Backbone Brackets and Arginine Tweezers delineate Class I and Class II aminoacyl tRNA synthetases

Florian Kaiser[1,2]◕*, Sebastian Bittrich[1,2]◕, Sebastian Salentin[2], Christoph Leberecht[1,2], V. Joachim Haupt[2], Sarah Krautwurst[1], Michael Schroeder[2], Dirk Labudde[1]

1 University of Applied Sciences Mittweida, Technikumplatz 17, 09648, Mittweida, Germany
2 Biotechnology Center (BIOTEC), TU Dresden, Tatzberg 47-49, 01307, Dresden, Germany

◕These authors contributed equally to this work.
* florian.kaiser@hs-mittweida.de

Abstract

The origin of the machinery that realizes protein biosynthesis in all organisms is still unclear. One key component of this machinery are aminoacyl tRNA synthetases (aaRS), which ligate tRNAs to amino acids while consuming ATP. Sequence analyses revealed that these enzymes can be divided into two complementary classes. Both classes differ significantly on a sequence and structural level, feature different reaction mechanisms, and occur in diverse oligomerization states. The one unifying aspect of both classes is their function of binding ATP. We identified Backbone Brackets and Arginine Tweezers as most compact ATP binding motifs characteristic for each Class. Geometric analysis shows a structural rearrangement of the Backbone Brackets upon ATP binding, indicating a general mechanism of all Class I structures. Regarding the origin of aaRS, the Rodin-Ohno hypothesis states that the peculiar nature of the two aaRS classes is the result of their primordial forms, called Protozymes, being encoded on opposite strands of the same gene. Backbone Brackets and Arginine Tweezers were traced back to the proposed Protozymes and their more efficient successors, the Urzymes. Both structural motifs can be observed as pairs of residues in contemporary structures and it seems that the time of their addition, indicated by their placement in the ancient aaRS, coincides with the evolutionary trace of Proto- and Urzymes.

Author summary

Aminoacyl tRNA synthetases (aaRS) are primordial enzymes essential for interpretation and transfer of genetic information. Understanding the origin of the peculiarities observed with aaRS can explain what constituted the earliest life forms and how the genetic code was established. The increasing amount of experimentally determined three-dimensional structures of aaRS opens up new avenues for high-throughput analyses of molecular mechanisms. In this study, we present an exhaustive structural analysis of ATP binding motifs. We unveil an adversative implementation of enzyme substrate binding in each aaRS Class. While Class I binds via interactions mediated by backbone hydrogen bonds, Class II uses a pair of arginine residues to establish salt bridges to its ATP ligand. We show how nature realized the binding of the same ligand species with completely different mechanisms. In addition, we demonstrate that

sequence or even structure analysis for conserved residues may miss important functional aspects which can only be revealed by ligand interaction studies. Additionally, the placement of those key residues in the structure supports a popular hypothesis, which states that prototypic aaRS were once coded on complementary strands of the same gene.

Introduction

The synthesis of proteins is fundamental to all organisms. It requires a complex molecular machinery of more than 100 entities to ensure efficiency and fidelity [1–3]. The ribosome pairs an mRNA codon with its corresponding anticodon of a tRNA molecule that delivers the cognate amino acid. Aminoacyl tRNA synthetases (aaRS) ligate amino acids to their corresponding tRNA [4], which is why they are key players in the transfer of genetic information. The mere existence of proteins and nucleic acids is a chicken or the egg dilemma. The sequential succession of amino acids in each protein is encoded by nucleic acid blueprints. In turn, these proteins are indispensable to replicate and translate nucleic acids. It is debated how this reflexive system came to be [5] and which polymer type constituted the earliest living systems. The RNA world hypothesis assumes nucleic acids were the sole basis of primordial life. RNA molecules can store and interpret genetic information, while also allowing for catalytic activity. In succession, proteins emerged to implement more elaborate, specific, and efficient catalytic activity [6]. However, the limited catalytic repertoire of RNA molecules [7] raises concerns that such a primordial world was based on a single polymer type. The peptide-RNA world hypothesis assumes that life and genetic information originated from a system in which RNA and peptides coexisted and complemented each other from the very beginning [7–10]. It is argued that only this interleaving of the two types of macromolecules can account for the speed with which the genetic code developed [9, 11, 12]. Both hypotheses were reviewed recently [11, 13]. Either way, aaRS are the entities which most prominently reflect that early episode of life.

The unique interface between gene and gene products is shaped by aaRS as they attach the amino acid to the corresponding tRNA molecule [4, 11]. Three main theories have been proposed to explain the emergence of the self-encoding translational machinery, namely: coevolution [14], ambiguity reduction [15, 16], and stereochemical forces [17]. The interaction between amino acid and nucleic acid lies at the basis of each theory and is linked to the emergence of aaRS [7, 18]. There is strong evidence for two archaic proto-enzymes as the origin of all aaRS, which were among the earliest proteins that enabled the development of life [8, 19–21]. Since then, these predecessors have evolved divergently into Class I and Class II (Figure 1), where each is responsible for a distinct set of amino acids [22–24]. The physicochemical properties of amino acids are distributed evenly between both classes, even though amino acids handled by Class I were shown to be slightly bigger [25]. This suggest a concurrent emergence of both classes and that archaic aaRS substrates have differed sufficiently to require two specialized kinds of aaRS [11]. Both classes are, on several levels, as distinct as possible from each other [11].

Every aaRS recognizes an amino acid and prevents misacylation of tRNAs by maximizing ligand specificity. The discrimination mechanisms between similar amino acids are well-studied [4, 26–28]. During the enzymatic reaction the designated amino acid is activated, forming an aminoacyl adenylate, before it is linked to the cognate tRNA [29, 30]. For example, the fusion of aspartic acid and its corresponding $\mathrm{tRNA}^{\mathrm{Asp}}$ by the aspartyl-tRNA synthetase (AspRS) follows the two-step reaction:

$$\mathrm{Asp} + \mathrm{ATP} \xrightarrow{\mathrm{AspRS}} \mathrm{Asp{-}AMP} + \mathrm{PP_i}$$

$$\text{Asp}-\text{AMP} + \text{tRNA}^{\text{Asp}} \xrightarrow{\text{AspRS}} \text{Asp}-\text{tRNA}^{\text{Asp}} + \text{AMP}.$$

Today most organisms feature 20 concrete realizations, each handling one specific amino acid [31, 32] – throughout the paper they are referred to as aaRS Types.

The modular architecture of aaRS has evolved well-orchestrated and was optimized for its specific requirements [23, 33]. Frequent domain inserts [9, 11] can render the evolutionary origin hard to track [34]. In principle, all aaRS have to conserve three functions: the correct recognition of the tRNA identity and amino acid as well as the ligation of both. Commonly, the anticodon binding domain ensures tRNA integrity by recognizing particular features of the anticodon [35, 36]. The identification and transfer of amino acids is then mediated by the catalytic domain, which differs in topology between the two classes (Fig 1). To minimize errors in protein biosynthesis, pre- and post-transfer editing mechanisms are conducted by approximately half of the aaRS Types [26, 37, 38].

Fig 1. The two aaRS classes and amino acids they ligate to the cognate tRNA. Based on the physicochemical properties of the amino acids (colored according to [39]) no distinction can be made between the two classes. However, statistically significant differences based on amino acid side chain size [25] and binding site size [40, 41] are evident. Lysine is mostly processed by Class II aaRS, but in some archaic organisms a Class I aaRS is responsible for lysine [42]. Prior to tRNA ligation, the amino acid ligand is converted to its activated form: aminoacyl adenylate.

Sequences of aaRS proteins are highly diverse and result from fusion, duplication, recombination, and horizontal gene transfer [43, 44]. However, two sets of Class-specific and mutually exclusive sequence motifs were identified, which are responsible for interactions with adenosine phosphate as well as catalysis [4, 22, 45]. Class I features the conserved HIGH and KMSKS motifs [4, 22]. The functional key motifs in Class II are referred to as Motif "1", Motif "2", and Motif "3" [4]. Both HIGH and KMSKS stabilize the transition state, whereby the latter constitutes a mobile loop in the folded structure [4]. The binding of ATP and the transition state of the reaction of individual Class I proteins was demonstrated to be stabilized by a structural rearrangement [8, 46–53], which stores energy in a constrained conformation of the KMSKS motif [54]. The Class II motifs are less conserved [34] and more variable in their relative arrangement [22]. Motif "1" mediates the dimerization of protein structures, commonly found in Class II aaRS [4, 55]. Motif "2" and "3" are essential for the reaction mechanism and feature two highly conserved arginine residues [22, 56, 57].

The catalytic domain of Class I adapts a Rossmann fold [29], whereas Class II possesses a unique fold [44, 58, 59]. To assert the global structural similarity, two major structural alignments were calculated for Class I and Class II, respectively, that revealed high structural similarity within each Class with average sequence identity below 10% [60]. On a functional level, both aaRS classes exhibit distinct ATP binding site architectures and reaction mechanisms. Class I aaRS proteins attach the amino acid to the 2'OH-group of the tRNA's 3'-terminal adenosine, whereas Class II proteins use the 3'OH-group as the attachment location [61].

Rodin-Ohno hypothesis

In 1995, Rodin and Ohno proposed an elegant explanation for the peculiarities that are observed in contemporary aaRS: both classes were originally encoded on complementary strands of the same nucleotide fragment [8] (Fig 2). The Rodin-Ohno hypothesis is supported by an experimental deconstruction of aaRS sequences [9, 11]. In these studies, parts of contemporary aaRS proteins were removed and the catalytic strength of the

resulting transcripts was assessed. One representative sequence of each Class was reduced to a peptide of only 46 amino acids. The coding nucleotide sequences of these 46-residue peptide were paired complementarily. These so called "Protozymes" were investigated regarding their structural and catalytic properties; they form molten globules [9, 11] and – despite the lack of ordered tertiary structure – they are still capable of rate enhancements by orders of magnitude [9, 11]. It is essential that the efficiency of different enzyme families across the proteome increases at comparable rates [9, 11]. The phenomenon of anti-parallel coupling of two genes was also postulated for other families of proteins [62, 63] and seems to be a phenomenon that affects the whole genome [64, 65]. One contradicting theory is the coevolutionary theory of the genetic code [14]. This theory suggests two main groups of amino acids based on the connectedness of their biochemical pathways and that amino acid biosynthesis was the dominant factor that shaped the genetic code [18]. Other authors suggested that both classes evolved from unrelated ancestors and are of independent origin [22].

Fig 2. The Rodin-Ohno hypothesis states that both aaRS classes descended from the opposite strands of a single gene. The signature motifs of each class were fully complementary on this gene. Both Protozymes originated from the complementary "HIGH-Motif 2" region (shaded in red). Contemporary aaRS feature insertion domains (ID) and Connecting Peptides (CP1) as well as the addition of the anticodon binding domain (ABD). Figure adapted from [9, 66].

The Rodin-Ohno hypothesis can explain why ATP and tRNA binding sites of both classes seem to be mirror images of each other [67] as well as the fact that both classes share virtually no similarities [4, 11, 44] beside their actual function [8, 9, 11]. All of the contemporary aaRS Types are connected by the requirement to bind ATP. This basal unifying characteristic was found to involve hydrogen bonds in the Class I Protozymes [9].

Remarkably, the restrictions inherent with a complementary coding may explain why the middle base of a codon is the most distinctive base for the corresponding amino acid nowadays [21]. Other studies showed how slight differences in the substrate can result in a stable separation of aaRS into two classes [7, 10]. Potentially, the two Protozymes diverged into ten aaRS Types each (Fig 1) and simultaneously increased fidelity and incorporated additional domains when necessary [8, 9, 19–21]. Most of aaRS evolution took place before the "Darwinian threshold" [60]. Only a small number of amino acids, such as tryptophan, were gradually incorporated into the genetic code after the last universal common ancestor and inefficient proteins evolved over time [18]. While similar amino acids were once processed by the same aaRS, specificity may have required additional aaRS Types to cope with increasing complexity. It is still possible to observe such generic aaRS in some organisms [68, 69].

Motivation

A systematic delineation of aaRS active site residues is expedient [11]. The most conserved part of the aaRS reaction mechanism is the amino acid activation with ATP, since it represents the principal kinetic barrier for the creation of peptides in a pre-biotic context [66]. This fundamental mechanism is shared by all Class I and Class II aaRS enzymes, irrespective of their Type or the organism of origin. Furthermore, the catalytic domain has been predicted to constitute the ancestral aaRS precursors [9, 11, 63, 70]. The residues of the catalytic domain involved in amino acid binding were molded to meet specific requirements of the individual properties of each amino acid during evolution. In contrary, the ATP binding part includes the most conserved parts of the structure. To achieve a systematic delineation of available protein structures, this study

focuses on the most common element: the binding of the ATP substrate. Individual aaRS and their mechanism to discriminate similar amino acids have been extensively studied on the structural level [4, 26–28]. However, a comprehensive and comparative study of structural features in aaRS proteins is missing. There are no structural motifs known that capture the profound differences of the ligand recognition mechanism.

To unveil general adenosine phosphate binding properties of each aaRS Class, we have investigated the corresponding binding pockets of 972 aaRS protein molecules for each aaRS Type across all kingdoms of life. In total, 448 protein chains for Class I and 524 chains for Class II, available from the Protein Data Bank (PDB) [71], were analyzed. Previous studies have focused on comparing subsets of structures for each Class but to our knowledge no conclusive study was conducted that includes structures for every aaRS Type for both classes.

The results of this study outline the dichotomy between the two classes (Fig 3) on a functional level. A conserved pair of arginine residues is grasping the adenosine phosphate part of the ligand in nearly all Class II structures. Class I features no comparable structural pattern for adenosine phosphate binding, but interaction analysis divulged two highly conserved backbone hydrogen bonds, which seem to realize the same function without the need for conserved amino acid side chains. Due to their geometrical characteristics, we refer to the Class I and Class II motifs as Backbone Brackets and Arginine Tweezers, respectively. The Backbone Brackets motif demonstrates the limitations of sequence analysis and was, to our knowledge, never identified as a highly conserved interaction pattern prior to this study. Additionally, a novel geometrical characterization of these structural motifs demonstrates that significant structural rearrangements can be observed for all Class I structures upon ligand binding. The highly sensitive geometric characterization of side chain angles and alpha carbon distances is able to detect subtle differences in ligand binding and is potentially suitable to be applied on other conserved structural patterns as well.

Both structural motifs can be traced back to the Protozyme and Urzyme regions postulated in the studies based on the Rodin-Ohno hypothesis [8, 9, 11]. The analysis of codons in the corresponding regions accentuates existing insights and allows for an additional look behind the curtain of evolution.

Fig 3. Based on the analysis of 972 protein 3D structures (448 protein chains for Class I and 524 chains for Class II), Backbone Brackets and Arginine Tweezers were identified as structural motifs distinctive for their respective aaRS Class.

Results

This study presents a dataset of aaRS structures annotated with ligand information, which serves as a stepping stone to understand common and characteristic ligand interaction properties. It is composed of 972 individual chains containing 448 (524) Class I (Class II) catalytic aaRS domains and covers at least one ligand-bound structure for each aaRS type. The dataset is provided in S1 File and S2 File. The Class I chains originate from 256 biological assemblies and comprise 151 bacterial, 84 eukaryotic, 20 archaea, and one viral structure. The Class II chain set corresponds to 267 biological assemblies where 102 are of bacterial origin, 104 from eukaryotes, and 61 from archaea (for a detailed overview see S9 Fig). The sequence identity is below 33% (29%) for 95% of all Class I (Class II) structures, while pairwise structure similarity is high with a TM score [72] over 0.8 for 95% of the structures (S8 Fig). The high sequential diversity probably stems from the variety of covered organisms and domain insertions. In

contrast, the low structural diversity can be seen as a result of conserved function and the shared topology of the catalytic domain within each aaRS Class.

Sequence positions of all structures in the dataset were unified using a multiple sequence alignment (MSA) generated with the T-Coffee expresso pipeline [73] (Section Mapping of binding sites, S5 File and S6 File). This type of MSA is backed by the additional structural alignment of protein structures. Hence, the structurally conserved catalytic core region is preferred during alignment, since insertion domains and structurally diverse attachments do not align structurally across the whole dataset. The MSA allows the investigation of a plethora of structures independently of the concrete aaRS Type. This investigation is aided by a renumeration that effectively provides a mean to compare sequentially divergent, structurally similar proteins. All further referenced positions are given in accordance to this MSA. In figures where depictions of structures are shown, the original sequence positions of residues are listed. To infer original sequence positions from given renumbered sequence positions, tables S13 File and S14 File are provided. These tables contain the corresponding original sequence positions for each position of the MSA and for each structure in the dataset.

Backbone Brackets and Arginine Tweezers

In order to investigate the contacts between aaRS residues and their ligands, noncovalent protein-ligand interactions were annotated. This revealed two highly consistent interaction patterns between catalytic site residues and the adenosine phosphate part of the ligand: conserved backbone hydrogen bonds in Class I as well as two arginines with conserved salt bridges and side chain orientations in Class II.

Strikingly, the residues mediating the backbone interactions were mapped in 441 of 448 (98%) Class I renumbered structures at the two positions 274 and 1361. Closer investigation on the structural level revealed geometrically highly-conserved hydrogen bonds between the amino acid backbones and the adenosine phosphate part of the ligand (Fig 4A). These two residues mimic a bracket-like geometry (Fig 4B), enclosing the adenosine phosphate, and were thus termed Backbone Brackets. The interacting amino acids are not limited to specific residues as their side chains do not form any ligand contacts. Hence, position 274 of the Class I motif is not apparent on sequence level while position 1361 exhibits preference for hydrophobic amino acids, e.g. leucine, valine, or isoleucine (Fig 4C). Examples for the Backbone Brackets motif are residues 153 (corresponding to renumbered residue 274) and 405 (corresponding to renumbered residue 1361) in Class I ArgRS structure PDB:1f7u chain A.

In contrast, Class II aaRS structures show a conserved interaction pattern of two arginine residues at renumbered positions 698 and 1786, which were identified in 482 of 524 (92%) structures. The two arginine residues grasp the adenosine phosphate part of the ligand (Fig 4D) with their side chains, resembling a pair of tweezers (Fig 4E), and were thus named Arginine Tweezers. These two arginines are invariant in sequence (Fig 4F). Examples for the Arginine Tweezers motif are residues 217 (corresponding to renumbered residue 698) and 537 (corresponding to renumbered residue 1786) in Class II AspRS structure PDB:1c0a chain A. Additionally, a highly conserved glutamic acid is the most prevalent at renumbered position 700. This residue establishes hydrogen bonds to the adenine group of the ligand in SerRS, HisRS, ThrRS, LysRS, ProRS, and AspRS.

The Backbone Brackets and their counterpart, the Arginine Tweezers, are both responsible for the interaction with the adenosine phosphate part of the ligand (all ligand interactions are shown by example in S2 Fig). Mappings of the motif residues to original sequence numbers can be found in S7 File and S9 File. For some structures it was not possible to pinpoint the conserved motifs after unifying sequence positions (listed in S8 File and S10 File).

Fig 4. (**A**) Structural representation of the Backbone Brackets motif interacting with Tryptophanyl-5'AMP ligand in TrpRS (PDB:1r6u chain A). The ligand interaction is mediated by backbone hydrogen bonds (solid blue lines). Residue numbers are given in accordance to the structure of origin. (**B**) The geometry of the Backbone Brackets motif resembles brackets encircling the ligand. (**C**) WebLogo [74] representation of the sequence of Backbone Brackets residues (274 and 1361) and three surrounding sequence positions. Residue numbers are given in accordance to the MSA. (**D**) Structural representation of the Arginine Tweezers motif in interaction with Lysyl-5'AMP ligand in LysRS (PDB:1e1t chain A). Salt bridges (yellow dashed lines) as well as π-cation interactions are established. Residue numbers are given in accordance to the structure of origin. (**E**) The Arginine Tweezers geometry mimics a pair of tweezers grasping the ligand. (**F**) Sequence of Arginine Tweezers residues (698 and 1786) and surrounding sequence positions. The Backbone Brackets show nearly no conservation on sequence level since backbone interactions can be established by all amino acids, while the Arginine Tweezers rely on salt bridge interactions, always mediated by two arginines. Residue numbers are given in accordance to the MSA.

Further analysis of secondary structure elements for both motifs shows that residues of the Backbone Brackets are predominantly tied to unordered secondary structure elements (S3 Fig). However, the positions 275, 276, 277, 1359, and 1360 feature a consistently unordered secondary structure. A predominantly unordered state can also be observed for the N-terminal Arginine Tweezers residue 698, while the following three positions almost exclusively occur in strand regions (S4 Fig). Residue 1786 is always observed in α-helical regions, mostly at the third position of the α-helix element.

The high conservation of backbone or side chain geometry of these motifs suggests that their residues are indispensable for enzyme functionality. To substantiate this assumption, Backbone Brackets and Arginine Tweezers were characterized in greater detail and analyzed regarding their ligand interactions and geometric properties.

Interaction patterns

Contacts between ligands and proteins are established via a variety of noncovalent interaction types such as hydrogen bonds, π-stacking, or salt bridges. These interaction types were annotated using the Protein-Ligand Interaction Profiler (PLIP) [75] to investigate whether evolution adapted entirely different strategies or if some characteristics are shared between both aaRS classes.

Two sets of 29 and 40 representative complexes for Class I and Class II were composed to analyze adenosine phosphate-binding. For the comparison of commonly interacting residues between different aaRS Types, a matrix visualization was designed (Fig 5). This allows for the assessment of interaction preferences at residue level. Data for frequent interactions was available for 12 residues and 10 different aaRS Types for Class I as well as 13 residues and 11 aaRS Types for Class II. All sequence numbers shown in Fig 5 originate from the MSA renumbering and corresponding sequence numbers of all structures in the dataset can be derived from the tables provided in the S13 File and the S14 File.

While six different interaction types are used to bind the adenosine phosphate ligand, hydrogen bonds are the prevalent type of contact, especially for the recognition of the ribose moiety (see Fig 5). The aromatic ring system of adenine is recognized via hydrogen bonds and π-stacking interactions in both Class I and Class II complexes. Class II aaRS bind this part of the ligand also forming π-cation interactions with the charge provided by one guanidinium group of the Arginine Tweezers (residue 1786). Residue 698 interacts predominantly with the negatively charged phosphate group of

Fig 5. Protein-ligand contacts in representative adenosine phosphate-binding complexes for aaRS Class I and II. Residues are grouped according to the non-amino acid ligand fragment (phosphate, ribose, or adenine) that they are interacting with. Preferred interaction types for each aaRS Type and binding site residue are color-coded. Fields split into two triangles indicate two equally preferred interactions. The asterisk (*) indicates aaRS Types incorporating noncanonical amino acids. Automatically retrieved [76, 77] mutation effects [78–84] are shown as centered shapes. In essence, Class I interactions are mainly hydrogen bonds, while Class II adenosine phosphate-binding is realized by an array of different interaction types. All sequence numbers are given according to the MSA.

the ligand via salt bridges. This binding pattern is conserved in Class II and handled by the other guanidinium group featured by the Arginine Tweezers. In Class I, hydrogen bonding is essential for the binding of phosphate. Here, residue 274 binds to the phosphate and is part of the Backbone Brackets motif which embraces the phosphate and the aromatic ring at the other end (residue 1361) using backbone hydrogen bonds.

Both motifs share the tendency to form electrostatic interactions with the α-phosphate of the ligand. In general, the phosphate group predominantly participates in salt bridges and hydrogen bonds. The ribose moiety is almost exclusively stabilized by hydrogen bonds to its hydroxyl groups.

Geometric characterization

Backbone Brackets and Arginine Tweezers were analyzed at the geometrical level to further substantiate the profound differences in adenosine phosphate recognition. The side chains of the Backbone Brackets' residues are expected to exhibit higher degrees of freedom in comparison to the Arginine Tweezers. Furthermore, a significant change in alpha carbon distance of both motif residues indicates a conformational change during ligand binding. The state complexed with adenosine phosphate (M1) and the state in which no adenosine phosphate is bound (M2) were analyzed separately in order to quantify these aspects (see S1 Fig for a visual representation of M1 and M2). Structure alignments of both motifs in respect to their binding modes (provided in S7 Fig) visually support the differences in side chain orientation and variable amino acid composition of the Backbone Brackets.

The angle between side chains of the Backbone Brackets is continuously high a mean of 144.90 ± 20.93° for M1 and 141.40 ± 20.13° for M2, respectively. This emphasizes that the side chain orientation is indistinguishable between M1 and M2 as only the backbone participates in ligand binding. The alpha carbon distance is conserved for the majority of the Backbone Brackets observations, with a mean of 17.92 ± 0.86 Å for M1 and 18.41 ± 0.82 Å for M2, respectively. However, some observations (structures PDB:5v0i chain A, PDB:1jzq chain A, PDB:3tzl chain A, PDB:3ts1 chain A) exhibit higher alpha carbon distances of 20.54 Å, 19.74 Å, 19.10 Å, and 18.79 Å, respectively. In contrast, one occurrence of the Backbone Brackets motif in structure PDB:4aq7 chain A has a remarkably low alpha carbon distance of 16.50 Å. Nevertheless, alpha carbon distances between bound and unbound state differ significantly ($p<0.01$, S5 Fig). This indicates the substantial contribution of backbone interactions as well as the conformational change observed during adenosine phosphate-binding.

The side chain variation is marginal for the Arginine Tweezers if an adenosine phosphate ligand is bound. In contrast, the side chain angle of the apo form is highly variable with a mean of 91.82 ± 8.69° for M1 and 79.81 ± 21.67° for M2, respectively. The side chain angles between the bound and unbound state differ significantly ($p<0.01$, S6 Fig), reinforcing the pivotal role of highly specific side chain interactions during

ligand binding. This effect cannot be observed for the alpha carbon distances of the Arginine Tweezers, with a mean of 14.76 ± 0.66 Å for M1 and 14.93 ± 0.79 Å for M2, respectively.

Fig 6. Geometric analysis of the ligand recognition motifs responsible for the adenosine phosphate interaction for aaRS Class I and Class II representative and nonredundant structures. The alpha carbon distance is plotted against the side chain angle θ. Binding modes refer to states containing an adenosine phosphate ligand (M1) or not (M2). Backbone Brackets in M1 allow for minor variance with respect to their alpha carbon distance, constrained by the position of the bound ligand. In contrast, Arginine Tweezers in M1 adapt an orthogonal orientation in order to fixate the ligand.

Relations to known sequence motifs

Fig 7 encompasses structure and sequence motifs as well as the sequence conservation scores of the underlying MSA. Amino acids interacting with the adenosine phosphate of the ligand (ordinate in Fig 5) are annotated.

For Class I sequence motifs [24, 45, 85], the HIGH motif features sequence conservation and is located nine positions downstream of the N-terminal Backbone Brackets residue. The KMSKS motif exhibits no sequence conservation and can be observed downstream of the C-terminal Backbone Brackets residue. The five-residue motif contains ligand binding site residues 1414, 1415, 1434, and 1441 and is distributed within a corridor of around 70 aligned sequence positions.

For the Class II sequence motifs [24, 56, 57, 59, 85], Motif "1" is moderately conserved in sequence. However, it does not interact with the ligand according to our analysis. Motif "2" is conserved around the N-terminal Arginine Tweezers' residue and contains five additional ligand binding site residues of lower sequence conservation. Motif "3" exhibits high sequence conservation and comprises the C-terminal Arginine Tweezers' residue.

Further ligand binding site residues, which are not part of known sequence motifs, are mostly occurring in the sequence conserved regions which predominantly bind the ribose moiety.

Fig 7C relates the identified Backbone Brackets and Arginine Tweezers to the proposed Protozyme and Urzyme regions of both aaRS classes [8, 9, 11]. One Backbone Bracket residue is present in the Class I Protozyme, located upstream of the HIGH motif. The other Backbone Bracket residue is located close to the KMSKS motif and therefore part of the Urzyme. Regarding Class II, the N-terminal arginine residue is located in Motif "2" and close to the antisense coding position of the C-terminal Backbone Bracket residue in the Protozyme. The C-terminal Arginine Tweezers residue is located in Motif "3", that is neither part of the Urzyme nor the Protozyme region.

Urzyme regions and codon assignment

Rodin and Ohno proposed regions that are associated with each other across the Class division of aaRS [8]. The "HIGH-Motif 2" region was mapped to residues with numbers between 255 to 336 in the renumbered structures of Class I and to 648 to 718 in Class II (according the 46-mers generated by Martinez-Rodriguez et al. [9]). Further, the "KMSKS-Motif 1" region was mapped to residue numbers 1352 to 1452 for Class I and 347 to 371 for Class II in the renumbered structures (according to the alignments by Rodin and Ohno [8]).

Original codons have been mapped for key regions and consensus codons were generated for each of the residues of this region (see Table 1 and Table 2). The codons

Fig 7. Integrative sequence view for aaRS Class I (**A**) and II (**B**). Boxes delineate sequence motifs previously described in literature [45, 56, 57]. The trace depicts the sequence conservation score of each position in the MSA (S5 File and S6 File). These scores were computed with Jalview [39, 86], positions composed of sets of amino acids with similar characteristics result in high values. Furthermore, all positions relevant for ligand binding (Fig 5) are depicted. Backbone Brackets and Arginine Tweezers have been emphasized by their respective pictograms. Positions of low conservation or those not encompassed by sequence motifs were intangible to studies primarily based on sequence data. Especially backbone interactions might be conserved independently from sequence. (**C**) Sequence representation of the Rodin-Ohno hypothesis [8, 9, 11] with equivalents of the Backbone Brackets or Arginine Tweezers residues shown as green dots. The N-terminal residue of each, the Backbone Brackets and the Arginine Tweezers motif, is present in the Protozyme region (shaded red). Additionally, the C-terminal Backbone Brackets' residue is located in the Urzyme region.

are rather diverse, but for key positions the middle base exhibits conservation. In the "HIGH-Motif 2" region positions 274-698, 281-692, and 284-689 show complementary middle base pairing. For the "KMSKS-Motif 1" region only one conserved complementary middle base pairing is present at position 1414-365.

Table 1. "HIGH-Motif 2" codon assignment and base pairing. First and last row are consensus residues according to the structure-based MSA, "+" indicates gaps. Signature regions according to [8] are emphasized. Sequence numbers are given according to the MSA. Middle rows indicate consensus codons; matches are indicated by vertical lines and mismatches by "x". Arginine Tweezers' and Backbone Brackets' residues are framed by boxes.

	G	P	$\boxed{\text{Y}}$	P	S	G	+	S	L	**H**	**I**	**G**	**H**	A
C I	272	273	274	275	276	277	278	279	280	281	282	283	284	285
	xxx	xxx	xAx	xCx	Axx	Gxx	xxx	xxx	xTx	CAx	xTx	GGx	xAx	xxx
			x	|	|					|			|	
	xxx	GAx	xAx	xGx	TTx	xxx	xxx	xxx	xxx	xTx	xxx	xxx	xTx	xxx
C II	701	700	699	698	697	696	695	694	693	692	691	690	689	688
	N	E	N	$\boxed{\text{R}}$	F	C	P	G	G	**F**	**E**	**A**	**M**	R

Table 2. "KMSKS-Motif 1" codon assignment. First and last row are consensus residues according to the structure-based MSA, "+" indicates gaps. Signature regions according to [8] are emphasized. Sequence numbers are given according to the MSA. Middle rows indicate consensus codons; matches are indicated by vertical lines and mismatches by "x". The C-terminal Backbone Brackets' residue is framed by a box. Sequence positions were omitted if both complementary sequences feature low occupancy, and are therefore not necessarily consecutive.

	H	G	L	$\boxed{\text{L}}$	**K**	**M**	**S**	**K**	+	L	G	N
C I	1352	1356	1360	1361	1414	1415	1416	1417	1418	1444	1450	1452
	xxx	xxx	xxx	xTx	AAx	xTx	xxx	Axx	xxx	xxx	xxx	xxx
					|						
	xxx	CCx	xxx	xxx	xTx	xxx	xxx	xxx	GGx	xxx	xxx	xxx
C II	371	370	368	366	365	364	363	362	361	352	348	347
	M	P	T	E	**V**	**E**	**M**	**F**	G	R	D	L

Effect of mutagenesis experiments and natural variants

To estimate the importance of certain ligand interactions, one can exploit data derived from mutagenesis experiments and natural variants.

Fig 5 shows the effect of nine mutations on the enzymatic activity of aaRS. There is no obvious link between conserved interactions and outcomes of mutations. For example, there are loss of function mutations occurring in regions with observed interactions and equally many cases where no interactions were observed while the mutation still has a negative effect. All sequence positions are given according to the MSA.

For Class I TyrRS, mutations of any histidine of the HIGH motif [45] lead to a decrease in activity, since both residues contribute to the stabilization of the transition state of the reaction [78, 79]. The same holds true for Asp-1300 and Gln-1301 which interact with the ribose part of the ligand [82, 84]. Phe-1415 is part of a sequence motif responsible for ligand attachment to the tRNA; a mutation to leucine shows no effect on the binding affinity [84].

Cys-1458 in Class II AlaRS is part of a four residue zinc-binding motif [87] and an exchange with serine results in no effect whatsoever. It is assumed that the other three amino acids can compensate the mutation [81]. The single-nucleotide polymorphism (SNP) with no known effect is associated to position 1703 in AspRS (rs1803165 in dbSNP [88]).

Ile-703 in Class II GlyRS does not directly interact with the ligand – mutations, however, result in a negative effect and are most prominently linked to Charcot-Marie-Tooth disease as the amino acid is crucial for tRNA ligation [80]. Another SNP occurs at Gly-1783; the exchange with arginine prohibits ligand binding and was tied to a loss of activity as well as distal hereditary motor neuropathy type VA [83].

Discussion

The reflexive system of building blocks and building machinery implemented in aaRS is an intriguing aspect of the early development of living systems. There is evidence that proteins arose from an ancient set of peptides [89] and that these peptides were co-factors of the early genetic information processing by RNA.

Sequence-based analyses were among the first tools to investigate the transfer of genetic information. DNA and protein sequences comprise the developmental history of organisms, their specialization, and diversification [44]. However, following the "functionalist" principle in biology, sequence is less conserved than structure, which is, in turn, less conserved than function [90]. Therefore, structural features and molecular contacts have been recognized as key aspects in grasping protein function [91, 92] and evolution. Only if the necessary function can be maintained by compatible interaction architectures, the global role of the protein in the complex cellular system is ensured [93]. This is also eminent in aaRS precursor structures that were described to be molten globules but as long as the function of the protein is ensured, it is able to survive during evolution [9]. If evolution tries to conserve structure over function, the evolutionary progress might have been considerably slower and thresholds for the development of new functions would have been higher [90].

Each amino acid of a protein fulfills a certain role and can often be replaced by amino acids with compatible attributes [91]. By considering each amino acid in the context of its sequence, its structural surroundings, and finally its biological function, one can determine possible exchanges and the evolutionary pressure driving these changes [90, 94]. Up to this point, pure sequence or structure analysis methods – ignoring ligand interaction data – missed the functional relevance of the Backbone

Brackets entirely.

Backbone Brackets and Arginine Tweezers

The analysis of Backbone Brackets geometry showed a high variance of side chain angles for both binding modes. The distinction between these modes is significantly manifested in a change of the alpha carbon distance, which supports that the conformational change during ligand binding previously observed in ArgRS [95], TyrRS [46–49, 96], and TrpRS [50, 52–54] is a general mechanism in Class I aaRS. Furthermore, the C-terminal residue of the Backbone Brackets is located close to the KMSKS motif (Table 2). Thus, the structural rearrangement in the KMSKS motif upon ATP binding might indirectly affect the geometric orientation of the C-terminal residue of the Backbone Brackets – especially regarding the position of its alpha carbon.

In contrast to the Backbone Brackets, the Arginine Tweezers are highly restrained in side chain orientation if a ligand is bound, which shows that this orientation is key to adenosine phosphate recognition. If no ligand is bound, the Arginine Tweezers' geometry is less limited, which is reflected in a higher variability of side chain orientations. Conclusively, the distinction between the two binding modes can be made by taking the geometry of the motifs into account: alpha carbon distances for Backbone Brackets and side chain angles for Arginine Tweezers.

The conserved Arginine Tweezers motif resembles a common interaction pattern for phosphate recognition [91], which usually features positively charged amino acids [97]. However, the conformational space of ATP ligands was shown to be large throughout diverse superfamilies [98] and hence the geometry of binding sites involved in ATP recognition is manifold. The uniqueness of aaRS compared to other ATP-binding proteins was shown in AspRS, where the ligand binds in a compact form with a bent phosphate tail instead of the usually found extended form [98]. This conformation of ATP is energetically unfavorable but allows easy access of the α-phosphate for tRNA binding [99]. This suggests that the Arginine Tweezers motif possesses a unique geometry and is not a generalizable pattern for ATP binding, such as the frequently occurring P-loop domain [97].

As the function of fixating the adenosine phosphate part is crucial in aaRS enzymes, mutations of the Arginine Tweezers residues result in loss of function [100, 101]. However, to our knowledge, the Backbone Brackets motif was not identified in earlier literature and is herein described for the first time. The stunning balance of evolutionary diversification [102] and equality in function is underlined by profoundly different implementation of ligand recognition in terms of adjacent sequence (Fig 4C and 4F), embedding secondary structure elements (S3 Fig and S4 Fig), geometrical properties (Fig 6), and interaction characteristics (Fig 5).

The catalytic core of both aaRS classes is also hypothesized to consist of amino acids handled by the complementary aaRS Class [11, 103, 104]. The conserved residues of the Arginine Tweezers in Class II support that statement because ArgRS is a Class I aaRS. The contemporary implementations of the Backbone Brackets, however, are dominantly realized by amino acids handled by Class I. Further studies are necessary to test this hypothesis by a detailed investigation of the identified binding site residues for all proteins of the dataset.

Backbone Brackets are not conserved in sequence

The Backbone Brackets are remarkable, since backbone interactions are often neglected in structural studies. Nevertheless, backbone hydrogen bonds make up at least one-quarter of overall ligand hydrogen bonding [105]. In these cases, side chain properties may only play a minor role, e.g. for steric effects, and allow for larger

flexibility in implementation of a binding pattern as long as the correct backbone orientation is ensured. There are examples of protein-ligand complexes where backbone hydrogen bonds are a major part of the binding mechanism, e.g. in binding of the cofactor NAD to a CysG protein from *Salmonella enterica* (PDB:1pjs) as determined with PLIP [75]. In conclusion, the Backbone Brackets exhibit conservation on functional level rather than on sequence level, which renders sequence-based motif analysis infeasible. This motif is a prime example for conservation of function over structure or sequence [90]. When ligands can still be bound specifically by backbone interactions, these binding sites become significantly more resilient to mutations. The complementary codon pairing of both classes' Protozymes might not only have shaped the genetic code [104], but also required some positions in the Class I Protozyme to be highly variable to compensate changes in the complementary strand. Any amino acid can furnish the observed backbone hydrogen bonds to the ATP ligand, thus drastically increasing the evolvability of both Protozymes.

Complementary coding of Backbone Brackets and Arginine Tweezers

The isolated "HIGH-Motif 2" region has been shown to be catalytically active [9]. Interestingly, the Arginine Tweezer and the Backbone Bracket appear in very close proximity to each other, when considering the complementary coding according to the Rodin-Ohno hypothesis (see Table 1). This N-terminal Arginine Tweezers residue is oppositely arranged to a conserved proline residue in Class I at position 275. The mapped codons show a matching middle base pair at this position, which is conserved across all kingdoms of life. This further strengthens the evidence for the evolutionary constraints of these residues. Both amino acids fulfill a very important role for the function of aaRS in general. The role of the arginine is well established, it binds the γ-phosphate of the ATP molecule and enforces the crucial bent conformation of the phosphate tail [57]. The conserved proline acts as a wedge to open the amino acid binding site to provide access between adjacent strands of a β-sheet [31]. The proline residue does not interact directly with the ligand but is still conserved in the binding site, which is why a proposed structural role seems reasonable.

The region reconstructed by [9] is also considered to be the so called Protozyme – the minimal functional aaRS unit required in ancient protein biosynthesis. This region contains the N-terminal residue of both structural motifs identified in this study. This suggests that both N-terminal residues can fulfill their functional role in solitary, but with reduced efficiency. During evolution, the aminoacylation reaction was further improved by adding their other functionally equivalent counterpart.

This is substantiated by the occurrence of the second Backbone Brackets residue at position 1361 very close to the KMSKS mobile loop (residues 1414 to 1417). This C-terminal Backbone Brackets' residue is part of the region identified as Urzyme, which evolved from the Protozyme, and is more efficient in catalyzing the aminoacylation reaction. Despite the low conservation on sequence, both Backbone Brackets residues have conserved central codon base pairs. This is also the case for other residues that are highly conserved on amino acid sequence, such as the histidine residues in the HIGH motif. This underlines the functionalist principle that has recently been addressed in the context of the evolution of binding sites [90]. The attempt to find conservation on sequence or even structural level is in this case futile, since the interaction is mediated by backbone atoms and in principle this interaction can be realized by any amino acid. Yet, the middle base of the codon for both Backbone Brackets' residues is conserved. The C-terminal Backbone Brackets' residue shows a tendency for hydrophobic amino acids (see Fig 4C). This is reflected by the conserved thymine middle base that usually

codes for hydrophobic amino acids such as leucine, isoleucine, and valine. In contrast, the conserved adenine middle base of the N-terminal Backbone Brackets' residue codon encodes for many diverse amino acids, such as glutamic acid, lysine, or glutamine. This coincides with the low sequence conservation observed at this position.

The second Arginine Tweezers residue is situated in the Motif "3" region that has been described before as being important in the aminoacylation reaction [61]. Even though this region is not considered part of either the Ur- or Protozyme, it is present in most of the Class II structures. A comparison of the catalytic rate enhancement, relative to the uncatalyzed second-order rate for the Urzyme with added Motif "3", but without the preceding insertion domain (similar comparisons have been made previously [44, 106] and are concluded in [9, 11]) is reasonable. It seems that Class II compensated the lack of the second binding element of the ATP part by focusing on the dimerization associated to most of Class II synthetases [44, 107]. In contrast, Class I evolved the C-terminal Backbone Brackets' residue and did not develop mechanisms such as dimerization to match the reaction speed of Class II. During the course of evolution the ATP binding by two entities proved efficient and was adapted by Class II synthetases as well.

According to the Rodin-Ohno hypothesis [8], one can conclude the following chronological appearance of the Backbone Brackets and Arginine Tweezers motif. The N-terminal residues of both motifs seem to be the most ancient parts, both located in the Protozyme region. Over a prolonged period of time the C-terminal Backbone Brackets residue, which is located close to the KMSKS motif and hence part of the Urzyme, was introduced. The most recent residue seems to be the C-terminal Arginine Tweezers' residue, located in Motif "3", which is neither part of the Protozyme nor the Urzyme.

Disease implications

Due to the fundamental role of aaRS for protein biosynthesis, a systematic assessment of mutation effects in yeast was conducted by Cavarelli and coworkers [100]. Mutations of aaRS-coding genes can be drastic and may result in a variety of human diseases, even if the structural effect is unknown [108, 109].

Structural analysis of a GlyRS mutant (G526R) showed that the Charcot-Marie-Tooth disease may be caused by blockage of the ATP binding site. Furthermore, this mutation induces a larger contact area in the homo-dimer interface, which stems partially from the anticodon binding domain [83]. Other mutations result in a wider range of diseases and symptoms such as hearing loss, ovarian failure, or cardiomyopathy [110, 111]. Even for cellular processes unrelated to translation, aaRS play a pivotal role, e.g. for angiogenesis [112]. Due to the highly individual characteristics of aaRS enzymes between organisms, it is possible to create precisely targeted antibiotics with minimal side effects [113–115].

Unfortunately, automatically mapped mutational data does not cover the Backbone Brackets or Arginine Tweezers motif. It is expected that mutations of the Arginine Tweezers will cause a strong decrease in enzyme activity as shown in [100]. In contrast, the Backbone Brackets are expected to be more resilient to mutational events. However, bridging the gap between mutational studies and key interaction patterns will require further analysis beyond this study and needs to be substantiated by *in vitro* experiments. The provided high-quality aaRS dataset can serve as the basis for such work.

Limitations

The method used to unify residue numbering in all structures relies on the quality of the used MSA as well as the quality of local structure regions. Hence, the Backbone

Brackets and Arginine Tweezers were not successfully mapped for all structures of the dataset. On the one hand, some binding site regions were not experimentally determined (e.g. PDB:3hri) or the mapping of the motif residues failed (e.g. PDB:4yrc) due to ambivalent regions in the MSA. On the other hand, some aaRS may have evolved different strategies to bind the ligand, even for the same aaRS type [116].

However, the conserved ligand interactions were related to known sequence motifs (Fig 7). The sequentially high variance of the KMSKS motif was described before [45] and explains why the MSA algorithm distributes this motif over 70 positions. Another explanation is the differing conformation between the two binding modes [46–48, 50–54] which leads to a scattered structure-based sequence alignment in the KMSKS region [73]. The interacting residues 1352, 1360, and 1361 of Class I are located upstream of the KMSKS motif. In case of Class I, the AIDQ motif in TrpRS is known [108], yet no consensus for all aaRS Types was established. Class II sequence motifs exhibit high degeneracy and can hardly be identified without structural information [102]. Motif "1" is the only sequence motif which is not linked to any relevant ligand interaction site; its primary role lies in the stabilization of Class II dimers [56].

The geometric characterization of the two ligand recognition motifs (see Fig 6) highlighted some observations of the Backbone Brackets, which exhibit a substantial increase or decrease of the residue alpha carbon distance. For instance, chain A of an LeuRS of *Escherichia coli* (PDB:4aq7) is complexed with tRNA and the Backbone Brackets' alpha carbon distance is about 1 Å below the average. Manual investigation of this structure showed that there is no obvious conformational difference to other structures. Likewise, the annotated interactions were checked for consistency using PLIP and showed usual interactions with the adenine and the sulfamate group (the phosphate analogue) of the ligand. For the Backbone Brackets with higher alpha carbon extent (structures of IleRS, TrpRS, and TyrRS), interaction analysis revealed that residue 274 interacts with the amino acid side chain, as all of these structures contain a single aminoacyl ligand (PDB:3tzl chain A, PDB:3ts1 chain A, PDB:1jzq chain A) or two separate ligands (amino acid and AMP, PDB:5v0i chain A). This suggests that the structures resemble a partially changed conformation prior to tRNA ligation and a possible role of the Backbone Brackets motif in amino acid recognition. Likewise, these effects can arise from low quality electron density maps in the structure regions of interest. However, these hypotheses have to be addressed and validated in future work.

Interestingly, our analysis did not reveal a high count or conservation of interactions established with the well-known HIGH motif in Class I. Despite irregularly occurring salt bridges, hydrogen bonds, and one π-cation interaction in GluRS (see Fig 5A), no interactions were observed. This especially holds true for the first histidine residue of the HIGH motif, which only interacts with the ligand in GluRS. However, it was shown that the HIGH motif is mainly relevant for binding in the pre-acylation transition state of the reaction [45], i.e. HIGH interacts with the phosphate of ATP. This explains the irregular observations of interactions which are established only if an ATP ligand is present (e.g. GluRS PDB:1j09 chain A residue 15).

Prospects

Adaptations of the presented workflow to other protein families of interest might allow to study binding mechanisms in a new level of detail and by using publicly available data alone. Even if the geometric characterization is dependent on the quality of local structure regions, the comparison of alpha carbon distances and side chain angles is a simple yet valuable tool to separate different binding states. Geometrical properties can reveal the importance of conserved side chain orientations, the degree of freedom in unbound state, or shifts in backbone arrangement. However, choosing these two properties to compare residue binding motifs depends on the specific use case.

For the analysis of aaRS structures, the geometric characterization of the two conserved core interaction patterns was shown to be sufficiently sensitive to suggest the structural rearrangement of Class I aaRS to be a general mechanism. Hence, if structural motifs conserved in a larger number of protein structures are known, geometric analysis can reveal insights into global structural effects that occur during ligand binding without requiring any additional information.

In a similar way, the obtained interaction data proved as a valuable resource to understand fundamental aspects of aaRS ligand recognition. Despite the fact that interactions can not be determined for apo structures and do not take into consideration the dynamic nature of enzyme reactions, both, structure and interaction data conflates several aspects of evolution and proved to outperform pure sequence-based methods. Regarding the Rodin-Ohno hypothesis, structural investigation of the proposed Protozymes [8, 9, 11] and their ligand binding properties can further substantiate the importance of the Backbone Brackets and Arginine Tweezers as the primordial ATP binding site.

The designed approach was used to analyze aaRS from the different viewpoints: sequence backed by structure information, ligand interactions, and geometric characterization of essential ligand binding patterns. Additionally, this study provides the largest manually curated dataset of aaRS structures including ligand information available to date. This can serve as foundation for further research on the essential mechanisms controlling the molecular information machinery, e.g. investigate the effect and disease implications of mutations on crucial binding site residues. Further phylogenetic analyses can be conducted, based on the identified structural motifs. The sequence of aaRS proteins was shown to be highly variable [60] yet Backbone Brackets and Arginine Tweezers constituted a common pattern shared by almost all structures of the corresponding aaRS classes.

Alongside the aaRS-specific results, the workflow is a general tool for identification of significant ligand binding patterns and the geometrical characterization of such. Further studies may adapt the presented methodology to study common mechanisms in highly variable implementations of ligand binding, i.e. for nonribosomal peptide synthetases as another enzyme family that is required to recognize all 20 amino acids [117].

Materials and methods

Dataset preparation

Proteins with domains annotated to belong to aaRS families according to Pfam 31.0 [118] were selected (see S1 Appendix for a detailed list of Pfam identifiers) and their structures were retrieved from PDB. Additionally, structures with Enzyme Commission (EC) number 6.1.1.- were considered and included in the initial dataset. Structures with putative aaRS function were excluded.

For each catalytic chain the aaRS Class and Type, resolution, mutational status, the taxonomy identifier of the organism of origin, and its superkingdom were determined. For chains where a ligand was present, these ligands were added to the dataset and it was decided if this ligand is either relevant for amino acid recognition (i.e. contains an amino acid or a close derivate as substructure), for adenosine phosphate-binding (i.e. contains an adenosine phosphate substructure), or for both (e.g. aminoacyl-AMP).

As the presented study focuses on the binding of the adenosine phosphate moiety, two binding modes referred to as M1 and M2 (S1 Fig) were defined. M1 features an adenosine phosphate-containing ligand (e.g. aminoacyl-AMP, ATP), whereas M2 does not contain any ligand that binds to the adenosine phosphate recognition region of the binding pocket (e.g. plain amino acid, empty pocket).

To avoid the use of highly redundant structures for analysis, all structures in the dataset were clustered according to >95% sequence identity using Needleman-Wunsch [119] alignments and single-linkage clustering. For each of these clusters, a representative chain (selection scheme listed in S2 Appendix) was determined. The same procedure was used to define representative chains for the adenosine phosphate bound state M1 and no adenosine phosphate bound state M2. The final dataset is provided as formatted table in S1 File and as machine-readable JSON version in S2 File.

Mapping of binding sites

To allow a unified mapping of aaRS binding sites, an MSA of 81 (75) representative wild type sequences of Class I (Class II) (S3 File and S4 File) aaRS was performed. The alignment was calculated with the T-Coffee expresso pipeline [73], which guides the alignment by structural information. Using the obtained MSA (S5 File and S6 File), residues in all aaRS structures were renumbered with the custom script "MSA PDB Renumber", available under open-source license (MIT) at `github.com/vjhaupt`. All renumbered structures are provided in PDB file format (S11 File and S12 File). Only protein residues were renumbered, while chain identifiers and residue numbers of ligands were left unmodified. Lists of structures where the Backbone Brackets or Arginine Tweezers were not mapped successfully are found in S8 File and S10 File.

Annotation of noncovalent interactions

Annotation of noncovalent interactions between an aaRS protein and its bound ligand(s) was performed with the PLIP [75] command line tool v1.3.3 on all renumbered structures with default settings. The renumbered sequence positions of all residues observed to be in contact with the ligand were extracted. This resulting set of interacting residues was used to determine the position-identical residues from all aaRS structures in the dataset, even if no ligand is bound.

Generation of interaction matrix

Information on noncovalent protein-ligand interactions from renumbered structure files (see above) was used to prepare separate interaction matrices for aaRS Class I and Class II. First, only representative structures for M1 were selected. Second, only residues which are in contact with the non-amino acid part of the ligand (i.e. adenine, ribose moiety or the phosphate group) were considered. This was validated manually for each residue. Furthermore, residues relevant for only one aaRS Type were discarded. For each considered residue, the absolute frequency of observed ligand interactions was determined with respect to the PLIP interaction types (hydrophobic contacts, hydrogen bonds, salt bridges, π-stacking, and π-cation interactions). Additionally, the count of residues not interacting with any ligand ("no contact") was determined. In the interaction matrix (Fig 5), aaRS Types are placed on the abscissa and renumbered residue positions on the ordinate. The preferred interaction type for each residue and ligand species is color-coded. If two interaction types occurred with the same frequency, a dual coloring was used. Residues were grouped in the figure according to the ligand fragment they are mainly forming interactions with.

Annotation of mutagenesis sites and natural variants

For each chain, a mapping to UniProt [76] was performed using the SIFTS project [77]. Where available, mutation and natural variants data was retrieved for all binding site

residues from the UniProt [76] database. In total, 32 mutagenesis sites and 8 natural variants were retrieved.

Analysis of core-interaction patterns

All motif occurrences in M1 and M2 representative chains were aligned in respect to their backbone atoms (S7 Fig) using the Fit3D algorithm [120]. Additionally, the alpha carbon distances and the angle between side chains were determined. The side chain angle θ between two residues was calculated by abstracting each side chain as a vector between alpha carbon and the most distant carbon side chain atom. If $\theta = 0°$ or $\theta = 180°$ the side chains are oriented in a parallel way. Side chain angles were not calculated if one or both residues of the Backbone Brackets motif were glycine.

Furthermore, the sequential neighbors of the core-interaction patterns have been visualized with WebLogo graphics [74], regarding their sequence and secondary structure elements. Secondary structure elements were assigned according to the rule set of DSSP [121].

Codon assignment

The sequence regions proposed by Rodin and Ohno [8] were chosen as candidates for the codon assignment; the tangible positions are listed in Table 1 and Table 2. Cluster representative structures where chosen for the following analysis. In order to assign the original coding nucleotide sequence to each of the structures, the sequences of the structures were retrieved from the UniProt database [76] using the SIFTS project [77] to map PDB structures to UniProt entries. Afterwards, the corresponding codons were assigned to each amino acid by extracting them from the annotated coding sequences deposited in the European Nucleotide Archive [122]. Consensus codons were generated for each amino acid using WebLogo graphics [74] and choosing the most prominent nucleotide for positions with an entropy higher than one bit.

Acknowledgments

We thank Peter R. Wills for initially approaching us with the intriguing topic of the origin of genetic coding. Further, we appreciated the meetings and are grateful for guiding us the way through the entire project. Gratitude is owed to Lauren Adelmann, Hanna Siewerts, and Alexander Eisold for proofreading the manuscript. We thank the reviewers who proposed valuable improvements to the manuscript and helped us to strengthen the discussion substantially.

References

1. Mukai T, Lajoie MJ, Englert M, Soll D. Rewriting the Genetic Code. Annu Rev Microbiol. 2017;.

2. Lee JH, Choi SK, Roll-Mecak A, Burley SK, Dever TE. Universal conservation in translation initiation revealed by human and archaeal homologs of bacterial translation initiation factor IF2. Proc Natl Acad Sci USA. 1999;96(8):4342–4347.

3. Fox GE. Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol. 2010;2(9):a003483.

4. Ibba M, Söll D. Aminoacyl-tRNA synthesis. Annu Rev Biochem. 2000;69:617–650.

5. Di Giulio M. The origin of the genetic code: theories and their relationships, a review. Biosystems. 2005;80(2):175–184.

6. Gilbert W. Origin of life: The RNA world. nature. 1986;319(6055).

7. Wills PR. The generation of meaningful information in molecular systems. Phil Trans R Soc A. 2016;374(2063):20150066.

8. Rodin SN, Ohno S. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig Life Evol Biosph. 1995;25(6):565–589.

9. Martinez-Rodriguez L, Erdogan O, Jimenez-Rodriguez M, Gonzalez-Rivera K, Williams T, Li L, et al. Functional Class I and II Amino Acid-activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. J Biol Chem. 2015;290(32):19710–19725.

10. Wills PR. Spontaneous mutual ordering of nucleic acids and proteins. Orig Life Evol Biosph. 2014;44(4):293–298.

11. Carter CW. Coding of Class I and II Aminoacyl-tRNA Synthetases. Adv Exp Med Biol. 2017;966:103–148.

12. Wills PR, Carter CW. Insuperable problems of the genetic code initially emerging in an RNA world. Biosystems. 2017;.

13. Bernhardt HS. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others) a. Biology direct. 2012;7(1):23.

14. Wong JT. A co-evolution theory of the genetic code. Proc Natl Acad Sci USA. 1975;72(5):1909–1912.

15. Sonneborn T. Degeneracy of the genetic code: extent, nature, and genetic implications. Evolving genes and proteins. 1965; p. 377–397.

16. Woese CR. Order in the genetic code. Proceedings of the National Academy of Sciences. 1965;54(1):71–75.

17. Guimarães RC, Moreira CHC, de Farias ST. A self-referential model for the formation of the genetic code. Theory in Biosciences. 2008;127(3):249.

18. Wong JT. Coevolution theory of the genetic code at age thirty. Bioessays. 2005;27(4):416–425.

19. Brown JR, Doolittle WF. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proceedings of the National Academy of Sciences. 1995;92(7):2441–2445.

20. Schimmel P, Giege R, Moras D, Yokoyama S. An operational RNA code for amino acids and possible relationship to genetic code. Proceedings of the National Academy of Sciences. 1993;90(19):8763–8768.

21. Chandrasekaran SN, Yardimci GG, Erdogan O, Roach J, Carter CW. Statistical evaluation of the Rodin-Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases. Mol Biol Evol. 2013;30(7):1588–1604.

22. Eriani G, Delarue M, Poch O, Gangloff J, Moras D. Partition of tRNA Synthetases into Two Classes Based on Mutually Exclusive Sets of Sequence Motifs. Nature. 1990;347(6289):203.

23. Wolf YI, Aravind L, Grishin NV, Koonin EV. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome research. 1999;9(8):689–710.

24. Moras D. Structural and functional relationships between aminoacyl-tRNA synthetases. Trends Biochem Sci. 1992;17(4):159–164.

25. Carter CW, Wolfenden R. tRNA acceptor stem and anticodon bases form independent codes related to protein folding. Proceedings of the National Academy of Sciences. 2015;112(24):7489–7494.

26. Dock-Bregeon A, Sankaranarayanan R, Romby P, Caillet J, Springer M, Rees B, et al. Transfer RNA-mediated editing in threonyl-tRNA synthetase. The class II solution to the double discrimination problem. Cell. 2000;103(6):877–884.

27. Hadd A, Perona JJ. Coevolution of specificity determinants in eukaryotic glutamyl-and glutaminyl-tRNA synthetases. Journal of molecular biology. 2014;426(21):3619–3633.

28. Nair N, Raff H, Islam MT, Feen M, Garofalo DM, Sheppard K. The Bacillus subtilis and Bacillus halodurans Aspartyl-tRNA Synthetases Retain Recognition of tRNA Asn. Journal of molecular biology. 2016;428(3):618–630.

29. Arnez JG, Moras D. Structural and functional considerations of the aminoacylation reaction. Trends in biochemical sciences. 1997;22(6):211–216.

30. Ibba MP, Stange-Thomann N, Kitabatake M, Ali K, Söll I, Carter CW, et al. Ancient adaptation of the active site of tryptophanyl-tRNA synthetase for tryptophan binding. Biochemistry. 2000;39(43):13136–13143.

31. Burbaum JJ, Schimmel P. Structural relationships and the classification of aminoacyl-tRNA synthetases. J Biol Chem. 1991;266(26):16965–16968.

32. de Pouplana LR, Schimmel P. Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends in biochemical sciences. 2001;26(10):591–596.

33. Schimmel P, Ripmaster T. Modular design of components of the operational RNA code for alanine in evolution. Trends in biochemical sciences. 1995;20(9):333–334.

34. Chaliotis A, Vlastaridis P, Mossialos D, Ibba M, Becker HD, Stathopoulos C, et al. The complex evolutionary history of aminoacyl-tRNA synthetases. Nucleic Acids Res. 2017;45(3):1059–1068.

35. Rould M, Perona J, Steitz T. Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase. Nature. 1991;352(6332):213.

36. Normanly J, Abelson J. tRNA identity. Annual review of biochemistry. 1989;58(1):1029–1049.

37. Martinis SA, Boniecki MT. The balance between pre- and post-transfer editing in tRNA synthetases. FEBS Lett. 2010;584(2):455–459.

38. Splan KE, Ignatov ME, Musier-Forsyth K. Transfer RNA modulates the editing mechanism used by class II prolyl-tRNA synthetase. J Biol Chem. 2008;283(11):7128–7134.

39. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci. 1993;9(6):745–756.

40. Belrhali H, Yaremchuk A, Tukalo M, Berthet-Colominas C, Rasmussen B, Bosecke P, et al. The structural basis for seryl-adenylate and Ap4A synthesis by seryl-tRNA synthetase. Structure. 1995;3(4):341–352.

41. Fujinaga M, Berthet-Colominas C, Yaremchuk AD, Tukalo MA, Cusack S. Refined crystal structure of the seryl-tRNA synthetase from Thermus thermophilus at 2.5 A resolution. J Mol Biol. 1993;234(1):222–233.

42. Ambrogelly A, Söll D, Nureki O, Yokoyama S, Ibba M. Class I Lysyl-tRNA Synthetases. Landes Bioscience; 2013.

43. Diaz-Lazcoz Y, Aude J, Nitschke P, Chiapello H, Landes-Devauchelle C, Risler J. Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. Molecular biology and evolution. 1998;15(11):1548–1561.

44. Woese CR, Olsen GJ, Ibba M, Söll D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiology and Molecular Biology Reviews. 2000;64(1):202–236.

45. Schmitt E, Panvert M, Blanquet S, Mechulam Y. Transition state stabilization by the 'high'motif of class I aminoacyl-tRNA synthetases: the case of Escherichia coli methionyl-tRNA synthetase. Nucleic acids research. 1995;23(23):4793–4798.

46. First EA, Fersht AR. Involvement of threonine 234 in catalysis of tyrosyl adenylate formation by tyrosyl-tRNA synthetase. Biochemistry. 1993;32(49):13644–13650.

47. First EA, Fersht AR. Involvement of threonine 234 in catalysis of tyrosyl adenylate formation by tyrosyl-tRNA synthetase. Biochemistry. 1993;32(49):13644–13650.

48. First EA, Fersht AR. Mutational and kinetic analysis of a mobile loop in tyrosyl-tRNA synthetase. Biochemistry. 1993;32(49):13658–13663.

49. First EA, Fersht AR. Analysis of the role of the KMSKS loop in the catalytic mechanism of the tyrosyl-tRNA synthetase using multimutant cycles. Biochemistry. 1995;34(15):5030–5043.

50. Chandrasekaran SN, Das J, Dokholyan NV, Carter CW. A modified PATH algorithm rapidly generates transition states comparable to those found by other well established algorithms. Struct Dyn. 2016;3(1):012101.

51. Chandrasekaran SN, Carter CW. Augmenting the anisotropic network model with torsional potentials improves PATH performance, enabling detailed comparison with experimental rate data. Struct Dyn. 2017;4(3):032103.

52. Carter CW, Chandrasekaran SN, Weinreb V, Li L, Williams T. Combining multi-mutant and modular thermodynamic cycles to measure energetic coupling networks in enzyme catalysis. Struct Dyn. 2017;4(3):032101.

53. Weinreb V, Li L, Carter CW. A master switch couples Mg2+-assisted catalysis to domain motion in B. stearothermophilus tryptophanyl-tRNA Synthetase. Structure. 2012;20(1):128–138.

54. Weinreb V, Li L, Chandrasekaran SN, Koehl P, Delarue M, Carter CW. Enhanced amino acid selection in fully evolved tryptophanyl-tRNA synthetase, relative to its urzyme, requires domain motion sensed by the D1 switch, a remote dynamic packing motif. J Biol Chem. 2014;289(7):4367–4376.

55. Eriani G, Cavarelli J, Martin F, Dirheimer G, Moras D, Gangloff J. Role of dimerization in yeast aspartyl-tRNA synthetase and importance of the class II invariant proline. Proceedings of the National Academy of Sciences. 1993;90(22):10816–10820.

56. Åberg A, Yaremchuk A, Tukalo M, Rasmussen B, Cusack S. Crystal Structure Analysis of the Activation of Histidine by Thermus thermophilus Histidyl-tRNA Synthetase. Biochemistry. 1997;36(11):3084–3094.

57. Cusack S. Aminoacyl-tRNA synthetases. Current opinion in structural biology. 1997;7(6):881–889.

58. Cusack S, Berthet-Colominas C, Hartlein M, Nassar N, Leberman R. A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature. 1990;347(6290):249.

59. Cusack S, Hartlein M, Leberman R. Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases. Nucleic Acids Res. 1991;19(13):3489–3498.

60. O'Donoghue P, Luthey-Schulten Z. On the evolution of structure in aminoacyl-tRNA synthetases. Microbiology and Molecular Biology Reviews. 2003;67(4):550–573.

61. Banik SD, Nandi N. Mechanism of the activation step of the aminoacylation reaction: a significant difference between class I and class II synthetases. J Biomol Struct Dyn. 2012;30(6):701–715.

62. LeJohn HB, Cameron LE, Yang B, Rennie SL. Molecular characterization of an NAD-specific glutamate dehydrogenase gene inducible by L-glutamine. Antisense gene pair arrangement with L-glutamine-inducible heat shock 70-like protein gene. Journal of Biological Chemistry. 1994;269(6):4523–4531.

63. Carter CW, Duax WL. Did tRNA synthetase classes arise on opposite strands of the same gene? Molecular cell. 2002;10(4):705–708.

64. Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, et al. Over 20% of human transcripts might form sense–antisense pairs. Nucleic acids research. 2004;32(16):4812–4820.

65. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al. Antisense transcription in the mammalian transcriptome. Science. 2005;309(5740):1564–1566.

66. Carter CW, Li L, Weinreb V, Collier M, Gonzalez-Rivera K, Jimenez-Rodriguez M, et al. The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. Biol Direct. 2014;9:11.

67. Ruff M, Krishnaswamy S, Boeglin M, Poterszman A, Mitschler A, Podjarny A, et al. Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA (Asp). Science. 1991;252(5013):1682–1689.

68. Schulze JO, Masoumi A, Nickel D, Jahn M, Jahn D, Schubert WD, et al. Crystal structure of a non-discriminating glutamyl-tRNA synthetase. Journal of molecular biology. 2006;361(5):888–897.

69. Mailu BM, Ramasamay G, Mudeppa DG, Li L, Lindner SE, Peterson MJ, et al. A nondiscriminating glutamyl-tRNA synthetase in the Plasmodium apicoplast the first enzyme in an indirect aminoacylation pathway. Journal of Biological Chemistry. 2013;288(45):32539–32552.

70. Pham Y, Li L, Kim A, Erdogan O, Weinreb V, Butterfoss GL, et al. A minimal TrpRS catalytic domain supports sense/antisense ancestry of class I and II aminoacyl-tRNA synthetases. Molecular cell. 2007;25(6):851–862.

71. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–242.

72. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–2309.

73. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 2006;34(Web Server issue):W604–608.

74. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–1190.

75. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res. 2015;43(W1):W443–447.

76. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2017;45(D1):D158. doi:10.1093/nar/gkw1099.

77. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research. 2013;41(D1):D483. doi:10.1093/nar/gks1258.

78. Vidal-Cros A, Bedouelle H. Role of residue Glu152 in the discrimination between transfer RNAs by tyrosyl-tRNA synthetase from Bacillus stearothermophilus. J Mol Biol. 1992;223(3):801–810.

79. Xin Y, Li W, Dwyer DS, First EA. Correlating amino acid conservation with function in tyrosyl-tRNA synthetase. J Mol Biol. 2000;303(2):287–298.

80. Griffin LB, Sakaguchi R, McGuigan D, Gonzalez MA, Searby C, Zuchner S, et al. Impaired function is a common feature of neuropathy-associated glycyl-tRNA synthetase mutations. Hum Mutat. 2014;35(11):1363–1371.

81. Miller WT, Hill KA, Schimmel P. Evidence for a “cysteine-histidine box” metal-binding site in an Escherichia coli aminoacyl-tRNA synthetase. Biochemistry. 1991;30(28):6970–6976.

82. Xin Y, Li W, First EA. Stabilization of the transition state for the transfer of tyrosine to tRNA(Tyr) by tyrosyl-tRNA synthetase. J Mol Biol. 2000;303(2):299–310.

83. Xie W, Nangle LA, Zhang W, Schimmel P, Yang XL. Long-range structural effects of a Charcot-Marie-Tooth disease-causing mutation in human glycyl-tRNA synthetase. Proceedings of the National Academy of Sciences. 2007;104(24):9976–9981.

84. Xin Y, Li W, First EA. The 'KMSKS' motif in tyrosyl-tRNA synthetase participates in the initial binding of tRNA(Tyr). Biochemistry. 2000;39(2):340–347.

85. Carter CW. Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases. Annu Rev Biochem. 1993;62:715–748.

86. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–1191.

87. Berg JM. Potential metal-binding domains in nucleic acid binding proteins. Science. 1986;232(4749):485–487.

88. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–311.

89. Alva V, Soding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. Elife. 2015;4:e09410.

90. Najmanovich RJ. Evolutionary studies of ligand binding sites in proteins. Curr Opin Struct Biol. 2016;45:85–90.

91. Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. Trends in biochemical sciences. 2005;30:622–629.

92. Salentin S, Haupt VJ, Daminelli S, Schroeder M. Polypharmacology rescored: protein-ligand interaction profiles for remote binding site similarity assessment. Prog Biophys Mol Biol. 2014;116(2-3):174–186.

93. Samish I, Bourne PE, Najmanovich RJ. Achievements and challenges in structural bioinformatics and computational biophysics. Bioinformatics. 2015;31(1):146–150.

94. Caetano-Anolles G, Wang M, Caetano-Anolles D, Mittenthal JE. The origin, evolution and structure of the protein world. Biochem J. 2009;417(3):621–637.

95. Delagoutte B, Moras D, Cavarelli J. tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding. The EMBO Journal. 2000;19(21):5599–5610.

96. Kobayashi T, Takimura T, Sekine R, Kelly VP, Vincent K, Kamata K, et al. Structural snapshots of the KMSKS loop rearrangement for amino acid activation by bacterial tyrosyl-tRNA synthetase. J Mol Biol. 2005;346(1):105–117.

97. Barelier S, Sterling T, O'Meara MJ, Shoichet BK. The Recognition of Identical Ligands by Unrelated Proteins. ACS Chem Biol. 2015;10(12):2772–2784.

98. Stockwell GR, Thornton JM. Conformational diversity of ligands bound to proteins. J Mol Biol. 2006;356(4):928–944.

99. Schmitt E, Moulinier L, Fujiwara S, Imanaka T, Thierry JC, Moras D. Crystal structure of aspartyl-tRNA synthetase from Pyrococcus kodakaraensis KOD: archaeon specificity and catalytic mechanism of adenylate formation. EMBO J. 1998;17(17):5227–5237.

100. Cavarelli J, Eriani G, Rees B, Ruff M, Boeglin M, Mitschler A, et al. The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction. EMBO J. 1994;13(2):327–337.

101. Navarre WW, Zou SB, Roy H, Xie JL, Savchenko A, Singer A, et al. PoxA, yjeK, and elongation factor P coordinately modulate virulence and drug resistance in Salmonella enterica. Mol Cell. 2010;39(2):209–221.

102. Giege R, Springer M. Aminoacyl-tRNA Synthetases in the Bacterial World. EcoSal Plus. 2016;7(1).

103. Eigen M, Schuster P. A principle of natural self-organization. Naturwissenschaften. 1977;64(11):541–565.

104. Zull JE, Smith SK. Is genetic code redundancy related to retention of structural information in both DNA strands? Trends in biochemical sciences. 1990;15(7):257–261.

105. Gallina AM, Bork P, Bordo D. Structural analysis of protein-ligand interactions: the binding of endogenous compounds and of synthetic drugs. J Mol Recognit. 2014;27(2):65–72.

106. Wolf YI, Koonin EV. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. Biology Direct. 2007;2(1):14.

107. Cusack S. Sequence, structure and evolutionary relationships between class 2 aminoacyl-tRNA synthetases: an update. Biochimie. 1993;75(12):1077–1081.

108. Guo LT, Chen XL, Zhao BT, Shi Y, Li W, Xue H, et al. Human tryptophanyl-tRNA synthetase is switched to a tRNA-dependent mode for tryptophan activation by mutations at V85 and I311. Nucleic Acids Res. 2007;35(17):5934–5943.

109. Simons C, Griffin LB, Helman G, Golas G, Pizzino A, Bloom M, et al. Loss-of-function alanyl-tRNA synthetase mutations cause an autosomal-recessive early-onset epileptic encephalopathy with persistent myelination defect. Am J Hum Genet. 2015;96(4):675–681.

110. Datt M, Sharma A. Evolutionary and structural annotation of disease-associated mutations in human aminoacyl-tRNA synthetases. BMC Genomics. 2014;15:1063.

111. Stum M, McLaughlin HM, Kleinbrink EL, Miers KE, Ackerman SL, Seburn KL, et al. An assessment of mechanisms underlying peripheral axonal degeneration caused by aminoacyl-tRNA synthetase mutations. Mol Cell Neurosci. 2011;46(2):432–443.

112. Mirando AC, Francklyn CS, Lounsbury KM. Regulation of angiogenesis by aminoacyl-tRNA synthetases. Int J Mol Sci. 2014;15(12):23725–23748.

113. Randall CP, Rasina D, Jirgensons A, O'Neill AJ. Targeting Multiple Aminoacyl-tRNA Synthetases Overcomes the Resistance Liabilities Associated with Antibacterial Inhibitors Acting on a Single Such Enzyme. Antimicrob Agents Chemother. 2016;60(10):6359–6361.

114. Pham JS, Dawson KL, Jackson KE, Lim EE, Pasaje CF, Turner KE, et al. Aminoacyl-tRNA synthetases as drug targets in eukaryotic parasites. Int J Parasitol Drugs Drug Resist. 2014;4(1):1–13.

115. Chopra S, Palencia A, Virus C, Schulwitz S, Temple BR, Cusack S, et al. Structural characterization of antibiotic self-immunity tRNA synthetase in plant tumour biocontrol agent. Nat Commun. 2016;7:12928.

116. Merritt EA, Arakaki TL, Gillespie JR, Larson ET, Kelley A, Mueller N, et al. Crystal structures of trypanosomal histidyl-tRNA synthetase illuminate differences between eukaryotic and prokaryotic homologs. J Mol Biol. 2010;397(2):481–494.

117. Challis GL, Ravel J, Townsend CA. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chem Biol. 2000;7(3):211–224.

118. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–285.

119. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–453.

120. Kaiser F, Eisold A, Bittrich S, Labudde D. Fit3D: a web application for highly accurate screening of spatial residue patterns in protein structure data. Bioinformatics. 2016;32(5):792–794.

121. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–2637. doi:10.1002/bip.360221211.

122. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European nucleotide archive. Nucleic acids research. 2010;39(suppl_1):D28–D31.

Supporting information

S1 Fig. Binding mode definition. Binding modes M1 and M2 are defined based on the complexed ligand: ligands that bind to the adenosine phosphate moiety (highlighted in red, only in contact when adenosine phosphate is part of the ligand) of the binding site (M1), no ligands or ligands that bind exclusively to the aminoacyl part (green) of the binding site (M2).

S2 Fig. Core-interaction patterns. Both aaRS classes contain highly conserved patterns, responsible for proper binding of the adenosine phosphate part of the ligand. Class I aaRS share a highly conserved set of backbone hydrogen interactions with the ligand: the Backbone Brackets. Class II active sites contain a pattern of two arginine residues grasping the adenosine phosphate ligand: the Arginine Tweezers. Interactions

were calculated with PLIP [75] and are represented with colored (dashed) lines: hydrogen bonds (solid, blue), π-stacking interactions (dashed, green), π-cation interactions (dashed, orange), salt bridges (dashed, yellow), metal complexes (dashed, purple), and hydrophobic contacts (dashed grey). (**A**) Class I Backbone Brackets motif and interactions with the ligand Tryptophanyl-5'AMP as observed in TrpRS structure PDB:1r6u chain A. (**B**) Class II Arginine Tweezers motif and interactions with the ligand Lysyl-5'AMP as observed in LysRS structure PDB:1e1t chain A.

S3 Fig. Secondary structure of Backbone Brackets adjacent residues. WebLogo [74] representation of secondary structure elements around the Backbone Brackets residues (274 and 1361) annotated by DSSP [121]: helices (blues), strands (red), and unordered (black). Unassigned states are represented by the character 'C'. The height of each character corresponds to the relative frequency.

S4 Fig. Secondary structure of Arginine Tweezers adjacent residues. WebLogo [74] representation of secondary structure elements around the Arginine Tweezers residues (698 and 1786) annotated by DSSP [121]: helices (blues), strands (red), and unordered (black). Unassigned states are represented by the letter 'C'. The height of each character corresponds to the relative frequency.

S5 Fig. Distributions of alpha carbon distances for Backbone Brackets and Arginine Tweezers. Distributions of alpha carbon distances for Class I Backbone Brackets motif and Class II Arginine Tweezers motif in adenosine phosphate bound (M1) and unbound state (M2). The alpha carbon distance of the Backbone Brackets differs significantly between the two states (Mann-Whitney U $p<0.01$).

S6 Fig. Distributions of side chain angles for Backbone Brackets and Arginine Tweezers. Distributions of side chain angle θ for Class I Backbone Brackets motif and Class II Arginine Tweezers motif in adenosine phosphate bound (M1) and unbound state (M2). The side chain angles of the Arginine Tweezers differs differs significantly between the two states (Mann-Whitney U $p<0.01$).

S7 Fig. Alignments of Backbone Brackets and Arginine Tweezers. Structural backbone-only alignments computed with Fit3D [120] of relevant binding site residue motifs derived from M1 and M2 representative structures in respect to their binding modes for aaRS Class I and Class II. (**A**) The Class I Backbone Brackets motif aligned in respect to binding modes. A high side chain variance (gray line representation) is evident. However, backbone orientations are highly conserved to realize consistent hydrogen bond interaction with the adenosine phosphate part of the ligand. (**B**) The Class II Arginine Tweezers motif aligned in respect to binding modes. Low side chain variance can be observed if an adenosine phosphate ligand is bound (M1), whereas the absence of an adenosine phosphate ligand (M2) allows an increased degree of freedom for side chain movement. Averaged backbone and side chain RMSD values after all-vs-all superimposition are shown in S1 Table.

S8 Fig. Pairwise sequence and structure similarity. Structure and sequence similarity for pairs of cluster representative chains for aaRS Class I (**A**) and II (**B**). Depicted is the sequence similarity (% identity) after a global Needleman-Wunsch [119] alignment of both structures against the structure similarity determined by TMAlign [72]. For Class I (Class II) 95% of all pairs exhibit <33% (29%) sequence identity and <0.85 (0.84) TM score. The 95% quantile borders are depicted as red dashed lines.

S9 Fig. Origin organisms of aaRS Class I and Class II structures in the dataset. The organisms of origin for aaRS Class I (**A**) and Class II (**B**) structures in the dataset. The inner circles correspond to the superkingdom of the organism. The outer circle depicts the partition into specific species (combining different strains). Sections representing eukaryotic species are colored in violet, bacteria are colored in green, archaea are colored in orange and vira are colored in gray. Species, that are origin of less than two percent of the structures are condensed to the "other" cluster for each superkingdom. All superkingdoms are represented in both datasets. Class I contains more bacterial structures than Class II, but fewer originating from eukaryotes or archaea. Interestingly, Class I also contains one viral structure. Despite the diverse origins of the structures the conserved interaction patterns can be observed.

S1 Table. Backbone and side chain RMSD of Backbone Brackets and Arginine Tweezers after superimposition. Averaged backbone and side chain RMSD values after all-vs-all superimposition are shown in this table.

Table 3. Averaged backbone RMSD values of Backbone Brackets and Arginine Tweezers after superimposition.

motif	binding mode	observations	backbone RMSD [Å]
Backbone Brackets	M1	28	0.32
Backbone Brackets	M2	59	0.34
Arginine Tweezers	M1	39	0.24
Arginine Tweezers	M2	47	0.28

M1 contains an adenosine phosphate ligand, no adenosine phosphate ligand in M2.

S1 Appendix. Dataset preparation. All selected protein chains from the PDB carry one of the following protein family annotations, according to Pfam [118]: PF00133, PF00152, PF00579, PF00587, PF00749, PF00750, PF01406, PF01409, PF01411, PF02091, PF02403, PF02912, PF03485, PF09334. Additionally, structures annotated with an EC number indicating tRNA-ligation activity were selected: 6.1.1.1 (TyrRS), 6.1.1.2 (TrpRS), 6.1.1.3 (ThrRS), 6.1.1.4 (LeuRS), 6.1.1.5 (IleRS), 6.1.1.6 (LysRS), 6.1.1.7 (AlaRS), 6.1.1.9 (ValRS), 6.1.1.10 (MetRS), 6.1.1.11 (SerRS), 6.1.1.14 (GlyRS), 6.1.1.15 (ProRS), 6.1.1.16 (CysRS), 6.1.1.17 (GluRS), 6.1.1.18 (GlnRS), 6.1.1.19 (ArgRS), 6.1.1.20 (PheRS), 6.1.1.21 (HisRS), 6.1.1.22 (AsnRS), 6.1.1.23 (AspRS), 6.1.1.26 (PylRS).

For each of the resulting structures, the existence of a catalytic domain was checked manually and only the chains containing a domain with confirmed catalytic activity were retained. If there were ligands present in the structure, these ligands were annotated manually to avoid errors in the assignment of ligands to their catalytic chain. This procedure resulted in a high quality dataset of 972 individual aaRS chains containing a catalytic domain.

S2 Appendix. Selection of representative entries. In order to avoid redundancy, representatives were defined for each sequence cluster with >95% sequence identity and discriminated between three types: cluster representatives, representatives that contain an adenosine phosphate ligand (M1), and representatives that do not contain an adenosine phosphate ligand (M2).

The selection criteria for these categories were defined as follows:

- cluster representative:

    1. protein must be wild type (if wild type exists)
    2. best resolution
    3. longest sequence coverage

- representatives with an adenosine-relevant ligand
    1. chain must contain an adenosine phosphate ligand
    2. this ligand must be standard (adenosine phosphate or close derivate)
    3. no experimentally validated inhibitor ligand in the binding site
    4. protein must be wild type (if wild type exists)
    5. best resolution
- representatives without an adenosine phosphate ligand
    1. chain must not contain an adenosine phosphate ligand
    2. binding site must not contain an inhibitor ligand
    3. protein must be wild type (if wild type exists)
    4. best resolution

For ties in the selection, structures were sorted naturally ascending according to their PDB identifier and chain identifier and the first was chosen.

S1 File. Dataset as table. Summary table of all aaRS protein chains used for the analysis, including PDB identifier, chain identifier, superkingdom, taxonomy identifier, and ligand information (if any).

S2 File. Dataset as JSON file. Machine-readable JSON version of the dataset. Additionally enriched with protein sequence, sequence cluster identifier, and representative types for each dataset entry.

S3 File. Class I sequences in FASTA format. Protein sequences of Class I aaRS structures used to construct the structure-guided MSA in FASTA format.

S4 File. Class II sequences in FASTA format. Protein sequences of Class II aaRS structures used to construct the structure-guided MSA in FASTA format.

S5 File. Class I multiple sequence alignment. Structure-guided MSA of Class I sequences in FASTA format.

S6 File. Class II multiple sequence alignment. Structure-guided MSA of Class II sequences in FASTA format.

S7 File. Backbone Brackets residue mapping. Mapping of the Backbone Brackets Class I motif to sequence positions in origin structures.

S8 File. Backbone Brackets failed mapping. List of structures where the mapping of the Backbone Brackets motif was not possible.

S9 File. Arginine Tweezers residue mapping. Mapping of the Arginine Tweezers Class II motif to sequence positions in origin structures.

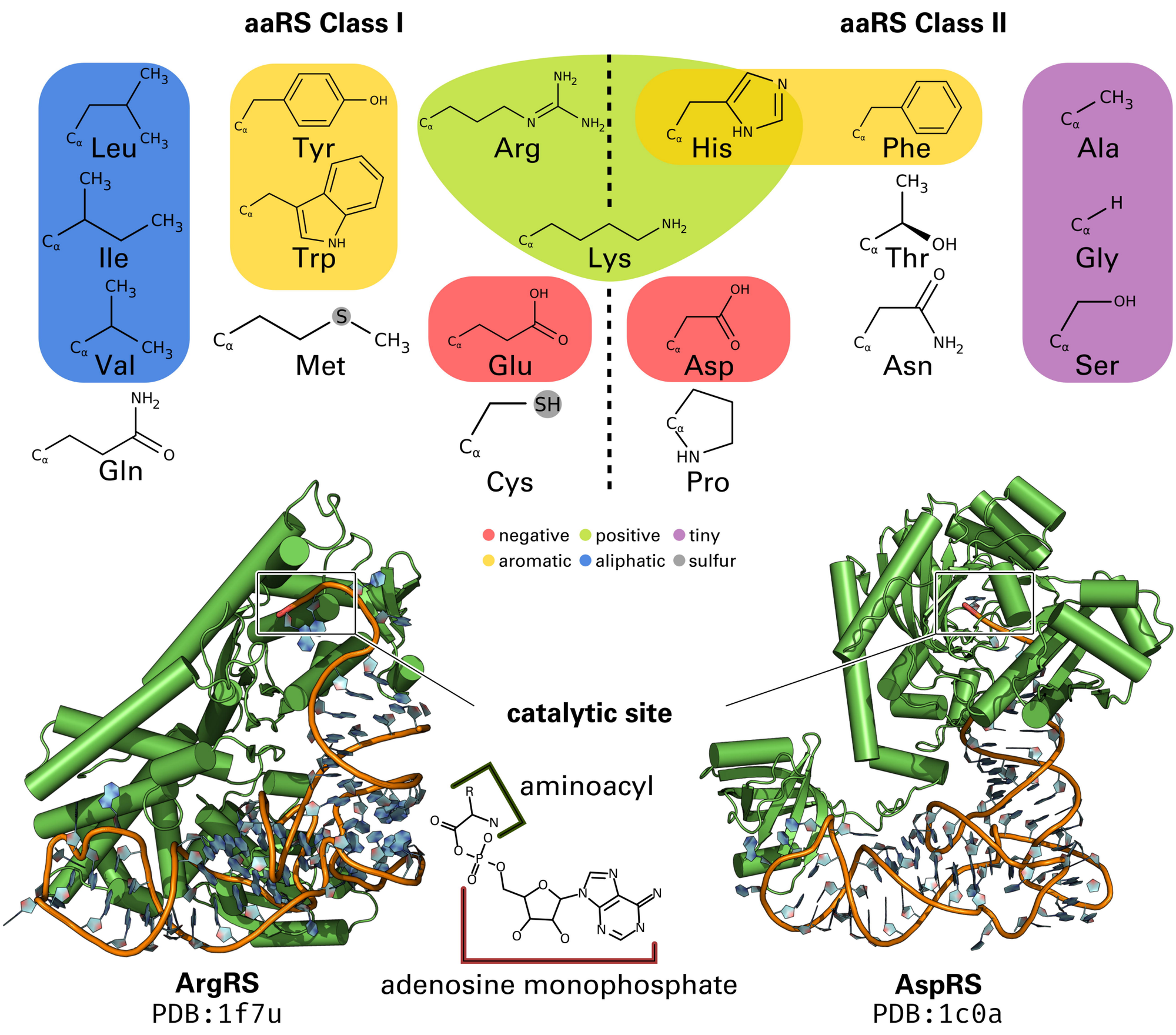
S10 File. Arginine Tweezers failed mapping. List of structures where the mapping of the Arginine Tweezers motif was not possible.

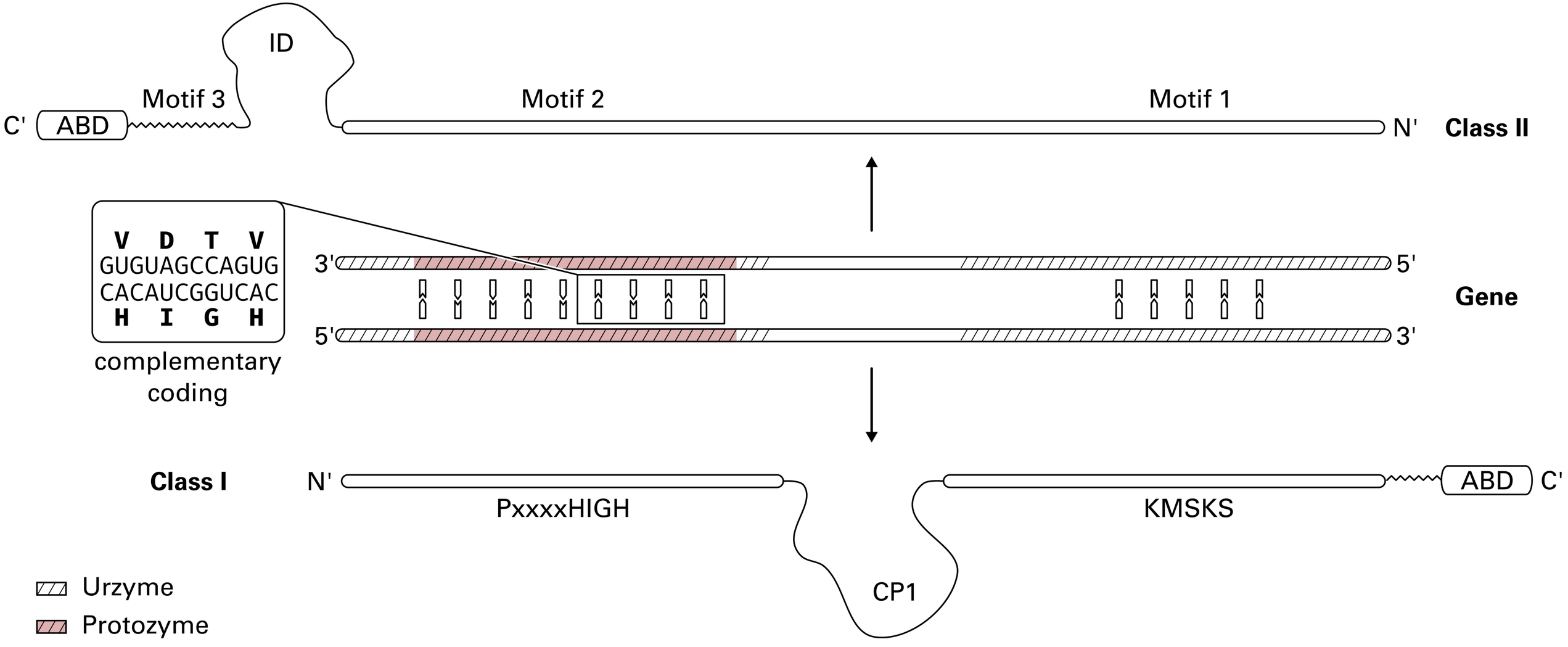
S11 File. Archive containing Class I renumbered structures. All structures of Class I aaRS with residues renumbered according to the MSA.

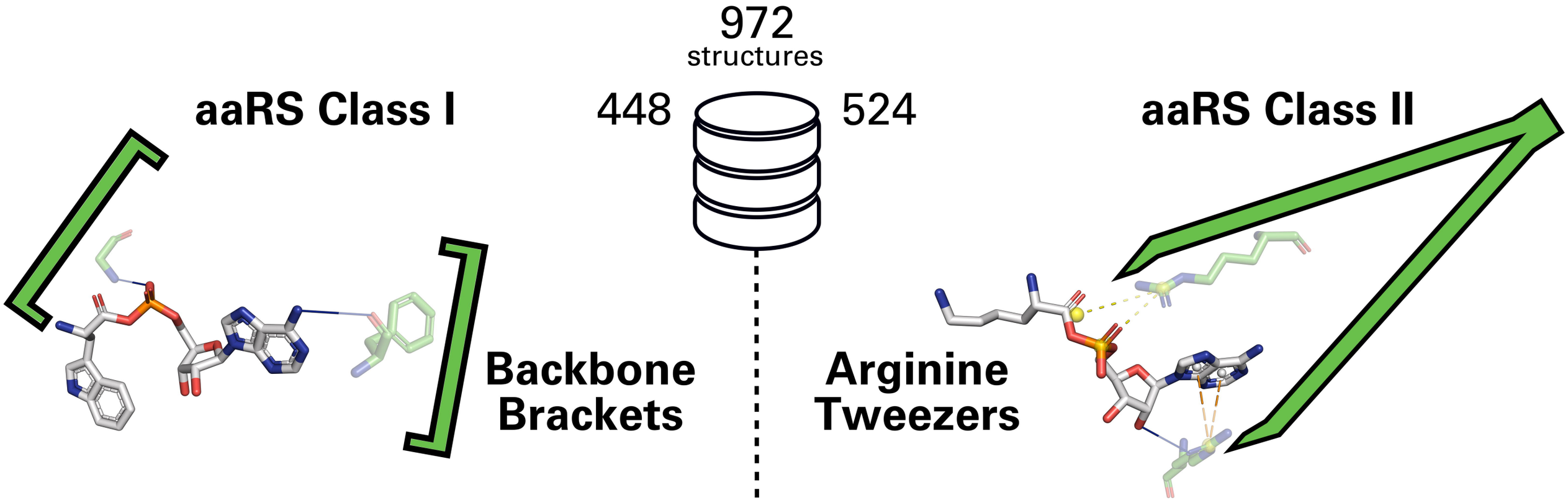
S12 File. Archive containing Class II renumbered structures. All structures of Class II aaRS with residues renumbered according to the MSA.

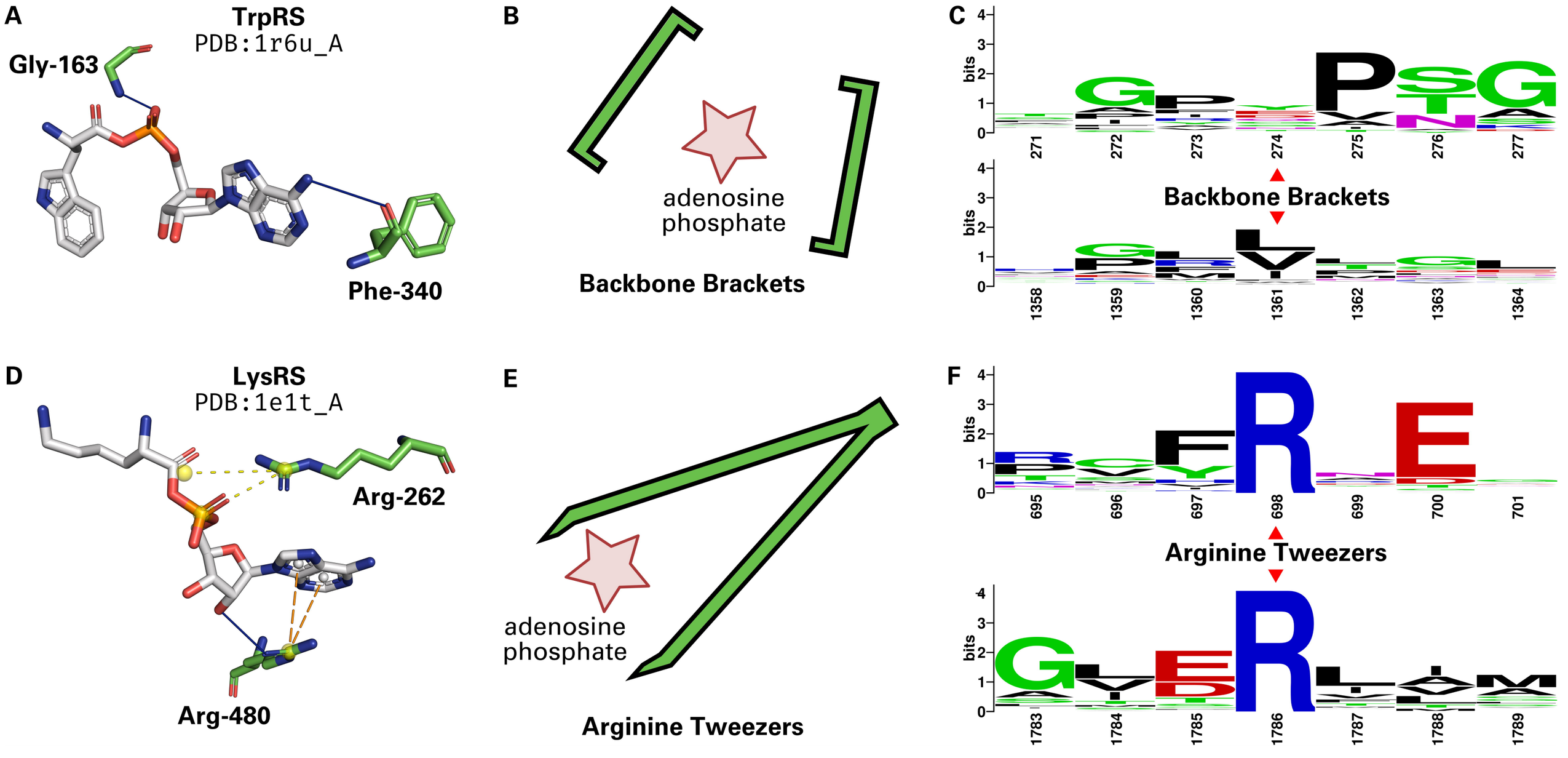
S13 File. Renumbering table for Class I structures. Formatted table that contains all sequence positions of the Class I MSA and annotations of sequence motifs, Backbone Brackets' residues, and ligand binding regions (rows). Each renumbered sequence position is related to its original sequence position for every structure in the dataset (columns).

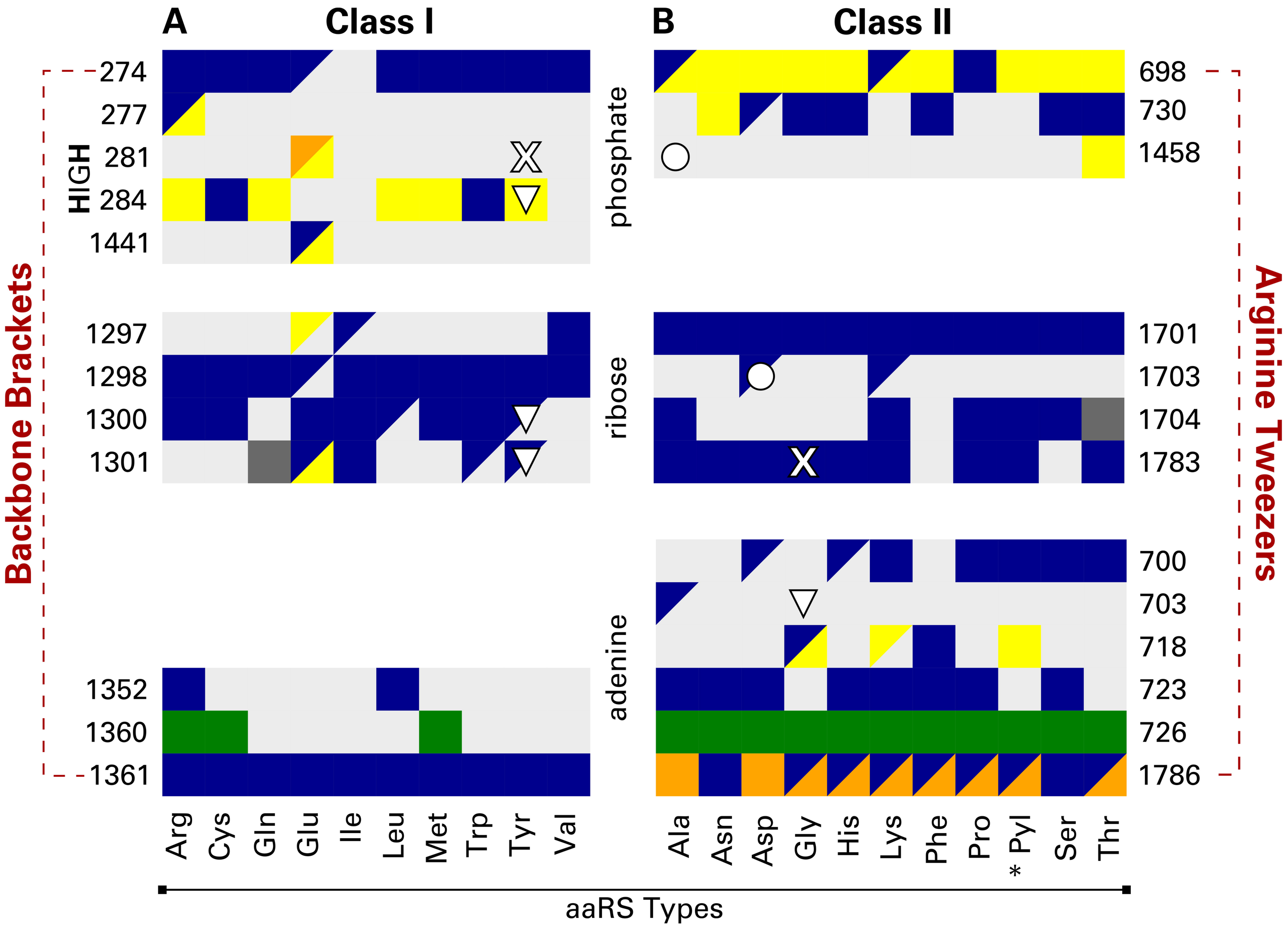
S14 File. Renumbering table for Class II structures. Formatted table that contains all sequence positions of the Class II MSA and annotations of sequence motifs, Arginine Tweezers' residues, and ligand binding regions (rows). Each renumbered sequence position is related to its original sequence position for every structure in the dataset (columns).

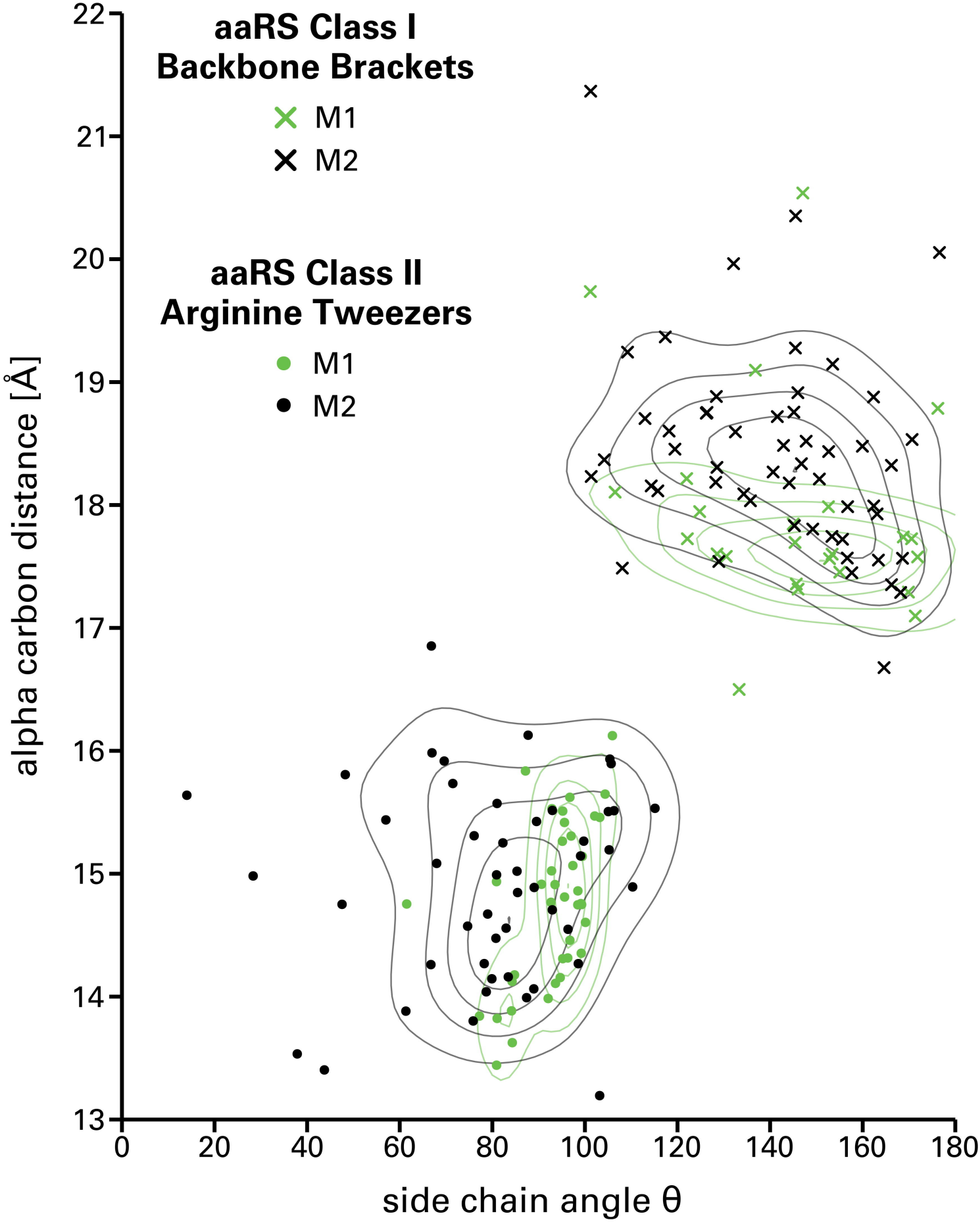

aaRS Class I
aaRS Class II
Leu
Ile
Val
Tyr
Trp
Arg
Lys
His
Phe
Thr
Ala
Gly
Ser
Met
Glu
Asp
Asn
Gln
Cys
Pro
negative
positive
tiny
aromatic
aliphatic
sulfur
catalytic site
aminoacyl
adenosine monophosphate
ArgRS
PDB:1f7u
AspRS
PDB:1c0a

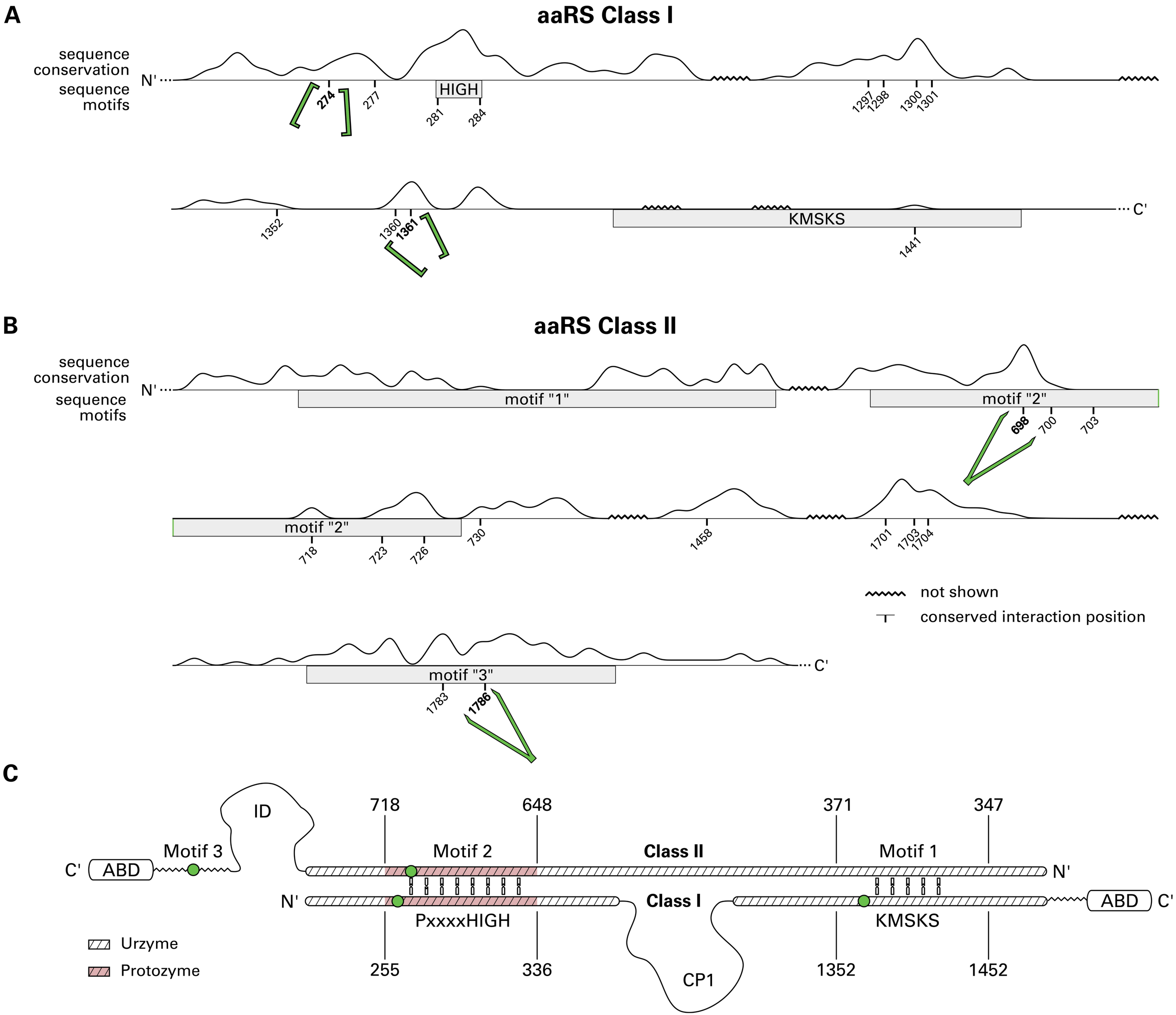

ID
Motif 3
Motif 2
Motif 1
C'
ABD
N'
Class II
V D T V
GUGUAGCCAGUG
CACAUCGGUCAC
H I G H
complementary
coding
3'
5'
5'
3'
Gene
Class I
N'
PxxxxHIGH
CP1
KMSKS
ABD
C'
Urzyme
Protozyme


972
structures
448
524
aaRS Class I
aaRS Class II
Backbone
Brackets
Arginine
Tweezers


A
TrpRS
PDB:1r6u_A
Gly-163
Phe-340
B
adenosine
phosphate
Backbone Brackets
C
bits
271
272
273
274
275
276
277
Backbone Brackets
1358
1359
1360
1361
1362
1363
1364
D
LysRS
PDB:1e1t_A
Arg-262
Arg-480
E
adenosine
phosphate
Arginine Tweezers
F
695
696
697
698
699
700
701
Arginine Tweezers
1783
1784
1785
1786
1787
1788
1789


A
Class I
B
Class II
Backbone Brackets
Arginine Tweezers
HIGH
274
277
281
284
1441
1297
1298
1300
1301
1352
1360
1361
phosphate
ribose
adenine
698
730
1458
1701
1703
1704
1783
700
703
718
723
726
1786
Arg
Cys
Gln
Glu
Ile
Leu
Met
Trp
Tyr
Val
Ala
Asn
Asp
Gly
His
Lys
Phe
Pro
* Pyl
Ser
Thr
aaRS Types
interaction types
hydrophobic
hydrogen bond
salt bridge
π-cation
π-stacking
no contact
mutations
no effect
decreased activity
loss of activity


aaRS Class I
Backbone Brackets
M1
M2
aaRS Class II
Arginine Tweezers
M1
M2
alpha carbon distance [Å]
side chain angle θ
22
21
20
19
18
17
16
15
14
13
0
20
40
60
80
100
120
140
160
180


A
aaRS Class I
sequence conservation
sequence motifs
N'
C'
HIGH
KMSKS
274
277
281
284
1297
1298
1300
1301
1352
1360
1361
1441
B
aaRS Class II
motif "1"
motif "2"
motif "3"
698
700
703
718
723
726
730
1458
1701
1703
1704
1783
1786
not shown
conserved interaction position
C
ID
Motif 3
ABD
Motif 2
Class II
Motif 1
PxxxxHIGH
Class I
KMSKS
CP1
718
648
371
347
255
336
1352
1452
Urzyme
Protozyme