

Supporting Information for

Imperfect symmetry facilitated the evolution of specificity and high-order stoichiometry in vertebrate hemoglobin

Carlos R. Cortez-Romero (1), Jixing Lyu (4), Arvind S. Pillai (2,5), Arthur Langanowsky (4), Joseph W. Thornton (2,3)

Affiliations: Department of (1) Cell and Molecular Biology, (2) Ecology and Evolution, and (3) Human Genetics, University of Chicago, Chicago, IL, 60637. (4) Department of Chemistry, Texas A&M University, College Station, TX, 77843. (5) Institute of Protein Design, University of Washington, Seattle, WA, 98195

Corresponding author: Joseph W. Thornton

Email: joet1@uchicago.edu

This PDF file includes:

Figures S1 to S13
Materials and methods (complete)

SUPPLEMENTARY FIGURES

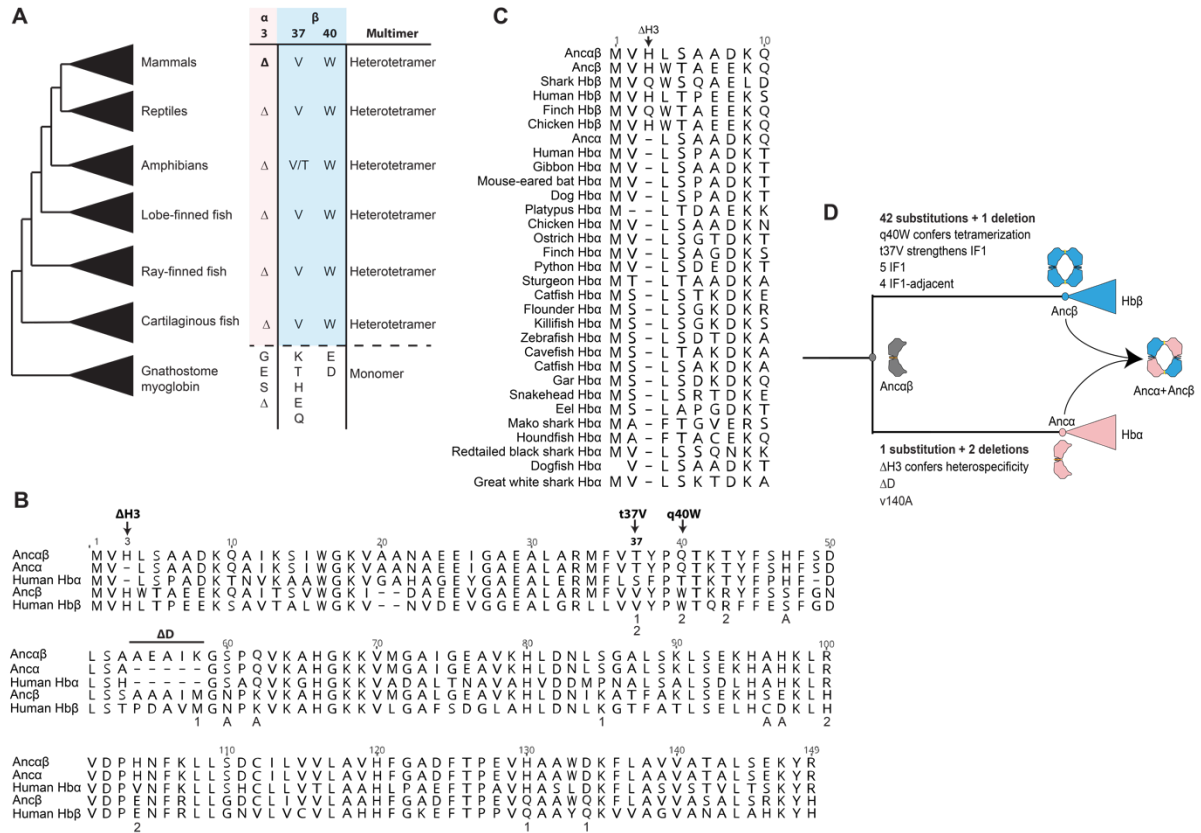


Fig. S1. Key sequence changes are conserved in extant hemoglobin subunits. A) Multimerization state and residues at key sequence sites in extant Hb subunits from major jawed vertebrate taxa. Myoglobin is the closest paralogous protein in the globin family. For complete alignment, see <https://tinyurl.com/yc6kvdb8>. **B)** Alignment of extant human and reconstructed ancestral Hb subunits is shown. Historical sequence changes that confer heterospecificity ($\Delta H3$ in Hb α) and tetramerization ($q40W$ in Hb β) are labeled and shown with arrows. Additional substitutions that occurred on the branch leading to Anca β and assayed in this paper are also labeled, including changes on the surface of IF1 (1), IF2 (2), and at sites adjacent to IF1 (A). **C)** Alignment of N-terminal portion of Hb subunits from species representative of major vertebrate taxa. The deletion of residue 3 in alpha subunits is marked. **D)** Summary of key sequence changes that occurred after the duplication of Anca $\alpha\beta$. Multimerization states of reconstructed ancestral proteins is shown.

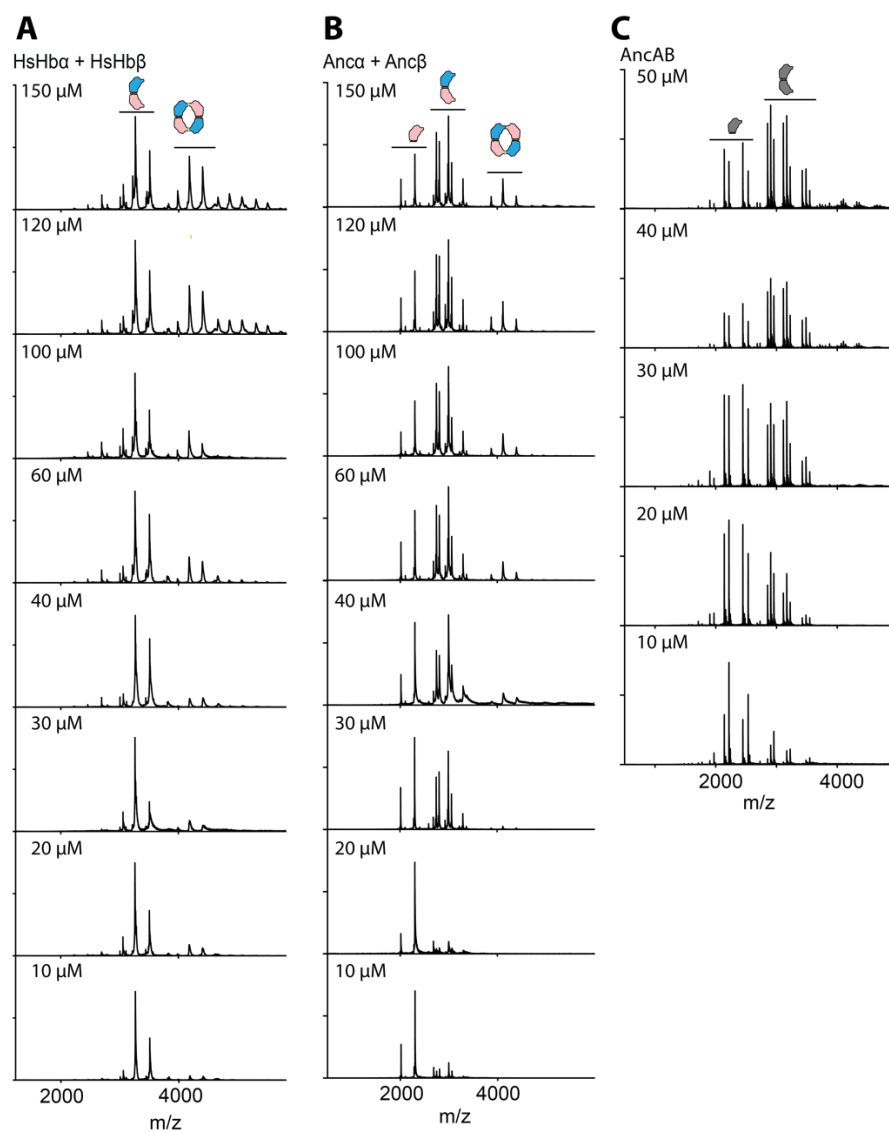


Fig. S2. Native mass spectrometry spectra. nMS spectra across a concentration series is shown for A) human Hb, B) Anca + Anc β , and C) AncAB. Peaks corresponding to monomers, dimers, and tetramer are labeled.

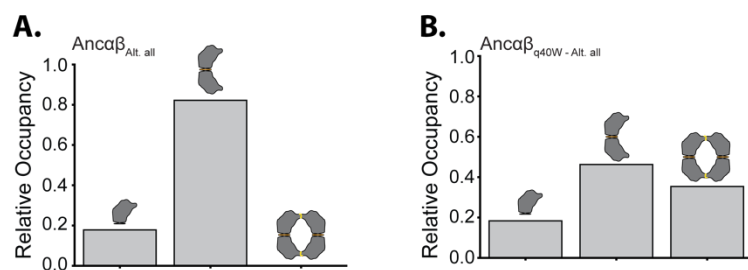


Fig. S3. Effect of q40W tetramerization is robust to statistical uncertainty. (A) Relative occupancy of monomer, dimer, and tetramer of Ancaβ_{Alt. all}, an alternative reconstruction of Ancaβ that contains the second most likely state at all ambiguously reconstructed sites, measured at 20 μM total protein using native MS. (B) Relative occupancy Ancaβ_{Alt. all} with substitution q40W.

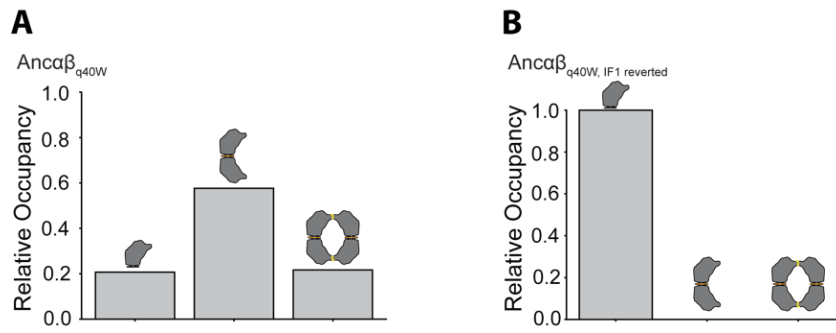


Fig. S4. The effect of q40W on tetramerization depends on IF1. (A) Relative occupancy of Ancaβ_{q40W}, measured by native MS at 20 μM total protein. (B) Relative occupancy of Ancaβ_{q40W_IF1-reverted}, which contains mutation q40W, as well as reversions to the ancestral state found in AncMH of all residues that were substituted between AncMH and Ancaβ .

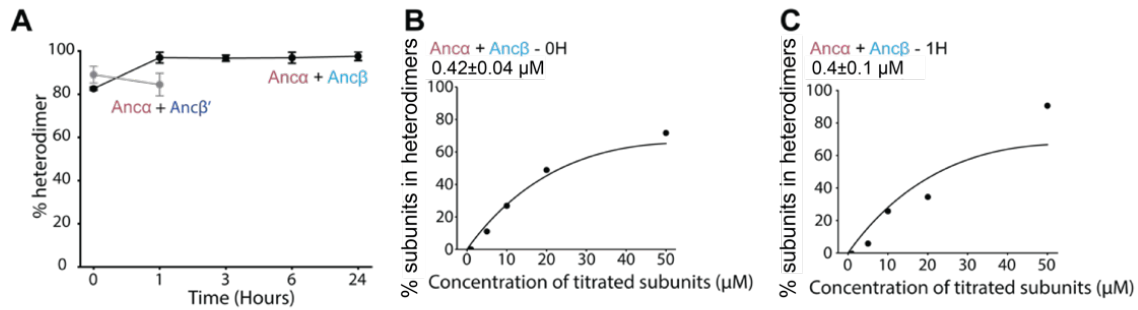


Fig. S5. Heterodimer occupancy of Anca and Ancβ is near equilibrium after mixing. (A) The percent of all dimers that are heterodimers, measured by nMS when proteins are mixed at 50 μM each and allowed to incubate for 0, 1, 3, 6, or 24 hours. Black line and points, Anca + Ancβ (which only dimerize when expressed separately and then mixed). Grey line and points, Anca + Ancβ' (Ancβ in which IF2 surface substitutions are reverted to their ancestral state in Ancaβ, thus preventing tetramerization). Each dot shows the mean of three replicates; error bars, SEM. (B) Affinity of monomer-to-heterodimer assembly measured by nMS immediately upon mixing of Anca and Ancβ. Anca was kept constant at 50 μM , while the concentration of Ancβ varied. Points, fraction of all subunits in the mixture that are incorporated into heterodimers. Line, best-fit binding curve. Estimated K_d and 95% confidence interval are shown. (C) Estimated heterodimerization affinity measured as in panel B, but 1 hour after mixing.

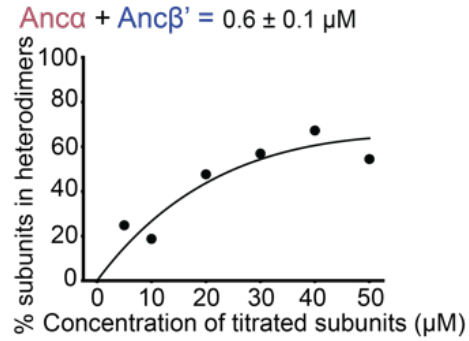


Fig. S6. Heterodimerization by $\text{Anc}\alpha + \text{Anc}\beta'$. Monomer-to-heterodimer assembly measured by nMS. $\text{Anc}\alpha$ was kept constant at $50 \mu\text{M}$ while $\text{Anc}\beta'$ was at variable concentration. Points, fraction of all subunits in the mixture that are incorporated into heterodimers at each concentration. Line, best-fit binding curve. Estimated K_d and 95% confidence interval are shown.

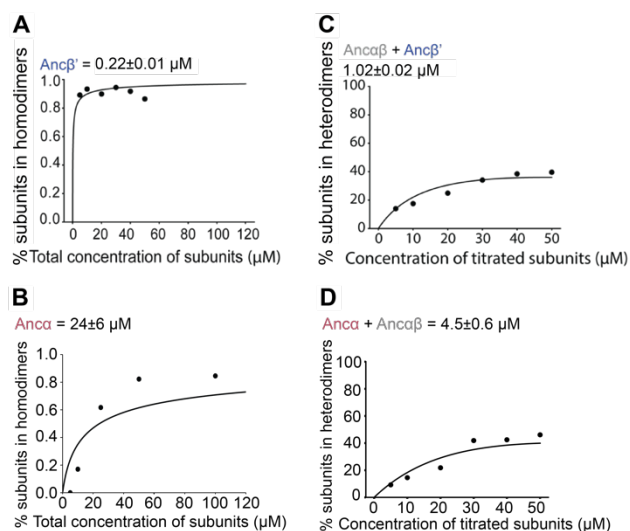


Fig. S7. Dimerization by Anca and Ancβ' (A-B) Homodimerization by Ancβ' (panel A) and by Anca (B). measured by nMS across a titration series. Each point shows the fraction of subunits incorporated into dimers as the concentration of protein varied. Best-fit binding curve, K_d , and 95% confidence interval are shown. (C-D) Heterodimerization by mixtures of Ancαβ'+ Ancβ' (C) and Ancαβ'+Anca and Ancaβ+Ancαβ (D). Each point shows the fraction of all subunits incorporated into heterodimers. In each case, one protein was held constant at 50 μM while the other was varied.

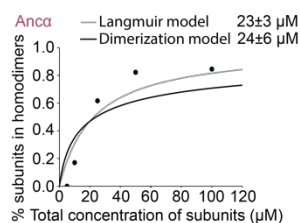


Fig. S8. Estimated affinity of the monomer-dimer transition by the Anc α homodimer is robust to the binding model used. We fit two different binding equations (curves) to the stoichiometries of **Anc α** measured using nMS. The Langmuir-Hill model (black line) assumes that the concentration of free monomeric subunits is not depleted by dimerization. The dimerization model used throughout the rest of this paper accounts for this depletion (described in Materials and Methods). The estimated Kds (shown with their 95% confidence intervals) are statistically indistinguishable.

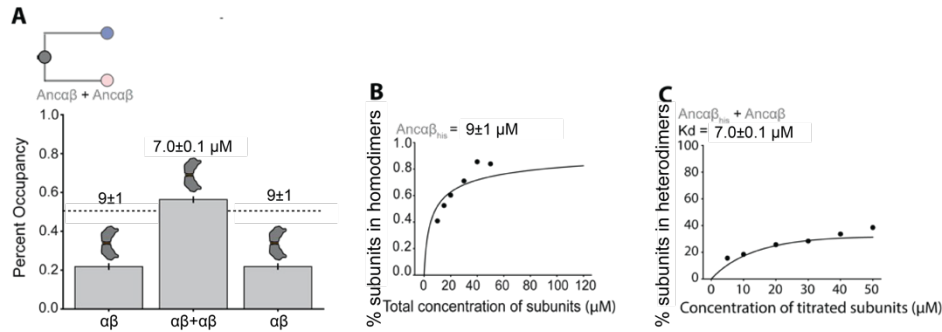


Fig. S9. Dimerization affinities and occupancy of Ancaβ. (A) Expected fractional occupancies of homodimer and heterodimers when Ancaβ is mixed at equal concentrations with Ancaβ_{his} (500 μM each), given the measured dimerization affinities (shown above each column, with 95% confidence interval). Ancaβ_{his} is Ancaβ with an N-terminal polyhistidine tag, which allows the masses of the three kinds of dimer to be distinguished. (B-C) Homodimerization by Ancaβ_{his} and heterodimerization by affinity of Ancaβ + Ancaβ_{his}, measured and represented as in Fig. S5.

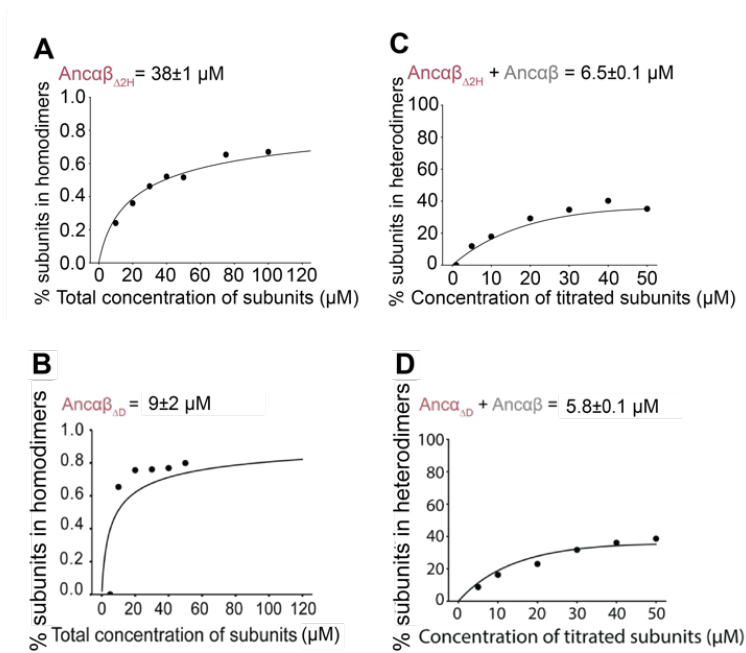


Fig. S10. Effect of historical deletions on dimerization. (A-B) Homodimerization and (C-D) Heterodimerization by mixtures, measured and represented as in Fig. S5.

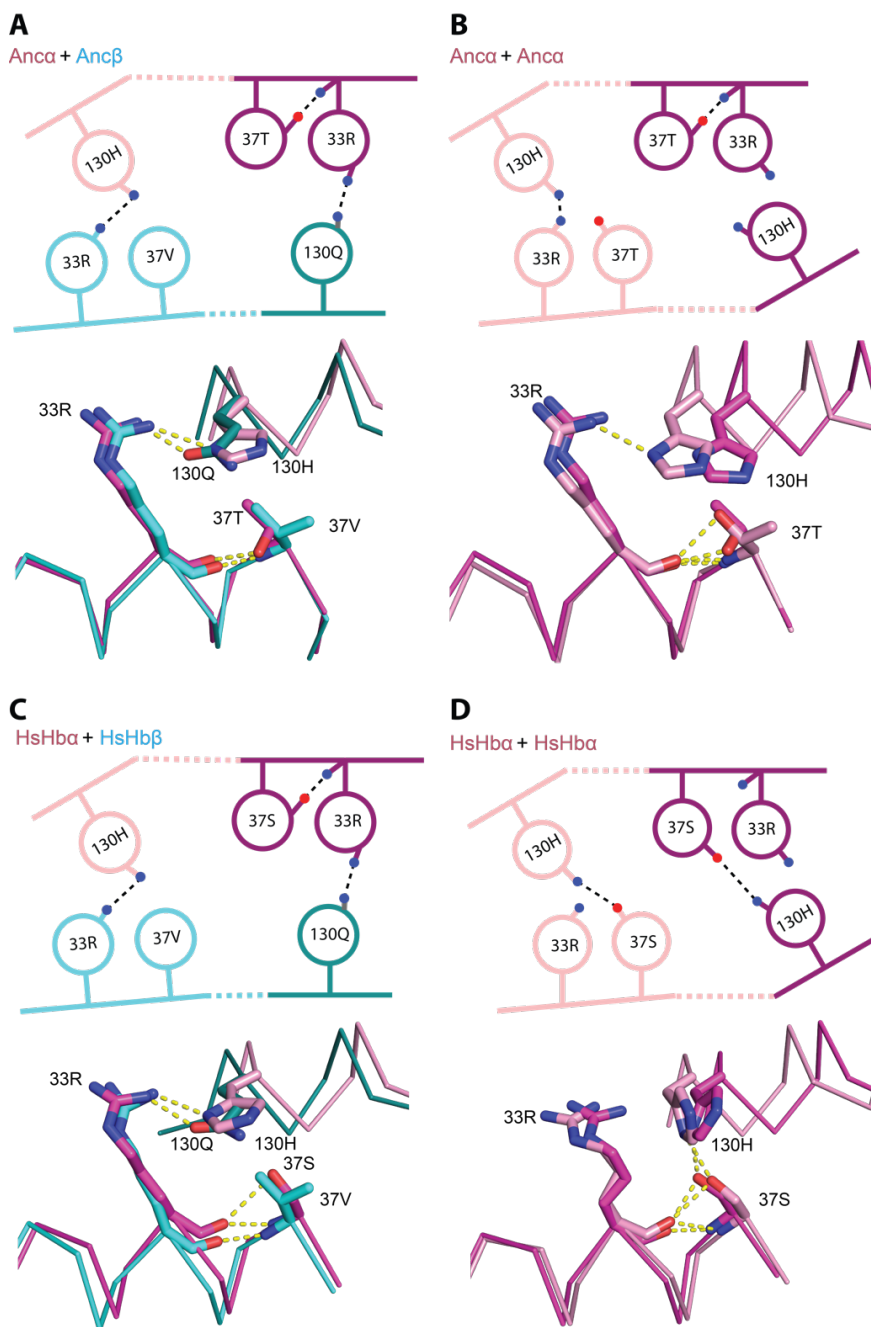


Fig. S11. Nonadditive interactions that contribute to specificity are conserved in derived Hb complexes. In the modeled homodimers and heterodimers of Anca+Anc β (panels A, B) and X-ray crystal structure of human hemoglobin (PDB 4HHB and 3S48), the figure shows the key IF1 residues with nonadditive interactions in Anca β +Anca $\beta_{\Delta H3}$ (see Fig. 5G for comparison). Top, cartoon of key contacts. The two iterations of these interactions across the isologous interface are shown, one each in light or dark hue. Blue and red, nitrogen and oxygen atoms, respectively. Dotted lines, hydrogen bonds. Bottom, structural alignment of the two iterations of the isologous interface in each dimer. Each dimer structure was duplicated exactly and then aligned to the original by targeting one subunit of the copy to align to the other subunit of the original. Hues correspond to the isologous iterations in the cartoon above

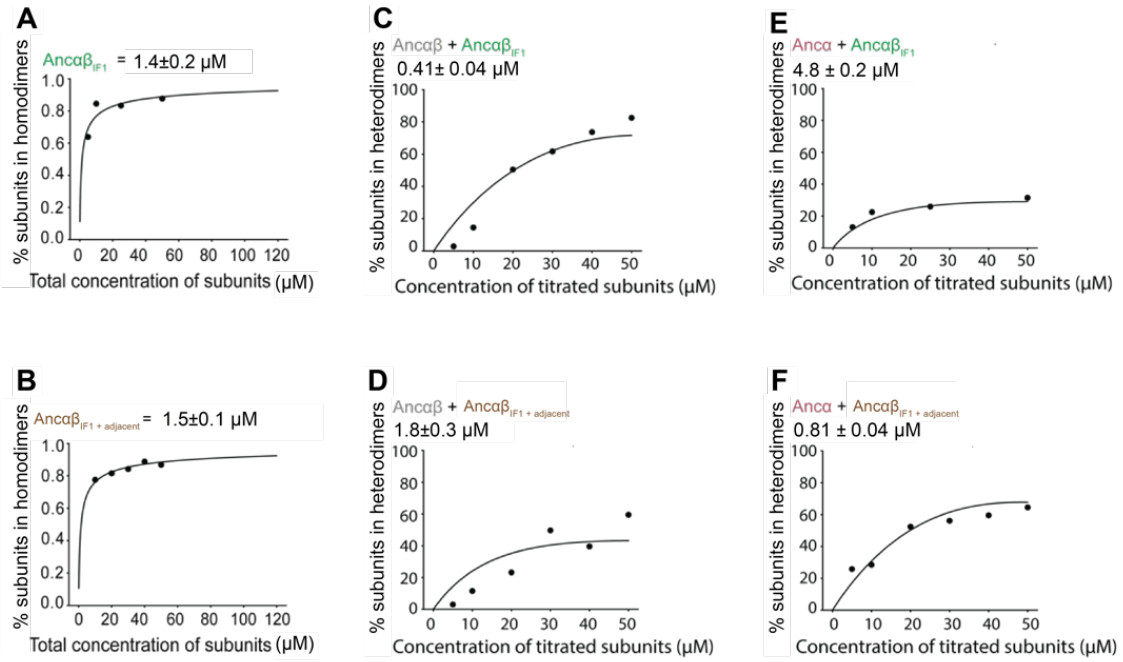


Fig. S12. Homodimerization by $Anca\beta_{IF1}$ and $Anca\beta_{IF1 + Adjacent}$ (A,B) and heterodimerization by those proteins when mixed with $Anca\beta$ (C,D) or $Anca\alpha$ (E,F). Measurements and representation as in Fig. S5.

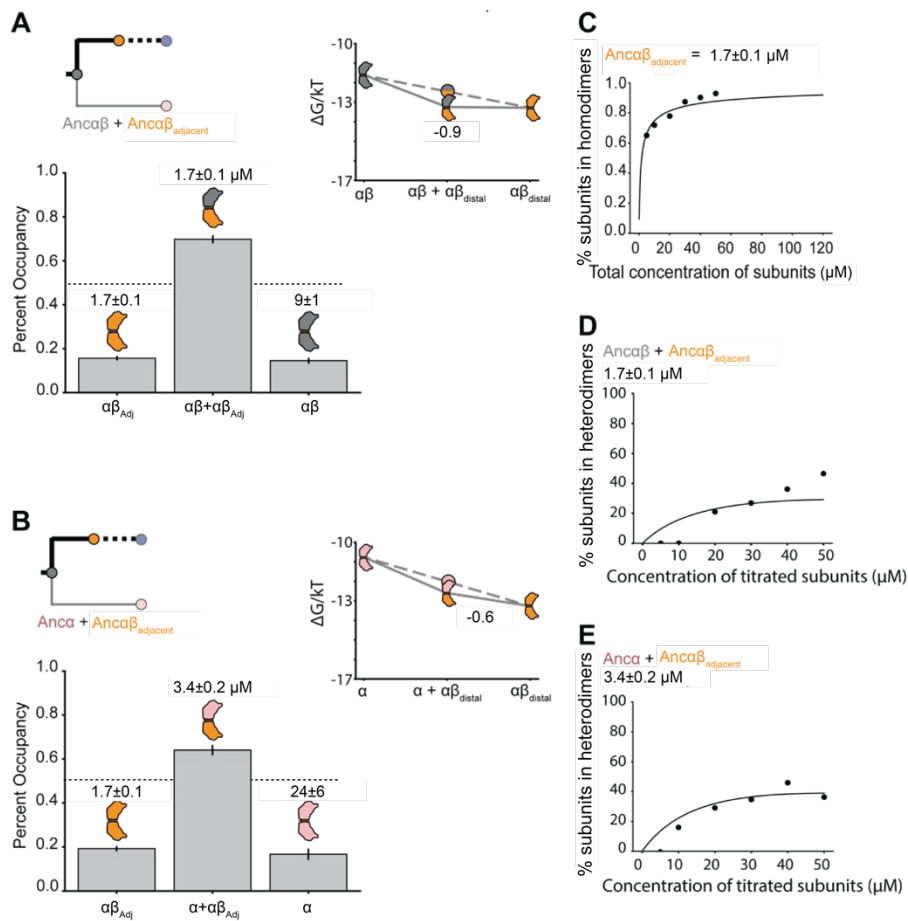


Fig. S13. Dimerization affinity and occupancies for $Anca\beta_{Adjacent}$. Expected fractional occupancies of homodimer and heterodimers when $Anca\beta_{Adjacent}$ is mixed with $Anca\beta$ (A) or $Anca\alpha$ (B), each at (500 μM), given the measured dimerization affinities (shown above each column, with 95% confidence interval). Inset, ΔG of each dimerization (measured in units of kT), with ΔG_{spec} of the heterodimer shown. (C,D,E) Measurement of binding affinities, measured and represented as in Fig. S5.

METHODS

Sequence data, alignment, phylogeny, and ancestral sequence reconstruction.

The reconstructed ancestral sequences used here are the same as those reported previously (1). Briefly, 177 amino acid sequences of hemoglobin and related paralogs were collected and aligned. The maximum likelihood (ML) phylogenetic tree was inferred using the AIC best-fit model, LG+G+F (57,58). The phylogeny was rooted using as outgroups neuroglobin and globin X, which are found in both deuterostomes and protostomes and diverged prior to the gene duplications that produced vertebrate myoglobin and the hemoglobin subunits. Ancestral sequence reconstruction was performed using the empirical Bayes method (2), given the alignment, ML phylogeny, ML branch lengths, and ML model parameters. Reconstructed ancestors that were used in this study have been deposited previously in GenBank (IDs MT079112, MT079113, MT079114, MT079115).

The historical mutations that we introduced into those ancestral proteins are the following. For the set *IF1-reverted*, all sites in IF1 that were substituted on the branch leading to Ancb are reverted to the ancestral state found in Ancab; the mutations introduced are V36t, Y38h, V115a, V119e, H130r, D134e. For the set *IF2-reverted*, all sites that were substituted in IF2 on the branch leading to Ancb are reverted to the ancestral state found in Ancab; the mutations introduced are T37v, W40q, R43t, H100r, E104h. For the set *IF1*, all sites at IF1 that were substituted between Ancab and Ancb are changed to the derived state found in Ancb; the mutations introduced are t37V, k58M, r107K, h130Q, d134Q4. For the set *Adjacent*, five sites adjacent to IF1 that were substituted between Ancab and Ancb are changed to the derived state found in Ancb; the mutations introduced are h47S, s60N, q62K, a96S, h97E. The set *IF1+Adjacent* is the union of the sets *IF1* and *Adjacent*. Deletion DD deletes residues a54, e55, a56, i57, and k58 from Ancab.

Recombinant protein expression. Coding sequences for reconstructed ancestral proteins were optimized for expression in *Escherichia coli* using IDT Codon Optimization and synthesized *de novo* as gBlocks (IDT). Coding sequences were cloned by Gibson assembly into vector pLIC (3) under control of a T7 polymerase promoter. For co-expression of Anc α +Anc β , a polycistronic operon was constructed under control of a T7 promoter and separated by a spacer containing a stop codon and ribosome binding site, as described in (4).

BL21 (DE3) *Escherichia coli* cells (New England Biolabs) were heat-shock transformed and plated onto Luria broth (LB) containing 50 ug/mL carbenicillin. For the starter culture, a single colony was inoculated into 50 mL of LB with 1:1000 dilution of working-stock carbenicillin and grown overnight. 5 mL of the starter culture were inoculated into a larger 500-mL terrific broth (TB) mixture containing the appropriate antibiotic

concentration. Cells were grown at 37° C and shaken at 225 rpm in an incubator until they reached an optical density at 600 nm of 0.6-0.8.

For expression of single globin proteins, 100 μ M of isopropyl- β -D-1-thiogalactopyranoside (IPTG) and 25 mg/500 mL of hemin were added to each culture. Expression of single proteins in culture were done overnight at 22° C. Cells were collected by centrifugation at 4,000g and stored at -80° C until protein purification. Coexpressed proteins were induced using 500 mM IPTG expression with 25 mg/500 mL hemin for 4 hours at 37°C. Cells were collected by centrifugation at 4,000g, immediately followed by purification.

Human hemoglobin was bought commercially (Sigma-Aldridge) and resuspended in PBS.

We attempted to co-express and purify Anca $\beta_{\Delta H3}$ in complex with Anca β_{40W} , but we were not able to identify conditions at which the two species could be expressed and purified to near-equal concentrations.

Protein purification by ion exchange. All singly expressed proteins (all ancestral globins except Anca α +Anca β) were purified using ion exchange chromatography. All buffers were vacuum filtered through a 0.2 μ M PTFE membrane (Omnipore). After expression, cells were resuspended in 30 mL of 50 mM Tris-Base (pH 6.88). The resuspended cells were placed in a 10 mL falcon tube and lysed using a FB505 sonicator (1s on/off for three cycles, each 1 minute). The lysate was saturated with CO, transferred to a 30 mL round bottom tube, and centrifuged at 20,000g for 60 minutes to separate supernatant from non-soluble cell debris. The supernatant was collected and syringe-filtered using HPX Millex Durapore filters (Millipore) to further remove debris. A HiTrap SP cation exchange (GE) column was attached to an FPLC system (Biorad) and equilibrated in 50 mM Tris-Base (pH 6.88). The lysate was passed over the column. 50 mL of 50 mM Trise-Base (pH 6.88) was run through the SP column to remove weakly bound non-target soluble products. Elution of bound ancestral Hbs was performed with 100-mL gradient of 50mM Tris-Base 1 M NaCl (pH 6.88) buffer which was run through the column from 0% to 100%. 1.5 mL fractions were captured during the gradient process, all fractions containing red eluant were put into an Amicon ultra-15 tube and concentrated by centrifugation at 4,000g to a final volume of 1 mL. For additional purification, concentrated sample was injected into a HiPrep 16/60 Sephacryl S-100 HR size exclusion chromatography (SEC) column. The column was equilibrated in phosphate buffered saline (PBS) at pH 7.4. Purified ancestral globins elute at different volumes depending on the protein's complex stoichiometry: 48-52 for tetramers, 56-60 for dimers, and 65-67 for monomers. The purified proteins were concentrated as mentioned above and then flash frozen with liquid nitrogen.

For experiments in which two proteins were singly expressed and purified and then mixed together, expression and purification of each protein were performed as described

above. The concentration of each protein was then quantified using the Hemoglobin Assay Kit (Sigma). Proteins were then mixed together at 50 μ M each for nMS. This procedure was performed in triplicate to assess technical error introduced by the quantification and mixing process.

Protein purification by zinc affinity chromatography. Coexpressed proteins Anc α + Anc β were purified using zinc-affinity chromatography, which was performed using a HisTrap metal affinity column (GE) on a Biorad NGC Quest. Nickel ions were stripped from the column (buffer 100 mM EDTA, 100 mM NaCl, 20 mM TRIS, pH 8.0), followed by five column volumes of water. To attach zinc to the column, 0.1 M ZnSO₄ was passed over until conductance was stable, approximately 5 column volumes, followed by five column volumes of water. After expression, cells were resuspended in a 50 mL lysis buffer (20 mM Tris, 150 mM NaCl, 10% glycerol (v/v), 1mM BME, 0.05% Tween-20, and 1 Roche Protease EDTA-free inhibitor tablet, pH 7.40), sonicated as described above, and the lysate passed through the prepared column. To remove non-specifically bound protein, the column was washed with 50 mL of lysis buffer. Bound protein was then eluted across a gradient of imidazole concentrations (0 to 500 mM) in a total of 100 mL elution buffer (20 mM Tris, 150 mM NaCl, 500 mM imidazole, 10% glycerol, and 1 mM BME, pH 7.4). 1 mL fractions were collected. The fraction corresponding to the second peak of UV absorbance at 280 nm has a visible red color and was collected and concentrated as described above. The concentrated solution was injected into a Biorad ENrich 650 10 x 300 columns for additional purification and eluted in PBS buffer.

Size exclusion chromatography assay. For protein concentrations from 0 to 500 μ M, size exclusion chromatography was performed using a Superdex 75 increase 10/300 GL column (GE) equilibrated in PBS, then injected with 250 μ L of sample using a 2 mL injection loop on an Biorad NGC Quest FPLC and monitored by absorbance at 280 nm. For proteins at concentration 1 mM, a HiPrep 16/60 Sephacryl S-100 HR was equilibrated in PBS using an AKTApriime FPLC, then injected with 1mL sample and monitored by absorbance at 280 nm.

Native Mass Spectrometry. Protein samples were buffer exchanged into 200mM ammonium acetate using either a centrifugal buffer exchange device (Micro Bio-Spin P-6 Gel, Bio-Rad) or a dialysis device (Slide-A-Lyzer MINI Dialysis Unit, 10000 MWCO, Thermo) prior to native MS experiments. Samples were loaded into gold-coated glass capillaries made in-house and introduced to Synapt G1 HDMS instrument (Waters corporation) equipped with a 32k RF generator (5). The instrument was set to a source pressure of 5.47 mbar, capillary voltage of 1.75 kV, sampling cone voltage of 20 V, extractor cone voltage of 5.0 V, trap collision voltage of 10 V, collision gas (Argon) flow rate of 2 mL/min (2.65×10^{-2} mbar), and T-wave settings (velocity/height) for trap, IMS and transfer of 100 ms -1 /0.2 V, 300 ms -1 /16.0 V, and 100 ms -1 /10.0 V, respectively. The source temperature (70 °C) and trap bias (30 V) were optimized. Part of the native MS experiments were conducted by Thermo Scientific Exactive Plus Orbitrap with Extended Mass Range (EMR) with tuning as follow: source DC offset of 15 V, injection

flatapole DC to 13 V, inter flatapole lens to 5, bent flatapole DC to 4, transfer multipole DC to 3 and C trap entrance lens to 0, trapping gas pressure to 5.0 with the CE to 10, spray voltage to 1.50 kV, capillary temperature to 100 °C, maximum inject time to 100 ms. Mass spectra were acquired with a setting of 8750 resolution, microscans set to 1 and averaging set to 100. Mass spectra were deconvoluted using Unidec (6).

Calculating multimerization affinity of homodimers. To estimate K_d of the monomer-to-homodimer transition of singly expressed proteins, we performed nMS at variable protein concentrations (P_{tot}). The occupancy of each oligomeric state at each concentration was calculated as the proportion of all globin subunits in that state, based on the summed areas under the corresponding peaks in the native MS spectrum. The fraction of subunits assembled into dimers (F_d) includes dimers and tetramers and is defined as

$$F_d = \frac{2x_d + 4x_t}{(x_m + 2x_d + 4x_t)},$$

where x_m , x_d , and x_t are the total signal intensities of all peaks corresponding to the monomeric, dimeric and tetrameric stoichiometries, respectively. Nonlinear regression was used to find the best-fit value of K_d of dimerization using the equation:

$$F_d = \frac{1}{P_{tot}} * \frac{(4P_{tot} + K_d) - \sqrt{(4P_{tot} + K_d)^2 - 16P_{tot}^2}}{4}$$

As an alternative model of homodimerization, we also used a version of the Hill-Langmuir equation:

$$F_d = \frac{P_{tot}}{P_{tot} + K_d}$$

The Hill-Langmuir model, which is typically used for ligand-receptor binding, does not account for depletion of free monomeric subunits as homodimerization takes place and is therefore not a valid model in this case; we used it solely to determine the robustness of the estimated K_d to the specific binding equation used.

Calculating multimerization affinity of homotetramers. To estimate the K_d of the homodimer-homotetramer transition, the fraction of subunits assembled into tetramers is defined as

$$F_t = \frac{4x_t}{(2x_d + 4x_t)}.$$

The concentration of all dimers is defined as

$$P_d = F_d \times P_{tot}.$$

Nonlinear regression was then used to find the K_d of tetramerization using the equation:

$$F_t = \frac{1}{P_d} * \frac{(4P_d + K_d) - \sqrt{(4P_d + K_d)^2 - 16P_d^2}}{4}$$

Parameters were estimated using the curvefit script in the Scipy package (7). The 95% confidence interval on the K_d was estimated as 1.96 times the estimated standard error.

Calculating multimerization affinity of heteromers. To determine the K_d of heterodimerization, we used nMS to measure stoichiometries across a titration series in which one protein's concentration was held constant at 50 μ M and the other was added at variable concentration (1 to 50 μ M). From the nMS spectrum, we estimated the proportion of the heterodimer and the two homodimers as

$$F_{\alpha\alpha} = \frac{2x_{\alpha\alpha}}{(2x_{\alpha\alpha} + 2x_{\alpha\beta} + 2x_{\beta\beta} + x_{\alpha} + x_{\beta})}$$

$$F_{\alpha\beta} = \frac{2x_{\alpha\beta}}{(2x_{\alpha\alpha} + 2x_{\alpha\beta} + 2x_{\beta\beta} + x_{\alpha} + x_{\beta})}$$

$$F_{\beta\beta} = \frac{2x_{\beta\beta}}{(2x_{\alpha\alpha} + 2x_{\alpha\beta} + 2x_{\beta\beta} + x_{\alpha} + x_{\beta})}$$

where each x represents the signal intensity of all peaks corresponding to the species denoted in the subscript. The dissociation constant for each dimer is defined as $Kd_1 =$

$\frac{x_{\alpha}^2}{x_{\alpha\alpha}}$, $Kd_2 = \frac{x_{\beta}^2}{x_{\beta\beta}}$, and $Kd_3 = \frac{x_{\alpha}x_{\beta}}{x_{\alpha\beta}}$. By substitution, $F_{\alpha\beta}$ can be expressed as

$$F_{\alpha\beta} = \frac{\sqrt{Kd_1 * Kd_2 * F_{\alpha\alpha} * F_{\beta\beta}}}{Kd_3}$$

Kd_3 was estimated using this equation by nonlinear regression, where $F_{\alpha\alpha}$, $F_{\alpha\beta}$ and $F_{\beta\beta}$ were measured using the titration series, and the affinities Kd_1 and Kd_2 were assigned the values estimated in the homodimerization experiments described above.

Prediction of homodimer and heterodimer occupancy at high concentrations. The occupancy of each dimer at physiologically relevant concentrations (1 mM total globin subunits) was predicted as follows, because nMS is limited to concentrations $<100\mu$ M. In

a mixture of two types of globins *A* and *B*, the total concentration of each subunit can be expressed in terms of the concentration of monomers [A] and [B] in the mixture:

$$[A]_{\text{tot}} = [A] + [AB] + 2[AA] = [A] + \frac{[A][B]}{Kd_3} + \frac{2[A]^2}{Kd_1}$$

$$[B]_{\text{tot}} = [B] + [AB] + 2[BB] = [B] + \frac{[A][B]}{Kd_3} + \frac{2[B]^2}{Kd_2}$$

We used these equations to predict [A] and [B] at any value of C_A and C_B given the experimentally estimated Kds. The concentration of each dimer was then estimated using the equations $[AA] = \frac{[A]^2}{Kd_1}$, $[BB] = \frac{2[A][B]}{Kd_3}$, and $[BB] = \frac{[B]^2}{Kd_2}$.

Establishing the upper limit of IF2 Kd. We estimated the minimum Kd of assembly across IF2 by Anca β _{37V+40W}; IF1 removed, because no homotetramer was observed using nMS at a protein concentration of 20 μ M. The minimum detection limit for dimers in the nMS assay is 1 μ M. Kd is defined as $Kd = \frac{[M]^2}{[D]}$, where [M] and [D] are the equilibrium concentrations of monomer and dimer, respectively. Therefore

$$Kd_{\text{min}} = \frac{(20 * 10^{-6})^2 M}{1 * 10^{-6} M} = 400 \mu M$$

Determining $\Delta\Delta G$ of specificity. Specificity for heterodimer assembly between two paralogs can be defined as the difference between the additive affinity of the heterodimer and the measured affinity of the heterodimer, using ΔG s derived measured dimerization affinity for two homodimers and their respective heterodimer. The additive affinity of the heterodimer is defined as the averaged ΔG of both homodimers:

$$\Delta G_{\text{heterodimer}}^{\text{additive}} = \frac{\Delta G_{\text{homodimer 1}} + \Delta G_{\text{homodimer 2}}}{2}$$

Specificity is then the difference between the additive and measured heterodimer ΔG .

$$\Delta\Delta G_{\text{spec}} = \Delta G_{\text{heterodimer}}^{\text{measured}} - \Delta G_{\text{heterodimer}}^{\text{additive}}$$

This metric is analogous to the coupling energy, which expresses the deviation of the measured DG for a double mutant from that expected given the DGs of two single mutants assuming additivity (8-10).

Quantifying non-additive effect on specificity between Anca α and Anc β . The non-additive effect on specificity can be defined as the difference between the predict and measured $\Delta\Delta G$ of the derived complex Anca α + Anc β .

$$\Delta\Delta\Delta G = \Delta\Delta G_{\alpha+\beta} - (\Delta\Delta G_{\alpha} + \Delta\Delta G_{\beta}).$$

Prediction of monomer, dimer, and tetramer occupancies with no IF2 specificity.

The occupancy of monomers, dimers, and tetramers between 1 mM and 4 mM predicted was calculated as follows. The concentration of subunit in each stoichiometric species can be expressed in terms of the concentration of monomers [A] and [B]:

$$\begin{aligned} [A]_{\text{tot}} &= [A] + [AB] + 2[AA] + [ABBB] + 2[AABB] \\ &= [A] + \frac{[A][B]}{Kd_3} + \frac{2[A]^2}{Kd_1} + \frac{\frac{[A][B]^3}{Kd_2 * Kd_3}}{Kd_4} + \frac{\frac{2[A]^2[B]^2}{Kd_2^2}}{Kd_4} \end{aligned}$$

$$\begin{aligned} [B]_{\text{tot}} &= [B] + [AB] + 2[BB] + 2[AABB] + 3[ABBB] + 4[BBBB] \\ &= [B] + \frac{[A][B]}{Kd_3} + \frac{2[B]^2}{Kd_2} + \frac{\frac{2[A]^2[B]^2}{Kd_2^2}}{Kd_4} + \frac{\frac{3[A][B]^3}{Kd_2 * Kd_3}}{Kd_4} + \frac{\frac{4[B]^4}{Kd_3^2}}{Kd_4} \end{aligned}$$

We used these equations to predict [A] and [B] across a range of $[A]_{\text{tot}}$ and $[B]_{\text{tot}}$ values given previously measured equilibrium constants. Predicted [A] and [B] concentrations were used to calculate the concentration of homodimers and heterodimers as described above, and the concentration of tetramers were calculated using the following equations:

$$[BBBB] = \frac{\frac{[B]^4}{Kd_2^2}}{Kd_4}$$

$$[ABBB] = \frac{\frac{[A][B]^3}{Kd_2 * Kd_3}}{Kd_4}$$

$$[AABB] = \frac{\frac{[A]^2[B]^2}{Kd_3^2}}{Kd_4}$$

where [BBBB] corresponds to the concentration of homotetramer, [ABBB] is concentration of $\alpha_1\beta_3$ tetramers, and [AABB] is the concentration of $\alpha_2\beta_2$ heterotetramers.

Homology models. SWISS-Model was used to generate a structural model of the An $\alpha\beta_{q40W}$ homotetramer using the crystal structure of the human Hb β homotetramer (PDB 1CMB) as template, which was then refined using Rosetta's Fast Relax protocol, which energetically minimizes the initial structure via small adjustments to the backbone

and side chain torsion angles (11). PyMOL V2.1 was used to visualize the proteins and capture images.

IF1-mediated homodimers were generated by the same procedure, except for homodimers of An α or An $\alpha\beta_{AD}$, for which the homodimer of human Hb α (PDB 3S48) was used as template. IF1-mediated heterodimers were generated by the same procedure but using the heterotetramer of human Hb (PDB 4HHB). For PyMol PSE file containing these models, see Supplementary file.

Works cited

1. A. S. Pillai *et al.*, Origin of complexity in haemoglobin evolution. *Nature* **581**, 480-485 (2020).
2. Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-556 (1997).
3. J. Yang, Z. Zhang, X. A. Zhang, Q. Luo, A ligation-independent cloning method using nicking DNA endonuclease. *Biotechniques* **49**, 817-821 (2010).
4. C. Natarajan *et al.*, Expression and purification of recombinant hemoglobin in *Escherichia coli*. *PLoS One* **6**, e20176 (2011).
5. A. C. Leney, A. J. R. Heck, Native mass spectrometry: what is in the name? *Journal of the American Society for Mass Spectrometry* **28**, 5-13 (2016).
6. M. T. Marty *et al.*, Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Analytical Chemistry* **87**, 4370-4376 (2015).
7. P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261-272 (2020).
8. P. J. Carter, G. Winter, A. J. Wilkinson, A. R. Fersht, The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (*Bacillus stearothermophilus*). *Cell* **38**, 835-840 (1984).
9. A. Horovitz, Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Folding and Design* **1**, R121-R126 (1996).
10. J. A. Wells, Additivity of mutational effects in proteins. *Biochemistry* **29**, 8509-8517 (1990).
11. L. G. Nivón, R. Moretti, D. Baker, A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLOS ONE* **8**, e59004 (2013).