Determinants and prediction of esterase substrate promiscuity patterns

Martínez-Martínez, Mónica ; Coscolín, Cristina; Santiago, Gerard; Chow, Jennifer; Stogios, Peter J.; Bargiela, Rafael ; Gertler, Christoph; Navarro-Fernández, J.; Bollinger, Alexander; Thies, Stephanie; Méndez-García, Celia; Popovic, Anna; Brown, Greg; Chernikova, Tatyana; García-Moyano, Antonio; Bjerga, Gro E.K.; Perez-Garcia, Pablo; Hai, Tran; del Pozo, Mercedes V.; Stokke, Runar; Steen, Ida H.; Cui, Hong; Xu, Xiaohui; Nocek, Boguslaw; Alcaide, Maria; Disasto, Marco; Mesa, Victoria; Pelaez, Ana I.; Sanchez, Jesus; Buchholz, Patrick C.F.; Pleiss, Jurgen; Fernández-Guerra, Antonio; Glockner, Frank O.; Golyshina, Olga; Yakimov, Michail M.; Savchenko, Alexei; Jaeger, Karl-Erich; Yakunin, A. F.; Streit, Wolfgang R.; Golyshin, Peter; Guallar, Victor; Ferrer, Manuel

ACS Chemical Biology

DOI:
10.1021/acschembio.7b00996

Published: 01/01/2018

Peer reviewed version

Determinants and prediction of esterase substrate promiscuity patterns

Mónica Martínez-Martínez[†,○], Cristina Coscolín[†,○], Gerard Santiago[‡,○], Jennifer Chow[§], Peter J. Stogios[‖], Rafael Bargiela[†,▽], Christoph Gertler[⊥,Δ], José Navarro-Fernández[†], Alexander Bollinger[#], Stephan Thies[#], Celia Méndez-García[⊤,▲], Ana Popovic[‖], Greg Brown[‖], Tatyana N. Chernikova[⊥], Antonio García-Moyano[¥], Gro E.K. Bjerga[¥], Pablo Pérez-García[§], Tran Hai[⊥], Mercedes V. Del Pozo[†], Runar Stokke[Ψ], Ida H. Steen[Ψ], Hong Cui[‖], Xiaohui Xu[‖], Boguslaw P. Nocek[ξ], María Alcaide[†], Marco Distaso[⊥], Victoria Mesa[⊤], Ana I. Peláez[⊤], Jesús Sánchez[⊤], Patrick C. F. Buchholz[ϕ], Jürgen Pleiss[ϕ], Antonio Fernández-Guerra[¶,J,ʃ], Frank O. Glöckner[¶,J], Olga V. Golyshina[⊥], Michail M. Yakimov[♯,₸], Alexei Savchenko[‖], Karl-Erich Jaeger[#,ƨ], Alexander F. Yakunin[‖,*], Wolfgang R. Streit[§,*], Peter N. Golyshin[⊥,*], Víctor Guallar[‡,ɾ,*], Manuel Ferrer[†,*]. The INMARE Consortium

[†]Institute of Catalysis, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain

[‡]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain

[§]Biozentrum Klein Flottbek, Mikrobiologie & Biotechnologie, Universität Hamburg, 22609 Hamburg, Germany

[‖]Department of Chemical Engineering and Applied Chemistry, University of Toronto, M5S 3E5 Toronto, ON, Canada

[⊥]School of Biological Sciences, Bangor University, LL57 2UW Bangor, UK

[#]Institut für Molekulare Enzymtechnologie, Heinrich-Heine-Universität Düsseldorf, 52425 Jülich, Germany

[⊤]Department of Functional Biology-IUBA, Universidad de Oviedo, 33006 Oviedo, Spain

[¥]Uni Research AS, Center for Applied Biotechnology, 5006 Bergen, Norway

[Ψ]Department of Biology and KG Jebsen Centre for Deep Sea Research, University of Bergen, 5020 Bergen, Norway

[ξ]Structural Biology Center, Biosciences Division, Argonne National Laboratory, Argonne, 60439 Illinois, USA

[ϕ]Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, 70569 Stuttgart, Germany

[¶]Jacobs University Bremen gGmbH, Bremen, Germany

[J]Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

[ʃ]University of Oxford, Oxford e-Research Centre, Oxford, United Kingdom

[♯]Institute for Coastal Marine Environment, Consiglio Nazionale delle Ricerche, 98122 Messina, Italy

[₸]Immanuel Kant Baltic Federal University, 236041Kaliningrad, Russia

[ƨ]Institute for Bio- and Geosciences IBG-1: Biotechnology, Forschunsgzentrum Jülich GmbH, 52425 Jülich, Germany

[ɾ]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

Keywords: biodiversity | esterase | lipase | metagenomics | phosphatase | substrate promiscuity

ABSTRACT: Esterases receive special attention because their wide distribution in biological systems and environments and their importance for physiology and chemical synthesis. The prediction of esterases substrate promiscuity level from sequence data and the molecular reasons why certain such enzymes are more promiscuous than others, remain to be elucidated. This limits the surveillance of the sequence space for esterases potentially leading to new versatile biocatalysts and new insights into their role in cellular function. Here we performed an extensive analysis of the substrate spectra of 145 phylogenetically and environmentally diverse microbial esterases, when tested with 96 diverse esters. We determined the primary factors shaping their substrate range by analyzing substrate range patterns in combination with structural analysis and protein-ligand simulations. We found a structural parameter that helps ranking (classifying) promiscuity level of esterases from sequence data at 94% accuracy. This parameter, the active site effective volume, exemplifies the topology of the catalytic environment by measuring the active site cavity volume corrected by the relative solvent accessible surface area (SASA) of the catalytic triad. Sequences encoding esterases with active site effective volumes (cavity volume/SASA) above a threshold show greater substrate spectra, which can be further extended in combination with phylogenetic data. This measure provides also a valuable tool for interrogating substrates capable of being converted. This measure, found to be transferred to phosphatases of the haloalkanoic acid dehalogenase superfamily and possibly other enzymatic systems, represents a powerful tool for low-cost bioprospecting for esterases with broad substrate ranges, in large scale sequence datasets.

Enzymes with outstanding properties in biological systems and the conditions favoring their positive selection are difficult to predict. One of these properties is substrate promiscuity, which typically refers to a broad substrate spectrum and acceptance of larger substrates. This phenomenon is important from environmental,[1] evolutionary,[2-5] structural,[6-8] and biotechnological[9,10] points of view. The relevance of substrate promiscuity is indisputable as the operating basis for biological processes and cell function. As example, the evolutionary progress of enzymes from lower to higher substrate specificity allows the recruitment of alternate pathways for carbon cycling and innovations across metabolic sub-systems and the tree of life by maximizing the growth rate and growth efficiency.[11] Promiscuous enzymes are energetically more favorable than specialized enzymes,[4] and therefore, the cell does not require many different enzymes to take up substrates, favoring genome minimization and streamlining.[12] In addition, the acquisition of new specificities without compromising primary or ancestral ones is a major driver of microbial adaption to extreme habitats.[13] From a more practical standpoint, along with requirements of a technical nature such as selectivity, scalability and robustness, a narrow substrate spectrum is one of the most frequent problems for industrial enzyme applications.[14] A consensus exists that “the more substrates an enzyme converts the better”, opening application ranges with consequent reduction of the production cost of multiple enzymes.[10,14,15]

Enzymes with wide substrate ranges occur naturally, as systematically investigated for halo-alkane dehalogenases,[16] phosphatases,[1] beta-lactamases[2,17] and hydroxyl-nitrile lyases.[5] Some enzymes are more

promiscuous than others simply due to their fold or degree of plasticity or the presence of structural elements or mutations occurring under selection in the proximity of the active-site cavity and access tunnels favoring promiscuity. However, the general explanation, if any, by which an enzyme binds and converts multiple substrates are unknown, although molecular insights have been reported for single enzymes.[18] A tool that can clearly distinguish promiscuous vs non-promiscuous enzymes and suggest substrates potentially being converted or not by them, might therefore be valuable to apply low-cost sequencing in discovery platforms in any biological context.

In an ideal scenario, functional characterization of enzymes with genomics[19] and metagenomics[10,20] techniques using a large library of substrates would guide the analysis of sequence-to-promiscuity relationships and explore the mechanistic basis of promiscuity. In addition, such studies may help identify a new generation of highly promiscuous microbial biocatalysts. However, extensive bioprospecting and biochemical studies are rare,[10] despite the growing number of sequences available through low-cost sequencing efforts[21] and the growing number of enzymes that are typically characterized with limited substrate sets.[14] To address this knowledge gap, we functionally assessed the substrate specificity of a set of 145 phylogenetically, environmentally and structurally diverse microbial esterases (herein referred to as 'EH', which means Ester Hydrolase) against a customized library of 96 different substrates to find predictive markers of substrate promiscuity rather than discrete determinants of substrate specificity that may differ from protein to protein. EHs were selected for an analysis of substrate promiscuity because they typically have specific definitions of molecular function, can be easily screened in genomes and metagenomes compared with many other classes of proteins, are among the most important groups of biocatalysts for chemical synthesis and are widely distributed in nature, with at least one EH per genome.[14]

Our work adds important insights and empirical, structural and computational data to facilitate the elucidation of the molecular basis of substrate promiscuity in EHs, which was further extended to phosphatases from the haloalkanoic acid dehalogenase (HAD) superfamily. This was achieved by deciphering what we consider a predictive structural marker of substrate promiscuity and by stablishing the reasons why certain such enzymes are more promiscuous than others and can convert substrates that others cannot. This study does not pretend to generate a quantitative measure to predict the number of compounds that an enzyme will hydrolyze, but a tool and a parameter that will help in ranking (classifying) promiscuity level. Following on from that, we propose in this work the first molecular classification method of this kind derived from first principle molecular simulations and with clear physical/structural interpretation. This work also provides an example of the utility of this parameter to screen the sequence space for highly promiscuous EHs that may compete with best commercial EH preparations. We also provide first preliminary evidences of a number of underexplored microbial phylogenetic lineages containing EHs with prominent substrate range.

RESULTS AND DISCUSSION

The substrate range of 145 diverse EHs. A total of 145 EHs were investigated. Extensive details of the sources and screen methods are provided in Supporting Information, Methods and Table S1. In an environmental context, the source of enzymes was highly diverse because they were isolated from bacteria from 28 geographically distinct sites (125 EHs in total) and from 6 marine bacterial genomes (20 EHs) (Supporting Information, Fig. S1). A phylogenetic analysis also indicated that sequences belong to bacteria distributed across the entire phylogenetic tree (Supporting Information, Results and Fig. S2).

The 145 putative proteins exhibited maximum amino acid sequence identities (Supporting Information, Table S1) ranging from 29.1 to 99.9% to uncharacterized homologous proteins in public databases, with an average value (reported as %, with the interquartile range (IQR) in parentheses) of 74.3% (40.3%). The pairwise amino acid sequence identity for all EHs ranged from 0.2 to 99.7% (Supporting Information, Table S2), with an average value of 13.7% (7.6%). BLAST searches were performed for all query sequences by running NCBI BLASTP against the current version of the Lipase Engineering Database[22] using an E-value threshold of 10^{-10} and were successful for all but 9 candidates. A total of 120 EH sequences were unambiguously assigned to some of the 14 existing families (F) of the Arpigny and Jaeger classification, which are defined based on amino acid sequence similarity and the presence of specific sequence motifs.[14,23] These EHs included sequences with a typical α/β hydrolase fold and conserved G-X-S-X-G (FI: 20, FIV: 36, FV: 33, FVI: 5, and FVII: 6) or G-X-S-(L) (FII: 9) motifs and sequences with a serine beta-lactamase-like modular (non α/β hydrolase fold) architecture and a conserved S-X-X-K motif (FVIII: 11). An additional set of 9 sequences were assigned to the *meta*-cleavage product (MCP) hydrolase family[24] and 6 to the so-called carbohydrate esterase family,[25] both with typical α/β hydrolase folds. Finally, one was a cyclase-like protein from the amido-hydrolase superfamily.[26] Sequences-to-family assignments are summarized in Supporting Information, Table S1. Taken together, the primary sequence analysis suggests that the diversity of polypeptides is not dominated by a particular type of protein or highly similar protein clusters but consists of diverse non-redundant sequences assigned to multiple folds and sub-families, which are distantly related to known homologs in many cases.

The substrate profiles of all EHs were examined using a set of 96 chemically and structurally distinct esters (Supporting Information, Table S3). We are aware that the number of compounds hydrolyzed may be an ambiguous indicator of promiscuity, because the size and composition of the library may influence the results. For this reason, the composition of the library was not random but based on including esters with variation in size of acyl and alcohol groups and with growing residues (aromatic, aliphatic, branched and unbranched) at both sides leading to more challenging substrates because a larger group adjacent to the ester bond increases the difficulty of conversion. Halogenated, chiral and sugar esters, lactones and an alkyl di-ester, were also included. Esters with nitro substituents were not included. We used the partitioning coefficient (log P value) to indicate the chemical variability of the esters because this parameter reflects electronic and steric effects and hydrophobic and hydrophilic characteristics. Log P was determined with the software ACD/ChemSketch 2015.2.5. Log P values (Supporting Information, Table S3) ranged from -1.07 (for methyl glycolate) to 23.71 (for triolein), with an average value (IQR in parentheses) of 3.13 (2.86),

which indicates that the ester library used in this study had broad chemical and structural variability. Nevertheless adding new substrates could surely help (and even change) the ranking of the EHs herein analyzed. The dynamic range of the assay may also influence the results. For this reason, to detect enzyme-substrate pairs for a given EH, the ester library was screened with each of the 145 EHs in a kinetic pH indicator assay in 384-well plates[24,27,28] which unambiguously allow quantifying specific activities at pH 8.0 and 30°C, using a substrate concentration above 0.5 mM (see Supporting Information, Results). Two commercial lipases, CalA and CalB from *Pseudozyma aphidis* (formerly *Candida antarctica*), were included in the assays for comparison. Using this dataset, we linked the biocatalytic data to the sequence information for the respective enzyme. In this study, sequence information meant any sequence that encoded an EH of interest. Biocatalytic data meant experimental data on substrate conversion (i.e., units g^{-1} or U g^{-1}) followed for 24 h.

We determined the probability of finding an EH with a broad substrate profile by plotting the number of esters that were hydrolyzed by all preparations. Fig. 1 shows that the number of esters hydrolyzed by all 147 EHs (including CalA/B) fits to an exponential distribution ($r^2 = 0.99$; *p*-value $3.2e^{-38}$; Pearson's correlation coefficient) with a median of 18 substrates per enzyme, 9 hits at the 25th percentile and 29 hits at the 75th percentile. Based on this distribution and a previously established criterion,[1] we considered an enzyme specific if it used 9 esters or fewer (27% of the total), as showing moderate substrate promiscuity if it used between 10 and 29 esters (51% of the total), and as showing high-to-prominent promiscuity if it used 30 or more esters (22% of the total). This criterion indicated a percentage of EHs with a prominent substrate range similar to that found for HAD phosphatases (24%).[1]

Phylogeny is a predictive marker of substrate promiscuity. Hierarchical clustering was performed to evaluate the differences in substrate range patterns (Fig. 2). For the sake of simplicity, clustering was performed for those EHs that hydrolyzed 10 or more esters (i.e., 107 total EHs). We first observed a large percentage of enzymes with presumptive broad active site environments that accommodated large aromatic and sterically hindered esters such as benzyl (*R*)-(+)-2-hydroxy-3-phenylpropionate (49% of the total), benzoic acid-4-formyl-phenylmethyl ester (27%), 2,4-dichlorophenyl 2,4-dichlorobenzoate (~8%), 2,4-dichlorophenyl 2,4-dichlorobenzoate (~5%) and diethyl-2,6-dimethyl 4-phenyl-1,4-dihydro pyridine-3,5-dicarboxylate (~1%). Therefore, even though the EHs in this study were identified by a selection process based on the utilization of short esters (see Supporting Information, Methods), the isolation of EHs with ample substrate spectra and the ability to hydrolyze very large substrates was not compromised.

We detected drastic shifts in substrate specificity (Fig. 2), with glyceryl tri-propionate as the only substrate hydrolyzed by all EHs. This is consistent with the high sequence variability within EHs, with an average pairwise identity of 13.74%. We then sought to determine the primary factors shaping the substrate range and thus defined different functional clusters. First, we observed that global sequence identity was of limited relevance for inferring the substrate range because no correlation was found ($r^2 = 0.25$) between the differences in identity and the number of esters that were hydrolyzed (Supporting Information, Tables S1 and S2). Second, comparisons of the substrate range and the hydrolysis rate (U g^{-1} for the best substrates)

were performed (Supporting Information, Table S1). No correlation existed ($r^2 = 0.073$), suggesting that our assay conditions allow evaluating promiscuity level whatever is the hydrolytic rate of the EH. In addition to the low correlation values, no threshold above or below which one could qualitatively classify substrate range was observed in both cases, so that sequence identity and hydrolytic rate are neither predictive nor classification parameters of promiscuity. Additionally no link between substrate range and habitat was found because EHs from the same bio-source fell into separate clusters (Fig. 2). Phylogeny-substrate spectrum relationships were further examined. Fig. 2 indicates that the broad substrate-spectrum EHs did not cluster in a single phylogenetic branch, yet substrate promiscuity was mostly found for members of one of 10 sub-families covered. Indeed, 67% of the EHs that could hydrolyze 30 or more esters (mostly located in Clusters C1 and C2 in Fig. 2) were assigned to FIV,[14,23] and this percentage increased to 84% when considering only those EHs that could hydrolyze 42 to 72 esters (Fig. 2; Cluster C1). In addition to FIV members, a FVIII serine beta-lactamase showed prominent substrate spectra (see Cluster 1). Members of both families (FIV: 8; FVIII: 1; see Cluster C1) hydrolyzed as many esters (from 61 to 72) as the yeast family member CalB (68 esters), the most promiscuous commercially available lipase preparation used for the production of fine chemicals.[29]

Phylogeny was thus indicated as a predictive marker of the substrate range of EHs, as although a broad substrate scope was assigned to several sequence clusters, this feature was prevalent in members of FIV. A query sequence that matched FIV could be easily identified by means of the consensus motif GDSAGG around the catalytic serine; this family is also called the hormone-sensitive lipase (HSL) family because a number of FIV EHs display a striking similarity to the mammalian HSL.[14,23] Noticeably, the location of some FIV members in functional clusters with narrow substrate spectra (Fig. 2) suggests that factors other than phylogeny contribute to the substrate spectra of EHs.

The active site effective volume is a prominent marker of EH promiscuity. Structural-to-substrate spectrum relationships were further examined by protein-ligand simulations to find additional markers of promiscuity. Crystals from recombinant EH1,[28] the protein with the broadest substrate range under our assay conditions, were obtained as described in Supporting Information, Methods. The enzyme with the widest substrate range was considered the best candidate for understanding the nature of promiscuity. This enzyme seems to have a wide active site environment as, under our assay conditions, it accepted 72 esters ranging from short (e.g., vinyl acetate) to large (e.g., 2,4-dichlorobenzyl-2,4-dichlorobenzoate) (Fig. 2). We also obtained crystals of recombinant EH102, which was isolated from the same habitat[28] but had a restricted substrate range, hydrolyzing only 10 of the 96 esters tested (Fig. 2). Crystallographic data and refinement statistics for the two structures are given in Supporting Information, Table S4.

To rationalize the substrate range shown by EH1 and EH102, we performed substrate migration studies using the software Protein Energy Landscape Exploration (PELE), which is an excellent tool to map ligand migration and binding, as shown in studies with diverse applications.[30-32] To map the tendency of a substrate to remain close to the catalytic triad, the substrate was placed in a catalytic position, within a proton abstraction distance from the catalytic serine, and allowed to freely explore the exit from the active site. The

PELE results for both proteins and glyceryl tri-acetate are shown in Fig. 3a. Clearly, EH1 has a significantly better binding profile, with an overall lower binding-energy and a better funnel shape, whereas EH102 had a qualitatively unproductive binding-energy profile. This difference in the binding mechanism can be explained by the catalytic triad environment. EH1 has a somewhat wide but buried active site, whereas EH102 has a surface-exposed catalytic triad (Fig. 4a). These structural differences translate into significant changes in the active site volume, as defined using Fpocket; the active site cavity of EH1 is 3-fold larger than that of EH102. Moreover, important changes are observed when inspecting the solvent exposure of the cavity. Fig. 3b shows the relative solvent accessible surface area (SASA) for the substrate along the exploration of PELE, computed as a (dimensionless) percentage (0-1) of the ligand SASA in solution. Even at catalytic positions (distance Ser(O)-substrate(C) ~3-4 Å), in EH102 we observe that ~40% of the surface of the substrate is accessible to the solvent, which greatly destabilizes the substrate and facilitates escape to the bulk solvent. By contrast, EH1 has a larger but almost fully occluded site, with relative SASA values of approximately 0-10%, which can better stabilize the substrate.

After defining key points underlying the promiscuity of EH1, i.e., a larger active site volume and a lower SASA (Fig. 4a), we extended the analysis to other EHs. First, we collected all 11 available crystal structures (Supporting Information, Table S1) and computed the active site volume and relative SASA of the catalytic triad (Fig. 5, square symbols). We next extended the analysis to the rest of the EHs using homology modeling (using the 11 crystals available) and produced a structural model for 84 additional enzymes. The missing ones were those with sequence identities of less than 25% (to an existing crystal) or those for which the catalytic triad could not be unambiguously identified (i.e., not suitable alignments). Fig. 5 (circle symbols) shows the active site effective volume data for all structural models. The analysis indicated a ratio threshold of 62.5 $Å^3$ for qualitatively classifying substrate promiscuity. Note that the relative SASA of the catalytic triad (derived from the GetArea server, see Supporting Information, Methods) adopts values of 0-100; the actual value of the effective volume threshold will depend on the chosen range. We observed that values equal to or higher than 62.5 $Å^3$ corresponded to EHs with activity for 20 or more of the 96 substrates tested and opposite. There were only 6 outliers out of 95 EHs that did not follow this rule. Thus performance is excellent (with 94%) of accuracy if used as a classifier. The effective volume, however does not have quantitative predictions for the exact number of esters hydrolyzed (r^2=0.16 for data in Fig. 5), most likely because above the 62.5 $Å^3$-threshold, the capability to hydrolyze more or less substrates may specifically depend on the topology of the catalytic environment (Fig. 4a-c), which may differ within families. Particularly, none of the different family members that conformed to the ≥62.5 $Å^3$-threshold except those from FIV (i.e., at least 50% of its members as shown in Fig. 5, grey circle symbols) and CalB, could hydrolyze 42 or more esters. Therefore, the classification potential of the effective volume measure increased when combined with phylogenetic data. Noticeably, we observed that the predictive capacity of cavity volume/SASA is not influenced by the presence of flexible elements in the structure (Supporting Information, Results).

The active site effective volume is also indicative of molecules being accepted as substrates. We further used the active site cavity volume/SASA to also dissect its role in substrate specificity. We restricted the analysis to the 96 EHs for which this value could be unambiguously calculated (see above). The analysis indicated that the conversion of 34 esters was only observed for EHs conforming to the ≥62.5 $Å^3$-threshold (Supporting Information, Fig. S3). All but two (vinyl crotonate and ethyl acetate) could be considered large alkyl or hindered aromatic esters, and included important molecules in synthetic organic chemistry such paraben esters. This suggests that active sites with larger volume and a lower SASA (i.e. cavity less exposed to the surface) will most likely support hydrolysis of these esters. Therefore, the effective volume measure could be used to some extent as an indicator of substrates that may or may not be hydrolyzed by EHs. However, not all EHs fitting the ≥62.5 $Å^3$-threshold could convert all these 34 esters, implying that this measure does not allow deepening into substrate specificity, which may depend on the topology of the catalytic environments as mentioned previously (Fig. 4a-c). However, we found that the probability that benzyl-, butyl- and propyl-paraben esters, major intermediates in chemical synthesis, are converted by members of the FIV with an effective volume ≥62.5 $Å^3$ is significantly higher (~35%) than that of EHs from FIV but < 62.5 $Å^3$ and EHs from other families whatever the value of the effective volume (approaching zero percent in our study); for those EHs for which effective volume could not be calculated this probability is as low as 1.9% (Supporting Information, Fig. S4). This again exemplifies that the effective volume measure, when combined with phylogenetic information, is not only indicative of a promiscuity level but also can be used to predict the capacity to hydrolyze esters such as paraben esters. Screen programs to find EHs capable of converting paraben esters should most likely be directed to find those assigned to FIV and with cavity volume/SASA ≥62.5 $Å^3$.

The effective volume is also a marker of substrate promiscuity in proteins others than EHs. In order to evaluate the possibility that the active site effective volume may be a marker of substrate promiscuity in other enzymes, substrate spectra-effective volume relationships should be investigated in other protein families. In this line, Huang *et al.*[1] recently performed a systematic analysis of the substrate spectra of 200 phosphatases of the HAD superfamily, when tested against a set of 167 substrates. We collected the available crystal structures of each of the HAD phosphatases (Supporting Information, Table S5) and computed the active site effective volume. We restricted the analysis to C2 cap members as they were reported to have a broader substrate spectra[1] and crystal structures with low to high effective volume are available. Interestingly, we observed that the effective volume (using the two conserved aspartic catalytic residues as the corrective SASA factor) was highly correlated (r^2=0.92) with the substrate range (Fig. 6). Thus, the effective volume can be used as a molecular classification parameter of substrate promiscuity of phosphatases of the HAD superfamily when crystal structures are available. When this analysis was extended to the rest of the enzymes using homology modeling, we observed a similar trend to that of EHs (Supporting Information, Fig. S5). That is, no correlation existed (r^2 = 0.043) but still the effective volume can be used as a classifier of the substrate range as for EHs. Indeed, although a threshold could not be

unambiguously established, sequences with the top 10 effective volumes belong to moderate-high to high promiscuity enzymes.

In conclusion, we found that the topology around the catalytic position, by meaning of an active site effective volume (cavity volume/SASA) threshold, is a dominant criterion of substrate promiscuity in EHs, which can be further extended by adding phylogenetic analysis. The rationale behind this parameter is as follows. Large volumes increase promiscuity until a certain value at which the cavity becomes too exposed and is not capable of properly accommodating and, importantly, retaining the substrate in specific catalytic binding interactions. This point is well captured by the SASA percentage of the catalytic triad, a dimensionless ratio that corrects for large volume measures in exposed sites. Importantly, the parameters of active site volume and relative SASA can be easily transferred to other systems. Indeed, the fact that the EHs investigated herein have different folds and that this parameter was also a marker of substrate spectra for phosphatases of the HAD superfamily, opens the possibility of applying the effective volume measure to other enzymes requiring substrate anchoring. In all cases, the effective volume threshold-to-substrate relationships must be established. We would like to notice that the active site volume is not a static property, as the active site will breathe, depending on how flexible the protein is. In addition to that, the 62.5 $Å^3$-threshold for qualitatively classifying substrate promiscuity is based on the analysis of 147 EHs when tested against 96-esters. Although, increasing the number of EHs and esters may influence this threshold and increase accuracy, it will not affect the fact that the measurement of the effective volume (cavity volume/SASA) can be used as the first molecular classification method of substrate promiscuity in EHs.

Our measurement is not a quantitative one, but rather a qualitative ranking (classification) procedure that will allow, for example, selecting sequences in databases for expression, particularly, those encoding promiscuous enzymes capable of converting multiple substrates. This will substantially reduce reagent and labor costs compared to methods requiring the extensive cloning of all genes, and the expression and characterization of all enzymes in databases to later find those being promiscuous.[33] This possibility was herein examined by successfully mapping the open reading frames from the TARA Oceans project assemblies,[34] and by identifying a high number of sequences encoding EHs with presumptive prominent substrate promiscuity (Supporting Information, Results, Fig. S6, Fig. S7). Application of the effective volume measure to examine the sequences daily generated or deposited in databases requires having some crystals or X-ray structures for the model production. This limitation prevents predicting promiscuity from sequences lacking any structural information. Indeed 36% of the EHs in this study (52 of the 147, including CalA/B) could not be included in the correlation because no calculation was possible. Accumulation of structural information and design and application of better modelling algorithms in the future will help solving this limitation.[35] Future studies might also explore molecular dynamics (MD) simulations to measure also the flexibility of the active site and not just the size of the cavity. By using this strategy it was recently reported that the broad promiscuity of the members of the alkaline phosphatase superfamily arises from cooperative electrostatic interactions in the active site, allowing each enzyme to adapt to the electrostatic needs of different substrates.[36] In the particular case of EHs phylogeny, a marker which does not require a

three dimensional structure, was also suggested as a predictive classification marker of the substrate range. Indeed, this study suggests that in case of an unknown EH for which a crystal structure is not available or a homology model could not be established, then its assignation to Family IV[14,23] increases the likelihood that this EH is promiscuous.

The present study not only provides clear evidence that substrate promiscuity in EHs has evolved from different core structural domains fitting an effective volume around the active site, albeit with a bias toward that occurring in FIV members, but also from different phylogenetic lineages, many of which remain underexplored to date (Supporting Information, Results and Fig. S2). These are new findings as it was previously thought that the substrate range in a superfamily increased from a single ancestral core domain,[1] and because the identities of some microbial groups containing promiscuous enzymes, herein EHs, were previously unknown. Finally, this study also enabled the selection of a set of EH candidates that can compete with best commercial EHs such as CalB, as they show a broader substrate profile and specific activities up to 3-fold higher (Supporting Information, Table S6). Their sequences can be used to search databases for similar promiscuous EHs. Further investigations should also determine the occurrence of other types of promiscuous EH phenotypes with broader substrate ranges than those identified in this study. For example, at least the stability of substrate-promiscuous EHs at different temperatures and with various solvents, along with the occurrence and evolution of secondary reactions, should be investigated in terms of condition and catalytic promiscuity.

METHODS

Protein samples. Two main sources of EHs were used in the present study, all of them isolated via naïve and sequence-based screens in genomes and metagenomes. A first set of samples were EHs previously reported in the bibliography (69 in total) and that were herein substrate-profiled for first time. A second set were EHs (77) that are herein reported for first time. The extensive details of the source, cloning, expression and purification of each of the active and soluble EHs are provided in Supporting Information, Methods and Table S1.

Ester bond hydrolysis activity assessment: substrate profiling tests with 96 esters. Hydrolytic activity was assayed at 550 nm using 96 structurally diverse esters in 384-well plates as previously described.[24,27,28] Before the assay, a concentrated stock solution of the esters was prepared at a concentration of 100 mg/mL in acetonitrile and dimethyl sulfoxide (DMSO). The assays were conducted according to the following steps. First, a 384-well plate (Molecular Devices, LLC, CA, USA) was filled with 20 μL of 5 mM *N*-(2-hydroxyethyl)piperazine-*N*'-(3-propanesulfonic acid (EPPS) buffer, pH 8.0, using a QFill3 microplate filler (Molecular Devices, LLC, CA, USA). Second, 2 μL of each ester stock solution was added to each well using a PRIMADIAG liquid-handling robot (EYOWN TECHNOLOGIES SL, Madrid, Spain). The ester was dispensed in replicates. After adding the esters, the 384-well plate was filled with 20 μL of 5 mM EPPS buffer, pH 8.0, containing 0.912 mM Phenol Red (used as a pH indicator) using a QFill3 microplate filler. The final ester concentration of the ester in each well was 1.14 mg/mL, and the final concentration of

Phenol Red was 0.45 mM. A total of 2 μL of protein extract (containing 1-5 mg/mL pure protein or 200 mg/mL wet cells expressing proteins) was immediately added to each well using an Eppendorf Repeater M4 pipette (Eppendorf, Hamburg, Germany) or a PRIMADIAG liquid-handling robot. Accordingly, the total reaction volume was 44 μL, with 4.5% (*v*/*v*) acetonitrile or DMSO in the reaction mixture. After incubation at 30 °C in a Synergy HT Multi-Mode Microplate Reader, ester hydrolysis was measured spectrophotometrically in continuous mode at 550 nm for a total time of 24 h. Commercially available CALA L and CALB L (Novozymes A/S, Bagsvaerd, Denmark) were diluted tenfold with 5 mM EPPS buffer, pH 8.0, and 2 μL of this solution was used immediately for reaction tests under the conditions described before. In all cases, specific activities (in U g^{-1} protein) were determined. One unit (U) of enzyme activity was defined as the amount of wet cells expressing EHs or pure EHs required to transform 1 μmol of substrate in 1 min under the assay conditions using the reported extinction coefficient ($\varepsilon_{\text{Phenol red}}$ at 550 nm = 8,450 M^{-1} cm^{-1}). All values were corrected for non-enzymatic transformation (i.e., the background rate) and for the background signal using *E. coli* cells that did not express any target protein (control cells included empty vectors). Note that a positive reaction was indicated by the restrictive criterion of a change greater than 6-fold above the background signal. Specific activity determinations (in U g^{-1}) for wet cells expressing each of the selected EHs or pure or commercial proteins are available in Supporting Information, Table S3 and Table S6, respectively.

Structural determinations and homology modeling. The proteins EH1 and EH102 were expressed, purified and crystallized using the sitting-drop method in Intelliplate 96-well plates and a Mosquito liquid-handling robot (TTP LabTech) according to previously described procedures.[37] For EHs for which crystal structures were not available, homology models were developed using Prime software from Schrödinger. Prime uses BLAST (with BLOSUM62 matrix) for homology search and alignment and refines the results using the Pfam database and pairwise alignment with ClustalW.

Protein Energy Landscape Exploration (PELE) simulations. We used Protein Energy Landscape Exploration (PELE) software to sample the binding modes of glyceryl tri-acetate with EH1 and EH102.[38,39] The initial structures were taken from the coordinates of the EH1 and EH102 crystal structures (PDB codes: 5JD4 and 5JD3, respectively). The protonation state of titratable residues was estimated with the Protein Preparation Wizard (PROPKA)[40] and the H++ server (http://biophysics.cs.vt.edu/H++) followed by visible inspection. At pH 8 (the pH at which the activity assays were performed), the catalytic triad histidine residues were δ-protonated, and the catalytic triad aspartic acid residues were deprotonated, resulting in the formation of a histidine-serine and histidine-aspartic hydrogen-bonding network. The glyceryl acetate structure was fully optimized with Jaguar[41] in an implicit solvent, and the electrostatic potential charges were computed with the density functional M06 at the 6-31G* level of theory. The ligand parameters were extracted from these for the classic simulations.

Cavity Volume and Solvent Accessible Surface Area (SASA) calculation. The relative Solvent Accessible Surface Area (SASA) for a residue was obtained using the GetArea web server.[42] Cavity volumes

were computed with Fpocket,[43] a very fast open-source protein pocket (cavity) detection algorithm based on Voronoi tessellation. Fpocket includes two other programs (dpocket and tpocket) that allow the extraction of pocket descriptors and the testing of owned scoring functions, respectively.

For the extensive details of the Methods, see Supporting Information, Methods.

AUTHOR INFORMATION

Corresponding Author

*(V.G.) E-mail: victor.guallar@bsc.es.

*(M.F.) E-mail: mferrer@icp.csic.es.

Present Addresses

[▽]Current address School of Chemistry, Bangor University, LL57 2UW Bangor, UK.

[Δ]Current address Lehrstuhl für Biotechnologie, RWTH Aachen University, Aachen, Germany.

[▲]Current address Carl R. Woese Institute for Genomic Biology, Urbana, USA.

Author Contributions

[○]These authors contributed equally to this work.

ORCID

Manuel Ferrer: 0000-0003-4962-4714.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

C. Coscolín thanks the Spanish Ministry of Economy, Industry and Competitiveness for a PhD fellowship (Grant BES-2015-073829). V. Mesa thanks the Francisco José de Caldas Scholarship Program (Administrative Department of Science, Technology and Innovation, COLCIENCIAS). The authors acknowledge the members of the MAMBA, MAGICPAH, ULIXES, KILLSPILL and INMARE Consortia for their support in sample collection. David Rojo is also acknowledged for his valuable help with log P calculations.

This project received funding from the European Union's Horizon 2020 research and innovation program [Blue Growth: Unlocking the potential of Seas and Oceans] under grant agreement no. [634486] (project acronym INMARE). This research was also supported by the European Community Projects MAGICPAH (FP7-KBBE-2009-245226), ULIXES (FP7-KBBE-2010-266473) and KILLSPILL (FP7-KBBE-2012-312139) and grants BIO2011-25012, PCIN-2014-107, BIO2014-54494-R and CTQ2016-79138-R from the Spanish Ministry of Economy, Industry and Competitiveness. The present investigation

was also funded by the Spanish Ministry of Economy, Industry and Competitiveness within the ERA NET IB2, grant no. ERA-IB-14-030 (MetaCat), the UK Biotechnology and Biological Sciences Research Council (BBSRC), grant nr. BB/M029085/1, and the German Research Foundation (FOR1296). R.B. and P.N.G. acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government. O.V.G. and P.N.G. acknowledge the support of the Centre of Environmental Biotechnology Project funded by the European Regional Development Fund (ERDF) through Welsh Government. A.Y. and A.S. gratefully acknowledge funding from Genome Canada (2009-OGI-ABC-1405) and the NSERC Strategic Network grant IBN. A.I. Pelaez was supported by the Counseling of Economy and Employment of the Principality of Asturias, Spain (Grant FC-15-GRUPIN14-107). V.G. acknowledges the joint BSC-CRG-IRB Research Program in Computational Biology. The authors gratefully acknowledge financial support provided by the European Regional Development Fund (ERDF).

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

Supporting Results, Methods, Figures S1-S7, and Table S4 (PDF)

Supporting Tables S1-S3, S5 and S6 (Excel)

REFERENCES

(1) Huang, H., Pandya, C., Liu, C., Al-Obaidi, N. F., Wang, M., Zheng, L., Toews Keating, S., Aono, M., Love, J. D., Evans, B., Seidel, R. D., Hillerich, B.S., Garforth, S. J., Almo, S. C., Mariano, P. S., Dunaway-Mariano, D., Allen, K. N., and Farelli, J. D. (2015) Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc. Natl. Acad. Sci. USA 112,* E1974-1983.

(2) Huang, R., Hippauf, F., Rohrbeck, D., Haustein, M., Wenke, K., Feike, J., Sorrelle, N., Piechulla, B., and Barkman, T. J. (2012) Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proc. Natl. Acad. Sci. USA 109,* 2966-2971.

(3) Yip, S. H., and Matsumura, I. (2013) Substrate ambiguous enzymes within the *Escherichia coli* proteome offer different evolutionary solutions to the same problem. *Mol. Biol. Evol. 30,* 2001-2012.

(4) Price, D. R., and Wilson, A. C. (2014). Substrate ambiguous enzyme facilitates genome reduction in an intracellular symbiont. *BMC Biol. 12*, 110.

(5) Devamani, T., Rauwerdink, A. M., Lunzer, M., Jones, B. J., Mooney, J. L., Tan, M. A., Zhang, Z. J., Xu, J. H., Dean, A. M., and Kazlauskas, R. J. (2016). Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J. Am. Chem. Soc. 138*, 1046-1056.

(6) Hult, K., and Berglund, P. (2007). Enzyme promiscuity: mechanism and applications. *Trends Biotechnol 25*, 231-238.

476 (7) Copley, S. D. (2015). An evolutionary biochemist's perspective on promiscuity. *Trends Biochem. Sci. 40*, 72-78.

478 (8) London, N., Farelli, J. D., Brown, S. D., Liu, C., Huang, H., Korczynska, M., Al-Obaidi, N. F., Babbitt, P. C., Almo, S. C., Allen, K. N., and Shoichet, B. K. (2015) Covalent docking predicts substrates
480 for haloalkanoate dehalogenase superfamily phosphatases. *Biochemistry 54*, 528-537.

(9) Nobeli, I., Favia, A. D., and Thornton, J. M. (2009) Protein promiscuity and its implications for
482 biotechnology. *Nat. Biotechnol. 27*, 157-167.

(10) Ferrer, M., Martínez-Martínez, M., Bargiela, R., Streit, W. R., Golyshina, O. V., and Golyshin, P. N.
484 (2016) Estimating the success of enzyme bioprospecting through metagenomics: current status and future trends. *Microb. Biotechnol. 9*, 22-34.

486 (11) Braakman, R., and Smith, E. (2014) Metabolic evolution of a deep-branching hyperthermophilic chemoautotrophic bacterium. *PLoS One 9*, e87950.

488 (12) Giovannoni, S. J., Cameron Thrash, J., and Temperton, B. (2014) Implications of streamlining theory for microbial ecology. *ISME J. 8*, 1553-1565.

490 (13) Lan, T., Wang, X. R., and Zeng, Q. Y. (2013) Structural and functional evolution of positively selected sites in pine glutathione S-transferase enzyme family. *J. Biol. Chem. 288*, 24441-24451.

492 (14) Ferrer, M., Bargiela, R., Martínez-Martínez, M., Mir, J., Koch, R., Golyshina, O. V., and Golyshin, P. N. (2015) Biodiversity for biocatalysis: A review of the α/β-hydrolase fold superfamily of esterases-
494 lipases discovered in metagenomes. *Biocatal. Biotransform. 33*, 235-249.

(15) Schmid, A., Dordick, J. S., Hauer, B., Kiener, A., Wubbolts, M., and Witholt, B. (2001) Industrial
496 biocatalysis today and tomorrow. *Nature 409*, 258-268.

(16) Koudelakova, T., Chovancova, E., Brezovsky, J., Monincova, M., Fortova, A., Jarkovsky, J., and
498 Damborsky, J. (2011) Substrate specificity of haloalkane dehalogenases. *Biochem. J.* 435, 345-354.

(17) Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A., and Sanchez-Ruiz, J. M. (2013)
500 Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc. 135*, 2899-2902.

502 (18) Amin, S. R., Erdin, S., Ward, R. M., Lua, R. C., and Lichtarge, O. (2013) Prediction and experimental validation of enzyme substrate specificity in protein structures. *Proc. Natl. Acad. Sci. USA* 110,
504 E4195-4202.

(19) Anton, B. P., Chang, Y. C., Brown, P., Choi, H. P., Faller, L.L., Guleria, J., Hu, Z., Klitgord, N.,
506 Levy-Moonshine, A., Maksad, A., Mazumdar, V., McGettrick, M., Osmani, L., Pokrzywa, R., Rachlin, J., Swaminathan, R., Allen, B., Housman, G., Monahan, C., Rochussen, K., Tao, K., Bhagwat, A. S., Brenner,
508 S. E., Columbus, L., de Crécy-Lagard, V., Ferguson, D., Fomenkov, A., Gadda, G., Morgan, R. D., Osterman, A. L., Rodionov, D. A., Rodionova, I. A., Rudd, K. E., Söll, D., Spain, J., Xu, S. Y., Bateman, A.,
510 Blumenthal, R. M., Bollinger, J. M., Chang, W. S., Ferrer, M., Friedberg, I., Galperin, M. Y., Gobeill, J., Haft, D., Hunt, J., Karp, P., Klimke, W., Krebs, C., Macelis, D., Madupu, R., Martin, M. J., Miller, J. H.,
512 O'Donovan, C., Palsson, B., Ruch, P., Setterdahl, A., Sutton, G., Tate, J., Yakunin, A., Tchigvintsev, D.,

Plata, G., Hu, J., Greiner, R., Horn, D., Sjölander, K., Salzberg, S.L., Vitkup, D., Letovsky, S., Segrè, D.,
514 DeLisi, C., Roberts, R. J., Steffen, M., and Kasif, S. (2013) The COMBREX project: design, methodology, and initial results. *PLoS Biol. 11*, e1001638.

516 (20) Colin, P. Y., Kintses, B., Gielen, F., Miton, C. M., Fischer, G., Mohamed, M. F., Hyvönen, M., Morgavi, D. P., Janssen, D. B., and Hollfelder, F. (2015) Ultrahigh-throughput discovery of promiscuous
518 enzymes by picodroplet functional metagenomics. *Nat. Commun.* 6, 10008.

(21) Chen, C., Huang, H., and Wu, C. H. (2017) Protein bioinformatics databases and resources. *Methods*
520 *Mol. Biol. 1558*, 3-39.

(22) Fischer, M., and Pleiss, J. (2003) The Lipase Engineering Database: a navigation and analysis tool
522 for protein families. *Nucleic Acids Res. 31*, 319-321.

(23) Arpigny, J. L., and Jaeger, K. E. (1999) Bacterial lipolytic enzymes: classification and properties.
524 *Biochem. J. 343*, 177-183.

(24) Alcaide, M., Tornés, J., Stogios, P.J., Xu, X., Gertler, C., Di Leo, R., Bargiela, R., Lafraya, A.,
526 Guazzaroni, M. E., López-Cortés, N., Chernikova, T. N., Golyshina, O. V., Nechitaylo, T. Y ., Plumeier, I., Pieper, D. H., Yakimov, M. M., Savchenko, A., Golyshin, P. N., and Ferrer, M. (2013) Single residues
528 dictate the co-evolution of dual esterases: MCP hydrolases from the α/β hydrolase family. *Biochem. J. 454*,157-166.

530 (25) Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M., and Henrissat, B. (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochem. J.* 432, 437-444.

532 (26) Popovic, A., Hai, T., Tchigvintsev, A., Hajighasemi, M., Nocek, B., Khusnutdinova, A. N., Brown, G., Glinos, J., Flick, R., Skarina, T., Chernikova, T. N., Yim, V., Brüls, T., Paslier, D.L., Yakimov, M. M.,
534 Joachimiak, A., Ferrer, M., Golyshina, O. V., Savchenko, A., Golyshin, P. N., and Yakunin, A. F. (2017) Activity screening of environmental metagenomic libraries reveals novel carboxylesterase families. *Sci. Rep. 7*, 44103.

536 (27) Janes, L. E., Löwendahl, C., and Kazlauskas, R. J. (1998) Rapid quantitative screening of hydrolases
538 using pH indicators. Finding enantioselective hydrolases. *Chem. Eur. J. 4*, 2317-2324.

(28) Martínez-Martínez, M., Alcaide, M., Tchigvintsev, A., Reva, O., Polaina, J., Bargiela, R.,
540 Guazzaroni, M. E., Chicote, A., Canet, A., Valero, F., Rico Eguizabal, E., Guerrero, Mdel C., Yakunin, A. F., and Ferrer, M. (2013) Biochemical diversity of carboxyl esterases and lipases from Lake Arreo (Spain): a
542 metagenomic approach. *Appl. Environ. Microbiol.* 79, 3553-3562.

(29) Daiha, Kde G., Angeli, R., de Oliveira, S. D., and Almeida, R. V. (2015) Are lipases still important
544 biocatalysts? A study of scientific publications and patents for technological forecasting. *PLoS One 10*, e0131624.

546 (30) Borrelli, K. W., Cossins, B., and Guallar, V. (2010) Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility. *J. Comput. Chem. 31*, 1224-1235.

548 (31) Hernández-Ortega, A., Borrelli, K., Ferreira, P., Medina, M., Martínez, A. T., and Guallar, V. (2011) Substrate diffusion and oxidation in GMC oxidoreductases: an experimental and computational study on
550 fungal aryl-alcohol oxidase. *Biochemical. J. 436*, 341-350.

(32) Santiago, G., de Salas, F., Lucas, F., Monza, E., Acebes, S., Martínez, A., Camarero, S., and Guallar,
552 V. (2016) Computer-aided laccase engineering: toward biological oxidation of arylamines. *ACS Catalysis 6*, 5415-5423.

554 (33) Barak, Y., Nov, Y., Ackerley, D. F., and Matin, A. (2008) Enzyme improvement in the absence of structural knowledge: a novel statistical approach. *ISME J. 2*,171-179.

556 (34) Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F.,
558 Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M.,
560 Searson, S., Kandels-Lewis, S.; Tara Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan,
562 M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., and Bork P. (2015) Structure and function of the global ocean microbiome. *Science 348*, 1261359.

564 (35) Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A. (2016) Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins 84 Suppl 1*:4-
566 14.

(36) Barrozo, A., Duarte, F., Bauer, P., Carvalho, A. T. P., and Kamerlin, S. C. L. (2017) Cooperative
568 electrostatic interactions drive functional evolution in the alkaline phosphatase superfamily. *J. Am. Chem. Soc. 137,* 9061−9076.

570 (37) Alcaide, M., Stogios, P. J., Lafraya, Á., Tchigvintsev, A., Flick, R., Bargiela, R., Chernikova, T. N., Reva, O. N., Hai, T., Leggewie, C. C., Katzke, N., La Cono, V., Matesanz, R., Jebbar, M., Jaeger, K. E.,
572 Yakimov, M. M., Yakunin, A. F., Golyshin, P. N., Golyshina, O. V., Savchenko, A., Ferrer, M. (2015) Pressure adaptation is linked to thermal adaptation in salt-saturated marine habitats. *Environ. Microbiol. 17*,
574 332-345.

(38) Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001) Evaluation and
576 reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B 105*, 6474-6487.

578 (39) Borrelli, K. W., Vitalis, A., Alcantara, R., and Guallar, V. (2005) PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *Chem. Theory Comput. 1*, 1304-1311.

580 (40) Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013) Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided*
582 *Mol. Des. 27*, 221-234.

(41) Bochevarov, A. D., Harder, E., Hughes, T. F., Greenwood, J. R., Braden, D. A., Philipp, D. M.,
584 Rinaldo, D., Hall, M. D., Zhang, J., and Friesner, R. A. (2013) Jaguar: A high-performance quantum

chemistry software program with strengths in life and materials sciences. *Int. J. Quantum. Chem. 113*, 2110-
586 2142.

(42) Fraczkiewicz, R., and Braun, W. (1998) Exact and efficient analytical calculation of the accessible
588 surface areas and their gradients for macromolecules. *J. Comput. Chem. 19*, 319.

(43) Guilloux, V. L., Schmidtke, P., and Tuffery, P. (2009) Fpocket: An open source platform for ligand
590 pocket detection. *BMC Bioinformatics 10*, 168.

592 **Fig. legends**

594 **Figure 1.** Number of ester substrates hydrolyzed by each of the 145 EHs investigated in this study. The commercial preparations CalA and CalB (marked with filled square) are also included. This figure is created
596 from data in Supporting Information, Table S1. The activity protocol established and used to identify the esters hydrolyzed by each EH was based on a 550-nm follow-up pH indicator assay described in Supporting
598 Information, Methods. The list of the 96 structurally different esters tested is shown in Fig. 2. Full details of the activity protocol are provided in Supporting Information, Methods. The trend line shows a not-single
600 exponential fit of the experimental data. The fit was obtained using R script and the "Im" function, to extract a polynomial regression of degree 6 with the following line "model<-
602 lm(MM[,1]~poly(MM[,2],6,raw=TRUE))", where MM[,1] corresponds to the number of esters hydrolyzed, and MM[,2] the position in the x axe (from 1 to 147).

604

Figure 2. Hierarchical clustering of the substrate ranges of the EHs. Only EHs that hydrolyzed 10 or more
606 esters were considered (107 in total, including CalA/B). This figure is created from data in Supporting Information, Table S3. The specific activities of the EHs for each of the 96 esters were determined as
608 described in Fig. 1. The list of the 96 esters tested and the frequency of each ester considered as a hit (in brackets) are shown on the left side. The ID code representing each EH is given at the bottom. Each
610 hydrolase is named based on the code 'EH', which means Ester Hydrolase, followed by an arbitrary number from 1 to 145 for the most to least promiscuous enzyme. The number in brackets indicates the number of
612 esters hydrolyzed by each enzyme. The bio-source of each EH is indicated at the bottom with a number in white or black squares that follows the nomenclature in Supporting Information, Fig. S1. The Fig. was
614 created with the R language console using a binomial table with information about the activity/inactivity (1/0) of the analyzed enzymes against the 96 substrates as a starting point. For the central graphic, which
616 shows the data in Supporting Information, Table S3, we used the drawing tools provided by the basic core packages of R. The hierarchical clusters of the enzymes (shown at the top) and substrates (shown on the
618 right side) were generated by calculating a distance matrix using a "binomial" method and the hclust function to generate the tree. Using the functions as.phylo and plot.phylo from the ape package, the clusters
620 were added to the top and right of the figure. A combination of the Set1 palette from the R package RColorBrewer and colors from the basic palette from R were used as the color palette for sequences
622 assigned to each family (F) (see inset), including FI to FVII, carbohydrate esterase (CE), and carbon-carbon *meta*-cleavage product hydrolase (C-C MCP) families, all with a typical α/β hydrolase fold; FVIII serine
624 beta-lactamase with non α/β hydrolase fold; and cyclase-like protein from the amido-hydrolase superfamily. Sequences that were not unambiguously ascribed to existing families were referred to as "Unclassified", and
626 those of yeast origin were assigned to "yeast class". The two "clusters" C1 and C2 that contained the most substrate-promiscuous EHs are color-coded under a shadowed background.

628

Figure 3. Protein Energy Landscape Exploration (PELE) analysis. Panel (a) shows the protein-substrate interaction plots for EH1 (red) and EH102 (blue). Panel (b) shows the relative SASA for glyceryl tri-acetate in EH1 (red) and EH102 (blue) computed as a dimensionless ratio (0-1) using PELE.

Figure 4. Catalytic triad exposure of selected EHs with the broadest and lowest substrate ranges. (a) The catalytic triad (ball-and-sticks) and the main adjacent cavity (gray clouds) as detected by SiteMap are underlined to demonstrate the differences between a promiscuous (EH1) and non-promiscuous (EH102) EHs. EH1 can hydrolyze 72 esters and has a defined hidden binding cavity (effective volume: 166.7 $Å^3$). EH102, by contrast, can hydrolyze only 10 esters and has a surface-exposed triad (high SASA) and an almost negligible binding cavity (38.5 $Å^3$). The 3 top EHs with the broadest substrate ranges (b), positioned in the ranking after EH1, and the commercial CalB and CalA lipases (c), are also represented. On each panel, we highlight the catalytic triad and the main adjacent cavity as detected by SiteMap, demonstrating the differences in active site topology. EH2, EH3 and EH4, all assigned to FIV (as EH1), hydrolyzed 71, 69, and 67 esters and have defined but distinct hidden binding cavities (500, 200 and 200 $Å^3$, in the same order), as EH1. CalB, which was capable of hydrolyzing 68 esters, has a binding cavity (200 $Å^3$) that is also hidden but highly different from those of the other EHs. CalA, by contrast, hydrolyzed only 36 esters and has a low surface-exposed triad (SASA), with restrictive access to the catalytic triad (1000 $Å^3$).

Figure 5. The topology of the catalytic environment defines the substrate range of the EH. The figure shows the relationships between the active site effective volume (in $Å^3$) and enzyme promiscuity (number of substrates hydrolyzed). Note that the presented data were obtained using the active site cavity volume computed in $Å^3$ and SASA as a dimensionless ratio from 0 to 100 using the GetArea server (http://curie.utmb.edu/getarea.html). The panel contains information for EHs for which crystal structures (square) and homology models (circles) could be unambiguously established (sequence identity ≥25%) and the catalytic triad identified. Gray circles and squares indicate the EHs assigned to FIV. The analysis indicated a threshold ratio (indicated by the dashed gray line) at which it is possible to qualitatively classify substrate promiscuity based on hydrolysis of at least 20 substrates. Phylogenetic analysis further extended the substrate spectra to ≥42 esters, as only enzymes assigned to FIV and conforming to the 62.5 $Å^3$-threshold, together with CalB, were capable of converting such a high number of esters. The positioning for the commercial CalA and CalB lipases are indicated.

Figure 6. Relationships between the active site effective volume (in $Å^3$) and enzyme promiscuity (number of substrates hydrolyzed) of C2 members of HAD phosphatases. The number of substrates converted by each HAD phosphatase was obtained from Huang *et al.*[1] and is summarized in Supporting Information, Table S5. The panels contain information for HAD phosphatases for which crystal structures were available and the catalytic residues identified. The active site effective volume (in $Å^3$) was calculated as described in Fig. 5.

FIG 1

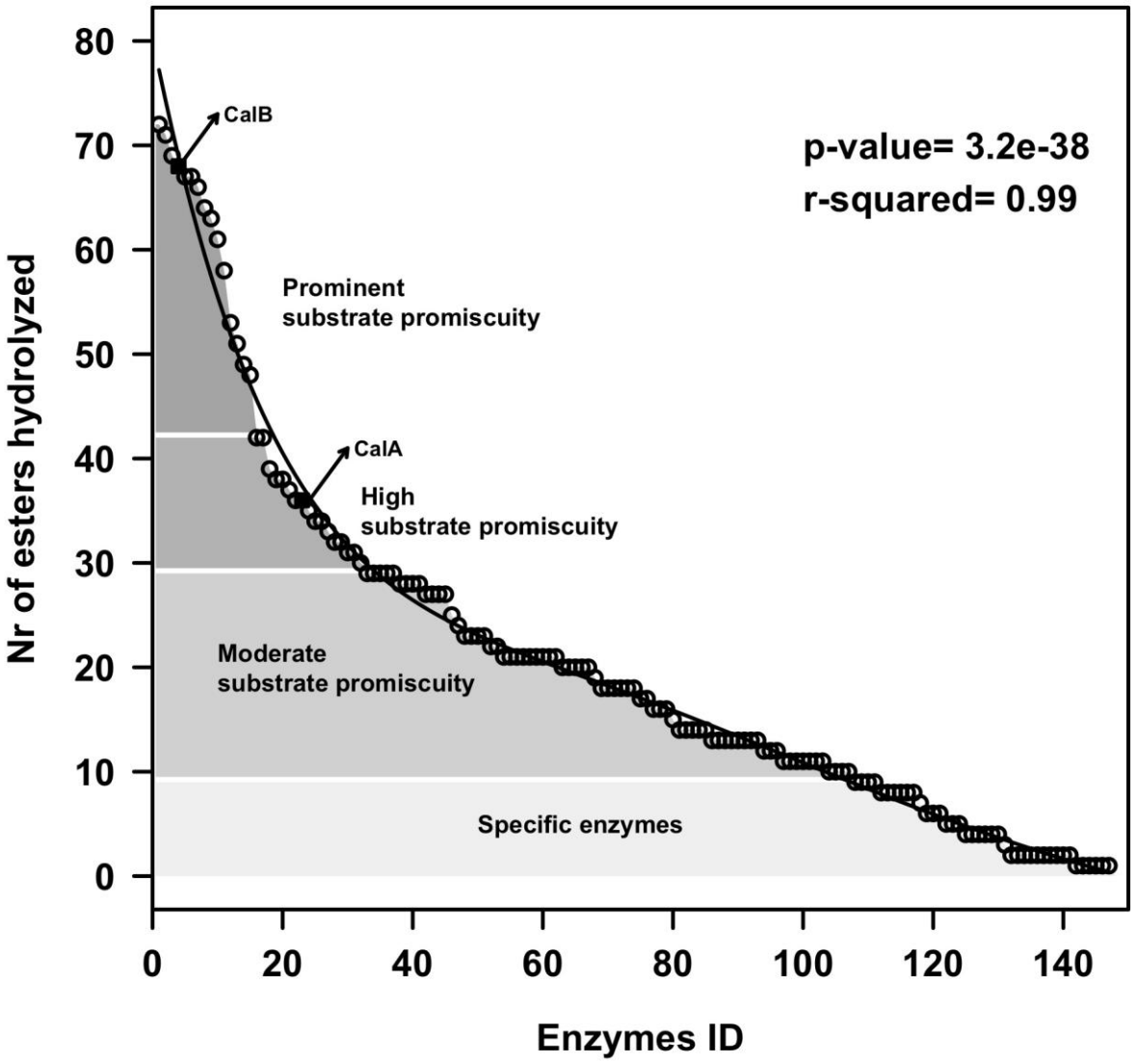

80
70
60
50
40
30
20
10
0
Nr of esters hydrolyzed
CalB
p-value= 3.2e-38
r-squared= 0.99
Prominent
substrate promiscuity
CalA
High
substrate promiscuity
Moderate
substrate promiscuity
Specific enzymes
0
20
40
60
80
100
120
140
Enzymes ID

FIG 2

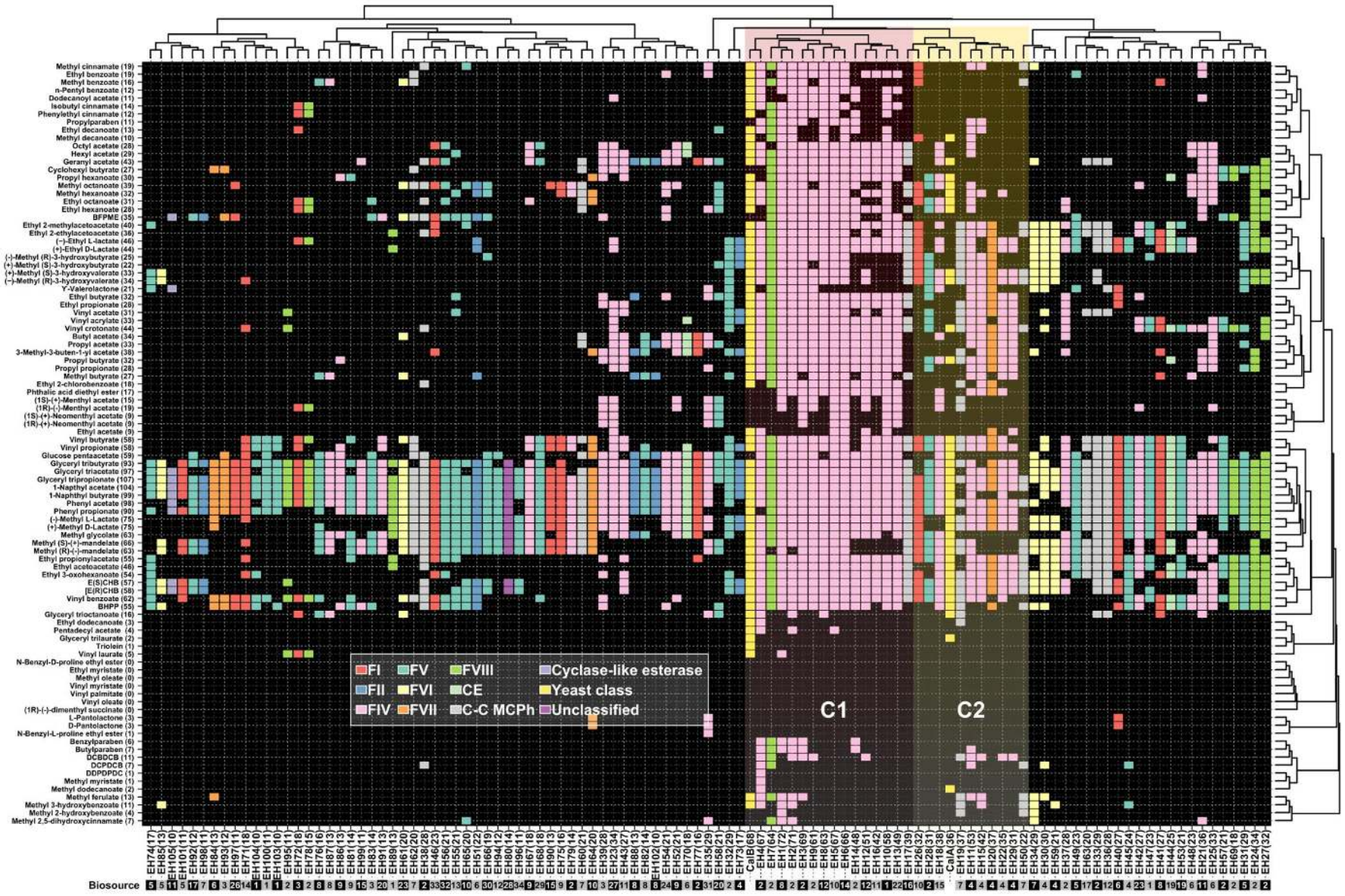

FI
FII
FIV
FV
FVI
FVII
FVIII
CE
C-C MCPh
Cyclase-like esterase
Yeast class
Unclassified
C1
C2
Biosource

FIG 3

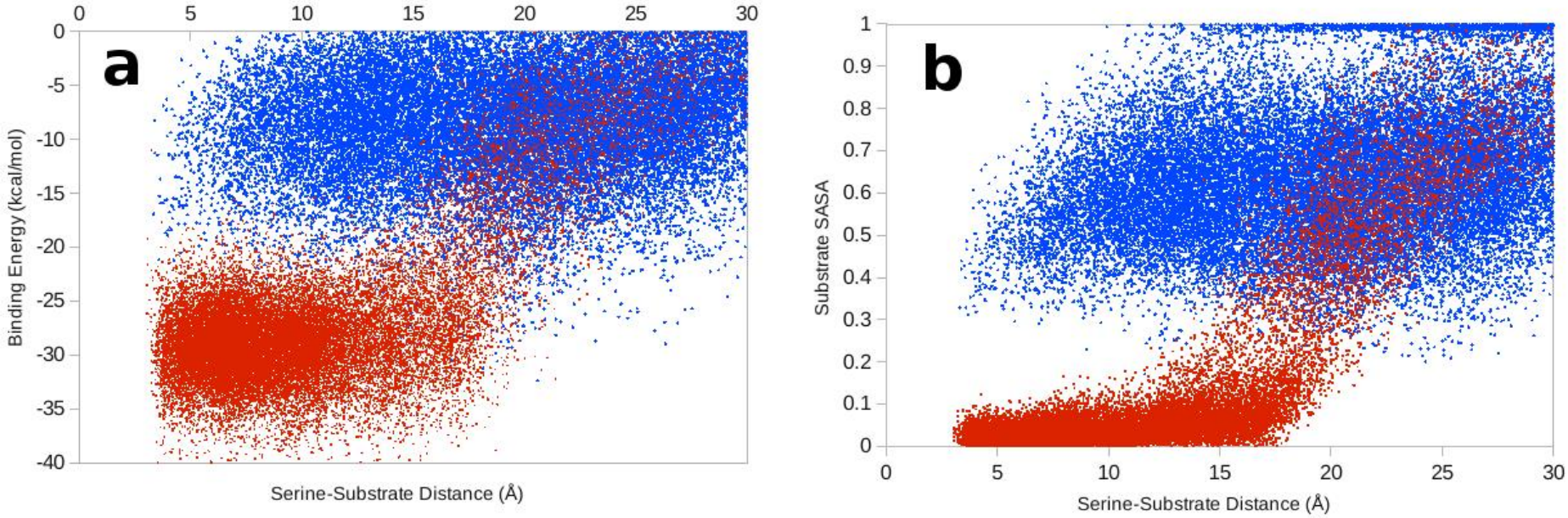

a
Binding Energy (kcal/mol)
0
-5
-10
-15
-20
-25
-30
-35
-40
0
5
10
15
20
25
30
Serine-Substrate Distance (Å)
b
Substrate SASA
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
Serine-Substrate Distance (Å)

FIG 4

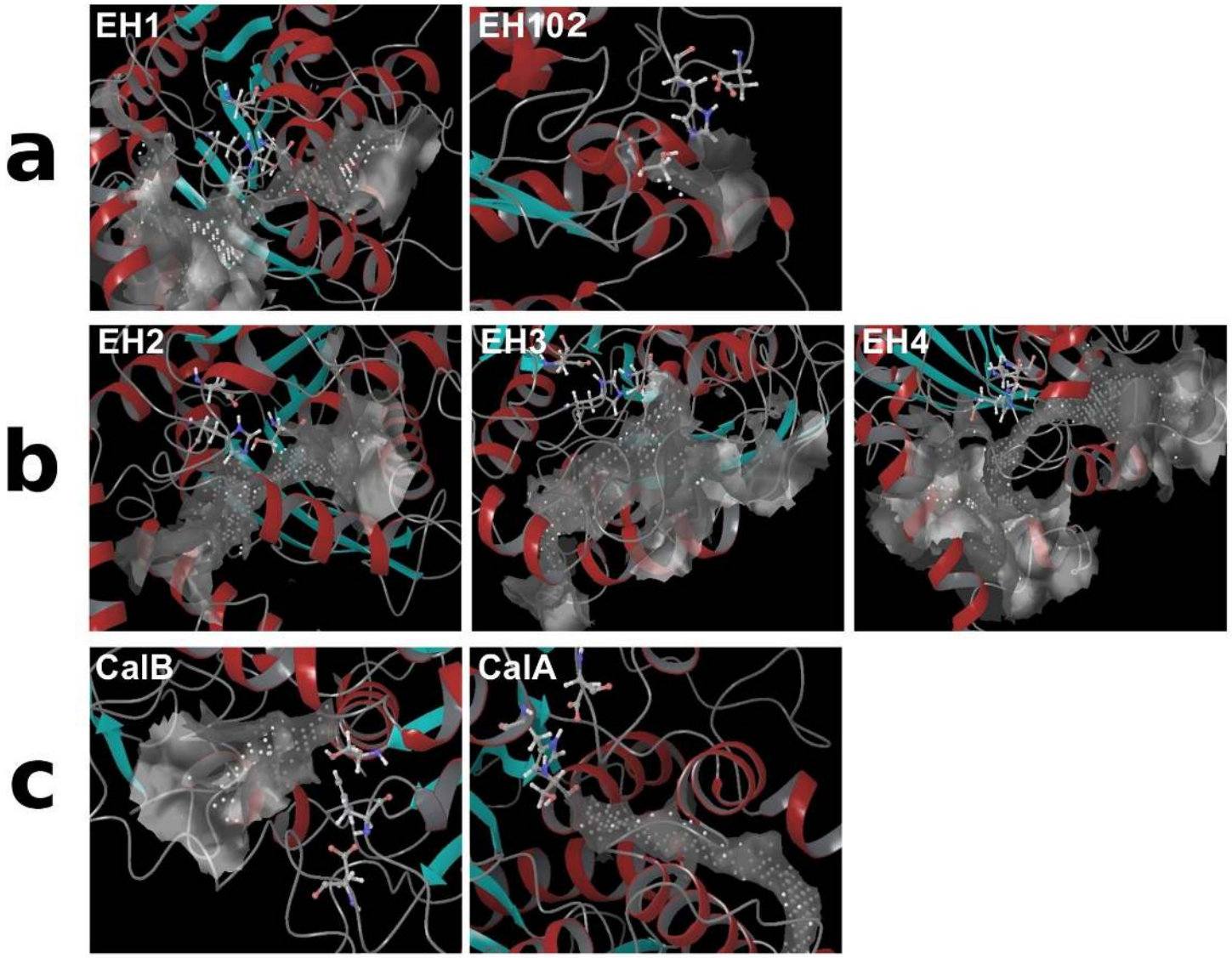


FIG 5

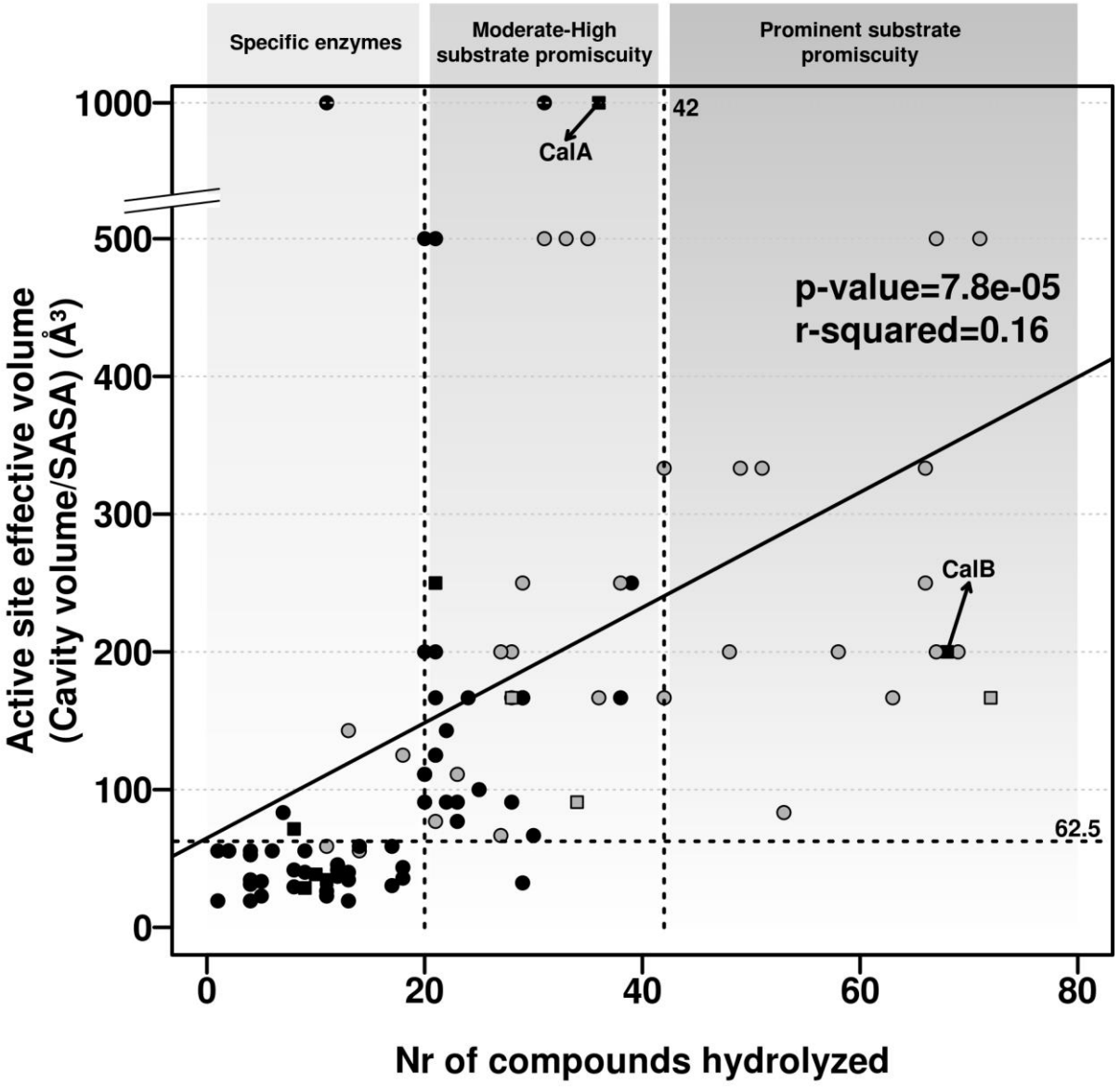
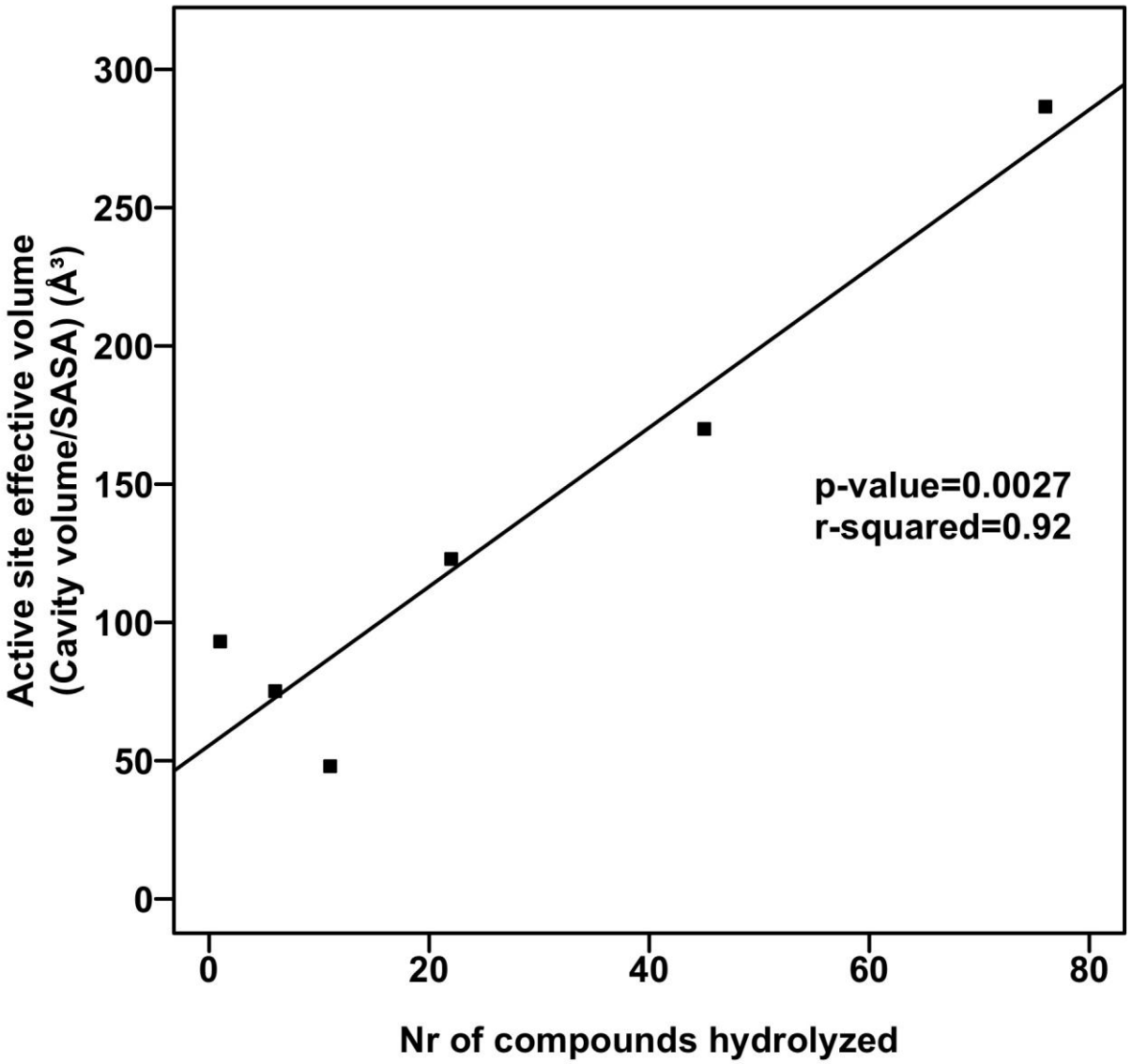

Specific enzymes
Moderate-High substrate promiscuity
Prominent substrate promiscuity
42
CalA
CalB
62.5
p-value=7.8e-05
r-squared=0.16
Active site effective volume (Cavity volume/SASA) (Å³)
Nr of compounds hydrolyzed
1000
500
400
300
200
100
0
0
20
40
60
80

FIG 6


Active site effective volume
(Cavity volume/SASA) (Å³)
300
250
200
150
100
50
0
0
20
40
60
80
Nr of compounds hydrolyzed
p-value=0.0027
r-squared=0.92