


An approach to functionally relevant clustering of the protein universe: Active site profile-based clustering of protein structures and sequences

Stacy T. Knutson,^{1,2} Brian M. Westwood,^{1,2} Janelle B. Leuthaeuser,³ Brandon E. Turner,¹ Don Nguyendac,¹ Gabrielle Shea,¹ Kiran Kumar,¹ Julia D. Hayden,⁵ Angela F. Harper,¹ Shoshana D. Brown,⁴ John H. Morris,⁴ Thomas E. Ferrin,⁴ Patricia C. Babbitt,⁴ and Jacquelyn S. Fetrow ^{6*}

¹Department of Physics, Wake Forest University, Winston-Salem, North Carolina 27106

²Department of Computer Science, Wake Forest University, Winston-Salem, North Carolina 27106

³Molecular Genetics and Genomics Program, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157

⁴Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158

⁵Biochemistry Program, Dickinson College, Carlisle, Pennsylvania 17013

⁶Department of Chemistry, University of Richmond, Richmond, Virginia 23173

Received 10 December 2016; Accepted 22 December 2016

DOI: 10.1002/pro.3112

Published online 5 January 2017 proteinscience.org

Abstract: Protein function identification remains a significant problem. Solving this problem at the molecular functional level would allow mechanistic determinant identification—amino acids that distinguish details between functional families within a superfamily. Active site profiling was developed to identify mechanistic determinants. DASP and DASP2 were developed as tools to search

Abbreviations: Chi-MLE, chloromuconate cycloisomerase; DGalN, d-galactonate dehydratase; DiPepEp, dipeptide epimerase; DTartD, d-tartrate dehydratase; GalD, galactarate dehydratase; GlucD, glucarate dehydratase; LFucD, l-fuconate dehydratase; LTalGalD, l-talarate/galactarate dehydratase; ManD, mannonate dehydratase; MAL, methylaspartate ammonia lyase; MLE, muconate cycloisomerase; MLEanti, muconate cycloisomerase-anti; MLEsyn, muconate cycloisomerase-syn; MR, mandelate racemase; NSAR, N-succinylaminoacid racemase; NSAR2, N-succinylaminoacid racemase 2; OSBS, O-succinylbenzoate synthase; Prx, peroxiredoxins; RhamD, rhamnonate dehydratase

Additional Supporting Information may be found in the online version of this article.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes. S.T. Knutson and B.M. Westwood are co-first authors on this manuscript.

Availability of Data and Materials: All data generated or analyzed during this study are included in this published article (and its supplementary information files) or available from corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Broader Audience Statement: Enzyme functional site analysis is a major unsolved problem. Experiments are difficult and expensive and algorithms to identify functional sites from sequence are plagued with issues of mis- and over-annotation. Here, TuLIP, a novel approach to clustering proteins of known structure based on active site information, is shown to produce functionally relevant clusters for proteins of known structure. This algorithm is a major step towards producing an automated method for functionally relevant clustering of the protein universe.

Grant sponsor: National Institutes of Health; Grant numbers: T32-GM095440 to JBL, grant GM60595 to PCB, and grant P41-GM103311 to TEF, JHM.

*Correspondence to: Jacquelyn S. Fetrow; Maryland Hall, Suite 202, 28 Westhampton Way, University of Richmond, VA 23173. E-mail: jfetrow@richmond.edu

sequence databases using active site profiling. Here, TuLIP (Two-Level Iterative clustering Process) is introduced as an iterative, divisive clustering process that utilizes active site profiling to separate structurally characterized superfamily members into functionally relevant clusters. Underlying TuLIP is the observation that functionally relevant families (curated by Structure-Function Linkage Database, SFLD) self-identify in DASP2 searches; clusters containing multiple functional families do not. Each TuLIP iteration produces candidate clusters, each evaluated to determine if it self-identifies using DASP2. If so, it is deemed a functionally relevant group. Divisive clustering continues until each structure is either a functionally relevant group member or a singlet. TuLIP is validated on enolase and glutathione transferase structures, superfamilies well-curated by SFLD. Correlation is strong; small numbers of structures prevent statistically significant analysis. TuLIP-identified enolase clusters are used in DASP2 GenBank searches to identify sequences sharing functional site features. Analysis shows a true positive rate of 96%, false negative rate of 4%, and maximum false positive rate of 4%. *F*-measure and performance analysis on the enolase search results and comparison to GEMMA and SCI-PHY demonstrate that TuLIP avoids the over-division problem of these methods. Mechanistic determinants for enolase families are evaluated and shown to correlate well with literature results.

Keywords: functional site profile; active site profile; mechanistic determinants; isofunctional clusters; function annotation; functionally relevant clustering; misannotation

Introduction

Since the first sequencing of the *Haemophilus influenzae* genome in 1995,¹ databases have exploded with gene and protein sequences; however the majority of this ever-increasing number of proteins and predicted proteins lack function annotation. Despite development of some large scale experimental assays for evaluating protein function, experimental determination remains expensive and time consuming: it has been estimated that less than 5% of protein functions have been experimentally determined.² Computational methods to efficiently identify protein function are required; however, it is essential that such methods be accurate, as mis-annotation is a well-documented problem.^{3–5} The issue of over-annotation—annotating function to a level of detail beyond which the method can address—has also been clearly demonstrated.^{6–8}

The most useful function annotation methods will identify molecular functional details and divide protein superfamilies into functionally relevant groups based on those details. Accurate clustering of proteins centered on molecular functional details would address the problem of over-annotation. In this manuscript, we use the term “isofunctional family” to describe a protein family in which all family members share mechanistic details. Our goal is to identify isofunctional families within protein superfamilies. Importantly, mechanistic determinants—those residues that are involved in the mechanism or substrate specificity within an isofunctional family—could be identified. Mechanistic determinants drive function diversification of biological processes in protein superfamilies; thus, simplifying their recognition would be valuable.

Although methods that focus on molecular function details are limited, protein clustering has a long

history. Large databases, such as CATH^{9,10} and PFAM,^{11–13} divide protein superfamilies into groups based on structural and/or sequence similarity. FunFams¹⁴ are an extension to CATH that use an automatic hierarchical clustering approach and hidden Markov models (HMMs) to identify functional families within superfamilies exhibiting the same Enzyme Commission (EC) number. There is no tuning to a particular family; however, the advantage of this method is automation. The Enzyme Function Initiative approach classifies large numbers of homologues using sequence similarity networks,¹⁵ providing a way to analyze an entire superfamily. Yet, the correlation between full sequence similarity and the details of molecular function are not well understood.¹⁶ Indeed, it is this relationship between molecular function and computational clustering that is critical to such methods. Validation of functionally relevant clustering methods must be accomplished on groups for which the functional relationship is known (such as those in the Structure-Function Linkage Database, SFLD).^{17,18}

In 2007, the Subfamily Classification in Phylogenomics (SCI-PHY) algorithm emerged as a large-scale method to group proteins into functional families using phylogenomic classifications.¹⁹ A hierarchical tree is created from a starting set (typically a Pfam family or SFLD superfamily) based on profile comparisons; the tree is cut to optimize the balance between number of clusters and sequence diversity within the clusters. Subsequently, HMMs are created for each cluster¹⁹ and new sequences are evaluated against each HMM to identify the cluster to which each sequence belongs (if any). SCI-PHY demonstrated superior performance compared to other sequence-based methods in clustering functionally related proteins within a superfamily. However, SCI-

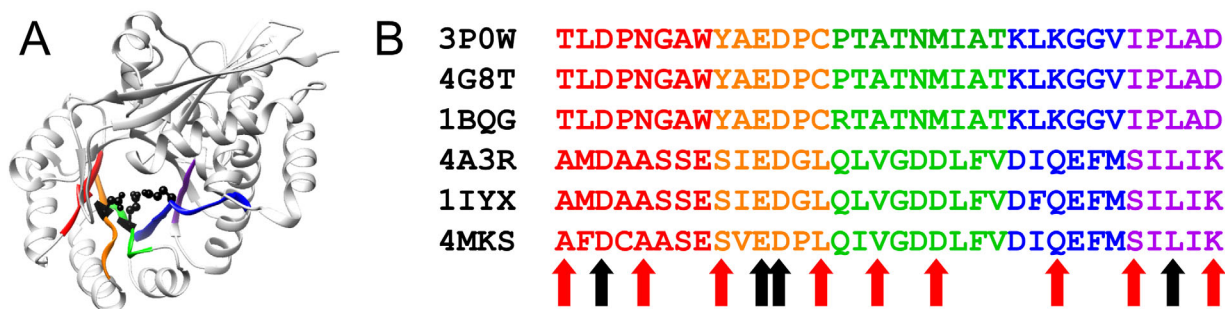


Figure 1. Active Site Profiling (ASP) approach to identify features of the active site microenvironment. In a protein structure, key residues (represented as black side chains) important for molecular function are identified. All residues within 10 Å of a key residue are selected (colored fragments, A and B). The sequences of these discontinuous fragments or motifs are concatenated N-terminus to C-terminus to create the *active site signature* for each protein (B, individual lines) which we hypothesize contains all or most of the active site's functionally relevant features; letter colors represent continuous fragments or motifs extracted from the structure. Multiple signatures are aligned to create an *active site profile*, ASP (B). An ASP score (see Methods)²² quantifies the sequence similarity across the profile. Black arrows represent positions conserved throughout the entire profile. Red arrows represent some of the positions that differ between functional groups and may indicate specificity determining positions.

PHY requires a multiple sequence alignment (MSA) of all proteins in the superfamily; thus, the method is limited by the attendant issues of alignment accuracy and gap placement.

Three years later, the Genome Modelling and Model Annotation (GeMMA) method was developed.²⁰ GeMMA utilizes pairwise sequence comparisons (rather than a MSA) to create a hierarchy of proteins from a known protein superfamily. In addition, the hierarchical clustering is considered complete once the profile comparisons no longer pass the significance threshold. Although the execution of GeMMA results in high performance scores,²⁰ GeMMA subdivides superfamilies into many clusters with little evidence to suggest these many small groups each represent distinct functions. Both GeMMA and SCIPHY utilize alignment of full sequence within a large family, rather than focusing on functional sites; consequently, mechanistic determinants of subfamilies might be evaluated after the completion of the process, but are not identified during the process.

Active Sites Modeling and Clustering (ASMC) was also developed in 2010 to identify and cluster proteins into isofunctional groups by focusing on the functional site details.²¹ In ASMC, active sites are identified in proteins of known structure and homology models are built for proteins of unknown structure within the superfamily. Features of the homology models are compared by creating a MSA of active site fragments, similar to the approach used in active site profiling,²² and then hierarchically clustering those fragments. ASMC demonstrated high performance on the validation test set. Subsequently, ASMC was used in conjunction with phylogenetic analyses, genomic context, and sequence similarity to characterize the BKACE superfamily into functionally relevant groups based on active site similarity.²³ Because of homology modeling accuracy, the approach is limited to sequences with

greater than 30% sequence identity to a characterized superfamily structure. Though mechanistic determinants of functional sites can be identified, the lack of widespread structural characterization in some superfamilies diminishes ASMC's ability to cluster all protein superfamilies.

To easily identify and compare mechanistic determinants in known functional families, we previously developed active site profiling.²² In this approach, key functional residues are identified in a given protein [Fig. 1(A), black side chains]. All residues within 10 Å of a key residue are identified [Fig. 1(A), colored fragments]. These active site fragments are concatenated N- to C-terminus to create an active site signature for each protein; signatures are aligned to create an active site profile (ASP) [Fig. 1(B)]. As defined by Cammer and coworkers,²² an ASP score defines the similarity between the signatures in the profile. A profile allows identification of potential molecular determinants common across the whole family [Fig. 1(B), black arrows] or consistent within subfamilies [Fig. 1(B), red arrows].

In 2005, we subsequently developed an approach called DASP (Deacon Active Site Profiler) that uses the active site profiling concept to search sequence databases^{24–26} (Fig. 2). We have applied ASP-based sequence searching (DASP) to the peroxiredoxin (Prx)²⁵ and cytochrome P450²⁷ superfamilies with good success. Like DASP, PSI-BLAST uses profiling of multiple sequences; however, profiles are calculated across the entire sequence, rather than across the active site fragments. Comparison of DASP to PSI-BLAST in Prx searches has demonstrated the superior accuracy of DASP towards functional family-specific searches (Supporting Information File 1, Fig. S1).

Despite these developments, there remains a need to develop automatable methods to cluster proteins in functionally relevant ways. We are working

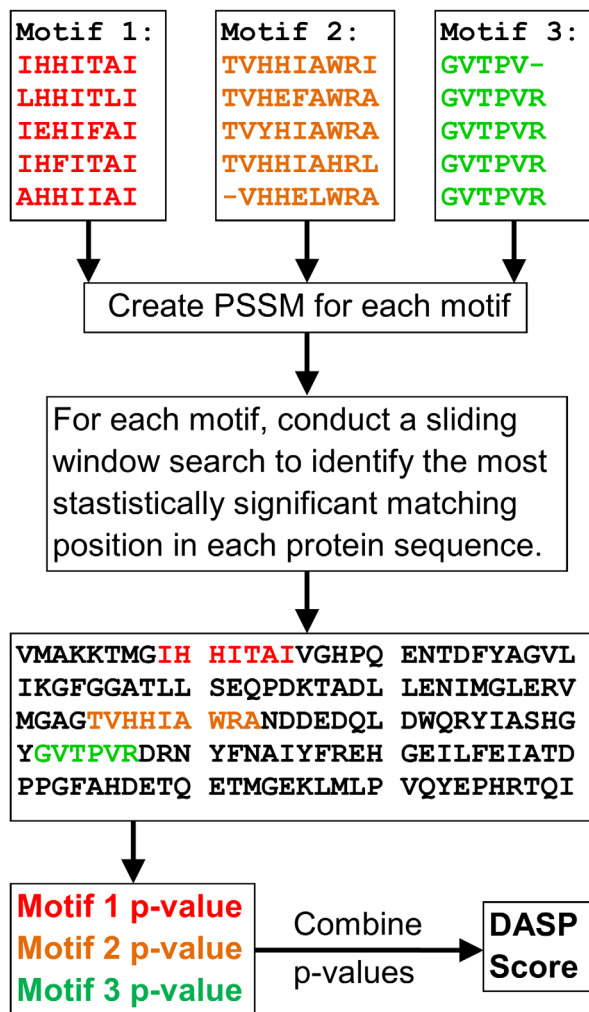


Figure 2. DASP method using ASPs to search for sequences that contain functional site features similar to those represented in the original profile. Sequence databases can be searched for proteins containing active site features similar to those in the original ASP using a tool called DASP^{24,30}; DASP2 was developed to improve input and algorithmic details³⁵. As described in Methods, the ASP is split into individual motifs (representing the colored, structurally continuous fragments in Fig. 1) (top). A position-specific scoring matrix (PSSM) is created for each motif. For each sequence in the database, the best match (most-significant p -value) of each motif-specific PSSM to a sequence fragment is identified using a sliding window search. The individual motif p -values are combined using QFAST, which produces the DASP score, which represents the statistical significance of the fragment matches to all ASP motifs. This process of motif matching and significance calculation is repeated for every sequence in the database.

to develop such a method using active site profiling, in which active site similarities drive clustering of functional groups. One limitation of the approach applied to the Prxs is that identification of the input clusters was manual, relying on expert analysis. Thus, the method identifies mechanistic determinants, but only with *a priori* knowledge of group

members. In this contribution, we address this limitation. We introduce a process called TuLIP (Two-Level Iterative clustering Process), an iterative, divisive clustering process that utilizes pairwise ASP scores as the edge metric in a similarity network approach to cluster the structurally characterized members of a superfamily into groups. Our previous work¹⁶ suggested such clusters should correlate with molecular function, and in this contribution we show that TuLIP-identified clusters do, in fact, correlate with functional families identified by SFLD curators. TuLIP is part of a two-step process for clustering both protein structures (TuLIP) and protein sequences (MISST).²⁸

Results and Discussion

ASPs discretely identify most SFLD-identified enolase superfamily subgroups and families

We hypothesize that ASPs capture key functional site features or functional determinants, a hypothesis supported by previously published results on the Prxs^{25,29} and other protein families.¹⁶ If this hypothesis is correct, database searches with ASPs created for functionally relevant groups should identify only group members and no other proteins at significant scores. Because subgroups and families within each SFLD enzyme superfamily represent curated functionally relevant groups, we initially asked whether ASPs created for proteins of known structure in each SFLD-identified subgroup in the enolase superfamily would identify subgroup members only, but not other superfamily members. A summary of the enolase superfamily hierarchy used throughout this manuscript are provided in Supporting Information File 2.

ASPs were created as previously described²² (process outlined in Fig. 1) for each of the seven subgroups in the enolase superfamily: enolase, mannate dehydratase (ManD), glucarate dehydratase (GlucD), galactarate dehydratase (GalD), methylaspartate ammonia lyase (MAL), mandelate racemase (MR), and muconate cycloisomerase (MLE). Profiles were created using structures present in the 2011 SFLD (Table I, column 3). DASP searches^{26,30} (method outlined in Fig. 2) of the 2013 Protein Databank (PDB) (Table I, column 4) were then performed using these profiles.

Five of the seven subgroups (enolase, ManD, GlucD, GalD, and MAL) produced ASPs with positive profile scores (Fig. 3). Searching the PDB with these profiles identified all subgroup members, including those not used to build the profile, at scores more significant than the trusted DASP score threshold ($1e-10$, dashed line, Fig. 3; “trusted score threshold” defined in Methods), and subgroup members were distinct from members of other subgroups

Table I. Subgroup and Family Members of the Enolase Superfamily for DASP Searches and TuLIP

Subgroup	Family	Structs (Aug 2011)	Structs (June 2013)	TuLIP Structs ^a
Enolase	Enolase	37	56	18
Mannosate Dehydratase (ManD)	Uncharacterized	2	14	8
	Mannosate Dehydratase (ManD)	4	22	9
Glucuronate Dehydratase (GlucD)	Uncharacterized	1	8	4
	Glucuronate Dehydratase (GlucD)	7	22	12
Galacturonate Dehydratase (GalD)	Galacturonate Dehydratase (GalD)	5	6	1
Methylaspartate Ammonia-Lyase (MAL)	Methylaspartate Ammonia-Lyase (MAL)	4	6	3
Mandelate Racemase (MR)	Uncharacterized	24	76	37
	D-Tartrate Dehydratase (DTartD)	3	3	1
	L-Fuconate Dehydratase (LFucD)	4	5	2
	L-Talarate/Galacturonate Dehydratase (LTalGalD)	5	5	2
	Mandelate Racemase (MR)	6	8	2
	Rhamnosate Dehydratase (RhamD)	11	11	3
	D-Galactonate Dehydratase (DGalnD)	0	0	1
Muconate Cycloisomerase (MLE)	Uncharacterized	7	11	7
	Muconate Cycloisomerase – anti (MLEanti)	3	3	1
	Muconate Cycloisomerase – syn (MLEsyn)	9	10	3
	Chloromuconate Cycloisomerase (Chl-MLE)	2	2	2
	Dipeptide Epimerase (DipepEp)	21	29	9
	N-Succinylamino Acid Racemase (NSAR)	9	9	2
	N-Succinylamino Acid Racemase 2 (NSAR2)	3	3	1
	O-Succinylbenzoate Synthase (OSBS)	24	27	12

^a Nonredundant structures utilized in the TuLIP profiles.

(blue bars distinct from red bars, Fig. 3). Thus, ASPs built from SFLD-identified subgroup members capture relevant active site features and utilize those features to identify other members of the subgroup.

Members of two subgroups of the enolase superfamily, MR and MLE, were not discretely identified in DASP searches using subgroup-specific ASPs—meaning enolase superfamily sequences which were not subgroup members were identified from the subgroup-specific searches (interwoven red and blue bars, Fig. 3, bottom). These subgroups are well-represented in the PDB (Fig. 3, column 3). Such representation results in diversity of the active site, producing negative ASP scores (Fig. 3, column 4), suggesting further subdivision of these two subgroups is required. In the SFLD hierarchy, a finer level of functional detail is represented by family¹⁷; therefore, ASPs were built for each family in the MR and MLE subgroups. These profiles were used as input to DASP searches of the PDB.

In the MR subgroup, ASPs were built for the D-tartrate dehydratase (DTartD), L-fuconate dehydratase (LFucD), L-talarate/galacturonate dehydratase (LTalGalD), mandelate racemase (MR), and rhamnosate dehydratase (RhamD) families. The higher ASP scores [Fig. 4(A), column 4] indicate these family-specific profiles are less diverse than the overall subgroup ASP. In DASP searches, every family identifies its members discretely, as the family members [blue bars, Fig. 4(A)] are distinguished from the nonfamily members [red bars, Fig. 4(A)] by

at least two DASP score orders of magnitude. Thus, ASPs built from each MR family capture functional site features that are sufficient to distinguish family members from nonfamily members, similar to what was observed for the other five enolase superfamily subgroups (Fig. 3).

ASPs were also built for each family in the MLE subgroup: muconate cycloisomerase-anti (MLE-anti); muconate cycloisomerase-syn (MLE-syn); chloromuconate cycloisomerase (Chl-MLE); dipeptide epimerase (DipepEp); o-succinylbenzoate synthase (OSBS); N-succinyl-amino acid racemase (NSAR); and N-succinyl-amino acid racemase 2 (NSAR2). ASPs for three of the seven families (MLE-anti, NSAR, and NSAR2) distinguished family members from the other MLE and enolase superfamily families [blue bars separated from red bars, Fig. 4(B)]. Searches using ASPs for the other four MLE families were less robust. DASP searches with these ASPs identified family members [blue bars, Fig. 4(B)] as well as members of other MLE and enolase superfamily families [red bars, Fig. 4(B)] at significant scores. Such cross-hits showed that the MLE-syn, Chl-MLE, DipepEp, and OSBS families are less easily distinguished using the ASP approach. These families were also difficult to distinguish by expert curators.^{31–33}

TuLIP: An ASP-based method to cluster superfamilies into functionally relevant groups

The ability to use ASPs to identify and retrieve 13 of 17 SFLD-identified enolase superfamily subgroups

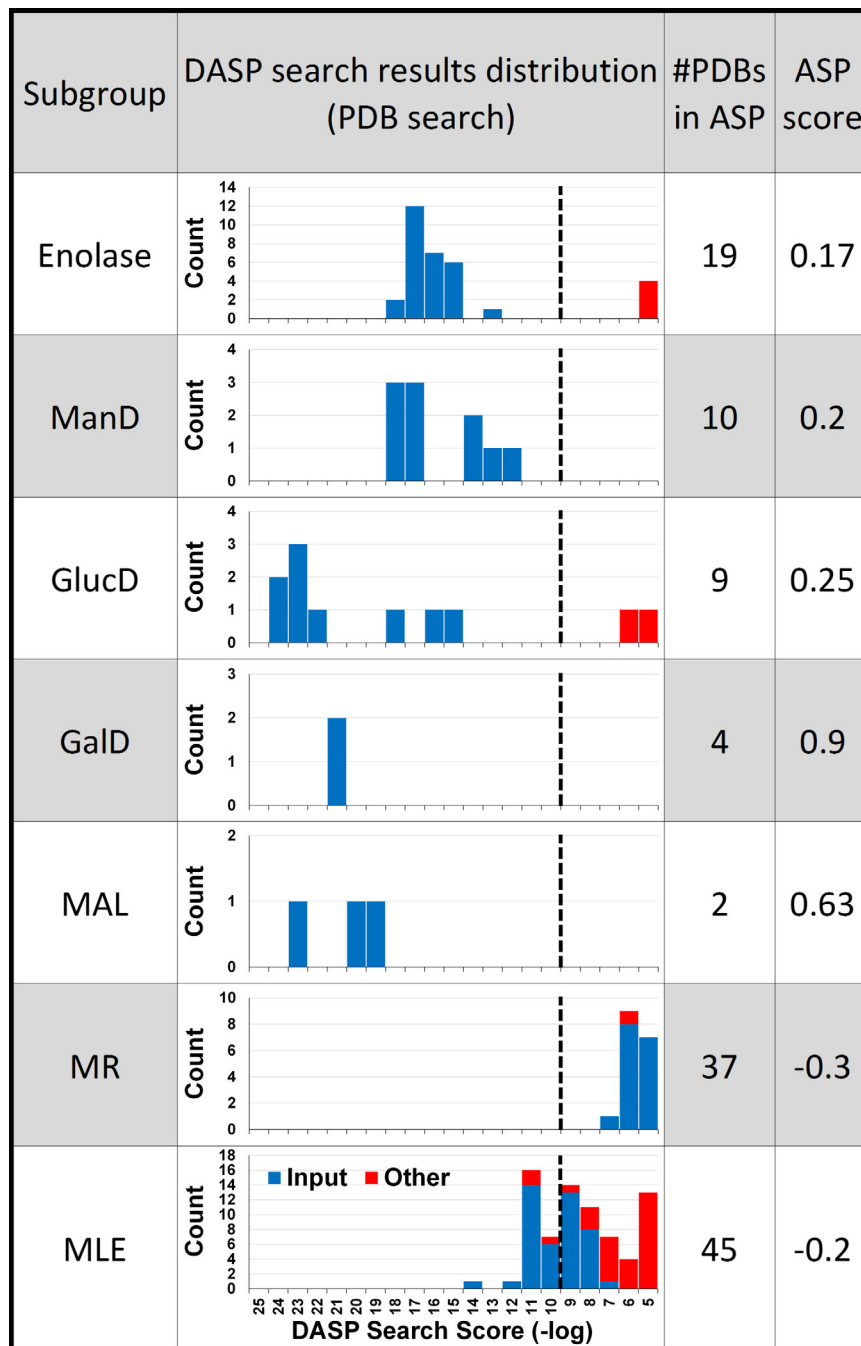


Figure 3. ASPs distinguish subgroup members from all other enolase superfamily proteins in five of seven subgroups. An ASP was built for the nonredundant proteins of known structure (April 2011) in each enolase superfamily subgroup. Each ASP was used to search the 2013 PDB using DASP. Search results are displayed as a distribution of DASP search scores. Blue bars indicate proteins that are SFLD-identified subgroup members. Red bars represent enolase superfamily proteins that are not members of the subgroup used for the search. The black dotted line indicates the original trusted DASP score threshold of $1e-10$. The third and fourth columns indicate the number of protein sequences used to create the profile and the ASP score (see Methods) for that profile, respectively. Domains that are 100% identical are counted only once in a given bin.

and families suggests the ASPs pinpoint molecular details that distinguish functionally relevant groups, a result that prompts the following question: Is it possible to create an automatable procedure that uses active site profiling to divisively cluster protein superfamily members of known structure into functionally relevant groups?

TuLIP was developed to accomplish this goal. This process, described in detail in Methods and outlined in Figure 5, begins by creating an active site signature for each superfamily member whose structure is known. A pairwise ASP score is calculated for each pair of signatures, creating an all-by-all pairwise score matrix. These pairwise ASP scores

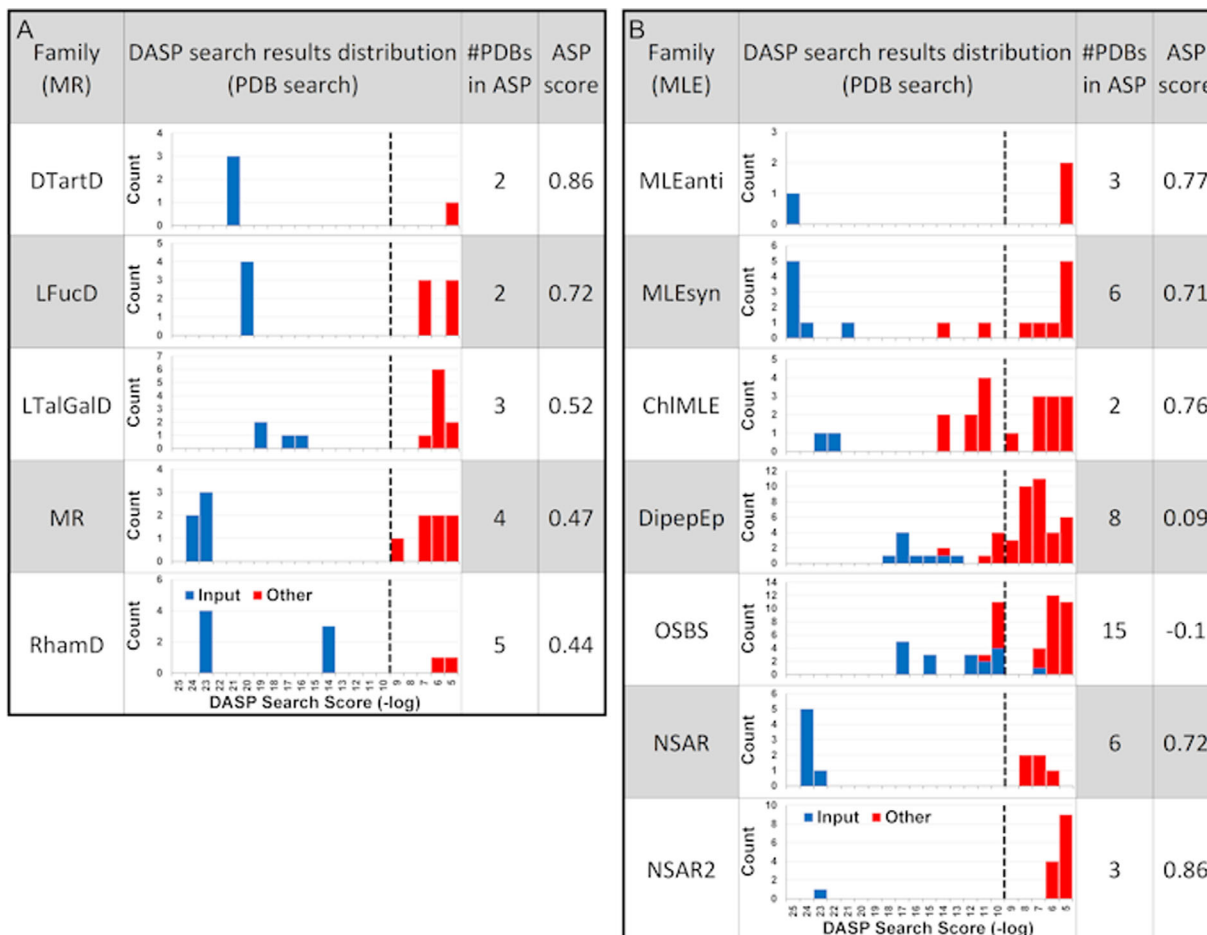


Figure 4. ASPs can distinguish all five MR families (A) and three of seven MLE families (B) from among all PDB sequences. Columns and axes are as in Figure 3. Blue bars indicate proteins that are SFLD-identified family members in either the MR (A) or MLE (B) subgroups. Red bars are proteins that are not members of the family used for the search. The black dotted line indicates the original DASP score threshold of $1e-10$. Domains that are 100% identical are counted only once in a given bin.

serve as the edge metric for network-based clustering of active site signatures. Discrete clusters are identified by gradually increasing the pairwise score threshold and executing MCL clustering,³⁴ a process which isolates clusters of signatures that are more similar to each other than to any of the other signatures [Fig. 5(A)]. Each discrete cluster is examined for *validity as a functionally relevant cluster* by creating a profile for the cluster members and analyzing if a DASP search of PDB distinctly identifies only cluster members. To support TuLIP, a new version of DASP, DASP2,³⁵ was developed to improve the efficiency of database searching and overcome certain edge case anomalies noted in the original DASP implementation (see Methods).

A two-stage process identifies *validated clusters*, first using stringent criteria (strict, or Sct, clusters), and second using less stringent criteria (relaxed, or Rlx, clusters) [Fig. 5(B); see Methods]. The relaxed stage identifies weaker relationships in which proteins share active site similarity, but which may have been obscured by the strong relationships identified in the strict clustering stage.

TuLIP was validated by applying the process to a set of 160 nonredundant enolase superfamily proteins of known structure present in the 2014 SFLD (representing 339 enolase superfamily sequences; Table I; Supporting Information File 2). Twenty three functionally relevant groups were identified: 16 and 7 groups from the strict (Sct) and relaxed (Rlx) stages, respectively (groups listed in Table II). TuLIP placed 23 proteins into their own groups as singlet proteins, including the lone nonredundant structures representing the GalD, NSAR2, and MLE-antifamilies (Supporting Information File 1, Fig. S2, blue bracket). Singlet identification indicates the active site features of each protein were dissimilar from all other superfamily members of known structure.

The SFLD-defined annotation of each protein was compared to the TuLIP groupings (Supporting Information File 1, Fig. S2). High correspondence between the TuLIP and SFLD groups is observed, including five of the seven subgroups identified distinctly and four of the six MR families identified distinctly. Notably, 12 MR subgroup

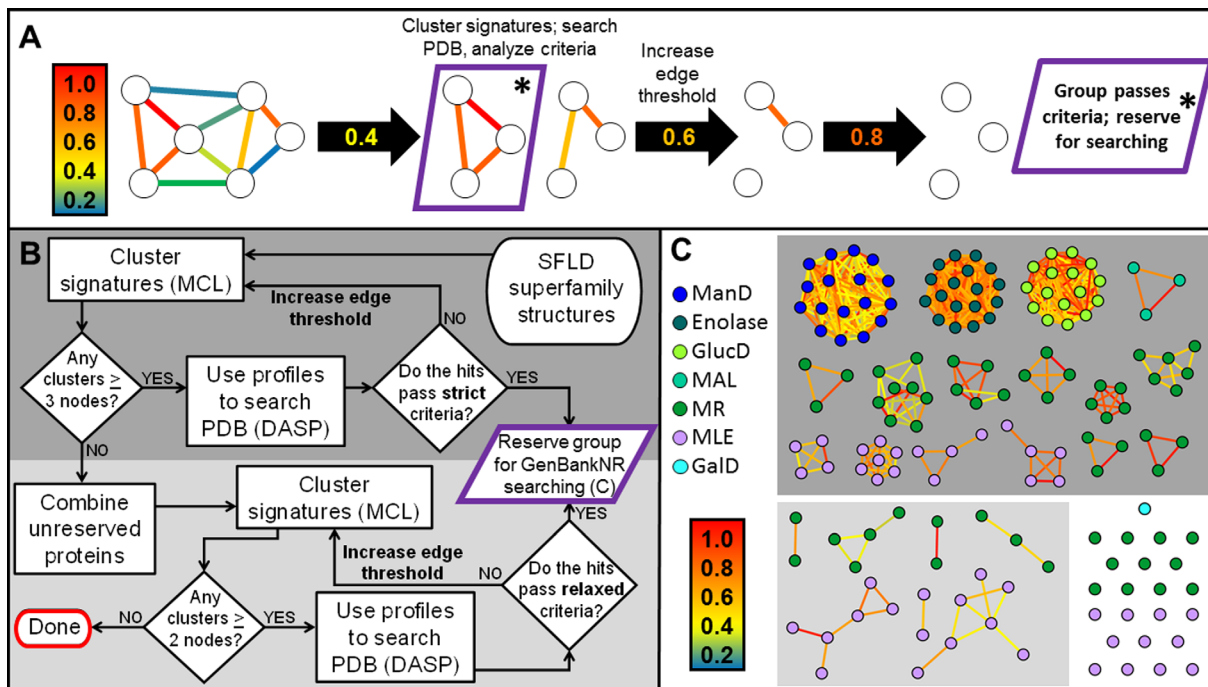


Figure 5. TuLIP, an automatable process of functionally relevant clustering of proteins of known structure, identifies functionally relevant enolase protein clusters. (A) An illustration of MCL clustering³⁴ through stepwise increases in ASP score thresholds to identify groups in which the active sites of group members are more similar within the cluster than they are to all other proteins. This concept underlies the TuLIP algorithm. (B) The TuLIP algorithm outlines the iterative, two-stage process. The two stages, strict and relaxed (defined in Methods), are shown on dark and light gray backgrounds respectively. In each stage, the algorithm proceeds through iterative MCL clustering at increasing edge thresholds, as illustrated visually in A. Upon completion of the algorithm, all proteins are either members of a functionally relevant group (purple parallelogram) or have been subdivided into singlets (red circle). All steps outlined in (B) are automatable, meaning that each step uses objective criteria and no step requires human interpretation of data. (C) TuLIP-identified clusters for the enolase superfamily. The node color represents SFLD subgroup designation of the protein, according to the legend. Clusters of the enolase superfamily satisfying either strict or relaxed criteria, and thus identified as functionally relevant clusters, are shown on the dark (strict) and light gray (relaxed) backgrounds, respectively. Singlet proteins not included in any cluster by TuLIP are displayed on white background. Cluster edge colors (A and C) represent pairwise ASP scores (scoring function as previously defined²²), according to the color legend (A).

uncharacterized proteins were assigned by TuLIP to clusters that correspond to SFLD families. TuLIP also identifies new potentially functionally relevant clusters not previously identified by SFLD curators: 25 of the 48 MR proteins which SFLD labels as “uncharacterized MR family” were clustered into seven newly identified TuLIP groups (Supporting Information File 1, Fig. S2, orange bracket).

Two exceptions to the strong qualitative correspondence between SFLD and TuLIP groups are observed. In the MR subgroup, TuLIP combines the L_{FucD} and L_{TalGalD} families into one cluster, Sct23 (Supporting Information File 1, Fig. S2, black arrow). In the MLE subgroup, TuLIP combines the MLE-syn and Chl-MLE families into one group (Sct30) (Supporting Information File 1, Fig. S2, purple arrow) and subdivides both the OSBS and Dipep families into multiple clusters (Supporting Information File 1, Fig. S2, orange arrows), similar

to the results observed for the DASP searches using ASPs built from the expert-identified groups (Fig. 4). Notably, in the most recent curation of SFLD, MLE-syn and Chl-MLE are no longer distinguished as two separate families, consistent with Sct30 identified by TuLIP.

Strong correspondence with SFLD subgroups and families suggests that TuLIP-identified clusters are functionally relevant; however, the number of proteins in the PDB is insufficient to perform meaningful statistical analysis. Previous work on the Prxs shows that DASP searches of GenBank using ASPs of known functionally relevant groups distinctly identify other members of those groups (Supporting Information File 1, Fig. S1).²⁵ Thus, ASPs constructed from proteins in each TuLIP-identified group were used to search GenBank, a significantly larger sequence database, to determine the quantitative correlation between TuLIP groups and SFLD annotations.

Table II. *DASP2 Sequences Identified, and SFLD Mapping, for Each TuLIP Group*

	TuLIP group	Known structures ^a	Total DASP2 hits ^b	Hits in SFLD ^c	SFLD subgroup	Hits in SFLD subgroup	SFLD family	Hits in SFLD family	% Coverage ^d	Hits not in SFLD
TuLIP groups mapped to SFLD Sub-groups and Families	Sct3	17	795	778	ManD	746	ManD ^e	575	91.3	17
	Sct5	18	7338	7175	Enolase	7142	Enolase ^e	7021	88.6	163
	Sct6	16	1779	1721	GlucD	1701	GlucD ^e	1473	99.1	58
	GalD-eng	1	10	10	GalD	3	GalD ^e	2	100	0
	Sct7	3	204	59	MAL	59	MAL ^e	37	84.1	145
	Sct13	8	380	369	MR	366	RhamD ^e	241	60.0	11
	Sct23	6	486	473	MR	467	LFucD ^e	134	27.5	13
							LTalGalD ^e	137	99.3	
	Sct27	3	38	38	MR	38	MR ^e	7	100	0
	Rlx39	4	57	57	MR	55	DGalnD ^e	52	6.9	0
	Rlx45	2	235	229	MR	228	DTartD ^e	135	100	6
	Sct20	4	688	672	MLE	672	OSBS ^e	670	30.2	16
	Sct22	7	598	589	MLE	588	NSAR ^e	9	100	9
							OSBS	161	7.3	
	Sct30	5	641	630	MLE	630	MLEsyn ^e	594	87.5	11
							Chl-MLE ^e	32	76.2	
							DipepEp	1	0.05	
	Sct31	4	66	60	MLE	60	DipepEp ^e	57	2.6	6
							OSBS	1	0.05	
	Rlx48	2	1025	953	MLE	953	DipepEp ^e	951	43.3	72
	Rlx50	6	572	555	MLE	555	DipepEp ^e	516	23.5	17
							OSBS	4	0.2	
	NSAR2-eng	1	133	129	MLE	129	NSAR2 ^e	128	93.4	4
MLEanti-eng	1	58	56	MLE	56	MLEanti ^e	47	90.4	2	
						OSBS	4	0.2		
TuLIP groups not mapped to SFLD	Sct10	3	96	93	MR	93	N/A	0	N/A	3
	Sct14	5	56	55	MR	55	N/A	0	N/A	1
	Sct15	4	195	173	MR	169	N/A	0	N/A	22
	Sct16	3	34	34	MR	7	N/A	0	N/A	0
	Sct25	5	139	139	MR	69	N/A	0	N/A	0
	Rlx36	2	9	9	MR	9	N/A	0	N/A	0
	Rlx44	3	27	27	MR	27	N/A	0	N/A	0
	Rlx42	7	78	61	MLE	58	OSBS	4	0.2	17
							DipepEp	2	0.1	
	Total	140	15737	15144						

^a Known structures is the count of nonredundant structures in the 2014 SFLD, based on 95% full sequence identity.

^b All search results for each group, given the trusted DASP score ($\leq 1e-12$).

^c Hits are defined as the number of proteins identified by DASP2 that are in the SFLD enolase superfamily as of 2/18/14.

^d Percent of all SFLD-identified proteins that are identified by DASP2 search using TuLIP-identified clusters.

^e Identifies an SFLD family mapped to TuLIP group.

Proteins identified by DASP2 searches of GenBank using TuLIP-based profiles correlate with SFLD subgroups and families

ASPs were created for each of the 23 TuLIP-identified groups and used in DASP2 searches of the GenBankNR sequence database (ASPs provided in Supporting Information File 3). For purposes of comparison to SFLD, engineered ASPs (see Methods and ²⁵) were constructed for three singlet proteins. These singlets are the only nonredundant structural representatives of their respective families (GalD, MLEanti, and NSAR2); those engineered ASPs were also used to search GenBankNR (GalD-eng, MLEanti-eng, and NSAR2-eng). The search tool DASP2 allowed us to identify additional proteins with functional site features similar to those in each TuLIP group.

These 26 GenBank searches (one for each of the 23 TuLIP-identified groups and each of the three engineered singlet ASPs) identified 15,737 proteins at the trusted DASP score threshold of $1e-12$ (Table II), a 50-fold increase over the coverage in the PDB database. (All sequences identified at DASP2 score threshold of $\leq 1e-8$ are provided in Supporting Information File 4.) Comparison of each DASP2-identified protein to its respective SFLD classification indicates that most of the DASP2-identified proteins are annotated as enolase superfamily members: of the 15,737 proteins identified, 96% were present in February 2014 SFLD (15,144 sequences) and 4% were not (593 sequences; Table II and Supporting Information File 4), representing a modest false positive rate of less than 4%. Membership of these 593 sequences in the enolase

superfamily is inconclusive, though, so the 4% represents a maximum false positive rate.

The extent to which each DASP2 search covered individual SFLD subgroups and families was also evaluated (Supporting Information File 1, Fig. S3). At a score threshold of $1e-12$, the DASP2 searches identified 100% of the GalD subgroup, the MR family, the NSAR family, and the DTartD family. Over 65% of the SFLD sequences were identified in most remaining subgroups and families (Supporting Information File 1, Fig. S3). Coverage of the RhamD and LFucD families (MR subgroup) is less complete—60.0% and 27.5% of these SFLD families, respectively, were identified. Family coverage in the MLE subgroup ranges from 37% of the OSBS family to 100% of the NSAR family, with coverage in six of seven families over 65%. Median percent coverage of all families was 94.52% and 88.03% at threshold scores of $\leq 1e-8$ and $\leq 1e-12$, respectively.

TuLIP groups were mapped to SFLD subgroups and families (Table II) in detail by identifying the percent agreement between SFLD annotations for each protein and each TuLIP group (heat map visualization, Fig. 6). In the results section of Supporting Information File 1, we describe in detail the qualitative comparison between the TuLIP-identified groups and the SFLD subgroups and families.

These comparisons of DASP2-identified enolases in GenBank to SFLD-identified members of the enolase superfamily demonstrate that the ASP-based automatable and SFLD knowledge-based approaches to functionally relevant clustering track well with each other. One limitation of the ASP-based approach is that if a family or subfamily is not represented among known structures, it will not be represented in the TuLIP clusters.

TuLIP groups are distinct, with little overlap at a DASP score threshold of $1e-13$

DASP2 GenBank searches increased coverage of sequences in the enolase superfamily almost 50-fold from the limited representation in the structure database. The search quality depends on whether or not a single molecular function is identified for most groups (unless there is biological justification, such as a sequence with more than one molecular functional site). Network-based plots that allow visualization of cross-hits between the groups (Supporting Information File 1, Fig. S4, black edges) show, as expected, more cross-hits are observed at a DASP2 score threshold of $1e-8$ than at the more significant score, $1e-12$; cross-hits are mostly between MR and MLE families (Supporting Information File 1, Fig. S4, blue and red nodes). Unsurprisingly, some of these same subgroups were also challenging for expert SFLD curators.^{31–33}

We have previously shown that a DASP score threshold of $1e-10$ is sufficient to distinguish the six

subgroups in the Prx superfamily²⁵ and, thus, was deemed a “trusted” DASP score threshold for that superfamily. In this work using DASP2, the trusted score threshold was more significant, $1e-12$. At this threshold, only three cross-hits were observed among 26 functionally relevant clusters containing over 15,000 enolase superfamily sequences (0.02%) [Fig. 7(A); Supporting Information File 1, Fig. S4(B)]. All three cross-hits were identified by DASP2 searches Rlx48 and Rlx50, two of the three TuLIP groups mapping to the SFLD-annotated DipepEp family [Supporting Information File 1, Fig. S4(B), blue nodes]. At a DASP2 score of $1e-13$, no cross-hits are observed between groups, while at the less stringent threshold of $1e-11$, 17 cross-hits are observed [Fig. 7(A)]. Overall, these results indicate that DASP2 can differentiate molecular functional groups based on functional site features.

Quantitative analysis demonstrates good performance of DASP2 GenBank searches compared to SFLD annotations

To quantitatively compare the ASP-based identification of sequences to SFLD annotations, 18 TuLIP groups were mapped to SFLD subgroups and families using 16 mappings (Table II; Supporting Information File 1, Fig. S5) such that true and false positive and negative counts could be calculated (see Methods and Supporting Information File 5). True and false positive rates, precision-recall, and performance (combined measure of purity, edit distance, and VI distance, defined in Methods and Supporting Information File 5) were calculated for the combined results for the 16 mappings. The *F*-measure (an integrated measure of precision and recall) was also calculated individually for each of the 16 mappings.

A standard ROC curve for the mapped groups displays extremely high true positive rates and extremely low false positive rates [Supporting Information File 1, Fig. S5(A)], consistent with what was previously observed for the Prx superfamily (Supporting Information File 1, Fig. S1²⁵). A blow-up [Supporting Information File 1, Fig. S5(A), inset] shows the more typical ROC curve for the MLEsyn/ChlMLE, DipepEP, NSAR, OSBS, and LFucD/LTalGalD families—these are the groups discussed previously for being combined or subdivided by the TuLIP process. In addition, precision-recall curves demonstrate the high sensitivity and specificity of the searches, with the exception of Sct22 [Supporting Information File 1, Fig. S5(B)]. Sct22 contains the nine (100%) NSAR proteins, as well as 161 (7.1%) OSBS proteins. This TuLIP group was mapped to the NSAR family; thus, all OSBS proteins identified are considered false positives for these calculations. Consequently, low specificity is observed for Sct22 in the precision-recall curves. Previous work shows some NSARs may catalyze OSBS activity as

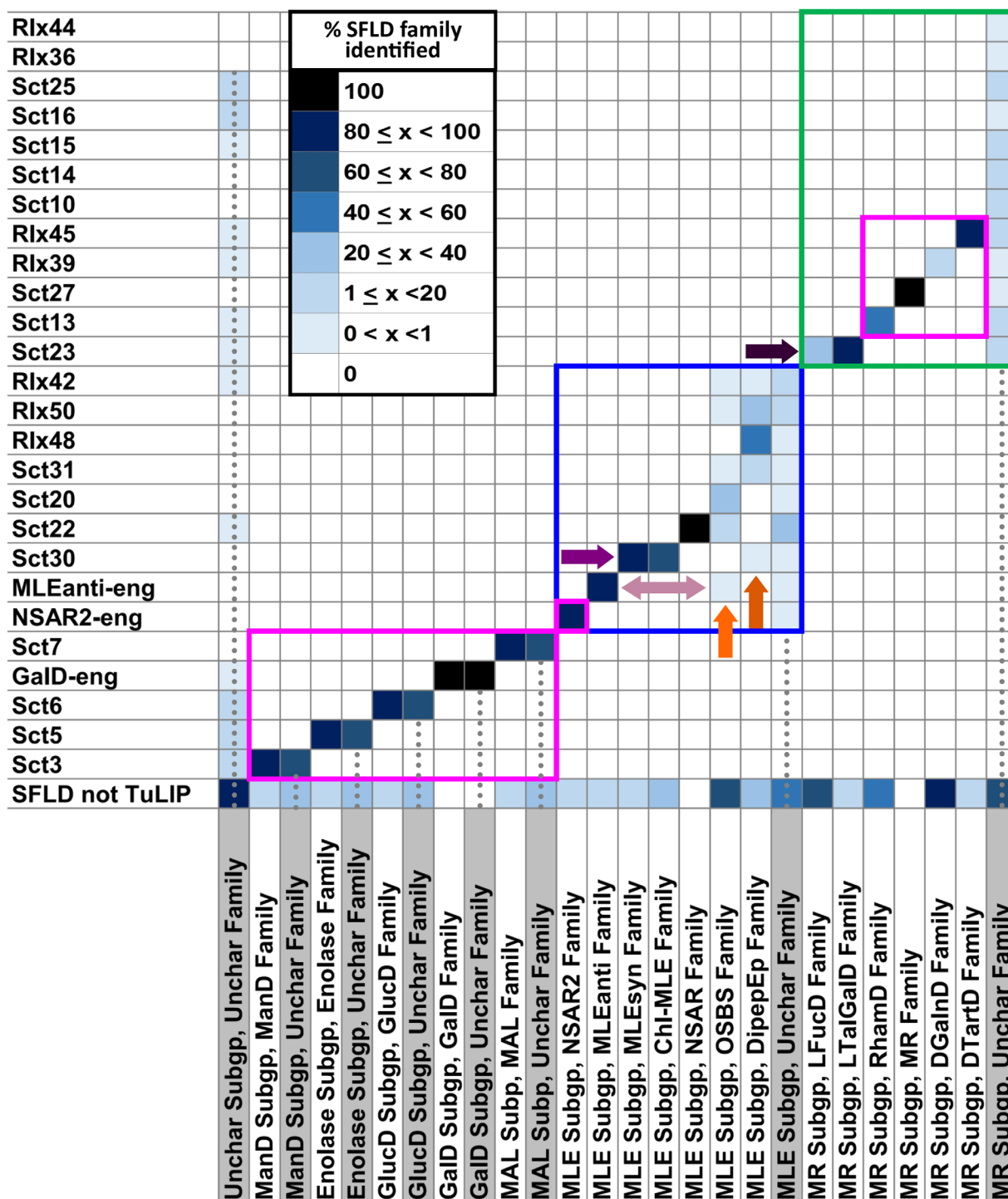


Figure 6. Comparison of DASP2-identified to SFLD-identified enolases in GenBank demonstrates a high correspondence between the automatable and knowledge-based approaches to functionally relevant clustering. In this heat map, columns represent the SFLD subgroup and family assignments for these proteins and rows represent the TuLIP group to which each individual protein sequence was assigned at a trusted DASP2 score threshold of $1e-12$: Sct indicates groups identified in the first TuLIP stage (strict criteria) and Rlx indicates groups identified in the second stage (relaxed criteria). Grid color represents the percent of proteins of a specific subgroup or family identified by each TuLIP group search, according to the legend. Gray-highlighted column labels and dotted lines indicate sequences labeled uncharacterized by SFLD. Pink boxes indicate TuLIP groups that map one-to-one to SFLD subgroups or families, the green box indicates all families in the MR subgroup, and the blue box indicates all families in the MLE subgroup. Black, purple, and orange arrows identify SFLD families that are either combined or subdivided by TuLIP, as discussed in text.

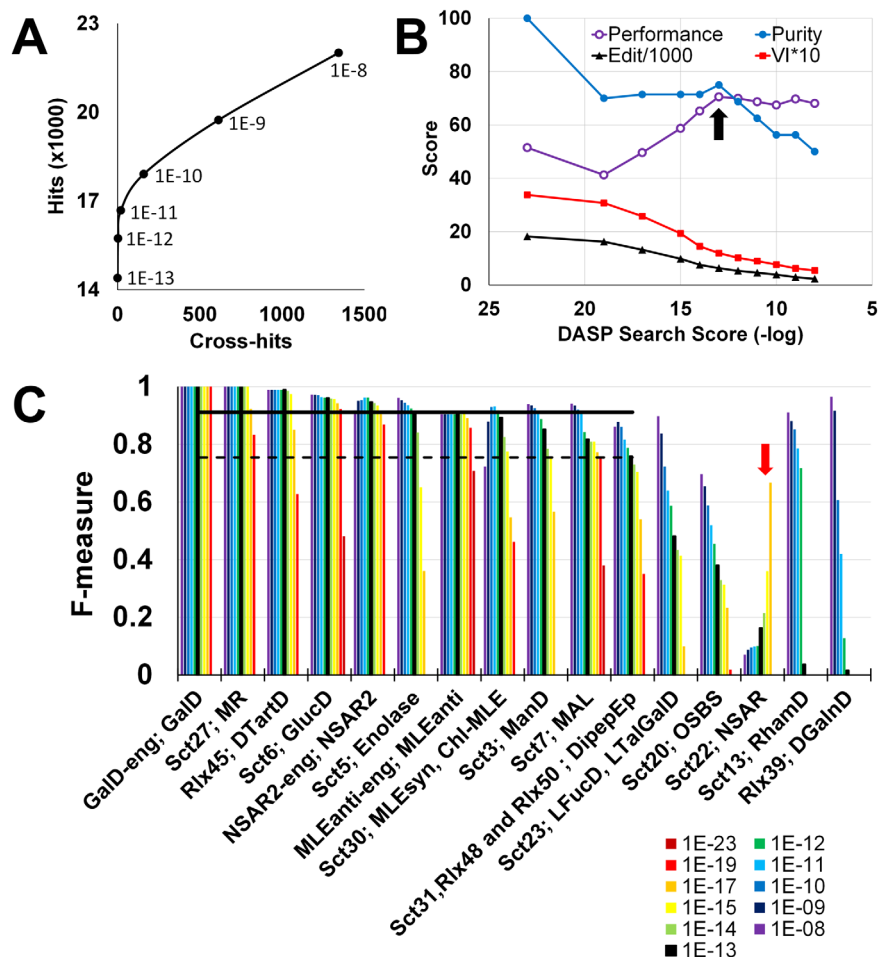


Figure 7. Quantitative analyses demonstrate the quality of DASP2 assignment of proteins to TuLIP-identified groups and support the optimal DASP2 score threshold. (A) Cross-hits (protein sequences identified by more than one DASP2 search of GenBank at the given DASP2 score threshold) were plotted against total hits for the 23 TuLIP groups and three engineered groups used to search GenBank. Only three cross-hits of 15,737 identified sequences (0.02%) were identified at a trusted DASP2 score threshold of 1e-12, and no cross-hits at 1e-13. (B) Performance, edit distance, purity and VI distance were calculated for the 16 TuLIP groups which could be mapped to SFLD subgroups or families. The performance curve, a composite of edit, purity, and VI distance, shows peak performance at a DASP2 score threshold of 1e-13 (black arrow). (C) The *F*-measure, a composite of precision and recall (see Methods) was plotted for each of the 16 TuLIP groups mapped to SFLD subgroups or families. Rainbow colored bars represent DASP2 score thresholds, according to the legend. Solid and dashed lines indicate the mean and two standard deviations of the 11 groups which *F*-measure is above 0.75 at 1e-13. The red arrow indicates Sct22, as discussed in the text.

bifunctional enzymes³⁶; thus, this result may not be a false positive, but may be biologically relevant.

To further evaluate the utility of the search results, purity, edit distance, and VI distance were calculated and combined into a performance measure based on the mapping of TuLIP groups to SFLD families. Optimal performance is observed at a DASP2 score of 1e-13 [Fig. 7(B), black arrow], which is more significant than the optimal performance score of 1e-10 previously identified for Prx subgroups.²⁵ 1e-13 is also the first DASP2 score threshold in which no cross-hits are observed [Fig. 7(A)].

If the search results correlated perfectly to SFLD, the performance score would be 100. The performance score of 70.5 calculated at 1e-13 represents

a high level of correlation to known functions with room for improvement. As a comparison, SCI-PHY¹⁹ and GeMMA²⁰ score 91.70 and 90.59, respectively, for the enolase superfamily; however, reported by these researchers, these performances are very dependent on the specific superfamily. Because both SCI-PHY and GeMMA start with the entire enolase superfamily of sequences and subdivide the superfamily, one would expect a high performance score. Both methods subdivide the superfamily well beyond SFLD subgroups and families; this finer subdivision is only weakly penalized in the performance score, as it was calculated. In contrast, DASP is agglomerative: it starts with a simple representation of each family (based on structures represented in the PDB), and then attempts to identify all other

members of the family from GenBank. The DASP agglomerative approach of family member identification from GenBank is more difficult, akin to how SFLD curators identify subgroups and families using sequence analysis, HMMs, and a high level of expert curation.^{17,18} Nonidentification of some superfamily members by DASP is more heavily penalized in the performance calculation than the over-division demonstrated by GeMMA and SCI-PHY. Results suggest that further iterative searches using ASPs from the DASP-identified sequences in each group identify more proteins belonging to these functionally relevant groups,²⁸ which would, thus, increase the performance score.

Utility of TuLIP clusters in identifying sequences from GenBank was also evaluated using the *F*-measure, a composite of precision and recall that identifies sensitivity and specificity at each score threshold. An *F*-measure value of 1 indicates perfect precision and perfect recall. At the DASP score threshold of 1e-13, the *F*-measure is greater than 0.9 for seven TuLIP groups (GalD-eng, Sct27, Rlx45, Sct6, NSAR2-eng, Sct5, and MLEanti-eng), between 0.75 and 0.9 for four TuLIP groups (Sct30, Sct3, Sct7, and the Sct31/Rlx48/Rlx50 DipepEp mapping), and below 0.75 in only five cases: Sct23, Sct20, Sct22, Sct13, and Rlx39, [Fig. 7(C)].

It is worthwhile exploring the reasons underlying the very different *F*-measures for the different functional groups, especially those that score more poorly. Sct23 is the TuLIP group that combined 99.3% of LTalGalD proteins and 27.5% of LFucD proteins. It is mapped to LTalGalD and LFucD (Table II); consequently, the large number of unidentified LFucD proteins count as false negatives. Sct20 is mapped to OSBS, as it is composed only of OSBS. As discussed above, the OSBS group is subdivided into three TuLIP groups and Sct20 only contains 30.2% of them, thus decreasing the *F*-measure because of many false negatives. Sct13 maps solely to RhamD, but identifies only 60.0% of these proteins in this single DASP2 search (Table II). Rlx39 contains 52 DGalnD proteins, which accounts for only 6.8% of the family annotated in SFLD. Notably, DGalnD contains no structurally characterized proteins, so it is unsurprising most of the family is not identified. These four TuLIP groups, which all exhibit an *F*-measure below 0.75, are comprised of families which were combined or subdivided by TuLIP. Again, an iterative GenBank search process is an essential process to identify additional sequences and, ideally, subdivide clusters in functionally relevant ways.

Notably, the *F*-measure mapped across DASP2 score thresholds demonstrates a different behavior for Sct22 [Fig. 7(C), red arrow]. The largest *F*-measure is observed at a DASP2 score of 1e-17 and decreases as the DASP score threshold either

increases or decreases. As mentioned above, this group is mapped to the NSARs. At 1e-17, the recall is perfect—all NSARs are identified. As the DASP2 score threshold becomes less significant, more OSBS proteins are identified and these proteins count as false positives. On the other hand, at DASP2 scores more significant than 1e-17, NSAR proteins begin to be lost, thus resulting in false negatives and negatively impacting the *F*-measure. The DASP2 score 1e-17, at which both OSBS and NSAR proteins are identified, is more stringent than any trusted score threshold we have yet observed, which suggests that DASP2 is identifying the functional features described by the experimental work in which the NSAR protein from *G. kaustrophilus* was shown to efficiently catalyze the OSBS reaction.^{37,38}

Overall these quantitative analyses demonstrate the functionally relevant clusters identified by TuLIP correlate with a detailed level of molecular function at the SFLD subgroup or family level. High quality identification of GenBank enolase superfamily sequences is observed in most groups, despite the fact that these TuLIP groups were constructed from the limited coverage found in the structure database. The over-division demonstrated by GEMMA and SCI-PHY is avoided. These results establish the utility of using TuLIP to cluster proteins of known structure into groups that are both functionally relevant and discrete, followed by using DASP2 to identify members of these functionally relevant groups from large sequence databases.

Enhancing detailed molecular annotations in GenBank

Annotation at the level of molecular functional detail without sufficient evidence is a source of mis-annotation (or over-annotation) in large sequence databases.⁶ The ASP-based approach presented here clusters proteins based on molecular functional detail, and therefore can add new, accurate sequence annotations. Thus, we compared GenBank annotations with the detailed molecular annotations provided by the current process (Table III).

Between 0% (NSAR2) and 90-95% (MLEsyn, Chl-MLE, and OSBS) of the GenBank annotations are accurate to the level of detail provided by these DASP2 searches. Between 0% (MLEanti and DGalnD) and 20% (GalD) of the sequences have no annotation in GenBank. The place where this process makes the largest contribution is in adding detail to the more vague GenBank annotations, where contributions range from 0% (MAL and GalD) to 89% (DTartD) of the sequences, and correcting incorrect or over-annotated sequences, which account for between 0% (MAL) and 70% (NSAR2 and GalD).

Table III. GenBank Annotations of Proteins Identified by Searching with TuLIP-Identified Groups

TuLIP group ^a	SFLD family mapping	GenBank annotations that match SFLD family (%)	GenBank annotation is correct but vague ^b (%)	GenBank incorrect annotation ^c (%)	No annotation in GenBank (%)
Sct3	ManD	2.9	68.4	24.3	4.4
Sct5	Enolase	80.3	15.9	0.1	3.7
Sct6	GlucD	82.3	4.0	10.0	3.7
GalD-eng	GalD	10.0	0.0	70.0	20.0
Sct7	MAL	83.3	0.0	0.0	16.7
Sct13	RhamD	51.1	20.5	16.8	11.6
Sct27	MR	65.8	23.7	0.0	10.5
Rlx39	DGalnD	84.2	10.5	5.3	0.0
Rlx45	DTartD	0.4	88.9	3.0	7.7
Sct20	OSBS	93.5	4.1	0.6	1.9
Sct22	NSAR	21.1	8.7	63.0	7.2
Sct30	MLEsyn, Chl-MLE	93.1	1.2	3.1	2.5
Sct31, Rlx48, Rlx50	DipepEp	39.8	25.7	26.2	8.3
MLEanti-eng	MLEanti	37.9	3.4	58.6	0.0
NSAR2-eng	NSAR2	0.0	26.3	69.9	3.8

^a Sct23 maps to both LFucD and LTalGalD and is not included in this evaluation, as the relationship between these two families is not understood; Sct30 maps to both MLE-syn and Chl-MLE and both carry out the same reaction (on different substrates), thus, both included as a “match” for this evaluation.

^b Correct superfamily or subgroup, but not specified to family level or noncommittal modifiers to the family specification, such as ‘hypothetical ...’; and ‘... like’.

^c Wrong subgroup/family designation or nonenolase superfamily name.

The enolase family is very well-studied, so the amount of detailed, molecular functional information that would be added to GenBank is somewhat less than what was observed for the Prxs.²⁵ Results from both the enolase and Prx superfamilies demonstrate that significant new and more detailed molecular functional information would be added to GenBank, providing an enabling solution to the under-annotation and over-annotation problems.

Mechanistically important details can be elucidated from residue conservation in functionally relevant groups

Given the quality of the described approach, it is important to ask whether conserved residues identified in these clusters provide insight into important functionally relevant or mechanistically determinant residues. The ASP approach simplifies extraction of such well-conserved residues in the protein’s active sites. Do these conserved residues identify the molecular functional determinants in each group?

Mechanistically important residues for families in several superfamilies were previously hypothesized using this approach.^{16,25} In the Prx superfamily, the roles of identified residues were substantiated by molecular dynamics and electrostatics calculations,²⁹ then validated by experimental work (Nelson and Poole, unpublished results). Network-based clustering of ASPs allowed identification of mechanistically important residues in the ManD family¹⁶ that had previously been identified by experimental work.³⁹ Others have also suggested a similar approach on the carbohydrate kinases.⁴⁰ We here follow this approach to hypothesize mechanistic

determinants for the OSBS and DipepEp isofunctional groups.

TuLIP divides the SFLD OSBS family into two main groups: Sct20 and Sct22 (Table II; Fig. 6, orange arrow). Sct20 is composed of 670 OSBS proteins and two MLE-uncharacterized proteins. Sct22 is composed of 161 OSBS proteins, all nine NSAR proteins (Table II), and 419 uncharacterized MLE subgroup proteins. Sct20 and Sct22 contain distinguishing features at their active sites that can be compared to the features in the SFLD-identified OSBS and NSAR families (Fig. 8). Six particular features distinguish Sct20 and Sct22 (Fig. 8, peach shading). Most obvious is signature position 26 (residue 177 in 2QVH and 237 in 1SJA), which is an Ala in Sct20 [Fig. 8(A)] and a Cys in Sct22 [Fig. 8(C)]. The analogous position in the NSAR family is invariant as a Cys [Fig. 8(D)], while both Cys and Ala are found in the OSBS proteins [Fig. 8(B)]. The second most obvious distinguishing feature is the invariant Val and Asn in signature position 36 of Sct20 and Sct22, respectively (residue 201 in 2QVH and 261 in 1SJA). Asn is invariant in the NSAR proteins [Fig. 8(D)], while both residue types are observed at this position in the OSBS proteins [Fig. 8(B)]. Signature position 4 (residue 100 in 2QVH and 164 in 1SJA) also distinguishes Sct20 (Val) and Sct22 (Ile). Distinguishing features are also found at signature positions 7, 12, and 13 (Fig. 8). These features are oriented towards the substrate binding site [Fig. 8(E)], allowing us to hypothesize functional relevance. Similar NSAR and OSBS clustering was reported using full sequence-, full structure-, and active site-based analyses.¹⁶ The current results,

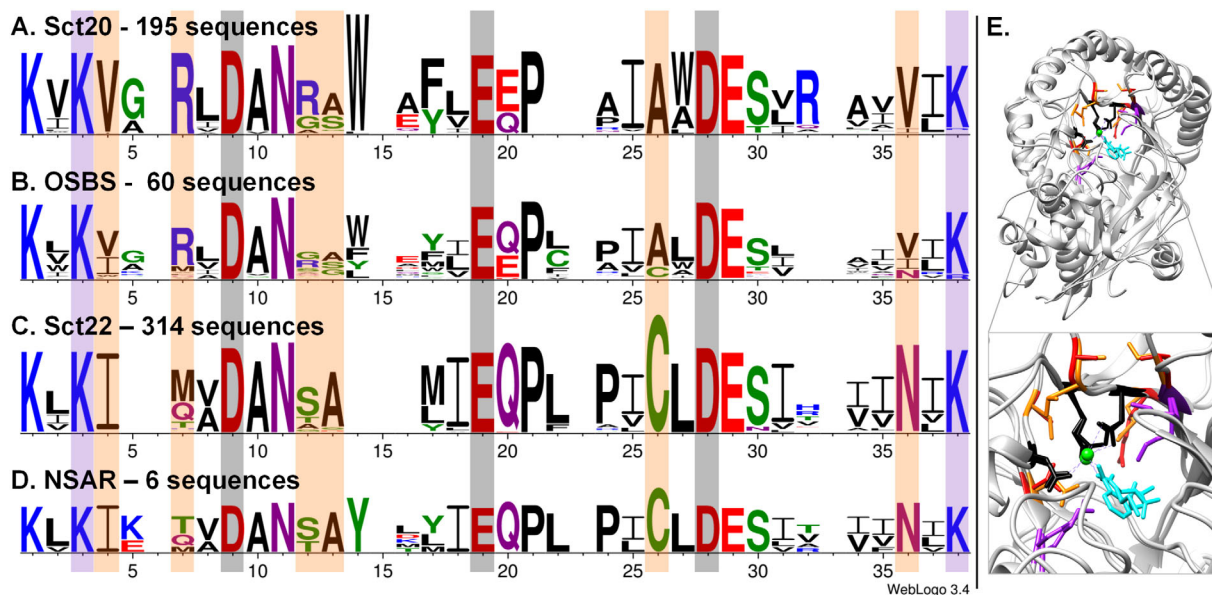


Figure 8. Conserved residues in the active sites of Sct20, Sct22, OSBS, and NSAR illustrate the functional site features that distinguish Sct20 and Sct22 from the SFLD families. Weblogos⁵⁵ were generated for TuLIP groups Sct20 (A) and Sct22 (C) and for the structurally analogous positions in SFLD-identified OSBS (B) and NSAR (D) families. Gray highlights indicate metal ligand binding residues conserved across these proteins used for active site profile construction. Purple highlights are lysine residues that SFLD designates as residues important for proton abstraction. Peach highlights illustrate significant differences between Sct22 and Sct20 that are discussed in the text. Representative structures are overlaid (E) for 1SJA (Sct22), an OSBS family member co-crystallized with N-acetyl-methionine, and 2QVH, an OSBS with o-succinyl benzoate (both substrates displayed as teal; metal ion is green sphere). Key residue side chains are black, while conserved lysines are displayed as purple. Residues conserved in the ASPs are displayed as orange (Sct20) and red (Sct22) side chains.

combined with those results, suggest that a subset of the SFLD-identified OSBS proteins share more functional site similarities with the NSAR proteins than the other OSBS proteins.

The SFLD DipepEp family is divided by TuLIP into three main groups (Table II; Fig. 6, orange arrow): Sct31, containing 57 DipepEp proteins, one OSBS protein, and two MLE-uncharacterized proteins; Rlx48, containing 951 DipepEp proteins; and Rlx50, containing 516 Dipep proteins, four OSBS protein, and 35 MLE-uncharacterized proteins (Table II). As with Sct20 and Sct22, these TuLIP-identified DipepEp groups can be distinguished by features at their active sites (Fig. 9, peach shading). The three most distinguishing positions are signature positions 4, 26, and 32 (V162, A239 and R245 in 2ZAD; L152, C223, and H229 in 1JPD; V163, M242, and F248 in 1TKK). At position 4, Val is found in Sct31 and Rlx50 [Fig. 9(B,D)], while Leu is found in Rlx48 [Fig. 9(C)]. At position 26, Ala, Cys, and either Met or Leu are found in Sct31, Rlx48, and Rlx50, respectively. Finally, Arg, His, and Phe are almost invariant at signature position 32 in Sct31, Rlx48, and Rlx50, respectively. Distinguishing features are also found at signature positions 7, 12, 31, and 34 [Fig. 9(B–D)]; these features are oriented towards the substrate binding site as well [Fig. 9(E)]. The residue conservation analysis between the TuLIP-identified DipepEp groups suggests three

distinct groups can be identified in this family, each with distinct functional site features.

TuLIP process applied to the glutathione transferase superfamily

TuLIP was also applied to the cytosolic glutathione transferase (GST) protein superfamily, another large superfamily in which members play multiple and important roles in metabolism and detoxification in eukaryotic organisms.^{41–43} Like the enolases, the GSTs are ubiquitous and diverse. The GST proteins are members of the thioredoxin fold family.⁴⁴ Their functions have traditionally been organized by Greek letters that were assigned as new superfamily members were identified; however, more recently the Enzyme Function Initiative has focused computational and experimental work on GST functional analysis.¹⁵ Comprehensive sequence and structure analysis has allowed mapping of these traditional functional groups onto the superfamily,⁴⁵ and docking has identified potential substrates across this superfamily.⁴⁶ Here we compare the TuLIP classification of GSTs of known structure to both SFLD and to the traditional classifications as reported in Swiss-Prot.

Using TuLIP, 155 nonredundant GST structures identified from SFLD were clustered into 24 groups and 24 singlets; these groups were compared to the level 2 subgroup classifications in SFLD. The

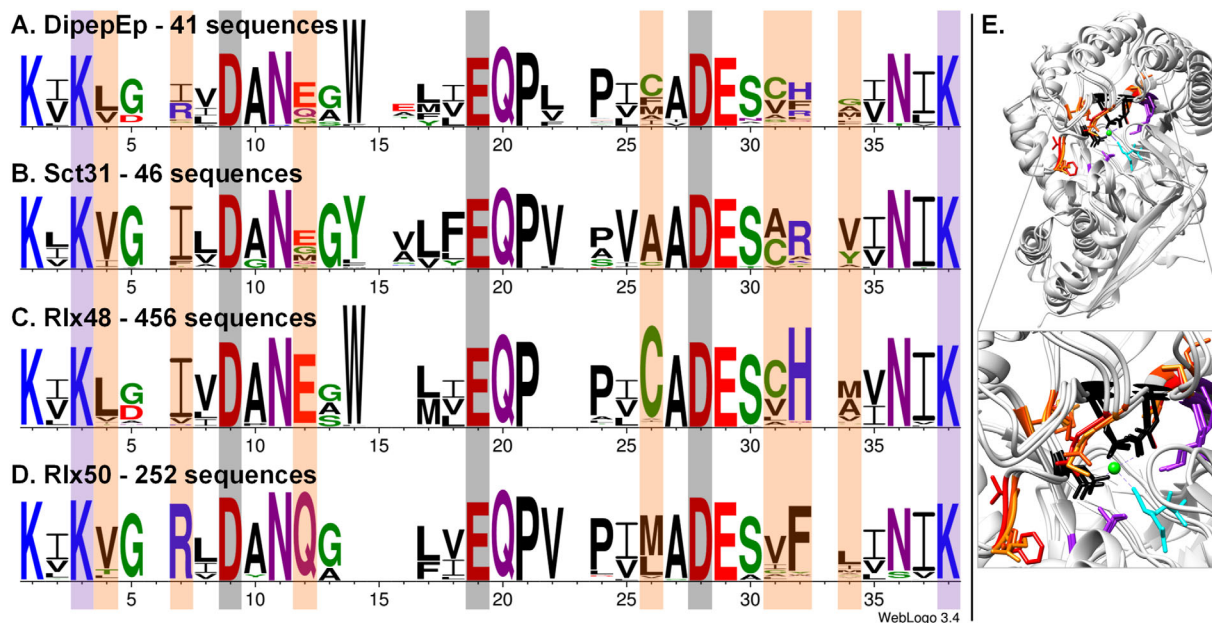


Figure 9. Conserved residues in the active sites of Sct31, Rlx48, Rlx50, and DipepEp illustrate the functional site features that distinguish TuLIP-identified subfamilies from the SFLD-identified DipepEp family. WebLogos⁵⁵ were created from the ASPs for Sct31 (B), Rlx48 (C), and Rlx50 (D) and for the structurally analogous residues in the SFLD-identified DipepEp (A) family. Gray highlights indicate metal ligand binding residues that were used for active site profile construction. Purple highlights are lysine residues that SFLD designates as residues important for proton abstraction. Peach highlights illustrate significant differences between Sct31, Rlx48, and Rlx50 that are discussed in the text. Representative structures are overlaid (E) for 3DER (Sct31), a DipepEp family member co-crystallized with alanine-lysine, and 4GFI (Rlx48) and 3R1Z (Rlx50), both DipepEp with alanine-glutamate. Substrates are displayed as teal; metal ion as a green sphere. Key residue side chains are black, while conserved Lys are displayed as purple. Residues conserved in the ASPs are displayed as orange (Sct31), orange-red (Rlx48) and red (Rlx50) side chains.

numbers of structures are insufficient for quantitative analysis; however, qualitative analysis is informative. For the GST superfamily, SFLD identifies two levels of subgroups, which we identify here as level 1 and level 2. In their representative networks of the GST superfamily, level 2 subgroups are defined by Mashiyama and coworkers as distinct clusters in a sequence similarity network where the edge threshold is more significant than a BLAST *E*-value of $1e-25$.⁴⁵

The overall correspondence between TuLIP group and SFLD level 2 subgroup is striking [Fig. 10(A)]. There is a one-to-one correlation between TuLIP groups and SFLD subgroups for eight TuLIP groups [Sct3, Sct8, Sct9, Sct13, Sct14, Sct17, Rlx42, and Rlx43; Fig. 10(A), fuchsia box]. Some SFLD level 2 subgroups are combined by TuLIP [Fig. 10(A), purple arrows]: Sct4 (Xi and Main.27); Sct6 (Main.4 and Main.22); Rlx40 (Main.6 and Main.7; includes one protein from Main.1); and Sct12 (Main.19, Main.9, and ProstE). Combination of some subgroups is likely because of the limited structural representation of the superfamily, preventing a finer level of functional distinction. Subsequent GenBank searches to identify proteins that share active site similarity with each group will increase group

membership significantly, thus, allowing for potential subdivision based on functional characteristics.

Three SFLD level 2 subgroups are subdivided by TuLIP [Fig. 10(A), orange arrows]: Main.1 (Sct16, Sct7, Rlx40); Main.3 (Sct15, Rlx41); and Main.2 (Sct12, Rlx59). Notably, the three Main groups subdivided by TuLIP are the three largest level 2 representative networks presented by Mashiyama and coworkers.⁴⁵ All three show two distinct subnetworks which contain a structural representative. Given that TuLIP clusters proteins of known structure, TuLIP is potentially identifying clusters similar to those previously published, though further investigation is required to determine the functional relevance of these subdivisions.

The AMPS subgroup is also subdivided by TuLIP into multiple clusters [Sct19, Sct21, Sct22, Rlx50, Rlx51, Rlx54, Rlx55; Fig. 10(A), orange bracket]. The AMPS subgroup contains the traditionally defined alpha, mu, pi, and sigma classes. Mammalian cytosolic GSTs were originally classified into alpha, mu, and pi classes^{47,48}; sigma was later identified to be a glutathione-dependent prostaglandin D2 synthase.⁴⁸ In Swiss-Prot, proteins in these four traditional classes are identified and we compare these traditional classes to the TuLIP-identified

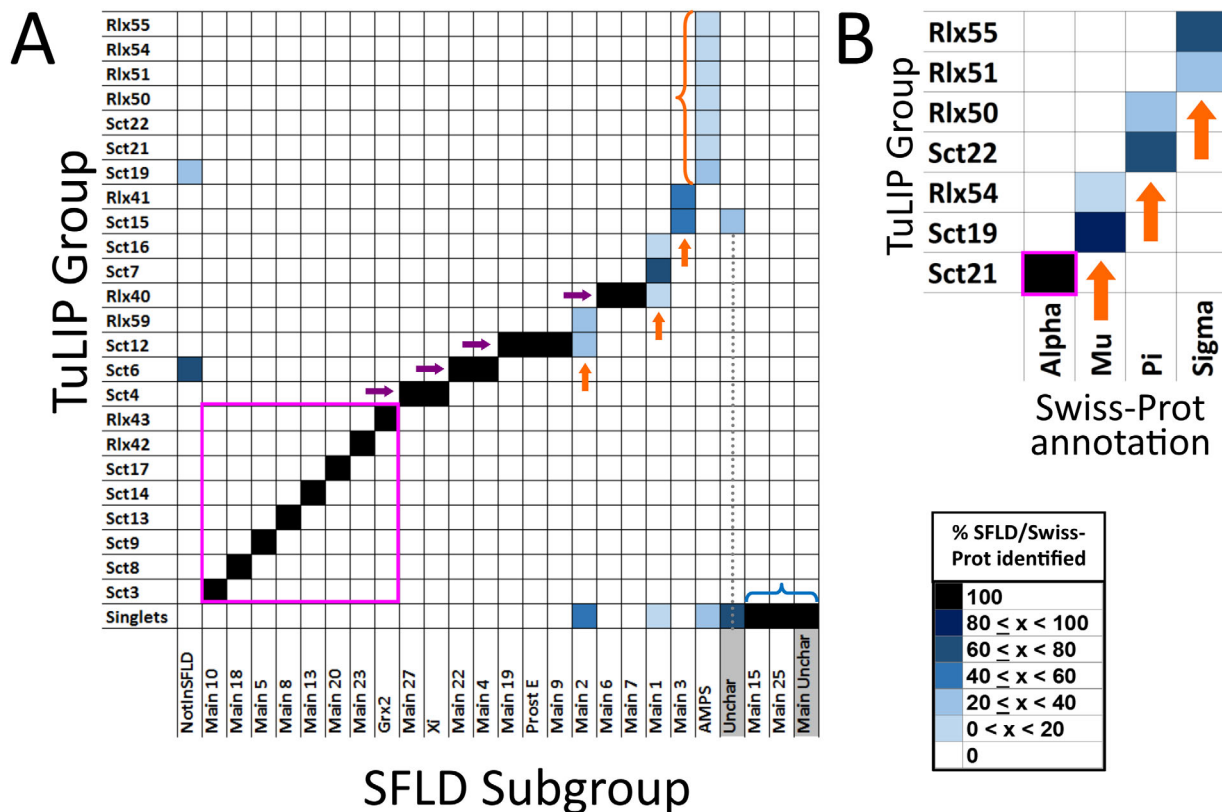


Figure 10. TuLIP clusters the GST superfamily in a functionally relevant manner. Heat maps show the comparison of the TuLIP-identified groups with the SFLD-identified subgroups in the GST superfamily (A) and the traditional alpha, mu, pi, and sigma classifications, as annotated in Swiss-Prot (B). Columns represent the SFLD (A) or SwissProt (B) annotations and rows represent the TuLIP group to which each protein structure was clustered. Sct and Rlx (*y*-axis) indicate groups identified in the first (strict criteria) and second (relaxed criteria) stages. Grid color represents the percent of proteins of a specific subgroup or family identified in each TuLIP group, according to the legend. Gray-highlighted column labels and dotted lines indicate sequences labeled as uncharacterized by SFLD. Pink boxes indicate TuLIP groups that map one-to-one to SFLD subgroups or Swiss-Prot annotations. Purple and orange arrows identify SFLD subgroups or Swiss-Prot annotations either combined or subdivided, respectively, by TuLIP. The orange bracket indicates the AMPS subgroup that is subdivided by TuLIP.

clusters [Fig. 10(B)]. The correspondence between the seven TuLIP AMPS groups and the four Swiss-Prot AMPS classes is significant: TuLIP does not combine any of the four classes but does subdivide the mu, pi, and sigma classes [Fig. 10(B), orange arrows]. These results demonstrate that TuLIP can identify these traditional functionally distinct classes. Whether the subdivisions of the mu, pi, and sigma classes is biologically relevant remains to be determined.

These results show some correspondence to the protein similarity networks that used sequence similarity as the edge metric. These representative networks identified at least six subnetworks that might be distinguished in the AMPS level 2 network.⁴⁵ These researchers mapped solved structures to their level 2 networks and showed that both alpha and mu structures mapped to two different subnetworks. TuLIP is possibly recognizing these subnetworks.

Importantly, TuLIP replicates what was previously shown with similarity-based networks using

pairwise active site signature scores as the edge metric. These similarity networks divided the AMPS subgroup into four clusters: one cluster corresponded to alpha, one cluster to mu, and one cluster combined pi and sigma (the last cluster was too small for meaningful analysis).¹⁶ Thus, the TuLIP process can replicate what was observed by manually curated similarity networks.

These GST results demonstrate the generalizability of TuLIP. Remarkably, no superfamily-specific thresholds are required to complete TuLIP clustering—all parameters used for TuLIP clustering of the GSTs were the same used for clustering of the enolases. The only expert work required is the identification of key functional residues from known structural representatives. This generalizability lays the foundation for clustering of superfamilies for which little data are previously known.

Conclusions

The main contributions of this work are the development of an automatable approach to divide the

proteins of known structure within a superfamily into functionally relevant clusters that correlate with known functional families. In addition, we show that these clusters can be used to search GenBank to identify other members of the functional family, an agglomerative approach to identifying iso-functional clusters. In contrast to most other methods of functional classification which require *a priori* knowledge of all superfamily members and then divisively cluster them, this approach is agglomerative and does not require *a priori* knowledge of all superfamily members. Thus, newly deposited family members can easily be identified. The method pinpoints mechanistic determinants that distinguish each functional family, which provides insight into distinguishing mechanistic details.

Methods

Protein superfamilies

Validation was performed by comparing the TuLIP-identified clusters to subgroups and families in the SFLD (<http://sflld.rbvi.ucsf.edu>), a manually curated, hierarchical scheme for classification of molecular function.¹⁷ The top level of the SFLD hierarchy is the superfamily, a group of proteins which are evolutionarily related and share at least a partial step in enzyme mechanism. Superfamilies are split into discrete subgroups (level 1, and more specifically, level 2) which are further divided into one or more families. A family is composed of enzymes which share a complete mechanism and, thus, represents fine detail of molecular function classification.¹⁷ Because of this detailed molecular function-specific hierarchy, the SFLD annotations have been used as a “gold standard” by other functional clustering and annotation methods.^{6,20}

One SFLD superfamily in particular, the enolase superfamily, was used to develop and initially validate the clustering method presented here. A summary of enolase superfamily subgroups and families is provided in Table I. To assess generalizability, the glutathione transferase (GST) superfamily was also utilized. Structures comprising the subgroups and families within each superfamily are provided in Supporting Information File 2.

Structurally characterized members of each superfamily were downloaded from SFLD. To avoid bias introduced by over-representation, redundancy within each superfamily was reduced by clustering proteins with either 99% (initial validation) or 95% (TuLIP) or greater sequence identity and selecting one representative from each cluster (C++ script written in-house). Representatives used in this work are provided in Supporting Information File 2.

Active site profiling: a method to identify feature similarity at protein functional sites

Active site profiling, originally described by Cammer and coworkers²² and represented in Figure 1, is a method to capture the features of a functional site. Briefly, key residues (listed in Supporting Information File 2 for the superfamilies discussed herein), which are involved in catalytic function and are structurally analogous across the superfamily, are identified by analysis of literature, database annotations such as Catalytic Site Atlas⁴⁹ and SFLD,^{17,18} and structure conservation. This key residue selection is the only step of the process that requires manual intervention and is prior to the beginning of the TuLIP process. All residues whose center of geometry lies within 10 Å of the center of geometry of any key residue are identified [colored ribbons, Fig. 1(A)]. DASP creates an *active site signature* for the functional site by extracting and aligning (in N- to C-terminal order) the sequence fragments (length at least three residues) composed of the residues within the 10 Å sphere [Fig. 1(B)]. Previous work has shown that these signatures contain all or most of the functionally relevant features from the active site.²²

For multiple proteins containing related functional sites, the signatures are aligned to create an ASP [Fig. 1(B); enolase superfamily TuLIP profiles are provided in Supporting Information File 3]. An *ASP score* is then calculated,²² which takes into account sequence identity and similarity and includes a negative contribution for gaps. ASP scores can range from 1.0, for a profile in which all signatures are identical, to 0 and negative numbers for profiles composed of signatures that are highly dissimilar or poorly aligned.

DASP and DASP2: tools which utilize ASPs for searching sequences

ASPs can be used to search sequence databases for proteins containing sites with features similar to those in the ASP using DASP (Deacon Active Site Profiler), first described by Huff et al.^{24,30} and described in more detail in the Supporting Information of Nelson et al.²⁵ The process is outlined in Figure 2. Briefly, a profile is broken into each of its component fragments, the fragments are aligned into motifs, and a position-specific scoring matrix (PSSM) is calculated for each motif by iterating over the columns of the alignment and tallying the observed counts (the number of occurrences of each residue) and the pseudocounts (based on the overall frequency of the amino acid in the background database) in each column. Each motif-specific PSSM is applied in a sliding window procedure across each sequence in the database. At each position, a *p*-value is calculated for the alignment of the PSSM to that

protein fragment, which represents the probability of finding a match as good as the observed match in a random position of a random sequence. The protein fragment which exhibits the most significant p -value is the “best match” to the PSSM. Individual p -values for the best match of each motif PSSM to a given protein sequence are combined using QFAST.⁵⁰ This combined p -value, the *DASP score*, represents the statistical probability of the sequence containing fragments that exhibit features similar to those found in the input profile.

DASP scores are calculated for all sequences in the database (in the work described here, either sequences of proteins of known structure from the PDB or all sequences in GenBankNR). Previous work²⁵ and unpublished data on searches of sequences in the PDB suggest a *trusted* DASP score threshold of $1e-10$, where searches identify only those sequences used to construct the profile. A *generous* DASP score threshold of $1e-8$ includes more family members related to the profile used for searching, but also can include a small number of false positives (<1.3% in previous work). In this work, a trusted cutoff of $1e-10$ was used for PDB searches and $1e-12$ for GenBank searches, in which the number of false positives is <0.02%.

The DASP tool is publicly available.²⁶ For much of the current work we used a modified version, DASP2, which runs on the Resource for Biocomputing, Visualization, and Informatics; (RBVI) Linux cluster at UCSF. This version implements more efficient searching and, thus, a decreased search time.³⁵ Additionally, the PDB and GenBank databases are updated weekly on the RBVI cluster, so searches are consistently run on the most up-to-date databases. Finally, modified amino acids (e.g. selenomethionine) are labeled as their more common counterpart rather than Xs in the signatures, an important modification for some superfamilies (such as the Prxs). These differences between DASP and DASP2 do not significantly affect search results, but allow us to complete multiple searches more efficiently.³⁵

TuLIP: an objective and automatable approach to functionally relevant clustering

TuLIP is an iterative, divisive clustering process (Fig. 5) that utilizes active site profiling to separate the structurally characterized members of a superfamily into clusters based on functional site features. Active site profiling identifies features at protein functional sites (Fig. 1), thus TuLIP-identified clusters are hypothesized as functionally relevant clusters.

Key residue identification and active site signature calculation (using DASP) are input into TuLIP. In the first step in TuLIP, a pairwise ASP score is calculated for each pair of functional site signatures in the group; this complete set of pairwise scores

comprises an all-by-all network in which nodes and edges represent proteins and pairwise ASP scores, respectively [Fig. 5(A)]. Starting with an ASP score of 0.0, the ASP score threshold is iteratively increased by 0.05, removing edges lower than the threshold. At each threshold, discrete subnetworks are identified by clustering using the MCL method.³⁴ These candidate subnetworks contain proteins in which the features of their respective functional sites are more similar within the subnetwork than to proteins outside of the subnetwork [Fig. 5(A)]. Each candidate subnetwork is evaluated as a *validated cluster* using strict or relaxed criteria, as described subsequently. Following this evaluation, remaining subnetworks are subject to further clustering by increasing the pairwise ASP score. Subnetworks are evaluated at each stage, until all proteins are assigned to a valid cluster or until all remaining subnetworks consist of only one protein.

This process of stepwise increase in the score threshold in MCL clustering is completed twice: strict and relaxed. Subnetworks are evaluated by a defined set of *strict criteria* in the first stage [dark gray background, Fig. 5(B)] and by a set of *relaxed criteria* in the second stage [light gray background, Fig. 5(B)]. At each score threshold, subnetworks are evaluated for self-identification in a DASP2 search, as follows. An ASP is created for all proteins within the subnetwork (Fig. 1). This profile is used to search the sequences in the PDB using DASP2 (Fig. 2). The distribution of output DASP2 scores is evaluated to determine if it meets the following *strict criteria* (Fig. 5, dark gray background): (1) all input proteins are identified at a DASP score more significant than $1e-10$; (2) no noninput protein is identified at a DASP score more significant than $1e-10$; (3) two orders of magnitude separate input proteins from noninput proteins; and (4) minimum cluster size is three proteins. (Note: all strict criteria are objective and none require human interpretation.) If the DASP2 search results using the subnetwork profile meet these strict criteria, the subnetwork is considered *validated* and, thus, functionally relevant [Fig. 5(B), purple parallelogram]; its proteins are removed from any subsequent iterative clustering. Subnetworks of only one or two proteins (singlets and doublets) are reserved for the *relaxed* stage of the process.

Proteins from subnetworks that do not pass strict criteria continue to the next iteration. If new, non-SFLD proteins are identified at a DASP2 search score more significant than $1e-10$ (with all other strict criteria met) those proteins are added to the input set (given they are less than 95% identical to other inputs). In the next iteration, the score threshold is increased by 0.05. If any resulting subnetworks pass strict criteria, the subnetwork, including the new proteins, is considered functionally relevant;

if any resulting subnetworks do not pass strict criteria, the subnetwork remains in the full network and the new proteins are included in the next clustering iteration.

When all validated clusters have been identified by the strict criteria, the remaining proteins not assigned to a validated or functionally relevant cluster progress through a second stage of iterative clustering. In this second stage, functionally relevant groups are identified by *relaxed* criteria [Fig. 5(B), light gray background]. As in the strict stage, all pairwise ASP scores between these proteins are calculated to form a complete network and the pairwise ASP score threshold is iteratively increased by 0.05 and proteins clustered using MCL clustering.³⁴

At each iteration, subnetworks are identified and their ASPs are created and used in a DASP2 search of PDB sequences. The ASP score distributions are analyzed for meeting the following *relaxed criteria* for self-identification: (1) all input proteins are identified at a DASP score more significant than 1e-8; (2) no noninput proteins identified at a DASP score more significant than 1e-12; (3) two orders of magnitude separate input proteins from noninput proteins; and (4) minimum subnetwork size is two proteins. (Note, again, all criteria are objective; none require human interpretation.) Subnetworks meeting these criteria are considered *validated* as functionally relevant clusters and are removed from subsequent iterative clustering [Fig. 5(B), purple parallelogram].

When the network contains only singlet proteins [Fig. 5(C), white background], the TuLIP process is complete [Fig. 5(B), red circle “Done”]. An ASP is created for each functionally relevant cluster identified in both strict and relaxed stages and each profile is used to search GenBank.

Profile engineering to produce profiles for the singlet clusters

Singlets, which cannot be used to create a profile for a DASP2 GenBank search, are retained for a “profile engineering” process. In this work, we engineered profiles for three enolase superfamily families that contained just one nonredundant structure each: MLE-anti, NSAR2, and GalD. A modification of the profile engineering process previously described²⁵ was used. A small number of sequentially similar proteins are identified for each structure by performing a BLASTP⁵¹ search of GenBank using each singlet protein structure as the query sequence (3DG3, 2P88, and 3FYY, respectively). A procedure was developed to identify five sequences with high sequence identity to each query sequence, but enough diversity to create ASPs useful in a GenBank search. An initial list of similar sequences is created from the BLASTP search results in default order. Proteins with 100% sequence identity are not

used. To create a representative set of related proteins, the first protein listed at each percent identity down to 80% is added to the set. Sequences with less than 80% identity are added with no restrictions until the set contains twenty proteins. To ensure the profile is not too diverse, no sequences below 60% identity are included. From this initial set, five sequences are selected (for the work here, those proteins are 4th, 8th, 12th, 16th, and 20th in list order) to form the final set of sequences used in the engineered profile. Ideally, this set will contain proteins with similar functional sites to the query protein but enough diversity to identify all proteins in the family with a GenBank search.

To create the engineered ASP from these proteins, the six sequences (one PDB query sequence and five sequences selected from the BLASTP search) were aligned using ClustalW2.⁵² The active site signature motifs (at least four residues in length) were identified from the original PDB structure [Fig. 1(A,B)]; these motifs were used to extract the analogous motifs from each sequence in the multiple sequence alignment. The resulting “engineered profile” was used to search GenBank.

Evaluation of results of DASP2 search of the GenBank sequence database

ASPs of 23 TuLIP-identified enolase superfamily groups, as well as the three engineered profiles, were utilized to search the GenBank database using DASP2 (Fig. 2), as described above and previously.^{24,25} Search results were evaluated for true positives by comparison to SFLD, for distinctiveness by identifying cross-hits (where one sequence is identified in more than one search), as well as for precision, recall, *F*-measure,⁵³ purity, edit distance, VI distance,⁵⁴ and performance measures used by others in evaluation of protein function mapping.^{19,20} Details of each calculation are described in Supporting Information File 5.

To calculate these measures, we must first define the “positive” and “negative” sets of known sequences. The “universe” of sequences were those sequences in the SFLD-defined enolase superfamily as of February 18th, 2014, the same day the GenBank searches were completed. The goal is to demonstrate that the approach can identify molecular functional details; thus, we rigorously evaluated its ability to distinguish between members of an individual SFLD subgroup or family and all other members of the SFLD superfamily. Accordingly, the “positive” set included those proteins in a given SFLD subgroup or family, and the “negative” set was composed of all other sequences in the SFLD-defined enolase superfamily not part of the subgroup or family being considered. (Nonsuperfamily sequences were never identified at DASP score thresholds evaluated in this study.)

The evaluation of true positives and false positives depends explicitly on how TuLIP groups are mapped to SFLD subgroups and families. TuLIP produced 23 functionally relevant groups and three engineered groups. Eight of those identified one SFLD subgroup or family exclusively and were thus mapped to that subgroup or family (Table II, peach cells). Two TuLIP groups mapped uniquely to two SFLD subgroups or families each (Table II, blue cells). In these instances, both SFLD groups were mapped to each TuLIP group; therefore, proteins from either SFLD groups would count as true positives. One TuLIP group (Sct20) contained only OSBS; another (Sct22) included many OSBS and 100% of the NSAR proteins (Table II, yellow cells); thus, Sct20 was mapped to OSBS and Sct22 was mapped to NSAR. The SFLD DipepEp family was mapped to three individual TuLIP groups: Sct31, Rlx48, and Rlx50 (Table II, purple cells) and DipepEp was counted as true positive for each of them. The three engineered groups each mapped uniquely to the family they represented (Table II, green cells). Overall, 16 mappings were created using 18 TuLIP groups; the remaining eight TuLIP groups contained mostly uncharacterized proteins and were not mapped to any SFLD subgroups or families.

GenBank sequences identified with a DASP2 search score equal to or more significant than the threshold were tested. For each sequence, a true positive was a sequence found both in the SFLD group and the corresponding TuLIP group. A sequence in the SFLD subgroup or family, but not identified by the corresponding TuLIP group (given the score threshold being considered), was defined as a false negative. Sequences identified by a TuLIP group and not in the mapped SFLD subgroup or family were false positives. Sequences present in the enolase superfamily, but not in the SFLD group and not in the corresponding TuLIP group were true negatives.

Acknowledgments

Molecular graphics and analyses were performed using the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (<http://www.rbvi.ucsf.edu>).

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Field C, Gocayne JD, Scott J, Shirley R, Liu L-I, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs J, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512.
2. Valencia A (2005) Automatic annotation of protein function. *Curr Opin Struct Biol* 15:267–274.
3. Bork P, Bairoch A (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet* 12:425–427.
4. Karp PD (1998) What we do not know about sequence analysis and sequence databases. *Bioinformatics* 14:753–754.
5. Kyrpides NC, Ouzounis CA (1999) Whole-genome sequence annotation: Going wrong with confidence. *Mol Microbiol* 32:886–887.
6. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 5:e1000605.
7. Pallen M, Wren B, Parkhill J (1999) Going wrong with confidence”: misleading sequence analyses of CiaB and clpX. *Mol Microbiol* 34:195.
8. Fetrow JS, Siew N, Skolnick J (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J* 13:1866–1874.
9. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—A hierarchic classification of protein domain structures. *Structure* 5:1093–1108.
10. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, Lehtinen S, Studer RA, Thornton J, Orengo CA (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res* 43:D376–D381.
11. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins Struct Funct Gen* 28:405–420.
12. Bateman A, Birney E, Cerruti L, Durbin R, Ewinger L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Res* 30:276–280.
13. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44:D279–D285.
14. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res* 41:D490–D498.
15. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W, Poulter CD, Raushel FM, Sali A, Shoichet BK, Sweedler JV (2011) The enzyme function initiative. *Biochemistry* 50:9950–9962.
16. Leuthaeuser JB, Knutson ST, Kumar K, Babbitt PC, Fetrow JS (2015) Comparison of topological clustering within protein networks using edge metrics that evaluate full sequence, full structure, and active site micro-environment similarity. *Protein Sci* 24:1423–1439.
17. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC (2006) Leveraging enzyme structure-function relationships for functional inference and experimental

- design: the structure-function linkage database. *Biochemistry* 45:2545–2555.
18. Akiva E, Brown S, Almonacid DE, Barber AE, Custer AF, Hicks MA, Huang CC, Lauck F, Mashiyama ST, Meng EC, Mischel D, Morris JH, Ojha S, Schnoes AM, Stryke D, Yunes JM, Ferrin TE, Holliday GL, Babbitt PC (2014) The structure–function linkage database. *Nucleic Acids Res* 42:D521–D530.
 19. Brown DP, Krishnamurthy N, Sjölander K (2007) Automated protein subfamily identification and classification. *PLoS Comput Biol* 3:e160.
 20. Lee DA, Rentzsch R, Orengo C (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Res* 38:720–737.
 21. de Melo-Minardi RC, Bastard K, Artiguenave F (2010) Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinforma Oxf Engl* 26:3075–3082.
 22. Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S, Gallina M, Baxter SM, Fetrow JS (2003) Structure-based active site profiles for genome analysis and functional family subclassification. *J Mol Biol* 334:387–401.
 23. Bastard K, Smith AAT, Vergne-Vaxelaire C, Perret A, Zapparucha A, De Melo-Minardi R, Mariage A, Boutard M, Debard A, Lechaplais C, Pelle C, Pellouin V, Perchat N, Petit J-L, Kreimeyer A, Medigue C, Weissenbach J, Artiguenave F, De Berardinis V, Vallenet D, Salanoubat M (2014) Revealing the hidden functional diversity of an enzyme family. *Nat Chem Biol* 10:42–49.
 24. Huff RG, Bayram E, Tan H, Knutson ST, Knaggs MH, Richon AB, P. Santago I, Fetrow JS (2005) Chemical and structural diversity in cyclooxygenase protein active sites. *Chem Biodivers* 2:1533–1552.
 25. Nelson KJ, Knutson ST, Soito L, Klomsiri C, Poole LB, Fetrow JS (2011) Analysis of the peroxiredoxin family: using active-site structure and sequence information for global classification and residue analysis. *Proteins* 79:947–964.
 26. Fetrow JS (2006) Active site profiling to identify protein functional sites in sequences and structures using the Deacon Active Site Profiler (DASP). *Curr Protoc Bioinformatics* Chapter 8.
 27. Gober JG, Rydeen AE, Gibson-O’Grady EJ, Leuthaeuser JB, Fetrow JS, Brustad EM (2016) Mutating a highly conserved residue in diverse cytochrome P450s facilitates diastereoselective olefin cyclopropanation. *Chembiochem Eur J Chem Biol* 17:394–397.
 28. Harper AF, Leuthaeuser JB, Babbitt PC, Morris JH, Ferrin TE, Poole LB, et al. (2017) An Atlas of Peroxiredoxins Created Using an Active Site Profile-Based Approach to Functionally Relevant Clustering of Proteins. *PLoS Comput Biol*. 13:e1005284.
 29. Yuan Y, Knaggs MH, Poole LB, Fetrow JS, Salsbury FR Jr. (2010) Conformational and oligomeric effects on the cysteine pK(a) of tryparedoxin peroxidase. *J Biomol Struct Dyn* 28:51–70.
 30. Huff RG (2005) DASP. Active site profiling for identification of functional sites in protein sequences and structures. Masters thesis. Wake Forest University.
 31. Kalyanaraman C, Imker HJ, Fedorov AA, Fedorov EV, Glasner ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP (2008) Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. *Structure* 16:1668–1677.
 32. Lukk T, Sakai A, Kalyanaraman C, Brown SD, Imker HJ, Song L, Fedorov AA, Fedorov EV, Toro R, Hillerich B, Seidel R, Patskovsky Y, Vetting MW, Nair SK, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proc Natl Acad Sci USA* 109:4122–4127.
 33. Glasner ME, Fayazmanesh N, Chiang RA, Sakai A, Jacobson MP, Gerlt JA, Babbitt PC (2006) Evolution of structure and function in the o-succinylbenzoate synthase/N-acylamino acid racemase family of the enolase superfamily. *J Mol Biol* 360:228–250.
 34. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584.
 35. Leuthaeuser JB, Morris JH, Harper AF, Ferrin TE, Babbitt PC, Fetrow JS (2016) DASP3: identification of protein sequences belonging to functionally relevant groups. *BMC Bioinformatics* 17:458.
 36. Sakai A, Xiang DF, Xu C, Song L, Yew WS, Raushel FM, Gerlt JA (2006) Evolution of enzymatic activities in the enolase superfamily: N-succinylamino acid racemase and a new pathway for the irreversible conversion of d- to l-amino acids. *Biochemistry* 45:4455–4462.
 37. Sakai A, Fedorov AA, Fedorov EV, Schnoes AM, Glasner ME, Brown S, Rutter ME, Bain K, Chang S, Gheyi T, Sauder JM, Burley SK, Babbitt PC, Almo SC, Gerlt JA (2009) Evolution of enzymatic activities in the enolase superfamily: stereochemically distinct mechanisms in two families of cis,cis-muconate lactonizing enzymes. *Biochemistry* 48:1445–1453.
 38. Palmer DR, Garrett JB, Sharma V, Meganathan R, Babbitt PC, Gerlt JA (1999) Unexpected divergence of enzyme function and sequence: “N-acylamino acid racemase” is o-succinylbenzoate synthase. *Biochemistry* 38:4252–4258.
 39. Rakus JF, Fedorov AA, Fedorov EV, Glasner ME, Vick JE, Babbitt PC, Almo SC, Gerlt JA (2007) Evolution of enzymatic activities in the enolase superfamily: d-Mannonate dehydratase from *Novosphingobium aromaticivorans*. *Biochemistry* 46:12896–12908.
 40. Zhang Y, Zagnitko O, Rodionova I, Osterman A, Godzik A (2011) The FGGY carbohydrate kinase family: insights into the evolution of functional specificities. *PLoS Comput Biol* 7:e1002318.
 41. Armstrong RN (1997) Structure, catalytic mechanism, and evolution of the glutathione transferases. *Chem Res Toxicol* 10:2–18.
 42. Sheehan D, Meade G, Foley VM, Dowd CA (2001) Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochem J* 360:1–16.
 43. Hayes JD, Flanagan JU, Jowsey IR (2005) Glutathione transferases. *Annu Rev Pharmacol Toxicol* 45:51–88.
 44. Atkinson HJ, Babbitt PC (2009) An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. *PLoS Comput Biol* 5:e1000541.
 45. Mashiyama ST, Malabanan MM, Akiva E, Bhosle R, Branch MC, Hillerich B, Jagessar K, Kim J, Patskovsky Y, Seidel RD, Stead M, Toro R, Vetting MW, Almo SC, Armstrong RN, Babbitt PC (2014) Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biol* 12:e1001843.
 46. Dong GQ, Calhoun S, Fan H, Kalyanaraman C, Branch MC, Mashiyama ST, London N, Jacobson MP, Babbitt PC, Shoichet BK, Armstrong RN, Sali A (2014) Prediction of substrates for glutathione transferases by covalent docking. *J Chem Inf Model* 54:1687–1699.

47. Mannervik B, Awasthi YC, Board PG, Hayes JD, Di Ilio C, Ketterer B, Listowsky I, Morgenstern R, Muramatsu M, Pearson WR (1992) Nomenclature for human glutathione transferases. *Biochem J* 282:305–306.
48. Mannervik B, Board PG, Hayes JD, Listowsky I, Pearson WR (2005) Nomenclature for mammalian soluble glutathione transferases. *Methods Enzymol* 401: 1–8.
49. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32:D129–1D133.
50. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14:48–54.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
52. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
53. Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit Lett* 27:861–874.
54. Meilă M Comparing clusterings: an axiomatic view. In: ACM Press; 2005. pp. 577–584. Available from: <http://portal.acm.org/citation.cfm?doid=1102351.1102424>
55. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14: 1188–1190.