Flexibility of Short-chain dehydrogenase is interconnected to its promiscuity for the reduction of multiple ketone intermediates

Affiliations:

Anirudh P Shanbhag[1,2], Sreenath Rajagopal[2], Arindam Ghatak[1,3], Nainesh Katagihallimath[2], Ramswamy S.[4] and Santanu Datta[2]

1. Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, Kolkata 700129
2. Bugworks Research India Pvt. Ltd., C-CAMP, UAS GKVK Campus, Bangalore 560065
3. Biomoneta Research Pvt. Ltd., C-CAMP, UAS GKVK Campus, Bangalore 560065
4. Institute of Stem-cell Science and Regenerative Medicine (InStem), UAS GKVK Campus, Bangalore 560065

Corresponding Author:

Dr. Santanu Datta, Ph.D.

Email: santanu@bugworksresearch.com

Abstract

Short-chain dehydrogenases/reductases (SDRs) are a convenient class of enzymes used to synthesize enantiopure alcohols. Several studies describe native or engineered SDRs for converting substrates of interest using cost and time-intensive high-throughput approaches. The classification of SDRs is based on chain length and cofactor binding site. Of these, the shorter 'Classical' and the longer 'Extended' enzymes participate in ketoreduction. However, comparative analysis of various modelled SDRs reveals a length independent conserved N-terminal Rossmann fold and a variable C-terminus region for both types. The general hypothesis is that the latter domain influences the enzyme's flexibility that may affect the observed promiscuity of the enzyme. We have used a machine learning algorithm on this flexible domain to build a rationale to screen promiscuous candidates. We have built a data set consisting of physicochemical properties derived from the amino-acid composition of enzymes to select closely associated promiscuous mesophilic enzymes. The resulting in vitro studies on pro-pharmaceutical substrates illustrate a direct correlation between the C-terminal lid-loop structure, enzyme melting temperature and the turnover number. We present a walkthrough for exploring promiscuous SDRs for catalyzing enantiopure alcohols of industrial importance.

Keywords:

Short-chain dehydrogenases, Ketoreductases, Catalytic efficiency, Enzyme turnover, Active pharmaceutical intermediates

Introduction

Short-chain dehydrogenases (SDRs) form a promiscuous class of enzymes that utilize NAD(P)H as a cofactor for catalyzing a wide array of prochiral ketones into enantiopure alcohols. These enzymes have been studied by techniques employing site-directed mutagenesis, site saturation mutagenesis, substrate channeling, and gene shuffling. (1)(2)(3). SDRs encompass functions like detoxification, hormone metabolism, secondary metabolite and biopolymers production (4)(5)(6)(7). Such broad substrate promiscuity is often explored to prospect an enzyme with superior kinetic parameters for converting substrates of industrial importance. Many studies have reported engineered enzymes demonstrating better stability, stereoselectivity and turnover characteristics. However, these upgrades often focus on a single substrate of interest rather than a generic improvement with other substrates of similar markush structures (8)(9). Furthermore, substrate catalysis and turnover rates are trade-offs unless the enzymes are specifically engineered (10). Therefore, it is essential to analyse various enzymes to obtain a candidate to convert multiple substrates at a higher turnover.

Structurally, SDRs have a ubiquitous cofactor binding Rossmann fold which is similar to other dehydrogenases such as Medium-chain dehydrogenases, Iron-containing alcohol and Aldehyde dehydrogenases, signifying convergent evolution (11)(12)(13). They are classified into five types, namely, Classical, Extended, Intermediate, Complex and Divergent. Most of them fall under the first two classes and contain the active site SxnYxxxK. However, 'Classical' SDRs are ~250 residues long with TGxGx3G motif as a cofactor binding region. In contrast, the 'Extended' counterparts have an Alanine instead of the third Glycine and are ~350 residues long. Interestingly, another study has analysed thousands of members showing a length independent and extensively interconnected network of similar proteins (14). Such a superfamily with closely related members with natural substrate promiscuity is ideal for understanding multiple substrate catalysis.

SDRs being a large superfamily of proteins (with a monotonically increasing number of members every year), screening for a promiscuous candidate could be technically overwhelming. However, machine learning algorithms have helped decipher patterns among complex data sets and protein data (15)(16). Previously, SDRs have been classified using Hidden Markov Models (HMMs), which are prominently used in protein sequence classification in SCOP, CATH and Uniprot databases (17)(18). The algorithm has powerful statistical functions and can handle variable lengths of protein sequences inputs. However, it has many unstructured parameters and cannot capture higher-order correlations among amino acids in a protein. Also, HMM has to be reasonably constrained and iterative to represent a small fraction of space over a vast number of sequences. Hence, it is optimal in identifying active sites, secondary structure similarities (used in multiple sequence alignment) and native substrates based on existing data.

Additionally, HMM is often unable in sorting proteins based on other overarching factors such as physicochemical properties, taxonomical origins and other non-canonical classifiers. Therefore, instead of engineering a protein for substrate-specific conversion, we have focused on finding a promiscuous SDR which can convert multiple substrates of interest. We wish to classify SDRs using machine learning algorithms to generate a 'signature' of multi-substrate catalysis. Unlike traditional prospecting that uses phylogenetic classification to ascertain the closest homolog to a reported promiscuous enzyme, we classify the proteins based on their respective phylum. This strategy eliminates outliers and gives a homogenous dataset for further clustering to ascertain promiscuous candidates based on their proximity to reported ones. Here, we present a rationale and a 'Proof of Concept' for deriving an enzyme that can catalyze different active pharmaceutical intermediates.

Results

(i) Modelling and architecture of SDRs

Modelling different sized Short-chain dehydrogenases can help in understanding the architecture of the proteins. There is a soft hypothesis that the lid-loop structure, a helix-loop helix-motif, plays an essential role in the enzyme's flexibility. It is most often the α7 or α6 helix in SDRs. It is a well-characterized structural entity and was first reported in Thiohydorxy naphthalene reductase from *Magnaporthae grisiae*. (19)(20). The dynamic closing of the lid loop upon cofactor and substrate binding restricts substrate access to the active site to a direction that is compatible with the observed stereochemical outcome of ketoreduction (21). We modelled SDRs of varying lengths to determine the influence of the C-terminus on the overall architecture of the protein. Classical and Extended SDRs exist as dimers (or tetramers) and monomers, respectively. However, the models illustrate a different narrative. The N-terminal Rossmann fold is structurally conserved, whereas; the C-terminal forms a lid-loop structure and consists of residues that participate in oligomerization. Some of the modelling data from the 'longer' Classical SDRs from other organisms showed a change in loop architecture and the oligomerization of the enzymes (Figure 2).

Upon modelling, two distinct sets of SDRs were observed, based on the accessibility of α-helices. Invariably the amino acid composition of the protein is a determines this characteristic. In SDRs which are monomers, the α-helices are not accessible, thus preventing oligomerization. Furthermore, the C-terminal part of the protein participates in changing its shape. Figure 2 demonstrates this change from *Mycobacterium* sp. 301 AA to *Mycobacterium* sp. 311 AA, where the C-terminal loop goes around the P interface (Tetramerization interface) towards the Q interface (Dimer interface), thereby changing the exposure of the α-helix residues imperative for oligomerization. The significance of the orientation of both interfaces was demonstrated by Biswas *et al.*, where swapping of the Q-interface domains between the naturally tetrameric *Staphylococcus* FabG and dimeric *Escherichia coli* FabG led to a tetrameric chimeric variant of *E. coli* FabG (22). Another important fact is that the active site is more accessible due to a floppy lid-loop structure (red).

Further, upon comparison with an extended SDR, it is seen that both accessibilities of the active site of the non-accessibility of the Q-interface are evident (Figure 2 *Z. rouxii* 338 AA). In classical SDRs, the α4 and α5 helices are unrestricted and available for oligomerization. However, in extended SDRs, the helices are hindered and therefore, oligomerization fails to occur in these enzymes (23). The change in the cofactor binding site of the Extended SDRs is not a product of the protein's architecture; instead, it is the composition of the Amino acids. The active site is nearly the same among all the classical SDRs, as seen in Supplementary figure S2.

(ii) The catalytic efficiency of an essential SDR and its homologs from *E. coli*

FabG is one of the essential SDRs along with the FabI gene in E. coli. Most essential enzymes are specific and have low promiscuity, and therefore, we used the enzyme as a control. UcpA and IdnO are the closest non-essential counterparts of FabG in the organism. They were selected to discern the trend towards multi-substrate catalysis and delineate the relationship between gene essentiality, flexibility, and catalytic promiscuity. These enzymes are similar to each other (Figure 3B). Further, the ketoreductase ZRK was used as a positive control for ascertaining the spectrum of substrate promiscuity.

Differential scanning fluorimetry (Thermal shift assay or DSF) was performed to test the enzymes' rigidity, where the flexibility of the enzymes is inversely proportional to the melting temperature of the proteins (24). Many reports on enzymes belonging to the same family have demonstrated protein flexibility as a characteristic dependent on the optimum functional temperature. In psychrophiles, the cold-adapted proteins Uracil DNA glycosylases and Endonucleases I are more flexible than their mesophilic counterparts (25)(26). Similarly, thermophilic proteins such as malate dehydrogenases were studied protein rigidity using both differential scanning fluorimetry and Molecular Dynamics (MD) simulations (27)(28).

The dynamicity of the SDR models were checked using various methods. The B-factor was measured on AlphaFold predicted models in ChimeraX tool (Supplementary Figure S4A). Further, the deep-learning server Medusa showed an increase in number of flexible residues in proteins with larger lid-loop structure (Supplementary Figure S4B). We also cross validated this using both the primary sequence and AlphaFold models using PredyFlexy and FlexServ (Supplementary Figure S4 C and D). Thus, it is not overly optimistic to

assume that protein flexibility can be further augmented by Differential scanning fluorimetry data. During a DSF assay, the buried hydrophobic residues bind to SYPRO orange as the protein unfolds due to increasing temperature, producing a melt curve. The slope of the melt curve and the midpoint of transition (Tm) indicate the enzyme's overall rigidity/flexibility (Figure 3A). The models depict a relationship between the lengths of the lid-loop and the proteins' melting temperature (Figure 3B). The Eco_FabG is relatively rigid and has the highest melting temperature (63.8 ^{o}C), while UcpA and IdnO has a lower Tm (54-56 ^{o}C), and ZRK has the lowest (33.55 ^{o}C) melting temperature (Figure 3A). Furthermore, ZRK shows the highest catalytic efficiency, followed by IdnO and UcpA. Therefore, Eco_FabG, which has the minimum catalytic efficacy, establishes an inverse correlation between rigidity as a characteristic and multi-substrate catalysis (Figure 3C).

(iii) Machine learning

a. Dataset preparation

The dataset was prepared by downloading SDR sequences from UniProt, CATH and SCOP databases. They were then filtered using the Batch CDD (Cluster Domain Database) to identify the family and superfamily, respectively. Furthermore, the taxonomic origin for every protein sequence (at the phylum level) was manually curated while eliminating entries with unknown origins. The features from the ProtScale and ProtParam (ExPasy server) were used to build the dataset. A total of 81,014 curated SDRs were finally deployed for classification.

b. Feature selection

In data processing, intelligent feature selection is an essential step for filtering out redundant or low-affecting factors. This step increases the efficiency of sorting through supervised or unsupervised learning. A Chi-square test was performed, and a correlation matrix was produced (Figure 4). Theoretically, a very low or a high number of features often leads to wrong classification or overfitting. In this study, the correlation matrix helped determine 20 features for classifying SDRs; namely, Molecular Wt., GRAVY, the amino acids A, C, D, E, F, G, I, K, L, N, Q, R, S, V, Y, Aliphatic Index, % Negatively charged amino acids, % of Positively charged amino acids were chosen as features for classification.

c. Classification of enzymes

A refined dataset exemplifies the accuracy and performance of the algorithms in question. To this end, the proteins must be refined by considering the gradual change influenced by speciation. Genes (and proteins by extension) are known to evolve from gene duplication, gene transfer and mutation events. Compared to these two occurrences, gene insertions are rare (29)(30)(31). Therefore, classifying sequences based on their phylum of origin can help obtain candidates that have undergone a gradual change in the line of evolution and speciation. SDRs have many members due to gene duplication and translocation events throughout various phyla across geological time. Hence, considering enzymes that have existed in the line of evolution would enable in eliminating strong outliers (which have arisen from gene insertion or multiple gene truncation events) from the dataset. Additionally, most of these events are across closely related species in a phylum, subphylum, and class. Thus, classification was done to avoid extraordinary circumstances in which these genes which might have undergone major deletions or insertions. This process refines the dataset for the subsequent clustering analysis.

Machine learning algorithms such as k-Nearest Neighbour, Extra tree, Random Forest, and Support Vector machines were used for classifying enzymes to their respective derivative organism's phyla based on the physicochemical parameters. This process helped retain only true positives using the confusion matrix (Example shown in Figure 5A and complete data is depicted in Supplementary figures S7-S14). We observed that True positives were (marginally) better classified using Support vector machines (SVM) than other data-mining algorithms. Each machine learning algorithm underwent 1500 iterations, beyond which the accuracy could not be improved to greater than 80% for SDRs (Figure 5B).

d. Clustering of enzymes

The number of clusters was determined using the Elbow (or knee of a curve) method, a heuristic approach for determining the number of optimum clusters for any given dataset. K-Means clustering was used for finding enzymes that lie close to the reported promiscuous enzymes. Since k-means clustering uses a mean between

three data points, i.e., the cluster's centroid defined as epsilon value (the sum of squares of the distance between data points and the respective centroid of the cluster to which the data point belongs) was measured using the squares of the distances between them. The elbow method uses a two-dimensional plot between the epsilon value and the number of iterations to determine the number of clusters (Supplementary figure S5).

The clustering results show that enzymes group together based on their catalytic potential (promiscuity) and melting temperature. Thermophilic enzymes belonging to clusters 7, 12 and 19 are close to each other. We used principal component analysis (PCA) for graphically depicting these results (Figure 5C). Similarly, clusters 4, 7, 11, 12, 19 have promiscuous enzymes grouped in the neighbourhood. Except for clusters 12 and19, most of the enzymes from the remaining clusters contain promiscuous candidates. Potential unknown enzymes clustered with reported promiscuous SDRs were chosen for in vitro analysis (Table 1). ByueD (*Bacillus subtilis* ketoreductase) and DHK (*Debaryomyces hansenii* ketoreductase) were selected for further analysis. As a negative control, another unverified enzyme Hketo (*Hansenula polymorpha* ketoreductase) was chosen from cluster 17 (a low promiscuous cluster in Table 1).

(iv) In vitro analysis

For In vitro analysis, enzymes were chosen from the clusters based on the different architecture. As aforementioned, SDRs chosen for ketoreduction are predominantly from classical and 'extended' types. Therefore, from the cluster, three known and three unknown enzymes were chosen for in vitro analysis. Classical SDRs are known for producing Active Pharmaceutical Intermediates (API), namely, *Synechococcus sp.* PCC 7094 FabG, ScoRed (having a classical SDR structure similar to Eco_FabG) were selected, and ZRK were chosen from the cluster. Additionally, candidates not reported for the production of Active Pharmaceutical Intermediates such as ByueD (*Bacillus subtilis),* Hketo *(Hansenula polymorpha)* and DHK (*Debaryomyces hansenii*) were chosen based on the clusters predicted (Figure 6B).

The DSF experiments exhibited *Synechococcus* FabG to have a higher melting temperature than the rest of the SDRs. Meanwhile, DHK has a similar melting temperature as ZRK at the lower end, followed by ByueD and Hketo, respectively. Conversely, ScoRed had a Tm like Hketo and ByueD despite having a smaller lid-loop structure as FabG (Figure 6A and 6B). Once again, the DSF data shows a direct correlation between the Tm and the lid-loop structure in SDRs (Figure 6A and 6B).

Enzyme kinetics was performed against substrates consisting of commonly published ketones such as acetophenone, α-chloroacetophenone and Ethyl-4-chloro acetoacetate. Further, other pharmaceutical ketone intermediates (K-INT) such as Talampanel K-INT, Aprepitant K-INT, Sitagliptin K-INT, Rivastigmine K-INT, Dolastatin K-INT, Sulopenem K-INT, MA-20565 K-INT and Crizotinib K-INT useful in drug industries were introduced as well (Supplementary figure S1).

Enzyme kinetics data reveal that *Synechococcus* FabG has the highest km and low kcat (thereby having the highest catalytic efficiency or kcat/km ratio except for Eco_FabG) for most of the substrates. Comparatively, ScoRed is better than previously analysed UcpA and IdnO. Hketo has a slightly larger lid-loop structure compared to those above low-molecular-weight classical SDRs. It depicts a better catalytic efficiency despite being clustered as a low-promiscuous candidate. Comparatively, ByueD has a larger lid-loop and has better catalytic efficacy than previous enzymes. However, DHK with a similar structure and a larger lid-loop portray low km with high catalytic efficiency than other SDRs. Interestingly, ScoRed exhibited the best kcat/km values for Dolastatin K-INT and Acetophenone whereas; FabG showed comparably good activity against acetophenone, α-chloroacetophenone and Ethyl-4-chloro acetoacetate (COBE). Nevertheless, overall, ZRK and DHK proved superior for the catalytic conversion of industrially relevant pharmaceutical intermediates (Figure 6C).

The standard deviations of in vitro enzyme kinetics (Figure 7A) show that they do not vary as much as the catalytic efficiency. This lack of variation indirectly depicts the homogeneity of the active site among all the SDRs (Figure 7B). The extended SDRs are more flexible as lesser variation is present among the substrates. This characteristic is especially true for ZRK and DHK.

Discussion

Many studies demonstrate the production of chiral alcohols, but only a few have prevailed in making any economic advances in this field. Invariably, there are hurdles in obtaining an enzyme that can perform a function of interest because the substrates are non-native. Besides, a more straightforward method can alleviate the cost and time-intensive techniques like high throughput screening and genome/metagenome sequencing. However, the onus of being a catalyst in an industrial setup with low-cost input is a tough challenge and is necessary for a large-scale setup (32)(33). Additionally, finding a promiscuous enzyme is a superior alternative as it avoids protein engineering and high throughput screening. Therefore, using machine learning for unearthing a promiscuous candidate is a favourable solution in this framework.

The modelling results show a change in protein architecture due to the variations in the lid-loop structure (Figure 2). Biochemically, classical SDRs have evolutionarily pre-existed their extended counterparts as they are critical players in essential biochemical pathways (34). Therefore, classical SDRs were modelled to architecturally link them to the 'Extended' class of proteins. Although connecting them in terms of evolution is an arduous task, SDRs (especially from microbes and unicellular eukaryotes) have a conserved N-terminal structure and a variable C-terminus substrate interacting region that gives a perspective on the change in the architecture of SDRs as a whole. Also, Extended SDRs are naturally promiscuous contenders for detoxification and generation of secondary metabolites in yeast and higher eukaryotes. Therefore, this superfamily is an essential placeholder for ascertaining promiscuous ketoreductases.

The lid-loop structure of the short-chain dehydrogenases is speculated to participate in the entry and exit of the substrate (35). The structure stabilizes upon the binding of NADPH and helps maintain the protein's overall architecture (36). It has also been altered using site-directed mutagenesis for attaining a specific product of interest, e.g., the lid-loop FabG in *Synechococcus* sp. was engineered to mimic the flexibility of another native SDR, the PhaB, for increasing the production of Polyhydroxy butyrate. Conversely, the flexibility was reduced in *Serratia marcescens* SDR for the production of (R)-phenylephrine (37)(38). Contextually, it is possible to engineer SDRs to obtain a specific product of interest; however, for multi-substrate catalysis, the overall architecture, flexibility, and an estimate of catalytic activity need to be considered for establishing a viable candidate. Therefore, firstly, it is crucial to establish the relationship between flexibility and catalytic efficiency. Secondly, enzymes belonging to a similar group with comparable physicochemical properties with diverse lid-loop sizes are favourable for locating a sustainable enzyme.

To prove the first assumption, the metabolically essential SDR Eco_FabG (*E. coli* FabG) and other non-essential SDRs (IdnO and UcpA) with similar identity demonstrate a clear relationship between flexibility and catalytic efficiency (Figure 3A). Also, the general direction for prospecting a multi-substrate catalyzing SDR could be their non-essentiality. Therefore, a reasonably well established, non-essential SDR from ZRK was chosen as a positive control for this experiment. It demonstrated the relationship between flexibility and catalytic efficiency for multiple substrates (Figure 3C). Also, modelling data showed ZRK, an extended class SDR, to have a larger lid-loop (Figure 3B).

This size extension could contribute to the overall flexibility of the protein as well. Therefore, for the second assumption, we focused on sorting the proteins based on their physicochemical properties. To this end, we used machine learning to classify and cluster enzymes similar to reported promiscuous candidates. We also chose candidates with different molecular weights from FabG (~250 residues) to ZRK (~320 residues) for forming a spectrum of non-essential enzymes with increasing lid-loop architecture and expectantly flexibility.

Deep learning and machine learning have been used to gain a better perspective on protein structure and function. Also, protein sequences are a good metric for predicting the kingdom of origin and thermostability (39)(40). Both systems have advanced quite a lot in the last decade, e.g., Chakravorty *et al.* used structural data, protein, the encoding nucleotide sequence as features for classifying thermostable and mesothermal proteins. Many unsupervised and supervised methods were used to this end to determine the cause of thermostability at 'each level' (i.e., nucleotide, protein, structure level). The study showed that protein sequence data is the best measure for understanding protein thermostability with 90.97% accuracy (41).

Similarly, classification of protein stability using network features (in addition to structural and amino acid compositions) using network centrality values and sequence features (specifically di- and tri-peptide bonds) was

quite successful. The centrality depends on the distances between Cα atoms that correlate better with protein stability. These features improved classification significantly to 96% accuracy (42).

A processed dataset exemplifies the accuracy and performance of the algorithms in question. Therefore, we refined the dataset by considering the gradual change in genomes (and subsequently proteins) influenced by speciation. Genes (and proteins, by extension) evolve from gene duplication and mutation events (43). Compared to these two occurrences, gene insertions are rare. Genomic studies across the three domains of life have shown that successful gene shuffling has occurred only 48 times in higher eukaryotes (i.e., between *Caenorhabditis and Drosophila*) in half a billion years. Also, gene shuffling is one order of magnitude lesser than gene duplication in eukaryotes, unlike prokaryotes, where both incidences are similar. Therefore, this shows that the introns are not necessarily involved in gene shuffling as prokaryotes lack them and yet, gene shuffling is a common artefact among them (44). Therefore, SDRs were classified based on their phylum of origin. Also, as mentioned earlier, features based on protein sequences are superlative for predicting thermostability, an invariable measure of flexibility. Classification effectively removed the outliers (of evolution) and gave a refined dataset of true positives, which was further clustered to provide unknown candidates closer to the reported promiscuous members.

The clustered true positives (Hketo, ByueD, DHK) lying close to the reported mesophilic promiscuous proteins were selected. Also, ScoRed and *Synechococcus* FabG were chosen for establishing the in vivo capability of reported candidates similar to ZRK. Moreover, the desired enzymes were modelled to discern their lid-loop structures (Figure 6B). The DSF assay exhibited that lid-loop size is inversely proportional to the melting temperature. ScoRed is an exception as it showed a similar melting temperature as Hketo, ByueD and DHK (Figure 6A). However, the catalytic efficiency of all the enzymes established that lid-loop is an essential component for instituting flexibility and promiscuity in SDRs. The large lid-loop containing DHK and ZRK are superior for converting pharmaceutical ketone intermediates. Further, Hketo and ByueD have the next best choice catalytic efficiency (Figure 6C). Although *Synechococcus* FabG and ScoRed have good kcat/km values, they are relatively less than the SDRs with larger lid-loop.

The clustered true positives (Hketo, ByueD, DHK) lying close to the reported mesophilic promiscuous proteins were selected. Also, ScoRed and *Synechococcus* FabG were chosen for establishing the in vivo capability of reported candidates similar to ZRK. Moreover, the desired enzymes were modelled to discern their lid-loop structures as well (Figure 6B). The DSF assay exhibited that lid-loop size is inversely proportional to the melting temperature. ScoRed is an exception as it showed a similar melting temperature as Hketo, ByueD and DHK (Figure 6A). However, the catalytic efficiency of all the enzymes established that lid-loop is an essential component for instituting flexibility and promiscuity in SDRs. The large lid-loop containing DHK and ZRK are superior for converting pharmaceutical ketone intermediates. Further, Hketo and ByueD have the next best catalytic efficiency (Figure 6C). Although *Synechococcus* FabG and ScoRed have good kcat/km values, they are relatively less than the SDRs with larger lid-loop.

It is critical to understand that engineering proteins are subtle, and an angstrom difference in the right places in an enzyme can be a ‘make or break’ for catalysis (1)(45). Therefore, a strategy to identify a potential SDR based on flexibility and architecture is a less expensive and valuable strategy for converting multiple ketones of pharmaceutical importance. Further, it leads to costly high throughput screening as well. More than ~100,000 (and counting) SDRs and machine learning of primary sequences can help detect attractive test candidates.

Author contributions

SD and RS conceived and provided support via a grant for the study. APS, SR and AG performed all the preliminary and final experiments. NK was involved in designing and executing the in vitro and modelling experiments. All authors have read and approved the final manuscript.

Acknowledgement

The financial support from the Biotechnology Industry Partnership Program, Department of Biotechnology, India (DBT Sanction Order No: BT/BIP0475/13/11) and Grant BT/PR13429/BRB/10/767/2009 to Prof. S. Ramaswamy is gratefully acknowledged.

Methodology

1. Reagents and chemicals

A set of high value prochiral ketone intermediates namely, Talampanel ketone intermediate or K-INT (3,4-methylenedioxyphenyl acetone), Barnipidine K-INT (1-N-carbobenzoxy-3-pyrrolidione), Aprepitant K-INT (1-(3,5-bis-trifluoromethyl-phenyl)-ethanone), Crizotinib K-INT (1- (2,6-dichloro-3-fluoro-phenyl)-ethanone, 3-trifluoromethyl acetophenone), Dolastatin K-INT (2-phenyl-1-thiazol-2-yl-ethanone, 1- (3-methoxyphenyl) ethanone), Sitagliptin K-INT (3-oxo-4- (2,4,5-trifluoro-phenyl)-butyric acid methyl ester) and Rivastigmine K-INT (1- (3-methoxyphenyl) ethanone) were synthesized as reported elsewhere (46). The uses of these pharmaceutical intermediates are highlighted in Supplementary Figure 1.

2. Homology modeling

The homologs for the enzymes of interest were identified using Protein BLAST. The best hit was used as a template for homology modelling using SWISS-MODEL webserver and was validated by PROCHECK software. Further, the obtained model was then refined using AlphaFold through ChimeraX interface.

3. Flexibility prediction

Further, the flexibility of the modelled proteins was checked by using B-factor in ChimeraX. Furthermore, tools for predicting protein flexibility namely, Medusa (a server which uses deep learning to predict protein flexibility), PredyFlexy and FlexServ were used for validating the AlphaFold models (47)(48)(48).

4. Cloning of the genes

- The genes native to *Escherichia coli* viz. Ecoli_FabG, IdnO and UcpA were acquired from the ASKA collection. The enzymes are cloned in pCA24N plasmid which has T5 promoter which is compatible with T7 polymerase in BL21(DE3) strain (49).
- Genes encoding the other heterogenous enzymes namely, *Sulfolobus solfataricus* alcohol dehydrogenase (SSADH), *Zygosaccharomyces rouxii* SDR (ZRK), *Hansenula polymorpha* DL-1 peroxisomal 2,4-dienoyl-CoA reductase (Hketo), *Synechococcus sp.* PCC 7942 3-ketoacyl-[acyl-carrier-protein] reductase (FabG), *Bacillus sp.* ECU0013 ADH (ByueD) and *Debaryomyces hansenii* Ketoreductase (DHK) were codon optimized for *E. coli* and synthesized by Geneart (http://www.geneart.com).
- The gene sequences are deposited in DDBJ and Genbank with the following accession numbers namely, LC325171 for SSADH, LC325172 for ZRK, LC325173 for PAR, LC325174 for ByueD, LC325175 for Hketo, LC325176 for FabG, XP_458533.2 for DHK.
- These are then sub-cloned into pET28a vector (Novagen) using the restriction sites NdeI and XhoI with an N-terminal His-tag in frame.

5. Purification of protein

- The cells are harvested by centrifugation (8000g ×20 min) and resuspended in lysis buffer (50 mM Sodium Phosphate Dibasic, pH 8.0, 0.3 M NaCl, 10 mM imidazole). The cells are lysed on ice by sonication at 39% amplitude with ON and OFF cycle of 10 seconds for 7 rounds and the debris removed by centrifugation (22000 g × 30 min).
- The recombinant His tagged protein is purified by Ni-NTA affinity chromatography (4ml, Qiagen) and eluted by using elution buffer (50 mM Sodium Phosphate Dibasic, pH 8.0, 0.3 M NaCl, 250 mM imidazole).
- The protein is concentrated to 10 mg/mL with 10,000 MWCO centricons (Millipore). To remove imidazole the protein is diluted in 20mM Sodium Phosphate buffer pH 7.2 and concentrated again. This procedure is repeated 3 times until ~0.25mM imidazole remained.
- Soluble protein expression experiments are performed as in the protocol followed for whole-cell transformation. The cells are harvested 18h post induction and the soluble His-tagged ketoreductases are purified by Ni–NTA affinity chromatography.

- The SDS PAGE of the purified enzymes is depicted in Supplementary figure S2.

6. Buffer optimization and Thermal shift assay (TSA)

- The purified protein is obtained from -80°C stock (prepared after purification) and thawed in ice for ~30 minutes. The protein concentration is measured in a spectrophotometer (A_{215}/A_{225}) and the protein is diluted to 1mg/ml.
- For a single reaction, 50ug/ml of final protein concentration is used along with 20x SYPRO orange (Final concentration). The reaction is run in an RT PCR machine (Bio-Rad/Applied Biosystems 7900HT) at 2% ramp rate (~1°C per minute) and from 25°C to 90°C at Ex/Em (497/520).
- The run is approximately 40 minutes. The raw data is extracted into Graphpad Prism and the Tm (V_{50}) is calculated using Boltzmann sigmoidal fit to obtain the melting temperature of protein.

7. In Vitro Kinetic Assays

- Enzyme specific activity

The measurement of exam activity is done in the Tecan S200 spectrophotometer at A_{340}. The final concentration of enzyme used is 1ug/ml for the reaction. The substrate is added to approximately 200 times the concentration of the enzyme. A consecutive rate of equal conversion is used to establish the specific activity of the enzyme. It is then extrapolated for 1uM/min/mg (specific activity).

- Enzyme Kinetics

For ketoreduction, all assays are performed in optimal buffer conditions with the enzyme concentration at 1ug/ml. The km for each ketone of interest is obtained by varying the substrate concentration in the presence of 2000μM NADPH/NADH. The Km for NADPH/NADH is obtained by varying the cofactor concentration in the presence of saturating amounts of ketone substrate. The slope for initial steady-state reaction is used for obtaining the velocity (uM/s) at known substrate concentrations. These two parameters are plotted in GraphPad Prism to obtain a Michaelis Menten fit to obtain Km, Vmax and Kcat of an enzyme for a substrate of interest.

8. Data extraction

(i) Data collection:

A domain specific search is done on Uniprot server to extract all the data for SDR. Also sdr-enzymes.org, an HMM based classification website, was used to gather all the protein sequences. The data added in MS-Excel spreadsheet and redundancy of the sequences of removed using CD-HIT server with sequence identity cut off as 0.9 and also, conditional formatting in Excel was used to verify the same. Physicochemical parameters: NCBI taxonomy identifier, Algabase, Bacterio.net, Encyclopedia of Life (EOL) databases were used to identify the Species, Phylum and Environment of survival. These were added individually in the database.

ProtParam and Protscale were used to measure Total Amino acids, Enzyme name, Sequence, Various Amino Acids %, Molecular Weight, Theoretical PI, GRAVY (Hydrophobicity), Extinction coefficient, Aliphatic Index, Negative %, Positive % and Solubility were used as features for classifying proteins.

Normalization: Transformation of the data to a homogeneous and final dataset was performed applying the following equation:

$$N = \frac{O - m}{M - m}$$

Where (N) denotes normalized output, O represents the original data value, M and m refers to the maximum and minimum data values of a parameter, respectively.

(ii) Machine Learning: Feature selection, Classification and Clustering

Scikit, a free machine learning and clustering (Scipy) tool based on Python programming language, which allows machine learning and clustering through cluster-computing and iterative processes, was used for finding.

(iii) Supervised learning methods

In supervised learning methods, part of the dataset is used for training the system to recognize hidden patterns and classify them accordingly. In this exercise, the system was trained using 90% of the dataset and 10% was used for testing. This was iterated multiple times in order to get a semblance of true positives. These were then used for clustering using unsupervised learning. Following are the methods used in classification.

(iv) K-Nearest neighbour

This method utilizes distances between points as a measure for classification. It classifies the data points into different classes based on the characteristics of the nearest neighbours. The distances are measured using Minkowski, Manhattan, Cosine, Euclidean and Jaccard distance. Of these, Makowski and Cosine distances are used to measure similarities in vectors. In contrast, Jaccard's distance is useful for smaller sample size with binary inputs e.g., DNA gel analysis. The Manhattan distance is useful for a dataset consisting of real-world geometric problems. Therefore, in the current study, Euclidean distance has been used for classifying the enzymes as it is contextually superior to Manhattan distance. The number of weights used in the current study was 3 using Euclidean distance. About 1500 iterations were carried out using Random Forest to ascertain classification.

(v) Support Vector Machine (SVM)

The method utilizes a hyper-plane (a plane in *n* dimensions) that separates the data belonging to different classes. This is useful for classifying high dimensional data. The gamma value used for the current classification was 2.25, followed by 500 iterations.

(vi) Random Forest (RF)

A multidimensional dataset such as the current one cannot be classified using simple binary or decision trees. Although the computation time is significantly low, the classification is poor due to the multi-dimensionality of the dataset. Therefore, the random forest (an advance on the decision trees) is better suited for the problem. It makes predictions based on consensus rather than just a binomial decision of the decision trees. The algorithm works on the principle of navigating over random subsets of features for obtaining the best feature. Hence, the search space is diversified, resulting in robust prediction. Furthermore, it can handle imbalances in sample size among the classes. About 1500 iterations were carried out using Random Forest to ascertain classification.

(vii) Extra tree (ET)

Another algorithm that is based on decision trees is Extra trees, short for extremely randomized trees. It is an ensemble learning algorithm where, unlike random forest that develops each decision tree from a sample of the training dataset, the extra tree algorithm utilizes the complete training dataset. It works by making a large number of unpruned decision trees from the training dataset and then classifying them based on the majority. About 1500 iterations were carried out using Random Forest to ascertain classification.

(viii) Unsupervised learning methods

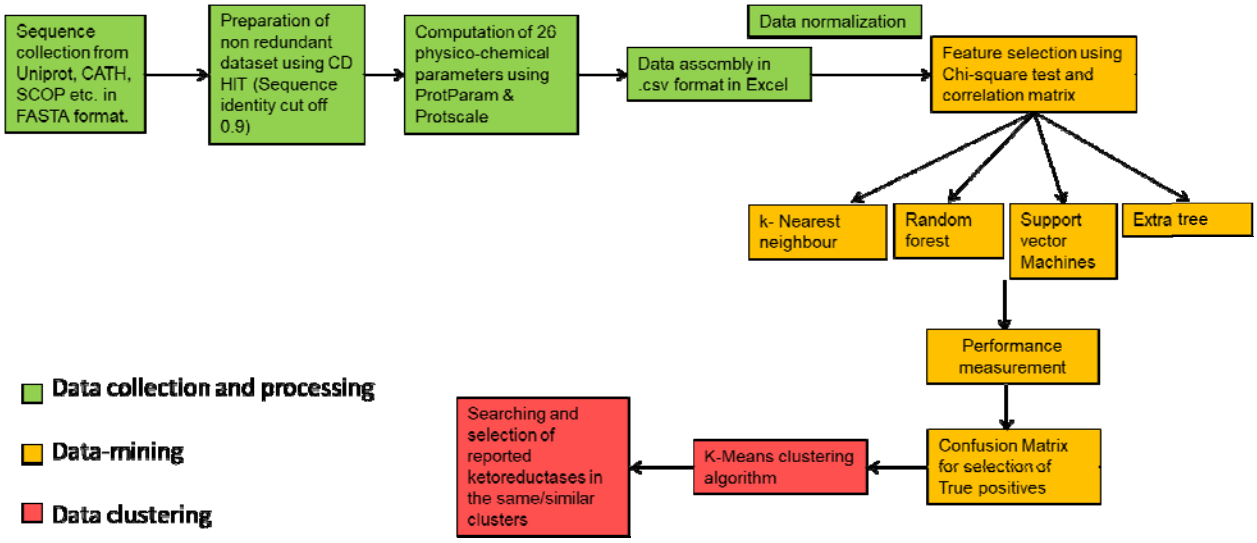
In unsupervised learning the data is clustered without a primary training set. It is useful for finding low-dimensional features from a high-dimensional input data. The unsupervised learning techniques used in this thesis are as follows:

(ix) k-means clustering

The k-means clustering is an unsupervised algorithm which is done by using a set of random numbers as *mean* denoted as '*k*'. After which the numbers closer in value to the *mean* are taken into consideration for forming an initial cluster. Further, the mean is continually changed until all the numbers are close to its value. The number of means depend on the number of clusters chosen for the study.

(x) Principal component analysis

The principle component analysis (PCA) is used for reducing dimensions of the data to two. Based on the '*n*' input data vectors, the PCA generates '*p*' singular vectors of the data matrix.



Overall schematic of obtaining proteins

References

1. Koesoema, A. A., Standley, D. M., Senda, T., and Matsuda, T. (2020) Impact and relevance of alcohol dehydrogenase enantioselectivities on biotechnological applications. *Appl. Microbiol. Biotechnol.* **104**, 2897–2909

2. Zhu, D., Yang, Y., and Hua, L. (2006) Stereoselective enzymatic synthesis of chiral alcohols with the use of a carbonyl reductase from Candida magnoliae with anti-prelog enantioselectivity. *J. Org. Chem.* **71**, 4202–4205

3. Sührer, I., Haslbeck, M., and Castiglione, K. (2014) Asymmetric synthesis of a fluoxetine precursor with an artificial fusion protein of a ketoreductase and a formate dehydrogenase. *Process Biochem.* **49**, 1527–1532

4. Moummou, H., Kallberg, Y., Tonfack, L. B., Persson, B., and van der Rest, B. (2012) The Plant Short-Chain Dehydrogenase (SDR) superfamily: Genome-wide inventory and diversification patterns. *BMC Plant Biol.* 10.1186/1471-2229-12-219

5. Persson, B., Kallberg, Y., Bray, J. E., Bruford, E., Dellaporta, S. L., Favia, A. D., Duarte, R. G., Jörnvall, H., Kavanagh, K. L., Kedishvili, N., Kisiela, M., Maser, E., Mindnich, R., Orchard, S., Penning, T. M., Thornton, J. M., Adamski, J., and Oppermann, U. (2009) The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.* **178**, 94–98

6. He, W., Gauri, M., Li, T., Wang, R., and Lin, S. X. (2016) Current knowledge of the multifunctional 17β-hydroxysteroid dehydrogenase type 1 (HSD17B1). *Gene*. **588**, 54–61

7. Yu, S., Sun, Q., Wu, J., Zhao, P., Sun, Y., and Guo, Z. (2021) Genome-Wide Identification and Characterization of Short-Chain Dehydrogenase/Reductase (SDR) Gene Family in Medicago truncatula. *Int. J. Mol. Sci. 2021, Vol. 22, Page 9498*. **22**, 9498

8. Itoh, N. (2014) Use of the anti-Prelog stereospecific alcohol dehydrogenase from Leifsonia and Pseudomonas for producing chiral alcohols. *Appl. Microbiol. Biotechnol.* **98**, 3889–3904

9. Qin, F., Qin, B., Zhang, W., Liu, Y., Su, X., Zhu, T., Ouyang, J., Guo, J., Li, Y., Zhang, F., Tang, J., Jia, X., and You, S. (2018) Discovery of a Switch between Prelog and Anti-Prelog Reduction toward Halogen-Substituted Acetophenones in Short-Chain Dehydrogenase/Reductases. *ACS Catal.* **8**, 6012–6020

10. O'Brien, P. J., and Herschlag, D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* 10.1016/S1074-5521(99)80033-7

11. Hedlund, J., Jörnvall, H., and Persson, B. (2010) Subdivision of the MDR superfamily of medium-chain dehydrogenases/reductases through iterative hidden Markov model refinement. *BMC Bioinformatics*. 10.1186/1471-2105-11-534

12. Jackson, B., Brocker, C., Thompson, D. C., Black, W., Vasiliou, K., Nebert, D. W., and Vasiliou, V. (2011) Update on the aldehyde dehydrogenase gene (ALDH) superfamily. *Hum. Genomics*. **5**, 283–303

13. Gaona-López, C., Julián-Sánchez, A., and Riveros-Rosas, H. (2016) Diversity and evolutionary analysis of iron-containing (Type-III) alcohol dehydrogenases in eukaryotes. *PLoS One*. 10.1371/journal.pone.0166851

14. Gräff, M., Buchholz, P. C. F., Stockinger, P., Bommarius, B., Bommarius, A. S., and Pleiss, J. (2019) The Short-chain Dehydrogenase/Reductase Engineering Database (SDRED): A classification and analysis system for a highly diverse enzyme family. *Proteins Struct. Funct. Bioinforma.* **87**, 443–451

15. Xu, Y., Verma, D., Sheridan, R. P., Liaw, A., Ma, J., Marshall, N. M., McIntosh, J., Sherer, E. C., Svetnik, V., and Johnston, J. M. (2020) Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* **60**, 2773–2790

16. Fang, J. (2019) A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Brief. Bioinform.* **21**, 1285–1292

17. Kallberg, Y., Oppermann, U., and Persson, B. (2010) Classification of the short-chain dehydrogenase/reductase superfamily using hidden Markov models. *FEBS J.* **277**, 2375–2386

18. Jörnvall, H., Krook, M., Persson, B., Atrian, S., Gonzàlez-Duarte, R., Jeffery, J., and Ghosh, D. (1995) Short-Chain Dehydrogenases/Reductases (SDR). *Biochemistry*. **34**, 6003–6013

19. Andersson, A., Jordan, D., Schneider, G., and Lindqvist, Y. (1997) A flexible lid controls access to the active site in 1,3,8-trihydroxynaphthalene reductase. *FEBS Lett.* **400**, 173–176

20. Hou, J., Wojciechowska, K., Zheng, H., Chruszcz, M., Cooper, D. R., Cymborowski, M., Skarina, T., Gordon, E., Luo, H., Savchenko, A., and Minor, W. (2012) Structure of a short-chain dehydrogenase/reductase from Bacillus anthracis. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **68**, 632–637

21. Bonnett, S. A., Whicher, J. R., Papireddy, K., Florova, G., Smith, J. L., and Reynolds, K. A. (2013) Structural and stereochemical analysis of a modular polyketide synthase ketoreductase domain required for the generation of a cis-alkene. *Chem. Biol.* **20**, 772–783

22. Biswas, R., Singh, B. K., Dutta, D., Basak, A., and Das, A. K. (2018) Domain swapping between FabGs deciphers the structural determinant for in-solution oligomerization and substrate binding. *Biophys. Chem.* **237**, 9–21

23. Reductase, S. D., Oa, S. R. W., and Brandt, W. (2007) Molecular Modeling and Site-Directed Mutagenesis Reveal the Benzylisoquinoline Binding Site of the. **143**, 1493–1503

24. Vihinen, M. (1987) Relationship of protein flexibility to thermostability. *Protein Eng. Des. Sel.* **1**, 477–480

25. Olufsen, M., Smalås, A. O., Moe, E., and Brandsdal, B. O. (2005) Increased Flexibility as a Strategy for Cold Adaptation: A COMPARATIVE MOLECULAR DYNAMICS STUDY OF COLD- AND WARM-ACTIVE URACIL DNA GLYCOSYLASE *. *J. Biol. Chem.* **280**, 18042–18048

26. Benrezkallah, D., Sediki, H., and Krallafa, A. M. (2020) The Structural Flexibility of Cold- and Warm-Adapted Enzymes (Endonucleases I) by Molecular Dynamics Simulation. *Chem. Proc. 2021, Vol. 3, Page 128*. **3**, 128

27. Závodszky, P., Kardos, J., Svingor, Á., and Petsko, G. A. (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl. Acad. Sci.* **95**, 7406–7411

28. Dong, Y. W., Liao, M. L., Meng, X. L., and Somero, G. N. (2018) Structural flexibility and protein adaptation to temperature: Molecular dynamics analysis of malate dehydrogenases of marine molluscs. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1274–1279

29. Choudhuri, S. (2014) Fundamentals of Molecular Evolution. *undefined*. 10.1016/B978-0-12-410471-6.00002-5

30. Mendivil Ramos, O., and Ferrier, D. E. K. (2012) Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. *Int. J. Evol. Biol.* **2012**, 1–10

31. Tria, F. D. K., and Martin, W. F. (2021) Gene Duplications Are At Least 50 Times Less Frequent than Gene Transfers in Prokaryotic Genomes. *Genome Biol. Evol.* 10.1093/GBE/EVAB224

32. Kara, S., Schrittwieser, J. H., Hollmann, F., and Ansorge-Schumacher, M. B. (2014) Recent trends and novel concepts in cofactor-dependent biotransformations. *Appl. Microbiol. Biotechnol.* **98**, 1517–1529

33. Uppada, V., Bhaduri, S., and Noronha, S. B. (2014) Cofactor regeneration - an important aspect of biocatalysis. *Curr. Sci.* **106**, 946–957

34. Shanbhag, A. P. (2019) FabG: from a core to circumstantial catalyst. *Biotechnol. Lett.* **41**, 675–688

35. Zheng, J., Taylor, C. A., Piasecki, S. K., and Keatinge-Clay, A. T. (2010) Structural and functional analysis of A-type ketoreductases from the amphotericin modular polyketide synthase. *Structure*. **18**, 913–922

36. Blaise, M., Van Wyk, N., Banères-Roquet, F., Guérardel, Y., and Kremer, L. (2017) Binding of NADP+

triggers an open-to-closed transition in a mycobacterial FabG β-ketoacyl-ACP reductase. *Biochem. J.* **474**, 907–921

37. Liu, J. S., Kuan, Y. C., Tsou, Y., Lin, T. Y., Hsu, W. H., Yang, M. Te, Lin, J. Y., and Wang, W. C. (2018) Structure-guided design of Serratia marcescens short-chain dehydrogenase/reductase for stereoselective synthesis of (R)-phenylephrine. *Sci. Rep.* **8**, 1–11

38. Taguchi, K., Aoyagi, Y., Matsusaki, H., Fukui, T., and Doi, Y. (1999) Co-expression of 3-ketoacyl-ACP reductase and polyhydroxyalkanoate synthase genes induces PHA production in Escherichia coli HB101 strain. *FEMS Microbiol. Lett.* **176**, 183–190

39. Banerjee, A. K., Ravi, V., Murty, U. S. N., Shanbhag, A. P., and Prasanna, V. L. (2013) Keratin protein property based classification of mammals and non-mammals using machine learning techniques. *Comput. Biol. Med.* **43**, 889–899

40. AlQuraishi, M. (2019) ProteinNet: A standardized data set for machine learning of protein structure. *BMC Bioinformatics*. **20**, 1–10

41. Chakravorty, D., Khan, M., and Patra, S. (2015) Thermostability of Proteins Revisited Through Machine Learning Methodologies: From Nucleotide Sequence to Structure. *Curr. Biotechnol.* **6**, 39–49

42. Sharma, A., Bagler, G., and Bera, D. (2019) Supervised learning of protein thermal stability using sequence mining and distribution statistics of network centrality. *BioRxiv*. 10.1101/777177

43. Glasner, M. E., Truong, D. P., and Morse, B. C. (2020) How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation. *FEBS J.* **287**, 1323–1342

44. Conant, G. C., and Wagner, A. (2005) The rarity of gene shuffling in conserved genes. *Genome Biol.* 10.1186/gb-2005-6-6-r50

45. Zhang, R., Xu, Y., and Xiao, R. (2015) Redesigning alcohol dehydrogenases/reductases for more efficient biosynthesis of enantiopure isomers. *Biotechnol. Adv.* **33**, 1671–1684

46. Haq, S. F., Shanbhag, A. P., Karthikeyan, S., Hassan, I., Thanukrishnan, K., Ashok, A., Sukumaran, S., Ramaswamy, S., Bharatham, N., Datta, S., Samant, S., and Katagihallimath, N. (2018) A strategy to identify a ketoreductase that preferentially synthesizes pharmaceutically relevant (S)-alcohols using whole-cell biotransformation. *Microb. Cell Fact.* **17**, 1–14

47. Vander Meersche, Y., Cretin, G., de Brevern, A. G., Gelly, J. C., and Galochkina, T. (2021) MEDUSA: Prediction of Protein Flexibility from Sequence. *J. Mol. Biol.* **433**, 166882

48. De Brevern, A. G., Bornot, A., Craveur, P., Etchebest, C., and Gelly, J. C. (2012) PredyFlexy: flexibility and local structure prediction from sequence. *Nucleic Acids Res.* **40**, W317

49. Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., and Mori, H. (2005) Complete set of ORF clones of Escherichia coli ASKA library (A complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res.* **12**, 291–299

Figure 1: Schematic explaining the prospecting of enzymes using machine learning for clustering candidates with broad substrate activity. Standard phylogenetic classification helps in finding orthologs/paralogs but may not classify enzymes for substrate activities. Due to the similarity in enzymes the short-chain dehydrogenases are classified using supervised learning algorithms to denote true and false positives. Subsequently, they are clustered into enzymes that have a higher probability of being promiscuous. Unknown members closer to the reported promiscuous candidates are purified for in vitro analysis.

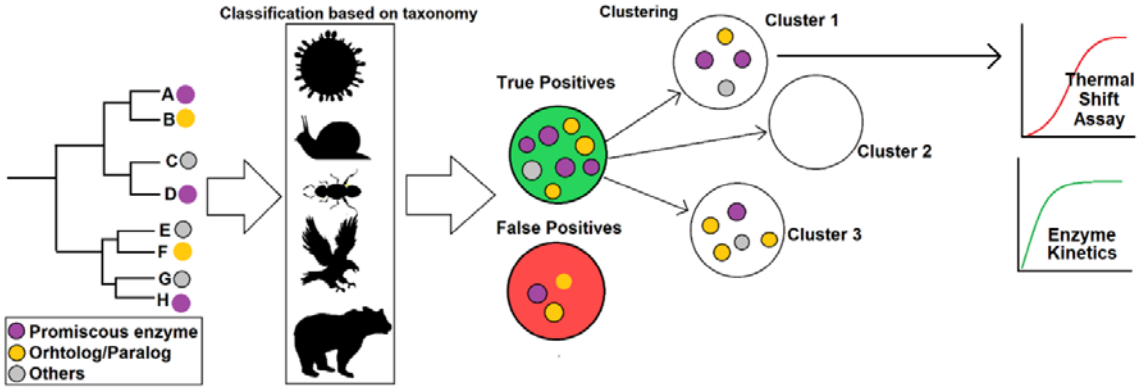


Figure 2: The changes in the lid-loop of the Short-chain dehydrogenases are invariably based on the amino acid composition of the loop affecting the overall architecture and oligomerization of the enzyme. They can exist as Dimers (subsequently Tetramers) in Classical type enzymes and as Monomers in Extended type enzymes

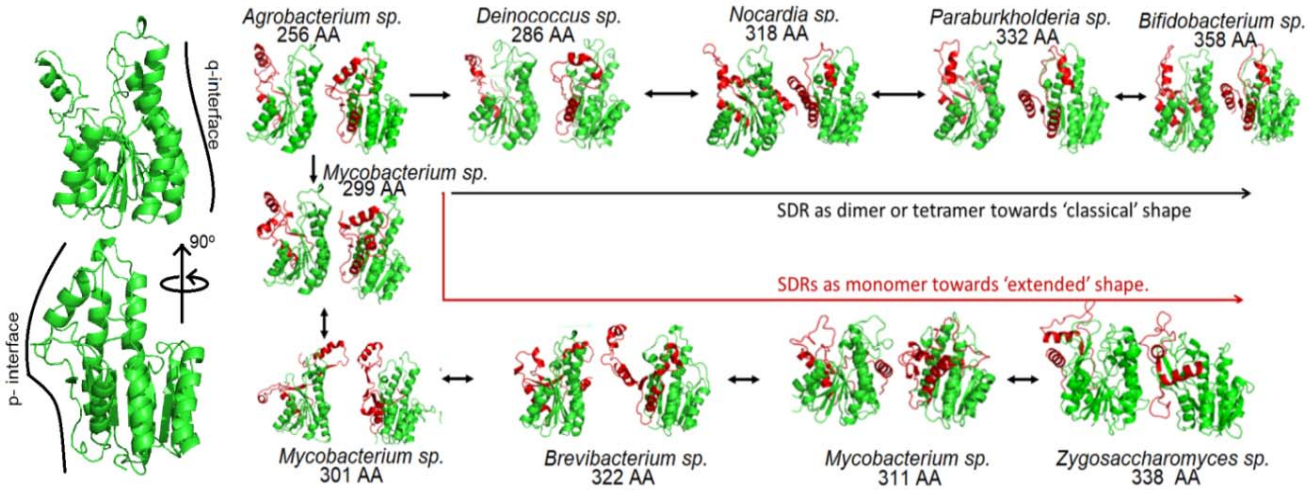
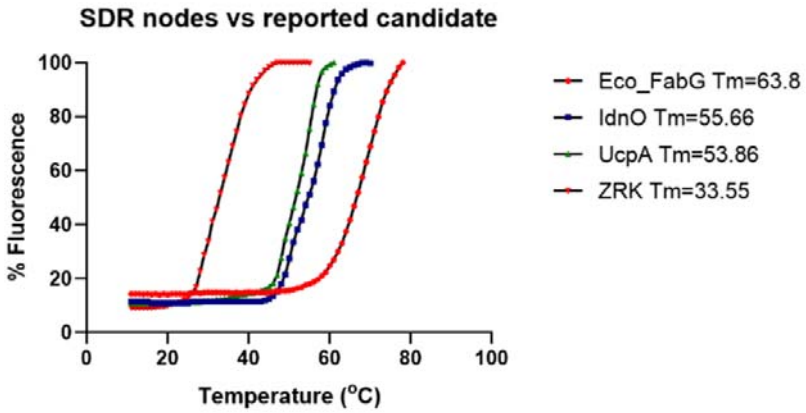
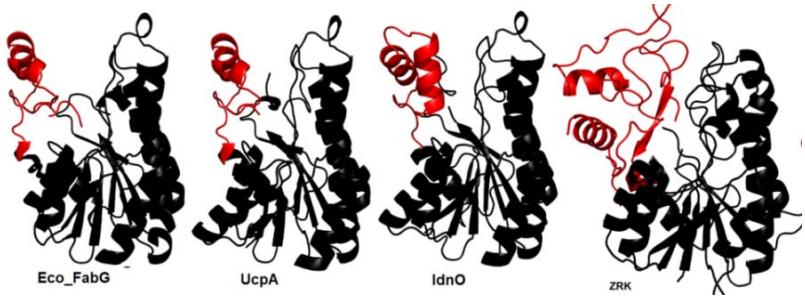


Figure 3: Repurposing of ketoreductases for production of chiral alcohols. A. Differential scanning fluorimetry of short chain dehydrogenases. B. Lid-loop structure (red) of chosen SDRs. C. Heat map of the catalytic efficiency of chosen SDRs Units: km=uM kcat=uM/min.

A.



B.



C.

	km				kcat				kcat/km			
Compound	**Ecoli_FabG**	**UcpA**	**IdnO**	**ZRK**	**Ecoli_FabG**	**UcpA**	**IdnO**	**ZRK**	**Ecoli_FabG**	**UcpA**	**IdnO**	**ZRK**
Acetone	412.1	284.5	707.8	462.8	61930.0	59290.0	84400.0	62110.0	150.3	208.4	119.2	134.2
Dolastatin	358.8	265.3	156.3	210.6	34500.0	34830.0	35750.0	67480.0	96.2	131.3	228.7	320.4
MA 20565	714.3	976.0	886.2	317.7	48470.0	72200.0	78580.0	66440.0	67.9	74.0	88.7	209.1
Barnipidine	444.0	445.2	429.8	189.7	17430.0	20000.0	24020.0	39960.0	39.3	44.9	55.9	210.6
Sulopenem	921.4	807.2	557.0	312.8	61040.0	55960.0	56670.0	70390.0	66.2	69.3	101.7	225.0
Rivastigmine	1919.0	780.6	440.5	495.9	81120.0	61530.0	57420.0	77970.0	42.3	78.8	130.4	157.2
Crizotinib	1041.0	1021.0	1635.0	338.5	40590.0	32100.0	51970.0	51580.0	39.0	31.4	31.8	152.4
COBE	42.3	64.7	36.9	5.4	114300.0	114900.0	81830.0	63950.0	2700.9	1776.4	2218.2	11785.8
Aprepitant	40.2	206.4	114.9	186.2	2003.0	2355.0	1951.0	22590.0	49.8	11.4	17.0	121.3
Sitagliptin	635.9	598.7	650.6	345.6	26570.0	13980.0	12620.0	74570.0	41.8	23.4	19.4	215.8
Acetophenone	284.2	467.5	378.7	760.9	51370.0	73780.0	63010.0	91040.0	180.8	157.8	166.4	119.6
Talampanel	81.4	49.4	8.2	289.1	19.8	29.3	7.5	74740.0	0.2	0.6	0.9	258.5

Figure 4: Feature selection of the physicochemical properties using correlation matrix for classifying SDRs. Positive integers depict a positive correlation and negative integers represent vice versa. Top 20 physicochemical properties having correlation values farther from zero were chosen for classification.

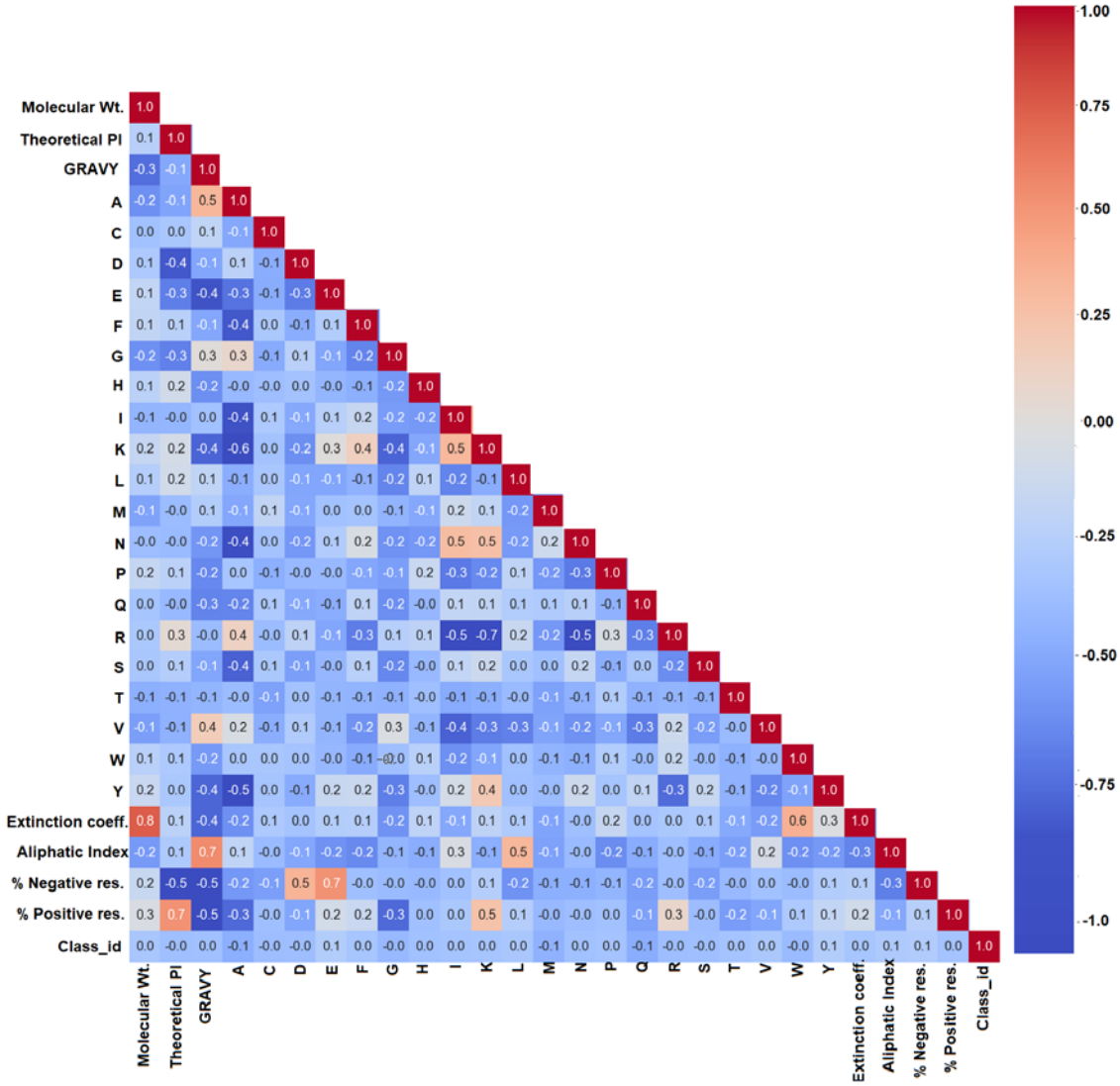
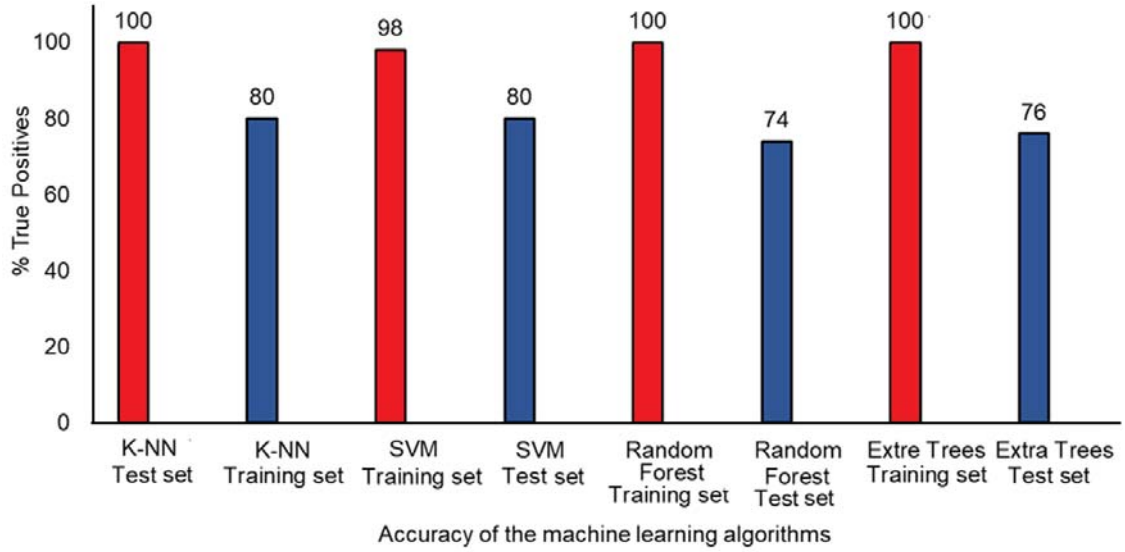


Figure 5: A. An example of a confusion matrix portraying true positives (reverse diagonal marked in red) obtained by SVM for SDRs. B: Graph depicting the accuracy of Support vector machines (SVM), Extra trees, Random Forest, and k-Nearest Neighbour method in classifying SDRs based on physicochemical parameters. C. Principal component analysis (PCA) of SDRs where the chosen members for in vitro analysis are encircled, and the thermophilic enzymes are depicted in orange

A.

	Class_ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Class_ID	Actual	Ma	Firm	Fun	Prot	Pla	Fis	Am	Art	Pro	Actin	Arc	Ne	An	Mo	Pla	Alg	Ba	Ch	Pla	Cya	Un	Fus	Ch	Ch	Ba	Sp	Ve	De	Sy	Th	Act	Nit	Ch	Arr	Te	Aves
0	Mammal	29	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Firmicutes	1	133	5	19	0	1	0	1	0	1	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2	Fungi	3	1	195	31	0	0	0	0	0	24	0	1	0	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
3	Proteobacteria	0	12	16	955	1	0	0	1	0	44	2	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0	1	0	1	0	2	0	0	1	0	0
4	Plant	0	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Fish	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
6	Amphibian	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Arthropod	3	0	1	0	1	2	0	7	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Protozoa	1	1	1	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Actinobacteria	0	1	24	42	0	0	0	0	0	341	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Archaea	0	0	1	4	0	0	0	0	0	1	23	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Nematode	0	0	1	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Annelida	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Molluscs	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Platyhelmint	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	Algae	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	Bacterioidetes	0	8	0	8	0	0	0	0	0	1	1	0	0	0	0	0	29	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
17	Chloroflexi	0	0	0	6	0	0	0	0	0	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Planctomycetes	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Cyanobacteria	0	0	3	5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Unknown	0	2	0	6	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Fusobacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	Chlamydiae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Chlorobi	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
24	Bacilli	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Spirochaete	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0
26	Verrucomicrobia	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	Deinococcus	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	Synergistetes	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Thermotogae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
30	Actinomycetes	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	Nitrospirae	0	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
32	Chlorofelxi	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	Armatimonadetes	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	Tenericutes	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	Aves	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5

B.



C.

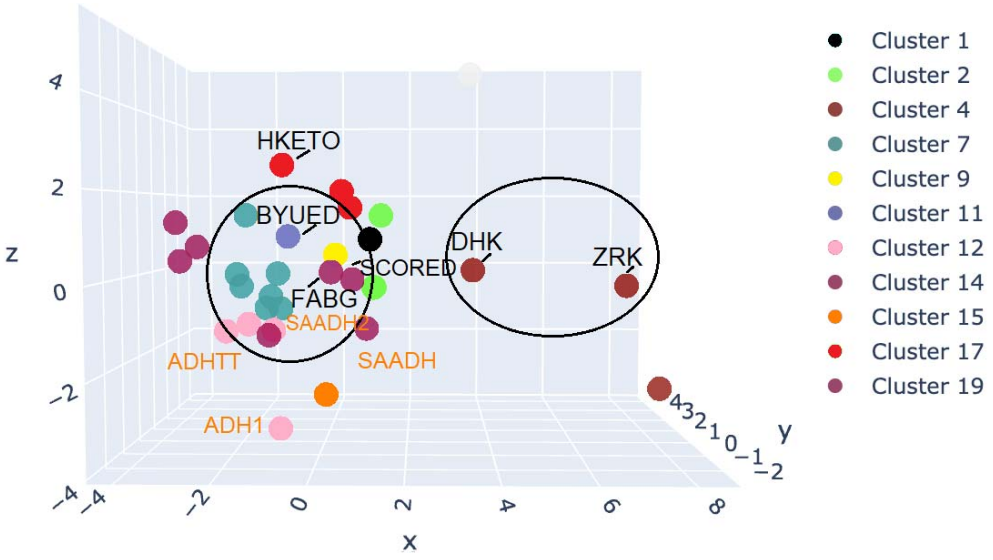
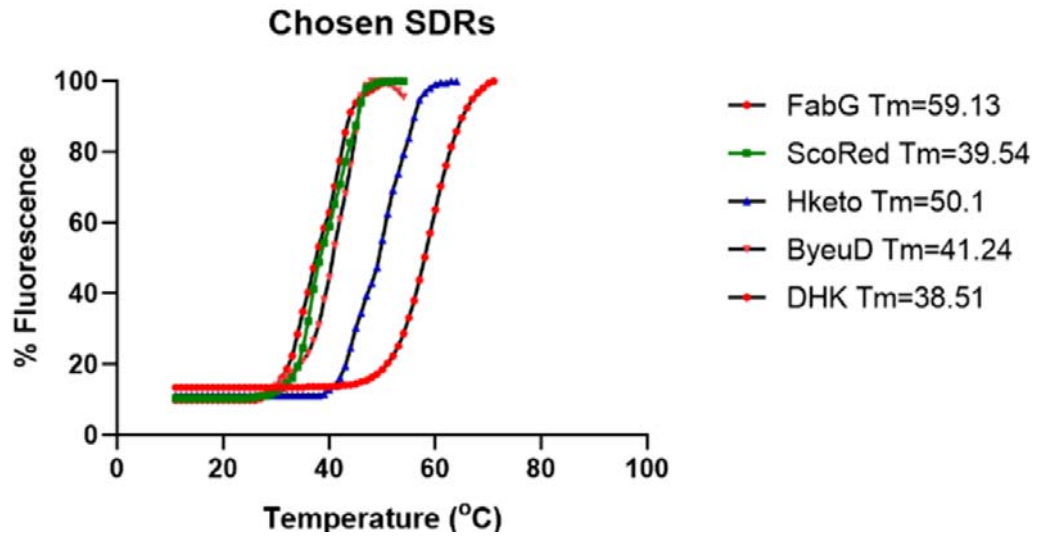
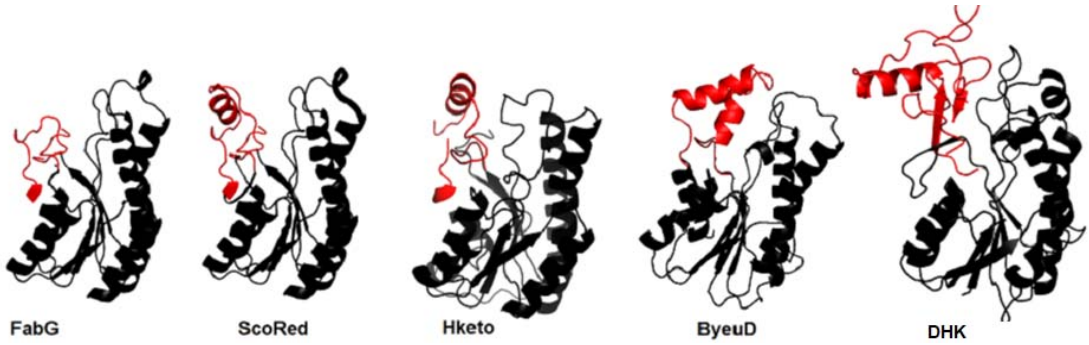

HKETO
BYUED
SCORED
DHK
ZRK
FABG
SAADH2
ADHTT
SAADH
ADH1
x
y
z
Cluster 1
Cluster 2
Cluster 4
Cluster 7
Cluster 9
Cluster 11
Cluster 12
Cluster 14
Cluster 15
Cluster 17
Cluster 19

Figure 6: Relationship between flexibility and activity of the data mined SDRs A. Differential scanning fluorimetry of Short-chain dehydrogenases. B. The increasing complexity of the Lid-loop structure (red) of chosen SDRs. C. Heat map of the catalytic efficiency of the datamined SDRs Units: km=uM kcat=uM/min.

A.



B.



C.

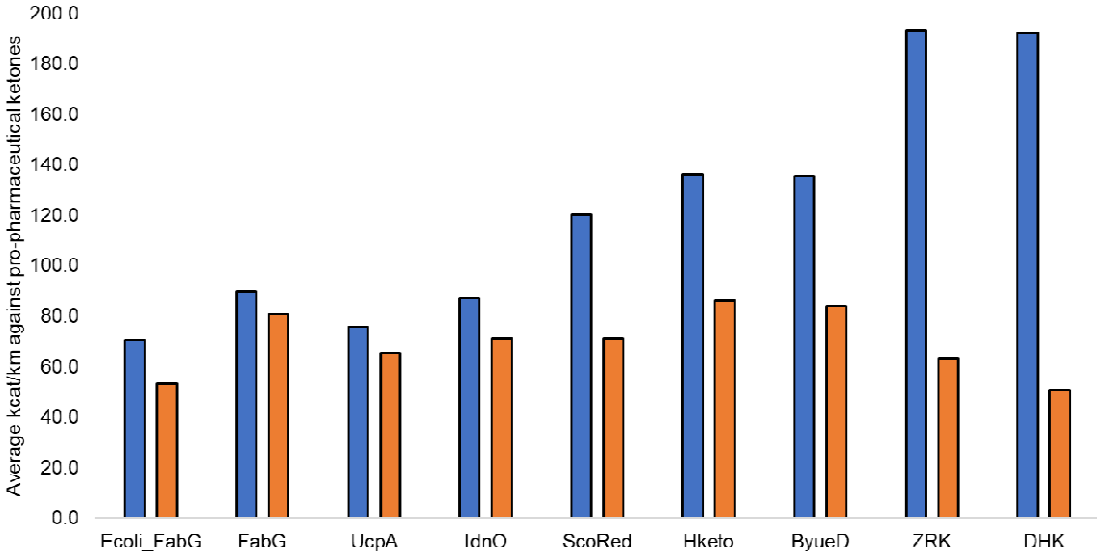
km	Ecoli_FabG	UcpA	IdnO	ScoRed	Hketo	ByueD	ZRK	DHK
Acetone	412.1	284.5	707.8	463.9	153.6	179.1	462.8	292.2
Dolastatin	358.8	265.3	156.3	148.3	312.4	501.2	210.6	356.9
MA 20565	714.3	976.0	886.2	710.7	513.8	635.5	317.7	296.6
Barnipidine	444.0	445.2	429.8	359.9	500.6	379.4	189.7	256.5
Sulopenem	921.4	807.2	557.0	353.7	612.1	517.0	312.8	420.8
Rivastigmine	1919.0	780.6	440.5	1516.0	973.3	754.0	495.9	273.5
Crizotinib	1041.0	1021.0	1635.0	815.2	752.7	554.3	338.5	247.0
COBE	42.3	64.7	36.9	11.1	14.3	14.6	5.4	9.6
Aprepitant	40.2	206.4	114.9	178.1	168.3	197.1	186.2	194.7
Sitagliptin	635.9	598.7	650.6	646.3	525.1	450.7	345.6	596.7
Acetophenone	284.2	467.5	378.7	358.3	315.2	429.1	760.9	328.2
Talampanel	81.4	49.4	8.2	74.2	390.8	536.7	289.1	319.0

kcat	Ecoli_FabG	UcpA	IdnO	ScoRed	Hketo	ByueD	ZRK	DHK
Acetone	61930	59290	84400	77210	56540	68810	62110	71910
Dolastatin	34500	34830	35750	42400	47510	47710	67480	69330
MA 20565	48470	72200	78580	56790	76310	72140	66440	70390
Barnipidine	17430	20000	24020	26940	40500	39640	39960	39980
Sulopenem	61040	55960	56670	53230	68190	66710	70390	69180
Rivastigmine	81120	61530	57420	99450	87760	84890	77970	71760
Crizotinib	40590	32100	51970	44820	60340	56330	51580	49780
COBE	114300	114900	81830	58250	72390	63490	63950	67810
Aprepitant	2003	2355	1951	10510	12800	16390	22590	18620
Sitagliptin	26570	13980	12620	76840	30360	52520	74570	86400
Acetophenone	51370	73780	63010	65830	52400	60930	91040	58830
Talampanel	19.77	29.34	7.51	6064	64370	58230	74740	74620

kcat/km	Ecoli_FabG	UcpA	IdnO	ScoRed	Hketo	ByueD	ZRK	DHK
Acetone	150.3	208.4	119.2	166.4	368.1	384.2	134.2	246.1
Dolastatin	96.2	131.3	228.7	285.9	152.1	95.2	320.4	194.3
MA 20565	67.9	74.0	88.7	79.9	148.5	113.5	209.1	237.3
Barnipidine	39.3	44.9	55.9	74.9	80.9	104.5	210.6	155.9
Sulopenem	66.2	69.3	101.7	150.5	111.4	129.0	225.0	164.4
Rivastigmine	42.3	78.8	130.4	65.6	90.2	112.6	157.2	262.4
Crizotinib	39.0	31.4	31.8	55.0	80.2	101.6	152.4	201.5
COBE	2700.9	1776.4	2218.2	5238.3	5058.7	4354.6	11785.8	7035.7
Aprepitant	49.8	11.4	17.0	59.0	76.1	83.2	121.3	95.6
Sitagliptin	41.8	23.4	19.4	118.9	57.8	116.5	215.8	144.8
Acetophenone	180.8	157.8	166.4	183.7	166.2	142.0	119.6	179.3
Talampanel	0.2	0.6	0.9	81.7	164.7	108.5	258.5	233.9

Figure 7: A. Average catalytic efficiency of ketoreductases with their respective standard deviations. The graph depicts an indirect measure of the flexibility of the active site as the change is quite similar to each other despite the large difference in catalytic efficiency B. The geometry of the active site and substrate-binding site between the chosen Short-chain dehydrogenases shows a similarity in the distances between the participating residues.

A.



B.

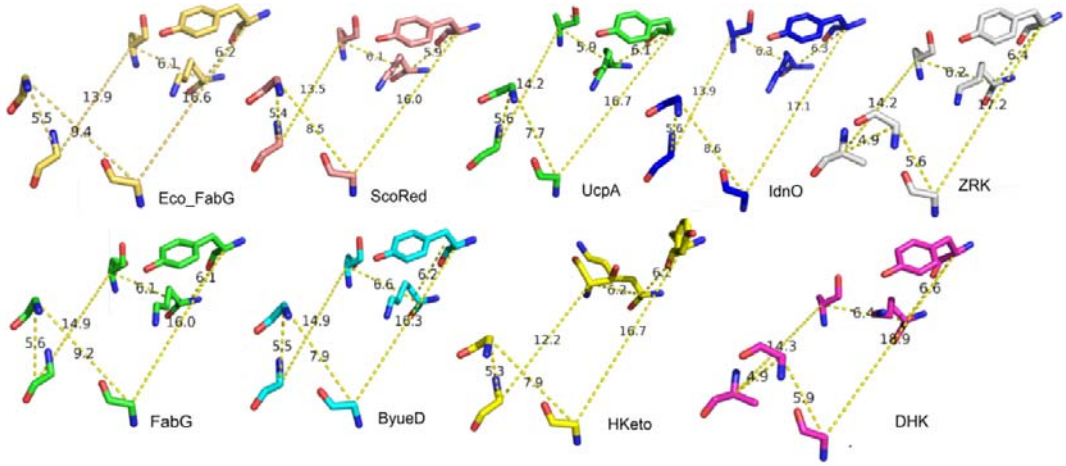


Table 1: Clustering of the reported 23 SDRs from the total 81,014 members. Characteristics such as promiscuity and sensitivity are present in the same clusters. Some of the unknown enzymes (highlighted in yellow) lying in clusters having promiscuous and non-promiscuous members were chosen for in vitro analysis.

Serial	Clusters	Activity	Sensitivity	Phylum
>pr|BgADH3	1	Very low promiscuity	Mesophilic	*Burkholderia gladioli BSR3 (CCTCC M 2012379)*
>pr|SmCR	2	Promiscuous	Mesophilic	*Starmerella magnoliae*
>pr|Kacr1	2	Unknown	Mesophilic	*Kluyveromyces aestuarii*
>pr|CmScr1/CaCR	2	Very low promiscuity	Mesophilic	*Candida magnoliae*
>pr|SsCR	4	Promiscuous	Mesophilic	*Sporobolomyces salmonicolor*
>pr|ZRK	4	Promiscuous	Mesophilic	*Zygosaccharomyces rouxii*
>pr|DHK	4	Unknown	Mesophilic	*Debaryomyces hansenii*
>pr|SaADH2	7	Low promiscuity	Thermophilic	*Sulfolobus acidocaldarius strain DSM 639*
>pr|LbADH	7	Promiscuous	Mesophilic	*Lactobacillus brevis ATCC 14869*
>pr|LkADH	7	Promiscuous	Mesophilic	*Lactobacillus kefiri/ Lactobacillus kefir DSM 20587*
>pr|AcCR	7	Promiscuous	Mesophilic	*Acetobacter sp. CCTCC M209061*
>pr|AcCR 2	7	Promiscuous	Mesophilic	*Acetobacter sp. CCTCC M209062*
>pr|LcSDR	7	Unknown	Mesophilic	*Lactobacillus composti*
>pr|Gox2036	7	Very low promiscuity	Mesophilic	*Gluconobacter oxydans 621H*
>pr|LbSDR	9	Unknown	Mesophilic	*Lactobacillus brevis ATCC 367*
>pr|AfPDH	11	Promiscuous	Mesophilic	*Candida maris*
>pr|ByeuD	11	Unknown	Mesophilic	*Bacillus subtils ECU0013*
>pr|ADH1	12	Low promiscuity	Thermophilic	*Pyrococcus furiosus DSM 3638*
>pr|AdhTt	12	Low promiscuity	Thermophilic	*Thermus thermophilus HB27*
>pr|RasADH	12	Promiscuous	Mesophilic	*Ralstonia sp. DSMZ 6428*
>pr|SmADH31	12	Very low promiscuity	Mesophilic	*Stenotrophomonas maltophilia*
>pr|LsADH	14	Promiscuous	Mesophilic	*Leifsonia sp. strain S749*
>pr|ScCR or ScoRed	14	Promiscuous	Mesophilic	*Streptomyces coelicolor*
>pr|FabG_Syn	14	Promiscuous	Mesophilic	*Synechococcus elongatus*
>pr|CaCR	15	Very low promiscuity	Mesophilic	*Candida albicans SC5314*
>pr|Hketo	17	Unknown	Mesophilic	*Hansenula polymorpha*
>pr|PsCRI	17	Low promiscuity	Mesophilic	*Scheffersomyces stipitis CBS 6054*
>pr|PsCRII	17	Low promiscuity	Mesophilic	*Scheffersomyces stipitis CBS 6054*
>pr|SaADH	19	Low promiscuity	Thermophilic	*Sulfolobus acidocaldarius strain DSM 639*
>pr|ChKRED20	19	Promiscuous	Mesophilic	*Chryseobacterium sp. CA49*
>pr|SSCR	19	Very low promiscuity	Mesophilic	*Scheffersomyces stipitis CBS 6054*
>pr|Scr2 or ScrII	19	Very low promiscuity	Mesophilic	*Candida parapsilosis*