Balancing Specificity and Promiscuity in Enzyme Evolution: Multidimensional Activity Transitions in the Alkaline Phosphatase Superfamily

Bert van Loo, Christopher D. Bayer, Gerhard Fischer, Stefanie Jonas, Eugene Valkov, Mark F. Mohamed, Anastassia Vorobieva, Celine Dutruel, Marko Hyvönen* and Florian Hollfelder*

Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

*For correspondence: fh111@cam.ac.uk; mh256@cam.ac.uk

Abbreviations: AS: arylsulfatase; PMH: phosphonate monoester hydrolase, NPP: nucleotide phoshodiesterase/pyrophosphatase; AP: alkaline phosphatase; PAS: *Pseudomonas aeruginosa* arylsulfatase.

ABSTRACT

Highly proficient, promiscuous enzymes can be springboards for functional evolution, able to avoid loss of function during adaptation by their capacity to promote multiple reactions. We employ systematic comparative study of structure, sequence and substrate specificity to track the evolution of specificity and reactivity between promiscuous members of clades of the alkaline phosphatase (AP) superfamily. Construction of a phylogenetic tree of protein sequences maps out the likely transition zone between arylsulfatases (ASs) and phosphonate monoester hydrolases (PMHs). Kinetic analysis shows that all enzymes characterized have four chemically distinct phospho- and sulfoesterase activities, with rate accelerations ranging from 10^{11}-10^{17}-fold for their primary and 10^{9}-10^{12}-fold for their promiscuous reactions, suggesting that catalytic promiscuity is widespread in the AP-superfamily. This functional characterization and crystallography reveal a novel class of ASs that is so similar in sequence to known PMHs that it had not been recognized as having diverged in function. Based on analysis of snapshots of catalytic promiscuity 'in transition' we develop possible models that would allow functional evolution and determine scenarios for trade-off between multiple activities. For the new ASs we observe largely invariant substrate specificity that would facilitate the transition from ASs to PMHs *via* trade-off-free molecular exaptation, i.e. evolution without initial loss of primary activity and specificity toward the original substrate. This ability to bypass low activity generalists provides a molecular solution to avoid adaptive conflict.

INTRODUCTION

The view that enzymes are specific for one substrate has been profoundly revised in recent years by the realization that many enzymes are catalytically promiscuous.[1-20] In some cases the observation of promiscuous activities can be attributed to the modest catalytic requirements of the promiscuous reactions, but more and more studies show promiscuous activities with large rate enhancements for chemically demanding reactions requiring substantial stabilization of the transition state.[14, 18, 21-25] This observation raises the question how enzymes can satisfy the stringent and, at the level of molecular recognition of ground and transition states, often divergent requirements for catalysis of multiple reactions.[5] Promiscuity has been postulated to play a role in the evolution of new enzymatic functions, by providing the organism with an initially moderate activity, which shortens the distance to the point at which this activity confers a selective advantage.[1, 9, 20, 26-27] A promiscuous enzyme may thus serve as a starting point for the evolution of a new enzyme with a different activity (possibly even prior to gene duplication), [9] so that catalysts e.g. for the breakdown of newly encountered compounds (e.g. xenobiotics)[28] can be created relatively rapidly.

O'Brien and Herschlag[1] (1999) were the first to link data on enzyme promiscuity to its mechanistic underpinnings. However, despite the increasing evidence for the generality of enzyme promiscuity,[3-4, 8] a systematic, quantitative analysis of this historically underreported phenomenon and its role in evolution is only starting to emerge. In this work, we studied a set of phylogenetically related phosphonate monoester hydrolases (PMHs) and newly identified class of arylsulfatases (ASs) that are part of the alkaline phosphatase (AP) superfamily in order to explore transitions between multiple enzymatic activities characteristic for this superfamily. This approach enables us to probe to what extent the phylogenetic annotation and the experimentally determined activities coincide. The diagnostic substrates **1**-**4** (top line, Figure 1) represent four distinct hydrolytic reaction types catalyzed by AP-superfamily members and differ in their reactive functional groups, substrate charges, sizes and hydrophobicities, and the nature of the transition states involved in their uncatalyzed hydrolyses. The reactions catalyzed by these enzymes are thermodynamically demanding, with half-lives ranging from 200 days to 10^5 years. The AP superfamily is characterized by crosswise catalytic promiscuity, i.e. the 'primary' reactions of one enzyme are found to be promiscuous reactions catalyzed by other family members[13-14, 18, 21, 23-26, 29-30] (Figure 1), as well as additional activities.[18, 31] We use substrates **1**-**4** to assign a function to sub-groups of the superfamily (e.g. potentially as phosphate monoesterases, phosphodiesterases,[30] phosphonate monoester hydrolases,[32] or arylsulfatases,[33-36]), leading to classification of newly identified enzymes according to their respective top activity, even though the natural substrates may not be known.

Studies with such representative substrates have been used to characterize trajectories of directed evolution experiments[37-39] and to infer specificity changes in evolution[1, 24, 40-42] and epistatic constraints.[43-44] This matrix of substrate classes and new superfamily members allows us to address the following questions for these hydrolases: (i) how is protein sequence phylogeny correlated to enzyme activity and the specificity determinants, (ii) how will multiple activities of these promiscuous enzymes trade-off against each other or will promiscuity be maintained in all members of this superfamily?

Structurally characterized members of the AP superfamily (Figure 1) highlight similarities in fold and active site residues.[29, 45-46] However, understanding the specificity determinants within this common fold is hampered by considerable sequence variation outside the active site. This ambiguity is highlighted by two pairwise comparisons: e.g. by *Bc*PMH and PAS (pairwise sequence identity: 27%). These two AP-superfamily members have *different* primary activities, but the majority (65%) of the amino acids align structurally and key active site residues (in particular the crucial formylglycine (fGly) nucleophile and the leaving group stabilizing residues) are in virtually identical positions (Figure 2, Table S2-S3). Conversely, for two superfamily members with the *same* primary activity (e.g. the sulfatases PAS *vs* human arylsulfatase C (*Hs*ASC)), similar pairwise structural alignment scores arise (67% alignment, 28% identity; Table S2-S3). These comparisons suggest that enzymes with different primary functions can be as closely related in sequence and structure as enzymes that share the same primary activity and raise the question which sequence, fold or active-site features are indicative for specificity and promiscuity. Beyond assignment of function, the identification of specificity determinants would also give insight into the transition of function within the superfamily. We focus on the sequence space where a transition between activities of the AP superfamily might occur. This space is populated by enzymes that are more closely related to PMHs (in terms of protein sequence phylogeny and structure) than to any other ASs of the AP superfamily, but their substrate specificity profiles show that they are ASs, correcting misannotation, and locating the point of functional divergence between the two activities.

MATERIALS AND METHODS

Materials. Phosphate monoester **1**, phosphonate monoester **3c** and sulfate monoester **4** were purchased from Sigma. Phosphate diesters **2a-2c** and phosphonate monoesters **3a** and **3b** were synthesized in a similar fashion as described previously[47-49] (for details see supporting information, SI). All restriction enzymes and T4 DNA ligase were from Fermentas. Vector pASKIBA5plus and Streptactin resin were purchased from Stratech scientific. *Pfu* turbo was from Agilent. Bacterial strains *Advenella kashmirensis* WT001, *Ralstonia metallidurans* CH34, *Stappia aggregata* IAM 12614 and *Silicibacter pomeroyi* DSS-3 were purchased from the DSMZ. Cell material of *Agrobacterium radiobacter* K84 and purified DNA from *Rhodopseudomonas palustris* CGA009 were obtained from ATCC.

Sequence retrieval, multiple sequence alignment and phylogenetic analysis. The sequences included in the phylogenetic tree of all arylsulfatases (ASs) and phosphonate monoester hydrolases (PMHs) of known activity (Figure 3A, Figure S1-S5), either from published data or found in this study (see Table S5 for details on the sequences included), were aligned using T-coffee (expresso mode), ClustalΩ, MAFFT and MUSCLE (all with default settings). These alignments were used as input to build a maximum likelihood phylogenetic tree with RAxML 8.2.10,[50] using various amino acid substitution matrices with a γ-model for rate heterogeneity, estimate proportion of invariable sites and empirical base frequencies (substitution matrix (LG, VT or WAG) +G+I+F, see legends to Figure S1-S4 for details) running on the CIPRES Science Gateway[51] (www.phylo.org/portal2/). The optimal amino acid substitution matrix and parameter configuration for tree-building was calculated using ProtTest 3.4.2.[52]. The T-coffee alignment (expresso mode) was also used to build a Bayesian maximum likelihood tree using MrBayes (Figure S5, see SI for details). The tree in Figure 3A is based on the T-coffee alignment (expresso mode), with Figure S1 serving as a legend to Figure 3A.

In order to obtain the genes encoding putative arylsulfatases (ASs) and phosphonate monoester hydrolases (PMHs), existing sequenced bacterial genomes were subjected to a BLAST-search, initially using the amino acid sequences of *B. caryophili*[18, 53] and *R. leguminosarum*[32]

PMHs as search sequences. During the course of this study sequences of the enzymes for which the primary putative activities were confirmed by experimental data were also used as search sequences. The protein sequences of the resulting hits were aligned using ClustalX and screened for the presence of putative active site residues. Enzymes in which any of the putative active site residues were missing were discarded. Based on the nature of the active site residues (Figure 2) the putative sequences were designated to be either arylsulfatases (ASs) or PMHs. Initial data indicated that the majority of the hits (>90%) originated from α- and β-proteobacteria. For simplicity, we restricted the BLAST-search to these two subgroups of bacteria for the data that were included in the final alignment. The guide tree from the ClustalX alignment was used to identify obvious outliers that were subsequently discarded. The final alignment was done with 85 (putative) PMHs, 95 (putative) ASs (for details see Table S6-S7) and 87 (putative) CSs using the 3D-coffee mode of T-coffee as described previously.[54] These alignments were used as input to build a maximum likelihood phylogenetic tree with RAxML BlackBox 8.0.24,[50] using the Le & Gascuel[55] amino acid substitution matrix (LG) with a γ-model for rate heterogeneity, estimate proportion of invariable sites and empirical base frequencies (LG+G+I+F) running on the CIPRES Science Gateway[51] (www.phylo.org/portal2/). Several additional methods (Clustal Ω, MAFFT or MUSCLE) were employed to create alternative phylogenetic trees, based on the same data and using the same settings (Figure S8). The 3D-coffee alignment was also used to build a Bayesian maximum likelihood tree (Figure S9B, SI). The tree shown in Figure S10 was built with RAxML 8.0.24[50] running on the CIPRES Science Gateway[51] (www.phylo.org/portal2/) based on 67 positions (as found in the complete alignment) representing all residues within a distance of 4.0 Å of the 11 putative active site residues in the X-ray structure (and included these 11 positions, see Figure S13 for details), using the Le & Gascuel[55] amino acid substitution matrix (LG) with a γ-model for rate heterogeneity and estimate proportion of invariable sites (LG+G+I). The optimal amino acid substitution matrices and parameter configuration for tree-building were calculated using ProtTest 3.4.2.[52]

Cloning of arylsulfatase (AS) and phosphonate monoester hydrolase (PMH) encoding genes. The genes encoding the various PMHs and ASs (Table 1) were amplified by PCR using the primer pairs described in Table S8. We used either commercially available genomic DNA (*Rhodopseudomonas palustris* CGA900) or whole cell material (all other organisms) of the respective organisms as a template for the PCR reactions. Primers were used at 0.4 nM in a reaction with 0.2 mM dNTPs and 0.05 U μL^{-1} *Pfu*-Turbo® DNA polymerase. The temperature program used for *Sa*AS was 15 min at 95 ºC without polymerase, followed by 30 cycles of 60 s at 95 ºC, 45 s at 58 ºC, 180 s at 72 ºC, and finished with 10 min at 72 ºC. For all other sequences, the temperature program consisted of 15 min at 95 ºC without polymerase, followed by 30 cycles of 60 s at 95 ºC, 45 s at 68 ºC - 0.5 ºC $cycle^{-1}$ (each cycle the temperature of this segment was lowered by 0.5 ºC), 180 s at 72 ºC, and finished with 10 min at 72 ºC. The PCR products were digested with various restriction enzymes (Table S8) and subsequently ligated into appropriately digested pASKIBA5plus plasmid DNA using T4 DNA ligase. The ligation mixture was transformed into *Escherichia coli* TOP10 using electroporation. Cells were plated on LB medium containing ampicillin and resulting colonies were checked for insert using a PCR reaction with *Taq* polymerase and colony material as the template. Plasmid DNA was extracted from positive clones and confirmed by sequencing.

Protein production and purification. Expression of the respective genes in the pASKIBA5plus vector results in a translational fusion with an N-terminal Strep-tag. Strep-tagged *Bc*MPH and *Rl*PMH were produced and purified as described previously.[18, 32] The other PMHs and ASs were produced in *E. coli* BL21 (DE3) by co-expressing the formyl glycine generating enzyme (FGE) from *Mycobacterium tuberculosis* H37v[56] (*Mtb*FGE) from pRSFDuet*Mtb*FGE plasmid.[18, 32] *E. coli* BL21(DE3) were typically grown in 750 mL of 2×YT medium with ampicillin (100 mg L^{-1}) and kanamycin (50 mg L^{-1}) at 37 °C to an A_{600nm} ~ 0.4, at which point the culture was cooled to 28°C. Once the culture had reached the desired temperature, expression of *Mtb*FGE, was induced by the addition of 1 mM IPTG, approximately 30 minutes prior to the induction of expression of the strep-tagged PMH/AS by adding up to 200 µg L^{-1} anhydrotetracycline followed by overnight growth at 28°C. Cells were harvested by centrifugation and resuspended in 50 mM Tris-HCl pH 8.0. The resuspended cells were lysed either by an Emulsiflex-C5 high pressure homogenizer or by sonication and cell-free extract was obtained by centrifugation at 40,000 × g for 90 minutes. The cell-free extract was subsequently loaded onto a Q-Sepharose anion exchange column and protein was eluted with a gradient of 0-1 M NaCl. Activity was tested towards either phosphonate monoester **3c** (PMHs) or sulfate monoester **4** (ASs). The overexpressed proteins typically eluted at 0.3-0.5 M NaCl. The active fractions were pooled, 1/10 of the pooled volume of 1 M Tris-HCl pH 8.0 + 1.5 M NaCl was added to the sample, which was subsequently loaded onto 1 mL Strep-Tactin column equilibrated with 100 mM Tris-HCl pH 8.0 + 150 mM NaCl. The column was washed with 100 mM Tris-HCl pH 8.0 + 150 mM NaCl to remove unbound protein and the tagged proteins were eluted with 2.5 mM *d*-desthiobiotin in 100 mM Tris-HCl pH 8.0 + 150 mM NaCl. The active protein containing fractions were pooled and concentrated to 10-15 mg mL^{-1} protein and loaded onto a Superdex 200 prep grade size exclusion column (GE Life Sciences) running in 100 mM Tris-HCl pH 8.0 + 150 mM NaCl. Active protein eluted at a molecular weight corresponding to either a dimeric (ASs) or tetrameric (PMHs) protein (see Figure S14, SI). Protein containing fractions were concentrated down to 100-200 μM, aliquoted, flash frozen in liquid nitrogen and stored at -20 °C.

The possibility that contaminants contribute to the observed activities was ruled out by the following considerations: the absolute levels of some of the promiscuous reactions are similar to previously described 'native' functions,[19, 30, 42, 57-58] making it unlikely that they are caused by a contaminating protein that co-purifies from the expression host (*E. coli*). Precautions taken to prevent and rule out cross-contaminations include protein purification with an affinity-tag (Strep-tag binding to Strep-Tactin), using fresh affinity resin for each variant and extensive cleaning of the other columns with 0.5 M sodium hydroxide in between purifications to remove and inactivate any AS or PMH from previous purifications. A wide range of K_M-values (5-100 fold differences for the same substrate) was measured, indicating that the promiscuous activities are not caused by a contaminant (in which case the promiscuous conversions would show similar K_M-values characteristic of that contaminant).

Catalytic effects of metals and metal content determination using microPIXE. All enzymes were activated most strongly by Mn^{2+} (compared to other divalent metal ions, Ca^{2+} or Mg^{2+}), and kinetic experiments were carried out under conditions of saturating Mn^{2+} (see SI).

The metal content of the various ASs and PMHs (Table S9) was determined using microPIXE (carried out at the University of Surrey Ion Beam Centre) as described previously,[18, 32] using sulfur atoms as the internal quantitative standard to calculate the metal ion occupancy.

Enzyme assays. Enzyme assays were performed in either 20 mM sodium acetate, 100 mM NaCl (phosphonate monoester **3a**-**c**, pH 4.8-5.6), 20 mM bis-tris propane, 100 mM NaCl (phosphonate monoesters **3b** and **3c**, pH 5.5-9.0) or SID buffer (44 mM succinic acid, 33 mM imidazole, 33 mM diethanolamine, all other substrates, pH 5.5-9.6). Ionic strength was maintained constant within ±10% in the pH range 4.8-9.6. The catalytic parameters towards the promiscuous substrates for the new PMHs and ASs were determined at their respective pH-optima of their primary activity[18, 32] (Table S10-S17 and Figure S8 and S9).

All enzymes were activated most strongly by Mn^{2+} (compared to other divalent metal ions, Ca^{2+} or Mg^{2+}), and kinetic experiments were carried out under conditions of saturating Mn^{2+}(see supporting information and Table S10-S28 for details). Enzymatic hydrolysis was followed by monitoring the release of 4-nitrophenolate at 400 nm using a SpectraMax Plus multiwell reader. The extinction coefficients of 4-nitrophenol at varying pH were determined for pH 5.5-9.6, increasing from 1,000 to ~20,000 M^{-1} cm^{-1} with increasing pH.

Initial rates (V_{obs} in M s^{-1}) were plotted against the substrate concentration ([S]) and kinetic parameters k_{cat}, K_{M} and/or $k_{\mathrm{cat}}/K_{\mathrm{M}}$ were determined using equations 1-3. For fitting with equation 2, $k_{\mathrm{cat}}/K_{\mathrm{M}}$ is treated as a single parameter. In order to obtain a reliable fit to equations 1-3, a minimum of 16 different initial rate measurements were performed in such a way that it included substrate concentrations ranging from at least 5-fold below the estimated K_{M}-value to at least 5-fold above the estimated K_{M}-value (with increments that increase with increasing substrate concentration), wherever substrate solubility allowed for it. If the latter was not the case, at least substrate 16 concentrations were chosen that range from 5-100% of the maximum solubility in 5-10% increments. The K_{M}-value was estimated from several initial rate measurements covering a wide range of substrate concentrations (varying by several orders of magnitude). Typical enzyme concentrations used were 0.0016-0.74 μM (for primary activities) or 0.16-10 μM (for promiscuous activities).

$$V_{obs} = \frac{k_{cat} \times [Enz] \times [S]}{K_M + [S]} \qquad \text{eq. 1}$$

$$V_{obs} = \frac{k_{cat}}{K_M} \times [Enz] \times [S] \qquad \text{eq. 2}$$

$$V_{obs} = \frac{k_{cat} \times [Enz] \times [S]}{K_M + [S] + S^2 / K_{SI}} \qquad \text{eq: 3}$$

k_{cat}/K_M refers to the first irreversible step of the enzymatic reaction, i.e. in this case to the formation of the covalent intermediate between the formylglycine nucleophile and the respective P or S center of the substrates. The phenolate product does not bind significantly to the enzyme, rendering the reverse reaction to starting material unlikely. All specificity comparisons are based on k_{cat}/K_M and therefore refer to the same molecular event, the attack of the nucleophile, regardless of the overall rate-limiting step.

The errors (δ) indicated in Table S10-S29 were either obtained directly from the curve fitting to equations 1-3 or calculated according to equation 4 (for k_{cat}/K_M if k_{cat} and K_M were obtained by fitting to equation 1 or 3).

$$\delta \frac{k_{cat}}{K_M} = \left| \frac{k_{cat}}{K_M} \right| \times \sqrt{\left(\frac{\delta k_{cat}}{k_{cat}} \right)^2 + \left(\frac{\delta K_M}{K_M} \right)^2} \qquad \text{eq. 4}$$

Crystallization, data collection, structure determination and analysis. All crystals were grown at 18 °C by hanging or sitting drop vapor diffusion with protein (~9-12 mg mL^{-1}) reservoir ratios (v/v) of 1:1 (*Sp*AS1 and *Sp*AS2), 2:1 (*Ar*PMH) or 1:3 (*Sp*PMH). *(i) Strep-SpAS1* crystals grew in 0.1 M Tris-HCl pH 8.0, 20% (w/v) PEG 6000, 0.75 M LiCl. Crystals were soaked in the mother liquor containing 10% (v/v) glycerol prior to flash cooling in liquid nitrogen. *(ii) Strep-*

SpAS2 crystals grew in 20% (w/v) PEG 3350, 0.2 M ammonium formate. Crystals were soaked in the mother liquor containing 20% (v/v) glycerol prior to flash cooling in liquid nitrogen. *(iii) Strep-ArPMH* crystals grew in 0.1 M Tris-HCl pH 7.0, 10% (w/v) PEG 8000 and 0.2 M $MgCl_2$. Crystals were cryo-protected by adding 3 M trimethylamine-*N*-oxide (TMAO) to the mother liquor and soaking the crystal for 1 minute before flash cooling in liquid nitrogen. *(iv) Strep-SpPMH* crystals grew 0.1 M bis-tris-propane pH 7.0, 9% (w/v) PEG 8000, 0.2 M $MgCl_2$. They were soaked in the mother liquor containing 20% (v/v) glycerol prior to flash cooling in liquid nitrogen.

Briefly, all crystallographic data were collected at ESRF or Diamond synchrotron sources and processed using standard crystallographic software. For details on processing and refinement see Table S32. Coordinates and experimental structure factors have been submitted to the Protein Data Bank with accession codes 4UPH (*Ar*PMH), 4UPI (*Sp*AS1), 4UPK (*Sp*PMH) and 4UPL (*Sp*AS2).

RESULTS AND DISCUSSION

Classifying adjacent AP-Superfamily clades with phosphonate monoester hydrolase and arylsulfatase primary activity. In order to assemble a test set of related members of the AP superfamily populating the transition zone between sulfatases and phosphonate monoester hydrolases, we ran a BLAST search using the amino acid sequences of the two previously characterized PMHs[18, 32] as search sequences. The genomes of α- and β-proteobacteria yielded well over 100 putative sequences that were originally annotated either as PMHs or arylsulfatases (ASs) (see Table S6-S7 for details). All of them contained the Cys-X-Pro/Ala-X-Arg recognition motif essential for the post-translational formation of cysteine into the formylglycine nucleophile found in ASs[59] and PMHs[18-19] (Figure 2). Based on the level of similarity to the active site residues of PMHs we divided the sequences into three groups. One group – termed T-PMHs (bearing a threonine next to the catalytic nucleophile; Thr107 in *Rl*PMH and *B*cPMH, homologous to position His^A in Figure 2, Table S4) - had identical active site residues to the original two PMHs. The second group, D-PMHs, characterized by an aspartate in place of a threonine (His^A in Figure 2). The classification of these two groups as PMHs (Table S4) was supported by high overall pairwise sequence similarity to the known PMHs (~50%). The third group of enzymes assumes an intermediate position between PMHs and known ASs. These enzymes have histidine and lysine at positions His^A and Lys^A ,(as in ASs), but two metal-coordinating residues are identical to the ones in PMHs (Gln at position Asp^B; His instead of Asn^A). The active site of this group of enzymes is therefore a hybrid between classical ASs and PMHs.

One T-PMH, two D-PMHs and six of the hybrid sulfatase-like proteins (Table 1) were expressed in *E. coli* and purified to homogeneity. In the final step of protein purification by size exclusion chromatography the PMHs eluted as tetramers, whereas the six sulfatase-like proteins eluted as dimers (Figure S14). Activity assays showed turnover for all four substrate classes **1**-**4** that represent primary activities in AP-superfamily members (Figure 1). T-PMHs and D-PMHs preferentially hydrolyzed phosphate diesters (**2a-2c**) and phosphonate monoesters (**3a**-**3c**). The sulfatase-like proteins preferentially hydrolyzed arylsulfate monoesters (**4**). This experimental result establishes them as a new class of ASs, even though they were identified from the databases using PMHs as search sequences. However, they do prefer Mn^{2+} as their active site metal ion (as in PMHs), instead of the Ca^{2+}/Mg^{2+}-ions typical for previously identified ASs (see SI and Table S4 for details). In these new ASs we observe high activities towards phosphoesters **1**-**3** (k_{cat}/K_M-values as high as 10^4 s^{-1} M^{-1}), similar to the bona fide *sulfatase* PAS,[21] known to hydrolyze sulfate esters *in vivo*.[60] We conclude that the identities of the active site residues at positions Lys^A and His^A are better predictors for functional classification of ASs and PMHs than overall sequence similarity.

Structure-guided alignment of the newly identified sulfatases with previously characterized PMHs[18, 32, 53] and sulfatases in the AP superfamily (see Table 1 and Table S5 for details on the sequences included) shows that the new ASs are more closely related to PMHs than they are to any other known ASs, such as PAS[60] or several ASs of bacterial[61-64] and human[65] origin (Figure 3A, Figure S1 and Table S5).

To identify additional sub-family members (see Table S6 and S7 for a list of these sequences) the five PMHs (*Bc*PMH, *Rl*PMH and the three newly identified ones) and the six newly identified ASs were used as search sequences in BLAST, leading to identification of 169 related enzymes. We discarded the annotations listed in the database for these enzymes (see Table S6-S7 for details) and re-classified them based on their active site residues (Figure 2), yielding 80 putative PMHs and 89 putative ASs. The phylogenetic relationship shows that the functional re-classification coincides with the genetic division between the new ASs and the PMHs (Figure 3B), although the original division was based solely on the conservation of the 11 active site residues.

Structural comparisons of PMHs and ASs. In order to determine whether the aforementioned phylogenetic division corresponds to structural features of the newly identified AP superfamily members, two D-PMHs, *Ar*PMH and *Sp*PMH, and sulfatases, *Sp*AS1 and *Sp*AS2, were crystallized and their structures were determined (Figure 4, Figure S17 and S18, Table S32).

All of the enzymes show the conserved α/β-fold[29, 45-46] (Figure 1), familiar from all AP-superfamily members, such as AP,[66] NPP,[30] PAS[67] and both previously described PMHs.[18, 32] *Pairwise* comparisons[68] between all the known crystal structures of ASs and PMHs show that

*Sp*AS1 and *Sp*AS2 are more closely related to the PMHs than they are to any of the classical ASs (Tables S1-S3), mirroring the phylogenetic relationship seen in Figure 3A. This is contradicted by the substrate specificity profile, demonstrating the perils of assigning function based on sequence and structural homology alone.

All conserved active-site residues (Table S4) align structurally, confirming the mechanistic relatedness between the various PMHs and ASs (Figure 5). With the structures of new enzymes available, we rebuilt the phylogenetic tree (Figure 3B) using only 67 positions from the original alignment that represent the conserved active site and its direct surroundings (see Figure S13). The resulting tree shows the same global division as the original tree in Figure 3B, i.e. a clear distinction between the PMHs and ASs (Figure S10), which means that the active site and its immediate surroundings contain sufficient information to classify these enzymes. The fact that the global division between both clades is the same in both trees therefore strongly suggests that the phylogenetic division observed in Figure 3B is *directly* connected to the difference in substrate specificity.

Comparison of the four new structures with *Rl*PMH,[32] *Bc*PMH,[18] and the monomeric PAS[67] (Figure 4A) reveals a sequential increase in the oligomeric state within the superfamily. *Sp*AS1 and *Sp*AS2 are homodimeric enzymes with their helical C-terminal extensions facilitating the dimerization (Figure 4B-C, Figure S17A-B). In contrast, *Ar*PMH and *Sp*PMH are tetrameric enzymes with dihedral symmetry and very similar to their functional homologues *Rl*PMH[32] and *Bc*PMH[18] (Figure 4D-E, Figure S17C-D). The dimeric structure of the novel sulfatases corresponds to one half of the PMH tetramer and this dimerization is mediated by the interface that is equivalent to the larger of the two protomer-protomer interfaces in PMHs (Figure 4). The C-terminal regions mediating the dimerization of *Sp*AS1 and *Sp*AS2 are predominantly α-helical with a highly conserved hydrophobic motif, in contrast to the short β-hairpins that mediate the homologous oligomerization interface of the PMHs (Figure 6A). Insertions and deletions (Figure 6B) reshape the interface between protomers and C-terminal tails, and may – in agreement with the analysis by Hashimoto *et al.,*[69] - have a role in governing oligomerization (see SI and Figures S19-S21).

PMHs and ASs are highly promiscuous enzymes. The previously observed catalytic promiscuity of members of the AP superfamily[13-14, 18, 21, 23-26, 29-30] raises the question whether this feature is reflected in the newly identified superfamily members and its quantification provides the opportunity to assess the transition between chemical preferences. We thus tested the ability to hydrolyze the four primary substrate classes for the AP superfamily, represented by phosphate mono- (**1**) and diesters (**2**), phosphonate monoesters (**3**) and sulfate monoesters (**4**) and report Michaelis-Menten parameters (Figure 7A-B, Table S18-S27) at the pH of maximal activity for the primary substrate (Table S10-S17, Figure S15-S16,[18-19]) with the most activating divalent metal (Mn^{2+}, present at saturating concentrations, see SI). In addition to changes in the reactive group, the size of the non-reactive substituents (previously shown to influence catalytic rates in PAS[21] (Table S29) and NPP[40]) was also addressed in measurements with phosphodiesters **2a**-**2c** and phosphonate monoesters **3a**-**3c**. All PMHs and new ASs are active towards all four substrate classes and are thus truly promiscuous, further establishing 'crosswise' catalytic promiscuity as a general trait of AP-superfamily enzymes. Catalysis for all four chemically distinct reactions is efficient, with large second order rate accelerations ($(k_{cat}/K_M)/k_2$) for both the primary (10^{11}-10^{17}-fold) and the promiscuous (10^9-10^{12}-fold) activities (Table S18-S27).

The relative differences between the phosphodiesterase and phosphonate monoesterase activities of each individual PMH fall within one order of magnitude (Table S18-S22). However, we observe an increased propensity for the T-PMHs to accommodate phosphonate monoesters with larger non-reactive substituents, i.e. the smallest phosphonate monoester (**3a**) is converted with the lowest efficiency. The D-PMHs show the opposite effect, preferring smaller substrates (e.g. favoring **3a** *vs.* **3b** and **3c)**. This observation is consistent with the defining feature that separates the D- and T-clans: the threonine in position His^{A} (Figure 2) accommodates the non-reactive hydrophobic substituent of the substrate more readily than a charged aspartate would. In addition, for both D-PMHs the difference in catalytic efficiencies toward phosphate monoester **1**, phosphodiesters **2a-c** and phosphonate monoesters **3a-c** is much smaller than for the three T-PMHs (Figure 7A1-A4). This lack of specificity could be explained if the protonated monoanionic form of phosphate monoester **1** reacts like a phosphate diester **2** (as postulated previously for phosphodiesterases NPP[30] and GpdQ[70]). Given the acidic pH at which the catalytic parameters for both D-PMHs were determined, the monoanionic form of phosphate monoester **1** is likely to be present at high levels. Therefore, the relative lack of specificity of the two D-PMHs toward these types of substrates could be explained by a pH-rate optimum that causes the D-PMHs to employ identical catalytic effects for the conversion of both phosphate mono- and diesters.

Trade-off between promiscuous activities. For ASs, the specificity between the primary and the multiple promiscuous activities (i.e. the ratio between the primary activity and promiscuous activity for a given promiscuous substrate) appeared to vary little between the five dimeric ASs (Figure S22B1-B6). Direct correlation plots of the promiscuous activities *vs.* the primary sulfatase activity (both expressed as log($k_{\mathrm{cat}}/K_{\mathrm{M}}$)), showed a linear correlation with a slope near unity for all substrates tested, except for phosphonate monoester **3c** (Figure 7B1-B7). Although ASs clearly prefer their "native" substrate, the ratio between the primary and a given promiscuous activity is conserved, irrespective of the absolute level of the primary activity. As the primary activity increases, so does the promiscuous activity, demonstrating near-constant specificity and lack of trade-off.

For the phosphonate monoesters **3a-c** the variation in specificities between the dimeric ASs increases with increasing size of the non-reactive substituent (R-groups in Figure 1; Figure 7B5-B7). PAS, which had already been shown to be highly proficient toward phosphodiesters,[21] showed the same behavior (Table S29), despite being phylogenetically much further removed from the PMHs than the dimeric ASs (Figure 3A). When PAS was included in the analysis, the conclusions that can be drawn from the promiscuity correlations remain the same, although the trend toward increased variation in substrate specificities with size of the R-group observed across phosphonate monoesters **3a-c** is much stronger (Figure 7B5-B7). These observations are consistent with the following consideration: the putative binding site for these promiscuous substrates is under selective pressure to accommodate sulfate monoesters. Phosphonate monoesters **3a-c** have the same overall charge as sulfates, but bear an additional group (R in Figure 1) and would be likely to experience additional steric constraints in an active site evolved toward accommodating the smaller sulfate monoester substrate. However, the relative promiscuity patterns of ASs towards phosphate diesters **2a-c** (which also have the same charge as sulfate monoesters) showed no particular trend with respect to the size of the non-reactive substituent. There is no simple explanation for these observations, except that the mixed effect of steric fit (including rotational degrees of freedom in the substrate) and interactions with the bridging oxygen of the non-reactive substituent force substrates into unique orientations, in which beneficial and detrimental binding contributions may counteract each other. Consistent with the trend to favor promiscuous substrates similar in size and charge to the primary substrate, for four (out of the five) new ASs, the smallest promiscuous substrate with the same total charge, methylphosphonate **3a**, was also the compound toward which the enzyme had the highest activity besides the primary substrate.

PMH and AS differ in trade-off between efficiency and specificity. Promiscuous catalysts, in particular generalists, i.e. enzymes that catalyze more than one reaction at levels high enough to support a functional role, have been postulated to be particularly versatile in their adaptive potential, raising the question to what extent the catalysts in the activity transition zone discussed here somehow remain 'generalists' or if they become specialists as they are subjected to selection pressure for *one* activity. Higher activity of PMHs toward their primary substrates (phosphate diesters **2a-c**/phosphonate monoesters **3a-c**) coincides with an improved degree of specialization for the primary activities over the promiscuous conversion of phosphate or sulfate monoesters (Figure 7A1-A4 and Figure S22A1-A4). The evidence for this hypothesis in PMH is compelling (correlation slope < 1, $p < 0.1$, see Table S34-S35). and supports the notion that high catalytic efficiency requires a more specialized enzyme. Likewise this scenario has also been observed during laboratory evolution campaigns,[38] in which case improvement in specificity was driven by the requirement for improved turnover.

However, the observations for ASs stand in contrast to PMH: specificity (measured by the ratio primary activity *vs* any given promiscuous activity) does *not* increase with increasing catalytic efficiency ($k_{\mathrm{cat}}/K_{\mathrm{M}}$) toward sulfate monoester **4**. Apparently in ASs activity increases for the primary substrate (sulfate monoester **4**) are invariably accompanied by an increase in any of the promiscuous activities by the same order of magnitude, suggesting that this effect is at the very least not selected against. Indeed, *all* ASs prefer sulfate monoester **4** by ~10^2-fold over phosphonate monoester **3a**, irrespective of the absolute level of the primary activity. This apparent lack of further specialization, previously suggested as a necessity for obtaining extreme catalytic efficiency for the native reaction,[71] seems not to hamper near-optimal catalytic performance toward its primary substrate, as the catalytic efficiency for PAS-catalyzed sulfate monoester hydrolysis is close to diffusion control ($k_{cat}/K_M = 5 \times 10^7$ s^{-1} M^{-1}).[25] These correlations also support the idea that these promiscuous activities are not just evolutionary relics of past activities from a generalist ancestor.[9] The cross-correlation of sulfatase *vs.* phosphodiesterase activities for ASs and PMHs together with AP and NPP confirms that this linear correlation is specific for enzymes that are under selective pressure for sulfatase activity (Figure 8A), notwithstanding the longer genetic distance between PAS and *Sp*AS1/*Sp*AS2 (sequence identity: ~22%, Table S3) compared to *Sp*AS1/*Sp*AS2 *vs.* PMHs (~33%, Table S3). The remarkable ability to improve catalysis toward multiple chemically distinct substrates is illustrated by the following examples: for ASs improvements for substrates with identical charges (yet different transition states)[72] correlate well, e.g. the monoanions of sulfate **4,** phosphate diesters **2a-c** and phosphonate monoester **3a** (Figure 7B2-B5). Conversely, dianionic phosphate monoester **1** and the monoanionic sulfate **4** correlate similarly (Figure 7B1) despite different ground state charges, but here the uncatalyzed reactions proceed through analogous dissociative (‘exploded’) transition states.[72]

The conservation of these specificities across a 10^4-fold difference in sulfatase k_{cat}/K_M-levels suggests that this correlation is general. It seems that escape from this constant specificity is difficult (or at least not selected for), when these sulfatases are under selective pressure for improved sulfatase activity. If improvements in phosphoesterase activity are selected for, escape from this correlation appears to be possible, but not in all directions, as shown recently for

*Sp*AS1.[26] As a result, sulfatases with low phosphodiesterase activity (i.e. enzymes located in the top left corner in Figure 8A) are notably absent. It is likely that the conserved protein fold or the intrinsic catalytic features required for enzyme-catalyzed sulfate monoester hydrolysis inherently also favor catalysis of phosphoester hydrolysis, which constrains the observed specificities. This observation suggests that there is a maximum degree of specialization between the primary sulfatase activity and the activity toward phosphate monoesters, diesters and phosphonate monoesters, when an enzyme is under selective pressure for sulfate monoester hydrolysis. It remains to be seen whether the observed inability to specialize a sulfatase to minimize phosphate transfer activity (and uncouple it from sulfatase improvements) is a general feature of these (or any) enzymes. If so, it might explain why biological phosphorylation and sulfation are often occurring in different, physically separated subcellular locations in mammals,[73] thus avoiding complications of intrinsic cross-reactivity.

Reactivity Correlations Across Evolutionary Snapshots. The characterization of active sites with k_{cat}/K_M values differing over 4 orders of magnitude in a structurally conserved context, provides a unique opportunity to systematically analyze enzyme reactivity across an evolutionary transition zone. Figure 7 shows one-against-one correlations that plot the k_{cat}/K_M values of sub-group members for each promiscuous activity against its name-giving activity. Two scenarios can be envisaged: (i) If reactions are promoted by catalytic effects that affect the chemically distinct promiscuous reactions differentially, scatter should emerge, due to idiosyncratic patterns of enzyme-substrate interactions and to different extents of transition state stabilization. In an alternative scenario (ii), the interactions are quantitatively the same for all substrates, leading to a linear correlation with a slope of 1.

We explored these scenarios based on our new data for correlations of the main activity of PMHs (Figure 7A1-A4) and for the new ASs (Figure 7B1-B7). For PMHs there is no discernible trend: each substrate and each catalyst seem to be characterized by unique contributions to binding and catalysis, indicative of particular arrangements for each pair. Scenario (ii) is observed for the new ASs. The linearity of all but one of the correlations in Figure 7B1-B7 with a near identical slope around unity suggest that within this group of enzymes, specificity for the primary substrate over a given promiscuous one is constant across a 10^4-fold variation in catalytic efficiency (k_{cat}/K_M) towards its primary substrate. Indeed, this group of catalysts exhibits a correlation expected of a 'chemical' catalyst that is dominated by reactivity rather than substrate recognition. These correlations hold despite diverging sequences (even at the level of the conserved active site positions, e.g. when comparing PAS *vs.* the new AS subfamily, Table S4). Only an increase in substrate size - from methyl (**3a**) to ethyl (**3b**) and phenyl phosphonates (**3c**) (Figure 7B5-B7; $p > 0.1$, see Table S35 for details) - compromises the otherwise ideal correlation observed between sulfate **4** and phosphonate **3a**, suggesting that the availability of sufficient room for substrate binding is a prerequisite for the correlated multi-reactivity improvements.

CONCLUSIONS AND IMPLICATIONS

Classification of new enzymes. Our data clearly show that a combination of sequence analysis and thorough experimental verification of the catalytic function of putative enzymes is necessary to avoid propagation of misannotations based on unverified functions.[74] For example, *Ak*PMH was previously annotated as a sulfatase (Table 1), while our data classify the enzyme as a phosphonate monoester hydrolase[1], based both on its phylogenetic relation to the known PMHs and ASs, the nature of its active site residues (Table S4), and its substrate specificity profile. The true substrates of the newly identified sulfo- and phosphoester hydrolases remain currently unknown. However, the use of the type of model substrates (with hydrophobic leaving groups) used in this study to assign the primary reaction types catalyzed by these enzymes has been used historically to identify the substrate class they most likely act on *in vivo*.[75] In no case has this initial assignment been rescinded. Similar fluorogenic substrates continue to be used[76-77] (e.g. as a diagnostic, generally representative substrate indicating bacterial sulfatase activity[61] or in screening of metagenomic libraries for sulfatases[78]), even though sulfates sugars are more likely candidates.[79] One reason for the successful prediction of sulfatase activity with model substrates may be that sulfate transfer is one of the thermodynamically most demanding natural reactions,[80] so that a merely accidental activity can be ruled out. Taken together these arguments suggest that the assignment to chemical classes of substrates is a widely used approximation and a basis for our analysis of chemical reactivity at the level of reaction type, even if the precise identity of the biological substrates remains to be elucidated.

Experimental verification of primary function is nevertheless complicated by the existence of promiscuous side reactions and simple testing of the assumed primary activity could result in incorrect functional annotation. The new class of sulfatases described in this study is a further

[1] Note that we refer to enzymes that show its strongest activity towards substrates **3** as ‘phosphonate monoester hydrolases (PMHs)’, because the first representative of this group was given this name.[53] We have previously suggested to name such enzymes ‘phosphonate monoester hydrolases/phosphodiesterases’,[32] but stick here with the simpler term PMH to group it with its historical antecedents. However, it is understood that name giving for these hydrolases does not imply assignment of a biological role, given that no natural substrate is known and the absence of phosphonates in the soil environment during the evolution of this enzyme makes an alternative assignment as a phosphate diesterase (referring to the second best activity) possible, especially because more substrate candidates of this type exist.

example of this pitfall. The new sulfatases clearly have sulfate monoester hydrolysis as their primary activity (Figure 7B), yet several of them catalyze phosphodiester/phosphonate monoester hydrolysis at levels that are similar to those of the 'native' activity levels for PMHs (Table S18-S22) and NPP.[30] Since all of the previously uncharacterized enzymes (6 dimeric ASs, 3 new PMHs) described in this study have either *Rl*PMH or *Bc*PMH as their closest homolog with experimentally described function, their annotation as PMHs would have been plausible based on their absolute activity levels for phosphonate monoester **3a**. These considerations highlight that (*i*) the large variations in absolute activity levels (>10^3-fold, Figure 7) combined with the widespread catalytic promiscuity in the AP-superfamily makes simply confirming activity by testing a single substrate class unreliable for functional annotation, (*ii*) the specificities (i.e. the activity ratios of primary *vs.* promiscuous substrates) are conserved and define the functional class the enzymes belong to. Therefore, highly promiscuous enzymes should be tested on a *panel* of representative substrates to establish the correct correlation between sequence (e.g. identity of active site residues, phylogeny), structure and function.

Are there limitations to the evolution of enzyme specificity? For PMHs, higher rates toward the primary phosphonate monoester substrate results in improved specificity toward its primary substrate(s) over its promiscuous ones (Figure 7A1-A4 and Figure S15A1-A4). ASs appear not to exhibit such an efficiency-specificity trade-off (Figure 7B1-B6 and Figure S15B1-B6). This stands in contrast to the suggestion that evolution toward higher catalytic efficiency, i.e. approaching levels close to diffusion control, will eventually result in an increase in specificity,[71] since the near-diffusion controlled PAS ($k_{\mathrm{cat}}/K_{\mathrm{M}}$ ~10^7 s^{-1} M^{-1}) is essentially as specific as the 10^4-fold less active *Sp*AS1 (Figure 7B).

What may explain the apparent limit to specificity? If specificity is conserved, this means that the many mutations separating the various sulfatases would need to affect all activities to the same degree, and that mutations that maximize sulfatase activity while suppressing phosphodiesterase activity have not been fixed during evolution. There are two features that sulfatases could exploit to differentiate (with everything else being almost equal, i.e. leaving group, smallest possible substituents for phosphodiesters) between sulfatase, phosphatase and phosphodiesterase substrates: overall substrate charge and degree of necessity for leaving group stabilization. Substrate specificity based on difference in charge of the substrate will inherently favor phosphate monoesterase activity over all the others, whereas the high degree of leaving group stabilization that will probably be required for sulfate monoester hydrolysis[80] would benefit all reactions, albeit to a different extent. Both properties are therefore extremely limited in providing features that would exclusively benefit sulfatase activity. Furthermore, the low reactivity of likely native sulfate ester substrates[80] (e.g. sulfated sugars, k_{uncat} ~10^{-26} s^{-1}) must require sulfatases to provide highly reactive functionalities in their active site.

The high levels of promiscuous phosphodiester hydrolysis that accompany high sulfatase activity for our set of representative substrates, are unlikely to be relevant *in vivo*: natural *aryl* phosphodiesters are rare and most naturally occurring phosphodiesters differ sufficiently in their leaving group from naturally occurring sulfoesters, which may provide distinguishing features that prevent cross-reactivity. Furthermore, phosphodiesters can have additional recognition anchors (e.g. a functional group that can provide a specific interaction) in the non-reacting substituent (R in Figure 1). The latter was observed for NPP, in which a large part of the specificity difference (phosphodiesters **2** *vs.* phosphate monoester **1**) was abolished when the non-reacting substituent in the phosphodiester substrate was shortened to a methyl group.[30, 81]

Direction of evolution. We interpolate data from extant enzymes related by protein sequence phylogeny,[21, 25, 78] which depict accessible functional space characterized by evolutionary snapshots. The close neighborhood in sequence space in nearby clades of ASs and PMH suggest a pathway for natural evolution, because a functional change can be brought with relatively few mutations. However, this does not imply a real evolutionary history (as the enzymes' substrates are often unknown), but options based on the classification of primary and promiscuous reaction types. With these caveats in mind we speculate on the possible evolutionary scenarios the functional characteristics we observe. Given that sulfatases appear to be restricted in the maximal specificity between their native sulfate monoester substrate and phosphoesters (Figure 7B1-B6), the question of the consequences for evolution arises (Figure 8). Adaptation toward higher sulfatase preference would locate a highly specialized enzyme closer to the top left corner, yet this activity space is apparently 'forbidden'. Yet none of the enzymes studied by us or others[23, 40] are located in the activity space above the fitted curve shaded in light red in Figure 8A. The specificity of enzymes far away from the specificity limit, such as *Ak*PMH (marked A, Figure 8A) has no apparent intrinsic restrictions to move in all directions of functional space. In response to the selective pressure applied when e.g. equal or improved sulfatase activity is required, only half of the trajectories will be accessible. The maximally specialized sulfatases such as *Ak*AS (marked B, Figure 8A) that have reached the border of the "forbidden" specificity region appear to be intrinsically constrained to half of the accessible directions in functional space compared to unconstrained movement. Since this restriction is more pronounced toward one particular activity, it is more likely for a specialized phosphodiesterase to emerge from a sulfatase than the other way around: 50% of all *accessible* trajectories starting from *Ak*PMH (A in Figure 8A) result in improved sulfatase activity, whereas for *Ak*AS (B in Figure 8A), 75% of the *accessible* trajectories result in improved phosphodiesterase activity, i.e. the local structure of the sequence-function space restricts the *accessible* genotypic space asymmetrically, constraining evolution across a fitness landscape in favor of one activity over the other. This asymmetry would agree with the difference in uncatalyzed rate constants (k_{uncat}) for both reaction types (k_{uncat} natural sulfate monoesters ~10^{-26} s^{-1},[80] k_{uncat} natural phosphodiesters ~10^{-16} s^{-1}): it might be expected that adaptation toward a specificity profile that requires less catalytic power (compared to its original primary activity) is more likely than the other way around.

The invariable increase in phosphodiesterase activity accompanying increased sulfatase activity can be classified as exaptation: adaption toward one trait (sulfatase activity) results in an increased level of another trait (phosphodiesterase activity) for which there is no apparent selection. In this scenario (Figure 8B), adaptation towards an improved promiscuous function would follow neither of the frequently discussed evolutionary trajectories,[4, 82] involving strong or weak initial negative trade-off, with the latter proceeding through a generalist intermediate. These strong or weak negative trade-off scenarios were based on observed activities of mutants obtained during directed evolution experiments.[37, 83]. Here, by contrast, both the primary and promiscuous functions would be enhanced to the same degree, which suggests an 'indiscriminate improver' scenario with no initial trade-off. This scenario is possible for the improvement of activity toward phosphoesters **1**-**3a** in ASs that maintain their existing sulfatase *specificity*, while improving their sulfatase *activity* (Figure 7B1-B5). Key elements of overall structure and active site are conserved (albeit across larger sequence variation than in laboratory evolution experiments), suggesting the 'indiscriminate improver' scenario is accessible. This scenario is likely to arise when higher sulfatase activity is required, whereas selection pressure toward improved phosphodiesterase activity appears unrestricted and can follow all possible pathways (Figure 8B). A major advantage of the ‘indiscriminate improver’ scenario is that it allows for improvement of a new activity without relying on neutral drift (*weak* initial negative trade-off) or the occurrence of intermediates with poor function (*strong* initial negative trade-off). Instead, increased selection pressure for the primary activity, e.g. sulfatase activity, lifts the level of the secondary activity, e.g. phosphodiesterase activity, above the selective threshold (Figure 9). This molecular exaptation scenario provides an additional adaptive route to previously suggested sub-functionalization models for protein evolution, such as the Escape from Adaptive Conflict (EAC) model[84] and the Innovation, Amplification and Divergence (IAD) model.[85] All three scenarios are based on the assumption that both old and new activities have to be able to co-exist in the same enzyme. However, for both the EAC and IAD models initial adaptation toward a new function is expected to occur at the cost of the original activity and requires a *generalist* intermediate that is brought about as a result of trade-off (and has low activity) and emerges through near-neutral evolution in the original activity. By contrast, our Exaptive Sub-Functionalization model (ESF, Figure 9) can start with and immediate adaptive phase for both original and new activity. It is valid even with a *specialist* with improved primary activity and its constant specificity avoids non-functional or low-activity intermediates during prior to adaptation towards a new activity. The ESF model requires no trade-off between optimal function and adaptability and a selectable level of the secondary activity is reached as a result of optimization toward the original primary activity.

Evolution of function in enzyme families. Prediction of structure-activity relationships and manipulation of catalytic activity based on evolutionary implications are inherently difficult, since only a small fraction of the sequence variation directly affects function. A larger fraction of the sequence variation may be neutral with regard to catalysis, and instead be important to stability and folding, allosteric control, protein interactions or other factors. Previous studies of catalytic promiscuity within an enzyme superfamily have focused on strongly diverged enzymes (i.e. with sequence identity ~15-25%).[74, 86-88] In the AP superfamily the sequence identity between previously studied family members is low: the similarity between human ASs and AP was not recognized by sequence comparison, but only when the first AS structure was solved.[89] The PMHs and ASs that were known prior to this study showed significant structural homology and the positions of the active site residues were largely identical. However, 5 of 11 active site positions were occupied by different residues and the metal ion in their active sites was different (Mn^{2+} in PMHs, Ca^{2+}/Mg^{2+} in ASs).

Despite sequence divergence, however, the maintenance of the AP superfamily fold and the promiscuity of all its known members enables the comparison of the 'classic' ASs and PMHs (Table S4) for which the new class of homodimeric sulfatases described in this work is a midway point. Here only 2 out the 6 variant active site groups (5 residues plus the metal ion) coincide with a change in primary activity. Assignment of the midway point and estimating the transition in specificity is only possible with experimental characterization. In terms of sequence the new class of sulfatases is closer to the PMHs than to the previously known ASs (Figure 3A, Table S1-S3), but their promiscuous activity patterns resemble those of classical ASs: the new class of ASs represents the phylogenetic clade closest to the point of functional divergence between PMHs and ASs. The availability of many experimentally characterized promiscuous family members that are close in sequence, but functionally divergent, provides a platform to study the specificity determinants of these enzymes and their phylogenetic relationship.

This analysis suggests that there can be larger genetic variation between two enzymes with the same primary function compared to the sequence variation between two enzymes that have diverged to have different primary functions (Figure 3A, Table S3). Therefore, somewhat counterintuitively, phylogenetic distances are not always good predictors for functional differences. This work provides a first example of a reactivity correlation across a possible transition in function. Such an analysis of chemical differentiation between substrate classes provides a framework that will complement sequence analysis in the understanding of evolution of functionally different enzymes, with possible impact on reconstruction of ancestral proteins and their evolutionary and chemical routes to and from extant enzymes.

SUPPORTING INFORMATION

The supporting information contains full synthetic procedures, additional information on other experimental procedures, protein purification, detailed kinetic measurements (Michaelis-Menten parameters, pH-rate profiles), X-ray crystallography processing statistics, additional figures with details on structural comparison and full description of sequence data and structural data used for phylogenetic analysis.

ACKNOWLEDGEMENTS

This research was funded by the Biological and Biotechnological Research Council (BBSRC) and the EU programme Horizon2020. CDB was supported by a BBSRC studentship and the Cambridge European Trust. FH is an ERC Advanced Investigator (695669). We are grateful for the European Synchrotron Radiation Facility and Diamond Light Source for provision of synchrotron radiation facilities and we would like to thank the beamline managers for the assistance in using beamlines ID23-1 and -2 at ESRF and I24 at DLS (proposal MX6889).

REFERENCES

(1) O'Brien, P. J.; Herschlag, D. *Chem. Biol.* **1999,** *6*, R91-R105.

(2) Bornscheuer, U. T.; Kazlauskas, R. J. *Angew. Chem. Int. Ed.* **2004,** *43*, 6032-6040.

(3) Nobeli, I.; Favia, A. D.; Thornton, J. M. *Nat. Biotechnol.* **2009,** *27*, 157-167.

(4) Khersonsky, O.; Tawfik, D. S. *Annu. Rev. Biochem.* **2010,** *79*, 471-505.

(5) Babtie, A.; Tokuriki, N.; Hollfelder, F. *Curr. Opin. Chem. Biol.* **2010,** *14*, 200-207.

(6) Busto, E.; Gotor-Fernandez, V.; Gotor, V. *Chem. Soc. Rev.* **2010,** *39*, 4504-4523.

(7) Copley, S. D. *Curr. Opin. Chem. Biol.* **2003,** *7*, 265-272.

(8) Copley, S. D. *J. Biol. Chem.* **2012,** *287*, 3-10.

(9) Copley, S. D. *Trends Biochem. Sci.* **2015,** *40*, 72-78.

(10) Gatti-Lafranconi, P.; Hollfelder, F. *ChemBioChem* **2013,** *14*, 285-292.

(11) Hult, K.; Berglund, P. *Trends Biotechnol.* **2007,** *25*, 231-238.

(12) Li, C.; Hassler, M.; Bugg, T. D. *ChemBioChem* **2008,** *9*, 71-76.

(13) Mohamed, M. F.; Hollfelder, F. *Biochim. Biophys. Acta* **2013,** *1834*, 417-424.

(14) O'Brien, P. J.; Herschlag, D. *Biochemistry* **2001,** *40*, 5691-5699.

(15) Patrick, W. M.; Quandt, E. M.; Swartzlander, D. B.; Matsumura, I. *Mol. Biol. Evol.* **2007,** *24*, 2716-2722.

(16) Poelarends, G. J.; Veetil, V. P.; Whitman, C. P. *Cell. Mol. Life Sci.* **2008,** *65*, 3606-3618.

(17) Rey, N. A.; Neves, A.; Bortoluzzi, A. J.; Pich, C. T.; Terenzi, H. *Inorg. Chem.* **2007,** *46*, 348-350.

(18) van Loo, B.; Jonas, S.; Babtie, A. C.; Benjdia, A.; Berteau, O.; Hyvonen, M.; Hollfelder, F. *Proc. Natl. Acad. Sci. U. S. A.* **2010,** *107*, 2740-2745.

(19) Jonas, S.; Hollfelder, F., Mechanism and Catalytic Promiscuity: Emerging Mechanistic Principles for Identification and Manipulation of Catalytically Promiscuous Enzymes. In *The Handbook of Protein Engineering*, Bornscheuer, U. T.; Lutz, S., Eds. Wiley-VCH: Chichester, 2008; Vol. 1, pp 47–72.

(20) Pandya, C.; Farelli, J. D.; Dunaway-Mariano, D.; Allen, K. N. *J. Biol. Chem.* **2014,** *289*, 30229-30236.

(21) Babtie, A. C.; Bandyopadhyay, S.; Olguin, L. F.; Hollfelder, F. *Angew. Chem. Int. Ed.* **2009,** *48*, 3692-3694.

(22) Ercan, A.; Park, H. I.; Ming, L. J. *Biochemistry* **2006,** *45*, 13779-13793.

(23) Lassila, J. K.; Herschlag, D. *Biochemistry* **2008,** *47*, 12853-12859.

(24) O'Brien, P. J.; Herschlag, D. *J. Am. Chem. Soc.* **1998,** *120*, 12369-12370.

(25) Olguin, L. F.; Askew, S. E.; O'Donoghue, A. C.; Hollfelder, F. *J. Am. Chem. Soc.* **2008,** *130*, 16547-16555.

(26) Bayer, C. D.; van Loo, B.; Hollfelder, F. *ChemBioChem* **2017,** *18*, 1001-1015.

(27) Jensen, R. A. *Annu. Rev. Microbiol.* **1976,** *30*, 409-425.

(28) Wackett, L. P. *Curr. Opin. Microbiol.* **2009,** *12*, 244-251.

(29) Jonas, S.; Hollfelder, F. *Pure Appl. Chem.* **2009,** *81*, 731-742.

(30) Zalatan, J. G.; Fenn, T. D.; Brunger, A. T.; Herschlag, D. *Biochemistry* **2006,** *45*, 9788-9803.

(31) Yang, K.; Metcalf, W. W. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101*, 7919-7924.

(32) Jonas, S.; van Loo, B.; Hyvonen, M.; Hollfelder, F. *J. Mol. Biol.* **2008,** *384*, 120-136.

(33) Berteau, O.; Guillot, A.; Benjdia, A.; Rabot, S. *J. Biol. Chem.* **2006,** *281*, 22464-22470.

(34) Miura, T.; Okamoto, K.; Yanase, H. *Biosci. Biotechnol. Biochem.* **2006,** *70,* 1509-1512.

(35) Shvetsova, S. V.; Zhurishkina, E. V.; Bobrov, K. S.; Ronzhina, N. L.; Lapina,

I. M.; Ivanen, D. R.; Gagkaeva, T. Y.; Kulminskaya, A. A. *J. Basic Microbiol.* **2015,** *55*, 471-479.

(36) Wiegmann, E. M.; Westendorf, E.; Kalus, I.; Pringle, T. H.; Lubke, T.; Dierks, T. *J. Biol. Chem.* **2013,** *288*, 30019-30028.

(37) Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Mc, Q. G. S.; Roodveldt, C.; Tawfik, D. S. *Nat. Genet.* **2005,** *37*, 73-76.

(38) Tokuriki, N.; Jackson, C. J.; Afriat-Jurnou, L.; Wyganowski, K. T.; Tang, R.; Tawfik, D. S. *Nat. Commun.* **2012,** *3*, 1257.

(39) Kaltenbach, M.; Emond, S.; Hollfelder, F.; Tokuriki, N. *PLoS Genet.* **2016,** \ *12*, e1006305.

(40) Zalatan, J. G.; Herschlag, D. *J. Am. Chem. Soc.* **2006,** *128*, 1293-1303.

(41) Baier, F.; Chen, J.; Solomonson, M.; Strynadka, N. C.; Tokuriki, N. *ACS Chem. Biol.* **2015,** *10*, 1684-1693.

(42) Baier, F.; Tokuriki, N. *J. Mol. Biol.* **2014,** *426*, 2442-2456.

(43) Starr, T. N.; Picton, L. K.; Thornton, J. W. *Nature* **2017,** *549*, 409-413.

(44) Podgornaia, A. I.; Laub, M. T. *Science* **2015,** *347*, 673-7.

(45) Galperin, M. Y.; Bairoch, A.; Koonin, E. V. *Protein Sci.* **1998,** *7*, 1829-1835.

(46) Galperin, M. Y.; Jedrzejas, M. J. *Proteins* **2001,** *45*, 318-324.

(47) Hendry, P.; Sargeson, A. M. *J. Am. Chem. Soc.* **1989,** *111*, 2521-2527.

(48) McWhirter, C.; Lund, E. A.; Tanifum, E. A.; Feng, G.; Sheikh, Q. I.; Hengge, A. C.; Williams, N. H. *J. Am. Chem. Soc.* **2008,** *130*, 13673-13682.

(49) Padovani, M.; Williams, N. H.; Wyman, P. *J. Phys. Org. Chem.* **2004,** *17*, 472-477.

(50) Stamatakis, A. *Bioinformatics* **2014,** *30*, 1312-1313.

(51) Miller, M. A.; Pfeiffer, W.; Schwartz, T. *2010 Gateway Computing Environments Workshop (GCE 2010)* **2010**, pp 1-8.

(52) Abascal, F.; Zardoya, R.; Posada, D. *Bioinformatics* **2005,** *21*, 2104-5.

(53) Dotson, S. B.; Smith, C. E.; Ling, C. S.; Barry, G. F.; Kishore, G. M. *J. Biol. Chem.* **1996,** *271*, 25754-25761.

(54) van Loo, B.; Schober, M.; Valkov, E.; Heberlein, M.; Bornberg-Bauer, E.; Faber, K.; Hyvonen, M.; Hollfelder, F. *J Mol Biol* **2018,** *430*, 1004-1023.

(55) Le, S. Q.; Gascuel, O. *Mol. Biol. Evol.* **2008,** *25*, 1307-1320.

(56) Carrico, I. S.; Carlson, B. L.; Bertozzi, C. R. *Nat. Chem. Biol.* **2007,** *3*, 321-322.

(57) Ghanem, E.; Li, Y.; Xu, C.; Raushel, F. M. *Biochemistry* **2007,** *46*, 9032-9040.

(58) Vogel, A.; Schilling, O.; Niecke, M.; Bettmer, J.; Meyer-Klaucke, W. *J. Biol. Chem.* **2002,** *277*, 29078-29085.

(59) Hanson, S. R.; Best, M. D.; Wong, C. H. *Angew. Chem. Int. Ed.* **2004,** *43*, 5736-5763.

(60) Beil, S.; Kehrli, H.; James, P.; Staudenmann, W.; Cook, A. M.; Leisinger, T.; Kertesz, M. A. *Eur. J. Biochem.* **1995,** *229*, 385-394.

(61) Beatty, K. E.; Williams, M.; Carlson, B. L.; Swarts, B. M.; Warren, R. M.; van Helden, P. D.; Bertozzi, C. R. *Proc. Natl. Acad. Sci. U. S. A.* **2013,** *110*, 12911-12916.

(62) Myette, J. R.; Soundararajan, V.; Behr, J.; Shriver, Z.; Raman, R.; Sasisekharan, R. *J. Biol. Chem.* **2009,** *284*, 35189-35200.

(63) Myette, J. R.; Soundararajan, V.; Shriver, Z.; Raman, R.; Sasisekharan, R. *J. Biol. Chem.* **2009,** *284*, 35177-35188.

(64) Ulmer, J. E.; Vilen, E. M.; Namburi, R. B.; Benjdia, A.; Beneteau, J.; Malleron, A.; Bonnaffe, D.; Driguez, P. A.; Descroix, K.; Lassalle, G.; Le Narvor, C.; Sandstrom, C.; Spillmann, D.; Berteau, O. *J. Biol. Chem.* **2014,** *289*, 24289-24303.

(65) Diez-Roux, G.; Ballabio, A. *Annu. Rev. Genomics Hum. Genet.* **2005,** *6*, 355-379.

(66) Kim, E. E.; Wyckoff, H. W. *J. Mol. Biol.* **1991,** *218*, 449-464.

(67) Boltes, I.; Czapinska, H.; Kahnert, A.; von Bulow, R.; Dierks, T.; Schmidt, B.; von Figura, K.; Kertesz, M. A.; Uson, I. *Structure* **2001,** *9*, 483-491.

(68) Krissinel, E.; Henrick, K. *Acta Crystallogr. D Biol. Crystallogr.* **2004,** *60*, 2256-2268.

(69) Hashimoto, K.; Panchenko, A. R. *Proc. Natl. Acad. Sci. U. S. A.* **2010,** *107*, 20352-20357.

(70) Hadler, K. S.; Tanifum, E. A.; Yip, S. H.; Mitic, N.; Guddat, L. W.; Jackson, C. J.; Gahan, L. R.; Nguyen, K.; Carr, P. D.; Ollis, D. L.; Hengge, A. C.; Larrabee, J. A.; Schenk, G. *J. Am. Chem. Soc.* **2008,** *130*, 14129-14138.

(71) Tawfik, D. S. *Curr. Opin. Chem. Biol.* **2014,** *21C*, 73-80.

(72) Lassila, J. K.; Zalatan, J. G.; Herschlag, D. *Annu. Rev. Biochem.* **2011,** *80*, 669-702.

(73) Buono, M.; Cosma, M. P. *Cell. Mol. Life Sci.* **2010,** *67*, 769-780.

(74) Gerlt, J. A.; Babbitt, P. C.; Jacobson, M. P.; Almo, S. C. *J. Biol. Chem.* **2012,** *287*, 29-34.

(75) Stein, C.; Gieselmann, V.; Kreysing, J.; Schmidt, B.; Pohlmann, R.; Waheed, A.; Meyer, H. E.; O'Brien, J. S.; von Figura, K. *J. Biol. Chem.* **1989,** *264*, 1252-1259.

(76) Levy-Frebault, V. V.; Portaels, F. *Int J Syst Bacteriol* **1992,** *42*, 315-23.

(77) Smith, E. L.; Bertozzi, C. R.; Beatty, K. E. *ChemBioChem* **2014,** *15*, 1101-5.

(78) Colin, P. Y.; Kintses, B.; Gielen, F.; Miton, C. M.; Fischer, G.; Mohamed, M. F.; Hyvonen, M.; Morgavi, D. P.; Janssen, D. B.; Hollfelder, F. *Nat. Commun.* **2015,** *6*, 10008.

(79) Mougous, J. D.; Green, R. E.; Williams, S. J.; Brenner, S. E.; Bertozzi, C. R. *Chem Biol* **2002,** *9*, 767-76.

(80) Edwards, D. R.; Lohman, D. C.; Wolfenden, R. *J. Am. Chem. Soc.* **2012,** *134*, 525-531.

(81) Wiersma-Koch, H.; Sunden, F.; Herschlag, D. *Biochemistry* **2013,** *52*, 9167-9176.

(82) Khersonsky, O.; Roodveldt, C.; Tawfik, D. S. *Curr. Opin. Chem. Biol.* **2006,** *10*, 498-508.

(83) Nedrud, D. M.; Lin, H.; Lopez, G.; Padhi, S. K.; Legatt, G. A.; Kazlauskas, R. J. *Chem. Sci.* **2014,** *15*, 1931-1938

(84) Sikosek, T.; Chan, H. S.; Bornberg-Bauer, E. *Proc. Natl. Acad. Sci. U. S. A.* **2012,** *109*, 14888-14893.

(85) Nasvall, J.; Sun, L.; Roth, J. R.; Andersson, D. I. *Science* **2012,** *338*, 384-387.

(86) Glasner, M. E.; Gerlt, J. A.; Babbitt, P. C. *Curr. Opin. Chem. Biol.* **2006,** *10*, 492-497.

(87) O'Maille, P. E.; Malone, A.; Dellas, N.; Andes Hess, B., Jr.; Smentek, L.; Sheehan, I.; Greenhagen, B. T.; Chappell, J.; Manning, G.; Noel, J. P. *Nat. Chem. Biol.* **2008,** *4*, 617-623.

(88) Sunden, F.; AlSadhan, I.; Lyubimov, A. Y.; Ressl, S.; Wiersma-Koch, H.; Borland, J.; Brown, C. L., Jr.; Johnson, T. A.; Singh, Z.; Herschlag, D. *J. Am. Chem. Soc.* **2016,** *138*, 14273-14287.

(89) Bond, C. S.; Clements, P. R.; Ashby, S. J.; Collyer, C. A.; Harrop, S. J.; Hopwood, J. J.; Guss, J. M. *Structure* **1997,** *5*, 277-289.

(90) Kintses, B.; Hein, C.; Mohamed, M. F.; Fischlechner, M.; Courtois, F.; Laine, C.; Hollfelder, F. *Chem. Biol.* **2012,** *19*, 1001-1009.

(91) Hernandez-Guzman, F. G.; Higashiyama, T.; Pangborn, W.; Osawa, Y.; Ghosh, D. *J. Biol. Chem.* **2003,** *278*, 22989-22997.

(92) Lukatela, G.; Krauss, N.; Theis, K.; Selmer, T.; Gieselmann, V.; von Figura, K.; Saenger, W. *Biochemistry* **1998,** *37*, 3654-3664.

(93) Nikolic-Hughes, I.; O'Brien P, J.; Herschlag, D. *J. Am. Chem. Soc.* **2005,** *127*, 9314-9315.

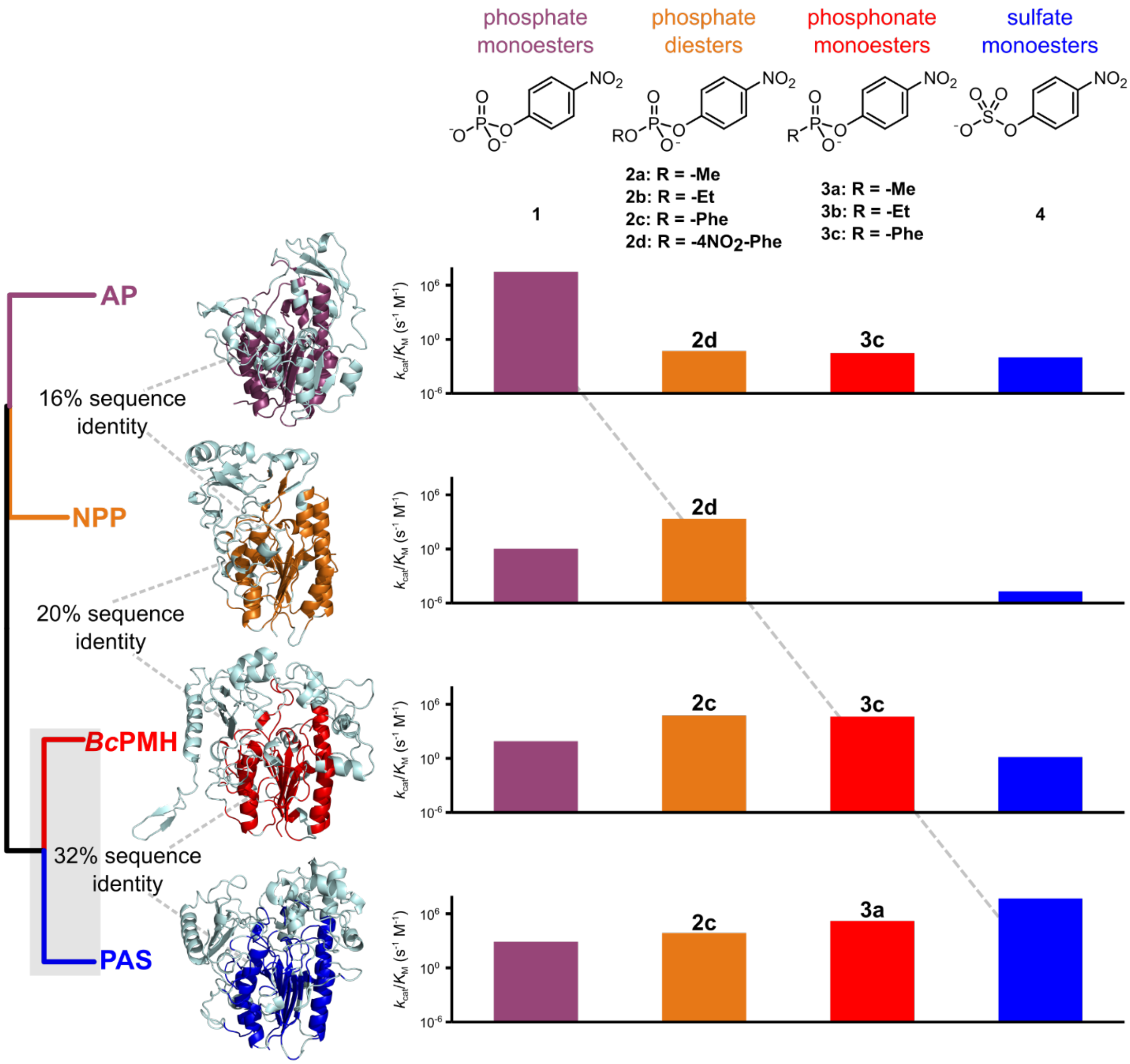


Figure 1. Catalytic promiscuity in the AP superfamily.

Matrix representation of the crosswise catalytic promiscuity in the AP superfamily, in which the primary activity (on the diagonal) of one enzyme is often a promiscuous activity for its family members.[13, 29] The color scheme matches the classification of the enzyme with the respective main activity (phosphate monoesters **1**: purple; phosphate diesters **2**: orange; phosphonate monoesters **3**: red; sulfate monoesters **4**: blue). In phosphodiesters **2a**-**2c** and phosphonate monoesters **3a-3c** the respective R-groups are: **a**:-CH_3 (methyl), **b**: C_2H_5 (ethyl), and **c**: C_6H_5 (phenyl) and the leaving group was 4-nitrophenolate in each case. All X-ray structures are shown as single protomers. The grey-shaded clade represents the AS/PMH subgroup that is expanded with several other ASs and

PMHs in Figure 3A. The representative enzymes structures shown are *E. coli* alkaline phosphatase (AP; PDB: 1ED9),[1, 14] *Xanthomonas axonopodis* nucleotide phosphodiesterase/pyrophosphatase (NPP; 2GSN),[23, 30] *Burkholderia caryophili* phosphonate monoester hydrolase (*Bc*PMH; 2W8S)[18] and *Pseudomonas aeruginosa* arylsulfatase (PAS; 1HDH).[21, 25, 90]

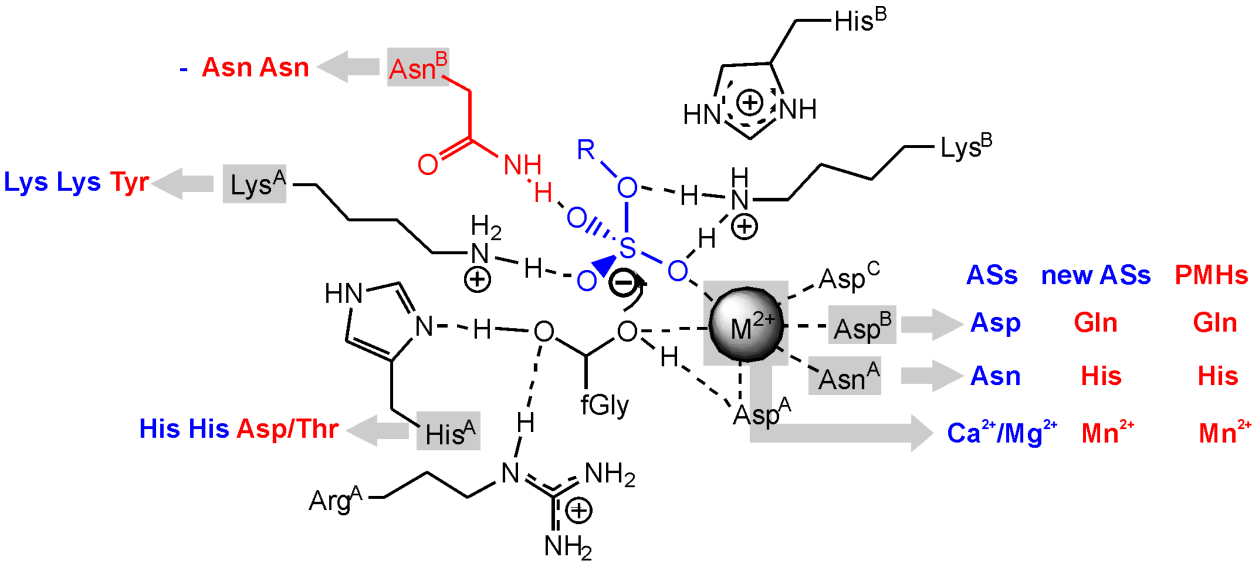


Figure 2. A unifying active site description of previously known arylsulfatases (ASs), the newly identified ASs (‘new ASs’) and phosphonate monoester hydrolases (PMHs).
The variation of active site residues within the AS/PMH sub-family (part of the AP superfamily of enzymes) is described using nomenclature introduced for ASs by Hanson *et al.*.[59] The reaction involves nucleophilic attack by a formylglycine (fGly) residue that is formed as the result of the post-translational modification of a cysteine embedded in the conserved Ser/Cys-X-Pro/Ala-X-Arg recognition motif that is present in all AP-type ASs and PMHs. coordinated by a bivalent metal ion (Mn^{2+} in PMHs[18, 32, 53] and Mg^{2+} or Ca^{2+}in previously described ASs[67, 89, 91-92]) and a conserved arginine (Arg^A), assisted by leaving group stabilization by a histidine-lysine pair (His^B and Lys^B). In addition to these invariant features, a number of metal-coordinating (Asn^A and Asp^B) and hydrogen bonding or cationic (His^A, Lys^A and Asn^B) residues are characteristic identifiers of the three enzyme classes described in this paper (see the color coding: PMHs in red, the ASs in blue). Of these Asn^B (highlighted in pink) is an addition to the Hanson nomenclature (as the structure of *Rl*PMH[32] shows it in the active site vicinity, possibly interacting with one the non-bridging oxygens in the substrate). The metal coordinating residues Asn^A and Asp^B are known to vary among known ASs. In PMH Asn^A is a histidine, but most likely has the same role as in PMHs (metal coordination), whereas Asp^B is a glutamine that does not interact with the metal ion, but instead interacts with Lys^B. In *Rl*PMH and *Bc*PMH, His^A is a threonine, whereas Lys^A is a tyrosine that is rotated away from the position occupied by Lys^A in ASs.[32]. For details on the nature of the

conserved active site residues in all PMHs and dimeric ASs described in this study, as well as PAS, see Table S4.

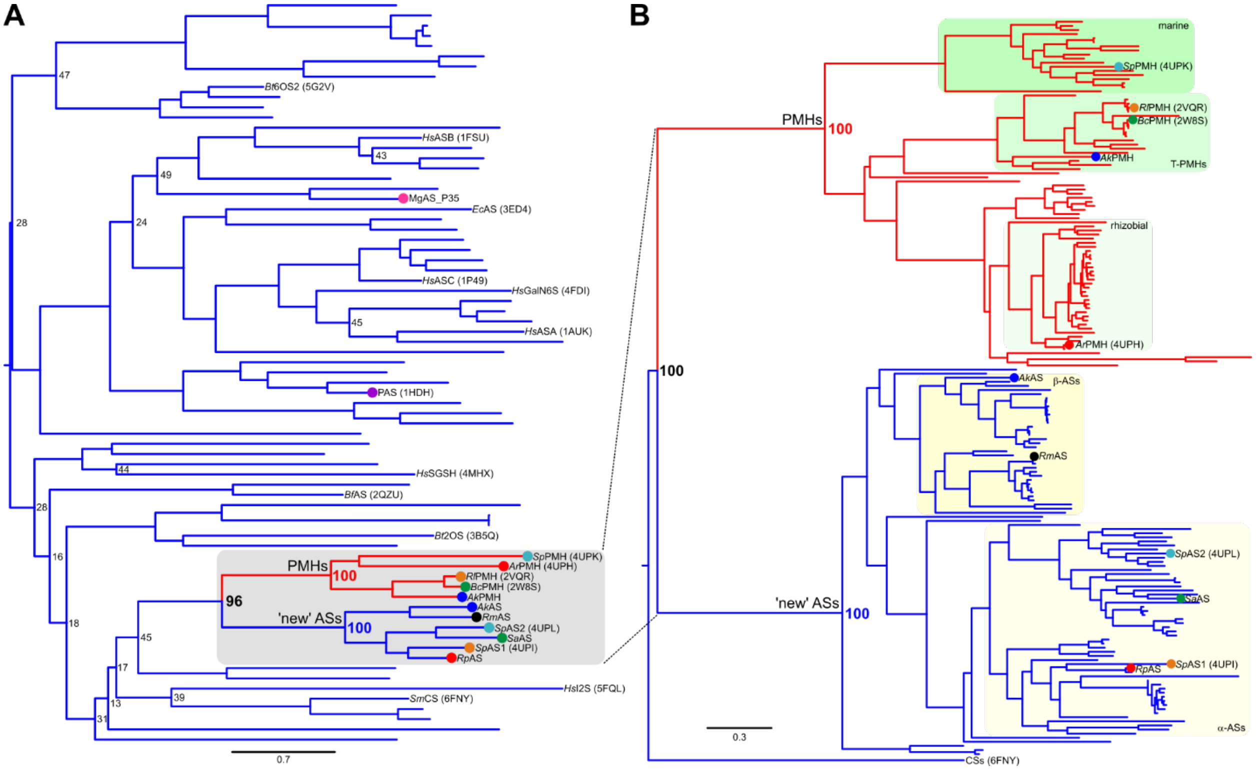


Figure 3. Maximum likelihood phylogenetic relationships between phosphonate monoester hydrolases (PMHs, red) and arylsulfatases (ASs, blue).

(**A**) Phylogenetic relationship of the newly identified ASs ('new' ASs) and PMHs presented in this study (Table 1) and the previously described *Bc*PMH[18, 53] and *Rl*PMH[32] as well as *Pseudomonas aeruginosa* AS (PAS),[21, 25, 60, 67] and an AS recently discovered in a metagenomic library (mAS_P35)[78] (as indicated by a purple and pink circle respectively) with all other ASs with experimentally verified function (Table S5). The sequences were aligned using the structure-guided expresso mode of T-coffee. Sulfatases for which the X-ray structure is known are indicated with their respective PDB IDs. The newly identified dimeric ASs ('new' ASs) are phylogenetically more closely related to the functionally different PMHs than they are to the other known ASs, although they are identified in this paper as sulfatases (See Figure S1 and Table S5 for details on the sequences included). The bootstrap support for the node from which the PMHs and new ASs diverge is high, however the exact placement of the AS/PMH clade in relation to all other known ASs is currently unresolved, as indicated by the low bootstrap support values for the nodes preceding the node from which the new ASs and the PMHs diverge. Trees based on multiple sequence alignments generated with other methods (Figure S2-S4) and a tree built using Bayesian maximum likelihood (Figure S5) all support the latter two observations. (**B**) Expansion of the phylogenetic clade indicated in panel (**A**) with putative 'new' ASs and PMHs, rooted with an expanded clade of choline sulfatases (represented as CSs), as described previously[54]). The average sequence identity was 51±12% and 49±18% within the AS and PMH clade respectively (average overall sequence identity between ASs and PMHs is 38±16%). The 'new' ASs all have active site residues that are identical to those found in the six new ASs listed in Table 1 (Table S4). The new AS-clade shows two distinct subclades, which both correspond to one or more of the indicated experimentally characterized enzymes (colored circles). These two subclades largely correspond to the taxonomic/phylogenetic classification of the α- and β-proteobacteria (respectively named α-ASs and β-ASs). The PMH clade (85 PMH sequences in total) shows three distinct phylogenetic subclades that each correspond to one or more of the experimentally characterized enzymes (colored circles). PMHs originating from the same type of environment (marine or rhizobial) show a tendency to cluster in the same subclade. The T-PMHs also cluster together and appear to have evolved out of the D-PMHs. For details on the sequences involved see Figure S6 and S7 and Table S6 and S7. The procedures for constructing both trees are described in the materials and methods. Trees based on alternative multiple sequence alignments (Figure S8B-D) or built using Bayesian maximum likelihood (Figure S9B) showed highly similar topologies.

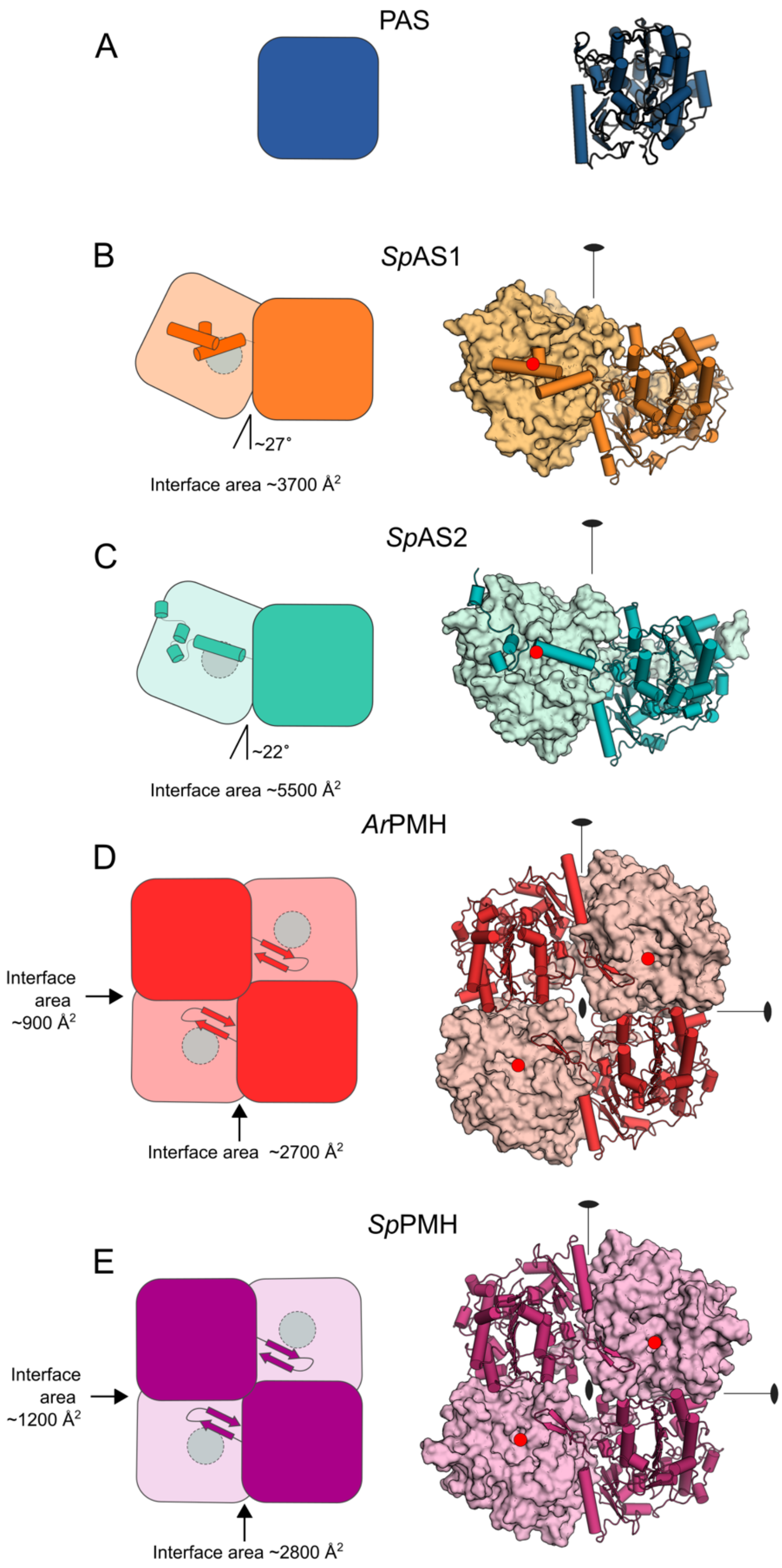
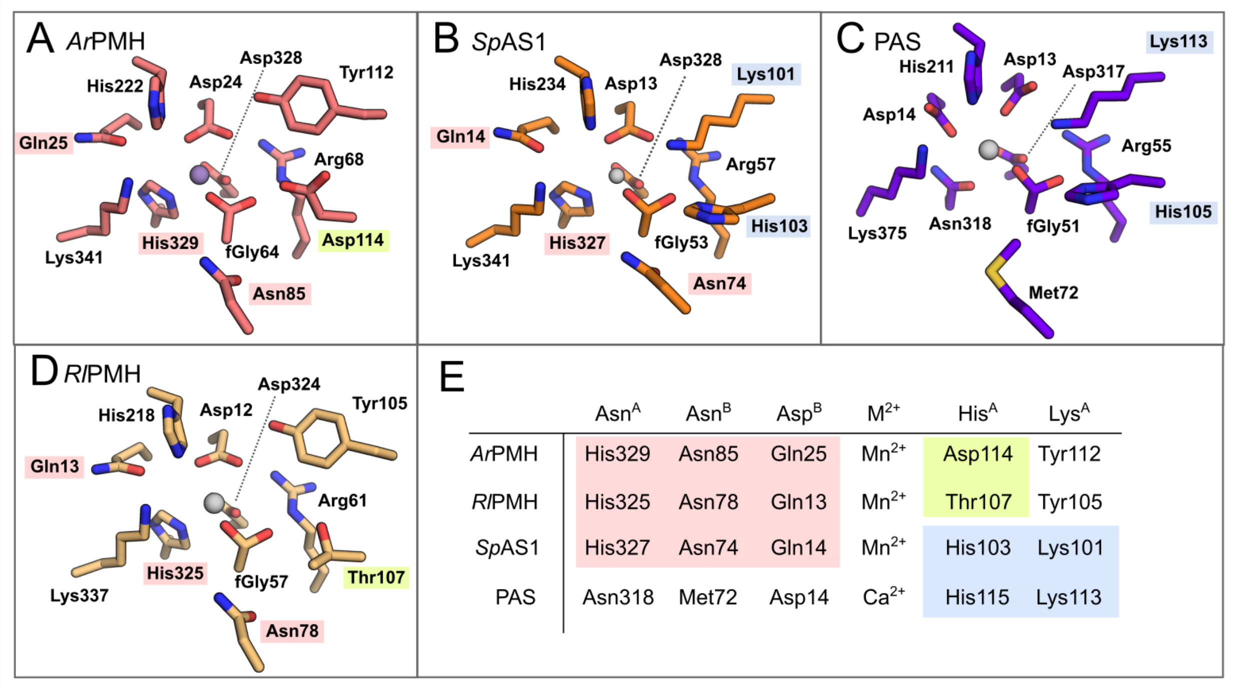

A
PAS
B
*Sp*AS1
~27°
Interface area ~3700 Å^2
C
*Sp*AS2
~22°
Interface area ~5500 Å^2
D
*Ar*PMH
Interface area ~900 Å^2
Interface area ~2700 Å^2
E
*Sp*PMH
Interface area ~1200 Å^2
Interface area ~2800 Å^2

Figure 4. Overall structural organization of the hydrolases described in this work, highlighting the organization of structural features responsible for oligomerization. The characteristic fold of members of the AP superfamily consists of a central β-sheet sandwiched between α-helices. The active site is largely formed by the side chains of residues that are part of loop regions on top of the central β-sheet and nucleophile and several positively charged groups that bind and activate the substrate are located here. This figure illustrates how such protomers are assembles in the following AP superfamily members: (**A)** Monomeric PAS**,** (**B)** dimeric *Sp*AS1, **(C)** dimeric *Sp*AS2, (**D**) tetrameric *Ar*PMH and **(E)** tetrameric *Sp*PMH. The right of each panel shows one protomer for each dimeric unit as a cartoon in darker color, with the molecular surface shown for the other protomer. The oligomeric structure is shown schematically on the left. The red dot in each of the surface-rendered protomers indicates the position of the active site fGly residue, which is located in the vicinity of the incoming C-terminal extensions. The oligomeric state observed in the X-ray structures is mirrored by size-exclusion chromatography experiments, where the PMHs elute as tetramers and the 'new' ASs as dimers (Figure S14).



	Asn^{A}	Asn^{B}	Asp^{B}	M^{2+}	His^{A}	Lys^{A}
*Ar*PMH	His329	Asn85	Gln25	Mn^{2+}	Asp114	Tyr112
*Rl*PMH	His325	Asn78	Gln13	Mn^{2+}	Thr107	Tyr105
*Sp*AS1	His327	Asn74	Gln14	Mn^{2+}	His103	Lys101
PAS	Asn318	Met72	Asp14	Ca^{2+}	His115	Lys113

Figure 5. Structural alignment of the active sites of arylsulfatases and phosphonate monoester hydrolases.

Active site residues for the classical ASs (e.g. PAS, panel **C**), newly identified ASs (*Sp*AS1, **B**), D-PMHs (*Ar*PMH, **A**) and T-PMHs (*Rl*PMH, **D**). All images are kept in the same orientation based on a multiple structural alignment that included all structures determined in this study and those of PAS,[67] *Bc*PMH[18] and *Rl*PMH.[32] The table (panel **E**) lists the active site residues for these four enzymes that differentiate the classical ASs from PMHs. *Sp*AS1 shows similarities to both the classical PAS (blue shading) and *Ar*PMH and *Rl*PMH (red shading), emphasizing its intermediate position. Structural alignment data (similarity scores, conserved residues) between ASs and PMHs of known structure are listed in Table S1-S4.

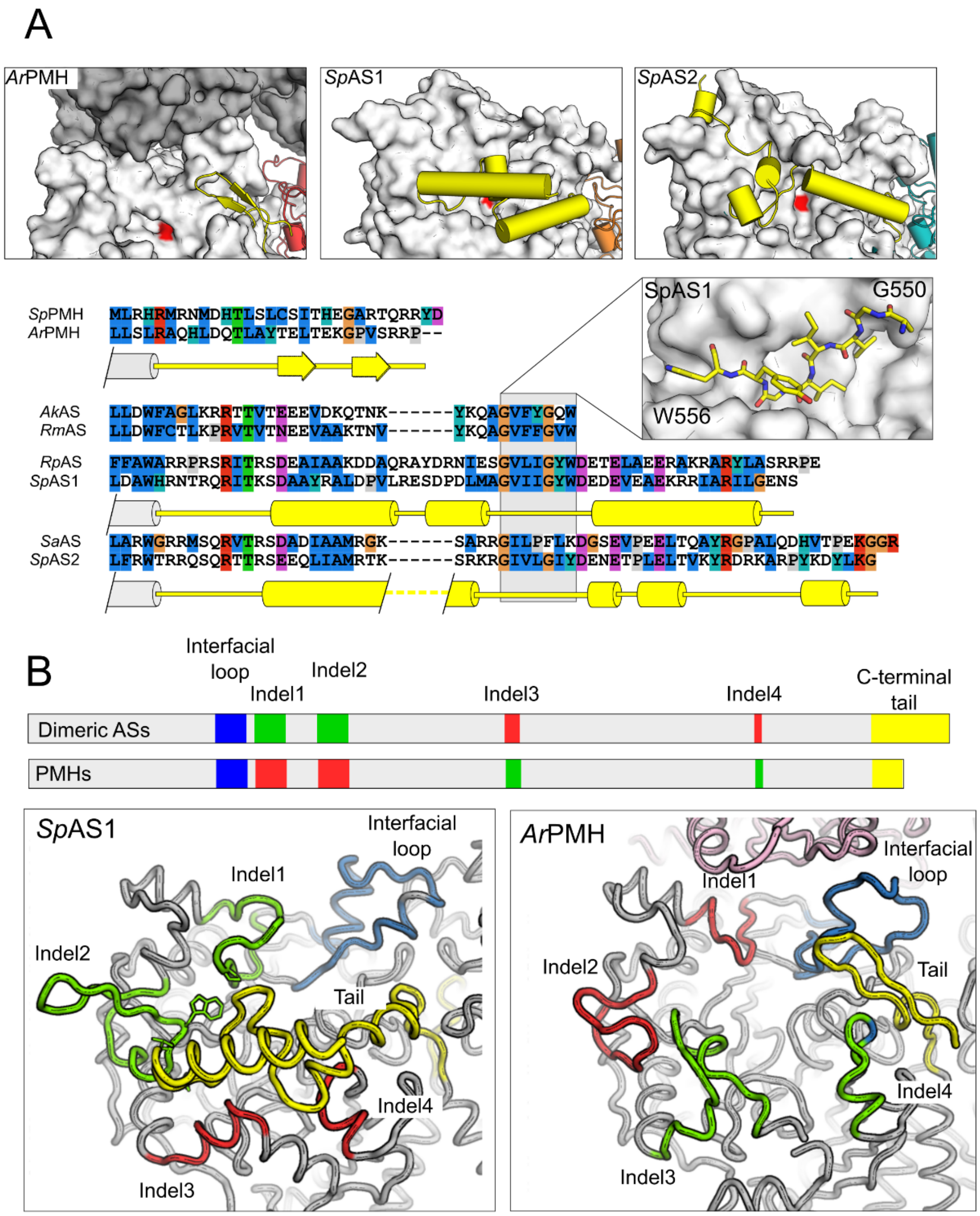

A
ArPMH
SpAS1
SpAS2
SpAS1
G550
W556
SpPMH MLRHRMRNMDHTLSLCSITHEGARTQRRYD
ArPMH LLSLRAQHLDQTLAYTELTEKGPVSRRP--
AkAS LLDWFAGLKRRTTVTEEEVDKQTNK------YKQAGVFYGQW
RmAS LLDWFCTLKPRVTVTNEEVAAKTNV------YKQAGVFFGVW
RpAS FFAWARRPRSRITRSDEAIAAKDDAQRAYDRNIESGVLIGYWDETELAEERAKRARYLASRRPE
SpAS1 LDAWHRNTRQRITKSDAAYRALDPVLRESDPDLMAGVIIGYWDEDEVEAEKRRIARILGENS
SaAS LARWGRRMSQRVTRSDADIAAMRGK------SARRGILPFLKDGSEVPEELTQAYRGPALQDHVTPEKGGR
SpAS2 LFRWTRRQSQRTTRSEEQLIAMRTK------SRKRGIVLGIYDENETPLELTVKYRDRKARPYKDYLKG
B
Interfacial loop
Indel1
Indel2
Indel3
Indel4
C-terminal tail
Dimeric ASs
PMHs
SpAS1
Interfacial loop
Indel1
Indel2
Tail
Indel4
Indel3
ArPMH
Interfacial loop
Indel1
Indel2
Tail
Indel4
Indel3

Figure 6. Oligomerization of novel dimeric ASs and PMHs.

(A) *C-terminal tail sequences are involved in dimerization*. The 3D-representation (top panel) contains helical features of C-termini (yellow) that bind in a well-defined groove on the surface of the other protomer. The alignments below combine the C-terminal sequences for selected members of each sub-group of AP-superfamily members (with secondary structure elements indicated underneath the sequences). The alignment starts at the end of the structurally conserved α-helix and the sulfatase sequences are grouped by conservation to the three different types of extensions observed in the complete alignment. The alignment highlights conservation: blue for conservation of hydrophobic residues, red for positively charged and purple for negatively charged residues; brown for glycines, green for Ser/Thr and grey for prolines. The conserved hydrophobic motifs in the sulfatases is boxed in gray and the binding of that sequence in *Sp*AS1 is shown in the blow-up diagram (with the motif shown as sticks; the rest of the tail structure is removed for clarity)**.** (**B**) *Oligomerization is enabled by insertions and deletions (indels) at the interface*. A schematic representation of the novel PMHs and ASs sequences highlights the features involved in tetramerization: red and green boxes respectively indicate the absence or presence of a stretch of amino acids, and the blue box indicates the interfacial loop that has a different conformation in the respective enzyme classes (see Figure S20 for the detailed multiple sequence alignment that underpins this analysis). The C-terminal tail (see also panel **A**) is shown in yellow. The tube cartoons of the two structures in the lower panels are colored in identical way to the schematic diagram, showing how the indels map onto the dimer interface. The blue loop shows significant difference in conformation between *Sp*AS1 and *Ar*PMH. The pink molecule seen at the top of the *Ar*PMH panel is the next protomer in the tetramer.

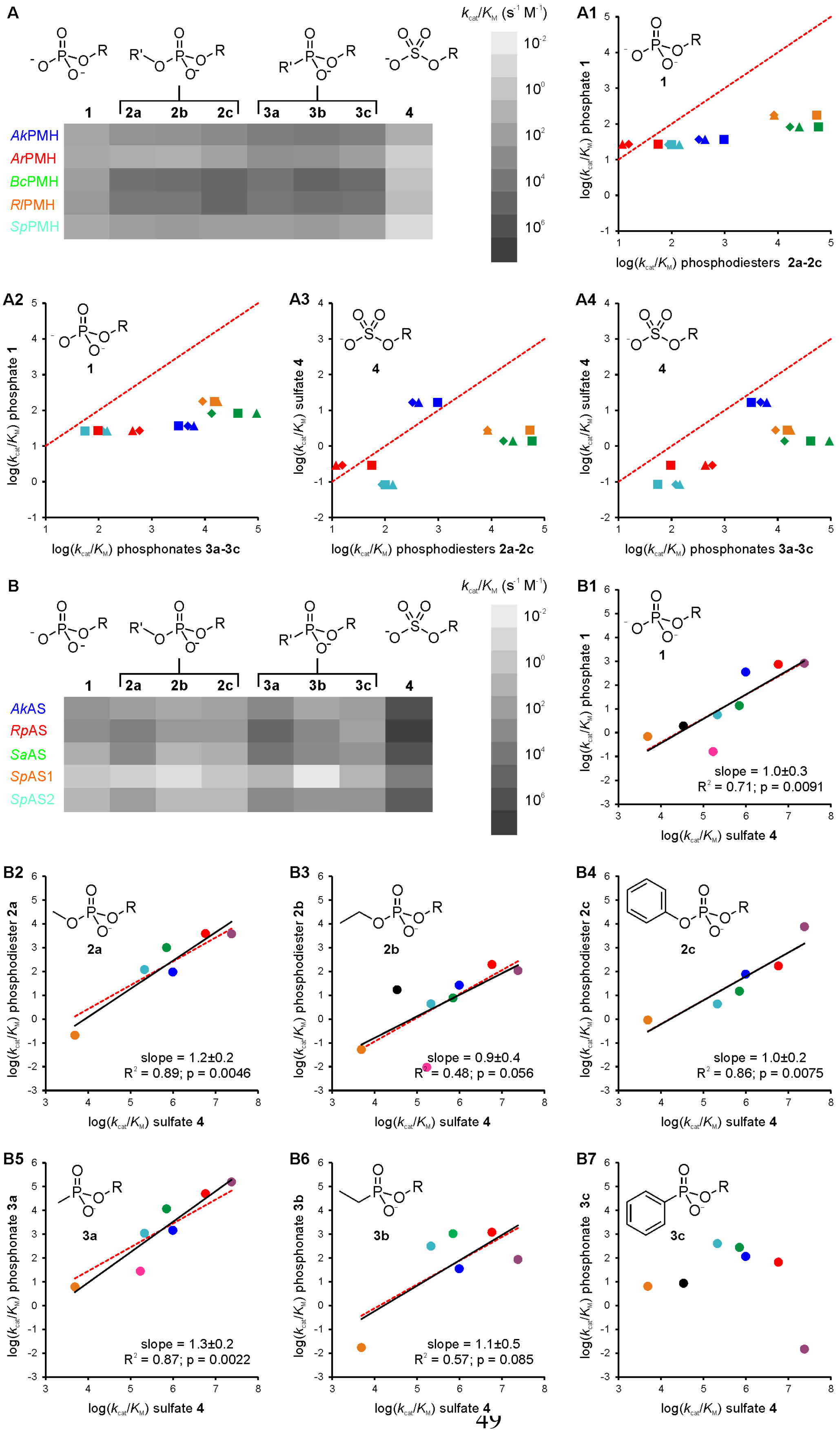

A
1
2a
2b
2c
3a
3b
3c
4
AkPMH
ArPMH
BcPMH
RlPMH
SpPMH
k_{cat}/K_M (s^{-1} M^{-1})
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}
A1
log(k_{cat}/K_M) phosphate 1
log(k_{cat}/K_M) phosphodiesters 2a-2c
A2
log(k_{cat}/K_M) phosphate 1
log(k_{cat}/K_M) phosphonates 3a-3c
A3
log(k_{cat}/K_M) sulfate 4
log(k_{cat}/K_M) phosphodiesters 2a-2c
A4
log(k_{cat}/K_M) sulfate 4
log(k_{cat}/K_M) phosphonates 3a-3c
B
AkAS
RpAS
SaAS
SpAS1
SpAS2
B1
log(k_{cat}/K_M) phosphate 1
log(k_{cat}/K_M) sulfate 4
slope = 1.0±0.3
R^2 = 0.71; p = 0.0091
B2
log(k_{cat}/K_M) phosphodiester 2a
slope = 1.2±0.2
R^2 = 0.89; p = 0.0046
B3
log(k_{cat}/K_M) phosphodiester 2b
slope = 0.9±0.4
R^2 = 0.48; p = 0.056
B4
log(k_{cat}/K_M) phosphodiester 2c
slope = 1.0±0.2
R^2 = 0.86; p = 0.0075
B5
log(k_{cat}/K_M) phosphonate 3a
slope = 1.3±0.2
R^2 = 0.87; p = 0.0022
B6
log(k_{cat}/K_M) phosphonate 3b
slope = 1.1±0.5
R^2 = 0.57; p = 0.085
B7
log(k_{cat}/K_M) phosphonate 3c
log(k_{cat}/K_M) sulfate 4

Figure 7. Correlation of promiscuous and primary activities of phosphonate monoester hydrolases (A) and arylsulfatases (B).

Catalytic efficiencies (k_{cat}/K_M) measured for various promiscuous substrates are shown as heat maps for PMHs (in panel **A**) and ASs (in panel **B**). The correlation between the activities towards primary and promiscuous substrates is shown as plots that display k_{cat}/K_M-values of the respective primary substrates on the x-axis against the observed promiscuous substrates on the y-axis. The PMHs in **panels A1-A4** [*Ak*PMH (blue), *Ar*PMH (red), *Bc*PMH (green), *Rl*PMH (orange) and *Sp*PMH (turquoise)] were tested with the primary substrates phosphodiesters **2a**/**b**/**c** (**A1** and **A3)** or phosphonate monoesters **3a**/**b**/**c**) **A2** and **A4)** and correlated with the data for phosphate monoesters **1** and sulfates **4**. For the PMHs the respective specificity $(k_{cat}/K_M)_{primary}/(k_{cat}/K_M)_{promiscuous}$ increases with higher levels of primary activity (as shown in Figure S22A1-A4), resulting in an increasing propensity for the data points in panels **A1-4** to deviate from the simulated curve with a slope of 1. The ASs **in panels B1-B7** are shown as circles [*Ak*AS (blue), *Rp*AS (red), *Sa*AS (green), *Sp*AS1 (orange), *Sp*AS2 (turquoise), *Rm*AS (black), mAS_P35 (pink) and PAS (purple)]. The correlation between primary and promiscuous activities for the ASs could be fitted with a slope near unity (black solid line, slope indicated at the bottom right of the graph) for panels **B1-6**, indicating that the *specificity* of ASs does not change with increasing primary activity. (The constant specificity is also directly shown in Figure S22B1-B6, based on the same data as this Figure). The scenarios in panels **B1-6** suggest that two activities can increase simultaneously without trade-off, revealing them as 'indiscriminate improvers' (see Figure 8B). To guide the eye, red dotted curves were drawn that represent scenarios for the correlation between the activity toward promiscuous and primary substrates in which the specificity $(k_{cat}/K_M)_{primary}/(k_{cat}/K_M)_{promiscuous}$ is constant, i.e. the plot of $\log[(k_{cat}/K_M)_{promiscuous}]$ *vs.* $\log[(k_{cat}/K_M)_{primary}]$ has a slope of 1. Their position was chosen based on the assumption that the enzymatic reaction is either non-specific (y-axis intercept = 0, panels **A1** and **A2**), has a specificity of 10^2-fold (y-axis intercept = -2, panels **A3** and **A4**) or is equal to 10^2-10^5fold (y-axis intercept = mean of $\log[(k_{cat}/K_M)_{promiscuous}/(k_{cat}/K_M)_{primary}]$ (panels **B1-6**). The color scheme matches the color code used in Figure 3. For a detailed list of all catalytic parameters (k_{cat}, K_M and k_{cat}/K_M) see Table S18-30.

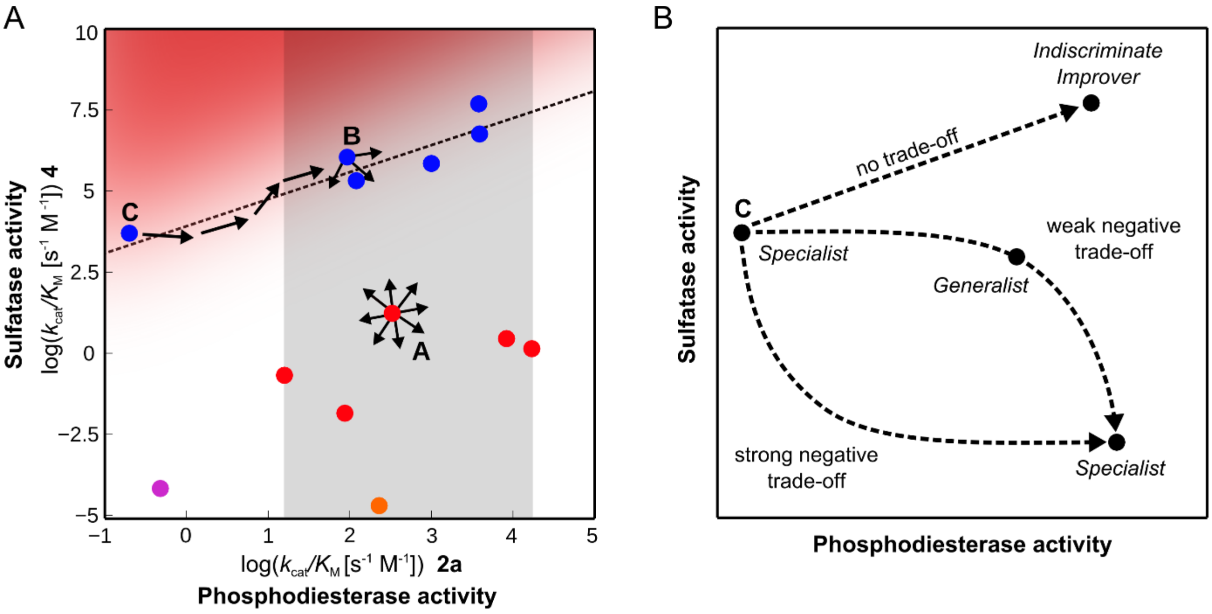


Figure 8. Trade-off analysis between promiscuous activities reveals empirical specificity limits.

(A) The trade-off plot of k_{cat}/K_M values for phosphodiester **2a** and sulfate monoester **4** shows several PMHs (red), ASs (blue), *E. coli* alkaline phosphatase (AP, mutant R166S,[93] purple) and *X. axonopodis* nucleotide pyrophosphatase/phosphodiesterase[23, 30] (NPP, orange). This ‘promiscuity landscape’ visualises the apparent specificity limit observed for arylsulfatases (Figure 7B1-6): none of the family members are found beyond the fitted correlation line, i.e. the dark red-shaded region appears ‘forbidden’ for this group of enzymes, despite their activities varying by more than three orders of magnitude. The lightly-shaded area by contrast represents empirically a combination of k_{cat}/K_M-values where activities for both reaction types can vary independently. Arrows indicate possible directions of adaptation for enzymes A (*Ak*PMH) and B (*Ak*AS). In one scenario B should be unable to evolve towards higher sulfatase specificity (i.e. ratio between sulfatase and phosphdiester activities for the same enzyme). By contrast, in a scenario for A this enzyme should not be constrained and thus able to change specificity in all directions. Several of the ASs catalyze the hydrolysis of phosphodiester **2a** at a similar level as PMHs and NPP, suggesting that ASs catalyze phosphodiester hydrolysis at ‘native’ levels. (The range of k_{cat}/K_M-values for ‘native’ phosphodiesterases is indicated by gray shading.) An adaptive trajectory from the relatively inefficient sulfatase *Sp*AS1 (denoted as “C” in the plot, $k_{cat}/K_M = 5.0 \times 10^3$ s^{-1} M^{-1}) will result in variants that improve in sulfatase, but also catalyze phosphodiester hydrolysis at 'native-like' levels, i.e. with k_{cat}/K_M-values similar to those of PMHs and NPP for phosphodiester **2a** (e.g. enzyme B). This can be seen as an example of molecular exaptation, i.e. selective pressure for one trait gives rise to another selectable trait (see Figure 9). (**B)** The transition between sulfatase and phosphodiesterase activities that is suggested by our mapping of the activity space of the extant enzymes (Figure 7B1-B6) is compared to a frequently used representation of trade-off that distinguishes between scenarios of initial weak and strong negative trade-off. Our observations differ from this familiar scenario: for example, evolution of a phosphodiesterase from e.g. a low-activity arylsulfatase (such as *Sp*AS1, C) may proceed *via* a pathway without a trade-off, leading to an enzyme that is improved in *both* activities (indiscriminate improver). This novel scenario of exaptive subfunctionalisation (ESF) implies that there is no selection pressure for increased *specificity*, but only for improved rates for one of the two reactions and the other one follows suit.

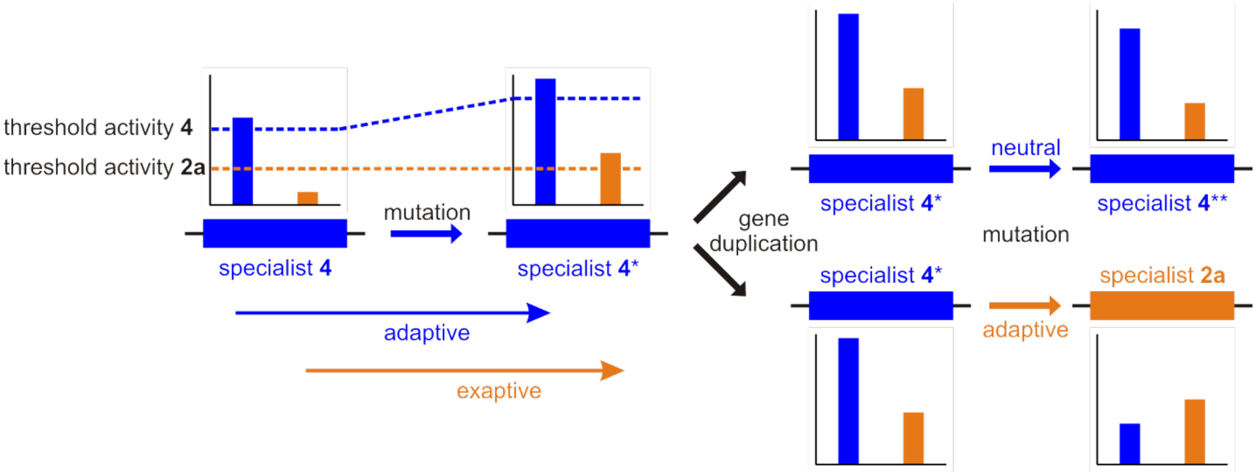


Figure 9. Possible evolution of a phosphodiesterase from a sulfatase rationalized by the exaptive sub-functionalization (ESF) model.

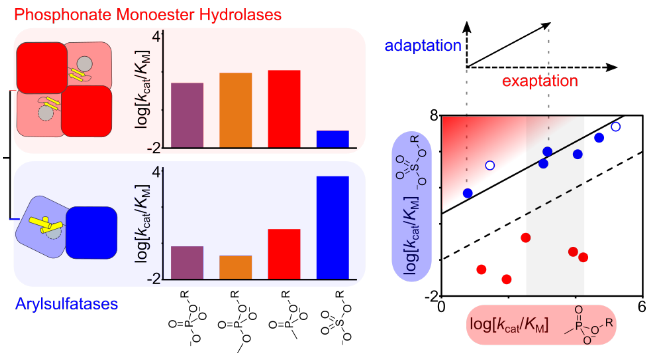
The evolution of a new primary function, in this case phosphodiesterase activity, could occur via gene duplication followed by accumulation of mutations in each copy that is under selection for either phosphodiesterase (**2a**) or sulfatase (**4**) activity. Since non-functional genes are easily lost during evolution, the new function has to evolve to a sufficient level prior to gene duplication to assure both gene copies are maintained after the duplication. Here we propose an exaptive sub-functionalization (ESF) model and exemplify it with the evolution of a phosphodiesterase from a sulfatase (e.g. *Sp*AS1): an increased selective pressure for sulfatase function results in an increase in sulfatase (adaptive) and phosphodiesterase (exaptive) activity. This scenario mimics the evolutionary pathway from variant C to B indicated in the cartoon of 'functional space' in Figure 8A. This model is distinguished from other sub-functionalization models such as EAC and IAD (where a trade-off between optimal performance and adaptability is invoked involving a low activity generalist, i.e. a specialist first needs to re-generalize or a generalist needs to duplicate and morph into two specialists in order satisfy multiple evolutionary requirements). The ESF model assumes than an existing activity can drive the promiscuous activity up to a selectable level (and no initial trade-off needed). The phase preceding possible gene duplication and functional divergence in the EAC and IAD models is near-neutral for the original activity, whereas in the ESF model this phase is adaptative for *both* activities. A scenario in which the specificity is reversed and the enzyme is, in first instance, under selection pressure for phosphodiesterase activity is far less likely to follow the exaptive scenario when evolving toward increased sulfatase activity.

Table 1: The various phosphonate monoester hydrolase (PMH) arysulfatase (AS) genes characterized in this study.

Name	Source	Accession number[a]	Annotation in NCBI database
*Ak*AS	*Advenella kashmirensis* WT001	AJA37533[b]	dimeric sulfatase
*Rm*AS	*Ralstonia metallidurans* CH34	ABF08681	sulfatase
*Rp*AS	*Rhodopseudomonas palustris* CGA009	CAE26808	putative sulfatase
*Sa*AS	*Stappia aggregata* IAM 12614	EAV45217	sulfatase family protein
*Sp*AS1	*Silicibacter pomeroyi* DSS-3	AAV97258	sulfatase family protein
*Sp*AS2	*Silicibacter pomeroyi* DSS-3	AAV96818	sulfatase family protein
*Ak*MPH	*Advenella kashmirensis* WT001	AFK62654	sulfatase
*Ar*PMH	*Agrobacterium radiobacter* K84	WP_012652905	phosphonate monoester hydrolase
*Bc*PMH	*Burkholderia caryophili* PG2982	AAC44467	phosphonate monoester hydrolase[18, 53]
*Rl*PMH	*Rhizobium leguminosarum bv. viciae* 3841	CAK03956	putative sulfatase[c]
*Sp*PMH	*Silicibacter pomeroyi* DSS-3	AAV97522	phosphonate monoester hydrolase, putative

[a] The Protein database at NCBI (as of March 2018). [b] This protein was originally identified under the now defunct accession number ZP_09480898. A tblastn search on the revised *A. kashmirensis* WT001 genome data, using the sequence now identified with accession number AJA37533, suggested the deletion of a single base that would have originally been located between position 3846570 and 3846569 bp (gene coded in reverse complement DNA strand), causing the open reading frame to be no longer identifiable. Our experiments show that the original data for the defunct accession number ZP_09480898 were correct. We therefore submitted the correct coding sequence to GenBank (http://www.ncbi.nlm.nih.gov/genbank) under accession number KM597480. [c] Experimentally shown to be a phosphonate monoester hydrolase/phosphodiesterase by Jonas *et al.* [32]

TABLE OF CONTENTS GRAPHIC


Phosphonate Monoester Hydrolases
Arylsulfatases
$\log[k_{cat}/K_M]$
adaptation
exaptation