SCIENTIFIC REPORTS
nature research

OPEN

The structural basis of the genetic code: amino acid recognition by aminoacyl-tRNA synthetases

Florian Kaiser[1,2,4]✉, Sarah Krautwurst[3,4], Sebastian Salentin[1], V. Joachim Haupt[1,2], Christoph Leberecht[3], Sebastian Bittrich[3], Dirk Labudde[3] & Michael Schroeder[1]

Storage and directed transfer of information is the key requirement for the development of life. Yet any information stored on our genes is useless without its correct interpretation. The genetic code defines the rule set to decode this information. Aminoacyl-tRNA synthetases are at the heart of this process. We extensively characterize how these enzymes distinguish all natural amino acids based on the computational analysis of crystallographic structure data. The results of this meta-analysis show that the correct read-out of genetic information is a delicate interplay between the composition of the binding site, non-covalent interactions, error correction mechanisms, and steric effects.

One of the most profound open questions in biology is how the genetic code was established. While proteins are encoded by nucleic acid blueprints, decoding this information in turn requires proteins. The emergence of this self-referencing system poses a chicken-or-egg dilemma and its origin is still heavily debated[1,2]. Aminoacyl-tRNA synthetases (aaRSs) implement the correct assignment of amino acids to their codons and are thus inherently connected to the emergence of genetic coding. These enzymes link tRNA molecules with their amino acid cargo and are consequently vital for protein biosynthesis. Beside the correct recognition of tRNA features[3], highly specific non-covalent interactions in the binding sites of aaRSs are required to correctly detect the designated amino acid[4–7] and to prevent errors in biosynthesis[5,8]. The minimization of such errors represents the utmost barrier for the development of biological complexity[9] and accurate specification of aaRS binding sites is proposed to be one of the major determinants for the closure of the genetic code[10]. Beside binding side features, recognition fidelity is controlled by the ratio of concentrations of aaRSs and cognate tRNA molecules[11] and may involve spatial secondary structures motifs in addition to side chain configurations[12,13].

Evolution. The evolutionary origin of aaRSs is hard to track. Phylogenetic analyses of aaRS sequences show that they do not follow the standard model of life[14]; the development of aaRSs was nearly complete before the Last Universal Common Ancestor (LUCA)[15,16]. Their complex evolutionary history included horizontal gene transfer, fusion, duplication, and recombination events[14,17–21]. Sequence analyses[22] and subsequent structure investigations[23,24] revealed that aaRSs can be divided into two distinct classes (*Class I* and *Class II*) that share no similarities at sequence or structure level. Each of the classes is responsible for 10 of the 20 proteinogenic amino acids and can be further grouped into subclasses[15]. One exception to this class separation rule is lysyl-tRNA synthetase (LysRS), where euryarchaeal genomes were shown to contain a Class I form[25] instead of the standard Class II form. Most eukaryotic genomes contain the complete set of 20 aaRSs. However, some species lack certain aaRS-encoding genes and compensate for this by post-modifications[7,26–28] or alternative pathways[29–31]. A scenario where Class I and Class II originated simultaneously from opposite strands of the same gene[32,33] is among the most popular explanations for the origin of aaRSs. This so-called Rodin-Ohno hypothesis (named after Sergei N. Rodin and Susumu Ohno[32]) is supported by experimental deconstructions of both aaRS classes[34–36]. At the dawn of life the concurrent duality could have allowed to implement an initial binary choice, which is the minimal requirement to establish any code[9].

Origin of genetic coding. Several theories exist (for a summary see reference[2]) that aim to explain the origin of the genetic code and its self-translating machinery. The theory of co-evolution[37], states that the appear-

[1]Biotechnology Center (BIOTEC), TU Dresden, 01307 Dresden, Germany. [2]PharmAI GmbH, Tatzberg 47, 01307 Dresden, Germany. [3]University of Applied Sciences Mittweida, 09648 Mittweida, Germany. [4]These authors contributed equally: Florian Kaiser and Sarah Krautwurst. ✉email: florian.kaiser@tu-dresden.de

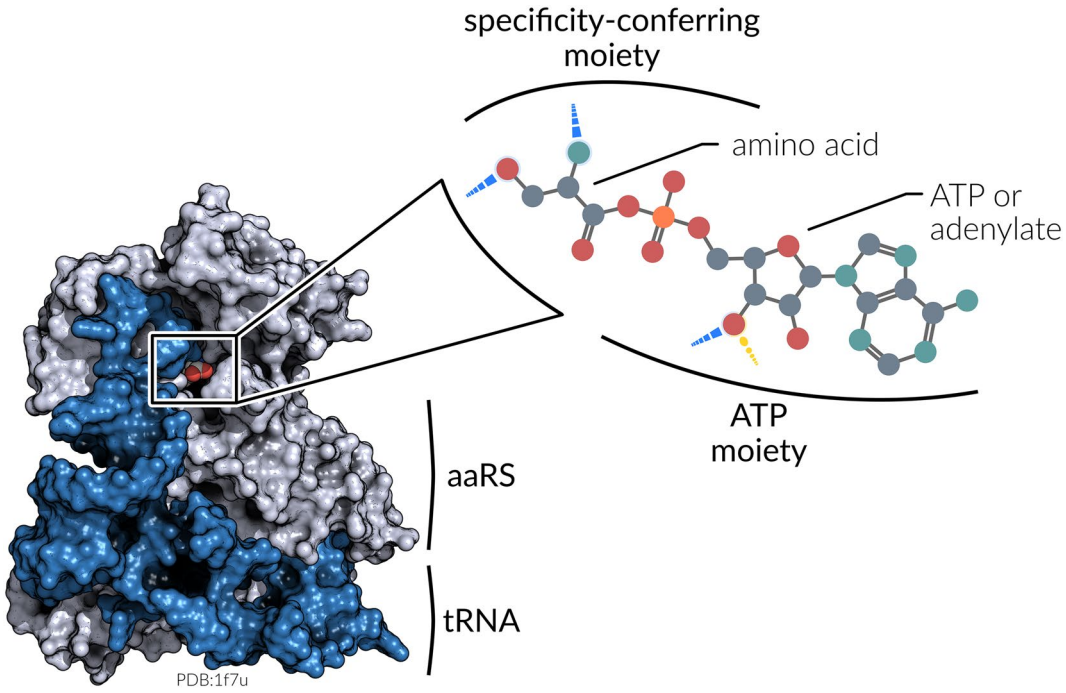

Figure 1. The aaRS·tRNA complex (PDB:1f7u) and the architecture of its active site. The enzyme catalyzes the covalent attachment of an amino acid to the 3' end of a tRNA molecule. The binding site itself can be divided into two moieties. While the ATP moiety is responsible for the fixation of ATP, which is consistent within each aaRS class[43], the specificity-conferring moiety differs between each aaRS and forms highly specific non-covalent interactions with the amino acid ligand. Depending on the reaction state (pre- or post-activation), the ATP moiety contains ATP or an adenylate group.

ance of amino acids via new biochemical synthesis pathways was strongly coupled with their integration into the genetic code. Thus, the co-evolution theory takes the age of amino acids—defined by the complexity of their biochemical pathways[38]—into account. Another theory, ambiguity reduction of physicochemical properties[39,40], considers the major selective pressure for genetic code emergence to be the minimization of deleterious effects of mutations. According to this theory, codons that differ only in a single base, encode for amino acids with comparable physicochemical properties to mitigate the effect of translation errors. The role of stereochemical forces in genetic code formation, is supported by numerous studies[2]. Here, the role of primordial amino acid binding structures is seen as a major determinant of which amino acids were embedded in the genetic code. This theory is of special interest in the light of our study, as amino acid binding sites of aaRSs might have composed such an ancient recognition structure.

Biochemical function. In order to fulfill their biological function aaRSs are required to catalyze two distinct reaction steps. Prior to its covalent attachment to the 3' end of the tRNA molecule, the designated amino acid is activated with adenosine triphosphate (ATP) and an aminoacyl-adenylate intermediate is formed[41,42]. In general, the binding sites of aaRSs can be divided into two moieties: the part where ATP is bound as well as the part where specific interactions with the amino acid ligand are established (Fig. 1). Is is assumed that the amino acid activation with ATP constituted the principal kinetic barrier for the creation of peptides in the prebiotic context[36]. Due to the fundamental importance of this first reaction step, highly conserved sequence[4] and structural motifs[43] exist, which are likely to be vital for the aminoacylation reaction. These structural motifs were detected in our previous study[43], reinforcing the Class I and Class II separation of aaRSs. In structures of Class I aaRSs, ATP is bound via backbone hydrogen bonds. This motif, termed Backbone Brackets, undergoes structural rearrangement upon ATP binding and is only revealed at functional interaction level. Class II aaRSs ensure ATP binding with a pair of arginine residues, forming salt bridges towards the ATP molecule. These Arginine Tweezers are observable at sequence as well as structure level. While the activation of amino acids with ATP is the basic requirement of all aaRSs and is consistent within each aaRS class[43], the recognition mechanism of individual amino acids differs substantially between each aaRS. These differences are among the key drivers to maintain a low error rate during the translational process.

Non-covalent binding site interactions. Non-covalent protein-ligand interactions play an important role for the specific binding of any ligand. These interactions are generally reversible and correspond to an energy of binding between $-80\ kJ \cdot mol^{-1}$ and $-10\ kJ \cdot mol^{-1}$, which is less compared to covalent interactions[44]. Several types of non-covalent interactions exist that can add energetic contribution to the binding of a protein ligand-complex. Each type is constrained regarding interaction partners and geometry. Generally, directed

	Interaction type					
	Hydrogen bond	Hydrophobic	Salt bridge	π-stacking	Metal complex	Total
Class I	468 (37.96%)	**550** (44.60%)	153 (12.41%)	59 (4.79%)	3 (0.24%)	1233
Class II	**856** (59.23%)	193 (13.36%)	202 (13.98%)	144 (9.97%)	50 (3.46%)	1445

Table 1. Overview of observed interactions between aaRSs and their amino acid ligands. The most prevalent interactions are hydrophobic interactions for Class I aaRSs and hydrogen bonds for Class II aaRSs (typeset in bold). Relative frequencies in respect of all interactions of the aaRS class are given in parentheses.

hydrogen bonds are considered to be the strongest non-covalent interaction, followed by π-cation and π-stacking interactions, electrostatic (or salt bridge) interactions, and hydrophobic interactions[45]. Based on experimentally determined three-dimensional structures of protein-ligand complexes, non-covalent interactions can be studied computationally. However, this requires a detailed annotation of non-covalent interaction patterns. In this study, we use the rule-based Protein-Ligand Interaction Profiler (PLIP)[46] to characterize the amino acid binding in aaRSs.

Motivation. In our last study we identified two unique ATP binding motifs in Class I and Class II aaRSs[43], which are by now the minimal description of the two classes. Hence, a detailed study of the amino acid binding site is the logical next step to extend the picture of ligand binding in aaRSs. Protein structures of aaRSs from all kingdoms of life, co-crystallized with their amino acid ligands, are publicly available in the Protein Data Bank (PDB)[47]. Furthermore, there are tools such as PLIP[46] to characterize and map the interactions of proteins and their ligands. These rich data allow for the investigation of specific characteristics of amino acid recognition in individual aaRS. The overall aim is to contribute to the understanding of how aaRSs realize the correct mapping of the genetic code and to provide a compendium of binding site interactions relevant to maintain amino acid specificity. The results shed light on how evolution implemented a specific recognition via the amino acid composition of the binding size, non-covalent interaction patterns, pre- or post-transfer correction mechanisms, and steric effects such as the volume of the binding cavity. Moreover, the overall recognition strategies for Class I and Class II aaRSs differ, suggesting that the existence of the classes allowed the enzymes to cover a broader ligand diversity and thus the gradual incorporation of new amino acids into the genetic code.

Results

Dataset. Based on all available structures in the PDB, 424 (189 Class I, 235 Class II) three-dimensional structures of aaRSs co-crystallized with their corresponding amino acid ligands were analyzed. The selected data covers aaRSs of 56 different species in total, 180 from eukaryotes, 213 from bacteria, and 31 from archaea (SI Appendix Fig. S1). In total, 70 human structures are part of the dataset. Each protein chain that contains a protein-ligand complex of a catalytic aaRS domain was considered. Data was available for each of the 20 aaRSs, plus the non-standard aaRSs pyrrolysyl-tRNA synthetase (PylRS) and phosphoseryl-tRNA synthetase (SepRS). Unfortunately, Class I LysRS could not be considered for analysis. The single structure of this enzyme from *Pyrococcus horikoshii* (PDB-ID: 1irx), which is part of the dataset, does not contain any co-crystallized amino acid ligand. The numbers of protein-ligand complexes available for each aaRS are given in SI Appendix Fig. S2. For twelve aaRSs, protein-ligand complexes were available in both pre-activation and post-activation reaction states, i.e. co-crystallized with either amino acid or aminoacyl ligand (SI Appendix Fig. S3). Out of all analyzed structures, 240 are in pre-activation and 184 in post-activation state. Out of the post-activation complexes, 72 are adenosine monophosphate (AMP) esters and 112 are non-hydrolysable analogs, mainly sulfamoyl derivatives.

Interaction features. The frequencies of observed non-covalent binding site interactions in respect of the aaRS class and the type of interaction are shown in Table 1. In general, hydrophobic interactions are the most prevalent interactions for Class I aaRSs with a frequency of 44.60% with respect to the total number of interactions, while hydrogen bonds are most frequently observed in Class II aaRSs with 59.23% frequency. Five (hydrogen bonds, hydrophobic interactions, salt bridges, π-stacking, and metal complexes) interaction types were observed in aaRSs. No π-cation interactions were observed to be involved in amino acid binding. Water bridges were excluded from the interaction analysis. Some aaRS structures deposited in the PDB are resolved including water, but other structures do not contain water molecules. In these cases, no water bridges can be detected using PLIP, despite them existing *in vivo*, which would lead to an experimental bias. Nonetheless, water molecules are known to mediate important interactions for ligand recognition[48] and their role should not be underestimated.

Amino acid recognition. The annotation of non-covalent protein-ligand interactions allowed to characterize interaction preferences of each aaRS at the level of individual atoms of their amino acid ligands. This analysis highlights the preferred modes of binding for each of the 22 amino acid ligands. Figure 2 shows the occurring interactions for each aaRS based on the analysis with PLIP. Each interaction is annotated with its occupancy, i.e. the relative frequency of occurrence in respect of the total number of structures for this aaRS. Binding site features are neglected at this point and all interactions are shown with respect to the amino acid ligand.

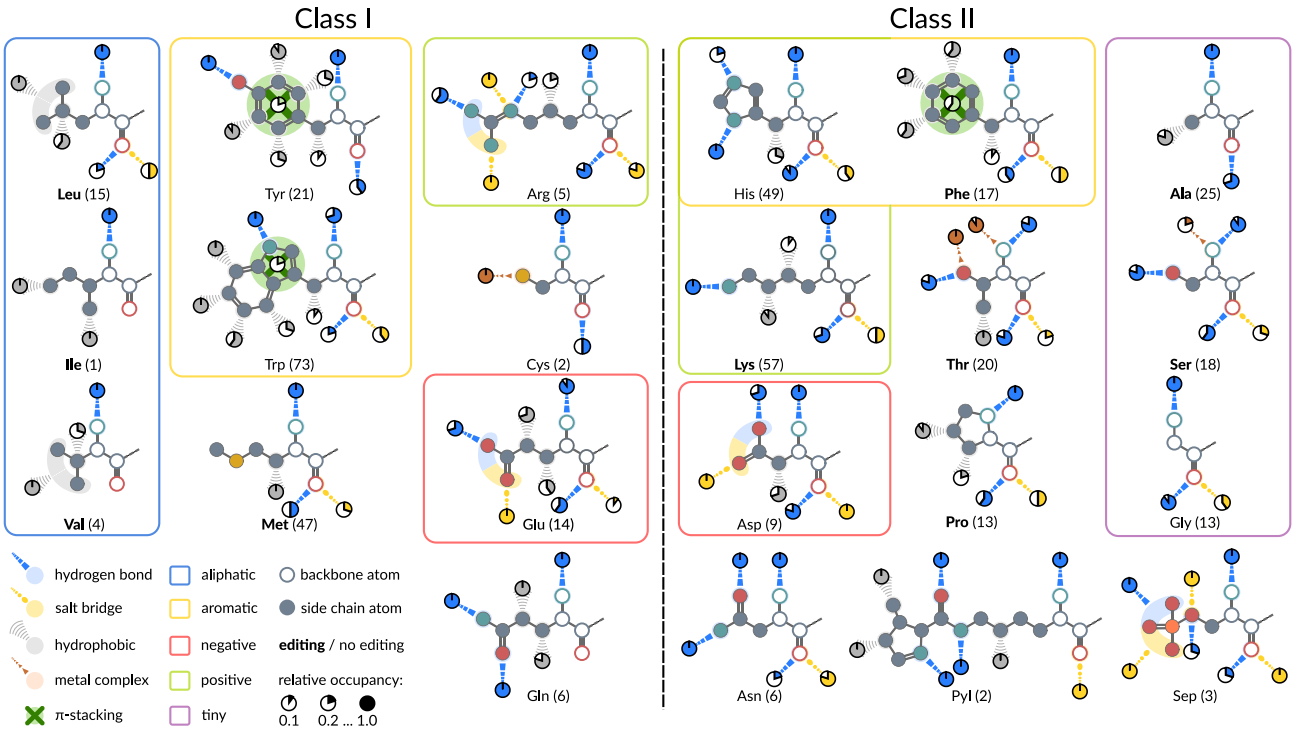

Figure 2. The recognition of individual amino acids by aaRSs mapped to their ligands. The ligands are grouped by physicochemical properties[49] and aaRS class. Different types of non-covalent protein-ligand interactions were determined with PLIP[46] and assigned to individual atoms of the ligand using subgraph isomorphism detection[50]. Backbone atoms of the ligand are depicted as circles without filled interior. The relative occupancy of each interaction in respect of the total number of investigated structures (number in parentheses for each aaRS) is given by pie charts. Interactions with an occupancy below 0.1 are neglected. Interactions for which a unique mapping to an individual atom is not possible due to ambiguous isomorphism, e.g. for the side chain of valine, were assigned to multiple atoms. π-stacking interactions are shown in dark green and refer to all atoms of the aromatic ring structures in TyrRSs, TrpRSs and PheRSs. Some aaRSs prevent the mischarging of their tRNAs via error correction mechanisms ("editing")[51]. The aaRSs conducting error correction are typeset in bold.

Class I. In general, Class I aaRSs interact mainly via hydrogen bonds and hydrophobic interactions with the ligand. The backbone atoms of all Class I ligands feature hydrogen bonding with the primary amine group. The occupancy of this interaction is high throughout all Class I aaRSs, indicating a pivotal role of this interaction for ligand fixation. Additionally, the oxygen atom of the ligand's carboxyl group is involved in hydrogen bonding except for glutaminyl-tRNA synthetase (GlnRS), isoleucyl-tRNA synthetase (IleRS), and valyl-tRNA synthetase (ValRS). The same atom forms additional salt bridges in leucyl-tRNA synthetase (LeuRS), arginyl-tRNA synthetase (ArgRS), methionyl-tRNA synthetase (MetRS), and glutamyl-tRNA synthetase (GluRS). The side chains of the aliphatic amino acids leucine, isoleucine, and valine are exclusively bound via hydrophobic interactions. ArgRS and GluRS form salt bridges between binding site residues and the charged carboxyl and guanidine groups of the ligand, respectively. Glutamine is bound by GlnRS via conserved hydrogen bonds to the amide group and hydrophobic interactions with beta and delta carbon atoms. The two aromatic amino acids tyrosine and tryptophan are recognized by π-stacking interactions and extensive hydrophobic contact networks. Tryptophan is bound preferably from one side of its indole group at positions one, six, and seven. The sulfur atom of the cysteinyl-tRNA synthetase (CysRS) ligand forms a metal complex with a zinc ion in both structures. MetRSs bind their ligand with a highly conserved hydrophobic interaction with the beta carbon atom.

Class II. Class II aaRSs consistently interact with the backbone atoms of the ligand via hydrogen bonds and salt bridges. The primary amine group forms hydrogen bonds with high occupancy and is involved in metal complex formation in threonyl-tRNA synthetases (ThrRSs) and seryl-tRNA synthetases (SerRSs). The carboxyl oxygen atoms of the ligands are bound by a combination of hydrogen bonding and electrostatic salt bridge interactions. The overall backbone interaction pattern is highly conserved within Class II aaRSs. Closer investigation revealed that a previously described structural motif of two arginine residues[43], responsible for ATP fixation, seems to be involved in stabilizing the amino acid carboxyl group with its N-terminal arginine residue. The charged amino acid ligands in histidyl-tRNA synthetase (HisRS) and LysRS form highly conserved hydrogen bonds with the binding site residues. Other specificity-conferring interactions include π-stacking interactions and hydrophobic contacts observed for phenylalanine-tRNA synthetase (PheRS), metal complex formation for ThrRS and SerRS with zinc, and salt bridges as well as hydrogen bonds for aspartyl-tRNA synthetase (AspRS). The amino acids alanine and proline are bound by alanyl-tRNA synthetases (AlaRSs) and prolyl-tRNA syn-

thetases (ProRSs) via hydrophobic interactions. No specificity-conferring interactions can be described for the smallest amino acid glycine due to absence of a side chain. Hence, glycyl-tRNA synthetase (GlyRS) can only form interactions with the backbone atoms of the ligand. Furthermore, asparaginyl-tRNA synthetases (AsnRSs) mediate highly conserved hydrogen bonds with the amide group of their asparagine ligand. The non-standard amino acid pyrrolysine is bound by PylRS via several hydrogen bonds and hydrophobic interactions with the pyrroline group. SepRSs employ mainly salt bridge interactions to fixate the phosphate group of the phosphoserine ligand.

Conserved Interaction Patterns. Class I aaRSs show a strong conservation of hydrogen bonds with the primary amine group of the amino acid ligand with 83.16% of all structures forming this interaction. Interactions with the carboxyl group are less conserved with a frequency of 32.65% for hydrogen bonds and 28.57% for salt bridges, respectively. In this context, the salt bridges with the carboxyl group are a form of extra strong hydrogen bonding[52]. Interaction patterns with the backbone atoms of the amino acid ligand are strikingly consistent within Class II aaRSs. This class forms hydrogen bonds with the primary amine group in 92.15% of all structures. Additionally, hydrogen bonds with the oxygen atom of the carboxyl group occur in 65.70% of all structures and salt bridges with the same atom are formed in 39.26% of all Class II protein-ligand complexes.

Similar recognition requires editing mechanisms. Various aaRSs are known to conduct pre- or post-transfer editing (see the work of Perona and Gruic-Sovulj[51] for a detailed discussion of editing mechanisms) in order to ensure proper mapping of amino acids to their cognate tRNAs. The similarity of interaction preferences depicted in Fig. 2 suggests that groups of very similar amino acids require editing mechanisms for their correct handling. Especially the three aliphatic amino acids isoleucine, leucine, and valine are bound via unspecific and weak hydrophobic interactions, substantiating the necessity of editing mechanisms observed for their aaRSs[53] and that substrate hydrophobicity cannot entirely account for specificity[54]. Distinction between those three similar amino acids is proposed to happen via the "double sieve"[52] mechanism. Exemplarily for IleRS, amino acids larger than isoleucine are excluded with the "first sieve" at the aminoacylation site, whereas smaller amino acids (like valine and leucine) are sorted out by the editing domain, functioning as a finer "sieve". Specificity can therefore be accomplished by steric selection based on side chain length and shape at the editing site[53]. A similar trend can be observed, e.g., for AlaRS[55] in order to distinguish alanine from serine or glycine.

Binding site geometry and cavity volume. We investigated binding site geometry and cavity volume in order to quantify their potential contribution to amino acid recognition. Known editing mechanisms in aaRSs are focused on the prevention or correction of tRNA mischarging within one aaRS class (intra-class), e.g. the amino acids isoleucine, leucine, and valine belong to Class I. However, GluRSs and AspRSs have a highly similar interaction pattern of hydrogen bonds and salt bridges with the carboxyl group and weak hydrophobic interactions. Both aaRSs do not use editing and are handled by different aaRS classes. In this case, the geometry and size of the binding site can act as an additional layer of selectivity; a mechanism also exploited by ValRS[53,56]. To quantify the contribution of binding site geometry, seven structures of GluRS and six structures of AspRS were superimposed with respect to their common adenine substructure using the Fit3D[57] software. As this superimposition can solely be computed for protein-ligand complexes which resemble the post-reaction state, only a subset of the structures was used. The results show that the ligands of GluRSs and AspRSs are oriented towards different sides of a plane defined by their common adenine substructure (Fig. 3A). There is a significant difference (Mann-Whitney U $p<0.01$) in ligand orientation, described by the torsion angle between phosphate and the amino acid substructure of the ligand (Fig. 3B). Class I GluRSs feature a torsion angle of $54.64 \pm 7.12°$, whereas the torsion angle of Class II AspRSs is $-65.02 \pm 7.40°$. Furthermore, the volume of the specificity-conferring moiety of the binding site (see Fig. 1) was estimated with the POVME[58] algorithm. It differs significantly (Mann-Whitney U $p<0.01$) between GluRS (147.00 ± 22.31 Å^3) and AspRS (73.34 ± 17.12 Å^3). This trend can be observed for all Class I and Class II structures, respectively. An analysis of all representative structures for Class I and Class II aaRSs shows that Class I binding sites are significantly (Mann-Whitney U $p<0.01$) larger on average (Fig. 3C). While Class I binding cavities have a mean volume of 143.40 ± 39.62 Å^3, Class II binding sites are on average 90.36 ± 32.09 Å^3 in volume.

Interaction patterns of individual aaRSs. In addition to the investigation of interaction preferences from the ligand point-of-view, the binding sites of each aaRS were analyzed regarding the residues that form interactions with the amino acid ligand. Because each aaRS is backed by multiple proteins from diverse organisms with considerably divergent sequences, we devised a computational abstraction to allow the reader to infer amino acids of individual proteins via a structure-driven multiple sequence alignments (MSAs) (see "Methods" section). Original sequence numbers for each position can be inferred with the mapping tables published along with this manuscript (see Data Availability). Each row in the table corresponds to the artificial sequence position, whereas each column gives the original position for each structure in our dataset as defined by the PDB. Figure 4A shows a sequence logo[59] representation of binding site interactions for AlaRS. Each colored position in the sequence logo represents interactions occurring at this position. Highly conserved interactions can be observed at renumbered position 135. The corresponding hydrogen bond and salt bridge interactions are formed with the backbone atoms of the ligand. On the protein side, this interaction is mediated by a conserved arginine residue that corresponds to the N-terminal residue of the previously described Arginine Tweezers motif[43]. Another prominent interaction is formed by valine at renumbered position 293. This residue interacts with the beta carbon atom of the alanine ligand via hydrophobic interactions. In some structures, this hydrophobic interaction is complemented by an alanine residue at renumbered position 325. Aspartic acid at renumbered position

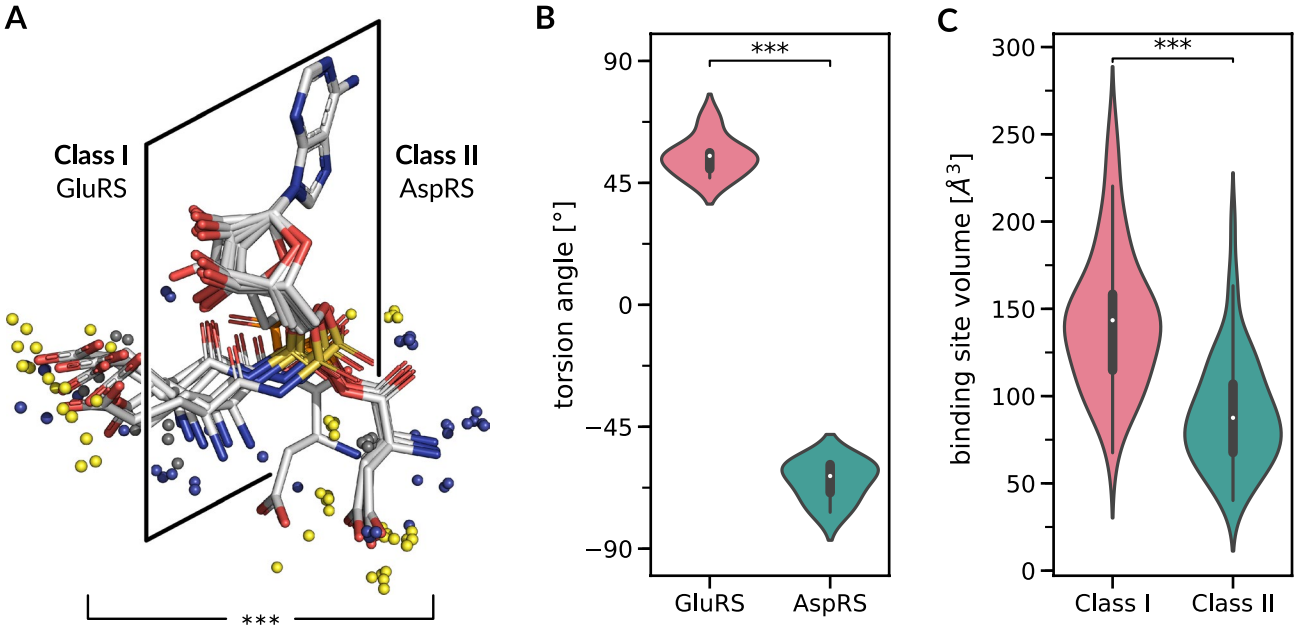

Figure 3. Binding geometry and binding cavity volume analysis. (**A**) Binding geometry of GluRSs and AspRSs. Aminoacyl ligands of Class I GluRSs and Class II AspRSs in post-activation state aligned with Fit3D[57] with respect to their adenine substructure. The midpoints of non-covalent interactions[46] with binding site residues are depicted as small spheres. Blue is hydrogen bond, yellow is salt bridge, and gray is hydrophobic interaction. (**B**) Distribution of torsion angles between the phosphate and amino acid substructure of the ligand. The orientation of the ligand in the binding site differs significantly (Mann–Whitney U $p < 0.01$) between GluRSs and AspRSs. (**C**) The volume of the specificity-conferring moiety of the binding site, estimated with the POVME algorithm[58], differs significantly between Class I and Class II aaRSs (Mann-Whitney U $p < 0.01$).

323 is highly conserved in AlaRSs and seems to be involved in amino acid fixation via hydrogen bonding of the primary amine group. Overall, the specificity-conferring interactions with the small side chain of alanine are hydrophobic contacts. An example for amino acid recognition in AlaRSs is given in Fig. 4B. The structure of bacterial *Escherichia coli* AlaRS forms the whole array of observed interactions. Sequence logos of the remaining aaRSs are given in SI Appendix Figs. S4–S24. Based on the interactions between binding site residues and the ligand, a qualitative summary of specificity-conferring mechanisms and key residues was composed (Table 2). Moreover, the ligand size and count of observed interactions was checked for dependence. There is a weak but significant positive correlation between the average number of interacting binding site residues for each aaRS and the number of all non-hydrogen atoms of the amino acid ligand (Pearson r=0.32, p<0.01). This indicates that the number of formed interactions generally increases with ligand size. However, smaller amino acids do not necessarily have a less complex recognition pattern. ThrRSs, for example, bind their amino acid ligand with on average more than a dozen binding site residues, while ValRSs employ on average five binding site residues. The hydroxyl group of threonine allows for an extended range of non-covalent interactions to be formed with binding site residues compared to valine, where only hydrophobic contacts can be established. Distributions of interacting binding site residues for each aaRS are given in SI Appendix Fig. S25.

Quantitative comparison of ligand recognition. To allow for a quantitative analysis and comparison of ligand recognition between several aaRSs, interaction and binding site features were represented as binary vectors, so-called interaction fingerprints (see "Methods" section). Based on these fingerprints, the Jaccard distance was computed for each pair of structures to represent the dissimilarity in ligand recognition. Subsequently, the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm[61] was used for dimensionality reduction and embedding of the high-dimensional fingerprints into two dimensions for visualization. This embedding is considered to be the *recognition space* of aaRSs. The two-dimensional visualization of this recognition space (Fig. 5) can be seen as a map describing the similarity in ligand recognition across all aaRSs. Thereby, each data point corresponds to a single amino acid binding site that was characterized by interaction and binding site features. In general, a similar recognition mechanism between two aaRSs can be assumed if they are located close to each other in this map. The more distant two aaRSs are from each other, the less similar their amino acid recognition. However, it has to be noted that the applied dimension reduction does not perfectly conserve distances. Figure 5A shows the embedding results for all aaRSs in the dataset colored according to the aaRS classes. A Principal Component Analysis (PCA) of the same data is given in SI Appendix Fig. S26. For each aaRS the average position of all data points in the embedding space was calculated and is shown as one-letter code label. Figure 5B shows the same data colored according to the physicochemical properties of the amino acid ligand, i.e. positive (lysine, arginine, and histidine), aromatic (phenylalanine, tyrosine, and tryptophan), negative (aspartic acid and glutamic acid), polar (asparagine, cysteine, glutamine, proline, serine, and threonine), and unpolar (glycine, alanine, isoleucine, leucine, methionine, and valine).

	Subclass	aaRS	Recognition mechanism	Involved residues (*)
Class I	IA	MetRS	HP with C_β	Trp-319, Ile-365
		IleRS	HP network with aliphatic side chain	Glu-567, Trp-575
		LeuRS	HP network with aliphatic side chain	Met-50, Phe-51, Phe/Leu/Trp-562, Tyr-568, His-650
		ValRS	HP with side chain methyl groups	Pro-41, Trp-456, Trp-495
	IB	CysRS	Cys-Cys-Cys-His tetrahedral MC with Zn	Cys-31, Cys-215, His-240
		GlnRS	HB[a] with amide group, HP[b] with C_β, C_γ	Arg-234[a], Tyr-417[a], Pro-236[b], Phe-439[b]
		GluRS	Arg-mediated SB[a] coordination with carboxylate group, HP[b] with C_γ	Arg-15[a], Arg-49[a], Arg-236[a], Tyr-218[b]
	IC	TrpRS	HP[a] network with indole, HB[b] to indole amine	Leu/Tyr/Phe-94[a], Val/Ile-289[a], His/Glu-135[b]
		TyrRS	HB[a] and HP[b] with phenole, (PS[c] with phenole)	Tyr-74[a,b], Asp-271[a], Leu-108[b], Gln-268[b], (His-113[c])
	ID	ArgRS	double SB[a] and HB[b] with guanidine group, HP[c] with C_γ	Asp/Glu-203[a], Asp-414[a], Tyr-410[b,c]
Class II	IIA	AlaRS	HP with C_β	Val-293
		GlyRS	n/a	n/a
		HisRS	HB[a] with imidazole group, (HP[b] with C_β)	Thr-98[a], Glu/Asp-148[a], Tyr-459[a], (Ala-507[b])
		ProRS	HB[a] and HP[b] with pyrrolidine ring	Thr-127[a], Asp/Glu-178[b], Trp/Met/Phe-176[b]
		SerRS	tetrahedral MC[a] with Zn, HB[b] with hydroxyl group	Glu-413[a], Lys/Arg-411[b], Ser-500[b]
		ThrRS	Cys-His-His-Thr tetrahedral MC[a] with Zn, HB[b] with hydroxyl group, HP[c] with methyl group	Cys-346[a], His-397[a], His-537[a], Arg-538[b], Thr-507[c]
	IIB	AspRS	SB coordination with carboxylate group	Lys-267, Arg-661, (His-261), (His-262)
		AsnRS	HB with amide group	Glu-233, Arg-377
		LysRS	HB[a] with side chain amino group, HP[b] with C_δ	Tyr-283[a], Glu-509[a], Tyr/Phe-507[b]
	IIC	PheRS	sandwich PS[a] and HP[b] with phenyl group	Phe-520[a], Phe/Tyr-522[a], Thr/Val-523[b], Ala-578[b]
	n/a	PylRS	HB[a] and HP[b] with pyrroline group, HB[c] with hydroxyl group and side chain amine group, HP[d] with C_δ	Tyr-208[a], Leu-126[b], Tyr-127[b], Asn-167[c], Gly-243[c], Ala/Val-225[d]
	n/a	SepRS	(backbone) HB[a] network, SB[b] with phosphate group	Met-25[a]8, Thr-259[a], His-257[b], Ser-302[b], Ser-304[b], Asn-396[b]

Table 2. Overview of specificity-conferring recognition mechanisms for all aaRSs grouped by aaRS class and subclass[15]. Only interactions with side chain atoms of the amino acid ligand were included in this summary. HB is hydrogen bond, SB is salt bridge, HP is hydrophobic, MC is metal complex, and PS is π-stacking interaction. Correspondences between interactions and residues are indicated by superscript letters. Entries in parentheses were only observed in certain structures and are no general pattern. (*) Residue numbers are given according to the respective MSA (see "Methods" section). Original residue numbers can be inferred with tables published along with this manuscript (see Data Availability).

Class I. In terms of amino acid binding both aaRS classes seem to employ different overall mechanism; they separate almost perfectly in the embedding space. Especially aromatic amino acid recognition in Class I tryptophanyl-tRNA synthetases (TrpRSs) and tyrosyl-tRNA synthetases (TyrRSs) is distinct from Class II aaRSs and forms two outgroups in the embedding space. Remarkably, two different recognition mechanisms exist for TrpRSs, indicated by two clusters approximately at positions (−2.0,6.0) and (1.0,8.5) of the embedding space, respectively. The cluster at position (−2.0,6.0) is formed by structures from bacteria and archaea, while the cluster at position (1.0,8.5) is formed by eukaryotes and archaea and is in proximity to TyrRSs. Closer investigation of two representatives from these clusters shows two distinct forms of amino acid recognition for TrpRSs. Human aaRSs employ a tyrosine residue in order to bind the amine group of the indole ring, while prokaryotes employ different residues (SI Appendix Fig. S27). The Class I aaRSs that are closest to Class II are GluRSs and CysRSs. A cluster of high density is formed by Class I IleRS, MetRS, and ValRS, which handle aliphatic amino acids. This indicates closely related recognition mechanisms and difficult discrimination between these amino acids.

Class II. For Class II aaRSs the recognition space is less structured. Nonetheless, clusters are formed that coincide with individual Class II aaRSs, e.g. a distinct recognition mechanism in AlaRSs. The aaRSs handling the small and polar amino acids threonine, serine, and proline are closely neighbored in the embedding space. Recognition of GlyRSs seems to be diverse; GlyRSs are not grouped in the embedding space. However, the recognition of glycine, which has no side chain, is limited by definition and thus the fingerprinting approach might fail to capture subtle recognition features. AspRSs and AsnRSs are located next to each other in the embedding space. Their recognition mechanisms seem to be very similar as the only difference between these two amino acids is the carboxylate and amide group, respectively.

Mechanisms that drive specificity. In order to quantify the influence of different aspects of binding site evolution on amino acid recognition by aaRSs, different interaction fingerprint designs were compared against each other. Each design includes varying levels of information and combinations thereof: the sequence composition of the enzyme's binding site (Seq), non-covalent interactions formed between side chains of the enzyme's binding site and the amino acid ligand (Int), whether pre- or post-transfer correction (i.e. "editing") is conducted

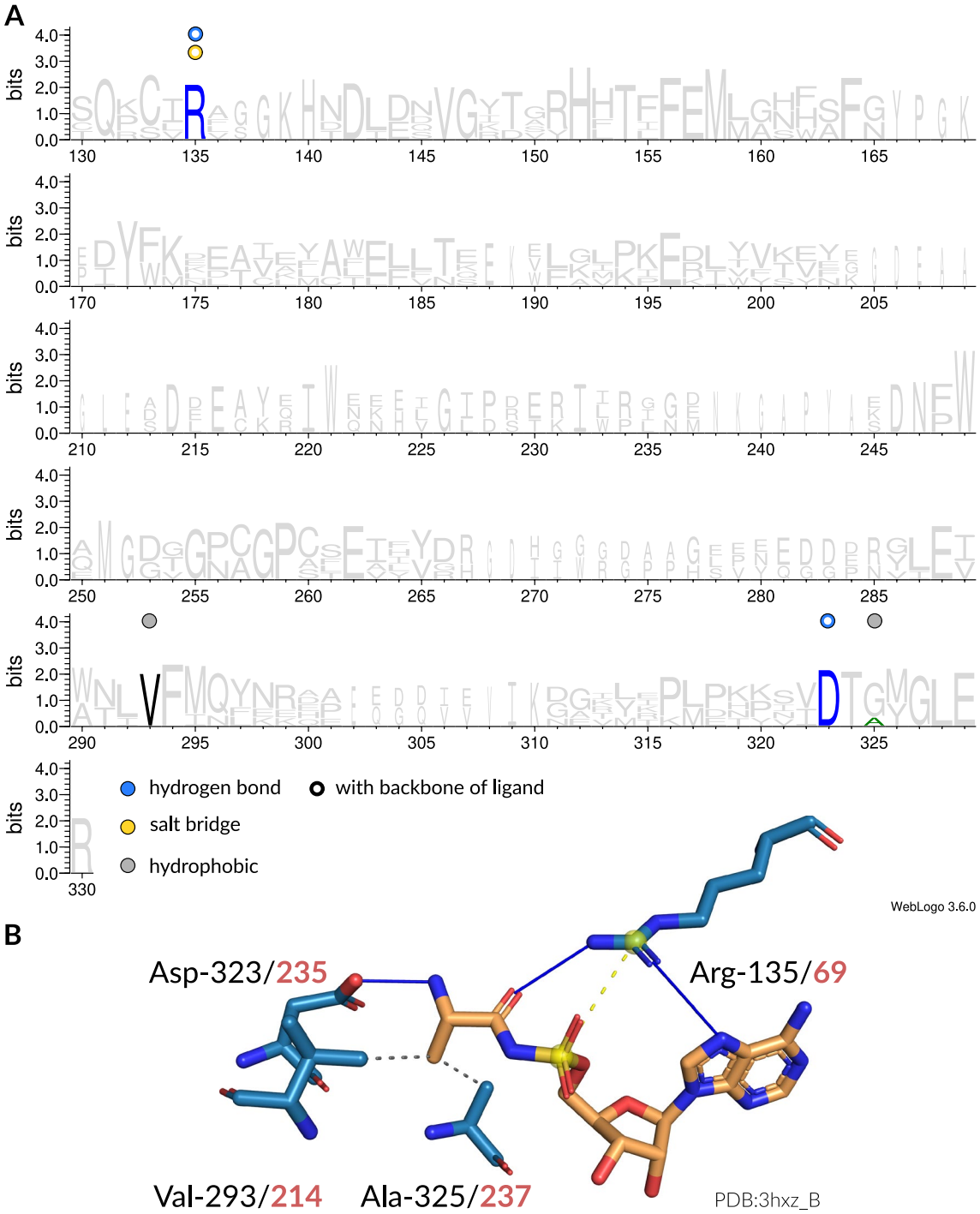

Figure 4. Interaction patterns of AlaRS. (**A**) Sequence logo[59] of representative sequences for AlaRSs. Non-covalent interactions with the amino acid ligand occurring at certain positions are indicated by colored circles. Filled circles are interactions with the side chain atoms, while hollow circles are interactions with any of the backbone atoms of the amino acid ligand. Blue is hydrogen bond, yellow is salt bridge, gray is hydrophobic interaction. (**B**) Depiction of interactions in the binding site (blue stick model) of an AlaRS from *Escherichia coli* (PDB:3hxz chain A) with its ligand (orange stick model). Here, hydrogen bonds (solid blue lines) and hydrophobic interactions (dashed gray lines) are established. The sequence positions of the interacting residues are given in accordance to the MSA (black) as well as the original structure (red). Figure created with PyMol[60]. Double bonds are indicated by parallel line segments, aromatic bonds by circular dashed lines.

(Ed), and the overall volume of the enzyme's binding cavity (Vol). To assess the segregation power of each fingerprint variant, the mean silhouette coefficient[62], a quantification for the error in clustering methods, over all

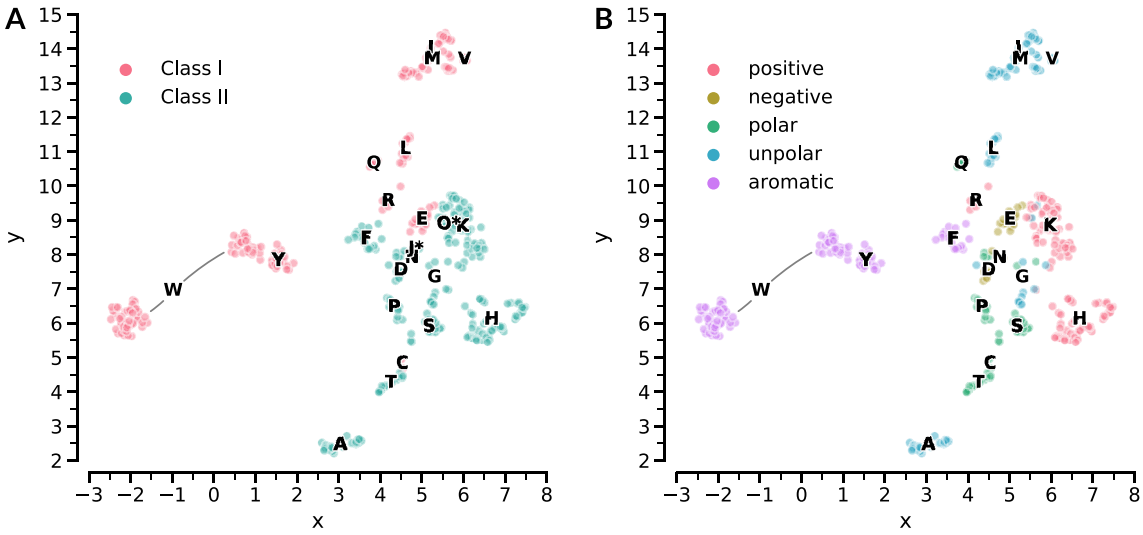

Figure 5. Recognition space analysis of all aaRSs. (**A**) Embedding[61] space of interaction fingerprints for all aaRS structures in the dataset. Scaling is in arbitrary units. The data points are colored according to the aaRS class. One letter code labels are given for each aaRS based on the averaged coordinates in the embedding space. An asterisk indicates the non-standard amino acids phosphoserine (J*) and pyrrolysine (O*). (**B**) Embedding space of interaction fingerprints for all aaRS structures in the dataset except phosphoserine and pyrrolysine. Scaling is in arbitrary units. One-letter codes of amino acid ligands are used to identify each aaRS. Every data point represents an individual protein-ligand complex. The color of the data points encodes the physicochemical properties[49] of the ligand.

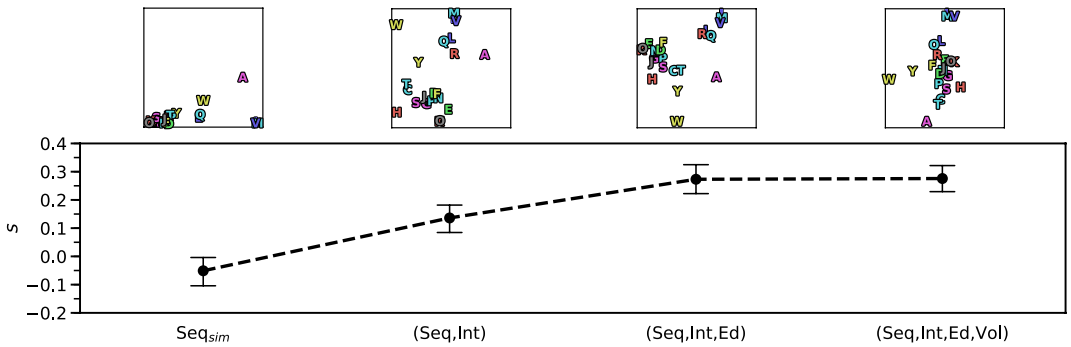

Figure 6. Comparison of different fingerprint designs that include the sequence composition of the enzyme's binding site (Seq), non-covalent interactions formed between side chains of the enzyme's binding site and the amino acid ligand (Int), pre- or post-transfer correction (i.e. "editing") mechanisms (Ed), and volume of the enzyme's binding cavity (Vol). Simple sequence-based fingerprints (Seq_{sim}) are a 20-dimensional representation of binding site composition. The line plot shows the silhouette coefficient[62] for each embedding. Points represent mean values, error bars are calculated based on all silhouette coefficients for each data point.

data points was calculated. This score allows to assess to which extent the recognition of one aaRS differs from other aaRSs and how similar it is within its own group. Perfect discrimination between all amino acids would give a value close to one, while a totally random assignment corresponds to a value of zero. Negative values indicate that the recognition of a different aaRS is rated to be more similar than the recognition of the same aaRS. Figure 6 shows the results of this comparison. When using fingerprints describing the sequence composition of the enzyme's binding site (Seq_{sim}), the mean silhouette coefficient over all samples is −0.0510, which indicates many overlapping data points and unspecific recognition. By including non-covalent interactions (Seq, Int) the value increases to 0.1361. If pre- or post-transfer correction mechanisms are considered (Seq, Int, Ed), the silhouette coefficient improves further to 0.2731. Adding information about the binding cavity volume (Seq, Int, Ed, Vol) slightly increases the quality of the embedding to 0.2757. The silhouette coefficients for error correction and volume-based fingerprints were calculated as baseline comparison. If only pre- or post-transfer correction mechanisms (Ed) are considered the mean silhouette coefficient amounts to −0.3027. For binding cavity volume (Vol) the mean silhouette coefficient is −0.4682.

Relation to physicochemical properties of the ligands. In order to investigate whether the fingerprinting approach is a simple encoding of the physicochemical properties of the amino acids, the results were related to experimen-

tally determined phase transfer free energies for the side chains of amino acids from water ($\Delta G_{w>c}$) and vapor ($\Delta G_{w>c}$) to cyclohexane[3,63]. These energies are descriptors for the size and polarity of amino acid side chains and underlie both, the rules of protein folding and the genetic code[64]. The Spearman's rank correlation between pairwise distances for each aaRS in the recognition space and physicochemical property space is weak with ρ=0.2564 and $p< 0.01$ (see SI Appendix Fig. S28). This indicates that the fingerprinting approach used in this study is a true high-dimensional representation of the complex binding mechanisms of amino acid recognition in aaRSs. This assumption is supported by a PCA (SI Appendix Fig. S26) of the fingerprint data, where the first two principal components account for only 9.24% and 8.44% of the covered variance, respectively.

Discussion

The correct recognition of individual amino acids is a key determinant for evolutionary fitness of aaRSs and considered to be one of the major determinants for the closure of the genetic code[10]. The results of this study emphasize the multitude of mechanisms that lead to the identification of the correct amino acid ligand in the binding sites of aaRSs. Based on available protein structure data, a thorough characterization of binding site features and interaction patterns allowed to pinpoint the most important drivers for the correct mapping of the genetic code. The main findings of this analysis can be summarized as follows: (i) Class I and Class II aaRSs employ different overall strategies for amino acid recognition. (ii) Interaction patterns and binding site composition are the most important drivers to mediate specificity. However, very similar amino acids require additional selectivity through steric effects or editing mechanisms. (iii) The analysis of interaction fingerprints suggests that error-free recognition is a delicate task demanding a complex interplay between binding site composition, interaction patterns, editing mechanisms, and steric effects. The results point towards a gradual diversification of amino acid recognition and, hence, a gradual extension of the genetic code.

Genetic code formation. We propose that the ancient aaRS binding sites might have formed the structural basis of the genetic code on the protein side. The exploration of stereochemical possibilities in the binding sites of aaRSs was likely to be vital for a stable and successful integration of amino acids into the genetic code. For the case of an RNA world, nucleotide sequences that bind specific amino acids were already suggested[65]. However, the limited conformational and catalytic repertoire of RNA molecules[66] is an underestimated factor. The amino acid binding sites of ancient aaRS precursors could have created a much broader recognition space, which was gradually gaining more complexity upon the addition of new amino acids to the translational system. Combined with the findings of our previous structural analyses[43], a modularity may be proposed for the substrate binding in aaRSs. This modularity allowed to re-use recognition patterns across different aaRSs, yet achieving a sufficient separation of the amino acid entities. ATP fixation differs substantially between the aaRS classes, implemented as the Backbone Brackets for Class I and the Arginine Tweezers for Class II. Binding of the amino acids still shows a general trend to employ different kinds of interaction for the two classes. Nonetheless, recognition of each amino acid is realized less class-specific, but more determined on slight differences between this highly specific ligand part. We find these considerations being compatible to the idea of sterochemically driven genetic code formation[39] and support the hypothesis that peptides and RNA coexisted and complemented each other from the very beginning[32,33,66,67].

Generation of orthologous aaRS·tRNA pairs. According to our analysis and the representation of binding site features in a high-dimensional vector space, the recognition space of aaRSs seems to be not yet fully explored, i.e. there are "blank areas" (Fig. 5). Whether these spots are tangible to the enzymes by binding site evolution can only be speculated. However, engineering aaRS·tRNA pairs in order to create an artificially extended genetic and subsequently to generate novel biopolymers is of high interest[68,69]. Beside the requirement of new codons and engineered ribosomes with broader substrate compatibility, the choice of an appropriate aaRSs·tRNA pair is of great importance. The major goal at the aaRS level is hereby to engineer specificity towards the new substrate but not to interfere with canonical aaRSs. According to our analysis there are several interesting candidates which are separated from other aaRSs in terms of their amino acid recognition as described by the high-dimensional fingerprints (see Fig. 5), namely bacterial TrpRSs, AlaRSs, HisRSs, GlnRSs, and LeuRSs. These aaRSs, especially bacterial TrpRSs (SI Appendix Fig. S27), form distinct clusters in the recognition space analysis and thus might be interesting targets for directed evolution of binding sites. TrpRSs were already successfully used to accomplish this goal[70]. We envision that an approach similar to the one presented in this study, might be helpful to estimate the success for generating novel aaRSs binding sites *in silico*. The characterization of key interactions for each aaRS (SI Appendix Fig. S4–S24) provides a valuable resource for predicting which mutations in the binding site are expected to alter specificity.

Coupling between tRNA and amino acid recognition. The specific detection of the cognate amino acid by aaRSs investigated here is only part of the whole reaction. aaRSs need to discriminate the tRNA molecule as well, to ensure correct coupling of amino acid and tRNA. This happens based on the anticodon and the acceptor stem of the tRNA, being recognized by the aaRS anticodon binding domain and the catalytic domain, respectively. Mischarged tRNAs due to failed cognate tRNA detection can hardly be corrected by the respective aaRS, but mistranslation may still be avoided by cross-editing of the cognate aaRS[71]. The evolutionary older[72] acceptor stem is highly necessary for specificity, whereas tRNAs are still correctly detected when the evolutionary younger anticodon information is masked[73]. Additionally, certain informative nucleotide motifs in the tRNA are relevant for the aaRS to couple the cognate tRNA and amino acid, differing only in as few as one position between the 20 types[15,73,74]. These specific discrimination are thought to have been incorporated and extended over time as

new amino acids joined the genetic code[75]. Combined with our results on amino acid recognition, it is therefore conceivable that aaRS specificity towards the cognate amino acid and tRNA developed simultaneously.

Class duality extends possibilities. The aaRS class duality allowed to broaden the amino acid recognition space significantly. In general, the recognition of amino acids with low side chain complexity seems to be complemented by allosteric interactions and cannot be exclusively implemented by configuring side chains. Although the volumes of Class I and Class II binding sites differ significantly, they are probably not the major determinants for amino acid selectivity. In general, Class I aaRSs handle most of the hydrophobic and larger amino acids[3] and thus the binding site volume of Class I aaRSs is expected to match the volumes of their larger ligands. Nonetheless, binding site volume and geometry may act as additional layers of selectivity. An example are the two negatively charged amino acids glutamic acid and aspartic acid, handled by a Class I and Class II aaRS, respectively. In this case, overall interactions are highly similar but binding geometry and binding site volume is significantly different. Both ligands are attacked from the opposite side[76] as highlighted by significantly different conformations (Fig. 3B). There is evidence that both amino acids were among the first to exist in the prebiotic context[37,77–81]. It is conceivable that the discrimination between glutamic and aspartic acid was based on tertiary contacts between secondary structures elements and size selectivity rather than on specific side chain interactions[82]. The recent identification of a protein folding motif[83] strengthens this assumption. This is further supported by the observation that ancient proteins, based on a limited set of amino acids, were still capable to exhibit secondary structures[81,84,85]. One canonly speculate whether a simultaneous emergence of two different aaRS classes and secondary structure formation allowed to incorporate these early – but highly similar – amino acids into the genetic code. An interesting hypothesis for the existence of two aaRS classes describes that ancestral Class I and II aaRS paired together with a tRNA molecule, contacting it from the opposite site to form a ternary complex. These pairs can be traced to belong to the now known aaRS subclasses[86]. This coincides with GluRS and AspRS, belonging to subclass Ib and IIb, respectively. The distribution of the aaRSs in Fig. 5 might support this idea of the development of two symmetrical classes as well, since the proposed pairs[86] are separated from each other. According to the biochemical pathway hypothesis[77], GluRS and AspRS might have been the first Class I and Class II representatives, with other aaRSs evolving from them[77,87]. However, the decreased usage of aspartic acid and the enrichment of glutamic acid in modern species, compared to the LUCA, points towards a different direction[88]. According to these usage frequencies, aspartic acid was incorporated into the genetic code prior to glutamic acid. This temporal order was equally concluded by the evaluation of various criteria to derive a consensus order of amino acid appearance[89].

Glutamine and asparagine followed glutamic acid and aspartic acid. Glutamine and asparagine are chemically closely related to glutamic and aspartic acid, respectively. It is likely that GlnRSs[6] and AsnRSs[7] mutually co-evolved from the evolutionary old GluRSs and AspRSs through recent gene duplication and were distributed via horizontal gene transfer (HGT)[15, 90]. A theory for this fast change in recognition was recently proposed by Carter *et al.*[91], according to which the HGT resulted in distinct clades in the phylogenetic tree of aaRS, distinguishing an aaRS like GlnRS strictly from the others. This resulted in a rapid change of specificity towards the amino acid ligand. In contrast, older aaRSs did not go through HGT, allowing their sequences to "wander" more during evolution, leading to clades which are not as clearly separated from each other and therefore inferior in discriminating specific amino acids[91]. Although the ligands of GluRS and GlnRS are rather similar, interaction patterns and binding site compositions differ between these two enzymes. These differences coincide with the analysis of the recognition space (Fig. 5), where GluRSs and GlnRSs are not neighbored in the embedding. Hence, they evolved to distinguish between these amino acids without editing mechanisms[92] or the exploitation of the negative charge of glutamic acid[93,94]. The discrimination of glutamine and glutamic acid by GlnRS cannot be attributed entirely to the composition of the binding site; changing the specificity from glutamine and glutamic acid could not be achieved by mutating only first order binding site residues[95]. This emphasizes the role of subtler interactions and allosteric effects within the catalytic domain as it was shown to be the case for TrpRS[96]. In contrast to the observed differences between GlnRS and GluRS, AspRS and AsnRS are directly neighbored in the embedding space and share a greater similarity in their recognition mechanism. However, as for GlnRS and GluRS, the discrimination between aspartic acid and asparagine is not entirely driven by specific interactions with binding site residues. Correct recognition depends on a water molecule that forms water-assisted hydrogen bonding between a binding site leucine in AsnRS and the amide group of the activated asparagine. Additionally, specificity of AsnRS to discriminate asparagine against aspartic acid is supported by two water molecules, forming the binding pocket to perfectly fit the asparagine side chain[48]. Although such indirect aspects may not be detected with our interaction-based investigation of static structures, they still contribute to specificity. The vicinity in the recognition space might be due to the limitation of interaction data, since water bridges were excluded from our analyses. Multiple aaRS structures were not determined with co-crystallized water molecules, making a detection with PLIP impossible. To avoid an overall bias due to this imbalance, water-mediated interactions were not considered during analysis. We conclude that for both Class I GlnRS/GluRS and Class II AsnRS/AspRS, the role of allosteric effects and other subtle interactions should not be underestimated.

Distinct recognition of arginine and lysine. Another interesting example are the two positively charged amino acids lysine and arginine. Interaction data suggests two unrelated ways to achieve ligand recognition in Class II LysRSs and Class I ArgRS, i.e. the two enzymes are well separated in the embedding space. The poor editing capabilities for LysRS regarding arginine[97] might have required a good separation of the two recognition mechanisms. Even if a relation of ArgRSs to aaRSs of hydrophobic amino acids was proposed[98], a separate subclass grouping for ArgRSs[15] seems to be reasonable and is in accordance with the observed data; the recognition

mechanism differs substantially from the hydrophobic amino acids. Furthermore, based on the consensus of all analyzed ArgRS structures, the characteristic Class I HIGH motif[4] seems to play an important role for stabilization of the arginine ligand in pre-activation state (see SI Appendix Fig. S4). For both histidine residues of the HIGH motif highly conserved salt bridges are observed that bind to the carboxyl group of the ligand.

Glycine recognition is not interaction-driven. Based on interaction data, the recognition of the smallest amino acid glycine seems to be rather unspecific; a large spread in the embedding space can be observed for individual protein-ligand complexes of GlyRS. This is to be expected as GlyRS is known to maintain its specificity not due to interactions with glycine – it has no side chain to interact with – but rather due to active site geometry that blocks larger amino acids[10,99].

Alanine recognition is crucial. Alanine is the second smallest amino acid with only a single heavy side chain atom. The idiosyncratic architecture of AlaRS is different from other Class II aaRSs[100]. Still, the confusion with glycine and serine[55], or non-proteinogenic amino acids[8], poses a challenge for correct recognition of alanine and a loss of specificity is associated with severe disease outcomes[101]. The recognition mechanism in AlaRSs seems to differ substantially from other Class II aaRSs (see Fig. 5), indicating evolutionary endeavor to develop a unique recognition mechanism.

Discrimination of hydrophobic amino acids requires editing. The hydrophobic amino acids isoleucine, leucine, valine, and methionine likely entered the genetic code at the same time[20,37,98]. The highly similar interaction patterns for IleRS, ValRS, and MetRS substantiate this assumption. Due to their difficult discrimination, editing functionality is key[5,56,92,102,103] for these aaRSs.

Tryptophan recognition suggests late addition to the genetic code. The emergence of TrpRSs and TyrRSs isconsidered to have happened at a later stage of evolution. The two aaRSs are likely to be of common origin[42,104] and constitute their own subclass, which is supported by sequence and structure studies[15,18,19,105,106]. PheRS supposedly evolved from the same precursor as TrpRS and TyrRS[21]. In general, TrpRSs and TyrRSs separate well from other aaRSs in the recognition space, which is likely due to the unique utilization of π-stacking interactions with binding site residues. Beside specific interactions in the binding site, allosteric effects and interdomain cooperativity[107,108] are drivers for TrpRS specificity. Furthermore, mutations in the dimerization interface of TrpRSs were shown to reduce specificity[96]. Remarkably, two distinct ways of recognition are apparent for TrpRSs in bacteria and eukaryotes. These differences support the previous described separation of eukaryotic TrpRSs and TyrRSs from their prokaryotic counterparts[109] and late addition of these amino acids to the genetic code[110]. However, structures from archaea do not follow this pattern and feature both recognition variants.

Methods

Data acquisition. The dataset from our last study[43] served as the basis for all analysis. As all structures in the dataset are annotated with ligand information, only entries containing ligands relevant for amino acid recognition were considered, i.e. they bind to the specificity-conferring moiety of the binding site (see Fig. 1). Every protein chain of the entry was considered that: (i) comprises a catalytic aaRS domain, (ii) contains a co-crystallized specificity-relevant ligand in the active site, and (iii) the ligand must contain an amino acid substructure. Filtering of the data resulted in 189 (235) structures for Class I (Class II) aaRSs that contain ligands with relevance for specificity. The number of structures in respect of the pre- or post-activation state of the catalyzed reaction is shown in SI Appendix Fig. S3. Furthermore, sequences of the dataset entries were clustered using single-linkage clustering with a sequence identity cutoff of 95% according to a global Needleman-Wunsch[111] alignment with BLOSUM62 substitution matrix computed with BioJava[112]. Representative chains for each cluster were selected, preferring wild type and high-quality structures. In total, 47 (54) protein chains were selected to be representatives for Class I (Class II) aaRSs. The dataset covers structures of all known aaRSs from species across all kingdoms of life (SI Appendix Fig. S1).

Mapping of sequence positions. Amino acid sequences were derived from the set of representative structures of the respective aaRS. To allow a unified mapping of sequence positions, an MSA was computed for each aaRS using the T-Coffee[113] Expresso pipeline. The quality of each MSA in the specificity-conferring region of the binding site was assessed regarding the correct mapping of the Backbone Brackets and Arginine Tweezers structural motifs[43], and the conservation of the respective sequence signature motifs[4,22]. All MSAs preserved the considered regions and passed the quality checks. The sequence positions for each aaRS were then unified according to the resulting MSA in order to investigate conserved interaction patterns. For this purpose the custom script "MSA PDB Renumber", available under open-source license (MIT) at github.com/vjhaupt, was used.

Annotation of non-covalent protein-ligand interactions. Non-covalent protein-ligand interactions were annotated for all entries in the dataset that contained a valid ligand using PLIP v1.3.3[46] with default parameters.

Determination of interactions relevant for specificity. Only interactions formed between the amino acid substructure of the ligand and binding site residues were considered for analysis. For this purpose subgraph isomorphism detection with the RI algorithm[50] was applied. The RI implementation of the SiNGA framework v0.5.0[114] was used. Each amino acid scaffold was represented by a graph created from the amino acid's SMILES

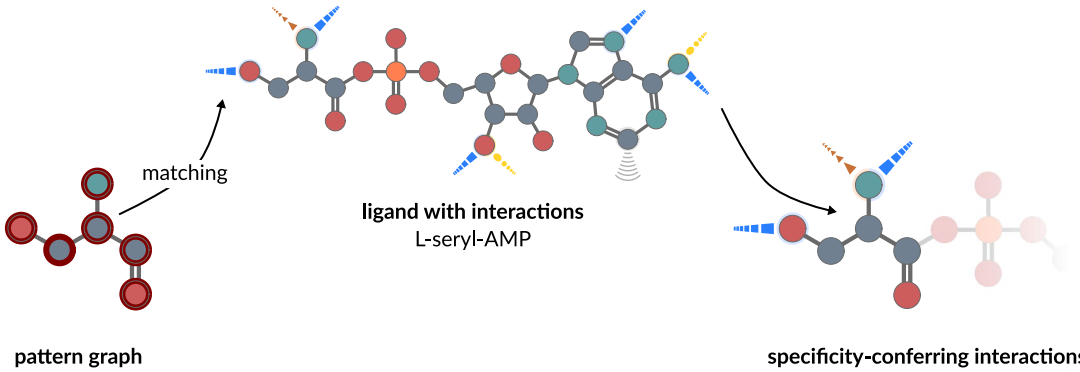

Figure 7. The identification of specificity-conferring interactions in SerRS. For each aaRS a pattern graph is used to map interactions. This patterns graph resembles the amino acid without its terminal hydroxyl group and is matched against the full ligand with annotated interactions using subgraph isomorphism detection[50]. The interactions formed between matched atoms and binding site residues are considered to be specificity-conferring interactions.

string taken from PubChem[115]. The full amino acid graph was modified using MolView v2.4 (available at molview.org) in order to remove the terminal hydroxyl group, which is cleaved during the enzymatic reaction and must thus be ignored for subgraph matching. For each dataset entry that contained a valid ligand, the corresponding amino acid graph was matched against the ligand in order to identify the atoms involved in the formation of specificity-conferring interactions. A depiction of the workflow to determine specificity-conferring interactions is given in Fig. 7.

Generation of interaction fingerprints. To allow for a quantitative comparison of recognition mechanisms, each protein-ligand complex was represented by a structure-invariant binary interaction fingerprint (see for example the paper of Salentin *et al.*[45] about the idea of interaction fingerprinting). Different fingerprint designs were chosen for comparison: a simple 20-dimensional fingerprint on binding site composition and a 500-dimensional fingerprint based on binding site composition and interaction information. The latter was further enriched with editing and binding site volume information.

Simple binding site based fingerprints. Binary and structure-invariant fingerprints that represent binding site compositions (used as baseline for the comparison of different fingerprint designs, Fig. 6) were constructed as follows. Each residue predicted to be in contact with any specificity-relevant atom of the ligand was considered for fingerprint generation. A 20-dimensional binary vector was used to represent the occurrence of individual residue types in the binding site. For each of the interacting residues the corresponding bit was set to active. Hence, multiple occurrences of the same residue type were not taken into account.

Binding site and interaction-based fingerprints. Single three-dimensional vectors of non-covalent interactions were encoded into a binary vector by considering the type of interaction, the interacting group in the ligand and the interacting amino acid residue. One such feature could be a hydrogen bond between an oxygen atom in the ligand and tyrosine in the protein. Each of these features is hashed to a number between 1 and 500 so that the resulting fingerprint has 500 bits.

Encoding of editing mechanisms and binding site volume. Information about the editing mechanisms performed by some aaRSs were taken from the paper of Perona and Gruic-Sovulj[51] and encoded by appending a 22-dimensional bit vector to the 500-dimensional fingerprint. Each active bit represents a ligand against which editing is performed, e.g. for structures of ThrRS the bit for serine is set. In addition to editing information the binding site volume, estimated with the POVME[58] algorithm, was encoded. Twelve bins were created that represent binding site volumes ranging from 30–270 $Å^3$ in steps of 20 $Å^3$. For example, if a structure has a binding site volume of 45 $Å^3$ the first bit was set to active. For a binding site volume of, e.g., 52 $Å^3$ the second bit was set to active and so on. The fingerprints were concatenated to contain the binding site and interaction features (500 bits), editing mechanisms (22 bits), and binding site volume (12 bits). The final fingerprint has a size of 534 bits.

Embedding of interaction fingerprints. To allow for a quantitative comparison of the interactions between individual aaRSs, the high-dimensional interaction fingerprints were embedded using UMAP version 0.3.2[61]. The parameters for all embeddings given in this manuscript were set as follows: $|\mathtt{n_{neighbors}}| = 60$, $|\mathtt{min_{dist}}| = 0.1$, $|\mathtt{n_{components}}| = 2$. The Jaccard distance was used to describe the dissimilarity between two fingerprints a and b:

$$d(a,b) = 1 - \frac{n_{a\wedge b}}{n_a + n_b - n_{a\wedge b}} \tag{1}$$

with $n_{a \wedge b}$ being the count of active bits common between fingerprints a and b, n_a the number of active bits in fingerprint a, and n_b the number of active bits in fingerprint b. This distance metric was used as input for UMAP.

Data availability

All accompanying data is made publicly available under https://doi.org/10.5281/zenodo.3598250. This repository contains the MSA files of representative structures for each aaRS that were used for consistent renumbering as well as Excel tables to infer original sequence positions from renumbered positions for each aaRS. Rows are renumbered positions, columns are sequence positions of individual structures.

Received: 17 April 2020; Accepted: 6 July 2020
Published online: 28 July 2020

References

1. Bernhardt, H. S. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a). *Biol. Direct* **7**, 23 (2012).
2. Di Giulio, M. The origin of the genetic code: theories and their relationships, a review. *BioSyst.* **80**, 175–184 (2005).
3. Carter, C. W. & Wolfenden, R. tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proc. Natl. Acad. Sci. USA* **112**, 7489–7494 (2015).
4. Ibba, M. & Söll, D. Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* **69**, 617–650 (2000).
5. Dock-Bregeon, A. *et al.* Transfer RNA-mediated editing in threonyl-tRNA synthetase. The class II solution to the double discrimination problem.. *Cell* **103**, 877–884 (2000).
6. Hadd, A. & Perona, J. J. Coevolution of specificity determinants in eukaryotic glutamyl- and glutaminyl-tRNA synthetases. *J. Mol. Biol.* **426**, 3619–3633 (2014).
7. Nair, N. *et al.* The Bacillus subtilis and Bacillus halodurans Aspartyl-tRNA Synthetases Retain Recognition of tRNA(Asn). *J. Mol. Biol.* **428**, 618–630 (2016).
8. Song, Y. *et al.* Double mimicry evades tRNA synthetase editing by toxic vegetable-sourced non-proteinogenic amino acid. *Nat. Commun.* **8**, 2281 (2017).
9. Carter, C. W. & Wills, P. R. Interdependence, reflexivity, fidelity, impedance matching, and the evolution of genetic coding. *Mol. Biol. Evol.* **35**, 269–286 (2018).
10. Pak, D., Kim, Y. & Burton, Z. F. Aminoacyl-tRNA synthetase evolution and sectoring of the genetic code. *Transcription* **1–15**, (2018).
11. Swanson, R. *et al.* Accuracy of in vivo aminoacylation reguires proper balance of tRNA and aminoacyl-tRNA synthetase. *Science* **242**, 1548–1551 (1988).
12. Pham, Y. *et al.* A minimal TrpRS catalytic domain supports sense/antisense ancestry of Class I and II aminoacyl-tRNA Synthetases. *Mol. Cell* **25**, 851–862 (2007).
13. Yu, Y. *et al.* Crystal structure of human tryptophanyl-tRNA synthetase catalytic fragment: Insights into substrate recognition, tRNA binding, and angiogenesis activity. *J. Biol. Chem.* **279**, 8378–8388 (2004).
14. Doolittle, R. F., Handy, J. & Bada, J. L. Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.* **8**, 630–636 (1998).
15. O'Donoghue, P. & Luthey-Schulten, Z. On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol. Mol. Biol. Rev.* **67**, 550–573 (2003).
16. Davis, B. K. Molecular evolution before the origin of species. *Prog. Biophys. Mol. Biol.* **79**, 77–133 (2002).
17. Diaz-Lazcoz, Y. *et al.* Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. *Mol. Biol. Evol.* **15**, 1548–1561 (1998).
18. Woese, C. R., Olsen, G. J., Ibba, M. & Söll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64**, 202–236 (2000).
19. Chaliotis, A. *et al.* The complex evolutionary history of aminoacyl-tRNA synthetases. *Nucleic Acids Res.* **45**, 1059–1068 (2017).
20. Brown, J. R. & Doolittle, W. F. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92**, 2441–2445 (1995).
21. Carter, C. W. Coding of Class I and II aminoacyl-tRNA synthetases. *Adv. Exp. Med. Biol.* **966**, 103–148 (2017).
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* **347**, 203–206 (1990).
23. 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* **347**, 249–255 (1990).
24. Cusack, S. Aminoacyl-tRNA synthetases. *Curr. Opin. Struct. Biol.* **7**, 881–889 (1997).
25. Ibba, M. *et al.* A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* **278**, 1119–1122 (1997).
26. Curnow, A. W. *et al.* Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl. Acad. Sci. USA* **94**, 11819–11826 (1997).
27. Becker, H. D. & Kern, D. Thermus thermophilus: a link in evolution of the tRNA-dependent amino acid amidation pathways. *Proc. Natl. Acad. Sci. USA* **95**, 12832–12837 (1998).
28. Hartlein, M. & Cusack, S. Structure, function and evolution of seryl-tRNA synthetases: implications for the evolution of aminoacyl-tRNA synthetases and the genetic code. *J. Mol. Evol.* **40**, 519–530 (1995).
29. Leinfelder, W., Zehelein, E., Mandrand-Berthelot, M. A. & Bock, A. Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine. *Nature* **331**, 723–725 (1988).
30. Sheppard, K. *et al.* From one amino acid to another: tRNA-dependent amino acid biosynthesis. *Nucleic Acids Res.* **36**, 1813–1825 (2008).
31. Sauerwald, A. *et al.* RNA-dependent cysteine biosynthesis in archaea. *Science* **307**, 1969–1972 (2005).
32. Rodin, S. N. & Ohno, S. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. *Orig. Life Evol. Biosph.* **25**, 565–589 (1995).
33. Martinez-Rodriguez, 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.* **290**, 19710–19725 (2015).
34. Chandrasekaran, S. N., Yardimci, G. G., Erdogan, O., Roach, J. & Carter, C. W. Statistical evaluation of the Rodin-Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases. *Mol. Biol. Evol.* **30**, 1588–1604 (2013).
35. Carter, C. W. Urzymology: experimental access to a key transition in the appearance of enzymes. *J. Biol. Chem.* **289**, 30213–30220 (2014).
36. Carter, C. W. *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* **9**, 11 (2014).
37. Wong, J. T. F. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci.* **72**, 1909–1912 (1975).

38. Griffiths, G. Cell evolution and the problem of membrane topology. *Nat. Rev. Mol. Cell Biol.* **8**, 1018–1024 (2007).
39. Sonneborn, T. Degeneracy of the genetic code: extent, nature, and genetic implications. *Evolving genes and proteins* **377–397**, (1965).
40. Woese, C. R. Order in the genetic code. *Proc. Natl. Acad. Sci.* **54**, 71–75 (1965).
41. Arnez, J. G. & Moras, D. Structural and functional considerations of the aminoacylation reaction. *Trends Biochem. Sci.* **22**, 211–216 (1997).
42. Praetorius-Ibba, M. *et al.* Ancient adaptation of the active site of tryptophanyl-tRNA synthetase for tryptophan binding. *Biochemistry* **39**, 13136–13143 (2000).
43. Kaiser, F. *et al.* Backbone brackets and Arginine Tweezers delineate Class I and Class II aminoacyl tRNA synthetases. *PLoS Comput. Biol.* **14**, e1006101 (2018).
44. Klebe, G. & Bohm, H. J. Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J. Recept. Signal Transduct. Res.* **17**, 459–473 (1997).
45. Salentin, S., Haupt, V. J., Daminelli, S. & Schroeder, M. Polypharmacology rescored: protein-ligand interaction profiles for remote binding site similarity assessment. *Prog. Biophys. Mol. Biol.* **116**, 174–186 (2014).
46. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated protein-ligand interaction profiler. *Nucleic Acids Res.* **43**, W443-447 (2015).
47. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
48. Iwasaki, W. *et al.* Structural basis of the water-assisted asparagine recognition by asparaginyl-tRNA synthetase. *J. Mol. Biol.* **360**, 329–342 (2006).
49. Livingstone, C. D. & Barton, G. J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756 (1993).
50. Bonnici, V., Giugno, R., Pulvirenti, A., Shasha, D. & Ferro, A. A subgraph isomorphism algorithm and its application to biochemical data. *BMC Bioinform.* **14 Suppl 7**, S13 (2013).
51. Perona, J. J. & Gruic-Sovulj, I. Synthetic and editing mechanisms of aminoacyl-tRNA synthetases. *Top. Curr. Chem.* **344**, 1–41 (2014).
52. Fersht, A. R. *et al.* Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* **314**, 235–238 (1985).
53. Fukai, S. *et al.* Structural basis for double-sieve discrimination of L-valine from L-isoleucine and L-threonine by the complex of tRNA(Val) and valyl-tRNA synthetase. *Cell* **103**, 793–803 (2000).
54. Zivkovic, I., Moschner, J., Koksch, B. & Gruic-Sovulj, I. Mechanism of discrimination of isoleucyl-tRNA synthetase against nonproteinogenic -aminobutyrate and its fluorinated analogues. *FEBS J* (2019).
55. Guo, M. *et al.* Paradox of mistranslation of serine for alanine caused by AlaRS recognition dilemma. *Nature* **462**, 808–812 (2009).
56. Fersht, A. . R. & Dingwall, C. Evidence for the double-sieve editing mechanism in protein synthesis. Steric exclusion of isoleucine by Valyl-tRNA synthetases. *Biochemistry* **18**, 2627–2631 (1979).
57. 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* **32**, 792–794 (2016).
58. Durrant, J. . D., Votapka, L., Sørensen, J. & Amaro, R. . E. POVME 2.0: an enhanced tool for determining pocket shape and volume characteristics. *J. Chem. Theory Comput.* **10**, 5047–5056 (2014).
59. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
60. Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8 (2015).
61. McInnes, L. *& Healy, J* (Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints, UMAP, 2018).
62. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
63. Wolfenden, R., Lewis, C. A., Yuan, Y. & Carter, C. W. Temperature dependence of amino acid hydrophobicities. *Proc. Natl. Acad. Sci. USA* **112**, 7484–7488 (2015).
64. W. Carter, C. & Wills, P. Did gene expression co-evolve with gene replication? In *Evolutionary Biology: Origin and Evolution of Biodiversity*, 293–313, https://doi.org/10.1007/978-3-319-95954-2_16 (Springer International Publishing, 2018).
65. Yarus, M. Primordial genetics: phenotype of the ribocyte. *Annu. Rev. Genet.* **36**, 125–151 (2002).
66. Wills, P. R. The generation of meaningful information in molecular systems. *Philos. Trans. R. Soc. A* **374**, 20150066 (2016).
67. Wills, P. R. Spontaneous mutual ordering of nucleic acids and proteins. *Orig. Life Evol. Biosph.* **44**, 293–298 (2014).
68. Chin, J. W. Expanding and reprogramming the genetic code. *Nature* **550**, 53–60 (2017).
69. Dunkelmann, D. L., Willis, J. C. W., Beattie, A. T. & Chin, J. W. Engineered triply orthogonal pyrrolysyl-tRNA synthetase/tRNA pairs enable the genetic encoding of three distinct non-canonical amino acids. *Nat. Chem.* **12**, 535–544 (2020).
70. Chatterjee, A., Xiao, H., Yang, P. . Y., Soundararajan, G. & Schultz, P. . G. A tryptophanyl-tRNA synthetase/tRNA pair for unnatural amino acid mutagenesis in E. coli. *Angew. Chem. Int. Ed. Engl.* **52**, 5106–5109 (2013).
71. Chen, M. *et al.* Cross-editing by a tRNA synthetase allows vertebrates to abundantly express mischargeable tRNA without causing mistranslation. *Nucleic Acids Research* **gkaa469** (2020).
72. Sun, F. J. & Caetano-Anollés, G. The origin and evolution of tRNA inferred from phylogenetic analysis of structure. *J. Mol. Evol.* **66**, 21–35 (2008).
73. Galili, T., Gingold, H., Shaul, S. & Benjamini, Y. Identifying the ligated amino acid of archaeal tRNAs based on positions outside the anticodon. *RNA* **22**, 1477–1491 (2016).
74. Tamaki, S., Tomita, M., Suzuki, H. & Kanai, A. Systematic analysis of the binding surfaces between tRNAs and their respective aminoacyl tRNA synthetase based on structural and evolutionary data. *Frontiers Genet.* **8**, (2018).
75. Carter, C. W. & Wills, P. R. Hierarchical groove discrimination by Class I and II aminoacyl-tRNA synthetases reveals a palimpsest of the operational RNA code in the tRNA acceptor-stem bases. *Nucleic Acids Res.* (2018).
76. Dutta, S., Choudhury, K., Banik, S. D. & Nandi, N. Active site nanospace of aminoacyl tRNA synthetase: difference between the class I and class II synthetases. *J. Nanosci. Nanotechnol.* **14**, 2280–2298 (2014).
77. Davis, B. K. Evolution of the genetic code. *Prog. Biophys. Mol. Biol.* **72**, 157–243 (1999).
78. Klipcan, L. & Safro, M. Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. *J. Theor. Biol.* **228**, 389–396 (2004).
79. Weber, A. L. & Miller, S. L. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* **17**, 273–284 (1981).
80. Rogers, S. O. Evolution of the genetic code based on conservative changes of codons, amino acids, and aminoacyl tRNA synthetases. *J. Theor. Biol.* **466**, 1–10 (2019).
81. Newton, M. S., Morrone, D. J., Lee, K. H. & Seelig, B. Genetic code evolution investigated through the synthesis and characterisation of proteins from reduced-alphabet libraries. *Chembiochem* **20**, 846–856 (2019).
82. Cammer, S. & Carter, C. W. Six rossmannoid folds, including the class I aminoacyl-tRNA synthetases, share a partial core with the anti-codon-binding domain of a Class II aminoacyl-tRNA synthetase. *Bioinformatics* **26**, 709–714 (2010).
83. Bittrich, S. *et al.* Application of an interpretable classification model on Early Folding Residues during protein folding. *BioData Min.* **12**, 1 (2019).

84. Lu, M. F., Xie, Y., Zhang, Y. J. & Xing, X. Y. Effects of cofactors on conformation transition of random peptides consisting of a reduced amino acid alphabet. *Protein Pept. Lett.* **22**, 579–585 (2015).
85. Kang, S. K. *et al.* ATP selection in a random peptide library consisting of prebiotic amino acids. *Biochem. Biophys. Res. Commun.* **466**, 400–405 (2015).
86. Ribas de Pouplana, L. & Schimmel, P. Two classes of tRNA synthetases suggested by sterically compatible dockings on tRN. *Cell* **104**, 191–193 (2001).
87. Wong, J. T. F. Coevolutlon theory of genetic code at age thirty. *BioEssays* **27**, 416–425 (2005).
88. Brooks, D. J., Fresco, J. R., Lesk, A. M. & Singh, M. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19**, 1645–1655 (2002).
89. Trifonov, E. N. The triplet code from first principles. *J. Biomol. Struct. Dyn.* **22**, 1–11 (2004).
90. Lamour, V. *et al.* Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. *Proc. Natl. Acad. Sci. USA* **91**, 8670–8674 (1994).
91. Carter, C. W., Popinga, A., Bouckaert, R. & Wills, P. R. Class i aminoacyl-trna synthetase urzyme and cp1 modules have distinct genetic origins. *bioRxiv* https://doi.org/10.1101/2020.04.09.033712 (2020).
92. Martinis, S. A. & Boniecki, M. T. The balance between pre- and post-transfer editing in tRNA synthetases. *FEBS Lett.* **584**, 455–459 (2010).
93. Schulze, J. O. *et al.* Crystal structure of a non-discriminating glutamyl-tRNA synthetase. *J. Mol. Biol.* **361**, 888–897 (2006).
94. Perona, J. J., Rould, M. A. & Steitz, T. A. Structural basis for transfer RNA aminoacylation by *Escherichia coli* Glutaminyl-tRNA synthetase. *Biochemistry* **32**, 8758–8771 (1993).
95. Bullock, T. L., Uter, N., Nissan, T. A. & Perona, J. J. Amino acid discrimination by a class I aminoacyl-tRNA synthetase specified by negative determinants. *J. Mol. Biol.* **328**, 395–408 (2003).
96. Sever, S., Rogers, K., Rogers, M. J., Carter, C. & Söll, D. Escherichia coli tryptophanyl-tRNA synthetase mutants selected for tryptophan auxotrophy implicate the dimer interface in optimizing amino acid binding. *Biochemistry* **35**, 32–40 (1996).
97. Jakubowski, H. Misacylation of tRNA(Lys) with noncognate amino acids by Lysyl-tRNA synthetase. *Biochemistry* **38**, 8088–8093 (1999).
98. Nagel, G. M. & Doolittle, R. F. Evolution and relatedness in two aminoacyl-tRNA synthetase families. *Proc. Natl. Acad. Sci. USA* **88**, 8121–8125 (1991).
99. Qin, X. *et al.* Cocrystal structures of glycyl-tRNA synthetase in complex with tRNA suggest multiple conformational states in glycylation. *J. Biol. Chem.* **289**, 20359–20369 (2014).
100. Naganuma, M., Sekine, S., Fukunaga, R. & Yokoyama, S. Unique protein architecture of alanyl-tRNA synthetase for aminoacylation, editing, and dimerization. *Proc. Natl. Acad. Sci. USA* **106**, 8489–8494 (2009).
101. Nakayama, T. *et al.* Deficient activity of alanyl-tRNA synthetase underlies an autosomal recessive syndrome of progressive microcephaly, hypomyelination, and epileptic encephalopathy. *Hum. Mutat.* **38**, 1348–1354 (2017).
102. Splan, K. E., Ignatov, M. E. & Musier-Forsyth, K. Transfer RNA modulates the editing mechanism used by class II prolyl-tRNA synthetase. *J. Biol. Chem.* **283**, 7128–7134 (2008).
103. Rayevsky, A., Sharifi, M. & Tukalo, M. A molecular dynamics simulation study of amino acid selectivity of LeuRS editing domain from Thermus thermophilus. *J. Mol. Graph. Model.* **84**, 74–81 (2018).
104. Doublie, S., Bricogne, G., Gilmore, C. & Carter, C. W. Tryptophanyl-tRNA synthetase crystal structure reveals an unexpected homology to tyrosyl-tRNA synthetase. *Structure* **3**, 17–31 (1995).
105. Wolf, Y. I., Aravind, L., Grishin, N. V. & Koonin, E. V. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9**, 689–710 (1999).
106. Fournier, G. P. & Alm, E. J. Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. *J. Mol. Evol.* **80**, 171–185 (2015).
107. Weinreb, V. *et al.* 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.* **289**, 4367–4376 (2014).
108. Li, L. & Carter, C. W. Full implementation of the genetic code by tryptophanyl-tRNA synthetase requires intermodular coupling. *J. Biol. Chem.* **288**, 34736–34745 (2013).
109. Ribas de Pouplana, L., Frugier, M., Quinn, C. . L. & Schimmel, P. Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria. *Proc. Natl. Acad. Sci. USA* **93**, 166–170 (1996).
110. Yang, X. L. *et al.* Crystal structures that suggest late development of genetic code components for differentiating aromatic side chains. *Proc. Natl. Acad. Sci. USA* **100**, 15376–15380 (2003).
111. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
112. Prlic, A. *et al.* BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **28**, 2693–2695 (2012).
113. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
114. Leberecht, C., Kaiser, F., Bittrich, S. & Krautwurst, S. cleberecht/singa: singa-all release v0.4.0, https://doi.org/10.5281/zenodo.1320146 (2018).
115. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res.* **44**, D1202-1213 (2016).

Author contributions

F.K. and S.K. prepared and analyzed data. F.K., S.K, and S.S. wrote the manuscript. S.S. designed and computed interaction fingerprints, V.J.H. implemented renumbering of structures, C.L. and S.B. supported data curation, formal analysis, and conceptualization. D.L. and M.S supervised the project. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-69100-0.

Correspondence and requests for materials should be addressed to F.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.