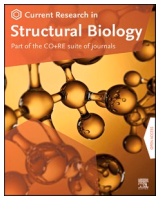
Current Research in Structural Biology 6 (2023) 100108

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology



Research Article

# Structural similarities between SAM and ATP recognition motifs and detection of ATP binding in a SAM binding DNA methyltransferase

Santhosh Sankar [a,1], Preeti Preeti [a,1], Kavya Ravikumar [a], Amrendra Kumar [a], Yedu Prasad [a], Sukriti Pal [a], Desirazu N. Rao [a], Handanahal S. Savithri [a], Nagasuma Chandra [a,b,*]

[a] Department of Biochemistry, Indian Institute of Science, Bangalore, 560012, Karnataka, India
[b] Department of BioEngineering, Indian Institute of Science, Bangalore, 560012, Karnataka, India

ARTICLE INFO

Handling Editor: Dr CS Verma

ABSTRACT

S-adenosylmethionine (SAM) is a ubiquitous co-factor that serves as a donor for methylation reactions and additionally serves as a donor of other functional groups such as amino and ribosyl moieties in a variety of other biochemical reactions. Such versatility in function is enabled by the ability of SAM to be recognized by a wide variety of protein molecules that vary in their sequences and structural folds. To understand what gives rise to specific SAM binding in diverse proteins, we set out to study if there are any structural patterns at their binding sites. A comprehensive analysis of structures of the binding sites of SAM by all-pair comparison and clustering, indicated the presence of 4 different site-types, only one among them being well studied. For each site-type we decipher the common minimum principle involved in SAM recognition by diverse proteins and derive structural motifs that are characteristic of SAM binding. The presence of the structural motifs with precise three-dimensional arrangement of amino acids in SAM sites that appear to have evolved independently, indicates that these are winning arrangements of residues to bring about SAM recognition. Further, we find high similarity between one of the SAM site types and a well known ATP binding site type. We demonstrate using in vitro experiments that a known SAM binding protein, *Hpy*AII.M1, a type 2 methyltransferase can bind and hydrolyse ATP. We find common structural motifs that explain this, further supported through site-directed mutagenesis. Observation of similar motifs for binding two of the most ubiquitous ligands in multiple protein families with diverse sequences and structural folds presents compelling evidence at the molecular level in favour of convergent evolution.

## 1. Introduction

S-Adenosylmethionine (SAM) is an important molecule, essential for cellular metabolism and viability, found ubiquitously in living organisms. It is one of the most frequently encountered small molecule ligands that enzymes utilize. In cellular metabolism, a number of enzymes that perform essential functions, use SAM as a functional group donor. The primary function of SAM is to facilitate methylation reactions by covalently attaching methyl groups to DNA, protein and other small molecules (Greer and Shi, 2012). Besides this, SAM is also known to be involved in the donation of its other chemical moieties in different reactions (Fontecave et al., 2004). For example, it is known to be a methylene group donor to fatty acids in bacteria (Taylor and Cronan, 1979), an amino group donor in biotin biosynthesis catalyzed by 7, 8-diaminopelargonic acid (DAPA) synthase (Stoner and Eisenberg, 1975), and a ribosyl group donor in tRNA modification (Slany et al., 1994). SAM-mediated enzymatic reactions are a key component of vitamin synthesis, polyamine synthesis and DNA synthesis (Wang and Frey, 2007; Pegg, 1986).

To achieve a wide array of functions, SAM has to be recognized by a variety of proteins. SAM is indeed known to exhibit high versatility in the types of proteins it binds to (Bauerle et al., 2015; Trausch et al., 2014). Proteins that bind to SAM are known to belong to completely different sequence families and adopt different structural folds. The methyltransferases form a major superfamily, but in addition there are several other proteins that recruit SAM ligands regularly for methylation of biomolecules. Methyltransferases (MTases) themselves have been grouped into 5 distinct structural classes based on their catalytic

* Corresponding author. Department of Biochemistry, Indian Institute of Science, Bangalore, 560012, Karnataka, India.
*E-mail address:* nchandra@iisc.ac.in (N. Chandra).
[1] Equal Contribution from the authors.

https://doi.org/10.1016/j.crstbi.2023.100108
Received 8 July 2023; Received in revised form 5 October 2023; Accepted 24 October 2023
Available online 17 November 2023

domains, which are structurally different from one another (Rossmann-like α/β, TIM barrel, tetrapyrrole methylase α/β, SPOUT α/β and SET domain β folds) (Schubert et al., 2003). Even within a given class, the sequence similarity between proteins are reported to be as low as 10%. Despite these differences, it is intriguing that these diverse proteins bind SAM. To understand what generates SAM binding capability in these diverse proteins, it is important to identify the determinants of SAM recognition. Knowledge of such determinants will enable large-scale identification of SAM-binding proteins, genome-wide annotation and in fact lay a foundation for seeking a holistic understanding of SAM physiology. It will also enable us to address questions on the origin of specificity of recognition of SAM. Further, structural motifs guide the design for incorporating SAM binding in proteins of interest for use in synthetic biology and biotechnology fields. Specifically, it will enable protein engineering design for altering the substrate specificity to bind a new class of ligands in a given protein, which has remained to be a challenging problem, despite extensive efforts in that direction.

One logical explanation for a given small molecule ligand to bind to diverse proteins can arise if such proteins, despite differences in their sequences and overall structural folds, share significant commonalities in their binding sites. Similarities merely at the binding site level are well documented in several proteins, suggestive of convergent evolution (eg., trypsin, chymotrypsin and nucleotide binding proteins) (Gherardini et al., 2007; Gorbalenya et al., 1989; Narunsky et al., 2020). In SAM binding proteins too, this aspect has been tested earlier through a search for conserved sequence motifs, and two different motifs have been identified in two separate studies, Motif-1, a glycine rich motif GxGxG and a hallmark of SAM binding sites, observed to interact with the methionine moiety of SAM ligands and Motif-2, disconnected from the first, comprising of acidic residues that makes bidentate hydrogen bonding with the O3' and O4' oxygen atoms of SAM ribose group (Liscombe et al., 2012; Cheng, 1995; Laurino et al., 2016). A few other independent motifs have also been reported (Kagan and Clarke, 1994; Korolev et al., 2002; Bügl et al., 2000), most of them are small sequence stretches and not generally conserved across all SAM binding proteins. Most of these studies have largely utilized sequence information, which have limited scope when it comes to proteins of diverse families. Structural information on the other hand is much better at identifying common patterns, if any, in a diverse set of proteins. There are 800 structures of proteins bound to SAM available in PDB, providing highly reliable and higher resolution information than sequences alone. A structural analysis has also been explored earlier from a ligand-centric perspective, which identified conserved structural motifs at three broad levels of sequence, structure and the ligand binding conformation (Gana et al., 2013), with an objective of facilitating functional annotation of SAM binding proteins. However, their analysis was centered on comparing proteins within each fold class and identifying conserved residues within each type and does not address why different folds are capable of binding the same ligand, nor does it account for variation in ligand presentation across folds. It also does not account for identifying motifs from spatially adjacent, but sequentially non-contiguous residues at the binding site. Development of algorithms for binding site comparisons in a fold-independent manner in the last few years have enabled us to factor in these aspects in this study that aims to decipher the structural determinants of SAM recognition.

Here we address these limitations and report a comprehensive unbiased structural bioinformatics study of SAM recognition, using all known proteins in PDB complexed with SAM. Our analysis indicates that there are largely 4 different site types capable of recognizing SAM, of which one belonging to the methyltransferases is well characterized. The other 3 motifs largely correspond to Radical SAM, SET domain and rRNA methyltransferase families. We tested the validity of the structural motifs and found they were sufficient to characterize the SAM binding proteins. Further, we found the predominant SAM motif to be closely related to ATP binding motifs and experimentally demonstrated a dual SAM and ATP binding ability in a selected protein- *H. pylori* methyltransferase, which poses a question if the two functionalities could have had a common evolutionary origin. Guided by the motif, we further enhanced ATP binding in this protein through site-directed mutagenesis. Put together, our analysis identifies multiple site architectures that can generate SAM binding ability and strongly points at convergent evolution centered on the binding functionality and provides a basis to design altered substrate specificity.

## 2. Methods

### 2.1. Dataset preparation

A search through PDB for protein structures complexed with SAM or SAH, solved by X-ray crystallography, resulted in a set of 800 entries (Burley et al., 2017). The ligand SAH (S-Adenosyl homocysteine) was also included as several proteins are crystallized with this instead of SAM. It binds to the same binding site since it is a structural analog of SAM, the by-product of SAM utilization and in some cases also a competitive inhibitor, but for simplicity, they are together referred to as SAM in the manuscript. Sequence redundancy was removed using the EMBOSS-Needle program keeping a sequence identity threshold of 50% (Rice et al., 2000). Redundancy at the fold was removed using the TMalign program (Zhang and Skolnick, 2005). This resulted in a final set of 96 proteins with diverse sequences and structures (SAM-set), which was utilized for further analysis (Fig. 1a). Pfam and SCOP databases were used to fetch protein annotations (Mistry et al., 2021; Lo Conte et al., 2000). Where fold annotations were available from the SCOP database, they are assigned directly. For those PDB entries where SCOP had no fold information, the SUPERFAMILY database was consulted to make the fold assignments.

### 2.2. Extraction and comparison of SAM binding sites

All residues in the binding site whose atom(s) were located with a distance threshold of 4.5 Å from any ligand atom were considered to define the ligand binding site in each protein (Fig. 1b). The binding site is defined as a group of residues that lie within a zone of 4.5 Å from any atom of the corresponding ligand (SAM/SAH). SiteMotif was used for comparing binding sites and estimating similarities among them (Sankar and Chandra, 2022). SiteMotif is a 3D pocket comparison algorithm which transforms input binding sites into a 2D matrix of distance differences. It then applies the graph clique detection algorithm to the distance matrices which estimates how similar two binding sites are. SiteMotif reports two scores namely M-dist$_{min}$ and M-dist$_{max}$. M-dist$_{max}$ captures the global similarities wherein M-dist$_{min}$ finds the local similarities between pockets. We previously showed that M-dist$_{min}$ $>$ 0.5 sufficiently captures the overall similarity among sites and hence we use the same cutoff throughout this work (Sankar and Chandra, 2022).

### 2.3. Site clustering and deriving SAM binding motif

Using M-dist$_{min}$ score, a projection network was constructed, where nodes represented binding sites and edges represented site similarity scores (Shannon et al., 2003). From the network, different clusters were delineated using the ClusterOne algorithm and a highest degree node in each cluster was identified as the cluster representative (Nepusz et al., 2012). SiteMotif was used to obtain the residue-wise alignment list between all sites and their cluster representatives. The alignments were used to derive conserved residue motifs for each cluster. The motif accuracy was examined by comparing it with the known SAM/SAH binding sites. The motif was used as an input for a 3D screen to identify new or previously unannotated SAM sites (Fig. 1c). For carrying out sensitivity and specificity analysis, we use FLAPP, a superfast site matching algorithm that we developed during the course of this work, in our laboratory (Sankar et al., 2022). FLAPP produces very rapid atomic level alignments, aligning a typical pair of binding sites at ~12.5 ms on a

single CPU core, and performs a PDB-wide scan in 1–2 min on a standard desktop with multiple cores. It has been rigorously validated and shown to produce accurate alignments and has also been benchmarked against the best site matching algorithms in the field. This enabled us to carry out a comparison of each motif at the PDB scale, using the entire set of non-redundant pockets derived from PDB, making the sensitivity and specificity analysis very comprehensive.

### 2.4. Cloning and site-directed mutagenesis

The gene *Hpy*AII.M1 (NCBI sequence accession NC_000915.1) of 783 bps was amplified from the genomic DNA of *H. pylori* 26695 using sequence-specific forward 5′-TTGATTTTGAATAAGATTTATATAG-3′ and reverse primers 5′-TTATTCGCATTCATTATAC-3'. The amplified PCR product was digested with restriction enzymes *BamH*1 and *Xho*1 and ligated into the multiple cloning site of the pGEX-4T2 vector. The obtained recombinant plasmid pGEX-4T2 with the insert of *Hpy*AII.M1 was entitled as (pGEX*Hpy*AII.M1) and subjected to propagation in *E. coli*. Cloned plasmid (pGEX*Hpy*AII.M1) was checked and confirmed by DNA sequencing.

The site-specific mutant of *Hpy*AII.M1 was generated using pGEX*Hpy*AII.M1 as the DNA template. Mutant I196R was constructed following the instructions provided with the QuickChange site-directed mutagenesis kit (Agilent Technologies). Mutagenic primers were designed with the introduction of a restriction site near the mutation site to confirm the desired mutant plasmid. PCR reaction was set up with Phusion HF DNA polymerase (NEB) and the mutant-specific oligonucleotides to amplify the desired mutation in the pGEX*Hpy*AII.M1 plasmid. The amplified PCR product was subjected to DpnI digestion before transformation into *E. coli Dh*5α and the mutant plasmid was confirmed by sequencing.

### 2.5. Protein expression and purification

The recombinant *Hpy*AII.M1 was transformed into Rosetta (DE3) pLysS competent cells and plated on the LB agar containing chloramphenicol and ampicillin of 34 $\mu gml^{-1}$ and 100 $\mu gml^{-1}$, respectively. Single colony harboring *Hpy*AII.M1 plasmid was used for seeding the primary culture supplemented with antibiotics. An amount of 0.5% overnight grown culture was used as inoculum for 4l LB broth and grown at 37 °C with continuous shaking at 180 rpm. Once the optical density $OD_{600}$ reached 0.8, the cells were stored at 4 °C for 1 h before induction. The secondary culture was then induced with 1 mM (isopropyl-β-$_D$-thiogalactopyranoside) IPTG and grown further for 18 h at 16 °C with agitation at 100 rpm. Expressed cells were harvested by centrifugation and compared with uninduced cells on 10% SDS PAGE gel.

The GST-tagged *Hpy*AII.M1 protein was purified using GST affinity chromatography followed by size exclusion chromatography. The cells obtained after induction, were resuspended in the lysis buffer (50 mM Tris pH 7.5, 300 mM NaCl, 5 mM beta-mercaptoethanol, 10% (v/v) glycerol, 0.5 mg $ml^{-1}$ lysozyme), sonicated on ice with the pulse of 5 s on and 10 s off for 5 min. After centrifugation, the clarified lysate was incubated with GST binding beads pre-equilibrated with M1 buffer (50 mM Tris pH 7.5, 300 mM NaCl, 5 mM beta-mercaptoethanol, 10% (v/v) glycerol) and left in the end-to-end rotor for 4 h at 4 °C. Thereafter, beads were subjected to washing with M1 buffer. The protein was eluted with an M1 buffer containing 10 mM reduced glutathione pH 7.5. After analysis on 10% SDS-PAGE pure fractions were pooled and dialysed with buffer (50 mM Tris pH 7.5, 100 mM NaCl, 5 mM beta-mercaptoethanol, 10% (v/v) glycerol) at 4 °C. The purified protein was subjected to size exclusion chromatography (superdex 200 increase 10/300 gl (GE healthcare)) and major peak fractions were concentrated (10 kDa cutoff Amicon ultra 15 concentrator unit (EMD, Millipore). The purified protein, nearly homogenous as observed, was used for further characterization. The molecular mass of the purified *Hpy*AII.M1 was estimated by matrix-assisted laser desorption/ionization MALDI mass spectroscopy. The yield of the *Hpy*AII.M1 protein was estimated to be 0.8 $mgL^{-1}$ culture medium. The I196R mutant *Hpy*AII.M1 was also purified using the same protocol.

### 2.6. Circular dichroism spectroscopy

To estimate the overall protein conformation and secondary structure of both the wild type *Hpy*AII.M1 and mutant I196R, the homogeneously purified proteins were subjected to CD spectroscopy analysis. The proteins were first buffer exchanged with 20 mM PBS buffer pH 7.5 at 4 °C. The far UV CD spectra was recorded using a Jasco-J815 spectrophotometer equipped with a quartz cell of 1 mm pathlength under constant purging with nitrogen. The ellipticity was measured at room temperature from 190 nm to 250 nm with scanning speed of 50 nm/min. The proteins were exchanged with 20 mM PBS buffer pH 7.5 at 4 °C prior to analysis.

### 2.7. ATP binding assay

To assess the ATP binding ability of H*py*AII.M1, a modified electrophoretic mobility shift assay (EMSA) assay was performed with 2 μM of protein in the presence of γ-$P^{32}$ ATP. Briefly, the protein was incubated with 0.5 μM of γ-$P^{32}$ ATP in a reaction buffer of 20 mM Tris pH 7.5, 12 mM MgCl2. In addition, a parallel reaction was set with 1 mM ATP. All the reactions were UV-crosslinked for 10 min in UV-crosslinker ((Hoefer)) with energy 1200 × 100 $\mu J/cm^2$ and further incubated at 37 °C for 10 min. Reaction was mixed with 2 μl of SDS-PAGE gel loading dye and loaded on PAGE gel separately for cold ATP and γ-$P^{32}$ ATP reactions. The gel was transferred to a Whatman filter paper after electrophoresis and kept for drying on the gel drier under vacuum at 40 °C for 4 h. The dried gel was used for radio-phosphoimage analysis on typhoon FLA 9500 phosphorimager, whereas the gel with cold ATP sample was analyzed by Coomassie blue staining method.

### 2.8. ATPase assay

The ATP hydrolysis by *Hpy*AII.M1 was measured by thin layer chromatography on polyethylenimine cellulose plates. Briefly, a reaction mixture (10 μL) contained 50 mM Tris-HCl (pH 7.5), 2 mM $MgCl_2$, 2 mM DTT, 50 μg/mL BSA, 50 μM cold ATP, 0.008 μCi γ-$P^{32}$ATP with increasing amount of *Hpy*AII.M1 or I196R mutant (0.3 μg–5 μg). Heat denatured protein was used as a negative control and calf intestinal alkaline phosphatase was considered as a positive control to ensure the 100% hydrolysis of ATP as standard. The reaction mixtures were subjected to incubation at 37 °C for 30 min and the reaction was terminated by adding 10 mM EDTA, and 1 μL aliquots were spotted on the TLC plates. The plates were developed in a chromatography chamber pre-equilibrated with a mobile phase solution containing 0.5 M LiCl, 1 M HCOOH, and 1 mM EDTA. The plates were visualized using a Typhoon FLA-9500 phosphorimager and quantified with multigauge software. For the analysis of kinetic parameters, a constant amount of protein *Hpy*AII.M1 and I196R mutant (3 μg each) was incubated with an increasing concentration of cold ATP ranging from 3 to 80 μM, spiked with radiolabeled γ-$P^{32}$ATP. Kinetic parameters $V_{max}$ and $K_M$ were determined by plotting velocity (1/v) versus substrate concentration (1/[S]) in a double-reciprocal Lineweaver−Burk plot. The turnover number ($k_{cat}$) was calculated by the ratio of $V_{max}$ to the total enzyme concentration used. Data was averaged from three independent sets of experiments. The product formation was estimated by considering the intensity of the hydrolysed substrate as compared to the ATP hydrolysis in the negative control (Image Gauge software (fujifilm)).

Reaction velocity was calculated as follows:

$$V(nM/min/\mu g) = \frac{F^{*} \times Conc\ of\ Substrate\ (nM\ of\ Cold\ ATP)}{Time\ of\ incubation\ (min) \times Conc\ of\ protein\ (\mu g)}$$

$$F^{*}\ (Fractional\ cleavage) = \frac{Intensity\ of\ Pi\ (Product)}{Intensity\ of\ Pi + IntensityofATP}$$

## 3. Results

### 3.1. Diversity of SAM binding proteins

A total of 800 SAM/SAH complexed structures were obtained from the PDB, which represented 193 Gene Ontology (GO) functional classes, 87 Pfam families, 59 functional EC classes with majority of them being transferases and lyases. At the domain level too, of the 87 domains present in this set, the most common domains were the SET domain, the methyltransferase domain, and the TP methylase domain, all involved in the methylation of diverse substrates. Fold assignment using SCOP revealed a total of 19 folds spanning the entire repertoire of SAM complexed structures (Fig. 2a and Supplementary Figure S7). This clearly shows that proteins recognizing SAM ligands are highly diverse in their sequences and folds and hence evolved independently.

Our interest was in understanding how such diverse proteins can recognize the same ligand. To address this, we tested if the proteins had similar structures at the binding sites. We prepared a non-redundant dataset by removing redundancy in two steps, for which we performed sequence filtering of the 800 proteins by an all-vs-all global sequence alignment and clustering at 50% sequence identity. We identified 240 clusters and chose one representative (one that had the highest similarity to all other members in the set) per cluster. Next, we aligned the whole structures of the 240 proteins in a similar all-vs-all manner using TMalign and clustered at a TMscore threshold of >0.6. 96 clusters were identified, from which one representative subunit per cluster (with highest similarity to all other members) was picked. The resulting set of 96 proteins were completely independent of each other in terms of sequence and fold, and was taken as the final dataset (SAM-set) for all further analysis. The following major domains were represented in the SAM-set: Methyltransferase domain (13.5%), SET domain (6.2%) and Radical SAM superfamily (5.2%). The remaining domains that occurred less frequently include FtsJ-like methyltransferase (4.17%), C-5 cytosine-specific DNA methylase (3.12%) and many others represented by individual members in our non-redundant set. The SAM-set and the initial list of proteins annotated by their SCOP, PFam families are listed in Supplementary File (SAMDatas.xlsx).

### 3.2. Binding sites of SAM are grouped into 4 major types

Next, we exhaustively compared the structures of the known SAM binding sites from each of the 96 representatives in the SAM-set, using SiteMotif. SiteMotif, developed by us recently, identifies similarities in multiple binding sites and places them in a common framework akin to a multiple alignment. Using SiteMotif-based distances, we constructed a projection network where each site is a node and an edge is drawn between a pair of sites if their sites are similar ($M\text{-}dist_{min} > 0.5$). The resulting network from 96 sites had 1118 edges, of which 78 of them grouped into four definable site-clusters (Fig. 2b) with a high extent of interconnectedness within each cluster. This indicates that the 78 sites, representing 750 sites from the redundant set can be clearly grouped into 4 major types, with the giant cluster encompassing 65% (62 sites) further grouping into 2 subtypes as seen from the projection network (Fig. 2b). In addition, a few singletons are also apparent, indicative of new site types of their own kind.

### 3.3. Deriving SAM recognition motifs

Next, we set out to derive structural motifs within each type by analysing the spatial conservation of residues among the SAM-set. A conservation profile was derived from each cluster by aligning all members to its representative using SiteMotif. The motifs reflect the importance of residues towards SAM recognition. The SAM/SAH ligand can be described to be made of 3 components - the a (adenine), r (ribose) and m (methionine) parts. We observed a significant conservation of the GxG stretch interacting with the m component. For example, cluster_1a comprises 36 binding sites with PDB: 4AX8 as the representative. The average sequence identity of 4AX8 with 35 members was only 11.92% and the average RMSD of the Cα atoms of the members with the representative was 13.06 Å, indicating that these proteins belonged to different sequence and structural families and yet their binding sites

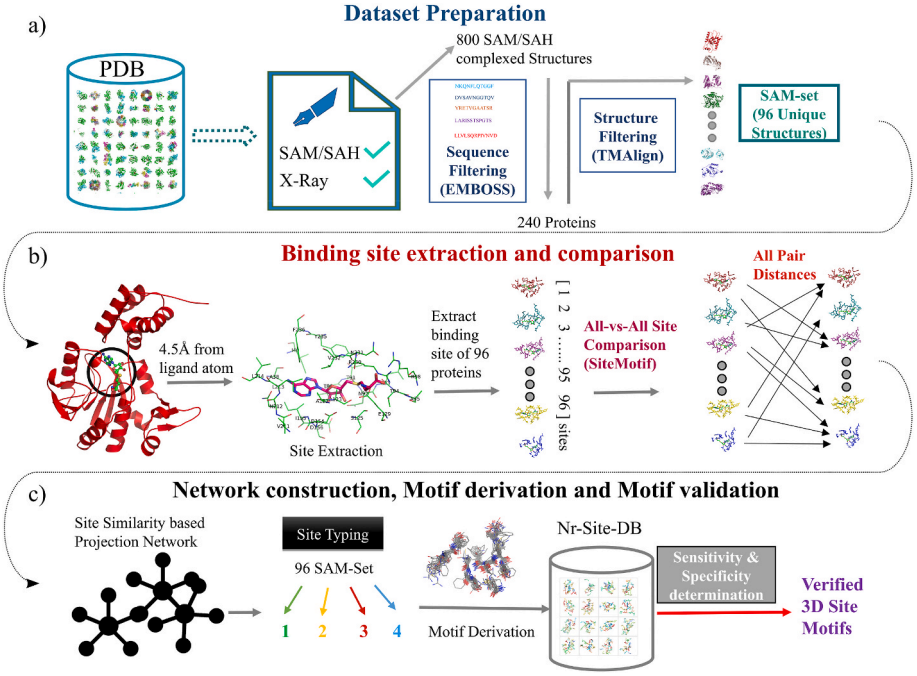



**Fig. 1.** An illustration of the workflow of the study including steps for (a) dataset preparation, (b) binding site extraction, all-vs-all comparison of binding sites, (c) construction of projection networks based on site similarities followed by site-typing and derivation of structural motif(s) for SAM recognition. The workflow includes a systematic validation protocol for testing sensitivity and specificity of the motifs.

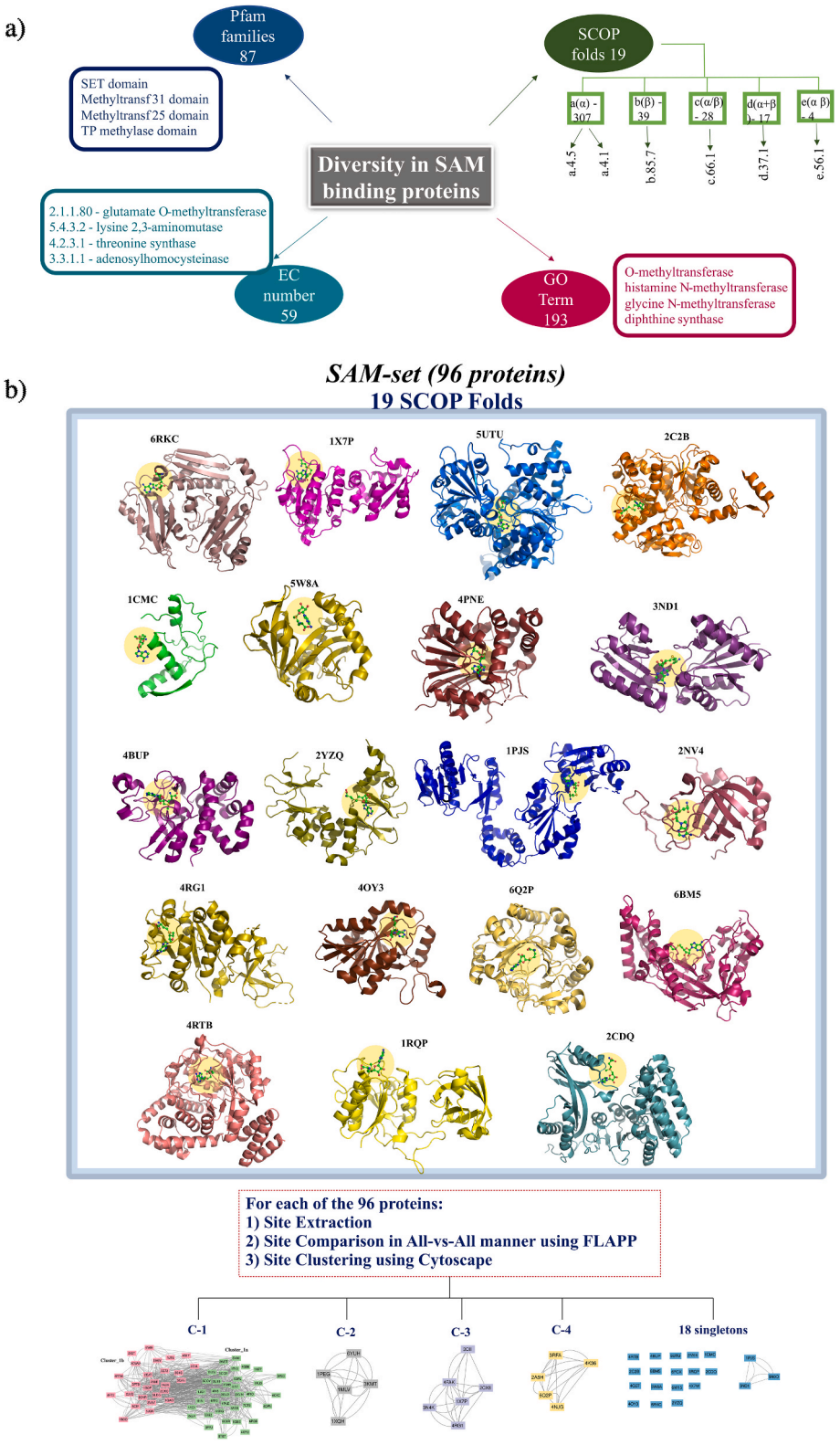



**Fig. 2.** An illustration of the SAM-set (A) describing the diversity in their sequences, structures and functions. The total number of Pfam families in the SAM-set and examples of 4 families are listed, which describes diversity in sequence space. Similarly the total number of SCOP folds adopted by proteins in the SAM_set as well as 4 most frequent folds are shown, to illustrate diversity in structural space. The range of functional roles played by the SAM-set is indicated in terms of the number of Gene-ontology (GO) terms and the number of enzyme (EC) classes with most frequent examples in each. (B) Construction of a projection network based on binding site similarities, where nodes represent sites and edges represent similarities between them over the chosen threshold. At M-$dist_{min} > 0.5$, 78 out of 96 proteins in the SAM-set were seen to share similarities among themselves, with 56 proteins forming one large cluster (cluster 1). In addition, there were three smaller-sized clusters (clusters 2, 3 and 4; $n \leq 6$). The remaining 18 of the 96 did not share any similarities and appear as singletons or have very few edges. The largest cluster was further separated into two clusters (1a and 1b), which represent 2 sub-types within cluster-1.

showed high similarities (Fig. 3). SiteMotif, through its ability to place all sites in a comparable framework and thereby effectively achieve a multiple site alignment, identified the motif GCG and QF[D/E] in cluster_1a. The GCG part was located near the methionine moiety of SAM and has been reported to play a major role in SAM catalysis. Phenylalanine in QF[D/E], where present, exhibited a prominent stacking interaction with the planar adenine ring and the side chain of negative charged residue (D or E), forming bidentate hydrogen bonding with

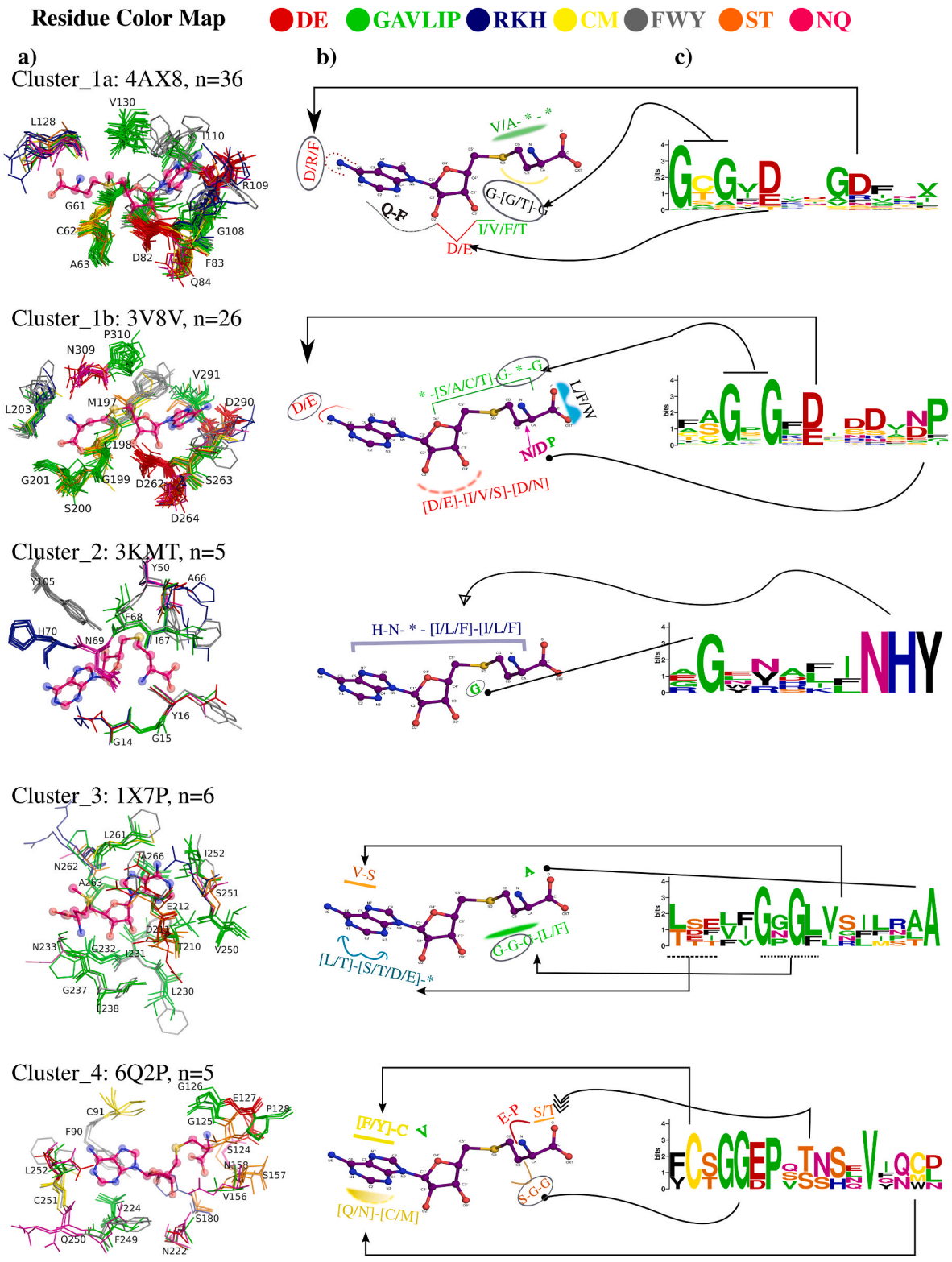



**Fig. 3.** Structural motifs of SAM binding. (a) Multiple alignments of SAM binding sites for each cluster are shown (left column), with (b) graphical representations of residues around SAM/SAH atoms (middle column) and (c) their corresponding structure-based sequence logos (right column). A representative protein for each cluster was identified as a node sharing a maximum number of edges with every other node, onto which, all other sites were superposed using SiteMotif and visualized using pymol. SAH was shown in ball and sticks, while conserved residues were represented in lines.

ribosyl oxygen (O3' and O4'). Besides these, we also found the abundance of aspartate and arginine near N6 of SAM. A GSG stretch was seen in cluster_1b as well, in a similar location with respect to the ligand. In addition, an N/D was seen to be interacting with amide nitrogen of SAM. Since, the sequences are very diverse, folds are diverse and the order of amino acids at the sites also not conserved, these can be regarded to have independently evolved, converging into a common function.

### 3.4. Presence of some common motif elements among clusters

The remaining 3 clusters which had fewer members in each (<6), also showed the presence of the GxG stretch, and in all cases close to the methionine moiety of SAM, which in essence is a variation of the GGG motif. For example, in cluster_1a, residue C or T was seen to replace the middle glycine generating a new variant (G-[C/T]-G). But in cluster 3, the GGG was preserved as such. Likewise, the glycine motif in cluster_4 reads as SGG. This is illustrated in the structure based sequence logos in Fig. 3. In addition, when we take all such motif residues that occurred in one or more clusters, we observe a strong preference for certain amino acids that make similar interactions with the ligand, regardless of the cluster. For instance, the residue D/E interacting with the O3' and O4' atom of 'r' component was highly conserved across many SAM binding proteins and further a D/E/N interacting with the N atom of the 'm' moiety of SAM was conserved between cluster1_b and cluster_4. In addition, a D/E in close proximity to the 'a' moiety was prominent in all clusters. These can be regarded as supermotifs (Fig. 3b - gray circled) since they exist even when the binding sites arise from very different folds, indicating their importance towards the recognition of SAM and SAH molecules.

It may be recalled that of the 96 sites, only 78 of them formed clusters, while the other 18 remained as singletons. We checked if the 18 singletons contained any similarities with any of the site motifs by structurally scanning each of the motifs against the 18 sites. We found that a majority of them indeed contained one or more of the super-motifs (GxG, D/E/N, hydrophobic stretch), albeit with some variation. As the overall similarity was lower than our chosen threshold and as these were all singletons, we have not considered them further in our analysis.

### 3.5. Sensitivity and specificity of the derived SAM motifs

Next, we tested how accurately the identified motifs capture the known SAM binding sites and how specific they are with respect to SAM binding. To address these, we carried out the following analysis: (a) recovery of the 78 sites (96 in SAMset - 18 singletons) by using only the motifs (from Fig. 3), (b) a PDB-wide scan of each of the motifs against a previously curated comprehensive set of non-redundant binding pockets (NRSiteDB), which also contains the SAM-set. With the first scan, all 78 sites were correctly identified as expected, confirming that the motifs indeed represented all of them, leading to 100% recovery. With the second scan, we aimed to test the recovery in the entire initial set of 800 SAM binding proteins. The NRSiteDB contains 60,502 binding sites extracted from PDB co-crystallized structures, after pruning at 70% sequence identity cutoff. The motifs-vs-all NRSiteDB sites scan (constituting 3,02,510 pairwise alignments) were carried out using FLAPP, a

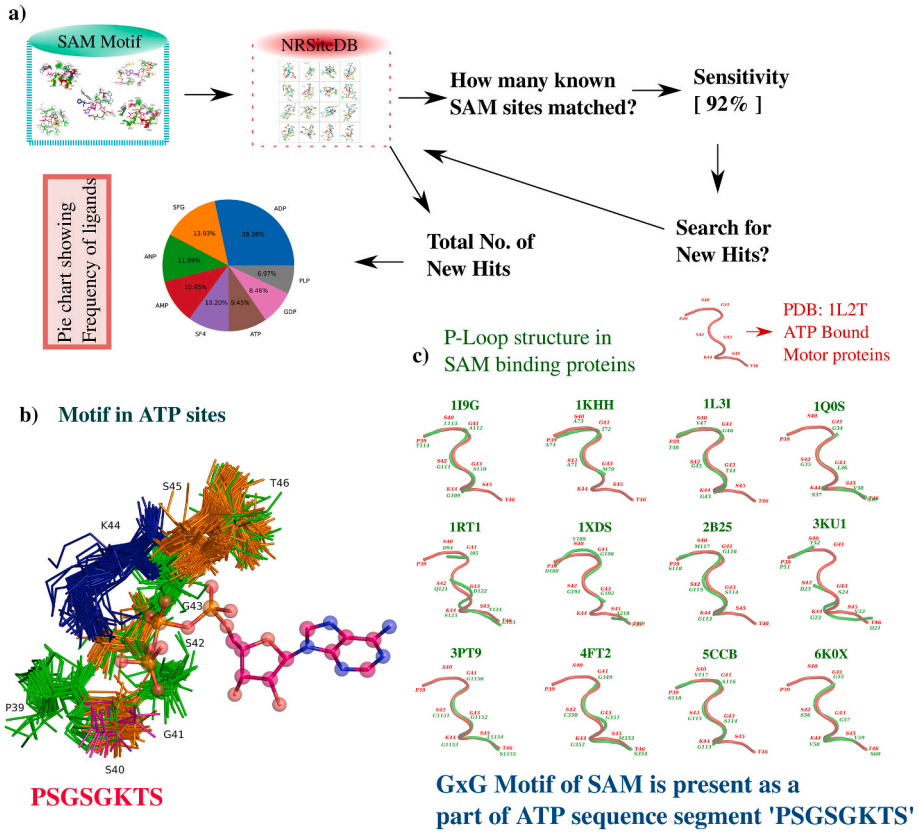



**Fig. 4.** Validation of structural motifs for SAM binding by comparison of SAM motifs against NRSiteDB (Non-Redundant binding Site DataBase) measuring the sensitivity of the derived motif in detecting known SAM complexes. (a) A workflow to scan SAM motifs against NRSiteDB (Non-Redundant binding Site DataBase). NRSiteDB features 60,502 binding sites extracted from PDB complexed structures. FLAPP was used to compute sensitivity of SAM motifs against NRSiteDB sites. Along with known hits, tests for specificity uncovered a number of ATP binding sites having SAM motifs in them. (b) Motif derivation for the ATP binding sites. c) Lastly, predominance of conserved ATP P-Loop motif in the SAM sites in different proteins. Overall, serine and glycine of the Walker motif were predominantly conserved in the SAM motifs. The lysine however was not seen in any of the SAM sites but was substituted to either serine or glycine. The overall secondary structure seemed to remain in the SAM binding sites too indicating the SAM and ATP binding sites share significant similarity.

recent fast pocket comparison tool developed by us (Sankar et al., 2022). FLAPP (With $F_{min} > 0.5$ and minimum number of aligned residues >5) outputs 732 known SAM/SAH complexes from 800 (92% recovery) as the top hits, reiterating the specificity of the motif in capturing true positives (Fig. 4a). The remaining 8% of the SAM/SAH complexes were observed to be part of 15 singletons and one smaller cluster (3 nodes), which explains why they were not picked in the first place. Together, they add up to 100% recovery, indicating high sensitivity of the motifs.

*A PDB wide motif scan reveals the presence of a SAM motif in a number of protein-ATP complexes:* Next, to evaluate motif specificity, we tested if any other ligand sites also possessed these motifs. In addition to picking known SAM hits, PDB-wide scans with SAM sites (clusters 1a and 1b) revealed a number of proteins that were complexed with ATP and ADP as the next-ranked hits (at a slightly lower but still significant similarity threshold of Fmin >0.4). This seems to suggest that ATP/ADP could potentially bind to SAM binding proteins and perhaps share a common point in their evolutionary trajectories. It must be noted that ATP and SAM share 2 of the 3 chemical moieties between them, ATP containing 'a', 'r' and 'p' (phosphate) moieties while SAM contains the 'a', 'r' and 'm' moieties. It was interesting to see that nearly 22% (138/630) of the 630 ATP complexes present in PDB were identified as hits with the motif of either cluster 1a or 1b. We constructed a projection graph based on site-similarity scores of the 630 ATP bound complexes from PDB, and identified the protein Bacterial ABC Transporter Cassette (PDB: 1L2T) to have the highest similarity to many SAM sites (Supplementary File: ATPDatas.xlsx). We create the ATP-set by first retrieving all PDB files through an advanced search in PDB using ATP as the query keyword, which yielded 2393 hits. These were clustered into sequence families at 70% sequence identity threshold using MMSeq2, which yielded 630 proteins, which we refer to as the ATP-set.

214 sites from the ATP-set had similarities ($F_{min} > 0.5$) with 1L2T from which we derived a site motif (Fig. 4b), which illustrates the presence of GxG motif in ATP binding sites as well, which forms a part of the well known Walker-A motif also known as the P-loop motif.

As the motifs responsible for the recognition of ATP and SAM were derived independently, we investigated if any known SAM binding proteins have ATP walker motifs in them. The structure corresponding to the segment 'PSGSGKST', taken from the ATP representative (1L2T), was aligned against 96 sites of the SAM-set and found several of the known SAM binding proteins to share good structural alignment. In particular, Fig. 4c shows that a large proportion of SAM binding sites adopt a loop structure similar to the P-loop in ATP-binding sites. Over the common scaffold of the P-loop, striking variations appear to be present at a few positions in the loop region in the SAM sites, providing a basis to explain their new functionality. For example, in SAM sites, the lysine from Walker motif has been seen to be replaced by glycine, serine, and rarely methionine or cysteine. This is to be expected, given that ATP with its negative charge on the phosphates, requires positively charged amino acids in its vicinity. The same does not necessarily apply for SAM or SAH. Meanwhile, the alternating serine and glycine in 'PSGSGKST' was found to be reasonably conserved in SAM binding proteins as well. We also found other residues of 'PSGSGKST' amino acids of the P-loop motif can be replaced (by alanine, valine, aspartate, methionine, cysteine and threonine) in SAM binding proteins, yet retaining the P-loop scaffold. The overall P-loop Walker secondary structure has been observed across a number of SAM binding sites. This leads to hypothesizing that ATP binding characteristics could theoretically be possible among SAM binding proteins such as methyltransferases.

### 3.6. Testing the theoretical feasibility of cross-binding of ATP and SAM in the datasets

To test the hypothesis that ATP and SAM binding sites share similarities, we tested the theoretical feasibility of ATP binding to proteins in the SAM-set and SAM binding to proteins in the ATP_set. For this, we first tested the feasibility of ATP binding to SAM sites computationally through ligand docking and comparing predicted binding energies with respect to the original ligand SAM. We computationally docked ATP and SAM independently to the ATP-set and to the SAM-set. AutoDock was used for this and a docking protocol was defined by specifying the centroid of known binding site residues as the grid center. The comprehensive exercise involved 3504 protein structures, 1825 known ATP binders and 1679 known SAM/SAH binders, which resulted in an exhaustive combination of 10,512 docking analyses. The binding energy of each protein-ligand complex was then assessed. We first verified that docking ATP onto ATP sites and SAM onto SAM sites largely produced the same ligand position and orientation in most cases. We also verified that docking ATP onto SAM sites and SAM onto ATP sites, showed that both ligands bound largely at the same site involving the same set of residues. The known ATP bound proteins were found to have a mean binding energy of -6.5 kcal/mol with ATP and -5.4 kcal/mol with SAM/SAH (Fig. 5), indicating that both ATP and SAM ligands were preferred nearly equally by kinases and hydrolases. On the other hand, known SAM binders showed a mean binding energy with ATP of -6.8 kcal/mol, whereas a mean binding energy of -7.5 kcal/mol for SAM. Overall, ATP-binding proteins showed higher affinity to ATP than SAM binding proteins while SAM-binding proteins showed higher affinity to SAM than ATP, although the overall difference was not high (Supplementary File: DockResult.xlsx). Put together, the analysis shows (a) overall similarities in the two groups, (b) each set has evolved to bind their cognate ligands with slightly better affinity.

### 3.7. Experimental testing of ATP binding to a SAM binding protein

To test further whether a SAM binding protein can indeed bind ATP, we carried out in vitro ligand binding experiments, for which we selected *Helicobacter pylori* methyltransferase HP1367 (*Hpy*AII.M1) whose clone was available from one of our laboratories (Kumar et al., 2018). *Hpy*AII.M1 is a type 2 restriction enzyme, which methylates DNA using SAM as a substrate. *Hpy*AII.M1 has a close homologue in cluster-1 (PDB: 1BOO) in which the highest number of matches with ATP binding sites was seen. We first computationally analyzed the binding feasibility and binding strength in a homology model of *Hpy*AII.M1 (Fig. S2). This entailed modeling the structure of *Hpy*AII.M1, locating the binding sites, and examining the SAM/ATP motifs. *Hpy*AII.M*1* shared 58% sequence identity with 92% query coverage with the identified template structure (PDB: 1G60, methyltransferase from MboIIA - complexed to SAH). *Hpy*AII.M1 shared all residues in common at the binding site with the template, making the site identification a trivial exercise. In one of our laboratories, we have previously demonstrated through biochemical experiments that this protein binds to SAM with a Km of 2.35 μM (Bangru et al., unpublished). The predicted binding affinity with SAM was -8.0 kcal/mol. A loop region 191–197 residues with sequence SGSGMTS at the SAM site, aligned with SGSGKST of the P-loop in P-loop containing ATP binding proteins. Docking ATP onto the site indicated a predicted interaction strength of -6.6 kcal/mol, which was comparable to that of SAM. With an aim of enhancing ATP binding, I196 residue located close to the γ-phosphate, as informed by the model, was selected for site directed mutagenesis to introduce an arginine residue (Fig. S3). The structure of the mutant protein was also modeled, which indicated that the arginine residue would be well positioned to interact with the phosphates and that it would not lead to any other significant perturbation in the structure. Both the wildtype *Hpy*AII.M1 and the mutant *Hpy*AII.M1-I196R were purified and investigated experimentally.

### 3.8. Purification and characterization of HpyAII.M1 and HpyAII.M1-I196R

The *Hpy*AII.M1 gene was cloned into the pGEX4T2 expression vector and used for overexpression of the protein *Hpy*AII.M1 in *E. coli* Rosetta (DE3) pLysS as represented in the methods section. Furthermore, the authenticity of the construct was confirmed by sequencing. The protein

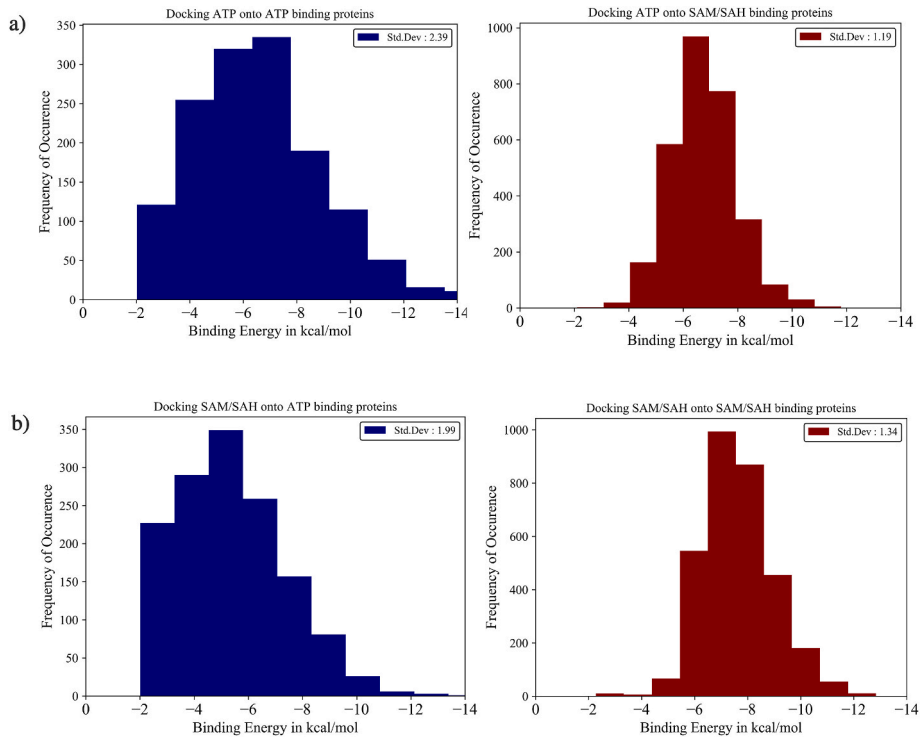



**Fig. 5.** Distribution of the binding affinities of ATP to proteins in the ATP-set and SAM in the SAM-set (right) and ATP to the SAM-set and SAM to the ATP-set (left) (a) distribution of the binding energy of ATP, SAM and SAH ligands with known ATP binding proteins such as kinases or hydrolases. The histogram (a) indicates that the interaction energy of SAM or SAH was comparable to the docking energy of ATP. Similarly the histograms at (b) indicate that the predicted interaction energies and hence the affinities are comparable between the two ligands for the same set of proteins in the SAM-set. However, a marginal preference was seen for SAM to SAM-set where SAM binding proteins bind to SAM slightly better than with ATP. Nevertheless the mean binding energy of ATP and SAM with known ATP binding was (-6.5 and -5.4 kcal/mol) and with known SAM binders was (-6.8 and -7.5 kcal/mol) respectively. Overall, the affinity for ATP and SAM are comparable for both sets.

*Hpy*AII.M1 of molecular weight 58 kDa was purified using two steps of purification including GST-affinity followed by size exclusion chromatography and checked on SDS-PAGE using Coomassie blue staining (Fig. S4). Further the purified *Hpy*AII.M1 was subjected to matrix assisted laser desorption ionization time of flight (MALDI-TOF) mass spectrometry and the protein was identified as site-specific DNA methyltransferase from *H. pylori* by peptide mass fingerprint analysis on the MASCOT server. The sequence coverage was 32% and the score was 82, which confirmed the identity (Fig. S4). In addition to the wild type protein, the I196R mutant was also overexpressed and purified using two steps of GST affinity and SEC chromatography as described for the wild type *Hpy*AII.M1. The fractions of major peaks were analyzed on 10% SDS-PAGE and the obtained mutant I196R was used for measuring ATP binding and ATPase activity.

### 3.9. CD spectroscopy analysis

The far UV CD spectra for *Hpy*AII.M1 wildtype and I196R mutant revealed that there were two prominent negative peaks at 212 nm and 224 nm in addition to a positive peak at 190–196 nm, indicative of α/β secondary structures in the proteins (Fig. S6). The CD results confirmed that the proteins exist in the well folded conformation in the solution. CD data deconvoluted to estimate the secondary structures showed 30% α-helices, 14% β-strands, which was in agreement with the structural model and structures of homologous proteins from PDB.

### 3.10. ATP binding assay reveals ATP binding property of HpyAII.M1 and HpyAII.M1-I196R

The ATP binding ability of *Hpy*AII.M1 was assessed by a modified EMSA assay (Hellman and Fried, 2007). 2 μg protein with a buffer containing 0.008 μCi γ-$P^{32}$ATP was incubated for 10 min in the presence of UV light and a portion of the resultant reaction was loaded on the SDS-PAGE and transferred on the Whatman filter paper that was subjected to drying under vacuum. The 0.008 μCi of radiolabeled γ-$P^{32}$ATP was used as a tracer to monitor the binding of the ATP and *Hpy*AII.M1 showed a sharp band when analyzed and compared with the protein without labeled ATP (Fig. S5). These results clearly demonstrated the ability of *Hpy*AII.M1 to bind to the ATP.

### 3.11. ATPase assay confirms ATP hydrolysis property of HpyAII.M1 and HpyAII.M1-I196R

We next tested if *Hpy*AII.M1 and the mutant were capable of hydrolysing ATP. The model suggested that an aspartic acid (D29) positioned near γ-phosphate could serve as the catalytic base for nucleophilic attack (Fig. S3). The *Hpy*AII.M1 and the I196R mutant were investigated for ATPase activity using γ-$P^{32}$ATP for tracing Pi released after hydrolysis of ATP (Fig. 6). The ATPase assay was performed using a range of protein concentrations (0.3–5 μg) with the wild type as well as the mutant. The product formed, that is the amount of radiolabeled Pi released, was monitored on PEI cellulose plate by thin layer chromatography. The *Hpy*AII.M1 was observed to hydrolyse the ATP in a concentration dependent manner (Fig. 6a). Increased concentration of protein resulted in increased Pi formation as compared to control reactions with no enzyme or with a heat denatured enzyme. The mobility of the product was comparable to the product Pi released with alkaline phosphatase (CIAP). The rate of product formation, the substrate saturation curve for the ATPase activity and Lineweaver-Burk plot clearly show ATPase activity by the enzyme (Fig. 7). Both proteins were found to hydrolyse ATP. However, the wild type *Hpy*AII.M1 showed higher activity as compared to the I196R mutant. The Km for the ATP for wild-

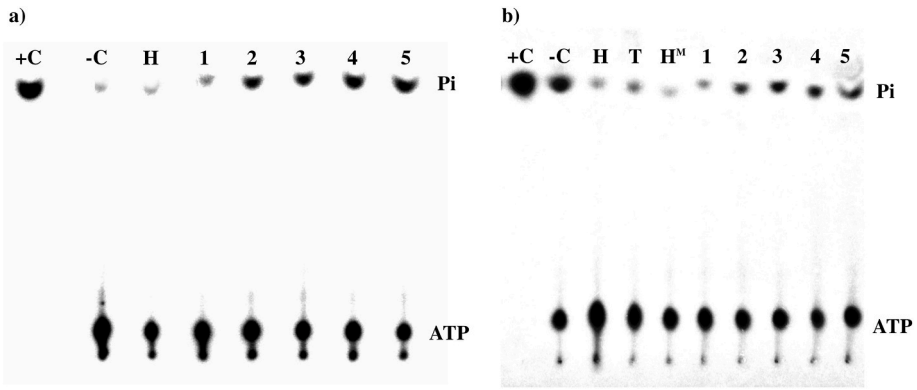



**Fig. 6.** ATPase assay: Hydrolysis of ATP by the enzyme was monitored by the product (γ-$P^{32}$ Pi) released (using γ-$P^{32}$ATP as a radiolabeled tracer) (a) *Hpy*AII.M1: lanes + C: positive control hydrolysis by CIAP and –C: negative control without enzyme, H: heat inactivated *Hpy*AII.M1; lanes 1–5 indicate the hydrolysis at 0.3 μg, 0.6 μg, 1.2 μg, 2.5 μg and 5 μg *Hpy*AII.M1 respectively. (b) I196R mutant: +C: positive control CIAP, –C: negative control without enzyme, H: heated *Hpy*AII.M1, T: GST tag protein only and $H^M$: heat inactivated I196R, lanes 1 to 5 shows increased Pi formation with increasing concentrations of I196R mutant enzyme (0.3 μg, 0.6 μg, 1.2 μg, 2.5 μg and 5 μg) respectively.

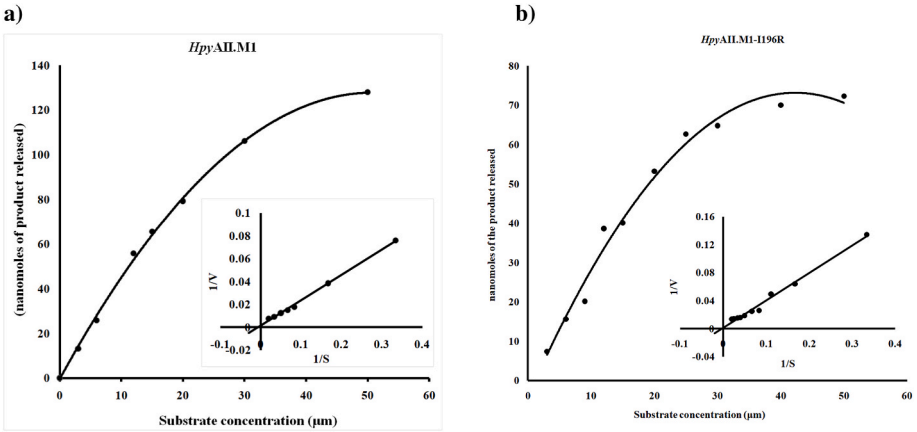



**Fig. 7.** Substrate saturation curve and Lineweaver-Burk plot with ATP, (a) *Hpy*AII.M1 showing product formation with various concentrations of ATP, (b) I196R mutant. A stock of enzyme (3 μg) in a reaction with cold ATP containing 0.008 μCi of labeled γ-$P^{32}$ ATP were used as substrate and tracer respectively. The reaction mix contains various substrate concentrations from 3 μM to 80 μM. *Hpy*AII.M1 enzyme was incubated with the reaction mixture at 37 °C for 30 min. The product Pi released was analyzed by TLC using PEI cellulose F plate and enzyme activity (v) was calculated using the formula.

type was estimated to be 39 μM and Vmax 233 μmol/s, whereas the Km and Vmax for the mutant were found to be 31 μM and 125 μmol/s respectively (Fig. 7). GST alone used as a tag did not show any activity (lane-4 of Fig. 6b). These results clearly show that *Hpy*AII.M1 in addition to binding SAM, is capable of binding ATP and also hydrolysing ATP and the I196R mutant shows a trend of enhanced ATP binding and reduced hydrolysis ability.

$$V(nanomoles\,/\,30min\,/\,3\mu g) = [F^{*} \times nanomoles\ of\ cold\ ATP]$$

$$where F^{*}(Fractional\ cleavage) = \frac{Intensity\ of\ product\ released}{Intensity\ of\ substrate + Intensity\ of\ product\ release}$$

The Lineweaver-Burk plot was generated by taking the inverse of both velocity and substrate $\frac{1}{v}$ vs $\frac{1}{s}$.

## 4. Discussion

Spatially proximal sets of amino acid residues that form binding sites for recognizing small molecule ligands, often referred to as motifs, impart functional capability to proteins. While it is known that motifs within a protein family are highly conserved, it is of great interest to identify if there are characteristic motifs for a given ligand in diverse proteins. Recognizing such motifs serve three main purposes - (a) functional annotation of a protein, given its structural model, (b) identification of cognate receptors for a given ligand, (c) understanding key residues in the binding sites which are critical determinants of ligand binding and of generating specificity. A subset of these motifs will be contiguous in sequence space, which are identified as characteristic sequence patterns for specific ligands. Such sequence motifs have been well studied, as they are easier to detect, requiring alignments based on sequence similarities of short stretches. Structural motifs without sequence adjacency are much harder to detect. In this work we identify structural motifs for SAM recognition. SAM, like ATP, is a ubiquitous

small molecule ligand, recognized by a wide array of proteins of different structural folds and sequence families, making it important to approach the problem of motif detection through structural data rather than through sequence data alone. Use of fast binding site three-dimensional comparison algorithms such as FLAPP has enabled clustering of all known sites into distinct site types, thereby enabling the derivation of multiple motifs that provide a much larger coverage of SAM binding proteins that was known from previous efforts to identify motifs. Although less convenient in terms of speed, alternative binding site detection methods such as PocketAlign, G-LoSA, SiteMotif could also be used in principle and can be expected to yield similar results (Sankar and Chandra, 2022; Yeturu and Chandra, 2011; Lee and Im, 2016).

We expect that the SAM motifs described in this work, will facilitate genome-wide identification of SAM-binding proteins, thus enabling a comprehensive definition of the range of proteins regulated or influenced in some way by SAM in a cell, which in turn can provide systematic insights into the range of methylation activity. Methylation is indeed an activity essential for cell viability, involving the transfer of a methyl group from a donor to one of the biological macromolecules such as proteins and DNA is critical for generating a dynamic update of switching individual components on or off in their constituent functions. S-Adenosylmethionine (SAM) is the sole substrate responsible for these methylation reactions. SAM's presence is ubiquitous across the three domains of life. Besides its role in methylation, there is ample evidence now to show that SAM is involved in many other important roles in the cell such as maintaining cellular homeostasis. There are a myriad of cellular proteins each possessing a distinct function requiring SAM either as a cofactor or a functional group donor or as a ligand. It is of interest to understand how SAM binds to a wide array of proteins and whether there are any general sequence or structural features that are characteristic of SAM binding. Identifying SAM motifs in poorly characterized proteins will form a basis to investigate their functional roles.

Knowledge of the function generating features in a protein also serve as valuable pointers to gain evolutionary insights. Our results show the presence of multiple site motifs in a diverse set of proteins clearly pointing to convergent evolution. Further the fact that the motifs are at known functional sites with known function in proteins with sequences that are significantly different, satisfy the general tests for convergence (Gherardini et al., 2007; Gorbalenya et al., 1989). There are many examples illustrating proteins with similar functions or structures having evolved independently in different lineages, showcasing the phenomenon of convergent evolution. Our findings show that this phenomenon is true for SAM as well, which is a ubiquitous cofactor. A limitation of our analysis is that there is an element of arbitrariness in cluster definition as the process of clustering is sensitive to the number of proteins that are being analyzed. As in any structural study of this kind, the analysis is restricted to available data which can have an inherent bias. As more structures become available, the exact membership may vary although we expect the overall pattern to remain. This challenge is seen with sequence-based methods too. and there are many instances showing that it is difficult to establish clear relationships based solely on sequence data. Nevertheless, structural analysis, particularly when comparing binding sites, offers a more robust approach to studying similarities in proteins as compared to analyzing sequences or entire protein folds. Focusing on binding sites provide a more nuanced understanding of how proteins with distinct evolutionary origins can evolve convergently to perform similar functions.

Further, observation of high similarity between SAM binding and ATP binding sites, which are the two most ubiquitous ligands and essential cofactors indicate that the sites for the two ligands have evolved multiple times independently as there is no detectable sequence similarity between the SAM-binding proteins and ATP-binding proteins. However, observing both ATP and SAM binding capacities in the same protein appears to suggest that possibly originally an ATP binding protein, small changes in sequence at the key residues at the binding site may have selected a SAM binding motif, to generate a new functionality. We differentiate our finding from proteins that can bind SAM and ATP binding sites at two different locations in the protein. While our results show that the same binding site can bind both SAM and ATP ligands, the protein WbdD presents an example with two independent domains with two separate binding sites for both ATP and SAM. Further experiments with mutant proteins that have abrogated SAM and enhanced ATP binding abilities, would help in probing functional significance if any in the cellular context. From an application point of view, knowledge of the motifs also provide a basis for protein engineering that will have applications in a wide range of causes.

## CRediT authorship contribution statement

**Santhosh Sankar:** Conceptualization, Investigation, Methodology, Software, Visualization, Writing – original draft. **Preeti Preeti:** Methodology, Validation. **Kavya Ravikumar:** Methodology, Validation. **Amrendra Kumar:** Methodology, Resources. **Yedu Prasad:** Resources. **Sukriti Pal:** Visualization, Writing – review & editing. **Desirazu N. Rao:** Investigation. **Handanahal S. Savithri:** Investigation. **Nagasuma Chandra:** Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing, Project administration.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

All required files are provided in supplementary

## Acknowledgements

We acknowledge support from the Bioinformatics grant, Department of Biotechnology, Government of India (BT/PR40186/BTIS/137/3/2021).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crstbi.2023.100108.

## References

Bauerle, M.R., Schwalm, E.L., Booker, S.J., 2015. Mechanistic diversity of radical S-adenosylmethionine (SAM)-Dependent methylation. J. Biol. Chem. 290 (7), 3995–4002. https://doi.org/10.1074/jbc.R114.607044.

Bügl, H., Fauman, E.B., Staker, B.L., Zheng, F., Kushner, S.R., Saper, M.A., Bardwell, J.C., Jakob, U., 2000. RNA methylation under heat shock control. Mol. Cell 6 (2), 349–360. https://doi.org/10.1016/s1097-2765(00)00035-6.

Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H., Velankar, S., 2017. Protein data bank (PDB): the single global macromolecular structure archive. Methods Mol. Biol. Clifton NJ 1607, 627–641. https://doi.org/10.1007/978-1-4939-7000-1_26.

Cheng, X., 1995. Structure and function of DNA methyltransferases. Annu. Rev. Biophys. Biomol. Struct. 24, 293–318. https://doi.org/10.1146/annurev.bb.24.060195.001453.

Fontecave, M., Atta, M., Mulliez, E., 2004. S-Adenosylmethionine: Nothing goes to waste. Trends Biochem. Sci. 29 (5), 243–249. https://doi.org/10.1016/j.tibs.2004.03.007.

Gana, R., Rao, S., Huang, H., Wu, C., Vasudevan, S., 2013. Structural and functional studies of S-Adenosyl-L-Methionine binding proteins: a ligand-centric approach. BMC Struct. Biol. 13, 6. https://doi.org/10.1186/1472-6807-13-6.

Gherardini, P.F., Wass, M.N., Helmer-Citterich, M., Sternberg, M.J.E., 2007. Convergent evolution of enzyme active sites is not a rare phenomenon. J. Mol. Biol. 372 (3), 817–845. https://doi.org/10.1016/j.jmb.2007.06.017.

Gorbalenya, A.E., Donchenko, A.P., Blinov, V.M., Koonin, E.V., 1989. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. FEBS Lett. 243 (2), 103–114. https://doi.org/10.1016/0014-5793(89)80109-7.

Greer, E.L., Shi, Y., 2012. Histone methylation: a dynamic mark in health, disease and inheritance. Nat. Rev. Genet. 13 (5), 343–357. https://doi.org/10.1038/nrg3173.

Hellman, L.M., Fried, M.G., 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat. Protoc. 2 (8), 1849–1861. https://doi.org/10.1038/nprot.2007.249.

Kagan, R.M., Clarke, S., 1994. Widespread occurrence of three sequence motifs in diverse S-Adenosylmethionine-Dependent methyltransferases suggests a common structure for these enzymes. Arch. Biochem. Biophys. 310 (2), 417–427. https://doi.org/10.1006/abbi.1994.1187.

Korolev, S., Ikeguchi, Y., Skarina, T., Beasley, S., Arrowsmith, C., Edwards, A., Joachimiak, A., Pegg, A.E., Savchenko, A., 2002. The crystal structure of spermidine synthase with a multisubstrate adduct inhibitor. Nat. Struct. Biol. 9 (1), 27–31. https://doi.org/10.1038/nsb737.

Kumar, S., Karmakar, B.C., Nagarajan, D., Mukhopadhyay, A.K., Morgan, R.D., Rao, D.N., 2018. N4-Cytosine DNA methylation regulates transcription and pathogenesis in Helicobacter pylori. Nucleic Acids Res. 46 (7), 3429–3445. https://doi.org/10.1093/nar/gky126.

Laurino, P., Tóth-Petróczy, Á., Meana-Pañeda, R., Lin, W., Truhlar, D.G., Tawfik, D.S., 2016. An ancient fingerprint indicates the common ancestry of rossmann-fold enzymes utilizing different ribose-based cofactors. PLoS Biol. 14 (3), e1002396 https://doi.org/10.1371/journal.pbio.1002396.

Lee, H.S., Im, W., 2016. G-LoSA: an efficient computational tool for local structure-centric biological studies and drug design. Protein Sci. Publ. Protein Soc. 25 (4), 865–876. https://doi.org/10.1002/pro.2890.

Liscombe, D.K., Louie, G.V., Noel, J.P., 2012. Architectures, mechanisms and molecular evolution of natural product methyltransferases. Nat. Prod. Rep. 29 (10), 1238–1250. https://doi.org/10.1039/c2np20029e.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., Chothia, C., 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res. 28 (1), 257–259. https://doi.org/10.1093/nar/28.1.257.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: the protein families database in 2021. Nucleic Acids Res. 49 (D1), D412–D419. https://doi.org/10.1093/nar/gkaa913.

Narunsky, A., Kessel, A., Solan, R., Alva, V., Kolodny, R., Ben-Tal, N., 2020. On the evolution of protein-adenine binding. Proc. Natl. Acad. Sci. U.S.A. 117 (9), 4701–4709. https://doi.org/10.1073/pnas.1911349117.

Nepusz, T., Yu, H., Paccanaro, A., 2012. Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods 9 (5), 471–472. https://doi.org/10.1038/nmeth.1938.

Pegg, A.E., 1986. Recent advances in the biochemistry of polyamines in eukaryotes. Biochem. J. 234 (2), 249–262. https://doi.org/10.1042/bj2340249.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. TIG 16 (6), 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

Sankar, S., Chandra, N., 2022. SiteMotif: a graph-based algorithm for deriving structural motifs in protein ligand binding sites. PLoS Comput. Biol. 18 (2), e1009901 https://doi.org/10.1371/journal.pcbi.1009901.

Sankar, S., Chandran Sakthivel, N., Chandra, N., 2022. Fast local alignment of protein pockets (FLAPP): a system-compiled program for large-scale binding site alignment. J. Chem. Inf. Model. 62 (19), 4810–4819. https://doi.org/10.1021/acs.jcim.2c00967.

Schubert, H.L., Blumenthal, R.M., Cheng, X., 2003. Many paths to methyltransfer: a chronicle of convergence. Trends Biochem. Sci. 28 (6), 329–335. https://doi.org/10.1016/S0968-0004(03)00090-2.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

Slany, R.K., Bösl, M., Kersten, H., 1994. Transfer and isomerization of the ribose moiety of AdoMet during the biosynthesis of queuosine tRNAs, a new unique reaction catalyzed by the QueA protein from Escherichia coli. Biochimie 76 (5), 389–393. https://doi.org/10.1016/0300-9084(94)90113-9.

Stoner, G.L., Eisenberg, M.A., 1975. Biosynthesis of 7, 8-diaminopelargonic acid from 7-Keto-8-aminopelargonic acid and S-Adenosyl-L-Methionine. The kinetics of the reaction. J. Biol. Chem. 250 (11), 4037–4043.

Taylor, F.R., Cronan, J.E., 1979. Cyclopropane fatty acid synthase of Escherichia coli. Stabilization, purification, and interaction with phospholipid vesicles. Biochemistry 18 (15), 3292–3300. https://doi.org/10.1021/bi00582a015.

Trausch, J.J., Xu, Z., Edwards, A.L., Reyes, F.E., Ross, P.E., Knight, R., Batey, R.T., 2014. Structural basis for diversity in the SAM clan of riboswitches. Proc. Natl. Acad. Sci. U.S.A. 111 (18), 6624–6629. https://doi.org/10.1073/pnas.1312918111.

Wang, S.C., Frey, P.A., 2007. S-adenosylmethionine as an oxidant: the radical SAM superfamily. Trends Biochem. Sci. 32 (3), 101–110. https://doi.org/10.1016/j.tibs.2007.01.002.

Yeturu, K., Chandra, N., 2011. PocketAlign a novel algorithm for aligning binding sites in protein structures. J. Chem. Inf. Model. 51 (7), 1725–1736. https://doi.org/10.1021/ci200132z.

Zhang, Y., Skolnick, J., 2005. TM-Align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33 (7), 2302–2309. https://doi.org/10.1093/nar/gki524.