Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

Biophysical Journal
Article

Biophysical Society

Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin

Jaime E. Martinez Grundman,[1] Eric A. Johnson,[1] and Juliette T. J. Lecomte[1,*]
[1]T.C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland

ABSTRACT Artificial proteins representing the consensus of a set of homologous sequences have attracted attention for their increased thermodynamic stability and conserved activity. Here, we applied the consensus approach to a *b*-type heme-binding protein to inspect the contribution of a dissociable cofactor to enhanced stability and the chemical consequences of creating a generic heme environment. We targeted the group 1 truncated hemoglobin (TrHb1) subfamily of proteins for their small size (~120 residues) and ease of characterization. The primary structure, derived from a curated set of ~300 representative sequences, yielded a highly soluble consensus globin (cGlbN) enriched in acidic residues. Optical and NMR spectroscopies revealed high-affinity heme binding in the expected site and in two orientations. At neutral pH, proximal and distal iron coordination was achieved with a pair of histidine residues, as observed in some natural TrHb1s, and with labile ligation on the distal side. As opposed to studied TrHb1s, which undergo additional folding upon heme binding, cGlbN displayed the same extent of secondary structure whether the heme was associated with the protein or not. Denaturation required guanidine hydrochloride and showed that apo- and holoprotein unfolded in two transitions—the first (occurring with a midpoint of ~2 M) was shifted to higher denaturant concentration in the holoprotein (~3.7 M) and reflected stabilization due to heme binding, while the second transition (~6.2 M) was common to both forms. Thus, the consensus sequence stabilized the protein but exposed the existence of two separately cooperative subdomains within the globin architecture, masked as one single domain in TrHb1s with typical stabilities. The results suggested ways in which specific chemical or thermodynamic features may be controlled in artificial heme proteins.

SIGNIFICANCE For decades, studies of hemoglobins have produced fundamental knowledge in the field of protein biophysics. Yet, the relationships between primary structure and fold stability, and the contribution of the heme group to the latter, remain incompletely explained and difficult to generalize. In this work, we focused on group 1 truncated hemoglobins, an ancient branch of the superfamily, and generated an artificial protein that contains the most represented residues at each position in a large set of homologous sequences. Our structural and thermodynamic analysis of this consensus globin provides insights into the architecture of truncated hemoglobins and, more generally, *b*-type heme-binding proteins.

INTRODUCTION

In recent years, much interest has been devoted to "consensus" proteins (1–4). In the simplest form of consensus design, the primary structure is built by selecting the most frequent amino acid at each position in a well-curated multiple sequence alignment (MSA) of homologous proteins. Consensus globular proteins tend to have three characteristics: they fold into the same structure as the parent proteins, they have enhanced thermodynamic stability, and they conserve some level of activity. Although several examples of consensus proteins have been reported, few represent tight protein-cofactor or protein-prosthetic group complexes, leaving the effects of sequence "averaging" for many protein families incompletely defined. Proteins that naturally bind the dissociable *b*-type heme group (iron protoporphyrin IX or Fe-PPIX, Fig. 1 *A*) fall in this little-studied category.

As a test case to explore the consequences of a consensus primary structure on the properties of a heme protein, we chose the truncated hemoglobin (TrHb) family. TrHbs form a distinct lineage within the hemoglobin superfamily (6). They are widely distributed in bacteria, unicellular eukaryotes, and plants (7). Unlike their better-known animal

Submitted March 12, 2023, and accepted for publication June 20, 2023.
*Correspondence: lecomte_jtj@jhu.edu
Editor: Ronald Koder.
https://doi.org/10.1016/j.bpj.2023.06.016

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

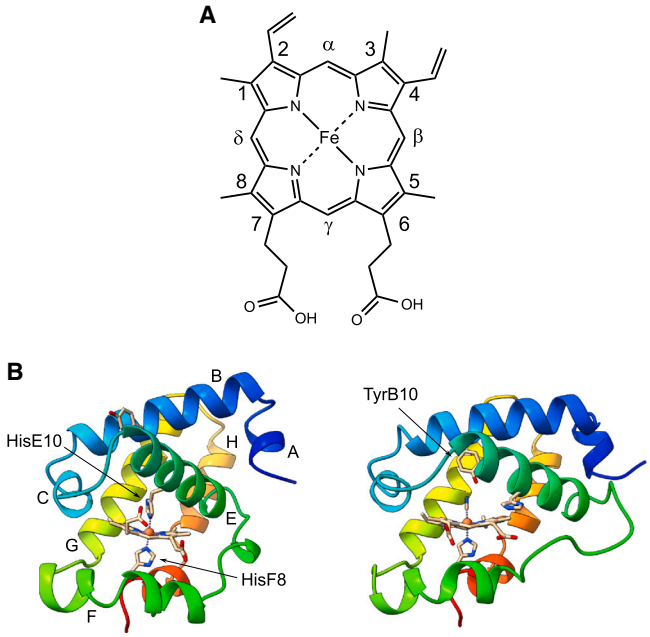

FIGURE 1 (*A*) Structure of the *b*-type heme. (*B*) Crystal structure of *Synechococcus* GlbN (5) shown in two conformations: the *bis*-histidine state (*left*, PDB: 4MAX) and the cyanide-bound state (*right*, PDB: 4L2M). Helical nomenclature (*A–C* and *E–H*) and residue positions (B10, E10, and F8) follow the myoglobin (Mb) convention. To see this figure in color, go online.

globin relatives, which adopt a 3-on-3 α-helical sandwich, TrHbs have a smaller 2-on-2 topology (8) (Fig. 1 *B*) and therefore their name. The studied specimens associate with a *b*-type heme and are generally capable of binding diatomic molecules (e.g., O_2, NO, CO) like their 3-on-3 counterparts. Many exist as monomers and perform enzymatic functions such as reactive nitrogen and oxygen species processing (9). Other physiological roles particular to the truncated family include gas sensing and redox reactions (10). The large number of TrHb sequences has allowed for a robust classification into four groups, historically labeled N to Q or 1 to 4 (9). The ancient origin of the family (7), the phylogenetic diversity of available sequences (7,9), and the simplification of the fold make TrHbs excellent subjects for analysis. We specifically targeted the consensus design to group 1 or N TrHbs (TrHb1s) (6,9,11), one of four TrHb branches for which experimental and computational data are available.

As in known functional globins, TrHbs bind heme using a histidine in the F helix (position F8, proximal). Besides this residue, each TrHb branch has strongly conserved sequence features (9,11). In TrHb1s, they include a Gly-Gly pair between the A and B helices, a second Gly-Gly pair at the beginning of the EF turn, an Asp N-capping the E helix, and a His N-capping the G helix. TrHb1s can be further split into two phylogenetic clades, subgroups 1 and 2 (11), the former of which has the most studied proteins. Subgroup 1 has two distinctive side chains on the distal side, namely TyrB10 and GlnE11. These residues form a stabilizing hydrogen bond network for exogenous ligands (8), TyrB10 interacting directly with dioxygen (Fe(II)-PPIX) or cyanide (Fe(III)-PPIX). In addition and in contrast to the 3-on-3 proteins, TrHb1s present tunnels (12), one of which is gated (13). These cavities are thought to facilitate ligand access to the distal side of the heme group.

The practical goals of this work were to 1) prepare cGlbN, a consensus TrHb1, 2) study its ability to bind heme, and the ability of the heme to bind diatomic ligands, 3) describe the fold without and with heme, and 4) assess the stability of the fold with denaturation experiments. With this system, we sought to determine how an increase in stability conveyed by the artificial sequence manifests itself in the two extreme and essential states assumed by a heme protein, i.e., with (holoprotein) and without (apoprotein) its prosthetic group. We also intended to define the heme-protein interactions obtained when averaging the active site and describe how the properties of individual existing proteins differ from those afforded by the generic consensus environment.

Besides informing on the TrHb1 subfamily and adding to the repertory of consensus proteins, cGlbN offers insights relevant to other *b*-type heme proteins. In cold-adapted globins, high conformational flexibility in the EF loop and helices B and E has been linked to decreased stability (14), whereas computational analyses of 3-on-3 globins have implicated high conformational flexibility of the CD corner and helices C and D in higher thermal stability (15). The presence of TrHbs in extremophiles adds urgency to a clarification of the relationship between primary structure and of apo- and holoprotein conformational dynamics as well as thermodynamic stability (16). Such clarification will be helpful to fashion artificial heme proteins into functional complexes intended to manage ligand storage, enzyme activity, and redox chemistry (17–21).

MATERIALS AND METHODS

Multiple sequence analysis and consensus design

The sequence of the TrHb1 from *Synechococcus* sp. PCC 7002, hereafter *Synechococcus* GlbN (UniProtKB: Q8RT58), was used to perform a BLAST search (22), from which a total of 1565 homologous TrHb1s were obtained (accession December 2017). After alignment with MAFFT (23), sequences shorter than 90 residues, not covering the entire TrHb1 domain length, lacking the proximal histidine, or having more than 90% pairwise identity with any other in the set were excluded. In addition, only sequences adhering to TrHb1 subgroup 1 definition, i.e., containing TyrB10 and the Gly-Gly motif at homologous positions between the A and B helices (11), were retained. The final set of 341 sequences was realigned with Clustal Omega (24,25). N- and C-terminal extensions and loop insertions not aligning to the conserved TrHb1 domain were manually trimmed using Jalview (26), generating a final MSA consisting of 341 sequences (rows) and 120 positions (columns). The percent composition of each amino acid *aa* was computed as $p_{aa,i} = 100 \times (n_{aa,i}/N_i)$ for every sequence *i* in the MSA, where $n_{aa,i}$ is the number of *aa* residues and N_i is

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

the total number of residues including gaps. The sequence of cGlbN was generated with the most frequent amino acid at each of the 120 MSA columns. A comprehensive analysis of a more recent (January 2021) set of 3024 TrHb1 sequences supported the subgroup distinction applied during MSA curation (supporting methods and Fig. S1).

Protein preparation

The gene for cGlbN with codons optimized for expression in *Escherichia coli* was synthesized by ATUM (Newark, CA) and delivered in the vector pJExpress414 (inducible T7 promoter; ampicillin resistance). The original sequence contained an Asp28-Pro29 dyad that was susceptible to cleavage during preparation (Fig. S2). To prevent proteolysis, Pro29 was replaced with Asp, the next most frequent amino acid at position 29 (26.1% compared with 27.9% for Pro; see supporting methods). The preparation and purification of cGlbN involved overexpression in *E. coli* BL21(DE3) cells as per established protocols (27–29), except that fully soluble cGlbN protein was obtained from fractions released by osmotic shock without cell lysis (Fig. S3). Chromatographic separation yielded pure apo-cGlbN as per the absence of a Soret band (∼400 nm) in electronic absorption spectra, and reconstitution with excess heme generated holo-cGlbN (see supporting methods for details).

Purified apo- and holo-cGlbN samples were flash-frozen in liquid nitrogen, lyophilized, and stored at −20°C for further use. Intact protein ultra-performance liquid chromatography-mass spectrometry (UPLC-MS) was performed on an Acquity/Xevo-G2 system and processed using BiopharmaLynx (Waters, Milford, MA). The deconvoluted mass of 13,210.9 Da obtained with unlabeled protein was consistent with initial Met cleavage during overexpression in *E. coli*. Isotopically labeled proteins had the expected masses.

Electronic absorption spectroscopy

A Varian Cary50 spectrophotometer (Agilent, Santa Clara, CA) was used for most electronic absorption data collected at room temperature. UV-vis spectra were recorded over the 250–750 nm range in 1-nm steps and with 0.1 s/nm averaging time. Holoprotein spectra as a function of temperature were collected on an Aviv 14DS spectrophotometer (Aviv Biomedical, Lakewood, NJ) equipped with a Peltier temperature-controlled cuvette holder, spanning 260–750 nm in 1-nm steps and 0.2 s/nm averaging time. Unless otherwise noted, typical samples were prepared in 20–25 mM potassium phosphate buffer (pH 7), with protein concentrations at 5–10 μM in 1-cm quartz cuvettes.

Circular dichroism spectroscopy

Aviv circular dichroism (CD) spectrometers models 410 or 420 with Peltier temperature control were used for all experiments. Typical far-UV spectra spanned 200–300 nm in 1-nm steps and 1–3 s/nm averaging time, using quartz cuvettes with 0.2-cm pathlength and protein concentrations of ∼20 μM, except for the variable temperature data and denaturant titrations, which consisted of ∼5 μM protein samples in 1-cm cuvettes with stir bars. Near-UV CD spectra spanned 250–350 nm in 0.5-nm steps with 1-s averaging time and five scans per spectrum for ∼92 μM protein samples in 1-cm cuvettes with stir bars. Owing to the low signal/noise ratio of the aromatic signal in the near-UV spectra, a Savitzky-Golay filter (30) was applied for smoothing of blank and sample spectra with a polynomial order of 3 and smoothing window of 5 nm. All CD spectra shown in this work were deposited into the Protein Circular Dichroism Data Bank (PCDDB), with accession codes indicated in each figure legend.

Extinction coefficient determination

The extinction coefficient of holo-cGlbN was determined on a heme basis with the pyridine hemochromogen assay (31–33). Triplicate measurements yielded $\varepsilon_{409nm} = 117 \pm 1$ mM^{-1} cm^{-1} at the maximum of the Soret band for the Fe(III) (i.e., ferric) form at pH 7. The extinction coefficient of apo-cGlbN was initially determined with the Edelhoch method (34) at 6 M guanidine-HCl (GdnHCl, UltraPure grade, Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) and using published residual values (35,36). However, incomplete unfolding and heme addition experiments indicated a >10% error; scaled spectra of apoprotein with added heme relative to that of pure holoprotein resulted in a corrected $\varepsilon_{279nm} = 5.29 \pm 0.08$ mM^{-1} cm^{-1}.

Heme binding and holoprotein properties

A heme titration of apo-cGlbN was performed by UV-vis spectroscopy with both titrant and protein samples containing 25 mM caffeine to maintain unbound heme in a soluble monomeric form (37–39). The protein concentration used for adequate sensitivity was 8 μM. Accurate heme concentrations in the caffeine-containing buffer were determined using $\varepsilon_{403nm} = 86.0 \pm 1.5$ mM^{-1} cm^{-1} by hemochromogen assay. Equilibrium was ensured at each point by unchanged spectra after 10 min. Difference spectra were obtained by subtracting the calculated total free heme spectrum corrected for sample dilution, and the extent of binding was monitored at the Soret band. The affinity was too high to be determined by this method or multivariate analysis (40).

A heme release kinetic experiment by UV-vis spectroscopy was performed using excess apomyoglobin (apoMb) prepared with the acid butanone method (41,42) from commercial horse heart Mb (Sigma). Samples were in 50 mM Tris-HCl, 1 mM EDTA (pH 7.0) buffer, with ∼4.7 μM holo-cGlbN and ∼50 μM apoMb. Singular value decomposition (SVD) was applied to the spectral data set, and a sum of two exponentials was globally fit to two resulting significant time-dependent vectors to determine apparent heme dissociation rate constants. The calculated final spectrum obtained from the SVD basis spectra yielded the expected Soret extinction difference as well as optical features in the visible region of aquomet holoMb. A control experiment of cGlbN under the same conditions without apoMb was stable over the course of several days (data not shown).

Far-UV CD spectral changes upon heme incorporation to apo-cGlbN were measured in a ∼20 μM sample into which an equimolar amount of hemin chloride solubilized in 0.1 M NaOH was added and equilibrated at 25°C for 15 min. The resulting holoprotein spectra were corrected for sample dilution.

Binding of exogenous ligands to ferric holo-cGlbN was determined by addition of small amounts of concentrated solutions (e.g., 1 M KCN) to UV-vis samples. For binding to Fe(II) (i.e., ferrous) holo-cGlbN, samples were reduced by addition of freshly prepared sodium dithionite (Sigma-Aldrich) to final concentrations of 0.5–2 mM and left to equilibrate until unchanging electronic absorption spectra were observed. Dioxygen binding was assessed by bubbling in 99.9% O_2 gas directly into ferrous samples for 2 min before data collection.

The ferric holoprotein was also subjected to varying conditions of pH, salt (e.g., KCl, $MgSO_4$), and denaturant (e.g., urea [J.T. Baker], GdnHCl) by preparing phosphate-buffered samples containing the desired additives. A pH titration by UV-vis spectroscopy covering the pH 5–12 range was carried out, and the resulting pH-dependent SVD vectors were globally fitted to extract two apparent pK_a values as described previously (43,44).

Thermodynamic stability measurements

Variable temperature CD data of both apo- and holo-cGlbN consisted of spectra collected from 25 to 95°C at intervals of 5°C. Samples were heated at a rate of 5°C/min, then equilibrated with stirring for 5 min before data collection. Variable temperature UV-vis spectra of ferric holo-cGlbN were also collected over the 25–95°C range in 10°C intervals, 2°C/min heating, and equilibration with stirring for 5 min before data collection. Reversibility was assessed by cooling to 25°C and comparing initial and final CD or UV-vis spectra.

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

Thermodynamic stability parameters were obtained by chemical denaturation using GdnHCl. Preliminary experiments with an automated syringe titrator that combined sample and titrant solutions within a single cuvette were performed with equilibration times of 5–10 min. Holo-cGlbN unfolding curves so obtained showed three denaturation steps and were found to be dependent on equilibration time. To obtain valid thermodynamic data, manual titrations were performed by dispensing predetermined buffer and denaturant stock volumes into glass test tubes using a MicroLab 600 Series automated liquid syringe handler (Hamilton). Apo- or holoprotein was then added into each tube for a final concentration of ~4.5 μM. All GdnHCl concentrations were verified by refractometry measurements (45). Each sample along the 0–8 M GdnHCl range was allowed to equilibrate at room temperature for 40 h followed by CD signal recording at 222 nm with 30 s signal averaging. Apo- and holo-cGlbN exhibited apparent two-step denaturation reactions. For the holo-cGlbN samples, UV-vis spectra were also collected to follow the heme signal.

A three-state model was fitted to the unfolding curve of apo-cGlbN monitored by CD, as in classic studies of apoMb (46,47). This includes the native (N), partially unfolded intermediate (I), and fully unfolded (U) forms as shown in Fig. 2 *A*. Nonlinear least-squares regression was allowed to vary the following parameters (see supporting methods for details): the equilibrium folding constants K_{UI}^0 and K_{IN}^0 for the U $\rightarrow$ I and I $\rightarrow$ N processes, respectively, extrapolated to zero denaturant concentration, the linear dependencies m_{UI} and m_{IN} of the associated free energies on denaturant concentration, and the pure signal intensities S_U, S_I, and S_N, of which S_I was allowed to depend linearly with denaturant concentration.

A similar two-step unfolding curve for holo-cGlbN by CD was interpreted to consist of a four-state scheme with a single native heme-bound state NH in equilibrium with free heme and the heme-less N state (Fig. 2 *B*). The resulting model (see supporting methods) contained fixed stability and signal parameters of the N, I, and U states obtained independently from the apoprotein data and the following fitting parameters: the equilibrium dissociation constant K_{NH}^0 for the NH $\rightarrow$ N + H process extrapolated to zero denaturant, the associated m_{NH} value for the free energy linear dependency, and the pure signal intensity of the NH state, S_{NH}, with a linear dependence on [GdnHCl]. A similar model has been described for holoMb unfolding involving a complex six-state mechanism that includes both free and fixed equilibrium heme binding parameters for the I and U states (48). Modeling these potential IH and UH states to our data did not yield significant differences to predicted fitted parameters, and we did not pursue them given a lack of evidence for IH or UH population at any significant level in either the CD or UV-vis spectra. All regression fitting procedures were carried out in Mathematica (Wolfram, Champaign, IL), with weights included as the inverse of the standard deviations from signal averaging at each data point.

NMR spectroscopy

NMR samples were prepared from lyophilized proteins resuspended to concentrations of 0.2–2 mM in potassium phosphate buffer (pH 6.3–7.5), and either 10 or 99% ^2H_2O. Apo-cGlbN samples included 0.02–0.05% NaN_3, and cyanide-bound ferric holo-cGlbN (i.e., cyanomet) samples included 5–7.2 mM KCN. Variable pH spectra were collected on ferric holo-cGlbN samples in phosphate buffer to which small amounts of 1 M HCl or KOH were added to obtain the desired pH.

A $$\mathrm{U} \underset{}{\overset{K_{UI}}{\rightleftharpoons}} \mathrm{I} \underset{}{\overset{K_{IN}}{\rightleftharpoons}} \mathrm{N}$$

B $$\mathrm{U} + \mathrm{H} \overset{K_{UI}}{\rightleftharpoons} \mathrm{I} + \mathrm{H} \overset{K_{IN}}{\rightleftharpoons} \mathrm{N} + \mathrm{H} \underset{K_{NH}}{\rightleftharpoons} \mathrm{NH}$$

FIGURE 2 Denaturation scheme for (*A*) the apoprotein and (*B*) the holoprotein.

All data were obtained on either a Bruker (Billerica, MA) Avance II 600 MHz or Ascend 800 MHz spectrometer, equipped with triple-resonance cryoprobes. Two-dimensional (2D) ^1H–^1H experiments for assignment of heme resonances for unlabeled holo-cGlbN samples included nuclear Overhauser effect spectroscopy with τ_{mix} = 75–80 ms, double-quantum-filtered correlated spectroscopy, and total correlation spectroscopy with τ_{mix} = 45 ms. 1D [^{15}N,^1H]-TRACT data (49) were collected and analyzed as described previously (50). A standard set of heteronuclear spectra was collected using nonuniform sampling (51) in uniformly ^{15}N,^{13}C isotope-labeled apo- and holo-cGlbN samples for backbone and side-chain assignments (Table S1). Chemical shifts were directly and indirectly referenced to the ^1H_2O signal (4.77 ppm at 25°C). 1D data were analyzed in TopSpin (Bruker), while 2D and 3D data were processed using NMRPipe (52). Nonuniform sampling data were reconstructed using SMILE (53), and the resulting spectra were analyzed with CARA (54), Sparky (55), or CcpNmr Analysis Assign v3 (56) locally or within the NMRbox virtual machine (57). Chemical shift assignments were deposited to the Biological Magnetic Resonance Data Bank, with accession numbers BMRB: 51849, 51848, 51847 for apo-cGlbN, ferric *bis*-His holo-cGlbN, and cyanomet holo-cGlbN, respectively.

Structure modeling

Initial homology-based structural models of cGlbN (without heme) were generated with SWISS-MODEL (58), using PDB: 1S6A (*Synechocystis* sp. PCC 6803 GlbN, or *Synechocystis* GlbN hereafter, with azide ligand and covalently attached heme), and PDB: 1RTX (same as PDB: 1S6A, but in the *bis*-His state). AlphaFold2 (59) was also used to generate models, all of which matched closely the SWISS models of TrHb1s with bound heme and bound exogenous ligands. The deposited backbone chemical shifts of the apoprotein were used by the Biological Magnetic Resonance Data Bank to generate heme-free structures with CS-Rosetta (60).

RESULTS

Group 1 TrHb features

We chose TrHb1s as a test case of consensus heme protein because of the small size of the domain, the availability of diverse sequences, and a suitable body of experimental knowledge. Among already studied proteins, *Synechococcus* GlbN provides a convenient reference for structure and stability comparison (5,28,50,61,62). *Synechococcus* GlbN and its relatives are monomeric, contain a single *b*-type heme (Fig. 1 *A*), and display a structure composed of seven helices (A, B, C, E, F, G, and H) as shown in Fig. 1 *B*. Key features of a well-folded TrHb1 include a short, loosely packed A helix (63), a long loop connecting the E and F helices (5), a distorted F helix, and a kinked H helix. In several GlbN relatives, the heme iron has a coordination bond to the proximal histidine (F8 in the standard myoglobin [Mb] nomenclature) and a histidine or lysine occupying position E10 on the distal side (forming an endogenous hexacoordinate or 6c complex, Fig. 1 *B*, *left*) (64,65). The distal ligand is labile, readily replaced with dioxygen in the ferrous state or cyanide in the ferric state (forming an exogenous 6c complex, Fig. 1 *B*, *right*). However, not all TrHb1s exhibit endogenous hexacoordination when position E10 is occupied by a potential iron ligand. The same ambiguity applies to a consensus sequence, and

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

thus the determination of iron coordination is an essential component of the characterization.

Consensus sequence features

The purposeful curation of the MSA described in materials and methods gives rise to a consensus sequence (Fig. 3 *A*) with features common to subgroup 1 of TrHb1s: a Gly-Gly pair at the beginning of the EF turn, an Asp N-capping the E helix, and a His N-cap to the G helix. Position E10 is occupied by a His, a potential ligand to the iron; Lys is a close second. Also present are distal residues GlnE7 and GlnE11, presumably suitable for interaction with the MSA-imposed TyrB10 and a bound exogenous ligand at the distal site.

The average percent composition over all sequences of the curated MSA for each amino acid, or $p_{aa,\mathrm{MSA}}$ (Table S2 and Fig. S4), differs only slightly from the corresponding value in average bacterial proteomes (67): at most a 2–3% increase of Ala and Asp residues, and similar depletion of Pro and Ser residues. However, the sequence of cGlbN was significantly (>1 SD) enriched in Glu (+4.7%) and Arg (+3.1%), and depleted in Gln (−2.4%), when compared with $p_{aa,\mathrm{MSA}}$ values (Fig. 3 *B*). An expected consequence of the overall increase in charged residues is a highly soluble protein, which so far we have been unable to crystallize. In addition, the overall enrichment of acidic residues (+5.9%) over basic residues (+0.5%) results in a protein with a predicted acidic isoelectric point (∼4.6) at the low end of the bimodal pI distribution of proteomes from *Synechococcus* sp. PCC 7002 (calculated with ExPASy and confirmed in the Proteome-pI 2.0 database (68)), *Synechocystis* sp. PCC 6803, and *E. coli* K12 (69).

cGlbN preparation

Unlike the TrHb1s we have previously prepared (27–29), overexpression of the gene coding for cGlbN in *E. coli* resulted in a protein product released by osmotic shock without requiring cell lysis (see materials and methods). This property had the advantage of limiting contamination from other cellular components. We also found that the consensus sequence strictly derived from the MSA was prone to fragmentation. UPLC-MS of purified samples showed the presence of two polypeptides, one spanning the full sequence length minus the initial Met and the other starting at Pro29 (Fig. S2). We attribute this heterogeneity to the lability of the Asp28-Pro29 bond and its hydrolysis, either enzymatically in the cell or otherwise in vitro (70–72). To prevent this degradation, the consensus sequence was modified to contain Asp29, the second most abundant amino acid at that position based on the MSA. No fragmentation was observed after this substitution was made. Data collected on the original sequence and the modified one did not reveal differences in physicochemical behavior, but the observation illustrates the perils of unwittingly introducing detrimental features in the primary structure.

Heme-protein interactions

Heme binding and release

Although most of the cGlbN extracted from *E. coli* was in the apoprotein form, a slight red color in the periplasmic preparations indicated *in vivo* heme binding. From these extracts, pure apo-cGlbN was separated by chromatographic methods (Fig. 4 *A*, *black trace*). To assess heme affinity, we performed a heme titration monitored by UV-vis spectroscopy (Fig. 4 *B*). The sharp break in the curve at a 1:1 ratio of added heme to apoprotein supported tight binding, with an equilibrium heme dissociation constant far below that of the protein concentration of the experiment (∼8 μM) and an upper limit of nM that is consistent with the shape of the curve.

Periplasmic fractions containing native heme were fully reconstituted with hemin chloride to generate pure holo-cGlbN (Fig. 4 *A*, *red trace*). Heme dissociation from the polypeptide matrix was investigated by exposing the ferric holoprotein to an excess of apoMb (Fig. 4 *C*) used as a

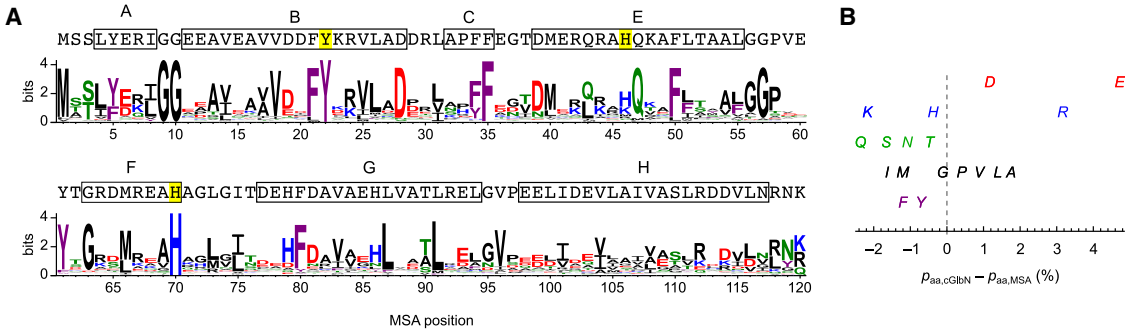

FIGURE 3 (*A*) Sequence of cGlbN and logo plot (66) of the MSA used to generate it. The labeled boxes correspond to the α-helices predicted from homology modeling. Positions B10, E10, and F8 are highlighted. Note the Pro29Asp replacement introduced in the cGlbN sequence (see text). (*B*) Graphical representation of amino acid percent composition in cGlbN, $p_{aa,\mathrm{cGlbN}}$, minus the average amino acid percent composition for all sequences in the MSA, $p_{aa,\mathrm{MSA}}$. Colors in (*A* and *B*) are red, acidic; blue, basic; purple, aromatic; green, polar; black, nonpolar. To see this figure in color, go online.

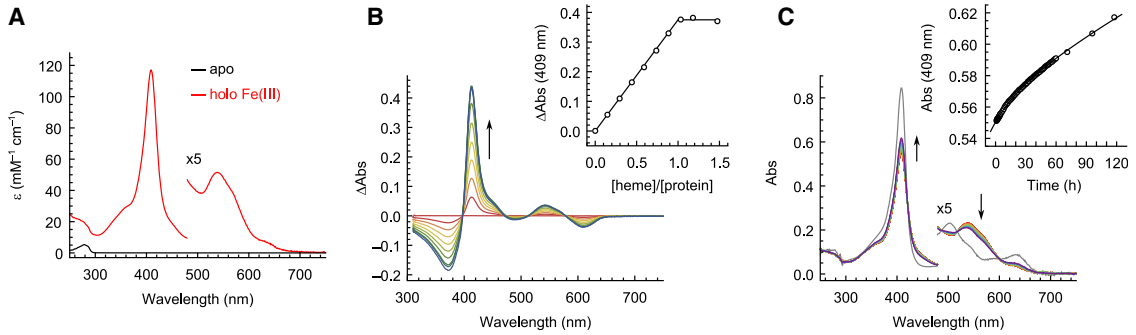

FIGURE 4 (*A*) UV-vis spectra of pure apo- (*black*) and ferric holo-cGlbN (*red*) at pH 7. (*B*) Equilibrium heme titration of apo-cGlbN (∼8 μM) in the presence of 25 mM caffeine, shown as difference spectra relative to the added heme contribution. Traces are colored from red to blue with increasing heme additions (*black arrow*). Inset: Soret absorbance intensity versus heme-to-protein ratio. The solid line is drawn to guide the eye. (*C*) Heme transfer from ferric holo-cGlbN (∼5 μM) to apoMb (10-fold excess). Traces are colored from red to blue as a function of time (*black arrows*). The calculated spectrum of fully reconstituted holoMb after SVD and global fitting is shown in gray. Inset: Soret absorbance intensity versus time. The solid line is the result of the global fit. To see this figure in color, go online.

high-affinity heme acceptor (73). UV-vis spectral changes were slow to develop and biphasic, and after SVD analysis and global fitting, the calculated final spectrum was consistent with the expected water-bound ferric (i.e., aquomet) holoMb species. The minor phase accounted for ∼5% of the signal and had a rate constant of $(1.3 \pm 0.1) \times 10^{-5}$ s^{-1}. The slower of the two phases, at $(5.0 \pm 0.1) \times 10^{-7}$ s^{-1} and with a signal amplitude contribution of ∼95%, was taken to be the effective heme dissociation rate constant (k_{off}). This constant is slightly slower than that measured for *Synechocystis* GlbN ($(\sim 3.0 \pm 0.1) \times 10^{-6}$ s^{-1} (74)).

Apoglobins, like many apoproteins meant to associate with a *b*-type heme, appear incompletely folded under native conditions. For those proteins, heme binding often causes some extent of additional folding (75). cGlbN was probed for this effect by measuring the far-UV CD spectrum of an apoprotein sample before and after addition of an equimolar amount of hemin chloride (Fig. 5). The initial apoprotein far-UV CD spectrum had strong signals at 208 and 222 nm. Surprisingly, a slight but reproducible *decrease* in negative value at 222 nm was observed upon heme addition. An estimate based on mean residue ellipticity (MRE) indicated that ∼55% of the residues are involved in α-helices in both apo and holo forms. Thus, the consensus sequence appeared to favor the formation of secondary structure independently of heme binding, in contrast to most existing globins.

Axial heme coordination

The UV-vis absorption spectrum of ferric holo-cGlbN without added exogenous ligand (Fig. 4 *A*) exhibited a Soret band (117 mM^{-1} cm^{-1} at 409 nm) and α-β bands (shoulder at ∼570 nm, peak at 538 nm) consistent with a low-spin (LS) ferric heme and therefore endogenous hexacoordination with two strong-field ligands. These features resemble those of the cyanobacterial GlbNs from *Synechococcus* (28) and *Synechocystis* (76), which show iron coordination with the proximal His70 (F8) and the distal His46 (E10) (i.e., 6c *bis*-His). To determine unambiguously the identity of the ligands in cGlbN, we collected NMR spectra using uniformly ^{15}N,^{13}C-labeled protein in the ferric form. Sequential assignments were obtained with the standard suite of triple-resonance experiments. Of note is the unique Ala45-His46-Gln47 stretch with a distinctively shifted histidine Cα signal at 82 ppm (Fig. 6 *A*), similar to the distal histidine Cα signal in both cyanobacterial GlbNs mentioned above (82.1 ppm (77) and 77.7 ppm (78), respectively). In cGlbN, the proximal His70 is embedded in a unique Ala-His-Ala sequence and also has a distinctively shifted Cα signal (71 ppm, not shown) in support of the *bis*-His scheme.

Not included in Fig. 6 *A* are signals from a second form of the protein. In fact, all holo-cGlbN NMR spectra contained

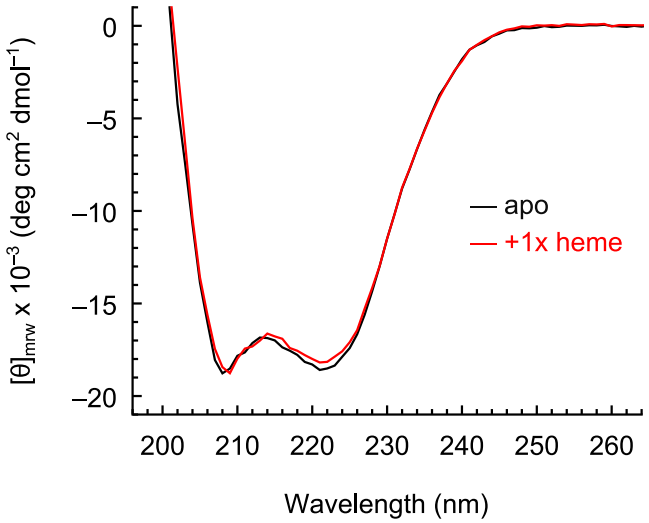

FIGURE 5 Far-UV CD spectra of a sample of apo-cGlbN before (*black*) and after (*red*) addition of an equimolar amount of ferric heme. Sample conditions were 25 mM potassium phosphate, pH 7. CD spectra are available at PCDCB: CD0006428000 (apo) and CD0006428003 (holo). To see this figure in color, go online.

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

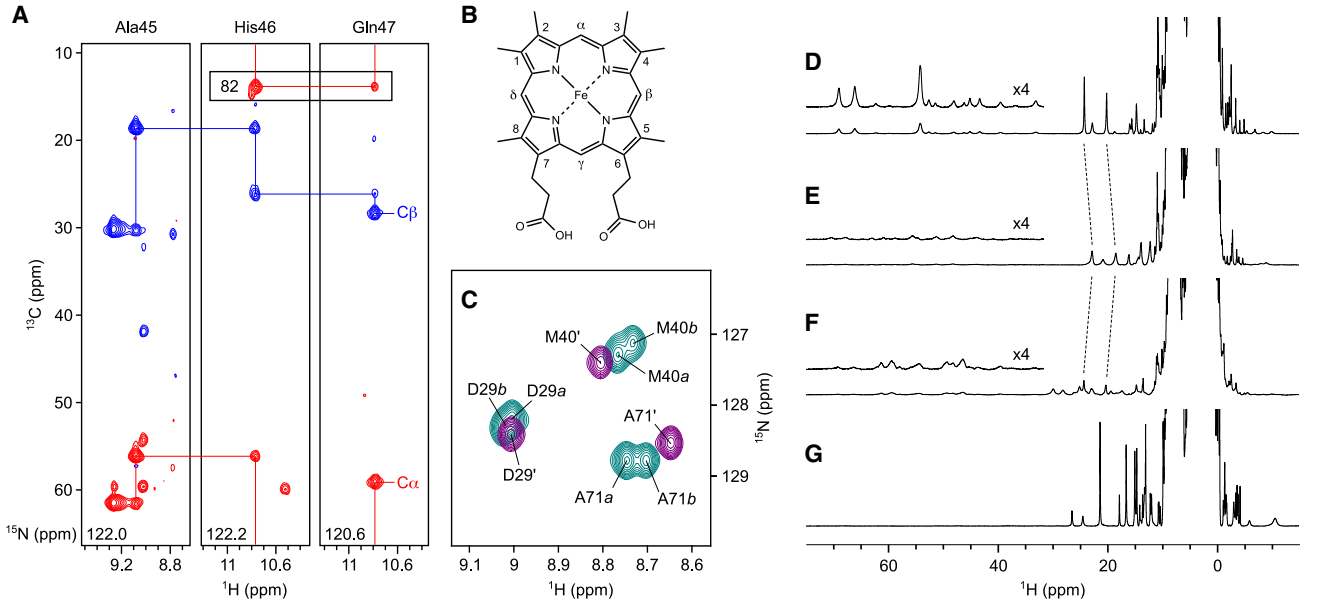

FIGURE 6 (*A*) 2D strip plot of HNCACB spectrum displaying the Ala45-His46-Gln47 stretch of one heme isoform of ferric cGlbN (sample conditions in Table S1). The His46 Cα resonance is folded along the ^{13}C dimension. (*B*) Structure of symmetric heme analogue 2,4-dimethyl-deuteroheme or DMDH. (*C*) Portion of ^{1}H-^{15}N HSQC NMR spectra of cyanide-bound ferric cGlbN reconstituted with *b*-type heme (*teal*) and DMDH (*purple*). Crosspeaks are indicated with *a* or *b* for heme isomerism or with an apostrophe in the DMDH-holoprotein. (*D–G*) ^{1}H NMR spectra of ferric cGlbN under different conditions: (*D*) pH 7.2 in 20 mM potassium phosphate and 10% $^{2}H_2O$; (*E*) similar to (*D*) except in 5 mM phosphate and pH-adjusted to pH 5.2 with HCl; (*F*) similar to (*E*) except at pH 9.8 with KOH; (*G*) same sample as in (*D*) after threefold excess KCN added. The dashed lines in (*D–F*) track two resonances from the LS *bis*-His form. To see this figure in color, go online.

several doubled heme and protein resonances. Reconstitution of cGlbN with a C2 symmetric iron-porphyrin (2,4-dimethyl-deuteroheme or DMDH, Fig. 6 *B*) resulted in spectra consistent with a single heme-bound species (Figs. 6 *C* and S5). The observations confirmed that spectral heterogeneity in holo-cGlbN spectra was caused by indiscriminate insertion of the heme group in two orientations closely related by a 180° rotation about the α-γ meso axis (79). Full ^{1}H-^{15}N HSQC spectra comparing holo-cGlbN with heme *b* versus DMDH are shown in Fig. S6 (ferric native forms) and Fig. S7 (ferric cyanide-bound forms).

Distal histidine pH-lability and ligand binding

A low-intensity shoulder was observed at ∼630 nm in UV-vis spectra of ferric cGlbN under neutral pH conditions (Fig. 4 *A*). This charge-transfer transition is typically attributed to a high-spin (HS) heme in which the distal His is displaced and the sixth coordination site either remains vacant (i.e., pentacoordinate, 5c) or harbors a water molecule. A small population of HS species was apparent in the ^{1}H NMR spectrum as well, which displayed broad and weak signals in the 70 to 30 ppm range (Fig. 6 *D*), in addition to strong LS signals in the 28 to 12 ppm and −2 to −14 ppm ranges attributed to the dominant *bis*-His form.

Functional heme proteins have roles conditioned by their coordination schemes. For example, redox proteins such as cytochrome b_5 maintain *bis*-His ligation even in the presence of exogenous ligands, whereas *Synechococcus* GlbN, a likely detoxification enzyme (61), alternates between endogenous and exogenous hexacoordination. To situate the properties of cGlbN among *bis*-His proteins, we subjected ferric cGlbN to different pH conditions as a probe of coordination lability. Lowering the pH from neutral to ∼5 broadened the ^{1}H NMR spectrum and gradually eliminated the HS signals (Fig. 6 *E*). Raising the solution pH decreased the intensity of both HS and LS resonances and produced a new set of lines (30–15 ppm) consistent with a hydroxide-bound ferric (i.e., hydroxymet) form (Fig. 6 *F*).

The pH response was also monitored by UV-vis spectroscopy (Fig. 7). Although acid-induced protein precipitation and heme loss occurred readily at pH < 5 (not shown), no such heme loss was observed up to pH 12. Analysis of the absorbance data of the intact holoprotein by SVD followed by nonlinear regression, as shown in Fig. S8, indicated a fully populated *bis*-His complex at low pH with a transition ($pK_{app,1} = 6.6$, $n_1 = 1.0$) to a mixture of *bis*-His and HS species at pH 7–9. Upon increasing the pH, the protein transitioned to a fully populated hydroxymet species ($pK_{app,2} = 10.2, n_2 = 1.0$). Based on linear combinations of the SVD-extracted component spectra, a contribution of ∼10% HS species was estimated at pH 7. Natural *bis*-His TrHb1s display a different behavior; for example, *Synechocystis* GlbN populates the *bis*-His state over a broad range of pH and switches to an HS state as the protein unfolds and releases heme at acidic pH (80). In cGlbN, this transition appears shifted to a pH at which the solubility is too low for observation.

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

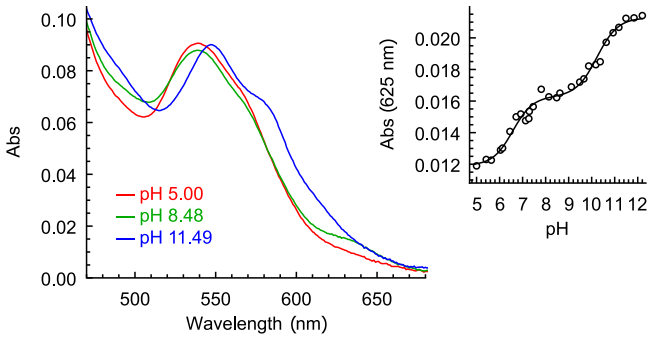

FIGURE 7 Absorbance spectra of ferric cGlbN in the visible wavelength range at the indicated pH values, representing the basis set from the titration data. Inset: absorbance at 625 nm as a function of pH with solid line generated by fixing the pK_{app} and n values obtained from the globally fitted vectors from SVD (see Fig. S8 for more details). To see this figure in color, go online.

The ability of added exogenous ligands to displace the distal His was also investigated. The cyanide-bound ferric (i.e., cyanomet) species was readily produced by addition of excess KCN and yielded a typical LS ^{1}H NMR spectrum (81), shown in Fig. 6 *G* to illustrate the hyperfine shifted lines. The UV-vis spectrum was also consistent with the formation of a cyanide complex in both redox states of the iron (Fig. S9 *A*). Reduction of ferric cGlbN to ferrous cGlbN using dithionite generated a mixed HS (5c) and LS (6c, likely *bis*-His) spectrum, suggesting weak distal His46 affinity in the ferrous state. Full conversion to an oxy spectrum was observed upon exposure to O_2 gas (Fig. S9 *B*). These observations are consistent with the documented ligand-binding behavior of other TrHb1s.

Fold determination

The evidence presented thus far is consistent with a helical protein able to bind the heme tightly and to coordinate it with two histidine residues, albeit with a low distal affinity. The next level of characterization was to compare the fold with that of known TrHb1s, with special attention to the distinctive structural features mentioned above. To this end, we applied NMR spectroscopy to three forms of cGlbN: apoprotein, ferric holoprotein, and cyanomet holoprotein.

Apo-cGlbN

Initial spectra of uniformly ^{15}N,^{13}C-labeled apo-cGlbN contained doubled resonances of uneven intensities and linewidths, unlike a well-folded, rigid protein. TRACT experiments (49) returned a correlation time of ∼6 ns (Fig. S10), a reasonable value for a 120-residue globular protein, eliminating the possibility of multimer formation. Intact protein UPLC-MS of the same sample returned a single molecular weight corresponding to the intact protein, which suggested that resonance multiplication was caused by kinetically trapped conformations or species exchanging slowly on the chemical shift timescale. To address the first possibility, we subjected the apoprotein to denaturation in 8 M GdnHCl followed by refolding. The annealing procedure resulted in improved spectral quality as shown by the ^{1}H-^{15}N HSQC in Fig. S11 but did not fully eliminate heterogeneity. Nonetheless, partial backbone assignments (∼75%) were obtained with standard 3D heteronuclear spectra. Chemical shift data, analyzed with TALOS+ (82), confirmed the location of most of the α-helices predicted from TrHb1 homology (Fig. 8 *A*). Regions of the apoprotein that were not assigned encompassed the EF loop, F helix, and part of the H helix. The stretches leading to these regions have low predicted order parameters (not shown). This pattern is common to other apoproteins, including apoMb (83), which maintains a well-folded distal side to the heme cavity but has fluctuating F helix and abutting end of H helix. The tertiary structure of apo-cGlbN was difficult to determine because of the heterogeneity mentioned above. Unambiguous ^{1}H–^{1}H nuclear Overhauser effects (NOEs) implicated Val25 (B13) in long-range interactions (Fig. S12), but overall, the spectra were indicative of a fluctuating, loosely packed structure.

Holo-cGlbN in the ferric bis*-His state*

The backbone assignments of ferric cGlbN used above to identify His46 and His70 as the heme ligands also helped delineate the secondary structure of one of the heme isomers (Fig. 8 *B*). As opposed to the apoprotein, signals from the F helix were detectable. Chemical shifts were consistent with all predicted TrHb1 helical elements. Partial heme assignments were obtained with homonuclear data, and NOEs between heme and protein agreed with the original hypothesis of cofactor adopting two orientations (Fig. S13). Notwithstanding the doubling due to heme isomerism, as well as some low-intensity resonances likely arising from the low-population HS species (Figs. 6 *D* and S6), the linewidths and dispersion of the spectra indicated a well-defined tertiary structure.

Holo-cGlbN in the cyanomet state

Cyanomet holo-cGlbN gave rise to the highest quality cGlbN NMR spectra, with additional signals assigned in the EF loop. The chemical shift information for one of the two holoprotein isomers was analyzed with TALOS+ (Fig. 8 *C*) and traced the same helical structure as the *bis*-His protein. Of interest for structural purposes is the appearance of two OH protons at 26 and 24 ppm (Fig. 6 *G*). These are manifestations of a hydrogen bond between the bound cyanide and TyrB10 (84). A network of NOEs establishes the following contacts: heme ↔ Val83 ↔Phe21 ↔ Tyr22(B10) (Fig. S14), in agreement with the geometry of other cyanide-bound TrHb1s. Other NOEs confirmed that the tertiary structure matched that of a TrHb1.

Thermodynamic stability of cGlbN

Because the structural features of cGlbN resembled those of existing TrHb1s, a thermodynamic analysis was expected to shed further light on stability and cofactor binding in

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

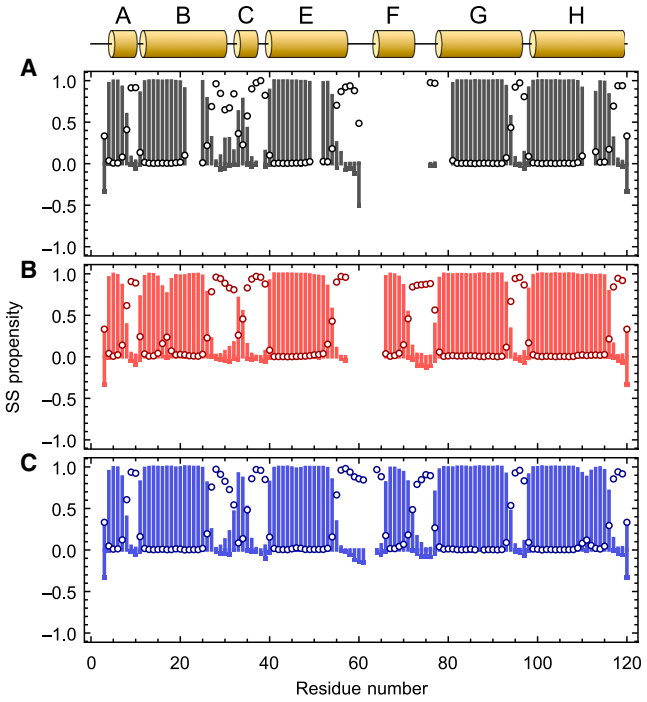

FIGURE 8 Secondary structure (SS) propensity plots from TALOS+ analysis of (*A*) apo-cGlbN, (*B*) ferric holo-cGlbN, and (*C*) cyanomet holo-cGlbN backbone NMR assignments. Positive bars indicate α-helices, negative bars indicate β-strands, and positive circles indicate coils. Cylinders above plots show homology-predicted helices (Fig. 3 *A*). To see this figure in color, go online.

comparison with other studied natural globins. We first attempted thermal unfolding, with changes in secondary structure and the heme environment monitored by CD and UV-vis spectroscopies, respectively (Fig. S15). Apo- and holoprotein showed no defined transition up to 95°C. Instead, the ellipticity at 222 nm increased linearly, reflecting a partial, noncooperative loss of helicity. At the highest temperature, the signal had undergone a decrease of only 30% in the apo form and 19% in the holo form, relative to values at 25°C. The initial CD spectra were recovered after cooling back to 25°C. This behavior is comparable with that of highly stable consensus proteins devoid of cofactor (85) and prompted the use chemical denaturants.

Incubation of holo-cGlbN in 8 M urea resulted in a small decrease in the HS spectral features (e.g., ∼630 nm) but was otherwise inconsistent with heme dissociation even after prolonged equilibration at room temperature (Fig. S16). The stronger ionic denaturant guanidine hydrochloride (GdnHCl) was required to disrupt the structure. In a titration monitored by UV-vis spectroscopy (Fig. 9 *A*), holo-cGlbN displayed a transition with a midpoint of ∼3.7 M GdnHCl giving way to the spectrum of free heme solubilized by GdnHCl. Interestingly, in the 0–2 M concentration range, a decrease in HS spectral features occurred as observed upon lowering solution pH, increasing urea concentration, or increasing $MgSO_4$ concentration (Fig. S17).

Complementary GdnHCl titrations of apo- and holoprotein samples were followed by CD spectroscopy. Apo-cGlbN unfolds in an apparent two-step process (Fig. 9 *B*), with an initial steep decrease in ellipticity (midpoint at ∼2 M GdnHCl) and a second, shallower transition (midpoint at ∼6.3 M GdnHCl). Thermodynamic stability parameters were obtained by fitting a three-state (N, I, and U) unfolding model to these data (Table S3). The Gibbs free energy changes extrapolated to 0 M GdnHCl describing the U → I and I → N processes were approximately −41 and −23 kJ mol^{-1}, respectively, with associated *m* values of 6.5 and 11.5 kJ mol^{-1} M^{-1}. The signal of I was allowed to depend linearly on denaturant concentration, which greatly improved the fit and allowed extrapolation of I state helical content of ∼58% at 0 M GdnHCl (40% at 4 M GdnHCl) relative to the N-state. The sum of the *m* values is consistent with that of a 120-residue folded protein (86) but higher than that of the reference *Synechococcus* GlbN apoprotein (6.7 kJ mol^{-1} M^{-1} (75)).

Denaturation of holo-cGlbN also produced a two-step curve, with an additional and reproducible feature in the 0 to 2 M GdnHCl concentration range. In this regime, the helicity of holo-cGlbN increased linearly to that of the apoprotein. The gradual change coincided with the decrease of the HS spectral contribution (Fig. S17 *B*) and conveniently resulted in a homogeneous holoprotein form (i.e., fully *bis*-His). For lack of a suitable analytical representation, the process was modeled as a baseline drift.

The midpoint of the first transition of holo-cGlbN denaturation was shifted to a higher GdnHCl concentration compared with apoprotein (Fig. 9 *B*), in concert with heme loss as per the UV-vis absorption data. The second step was common to both apo- and holo-cGlbN. Thus, a parsimonious model in which heme binding only occurs to the N-state was considered sufficient to analyze the ensemble of data. Using the independently obtained apoprotein parameters, the four-state model was fit to the holoprotein data to extract a heme dissociation constant, K_{NH}, of ∼5 pM (ΔG°_{NH} = ∼65 kJ/mol) and an associated *m* value of 3.9 kJ mol^{-1} M^{-1} (Table S3). The dissociation constant is only 50 times higher than that of apoMb under similar conditions (48).

Near-UV CD spectra are sensitive to the packing of aromatic residues (87). In cGlbN, five out of five Phe and two out of three Tyr were expected to reside close to the heme group (Fig. S18). Near-UV spectra of apo-cGlbN collected at 0, 4, and 8 M GdnHCl (Fig. 9 *C*) therefore provide markers of tertiary structure for N, I, and U. At 0 M GdnHCl, distinct negative peaks at 262 and 268 nm are consistent with Phe ^1L_b electronic transitions, while the overlapped 276–288 nm region can be attributed to Tyr transitions (88). At 4 M GdnHCl all signals were much diminished; at 8 M, there were no signal changes except some increase in negative Tyr intensity. The data indicated that most tertiary packing inhabited by Phe side chains was lost in the I state and to nearly the same extent as the fully unfolded protein at 8 M GdnHCl.

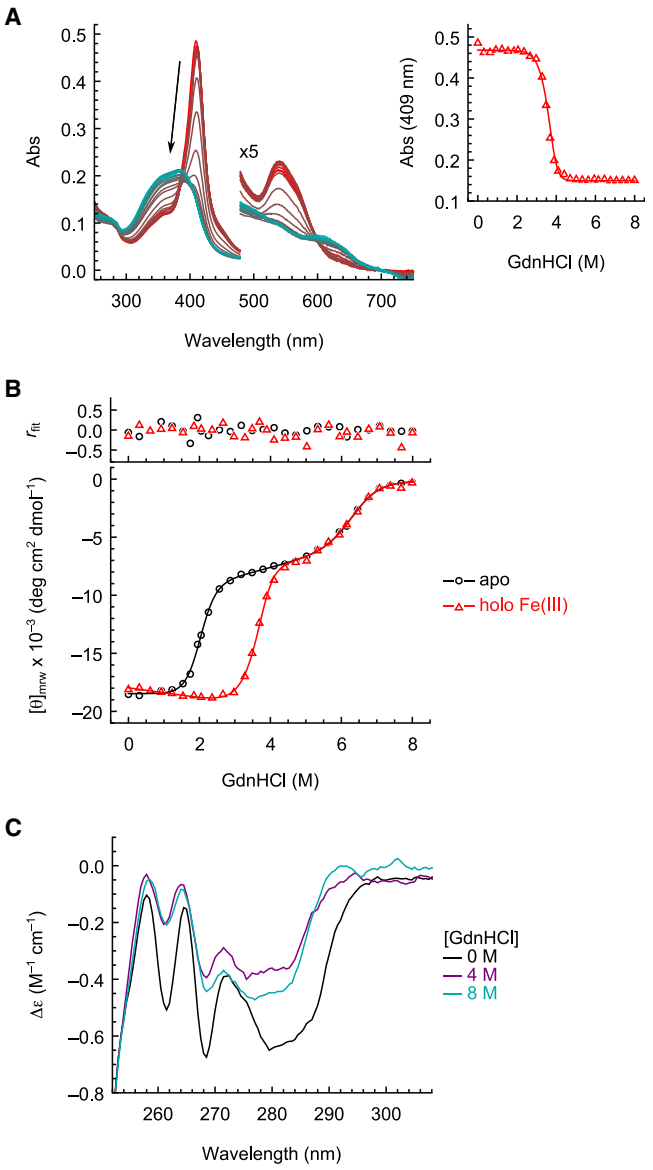

FIGURE 9 Thermodynamic stability and tertiary structure properties of cGlbN probed by GdnHCl-induced denaturation. (*A*) UV-vis absorbance spectra of holo-cGlbN. Traces are colored from red to teal with increasing denaturant concentration (*black arrow*). Inset: Soret absorbance versus GdnHCl concentration. The solid line was generated from the fitted stability parameters using the CD data in (*B*). The midpoint of the heme loss transition is ~3.7 M. (*B*) CD signal monitored at 222 nm of the apo-cGlbN (*black circles*) and holo-cGlbN (*red triangles*) titrations. Solid lines are the result of the fits, with residuals shown above. (*C*) Near-UV CD spectra of apo-cGlbN samples at the indicated denaturant concentrations. All samples in (*A–C*) were prepared in 25 mM potassium phosphate, pH 7.0. CD spectra are available at PCDDC: CD0006430000 (native), CD0006430001 (intermediate), and CD0006430002 (unfolded). To see this figure in color, go online.

DISCUSSION

Structural features of a consensus hemoglobin

As a consensus protein, cGlbN presented both expected and original features. Not surprisingly, the sequence folded into a recognizable TrHb1 structure. Also anticipated was its ability to bind the heme group in the correct location despite a low extent of residue conservation in the heme pocket (Fig. S19) and the potential of alternative binding sites (43). Given the largely unknown biological functions of TrHb1s, one interpretation of the low heme pocket conservation is that members of the MSA have a variety of roles, each determined by distinct sets of residues. Another possibility is that a common function is maintained by sequence correlations that are lost in the consensus selection. A simpler view holds that the variable residues serve to build a sufficiently hydrophobic pocket for tight binding within the TrHb1 architecture and play only a minor role in tuning the reactivity of the heme group. Finally, among TrHb1s, the heme orientation does not appear to be conserved or strongly selected (8); cGlbN illustrates this lack of preference for one face of the heme over the other.

It is noteworthy that the consensus sequence imparted sufficient main-chain flexibility in interhelical loops to allow for the two conformations accessible to the holo-TrHb1 fold, one in which exogenous ligands are stabilized in the distal heme pocket and another in which the distal E10 residue (e.g., His46) is able to coordinate the iron. However, the pH titration data revealed that *bis*-His cGlbN was only ~90% populated at neutral pH, in contrast to examples of natural *bis*-His GlbNs. At pH 7, the polypeptide is predicted to carry an approximate charge of −11. Unfavorable electrostatic repulsions imposed by the artificial primary structure could compromise distal histidine coordination leading to its lability. These repulsions would be screened by acidic pH, presence of multivalent salts, and low concentrations of GdnHCl, allowing for compaction, refolding (89), and full *bis*-His coordination. If needed, these properties could be fine-tuned in cGlbN with a limited number of back-substitutions as was done in the context of a consensus luciferase enzyme (90).

Detailed structural information on holo-cGlbN with cyanide bound was readily obtained because of high-quality spectra and similarity to previously studied TrHb1 proteins. In contrast, apo-cGlbN remains incompletely characterized. The high helicity detected by CD along with the nonuniform presentation of NMR linewidths and the missing resonances suggests that apo-cGlbN has holoprotein-like secondary structure but tertiary structure fluctuating on multiple chemical shift timescales. These features would qualify the apoprotein as a "bad folder" in a so-called gemisch state (91). Upon binding the heme group, the same level of secondary structure is maintained but a more rigid fold is adopted, characterized by a well-defined F-helix and stable tertiary interactions detected by NMR spectroscopy.

Modeling the different conformational states of cGlbN puts existing sequence-based prediction algorithms to a test. Using SWISS-MODEL (58), coordinates without an explicit heme molecule were templated with the azide-bound state of *Synechocystis* GlbN (PDB: 1S6A) or its

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

bis-His state (PDB: 1RTX). As such, these models serve as plausible representations of the conformational states observed experimentally in holo-cGlbN but not a priori for apo-cGlbN. The ligand-agnostic algorithm AlphaFold2 (59,92) returned a folded apo structure with high conformational similarity to holo-TrHb1 structures with exogenous ligand bound (e.g., root mean-square deviation of ∼1.2 Å against PDB: 1S6A; Fig. S20). Remarkably, when applied to the *Synechocystis* GlbN sequence, the resulting AlphaFold2 model also mimicked the exogenous ligand-bound state. However, according to CD and NMR data, *Synechocystis* apoGlbN is 30% less helical than the holoprotein (74) and not well folded. Prediction of cofactor-free structures for proteins containing a large, tightly bound moiety is one of the known challenges for AlphaFold (59). The bias toward holoprotein conformation, however, provides an opportunity to model ligand transplantation, as implemented in AlphaFill (93). For our purposes, it is unclear whether the AlphaFold2 model can be trusted for the interpretation of apo-cGlbN data even though the program appears to be successful with some heme proteins (94).

As an experimentally driven alternative to sequence-based predictions, the apoprotein backbone chemical shifts were submitted to CS-Rosetta (60). Convergence was reached, and a set of low-energy structures was obtained. The 10 lowest-energy structures have a root mean-square deviation of ∼1.3 Å from the lowest energy structure, and each member is consistent with a TrHb fold (Fig. S21). On the distal side there is obvious resemblance to the holoprotein with exogenous ligand bound but, on the proximal side, the F helix extends over two full helical turns, crowds the heme pocket, and interferes with the C-terminal end of the H helix, which CS-Rosetta labels as a flexible tail (Fig. S22). This best CS-Rosetta set has elements in common with holo and apo TrHb1s, perhaps forecasting the behavior of existing apoproteins. It is interesting that it also is reminiscent of the nonheme globin RsbR (95) (PDB: 2BNL), a protein in which the proximity of the E and F helices occludes the heme binding site. Further structural determination of apo-cGlbN will be necessary to validate these tentative models.

Thermodynamic stability

Consensus sequences tend to produce "hyperstable" proteins (96,97). For cGlbN, adoption of consensus residues across the entire TrHb1 domain resulted in stabilized apoprotein secondary structure. When probed by GdnHCl-induced denaturation, however, a segregation of two thermodynamic subdomains was revealed and required a three-state representation (Fig. 10 *A*). Extrapolation of the apoprotein data to 0 M GdnHCl suggested that ∼42% of native secondary structure was contained in the less stable of the two subdomains or "SD1." In the I state, which is fully populated at 4 M GdnHCl, the apoprotein contains the remaining helical content in the more stable subdomain, "SD2." Denaturation of the holoprotein monitored at the Soret band (Fig. 9 *A*) showed the unfolding of SD1 to be associated with heme loss. We interpret this observation, as well as the near-UV CD spectra (Fig. 9 *C*), to indicate that SD1 encompasses the heme cavity structure, which unfolds during the first transition, while SD2 constitutes a hydrophobic core remote from the heme binding site and corresponds to the second unfolding transition (Fig. 10 *B*).

The thermodynamic behavior of cGlbN contrasts with that of most consensus globular proteins, as they show high stability manifested in a single cooperative transition (2), when generated from MSAs of single- and multidomain protein families. The unusual behavior also contrasts with that of TrHb1s for which thermodynamic information is available. *Synechocystis* GlbN (98), which is 59% identical to cGlbN, has relatively low stability. Denaturation of the apoprotein is complete in 2 M urea with an apparent midpoint of ∼1 M, whereas denaturation of the holoprotein has a midpoint of ∼5 M (*m* value ∼6 kJ mol^{-1} M^{-1}) (98). Holoprotein data from CD and UV-vis are not perfectly coincident, hinting at three-state behavior. The apoprotein of *Synechococcus* GlbN (55% identical to cGlbN) has more secondary structure than the apoprotein of *Synechocystis* GlbN and exhibits an apparent two-state urea denaturation, with midpoint of ∼3.2 M (*m* value ∼7 kJ mol^{-1} M^{-1}) (75). Thus, although apoprotein stability spans a wide range of values, the N, NH, and I states of cGlbN are measurably more stable than the N and NH states of its known parents. We have reported that the apo form of *Synechococcus* GlbN has a cluster of slowly exchanging amide protons located at the interface of the B, G, and H helices, away from the heme cavity (62). The I state of cGlbN appears to contain the same structural core (SD2), one that may be inherent to the TrHb topology but thermodynamically unresolved in natural proteins owing to their more typical stabilities. In NMR samples of cGlbN solvated in >90% D_2O, we observed that the slowest amide protons to be replaced with deuterons are found in the presumed SD2 and SD1 (*bis*-His protein, Fig. S23) and SD2 (apoprotein, Fig. S24).

As recapitulated recently (99), structure prediction tools are able to produce models of the fully folded state, but are ineffectual at elucidating a folding path from amino acid sequence alone. Natural globin proteins illustrate the distinction; when expressed in their native organisms as apoproteins, they populate physiologically relevant but often arrested folding states governed solely by primary structure, until eventually bound to heme and fully folded. Structure prediction tools appear to skip these partially folded species in favor of the holoprotein conformation biased by the available holoprotein structures in databases despite the absence of a large number of protein-heme interactions.

cGlbN seems to exemplify a protein that undergoes less extensive secondary structure rearrangement upon cofactor

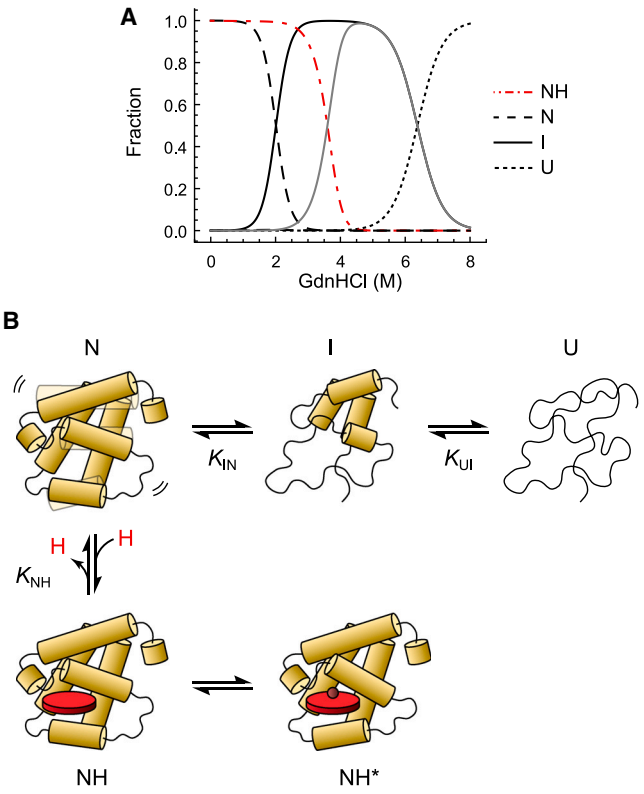

FIGURE 10 Species plot and thermodynamic (un)folding model of cGlbN based on the fitted values from GdnHCl denaturation CD data (see Fig. 9 *B*). (*A*) The apoprotein curves of the N state (*dashed*), I state (*solid*), and U state (*dotted*) are shown in black, while the holoprotein red curves involve a single heme-bound NH state (*dotted-dashed*). The apo I state curve resulting from NH heme release is shown in gray. (*B*) Structural diagrams of cGlbN (un)folding, with the associated stability and heme binding constants discussed in the text. The NH state represents the holoprotein in the dominant *bis*-His form, whereas NH* represents the minor high-spin (likely aquomet) form, which is present in low-salt samples at neutral pH and disappears at low GdnHCl concentrations. For lack of quantitative information, NH* is not included in the species plot. To see this figure in color, go online.

binding than less-stable congeners and for which modeling may be more accurate. However, aspects of the artificial protein (conformational fluctuations and heterogeneity) are not captured by prediction algorithms. It is also notable that the preferred models generated by ligand-agnostic tools for existing or artificial proteins are not those of the *bis*-His state. Endogenous coordination therefore appears to result from a subtle balance of energetic contributions that are entirely foreign to the programs.

Consolidating the heme release and denaturation data allows for an estimate of heme binding kinetics. cGlbN displays high heme affinity (K_{NH} = ~5 pM), which is consistent with the appearance of the titration in Fig. 4 *B*, and slow heme dissociation (k_{off} = ~5 × 10^{-7} s^{-1} or ~1.8 × 10^{-3} h^{-1}). The association rate constant, k_{on}, is therefore ~1 × 10^{5} M^{-1} s^{-1}, slower than the diffusion-limited rate constant of 10^{8} M^{-1} s^{-1} proposed for Mb relatives (47) but not out of the ordinary for a heme protein (39). The observation of slow NH ⇆ I + H equilibration in preliminary holoprotein denaturation experiments, but fast N ⇆ I equilibration in apoprotein experiments, supports slow heme association. The large *m* value (11.5 kJ mol^{-1} M^{-1}) of the N-to-I transition suggests substantial burial of surface area in the N-state, which we attribute to the heme binding region. This, as well as the heterogeneous and broad features of the apo-cGlbN NMR spectra, paint the apo structure as imperfectly poised for association with the cofactor. Whereas natural proteins fold locally on binding, cGlbN appears to require rearrangement of pre-organized structure, which could explain the relatively slow heme binding on-rate.

The behavior of cGlbN adds to available information on natural and artificial globins. For natural 3-on-3 proteins, sperm whale Mb has long served as a reference protein. In an extensive study, Culbertson and Olson have inspected the equilibrium unfolding of several Mb variants (48). They detected an I state in the 2–3 M GdnHCl range. The fractional population of this species remained low because of conversion to a heme-bound state, IH, attributed to a hexacoordinate complex recruiting the distal histidine or some other residue as sixth ligand to the iron. In contrast, the I state of cGlbN does not bind heme and is energetically well separated from the N and NH states.

Further comparison can be made to the artificial 3-on-3 globins prepared by Isogai and co-workers. The "redesign" approach used an iterative optimization of the sequence to fit the 3D structure of Mb (21,100). The designed globins (DGs) have less than 30% identity with the wild-type sequence and display high compositional redundancy. Systematic heme binding information is not available for the various DG proteins, except for DG1, which forms an endogenous hexacoordinate species. The DGs share some features with apo-cGlbN: increased stability, a gemisch N-state, and population of an I state. The consensus approach parallels the redesign approach in that the selected residues are expected to stabilize the fold. The consensus sequence, however, limits the palette at each position according to natural proteins. The population of I in cGlbN reinforces the proposal that natural sequences tend to disfavor intermediate states (101), but also suggests how a two-subdomain architecture can be exploited for efficient heme protein design.

Finally, the relationship between cGlbN and ancestral TrHb1s has yet to be explored, with the caveat that the sequence used here, which was constrained by the makeup of the MSA, the imposition of specific residues at certain positions, and various thresholds for selection, is only one of a multitude of consensus possibilities. To our knowledge, ancestral sequence reconstruction has been applied to study the evolution of animal hemoglobins (see, e.g., (102,103)) but not TrHbs. Design challenges such as the curation of a tree-aware MSA and the treatment of pervasive horizontal

gene transfer (104) will have to be overcome to generate a group of plausible ancestral proteins. Although many sequences obtained by ancestral sequence reconstruction show enhanced thermodynamic stability, such is not always the case (105). It will be interesting to discover if highly stable reconstructed ancestral TrHb1s resemble the consensus derived from the same MSA and if, like cGlbN, they resolve two independent thermodynamic domains. It does not escape us that the cofactor may not be a *b*-type heme in some of the early proteins (106). This additional uncertainty is expected to complicate future comparisons and functional conclusions.

CONCLUSION

The consensus TrHb1 characterized in this work resembled the proteins from which it was derived. cGlbN bound heme with high affinity, adopted the typical 2-on-2 α-helical fold, and bound small ligands like oxygen. However, it also presented distinctive properties such as a "conformationally challenged" apoprotein state, a slow heme on-rate, and heterogeneity in heme binding and iron coordination. Most strikingly, the consensus sequence revealed that the TrHb1 architecture is composed of two independent subdomains, one of which encompasses the heme pocket. In addition, cGlbN can be viewed as a stable and naive background protein for convenient manipulation of heme chemistry.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2023.06.016.

AUTHOR CONTRIBUTIONS

J.E.M.G. designed and performed research, analyzed data, and wrote the manuscript. E.A.J. designed and performed research, and wrote the manuscript. J.T.J.L. designed research, analyzed data, and wrote the manuscript.

ACKNOWLEDGMENTS

The authors acknowledge Dr. Katherine Tripp (Center for Molecular Biophysics, Johns Hopkins University), Dr. Phil Mortimer (Mass Spectrometry Facility, Johns Hopkins University), and Dr. Ananya Majumdar (Biomolecular NMR Center, Johns Hopkins University) for assistance with data collection, Dr. Aaron Robinson for experimental assistance, Drs. Doug Barrick and Christopher Falzone for helpful discussions, and Dr. Dillon Nye, Alana Sherer, Soumya Behera, Kevin Liu, and Thomas Schultz for help at various stages of the study. This work was supported by NSF grants CHE-2003950 to JTJL and GRFP-1746891 to JEMG.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPORTING CITATIONS

References (107–118) appear in the supporting material.

REFERENCES

1. Porebski, B. T., and A. M. Buckle. 2016. Consensus protein design. *Protein Eng. Des. Sel.* 29:245–251.
2. Sternke, M., K. W. Tripp, and D. Barrick. 2019. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. USA.* 116:11275–11284.
3. Goyal, V. D., B. J. Sullivan, and T. J. Magliery. 2020. Phylogenetic spread of sequence data affects fitness of consensus enzymes: Insights from triosephosphate isomerase. *Proteins.* 88:274–283.
4. Motoyama, T., N. Hiramatsu, …, S. Ito. 2020. Protein sequence selection method that enables full consensus design of artificial L-threonine 3-dehydrogenases with unique enzymatic properties. *Biochemistry.* 59:3823–3833.
5. Wenke, B. B., J. T. J. Lecomte, …, J. L. Schlessman. 2014. The 2/2 hemoglobin from the cyanobacterium *Synechococcus* sp. PCC 7002 with covalently attached heme: comparison of X-ray and NMR structures. *Proteins.* 82:528–534.
6. Wittenberg, J. B., M. Bolognesi, …, M. Guertin. 2002. Truncated hemoglobins: A new family of hemoglobins widely distributed in bacteria, unicellular eukaryotes and plants. *J. Biol. Chem.* 277:871–874.
7. Vinogradov, S. N., M. Tinajero-Trejo, …, D. Hoogewijs. 2013. Bacterial and archaeal globins — A revised perspective. *Biochim. Biophys. Acta.* 1834:1789–1800.
8. Pesce, A., M. Couture, …, M. Bolognesi. 2000. A novel two-over-two a-helical sandwich fold is characteristic of the truncated hemoglobin family. *EMBO J.* 19:2424–2434.
9. Bustamante, J. P., L. Radusky, …, M. A. Martí. 2016. Evolutionary and functional relationships in the truncated hemoglobin family. *PLoS Comput. Biol.* 12:e1004701.
10. Nardini, M., A. Pesce, and M. Bolognesi. 2022. Truncated (2/2) hemoglobin: Unconventional structures and functional roles in vivo and in human pathogenesis. *Mol. Aspects Med.* 84, 101049.
11. Vuletich, D. A., and J. T. J. Lecomte. 2006. A phylogenetic and structural analysis of truncated hemoglobins. *J. Mol. Evol.* 62:196–210.
12. Milani, M., A. Pesce, …, M. Bolognesi. 2004. Heme-ligand tunneling in group I truncated hemoglobins. *J. Biol. Chem.* 279:21520–21525.
13. Ouellet, Y., M. Milani, …, M. Guertin. 2006. Ligand interactions in the distal heme pocket of *Mycobacterium tuberculosis* truncated hemoglobin N: Roles of TyrB10 and GlnE11 residues. *Biochemistry.* 45:8770–8781.
14. Giordano, D., F. M. Boubeta, …, C. Verde. 2020. Conformational flexibility drives cold adaptation in *Pseudoalteromonas haloplanktis* TAC125 globins. *Antioxid. Redox Sign.* 32:396–411.
15. Julió Plana, L., A. D. Nadra, …, L. Capece. 2019. Thermal stability of globins: Implications of flexibility and heme coordination studied by molecular dynamics simulations. *J. Chem. Inf. Model.* 59:441–452.
16. Ando, N., B. Barquera, …, M. B. Watkins. 2021. The molecular basis for life in extreme environments. *Annu. Rev. Biophys.* 50:343–372.
17. Lin, Y.-W., E. B. Sawyer, and J. Wang. 2013. Rational heme protein design: All roads lead to Rome. *Chem. Asian J.* 8:2534–2544.
18. Koder, R. L., J. L. R. Anderson, …, P. L. Dutton. 2009. Design and engineering of an O_2 transport protein. *Nature.* 458:305–309.
19. Polizzi, N. F., Y. Wu, …, W. F. DeGrado. 2017. De novo design of a hyperstable non-natural protein–ligand complex with sub-Å accuracy. *Nat. Chem.* 9:1157–1164.
20. Schnatz, P. J., J. M. Brisendine, …, R. L. Koder. 2020. Designing heterotropically activated allosteric conformational switches using supercharging. *Proc. Natl. Acad. Sci. USA.* 117:5291–5297.

21. Isogai, Y., M. Ota, …, K. Nishikawa. 1999. Design and synthesis of a globin fold. *Biochemistry.* 38:7431–7443.

22. Altschul, S. F., W. Gish, …, D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

23. Katoh, K., and D. M. Standley. 2013. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

24. Sievers, F., A. Wilm, …, D. G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539.

25. Sievers, F., and D. G. Higgins. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 27:135–145.

26. Waterhouse, A. M., J. B. Procter, …, G. J. Barton. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics.* 25:1189–1191.

27. Scott, N. L., and J. T. Lecomte. 2000. Cloning, expression, purification, and preliminary characterization of a putative hemoglobin from the cyanobacterium *Synechocystis* sp. PCC 6803. *Protein Sci.* 9:587–597.

28. Scott, N. L., C. J. Falzone, …, J. T. J. Lecomte. 2002. Truncated hemoglobin from the cyanobacterium *Synechococcus* sp. PCC 7002: Evidence for hexacoordination and covalent adduct formation in the ferric recombinant protein. *Biochemistry.* 41:6902–6910.

29. Johnson, E. A., S. L. Rice, …, J. T. J. Lecomte. 2014. Characterization of THB1, a *Chlamydomonas reinhardtii* truncated hemoglobin: linkage to nitrogen metabolism and identification of lysine as the distal heme ligand. *Biochemistry.* 53:4573–4589.

30. Savitzky, A., and M. J. E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36:1627–1639.

31. de Duve, C. 1948. A spectrophotometric method for the simultaneous determination of myoglobin and hemoglobin in extracts of human muscle. *Acta Chem. Scand.* 2:264–289.

32. Berry, E. A., and B. L. Trumpower. 1987. Simultaneous determination of hemes *a, b*, and *c* from pyridine hemochrome spectra. *Anal. Biochem.* 161:1–15.

33. Barr, I., and F. Guo. 2015. Pyridine hemochromagen assay for determining the concentration of heme in purified protein solutions. *Bio. Protoc.* 5, e1594.

34. Edelhoch, H. 1967. Spectroscopic determination of tryptophan and tyrosine in proteins. *Biochemistry.* 6:1948–1954.

35. Gill, S. C., and P. H. von Hippel. 1989. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* 182:319–326.

36. Grimsley, G. R., and C. N. Pace. 2003. Spectrophotometric determination of protein concentration. *Curr. Protoc. Protein Sci.* 33:311–319.

37. Gallagher, W. A., and W. B. Elliott. 1968. Alkaline haematin and nitrogenous ligands. *Biochem. J.* 108:131–136.

38. Kawamura-Konishi, Y., and H. Suzuki. 1985. Binding reaction of hemin to globin. *J. Biochem.* 98:1181–1190.

39. Carter, E. L., Y. Ramirez, and S. W. Ragsdale. 2017. The heme-regulatory motif of nuclear receptor Rev-erbβ is a key mediator of heme and redox signaling in circadian rhythm maintenance and metabolism. *J. Biol. Chem.* 292:11280–11299.

40. Leung, G. C.-H., S. S.-P. Fung, …, A. J. Hudson. 2019. Precise determination of heme binding affinity in proteins. *Anal. Biochem.* 572:45–51.

41. Hughson, F. M., D. Barrick, and R. L. Baldwin. 1991. Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis. *Biochemistry.* 30:4113–4118.

42. Teale, F. W. J. 1959. Cleavage of heme-protein link by acid methylethylketone. *Biochim. Biophys. Acta.* 35:543.

43. Nye, D. B., E. A. Johnson, …, J. T. J. Lecomte. 2019. Replacement of the heme axial lysine as a test of conformational adaptability in the truncated hemoglobin THB1. *J. Inorg. Biochem.* 201, 110824.

44. Martinez Grundman, J. E., L. Julió Plana, …, J. T. J. Lecomte. 2021. Control of distal lysine coordination in a monomeric hemoglobin: A role for heme peripheral interactions. *J. Inorg. Biochem.* 111437

45. Nozaki, Y. 1972. [3] The preparation of guanidine hydrochloride. *In* Methods in Enzymology Elsevier, pp. 43–50.

46. Barrick, D., and R. L. Baldwin. 1993. Three-state analysis of sperm whale apomyoglobin folding. *Biochemistry.* 32:3790–3796.

47. Hargrove, M. S., S. Krzywda, …, J. S. Olson. 1994. Stability of myoglobin: a model for the folding of heme proteins. *Biochemistry.* 33:11767–11775.

48. Culbertson, D. S., and J. S. Olson. 2010. Role of heme in the unfolding and assembly of myoglobin. *Biochemistry.* 49:6052–6063.

49. Lee, D., C. Hilty, …, K. Wüthrich. 2006. Effective rotational correlation times of proteins from NMR relaxation interference. *J. Magn. Reson.* 178:72–76.

50. Pond, M. P., A. Majumdar, and J. T. J. Lecomte. 2012. Influence of heme post-translational modification and distal ligation on the backbone dynamics of a monomeric hemoglobin. *Biochemistry.* 51:5733–5747.

51. Hyberts, S. G., K. Takeuchi, and G. Wagner. 2010. Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. *J. Am. Chem. Soc.* 132:2145–2147.

52. Delaglio, F., S. Grzesiek, …, A. Bax. 1995. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR.* 6:277–293.

53. Ying, J., F. Delaglio, …, A. Bax. 2017. Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *J. Biomol. NMR.* 68:101–118.

54. Keller, R. L. J. 2004. Computer Aided Resonance Assignment Tutorial. Cantina.

55. Goddard, T. D., and D. G. Kneller. 2006. SPARKY 3. University of California.

56. Skinner, S. P., R. H. Fogh, …, G. W. Vuister. 2016. CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *J. Biomol. NMR.* 66:111–124.

57. Maciejewski, M. W., A. D. Schuyler, …, J. C. Hoch. 2017. NMRbox: A resource for biomolecular NMR computation. *Biophys. J.* 112:1529–1534.

58. Waterhouse, A., M. Bertoni, …, T. Schwede. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46:W296–W303.

59. Jumper, J., R. Evans, …, D. Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature.* 596:583–589.

60. Shen, Y., R. Vernon, …, A. Bax. 2009. De novo protein structure generation from incomplete chemical shift assignments. *J. Biomol. NMR.* 43:63–78.

61. Scott, N. L., Y. Xu, …, J. T. J. Lecomte. 2010. Functional and structural characterization of the 2/2 hemoglobin from *Synechococcus* sp. *Prev. Controle Cancerol.* 49:7000–7011.

62. Vuletich, D. A., C. J. Falzone, and J. T. J. Lecomte. 2006. Structural and dynamic repercussions of heme binding and heme−protein crosslinking in *Synechococcus* sp. PCC 7002 hemoglobin. *Biochemistry.* 45:14075–14084.

63. Falzone, C. J., B. Christie Vu, …, J. T. J. Lecomte. 2002. The solution structure of the recombinant hemoglobin from the cyanobacterium *Synechocystis* sp. PCC 6803 in its hemichrome state. *J. Mol. Biol.* 324:1015–1029.

64. Johnson, E. A., and J. T. J. Lecomte. 2013. The globins of cyanobacteria and algae. *Adv. Microb. Physiol.* 63:195–272.

65. Johnson, E. A., and J. T. J. Lecomte. 2015. The haemoglobins of algae. *Adv. Microb. Physiol.* 67:177–234.
66. Crooks, G. E., G. Hon, …, S. E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.
67. Hormoz, S. 2013. Amino acid composition of proteins reduces deleterious impact of mutations. *Sci. Rep.* 3:2919.
68. Kozlowski, L. P. 2022. Proteome-pI 2.0: proteome isoelectric point database update. *Nucleic Acids Res.* 50:D1535–D1540.
69. Schwartz, R., C. S. Ting, and J. King. 2001. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of Life. *Genome Res.* 11:703–709.
70. Piszkiewicz, D., M. Landon, and E. L. Smith. 1970. Anomalous cleavage of aspartyl-proline peptide bonds during amino acid sequence determinations. *Biochem. Biophys. Res. Commun.* 40:1173–1178.
71. Vlasak, J., and R. Ionescu. 2011. Fragmentation of monoclonal antibodies. *mAbs.* 3:253–263.
72. Young, K. Z., S. J. Lee, …, M. M. Wang. 2020. NOTCH3 is non-enzymatically fragmented in inherited cerebral small-vessel disease. *J. Biol. Chem.* 295:1960–1972.
73. Smith, M. L., J. Paul, …, K. G. Paul. 1991. Heme-protein fission under nondenaturing conditions. *Proc. Natl. Acad. Sci. USA.* 88:882–886.
74. Lecomte, J. T., N. L. Scott, …, C. J. Falzone. 2001. Binding of ferric heme by the recombinant globin from the cyanobacterium *Synechocystis* sp. PCC 6803. *Biochemistry.* 40:6541–6552.
75. Landfried, D. A., D. A. Vuletich, …, J. T. J. Lecomte. 2007. Structural and thermodynamic consequences of b heme binding for monomeric apoglobins and other apoproteins. *Gene.* 398:12–28.
76. Couture, M., T. K. Das, …, M. Guertin. 2000. Structural investigations of the hemoglobin of the cyanobacterium Synechocystis PCC 6803 reveal a unique distal heme pocket. *Eur. J. Biochem.* 267:4770–4780.
77. Pond, M. P., D. A. Vuletich, …, J. T. J. Lecomte. 2009. ^{1}H, ^{15}N, and ^{13}C resonance assignments of the 2/2 hemoglobin from the cyanobacterium *Synechococcus* sp. PCC 7002 in the ferric bis-histidine state. *Biomol. NMR Assign.* 3:211–214.
78. Falzone, C. J., and J. T. J. Lecomte. 2002. Assignment of the 1H, 13C, and 15N signals of Synechocystis sp. PCC 6803 methemoglobin. *J. Biomol. NMR.* 23:71–72.
79. La Mar, G. N., N. L. Davis, …, K. M. Smith. 1983. Heme orientational disorder in reconstituted and native sperm whale myoglobin. Proton nuclear magnetic resonance characterizations by heme methyl deuterium labeling in the met-cyano protein. *J. Mol. Biol.* 168:887–896.
80. Vu, B. C., D. A. Vuletich, …, J. T. J. Lecomte. 2004. Characterization of the heme-histidine cross-link in cyanobacterial hemoglobins from *Synechocystis* sp. PCC 6803 and *Synechococcus* sp. PCC 7002. *J. Biol. Inorg. Chem.* 9:183–194.
81. La Mar, G. N., J. D. Satterlee, and J. S. de Ropp. 2000. Nuclear magnetic resonance of hemoproteins. *In* The Porphyrin Handbook. K. M. Smith, K. Kadish, and R. Guilard, eds Academic Press, pp. 185–298.
82. Shen, Y., F. Delaglio, …, A. Bax. 2009. TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR.* 44:213–223.
83. Lecomte, J. T., Y.-H. Kao, and M. J. Cocco. 1996. The native state of apomyoglobin described by proton NMR spectroscopy: The A-B-G-H interface of wild-type sperm whale apomyoglobin. *Proteins.* 25:267–285.
84. Vu, B. C., H. J. Nothnagel, …, J. T. J. Lecomte. 2004. Cyanide binding to hexacoordinate cyanobacterial hemoglobins: Hydrogen bonding network and heme pocket rearrangement in ferric H117A *Synechocystis* Hb. *Biochemistry.* 43:12622–12633.
85. Wetzel, S. K., G. Settanni, …, A. Plückthun. 2008. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J. Mol. Biol.* 376:241–257.
86. Myers, J. K., C. N. Pace, and J. M. Scholtz. 1995. Denaturant *m* values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* 4:2138–2148.
87. Pain, R. 2004. Determining the CD spectrum of a protein. *Curr. Protoc. Protein Sci.* 38:7.6.1–7.6.24.
88. Li, Z., and J. D. Hirst. 2017. Quantitative first principles calculations of protein circular dichroism in the near-ultraviolet. *Chem. Sci.* 8:4318–4333.
89. Hagihara, Y., S. Aimoto, …, Y. Goto. 1993. Guanidine hydrochloride-induced folding of proteins. *J. Mol. Biol.* 231:180–184.
90. Loening, A. M., T. D. Fenn, …, S. S. Gambhir. 2006. Consensus guided mutagenesis of *Renilla* luciferase yields enhanced stability and light output. *Protein Eng. Des. Sel.* 19:391–400.
91. Dill, K. A., S. Bromberg, …, H. S. Chan. 1995. Principles of protein folding — A perspective from simple exact models. *Protein Sci.* 4:561–602.
92. Mirdita, M., K. Schütze, …, M. Steinegger. 2022. ColabFold: making protein folding accessible to all. *Nat. Methods.* 19:679–682.
93. Hekkelman, M. L., I. de Vries, …, A. Perrakis. 2023. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods.* 20:205–213.
94. Kondo, H. X., Y. Kanematsu, and Y. Takano. 2022. Structure of heme-binding pocket in heme protein is generally rigid and can be predicted by AlphaFold2. *Chem. Lett.* 51:704–708.
95. Murray, J. W., O. Delumeau, and R. J. Lewis. 2005. Structure of a nonheme globin in environmental stress signaling. *Proc. Natl. Acad. Sci. USA.* 102:17320–17325.
96. Steipe, B., B. Schiller, …, S. Steinbacher. 1994. Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* 240:188–192.
97. Lehmann, M., L. Pasamontes, …, M. Wyss. 2000. The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta.* 1543:408–415.
98. Knappenberger, J. A., S. A. Kuriakose, …, J. T. J. Lecomte. 2006. Proximal influences in two-on-two globins: Effect of the Ala69Ser replacement on *Synechocystis* sp. PCC 6803 hemoglobin. *Biochemistry.* 45:11401–11413.
99. Chen, S.-J., M. Hassan, …, G. D. Rose. 2023. Opinion: Protein folds vs. protein folding: Differing questions, different challenges. *Proc. Natl. Acad. Sci. USA.* 120, e2214423119.
100. Isogai, Y., A. Ishii, …, K. Nishikawa. 2000. Redesign of artificial globins: Effects of residue replacements at hydrophobic sites on the structural properties. *Biochemistry.* 39:5683–5690.
101. Isogai, Y. 2006. Native protein sequences are designed to destabilize folding intermediates. *Biochemistry.* 45:2488–2492.
102. Pillai, A. S., S. A. Chandler, …, J. W. Thornton. 2020. Origin of complexity in haemoglobin evolution. *Nature.* 581:480–485.
103. Natarajan, C., A. V. Signore, …, J. F. Storz. 2023. Evolution and molecular basis of a novel allosteric property of crocodilian hemoglobin. *Curr. Biol.* 33:98–108.e4.
104. Merkl, R., and R. Sterner. 2016. Ancestral protein reconstruction: techniques and applications. *Biol. Chem.* 397:1–21.
105. Trudeau, D. L., M. Kaltenbach, and D. S. Tawfik. 2016. On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol.* 33:2633–2641.
106. Kořený, L., M. Oborník, …, J. Lukeš. 2022. The convoluted history of haem biosynthesis. *Biol. Rev.* 97:141–162.
107. Lu, S., J. Wang, …, A. Marchler-Bauer. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48:D265–D268.
108. Altschul, S. F., T. L. Madden, …, D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
109. Camacho, C., G. Coulouris, …, T. L. Madden. 2009. BLAST+: architecture and applications. *BMC Bioinf.* 10:421.

Please cite this article in press as: Martinez Grundman et al., Architectural digest: Thermodynamic stability and domain structure of a consensus monomeric globin, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.06.016

110. Pei, J., B.-H. Kim, and N. V. Grishin. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* 36:2295–2300.

111. Criscuolo, A., and S. Gribaldo. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* 10:210.

112. Lemoine, F., D. Correia, …, O. Gascuel. 2019. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res.* 47:W260–W265.

113. Guindon, S., J.-F. Dufayard, …, O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307–321.

114. Letunic, I., and P. Bork. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics.* 23:127–128.

115. Vázquez-Laslop, N., H. Lee, …, A. A. Neyfakh. 2001. Molecular sieve mechanism of selective release of cytoplasmic proteins by osmotically shocked *Escherichia coli*. *J. Bacteriol.* 183:2399–2404.

116. Pei, J., and N. V. Grishin. 2001. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics.* 17:700–712.

117. Pettersen, E. F., T. D. Goddard, …, T. E. Ferrin. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* 30:70–82.

118. Salzmann, M., K. Pervushin, …, K. Wüthrich. 1998. TROSY in triple-resonance experiments: New perspectives for sequential NMR assignment of large proteins. *Proc. Natl. Acad. Sci. USA.* 95:13585–13590.