Computers in Biology and Medicine 190 (2025) 110033

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

The proteomic code: Novel amino acid residue pairing models "encode" protein folding and protein-protein interactions

Tareq Hameduh [a,b], Andrew D. Miller [a,b,c,d], Zbynek Heger [a,b], Yazan Haddad [a,b,*,1]

[a] Department of Chemistry and Biochemistry, Mendel University in Brno, Zemědělská 1665/1, CZ-613 00, Brno, Czech Republic
[b] MendelFOLD s.r.o., Zezulova 174/3, CZ-613 00, Brno, Czech Republic
[c] Veterinary Research Institute, Hudcova 296/70, CZ-621 00, Brno, Czech Republic
[d] KP Therapeutics (Europe) s.r.o., Purkyňova 649/127, CZ-612 00, Brno, Czech Republic

ARTICLE INFO

Keywords:
Proteomic code
Contact map
Sense-antisense
Protein 3D structure
Protein folding
Protein-protein interactions

ABSTRACT

Recent advances in protein 3D structure prediction using deep learning have focused on the importance of amino acid residue-residue connections (*i.e.*, pairwise atomic contacts) for accuracy at the expense of mechanistic interpretability. Therefore, we decided to perform a series of analyses based on an alternative framework of residue-residue connections making primary use of the TOP2018 dataset. This framework of residue-residue connections is derived from amino acid residue pairing models both historic and new, all based on genetic principles complemented by relevant biophysical principles. Of these pairing models, three new models (named the GU, Transmuted and Shift pairing models) exhibit the highest observed-over-expected ratios and highest correlations in statistical analyses with various intra- and inter-chain datasets, in comparison to the remaining models. In addition, these new pairing models are universally frequent across different connection ranges, secondary structure connections, and protein sizes. Accordingly, following further statistical and other analyses described herein, we have come to a major conclusion that all three pairing models together could represent the basis of a universal proteomic code (second genetic code) sufficient, in and of itself, to "encode" for both protein folding mechanisms and protein-protein interactions.

1. Introduction

Recent applications of deep learning algorithms in the prediction of protein 3D structures have drawn attention to the role of direct atomic connections (*i.e.*, contact maps and distance maps) in accurate 3D models. However, in spite of the increasing accuracy of these models, a significant gap remains in their biological interpretability. The inherent 'black box' nature of these models complicates the development of biomimetic, biophysically relevant theories that could explain the relationship between sequence and structure in more mechanistic ways [1–3]. This is unfortunate given that recent research, building on nearly four decades of investigations, has underscored the central mechanistic role of amino acid residue-residue intra-connections for protein folding prediction [4–6], and inter-connections for protein-protein interactions (PPIs) [7–9]. In addition, connections within protein sequences (technically called contacts) can also be predicted from correlations between patterns of amino acid residue substitutions that derive from homology studies. Since homology in sequence reflects homology in structure, then connections can be used as constraints for 3D protein structure modelling [10]. In a comparable way, multiple sequence alignment (MSA) analyses can be used for 3D protein structure prediction based on contact maps and deep learning. Indeed, homology studies and MSA analyses are central to most recent developments in protein 3D structure predictions, as documented in the Critical assessment of structure prediction (CASP) experiments [11], and most recently, multiple MSA inputs were shown crucial for higher performance predictions by many participating groups in the latest CASP15 assessment [12–14]. However, homology and MSA based structure predictions should be understood as a by-product of folding rather than a causal factor. Furthermore, 3D protein structure predictions based on homology will always be limited by the availability of data from homologous experimental protein structures in protein structure databases. Accordingly, in response to this situation and the current structure prediction state-of-the-art, we decided instead to focus our analyses on an alternative framework of

* Corresponding author.
E-mail address: yazan.haddad@mendelu.cz (Y. Haddad).
[1] Present Address: Department of Chemistry and Biochemistry, Mendel University in Brno, Zemědělská 1665/1, CZ-613 00 Brno, Czech Republic.

https://doi.org/10.1016/j.compbiomed.2025.110033
Received 5 August 2024; Received in revised form 11 March 2025; Accepted 13 March 2025
Available online 19 March 2025

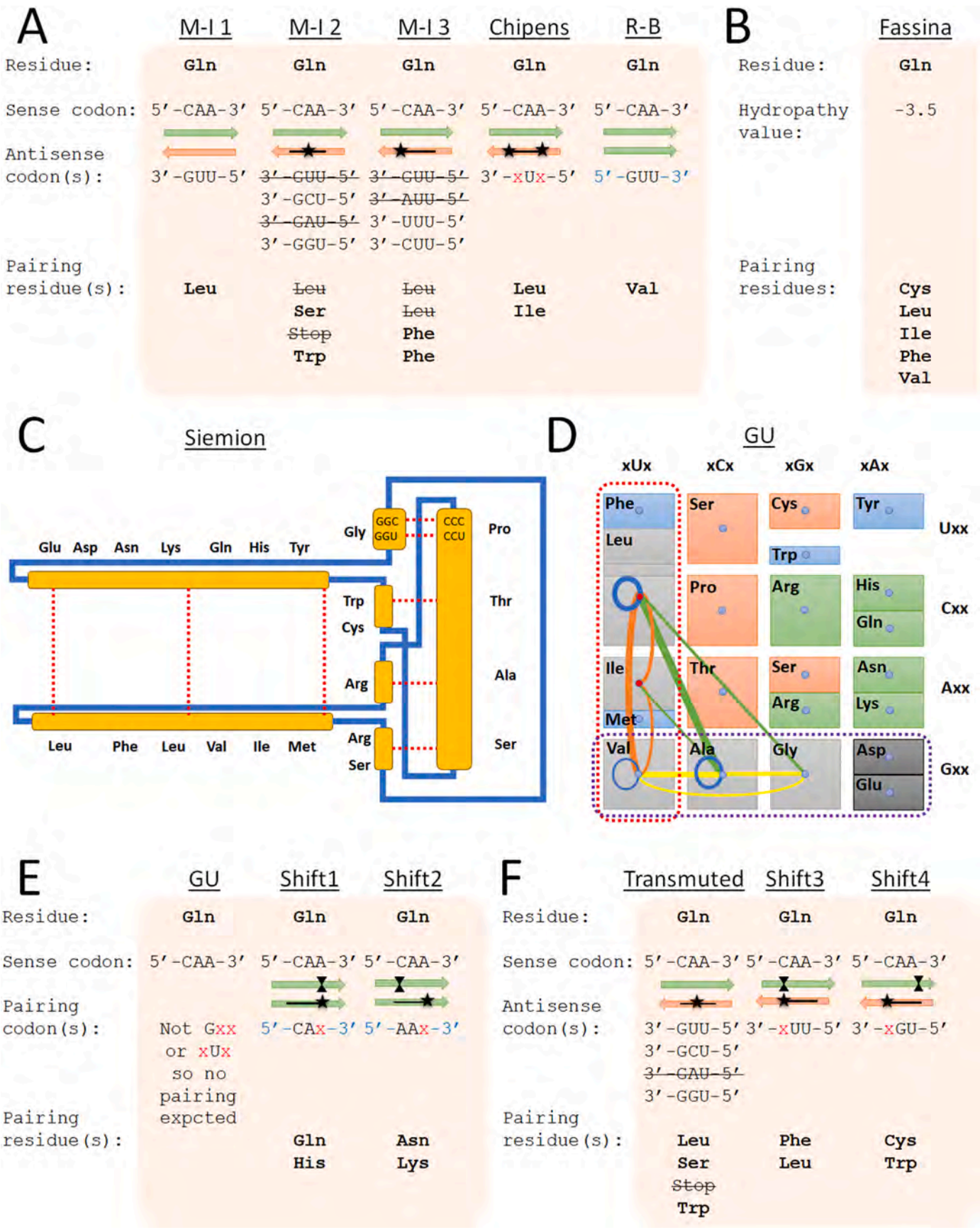
amino acid residue-residue connections based on amino acid residue pairing models both historic and new, all of which originate from genetic principles complemented by relevant biophysical principles. In so doing, we hoped to discover alternate ways to reach an understanding of protein structure/function without the data limitations imposed by homology and MSA approaches.

The historic amino acid residue pairing models selected for investigation are summarised here (Fig. 1A–C) followed by a summary of new pairing models of our own devising that were derived from alternative genetic principles admixed with relevant biophysical principles (Fig. 1D–F). In terms of the historic pairing models, Mekler [15], together later with Idlis [16,17], were the first to propose the possibility of a genetic principle to account for amino acid residue-residue connections, according to which codon-specified sense amino acid residues and corresponding antisense codon-specified antisense amino acid residues should be potentially capable of specific interactions through space with each other. Hence, for example, according to this hypothesis, an AUG sense codon-specified Met residue should be capable of specific interactions through space with a corresponding CAU antisense codon-specified His residue. Miller and team later described the complete set of amino acid residue-residue connections based on this genetic principle as "Mekler-Idlis (M-I 1) pairs" (Fig. 1A). In addition, Miller and team performed numerous biophysical and biological studies to characterise M-I 1 pairs in sense-antisense peptide interactions and evaluated their contributions to biology. These studies with M-I 1 pairs and sense-antisense peptides have been extensively reviewed elsewhere along with the work of other principal groups in the field [18,19]. One consequence of these studies was to highlight the fact that M-I 1 residue-residue connections are rather limited and incomplete from a biological point of view. Therefore, the Miller team [18,19] went on to propose the possibility of an expanded M-I 2 pairing model (Fig. 1A), based on 2nd base wobbles in corresponding antisense codons, and a M-I 3 pairing model (Fig. 1A), based on potential 3rd base wobbles in corresponding antisense codons [18,19]. By way of explanation, for example, according to the M-I 2 model, an AUG sense codon-specified Met residue should be capable of specific interactions through space with either a corresponding CUU antisense codon-specified Leu residue, a "2nd base wobble" CCU antisense codon-specified Pro residue, or a "2nd based wobble" CGU antisense codon-specified Arg residue. Similarly, for example, in the case of the M-I 3 model, an AUG sense codon-specified Met residue should be capable of specific interactions in space with either a "3rd base wobble" CAA/CAG antisense codon-specified Gln residue, or a "3rd base wobble" CAC antisense codon-specified His residue (a complete set of current pairing models is provided in Table S1).

One particular criticism of the M-I 1 pairing model has been that connections between Met and Ile are not allowed interchangeably via sense-antisense codons [20]. In addition, the M-I 1 pairing model was widely noted to lack sense-antisense self-complementarity within the open reading frames (ORFs) of genes, as shown by Zull and others [21]. Nevertheless, the M-I 1 pairing model has proved to be a remarkably durable model for studies into residue-residue connections not least because, as Root-Bernstein and Holsworth [22] have pointed out, Biro [23,24], and Blalock & Smith [25] were able to "reinvent" the M-I 1 pairing model with the help of the Kyte-Doolittle hydropathy scale [26], thereby using just biophysical principles to explain and account for sense-antisense peptide interactions. Given this, the M-I 1 model is also sometimes referred to as the Mekler-Biro-Blalock (MBB) hypothesis with embedded physicochemical properties. On the other hand, Chipens *et al.* [20,27,28] derived an alternative to the M-I 1 pairing model [15] based on their "roots of the codons" approach (Fig. 1A). In using this "roots of the codons" approach, the original 13 M-I 1 pairs were expanded to a possible 21 pairs depending on the existence of 2nd base G/C or A/U complementarities between sense and corresponding antisense codons. Basically, the Chipens pairing model is the opposite of the M-I 2 pairing model in the sense that it invokes partial 1st and 3rd base wobbles as opposed to 2nd base wobbles. Chipens *et al.* [20] have suggested that their model should perform better than the M-I 1 pairing model where alternative genetic code rules apply such as those found in mitochondrial genes in comparison to the *E. coli* genome, due to less allowed/prohibited connections in accordance with the evolution of natural amino acid residue families.

Subsequently, Root-Bernstein [22,29] proposed an alternative pairing model and genetic rationale involving the parallel reading of codons and anticodons, as based on the parallel nature of β-sheet strands in proteins, giving rise to 15 possible pairs (Fig. 1A). Essentially, the Root-Bernstein (R-B) model assumes that anticodons are read in 3′–5′ whereas M-I pairing models assume that anticodons are read in the 5′–3′ manner which explains the vast variation between the R-B and M-I pairing models. By way of explanation, for example, according to the R-B model, an AUG sense codon-specified Met residue should be capable of specific interactions with a corresponding UAC antisense codon-specified Tyr residue. This R-B model has the advantage that each amino acid residue will connect with either one or a group of structurally similar amino acid residues if their side-chains interact in a parallel β-sheet [22]. One further alternative to M-I 1 pairing is represented by the idiosyncratic Siemion model (Fig. 1C) which was based on the observed periodicity in codons when they are ordered following one-step point mutations (for example, a sequence of Gly codons -GGC-GGU-GGA-GGG- complement on the opposite side of a ring with a sequence of Pro codons -CCG-CCA-CCU-CCC-), which also cross-correlate with secondary structure properties (also known as Chou-Fasman conformational parameters) [30–32]. In fact, both R-B and Siemion models are very similar in the sense that they represent only a small number of amino acid residue-residue connections with only a few differences between them. Arguably, the R-B/Siemion models are not general to protein folds but may represent a subset of amino acid residue pairing models particular to parallel β-sheet formation. Finally, we introduce the work of Fassina and colleagues [33–36] who adopted a mathematical approach based on the Kyte-Doolittle hydropathy scale [26] to describe a pairing model based on 41 possible pairs of hydropathically complementary amino acid residues (Fig. 1B).

Over the years, the field of sense-antisense peptides has been pivotal in the development of concepts and theories concerning amino acid residue-residue connections and their roles in specific interactions between sense and corresponding antisense peptides. However, although the field might boast some genuine successes, there have been many unexpected and unexplained failures as well (reviewed in Refs. [19,37]). For example, recently, an R-B inspired antisense peptide was designed against insulin but failed to produce the anticipated interactions [38]. In contrast, an M-I 1 inspired antisense peptide against CD81 was found successful in achieving a measurable target site binding affinity and mediating treatment of metastatic cancer [39]. Such variability in biological efficacy may reflect variations in methods of validation, but may also reflect inbuilt limitations of the historic amino acid residue pairing models described above (Fig. 1A–C). Sense-antisense peptide studies have also been potentially pivotal in another way. In 2002, on the basis of their sense-antisense peptide research, Miller and team introduced a new term, namely the 'proteomic code', which was used to refer to the possibility that an amino acid residue interaction code may exist embedded in and/or deriving from the standard genetic code, based on the M-I pairing models [18]. By way of explanation, according to their proposal, assuming that amino acid residues coded for by sense codons (according to the standard genetic code) are indeed capable of specific side-chain interactions with amino acid residues coded for by their corresponding antisense codons (according to the same standard genetic code) [18], then logically a more general amino acid residue interaction code (proteomic code) should exist derived from the standard genetic code based on similar genetic principles. For this reason, the putative proteomic code was also referred to as a "second genetic code" [18]. Furthermore, assuming that the proteomic code exists, then the standard genetic code could also be said to possess not just 1) the well-known function to determine protein sequences from gene sequences, but also



(caption on next page)

Fig. 1. Amino acid residue pairing models. (A–C) Amino acid residue sense-to-antisense codon-specified pairing models previously described in literature. The Siemion model is based on a one-step mutation periodical ring where rows have complementary amino acid residues and columns have complementary amino acid residues as shown in panel **(C). (D–F)** Amino acid residue pairing models of our own devising here (GU, Shift1 and Shift2 involve sense-to-sense codon-specified pairing, while Transmuted, Shift3 and Shift4 involve sense-to-antisense codon-specified pairing). In panel **(D)**, the GU model invokes interactions between all amino acid residues specified by Gxx and xUx sense codons, as shown on the reduced codon table (top 12 connections are shown involving Leu, Ile, Val, Ala and Gly). Hence, the GU model defines interactions between hydrophobic, small and negative amino acid residues. **(E)** The Shift1 and Shift2 models are based on a variation of direct sense-to-sense codon-specified amino acid residue interactions. Shift1 occurs when the 3rd base of a given sense codon is deleted which results in effect in a 3rd base wobble in the sense codon, and hence new possible sense-to-sense codon-specified amino acid residue interactions. Shift2 occurs when the 1st base of given sense codon is deleted which results in different 3rd base wobbles in the sense codon, hence defining alternative possible sense-to-sense codon-specified amino acid residue interactions. **(F)** The Transmuted pairing model is in fact an amalgamation of the M-I 1 and M-I 2 pairing models. Shift3 occurs when the 1st base of a given sense codon is deleted which results in a 3rd base wobble in the corresponding antisense codon. Shift4 occurs when the 3rd base of given sense codon is deleted which results in a different possible 3rd base wobble in the corresponding antisense codon. Orientations of sense and antisense codons are important (unusual orientations are typed in blue). Antisense codons are crossed out if they were presented in previous M-I pairing models or if they represent a stop codon. Arrows represent DNA strands with wobble shown as star symbol and deletion shown as flowchart-collate symbol. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2) an additional function to determine through-space interactions between amino acid residues in a folding/folded polypeptide, and 3) a further function to provide a robust system of redundancies that permits evolution of amino acid residue connections through genetic variations (translocation, duplication, point-mutations and insertion/deletion) [18]. Both the second and third functions would represent a substantial increase in the functional scope and potential importance of the standard genetic code.

The new amino acid residue pairing models selected for investigation here are summarised (Fig. 1D–F). These consist of: 1) a GU pairing model based on the fact that Gxx or xUx codons, which occupy two whole sides of the standard genetic code table, typically code for hydrophobic and small or negatively charged amino acid residues that are known to be capable of specific pairwise side-chain/side-chain interactions with each other, according to previous data from contact map datasets (Fig. 1D) [40,41]; 2) a series of Shift pairing models, namely Shift1 and Shift2 derived from sense codon deletions of the 3rd and 1st base, respectively, and Shift3 and Shift4 derived from the impact on corresponding antisense codons of sense codon deletions of the 1st and 3rd base, respectively (Fig. 1E–F); 3) a Transmuted pairing model (formed by combining M-I 1 and M-I 2 models) to cover in one model all possible 2nd base variations in corresponding antisense codons (Fig. 1F). In all cases, Shift-amino acid residue-residue connections are defined in effect by the possibility that amino acid residues, coded for by sense codons, can interact specifically with corresponding amino acid residues specified by "Shift" NNx pairing/antisense codons. These new models were all devised for our studies here in response to the perceived limitations of the historic pairing models, including the fact that they are either limited by their low repertoire of possible amino acid residue-residue connections, are too specific in their scope (*e.g.*, R-B/Siemion models), and/or have been validated inconsistently in wet lab experiments.

Here, we now present our data from a series of biocomputational analyses that were performed using our alternative framework of amino acid residue-residue connections, and one primary high-quality dataset of protein 3D structures. Data were generated to indicate the significances or otherwise of all the amino acid residue-residue pairing models by means of frequency comparisons, observed/expected (O/E) ratios and correlations across six parsed datasets covering intra-connections, dimer inter-connections (both homodimers and heterodimers) and heterodimer inter-connections. Quantification of the connections are performed primarily using two atomistic contacts (standard C_β atom 8 Å cutoff and distant side-chains atom 8 Å cutoff) and three frequency modes (direct – F, normalised – NF, and relative – RF, frequencies) to avoid biases from the numbers of residues and numbers of atomic contacts. The NF mode derives from division of frequencies by a harmonic mean scale to balance the influence of residue numbers and numbers of atomic contacts (for each residue), particularly when these values are extreme. The RF mode represents a relative ratio between frequencies and the product of the numbers of residues and numbers of atomic contacts. In addition, since hydrophobicity is a key attractive force in driving amino acid residue-residue connections, correlations are included of residue-residue connections and pairing models with published hydrophobicity scales. Furthermore, we take advantage of the various amino acid residue relationship matrices deposited in the AAindex database [42], and use connection preferences in Dynameomics simulations [43], to search for apparent biorelevant correlations from our findings. In culmination, a case study is presented using a small model protein to visualize and illustrate the potential applicability of each main pairing model in accelerating protein folding during simulation studies.

2. Materials and methods

All methods were implemented in R (version 4.2.2) using RStudio (version 2023.09.1) for Windows 11. The R packages bio3d (version 2.4.4), Peptides (version 2.4.5), dplyr (version 1.1.1), and stringi (version 1.7.12) were used for structure analysis, hydrophobicity analysis, dataset manipulation, and string manipulation, respectively. Some figures were constructed in Microsoft 365 Excel/PowerPoint (version 2312), and some assembled in Adobe Photoshop CS6 (version 13.0.1).

2.1. Amino acid pairing models

A list of few models, compiled by A. E. Lyubarev (https://lyubarev.narod.ru/science/tabl1.htm last accessed in 24. 01. 2024), was very useful in constructing connections lists. However, several modifications were needed to allow for building symmetric matrices (where X amino acid connects with Y and Y connects with X). Final compiled models are shown in Table S1 in binary logical format where 0 is a non-connection and 1 is a connection (or categorical format in the case of Transmuted model). M-I 1 model refers to Mekler [15], M-I 2 and M-I 3 models refer to Miller's interpretations of second and third base wobbles in antisense codons [16–19], Chipens model refers to Chipens and colleagues [20,27,28], R-B model refers to Root-Bernstein [22,29], Siemion model refers to Siemion and colleagues [30–32] and Fassina model refers to Fassina and colleagues [33–36]. Transmuted model (Table S2) refers to combining M-I 1 and M-I 2, then dividing them into groups based on the type of second base substitution (0: non-connection; 1: Transversion – purine to pyrimidine and *vice versa*; 2: Transition – purine to purine and pyrimidine to pyrimidine; 3: Transition/Transversion). Other variations (Transmuted_transverse and Transmuted_transition) were tested to evaluate the importance of Transition by substituting their numeration giving a Transversion/Transition value of 3, respectively. Transmuted_bi refers to the binary model of combined M-I 1 and M-I 2 without further division [16,17]. To avoid confusion, only the best performing Transmuted variant is shown, while the remaining are listed in Table S1. GU model refers to all possible connections between amino acids Phe, Leu, Ile, Met, Val, Ala, Gly, Asp, and Glu. Shifts models were generated by deletion of terminal bases (either first or third base) in the sense or

antisense codons (computation of Shifts models is shown in Tables S3–S6)

2.2. 3D structures dataset

For all purposes in this work, the all-atom filtered Top2018 protein 3D structure dataset (https://zenodo.org/records/5773255 last accessed in 09. 12. 2024) was used at 70 % homology [44]. This dataset is one of the highest quality recent protein 3D structure datasets available. It is meticulously curated through advanced multi-level filtration, including crystal resolution, chain, residue and atom level filtering. This is done using high resolution of crystals, verified electron density, atomic-level filtering regarding B factor, geometric filtering using MolProbity that significantly reduces clashes, N/Q/H flips correction and the fact it is curated for low redundancy sequences according to level of interest. All these factors enhance the accuracy of atomic contact measurements. Based on our previous studies [45–47], the solvent accessibility plays significant role in conformations of the side-chains, and thus we have limited our protein folding study to intra-connection of single-chain only structures (Fig. 2). To study the interaction interfaces, and since the Top2018 dataset includes only separate chains, we have screened and filtered the double-chained origins to extract inter-connections. We have anticipated two issues, firstly, by downloading the double-chained origins PDB files from wwPDB databank, part of the quality filtering is lost (this was compensated by B factor and occupancy filtering), and secondly, it is known that most double-chained structures in wwPDB databank are homodimers (this was compensated by 95 % sequence identity filtering). Intra-connections were collected in two datasets named cmap_beta and cmap_side-chains, whereas inter-connections that include all dimers were collected in cmap_beta_dimer and cmap_side-chains_dimer, whereas those representing heterodimers (filtered at 95 % sequence identity) were designated cmap_beta_hetero and cmap_side-chains_hetero (Fig. 2).

2.3. Connections

To assess the relevance of models to contact maps, the frequency of connections is calculated after all the connections have been established in the folded protein in the crystal structure. In other words, we are not testing false negatives and false positives for all possible connections that can occur based on entire protein sequence. However, the frequency of connections (also their sum ratio to account for bias of frequency of residues) can give insight to the prevalence of particular connections that belong to particular model. For any connection between amino acid X and amino acid Y, the XY or YX pairs (variable name: XYorYX) are computed regardless of their order in protein sequence and written in the format "X|Y" with alphabetical sorting (*i.e.*, connections "X→Y" and "Y→X" are combined in one as "X|Y"). This yields a total of 210 X|Y connections instead of $20 \times 20 = 400$ possible amino acid combinations (the latter is described in literature in unidirectional format where "X→Y" and "Y→X" are calculated independently).

Atomic contact between two residues was studied in two formats: The cmap_beta connection follow the standard protocols, where distances between C_β atoms (and C_α atom in the case of Gly residue) are less than 8.0 Å [10,48]. The cmap_side-chains connection is focused on the distances between the side-chain terminal atoms or in some amino acids the heavy atoms at the extremity of the functional group or the atom at the centre of the terminal branch in the side chain (Table S7) at the same cutoff of 8.0 Å. The 8.0 Å cutoff reliably captures atomic contacts relevant for various physicochemical interactions, such as hydrogen bonds, van der Waals interactions, and electrostatic contacts, which are fundamental to protein stability and function [48,49].

The percentage of frequency for X|Y connection is calculated according to following equation:

$$Frequency^{X|Y}(F) = \frac{C^{X|Y}}{C^{All}}$$

Where $C^{X|Y}$ is the number of atomic contacts between X and Y and C^{All} is the total number of atomic contacts. The F frequency is biased by the number of residues (number of X and number of Y) and connections (atomic contacts made by X and atomic contacts made by Y).

The normalised frequency for X|Y connection is the number of atomic contacts between these pairs normalised for each residue in the pair by both the percentage of residue counts in dataset and the percentage of contacts as follows:

$$Normalised\_Frequency^{X|Y}(NF) = \frac{C^{X|Y}}{\left(\frac{1}{R^X \times C^X}\right) + \left(\frac{1}{R^Y \times C^Y}\right)}$$

Where $C^{X|Y}$ is the number of atomic contacts between X and Y, R^X is the percentage of occurrence of residue X in single chain dataset (Fig. 2), C^X is the percentage of number of atomic contacts made by residue X in relation to total number of atomic contacts, R^Y is the percentage of occurrence of residue Y in single chain dataset (Fig. 2), C^Y is the percentage of number of atomic contacts made by residue Y in relation to total number of atomic contacts.

Normalised frequency balances the possible bias coming from the

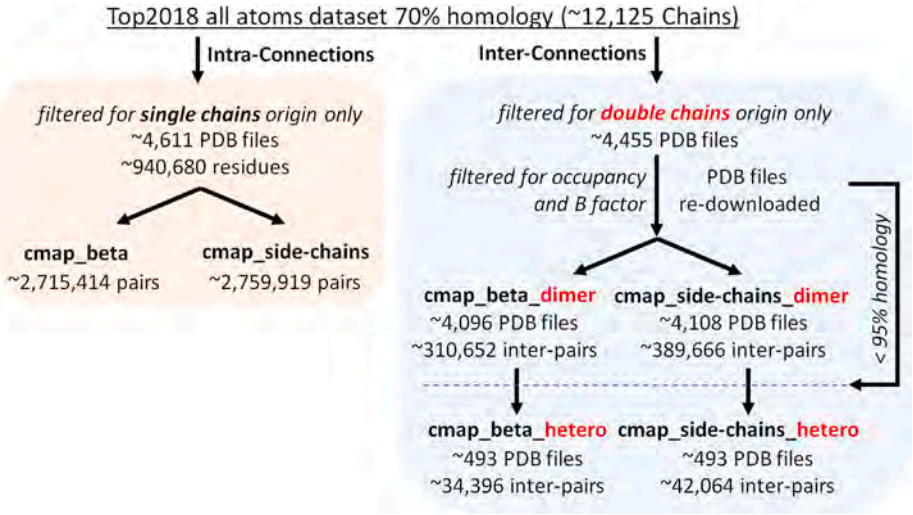


Fig. 2. Processing strategy employed for protein 3D structure analyses along with the corresponding number of identified residue-residue connections.

occurrence of residues and occurrence of connections of each amino acid in that particular pair. Previous literature also described relative frequency approach that computes frequency relative to the number of residues X, Y and to their number of connections [40]. Unfortunately, this simpler approach may not completely eliminate biases, especially if certain amino acids have significantly lower or significantly higher individual frequencies. The relative frequency of X|Y pair is calculated according to the following equation:

$$Relative\_Frequency^{X|Y}(RF) = \frac{C^{X|Y}}{R^X \times C^X \times R^Y \times C^Y}$$

$C^{X|Y}$ is the number of atomic contacts between X and Y, R^X is the percentage of occurrence of residue X in single chain dataset (Fig. 2), C^X is the percentage of number of atomic contacts made by residue X in relation to total number of atomic contacts, R^Y is the percentage of occurrence of residue Y in single chain dataset (Fig. 2), C^Y is the percentage of number of atomic contacts made by residue Y in relation to total number of atomic contacts.

The processing of PDB files and extraction of connections from atomic contact data was done using R language (The code is available online as described in the Availability of Data section). Briefly, a for() loop was used to process list of proteins using a sequence of functions using the Bio3D library [50]. Firstly, all variables were cleared at the start of the loop. Then, pdb files were read by read.pdb() function and saved in a dataframe. This function was embedded in a try() function to allow skipping analysis of current file due to complicated errors in a pdb (*e.g.*, duplicated atoms or consecutive residues). In the case of such errors, the loop would stop and would be manually restarted at the next pdb. Next, the pdb was cleaned using clean.pdb() function to remove water, ligands and hydrogens. Secondary structure data was collected using dssp() function. The pdb was trimmed according to atom selection (according to C_β atoms and side-chain atoms formats) using the following functions: atom.select(), combine.select(), and trim.pdb(). It is vital to keep only one representative heavy atom per residue, otherwise it will be challenging to estimate the correct number of connections. Contact map was collected via cmap() function and pairs were collected by converting the matrix into a list. The data were then parsed into one dataframe (*i.e.*, one dataset).

O/E ratios were directly calculated by dividing the observed frequency by the expected probability of each model. The expected probabilities were: 26/210, 83/210, 21/210, 52/210, 15/210, 13/210, 41/210, 109/210, 45/210, 27/210, 70/210, 50/210, and 20/210 for M-I 1, M-I 2, M-I 3, Chipens, R-B, Siemion, Fassina, Transmuted, GU, Shift1, Shift2, Shift3 and Shift4, respectively.

2.4. Other parameters

Protein size: Protein size was calculated based on chain length and categorized as S (small, ≤150 residues), M (medium, 151–300 residues) and L (large, >300 residues) groups.

Range: The range of connection between X and Y residues describes the pair proximity in the same chain sequence, calculated by difference in residue numbering, and categorized as S (short, ≤12 residues) and L (long, >12 residues). This cutoff for short range has been previously used in literature [10,51].

DSSP: Secondary structure information (DSSP is coded by 8 categories: α-helix, β-bridge, extended β-strand, 3−10 helix, pi helix, turn, bend, and coil) was extracted using the dssp() function of the bio3d package in R [50]. DSSP was subsequently converted to 3 coded categories to represent α-helix (H), extended β-strand (E) and the rest as coil (C). The DSSP pair were combined, alphabetically-sorted, and written in pair format in 6 categories: C|C, C|E, C|H, E|E, E|H, and H|H. The mkdssp version 2.0.4 program used here was a C++ adaptation by Maarten L. Hekkelman (https://github.com/ecapriotti/lb1-2/blob/master/dssp/last accessed in 24.01.2024), of the original source code written by Kabsch and Sander [52].

Sequence identity: Percent sequence identity for dimers was first obtained by extracting the sequences of both chains. Then the alignment and identity calculations were done using seqaln() and seqidentity() functions of the bio3d package which utilized the MUSCLE program. A 95 % threshold was used to distinguish homodimers from heterodimers. The MUSCLE version 3.8.31 program for Windows (https://drive5.com/muscle/downloads_v3.htm last accessed in 24.01.2024) was used here instead of later available versions owing to compatibility with bio3d package.

Hydrophobicity: Hydrophobicity scores for each pair (as 'XY' sequence) were calculated using hydrophobicity() function in Peptides R package [53]. Hydrophobicity scores table was described for amino acid pairs in the same previously mentioned X|Y format. Pearson correlation was used and tested via cor() and cor.test() functions in R, respectively.

Amino acid indices: Matrices for various physicochemical and biochemical properties of amino acid pairs were obtained from AAindex database (https://www.genome.jp/aaindex/last accessed in 24.01.2024) [42]. Only matrices that are symmetric were retained and converted from text to amino acid pair list in the same previously mentioned X|Y format.

2.5. Sensitivity analysis

To assess the robustness of the models, O/E ratios were calculated for three dataset collections: (1) Top2018 at 30 %, 50 % 70 % and 90 % homology datasets where distances between C_β atoms (and C_α atom in the case of Gly residue) or side-chain atoms are less than 8.0 Å and 5.0 Å, (2) CATH domain 20 % and 40 % homology datasets (http://download.cathdb.info/cath/releases/latest-release/last accessed in 09.12.2024) where distances between C_β atoms (and C_α atom in the case of Gly residue) or side-chain atoms are less than 8.0 Å, and (3) Orientations of Proteins in Membranes (OPM) database (https://opm.phar.umich.edu/ last accessed in 13.02.2025) where distances between C_β atoms (and C_α atom in the case of Gly residue) or side-chain atoms are less than 8.0 Å.

2.6. Molecular dynamics (MD)

All simulations were performed on GPU-accelerated workstation equipped with a 3.60 GHz six-core Intel® Core™ i5–8600 K CPU paired with a GeForce GTX 1080Ti 11 Gb graphics card. The software framework includes C/C++/CUDA development tools such as Nsight Eclipse Edition 11.2 from Nvidia in Santa Clara, CA, USA. Additionally, MPI (Message Passing Interface) and OpenMP (Open Multi-Processing) are enabled for parallel computing. All simulations are conducted with single precision, employing one MPI thread and six OpenMP threads for computational tasks.

Restrained MD was used in Gromacs (version 23) for demonstration of the role of predetermined connections in accelerating folding. The crystal 3D structure of (PDB ID: 2jku) was used without trimming. Extended form of the 3D structure was built in UCSF Chimera (version 1.16) via Tools > StructureEditing > BuildStructure > StartStructure > addpeptide modules. Using same sequence and with parallel β strand ($\varphi = -119°$ and $\psi = 113°$) and Dunbrack 2010 rotamer library. All simulations were performed under AMBER99SB force field (Table S8). The structure was placed in the centre of a decahedron box. Subsequently, the box was filled with water using the spc216 model and the system was made neutral using the Genion tool (2 Na^+ ions were added) of the Gromacs package with maximum 1000 steps to perform minimization using steepest descent.

Two steps of equilibration were performed, firstly in NVT system (constant number of particles, volume, and temperature) and secondly in NPT system (constant number of particles, pressure and temperature) as summarised in Table S8. Briefly, each equilibration was run for 100 ps with a time step of 2 fs using leap-frog integrator (short range electrostatic and Van der Waals cut-off value were 1.4 nm) with LINCS algorithm for hydrogen bonding constraints. The cutoff-scheme was

performed using Verlet for buffered neighbour grid searching (short range electrostatic and Van der Waals cut-off values were 1.0 nm). Particle Mesh Ewald (PME) was used for handling the long-range electrostatics in the periodic boundary conditions (cubic interpolation was used with Fast Fourier Transform (FFT) grid spacing at 0.16). Parrinello-Rahman pressure coupling was used only for the NPT equilibration with 1 bar reference pressure and compressibility of 4.5×10^{-5} bar^{-1}. The equilibration was maintained under modified Berendsen thermostat (tcoupl = V-rescale) with reference temperature around 300 K.

The MD production (Table S8) was run with a time step of 2 fs using leap-frog integrator (short range electrostatic and Van der Waals cut-off value were 1.2 nm) with LINCS algorithm for hydrogen bonding constraints. PME was used for handling the long-range electrostatics in the periodic boundary conditions (cubic interpolation was used with FFT grid spacing at 0.16). The system was maintained under modified Berendsen thermostat (tcoupl = V-rescale) with reference temperature around 300 K. Parrinello-Rahman pressure coupling was used with 1 bar reference pressure and compressibility of 4.5×10^{-5} bar^{-1}.

2.7. Statistical analysis

The O/E ratio was used to assist in the interpretation of the subsequent tests; however, it was not subjected to statistical testing due to the impracticality of analysing large categorical datasets. Significance of difference in mean rank frequency between connection and non-connection was tested using Mann-Whitney test for all models (except for Transmuted where Kruskal-Wallis test was used). Pearson's correlation was calculated for all models vs. other variables (frequencies, hydrophobicity scores, amino acids indices, dynamoeomics residue preference and all models). The p-value levels were considered significant at <0.05 and adjusted significant at <0.004 (*i.e.*, for alpha divided by 13 tested models).

3. Results

In this work, we have tested seven historic pairing models from the literature and six new pairing models that we have developed, all of which were designed to incorporate genetic principles supplemented by relevant biophysical principles for the purpose of explaining how amino acid residue-residue connections are formed amongst the chains of the protein 3D structures (both intra-and inter-connections). The performance of all these models was statistically tested using datasets of connection frequencies through: (1) the mean rank occurrence of connections in the model compared to non-connections (which can also be visualized using O/E ratios), and (2) the correlation between the model predictions and the observed frequency of connections.

3.1. Intra-connections

Overall, the GU, M-I 2, Transmuted and Shift2 pairing models were the most frequently observed intra-connections as compared to non-connections ($p < 0.004$) followed by M-I 1 ($p < 0.05$). In cmap_beta and cmap side-chains intra-connections, the significantly increased occurrence resulted in relatively high O/E ratios for the F and NF frequencies in GU, M-I 2, Transmuted, Shift2, and M-I 1 (Table 1). Surprisingly, RF frequencies were low for most pairing models and even significantly so despite the high observed occurrences and O/E ratios for both F and NF frequencies (Table 1). On the other hand, O/E ratios for NF frequencies were relatively higher than most reported O/E ratios of F frequencies of well performing models, indicating that normalization has improved the performance of well-performing pairing models and worsened the performance of most poorly-performing pairing models.

On taking a closer look at various groups of intra-connections, increased occurrence and high O/E performance was similar to the overall performance of GU, M-I 2, Transmuted and Shift2 pairing models according to range of connections ($p < 0.004$), secondary structure pairing ($p < 0.004$), and protein size ($p < 0.004$), in both cmap_beta (Table S9) and cmap_side-chains (Table S10) intra-connections.

Among the top 50 ranking amino acid residue-residue connections according to NF in cmap_beta, respective pairing models were observed as follows: 10 M-I 1 (20 %), 33 M-I 2 (66 %), 3 M-I 3 (6 %), 14 Chipens (28 %), 5 R-B (10 %), 11 Fassina (22 %), 43 Transmuted (86 %), 25 GU (50 %), 5 Shift1 (10 %), 22 Shift2 (44 %), 13 Shift3 (26 %) and 6 Shift4 (12 %) (Table S1). Again, GU, M-I 2, Transmuted and Shift2 pairing models were the most frequently observed.

When ranking amino acid residue-residue connections according to NF, we observed a steep decline for the first 20 top ranked connections, followed by a generally linear decline, which was correlated with F but not with RF frequency data (Fig. 3A–B). In fact, the highest ranked connections according to RF frequency data were clearly ranked lowest according to NF frequency. Otherwise, M-I 1 and M-I 2 models' connections were found to gravitate towards the first half of the NF ranking in all datasets, in contrast to connections of the M-I 3 model (Fig. 4A and C). Furthermore, both Transmuted and GU pairing models also gravitated towards the first half of the NF ranking in all datasets, to the greatest extent (Fig. 4B and D). On the other hand, Chipens, RB, Siemion and Fassina pairing models gravitated in the same way to much less of an extent (Fig. 4A and C), as did all the Shift pairing models (Fig. 4B and D).

The highest positive correlations were found between GU, Transmuted, M-I 2, and Shift2 pairing models across all connection datasets in F and NF datasets (Fig. 5). An intermediate positive correlation was found between cmap_beta (NF) and GU ($R^2 = 0.492$, $p < 0.004$), Transmuted ($R^2 = 0.334$, $p < 0.004$), M-I 2 ($R^2 = 0.289$, $p < 0.004$), and Shift2 ($R^2 = 0.179$, $p < 0.05$). RF frequency data gave only negative or weak correlations with GU ($R^2 = -0.195$, $p < 0.05$), Transmuted ($R^2 = -0.238$, $p < 0.004$), M-I 2 ($R^2 = -0.199$, $p < 0.004$), and Shift2 ($R^2 = -0.105$, $p > 0.05$).

3.2. Inter-connections

Similar to intra-connections, GU, M-I 2, Transmuted and Shift2 models were the most frequently observed inter-connections as compared to non-connections ($p < 0.004$) (Table 1). The same was observed across all secondary structure combinations in all four inter-connection datasets ($p < 0.004$) (Table S11). All four inter-connection datasets showed gradual albeit similar ranking to the cmap_beta NF with few noticeable exceptions (Fig. 3C and D). In the side-chains_dimer dataset, the connections E|R, D|R and G|R were ranked higher than the anticipated rank and thus showed as peaks in the curve (Fig. 3C). Similar observation was found in the side-chains_hetero dataset with peaks at F|L, E|R, D|R and D|K (Fig. 3D). The highest positive correlations were found between GU ($R^2 > 0.364$, $p < 0.004$), Transmuted ($R^2 > 0.363$, $p < 0.004$), M-I 2 ($R^2 > 0.288$, $p < 0.004$), and Shift2 ($R^2 > 0.247$, $p < 0.004$) pairing models across all four inter-connection datasets (Fig. 5).

3.3. Hydrophobicity

With regard to hydrophobicity, the GU and Fassina pairing models were expected to show a strong correlation. However, whilst the GU pairing model exhibited a strong positive correlation ($R^2 = 0.441$, $p < 0.004$) with the Kyte-Doolittle scale, the Fassina pairing model actually exhibited a less positive correlation ($R^2 = 0.156$, $p < 0.05$) followed by Chipens ($R^2 = 0.137$, $p < 0.05$), Shift3 ($R^2 = 0.096$, $p > 0.05$) and Transmuted ($R^2 = 0.096$, $p > 0.05$) pairing models (see Fig. S1 and Tables S12 and S13).

3.4. Amino acid indices

By searching for novel relationships among the symmetric matrices of amino acids indices, we identified various correlations with pairing models and residue-residue connections (Fig. S2 and Tables S14 and S15). Indices MIYS960103 and TANS760102 displayed the most

Table 1
Comparison of frequencies of residue-residue connections in different models in tested datasets. Significance is highlighted when frequency of observed connections is higher than that of observed non-connections.

Dataset		M-I 1	M-I 2	M-I 3	Chipens	R-B	Siemion	Fassina	Transmuted	GU	Shift1	Shift2	Shift3	Shift4
Observed Frequency														
cmap_beta	F	14.8 %[a]	53.4 %[b]	7.7 %	26.1 %	6.7 %	6.2 %	20.4 %	68.2 %[b]	40.4 %[b]	11.4 %[a]	42.8 %[b]	23.5 %	9.4 %
	NF	13.5 %[a]	69.9 %[b]	4.0 %	23.2 %	4.1 %	3.9 %	17.4 %	83.4 %[b]	64.8 %[b]	17.4 %	51.4 %[b]	18.1 %	9.9 %
	RF	7.5 %[a]	25.3 %[b]	8.2 %	17.6 %	4.8 %	3.2 %	16.4 %[a]	32.8 %[b]	9.8 %[b]	20.0 %	26.4 %[a]	16.9 %	11.1 %
cmap_side-chain	F	13.9 %[a]	53.7 %[b]	6.8 %	24.9 %	6.0 %	5.5 %	18.8 %	67.6 %[b]	41.4 %[b]	12.6 %	44.1 %[b]	21.7 %	9.0 %
	NF	12.8 %[a]	69.9 %[b]	3.4 %	22.7 %	3.6 %	3.4 %	16.7 %	82.6 %[b]	66.0 %[b]	18.9 %	53.1 %[b]	16.8 %	9.4 %
	RF	7.1 %[a]	25.4 %[b]	7.3 %	16.3 %[a]	4.5 %	3.0 %[a]	15.3 %[b]	32.4 %[b]	9.8 %[b]	20.8 %	27.5 %	15.6 %[a]	10.6 %
cmap_beta_dimer	F	15.4 %	50.7 %[b]	7.0 %[a]	25.8 %	7.4 %	6.8 %	19.3 %	66.1 %[b]	32.7 %[b]	13.2 %	41.2 %[b]	23.9 %	10.1 %
cmap_side-chain_dimer	F	14.2 %	49.0 %[b]	7.0 %[a]	24.3 %	7.0 %	6.5 %	18.0 %	63.2 %[b]	31.2 %[b]	13.7 %	41.7 %[b]	22.5 %	9.5 %
cmap_beta_hetero_dimer	F	15.5 %[a]	49.0 %[b]	7.6 %	26.4 %	7.5 %	7.0 %	20.3 %	64.5 %[b]	31.1 %[b]	10.0 %[a]	41.0 %[b]	24.4 %	9.6 %
cmap_side-chain_hetero_dimer	F	14.0 %	47.8 %[b]	7.5 %[a]	24.7 %	6.9 %	6.4 %	18.6 %	61.9 %[b]	30.2 %[b]	10.7 %[a]	41.8 %[b]	22.8 %	8.9 %
Expected Frequency (expected connections of each model divided by total 210 possibilities)														
		12.4 %	**39.5 %**	**10.0 %**	**24.8 %**	**7.1 %**	**6.2 %**	**19.5 %**	**51.9 %**	**21.4 %**	**12.9 %**	**33.3 %**	**23.8 %**	**9.5 %**
O/E														
cmap_beta	F	119.8 %	135.1 %	77.0 %	105.4 %	93.3 %	100.4 %	104.6 %	131.4 %	188.4 %	88.9 %	128.5 %	98.6 %	98.7 %
	NF	109.0 %	176.8 %	39.8 %	93.6 %	57.2 %	62.6 %	89.0 %	160.7 %	302.5 %	135.2 %	154.3 %	75.9 %	103.9 %
	RF	60.5 %	64.0 %	81.6 %	71.1 %	67.1 %	52.3 %	84.0 %	63.2 %	45.7 %	155.2 %	79.1 %	70.8 %	116.4 %
cmap_side-chain	F	112.5 %	135.9 %	68.1 %	100.7 %	84.5 %	88.8 %	96.3 %	130.3 %	193.4 %	97.7 %	132.3 %	91.2 %	94.3 %
	NF	103.1 %	176.8 %	34.4 %	91.7 %	50.7 %	54.7 %	85.4 %	159.2 %	308.1 %	147.2 %	159.4 %	70.7 %	99.2 %
	RF	57.1 %	64.2 %	73.5 %	65.8 %	62.9 %	48.6 %	78.3 %	62.5 %	45.9 %	162.0 %	82.4 %	65.6 %	111.3 %
cmap_beta_dimer	F	124.3 %	128.4 %	70.1 %	104.1 %	103.5 %	110.6 %	98.9 %	127.4 %	152.7 %	103.0 %	123.6 %	100.5 %	106.0 %
cmap_side-chain_dimer	F	115.1 %	123.9 %	70.0 %	98.3 %	98.2 %	104.6 %	92.2 %	121.8 %	145.4 %	106.2 %	125.1 %	94.7 %	99.5 %
cmap_beta_hetero_dimer	F	124.9 %	124.1 %	76.0 %	106.7 %	105.0 %	113.4 %	103.8 %	124.3 %	145.1 %	77.5 %	122.9 %	102.5 %	100.5 %
cmap_side-chain_hetero_dimer	F	113.2 %	121.1 %	75.4 %	99.8 %	96.4 %	103.1 %	95.5 %	119.2 %	141.1 %	83.5 %	125.3 %	95.7 %	93.0 %

[a] Significant at $p < 0.05$.
[b] Significant at $p < 0.004$.

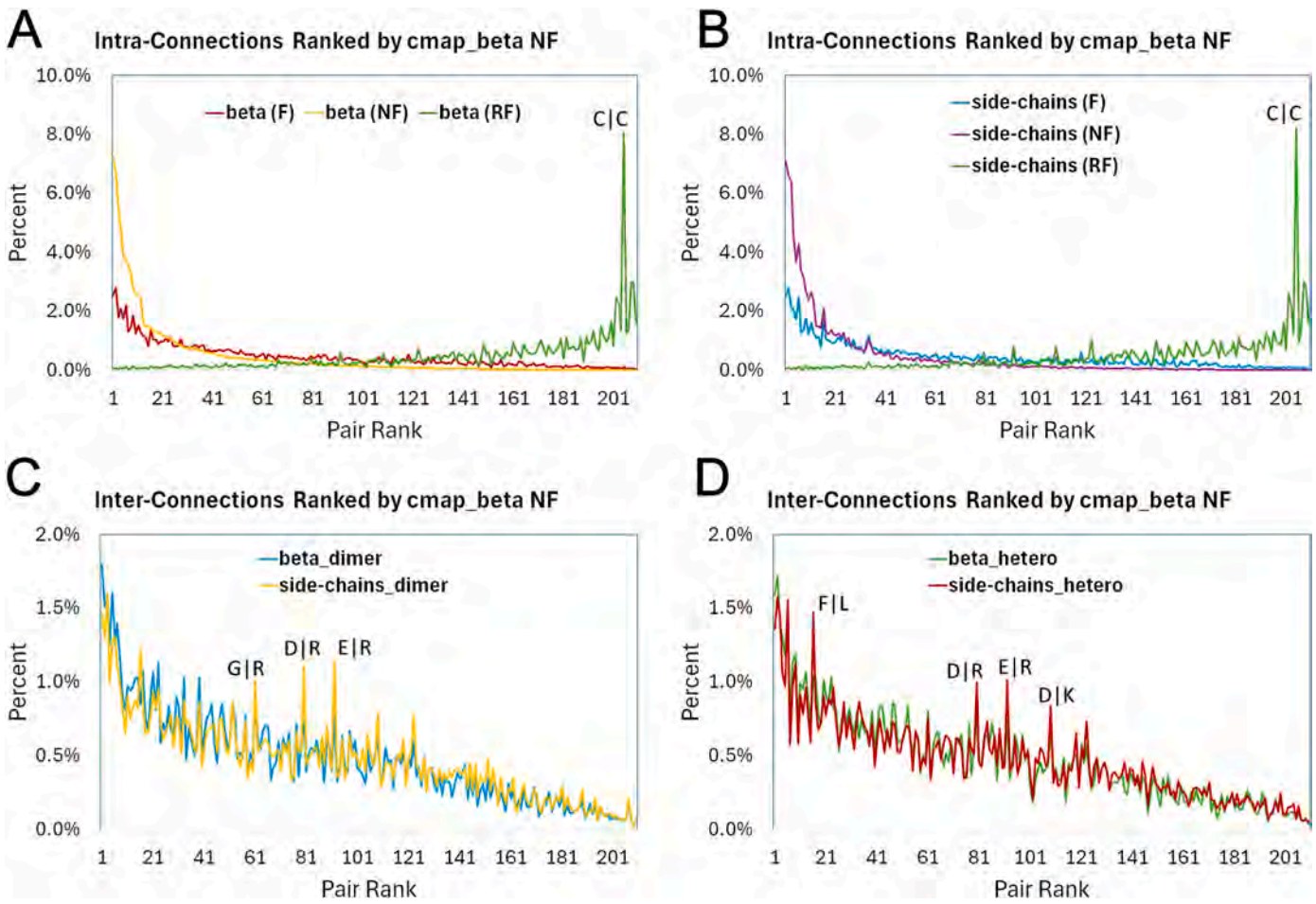


Fig. 3. Ranking of residue-residue connections according to cmap_beta NF in datasets. (A) Percent intra-connections of cmap_beta dataset ranked by cmap_beta NF showing steep decline in the F and NF top 20 connections. Intra-connections of beta (RF) showed highest percentage in the last 20 connections with C|C as highest. **(B)** Percent intra-connections of cmap_side-chains dataset ranked by cmap_beta NF showing steep decline in the NF top 20 connections. Intra-connections of side-chains (RF) showed highest percentage in the last 20 connections with C|C as highest. **(C)** Percent inter-connections from dimer datasets ranked by cmap_beta NF and showing similar yet more linear ranking to cmap_beta NF. A few connections like E|R, D|R and G|R showed higher percentages in side-chains than beta_dimer and relative to their ranking by cmap_beta NF. **(D)** Percent inter-connections from hetero dimer datasets ranked by cmap_beta NF and showing similar yet more linear ranking to cmap_beta NF. A few connections like F|L, E|R, D|R and D|K showed higher percentages in side-chains_hetero than beta_hetero and relative to their ranking by cmap_beta NF.

interesting correlations with GU, Transmuted, M-I 2 and Shift2 pairing models, although they were based on a previous connection frequency analysis. Index RISJ880101 (also a substitution matrix derived from structural data) was developed according to the observed side-chains interactions in only four groups: (i) Ile and Val, (ii) Leu and Met, (iii) Lys, Arg and Gln, and (iv) Tyr and Phe [54]. Interestingly, indices that were highly correlated with Transmuted connections also included BONM030104 (based on distances between centres of interacting side chains in the antiparallel orientation) where $R^2 = -0.330$ and $p < 0.004$, and also BONM030105 (based on distances between centres of interacting side chains in the intermediate orientation) where $R^2 = -0.302$ and $p < 0.004$ [55]. These correlations are in line with the higher correlations of Transmuted pairing model with side-chain residue-residue connections particularly in β-sheets. The remaining indices correlated with MI-3, Chipens, Fassina, and Shift3 pairing models in similar ways, primarily with MIYS960102 [56], MIYS990107 [57], GODA950101 [58], MIYS930101 [59], SIMK990102 [60], and SIMK990103 [60] all based on connection frequency or energy. Other indices were focused on residue interactions in highly conserved domains such as AZAE970101 [61], AZAE970102 [61] and NAOD960101 [62].

3.5. Dynameomics connection preference

Negative correlations between the Dynameomics simulation connection preference and M-I 2 ($R^2 = -0.106$, $p > 0.05$), Shift1 ($R^2 = -0.206$, $p < 0.05$) and Shift2 ($R^2 = -0.175$, $p < 0.05$) pairing models were observed indicating that these pairing models have similar connection preferences to the simulation (Fig. S3). With the exception of Transmuted, GU and Shift4 pairing models (which exhibited near zero correlations, $p > 0.05$), the remaining pairing models exhibited significant positive correlations ($p < 0.05$) suggesting a reversed connection preference between these models and the simulation. Overall, the Dynameomics simulation connection preference only exhibited clear support for the Shift1 and Shift2 pairing models.

3.6. Relationships between pairing models

The correlations and overlaps between different models were used to shed some light on common origins and limitations of pairing models. Negative correlations such as those between M-I 1, M-I 2, and M-I 3 pairing model with the Chipens model are simply products of non-overlapping pairs between these models (Fig. S4). The M-I 1 pairing model correlates nicely with Shift3, Chipens, Siemion, RB, and Fassina

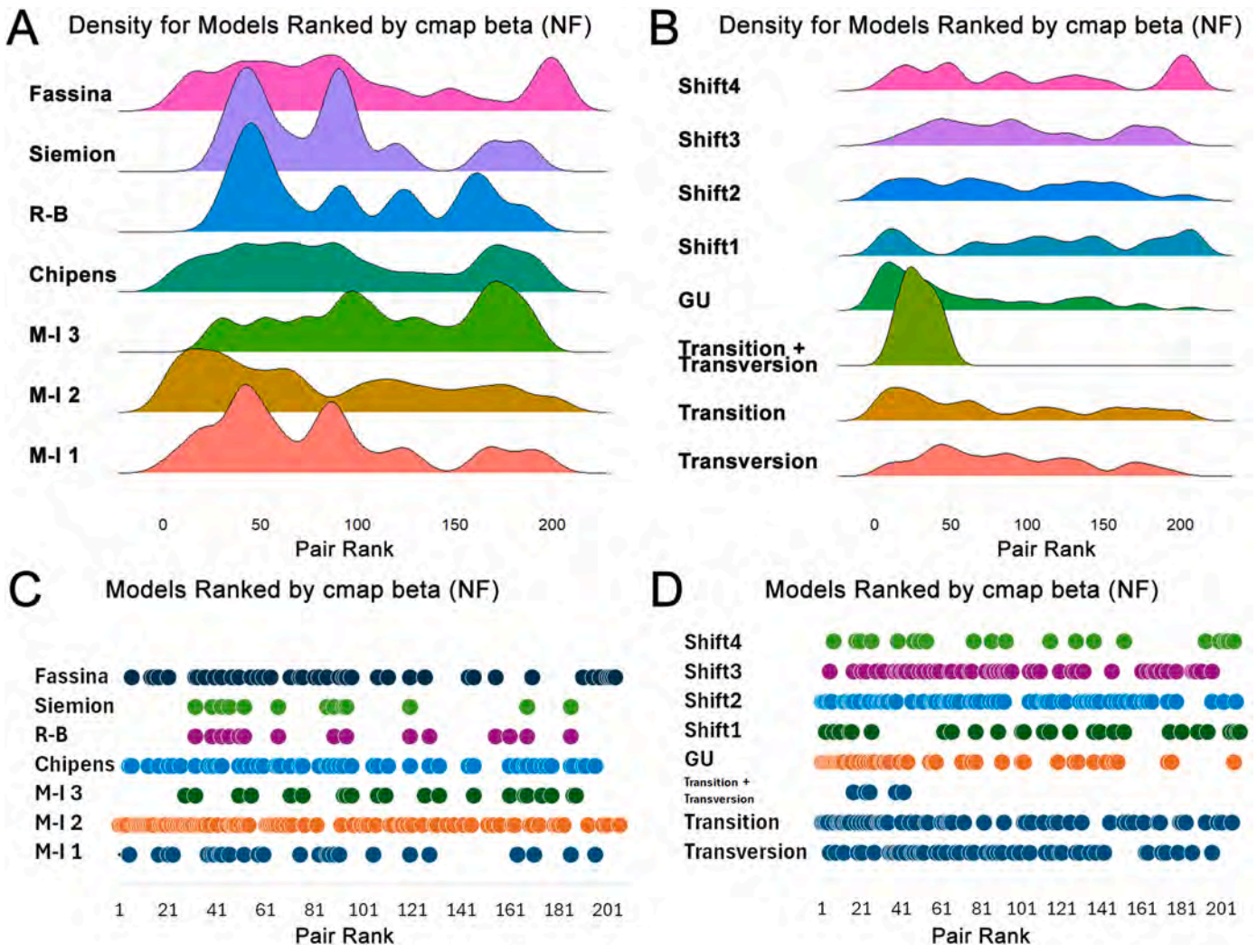


Fig. 4. Ranking of residue-residue connections for each pairing model according to cmap_beta NF in datasets. (A) Density of residue-residue connections in models in literature ranked by cmap_beta (NF). Noticeable skewing of curve to the left with minor peaks in the far right is mostly visible in M-I 1 and M-I 2 models. **(B)** Density of connections in our new models ranked by cmap_beta (NF). A noticeable leftward skewing of the curve with minor peaks on the far right is visible, primarily in GU, Transition, and Transition + Transversion pairing models. **(C)** Plot of individual residue-residue connections in literature models ranked by cmap_beta (NF). The widest distribution is mostly visible in M-I 2, Chipens and Fassina pairing models. **(C)** Plot of individual residue-residue connections for our new pairing models ranked by cmap_beta (NF). The most leftward skewing of the distribution is visible primarily with GU, Transition and Transition + Transversion models.

	M-I 1	M-I 2	M-I 3	Chipens	R-B	Siemion	Fassina	Transmuted	GU	Shift1	Shift2	Shift3	Shift4
beta (F)	0.081	0.308 **	-0.083	0.034	-0.020	0.001	0.024	0.381 **	0.502 **	-0.047	0.219 **	-0.008	-0.004
beta (NF)	0.016	0.289 **	-0.093	-0.017	-0.055	-0.045	-0.025	0.334 **	0.492 **	0.063	0.179 *	-0.063	0.006
beta (RF)	-0.102	-0.199 **	-0.042	-0.114	-0.063	-0.084	-0.054	-0.238 **	-0.195 *	0.146 *	-0.102	-0.112	0.036
side-chains (F)	0.050	0.306 **	-0.112	0.004	-0.045	-0.030	-0.019	0.375 **	0.514 **	-0.009	0.241 **	-0.052	-0.020
side-chains (NF)	0.005	0.285 **	-0.100	-0.022	-0.063	-0.053	-0.033	0.329 **	0.499 **	0.083	0.193 *	-0.075	-0.001
side-chains (RF)	-0.110	-0.197 *	-0.060	-0.134	-0.070	-0.090	-0.073	-0.238 **	-0.192 *	0.162 *	-0.085	-0.131	0.025
beta_dimer	0.135	0.340 **	-0.148 *	0.035	0.014	0.040	-0.008	0.447 **	0.408 **	0.017	0.247 **	0.004	0.029
side-chains_dimer	0.095	0.323 **	-0.167 *	-0.017	-0.008	0.020	-0.065	0.408 **	0.397 **	0.040	0.296 **	-0.050	-0.003
beta_hetero	0.147 *	0.306 **	-0.126	0.060	0.022	0.054	0.029	0.412 **	0.370 **	-0.136 *	0.254 **	0.022	0.003
side-chains_hetero	0.084	0.288 **	-0.139 *	-0.002	-0.017	0.014	-0.038	0.363 **	0.364 **	-0.108	0.304 **	-0.041	-0.039

Fig. 5. Correlations between pairing models and residue-residue connection frequencies in datasets. Highest positive correlations using direct and normalised frequencies (F and NF, respectively) were observed with GU, Transmuted, M-I 2, and Shift2 pairing models. Relative frequency (RF) data exhibited negative or low correlations in most pairing models. The p-value levels were considered significant at <0.05 (*) and adjusted significant at <0.004 (**).

pairing models (exhibiting correlation coefficient values of $R^2 = 0.672$, 0.622, 0.383, 0.345, and 0.216, respectively, and values of $p < 0.004$ for these models). On the other hand, the GU pairing model exhibited the weakest correlations with almost all the other models except the Shift1 model ($R^2 = 0.215$, $p < 0.05$). On the other hand, Shift3 showed the strongest, most significant correlation with all the historic pairing models at $p < 0.004$ (Fig. S4).

In reviewing overlaps, the GU pairing model was found to share 21, 5 and 5 connections with M-I 2, M-I 1 and M-I 3 leaving 14 residue-residue connections unrepresented (Fig. 6A). Among the other pairing models, Chipens and Fassina were found to share 13 connections between each other (Fig. 6B). The Transmuted pairing model (created by combining

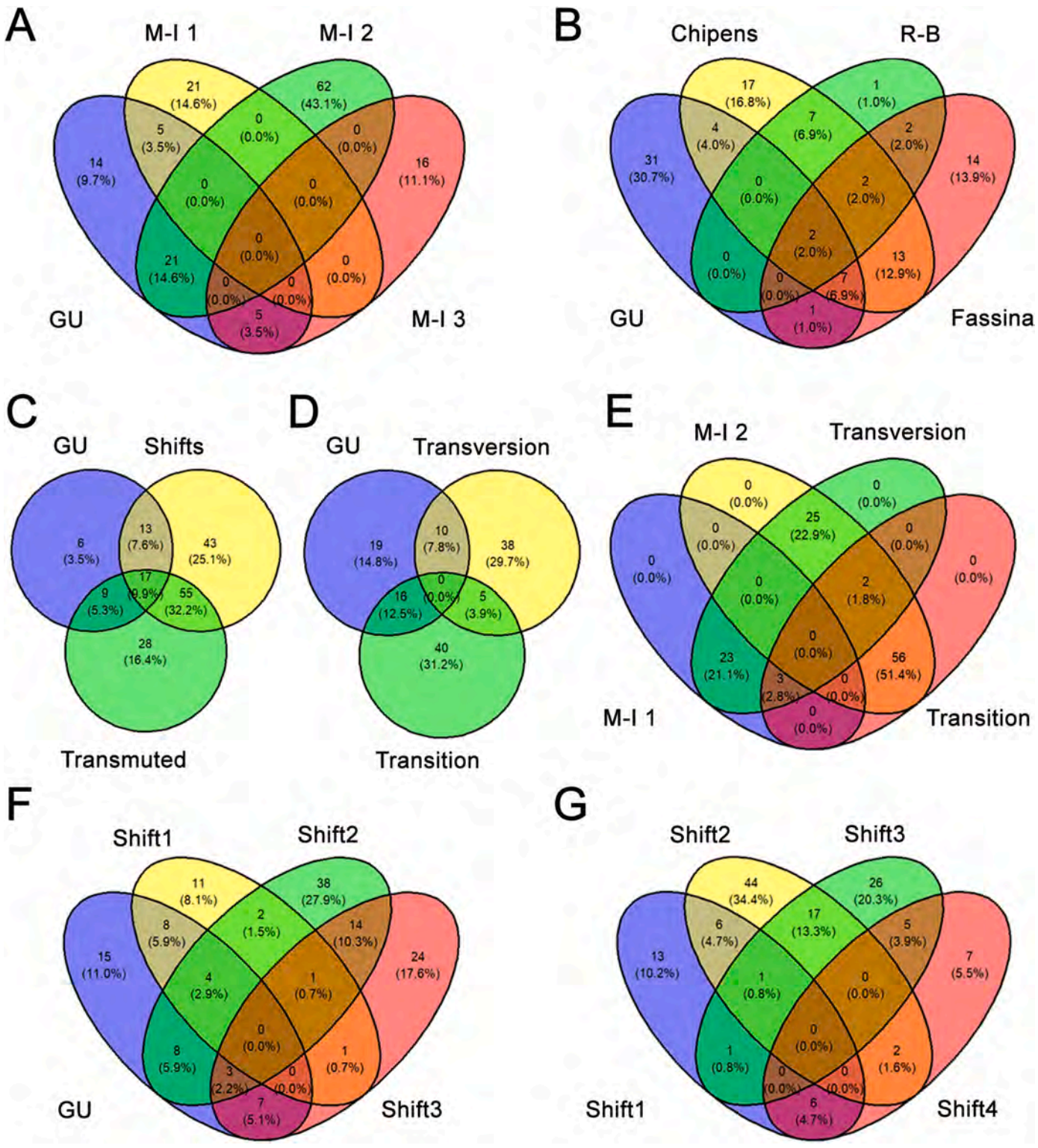


Fig. 6. Overlap between different pairing models. (A–B) Venn diagrams showing overlap between connections of the GU and historic pairing models. **(C–E)** Venn diagrams showing overlap between Transmuted and various other pairing models. **(F–G)** Venn diagrams showing overlap between the Shift pairing models.

M-I 1 and M-I 2) and Shift models share 39 residue-residue connections combined with the GU model (Fig. 6C). This highlights the fact that the GU, Transmuted and Shift2 pairing models, which are also the most statistically preferred, frequently overlap although not completely. Among the Transmuted residue-residue connections shared with the GU pairing model, there are 16 Transitions and 10 Transversions (Fig. 6D). Further deconstructing the Transmuted residue-residue connections, 56 (51.4 %) of Transitions belong to the M-I 2 pairing model, whilst two residue-residue connections are Transitions/Transversions (Fig. 6E). On the other hand, nearly 25 (22.9 %) of Transversions belong to the M-I 2 pairing model, whereas nearly 23 (21.1 %) of Transversions belong to the M-I 1 pairing model, leaving 3 residue-residue connections shared with the M-I 1 pairing model as Transitions/Transversions (Fig. 6E). With regard to the Shift pairing models, Shift1 and Shift2 pairing models

share 23 residue-residue connections with the GU pairing model, leaving 22 GU residue-residue connections unrepresented by either Shift1 or Shift2 models (Fig. 6F). Interestingly, the Shift2 pairing model exhibits the most unique connections followed by Shift3, Shift1 and Shift4 models, respectively (Fig. 6G). Overall, the clear presence of wholly unique connections and connections overlapping with one or more models, implies that a combination of redundancy and uniqueness exists within pairing models for reasons that should be discerned in the future.

3.7. Top most favoured and unfavoured residue-residue connections

Despite normalization, the top 50 most favoured intra-connections based on cmap_beta frequency (NF) ranking, were generally represented by the most frequent amino acid residues (Fig. 7A). Transition residue-residue connections seem particularly predominant, particularly the five self-connections, namely A|A, L|L, I|I, G|G, and V|V (Fig. 7A). Conversely, the top 50 most unfavoured intra-connections based on cmap_beta frequency (NF) ranking, were represented by the least frequent amino acid residues including six self-connections, namely W|W, C|C, M|M, H|H, K|K and Q|Q (Fig. 7B).

With respect to the top 50 most favoured inter-connections based on cmap_dimer_beta frequency ranking, a similar pattern was observed to the intra-connections with the addition of self-connection S|S (Fig. 7C). Once again, the top 50 most unfavoured inter-connections based on cmap_dimer_beta frequency ranking, also followed a similar pattern to the intra-connections, including self-connections C|C, W|W, H|H, M|M and K|K (Fig. 7D).

3.8. Sensitivity analysis

Testing the behaviour of the pairing models against different datasets provides indications as to their robustness. O/E ratios for intra- and inter-connections did not show strong divergence using different dataset collections including datasets based on other homology alignment-thresholds, alternative atomic contact distance cutoff of 5.0 Å, and also on membrane proteins (Fig. S6). O/E ratios for both GU and Transmuted pairing models remained high and dominant. For example, for C_β-based intra-connections, O/E ratios were 271–303 % and 147–169 % for GU and Transmuted pairing models, respectively. Similarly, for side-chain-based intra-connections, O/E ratios were 285–309 % and 149–159 % for GU and Transmuted pairing models, respectively. Furthermore, for C_β-based inter-connections involving dimers, O/E ratios were 150–160 % and 127–135 % for GU and Transmuted pairing models, respectively, whilst for side-chain-based inter-connections involving dimers, O/E ratios were 136–146 % and 116–122 % for GU and Transmuted pairing models, respectively.

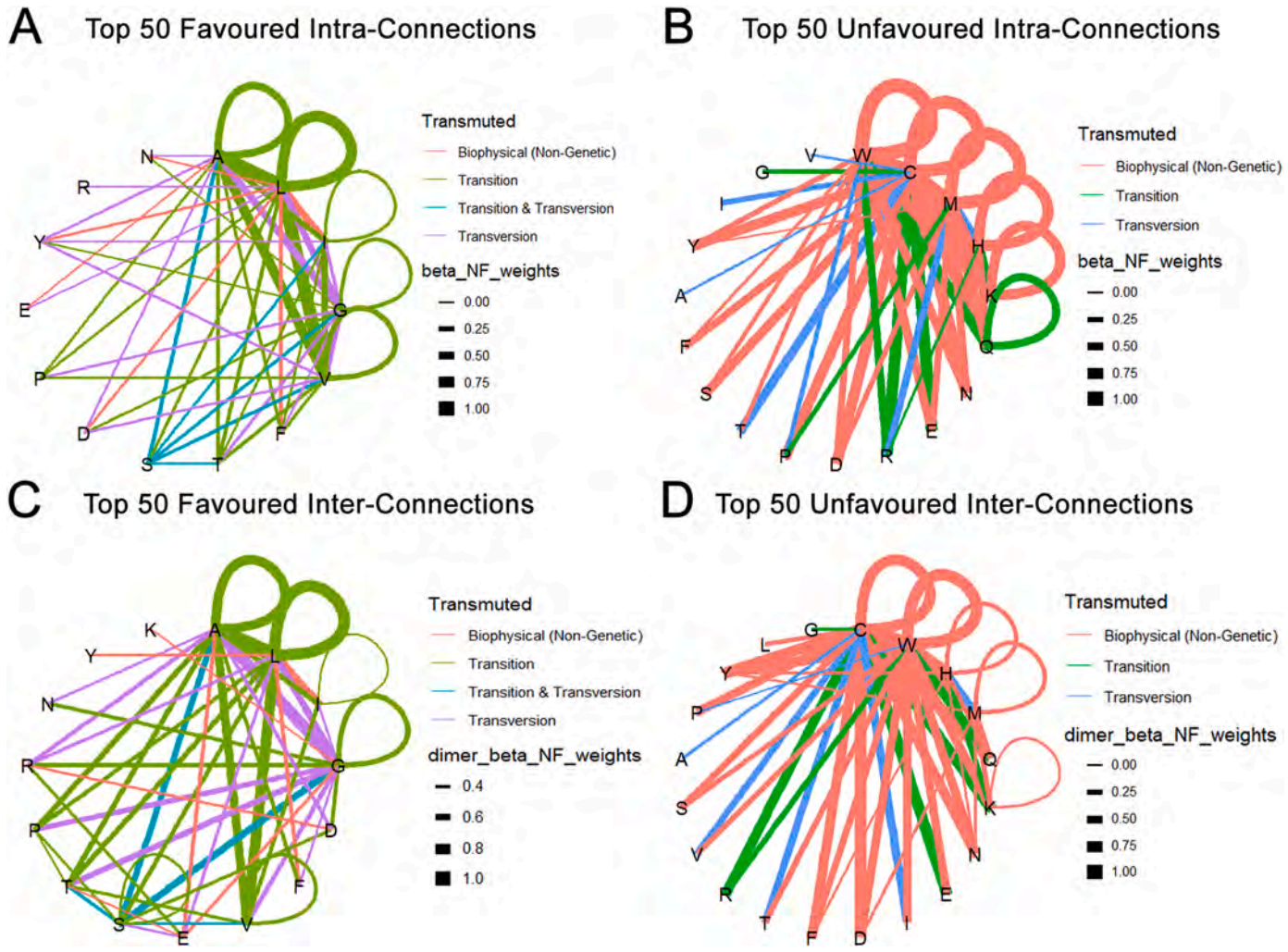
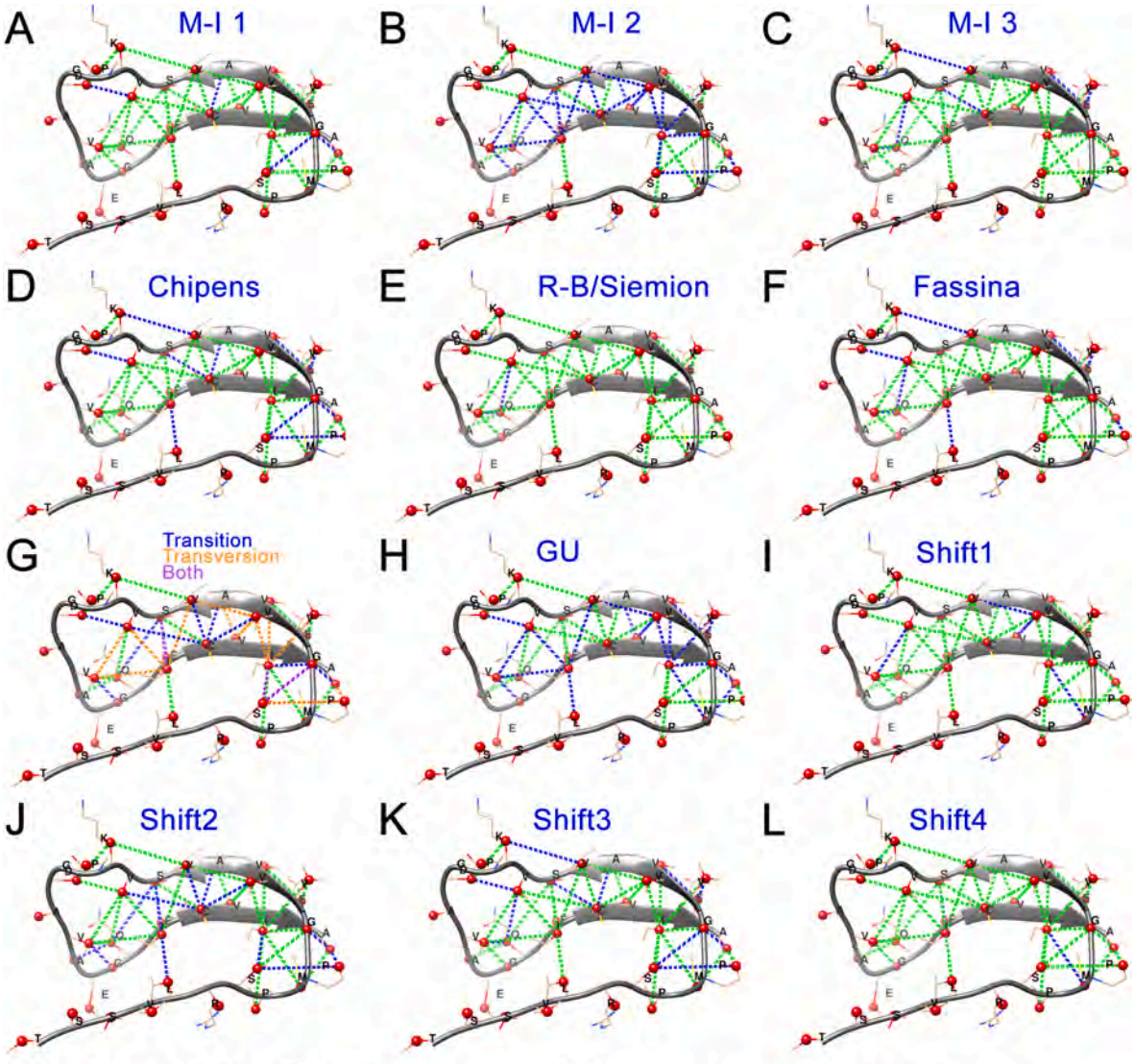


Fig. 7. Top favoured and unfavoured residue-residue connections in frequency rankings. (A) Top 50 most favoured intra-connections based on cmap_beta frequency (NF) ranking. **(B)** Top 50 most unfavoured intra-connections based on cmap_beta frequency (NF) ranking. **(C)** Top 50 most favoured inter-connections based on cmap_beta_dimer frequency (F) ranking. **(D)** Top 50 most unfavoured inter-connections based on cmap_beta_dimer frequency (F) ranking. Amino acid residues are identified by single letter abbreviations. Self-connections are indicated with loops. Line thickness increases with residue-residue connection frequencies.

3.9. Case study (restrained MD)

As an example of the distribution of residue-residue connections in protein 3D structure, a review of all the different pairing models (Cβ-based connections) found within the biotin acceptor domain of human propionyl-CoA carboxylase are illustrated (Fig. 8). With respect to the historic pairing models, the largest number of strategic connections (22 out of 40) were derived from the M-I 2 pairing model (Fig. 8A–F and 8N). With respect to the new models, the largest number of strategic connections (26 and 22, respectively) were from the Transmuted and GU



M Total 40 Intra-Connections in PDB (2JKU)

A|C	1	C|V	3	G|S	1	K|V	1
A|G	1	D|K	1	G|V	1	P|S	1
A|M	1	D|V	1	I|L	1	Q|S	1
A|P	1	E|G	1	I|M	1	Q|V	2
A|Q	1	E|S	1	I|P	1	V|V	4
A|V	1	E|V	3	I|S	2		
C|S	1	G|I	1	I|V	6		

N Model Distributions (Total 40 Intra-Connections)

M-I 1	M-I 2	M-I 3	Chipens	R-B/Siemion	Fassina
4	22	7	11	2	9
Transmuted	**GU**	**Shift1**	**Shift2**	**Shift3**	**Shift4**
26	22	5	15	12	2

Fig. 8. Examples of different models among beta carbon connections from PDB (2JKU) structure. (A–F) illustration of the presence of different historic pairing models (shown in blue) with the remaining contacts (shown in green). Out of the historic pairing models, residue-residue connections from the M-I 2 pairing model are dominant. Turning to the new pairing models; **(G)** illustration of the presence of the Transmuted pairing model, shown subdivided into Transitions (blue), Transversions (orange) and Transitions + Transversions (combined) (purple) connections, with the remaining contacts (shown in green); **(H-L)** illustration of the presence of GU and shift pairing models (shown in blue) with the remaining contacts (shown in green). Shown in **(M)** are a total of 40 intra-connections with I|V, V|V, C|V and E|V exhibiting the highest occurrence; illustrated in **(N)** is a distribution of the pairing models in the example structure indicating that Transmuted, GU, M-I 2 and Shift2 pairing models are the most frequently represented models amongst 40 identified residue-residue connections. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

pairing models (Fig. 8G-L and **8N**). The highest frequency residue-residue connections in the original crystal were 6 I|V, 4 V|V, 3 C|V, and 3 E|V (Fig. 8M). In order to demonstrate how this information can be useful in protein folding studies, MD simulations were then performed in triplicate using C_β contacts as restraints to guide and control the folding of this protein from an extended form to the final conformation (Fig. S5). All the restraints were specified by residue-residue connections predicted from the different pairing models. In doing this, M-I 2, Chipens, Transmuted, GU, Shift2 and Shift3 pairing models most effectively predicted the correct residue-residue connections relative to controls. Simulation of the original crystal structure fold without restraints stabilized in the RMSD range of 0.3–0.7 nm as compared to the crystal structure coordinates (Fig. S5A). These values describe the dynamic and flexible range of the RMSD for this protein in its soluble form without any restraints. An extended strand, also used as control without restraints, was able to attain a condensed form within 20 ns yet but still outside the RMSD threshold of 0.7 nm (Fig. S5B). The best performer from all pairing models was the GU model, which maintained a stable conformation within the 0.7 nm RMSD threshold in all three replicates (Fig. S5J). These findings show that using the GU residue-residue connections alone (representing 22 out of total 40 strategic residue-residue connections in the crystal structure) to specify the amino acid residue connections was sufficient to speed up the folding and maintain the predicted protein structure within the expected dynamic range.

4. Discussion

Protein contact maps, which offer simplified 2D illustrations of protein 3D structure, currently hold the key to the state-of-the-art accuracy of protein 3D structure prediction revolution [1,10]. Recent research has highlighted the role of both amino acid residue-residue intra-connections in protein folding prediction [4–6], and inter-connections in PPIs [7–9]. Given this, logic dictates that there must be some kind of underlying universal proteomic code: (1) at the core of protein folding, and (2) operating at protein-protein interaction interfaces.

Herein, we have focused on carrying out an extensive analysis of a novel framework of residue-residue connections derived from historic and new amino acid residue pairing models. Out of all these, the new Transmuted (M-I 1+M-I 2), GU and Shift1/Shift2 pairing models were the most significant and important. Otherwise, several other historic and new models in intra- and inter-connections were found much less significant, important and less effective, namely the M-I 3, Chipens, R-B, Siemion, Fassina, and Shift3/Shift4 pairing models, although their future relevance cannot be entirely ruled out for certain niche protein folding paradigms. At this point, we would like to re-emphasise that new GU pairing model is based on the genetic principle that all amino acid residues coded for by Gxx and xUx sense codons are capable of specific side-chain interactions with each other. In contrast, the new Shift1-4 pairing models are all based upon the genetic principle that amino acid residues coded for by sense codons are capable of specific side-chain interactions with corresponding amino acid residues specified by “Shift” NNx pairing/antisense codons. With the exception of the hydropathicity-inspired Fassina pairing model [33–36], all the other historic pairing models, namely M-I 1 [15], M-I 2 [16–19], M-I 3 [18,19], Chipens [20,27,28], Root-Bernstein [22,29], and Siemion [30–32], were based on the genetic principle that amino acid residues coded for by sense and corresponding complementary antisense codons are capable of specific side-chain interactions with each other both within and between proteins.

In order to reach an evaluation of all these different pairing models studied, a quantification of the atomic contacts was seen as the first challenge. This was resolved using two atomistic contacts (standard C_β 8 Å cutoff and side-chains 8 Å cutoff) and three frequency modes (direct, normalised, and relative frequencies). An 8 Å cut-off was used for both intra- and inter-connections. Duarte *et al.* [63] have explored contact maps based on C_α, C_β and C_α+C_β connection-types and found the optimal cut-off distances between 9 and 11 Å to be optimal for all the three connection types. However, the same study also showed that the region around an 8 Å cut-off better describes the distances between atoms in the range of the Van der Waals interactions. Alternatively, Kayikci *et al.* [64] analysed side-chain contacts making use of all side-chain atoms (with 4 Å cutoff) using a normalised weight to remove bias from residue frequency. In our analysis, we have taken a balanced perspective by considering both normalised and relative approaches to overcome occurrence biases in amino acid residue pairs X and Y and their connections. A normalised approach is more complex computationally but can balance out the effect of very low and very high frequencies of residues/contacts, giving a general overview of the data, whereas the relative approach is simpler, more sensitive to same amino acid residue-residue connections (*e.g.*, C|C, R|R, *etc.*), and more sensitive to significantly high/low frequencies of amino acid residues and contacts. We believe that the relative approach is more useful for describing unidirectional format pairing (*e.g.*, X→Y) compared to bidirectional pairing of the format used in this work (X↔Y, shown as X|Y for program scripting purposes) due to RF lower deviations. In this work, the standard deviations for frequencies in cmap_beta were 0.44 %, 1.03 % and 0.70 %, for F, NF and RF, respectively, and in cmap_side-chain were 0.45 %, 1.04 % and 0.70 % for F, NF, and RF, respectively. In our studies, RF top ranking pairs (C|C, C|W, C|M, M|W, C|H, H|W, W|W, and H|M) exhibited an extremely high bias towards amino acid residues of low-occurrence and consequently of limited reliability in offering up biorelevant explanations for intra- and inter-protein connections. Accordingly, the normalised approach (NF value) was chosen as the “gold standard” for our interpretations of connection frequency.

To assess the different pairing models, we employed frequency comparison, O/E ratios and correlations across six parsed datasets covering intra-connections, inter-connections of dimers (including both homodimers and heterodimers) and heterodimer inter-connections. Overall, the GU, M-I 2, Transmuted, Shift2, Shift1 and MI-1 pairing models exhibited the highest frequencies of occurrence (Table 1), best O/E performances (Table 1), and correlations (Fig. 5) across nearly all datasets, reflecting a universality in their representation across intra- and inter-connections. This may be because hydrophobic and small amino acid residues appear frequently among the top 50 most favoured amino acid residue pairs from all these models according to the NF in cmap_beta dataset (Table S1). This is visible in the steep slope comprising the 20 top ranking connections followed by linear decline (Fig. 3). Further evidence for the universality of these models is their similarity in performance over various parameters such as range between connections, secondary structure combinations in connections, and protein size (Table 1 and Tables S9–S10). These results taken together highlight the universality of the Transmuted (M-I 2/M-I 1), Shift2, and Shift1 pairing models across protein folding and PPIs, in common with the GU pairing model.

The top 12 most favoured amino acid residue pairs of the GU pairing model (namely: A|L, L|V, L|L, A|V, A|G, I|L, A|A, G|L, I|V, V|V, A|I and G|V) account for nearly 49.20 % of cumulative residue-residue connections according to NF in cmap_beta dataset (Table S1). This finding, combined with a strong positive correlation for the GU pairing model with the Kyte-Doolittle scale ($R^2 = 0.441$, $p < 0.004$), underscore how the GU pairing model ensures the recognition and inclusion of important biophysical as well as genetic principles in residue-residue connections (as noted above), most particularly with reference to hydrophobic side-chain/side-chain interactions which are of general importance as a means to provide stability to proteins in aqueous environments (with the notable exceptions of amyloid fibrils that are stabilized by backbone hydrogen bonding [65], or core-less proteins like plant defensins which are stabilized by cystine bridges [66]). Although high cystine frequencies have been observed in nearly half of the least favoured structures for GU and Transmuted pairing models (data not shown), caution is urged in interpreting pairing models that deviate substantially from

those associated with common amino acid residue frequencies.

In terms of the nature of our new pairing models, we note that the GU pairing model essentially excludes polar and charged amino acid residue-residue connections, but does nevertheless encompass key residue-residue connections found in protein cores. On the other hand, the Transmuted pairing model complements the GU pairing model nicely in encompassing a broader range of amino acid residue-residue connections with some preference given to Transition wobbles, in keeping with how point mutations may work at a mechanistic level. In addition, both Shift1 and Shift2 pairing models overlap with the GU model and provide access to residue-residue same self-connections. The Shift2 pairing model contributes to nearly 15 GU residue-residue connections while Shift1 contributes nearly 12 (Fig. 6F). Out of the top 12 GU residue-residue connections, 8 also belong to either the Shift1 or Shift2 pairing models (Table S1). Both Shift1 and Shift2 pairing models ensure that the genetic variations used to identify residue-residue connections are extended beyond point mutations to include indels within codons as well (only indels in first and third base can reproduce themselves in the rest of a reading frame). In general, both the historic and new pairing models all originate from sense strand DNA codons by definition. This fundamental could be tested experimentally by protein-protein interaction studies involving proteins coded for by the same sense strand ORFs. Multiple proteins coded for by the same sense strand ORFs are well documented at the transcriptional and translational levels in prokaryotes and viruses [67,68]. Another fundamental is that while the GU pairing model ensures the recognition of important biophysical principles in residue-residue connections, the Transmuted and Shift pairing models recognise the impact on potential residue-residue connections of genetic changes (including DNA duplication plus sense and antisense strand separation) plus evolutionary variations (associated with point mutations and indels).

An important latter part of our studies reported here was the analysis of membrane protein data sets. Membrane proteins are unique in their solubility given that they are “inside out” proteins with hierarchical structures suitable for differential interactions with bilayer membranes. Bilayer membranes comprise three distinct regions with clearly defined boundaries: the hydrophobic centre (25–35 Å depending on the length of hydrocarbon chains of the lipid), the tightly packed hydrophilic head groups on either side (each around 10–15 Å, creating pockets for aromatic amino acids), and the surrounding aqueous environments [69]. Folding and insertion of one or more protein domains (multipass) in the membrane is mediated by a transmembrane chaperone known as the translocon, although folding is obviously still controlled ultimately by factors such as Van der Waals forces and solvophobic exclusion. Whilst a transmembrane α-helix domain can exist in the range of 15–20 residues [69], distinct β-barrel transmembrane proteins also exist and follow a distinct folding mechanism. According to the budding model, β-barrel proteins transition between open and closed states. In the open state, the seam forms hydrogen bonds with substrate β-strands, facilitating β-sheet formation, enabling folded regions to pass through and integrate into the membrane. Closure occurs through the pairing of the *N*-terminal and *C*-terminal β-strands to form a continuous β-sheet [70]. Following an analysis of the OPM dataset of membrane proteins, we observed only minor deviations in the O/E ratios for GU, Transmuted and Shift1/Shift2 pairing models compared to ratios established with the TOP2018 dataset, even though these OPM-derived O/E ratios were generated using a much lower number of sampled structures (Fig. S6). This finding with membrane proteins is critical in terms of demonstrating the potential universality of our amino acid residue pairing models, as presented here, across all main protein classes in all organisms.

Obviously, protein stability is well-known to derive from a balance of the net forces and interactions between the protein functional groups and the environment. Two fundamental effects drive the folding process toward a stable conformation: the hydrophobic effect and complementary hydrogen bonding relationships. Experimental evidence underscores the necessity of both, offering valuable insights into protein folding. The spontaneous sequestration of nonpolar residues in the protein core facilitates backbone selection that ensures hydrogen bond satisfaction, ultimately leading to energetically stable conformations where polar residues remain exposed [71]. Additionally, the ribosomal channel plays a crucial role in the initial formation of alpha-helices, acting as a selection funnel for hydrogen bond satisfaction [72]. In a large experiment, Tsuboyama *et al.*, [73] studied various factors affecting protein folding stability: (1) intrinsic sequence and structure factors (*e.g.*, hydrophobic residue frequency, secondary structure promoting residues, packing/side-chain interactions and destabilizing mutations) (2) environmental and contextual factors (*e.g.*, pH, temperature, protease susceptibility), and (3) evolutionary/functional constraints (*e.g.*, trade-off between function and stability, thermodynamic coupling and energetic cooperativity between residues). Furthermore, certain protein designs can be considered such as disulfide bridges. Accordingly, these background observations and data support the need for amino acid residue-residue pairing models that can reasonably take account of this multitude of factors. Hence, in our opinion, the introduction of the GU, Transmuted and Shift1/Shift2 pairing models is an important next step away from historic towards new pairing models that could be better, more fundamental tools to link the standard genetic code and 3D protein structure/function.

Several studies have explored previously the nature of amino acid residue-residue interactions in datasets of experimentally determined protein 3D structures and attempted to identify the factors that influence these connections. Faure *et al.* [40], analysed a dataset of 1230 protein chains corresponding to 377,232 residues, and revealed variations in connections according to their proximity in the sequence, and according to protein size (which was attributed in smaller proteins to the fact that they possess amino acid residue frequencies slightly different from other proteins in the databank). Later, Esque *et al.* [41] analysed a dataset of 818 polypeptide chains corresponding to 187,433 residues, and revealed relationships between connections, residue volume and residue accessibility. Amala and Emerson [74], analysed a dataset of 768 protein structures belonging to the four structural classes (146, 196, 232, and 194 structures of α, β, α/β, and α+β, respectively) which confirmed the influence of diagonal motif variations on the contact maps. Cieslik and Derewenda [75], studied amino acid residue binding interactions by classifying residues into five groups, *i.e.*, aromatic (His, Phe, Trp, and Tyr), aliphatic (Leu, Ile, and Val), small (Ala, Gly, Ser, and Thr), charged (Arg, Asp, Glu, and Lys) and other (Asn, Gln, Met, and Pro). Their findings showed that charged residues are predominantly located on the surface, whereas small and aliphatic amino acids are often partly immersed, or are more predominant in the core, isolated from polar residues. Interestingly, one previous study was focused on the complementary amino acid residue-residue pairing in 82 PPIs and showed that such residue-residue connections form in clusters [76]. Unfortunately, these studies have only provided a global idea about amino acid residue-residue connections, often just describing the redundancies of amino acid residue properties, and hence are not very applicable to state-of-the-art predictions of protein 3D structure and PPIs.

With regards to PPIs, these may be categorized based on their localization within the cell and their binding affinity. Localization dictates functional roles, with interactions occurring in the cytoplasm, nucleus, or membranes, influencing processes like signalling, enzymatic regulation, and structural organization. Binding affinity, quantified by ΔG_{bind} (free energy change of binding), helps differentiate between transient and obligate interactions. Transient interactions, such as those involved in signalling pathways or enzymatic reactions, exhibit high negative values of ΔG_{bind} and are characterized by rapid association (on) and dissociation (off) rates. In contrast, obligate interactions exhibit comparably negative values of ΔG_{bind} associated with less rapid on and off rates, and hence play essential roles in structural integrity and multi-protein assemblies [77,78]. Clearly, new protein-protein interaction datasets that are well classified according to these two parameters (*i.e.*, ΔG_{bind} and localization) may also require other potential parameters

that take account of the nature and type of PPIs involved, for instance co-expression and co-evolution. Here, we have given attention to one important aspect only involving homo- and hetero-dimers in various levels of sequence alignment thresholds. The lower O/E ratios displayed by GU, Transmuted and other pairing models in inter-connections in comparison with intra-connections probably reflect the impact of different amino acid residue frequencies between the core and surface of protein binding sites.

In addition, to all our statistical studies, we also performed a next step, functional validation of the GU, Transmuted and Shift1/Shift2 pairing models by performing restrained MD studies with the biotin acceptor domain of human propionyl-CoA carboxylase, using different C_β restraints as defined by the different pairing models investigated during our studies. Importantly, whilst C_β restraints using all pairing models were able to promote substantial structural convergence relative to the simulation control performed without restraints (Figure S5), the GU, Transmuted and M-I 2 pairing models were the most effective which correlates with the fact that these models were the most frequent pairing models represented among 40 amino acid residue-residue connections seen in this protein structure (Fig. 8). Without doubt, amino acid usage (frequency) is a major factor in driving variations in residue-residue connections across different protein families.

Overall, although our investigations were not primarily intended to test for the existence or otherwise of a proteomic code, we would suggest that our data have in fact inadvertently demonstrated the reality of an evolved universal proteomic code based primarily on the GU, Transmuted and Shift1/Shift2 pairing models which are themselves derived from genetic principles complemented by biophysical principles of interaction. This universal proteomic code could represent a new fundamental in understanding protein structure/function, including protein folding and 3D structure prediction. In choosing TOP2018 dataset as our primary dataset for analyses leading to this conclusion, we opted for the most accurate protein 3D structure collection rather than the most representative of protein families. It is possible that further filtering has contributed inadvertently to certain protein subclasses being missed. However, the redundancy level in protein sequence identity can be considered adequate for drawing general conclusions. For individual proteins, we believe there should be a specific or refined proteomic code that is amino acid residue sequence-dependent, and comprises initial, interim and final residue-residue connections within a final, biologically active fold. These could be thought of as sequence-dependent "strategic connections" which might well be sufficient in and of themselves to guide folding of the given protein in the energetic landscape of the folding funnel. Clearly these sequence-dependent strategic connections must necessarily be a subset of the complete set of possible amino acid residue-residue connections mapped out by the pairing models that comprise the universal proteomic code.

The clear implication is that the universal proteomic code is in fact a general "all proteins" code comprising a basis set of amino acid residue pairing models for all proteins and PPIs, as underscored by the sensitivity analysis (Fig. S6). The formation of amino acid residue-residue connections originates from an interplay of various genetic and biophysical factors that must go through a refinement process for the most favoured connections. Initially, the number of all possible connections is far greater than the number of final connections in the final protein crystal structure. Only the final residue-residue connections, which are a subset of all possible connections for a particular model, will be strategic for the formation of a given final protein structure. The previous example (Fig. 8N) showed a final number of 40 connections of which 22 were from the GU pairing model. However, the GU model allows for about 231 possibilities in this protein and yet only a very small subset of 22 pairs has sustained the final fold. Chipens *et al.* [20] suggested that their pairing model and the M-I 1 pairing model may act in the early stages of folding to form initial connections. However, we propose that connections could be weighted instead depending on different roles with different pairing models employed to provide a particular selective advantage through the refinement process. Amino acid residue distribution rules (which can be interpreted as models that originate from periodicity in sequence) are also likely to be crucial for the determination of secondary and super-secondary structures [79,80]. Stambuk *et al.* [81] have investigated the practical applications of several pairing models (particularly, M-I 1, R-B, and Siemion) in addition to biophysical parameters (*e.g.*, hydrophobicity) in order to develop algorithms for the molecular design of functional sense/antisense peptides.

Without doubt, we are of the opinion that the foregoing amino acid residue pairing models could be readily incorporated into current machine learning for the prediction of protein structure and PPIs. Indeed, our models could be integrated at different stages: as input features, within machine learning model architectures, or during protein structure refinements. As input features, amino acid residue pairing models could be incorporated alongside sequence alignments to enhance feature extraction strategies, and improve structural predictions by providing additional constraints based on residue-residue interactions [82–84]. Alternatively, a hybrid approach might be proposed incorporating both machine learning and amino acid residue pairing models together. Furthermore, given recent advances, multi-chain templates can now be integrated into AlphaFold without masking inter-chain connections, making it feasible to use amino acid residue pairing models as template constraints for PPIs predictions [85]. Finally, existing machine learning models might be refined by enforcing contact map constraints derived from amino acid residue pairing models, either as an auxiliary loss function during training or as part of a post-prediction refinement strategy using energy minimization techniques. Future developments could involve integration of graph-based neural networks or generative approaches to further enhance predictive capabilities.

From our point of view, the future prospects for the universal proteomic code lie in developing a biologically relevant algorithm for protein 3D structure prediction that combines high accuracy with strong interpretability. This can be enhanced by analysing larger and more diverse datasets, and through the integration of explainable machine learning techniques. Overall, our investigation covered several amino acids pairing models, both historic and new. The new Transmuted pairing model, formed by merging M-I 1 and M-I 2 pairing models, was among the best performers with an emphasis on Transition. The new GU pairing model, defined to pull together sense codon-specified small and hydrophobic amino acid residues synonymous with residues associating readily by hydrophobic interactions, is the highest-ranking model and is evident in both intra- and inter-protein connections. Finally, the new Shift2 model, which links sense codon-specified amino acid residues with corresponding amino acid residues coded for by a subset of "Shift" NNx pairing codons, also performing well among the pairing models. Notably, the GU, Transmuted, and Shift2 pairing models exhibited superior performances in observed-over-expected ratios (Table 1) and correlations (Fig. 5) with diverse intra- and inter-connection datasets at both the C_β and side-chain levels. Taken together, the consistent performance of these pairing models across connection range, connection secondary structures, and protein size bolsters their candidacy as fundamental basis set components of a universally applicable proteomic code that derives fundamentally from core genetic principles complemented by relevant biophysical principles.

5. Conclusion

To the best of our knowledge, this study provides the first robust evidence for the existence of a universal proteomic code. By defining an expanded "basis set" of amino acid residue-residue connections, we propose a universal proteomic code that could be applied in a broad scope across protein tertiary and quaternary structures, even to protein-protein interfaces in multi-chained complexes. We also suggest that this universal proteomic code is able potentially to "encode" both protein folding and PPIs. In so saying, we draw back the veil on a missing dimension to the genetic code not under much previous active

consideration. By harnessing the properties of DNA codons such as complementarity and mutability, we are able to account for the redundancies in amino acid residue-residue connections and also potentially reduce the complexities of protein 3D structure prediction to simple genetic principles supplemented by relevant biophysical principles. The practical applications of a universal proteomic code are clear as a future potential tool for structural and synthetic biologists. For example, this could improve the scalability and accuracy of protein 3D structure prediction, advancing tools like AlphaFold and other predictive algorithms to new levels of precision. This knowledge could not only accelerate our ability to predict native protein structures but also provide a deeper mechanistic understanding of the rules underlying amino acid residue sequence-structure relationships. Moreover, insights gained from the universal proteomic code could guide targeted mutagenesis, enabling researchers to rationally design protein variants with improved stability, functionality, or specificity. Such advances have direct implications for engineering enzymes, optimizing metabolic pathways, and creating synthetic biological systems. Beyond this, the universal proteomic code could illuminate the principles underlying PPIs, fostering innovations in the design of novel therapeutics such as peptides, monoclonal antibodies, and other protein-based drugs. In summary, our next steps must now be to explore the implications of the universal proteomic code to better understand protein folding, protein structure/function, PPIs, the evolutionary basis of protein development, protein engineering, and synthetic biology.

CRediT authorship contribution statement

Tareq Hameduh: Writing – original draft, Formal analysis. **Andrew D. Miller:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Zbynek Heger:** Writing – review & editing, Project administration, Funding acquisition. **Yazan Haddad:** Writing – original draft, Project administration, Formal analysis, Conceptualization.

Availability of data

The data and code underlying this article are deposited in Zenodo (https://doi.org/10.5281/zenodo.14505653) under Creative Commons Attribution 4.0 International licence.

Ethical

We declare that no human or animal subject was used in this study. This work was purely computational whereas all procedures were performed in compliance with relevant laws and institutional guidelines in an area of research that does not require institutional approval by an ethical committee.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the revision stage of this work, YH used ChatGPT in order to correct language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zbynek Heger reports a relationship with Czech Health Research Council that includes: funding grants. Andrew D. Miller reports a relationship with Ministry of Education Youth and Sports of the Czech Republic that includes: funding grants. Andrew D. Miller reports a relationship with European Regional Development Fund that includes: funding grants. Andrew D. Miller reports a relationship with Interreg AT–CZ 2021–2027 that includes: funding grants. Andrew D. Miller is a shareholder in KP Therapeutics (Europe) s.r.o. All the authors are shareholders in MendelFOLD s.r.o.

Acknowledgements

The financial support from the Czech Health Research Council (project no. NU21J-08–00043) is gratefully acknowledged. ADM was supported by OPVVV Project FIT (Pharmacology, Immunotherapy, nanoToxicology) awarded by the Czech Ministry of Education, Youth and Sports (MŠMT) (CZ.02.1.01/0.0/0.0/15_003/0000495) with financial support from the European Regional Development Fund. Financial support from Interreg AT–CZ 2021–2027 (project NanoPrecMed, No. ATCZ00052) is also acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110033.

References

[1] T. Hameduh, Y. Haddad, V. Adam, Z. Heger, Homology modeling in the time of collective and artificial intelligence, Comput. Struct. Biotechnol. J. 18 (2020) 3494–3506, https://doi.org/10.1016/j.csbj.2020.11.007.
[2] M. Torrisi, G. Pollastri, Q. Le, Deep learning methods in protein structure prediction, Comput. Struct. Biotechnol. J. 18 (2020) 1301–1310, https://doi.org/10.1016/j.csbj.2019.12.011.
[3] L.M.F. Bertoline, A.N. Lima, J.E. Krieger, S.K. Teixeira, Before and after AlphaFold2: an overview of protein structure prediction, Front Bioinform 3 (2023) 1120370, https://doi.org/10.3389/fbinf.2023.1120370.
[4] J. Li, L. Wang, Z. Zhu, C. Song, Exploring the alternative conformation of a known protein structure based on contact map prediction, Eur Biophys J Biophy 64 (2024) 301–315, https://doi.org/10.1021/acs.jcim.3c01381.
[5] J. Jasmin Guven, N. Molkenthin, S. Muhle, A. Mey, What geometrically constrained models can tell us about real-world protein contact maps, Phys. Biol. 20 (2023) 046004, https://doi.org/10.1088/1478-3975/acd543.
[6] Z. Fakhoury, G.C. Sosso, S. Habershon, Generating protein folding trajectories using contact-map-driven directed walks, Eur Biophys J Biophy 63 (2023) 2181–2195, https://doi.org/10.1021/acs.jcim.3c00023.
[7] Y. Liu, Y. Liu, Z. Li, Protein-protein interaction prediction via structure-based deep learning, Proteins 92 (2024) 1287–1296, https://doi.org/10.1002/prot.26721.
[8] B. Harihar, K.M. Saravanan, M.M. Gromiha, S. Selvaraj, Importance of inter-residue contacts for understanding protein folding and unfolding rates, remote homology, and drug design, Mol. Biotechnol. 67 (2025) 862–884, https://doi.org/10.1007/s12033-024-01119-4.
[9] Y. Si, C. Yan, Improved inter-protein contact prediction using dimensional hybrid residual networks and protein language models, Briefings Bioinf. 24 (2023) 1–11, https://doi.org/10.1093/bib/bbad039.
[10] V.A. Jisna, P.B. Jayaraj, Protein structure prediction: conventional and deep learning perspectives, Protein J. 40 (2021) 522–544, https://doi.org/10.1007/s10930-021-10003-y.
[11] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)-Round XV, Proteins 91 (2023) 1539–1549, https://doi.org/10.1002/prot.26617.
[12] T. Oda, Improving protein structure prediction with extended sequence similarity searches and deep-learning-based refinement in CASP15, Proteins 91 (2023) 1712–1723, https://doi.org/10.1002/prot.26551.
[13] J. Liu, Z. Guo, T. Wu, R.S. Roy, C. Chen, J. Cheng, Improving AlphaFold2-based protein tertiary structure prediction with MULTICOM in CASP15, Commun. Chem. 6 (2023) 188, https://doi.org/10.1038/s42004-023-00991-6.
[14] Z. Peng, W. Wang, H. Wei, X. Li, J. Yang, Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15, Proteins 91 (2023) 1704–1711, https://doi.org/10.1002/prot.26570.
[15] L.B. Mekler, [Specific selective interaction between amino acid groups of polypeptide chains], Biofizika 14 (1969) 581–584.
[16] L.B. Mekler, R.G. Idlis, Построение Моделей трехМерных Молекул биологических полипептидов и нуклеопротеидов согласно об‡еМу коду, определяю‡еМу спёифическое линейное узнавание и связывание аМинокислотныМи остаткаМи полипептидов как друг друга, так и тринуклеотидов полинуклеотидов, VINITI Deposited Doc (1981) 1476–1481.
[17] L.B. Mekler, R.G. Idlis, Об‡ий стереохиМический генетический код-путь к биотехнологии и универсальной Медiйине XXI века уЖе сегодня, Природа 5 (1993) 29–63.
[18] J.R. Heal, G.W. Roberts, J.G. Raynes, A. Bhakoo, A.D. Miller, Specific interactions between sense and complementary peptides: the basis for the proteomic code, Chembiochem 3 (2002) 136–151, https://doi.org/10.1002/1439-7633(20020301)3:2/3<136::AID-CBIC136>3.0.CO;2-7.

[19] A.D. Miller, Sense-antisense (complementary) peptide interactions and the proteomic code; potential opportunities in biology and pharmaceutical science, Expet Opin. Biol. Ther. 15 (2015) 245–267, https://doi.org/10.1517/14712598.2015.983069.
[20] G.I. Chipens, N.G. Ievinia, R.B. Rudzish, [Code of codon roots, determining the intra- and intermolecular interaction of amino acids in peptide chains], Bioorg. Khim. 17 (1991) 1582–1584.
[21] J.E. Zull, R.C. Taylor, G.S. Michaels, N.B. Rushforth, Nucleic acid sequences coding for internal antisense peptides: are there implications for protein folding and evolution? Nucleic Acids Res. 22 (1994) 3373–3380, https://doi.org/10.1093/nar/22.16.3373.
[22] R.S. Root-Bernstein, D.D. Holsworth, Antisense peptides: a critical mini-review, J. Theor. Biol. 190 (1998) 107–119, https://doi.org/10.1006/jtbi.1997.0544.
[23] J. Biro, Comparative analysis of specificity in protein-protein interactions. Part II.: the complementary coding of some proteins as the possible source of specificity in protein-protein interactions, Med. Hypotheses 7 (1981) 981–993, https://doi.org/10.1016/0306-9877(81)90094-3.
[24] J.C. Biro, The Proteomic Code: a molecular recognition code for proteins, Theor. Biol. Med. Model. 4 (2007) 45, https://doi.org/10.1186/1742-4682-4-45.
[25] J.E. Blalock, E.M. Smith, Hydropathic anti-complementarity of amino acids based on the genetic code, Biochem. Biophys. Res. Commun. 121 (1984) 203–207, https://doi.org/10.1016/0006-291x(84)90707-1.
[26] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132, https://doi.org/10.1016/0022-2836(82)90515-0.
[27] G.I. Chipens, L.E. Gnilomedova, R.V. Rudzish, [Code of codon roots of amino acids and idiotypic networks], Bioorg. Khim. 17 (1991) 1449–1455.
[28] G.I. Chipens, R.V. Rudzish, [Interaction code for polar and nonpolar amino acids: "ice-breaker" model], Bioorg. Khim. 17 (1991) 1445–1448.
[29] R.S. Root-Bernstein, Amino acid pairing, J. Theor. Biol. 94 (1982) 885–894, https://doi.org/10.1016/0022-5193(82)90083-2.
[30] I.Z. Siemion, The regularity of changes of the Chou-Fasman parameters within the genetic code, Biosystems 32 (1994) 25–35, https://doi.org/10.1016/0303-2647(94)90016-7.
[31] I.Z. Siemion, R. Zbozien-Pacamaj, P. Stefanowicz, New hypothesis on amino acid complementarity and its evaluation on TGF-beta(2)-related peptides, J. Mol. Recogn. 14 (2001) 1–12, https://doi.org/10.1002/1099-1352(200101/02)14:1<1::AID-JMR512>3.0.CO;2.
[32] I.Z. Siemion, M. Cebrat, A. Kluczyk, The problem of amino acid complementarity and antisense peptides, Curr. Protein Pept. Sci. 5 (2004) 507–527, https://doi.org/10.2174/1389203043379413.
[33] G. Fassina, Complementary peptides as antibody mimetics for protein purification and assay, Immunomethods 5 (1994) 121–129, https://doi.org/10.1006/immu.1994.1046.
[34] G. Fassina, G. Cassani, A. Corti, Binding of human tumor necrosis factor alpha to multimeric complementary peptides, Arch. Biochem. Biophys. 296 (1992) 137–143, https://doi.org/10.1016/0003-9861(92)90555-b.
[35] G. Fassina, R. Consonni, L. Zetta, G. Cassani, Design of hydropathically complementary peptides for Big Endothelin affinity purification, Int. J. Pept. Protein Res. 39 (1992) 540–548, https://doi.org/10.1111/j.1399-3011.1992.tb00286.x.
[36] G. Fassina, M. Melli, Identification of interactive sites of proteins and protein receptors by computer-assisted searches for complementary peptide sequences, Immunomethods 5 (1994) 114–120, https://doi.org/10.1006/immu.1994.1045.
[37] N. Stambuk, P. Konjevoda, J. Pavan, Antisense peptide technology for diagnostic tests and bioengineering research, Int. J. Mol. Sci. 22 (2021) 9106, https://doi.org/10.3390/ijms22179106.
[38] K. Austin, J.A. Torres, J.D.V. Waters, E.R.M. Balog, J.M. Halpern, R.J. Pantazes, An orthogonal workflow of electrochemical, computational, and thermodynamic methods reveals limitations of using a literature-reported insulin binding peptide in biosensors, ACS Omega 9 (2024) 39219–39231, https://doi.org/10.1021/acsomega.4c06481.
[39] K. Xu, H. Gao, Y. Li, Y. Jin, R. Zhao, Y. Huang, Synthetic peptides with genetic-codon-tailored affinity for assembling tetraspanin CD81 at cell interfaces and inhibiting cancer metastasis, Angew Chem. Int. Ed. Engl. 63 (2024) e202400129, https://doi.org/10.1002/anie.202400129.
[40] G. Faure, A. Bornot, A.G. de Brevern, Protein contacts, inter-residue interactions and side-chain modelling, Biochimie 90 (2008) 626–639, https://doi.org/10.1016/j.biochi.2007.11.007.
[41] J. Esque, C. Oguey, A.G. de Brevern, Comparative analysis of threshold and tessellation methods for determining protein contacts, Eur Biophys J Biophy 51 (2011) 493–507, https://doi.org/10.1021/ci100195t.
[42] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205, https://doi.org/10.1093/nar/gkm998.
[43] D.A. Beck, A.L. Jonsson, R.D. Schaeffer, K.A. Scott, R. Day, R.D. Toofanny, D. O. Alonso, V. Daggett, Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations, Protein Eng. Des. Sel. 21 (2008) 353–368, https://doi.org/10.1093/protein/gzn011.
[44] C.J. Williams, D.C. Richardson, J.S. Richardson, The importance of residue-level filtering and the Top2018 best-parts dataset of high-quality protein residues, Protein Sci. 31 (2022) 290–300, https://doi.org/10.1002/pro.4239.
[45] T. Hameduh, M. Mokry, A.D. Miller, Z. Heger, Y. Haddad, Solvent accessibility promotes rotamer errors during protein modeling with major side-chain prediction programs, Eur Biophys J Biophy 63 (2023) 4405–4422, https://doi.org/10.1021/acs.jcim.3c00134.
[46] T. Hameduh, M. Mokry, A.D. Miller, V. Adam, Z. Heger, Y. Haddad, A rotamer relay information system in the epidermal growth factor receptor-drug complexes reveals clues to new paradigm in protein conformational change, Comput. Struct. Biotechnol. J. 19 (2021) 5443–5454, https://doi.org/10.1016/j.csbj.2021.09.026.
[47] Y. Haddad, V. Adam, Z. Heger, Rotamer dynamics: analysis of rotamers in molecular dynamics simulations of proteins, Biophys. J. 116 (2019) 2062–2072, https://doi.org/10.1016/j.bpj.2019.04.017.
[48] H. Zhang, Z. Bei, W. Xi, M. Hao, Z. Ju, K.M. Saravanan, H. Zhang, N. Guo, Y. Wei, Evaluation of residue-residue contact prediction methods: from retrospective to prospective, PLoS Comput. Biol. 17 (2021) e1009027, https://doi.org/10.1371/journal.pcbi.1009027.
[49] A. Vangone, A.M. Bonvin, Contacts-based prediction of binding affinity in protein-protein complexes, Elife 4 (2015) e07454, https://doi.org/10.7554/eLife.07454.
[50] B.J. Grant, A.P. Rodrigues, K.M. ElSawy, J.A. McCammon, L.S. Caves, Bio3d: an R package for the comparative analysis of protein structures, Bioinformatics 22 (2006) 2695–2696, https://doi.org/10.1093/bioinformatics/btl461.
[51] Z. Wang, J. Xu, Predicting protein contact map using evolutionary and physical constraints by integer programming, Bioinformatics 29 (2013) i266–i273, https://doi.org/10.1093/bioinformatics/btt211.
[52] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637, https://doi.org/10.1002/bip.360221211.
[53] D. Osorio, P. Rondón-Villarreal, R. Torres, Peptides: a package for data mining of antimicrobial peptides, R J 7 (2015) 4–14, https://doi.org/10.32614/RJ-2015-001.
[54] J.L. Risler, M.O. Delorme, H. Delacroix, A. Henaut, Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix, J. Mol. Biol. 204 (1988) 1019–1029, https://doi.org/10.1016/0022-2836(88)90058-7.
[55] M. Boniecki, P. Rotkiewicz, J. Skolnick, A. Kolinski, Protein fragment reconstruction using various modeling techniques, J. Comput. Aided Mol. Des. 17 (2003) 725–738, https://doi.org/10.1023/b:jcam.0000017486.83645.a0.
[56] S. Miyazawa, R.L. Jernigan, Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading, J. Mol. Biol. 256 (1996) 623–644, https://doi.org/10.1006/jmbi.1996.0114.
[57] S. Miyazawa, R.L. Jernigan, Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues, Proteins 34 (1999) 49–68, https://doi.org/10.1002/(sici)1097-0134(19990101)34:1<49::aid-prot5>3.0.co;2-l.
[58] A. Godzik, A. Kolinski, J. Skolnick, Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets, Protein Sci. 4 (1995) 2107–2117, https://doi.org/10.1002/pro.5560041016.
[59] S. Miyazawa, R.L. Jernigan, A new substitution matrix for protein sequence searches based on contact frequencies in protein structures, Protein Eng. 6 (1993) 267–278, https://doi.org/10.1093/protein/6.3.267.
[60] K.T. Simons, I. Ruczinski, C. Kooperberg, B.A. Fox, C. Bystroff, D. Baker, Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins, Proteins 34 (1999) 82–95, https://doi.org/10.1002/(sici)1097-0134(19990101)34:1<82::aid-prot7>3.0.co;2-a.
[61] E. Azarya-Sprinzak, D. Naor, H.J. Wolfson, R. Nussinov, Interchanges of spatially neighbouring residues in structurally conserved environments, Protein Eng. 10 (1997) 1109–1122, https://doi.org/10.1093/protein/10.10.1109.
[62] D. Naor, D. Fischer, R.L. Jernigan, H.J. Wolfson, R. Nussinov, Amino acid pair interchanges at spatially conserved locations, J. Mol. Biol. 256 (1996) 924–938, https://doi.org/10.1006/jmbi.1996.0138.
[63] J.M. Duarte, R. Sathyapriya, H. Stehr, I. Filippis, M. Lappe, Optimal contact definition for reconstruction of contact maps, BMC Bioinf. 11 (2010) 283, https://doi.org/10.1186/1471-2105-11-283.
[64] M. Kayikci, A.J. Venkatakrishnan, J. Scott-Brown, C.N.J. Ravarani, T. Flock, M. M. Babu, Visualization and analysis of non-covalent contacts using the Protein Contacts Atlas, Nat. Struct. Mol. Biol. 25 (2018) 185–194, https://doi.org/10.1038/s41594-017-0019-z.
[65] A.W. Fitzpatrick, T.P. Knowles, C.A. Waudby, M. Vendruscolo, C.M. Dobson, Inversion of the balance between hydrophobic and hydrogen bonding interactions in protein folding and aggregation, PLoS Comput. Biol. 7 (2011) e1002169, https://doi.org/10.1371/journal.pcbi.1002169.
[66] F.C.L. Almeida, K. Sanches, R. Pinheiro-Aguiar, V.S. Almeida, I.P. Caruso, Protein surface interactions-theoretical and experimental studies, Front. Mol. Biosci. 8 (2021) 706002, https://doi.org/10.3389/fmolb.2021.706002.
[67] I. Antonov, P. Baranov, M. Borodovsky, GeneTack database: genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences, Nucleic Acids Res. 41 (2013) D152–D156, https://doi.org/10.1093/nar/gks1062.
[68] A.E. Firth, I. Brierley, Non-canonical translation in RNA viruses, J. Gen. Virol. 93 (2012) 1385–1409, https://doi.org/10.1099/vir.0.042499-0.
[69] R. Qing, S. Hao, E. Smorodina, D. Jin, A. Zalevsky, S. Zhang, Protein design: from the aspect of water solubility and stability, Chem. Rev. 122 (2022) 14085–14179, https://doi.org/10.1021/acs.chemrev.1c00757.
[70] D. Tomasek, D. Kahne, The assembly of beta-barrel outer membrane proteins, Curr. Opin. Microbiol. 60 (2021) 16–23, https://doi.org/10.1016/j.mib.2021.01.009.
[71] G.D. Rose, From propensities to patterns to principles in protein folding, Proteins 93 (2025) 105–111, https://doi.org/10.1002/prot.26540.

[72] T. Yasuda, R. Morita, Y. Shigeta, R. Harada, Ribosome tunnel environment drives the formation of alpha-helix during cotranslational folding, Eur Biophys J Biophy 64 (2024) 6610–6622, https://doi.org/10.1021/acs.jcim.4c00901.
[73] K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani, J. J. Weinstein, N.M. Mangan, S. Ovchinnikov, G.J. Rocklin, Mega-scale experimental analysis of protein folding stability in biology and design, Nature 620 (2023) 434–444, https://doi.org/10.1038/s41586-023-06328-6.
[74] A. Amala, I.A. Emerson, Understanding contact patterns of protein structures from protein contact map and investigation of unique patterns in the globin-like folded domains, J. Cell. Biochem. 120 (2019) 9877–9886, https://doi.org/10.1002/jcb.28270.
[75] M. Cieslik, Z.S. Derewenda, The role of entropy and polarity in intermolecular contacts in protein crystals, Acta Crystallogr. D Biol. Crystallogr. 65 (2009) 500–509, https://doi.org/10.1107/S0907444909009500.
[76] C.K.E. Baek, C.H. Baek, Clustered complementary amino acid pairing (CCAAP) for protein-protein interaction, Biotechnol. Lett. 41 (2019) 79–90, https://doi.org/10.1007/s10529-018-2616-2.
[77] S. Akbarzadeh, O. Coskun, B. Guncer, Studying protein-protein interactions: latest and most popular approaches, J. Struct. Biol. 216 (2024) 108118, https://doi.org/10.1016/j.jsb.2024.108118.
[78] I.M. Nooren, J.M. Thornton, Diversity of protein-protein interactions, EMBO J. 22 (2003) 3486–3492, https://doi.org/10.1093/emboj/cdg359.
[79] A.E. Kister, A.V. Finkelstein, I.M. Gelfand, Common features in structures and sequences of sandwich-like proteins, P Natl. Acad. Sci. USA 99 (2002) 14137–14141, https://doi.org/10.1073/pnas.212511499.
[80] A. Kister, Amino acid distribution rules predict protein fold: protein grammar for beta-strand sandwich-like structures, Biomolecules 5 (2015) 41–59, https://doi.org/10.3390/biom5010041.
[81] N. Stambuk, P. Konjevoda, P. Turcic, K. Kover, R.N. Kujundzic, Z. Manojlovic, M. Gabricevic, Genetic coding algorithm for sense and antisense peptide interactions, Biosystems 164 (2018) 199–216, https://doi.org/10.1016/j.biosystems.2017.10.009.
[82] L. Xian, Y.S. Wang, Advances in computational methods for protein-protein interaction prediction, Electronics-Switz 13 (2024) 1059, https://doi.org/10.3390/electronics13061059.
[83] J. Zhang, J. Durham, C. Qian, Revolutionizing protein-protein interaction prediction with deep learning, Curr. Opin. Struct. Biol. 85 (2024) 102775, https://doi.org/10.1016/j.sbi.2024.102775.
[84] J. Wang, J.L. Watson, S.L. Lisanza, Protein design using structure-prediction networks: AlphaFold and RoseTTAFold as protein structure foundation models, Csh Perspect Biol. 16 (2024) a041472, https://doi.org/10.1101/cshperspect.a041472.
[85] C. Mirabello, B. Wallner, B. Nystedt, S. Azinas, M. Carroni, Unmasking AlphaFold to integrate experiments and predictions in multimeric complexes, Nat. Commun. 15 (2024) 8724, https://doi.org/10.1038/s41467-024-52951-w.