De Novo Design of a Four-Helix Maquette Protein for Haem b Binding

1. Protein design

De novo maquettes provide a minimal framework for exploring the relationship between sequence, structure, and function. The four-helix bundle (4HB) is the canonical scaffold, as its hydrophobic and electrostatic patterning leads to predictable folding. Here, a single-chain 4HB was engineered with weak haem-binding capacity to probe minimal sequence requirements for function.

The design applies binary heptad patterning: leucine at *a*/*d* core positions and glutamate/lysine at *e*/*g* sites form complementary inter-helical salt bridges, following established *de novo* principles (Gibney *et al.*, 1998; Huang *et al.*, 2016). This amphipathic organisation (Figure 1) shows leucine forming the hydrophobic core face and charged residues on the solvent-exposed surface, with high hydrophobic moments (μH = 0.312–0.716) consistent with strong coiled-coil helices. Each 22-residue helix (~6 turns) has a predicted length of 3.3 nm (0.15 nm per residue).

A minimal five-residue alphabet (L, K, E, G, H) balances simplicity with essential chemistry. Gly$_4$ linkers provide flexibility without disrupting helical packing. Shorter Gly$_2$ linkers produced misfolded topologies; while PGGP linkers restored correct geometry but expanded the alphabet. Gly$_4$ therefore optimises minimality while maintaining structural fidelity.

Full 106-residue sequence (helices underlined):

```
LKKLEEKLKKLEEKLKKLEEKL GGGG LKELEKKLKELEKKLKELEKKL HGG GGGG
LKKLEEKLKKLEEKLKKLEEKL GGGG LKELEKKLKELEKKLKELEKKL HGG
```

C-terminal His–Gly–Gly extensions on helices 2 and 4 introduce weak haem b coordination sites (Section 2). The Gly$_2$ spacers following each histidine provide conformational flexibility for haem coordination without disrupting helical structure.

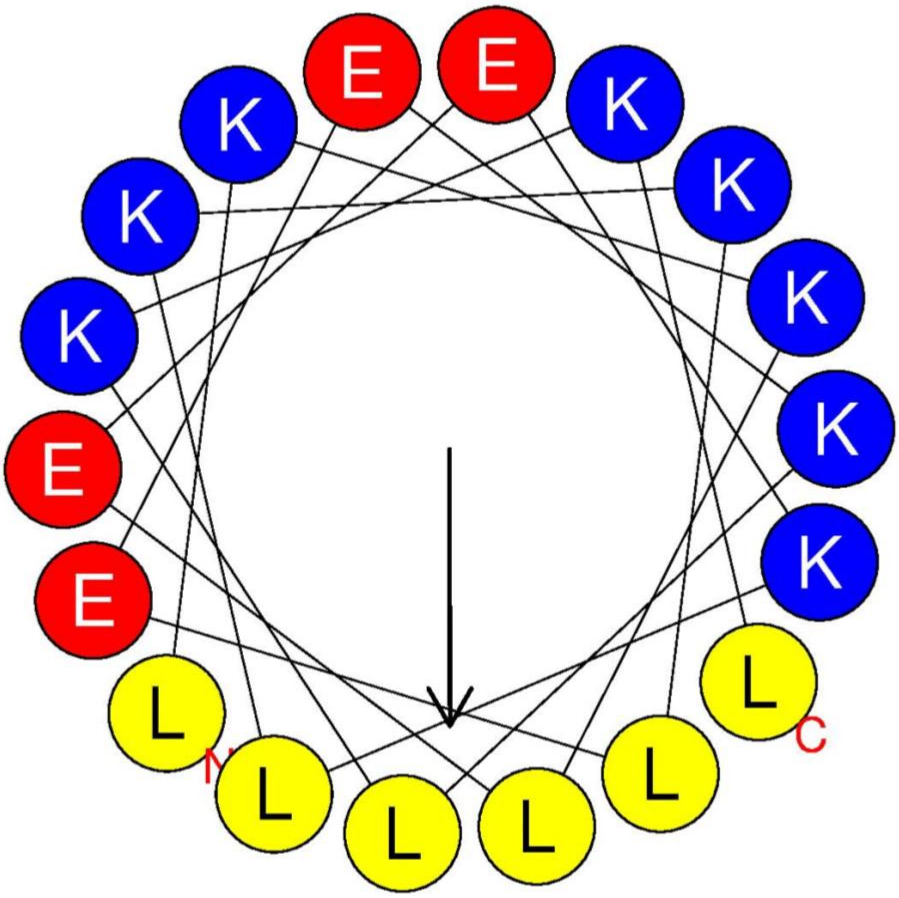

Figure 1. Helical-wheel projection. *HeliQuest projection showing leucine (yellow) segregated to the hydrophobic core face; lysine (blue) and glutamate (red) on the solvent-exposed face. μH = 0.31–0.72 (Eisenberg scale).*

2. Functional elements – haem b

Haem b is a ubiquitous redox cofactor whose iron centre coordinates with histidine imidazoles. To introduce weak, reversible haem binding while preserving the hydrophobic core, His–Gly–Gly (HGG) motifs were appended to helices 2 and 4 (Figure 2). Positioning ligands at helix termini avoids core disruption thus maintaining solvent accessibility, and the adjacent Gly_2 spacer confers rotational freedom that allows the histidine imidazole Nε2 to coordinate Fe(III) in a flexible, low-affinity bis-His geometry typical of weak maquette haem sites (Gibney *et al.*, 1998; Shifman *et al.*, 2000).

AlphaFold modelling positions the two His Nε2 atoms 14.6 Å apart, consistent with the 12–16 Å window observed in flexible maquettes (Gibney *et al.*, 1998; Moser *et al.*, 2016), supporting a deliberately low-affinity site (Shifman *et al.*, 2000). This could be verified via UV-spectroscopic observation of the Soret band shift from free to coordinated haem.

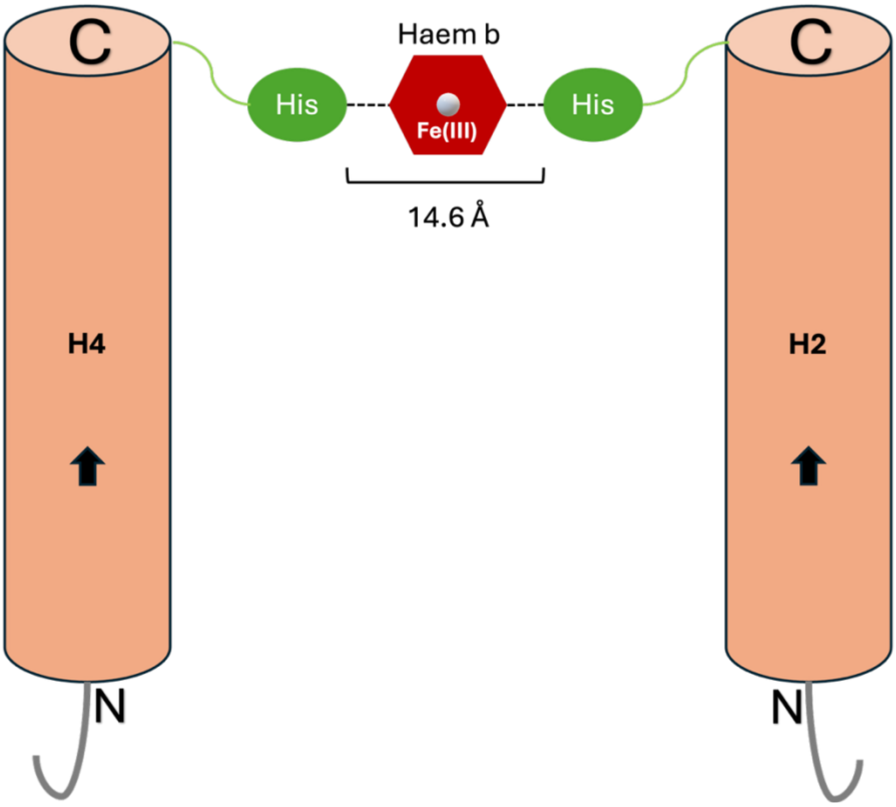

Figure 2. Haem b coordination geometry. *His–Gly–Gly tails (green) on H2 and H4 coordinate Fe(III) (red), forming flexible bis-His geometry (14.6 Å separation) characteristic of weak maquette haem binding.*

3. Structural representation

The designed sequence folds into a antiparallel single-chain 4HB bundle with each helix approximately 3.3 nm long and connected by Gly_4 linkers (Figure 3A). The topology follows a canonical up–down–up–down arrangement, producing a compact cylindrical bundle in which leucine residues create the buried hydrophobic core and glutamate/lysine provide complementary surface electrostatics.

His–Gly–Gly tails on helices 2 and 4 project toward one bundle end, positioning haem b for weak bis-His coordination (Figure 3B). *AlphaFold2* modelling reproduces this geometry with correct helix packing and continuous leucine core (Figure 3C). The His–His Nε2 distance of 14.6 Å (Figure 3D) falls within the 12–16 Å range characteristic of flexible haem coordination, consistent with weak binding (Shifman *et al.*, 2000).

Convergent *AlphaFold2* (pLDDT ≈ 79, pTM ≈ 0.70, core Cα RMSD ≤ 1 Å) and *Rosetta* models support a stable 4HB topology. These metrics, however, assess local geometry rather than ligand energetics or oligomeric state, so the model represents a structural hypothesis pending experimental validation (Leaver-Fay *et al.*, 2011; Jumper *et al.*, 2021).

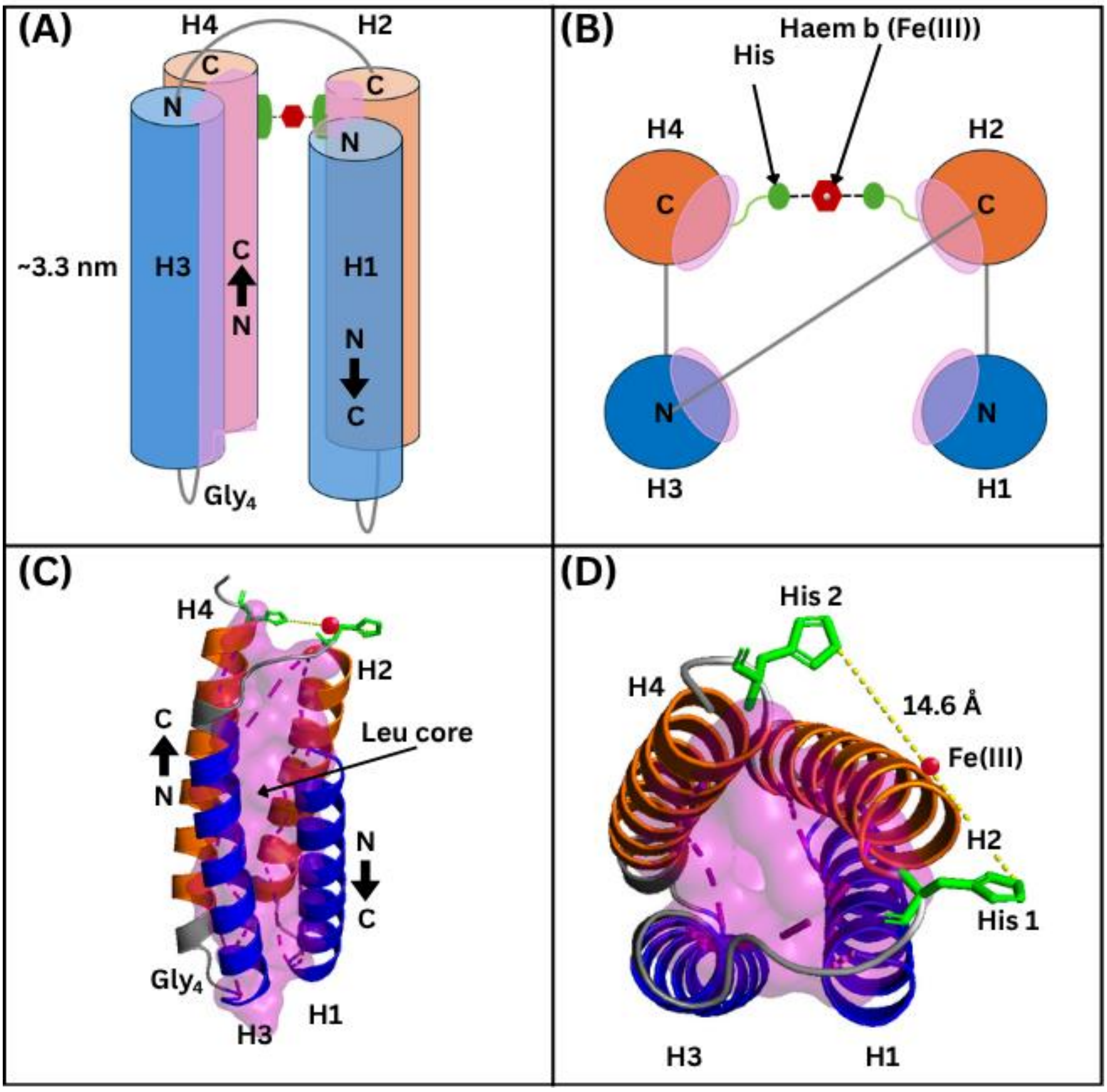

Figure 3. Structural representations of the four-helix maquette. *(A,B) Schematic views of the antiparallel bundle (H1–H4, 22 residues/helix) connected by Gly₄ linkers. HGG tails (green) coordinate haem b (red). (C,D) AlphaFold2 model showing buried Leu core (purple surface) and 14.6 Å His separation (yellow dashes).*

4. Sequence analysis

The design employs a five-residue alphabet (L, K, E, G, H), demonstrating how minimal chemical diversity can encode stable tertiary structure and function. Each residue performs a distinct role (Table 1): leucine forms the hydrophobic core (*a/d* positions), lysine and glutamate provide inter-helical salt bridges (*e/g* sites), glycine enables linker flexibility, and histidine coordinates haem b.

Table 1. Alphabet rationale

Residue	Role	Function
L (leucine)	Core hydrophobe	Buried *a/d* positions drive bundle stability
K (Lysine)	Positive charge	Inter-helical salt bridges; solubility
E (Glutamate)	Negative charge	Complements lysine for electrostatic pairing
G (Glycine)	Flexibility	Torsional freedom in linkers and HGG tails
H (Histidine)	Functional ligand	Coordinates haem b; enables redox activity

Each helix comprises 22 residues (~6 turns), corresponding to 3.3 nm in length (22 × 0.15 nm per residue), consistent with canonical α-helices (Pace and Scholtz, 1998). This minimal alphabet isolates essential folding determinants, making sequence–structure relationships interpretable and mirrors early *de novo* studies showing that binary patterning of hydrophobic and charged residues can yield autonomous folding (Regan and DeGrado, 1988; Riddle *et al.*, 1997; Woolfson *et al.*, 2015; Huang *et al.*, 2016).

5. DNA design for Escherichia coli expression

To enable efficient expression, the amino-acid sequence was reverse-translated and codon-optimised for *Escherichia coli* following established principles of codon bias (Sharp and Li, 1987; Gustafsson, Govindarajan and Minshull, 2004). Rare codons (< 5 % usage) were avoided to prevent translational pausing (Plotkin and Kudla, 2011). The rare leucine codon CTA (~ 4 % frequency) was replaced with CTG (> 40 %), while alternating high-frequency codons AAA/AAG, GAA/GAG, GGT/GGC, and CAT/CAC balanced GC content and reduced homopolymer runs (Kudla *et al.*, 2009). The optimised design achieved CAI = 0.90 and GC = 41 %, suitable for E. coli expression and synthesis. The gene can be inserted into a pET vector under T7 control for high-level overexpression (Salis, Mirsky and Voigt, 2009).

References

Gibney, B.R. *et al.* (1998) 'Effect of Four Helix Bundle Topology on Heme Binding and Redox Properties', *Biochemistry*, 37(13), pp. 4635–4643. Available at: https://doi.org/10.1021/bi971856s.

Gustafsson, C., Govindarajan, S. and Minshull, J. (2004) 'Codon bias and heterologous protein expression', *Trends in biotechnology*, 22(7), pp. 346–353.

Huang, P.-S. *et al.* (2016) 'The coming of age of de novo protein design', *Nature*, 537(7620), pp. 320–327. Available at: https://doi.org/10.1038/nature19946.

Jumper, J. *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *nature*, 596(7873), pp. 583–589.

Kudla, G. *et al.* (2009) 'Coding-sequence determinants of gene expression in Escherichia coli', *science*, 324(5924), pp. 255–258.

Leaver-Fay, A. *et al.* (2011) 'ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules', in *Methods in enzymology*. Elsevier, pp. 545–574.

Moser, C.C. *et al.* (2016) 'De Novo Construction of Redox Active Proteins', in *Methods in Enzymology*. Elsevier, pp. 365–388. Available at: https://doi.org/10.1016/bs.mie.2016.05.048.

Plotkin, J.B. and Kudla, G. (2011) 'Synonymous but not the same: the causes and consequences of codon bias', *Nature Reviews Genetics*, 12(1), pp. 32–42.

Regan, L. and DeGrado, W.F. (1988) 'Characterization of a helical protein designed from first principles', *Science*, 241(4868), pp. 976–978.

Riddle, D.S. *et al.* (1997) 'Functional rapidly folding proteins from simplified amino acid sequences', *Nature structural biology*, 4(10), pp. 805–809.

Salis, H.M., Mirsky, E.A. and Voigt, C.A. (2009) 'Automated design of synthetic ribosome binding sites to control protein expression', *Nature biotechnology*, 27(10), pp. 946–950.

Sharp, P.M. and Li, W.-H. (1987) 'The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias.', *Molecular biology and evolution*, 4(3), pp. 222–230.

Shifman, J.M. *et al.* (2000) 'Heme Redox Potential Control in de Novo Designed Four-α-Helix Bundle Proteins', *Biochemistry*, 39(48), pp. 14813–14821. Available at: https://doi.org/10.1021/bi000927b.

Woolfson, D.N. *et al.* (2015) ‘De novo protein design: how do we expand into the universe of possible protein structures?’, *Current opinion in structural biology*, 33, pp. 16–26.