical characteristics.

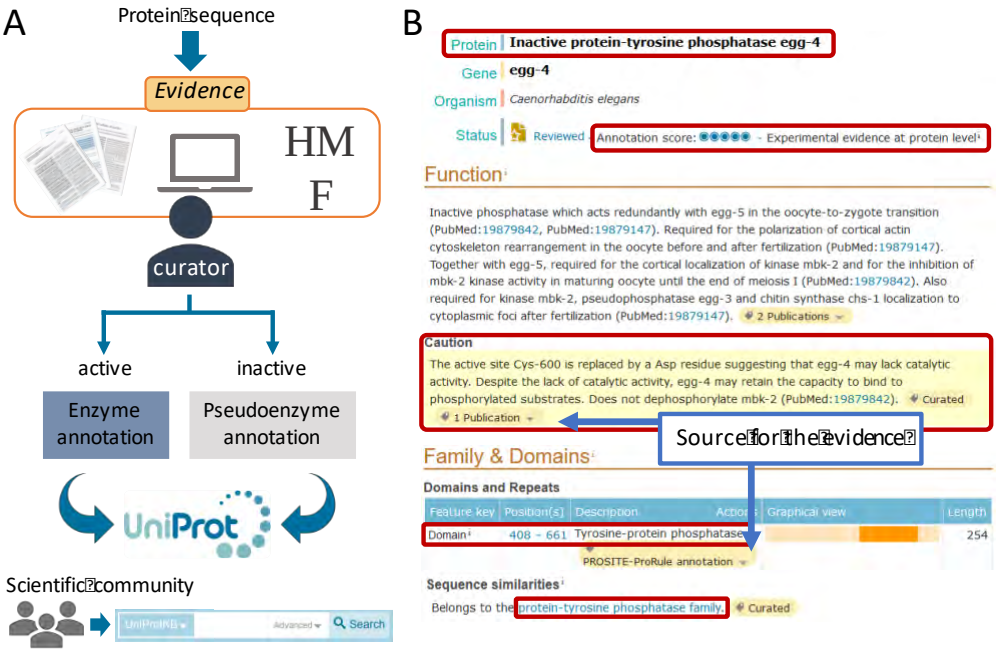
Figure 1



1053

1054

Figure 2



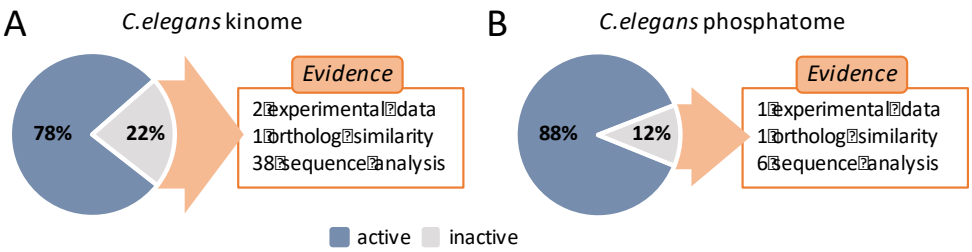
1055

1056

1057

1058

Figure 3



1059
1060

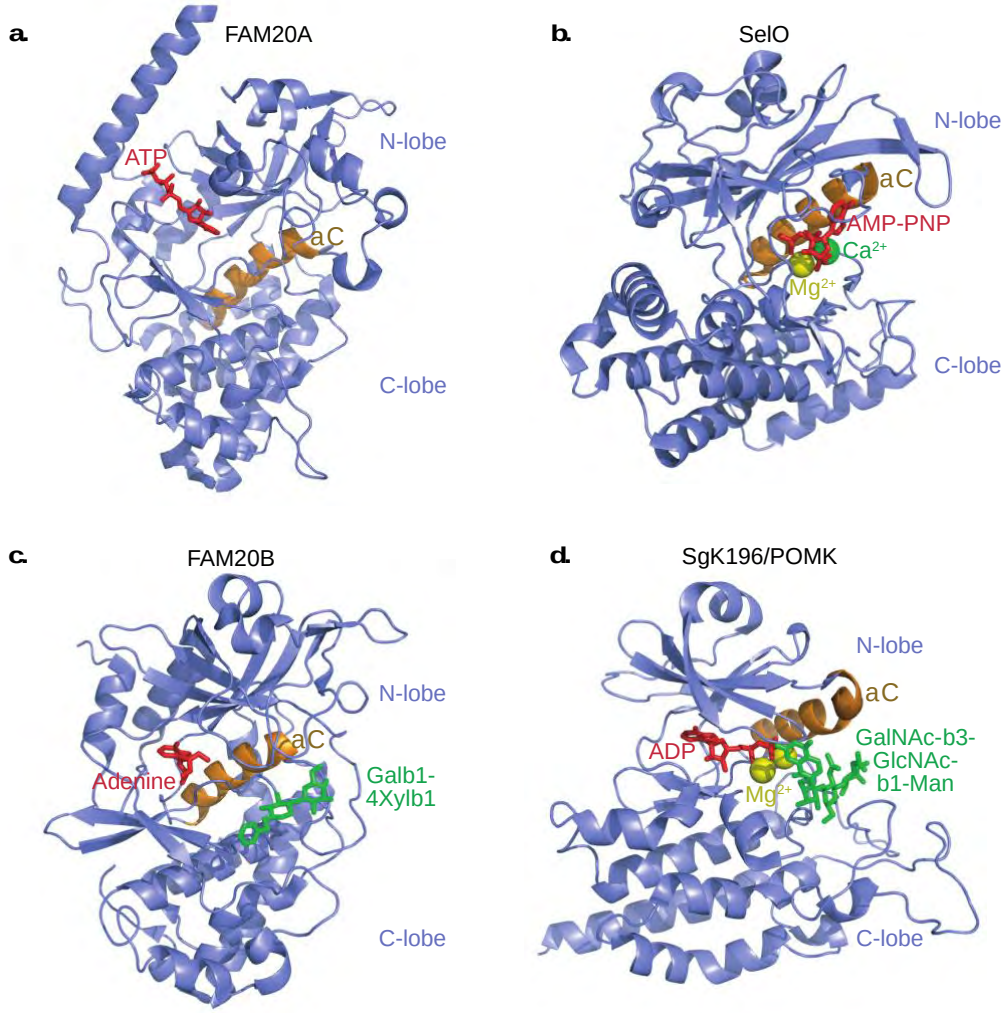
1061

1062

1063

1064

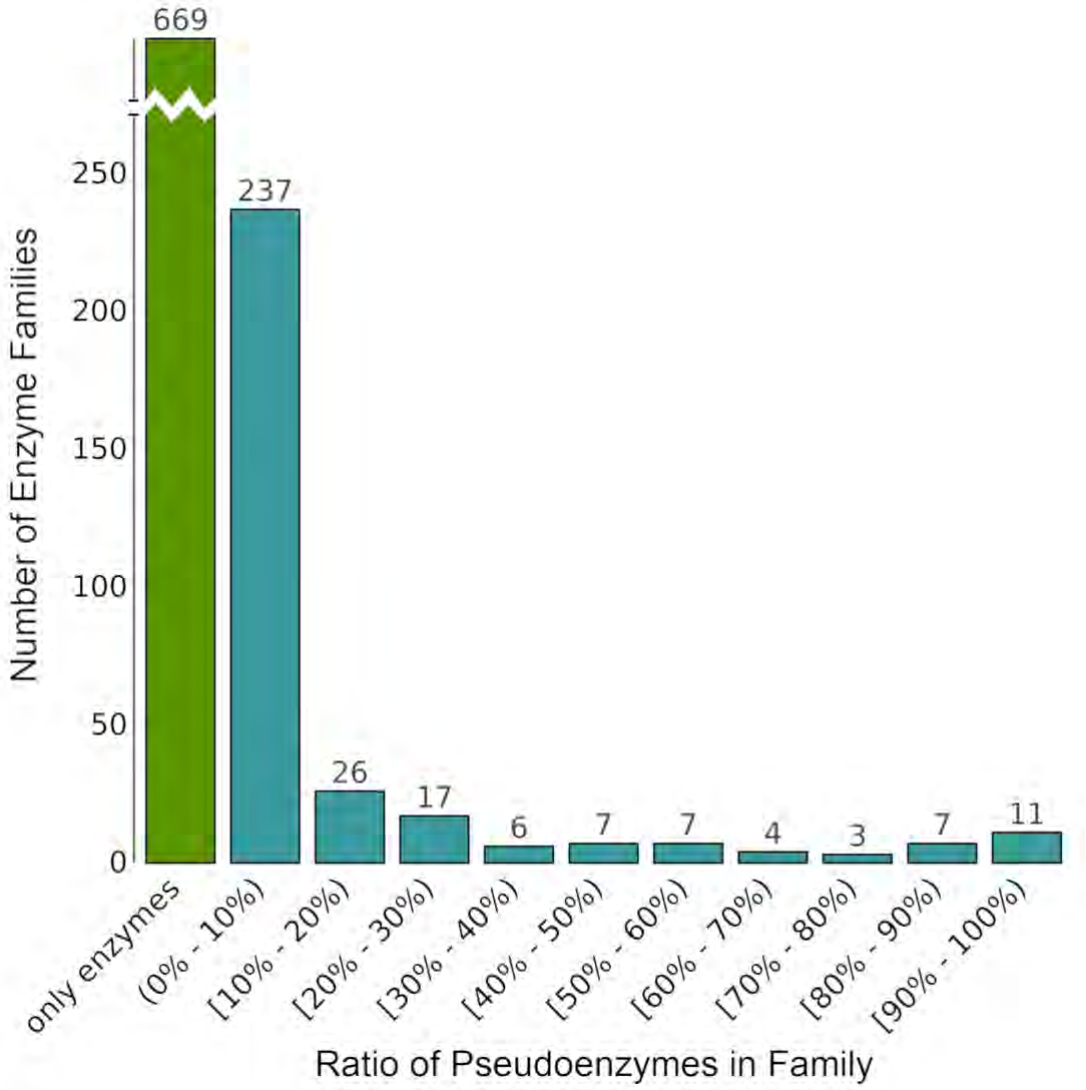
Figure 4



1065

1066

Figure 5


669
237
26
17
6
7
7
4
3
7
11
250
200
150
100
50
0
Number of Enzyme Families
only enzymes
(0% - 10%)
[10% - 20%)
[20% - 30%)
[30% - 40%)
[40% - 50%)
[50% - 60%)
[60% - 70%)
[70% - 80%)
[80% - 90%)
[90% - 100%)
Ratio of Pseudoenzymes in Family

1074

Figure 6

1075

1076

1077

1078

1079

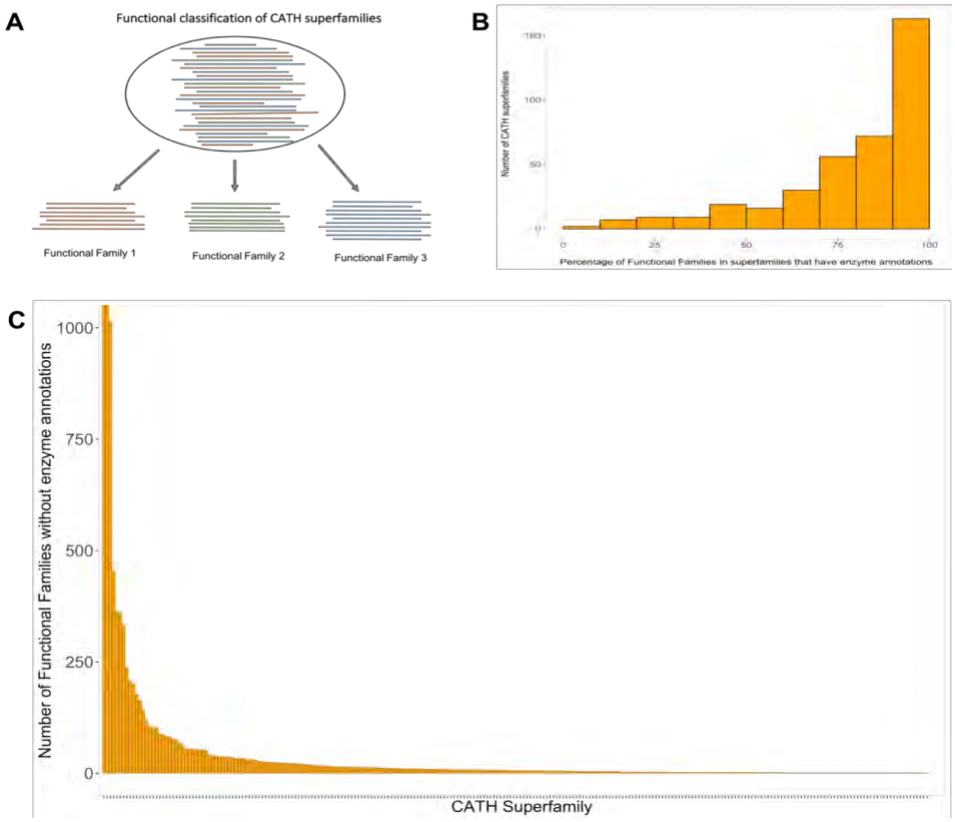
1080

1081

1082

1083

Figure 7

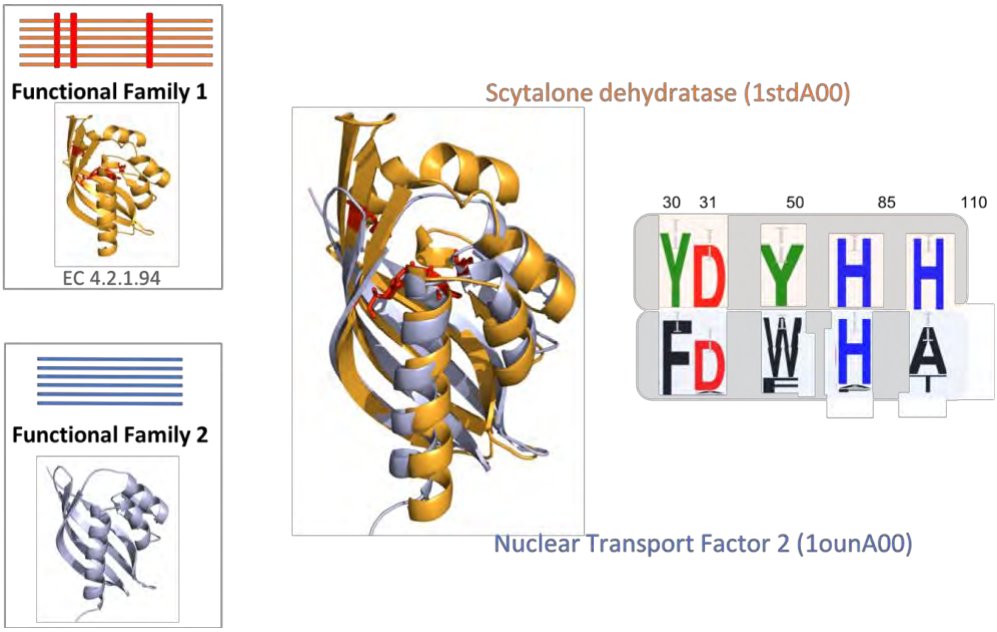


1084

1085

1086

Figure 8



1087

Table 1: The remarkable diversity of pseudoenzymes. Examples of pseudoenzymes from across the kingdoms of life, organized by class and function. Pseudoenzymes are highlighted in blue, whereas relevant conventional enzymes are in black. A broad selection of well-studied pseudoenzymes are discussed; the list is not meant to be comprehensive.

Class	Function	Examples	References
Pseudokinase	Allosteric regulation of conventional protein kinase	STRADα:LKB1, HER3:HER2, JAK JH2:JAK JH1, FAM20A:FAM20C, where a pseudokinase domain regulates activity of a conventional kinase in a cognate pair. Also a common mechanism in plant pseudokinases (including tandem pseudokinases) such as RKS1	[25-27, 35, 56, 78-80]
	Regulator of a phospholipid kinae	Vps15 is a probable pseudokinase that forms part of a multimeric complex for regulation of the Vps34 Phosphatidylinositol 3-kinase	[81]
	Allosteric regulation of other enzymes	Pseudokinase domains of Guanylyl cyclase-A and -B regulate activity of the tandem guanylyl cyclase domains	[82]
	Molecular switch	Phosphorylation of the MLKL pseudokinase domain triggers exposure of the executioner four-helix bundle domain and cell death	[47, 83]
	Protein interaction domain	MLKL pseudokinase domain is regulated by binding to the RIPK3 kinase domain and HSP90:Cdc37 co-chaperones	[73, 84]
	Scaffold for assembly of signalling complexes	Tribbles (TRIB) pseudokinases nucleate assembly of a complex between a substrate (e.g. C/EBPα) and the E3 Ubiquitin ligase, COP1, whose intrasubcellular localization is controlled by TRIB1	[74-76, 85] [73, 86]
		SgK223 (Pragmin)/SgK269 (PEAK1) form higher-order signaling assemblies that include Src-family kinases	[73, 87-91] [69, 92]

	Fundamental metabolic regulators of isoprenoid lipid production	UbiB pseudoenzyme family adopts an (inactive?) atypical protein kinase-like (PKL)-fold found in bacteria, archaea and eukaryotes. Human mitochondrial ADCK3 pseudokinase binds nucleotides such as ADP, but can be re-engineered into an ATP-dependent autophosphorylating enzyme. Also relevant to the yeast Coq8p ATPase.	
Pseudo-Histidine kinase	Protein interaction domain	*Caulobacter* DivL binds the response regulator, DivK, to regulate asymmetric cell division	[93]
Pseudophosphatase	Occlusion of conventional phosphatase access to substrate and subcellular localisation	STYX binds ERK1/2 kinase to occlude DUSP4 binding and anchors ERK1/2 in the nucleus	[23]
	Regulation of protein localisation in a cell	MTMR13 stabilizes lipid phosphatase MTMR2 and localizes MTMR2 to membranes	[94]
	Regulation of signalling complex assembly	STYX prevents substrate recruitment via FBXW7 to SCF E3 ligase complex	[95]
	Redox sensor as part of tandem enzyme:pseudoenzyme domains.	A Receptor Protein Tyrosine Phosphatase (RPTP) alpha redox-active Cys residue in the D2 (pseudophosphatase) domain controls the catalytic output of the D1 (canonical phosphatase) domain after exposure to ROS	[96]
Pseudoprotease	Allosteric regulator of conventional protease	cFLIP binds the cysteine protease, Caspase-8, to block apoptosis	[97]
	Regulation of protein localisation in a cell	Mammalian iRhom proteins bind and regulate membrane trafficking of receptors	[98, 99]
	Anoxygenic photosynthesis?	Mulitple pseudo-orthocaspases disclosed (prokaryotic)	[100]
	Activation of the prohenoloxidase cascade	Noncatalytic clip-domain SP family member PPAF-II from *H. diomphalia* binds and activates processed pro-phenoloxidases PPO1 and PPO2	[101]
	Prevention of zymogen activation	Caspase-like pseudoproteases csp-2 and csp-3 prevent ced-3 autoactivation in *C. elegans*	[102, 103]

Pseudodeubiquitinase (pseudoDUB)	Allosteric regulator of conventional DUB	Abraxas1 and 2 nucleate assembly of a higher order heterotetramer with BRCC36 active DUB (see also Fig. 1)	[104]
	Allosteric regulator of conventional enzyme	CSN6 supports activity of CSN5 and acts as a scaffold for the Cop9 signallosome complex	[105]
	Allosteric regulator of conventional DUB	PSMD7 (Rpn8 in yeast) supports the activity of the active DUB PSMD14 (Rpn11). Part of the 19S regulatory particle of the proteasome	[106, 107]
	Scaffold for assembly of enzyme complexes	PRP8 (Prpf8 in yeast) is central to the assembly of the spliceosome complex required for mRNA splicing. Prp8 contains three additional pseudoenzyme domains (pseudo-endonuclease, pseudo-reverse transcriptase, pseudo-RNaseH)	[16]
	Scaffold for assembly of enzyme complexes	USP39 (Sad1 in yeast) is a member of the spliceosome complex with presumed scaffolding roles	[16]
	Scaffold for assembly of enzyme complexes	eIF3f and eIF3h are both inactive and integral for the assembly of eIF3 complex required for protein synthesis	[16]
	Scaffold for assembly of enzyme complexes	USP52 (Pan2 in yeast) serves as a scaffold for the Pan2-Pan3 deadenylating complex	[16, 108]

	Unknown function	FAM105A is an OTU domain containing protein that is closest to OTULIN (a.k.a. FAM105B or Gumby). Likely functions as a protein interactin domain the ER/Golgi. The MINDY4b pseudoenzyme is also part of the broader MINDY DUB family.	[16, 109]
Pseudo Ubiqutin Conjugating enzyme (pseudo-E2)	Allosteric regulator of conventional E2 ligase	The ubiquitin E2 variant (UEV) domain of Mms2 binds the active E2, Ubc13, to enable catalytic activity and dictate K63-linked ubiquitin linkages	[104]
	Regulation of protein localization in a cell	The Tsg101 UEV domain exerts a chaperone function, including during viral replication	[110]
	Scaffold for assembly of signaling complexes	BRCC45 contains three tandem UEV domains which support protein-protein interactions within the BRISC-SHMT2 complex required for IFN signalling and the BRCA1-A complex required for DNA damage repair	[111]
Pseudonuclease	Allosteric regulator of conventional nuclease	CPSF-100 is a component of the pre-mRNA 3′ end processing complex containing the active counterpart, CPSF-73	[112]
	Nucleating and regulating assembly of protein:nucleic acid complexes	TEFM contributes to the elongation complex sliding clamp and sequesters ssDNA to maintain the mitochondrial transcription bubble	[113]
	Regulation of substrate localization	Exuperantia pseudo-exonuclease domain regulates cellular localization of the *bicoid* mRNA in *Drosophila* cells	[114]
PseudoATPase	Allosteric regulator of conventional ATPase	EccC comprises two pseudoATPase domains that regulate the N-terminal conventional ATPase domain	[115]

PseudoGTPase	Allosteric regulator of conventional GTPase	GTP-bound Rnd1 or Rnd3/RhoE bind p190RhoGAP to regulate the catalytic activity of the GTPase, RhoA	[116, 117]
	Allosteric control of conventional enzyme function	Three GTPase-like domains promote p190RhoGAP GAP activity	[118, 119]
	Scaffold for assembly of signaling complexes	CENP-M cannot bind GTP or switch conformations, but is essential for regulating kinetochore assembly	[120]
	Regulation of protein localization in a cell	Yeast light intermediate domain (LIC) is a pseudoGTPase, devoid of nucleotide binding, which connects the dynein motor protein to cargo proteins	[121]
Pseudochitinase	Substrate recruitment or sequestration	YKL-39 binds chitooligosaccharides via 5 binding subsites, but does not process these as substrates	[122, 123]
Pseudosialidase	Scaffold for assembly of signalling complexes	*P. falciparum* CyRPA nucleates assembly of the PfRh5/PfRipr complex that binds the host erythrocyte receptor, basigin, and mediates host cell invasion	[124]
Pseudoglycosidase	Scaffold for assembly of signalling complexes	β-Klotho tandem pseudoglycosidase domains act as a receptor for the hormone FGF21	[125]
Pseudolyase	Allosteric activation of conventional enzyme counterpart	Dead paralog binding to S-adenosylmethionine decarboxylase displaces an N-terminal inhibitory segment to activate catalytic activity	[126-128]
Pseudotransferase	Allosteric activation of cellular enzyme counterpart	Viral GAT recruits cellular phosphoribosyl-formylglycinamidine synthetase to deaminate RIG-I to block host antiviral defence	[129]
	Assembly of catalytically-active enzyme via composite active site	A composite active site is formed by dead and catalytic paralogs of *T. brucei* deoxyhypusine synthase heterotetramerize	[130, 131]

	Assembly of mitochondrial fission factor	MiD51 is an Inactive nucloetidyltransferase dynamin receptor that binds nucleotide diphosphates, and assembles Drp1, a regulator of mitochondrial fission	[132]
	Pantothenate kinase	Eukaryotic 'bifunctional' pantothenate kinase 4 (PANK4) phosphotransferases contain an N-terminal pantothenate pseudokinase domain fused to a phosphatase domain. Human PANK domain of PANK4 lacks catalytic Glu and Arg residues that are found in the PANK1-3 enzymes.	[133]
Pseudosulfotransferase	Unknown neuronal function?	Eukaryotic enzymatically-inactive soluble sulfotransferase (SULT4A1) fails to bind *in vitro* to the universal sulfate donor PAPS, or to sulfate a standard SULT substrate	[134]
Pseudo-histone acetyl transferase (pseudoHAT)	Possible scaffold for assembly of signalling complexes	Unlike bacterial *O. granulosus* counterpart, human C-terminal domain of putative bifunctional glycoside hydrolase and histone acetyltransferase (AT) *O*-GlcNAcase (OGA) lacks catalytic residues and detectable acetyl CoA binding	[135]
Pseudophospholipase	Scaffold for assembly of signalling complexes	DUF1669 pseudophospholipase D-like domains in FAM83 family proteins bind and regulate the canonical kinase CK1	[136, 137]
	Allosteric inactivation of conventional enzyme counterpart	Viper phospholipase A2 inhibitor attenuates the toxicity of the active phospholipase A2 paralog via heterodimerization	[138]
Pseudooxidoreductase	Allosteric inactivation of conventional enzyme counterpart	ALDH2*2 sequesters the active counterpart, ALDH2*1, to prevent its assembly into the tetrameric catalytic form	[139]
Pseudodismutase	Competition for enzyme's chaperone	Poxviral superoxide dismutase (SOD)-like proteins bind and sequester Copper chaperone for superoxide dismutase (CCS) to block activation of cellular SOD1	[140]
Pseudohydrolase	Integrin activation by protein with cryptic Nucleoside diiphosphate (Nudix) domain	KRIT1 antagonizes ICAP1 (negative regulator of integrin function) through a competitive mechanism. N-terminus of KRIT1 contains Nudix pseudoenzyme domain lacking canonical 'Nudix box'	[141]

	Pseudooligonucleotide synthetase	Anti-viral factor, loss induces resistance to West Nile Virus	Enzymatically-inactive Murine Oas1b (a 2'-5'-oligoadenylate synthease 1 paralog) possesses RNAse L-independent anti-viral properties towards flavivirus in mice.	[142, 143]

1092

1093

1094