De novo design of protein-mineral complexes

Paul Kwon

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2026

Reading Committee:

David Baker, Chair
Jennifer Kong
Cole DeForest

Program Authorized to Offer Degree:
Biochemistry

©Copyright 2026
Paul Kwon

University of Washington

Abstract

De novo design of protein-mineral complexes

Paul Kwon

Chair of the Supervisory Committee:
Dr. David Baker
Department of Biochemistry

Biomineralization is the process by which living organisms create functional hybrid materials, including shells, bones, teeth, and magnetosomes. These biominerals exhibit enhanced properties in aspects such as fracture toughness, mechanical resilience, and hierarchical structural organization compared to pure mineral counterparts. Understanding the molecular mechanisms of biomineralization, particularly how proteins assemble at solid-liquid interfaces, and mimicking these processes in the laboratory could one day lead to the development of novel hybrid materials with impactful applications ranging from carbon sinks for climate change mitigation to microelectronics. A significant barrier in biomineralization research is that the natural proteins involved are frequently unstructured, heterogeneous, or insoluble, complicating their structural characterization and engineering. To overcome this, de novo protein design provides a powerful strategy for creating tailored proteins with precise structures and functions. Within the Baker lab, de novo proteins have been successfully engineered to interface with minerals. These designed proteins can guide the nucleation, growth, and polymorph of mineral crystals. We further engineer these protein-mineral interfaces to self-assemble into complex higher-order structures, including two-dimensional sheets, cages, or fibers, bringing us closer to macroscopic material composites.

Table of Contents

Chapter 1: Controlling metal-carbonate phase, form, and function through de novo protein design 5

Chapter 2: Programmed mineralization of metal-oxides by de novo designed protein templates 94

Chapter 3: Templating and confining calcium phosphate mineralization within designed protein assemblies 119

Appendix 1: De novo design of RNA and nucleoprotein complexes 148

Chapter 1: Controlling metal-carbonate phase, form, and function through de novo protein design.

This chapter contains additional background for (and limited reproduction of) content submitted as: Paul S. Kwon, Xinqi Li, Le Tracy Yu, Harley Pyles, Todd H. Lewis, Connor Weidle, Andrew J. Borst, Catherine C. Bodinger, Alex Kang, Hannah Nguyen, Johnny Mendoza, Kenneth D. Carr, Brian Coventry, Yang Hsia, Zsombor Molnár, Dongsheng Li, Bo Zhang, Brandi M. Cossairt, Shuai Zhang, Asim K. Bera, James De Yoreo, David Baker. *bioRxiv* (2026).

Abstract: Biomineralization enables living systems to construct hybrid materials by controlling the location, orientation, and polymorph of inorganic crystals with proteins and other biomolecules. Despite decades of study, the molecular principles underlying these processes remain difficult to harness in engineered materials, in part because native biomineralization proteins are often intrinsically disordered, heterogeneous, or insoluble. Here, we establish a de novo protein-design framework that recapitulates hallmark biomineralization behaviors that were previously achieved only through ad hoc strategies. By fine-tuning RFdiffusion2 to generate lattice-matched repeat protein scaffolds, we enable the selective mineralization of aragonite, a metastable calcium carbonate polymorph, under the condition that otherwise yields mixed phases. We further engineer protein-calcite interfaces into reconfigurable two-dimensional arrays that template the growth of calcite crystals in architectures reminiscent of nacre. Extending beyond biominerals found in nature, we further apply this framework to cobalt carbonate, where lattice matched protein-mineral interfaces direct both mineral phase control and nanoscale mineral confinement within a soluble D3 protein cage. The resulting protein-mineral cages exhibit sustained electrocatalytic oxygen-evolution activity under alkaline conditions. Our results demonstrate the de novo protein design can genetically encode polymorph selection, spatial organization, and emergent functionality.

Introduction

Biomineralization enables living organisms to fabricate functional materials with exquisite precision, exemplified by shells[1], bones[2], teeth[3], and magnetosomes[4]. In mollusk shells, calcium carbonate ($CaCO_3$) deposition is controlled by organic matrices of chitin and proteins such as Pif[5], yielding composites with hierarchical organization[6], and polymorph selectivity[7] that results in a fracture toughness[6] greatly exceeding the mineral alone. Understanding the design rules underlying proteins that orchestrate biomineralization could enable the development of synthetic biomimetic materials with applications in sustainability[8], biotechnology[9], medicine[10], and energy[11]. Although individual biomineralization behaviours have been achieved using engineering biomolecular systems, these capabilities generally required ad hoc design strategies tailored to individual material systems. A molecular design framework capable of prospectively encoding these

distinct mineralization outcomes has remained elusive. Establishing such a framework would enable prospective molecular encoding of mineralization behaviors across multiple structure hierarchies, further extending biomineralization beyond naturally-occurring materials towards increasingly diverse material systems with emergent functionalities.

Native biomineralization proteins are often intrinsically disordered, heterogeneous, or insoluble, posing substantial challenges for structural analysis and structure-based engineering[12]. Previously, we showed that these limitations could be overcome by creating an epitaxial match between structurally defined *de novo* proteins[13] and crystal lattices[14] to encode a programmable mineral-binding interface[15] that controls mineral nucleation, growth and facet selection. Using *de novo* designed proteins, we exploited this lattice-matching approach originally inspired by native ice-binding proteins[16] to create templates that direct nucleation of calcite[15], hydroxyapitite[17], zinc oxide[18], iron oxide[18], and ice[19]. Despite demonstrating that lattice-matched protein-mineral interfaces can direct mineralization across diverse inorganic systems, existing design strategies have largely relied on empirical, trial-and-error approaches. Here, we establish a de novo protein-design framework by integrating programmable protein-mineral interfaces and supramolecular architectures to encode distinct mineralization outcomes. We validate this framework by encoding mineralization outcomes over mineral polymorph selection, spatial organization across protein assemblies, and nanoscale confinement within closed protein architectures. We focused on metal carbonates as particularly compelling targets because they provide an ideal model system for testing this framework while also being central both to the biogeochemical carbon cycle[20] and to broader materials applications[21].

Machine-learning (ML) design methods for inorganic mineral templating

We began by developing ML methods for designing proteins for inorganic mineral templating, leveraging the recent advances in generative protein design using RFdiffusion, and in atomic motif scaffolding[22]. We started from RFdiffusion2 (RFD)[22] , which enables direct generation of protein structures which precisely scaffold pre-specified atomic motifs interacting with non-protein atoms. RFD2 was developed for enzyme design, where the non-protein atoms are the reaction substrates and transition states, and the atomic motifs are the protein catalytic machinery.

In the design of mineralization, the non-protein atoms are in a coherent crystal array which extends over a length scale longer than a protein molecule (rather than a reaction substrate which typically only interacts with a small part of the protein) and the atomic motifs are regularly arranged over an extended region of a crystal surface, rather than confined to a localized active site (**Fig. 1a**). Furthermore, for enzyme design the goal is to generate a globular protein with an active site pocket hosting non-protein atoms, whereas in design of mineralization, proteins with repeating substructures that

match the periodicity of the inorganic crystal lattice are desirable because the size of the mineralizing interface can be readily increased or decreased by adding or removing repeat units, and because an optimal solution to the crystal templating challenge need only be found in one local region and can then be readily propagated.
To generate this propagation directly within a single designed chain, we fine-tuned RFD2 to natively produce repeat proteins (RFD2-RP): backbones built from tandem internal repeat units. To optimize RFD2 for non-protein ion templating, we incorporated conditioning on secondary structure and adjacency matrices, and included a large set of repeat proteins derived from the PDB[23] and AlphaFold (AF) database[24] for training (composition of training set 25% repeat proteins, 75% the overall PDB). RFD2-RP also incorporates conditioning on an "ori" token specifying the center of mass of the target motif, which improves generation accuracy relative to the original RFD2 used to scaffold side chains into a single helix as in our previous HAp design[17]. Together, these conditioning strategies allow secondary structure (α-helix or β-strand) to be specified directly at generation time, rather than emerging incidentally from the training distribution (**Fig. S1**). To bias designs toward fully helical architectures, 70% of each sequence was specified as α helices, while the remaining positions were left masked. This strategy produced predominantly helical designs, avoiding β-sheet–rich architectures, which can be prone to aggregation (**Fig. 1b**).

Design and characterization of protein-aragonite interface

To selectively template aragonite, a metastable $CaCO_3$ polymorph, we designed proteins complementary to the {110} facet (**Fig. 2a**)[25]. {110}-terminated facets are more prevalent than are {010} facets in the pseudohexagonal aragonite habit[25], and present carbonate groups in a geometry more readily matched by acidic amino-acid side chains on a designed protein surface. Glutamate residues were positioned to mimic the missing carbonate layer that would lie above the target mineral sublattice in the bulk crystal. We subsequently used tip-atom-guided RFD2-RP to generate protein scaffolds capable of presenting these carboxylate groups at the termini of aspartate or glutamate side chains. To build larger interfaces, we selected designs capable of propagating along the mineral facet through lattice-matched translations without steric clashes, then linked the propagated copies through another diffusion step (**Fig. 1a**).

To represent the aragonite surface during sequence design, we sampled protein sequences in the presence of the Ca^{2+} sublattice on aragonite {110} using LigandMPNN[26] while keeping the previously designed carboxyl-side-chains fixed. Three designs which AF2[27] structure predictions were close to the design models were tested in E coli (**Table S1**), two of

which were solubly expressed. One design, ARAG64, behaved as a monomer in size-exclusion chromatography (SEC) (**Fig. S2**) with circular dichromism (CD) exhibiting alpha-helical folds both at 25°C and 80°C (**Fig. 2b**).

We next investigated the effect of ARAG64 on $CaCO_3$ mineralization. Following incubation of 2.16 μM protein, 10 mM $CaCl_2$, and 10 mM $NaHCO_3$ at 80°C and pH 8.2 for 2 hours, X-ray diffraction (XRD) showed that crystals formed with ARAG64 matched an aragonite reference pattern, whereas crystals formed with FD31, a protein previously shown to drive calcite nucleation at low temperature, matched a calcite reference pattern (**Fig. 2c**). Optical microscopy further supported these assignments: ARAG64 produced needle-like aragonite crystals (**Fig. 2d**), consistent with the ARAG64 design model targeting aragonite with a 9.4 Å spacing between the indicated helices (**Fig. 2e**); while FD31 produced rhombohedral calcite crystals (**Fig. 2f**), consistent with the FD31 design model targeting calcite with a corresponding 12.1 Å spacing (**Fig. 2g**). Optical micrographs and XRD patterns of the no-protein control exhibited mixed polymorph formation under these conditions with a mixture of rhombohedral calcite and needle-like aragonite (**Fig. S3**) generating both a characteristic aragonite peak near 27.2°, assigned to the (021) reflection, and a calcite peak near 29.4°, assigned to the (104) reflection (**Fig. 2c**). Together, these results demonstrate that de novo designed proteins can direct polymorph outcome in $CaCO_3$ mineralization.

Templating calcite mineralization over designed 2D arrays

Having shown that designed proteins can control $CaCO_3$ polymorph selection, we next sought to direct $CaCO_3$ organization over longer length scales by designing 2D protein arrays incorporating a $CaCO_3$ mineralizing interface. To complement the aragonite promoting design in the preceding section, we chose to focus on templating calcite. To construct a triangular scaffold, calcite-nucleating repeat protein FD31[15] motifs were pre-aligned along specified eigenvectors to direct repeat propagation, then translated and rotated under C3 symmetry to define the target geometry. This pre-placement preserved the symmetry relationship between motifs while allowing RFD2-RP to connect them into continuous monomers (**Fig. 1c**). Because the intervening hydrophobic repeat interface is degenerate, the same packing interaction could be reused between adjacent monomers along each side of the triangle, enabling assembly of the intended C3-symmetric triangular architecture.

Next we designed an additional C2 interface onto the C3-oligomer to form a 2D hexagonal lattice. The triangles were circularly permuted to allow the mineralization motifs in the interior of the triangle to be part of one chain, the permuted

triangles were positioned close together and a new C2 symmetric interface was diffused[28]. We designed a p6 hexagonal lattice in which the direction of the local C3 symmetry axes alternate or "flip" between nearest neighbors, correcting for out-of-plane curvature and balancing strain throughout the array. The resulting backbones were subjected to ProteinMPNN[29] sequence design while fixing key glutamate residues in the calcite interface from FD31. Designs for which monomer predictions with AF2 matched the monomer design model, and hexamer predictions with AF3 were close to two adjacent lattice new cells, were selected for experimental characterization.

Of 96 designs screened experimentally (**Table S1**), one, CALC_HEX24 (**Fig. 3a**), formed well-ordered two-dimensional arrays. Arrays formed in samples containing 85.7 μM protein and of 500 mM NaCl, and separated into a dense bottom fraction after overnight incubation at room temperature (**Fig. S4**). Assembly was confirmed by negative-stain transmission electron microscopy after ~457-fold dilution of the bottom phase into a buffer containing the same salt concentration (**Fig. 3b**). Removal of NaCl by dialysis resulted in reversible disassembly of the arrays (**Fig. 3c**) Two-dimensional class averages of the assembled arrays were processed in cryoSPARC[30] and revealed a hexagonal lattice composed of repeating triangular units (**Fig. 3d**). The disassembly of the arrays at low salt (likely due to electrostatic repulsion between the highly charged monomers) enables convenient soluble expression and purification of the single-component array, in contrast to typical one-component arrays that accumulate in insoluble inclusion bodies and require denaturation and refolding[31]. This salt-dependent assembly combines the practical advantages of single-component systems with key benefits of two-component arrays[32], in which assembly is induced by mixing two individually soluble subunits.

Following NaCl-induced assembly, we transferred the arrays into pre-mineralization conditions by introducing 5 mM $CaCl_2$ and removing NaCl through dialysis. Although stacking of the array sheets was observed during this process, the overall ordered array structure was preserved (**Fig. S5a**). In contrast, when protein was directly introduced into $CaCl_2$-containing solution without NaCl-induced pre-assembly, no ordered arrays were observed and instead nonspecific aggregation occurred (**Fig. S5b).**

$CaCO_3$ mineralization of both oligomeric and pre-assembled array CALC_HEX24 was carried out in 5 mM $CaCl_2$ and 5 mM $NaHCO_3$ at a protein concentration of 0.4 μM. High-resolution transmission electron microscopy (HRTEM) of mineralized CALC_HEX24 oligomers revealed small nanoparticles approximately 5–10 nm in diameter (**Fig. 3e, Fig. S6a**). Fast Fourier Transform (FFT) analysis of the HRTEM region containing multiple nanoparticles showed spots

corresponding to real-space d-spacings of 3.02 Å and 2.41 Å, indexed as the (104) and (110) reflections of the calcite-type phase, respectively, matching the design (**Fig. S6b**). A representative high-resolution image further resolved lattice fringes with spacings of 3.02 Å and 2.41 Å, assigned to calcite (104) and (110) planes, with an interfringe angle of approximately 53° (**Fig. 3e**).

We next examined mineralization by pre-assembled CALC_HEX24 arrays, and TEM imaging showed the formation of large crystals, up to ~2 μm (**Fig. S7**). Further HRTEM of mineralized arrays and corresponding FFT analysis showed spots indexed as the calcite (113) and (202) reflections, with an inter-reflection angle of ~30° in the FFT pattern. (**Fig. 3f**). By contrast, reactions lacking protein or containing negatively charged BSA produced vaterite-like morphologies (**Fig. S8**), consistent with previous reports[15]. AFM of the mineralized CALC_HEX24 shows that the protein arrays remain intact and exhibit layered structures on the substrate (**Fig. 3g**). Rhombohedral crystals with edge angles of approximately 102° were observed on the protein array layers (**Fig. 3h**), whereas no comparable structures were detected on exposed silicon wafer regions (**Fig. S9**). The height profile measured along with the pink line in **Fig. 3g** reveals two discrete height steps of 4.5 nm, consistent with stacking of two individual array layers (**Fig. 3i**). Together with the HRTEM analyses, these results show that CALC_HEX24 arrays induce localized, protein-templated calcite mineralization.

Designed templating of carbonate mineral not found in biological systems

We next sought to test whether our lattice-matching strategy could be extended beyond calcium carbonate to compounds not found in biological systems. We targeted $CoCO_3$, as cobalt is a critical metal for energy-relevant materials, and protein-directed cobalt enrichment could provide a genetically encoded route to control cobalt-based composites with potential applications in electrocatalysis and photoelectrochemical energy conversion[11]. $CoCO_3$ adopts the calcite-group rhombohedral/trigonal lattice[33], preserving the general carbonate sublattice organization found in calcite, but substitution of Ca^{2+} by the smaller Co^{2+} ion contracts the lattice, altering the spacing of surface-exposed interaction sites while maintaining an analogous carbonate arrangement.

To target the {101}[33] facet of $CoCO_3$ (**Fig. 4a**), we adapted the lattice-matching strategy used for ARAG64 by positioning side-chain atomic motifs designed to interact with the mineral surface. Tip-atom-guided RFD2-RP was then used to scaffold these motifs. Designs capable of propagating along the {101} lattice without steric clashes were linked through a second diffusion step to create an extended protein-mineral interface. Sequences were subsequently designed using

LigandMPNN, which introduced additional histidine residues positioned to coordinate surface Co^{2+} ions. These coordination interactions differ from the predominantly electrostatic interactions used in the preceding calcium carbonate[15] and calcium phosphate[17] designs. This workflow yielded COCA46, a designed repeat-protein scaffold targeting the $CoCO_3$ {101} facet (**Fig. 4b**). COCA46 runs as a monomer in size-exclusion chromatography (SEC) (**Fig. S10**) with CD spectrum exhibiting alpha-helical folds at 25°C (**Fig. 4c**). During concentration, the pink color from cobalt ions remained with COCA46 in the retentate, suggesting ion binding (**Fig. S11**).

COCA46 was incubated at 25°C with 1 mM $CoCl_2$ and 1 mM $NaHCO_3$, and time-resolved Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) was used to monitor carbonate-associated signal evolution during mineralization (**Fig. 4d–f**). Protein-containing reactions showed clear amide I and amide II bands[34], confirming the presence of protein during measurement, whereas these bands were absent from the no-protein control (**Fig. 4d**). The signal in the carbonate-associated vibrational regions[35,36] was greater in the 0.5μM COCA46 case compared to 0.5 μM FD31 and no protein controls under the same precursor conditions (**Fig. 4d**). Quantification of integrated carbonate-associated signal profiles (**Fig. 4e**) revealed the concentration-dependent carbonate-species accumulation across 0.1, 0.5, or 1.0 μM COCA46 conditions, with 0.5 μM COCA46 H→N knockout mutant (**Table S1**), and no-protein condition (**Fig. 4f**) serving as controls. The 0.1 μM and 0.5 μM COCA46 conditions showed similar kinetic profiles — a rapid early accumulation followed by a gradual decrease in carbonate-associated signal, with the 0.5 μM condition producing a larger signal amplitude than the 0.1 μM condition. The 1.0 μM condition showed a more sustained carbonate-associated signal over the measurement period. The non-linear scaling of the precipitation kinetics with protein concentration likely reflects excess Co^{2+}-binding sites that reduce free Co^{2+} availability and lower the effective supersaturation, thereby stabilizing carbonate-containing species before their conversion into mineral products.

Following ATR-FTIR measurements, we examined the resulting mineralization products by SEM and TEM. After 3 hours of mineralization, HRTEM of the 0.5 μM COCA46 condition revealed particles approximately 5 nm in diameter with lattice spacings of 2.74 Å and 1.95 Å, assigned to (104) plane and (202) reflection of rhombohedral $CoCO_3$, respectively, with corresponding FFT rings indexed to the {104} and {202} reflections (**Fig. 4g, Fig. S12a**). Because the (202) reflection is a second-order associated with (101) plane, the appearance of (202) lattice fringes in HRTEM supports the engagement of the intended non-natural facet. Additional 2.74 Å fringes, intersected at ~85°, are consistent with symmetry-related {104} planes (**Fig. S12b**), further supporting the assignment to calcite-structured $CoCO_3$. Over 3 days of incubation, these particles grew into larger crystals containing cobalt (**Fig. S13a-b**). HRTEM image and FFT analysis of the 3-day products

showed spots consistent with the (202) and (110) reflections of $CoCO_3$ (**Fig. 4h-i**), indicating progressive structural ordering of the crystal phase. In contrast, the no-protein control (**Fig. S14a-b**) and the previously reported calcite-binding repeat protein FD31[15] both yielded amorphous precipitates (**Fig. S15-16**), highlighting the role of the designed protein interface in directing $CoCO_3$ mineralization. Disruption of the designed interface through the COCA46 H→N mutations led to amorphous or cobalt hydroxide particles after 3 days (**Fig. S17**), indicating the loss of $CoCO_3$ templating capability. To further assess whether the mineralization capability is design-specific, we examined design COCA54 which targets the same {101} facet along a different spacing direction (**Fig. S18a-b**). COCA54 contains fewer histidines at the protein–mineral interface and a greater number of negatively charged side chains, and produces predominantly amorphous precipitates (**Fig. S19-20**). Taken together, these results indicate that productive mineral ordering depends on the precise side-chain chemistry being presented at the protein–mineral interface.

Design and characterization of $CoCO_3$ templating dihedral cage

Having established that the lattice-matched protein interfaces can direct $CoCO_3$ nucleation and growth, we next evaluated whether this design strategy could be extended to template and encapsulate $CoCO_3$ in a defined crystal geometry. To address this, we designed a D3-symmetric protein architecture with interior surfaces matching the $CoCO_3$ {110} facet lattice with LigandMPNN[26]. (**Fig. 5a**). Out of the 12 dihedral designs screened (**Table S1**), one, COCA_BRL11, was confirmed with negative-stain TEM to form the intended protein cages (**Fig. S21**). We succeeded in solving the crystal structure of COCA_BRL11, and found it to be in close agreement with the computational design model (1.99 Å RMSD) (**Fig. 5b; Table S2**) with a barrel-like assembly surrounding a hollow interior cavity.

Time-resolved ATR-FTIR measurements of COCA_BRL11 and the COCA_BRL11 H→N mutant under the mineralization conditions established for COCA46 (1 mM $CoCl_2$ and 1 mM $NaHCO_3$) showed distinct carbonate-associated kinetic profiles (**Fig. 5c, Fig. S22**). The mutant initially showed higher intensity, but was overtaken by the original COCA_BRL11 design at equimolar protein concentrations after 80 minutes, suggesting that both cage-scaffolds enrich carbonate-species under the same conditions, but the designed histidine-mediated interface changes their subsequent evolution. We next examined the mineral products formed in the 0.5 μM COCA_BRL11 condition. After 3 hours, well-dispersed nanoparticles were observed (**Fig. S23a**), and HRTEM revealed lattice spacings consistent with the (110), (104), and (012) planes of $CoCO_3$, confirming the growth of crystalline $CoCO_3$ nanocrystals (**Fig. S23b**). After 3 days, Scanning Electron Microscopy with Energy-Dispersive X-ray (SEM-EDX) confirmed the presence of cobalt in the resulting

particles (**Fig. S24**). HRTEM imaging showed these particles did not evolve into large single-crystals, but instead formed polycrystalline aggregates of small particles, and FFT analysis revealed real-space lattice spacings and reflection features consistent with the (104), (113), (110), and (202) spacings of $CoCO_3$ crystals (**Fig. S25**). In contrast, the COCA_BRL11 H→N mutant failed to produce $CoCO_3$ crystals under the same conditions and instead yielded amorphous precipitates (**Fig. S26a-b**). Together, these results indicate that the designed histidine-mediated protein–mineral interface is required to direct crystalline $CoCO_3$ formation.

To evaluate the potency of COCA_BRL11 in nucleating $CoCO_3$, we experimented with lower supersaturation conditions with 0.25 mM $CoCl_2$ and 0.25 mM $NaHCO_3$. Under these conditions, 0.3 μM COCA_BRL11 promoted the formation of small crystalline particles within 1 day, with HRTEM and corresponding FFT revealing real-space lattice spacings and reflection rings assigned to $CoCO_3$ (012) and (104) planes (**Fig. 5d-e, Fig. S27**). In contrast, under identical conditions, the no-protein control (**Fig. S26c**) and COCA46 (**Fig. S26d**) produced amorphous precipitates.

To probe Co^{2+} binding within the designed scaffold, we first use cryogenic electron microscopy (cryo-EM) to determine a holo crystal structure of COCA_BRL11 after soaking them with CoCl2. The structure showed Co2+ incorporation predominantly at the intended mineral-facing interfaces of COCA_BRL11 (**Fig. S28a**), with the protein backbone closely matching the design model (2.32 Å RMSD) (**Fig. S28b**). These bound Co2+ suggest that COCA_BRL11 interfaces could direct the formation of cobalt-based inorganic phases by templating and enriching Co2+. We then prepared a monodisperse $CoCO_3$-mineralized COCA_BRL11 population by size-exclusion chromatography (**Fig. S28c**) and analyzed it using cryo-EM. Representative micrographs and 2D class averages showed nanoparticles with a range of internal densities (**Fig. S28d**). The 3D reconstruction (**Fig. 5f**) had a 1.76 Å RMSD with the design model (**Fig. S28e**), supporting the retention of the overall D3 cage geometry after mineralization (**Fig. S29; Fig. S30; Table S3**), and shows additional density that cannot be attributed to the protein in its internal cavity–presumably from the $CoCO_3$ nanocrystals–adjacent to the designed $CoCO_3$ templating interfaces (**Fig. 5g-j**). These data support a model in which COCA_BRL11 acts as a protein container that templates and confines $CoCO_3$ crystal formation within a defined oligomeric framework.

Dihedral cage-templated $CoCO_3$ forms electrocatalytically active protein–mineral composites

To evaluate the catalytic function of the protein-cage-$CoCO_3$ hybrid, we incubated 0.3 μM COCA_BRL11, 0.25 mM $CoCl_2$ and 0.25 mM $NaHCO_3$ in water for 1 day at 25°C to generate COCA_BRL11-mineralized $CoCO_3$ particles for

electrochemical testing. Electrochemical measurements showed that COCA_BRL11-mineralized $CoCO_3$ sustained oxygen evolution reaction (OER) activity under alkaline conditions as evidenced by continual detection of current transients over time (**Fig. 5k**). Event-level analysis gave mean peak amplitudes of 25 pA, integrated charges of 14 fC, and full width at half maximum (FWHM) values of 0.54 ms, indicating finite, sub-millisecond OER-associated detection events (**Fig. 5l–n, Fig S31**). The OER response varied with the concentration of COCA_BRL11-mineralized $CoCO_3$ (**Fig. S32**), consistent with a particle concentration-particle detection frequency dependence. COCA_BRL11 alone showed no measurable activity above background, and $CoCO_3$ formed without protein showed only low activity (**Fig. S33**). Several controls with the same mineralization conditions, including BSA, the COCA_BRL11 H→N mutant, and DHR46 showed initial OER responses but failed to sustain activity over time (**Fig. S32**). These results suggest that the homogeneous $CoCO_3$ nanocrystals confined by COCA_BRL11 cages prolonged OER activity.

Discussion:

This work demonstrates that a generative *de novo* protein design framework can control mineralization across multiple length scales and modes of organization, including polymorph selection, unbounded growth on reconfigurable protein arrays, and bounded geometric confinement within protein cages. ARAG64 selectively promotes aragonite while FD31 favors calcite under identical conditions that otherwise produced mixed phases, showing that precise chemical presentation on an ordered interface is sufficient to bias phase outcomes. COCA_BRL11 confined $CoCO_3$ formation within the cage interior, producing uniform crystalline nanoparticles with the protein architecture remaining intact after mineralization. This result shows that designed protein structures can actively constrain and shape mineral growth rather than merely templating nucleation along a surface.

Previous efforts to template non-biological minerals such as ZnO and FeO have largely relied on threading metal-interacting residues onto flat repeat-protein scaffolds, followed by larger scale experimental screening or yeast-display selection against naturally occurring crystal faces[15,18]. More recently, lattice-matching strategies using oligomeric assemblies have been explored[17], but these approaches produced crystals with heterogeneous sizes. Here, we move beyond these strategies by using RFD-based repeat-protein design to directly embed geometry-optimized atomic motifs into multi-helix scaffolds, enabling programmable targeting of metastable polymorphs and non-natural crystal facets without requiring high-throughput screening. Because mineral surfaces present binding motifs periodically over an extended interface, rather than within a single localized pocket, repeat-protein architecture is particularly well suited to

mineral design: propagation of the mineral-binding interface is encoded directly in the generated backbone, and the interface can be lengthened or shortened simply by adding or removing repeat units — a degree of modularity not available to globular, single-motif scaffolds. This approach yields successful designs after testing only tens of candidates, without requiring high-throughput screening of thousands of sequences. This design logic extends from calcite and aragonite to $CoCO_3$, a mineral with no known biological precedent, substantially broadening the scope of protein-directed mineralization. Moreover, confinement within the COCA_BRL11 cage yielded $CoCO_3$ nanoparticles with remarkably coherent sizes. Together with previous demonstrations of mineral confinement in designed protein scaffolds, including hydroxyapatite and zinc oxide[17,18], these results expand the possibility of using de novo protein architectures to template and confine the formation of coherent inorganic nanoparticles.

Overall, our results show that the polymorph selectivity and spatial control achieved by native mineralizing matrices can be reproduced and extended using ordered, programmable protein scaffolds designed from scratch. Thus the results establish de novo design as a tractable route to recreate some of the most distinctive features of biological mineralization. Beyond fundamental insight, the framework carries practical significance: genetically encodable control over carbonate mineralization has potential applications in biocement and carbon management[37], while the electrocatalytic activity of COCA_BRL11-derived cobalt carbonate particles demonstrates that designed biomineralization scaffolds can produce functional hybrid bioinorganic materials.

Acknowledgments: PSK thanks C. Fries, N. Hanikel, A. Favor, S. Wang, S. Rankovic, and Z. Jones for discussion on 2d arrays. PSK thanks A. Favor and A. Kubaney for discussion on RFdiffusion and MPNN. PSK thanks P. Lund-Anderson for sharing his 10 mM $Co(II)Cl_2$ stock. PSK thanks J. Decarreau for guidance on the IN Cell Analyzer. PSK thanks A. Swartz, E. Huddy, Y. Politanska and S. Moroz and the rest of Baker lab members for general discussions. We thank the IPD lab management team led by K. VanWormer, IPD IT team led by L. Goldschmidt, IPD Electron Microscopy Research Core (EMRC) led by A. Borst for technical support. Joel Quispe, Fangfang Zhang, and Wentao Jiang for management of the University of
Washington Cryo-EM facilities. We thank Howard Hughes Medical Institute (DB). Design, production, and characterization of the aragonite and $CoCO_3$ targeting designs was performed at the IPD in the Biochemistry Department at UW through support from the DOE SC BES, as part of the Energy Frontier Research Centers (EFRC): CSSAS, The Center for the Science of Synthesis Across Scales, under Award DE-SC0019288. AFM, ATR-FTIR, SEM, EDS/EDX, HRTEM characterizations were also supported by CSSAS at PNNL through FWP 72448. PNNL is a multiprogram national laboratory operated for DOE by Battelle under Contract No. DE-AC05-76RL01830.
This research used resources 17-ID-2 (FMX) of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704. The Center for BioMolecular Structure (CBMS) is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1605010).

Funding:
Howard Hughes Medical Institute
The Audacious Project at the Institute for Protein Design
The Open Philanthropy Project Improving Protein Design Fund
This material is based on work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0506
This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, as part of the Energy Frontier Research Centers program: The Center for the Science of Synthesis Across Scales under Award Number DE-SC0019288.

Author contributions:
Conceptualization: PSK, XL, LTY, HP, JDY, DB
Methodology: PSK, XL, LTY, HP, CW, AJB, CCB, THL, AK, HN, JM, KDC, BC, YH, ZM, DL
Investigation: PSK, XL, LTY, HP, CW, AJB, CCB, THL, AK, HN, JM, KDC
Data processing: PSK, XL, LTY, CW, JM, HN, AK, SZ
Visualization: PSK, XL, LTY, CW, JM, HN, AK, SZ
Funding acquisition: SZ, JDY, BZ, BMC, DB
Supervision: AJB, BZ, BMC, SZ, AKB, JDY, DB
Writing – original draft: PSK, XL, LTY, HP, CW, DB
Writing – review & editing: All authors

Competing interests: Authors declare that they have no competing interests.

Data and materials availability:
All data are available in the main text or the supplementary materials. Scripts for RFdiffusion 2 are available at https://github.com/RosettaCommons/RFdiffusion2. Scripts for RFdiffusion for symmetric motif scaffolding are available at https://github.com/RosettaCommons/RFdiffusion. Scripts for ProteinMPNN and LigandMPNN are available at https://github.com/dauparas/LigandMPNN. Scripts for AlphaFold2 and AlphaFold3 are available at https://github.com/google-deepmind/alphafold and https://github.com/google-deepmind/alphafold3 respectively. All detailed scripts for design pipelines and fine-tuned weights will be available on github at the time of publication.

References

1. Suzuki, M. *et al.* An Acidic Matrix Protein, Pif, Is a Key Macromolecule for Nacre Formation. *Science* **325**, 1388–1390 (2009).
2. Nudelman, F. *et al.* The role of collagen in bone apatite formation in the presence of hydroxyapatite nucleation inhibitors. *Nat. Mater.* **9**, 1004–1009 (2010).

3. Chen, C.-L., Bromley, K. M., Moradian-Oldak, J. & DeYoreo, J. J. In situ AFM Study of Amelogenin Assembly and Disassembly Dynamics on Charged Surfaces Provides Insights on Matrix Protein Self-Assembly. *J. Am. Chem. Soc.* **133**, 17406–17413 (2011).
4. Siponen, M. I. *et al.* Structural insight into magnetochrome-mediated magnetite biomineralization. *Nature* **502**, 681–684 (2013).
5. Bahn, S. Y., Jo, B. H., Choi, Y. S. & Cha, H. J. Control of nacre biomineralization by Pif80 in pearl oyster. *Sci. Adv.* **3**, e1700765 (2017).
6. Sun, J. & Bhushan, B. Hierarchical structure and mechanical properties of nacre: a review. *RSC Adv.* **2**, 7617 (2012).
7. Falini, G., Albeck, S., Weiner, S. & Addadi, L. Control of Aragonite or Calcite Polymorphism by Mollusk Shell Macromolecules. *Science* **271**, 67–69 (1996).
8. Liang, H., Liu, Y., Tian, B., Li, Z. & Ou, H. A sustainable production of biocement via microbially induced calcium carbonate precipitation. *Int. Biodeterior. Biodegrad.* **172**, 105422 (2022).
9. Magnetically sensitive proteins could lead to new imaging tools and remote-controlled drugs. https://doi.org/10.1126/science.zzfau8s (2026).
10. Tang, S., Dong, Z., Ke, X., Luo, J. & Li, J. Advances in biomineralization-inspired materials for hard tissue repair. *Int. J. Oral Sci.* **13**, 42 (2021).

11. Zhang, X., Yi, H., An, Q. & Song, S. Recent advances in engineering cobalt carbonate hydroxide for enhanced alkaline water splitting. *J. Alloys Compd.* **887**, 161405 (2021).

12. Boskey, A. L. & Villarreal-Ramirez, E. Intrinsically disordered proteins and biomineralization. *Matrix Biol.* **52–54**, 43–59 (2016).

13. Yang, W., Wang, S. & Lee, G. R. The past, present and future of de novo protein design. *Nature* **652**, 1139–1152 (2026).

14. Zhang, S., Pyles, H., Baker, D. & De Yoreo, J. J. Designing protein–material interfaces. *MRS Bull.* **50**, 1454–1464 (2025).

15. Davila-Hernandez, F. A. *et al.* Directing polymorph specific calcium carbonate formation with de novo protein templates. *Nat. Commun.* **14**, 8191 (2023).

16. Voets, I. K. From ice-binding proteins to bio-inspired antifreeze materials. *Soft Matter* **13**, 4808–4823 (2017).

17. Yu, L. T. *et al.* Templating and confining calcium phosphate mineralization within designed protein assemblies. Preprint at https://doi.org/10.64898/2026.01.14.699524 (2026).

18. Saragovi, A. *et al.* De novo design of metal-oxide templating proteins. Preprint at https://doi.org/10.1101/2024.06.24.600095 (2024).

19. De Haas, R. J. *et al.* Inhibition of ice recrystallization with designed twistless helical repeat proteins. *Proc. Natl. Acad. Sci.* **122**, e2514871122 (2025).

20. Ridgwell, A. & Zeebe, R. The role of the global carbonate cycle in the regulation and evolution of the Earth system. *Earth Planet. Sci. Lett.* **234**, 299–315 (2005).

21. McDonald, K., Khajuria, D. K. & Li, B. Potential healthcare applications of metal carbonate nanoparticles produced from emerging carbon capture technologies. *Nanomed.* **21**, 921–923 (2026).

22. Ahern, W. *et al.* Atom-level enzyme active site scaffolding using RFdiffusion2. *Nat. Methods* https://doi.org/10.1038/s41592-025-02975-x (2025) doi:10.1038/s41592-025-02975-x.

23. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

24. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

25. Caspi, E. N., Pokroy, B., Lee, P. L., Quintana, J. P. & Zolotoyabko, E. On the structure of aragonite. *Acta Crystallogr. B* **61**, 129–132 (2005).

26. Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *Nat. Methods* **22**, 717–723 (2025).

27. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

28. Tran, L. *et al.* Design of Orthogonal Far-Red, Orange and Green Fluorophore-binding Proteins for Multiplex Imaging. Preprint at https://doi.org/10.1101/2025.08.03.668343 (2025).

29. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

30. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

31. Gonen, S., DiMaio, F., Gonen, T. & Baker, D. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **348**, 1365–1368 (2015).

32. Ben-Sasson, A. J. *et al.* Design of biologically active binary protein 2D materials. *Nature* **589**, 468–473 (2021).

33. Zhao, Q. & Kulik, H. J. Where Does the Density Localize in the Solid State? Divergent Behavior for Hybrids and DFT+U. *J. Chem. Theory Comput.* **14**, 670–683 (2018).

34. Sizeland, K. H. *et al.* Nanostructure of electrospun collagen: Do electrospun collagen fibers form native structures? *Materialia* **3**, 90–96 (2018).

35. Henry, D. G., Watson, J. S. & John, C. M. Assessing and calibrating the ATR-FTIR approach as a carbonate rock characterization tool. *Sediment. Geol.* **347**, 36–52 (2017).

36. Grunenwald, A. *et al.* Revisiting carbonate quantification in apatite (bio)minerals: a validated FTIR methodology. *J. Archaeol. Sci.* **49**, 134–141 (2014).

37. Beatty, D. N., Williams, S. L. & Srubar, W. V. Biomineralized Materials for Sustainable and Durable Construction. *Annu. Rev. Mater. Res.* **52**, 411–439 (2022).

38. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

39. Gražulis, S. *et al.* Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res.* **40**, D420–D427 (2012).

40. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

41. Dang, B. *et al.* SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019).

42. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).

43. Nannenga, B. L., Iadanza, M. G., Vollmar, B. S. & Gonen, T. Overview of Electron Crystallography of Membrane Proteins: Crystallization and Screening Strategies Using Negative Stain Electron Microscopy. *Curr. Protoc. Protein Sci.* **72**, (2013).

44. Mastronarde, D. N. SerialEM: A Program for Automated Tilt Series Acquisition on Tecnai Microscopes Using Prediction of Specimen Position. *Microsc. Microanal.* **9**, 1182–1183 (2003).

45. Sun, M. *et al.* Practical considerations for using K3 cameras in CDS mode for high-resolution and high-throughput single particle cryo-EM. *J. Struct. Biol.* **213**, 107745 (2021).

46. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

47. Kim, K., Li, H. & Clarke, O. B. High-resolution ab initio reconstruction enables cryo-EM structure determination of small particles. Preprint at https://doi.org/10.1101/2025.09.08.674935 (2025).

48. Pettersen, E. F. *et al.* UCSF ChimeraX : Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

49. Emsley, P. & Cowtan, K. *Coot* : model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).

51. Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 136–153 (2015).

52. Croll, T. I. *ISOLDE* : a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 519–530 (2018).

53. Adams, P. D. *et al.* The Phenix software for automated determination of macromolecular structures. *Methods* **55**, 94–106 (2011).

54. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).

55. Winter, G. *xia2* : an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).

56. Winter, G. *et al. DIALS* : implementation and evaluation of a new integration package. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 85–97 (2018).

57. Winn, M. D. *et al.* Overview of the *CCP* 4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

58. Wojdyr, M., Keegan, R., Winter, G. & Ashton, A. *DIMPLE* - a pipeline for the rapid generation of difference maps from protein crystals with putatively bound ligands. *Acta Crystallogr. A* **69**, s299–s299 (2013).

59. Murshudov, G. N. *et al. REFMAC* 5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).

60. Smart, O. S. *et al.* Exploiting structure similarity in refinement: automated NCS and target-structure restraints in *BUSTER*. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 368–380 (2012).

61. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The *PDB_REDO* server for macromolecular structure model optimization. *IUCrJ* **1**, 213–220 (2014).

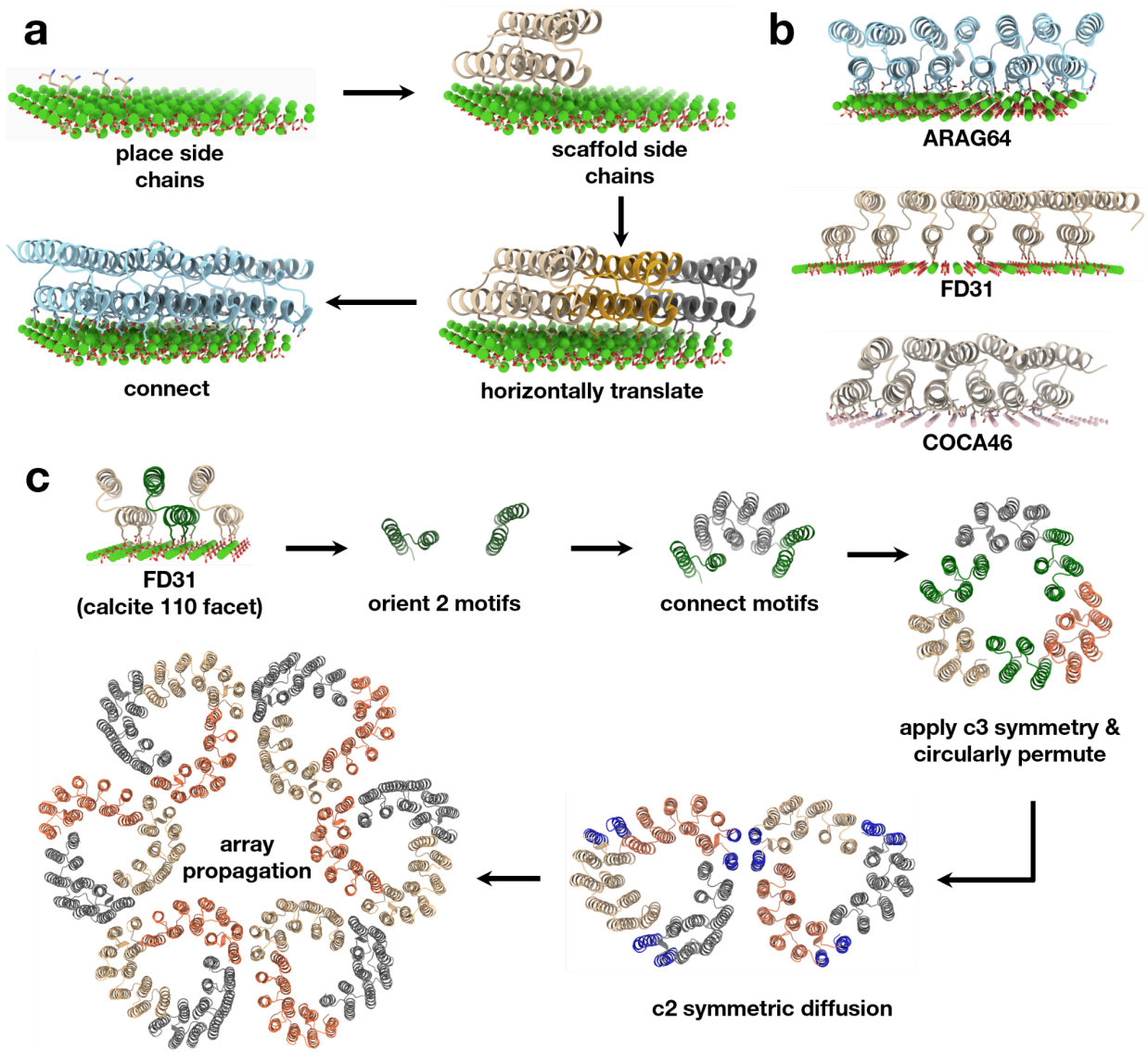


Figure 1: Design of protein–mineral interfaces and higher-order mineral-templating assemblies. a, Computational design strategy for generating extended protein–mineral interfaces against the aragonite {110} facet. Repeat-protein scaffolds were

generated using a fine-tuned RFdiffusion2 model that was trained on repeat proteins from the PDB and AlphaFold databases. **b,** Representative designs of de novo protein–mineral interface, ARAG64, FD31 and COCA46, targeting aragonite, calcite and cobalt carbonate, respectively. **c,** Strategy for converting the FD31 calcite-binding motif into an assembly array. FD31 protein–calcite motifs are shown in green and were used to design C3-symmetric triangular building blocks, followed by circular permutation and the design of a C2-symmetric inter-triangle interface, shown in blue, to assemble the array.

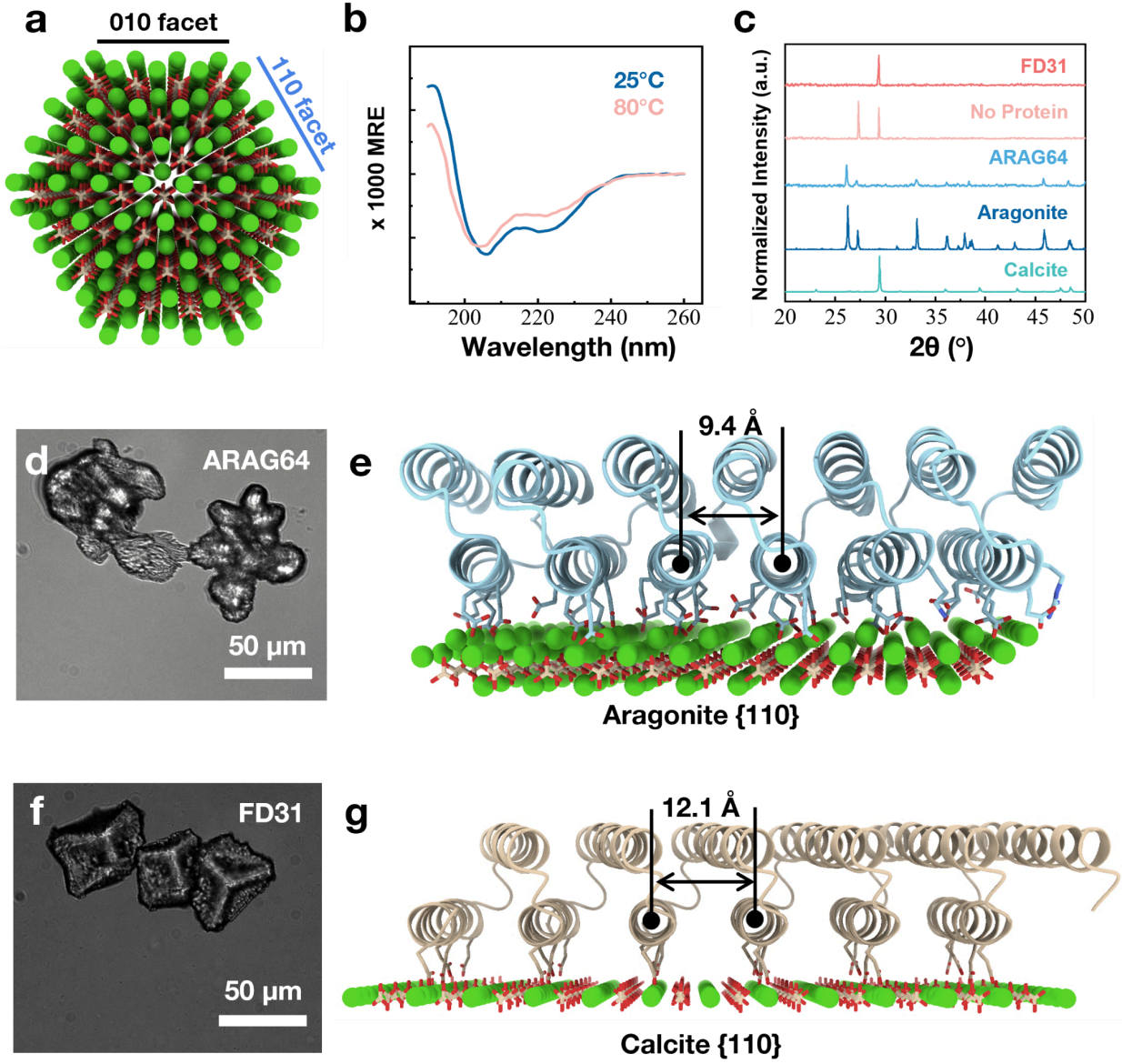

a
010 facet
110 facet
b
25°C
80°C
x 1000 MRE
200
220
240
260
Wavelength (nm)
c
FD31
No Protein
ARAG64
Aragonite
Calcite
Normalized Intensity (a.u.)
20
25
30
35
40
45
50
2θ (°)
d
ARAG64
50 µm
e
9.4 Å
Aragonite {110}
f
FD31
50 µm
g
12.1 Å
Calcite {110}

Figure 2: Aragonite and calcite templated by designed proteins. a, Atomic model of aragonite viewed along the c-axis, highlighting the {010} (black) and {110} (blue) facets. **b,** CD spectra of ARAG64 protein showing the α-helical structure at both 25°C and 80°C. **c,** Normalized XRD patterns of $CaCO_3$ crystal formed in the presence of ARAG64 and FD31 proteins, compared with a no protein control and the references of aragonite and calcite from RRUFF entries of R060195 and R050130, respectively. No-protein controls show a characteristic aragonite peak near 27.2°, assigned to the {021} reflection, and a calcite peak near 29.4°, assigned to the {104} reflection. **d,** Optical microscopy image of aragonite crystals templated by ARAG64 protein. **e,** Computational design model of ARAG64 protein positioned on the aragonite {110} surface, with the designed binding motifs at a distance of 9.4Å for lattice match. **f,** Optical microscopy image of calcite crystals templated by FD31 protein. **g,** Computational design model of FD31 protein positioned on the calcite {110} surface, with designed binding motifs at a 12.1 Å surface-matching spacing.

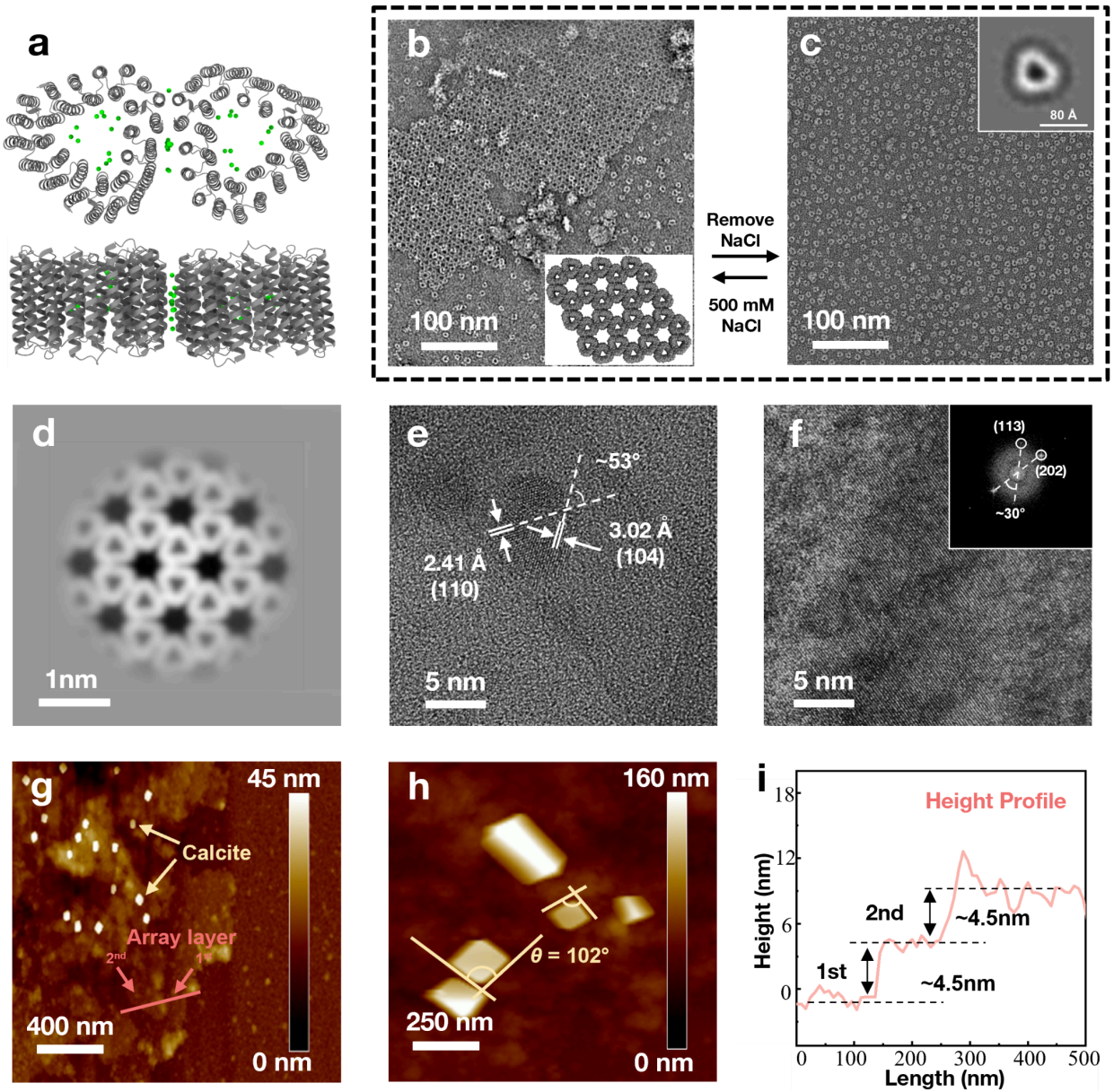

a
b
100 nm
c
80 Å
Remove NaCl
500 mM NaCl
100 nm
d
1nm
e
~53°
3.02 Å
(104)
2.41 Å
(110)
5 nm
f
(113)
(202)
~30°
5 nm
g
45 nm
Calcite
Array layer
2nd
1st
400 nm
0 nm
h
160 nm
θ = 102°
250 nm
0 nm
i
Height Profile
Height (nm)
18
12
6
0
2nd
~4.5nm
1st
~4.5nm
0
100
200
300
400
500
Length (nm)

Figure 3. Calcite biomineralization templated by reconfigurable CALC_HEX24 arrays. a, Axial and lateral views of the AlphaFold 3 model of CALC_HEX24. The calcium ions shown in green stabilize the inter-triangle interfaces. **b-c,** Negative-stain TEM images showing salt-dependent reversible assembly between arrays **(b)**, and oligomers **(c)**. The inset in panel **b** shows a representative 2D-class average of the oligomer, and the inset in panel **c** shows the propagated array model. **d,** Negative-stain TEM 2D-class average of the designed array. **e,** Unstained HRTEM image of calcite nanocrystals templated by the oligomer. The shown lattice fringes with spacings of 3.02 and 2.41Å and interfringe angle of ~53°, consistent with the calculated angle between calcite (104) and (110) planes. **f,** Unstained HRTEM image of calcite nanocrystals templated by the arrays. The inset FFT pattern shows an inter-reflection angle of ~30° between spots indexed to the calcite (113) and (202) reflections. **g,** AFM image of mineralized arrays. **h,** High-magnification AFM image of the crystals templated on the protein arrays with angles consistent with the 102° angle of rhombohedral calcite. **i,** corresponding height profile measured along the red line in Panel g, showing step heights of approximately 4.5 nm.

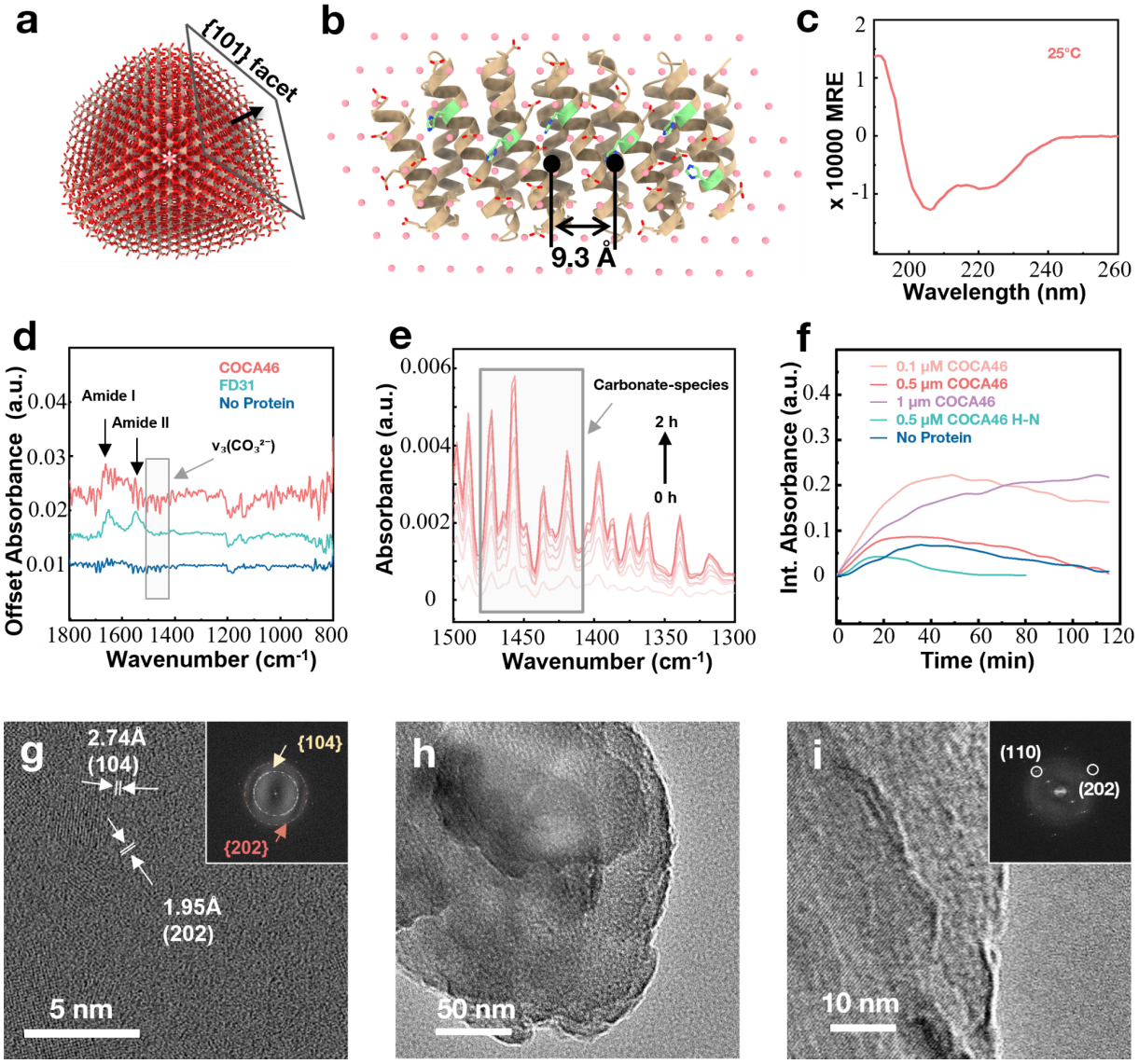

a
{101} facet
b
9.3 Å
c
25°C
x 10000 MRE
Wavelength (nm)
d
COCA46
FD31
No Protein
Amide I
Amide II
$v_3(CO_3^{2-})$
Offset Absorbance (a.u.)
Wavenumber (cm-1)
e
Carbonate-species
2 h
0 h
Absorbance (a.u.)
Wavenumber (cm-1)
f
0.1 μM COCA46
0.5 μm COCA46
1 μm COCA46
0.5 μM COCA46 H-N
No Protein
Int. Absorbance (a.u.)
Time (min)
g
2.74Å
(104)
1.95Å
(202)
{104}
{202}
5 nm
h
50 nm
i
(110)
(202)
10 nm

Figure 4. $CoCO_3$ templated by COCA46 protein. a, Crystal structure of $CoCO_3$, with an exposed {101} facet marked in grey. **b,** Design model of COCA46 protein targeting the $CoCO_3$ {101} facet with designed binding motifs at a 9.3 Å lattice-matching spacing. **c,** CD spectrum of COCA46 protein shows α-helical folds. **d,** ATR-FTIR spectra collected after 2 hours of mineralization with 0.5 μM COCA46 (red), 0.5 μM FD31 (green), and the no-protein control (blue) in 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25°C. Protein-containing reactions show amide I and amide II bands at 1650 and 1540 cm^{-1}, whereas these bands are absent from the no-protein control. The carbonate-associated vibrational region used for analysis is shaded. **e,** Representative time-resolved ATR-FTIR spectra showing the evolution of carbonate-associated peaks during mineralization from 0-2 hours. The shaded region (1410 to 1480 cm^{-1}) indicates the carbonate v_3 integration window used for carbonate quantification plotted in panel **f**. **f,** Integrated (Int.) carbonate-associated signal as a function of time for $CoCO_3$ mineralization with COCA46 protein at different concentrations (0.1, 0.5, and 1 μM), the 0.5 μM COCA46 H→N interface mutant, and the no-protein control. **g,** HRTEM image of CoCO3 nanocrystal after 3 hours of mineralization with 0.5 μM COCA46 in 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25°C. The lattice fringes with spacings of 2.74 and 1.95 Å are assigned to the $CoCO_3$ (104) and (202) planes. The inset FFT shows the rings indexed to {104} and {202} reflections. **h-i,** TEM images of $CoCO_3$ particles after 3 days of mineralization in the mineralization conditions of Panel g. The inset FFT panel **i** shows spots indexed as the $CoCO_3$ (110) and (202) reflections.

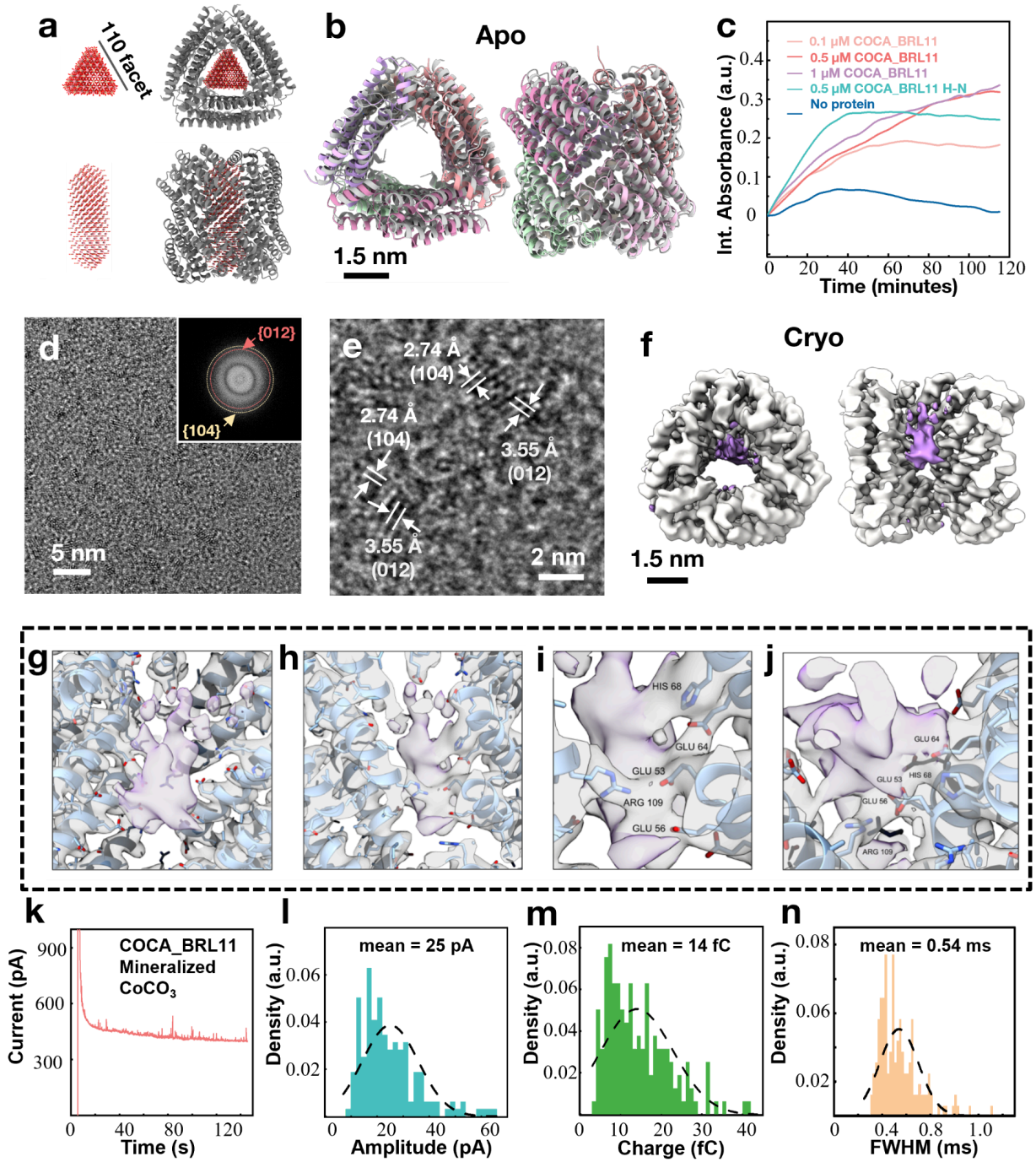
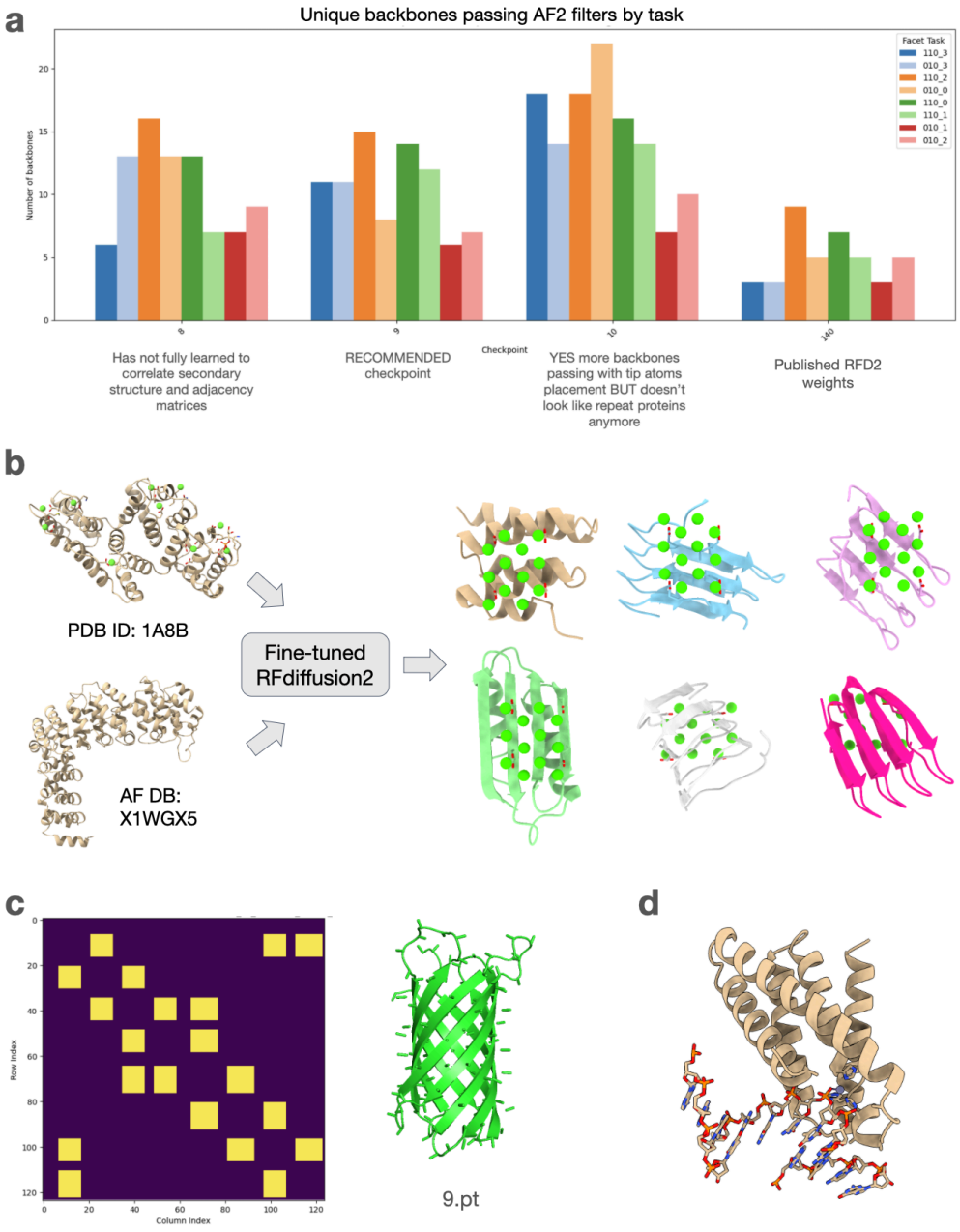
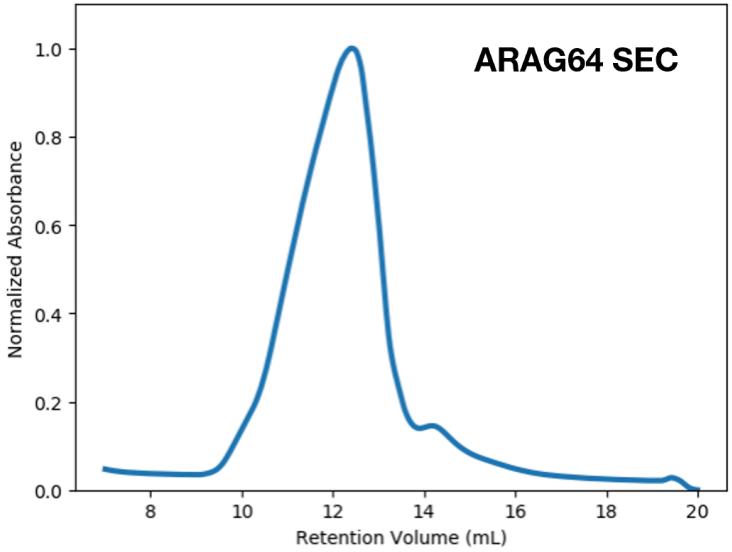

a
110 facet
b
Apo
1.5 nm
c
0.1 μM COCA_BRL11
0.5 μM COCA_BRL11
1 μM COCA_BRL11
0.5 μM COCA_BRL11 H-N
No protein
Int. Absorbance (a.u.)
Time (minutes)
d
{012}
{104}
5 nm
e
2.74 Å (104)
2.74 Å (104)
3.55 Å (012)
3.55 Å (012)
2 nm
f
Cryo
1.5 nm
g
h
i
HIS 68
GLU 64
GLU 53
ARG 109
GLU 56
j
GLU 64
GLU 53
HIS 68
GLU 56
ARG 109
k
COCA_BRL11
Mineralized
CoCO3
Current (pA)
Time (s)
l
mean = 25 pA
Density (a.u.)
Amplitude (pA)
m
mean = 14 fC
Density (a.u.)
Charge (fC)
n
mean = 0.54 ms
Density (a.u.)
FWHM (ms)

Figure 5. Cobalt carbonate templating and confinement by a designed protein cage. a, Design model of COCA_BRL11 protein, a D3-symmetric oligomer with protein interfaces that have lattices matched with the $CoCO_3$ {110} facets. **b,** Apo crystal structure of COCA_BRL11 aligned to the design model (gray). **c,** Time-resolved FTIR as a function of time for $CoCO_3$ mineralization with COCA_BRL11 protein at different concentrations (0.1, 0.5, and 1 μM), the 0.5 μM COCA_BRL11 H→N interface mutant, and the no-protein control in 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at 25°C and pH 8.2. **d,** Unstained HRTEM image of CoCO3 nanocrystals formed with COCA_BRL11 after 1 day mineralization. The inset FFT shows rings indexed to the {104} and {012} reflections. **e,** HRTEM image showing lattice fringes with spacings of 2.74 and 3.55 Å, assigned to $CoCO_3$ (104) and (012) planes, respectively. **f,** Cryo-EM maps in top and cutaway-side views with a 3.46 Å resolution refined without symmetry (C1). $CoCO_3$ and protein densities are shown in purple and gray, respectively. shown. **g,** Transparent cutaway-side view zoomed in Panel f. **h,** 180° view of Panel g showing the amino acid residues-mineral interfaces. **i,** Zoom in view of Panel h showing the labelled residues with continuous density with $CoCO_3$. **j,** Top view of Panel i showing the amino acid residues-mineral interface 90° view of (i). **k**, Representative current–time trace of COCA_BRL11-mineralized $CoCO_3$ particles recorded in 0.1 M KOH at 1.0 V versus Ag/AgCl using a carbon fiber microelectrode. **l–n**, Histograms of baseline-corrected peak amplitude (**l**), integrated charge (**m**), and FWHM (**n**) for isolated OER-associated current transients while excluding any signals attributed to the simultaneous collision of multiple particles or particle aggregates. Dashed curves serve as guides to the distributions. Values indicate mean values from the analyzed peaks.

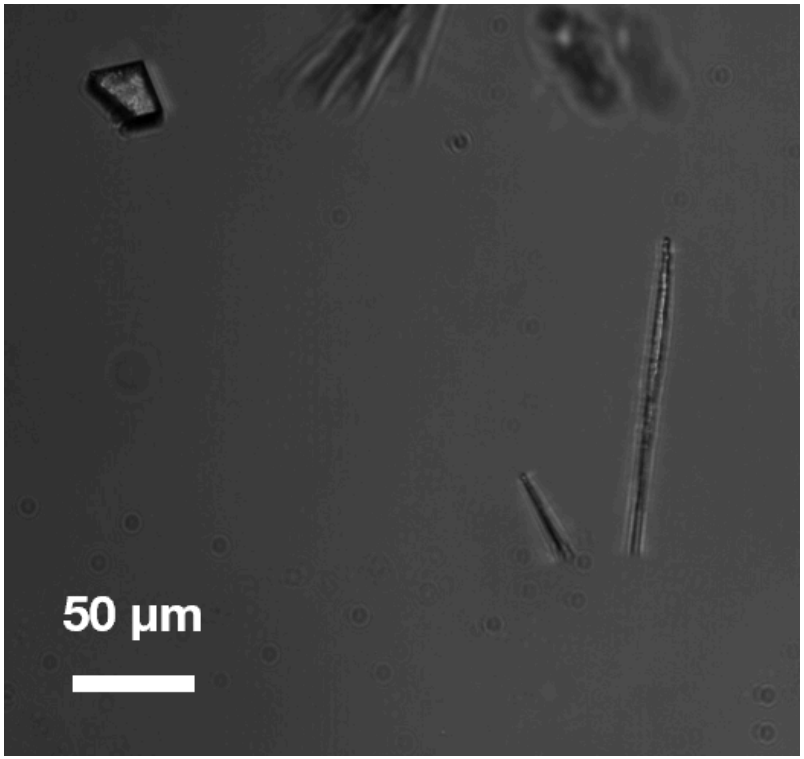
Supplemental data figures:


a
Unique backbones passing AF2 filters by task
Facet Task
110_3
010_3
110_2
010_0
110_0
110_1
010_1
010_2
Number of backbones
Checkpoint
Has not fully learned to correlate secondary structure and adjacency matrices
RECOMMENDED checkpoint
YES more backbones passing with tip atoms placement BUT doesn't look like repeat proteins anymore
Published RFD2 weights
b
PDB ID: 1A8B
AF DB: X1WGX5
Fine-tuned RFdiffusion2
c
Row Index
Column Index
9.pt
d

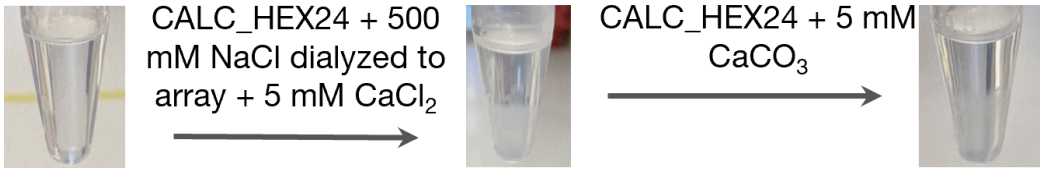
Supplementary Fig. 1. RFD2-RP. (a) Checkpoints 8, 9 and 10 on the x axis correspond to training epochs continued from checkpoint 140 of the original RFdiffusion 2 model[22]. All three fine-tuned checkpoints outperformed the original model weights across a range of biomineralization benchmarks, including scaffolding eight four-side-chain enumerations targeting the aragonite {010} and {110} facets, with four enumerations evaluated per facet. Successful scaffolds were defined by a side-chain RMSD < 1.5 Å, backbone RMSD < 2.0 Å, and pLDDT > 80. (b) Training examples for RFD2 enable the generation of designable protein backbones that engage an aragonite facet. (c) Checkpoint 9 shows better adherence to adjacency-matrix conditioning than checkpoint 10, illustrated here for a challenging β-barrel repeat-protein target. (d) Checkpoint 9 also generalizes to enzyme-design tasks. Shown is scaffolding of three histidine tip atoms coordinating a zinc ion for DNA cleavage.



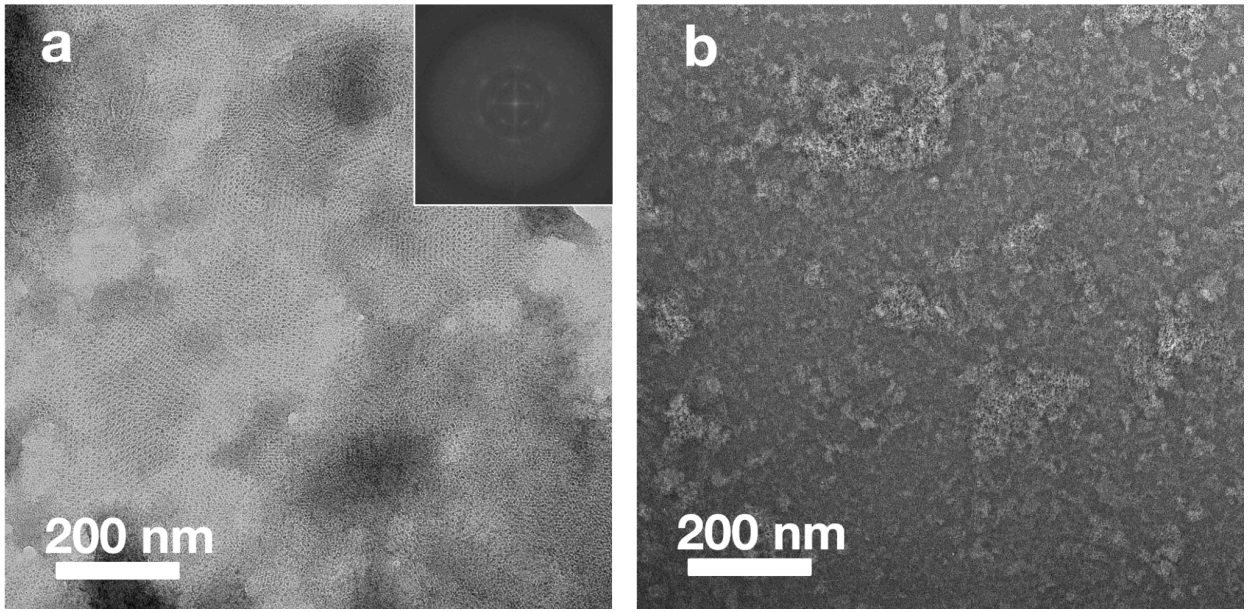
Supplementary Fig. 2. ARAG64 size-exclusion chromatography trace, monitored by UV absorbance at 280 nm, indicating predominantly monomeric species.



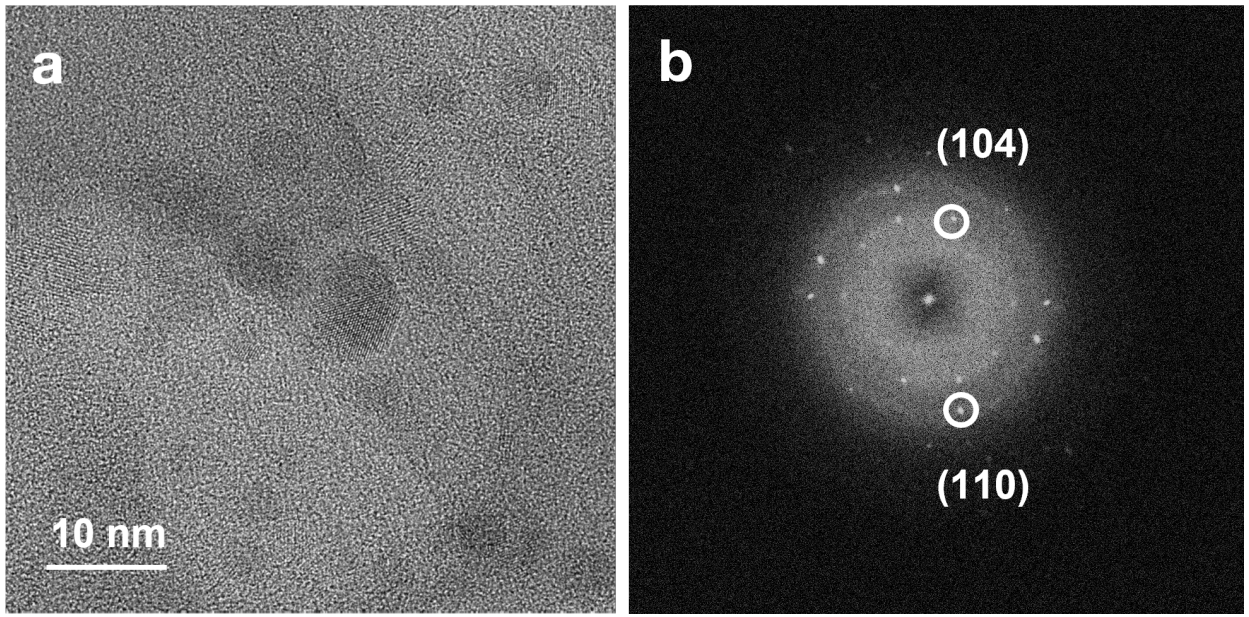
Supplementary Fig. 3. Optical microscopy of $CaCO_3$ mineralization without protein. Under mineralization conditions containing 10 mM $CaCl_2$ and 10 mM $NaHCO_3$ at 80°C and pH 8.2 for 2 hours, the no-protein control yields rhombohedral calcite and needle-like aragonite crystals.



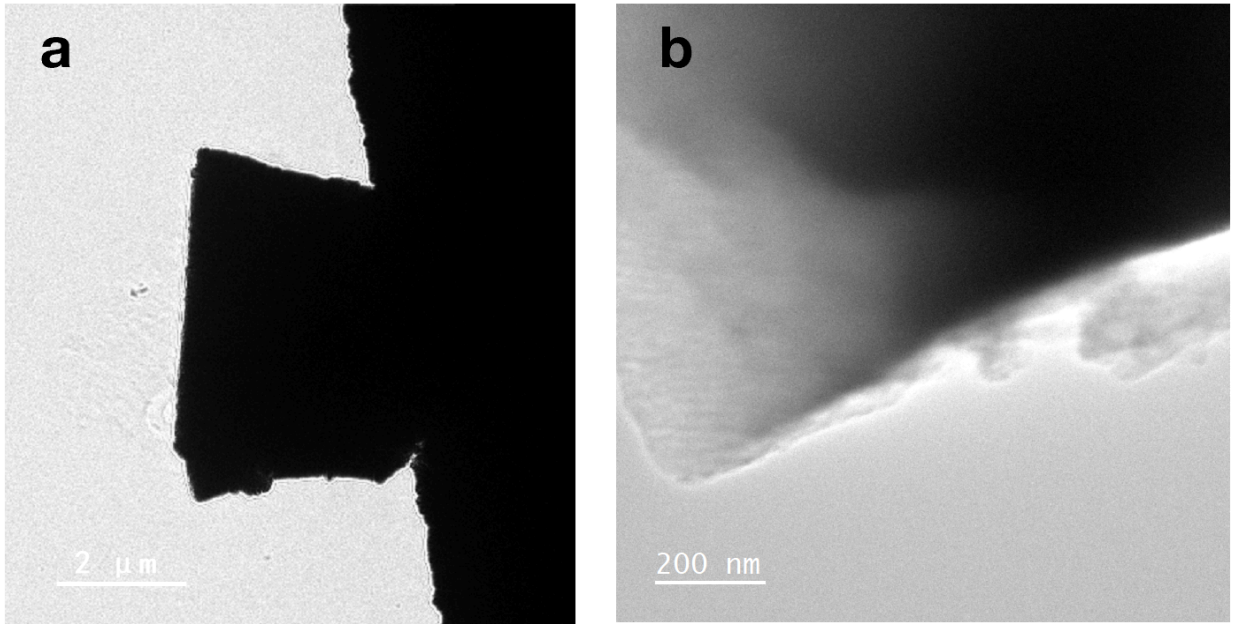
Supplementary Fig. 4. Appearance of unmineralized and mineralized bulk 2D array material under the indicated conditions.



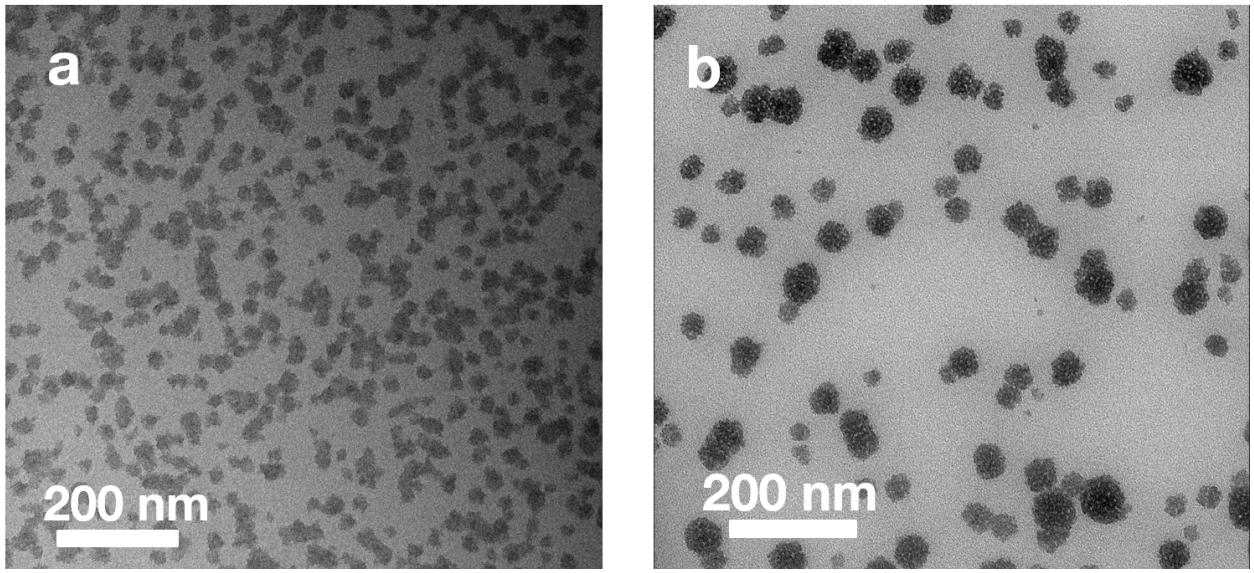
Supplementary Fig. 5. Salt-dependent reconfiguration of CALC_HEX24 arrays. **a.** Negative-stain TEM image of CALC_HEX24 after addition of 500 mM NaCl followed by dilution into 5 mM $CaCl_2$, with the corresponding FFT analysis, showing ordered array sheets. **b.** Negative-stain TEM image of arrays diluted directly into 1 mM $CaCl_2$,without NaCl-induced pre-assembly, showing nonspecific aggregation.



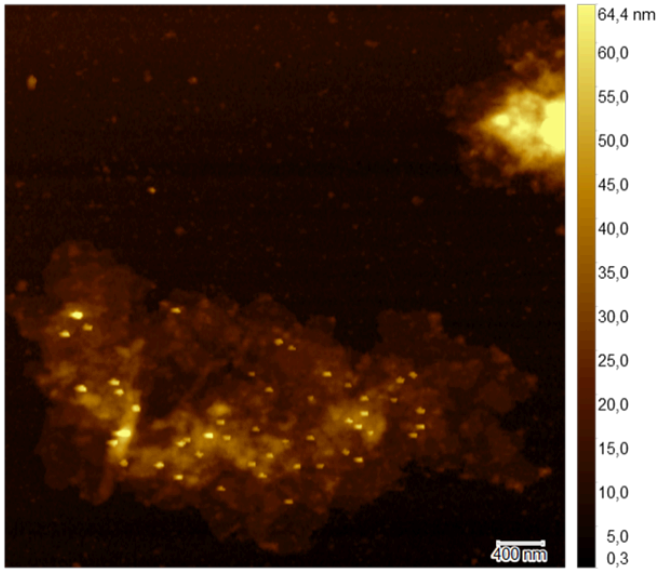
Supplementary Fig. 6. HRTEM and FFT analysis of $CaCO_3$ mineralization by non-assembled 0.4 μM CALC_HEX24 with 5 mM $CaCl_2$ and 5 mM $NaHCO_3$. **a,** Unstained HRTEM image of mineral formed with non-assembled CALC_HEX24, showing small nanoparticles. **b,** FFT pattern from the HRTEM image in a, showing the calcite (104) and (110) reflections, respectively.



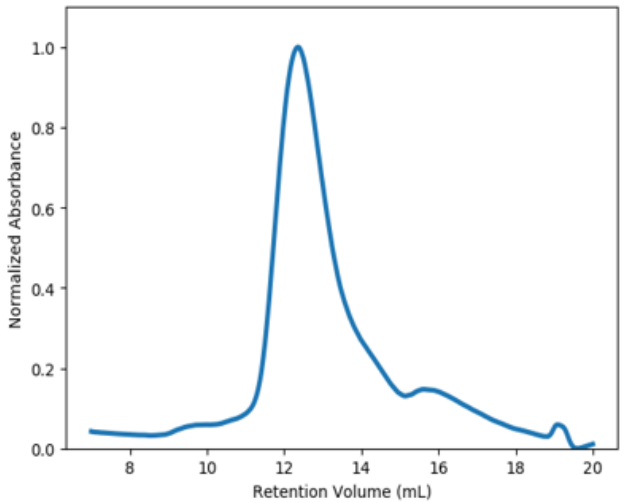
Supplementary Fig. 7. TEM image showing large crystals formed in the presence of CALC_HEX24 array. Low-(**a**) and higher-(**b**) magnification TEM images showing large crystals, up to ~2 µm,



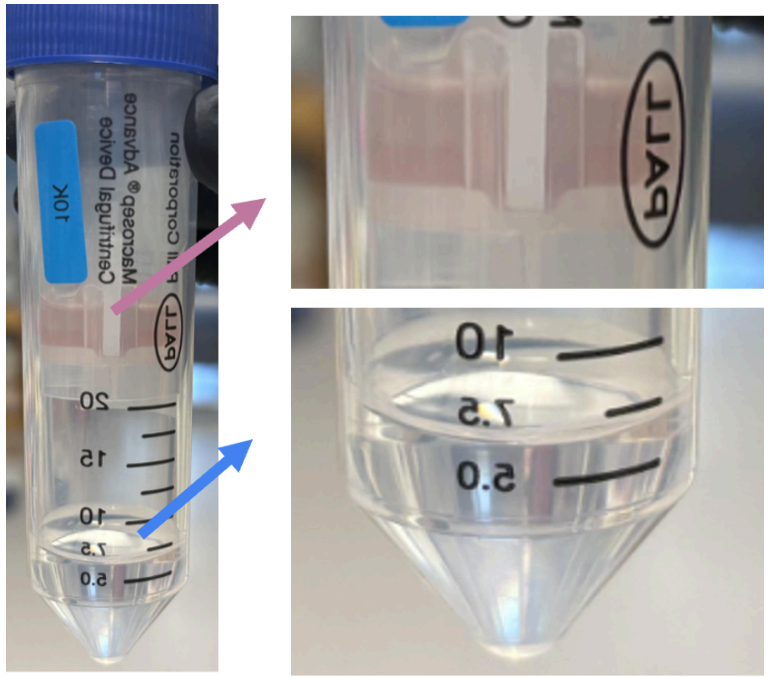
Supplementary Fig. 8. TEM images of $CaCO_3$ mineralization controls. **a.** Unstained TEM image of the no-protein control **a**, and the bovine serum albumin (BSA) control **b**, showing vaterite-like particles similar to that reported previously.[15]



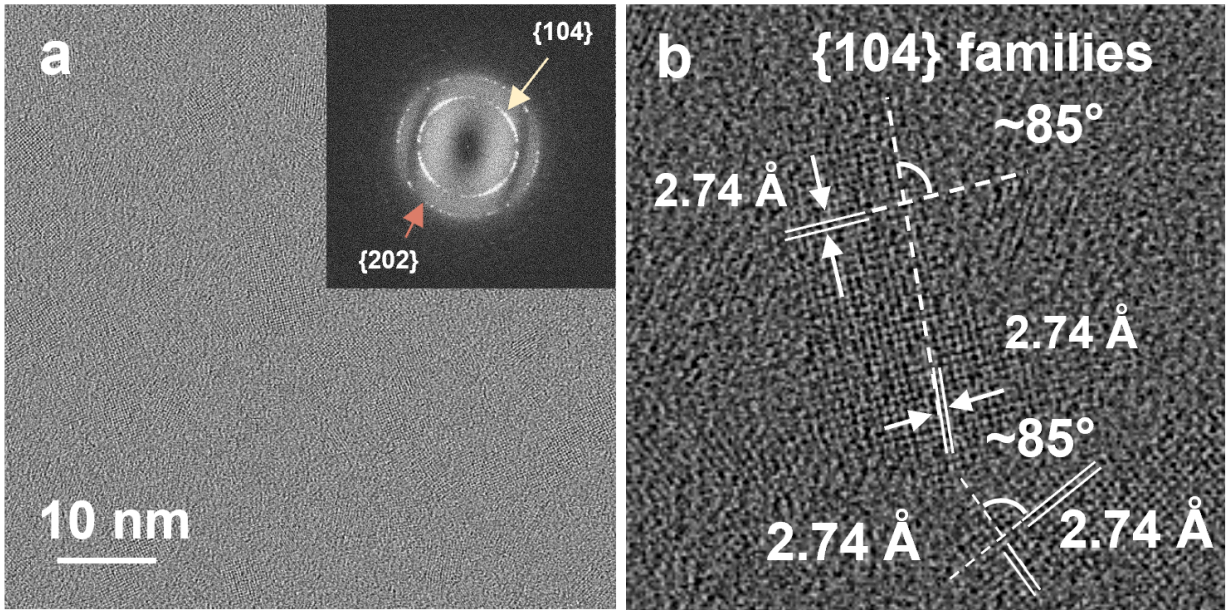
Supplementary Fig. 9. AFM overview of localized CaCO3 mineralization on CALC_HEX24 array. Zoomed-out AFM image of the mineralized array, showing $CaCO_3$ formation on the protein array with no crystals observed on the surrounding silicon wafer background.



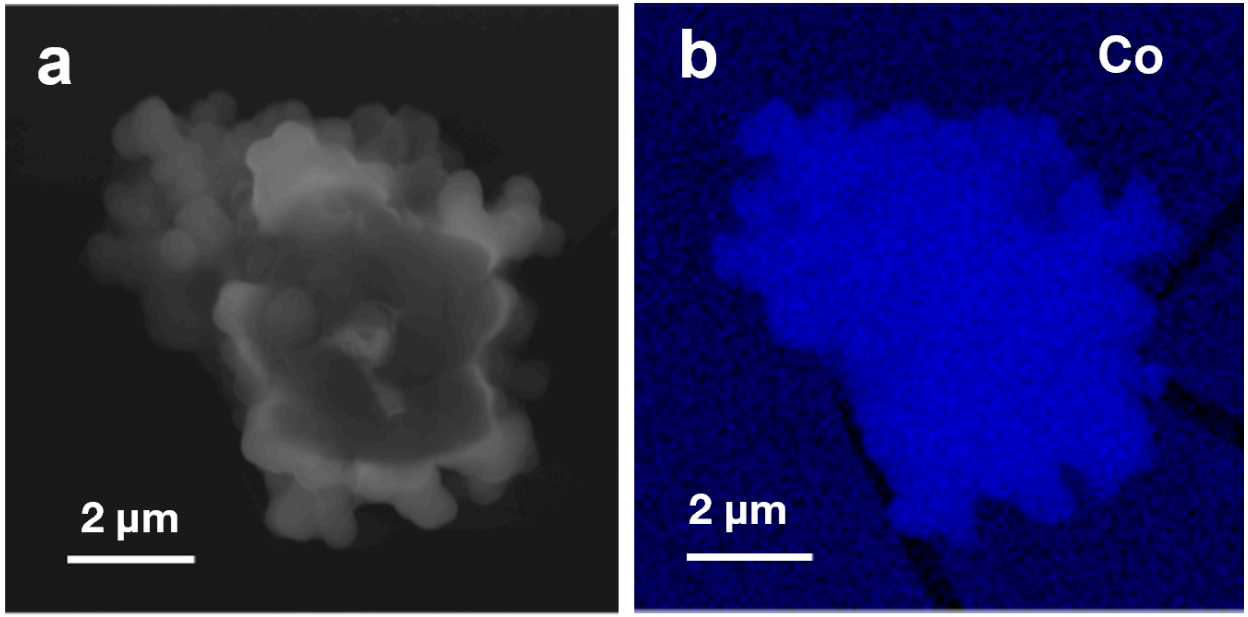
Supplementary Fig. 10. COCA46 size-exclusion chromatography trace monitored by UV absorbance at 280 nm, indicating a predominantly monomeric COCA46 species



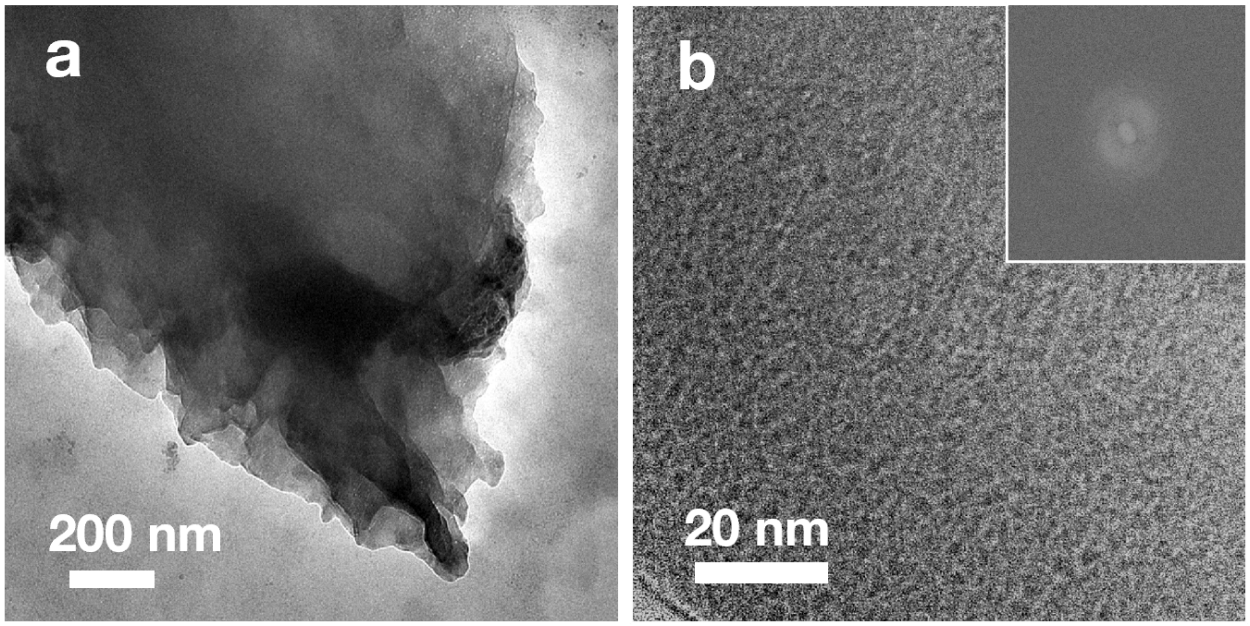
Supplementary Fig. 11. Visual evidence for Co(II) retention by COCA46 during concentration. Concentration of COCA46 in a 10-kDa molecular-weight cutoff filter in the presence of 250 μM $CoCl_2$ indicates Co(II) binding, as the retentate becomes increasingly pink while the flow-through remains clear.



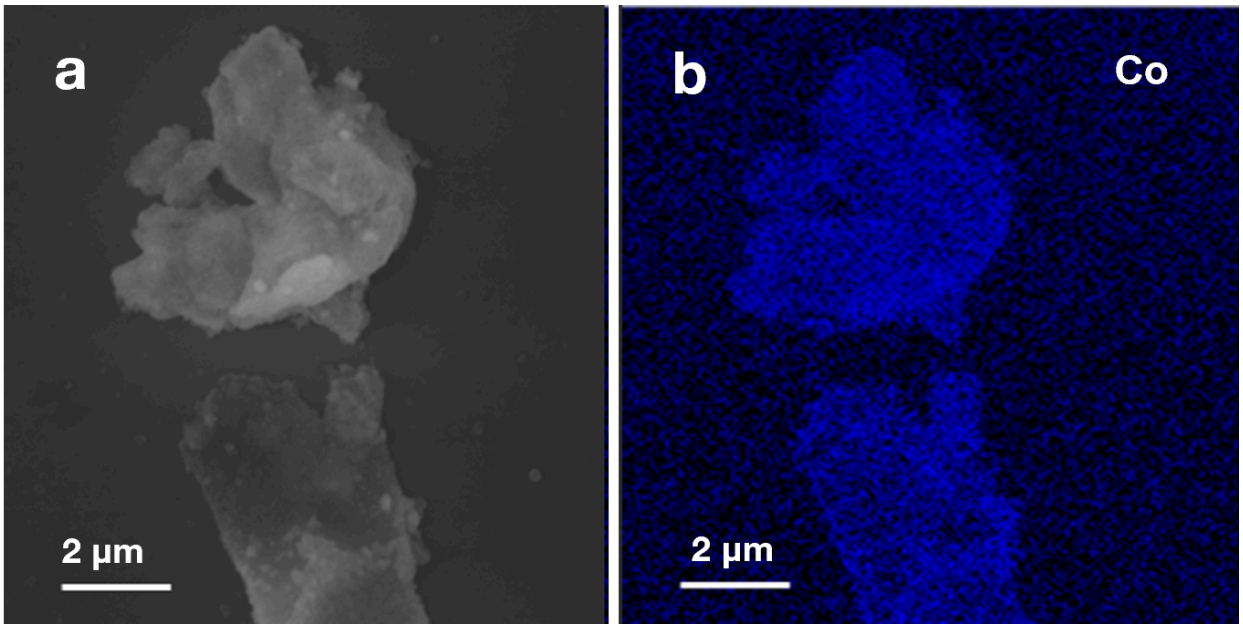
Supplementary Fig. 12. Additional HRTEM analysis of $CoCO_3$ formed with COCA46. **a,** Unstained HRTEM image of mineral formed with 0.5 μM COCA46 after 3 h of mineralization in 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C; the inset FFT pattern shows rings indexed to the $CoCO_3$ {104} and {202} reflections. **b,** Higher-magnification HRTEM image showing 2.74 Å lattice fringes assigned to symmetry-related $CoCO_3$ {104}-type planes, with interfringe angles of ~85°.



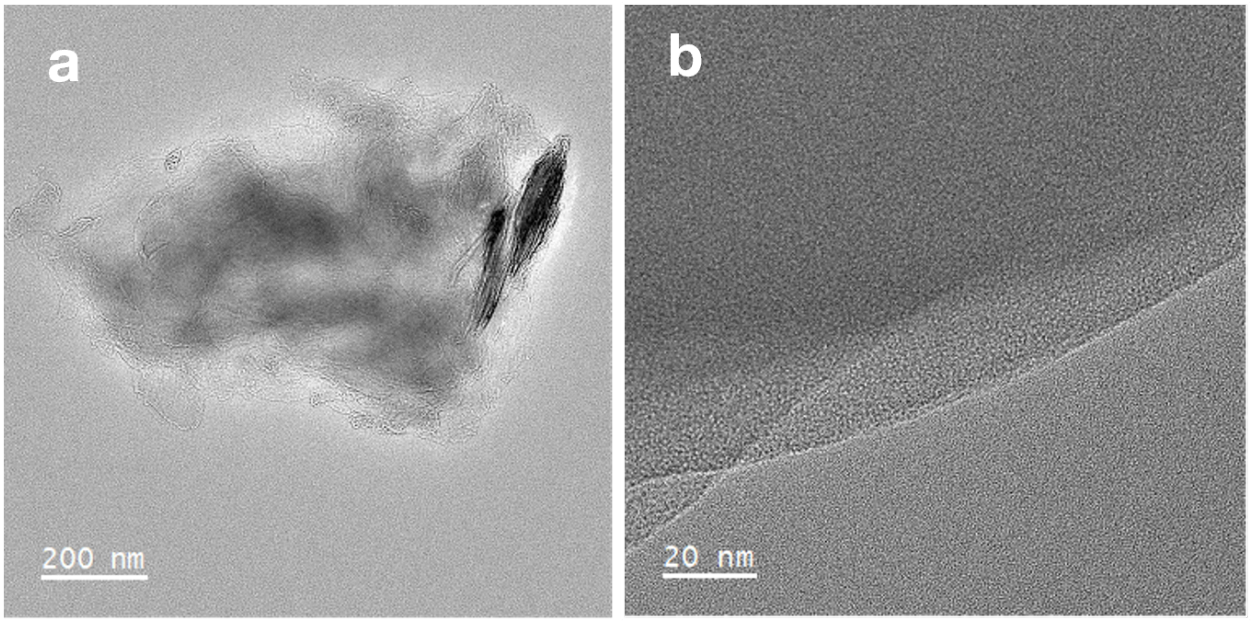
Supplementary Fig. 13. a, SEM-EDX analysis of cobalt-containing minerals formed with COCA46 after 3 d under mineralization conditions containing 0.5 µM COCA46, 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C. **b,** Corresponding EDX elemental map showing cobalt signal.



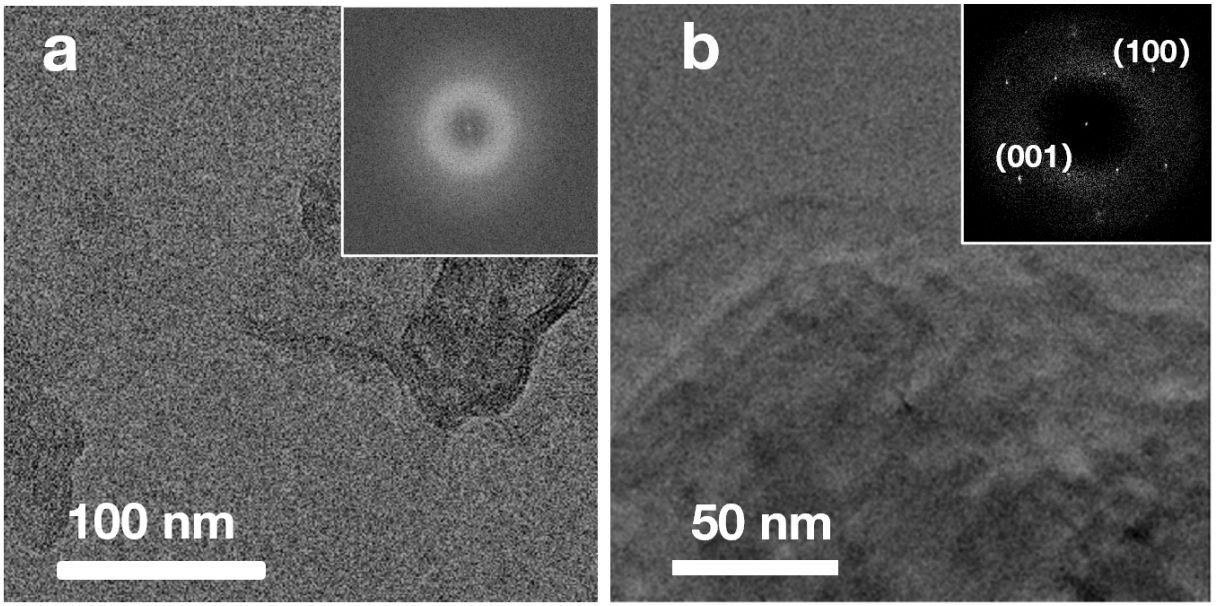
Supplementary Fig. 14. HRTEM images of a no-protein control after 3 days under the same mineralization conditions show amorphous precipitate. Low- (a) and higher- (b) magnification unstained TEM images



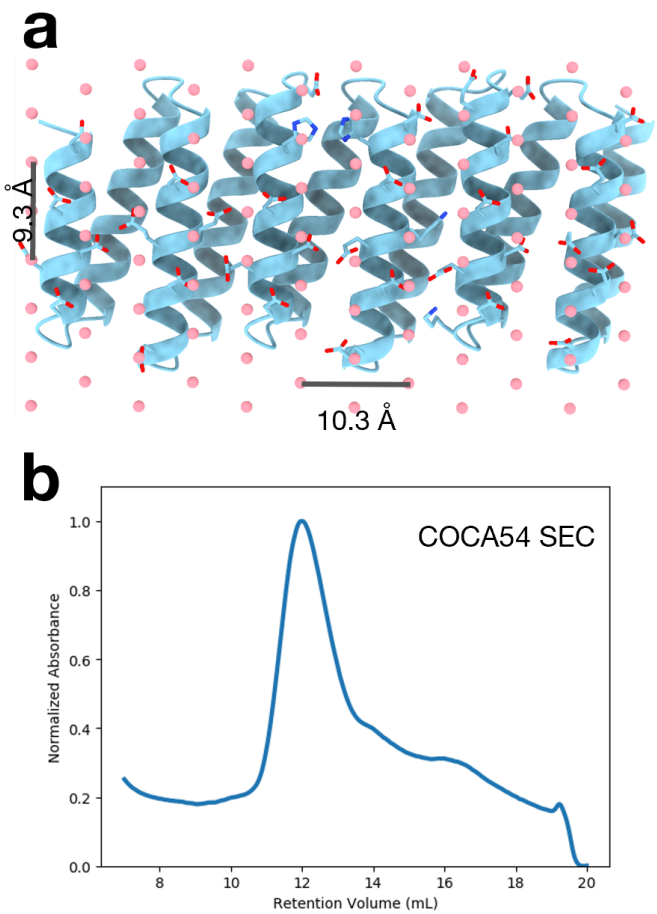
Supplementary Fig. 15. a, SEM analysis of cobalt-containing minerals formed with 0.5 µM FD31 after 3 days under mineralization conditions containing 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at 25°C and pH 8.2. **b,** corresponding EDX elemental map shows Cobalt signal.



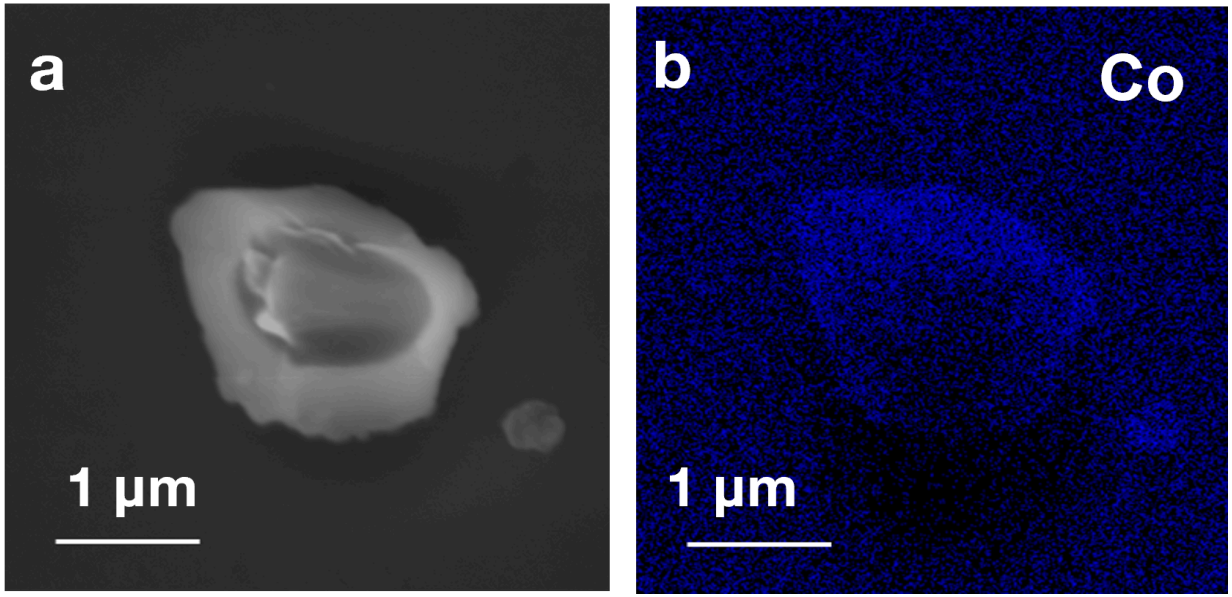
Supplementary Fig. 16. Unstained TEM and HRTEM images of the 0.5 μM FD31 condition after 3 days under mineralization conditions containing 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C, showing amorphous precipitates rather than crystalline $CoCO_3$.



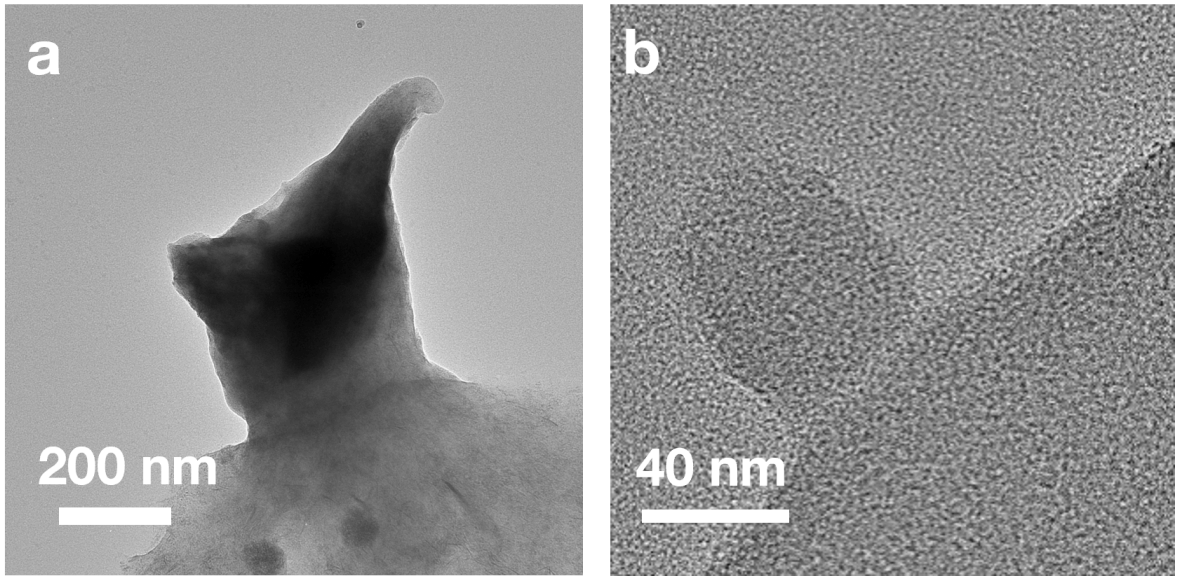
Supplementary Fig. 17. Negative-design control using the COCA46 H→N variant, in which the histidine residues predicted to interact with $CoCO_3$ were replaced with asparagine. After 3 days under the same mineralization conditions, this variant did not produce faceted $CoCO_3$ crystals. Instead, samples contained amorphous precipitate (a) or crystalline material with (001) and (100) reflections consistent with cobalt hydroxide (b). Insets show FFTs.



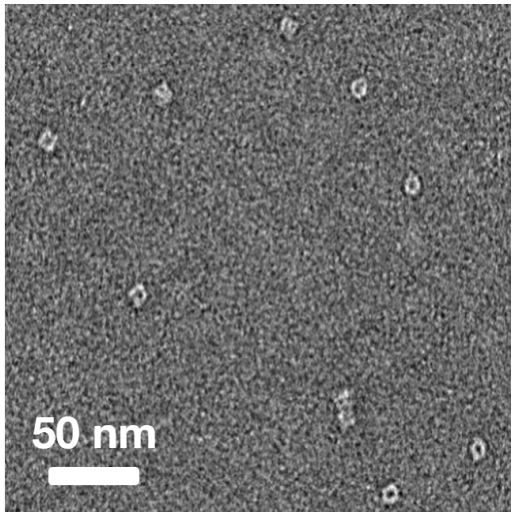
Supplementary Fig. 18. a, Design model of COCA54 targeting the $CoCO_3$ {101} facet along an alternative lattice-matching axis, in which cobalt ions are spaced by 10.3 Å rather than the 9.3 Å spacing targeted by COCA46 indicated in the figure. **b,** COCA54 size-exclusion chromatography trace monitored by UV absorbance at 280 nm, showing a dominant species.



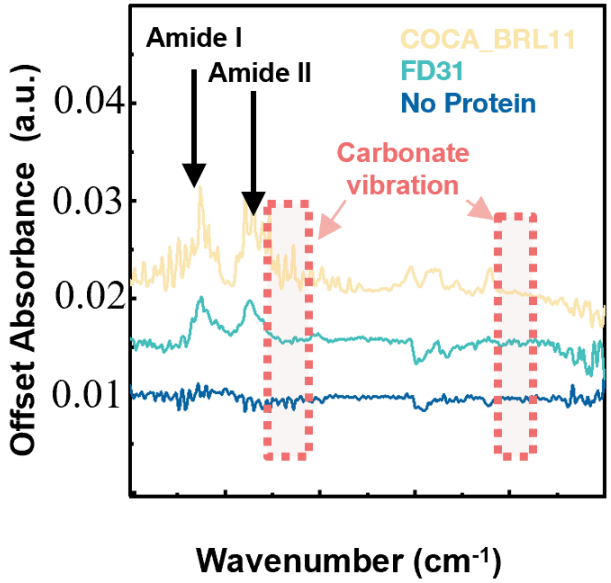
Supplementary Fig. 19. a, SEM–EDX analysis of cobalt-containing precipitates formed with COCA54 after 3 d under mineralization conditions containing 0.5 µM COCA54, 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C. **b,** Corresponding EDX elemental map showing cobalt signal.



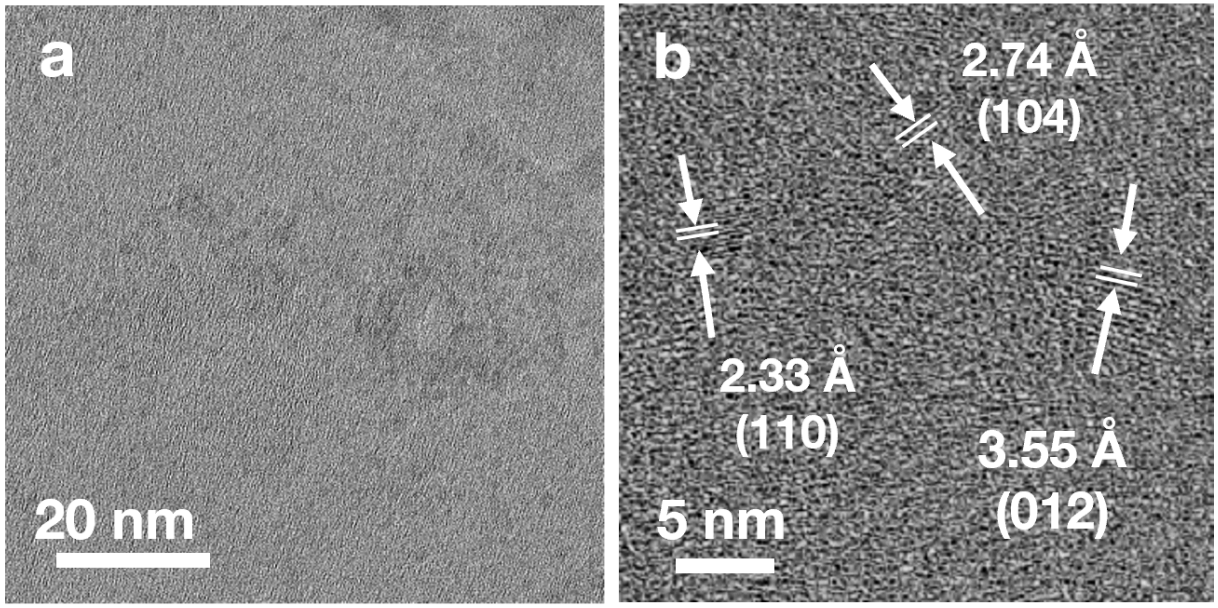
Supplementary Fig. 20. Unstained TEM and HRTEM images of 0.5 μM COCA54 after 3 d under mineralization conditions containing 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C show amorphous precipitates. Low- (a) and higher- (b) magnification images are shown.



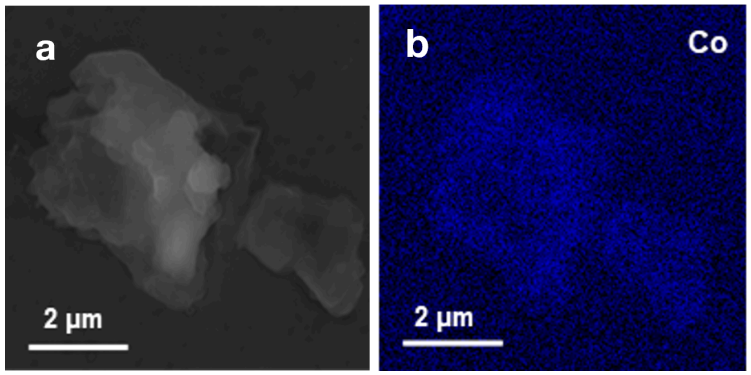
Supplementary Fig. 21. Negative-stain TEM of COCA_BRL11.



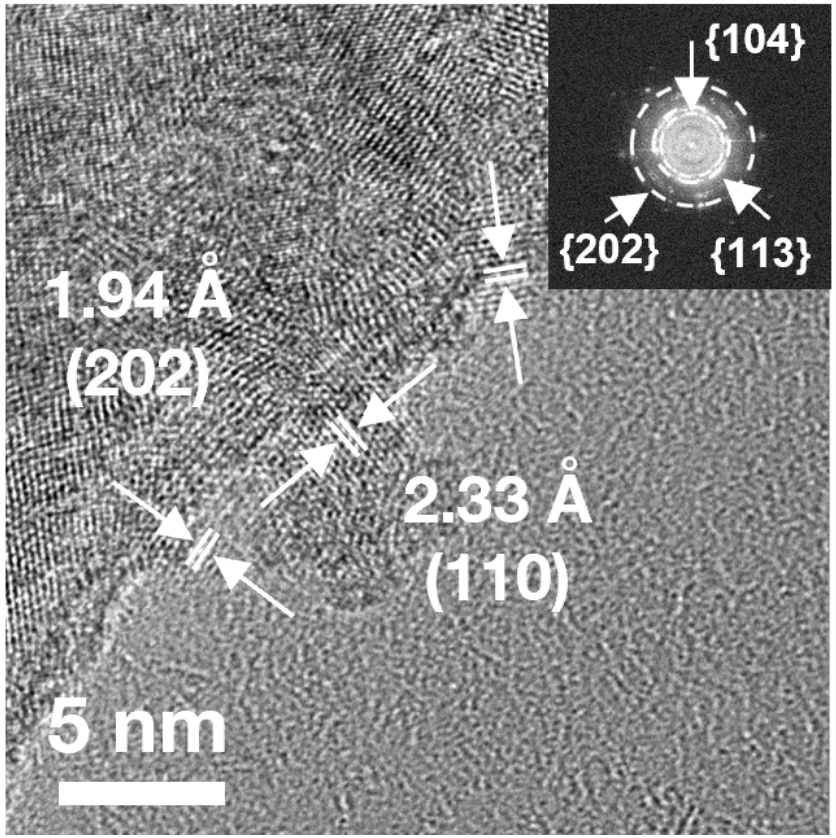
Supplementary Fig. 22. ATR-FTIR spectra of COCA_BRL11, FD31, and No protein control. Under $CoCO_3$ mineralization conditions containing 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C. Amide I and amide II bands indicate the presence of protein, and the highlighted regions indicate carbonate-associated vibrational bands.



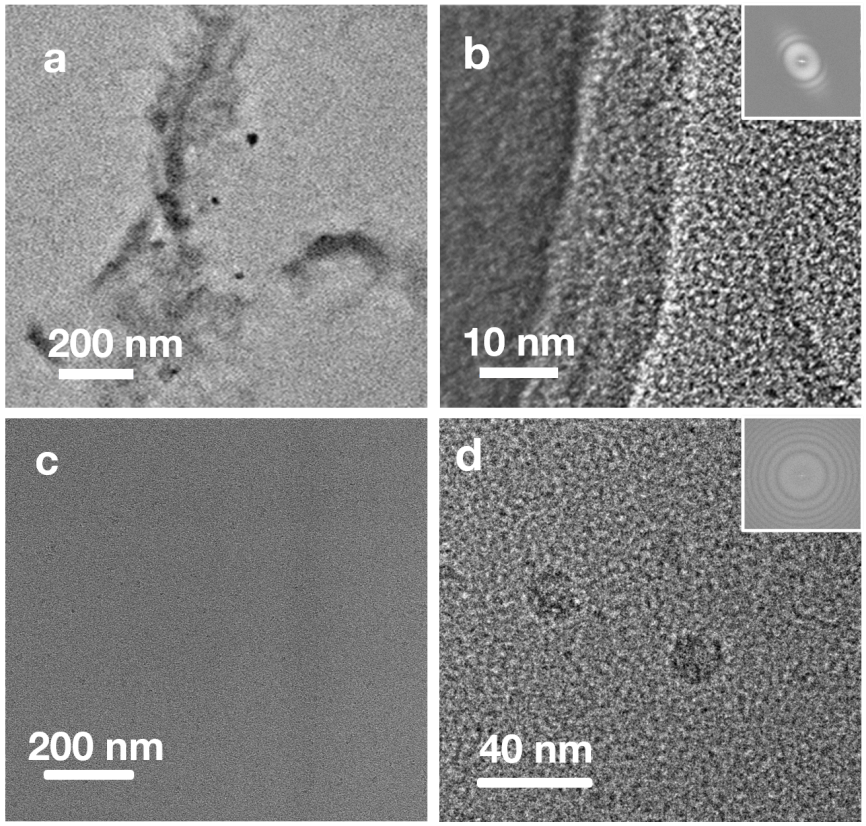
Supplementary Fig. 23. Unstained HRTEM image of 0.5 µM COCA_BRL11 after 3 hours of mineralization in 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25°C, showing crystalline $CoCO_3$ nanoparticles. Low- (**a**) and higher- (**b**) magnification images are shown, with lattice fringes in b assigned to the $CoCO_3$ (110), (104), and (012) planes.



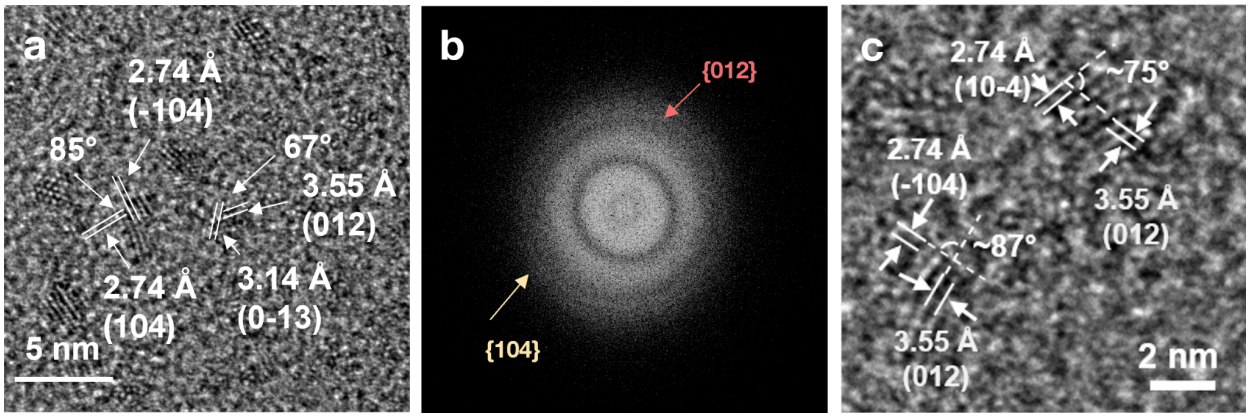
Supplementary Fig. 24. a, SEM–EDX analysis of cobalt-containing aggregates formed with 0.5 µM COCA_BRL11 after 3 d under mineralization conditions containing 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C. **b,** Corresponding EDX elemental map showing cobalt signal.



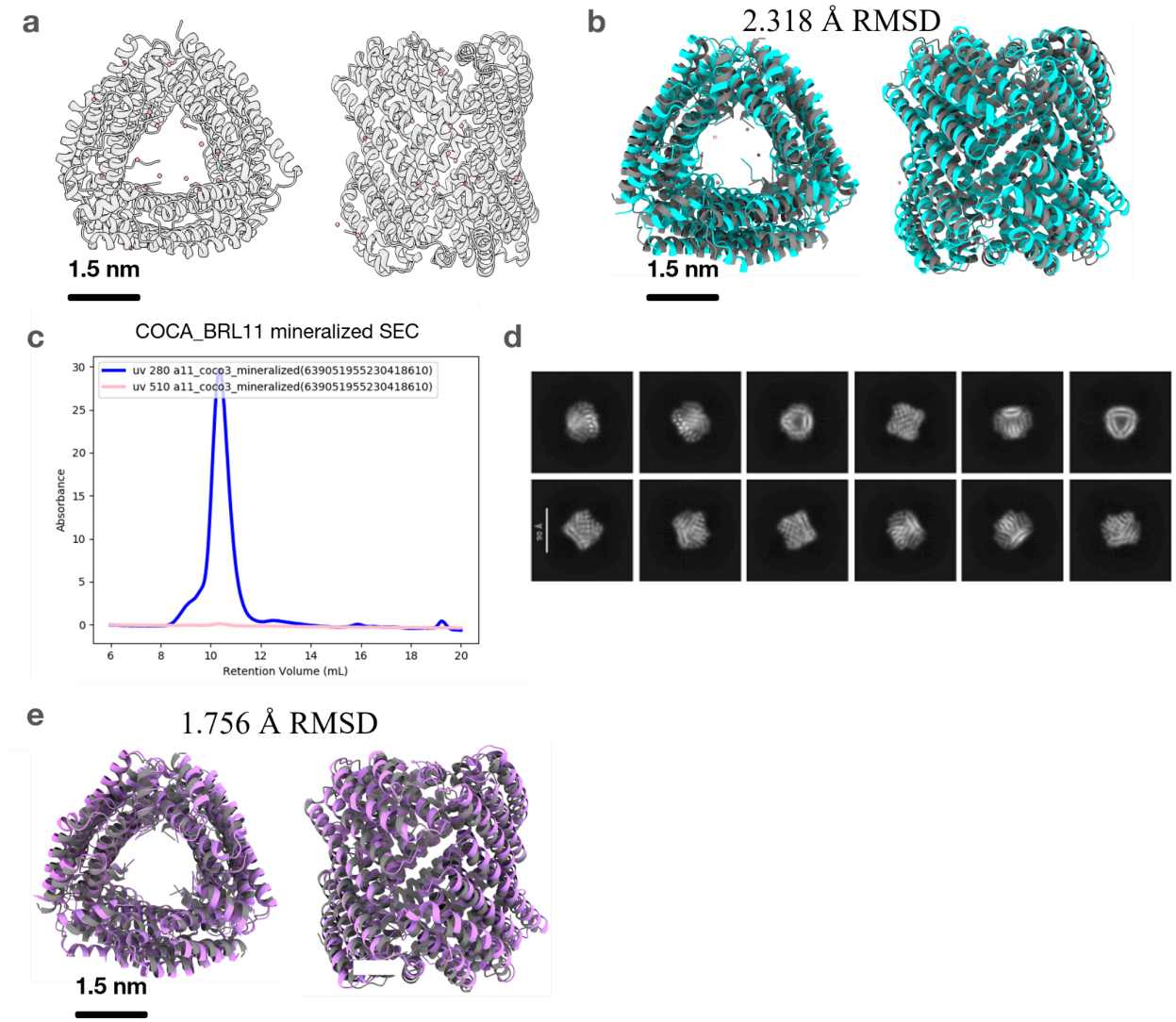
Supplementary Fig. 25. Unstained HRTEM image and corresponding FFT pattern of 0.5 µM COCA_BRL11 after 3 d under mineralization conditions containing 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C. Lattice fringes with spacings of 1.94 and 2.33 Å are assigned to the $CoCO_3$ (202) and (110) planes, respectively; the inset FFT pattern shows rings indexed to the {104}, {202}, and {113} reflections.



Supplementary Fig. 26: COCA_BRL11 $CoCO_3$ activity. Unstained TEM and HRTEM images of control reactions for COCA_BRL11 mineralization. **a,b,** Unstained TEM and HRTEM images of the COCA_BRL11 H→N mutant after mineralization in 1 mM $CoCl_2$ and 1 mM $NaHCO_3$ at pH 8.2 and 25 °C, showing amorphous precipitates; the FFT inset in b is consistent with amorphous material. **c,** Unstained TEM image of the no-protein control after 1 d in 250 µM $CoCl_2$ and 250 µM $NaHCO_3$ at pH 8.2 and 25 °C, showing amorphous precipitates. **d,** Unstained HRTEM image of 0.1 µM COCA46 after 1 d under the same low-concentration mineralization conditions, showing amorphous particles; the FFT inset is consistent with amorphous material.

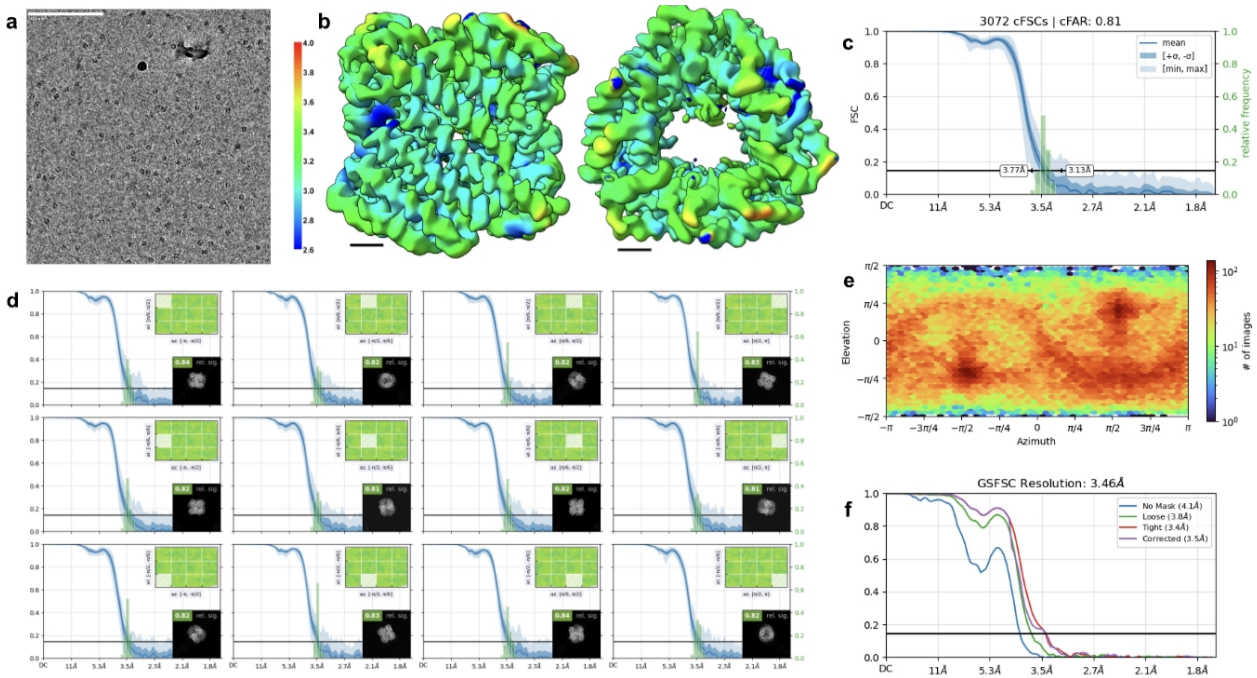


Supplementary Fig. 27. Additional HRTEM and FFT analysis of $CoCO_3$ nanocrystals formed with 0.3 μM COCA_BRL11 after 1 d under low-concentration mineralization conditions containing 250 μM $CoCl_2$ and 250 μM $NaHCO_3$ at pH 8.2 and 25 °C. **a,** HRTEM image showing lattice fringes with spacings of 2.74, 3.55 and 3.14 Å, with interfringe angles of ~85° between (104) and (-104) fringes, and ~67° between (012) and (0-13) fringes. **b,** Corresponding FFT pattern showing rings indexed to the $CoCO_3$ {104} and {012} reflections. **c,** HRTEM showing lattice fringes with spacings of 2.74 and 3.55 Å, consistent with $CoCO_3$ {104}- and {012}-type planes, respectively. The measured interfringe angles of ~75° between (10-4) and (012) and ~87° between (-104) and (012) further support the assignment to crystalline $CoCO_3$.

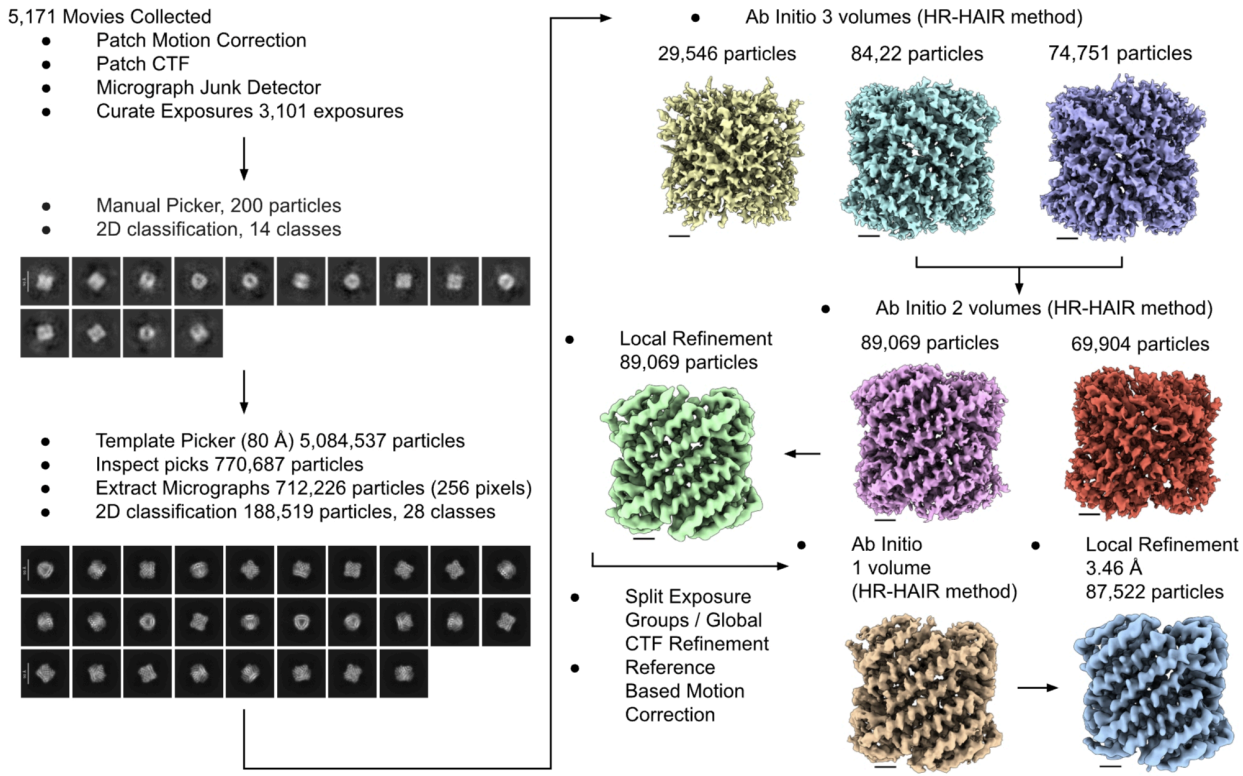


Supplementary Fig. 28. COCA_BRL11 SEC. a, Holo crystal structure of COCA_BRL11 after soaking with cobalt(II) chloride. **b,** Cobalt soaked holo crystal structure of COCA_BRL11 aligned to the design model in gray. **c,** Size exclusion chromatography traces

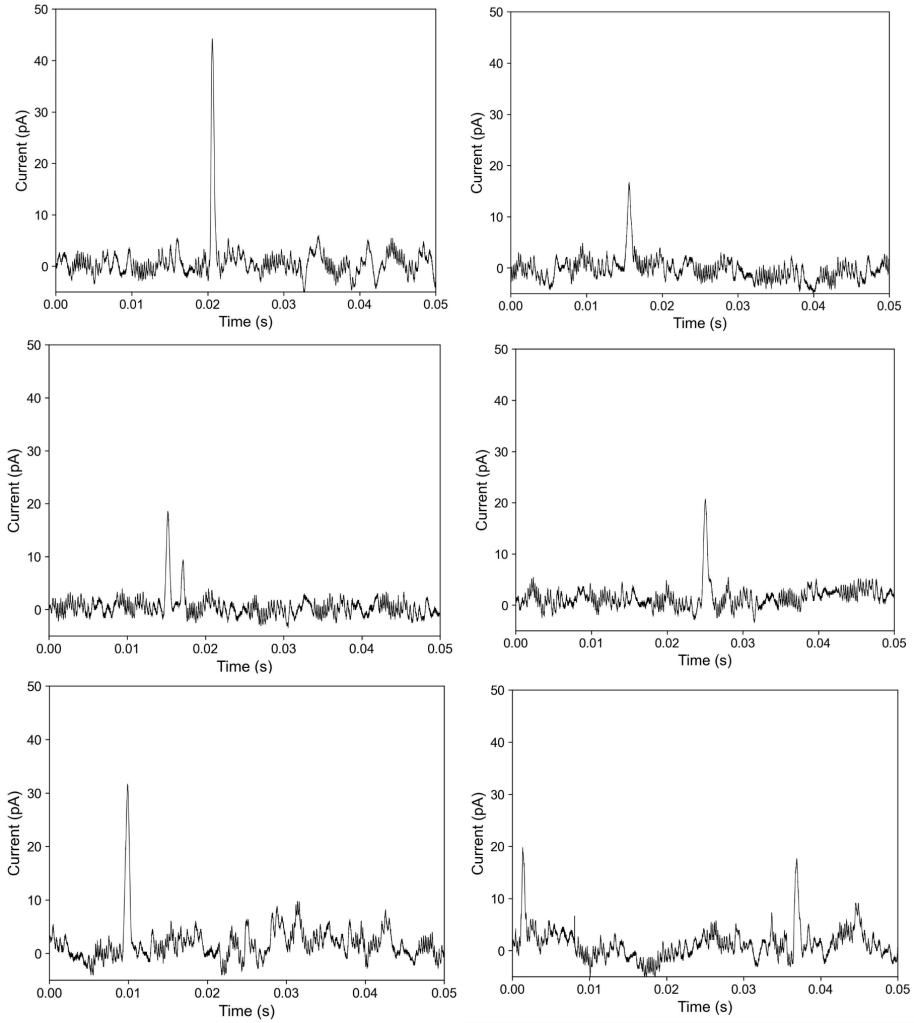
monitored at 280 nm and 510 nm show co-elution of the protein and Co(II) signals, respectively. **d,** Representative 2D class averages. **e,** CryoEM structure of COCA_BRL11 aligned to the design model in gray.



Supplementary Fig. 29. CryoEM analysis of $CoCO_3$-mineralized COCA_BRL11. (a) Representative Micrograph, scale bar 100 nm (b) Local Resolution Map, side view (left) and top view (right). 10 Å Scale bar (bottom left) (c) Conical Fourier Shell Correlations (FSC) Area Ratio (cFAR) evaluated with respect to 3072 viewing directions, resolution range 3.77-3.11 Å. (d) Orientational diagnostics data demonstrating full angular sampling (e) Orientational distribution plot demonstrating full angular sampling (f) Global FSC non uniform refinement

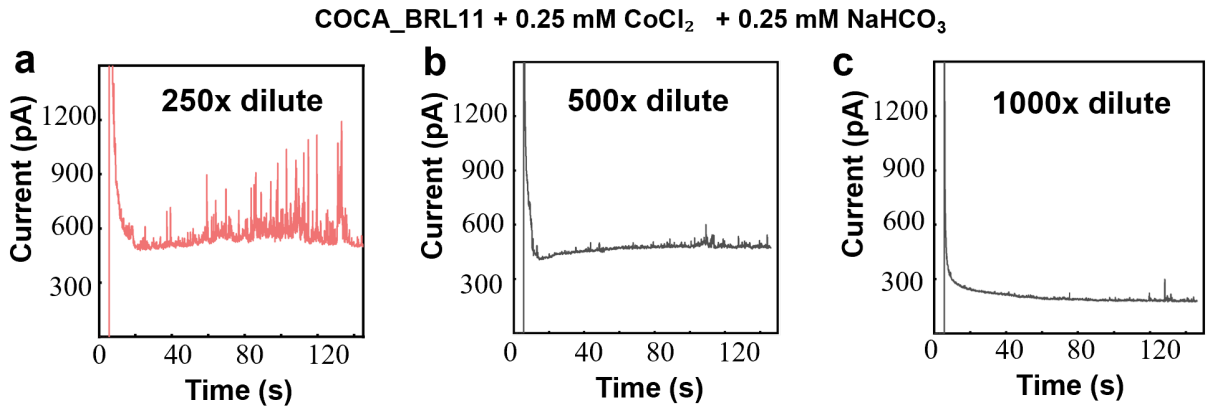


Supplementary Fig. 30. CryoEM data processing workflow, $CoCO_3$-mineralized COCA_BRL11: Schematic of movies collected and processing steps obtain the final map. With 2D and 3D classification shown, all maps refined without symmetry (C1). 10 Å Scale bar (bottom left of maps)

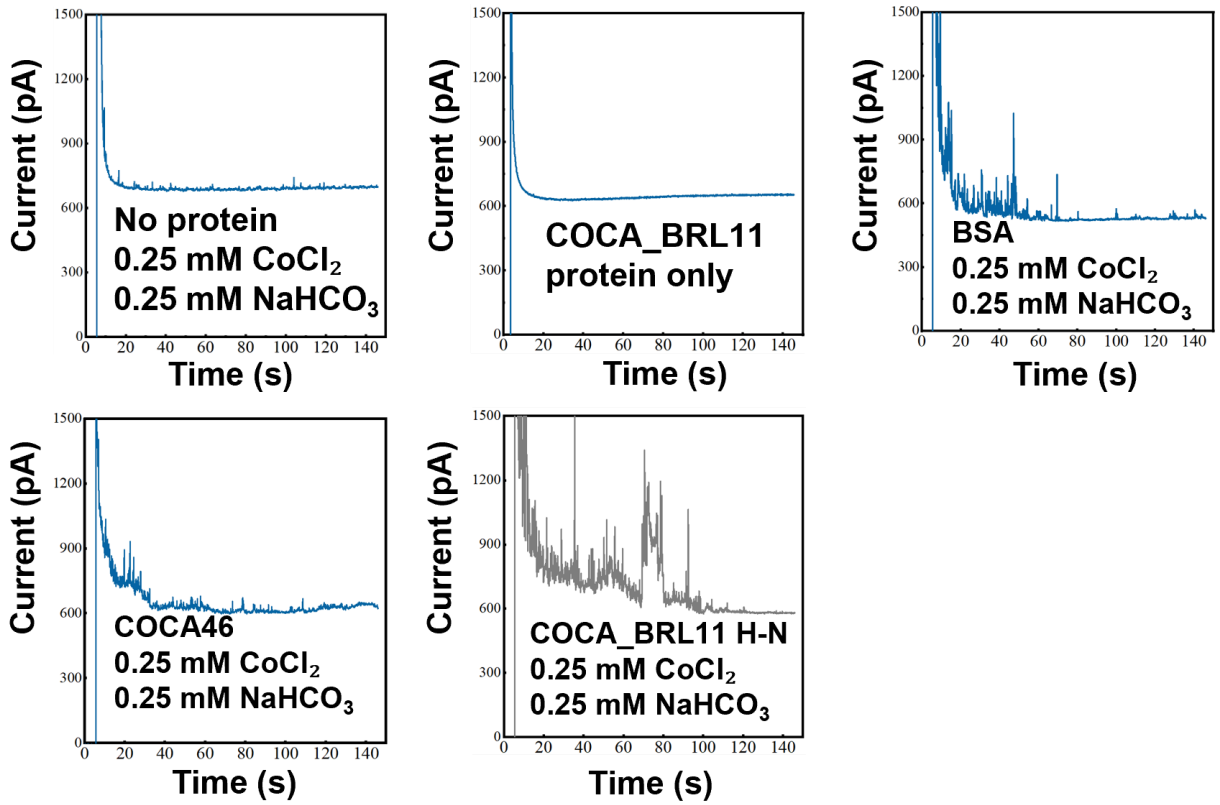


Supplementary Fig. 31. Representative isolated OER-associated current transients extracted from Fig. 5k. Representative baseline-corrected current spikes from COCA_BRL11-mineralized $CoCO_3$ particles recorded in 0.1 M KOH at 1.0 V versus Ag/AgCl using a 5 µm carbon fiber working electrode. Individual peaks were time-shifted to a

common zero-time origin for visualization. These isolated peaks are representative of the events used for the peak-amplitude, integrated-charge, FWHM, apparent TOF, and apparent TON analyses in Fig. 5l–n.



Supplementary Fig. 32. Representative chronoamperometric OER current–time traces of COCA_BRL11-mineralized CoCO3 samples recorded in 0.1 M KOH at 1.0 V versus Ag/AgCl using a 5 μm carbon fiber working electrode. COCA_BRL11 was mineralized with 1 mM $CoCl_2$ and 1 mM NaHCO3, and the resulting samples were measured after 250-fold (a), 500-fold (b) and 1000-fold (c) dilution.



Supplementary Fig. 33. Representative chronoamperometric OER current–time traces of control mineralization reactions recorded in 0.1 M KOH at 1.0 V versus Ag/AgCl using a 5 μm carbon fiber working electrode. No-protein $CoCO_3$, COCA_BRL11 protein-only, BSA-$CoCO_3$, COCA46-$CoCO_3$, and COCA_BRL11 H→N-$CoCO_3$ controls were measured under the same electrochemical conditions. The Co-containing controls were prepared with 0.25 mM $CoCl_2$ and 0.25 mM $NaHCO_3$. Compared with COCA_BRL11-mineralized $CoCO_3$, these controls showed background, transient, or less persistent current responses over the measurement window.

Supplementary Table 1: Designed protein sequences

Name	Protein sequence
FD31	MPEEALEEVEERIEELESALESNPTNEEELREILKKILEIFEELFREAKARNDTELLLEAVEAVIELLETLLELNPTNEELLREILKIILRIFELLFELAKKQNDTELLLEAVEAVIELLETLLELNPTNEELLREILKIILRIFELLFELAKKQNDTELLLEAVEAVIELLETLLELNPTNEELLREILKIILRIFELLFELAKKQNDTELLLEAVEAVIELLETLLELNPTNEELLREILKIILRIFELLFELAKKQNDTELLSEAKEAVAELLETLAELNPTNQELKEEIKKIQERIAELEKELAEKQNAGSGSHHWGSTLEHHHHHH
COCA_BRL11	SGEGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLEEAVELVVEALRHILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIRALETVSELLGHLIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVKAKEAGFADVADAFVDRVASAGSGSHHWGSTLEHHHHHH
COCA_BRL11 H→N	SGEGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLEEAVELVVEALRNILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIRALETVSELLGNLIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVKAKEAGFADVADAFVDRVASAGSGSHHWGSTLEHHHHHH
CALC_HEX24	SGDRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIFELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDQIKGIESICKAVKAFFEAIYEIRKKDFKDLSSIVAKTLGTALAEYVAAIEYGALKEPGIRRLFLRSLEAFLEIILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETEGSGSHHWGSTLEHHHHHH
COCA46	SGDEVVDLADEAAERGLERVEEGERAIELLARLAEEHADDEDTLAHLVDLLGDVVEAAGRAGRLEAVLKGLKAAAKALAAVGSDGVAGAAEHVADAAADAGLENPEAGLVALKGLVELTKGGAGNKDATEHFVDAAGDVLKALLVERIRPEEAVELAAQALAIAAAAGVGGDGADHLVDSAADAGLETPEAGLLAARLGLAAARAFAADAGTLAHIVDELEHVLEANGSPEVRRLLAEARALARGSGSHHWGSTLEHHHHHH
COCA46 H→N	SGDEVVDLADEAAERGLERVEEGERAIELLARLAEEHADDEDTLANLVDLLGDVVEAAGRAGRLEAVLKGLKAAAKALAAVGSDGVAGAAENVADAAADAGLENPEAGLVALKGLVELTKGGAGNKDATENFVDAAGDVLKALLVERIRPEEAVELAAQALAIAAAAGVGGDGADNLVDSAADAGLETPEAGLLAARLGLAAARAFAADAGTLANIVDELENVLEANGSPEVRRLLAEARALARGSGSHHWGSTLEHHHHHH
COCA54	SGKIDALLDKADDAADAGKIDEAVALYKEAIDAAAEAGRIGLLVDAAEEGVDAATEHGAAEKALELGKKATDLLAERAPHSEEAIHGLEDLGDELGDAGLAEGAAYVLKGAIALALEAGLVDHAADAAKEGVDLMAEHGREELAIELAKFAGQGLMEKAPYSEEAVDGLIDLADELADKGLIEGAVYVYKLAIQLAARAGLLDALREAAEEGKEAAGEHGREELAEELERLARELAARGSGSHHWGSTLEHHHHHH
ARAG64	SGEKEKKIKELIEKAKEDLKKITDAEEAREKLEEALERLLEESLEALSPQDATDLAKEALREGAKSGDEEAITETVEEAAEAFSEAAKARKATGTEAIDAATDLLEAAVSEVSDAKAATETVEEVGEAAEETDLSTASAEAITRLAEAGAEAAKKGDDEEARTETVEETAEEVAEIAERAKITGSEAVKIAARLAEEAAKAIKDGKKRKETAEEVLELAEDTDTSTVSEEALRELAEALAELAKKAKSEEDRREAFEEAREEIEEIAERAKLSESDLERILRELEESEKKAGSGSHHWGSTLEHHHHHH
COCA_BRL1	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVREALERILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIDALETVSELLGRLIERIQDLGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVRAKRAGFADVADAFVDRVASA
COCA_BRL2	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVVAALLKILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIHALETVSRLLNALIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKRAGFADVADAFVDRVASA
COCA_BRL3	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLERAVELVVEALEKILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIRALETVSRLLNRLIERIQDLGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKRAGFADVADAFVDRVASA
COCA_BRL4	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLERAVELVVAALERILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIRALETVSALLNRLIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKKAGFADVADAFVDRVASA
COCA_BRL5	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVVAALEKILRLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIYALETVSRLLNALIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKRAGFADVADAFVDRVASA
COCA_BRL6	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVVAALRKILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIHALETVSELLNHLIERIQDLGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKKAGFADVADAFVDRVASA
COCA_BRL7	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVTAALERILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIYALETVSRLLNRLIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKRAGFADVADAFVDRVASA
COCA_BRL8	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVRAALLHILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIYALETVSELLNHLIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKRAGFADVADAFVDRVASA

COCA_BRL9	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVVAALRRILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIHALETVSAL LNHLIERIQALGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKKAGFADVADAFVDRVASA
COCA_BRL10	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLRRAVELVVAALRHILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIHALETVSEL LNHLIERIQDLGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKRAGFADVADAFVDRVASA
COCA_BRL12	SGEIKELARLLKYFKIPEEDLAKLTADALAALRALTEAFRALAAAARASGDLELLERAVELVVEALEKILDLLPGDRAILEAAAELVRTLFEELAAIALERRDAELAIRALETVSELL NALIERIQDLGDEEAKEVLREVLERVSALMADLWAVFKDDVWAKLRALKALAGFVVAAKRAGFADVADAFVDRVASA
CALC_HEX1	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKDEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIKGIEAICKAVKAFFEAIYEIRKKDFKDLASIVAKTLGTALAEYVAAIESGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX2	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEFVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLAKELSSFTEDDIKGIEAICKAVKSFFEAIYEIRKKDFKDLSSIVSKTLGTALAEYVKAIESGALKNPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX3	DRETLKELIETIRELLEKLLERNPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLAKELSSFTEDQIKGIEAICKGVKAFFEAIYEIRKKDFKDLASIVAKTLGTALAEYVAAIEYGALKEPGILRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX4	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDQIEGIKAICEGVKSFFEAIYEIRKKDFKNLSSIVSKTLGTALSKYVEAIECGALKEPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX5	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKDEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIEGIKSICKAVKSFFEAIYEIRKKDFKNLASIVAKTLGTALAEYVAAIESGALKNPGILRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX6	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIKGIKAICKAVKAFFEAIYEIRKKDFKDLASIVSKTLGTALAEYVAAIESGALKNPGILRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX7	DRETLKELIETIRELLEKLLERFPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIKGIKAICKAVKAFFEAIYEIRKKDFKDLASIVAKTLGTALAEYVAAIEYGALKEPGILRLFLRSLESFLEII LEILESEKSISEEEKKLIIKLILLTLKTLLKLVEFIEEKE
CALC_HEX8	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDQIKGIKAICEAVKAFFEAIYEIRKKDFKDLSSIVAKTLGIALAEYVAAIEAGALKNPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX9	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPNFTEDQIKGIKAICEGVKAFFEAIYEIRKKDFKDLTSIVAKTLGKALAEYVAAIEAGALKEPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX10	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFDFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIKGIKAICEAVKAFFEAIYEIRKKDFKNLTSIVAKTLGKALAEYVAAIEAGALKKPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX11	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFDFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIKGIKAICEGVKSFFEAIYEIRKKDFKDLTSIVAKTLGNALACYVAAIEAGALKNPGIRRLFLRSLESFLEI ILEILESEKSISELEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX12	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDQIKGIKAICEAVKSFFEAIYEIRKKDFKDLSSIVAKTLGHALAEYVAAIESGALKEPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX13	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTVEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSVELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDDIKGIEAICKGVKSFFEAIYEIRKKDFKELTSIVAKTLGNALACYVAAIEYGALKRPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKIILETLKTLLKLVEFIEETE
CALC_HEX14	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDQIKGIKAICEAVKSFFEAIYEIRKKDFKDLASIVAKTLGKALAEYVAAIESGALKNPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX15	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKDEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIKGIEAICEAVKSFFEAIYEIRKKDFKNLTSIVSKTLGKALSQYVAAIESGALKNPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE

CALC_HEX16	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIEGIKAICKAVKAFFEAIYEIRKKDFKNLTSIVAKTLGTALAEYVAAIESGALKNPGIRRLFLRSLEAFLEII LEILESEKSISELEKKLIIKLILDTLKTLLKLVEFIEETE
CALC_HEX17	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTVEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSVELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIEGIKSICEAVKAFFEAIYEIRKKDFKELTSIVAKTLGKALAEYVAAIESGALKNPGIRRLFLRSLEAFLEII LEILESEKSISELEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX18	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTVEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSVELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIEGIKEICKAVKKFFEAIYEIRKKDFKNLTSIVSETLGEALSKYVAAIEKGALKNPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEEQE
CALC_HEX19	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPNFTKDQQEGIKAICEAVKSFFEAIYEIRKKDFKNLSSIVSKTLGIALSQYVAAIESGALKNPGIRRLFLRSLESFLE IILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX20	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPNFTEDQIKGIEAICKGVKAFFEAIYEIRKKDFKDLSSIVAKTLGIALAEYVAAIEYGALKEPGIRRLFLRSLEAFLEII LEILESEKSISELEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX21	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPNFTEDQQKGIEAICKAVKAFFEAIYEIRKKDFKELSSIVSKTLGEALAQYVAAIEYGALKNPGIRRLFLRSLEAFL EIILEILESEKSISELEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX22	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDQQKGIESICKAVKAFFEAIYEIRKKDFKELSSIVSKTLGIALAEYVAAIEYGALKEPGIERLFLRSLEAFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX23	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFDFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDDIKGIESICKAVKAFFEAIYEIRKKDFKELSSIVAKTLGIALAEYVAAIEYGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX25	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFDFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSKFTEDQIKGIESICKAVKSFFEAIYEIRKKDFKELSSIVAKTLGIALAEYVAAIEYGALKEPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX26	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPNFTEDQIKGIESICKGVKAFFEAIYEIRKKDFKELSSIVSKTLGIALSEYVAAIEYGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX27	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTTEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSTELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIKGIEAICKGVKAFFEAIYEIRKKDFKDLTSIVAKTLGNALAQYVAAIEYGALKNPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIDLILETLKTLLKLVEFIEETE
CALC_HEX28	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSLELALEFDFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIKGIEAICKAVKAFFEAIYEIRKKDFKDLSSIVSKTLGIALAEYVAAIESGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX29	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFDFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSKFTEDDIKGIEAICKGVKAFFEAIYEIRKKDFKDLSSIVAKTLGTALACYVAAIEAGALKNPGILRLFLRSLEAFLEII LEILESEKSISELEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX30	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKDEELLKEILETIKEILEKSYELALEFQFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDEQKGIAAICKGVKAFFEAIYEIRKKDFKDLASIVAKTLGEALAQYVAAIEAGALKNPGIRRLFLRSLEAFL EIILEILESEKSISELDKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX31	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTTEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSTELALEFQFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTKDEQEGIAAICKGVKAFFEAIYEIRKKDFKDLASIVAKTLGEALAQYVAAIESGALKNPGILRLFLRSLEAFL EIILEILESEKSISELEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX32	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKDEELLKEILETIKEILEKSYELALEFQFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDDQKGIAAICKGVKAFFEAIYEIRKKDFKDLSSIVAKTLGTALAEYVAAIESGALKNPGILRLFLRSLEAFLE IILEILESEKSISELEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX33	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTTEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSTELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDDIKGIESICKGVKSFFEAIYEIRKKDFKDLASIVAKTLGIALAEYVAAIESGALKRPGIRRLFLRSLESFLEII LEILESEKSISEEEKKLIIKLILETLKTLLKLVEFIEETE

CALC_HEX34	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIKGIEAICEGVKAFFEAIYEIRKKDFKDLASIVAETLGKANAKYVAAIESGALKREGIRRLFLRSLESFLE IILEILESEKSISEKEKKLIIKLILENLKTLLKLVEFIEETE
CALC_HEX35	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPNFTEDEIKGIEAICKGVKAFFEAIYEIRKKDFKELTSIVAKTLGTALAEYVEAIEYGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLELVEFIEETE
CALC_HEX36	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDEQKGIEAICKGVKSFFEAIYEIRKKDFKDLTSIVAETLGKALAEYVAAIEYGALKNPGIRRLFLRSLESFLE IILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX37	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFDFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSKFTEDEIKGIEAICKAVKKFFEAIYEIRKKDFKDLTSIVAKTLGEALAKYVEAIESGALKNPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLELVEFIEETE
CALC_HEX38	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDDIKGIEAICKAVKSFFEAIYEIRKKDFKDLTSIVAETLGKALAEYVKAIEYGALKRPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLELVEFIEETE
CALC_HEX39	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFDFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDDIKGIEAICKGVKAFFEAIYEIRKKDFKDLTSIVAKTLGTALAEYVEAIEYGALKRPGILRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLELVEFIEETE
CALC_HEX40	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFDFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSKFTEDEIKGIEAICKGVKAFFEAIYEIRKKDFKDLTSIVAKTLGTALAKYVEAIEYGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKDKKLIIKLILLTLKTLLELVEFIEETE
CALC_HEX41	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTTEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSTELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDEIKGIEAICKAVKAFFEAIYEIRKKDFKDLTSIVAKTLGTALAEYVAAIESGALKNPGIRRLFLRSLESFLEII LEILESEKSISELEKKLIIKIILLTLKTLLKLVEFIEETE
CALC_HEX42	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPKFTEDEIKGIKAICKGVKAFFEAIYEIRKKDFKDLASIVAKTLGEALAEYVAAIEAGALKNPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX43	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKDEELLKEILETIKEILEKSLELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDEIKGIESICKAVKAFFEAIYEIRKKDFKDLSSIVAKTLGNALAEFVAAIESGALKNPGILRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX44	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFDFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIKGIEEICKAVKSFFEAIYEIRKKDFKDLVSIVAKTLGEALAEYVAAIEYGALKNPGIRRLFLRSLESFLEII LEILESEKSISEKDKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX45	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDQIKGIEAICKGVKAFFEAIYEIRKKDFKDLTSIVAKTLGKALAEYVAAIEYGALKNPGIRRLFLRSLESFLEI ILEILESEKSISELEKKLIIKLILETLKTLLKLVEFIEETE
CALC_HEX46	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTIEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSIELALEFNFSVDEFCELASIILELIFE LAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDEIKGIEAICKGVKAFFEAIYEIRKKDFKDLTSIVAKTLGEALAQYVAAIESGALKNPGIRRLFLRSLESFLEIIL EILESEKSISEREKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX47	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDDIKGIEAICKAVKAFFEAIYEIRKKDFKDLTSIVAKTLGTALAEYVAAIEYGALKRPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX48	DRETLKELIETIRELLEKLLERHPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSLELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIEGIKEICKGVKSFFEAIYEIRKKDFKNLTSIVAKTLGEALACYVAAIESGALKNPGIRRLFLRSLESFLEI ILEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX49	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSLELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPNFTEDEIEGIKEICKGVKAFFEAIYEIRKKDFKNLTSIVAKTLGTALACYVAAIESGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX50	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTLEYVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSLELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELPSFTEDQIEGIKEICKAVKAFFEAIYEIRKKDFKDLTSIVAKTLGEALACYVAAIESGALKNPGIRRLFLRSLEAFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE

CALC_HEX51	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTTEFVKKTKDRELFKETLETIRELLEKLLERLPENEELLKEILETIKEILEKSTELALEFNFSVDEFCELASIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIEGIKEICKAVKSFFEAIYEIRKKDFKDLTSIVAKTLGEALACYVAAIESGALKNPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLKLVEFIEETE
CALC_HEX52	DRETLKELIETIRELLEKLLERYPKDEELLKEILETIKEILEKTYEYVKKTKDRELFKETLETIRELLEKLLERLPKNEELLKEILETIKEILEKSYELALEFNFSVDEFCELSSIILELIF ELAKIVLKKEKSEKGKEFIKEITELGKEFCDKLEKLSKELSSFTEDEIKGIEAICKGVKAFFEAIYEIRKKDFKDLTSIVAKTLGTALAQYVAAIEAGALKNPGIRRLFLRSLESFLEII LEILESEKSISEKEKKLIIKLILLTLKTLLELVEFIEETE
CALC_HEX53	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTKQGVEDVIEISEVASKCCSKAVKACADMNWPVEVLATIILCSNLISTLVLIYYSLVFGSKEELLATLKATLEFLEGLIESWYIISKTNYHSGIGKLFISAVKNFLL ILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX54	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNNKEVWLEYAEGVNKIFGKIISAALEMNWPVVATGIALINEIANLLIVLGNLIINNASNEELYAFLKFILENIEKKIENFFKLLLDTKEYNSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX55	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENAVEFLFLFEEAANKVARLAIKAAYERNYPLAAKGIRYCNQIVDTLIEVGITYSLKPELLFLACKEALEKIESLTELFFDYLIAQKGKDVSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX56	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENAVEFLFLFEEAANKVASLAKKAAYELNYLEAAKGIHYCNLIVDTLIEVGKVSALKPELLKESVLEALEKIEKLTELFFDYLIKLKGKDLSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX57	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTEEGKETLIEIKETADKIINGAIKMLVKRKWPVIYLLIFVFIKFIIDLLCIAGIAYINGGSYEEFKAILESFEKVVEAGFEALLLIPEGKESGIGKLFISAVKNFLLILLS IIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX58	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTEEGKKTLIKIKETADKIINKAIEMLVELGKHTIFMLIFVFLKHIIDLLCIAGIAYIDGGSYEEFKAILSFEKTVEAGFEALLLTPEGKLSGIGKLFISAVKNFLLILLS IIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX59	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTEEGKKTLVKIKETADKIINKAIEMLVELNWHPIFMLIFVFIKFIIDLLCILGISVISGGSYEEIKAILSFEKVVEAGFEALLLNPKGKLSGIGKLFISAVKNFLLILLS IIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX60	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTKEGIETLVEISETANKIINGAKKMLVKLDKPPIYILIMTFISFIIDLLCNFGIAVMKGGSYEELKAILLSFEKIVEAAFEAILLTPEGKESGIGKLFISAVKNFLLILLS IIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX61	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTIEGQLEIQELEGVANKNLSKLKKMLIEKKASPSLLNVMNAISSLISLLSEAGYILYKKGNFNQFKGIIISFEATVNAIFESLLLDPNALSSGIGKLFISAVKNFLLI LLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX62	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTKEGILQLIELNGVFNKNLSKAMKMLAERNESPSLYNALLAISNLVGNLVLAGIILYKNGNFNEWLGVIIQLENTILSAFESFLLDPNALSSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX63	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNLEIVKLAVKGIVKIAAKILNLGIEYAEKINNPALFFGISLAVKLVLELAVMAGLISLGKEKEAFLSALKNIELLIEIFADYCLENPEEINSGIGKLFISAVKNFLLILL SIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX64	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDKEIMKLFVKGTIKIVAKIYNLGLDYAEEKDNAALAFGISLASQLVLTLCNFAYLYSFGKLKEAVESALKNIELLIEIYAEYLLKNPEEINSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX65	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDKEIMKLMVKGINKIVAKILNLGLDYALKKNNAALAFALSLAATLVLTLLNFAFLYSFGKLKEAVNSALKMIETLIEIYADYLLENPEEINSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX66	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDLELTLLFSEGIIKIIAKMLSGGLKLASLLNDPSLFVGMFSCSLVVLLLIYFGVMYKMGKLYESAKMSLELLESIIESFFDYMLSIPGGANSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX67	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDLELMILFAEGILKIVAKMLSGALKYAAKIGNPEVFVGMYSCSLILLELIFFAMMYKMGKLKESAEMSLKLLESIIESFFDYMLSIPGGANSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK

CALC_HEX68	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDLELMILFAKGILKVIAKMLAGALEYAAKIGDPQLFVGMFSCSLILITLIYFAMLYKMGELKNAAKMSLKLLESIIESFFDYMLSIPGGMNSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX69	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDLELMLLFAEGILKILAKIYAGALKYAAKIGEPEVFVGIKSCILILTELINFAVLYKMGELKPAAEMSLKLLESIIESFFDYMLSIPGGANSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX70	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENLDEFLQLVLEAASKCKAYAIKACLAAGTEEYELAAICIEICHNILELGAFAGKALLKSKPELALILLEYLEVLAEKFFELLIKEKFKDVSGIGKLFISAVKNFLLI LLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX71	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNPELTLEACSKLKKILAEMESKAFKLFISRNEPDSLIGIMGTFLVGILCINFGVLYKVTKNKKELFLMSINLLEKIAEETFEYLKKLKVGGSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX72	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDPELTLEACSKLEEIVAKMKSGAFELFVKRNKPEELYGIFGTYLVVLLTINFGVLWKVTKNKKELFLMSINLLEKIASETFEILKKLKVGSSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX73	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNADSSIFNKISKINLKVCLGASEAIDELNPEHSSAATICMTGLACLGILCNAGSEYIKNKYEMMILLFELGLKVALLTLEALLKIGGIKNSGIGKLFISAVKNFLL ILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX74	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNASSEIFDKISLIAVKTCLKAAEAIDKLNPEHADAGTICMTGLANLSLILTNAAASYIKNRYKTFIILCNLGMENAILTLEALLSIGNLKHSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX75	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENIEEFSKLFLECILKILNLGIKACDEMNLPIVVKLTLLICCLMAIFLLLFSKELIKTNNPEYGLLAVEFLSILTELMFKAILKFGLGKDISGIGKLFISAVKNFLLILL SIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX76	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNNELYPLLSKGCKKCQSIADEASFEFLPISLELALAADTAARGLDLASDLLFAGAIFSKDKELFIEYIKYLLEKIEIDAKYVFKVCAKYKSSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX77	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDKEILSLISKGIVKCIEIASKGFAESKDNGVQLALKIAADCLLLAHDLVLEYMKGLSENNLYEALKRSLKKIENTTELVFDEILELKNKGVSGIGKLFISAVKNFL LILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX78	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTEEGIEFINELNEVALKICQKASKLFISQNNLTGAFCMSAISSLVSSLSLLGISYILNKNIYTSAKSALEKIENEIESFFNYMLDTNSCGNSGIGKLFISAVKNFLL ILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX79	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDPEMLKLMAEGVLKISSIIIKGFIELSLKQKDIAFTVAAVITNELLNIAINLIILYASKEIKLAIKLSIEKIEKSIELFEKLLKEAKCMGSSGIGKLFISAVKNFLLILLSII ESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX80	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNPEMLKLMAKGVLKISSIINEGFIELSLKQNKKEFTLAAVIVNKLLNILNNLIILFANNEIKAAFNLSIKEIENLIELFEELMKEMNVTNSSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX81	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNPEMLKLMAKGVLKISSIIIEGFIELSLKQNKKEFTLAGVITLKLLTIAISIIILFANNEIKAAISLGIEEIENLIELFEKLMKEMNVTDSSGIGKLFISAVKNFLLILLSII ESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX82	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNPEMLKLMAKGVLKISSIIIEGFIELSLKQNKKEFTLAAVITNKLLNIAINIIILYANNEIKAAISLGIEEIENLIELFEKLMKEMNVTDSSGIGKLFISAVKNFLLILLSII ESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX83	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKNNPEIYKLFQEGVFKVAKIGSLGSLELFSNPLESNCIQNIFFTSAKLGSMLLNLGLIINNKELLLISLNNILKFIEDESEKILSQISLVGNSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX84	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENAIEYALKFKKISSKIEVGAIEGVIKINPKNYMAVFGIENAGNICKNACNFAILYSMGQLKLAANISLNNLESLLDNFTDAALADPNGANSGIGKLFISAVKNF LLILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK

CALC_HEX85	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENAFEFLLKIEEVAAKGEILAKEAVYKIYPENSLAILTLSICNKIVSSLIDFGLISRLGYIYESAKSALKLLESLIEVFSEFKLKSKDKDNSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX86	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENLFEFLEKIRKVASKGEELAIKAINEIYPDNALAIYAMHICNNIVNNLIMFGIISSKGYIKESAESALKELENLIETFSEFKLKSPDKDNSGIGKLFISAVKNFLLI LLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX87	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKGNKESLLELYKAVLKIEEIGLKGCLALEDQDPVAMKLVAKVILAAALIARELIIDKMFGPLTKESALSLIDKIESESERIFDAMLNAKKKLSGIGKLFISAVKNFLLI LLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX88	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENFKEFSLKFLECAKKVCEGFIKASKKLFNDKISIAFMSIANTVISFINLLIMAGIAILNNDSKVLKKIIENAESIINNSFDMFEALEITNSGIGKLFISAVKNFLLILL SIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX89	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENFKEFSEKLLECAKKVCEGFIKAAKKLFNDKISLAFMSISLTVISLVNLLIMFGIAVLNNDSKVCKKIIENAESIINNSFDMFLALEITDSGIGKLFISAVKNFLLI LLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX90	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENFKEFSSLFLECANKICLGMIKTASELFDDKTTIVFMNIANTVISLVNLLIMFGIACRNNDEKVCKEIIENVEKLIESTFDMFLALKETNSGIGKLFISAVKNFLL ILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX91	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKENFEEFSEMFLECLEKVLLGALKGAKKMFDVKTQMAISNVLEAIKTLAELLILAGIAFRNNDTEAVKAVIEKVSKTVEETLEFLLKLNLPSSGIGKLFISAVKNF LLILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX92	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKTEEGAEALSKFKKVIHKISELASEYLKKKSLYAAKVVSLTGLIGSLLAELCSNYILNSDDKEFQLFEFKETLELLEKLIELAFKYISKLKSSGIGKLFISAVKNFLL ILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX93	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDKELLKEQIKEILKCVNKVYSLILETCMELNWITAMEIFIYIKEIVEELLLLAYTKATNPEKFLEKLESLLELIEEKIKLAFKNLLKGGYNSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX94	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDKELFKEQIKLIEECVNKVYLLALEGARELNWFTLMEIFIYIKEIVDKLLLLGYIKVTKPEEFLKELKNYLELLENKIEEAFEVLLKRGYNSGIGKLFISAVKNFLL ILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX95	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDKELFKEQIKLILECVNKVFKLIWEGAMELNWPTLMKIFTYITFIVNKLLLLGYIKVTNPELFLKELKSFLELLEKVIEEAFKVLLDYKPHSGIGKLFISAVKNFLL ILLSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
CALC_HEX96	DRELLKEAIETVRELLEKLLERNPNDKELLEKIVDIIDEILDKVYKYAEEEKDRELLKEAIETVRELLEKLLERNPNDKELIKKILEIVEKIFELTYKVAVAWQWSVEEFCELVSLILN LIFKLIEIILKNNKDKELFKEQIKLILECVEKVYKLIIEGAMELNWPTLMRIFIYIKFIVDKLLLLGYIKVTNPELFLKELKSFLELLEKVIEEAFEVLLEYKPHSGIGKLFISAVKNFLLIL LSIIESSPLSEEEKKELIKKILEIVEKIFELTYKVAVAEK
ARAG65	SVRERARRILREGEERVREITDVEEREEEVEEVAEELEELAEELSAEDGAEVAREVARLAAEAARGSEDEEEATETVEEAAEEASEILTTTEASADDAVDVATELTELAAAEV TDAEERTETTEETGEAALEAVDDTESKVSAETIRRFTDAGAAAAAKGTDEEERTEGFEETGEEVAELCTESEISATEAVDIFTELLEKGAAVVTDAEEALELAEETGEEAEDA CDDSESKASYESIRKFAEAGAEAATKGTDEEERREAFEETREEVEELCEESELSEEELEEIKRELEEREKRA
ARAG66	ETERRLREILEETEKEIEKITDADEREEKVEEGGEKILELVEENEDLSPEAGTEIARRILEIAARAAKRGTDDEERTETLEEAAEIASEILTETKATVTDAVEVFTEALKAATEVIS DADEVTETLEEALEAAEEVVDEDLEGDPEKAVKAVTEATKAAVDAAAKDGDDDETTETAEEASEALAEIATKLKATATTAVTALTDLTKAAARNITDADKVTETLEEALEAAEE VVDEGLESDPEAVVTAVTDAIKAALRAAARSGDEEKIEETAREGREALEELAEKLKASEELLERAREELERAEREV

*Names shown in green are more thoroughly characterized designs. Sequences shown in color correspond to the SNAC-tag and His-tags that were cleaved during IMAC purification. D3- are dihedral cage designs targeting $CoCO_3$, P6_32- designs are two-dimensional array designs targeting calcite, and PK designs are aragonite-targeting designs.

Supplementary Table S2: Data collection and refinement statistics

	COCA_BRL11 Apo (PDB ID: 13EH)	**COCA_BRL11 Holo (PDB ID: 13EG)**
Data collection		
Space group	*P 1*	*P* 2_1 *2*$_1$ *2*$_1$
Cell dimensions		
a, *b*, *c* (Å)	79.69, 90.66, 107.54	69.68, 85.61,181.67
α, β, γ (°)	89.19, 71.23, 65.57	90, 90, 90
Resolution (Å)	39.11 - 3.40 (3.50 - 3.40)*	41.67 - 4.15 (4.75 - 4.15)
R_{merge}	0.354 (1.889)	0.411 (4.624)
$I / \sigma I$	2.37 (0.38)	4.29 (0.39)
Completeness (%)	98.33 (95.83)	90.66 (71.83)
Redundancy	3.6 (3.6)	13.2 (13.6)
$CC_{1/2}$	0.942 (0.434)	0.997 (0.381)
Refinement		
Resolution (Å)	39.11 - 3.40 (3.50 - 3.40)	41.67 - 4.15 (4.75 - 4.15)
No. reflections	34835 (2850)	7868 (2032)
R_{work} / R_{free}	0.2418 (0.2907) /0.3006 (0.3384)	0.3019 (0.3497) /0.3578 (0.4190)
No. atoms		
Protein	17896	8901
Ligand/ion	n/a	17
Water	n/a	n/a
B-factors		
Protein	48	148

Ligand/ion	n/a	134
Water	n/a	n/a
R.m.s. deviations		
Bond lengths (Å)	0.002	0.003
Bond angles (°)	0.51	0.58

*Single Crystal used for each data collection. *Values in parentheses are for highest-resolution shell.

Supplementary Table S3: CryoEM Data Collection, Processing, and Modeling Statistics

$CoCO_3$-mineralized COCA_BRL11	
Data Collection	
Microscope	Titan Krios
Voltage (kV)	300
Detector	Gatan K3
Energy Filter	Gatan BioQuantum
Recording mode	Counting
Magnification	105,000 X
Movie micrograph pixel size (Å)	0.83
Dose rate ($e^-/Å^2/s$)	10.9
No. of frames per movie micrograph	100
Frame exposure time (s)	0.05
Movie micrograph exposure time (s)	5.00
Total dose ($e^-/Å^2$)	54.26
Under focus range (μm)	0.8 - 1.8
Number of movie micrographs	5,171
Map Processing	304
Symmetry Applied	None (C1)
Extraction Box Size (pix)	256
Initial particle images (no.)	712,226
Final particle images (no.)	87,522
Map resolution (Å)	3.46
FCS threshold	0.143
Map resolution range (Å)	3.13 - 3.77
Refinement	
Initial model used	Design Model
Map resolution (Å)	3.46
FCS threshold	0.143
Model resolution range (Å)	3.13 - 3.77
Map sharpening B factor	130.7
Model composition	
Non-hydrogen atoms	8,115
Protein Residues	1,169
Ligands	Na

Data Collection	
Microscope	Titan Krios
B factors (Å)	
Protein	187.29
Ligands	Na
R.M.S. deviations	
Bond lengths (Å)	0.002
Bond angles (°)	0.487
Validation	
MolProbity score	1.17
Clashscore	3.77
Rotamer Outliers (%)	0.32
Ramachandran plot	
Favored (%)	98.88
Allowed (%)	1.12
Disallowed (%)	0.00

Methods:

Computational methods

RFdiffusion 2 fine-tuning

Repeat protein annotations were retrieved from RepeatsDB via its public REST API and consolidated into a dataset. Corresponding structural models were obtained from either the RCSB Protein Data Bank (experimental structures, chain-specific) or AlphaFoldDB (predicted models). Structural preprocessing included removal of hydrogen atoms, exclusion of structures with average B-factors ≥30, deletion of ligands with no atoms within 4.0 Å of protein atoms, and removal of undefined ligands (UNK). Structures that failed parsing or processing were discarded.

Repeat annotations were then used to define the boundaries of structured repeat regions, and all residues outside these regions were removed before further analysis. For AlphaFoldDB-derived structures, only repeat regions with a mean pLDDT > 90 were retained. 38,274 structures were clustered at 30% sequence identity over 80% coverage using MMSeq2[38], yielding 8,712 unique clusters.

For model training, the new dataset was added at 25%, with the original dataset maintained at 75%. The model was trained for 10 epochs with weights initialized from the original RFdiffusion 2 paper[22]. The weight dimensions were reshaped to include secondary structure and adjacency matrix conditioning, increasing d_t1d by 5 and d_t2d by 4. Fine-tuned weight checkpoints were benchmarked on mineral facet tip-atom scaffolding tasks, with evaluation beginning between epochs 8-10 (Fig. S1).

Backbone generation

Mineral crystal structures were obtained from the Crystallography Open Database (COD)[39]. For atomic motif scaffolding, a subset of glutamate OCO atoms was superimposed onto selected carbonate OCO atoms before scaffolding with RFdiffusion 2. For larger motif scaffolding, motifs were first aligned along specified eigenvectors and then translated and rotated to achieve the desired symmetry. Various versions of RFdiffusion were used in this study. Detailed scripts and protocols will be made available on GitHub.

MPNN

Protein sequences were designed with the --temperature 0.3 and --pack_side_chains 1 flags enabled for all runs. When mineral facets were not modeled, ProteinMPNN[29] was used directly. In the presence of mineral facets, LigandMPNN[26] was employed. Residues designated as tip atoms were held fixed, unless specified otherwise, using --fixed_residues '{fixed_residues}' in combination with --ligand_mpnn_use_side_chain_context 1 and --repack_everything 0. When additional residues (for example,

selected glutamates) were held fixed during sequence assignment, they were excluded from the --redesigned_residues '{residues}' list. In some cases, we performed an additional iterative MPNN round on the AlphaFold model with mineral-interacting residues held fixed.

AlphaFold

These metrics were used as filters for AlphaFold 2[27] evaluation, with designs selected for ordering if they satisfied RMSD < 1.5 Å, pLDDT > 90, and PAE < 5. AlphaFold 3[40] was used to evaluate two-dimensional array designs by modelling 6 homo-oligomers, using thresholds of ipTM > 0.3 and pTM > 0.4. Cα coordinates were further analyzed by computing a best-fit plane defined by the first two principal components, and a flatness metric was defined as the variance of absolute Cα distances from this plane. Designs with flatness variance < 40 were selected for ordering. Oligomer designs were evaluated using AlphaFold3[40], with acceptance criteria of RMSD < 5 Å, pLDDT > 80, ipTM > 0.8, and pTM > 0.5.

Experimental methods

Construction of synthetic genes

Synthetic genes were ordered from and cloned by Genscript Inc. (Piscataway, NJ, USA) in pET29b+ E. coli expression vector in frame with a C-terminal 6x His-tag for affinity purification. A SNAC cleavage tag[41] was included before the purification tag.

Protein production

E. coli expression strain BL21(DE3*) (New England Biolabs, MA, USA) was transformed with plasmid for protein expression. After transformation and overnight growth on LB agar Kanamycin selection plates at 37°C, colonies were picked and transferred to 50mL autoinduction media[42] in 250 mL baffled flasks, where they would be incubated at 37°C overnight. At the end of this incubation period, cells were harvested by centrifuging at 4000g for 10 minutes at 25°C. Cultures were scaled to 1-4 L for x-ray crystallography.

Protein purification

Cell pellets were resuspended in 10-50 mL of lysis buffer (50 mM Tris pH 8.0, 150 mM NaCl, 30 mM Imidazole) supplemented with protease inhibitor (Thermo Fisher, MA, USA), lysozyme, and DNAse I; this was performed by vortexing and by mixing with serological pipette. This suspension was lysed by sonication (QSonica Sonicators, CT, USA) at 80% power for 5-10 minutes (10 sec on/10 sec off) with ¾” tips while keeping the cell suspension on ice. Soluble and insoluble lysates were separated by centrifugation at 25°C and 14,000 G for 40 minutes and applied to chromatography columns containing Ni-NTA (Qiagen, MA, USA) resin pre-equilibrated with lysis buffer. The columns were washed with 10x column volume of lysis buffer supplemented with 1M guanidine hydrochloride then 10x column volume of lysis buffer to remove remaining chaotrope. For SNAC cleavage of cobalt carbonate and X-ray crystallography designs, the resin was further washed with 5 column volumes of SNAC buffer (100 mM CHES, 100 mM Acetone oxime, 100 mM

NaCl, 500 mM GnCl, pH 8.6), incubated in 5 column volumes of SNAC buffer spiked with 2 mM $NiSO_4$ on an orbital shaker at room temperature overnight, and collected as the column flow-through. If uncleaved, a final 2 mL of elution buffer (50 mM Tris pH 8.0, 500 mM NaCl, 300 mM Imidazole) was added and collected for protein elution.

Concentration and buffer exchange was performed in a 10 kDa molecular weight cutoff spin concentrator. For the 2D array, dialysis into water was performed in a D-Tube Dialyzer Midi, MWCO 6-8 kDa (MilliporeSigma, MA, USA). Proteins that expressed well were typically further characterized and purified by size exclusion chromatography (SEC). SEC was done on a Superdex 200 increase 10/300 GL column in filtered 10 mM MOPS buffer pH 8.2 for mineralization or TBS buffer (50 mM Tris pH 8.0, 150 mM NaCl) for crystallization.

Negative stain electron microscopy

5 µL was applied on glow discharged, carbon-coated 400-mesh copper grids (01844- F, TedPella,Inc.), and stained using 0.75% uranyl formate[43]. Air-dried grids were then imaged on a FEI Talos L120C TEM (FEI Thermo Scientific, Hillsboro, OR) equipped with a 4K × 4K Gatan OneView camera at a magnification of 57,000x and pixel size of 2.47 Å. Micrographs collection was automated using EPU software (FEI Thermo Scientific, Hillsboro, OR) and were imported into cryoSPARC software[30]. Typically, micrographs were imported with constant CTF and then manual particle picking of 200-400 particles were done to make templates through 2D classing for automated picking. After full scale templated particle picking, additional 2D classing was done, and selected 2D classes were used for C1 (non-symmetrized) ab initio reconstruction.

Cryo-EM sample preparation

Purified protein sample of COCA_BRL11 was prepared at concentrations of 49 µm in 5 mM $NaHCO_3$. 2 µL of sample was applied to glow-discharged Quantifoil R2.0/2.0 300-mesh grids. All grids were vitrified using a Vitrobot Mark IV operated at 22°C and 100% humidity, with blot time 6.5-7.5 seconds and blot force of 0. All grids were plunge-frozen into liquid ethane, clipped following standard procedures, and stored in liquid nitrogen prior to data collection.

Cryo-EM data collection

COCA_BRL11 cryo-EM data was collected using SerialEM[44] on Thermo Fisher Scientific Titan Krios transmission electron microscope equipped with Gatan K3[45] direct electron detector with a BioQuantum energy filter (15 eV slit width) operated in counting mode. Movies were acquired using beam-image shift with multi-hole acquisition schemes and fractionated electron exposures. Movies were acquired with 5 shots per hole across nine holes per stage movement, with a total exposure of 54.26 $e^-/Å^2$ fractionated over 100 frames, yielding 5,171 movies. Movies were collected with nominal defocus ranges spanning −0.8 µm to −1.8 µm.

Cryo-EM data processing

All image processing for the COCA_BRL11 sample was carried out in cryoSPARC[46] (v4.7.1; Structura Biotechnology Inc.). Default parameters were used unless otherwise noted. Raw movies were subjected to Patch Motion Correction and Patch CTF estimation. Curate Exposures was used to remove 2,063 micrographs, leaving 3,101 good micrographs. Manual Picker was used to select 200 particles, followed by 2D classification with: number of 2D classes 20, Force max over poses/shifts set to false. 14 classes were used as templates for a round of template picking with: particle diameter 80 Å, and Maximum number of local maxima to consider set to 2000. 5,084,537 particles were initially selected, filtered down to 770,687 particles with the Inspect Picks Job. Extract Micrograph was used to extract 712,226 particles with a box size of 256 pixels. 2D classification was used with: Number of classes 300, Number of final full iterations 40, and Number of online-EM iterations 40. 28 classes were selected with 188,519 particles. Ab-Intio was carried out using HR-HAIR method[47] with: 3 classes, Maximum resolution 2.3 Å, Initial resolution 5 Å, Fourier radius step 0.005, Center structures in real space set to false, initial minibatch size 300, Final minibatch size 1000. 158,973 particles from the best two volumes were used for another round of HR-HAIR Ab-Initio, using the same settings but with 2 volumes generated. The best Ab Initio map was first downloaded and modified in ChimeraX[48] to remove excess noise with Gaussian filter standard deviation of 2, this modified map was uploaded to cryoSPARC with the Import Volumes Job and used to generate a mask for a local refinement with the Volume Tools job with: Type of output volume set to mask, Threshold 0.05, Dilation radius 3 pixels, Soft padding width 5 pixels. This mask, the best Ab Initio map with 89,069 particles, was used for local refinement to generate an initial map in C1. Exposure group utilities were used to split exposures into 45 groups, these particle groups and the local refinement map were used for a round of Global CTF refinement with: Fit Trefoil set to false. Local Refinement was run again as described above. Using this new map Reference based motion correction was run. Finally a last round of HR-HAIR Ab Initio was run with: 1 volume as the output. A final local refinement was run in C1 with this Ab-Initio map and a mask generated as described above, The final volume had a resolution of 3.46 Å using 87,522 particles. Final map was deposited in the EMDB under accession number: EMD-77159

Cryo-EM model building, refinement, and validation

The COCA_BRL11computational design model was used as initial references for atomic model building. Design models were docked into the final cryo-EM density maps using ChimeraX. Manual inspection and correction in Coot[49–51] and ISOLDE[52] and real-space refinement was performed in Phenix[53]. Multiple rounds of relaxation and real-space refinement were carried out until convergence of model geometry and agreement with the density were achieved. Final model quality was assessed using MolProbity[54]. Figures were generated using ChimeraX. Final structures were deposited in the Protein Data Bank under the following accession number: 35SI.

Crystallization and Structure Determination

Crystallization experiments were performed using the sitting-drop vapor diffusion method. Crystallization screening was performed using a Mosquito LCP by STP Labtech (200 nL drops in a 96-well plate format, at 20°C) and imaged using UVEX microscopes and a UVEX PS-256 system (JAN Scientific). Resulting crystals were harvested directly from the screening plate.

COCA_BRL11 apo crystals grew in 0.1 M sodium citrate, pH 5.5, 20% (w/v) PEG 3000, were harvested and flash frozen in liquid nitrogen. COCA_BRL11 apo crystals to be soaked with Co grew in 0.1 M BIS-TRIS, pH 6.5, 20% (w/v) PEG 5000 MME, were transferred to a solution consisting of 0.1 M BIS-TRIS, pH 6.5, 20% (w/v) PEG 5000 MME, 20% glycerol, 187.5 µM $CoCl_2$, and 250 µM $NaHCO_3$, in which they were stored for 12 hours prior to being flash-frozen in liquid nitrogen.

X-ray diffraction data were collected on the 17-ID-2 (FMX) beamline of the National Synchrotron Light Source II (NSLS2) at wavelength 0.97934 Å. Diffraction images were indexed and integrated using xia2[55]/DIALS[56] and scaled and merged using Pointless and Aimless within the CCP4 program suite[57]. Initial phases were obtained by molecular replacement using DIMPLE[58] with the designed model as the search model. Iterative model building and corrections were performed manually using COOT[49] following molecular replacement and subsequent structure refinement was performed with CCP4 Refmac5[59]. Initial refinement was conducted using BUSTER[60] to rapidly fix Ramachandran, rotamer, and density fit outliers, refining to convergence. PDB-REDO[61] was used to assess the model quality in between refinements and to fix any rotamer and density fit outliers. Final refinement was carried out in Phenix[53], with manual model building performed in COOT[49].The model quality was evaluated using MolProbity[54]. Data collection and refinement statistics are recorded in Supplementary Table S2. Data deposition, atomic coordinates, and structure factors reported for in this paper have been deposited in the Protein Data Bank (PDB), http://www.rcsb.org/ with accession codes 13EG and 13EH.

Mineralization

All precursor solutions were prepared in Milli-Q water and passed through 0.22 µm filters prior to use. Mineralization reactions were carried out in solution under the specified protein and precursor concentrations, temperatures, and incubation times. Every mineralization experiment was repeated 3-5 times with reproducible results.

$CaCO_3$ mineralization

$CaCO_3$ mineralization was carried out in Milli-Q water using $CaCl_2$ and $NaHCO_3$ as precursor solutions. For aragonite mineralization, $CaCl_2$ and $NaHCO_3$ were mixed to final concentrations of 10 mM each in the presence of 2.16 µM protein. Reactions were performed at 80°C and pH 8.2 for 2 hours. For calcite mineralization on protein arrays, pre-assembled CALC_HEX24 arrays were first prepared in the presence of 500 mM NaCl. Arrays were subsequently transferred into mineralization conditions containing 5 mM $CaCl_2$ and 5 mM $NaHCO_3$ following removal of NaCl by dialysis. Mineralization was carried out at a protein concentration of 0.4 µM at room temperature.

CoCO3 mineralization

$CoCO_3$ mineralization was carried out in Milli-Q water using $CoCl_2$ and $NaHCO_3$ as precursor solutions. Two concentrations were employed depending on the purpose of the experiment. For high-concentration mineralization, $CoCl_2$ and $NaHCO_3$ were mixed to final concentrations of 1 mM each in the presence of varying protein concentrations (0.1, 0.5, and 1.0 μM).

For low-concentration mineralization, $CoCl_2$ and $NaHCO_3$ were mixed to final concentrations of 0.25 mM each in the presence of 0.3 μM protein.

SEM

Scanning electron microscopy imaging was performed using an Apreo LowVac 2S microscope (Thermo Fisher Scientific) equipped with an Oxford Instruments energy-dispersive X-ray spectroscopy (EDS) system. Samples were prepared by depositing 5.0 μL of incubated reaction solution, collected under different experimental conditions, onto a silicon wafer. The samples were allowed to sit for 2 hours under ambient conditions to promote adsorption. The surface was subsequently dried using a gentle N_2 gas flow.

TEM

Typically, samples were prepared by depositing 0.6 μL of reaction solution, collected based on experiment, onto plasma-cleaned carbon-coated copper grids (300 mesh, Ted Pella). TEM analysis was carried out using a FEI Titan environmental TEM (ETEM) operated at an accelerating voltage of 300 kV. High-resolution TEM imaging and further FFT analysis were used to evaluate the crystallinity and structural characteristics of the mineral phases.

AFM

All AFM experiments were conducted using a Bruker MultiMode VIII atomic force microscope. Imaging was performed with PeakForce Tapping mode using ScanAsyst-Air probes (Bruker, CA). For sample preparation, 50 μL of mineralized protein solution was deposited onto a UV-ozone-treated Si substrate and incubated in a humid environment to promote the sample deposition while minimizing drying effects. After 30 minute incubation, excess solution was carefully wicked away, and the substrate was gently dried under a stream of N_2 gas prior to imaging. AFM data were processed using Bruker NanoScope Analysis software (v2.0).

ATR-FTIR

Attenuated total reflectance Fourier transform infrared (ATR–FTIR) spectroscopy was performed using a Bruker Alpha FTIR spectrometer equipped with an ATR module.

For kinetic measurements, freshly prepared reaction solutions were directly deposited onto the ATR crystal, ensuring full coverage of the active area. The liquid sample was then sealed using an additional glass cover to minimize gas (CO_2) exchange with the ambient

environment and to suppress solvent evaporation during measurement. Spectra were collected continuously over time to monitor the evolution of carbonate species.
The formation of carbonate was quantified by tracking the integrated absorbance of the v_3 vibrational band in the range of 1410–1480 cm^{-1}, which minimizes overlap with protein-amide and unreacted bicarbonate (~1380 cm^{-1}) signals.

X-ray diffraction
5 µL of sample was applied on a clean piece of silicon wafer (Ted Pella, INC.) and air dried. Measurements were collected on a Bruker D8 Discover instrument.

Electrochemical OER measurements

OER-associated current–time measurements were performed in 0.1 M KOH using a 5 µm carbon fiber working electrode biased at 1.0 V versus a Ag/AgCl quasireference electrode. The fabrication procedure used in this work to construct carbon fiber electrodes was reported previously.[1] The current was recorded using an Axopatch 200B amplifier (Axon Instruments) while sampling at 100 kHz using a 1440A digitizer (Axon Instruments). The current was filtered at 1 kHz with a low-pass Bessel filter. Current-time data was recorded using Axoscope 11.1 software (Molecule Devices) and analyzed with Clampfit 11.3 (Molecular Devices). All electrochemical experiments were performed in a custom-built, grounded Faraday cage and the electrode was polished on a custom-built microelectrode beveler following each measurement. The electrochemical response was monitored after introducing COCA_BRL11-mineralized $CoCO_3$-containing nanoparticles or control samples into the electrolyte. Under these measurement conditions, transient current spikes above the carbon electrode baseline were observed due to the collision and adhesion of single COCA_BRL11-mineralized $CoCO_3$-containing nanoparticles on the carbon electrode surface, catalyzing the OER upon electrical contact with the electrode surface generating a current spike. Under the applied anodic bias, the carbon electrode itself is partially active towards the OER resulting in a non-zero baseline current. However, because the $CoCO_3$-containing nanoparticle is far more catalytically active towards the OER than the underlying carbon electrode surface, we observe a sharp spike in current above the baseline. These current spikes were thus interpreted as collision-associated OER responses from $CoCO_3$-containing nanoparticles contacting and adhering to the carbon fiber electrode surface. Because high-amplitude events can arise from overlapping multi-particle collisions or particle aggregates, event-level analysis was performed only on isolated, well-defined current spikes that returned to baseline after each event.

For a representative peak analysis, individual current transients were baseline-corrected and time-shifted for visualization. Peak amplitude was determined from the baseline-corrected maximum current, the integrated charge was calculated by integrating the

baseline-corrected current over the event duration, and the full width at half maximum (FWHM) was used to quantify the transient lifetime of each event.

Turnover Frequency (TOF)

Using the average current amplitude, we can calculate the turnover frequency (TOF), representing the number of O_2 molecules produced per second at a single particle:

TOF = i / nF = (25 × 10^{-12} A) / [4 × (96485 C mol^{-1})] = 6.48 × 10^{-17} mol s^{-1} × (6.022 × 10^{23}) = **3.9 × 10^7 O_2 s^{-1} (at a single particle)**

where i is the current, n is the number of electrons transferred during the OER under basic conditions ($4OH^- \rightarrow O_2 + 2H_2O + 4e^-$) and F is Faraday's constant. We can also calculate this per cobalt active site.

If we assume 400 cobalt ions per particle and that all cobalt ions are active, which is an upper bound, this yields a TOF of 9.8 x 10^4 O_2/s per cobalt active site reasonably correlating with prior work.[2,3]

Total Turnover Number (TON)

Using the average charge transferred per collision event, we can calculate the total turnover number (TON), representing the number of O_2 molecules produced during a single current spike:

mol O_2 = Q / nF = (14 × 10^{-15} C) / [4 × (96485 C mol^{-1})] = 3.6 × 10^{-20} mol O_2 × (6.022 × 10^{23}) = **2.2 × 10^4 O_2 molecules per collision event**

where *Q* is the charge, *n* is the number of electrons transferred, and *F* is Faraday's constant.

Visualization and figures

All structural images for figures were generated using ChimeraX[48] or PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC). Figures were made in Google Slides.

References for SI:

1. Defnet, P. A., Han, C. & Zhang, B. Temporally-Resolved Ultrafast Hydrogen Adsorption and Evolution on Single Platinum Nanoparticles. Anal. Chem. 91, 4023–4030 (2019).

2. El Arrassi, A. et al. Intrinsic Activity of Oxygen Evolution Catalysts Probed at Single CoFe2 O4 Nanoparticles. J. Am. Chem. Soc. 141, 9197–9201 (2019).

3. Quast, T. et al. Single-Entity Electrocatalysis of Individual “Picked-and-Dropped” Co3 O4 Nanoparticles on the Tip of a Carbon Nanoelectrode. Angew. Chem. Int. Ed. 60, 3576–3580 (2021).

Chapter 2: Programmed mineralization of metal-oxides by de novo designed protein templates

This chapter contains additional background for (and limited reproduction of) content submitted as: Amijai Saragovi, Harley Pyles, Evelyn B. Huddy, Wenzhao Dai, Xinqi Li, Andrew J. Borst, Nikita Hanikel, Alexis Courbet, Paul S. Kwon, Fátima A. Dávila-Hernández, Ryan Kibler, Lukasz Goldschmidt, Dionne K. Vafeados, Aza Allen, Kenneth D. Carr, Asim K. Bera, Alex Kang, Evans Brackenbrough, Sakshi Yadav Schmid, Yuna Bae, Lance Stewart, Shuai Zhang, James J. De Yoreo, David Baker. *Science* (in revision).

Abstract: We describe a general strategy to design and screen de novo protein interfaces which guide the mineralization of specific metal-oxides, including technologically relevant materials that are not found in biological systems. Two interfaces promoted hematite growth in conditions typically yielding magnetite. Three interfaces promoted heterogenous ZnO nucleation under conditions where traditional ZnO-binding peptides and control proteins were ineffective. The composition and geometry of these interfaces suggests metal coordination and water structuring mechanisms. CryoEM analysis of a designed nanocage with ZnO nucleating interfaces lining interior cavities revealed atomic density, likely corresponding to growing ZnO, directly adjacent to the designed interfaces. These findings demonstrate that designed proteins can template mineralization through modes beyond electrostatics, and provide a route to protein-metal-oxide hierarchical architectures with tunable structure and function.

Introduction

The design of protein–mineral hierarchical assemblies, where proteins program the formation of defined compounds and custom geometries at atomic resolution, has not yet been achieved. Designed protein templates can use charge complementarity to manipulate the crystallization of calcium-containing minerals, native inorganic materials widely used in biology, such as calcium carbonate, by selectively nucleating specific polymorphs (e.g., calcite over vaterite;(*1*)) or regulating intermediate mineral species in solution(*2*). However, the vast majority of the universe of possible protein-inorganic hybrids, including technologically relevant metal oxides, are not explored by nature, and require a more general understanding and design approach. Previous work has leveraged native proteins, most notably ferritin, to encapsulate iron and other minerals(*3–5*). Random peptide sequences, typically 8-12 residues, have been selected from phage display libraries with biopanning and shown to bind a variety of inorganic materials, including sequences that modify semiconductor growth (in conditions which also allow homogeneous growth) and promote semiconductor nucleation when densely displayed on protein assemblies(*6–11*). However, native proteins and synthetic peptides do not provide a mechanistic understanding of how the structure of biomolecular interfaces controls inorganic crystallization, and therefore do not constitute a viable path to develop fully defined structured hybrid materials.

We reasoned that structurally and chemically defined protein interfaces could in principle bind metal oxides more strongly and specifically, and thereby exert greater control over nucleation and growth. The sequence space of the longer (> 50 amino acid)

polypeptides needed to create such proteins is too large to adequately explore with random libraries, necessitating a more directed approach for the design of protein-inorganic interfaces. De novo protein design enables the construction of molecular scaffolds with a wide variety of surface geometries(*12*, *13*), and the patterning of diverse sidechain chemistries on these surfaces. This enables specific, well-structured arrangements of chemical moieties that promote the binding to an inorganic phase to be identified and interrogated. We reasoned that proteins that strongly bind an inorganic crystal could lower the interfacial free energy of nascent nuclei and template heterogeneous growth of the material at supersaturation concentrations substantially below the homogeneous nucleation threshold. We reasoned further that should this be achievable, the interfaces could be incorporated into protein nanostructures lining a defined interior volume, with pores to allow the entry of precursors, enabling hierarchical assembly of protein–inorganic composites with technologically relevant materials.

We set out to design proteins presenting structured inorganic crystal-binding interfaces. The optimal surface for interacting with inorganic lattices is unknown, and hence we explored three design hypotheses and sought to generate proteins presenting 1) periodic arrays of charges(*1*, *14*), 2) residues partially coordinating metal ions(*15*), and 3) ordered hydration layers promoting adsorption(*16*, *17*)(**Figure 1A**). As the optimal geometry for protein-templated inorganic nucleation is also unknown, we presented these interfaces on designed scaffolds with different topologies, periodicity, and curvature. These scaffolds included primarily alpha helical designed helical repeat (DHR) proteins(*13*, *18*), designed alpha-beta proteins(*19–22*)(**Figure 1B, C**), and proteins containing flat beta-sheet surfaces: designed idealized beta-solenoid (DBS) repeat proteins (see Methods), a native antifreeze protein (RiAFP), and an idealized penta-peptide repeat (PPR) fold(*23*, *24*) (**Figure 1D**). We selected two target materials that can be synthesized in aqueous solutions but are not known to occur as primary mineral phases in native bio-inorganic hybrid materials: zinc oxide (ZnO) and hematite (alpha-Fe2O3)(*25*). ZnO has diverse applications in emerging technologies(*26*), energy scavenging nanosystems(*27*), thin film transistors(*28*), field emitters(*29*), and can be detected using photoluminescence (PL) spectroscopy(*30*). Hematite (alpha-Fe2O3) is a hydrogen-generating photocatalyst(*31*) that can be distinguished from other iron oxides by color(*32*).

We defined sets of interface amino-acid compositions based on the three above design hypotheses (**Figure 1A, Supplementary Figure 1**), histidines and cysteines were included to test metal coordination, threonine-rich surfaces were used to mimic water organizing motifs in ice-binding proteins, and acidic and basic residues were used to achieve periodic arrays of charges. We placed these potential mineral-interacting interfaces on each scaffold (**Figure 1B-D**) with each composition in both periodic and aperiodic arrangements (**Figure 1E-Q**), resulting in 7×10^3 designed proteins (**Figure 1R**). Outside of the intended binding surface, the sequence of the proteins remained fixed, and the original scaffold sequences were also included in the library.

We developed a yeast display flow cytometry (FC) screen that relies on the light scattering properties of nanoparticles (NPs) to identify and enrich yeast cells with increased binding propensity to ZnO or hematite NPs (**Figure 2A, B**; **Supplementary Figures 2-4**). Next-generation sequencing (NGS) was used to compare the abundance of designs in sixteen subpopulations of the library before

and after sorting and identify enriched sequences (**Figure 2C; Supplementary Figure 5-6**; see Methods). Individual enriched clones were retested by FC on yeast (**Figure 2C; Supplementary Figure 7-8; Supplementary Table 1-6**). In cases where mutated, truncated, or chimeric sequences were abundant in the sorted subpools, we tested the observed sequences as well as the design with the highest identity, and we also tested 'fiber' versions of the repeat protein designs where the sequence capping features on the terminal repeat were removed to promote head-to-tail interactions **(Supplementary Table 4;** see Methods). The identified ZnO binders included eight DBSs, two DHRs, two alpha-beta topologies, and a surface redesign of RiAFP. Five of these surfaces contained histidine and cysteine residues, which can coordinate zinc ions(*33*), six contained large hydrophobic patches, five contained threonine motifs observed in ice-binding proteins(*16*), and eight interfaces contained periodic changes (**Figure 2D-H; Supplementary Figure 9; Supplementary Table 1, 3, 5**). We also identified hematite binders: two DHRs containing histidine-rich surfaces, two DHRs with charge-rich surfaces, and three beta-solenoids containing hydrophobic surface patches (**Figure 2I; Supplementary Figure 10; Supplementary Table 2, 4, 6**). We expressed these proteins in *E. coli* with 6xHis tags and purified them with immobilized metal affinity chromatography (IMAC; **Figure 2D-F, I, Supplementary Figures 11-12; Supplementary Table 7-8**).

Biochemical characterization of metal-oxide binding proteins

We characterized five ZnO binding proteins (Z0, Z1, Z3, Z4, and Z5) that had substantial FC binding signals to ZnO NP and expressed well in *E. coli* as soluble proteins for further characterization. Z3 and Z5 were monodisperse by size exclusion chromatography (SEC), while Z0, Z1, and Z4 appeared to form large assemblies (**Supplementary Figure 11)**. Z0, Z3, Z4, and Z5 produced circular dichroism (CD) spectra consistent with their predicted secondary structure, while Z1 appeared disordered (**Supplementary Figure 11)**. Z0 formed long fibers and bundles of fibers visible with negative-stain transmission electron microscopy (ns-TEM), presumably through strand pairing head-to-tail interactions (**Supplementary Figure 13A**). To regularize these interactions, we designed and produced a version without these capping features, Z0-fiber, and used it in further analysis of this design (**Supplementary Figure 13B**). In a pull-down assay followed by SDS-PAGE, all of the purified ZnO designs bound ZnO NPs and did not bind to anatase, rutile, or hematite NPs, while bovine serum albumin (BSA) and lysozyme controls showed no binding (**Supplementary Figure 14**). Atomic force microscopy (AFM) of the zinc-terminated (001) surface incubated with the proteins showed binders had similar (Z3, Z4) or greater (Z0-fiber) coverage in respect to controls (**Supplementary Figure 15**). The three designs which produced the highest yields of purified protein, Z0-fiber, Z3, and Z4, were selected for further study.

We characterized two hematite binders, H7 and H9, which expressed as soluble protein in *E. coli* and demonstrated robust hematite NP binding signals when displayed as single clones on yeast (**Figure 2I, Supplementary Figure 8, 12; Supplementary Table 2).** In order to increase the yield and improve the monodispersity of these proteins, we mutated non-interfacial residues to negative residues, resulting in H7-neg and H9-neg (**Supplementary Figure 16; Supplementary Table 6**). No protein associated with hematite NP was

detected in a SDS-page pull-down assay (not shown); however, hematite nanoparticles incubated with H7-neg showed decreased flocculation compared to solutions without protein or BSA (**Supplementary Figure 17**).

We next tested if the designs coordinate metal ions. To eliminate the potential effects of His-tags, the designs were produced with strep tags instead and purified with Strep-Tactin beads (**Supplementary Table 7-8;** see Methods). Using isothermal titration calorimetry (ITC), Z0-fiber was found to bind Zn^{2+} exothermically with a Kd of 4.1 ± 1.4 μM, and Z3 was found to bind Zn^{2+} endothermically with a Kd of 16 ± 5.0 μM (**Figure 2G**). A ratiometric dye (Mag-Fura-2) assay indicated that Z0-fiber binds Zn^{2+} ions with a best fit model indicating six binding sites, three with 3.8 ± 0.2 nM affinity and three with 60 ± 6 nM affinity (**Figure 2D**), and Z3 showed three binding sites with 200 ± 60 nM affinity (**Figure 2E**; the difference in measured affinities is likely due to the lower pH for the ITC measurements (see Methods)). Z0-fiber interfacial amino acid substitutions indicated that the two rows of repeating histidine residues are critical for Zn^{2+} ion binding (**Figure 2D**). In the Z3 interface, mutating glutamate, cysteine, and histidine residues reduced Zn^{2+} ion binding, while mutation of basic residues did not (**Figure 2E**). Mutating histidines in the interfaces of both Z0-fiber and Z3 also resulted in reduced ZnO NP binding when displayed on yeast (**Supplementary Figure 18**). No heat signal was detected upon titration of Zn^{2+} ions to Z4 (**Figure 2G**), and the Kd was estimated to lie above the quantification limit of the dye competition assay using Mag-Fura-2 (> 1 μM; **Supplementary Figure 19**). Z4 contains a threonine motif found in alpha-helical ice-binding proteins (**Figure 2F, H**), and four other enriched proteins (Z8, Z9, Z12, and Z14) contain another threonine motif found in beta-solenoid ice-binding proteins (**Figure 2H, Supplementary Figure 9**)(*16*).

A ratiometric dye assay using Mag-Fura-2 showed that both H7 and H9 coordinate Co^{2+} ions, a common analog for Fe^{2+} (Figure 2I). Co^{2+} is typically used to assess affinity since Fe^{2+} often oxidizes to Fe^{3+}, forming ferrihydrite. H7 exhibited strong nanomolar binding affinity, while H9 only weakly coordinated Co^{2+} with micromolar affinity (**Figure 2I)**.

Assessment of metal oxide templating activity

We assayed the ability of the purified ZnO-binding proteins to promote the growth of ZnO NPs by incubating them in supersaturated solutions (3 mM $Zn(NO_3)_2$, 50 mM NaCl, 100 mM HEPES, pH 8.2), which allowed heterogeneous growth of ZnO, but not homogeneous growth (**Figure 3A**). We measured the photoluminescence (PL) of ZnO NPs to assess protein-induced growth in these conditions and used solutions seeded with 13 μg/mL of commercial ZnO NPs as a positive control, an ideal template. After 2 hours, neat solutions and solutions containing lysozyme showed no PL, while solutions seeded with 0.1 mg/mL of Z0-fiber, Z3, or Z4 had PL spectra consistent with ZnO NPs (**Figure 3B**)(*34*). X-ray powder diffraction of the material grown in the presence of Z0-fiber, Z3, and Z4 produced diffraction patterns indicative of ZnO (**Figure 3C**). Tracking PL over time showed the increase induced by the ZnO NPs control began immediately, while the signal increased in solutions with the proteins began after a delay of 30 to 60 minutes (**Figure 3D**). Additional control proteins BSA, DNase I, and DHR10-mica6(*14*) had no observable effect on the level of the PL signal

(**Supplementary Figure 20**). Z0-fiber exhibited inconsistent behavior among replicates, showing growth signals in only two out of six replicates, whereas all replicates containing Z3, Z4, and ZnO NP showed growth (**Supplementary Figure 21**). To confirm that the interfacial motifs we identified were responsible for the ZnO templating activity of the proteins, we tested the activity of the motifs when hosted by different scaffold proteins. Reasoning that the aggregation-prone behavior of Z0-fiber may be impeding its function, we extracted the putative functional motifs and designed topologies with RFdiffusion(*35*) and MPNN(*36*) to host the Z0 motifs in alpha-beta proteins (**Figure 3E**). We expressed eleven of these 'zab' proteins in *E. coli* and selected four, zab3, zab5, zab10, and zab11, with CD spectra and SEC traces consistent with the models (**Supplementary Figure 22**). Having observed threonines in the Z4 interface motifs (**Figure 3F**) which resemble a helical ice-binding motif (**Figure 2H)**, we tested a previously designed ice-binding twistless-helical repeat protein, b-iTHR-201 (**Figure 3G**), which presents a larger array of this motif(*37*). This ice-binding protein and two of the redesigns based on the Z0 interface (zab3 and zab10) showed strong templating signals in the ZnO NP growth assay (**Figure 3H, I**); to account for potential variability, six replicates were run for each of these conditions (**Figure 3J**). Four previously reported ZnO binding peptides(*7–10*), fused to Small Ubiquitin-like Modifier (SUMO) to aid solubility and expression, showed no activity (**Figure 3J**; **Supplementary Figure 23, Supplementary Table 9**). For replicates that triggered ZnO growth, we compared the times of nucleation, which we defined as the halfway point between the initial and maximum PL signal. Five design templates promoted an increase in photoluminescence at times greater than or equal to the control NPs, while the increase promoted by zab3 occurred substantially earlier (**Figure 3K**).

We examined if the H7 and H9 interfaces could direct the formation of hematite, the specific material they were selected for, over a variety of other iron oxide compounds. Operating under anaerobic growth conditions in a magnetic field which typically yield magnetite (**Supplementary Figure 24**), we found that supersaturated solutions containing H7 were more red in color after 20 minutes incubation, consistent with the presence of hematite, compared to solutions containing H9, lysozyme, no additive, or hematite or magnetite NPs (**Figure 3L, M; Supplementary Figure 25; Supplementary Movie 1**). The reaction was allowed to continue for 120 minutes, but the color was only compared within the first 20 minutes because subsequently, the magnetite material began to be pulled out of solution by the external magnetic field. (**Figure 3L, Supplementary Movie 1**). Raman spectroscopy of the reaction products indicated hematite was produced in solutions containing H7-neg and H9-neg, while magnetite was produced in solutions without protein or with lysozyme (**Figure 3N; Supplementary Figure 25)**. We next examined the effect of H7 on oxidation at the iron-water interface in aerobic conditions. After a 2-hour incubation on polished iron surfaces, solutions containing H7-neg accumulated less rust than samples incubated without protein, an effect not seen for a lysozyme control (**Figure 3O**).

Confinement of zinc oxide mineralization

We examined the effect of confinement on the rate of nucleation by lining the interior cavities of oligomers with the ZnO nucleating Z4 interface. We used RFdiffusion(*35*) and MPNN(*36*) to design cyclic oligomers that host motifs from the Z4 interface within the

interior of rings (**Figure 4A-D**, see Methods). We experimentally tested twenty-three oligomers and selected three designs forming monodisperse species of the expected size, Z4-C3i, Z4-C3ii, and Z4-C6, for further characterization (**Supplementary Figure 27**). The Z4-C3ii-XL trimer was derived from Z4-C3ii by adding another repeat of the Z4 interface to each monomer. The structures of Z4-C3i and Z4-C3ii were verified with X-ray crystallography (**Figure 4A, B**) and the structures of Z4-C6 and Z4-C3ii XL were verified with ns-TEM (**Figure 4C, D**). All four of these oligomer designs retained the templating activity of the Z4 monomer, while knockouts versions of Z4-C3i, Z4-C3ii, and Z4-C6 in which the threonine residues in the Z4 interfaces were mutated to glutamate (**Figure 4E)**, which disrupted the ZnO binding interface but not the assembly of the oligomer (**Supplementary Figure 28**), had no templating activity (**Figure 4F**). Cryo-EM of ZnO NPs grown with Z4-C6 showed both empty rings matching the designed hexamer and solid heterogeneous NP of the same diameter. (**Supplementary Figure 29**).

We next incorporated the ZnO templating interface into larger protein assemblies. To the Z4-C3i trimer, we incorporated a homodimeric protein-protein interface with an orientation that generates a 24-subunit octahedral nanocage. (**Figure 4G**). Z4-tube was designed by adding tetrameric domains to the N- and C-terminus of an extended version of Z4, such that the ZnO-interacting interface surrounds a rod-like cavity with a high aspect ratio (**Figure 4H**). These assemblies were expressed in *E. coli* and purified with strep tags and SEC (**Supplementary Figure 30**), and their structures confirmed with ns-TEM (**Figure 4G, H**). In the ZnO NP growth PL assay, Z4-cage promoted a large increase in PL, while Z4-tube showed a substantially lower increase (**Figure 4I)**. There was no clear relationship between the size of the cavity or number of interfaces and the magnitude or rate of ZnO NP nucleation (**Figure 4I, J**).

Cryo-EM characterization of protein-ZnO interface

We used Cryo-EM to investigate the structure of ZnO NP formed within the Z4-cage. While the majority of Z4-cage particles appeared empty in raw micrographs **(Figure 5A)**, Cryo-EM micrographs indicated stochastic NP growth originating from a subset of designed protein nanocages, with three distinct morphologies observed **(Figure 5A, B)**: first, within the cage ("filled"), second, growth extending outward ("blebbed"), and third, connecting sets of cages ("bridged") **(Figure 5B)**.

2D class averages of Z4-cage particles revealed secondary structure elements consistent with the design **(Figure 5C; Supplementary Figure 31)**, and 3D refinements with relaxed octahedral symmetry produced a ~3.8 Å Cryo-EM structure closely aligning with the design model **(Figure 5D, E; Supplementary Figure 31; Supplementary Table 10)**. Despite achieving a moderately high-resolution map, the asymmetric and stochastic growth of ZnO, combined with the imposition of the octahedral symmetry operator, hindered clear visualization and reliable analysis of ZnO density at this resolution. Consequently, a global C1 3D refinement was performed using a near-spherical mask, allowing for detailed and unbiased analysis of the pores along both the C3 and C4 axes. This approach ensured an overall averaged representation of the Z4-cage **(Supplementary Figure 32)**. The resulting ~4 Å asymmetric map was low-pass filtered to 8 Å to highlight lower-resolution features corresponding to observed stochastic ZnO nucleation and growth patterns **(Figure**

5A, B; Supplementary Figure 32). This approach revealed additional non-protein density, likely ZnO, localized along the center of each C3 axis and proximal to the designed ZnO binding interface **(Figure 5F; Supplementary Figure 32; Supplementary Table 10)**. In contrast, no extra density was observed along the C4 axes (distant from the designed ZnO binding interface) at any contour level **(Supplementary Figure 32)**. To evaluate whether ZnO nucleation occurred uniformly across all C3 trimers, we next performed octahedral symmetry expansion followed by 3D classification for all Z4-C3i trimers within the Z4-cage. As expected, a subset of C3 trimers displayed ZnO density along the C3 axis, with several 3D classes appearing empty, which is consistent with both the empty particle morphologies and asymmetry growth patterns of ZnO observed in raw micrographs **(Figure 5A, G)**.

3D local refinement of the threefold-symmetry axis of the mineralized Z4-cage yielded a near-atomic resolution cryo-EM map, ranging from 2.5 Å resolution in the center of the C3 axis to ~3.5 Å at the interface between adjacent Z4-C3i trimers in the nanocage **(Figure 5H; Supplemental Fig. 31; Supplementary Table 10)**. At this resolution, density consistent with solvent organization was visible along the ice-binding motifs in the Z4 designed interface(*16*) **(Figure 5H-J)**. Sixty-one waters were modeled into this density along the interface (see Methods), although the precise identity of these ordered species remains uncertain at this resolution. These results are consistent with stochastic ZnO NP nucleation within the trimers promoted by atomic ordering at the designed interface, followed by further ZnO growth into heterogeneous morphologies **(Figure 5)**.

Discussion

We present an approach to de novo design and screen de novo protein-mineral interfaces and apply it to identify proteins that bind ZnO and hematite NPs. We identified seven surfaces that promote hematite binding (**Supplementary Figures 10**) and demonstrate that two proteins, H7-neg and H9-neg, direct the formation of hematite in conditions that otherwise produce magnetite (**Figure 3L-N**), and appear to reduce the rate of oxidation at iron-water interfaces (**Figure 3O**). These proteins coordinate Co^{2+}, an established proxy for Fe^{2+} (**Figure 2I**), suggesting they function through metal-mediated interactions. We identified thirteen de novo ZnO binding protein interfaces (**Supplementary Figure 9**) and demonstrate that three of them, Z0-fiber, Z3, and Z4, promote ZnO growth (**Figure 3B-D**; **Supplementary Figure 9)**. Our results suggest these interfaces function by two distinct mechanisms. Firstly, there is evidence for direct ion-mediated interactions; Z0-fiber and Z3 coordinate Zn^{2+} ions and mutating metal-coordinating residues reduces their affinity to both ions and ZnO NPs (**Figure 2D, E, G**; **Supplementary Figure 18**). Secondly, ZnO binding interfaces, including Z4, contain water-structuring threonine motifs observed in ice-binding proteins (**Figure 2F, H; Supplementary Figure 9**), suggesting they may adsorb to ZnO via ordered water layers(*16*, *17*, *38*). Hydrophobic interfaces were also selected in our assay, but these proteins could not be purified from *E. coli* and the mechanism of their binding is unclear. While periodic charges were present in interfaces alongside other features (**Figure 2**, **Supplementary Table 1-2**), no interfaces composed entirely of arrays of charged groups, like those seen in previous mineral-templating and mineral-binding designed proteins(*1*, *14*), were observed. ZnO NP binding was more predictive of nucleation activity than binding to bulk zinc-terminated ZnO (**Fig. 3J, Supplementary Figures 14, 15**).

Our designed metal-oxide interacting proteins are larger, more ordered, and better structurally understood than the short peptides previously identified with inorganic-binding assays(*6*). The protein-ZnO interfaces template ZnO growth in supersaturated solutions under conditions where ZnO-binding peptides have no effect and maintain this function when grafted into new structures (**Figure 2 D-F, 3 E-K**). When lining the interior of protein assemblies, the Z4 interface promoted ZnO formation within the assembly (**Figure 4, Figure 5 F, G, Supplementary Figure 32**). CryoEM characterization of a designed 18 nm octahedral nanocage with the Z4 interface lining a triangular cavity along the threefold axis showed that NP nucleation occurs adjacent to this interface (**Figure 3 F, Figure 5 H-J**).

Nature has evolved biomineralizing proteins which control the formation of hybrid organic-inorganic materials, such as bone and nacre, with greater complexity than is accessible with conventional top-down manufacturing(*25*). Modern protein design tools have made protein structures and protein-protein interactions readily programmable, but inorganic-templating protein interfaces are inadequately understood. Our high-throughput design and evaluation approach provides a route to overcoming this limitation that can be readily extended to a wide range of inorganic crystals. In a complementary co-submission, we show that protein assemblies can template and confine calcium phosphate, a biologically relevant mineral, within defined nanoscale compartments. Together, these studies open the door to a new world of designed hybrid materials containing inorganic components–including those not explored by nature–that are controlled by de novo proteins.

References

1. F. A. Davila-Hernandez, B. Jin, H. Pyles, S. Zhang, Z. Wang, T. F. Huddy, A. K. Bera, A. Kang, C.-L. Chen, J. J. De Yoreo, D. Baker, Directing polymorph specific calcium carbonate formation with de novo protein templates. *Nat. Commun.* **14**, 8191 (2023).

2. B. Jin, Y. Chen, H. Pyles, M. D. Baer, B. A. Legg, Z. Wang, N. M. Washton, K. T. Mueller, D. Baker, G. K. Schenter, C. J. Mundy, J. J. De Yoreo, Formation, chemical evolution and solidification of the dense liquid phase of calcium (bi)carbonate. *Nat. Mater.* **24**, 125–132 (2025).

3. K. Iwahori, K. Yoshizawa, M. Muraoka, I. Yamashita, Fabrication of ZnSe nanoparticles in the apoferritin cavity by designing a slow chemical reaction system. *Inorg. Chem.* **44**, 6393–6400 (2005).

4. J. Zhang, D. Cheng, J. He, J. Hong, C. Yuan, M. Liang, Cargo loading within ferritin nanocages in preparation for tumor-targeted delivery. *Nat. Protoc.* **16**, 4878–4896 (2021).

5. K. Fan, C. Cao, Y. Pan, D. Lu, D. Yang, J. Feng, L. Song, M. Liang, X. Yan, Magnetoferritin nanoparticles for targeting and visualizing tumour tissues. *Nat. Nanotechnol.* **7**, 459–464 (2012).

6. U. O. S. Seker, H. V. Demir, Material binding peptides for nanotechnology. *Molecules* **16**, 1426–1451 (2011).

7. M. Umetsu, M. Mizuta, K. Tsumoto, S. Ohara, S. Takami, H. Watanabe, I. Kumagai, T. Adschiri, Bioassisted room‐temperature immobilization and mineralization of zinc oxide—the structural ordering of ZnO nanoparticles into a flower‐type morphology. *Adv. Mater.* **17**, 2571–2575 (2005).

8. M. M. Tomczak, M. K. Gupta, L. F. Drummy, S. M. Rozenzhak, R. R. Naik, Morphological control and assembly of zinc oxide using a biotemplate. *Acta Biomater.* **5**, 876–882 (2009).

9. Z. Wei, Y. Maeda, H. Matsui, Discovery of catalytic peptides for inorganic nanocrystal synthesis by a combinatorial phage display approach. *Angew. Chem. Int. Ed Engl.* **50**, 10585–10588 (2011).

10. P. Golec, J. Karczewska-Golec, M. Loś, G. Węgrzyn, Novel ZnO-binding peptides obtained by the screening of a phage display peptide library. *J. Nanopart. Res.* **14**, 1218 (2012).

11. M. T. Klem, D. Willits, D. J. Solis, A. M. Belcher, M. Young, T. Douglas, Bio‐inspired synthesis of protein‐encapsulated CoPt

nanoparticles. *Adv. Funct. Mater.* **15**, 1489–1494 (2005).

12. B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, D. Baker, Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

13. T. F. Huddy, Y. Hsia, R. D. Kibler, J. Xu, N. Bethel, D. Nagarajan, R. Redler, P. J. Y. Leung, C. Weidle, A. Courbet, E. C. Yang, A. K. Bera, N. Coudray, S. J. Calise, F. A. Davila-Hernandez, H. L. Han, K. D. Carr, Z. Li, R. McHugh, G. Reggiano, A. Kang, B. Sankaran, M. S. Dickinson, B. Coventry, T. J. Brunette, Y. Liu, J. Dauparas, A. J. Borst, D. Ekiert, J. M. Kollman, G. Bhabha, D. Baker, Blueprinting extendable nanomaterials with standardized protein blocks. *Nature* **627**, 898–904 (2024).

14. H. Pyles, S. Zhang, J. J. De Yoreo, D. Baker, Controlling protein assembly on inorganic crystals through designed protein interfaces. *Nature* **571**, 251–256 (2019).

15. S. Ciambellotti, C. Pozzi, S. Mangani, P. Turano, Iron biomineral growth from the initial nucleation seed in L-ferritin. *Chemistry* **26**, 5770–5773 (2020).

16. P. L. Davies, Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. *Trends Biochem. Sci.* **39**, 548–555 (2014).

17. D. Martin-Jimenez, E. Chacon, P. Tarazona, R. Garcia, Atomically resolved three-dimensional structures of electrolyte aqueous solutions near a solid surface. *Nat. Commun.* **7**, 12164 (2016).

18. T. J. Brunette, F. Parmeggiani, P.-S. Huang, G. Bhabha, D. C. Ekiert, S. E. Tsutakawa, G. L. Hura, J. A. Tainer, D. Baker, Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).

19. B. Koepnick, J. Flatten, T. Husain, A. Ford, D.-A. Silva, M. J. Bick, A. Bauer, G. Liu, Y. Ishida, A. Boykov, R. D. Estep, S. Kleinfelter, T. Nørgård-Solano, L. Wei, F. Players, G. T. Montelione, F. DiMaio, Z. Popović, F. Khatib, S. Cooper, D. Baker, De novo protein design by citizen scientists. *Nature* **570**, 390–394 (2019).

20. H. Jiang, K. M. Jude, K. Wu, J. Fallas, G. Ueda, T. J. Brunette, D. Hicks, H. Pyles, A. Yang, L. Carter, M. Lamb, X. Li, P. M. Levine, L. Stewart, K. C. Garcia, D. Baker, De novo design of buttressed loops for sculpting protein functions. *bioRxiv*, doi: 10.1101/2023.08.22.554384 (2023).

21. D. Baker, D. Zorine, Design of Repeat Alpha-Beta Proteins with Capping Helices, *bioRxiv* (2024)p. 2024.06.15.590358.

22. C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, Foldit Players, D. Baker, S. Ovchinnikov, Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).

23. A. Hakim, J. B. Nguyen, K. Basu, D. F. Zhu, D. Thakral, P. L. Davies, F. J. Isaacs, Y. Modis, W. Meng, Crystal structure of an insect antifreeze protein and its implications for ice binding. *J. Biol. Chem.* **288**, 12295–12304 (2013).

24. G. W. Buchko, S. Ni, H. Robinson, E. A. Welsh, H. B. Pakrasi, M. A. Kennedy, Characterization of two potentially universal turn motifs that shape the repeated five-residues fold--crystal structure of a lumenal pentapeptide repeat protein from Cyanothece 51142. *Protein Sci.* **15**, 2579–2595 (2006).

25. P. M. Dove, J. J. De Yoreo, S. Weiner, *Biomineralization* (Walter de Gruyter GmbH & Co KG, 2018).

26. B.-Y. Oh, M.-C. Jeong, T.-H. Moon, W. Lee, J.-M. Myoung, J.-Y. Hwang, D.-S. Seo, Transparent conductive Al-doped ZnO films for liquid crystal displays. *J. Appl. Phys.* **99**, 124505 (2006).

27. Y. Qin, X. Wang, Z. L. Wang, Microfibre-nanowire hybrid structure for energy scavenging. *Nature* **451**, 809–813 (2008).

28. K. Nomura, H. Ohta, K. Ueda, T. Kamiya, M. Hirano, H. Hosono, Thin-film transistor fabricated in single-crystalline transparent oxide semiconductor. *Science* **300**, 1269–1272 (2003).

29. Y. B. Li, Y. Bando, D. Golberg, ZnO nanoneedles with tip surface perturbations: Excellent field emitters. *Appl. Phys. Lett.* **84**, 3603–3605 (2004).

30. D. K. Sharma, S. Shukla, K. K. Sharma, V. Kumar, A review on ZnO: Fundamental properties and applications. *Materials Today: Proceedings* **49**, 3028–3035 (2022).

31. Y. Y. Avital, H. Dotan, D. Klotz, D. A. Grave, A. Tsyganok, B. Gupta, S. Kolusheva, I. Visoly-Fisher, A. Rothschild, A. Yochelis, Two-site H2O2 photo-oxidation on haematite photoanodes. *Nat. Commun.* **9**, 4060 (2018).

32. U. Schwertmann, R. M. Cornell, C. B. Koch, *Iron Oxides in the Laboratory* (Wiley-VCH Verlag, Weinheim, Germany, ed. 3, 2022).

33. W. Maret, Y. Li, Coordination dynamics of zinc in proteins. *Chem. Rev.* **109**, 4682–4707 (2009).

34. S. Repp, E. Erdem, Controlling the exciton energy of zinc oxide (ZnO) quantum dots by changing the confinement conditions.

Spectrochim. Acta A Mol. Biomol. Spectrosc. **152**, 637–644 (2016).

35. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

36. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

37. R. J. de Haas, H. Pyles, T. F. Huddy, J. van Ossenbruggen, C. Zheng, D. van den Broek, A. Carr, A. K. Bera, A. Kang, E. Brackenbrough, E. Joyce, B. Sankaran, D. Baker, I. K. Voets, R. de Vries, Inhibition of ice recrystallization with designed twistless helical repeat proteins, *bioRxiv* (2025). https://doi.org/10.1101/2025.03.09.642278.

38. C. Woll, The chemistry and physics of zinc oxide surfaces. *Prog. Surf. Sci.* **82**, 55–120 (2007).

39. J. K. Leman, B. D. Weitzner, S. M. Lewis, J. Adolf-Bryfogle, N. Alam, R. F. Alford, M. Aprahamian, D. Baker, K. A. Barlow, P. Barth, B. Basanta, B. J. Bender, K. Blacklock, J. Bonet, S. E. Boyken, P. Bradley, C. Bystroff, P. Conway, S. Cooper, B. E. Correia, B. Coventry, R. Das, R. M. De Jong, F. DiMaio, L. Dsilva, R. Dunbrack, A. S. Ford, B. Frenz, D. Y. Fu, C. Geniesse, L. Goldschmidt, R. Gowthaman, J. J. Gray, D. Gront, S. Guffy, S. Horowitz, P.-S. Huang, T. Huber, T. M. Jacobs, J. R. Jeliazkov, D. K. Johnson, K. Kappel, J. Karanicolas, H. Khakzad, K. R. Khar, S. D. Khare, F. Khatib, A. Khramushin, I. C. King, R. Kleffner, B. Koepnick, T. Kortemme, G. Kuenze, B. Kuhlman, D. Kuroda, J. W. Labonte, J. K. Lai, G. Lapidoth, A. Leaver-Fay, S. Lindert, T. Linsky, N. London, J. H. Lubin, S. Lyskov, J. Maguire, L. Malmström, E. Marcos, O. Marcu, N. A. Marze, J. Meiler, R. Moretti, V. K. Mulligan, S. Nerli, C. Norn, S. Ó'Conchúir, N. Ollikainen, S. Ovchinnikov, M. S. Pacella, X. Pan, H. Park, R. E. Pavlovicz, M. Pethe, B. G. Pierce, K. B. Pilla, B. Raveh, P. D. Renfrew, S. S. R. Burman, A. Rubenstein, M. F. Sauer, A. Scheck, W. Schief, O. Schueler-Furman, Y. Sedan, A. M. Sevy, N. G. Sgourakis, L. Shi, J. B. Siegel, D.-A. Silva, S. Smith, Y. Song, A. Stein, M. Szegedy, F. D. Teets, S. B. Thyme, R. Y.-R. Wang, A. Watkins, L. Zimmerman, R. Bonneau, Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **17**, 665–680 (2020).

40. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a

three-track neural network. *Science* **373**, 871–876 (2021).

41. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

42. J. C. Klein, M. J. Lajoie, J. J. Schwartz, E.-M. Strauch, J. Nelson, D. Baker, J. Shendure, Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2016).

43. A. Chevalier, D.-A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K.-H. Lam, G. Yao, C. D. Bahl, S.-I. Miyashita, I. Goreshnik, J. T. Fuller, M. T. Koday, C. M. Jenkins, T. Colvin, L. Carter, A. Bohn, C. M. Bryan, D. A. Fernández-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I. A. Wilson, D. H. Fuller, D. Baker, Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).

44. L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam, K. H. G. Verschueren, K. Verstraete, S. T. R. Walsh, N. Bennett, A. Phal, A. Yang, L. Kozodoy, M. DeWitt, L. Picton, L. Miller, E.-M. Strauch, N. D. DeBouver, A. Pires, A. K. Bera, S. Halabiya, B. Hammerson, W. Yang, S. Bernard, L. Stewart, I. A. Wilson, H. Ruohola-Baker, J. Schlessinger, S. Lee, S. N. Savvides, K. C. Garcia, D. Baker, Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

45. C. Camacho, T. Madden, *BLAST+ Release Notes* (National Center for Biotechnology Information (US), 2023).

46. T. G. M. Schmidt, A. Skerra, The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat. Protoc.* **2**, 1528–1535 (2007).

47. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

48. W. Kabsch, XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

49. M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

50. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

51. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

52. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

53. C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall 3rd, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, D. C. Richardson, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).

54. D. N. Mastronarde, Advanced data acquisition from electron microscopes with SerialEM. *Microsc. Microanal.* **24**, 864–865 (2018).

55. A. Punjani, Algorithmic advances in single particle cryo-EM data processing using CryoSPARC. *Microsc. Microanal.* **26**, 2322–2323 (2020).

56. A. Punjani, Real-time cryo-EM structure determination. *Microsc. Microanal.* **27**, 1156–1157 (2021).

57. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

58. T. I. Croll, ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology* **74**, 519–530 (2018).

59. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

Acknowledgments

We thank F. Praetorius and J. M. Galloway for helpful discussions. We acknowledge support from the Alexandria Venture Investments Translational Investigator Fund (GF116545), the Nan Fung Life Sciences Translational Investigator Fund (GF126554), and the

Washington Research Foundation Innovation Fellows Program (GF136997). This work was funded in part by the Open Philanthropy Project through the Improving Protein Design Fund (GF129460), The Audacious Project at the Institute for Protein Design (PG117879, PG117878, PG117866), the Alfred P. Sloan Foundation (G-2021-16899, GR017147) and the Howard Hughes Medical Institute (GR020267). Design and and screening of the mineral-binding library was conducted at UW and supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, as part of the Energy Frontier Research Centers program: CSSAS – The Center for the Science of Synthesis Across Scales under Award Number DE-SC0019288. AFM measurement was supported by CSSAS through FWP 72448 at PNNL, a multiprogram national laboratory operated for DOE by Battelle under Contract No. DE-AC05-76RL01830. Crystallographic data were collected at National Synchrotron Light Source II (NSLS-II) Beamline 17-ID-1 (AMX). The Center for BioMolecular Structure is primarily supported by the NIH-NIGMS through a Center Core P30 Grant (P30GM133893) and by the Department of Energy Office of Biological and Environmental Research (KP1607011). NSLS-II is a U.S. Department of Energy Office of Science User Facility operated under contract no. DE-SC0012704. This publication is based on data collected through the NECAT BAG proposal no. 311950.

Author contributions:
Conceptualization: AS, HP, FADH, DB
Data curation: AS, HP
Formal analysis: HP
Funding acquisition: LS, JJDY, DB
Investigation: AS, HP, EBH, PSK, AJB, WD, SYS, XL, SZ, AKB, AK, EB, AC, KDC, AA, DKV, YB, NH
Methodology: AS, HP, EBH, PSK
Resources: DKV, AA, RK, LG
Software: HP, EBH, PSK
Supervision: DB, JJDY, SZ
Visualization: HP, AJB, AS
Writing – original draft: HP, AS, DB
Writing – review & editing: All authors

Competing interests: Authors declare that they have no competing interests.

Data and materials availability:
All data are available in the main text or the supplementary materials. All protein design models and scripts are available at https://github.com/pylesharley/protein_mineral_library. The crystal structures of Z4-C3i and Z4-C3ii are available in the RCSB Protein Data Bank (ids: 9D92 and 9CC4) Cryo-EM structures of the Z4-cage and the Z4-C3i trimer structure within the cage are available in the RCSB Protein Data Bank (ids: 9O59 and 9O5Q) and EMDB (ids: EMD-70146 and EMD-70128).

Code availability:
All protein design scripts are available at https://github.com/pylesharley/protein_mineral_library. The Rosetta Macromolecular Modelling suite is available for non-commercial use at https://www.rosettacommons.org.

Supplementary Materials

Materials and Methods
Figs. S1 to S32
Tables S1 to S11
References 39 to 59
Movie S1

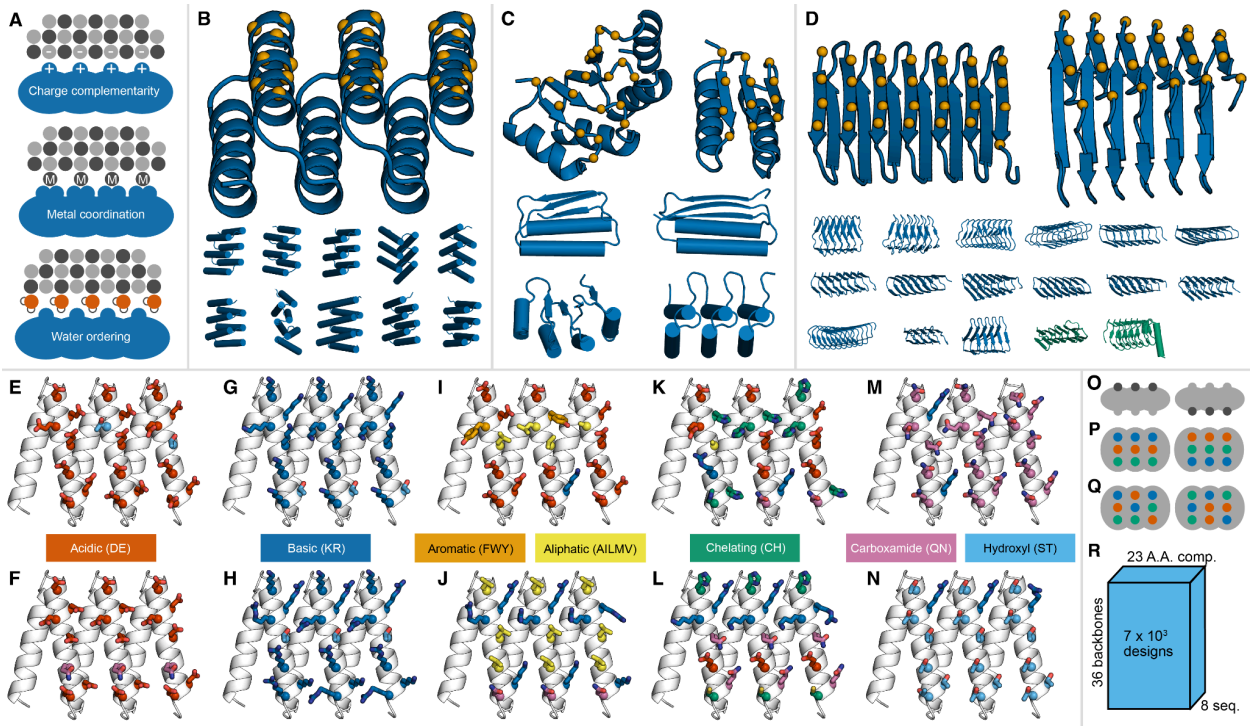


Fig. 1. Design principles for templating mineral growth. (**A**) Three hypothesized mechanisms of mineral binding: charge complementarity, metal coordination, and water ordering. (**B-D**) 36 protein backbones used in the library. Orange spheres show alpha carbons of residues in a designed interface. Scaffolds include (**B**) 11 designed helical repeats (DHR), (**C**) 6 alpha-beta topologies, and (**D**) 17 designed beta solenoids (DBS) and 2 native solenoids (in green). (**E-N**) Example designed surfaces defined by 10 (out of 23) amino acid compositions applied to the DHR backbone DHR14. The colors of the amino acid side chains correspond to chemical categories identified in rectangles. Compositions include (**E-H**) highly charged residues, (**I, J**) charged and hydrophobic moieties, (**K, L**) metal chelating residues, and (**M, N**) non-charged hydrophobic residues. (**O-Q**) Illustration of the designs generated for each combination of scaffold and composition. (**O**) Two sets of interface residues are selected on opposing surfaces of protein. For a given surface and amino acid composition, two repeat surfaces (**P**), with the same residue in analogous positions in each repeat subunit, and two non-repeat surfaces (**Q**), wherein residues are scrambled among repeats. (**R**) 4 designed interfaces for each of 2 surfaces yields 8 sequences per combination of 36 protein backbones and 23 amino acid composition, resulting in a library of approximately 7×10^3 designs.

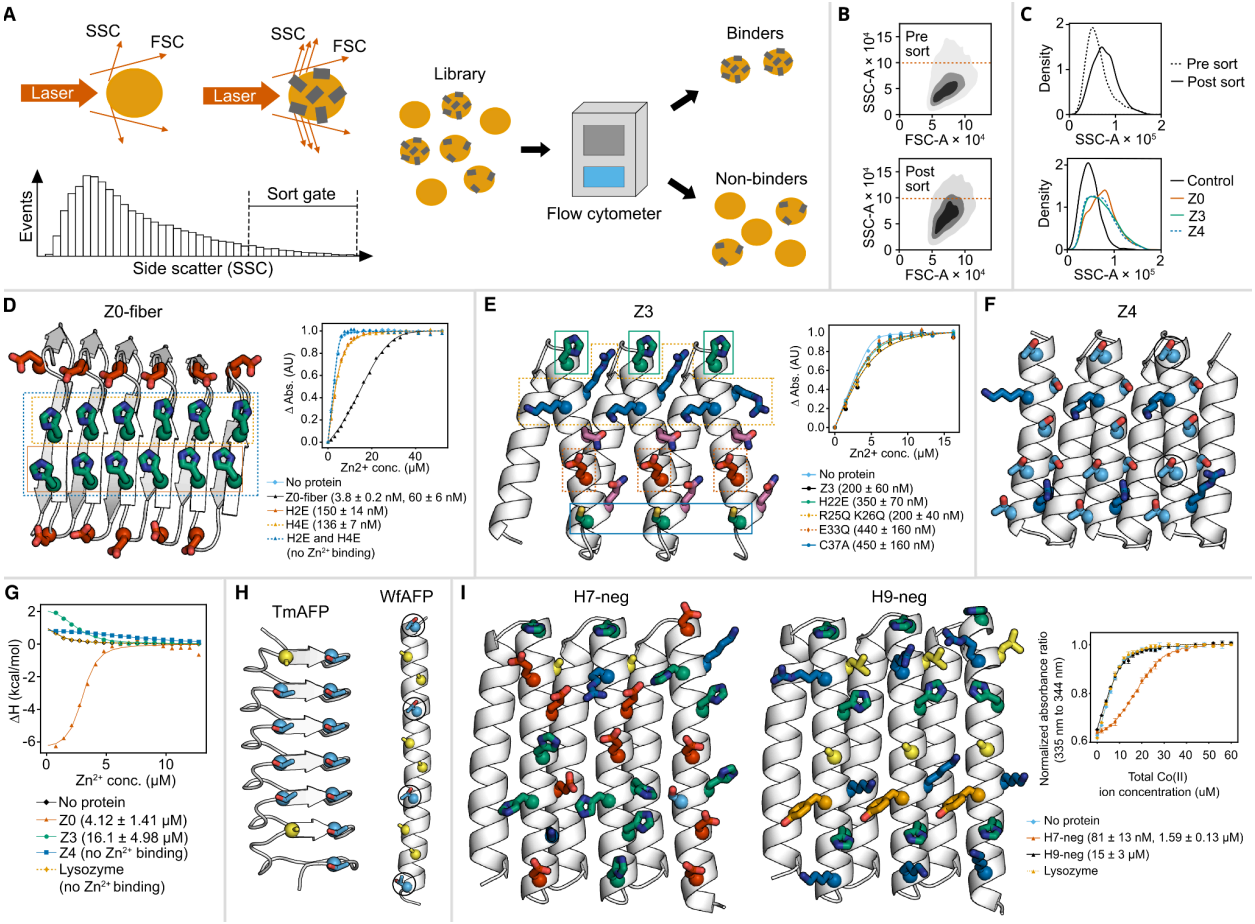


Fig. 2. Mineral binding screen. (**A**) Illustration of screening strategy. (**B**) SSC-A vs. FSC-A signals of yeast populations before and after ZnO enrichment sorting. Plots are kernel density plot estimates from 10,000 events. (**C**) SSC-A signals of a sample sublibrary before and after sorting, and yeast clones displaying three enriched designs vs. control yeast. (**D-E**) Z0-fiber and Z3 interfaces and mutational analysis of Zn^{+2} ion binding with ratiometric dye (Mag-Fura-2). Although mutants are named by the first residue that is mutated, all repeated instances of these residues, circumscribed by lines on the models, are mutated as well. (**F**) Model of the Z4 enriched design with circles highlighting threonines repeated in pattern resembling TXXXAXXXAXX motif in native antifreeze proteins (AFP). (**G**) Isothermal titration calorimetry analysis of Zn^{+2} against three designed proteins, lysozyme, and solutions without

protein (protein concentration = 30 μM). (**H**) Example native ice-binding proteins: beta solenoid Tenebrio molitor AFP (TmAFP) containing threonine-rich parallel beta-sheets (PDBid: 1ezg) and alpha-helical winter flounder AFP (WfAFP) with threonines in TXXXAXXXAXX motif circled (PDBid: 1wfa). (**I**) H7 and H9 interfaces and analysis of Co^{+2} ion binding with ratiometric dye (Mag-Fura-2). Colors of all amino-acid side chains correspond to the chemical categories defined in Figure 1.

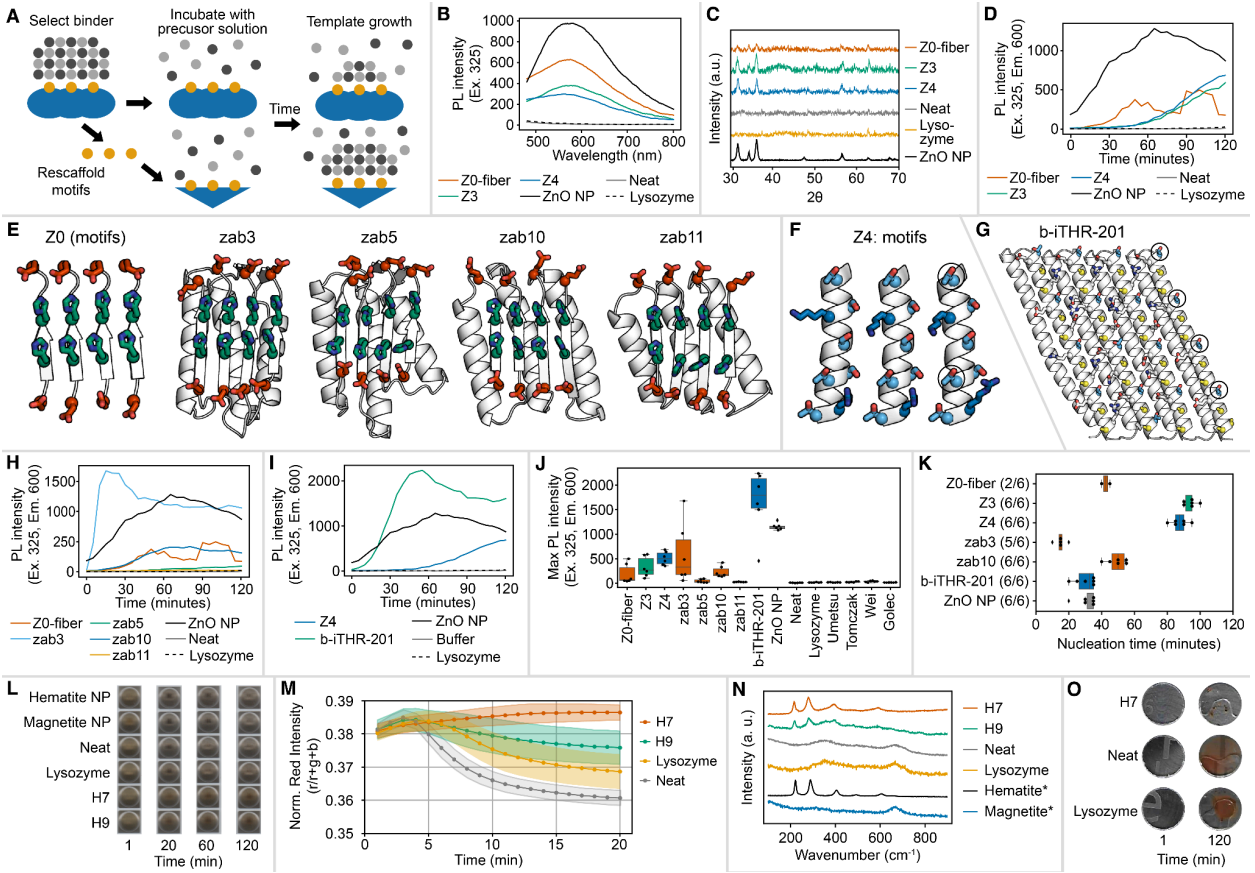


Fig. 3. Characterization of metal oxides nucleation by designed interfaces. (**A**) Schematic illustrating the nucleation assay and motif grafting methodology. (**B**) Emission spectra (Ex. = 325 nm) of solutions superstatured for ZnO seeded with selected ZnO-binding designs compared to ZnO NP, lysozyme and neat (no-additive) solution controls. (**C**) X-Ray diffraction analysis of synthesized products following two-hour incubation with designed proteins (Z0-fiber, Z3, Z4), control protein (lysozyme), or solutions without additive (neat). A reference spectra taken of ZnO nanoparticles (ZnO NP) is also provided. (**D**) Plot of PL emission over time (Ex. 325, Em. 600) of supersaturated solutions containing selected designs, lysozyme, ZnO NP, or neat solutions. (**E**) Models of interface motifs taken from Z0 and four Z0-grafted alpha-beta (zab) designs. (**F**) Models of the Z4 interface motifs with threonine side chains resembling TXXXAXXXAXX motifs observed in native α-helical ice-binding proteins circled in one repeat. (**G**) Designed Ice-binding Twistless Helical Repeat (iTHR) with threonines in TXXXAXXXAXX motifs circled in one repeat. Colors of amino acid

side chains in all panels correspond to the chemical categories defined in Figure 1. (**H-I**) PL emission over time (Ex. 325, Em. 600) of the zab redesigns vs. Z0-fiber and controls (**H**) and b-iTHR-201 vs. Z4 and controls (**I**). (**J-K**) Box plots showing 6 replicates of growth experiments using designed proteins, controls, and four previously reported ZnO-binding peptides, referred to by the last names of the first authors of the source publications (Umetsu, Tomczak, Wei, and Golec)(*7–10*). (**J**) Average maximum PL intensity (n=6). (**K**) Nucleation times, defined as the halfway point between the initial and maximum PL signal, for the subset of samples that showed a PL signal of nucleation (maximum PL > 100), n values as indicated. All ZnO growth solutions contain 3 mM $ZnNO_3$, 50 mM NaCl, 100 mM HEPES pH 8.2, and contain 0.1 mg/mL protein or 13 μg/mL ZnO NP as indicated. (**L**) Digital scans taken of iron oxide growth reactions over time with protein (0.15 mg /mL) or NP ($1.6x10^{-5}$ g/ml Hematite NP; $1.5x10^{-5}$ g/ml Magnetite NP) additives as indicated. Solutions contain ammonium iron(II) sulfate hexahydrate 0.05 mM, Iron(III) nitrate nonahydrate 0.1 mM, KNO_3 50 mM, CHES 100 mM. Normalized red-intensity of the reactions shown in panel **L**. Values shown are the mean red intensity divided by the sum of the mean read, mean green, and mean blue intensity within the area of each well (n=6). (**N**) Raman spectroscopy of the products of the H7-neg, H9-neg, lysozyme, and neat reactions shown in panels **L** and **N,** with spectra taken of reference hematite and magnetite nanoparticles for comparison. (**O**) Images from iron substrate scans prior and following two hours incubation with 10ul samples containing H7-neg (0.6 mg/ml), Lysozyme (0.6 mg/ml), or buffer only.

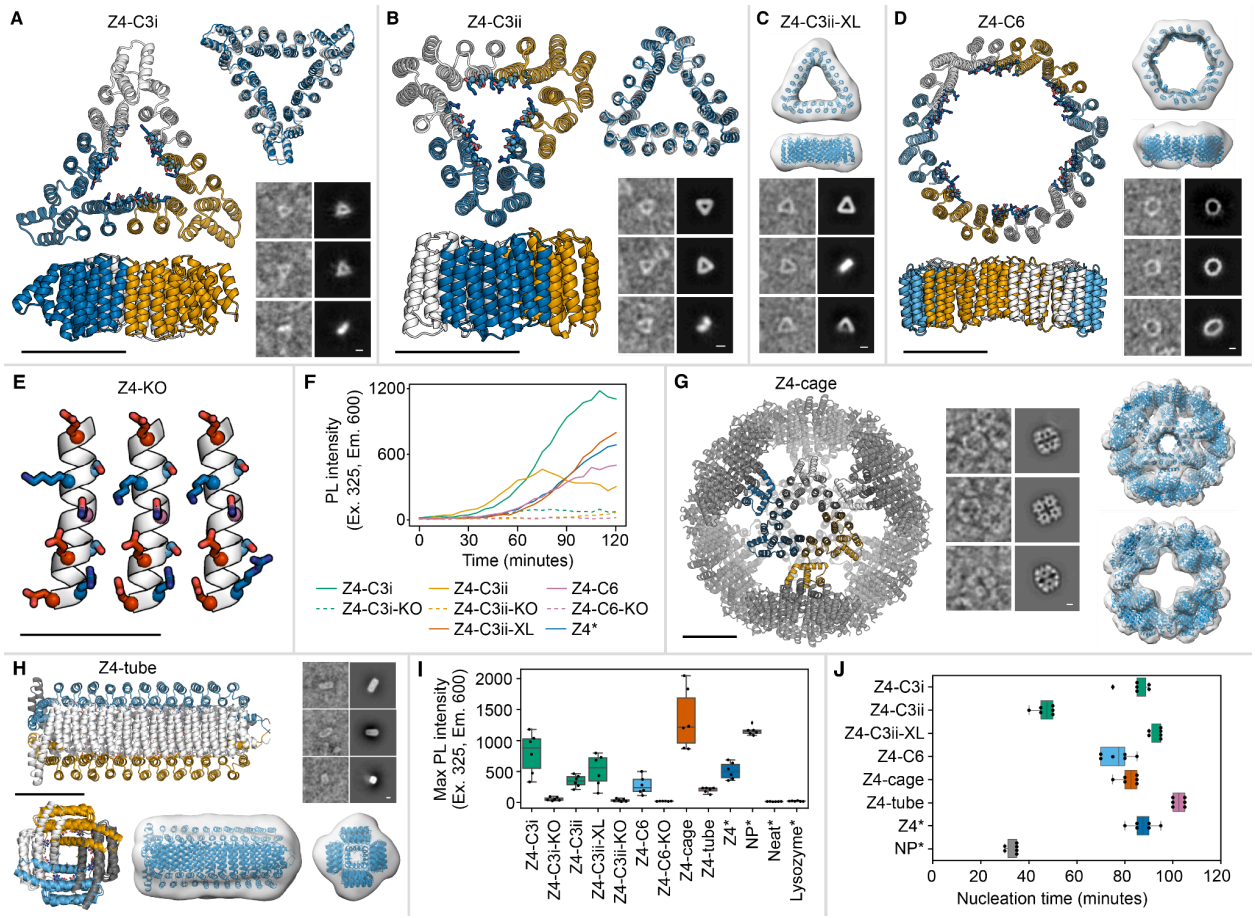


Fig. 4. ZnO nucleation by Z4 motif-containing designed oligomers. (**A-D**) Cyclic oligomers with Z4 motifs lining interior cavities. (**A-B**) Two C3 designs with design models colored by subunit with Z4 motifs shown as sticks. Design model backbones shown in white superimposed with X-ray crystal structures in blue (PDB IDs: 9d92, 9cc4). Images at bottom right are representative ns-TEM picked particles (left) and 2D class averages (right). (**C**) Z4-C3i-XL model superimposed with ns-TEM 3D reconstruction, and examples of picked particles and 2D classes. (**D**) Z4-C6 colored by subunit with motifs shown as sticks, model superimposed with ns-TEM 3D reconstruction, and examples of picked particles and 2D classes. (**E**) Model of the mutated Z4 knock out (Z4-KO) interface with threonines replaced with glutamates. (**F**) Plot of PL emission over time (Ex. 325, Em. 600) of Z4-motif containing cyclic oligomers compared to oligomers with interface KO mutants and the Z4 monomer (*sample replicated from Figure 3). (**G**)

Octahedral cage designed using Z4-C3i, ns-TEM picked particles and 2D class averages, and model superimposed with ns-TEM 3D reconstruction. (**H**) Z4-tube colored by subunit with Z4 motifs shown as sticks, ns-TEM picked particles, 2D-averages, and 3D reconstructions superimposed with models. (**I-J**) Effect of Z4-motif containing oligomers on ZnO nucleation compared to Z4 monomer, ZnO NP, and neat solutions (*samples are replicated from Figure 3). (**I**) Maximum PL intensity after nucleation assay (n=6). (**J**) Nucleation times, defined as the halfway point between the initial and maximum PL signal (n=6), for samples that showed a signal of nucleation (maximum PL > 100). Colors of amino acid side chains correspond to the chemical categories defined in Figure 1. All scale bars are 4 nm.

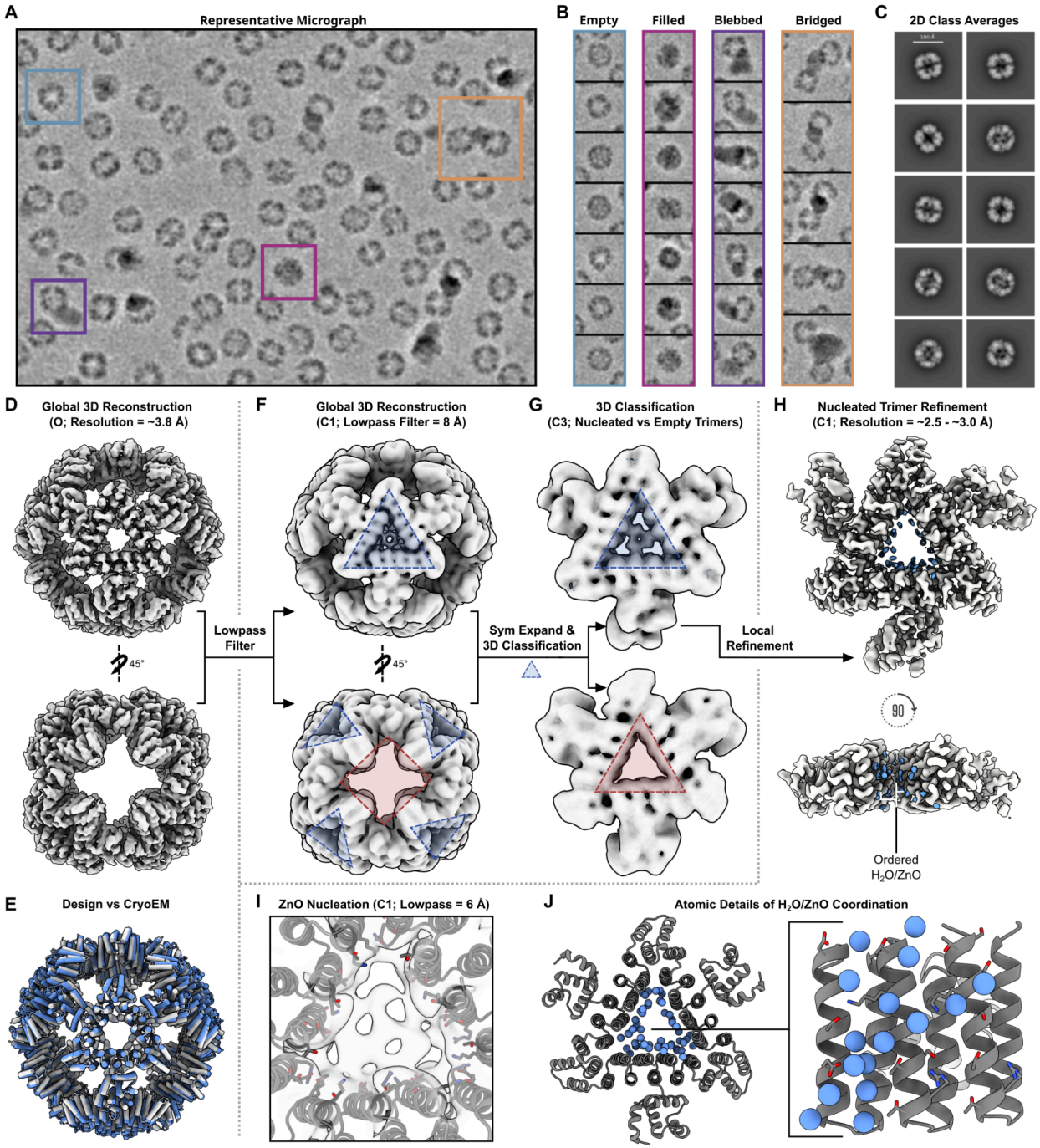

A
Representative Micrograph
B
Empty
Filled
Blebbed
Bridged
C
2D Class Averages
D
Global 3D Reconstruction
(O; Resolution = ~3.8 Å)
45°
Lowpass
Filter
F
Global 3D Reconstruction
(C1; Lowpass Filter = 8 Å)
45°
G
3D Classification
(C3; Nucleated vs Empty Trimers)
Sym Expand &
3D Classification
Local
Refinement
H
Nucleated Trimer Refinement
(C1; Resolution = ~2.5 - ~3.0 Å)
90
Ordered
H2O/ZnO
E
Design vs CryoEM
I
ZnO Nucleation (C1; Lowpass = 6 Å)
J
Atomic Details of H2O/ZnO Coordination

Fig. 5. Cryo-EM characterization of ZnO mineralized Z4-cage.
(A) Representative denoised micrograph of Z4-cage particles imaged in the presence of 3 mM $Zn(NO_3)_2$ for ~30 minutes prior to freezing. Highlighted are examples of empty (teal), "filled" (maroon), "blebbed" (purple), and "bridged" (orange) nanoparticle morphologies observed in a single micrograph. **(B)** Additional examples of each morphology observed in the dataset. **(C)** 2D class averages of Z4-cage. **(D)** 3.8 Å cryo-EM 3D reconstruction of Z4-cage refined with relaxed octahedral symmetry. **(E)** Comparison of the final built cryo-EM model of Z4-cage (grey) with the computational design model (blue). **(F)** C1 reconstruction of Z4-cage, low-pass filtered to 8 Å, showing ZnO nucleation localized along the C3 symmetry axis (blue), and no ZnO density along the C4 axis (red). **(G)** Octahedral symmetry expansion and 3D classification of all Z4-C3i trimers from all Z4-cage nanoparticles show both nucleated (blue) and empty (red) trimers exist within the data. **(H)** Near-atomic resolution 3D reconstruction of nucleated Z4-C3i trimers in the Z4-cage nanoparticle, highlighting density from coordinated H_2O or ions (blue) near the designed ice-binding motif. **(I)** Low-pass filtered view of (H), zoomed in on the C3 axis, illustrating stochastic ZnO growth within the Z4-C3i trimer. **(J)** Placement of atoms into the density (see Methods) suggests that initial H_2O/Zn ion coordination and nucleation occur along the designed Z4 ice-binding motif (sticks).

Chapter 3: Templating and confining calcium phosphate mineralization within designed protein assemblies

This chapter contains additional background for (and limited reproduction of) content submitted as: Le Tracy Yu, Xinqi Li, Harley Pyles, Anjali P. Patni, Andrew J. Borst, Neville P. Bethel, Paul S. Kwon, Connor Weidle, Ryan D. Kibler, Kenneth D. Carr, Yulai Liu, Dongsheng Li, Stanislav Moroz, Hannele Ruohola-Baker, Shuai Zhang, James De Yoreo, David Baker. *Science* (in revision).

Abstract: Bone formation involves the deposition of ordered hydroxyapatite (HAp) on collagen fibrils, but the underlying molecular mechanisms remain largely unresolved, limiting the design of protein-apatite hybrid materials. Here we show that computationally designed de novo proteins can template and confine HAp mineralization. We design C_3-symmetric oligomers with inner surfaces chemically complementary to the HAp {010} facet, and assemble these through additional interfaces into D_3 oligomers, one-dimensional nanotubes, and two-dimensional arrays. Electron microscopy revealed templated mineralization and confinement at each hierarchical level, with mineral shape guided by the underlying protein architecture. Conversion of an initially formed amorphous phase to HAp is driven by the designed protein-mineral interface. Our results establish a framework for programmable protein-guided mineralization, providing a foundation for next-generation hybrid materials.

Main Text:

Bone formation requires biomineralization of calcium phosphate within a collagen matrix(*1*–*3*), a process regulated by non-collagenous proteins that promote ordered HAp crystal formation on collagen fibrils.(*4*) However, the molecular mechanisms underlying this process remain largely unresolved due to the structural complexity of the extracellular matrix. Biomimetic mineralization systems have been developed to investigate the physicochemical principles governing mineral nucleation and organization. Studies of peptide(*5*, *6*), protein(*7*–*12*), DNA(*13*, *14*) and polymer(*15*, *16*) model systems have shown that charged residues play critical roles in directing mineralization. While these studies have provided insight into the role of charge, and simple flat negatively charged interfaces have been shown to promote calcium carbonate nucleation (*17*), confined mineral growth within specific locations of a biomolecular assembly has not yet been achieved.

We reasoned that the control provided by computational protein design(*18*), which enables the construction of well-defined protein architectures with programmable surface chemistry, could provide a solution to the challenge of templating and confining mineral growth. Recent advances in deep learning based protein design methods such as RFdiffusion(*19*–*21*), ProteinMPNN and related methods (*22*, *23*) now allow the creation of proteins with precisely defined shapes and side-chain chemistries (*24*–*29*). We set out to these methods to design protein subunits with interfaces

chemically and geometrically complementary to HAp lattice planes, and to explore the capability of these proteins to direct and confine HAp mineralization within a range of hierarchical assemblies.

Design strategy

Hydroxyapatite – $Ca_{10}(PO_4)_6(OH)_2$ – is the most stable phase of calcium phosphate in physiological conditions and has a hexagonal crystal structure (space group $P6_3/m$)(*30*). We built a model of a HAp crystal with a diameter of 3.8 nm as the design target. The model presents {001} facets along the c-axis and terminates at {010} facets along the a and b axes (Fig. 1A), which exposes calcium ions(*31*) to be templated using carboxylate sidechains (Fig. 1A). This HAp crystal has C_3 rotational symmetry, and we set out to design trimeric protein oligomers with matching C_3 symmetry with inner surfaces optimized to nucleate the target crystal. We reasoned that such HAp templating trimers could serve as building blocks for hierarchical architectures, enabling assembly into D_3 oligomers, one-dimensional nanotubes and two-dimensional arrays (Fig. 1B). Each assembly would present spatially arranged aspartate and glutamate residues to promote HAp formation at specific locations, giving rise to ordered protein-HAp composite materials.

Carboxylate residues were placed at coordinates matching the phosphate groups surrounding the outermost calcium ions of the target surfaces, mimicking their calcium-coordinating geometry (Fig. 1A, step 1), and we used RFdiffusion2(*21*) (RFD2) to generate protein backbone segments hosting those carboxylate positions in aspartate or glutamate side chains.(Fig. 1A, step 2). These segments were then propagated to equivalent positions; in total three additional copies were generated, each sequentially rotated by 60° and translated by 10.32 Å along the c-axis—the spacing between neighboring calcium ions in the HAp crystal (Fig. 1A, step 3). We connected these four segments using Ca-RFdiffusion(*32*) to generate the monomeric subunit and then applied C_3 symmetry around the crystal c-axis to generate a C_3 oligomer (Fig. 1A, steps 4-5). LigandMPNN(*22*) was used to design sequences that promote folding to the target structure while keeping the HAp interacting carboxylate residues fixed (Fig. 1A, Step 6). Designed trimers whose AlphaFold2(*33*) predicted structures deviated by less than 3 Å C_α Root Mean Square Deviation (RMSD) from the designed structures (Fig. 1A, Step 7) were selected for experimental characterization.

C_3 oligomer characterization and mineralization

73 designed C_3 trimers were expressed in *Escherichia coli* and purified (*34, see methods*). Of these, 42 were produced at high levels, and six folded into the intended oligomeric structure as determined by mass spectroscopy, Size Exclusion Chromatography (SEC), Circular Dichroism (CD), and negative stain transmission electron microscopy (nsEM) (Fig S1-S14, Table S1-S2).

To evaluate mineralization activity, we mixed the oligomers with a supersaturated $CaCl_2/K_2HPO_4$ solution. Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy (ATR-FTIR) of the HAp1_C_3 sample showed a strong signal at ~1045 cm^{-1} and a weak signal at ~1095 cm^{-1} (v_3 PO_4^{3-}) indicative of HAp formation(*35*) (Fig. S15A). This phosphate signal overlapped with that of commercial HAp nanoparticles, whereas control reactions lacking protein showed no such signal (Fig. S15A). The reactions with the other five C_3 oligomers produced broader phosphate stretching signals with peak splitting in the 1000-1200 cm^{-1} region, indicating less crystalline or mixed calcium phosphate phases (Fig. S15B).

We further characterized mineralization with electron microscopy. Prior to mineralization, HAp1_C_3 has a hollow ring structure matching the design model with a diameter of ~9 nm (Fig. 2A). After mineralization, additional density in the middle of the ring is evident (Fig. 2B). We characterized the mineralized sample under transmission electron microscopy without negative stain (TEM) and observed ~3.5 nm nanoparticles (Fig. 2C). High resolution transmission electron microscopy (HRTEM) revealed lattice fringes with d-spacings of ~2.73 Å and mutual angles of ~60°, matching the (300)-type reflections of HAp, consistent with the top-view orientation of the oligomers observed on the grids (Fig 2B), and the targeting {010} plane presented parallel to the protein-mineral interface. TEM screening of the other five C_3 oligomers showed no mineral confinement (Fig. S16); thus, HAp1_C_3 was selected for downstream investigations.

HAp1_C_3 has 4 repeats per monomer and each repeat has one aspartate and three glutamate residues designed to coordinate HAp calcium ions (Fig. 2F-G). To test their contribution to nucleation, five knockout mutants were generated with one or more carboxylates per repeat replaced with uncharged asparagine or glutamine (Fig. 2E and Table S1). All mutants retained the C_3 oligomeric structure by SEC (Fig. S17A). ATR-FTIR analysis of nucleation reactions showed similar amide I (1600-1700 cm^{-1}) and amide II (1500-1600 cm^{-1}) signals in all reactions containing protein, whereas the protein-free control showed no such signals. Only the original design produced the strong phosphate stretching at ~1063 cm^{-1} and a weaker signal at ~1092 cm^{-1} (Fig. S17B, C), characteristic of biological apatite(*36*), demonstrating that the designed carboxylate residues are essential for mineralization.

D_3 oligomer design, characterization and mineralization

We next converted the HAp1_C_3 oligomers into a D_3 assemblies as a stepping stone to extended nanotubes and 2D arrays. To generate D_3 assemblies, we stacked two C_3 oligomers by arranging either their N-termini or C-termini against each other with C_2 symmetry (Fig. 2G). We then used RFdiffusion(*20*) to extend the backbone under symmetry constraints to bridge the inter-ring gap, generating new D_3 interfaces. The amino acid sequences at the new D_3 Interface and adjacent regions of the structure were optimized by proteinMPNN(*23*), without consideration of HAp lattice matching.

Sequences predicted by AlphaFold2(*33*) and AlphaFold3(*37*) to form D_3 oligomers were selected for experimental characterization.

We experimentally screened 22 N-terminally stacked designs and 26 C-terminally stacked designs of D_3 oligomers. Among them, designs HAp1_D_3N and HAp1_D_3C were chosen for more detailed characterization based on SEC (Fig. S18-19), MP (Fig. S12C-F), and nsEM (Fig. 2H-L, and Fig S20-21) results indicating the correct oligomeric assembly, alpha-helical secondary structure, and hollow ring morphology with shape matching the design.

We next tested the nucleation activity of HAp1_D_3N and HAp1_D_3C. Prior to mineralization, the designs present as hollow rings under nsEM (Fig. 2H and K). After mineralization, we observed black dots under TEM for both D_3 oligomers (Fig. S22A and C). Staining the mineralized samples revealed filled rings (Fig. 2J and M; Fig. S22B and D), consistent with mineral occupying the protein cavity. HRTEM of the mineralized designs revealed crystalline lattice fringes matching with the HAp (040) lattice plane (Fig. S22E-F).

Protein nanotube design and characterization

Unbounded protein nanotube architectures were designed by combining the N- and C-termini interfaces of HAp1_D_3N and HAp1_D_3C respectively (Fig. 3A) to generate HAp1_Nt (Fig. S23). Cryogenic transmission electron microscopy (cryoEM) micrographs of HAp1_Nt and 2D class averages showed extended nanotube structures with clearly defined secondary structure elements (Fig. 3B). 3D reconstruction of these particles generated a 3.08 Å cryoEM density map with well-resolved structural details, enabling unambiguous assignment of side-chain rotamers. Consistent with the design model, the density for HAp1_Nt represented a hollow nanotube structure containing 10-25 Å periodic pores which were incorporated during computational design to facilitate molecular transport of mineral precursors into the nanotubes (Fig. 3C and Fig. S24). The resulting atomic model of HAp1_Nt (Fig. 3D) is in close agreement with the design model, (Fig. 3E) with a total monomer Cα RMSD of ~1.1 Å (Fig. 3F and Table S3).

The atomic model confirms the presence of four carboxylate-rich segments per monomer (Fig. 3F). The distances between carboxylate groups are 6.2–7.2 Å in one direction and 8.7–9.7 Å in the perpendicular direction, matching the spacing between calcium ions in the HAp {100}/{010}-type crystal lattice along the c-axis as 6.88 Å and a/b-axis as 9.42 Å, respectively (Fig. 3G). Each successive segment is related by ~60° and shifted 10.32 Å along the nanotube axis (1.5× the HAp c-axis repeat), matching the helical symmetry of HAp. The three-fold arrangement of monomers in the nanotube mirrors the hexagonal geometry of the HAp crystal face.

Nanotube mineralization

We mixed HAp1_Nt with $CaCl_2$ and K_2HPO_4 and observed predominantly extrafibrillar mineralization (*assay 1*) (Fig. 4A vs Fig. S25A), with nanotubes exhibiting polycrystalline features (Fig. S25B). Polymer-induced liquid phase precursors have been demonstrated to be critical for enhancing collagen intrafibrillar mineralization(*4*, *38–41*), we therefore prepared amorphous calcium phosphate (ACP) precursors by mixing $CaCl_2$/K_2HPO_4 with polyaspartic acid (pAsp) (*assay 2*). After incubating the pAsp-ACP precursor with HAp1_Nt, we observed electron-dense filaments with periodic banding (Fig. 4B). Staining of mineralized HAp1_Nt revealed strongly contrasting density within the nanotube lumen (Fig. 4C). HRTEM of intrafibrillarly mineralized nanotubes revealed HAp nanocrystals exposing (002) facets perpendicular to the designed {010} nucleation plane (Fig. 4D and Fig. S25C), matching the dimensions of HAp nanocrystals confined in the nanotube lumen and the orientation of the nanotubes which lie perpendicular to the beam on the TEM grids. The crystallinity is not uniform with amorphous regions coexisting along the nanotube axis. Precursor-only control showed no crystallinity (Fig. S25D). Volume density maps of mineralized HAp1_Nt with pAsp (unstained TEM) revealed coherent density of ~40 Å diameter, matching the dimension of designed protein cavity, consistent with mineral filling the nanotube interior while the overall map diameter expands to 107 Å relative to the unmineralized protein map diameter (92 Å), showing heterogeneous features coating the nanotube exterior surface (Fig. 4E). Combined with HRTEM (Fig. 4D), these reconstructions (Fig. 4E) suggest minerals occupying the nanotube lumen and nonspecific precursors associating with the exterior of the nanotube.

Scanning Transmission Electron Microscopy - Energy Dispersive X-Ray Spectroscopy (STEM-EDX) of mineralized HAp1_Nt revealed co-localization of calcium and phosphorus signals within nitrogen-rich protein nanotube regions (Fig. S26). Cryo-EM of mineralized HAp1_Nt showed increased filament electron density relative to unmineralized controls (Fig. S27 vs. Fig. 3B), consistent with mineral association along the scaffold. Heterogeneous mineral and precursor distribution precluded 3D reconstruction.

We next studied the kinetics of intrafibrillar mineralization with HAp1_Nt by examining the reactions at earlier timepoints. After 1 hour, TEM revealed nanoparticle clusters (Fig. 4F) that resemble the morphology of ACP nanoparticles observed in the literature(*38*), and the ATR-FTIR spectrum showed a broad peak at 1020–1100 cm^{-1}, confirming ACP formation(*42*) (Fig. 4I and Fig. S28A). After 3 hours, filamentous structures decorated with dark dots appeared (Fig. 4G), suggesting progressive adsorption of ACP onto the nanotubes and continued with infiltration (Fig. 4J). The ATR-FTIR spectrum showed multiple split bands at 1040-1056 cm^{-1} (v_1 PO_4), 1080 cm^{-1} (v_3 PO_4) and 990 cm^{-1} (HPO_4^{2-}) (Fig. 4I), characteristic of dicalcium phosphate dihydrate (DCPD or brushite) (*42–44*), suggesting conversion of ACP to DCPD. By 16 hours, the surface-bound particles decreased, leaving filamentous features (Fig. 4H), likely corresponding to further infiltration of

precursors into the nanotube interiors (Fig. 4J). The ATR-FTIR spectrum (Fig. 4I) showed that the initially split PO_4 bands evolved into two dominant peaks at ~1044 and ~1029 cm^{-1}, characteristic of poorly crystalline HAp.(*35*) After 24 hours, the filaments became thinner, consistent with densification of the mineral phase by dehydration of DCPD and HAp nucleation (Fig. 4B, D and Fig. 4J). The ATR-FTIR spectrum showed that the major peak at ~1040 cm^{-1} intensified (Fig. 4I), confirming crystalline HAp formation(*35*). Negative controls lacking protein showed no HAp formation (Fig. S28B and S29).

We investigated what aspects of HAp formation are controlled by the designed interface by characterizing the interface knockout mutants (Table S1). While several of the mutations interfered with fiber formation, HAp1_Nt_DtoN formed filaments similar to the original design (Fig. S30A). ATR-FTIR results showed multiply split v_3 PO_4 bands matching β-tricalcium phosphate (β-TCP)(*45*) with broad ACP bands at 1 and 3 hours, and better-resolved but phase-stalled β-TCP bands at 16 and 24 hours (Fig. S30 B and C). TEM showed low-electron density filamentous features after 16 and 24 hours of reaction but without mineral confinement in the nanotube (Fig. S30 D-G). These results suggest that although the HAp1_Nt_DtoN can template precursors, it fails to direct phase conversion to HAp or achieve spatial confinement of minerals, suggesting that the designed interface enhances the late-stage HAp conversion.

Liquid-phase atomic force microscopy (LP-AFM) was used to characterize the mechanical properties of HAp1_Nt before and after mineralization. Topographic imaging confirmed that both unmineralized and mineralized samples exhibit nanotubular structures, consistent with the TEM observations (Fig. 5A, B). Young's modulus mapping indicated moduli of approximately 5.1 MPa and 16.8 MPa for unmineralized and mineralized nanotubes (Fig. 5C and Fig. S31A, B). A relatively large standard deviation for the mineralized sample likely reflects the coexistence of mineralized and unmineralized regions within individual nanotubes. The ~3-fold increase in Young's modulus upon mineralization is consistent with HAp mineral reinforcement of the protein scaffold (Fig. S31C-E).

To evaluate biological relevance, we tested HAp1_Nt-driven mineralization in human induced pluripotent stem cell (hiPSC)-derived dental organoids differentiated to the induced secretory-stage maturation ameloblast (ismAM) (Fig 5D); in these organoids the apical side of the ameloblasts, which secrete the proteins inducing mineralization during normal tooth formation, surround a lumen like region in the organoids. We investigated mineralization in the presence of HAp1_Nt, calcein, and calcium phosphate. Compared to untreated controls (Fig. 5E), HAp1_Nt-treated organoids showed mineral deposition at the apical side of the epithelial ameloblasts (Fig. 5F), a site reflecting native enamel protein secretion and mineralization geometry (*46*). This was accompanied by a significant increase in mineralized foci per lumen of the epithelial layer ($p < 0.0001$, Fig. 5I) and Alizarin Red S (ARS) staining intensity ($p < 0.0001$, Fig. S32A), while lumen number per organoid remained unchanged (Fig. 5H), confirming that HAp1_Nt promotes spatially directed mineralization

without disrupting organoid architecture. ATR-FTIR of mineralized HAp1_Nt-treated organoids revealed a v_3 PO_4 band absent in untreated controls, consistent with calcium phosphate mineral deposition (Fig. S32B). The HAp1_Nt_DtoN mutant (Fig. 5G) showed no detectable calcification, demonstrating that mineralization-associated maturation depends on the designed nucleation interface rather than nonspecific protein effects.

2D array design, characterization and mineralization

We next explored lateral extension of the D_3 oligomer to form a two-dimensional (2D) protein lattice. We positioned two HAp1_D_3N oligomers side by side, thereby creating a D_2 symmetry between the two oligomers (Fig. 6A)(*47*). We then used RFdiffusion(*20*) to generate inter-oligomer contacts (Fig. 6B) and ProteinMPNN(*23*) to design amino acid sequences compatible with the new backbones (Fig. 6C). The resulting D_2 interface features hydrophobic interactions near the N-termini and salt bridges around the central contact region; geometric propagation generates the higher-order assembly (Fig. 6D). We screened 66 array designs among which 44 were soluble and were characterized with nsEM. HAp1_Ar had the most extended lattice pattern matching the design (Fig. 6E). The sharp peaks on the binned Fast Fourier Transform (FFT) confirm formation of ordered extended assemblies (Fig. 6E).

We explored the mineralization activity of the HAp1_Ar array (assay 2). After 24 hrs reaction, we observed formation of nanocrystals matching the lattice under TEM (Fig. 6F). nsEM of the mineralized array showed density inside the D_3 oligomers (Fig. 6G), suggesting templated nucleation; in contrast no particles were observed in the large hexameric regions between the D3 oligomers. As the amino acid sequence was designed to promote mineralization on the inside of the D3 assemblies, but not in the hexameric regions between them, these results provide a further demonstration of sequence specific templating of HAp formation. ATR-FTIR of mineralized HAp1_Ar revealed a strong v_3 PO_4 band at ~1020 cm^{-1} characteristic of HAp(*35*), which was absent in protein-free controls (Fig. S33).

Discussion

Our results demonstrate that HAp-templating interfaces can be computationally designed and used to both template and confine HAp growth at specific locations within designed protein assemblies. This ability to confine mineral growth goes beyond previous efforts with calcium carbonate, where designed proteins templated but failed to confine growth, resulting in large unbounded crystals with little memory of the shape of the nucleating protein(*17*). Structural insight into this confinement is provided by the cryo-EM structure of the apo HAp1_Nt, which shows that sidechain positions at the designed interface geometrically match the HAp crystal lattice. Mutation of the designed mineral templating sidechains at this interface abolishes the phase conversion from precursors to HAp, confirming that geometric positioning of functional groups is critical for mineralization.

The mineralization observed in our synthetic systems parallels that of collagen *in vitro* mineralization in key respects. Similar to collagen, we observed extrafibrillar and enhanced intrafibrillar mineralization of HAp1_Nt without and with polymer-induced precursors, respectively(*38*) as well as a heterogeneous, multistage nucleation process(*2*, *48*). While positively charged regions in collagen recruit negatively charged ACP clusters(*49*), our designed nanotubes feature negatively charged interfaces to recruit Ca^{2+} ions that could subsequently attract phosphate-rich ACP clusters. Although differences in scaffold dimensions preclude direct kinetic comparison, the earlier appearance of HAp in HAp1_Nt nanotubes relative to collagen fibrils(*49*) suggests the designed nucleation surface is at least comparably effective at templating mineral conversion.

References and Notes

1. M. J. Olszta, X. Cheng, S. S. Jee, R. Kumar, Y.-Y. Kim, M. J. Kaufman, E. P. Douglas, L. B. Gower, Bone structure and formation: A new perspective. *Mater. Sci. Eng. R Rep.* **58**, 77–116 (2007).

2. M.-H. Hong, J. H. Lee, H. S. Jung, H. Shin, H. Shin, Biomineralization of bone tissue: calcium phosphate-based inorganics in collagen fibrillar organic matrices. *Biomater. Res.* **26**, 42 (2022).

3. Y. Xu, F. Nudelman, E. D. Eren, M. J. M. Wirix, B. Cantaert, W. H. Nijhuis, D. Hermida-Merino, G. Portale, P. H. H. Bomans, C. Ottmann, H. Friedrich, W. Bras, A. Akiva, J. P. R. O. Orgel, F. C. Meldrum, N. Sommerdijk, Intermolecular channels direct crystal orientation in mineralized collagen. *Nat. Commun.* **11**, 5068 (2020).

4. S. Shan, Z. Tang, K. Sun, W. Jin, H. Pan, R. Tang, W. Yin, Z. Xie, Z. Chen, C. Shao, ACP-Mediated Phase Transformation for Collagen Mineralization: A New Understanding of the Mechanism. *Adv. Healthc. Mater.* **13**, 2302418 (2024).

5. J. D. Hartgerink, E. Beniash, S. I. Stupp, Self-Assembly and Mineralization of Peptide-Amphiphile Nanofibers. *Science* **294**, 1684–1688 (2001).

6. S. Wang, Y. Yang, R. Wang, X. Kong, X. Wang, Mineralization of calcium phosphate controlled by biomimetic self-assembled peptide monolayers via surface electrostatic potentials. *Bioact. Mater.* **5**, 387–397 (2020).

7. S. Akkineni, C. Zhu, J. Chen, M. Song, S. E. Hoff, J. Bonde, J. Tao, H. Heinz, S. Habelitz, J. J. De Yoreo, Amyloid-like amelogenin nanoribbons template mineralization via a low-energy interface of ion binding sites. *Proc. Natl. Acad. Sci.* **119**, e2106965119 (2022).

8. P.-A. Fang, J. F. Conway, H. C. Margolis, J. P. Simmer, E. Beniash, Hierarchical self-assembly of amelogenin and the regulation of biomineralization at the nanoscale. *Proc. Natl. Acad. Sci.* **108**, 14097–14102 (2011).

9. Y. Liu, D. Luo, X.-X. Kou, X.-D. Wang, F. R. Tay, Y.-L. Sha, Y.-H. Gan, Y.-H. Zhou, Hierarchical Intrafibrillar Nanocarbonated Apatite Assembly Improves the Nanomechanics and Cytocompatibility of Mineralized Collagen. *Adv. Funct. Mater.* **23**, 1404–1411 (2013).

10. A. Hasan, A. Chuvilin, A. Van Teijlingen, H. Rouco, C. Parmenter, F. Venturi, M. Fay, G. Greco, N. M. Pugno, J. Ruben, C. J. C. Edwards-Gayle, B. Myers, I. Dreveny, N. Cowieson, A. Winter, S. Gamea, X. F. Walboomers, T. Hussain, J. C. Rodríguez-Cabello, F. Rawson, T. Tuttle, S. Elsharkawy, A. Banerjee, S. Habelitz, A. Mata, Biomimetic supramolecular protein matrix restores structure and properties of human dental enamel. *Nat. Commun.* **16**, 9434 (2025).

11. Q. Song, K. Jiao, L. Tonggu, L. G. Wang, S. L. Zhang, Y. D. Yang, L. Zhang, J. H. Bian, D. X. Hao, C. Y. Wang, Y. X. Ma, D. D. Arola, L. Breschi, J. H. Chen, F. R. Tay, L. N. Niu, Contribution of biomimetic collagen-ligand interaction to intrafibrillar mineralization. *Sci. Adv.* **5**, eaav9075 (2019).

12. Y. Wang, T. Azaïs, M. Robin, A. Vallée, C. Catania, P. Legriel, G. Pehau-Arnaudet, F. Babonneau, M.-M. Giraud-Guille, N. Nassif, The predominant role of collagen in the nucleation, growth, structure and orientation of bone apatite. *Nat. Mater.* **11**, 724–733 (2012).

13. X. Liu, X. Jing, P. Liu, M. Pan, Z. Liu, X. Dai, J. Lin, Q. Li, F. Wang, S. Yang, L. Wang, C. Fan, DNA Framework-Encoded Mineralization of Calcium Phosphate. *Chem* **6**, 472–485 (2020).

14. D. Athanasiadou, K. M. M. Carneiro, DNA nanostructures as templates for biomineralization. *Nat. Rev. Chem.* **5**, 93–108 (2021).

15. Y. Zhang, Z. Liu, W. Zhang, X. Zhao, S. Ma, M. Li, L. Mei, A. Li, Y. Cheng, D. Pei, Non-collagenous Protein-inspired Hydrogels with Repeated Dicarboxylic Structure and High Matrix Strength for Amorphous Calcium Phosphate Stabilization to Promote Bone Defect Regeneration. *Adv. Funct. Mater.* **35**, 2500075 (2025).

16. J. Fang, P. Li, X. Lu, L. Fang, X. Lü, F. Ren, A strong, tough, and osteoconductive hydroxyapatite mineralized polyacrylamide/dextran hydrogel for bone tissue regeneration. *Acta Biomater.* **88**, 503–513 (2019).

17. F. A. Davila-Hernandez, B. Jin, H. Pyles, S. Zhang, Z. Wang, T. F. Huddy, A. K. Bera, A. Kang, C.-L. Chen, J. J. De Yoreo, D. Baker, Directing polymorph specific calcium carbonate formation with de novo protein templates. *Nat. Commun.* **14**, 8191 (2023).

18. H. Y. Koh, Y. Zheng, M. Yang, R. Arora, G. I. Webb, S. Pan, L. Li, G. M. Church, AI-driven protein design. *Nat. Rev. Bioeng.*, 1–23 (2025).

19. J. Butcher, R. Krishna, R. Mitra, R. I. Brent, Y. Li, N. Corley, P. Kim, J. Funk, S. Mathis, S. Salike, A. Muraishi, H. Eisenach, T. R. Thompson, J. Chen, Y. Politanska, E. Sehgal, B. Coventry, O. Zhang, B. Qiang, K. Didi, M. Kazman, F. DiMaio, D. Baker, De novo Design of All-atom Biomolecular Interactions with RFdiffusion3. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.09.18.676967.

20. J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

21. W. Ahern, J. Yim, D. Tischer, S. Salike, S. M. Woodbury, D. Kim, I. Kalvet, Y. Kipnis, B. Coventry, H. R. Altae-Tran, M. S. Bauer, R. Barzilay, T. S. Jaakkola, R. Krishna, D. Baker, Atom-level enzyme active site scaffolding using RFdiffusion2. *Nat. Methods*, 1–10 (2025).

22. J. Dauparas, G. R. Lee, R. Pecoraro, L. An, I. Anishchenko, C. Glasscock, D. Baker, Atomic context-conditioned protein sequence design using LigandMPNN. *Nat. Methods* **22**, 717–723 (2025).

23. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

24. N. R. Bennett, J. L. Watson, R. J. Ragotte, A. J. Borst, D. L. See, C. Weidle, R. Biswas, Y. Yu, E. L. Shrock, R. Ault, P. J. Y. Leung, B. Huang, I. Goreshnik, J. Tam, K. D. Carr, B. Singer, C. Criswell, B. I. M. Wicky, D. Vafeados, M. Garcia Sanchez, H. M. Kim, S. Vázquez Torres, S. Chan, S. M. Sun, T. T. Spear, Y. Sun, K. O'Reilly, J. M. Maris, N. G. Sgourakis, R. A. Melnyk, C. C. Liu, D. Baker, Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*, 1–11 (2025).

25. Y. Liu, C. Weidle, L. Mihaljević, J. L. Watson, Z. Li, L. T. Yu, S. Majumder, A. J. Borst, K. D. Carr, R. D. Kibler, T. M. Gamal El-Din, W. A. Catterall, D. Baker, Bottom-up design of Ca2+ channels from defined selectivity filter geometry. *Nature*, 1–9 (2025).

26. B. Liu, N. F. Greenwood, J. E. Bonzanini, A. Motmaen, J. Meyerberg, T. Dao, X. Xiang, R. Ault, J. Sharp, C. Wang, G. M. Visani, D. K. Vafeados, N. Roullier, A. Nourmohammad, D. A. Scheinberg, K. C. Garcia, D. Baker, Design of high-specificity binders for peptide–MHC-I complexes. *Science* **389**, 386–391 (2025).

27. C. Liu, K. Wu, H. Choi, H. L. Han, X. Zhang, J. L. Watson, G. Ahn, J. Z. Zhang, S. Shijo, L. L. Good, C. M. Fischer, A. K. Bera, A. Kang, E. Brackenbrough, B. Coventry, D. R. Hick, S. Qamar, X. Li, J. Decarreau, S. R. Gerben, W. Yang, I. Goreshnik, D. Vafeados, X. Wang, M. Lamb, A. Murray, S. Kenny, M. S. Bauer, A. N. Hoofnagle, P. Zhu, T. P. J. Knowles, D. Baker, Diffusing protein binders to intrinsically disordered proteins. *Nature* **644**, 809–817 (2025).

28. B. Zheng, Z. Lu, S. Wang, L. Liu, M. Ao, Y. Zhou, G. Tang, R. Wang, Y. Liu, H. Zhang, Y. Meng, J. Qiu, T. Feng, Z. Wang, R. Liu, Y. Xiao, Y. Liu, Z. Wang, Y. Huang, Y. Jiang, P. Zheng, Computational design of superstable proteins through maximized hydrogen bonding. *Nat. Chem.*, 1–10 (2025).

29. D. Kim, S. M. Woodbury, W. Ahern, D. Tischer, A. Kang, E. Joyce, A. K. Bera, N. Hanikel, S. Salike, R. Krishna, J. Yim, S. J. Pellock, A. Lauko, I. Kalvet, D. Hilvert, D. Baker, Computational design of metallohydrolases. *Nature*, 1–8 (2025).

30. J. C. Elliott, P. E. Mackie, R. A. Young, Monoclinic Hydroxyapatite. *Science* **180**, 1055–1057 (1973).

31. J. Zeglinski, M. Nolan, D. Thompson, S. A. M. Tofail, Reassigning the most stable surface of hydroxyapatite to the water resistant hydroxyl terminated (010) surface. *Surf. Sci.* **623**, 55–63 (2014).

32. L. Tran, S. Klein, S. Sharma, D. Juergens, J. Decarreau, B. Liu, Y. Wang, A. K. Bera, A. Kang, J. Woods, E. Joyce, D. K. Vafeados, N. Roullier, W. Chen, G. R. Lee, L. D. Lavis, J. Mahamid, L. An, D. Baker, Design of Orthogonal Far-Red, Orange and Green Fluorophore-binding Proteins for Multiplex Imaging. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.08.03.668343.

33. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

34. J. Qian, L. F. Milles, B. I. M. Wicky, R. J. Ragotte, A. Motmaen, A. J. Borst, R. Skotheim, S. Ols, B. Coventry, X. Li, R. D. Kibler, I. Goreshnik, M. Exposit, K. Lore, L. Stewart, D. Baker, Accelerating protein design by scaling experimental characterization. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.08.05.668824.

35. A. Sroka-Bartnicka, L. Borkowski, G. Ginalska, A. Ślósarczyk, S. G. Kazarian, Structural transformation of synthetic hydroxyapatite under simulated *in vivo* conditions studied with ATR-FTIR spectroscopic imaging. *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.* **171**, 155–161 (2017).

36. J. S. Zuluaga-Morales, M. V. Bolaños-Carmona, C. C. Cifuentes-Jiménez, P. Álvarez-Lloret, Chemical, Microstructural and Morphological Characterisation of Dentine Caries Simulation by pH-Cycling. *Minerals* **12**, 5 (2022).

37. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A.

Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

38. B. D. Quan, E. D. Sone, The effect of polyaspartate chain length on mediating biomimetic remineralization of collagenous tissues. *J. R. Soc. Interface* **15**, 20180269 (2018).

39. L. Niu, S. E. Jee, K. Jiao, L. Tonggu, M. Li, L. Wang, Y. Yang, J. Bian, L. Breschi, S. S. Jang, J. Chen, D. H. Pashley, F. R. Tay, Collagen intrafibrillar mineralization as a result of the balance between osmotic equilibrium and electroneutrality. *Nat. Mater.* **16**, 370–378 (2017).

40. L. Chen, Z. Zeng, W. Li, Poly(acrylic acid)-Assisted Intrafibrillar Mineralization of Type I Collagen: A Review. *Macromol. Rapid Commun.* **44**, 2200827 (2023).

41. U. Amornkitbamrung, Y. In, J. H. Lee, Z. Wang, S. H. Oh, H. Shin, D. S. Yoon, H. Shin, Interface-Controlled Biomimetic Intrafibrillar Mineralization of Collagen: Effect of Ca2+/[PO4]3− Concentration Ratio. *Adv. Mater. Interfaces* **10**, 2300384 (2023).

42. C. Drouet, Apatite Formation: Why It May Not Work as Planned, and How to Conclusively Identify Apatite Compounds. *BioMed Res. Int.* **2013**, 490946 (2013).

43. D. Chávez-Herrera, E. Rangel-Villanueva, M. Salazar-Hernández, A. Talavera-Lopez, A. N. Ardila A., R. Hernández-Soto, O. J. Solis-Marcial, J. A. Hernández, Morphology-Dependent Transformation of Dicalcium Phosphate Dihydrate (DCPD) to Octacalcium Phosphate (OCP) and Its Stability in Simulated Physiological Fluids. *Molecules* **30**, 3631 (2025).

44. A. Laskus-Zakrzewska, A. Zgadzaj, J. Kolmas, Synthesis and physicochemical characterization of Zn-doped brushite. *Ceram. Int.* **47**, 7798–7804 (2021).

45. M. H. AlRefeai, E. M. AlHamdan, S. Al-Saleh, A. S. Alqahtani, M. Q. Al-Rifaiy, I. F. Alshiddi, I. Farooq, F. Vohra, T. Abduljabbar, Application of β-Tricalcium Phosphate in Adhesive Dentin Bonding. *Polymers* **13**, 2855 (2021).

46. A. P. Patni, R. Mout, A. Alghadeer, R. H. Moore, S. Nademi, B. J. Ausk, P. Huber, R. Keshri, M. Regier, R. A. Cornell, T. Gross, B. S. Freedman, G. Q. Daley, D. Baker, J. Mathieu, H. Ruohola-Baker, Soluble Notch agonist enables human ameloblast maturation and enamel-like tissue formation for tooth regeneration. *Int. J. Oral Sci.* **18**, 25 (2026).

47. A. J. Ben-Sasson, J. L. Watson, W. Sheffler, M. C. Johnson, A. Bittleston, L. Somasundaram, J. Decarreau, F. Jiao, J. Chen, I. Mela, A. A. Drabek, S. M. Jarrett, S. C. Blacklow, C. F. Kaminski, G. L. Hura, J. J. De Yoreo, J. M. Kollman, H. Ruohola-Baker, E. Derivery, D. Baker, Design of biologically active binary protein 2D materials. *Nature* **589**, 468–473 (2021).

48. B. Xue, Y. Li, Z. Fu, H. Ping, K. Wang, Intrafibrillar Growth of Hydroxyapatite Nanocrystals in Multiscale Collagen. *Crystals* **13**, 692 (2023).

49. F. Nudelman, K. Pieterse, A. George, P. H. H. Bomans, H. Friedrich, L. J. Brylka, P. A. J. Hilbers, G. de With, N. A. J. M. Sommerdijk, The role of collagen in bone apatite formation in the presence of hydroxyapatite nucleation inhibitors. *Nat. Mater.* **9**, 1004–1009 (2010).

50. tracyyu2, tracyyu2/hap_mineralization: Templating and Confining Calcium Phosphate Mineralization within Designed Protein Assemblies, version HAp_mineralization_v1, Zenodo (2026); https://doi.org/10.5281/ZENODO.20334827.

51. The PyMOL Molecular Graphics System, Version 3.0 Schrödinger, LLC.

52. D. N. Mastronarde, Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

53. R. T. Kidmose, J. Juhl, P. Nissen, T. Boesen, J. L. Karlsen, B. P. Pedersen, Namdinator - Automatic molecular dynamics flexible fitting of structural models into cryo-EM and crystallography experimental maps. *IUCrJ* **6**, 526–531 (2019).

54. A. Casañal, B. Lohkamp, P. Emsley, Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. *Protein Sci. Publ. Protein Soc.* **29**, 1069–1078 (2020).

55. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

56. T. I. Croll, ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 519–530 (2018).

57. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

58. P. V. Afonine, B. K. Poon, R. J. Read, O. V. Sobolev, T. C. Terwilliger, A. Urzhumtsev, P. D. Adams, Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. Sect. Struct. Biol.* **74** (2018).

59. I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall III, J. Snoeyink, J. S. Richardson, D. C. Richardson, MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).

60. R. Mout, R. Jing, M. Tanaka-Yano, E. D. Egan, H. Eisenach, M. A. Kononov, R. Windisch, M. A. T. Najia, A. Tompkins, L. Hensch, T. Bingham, R. Gunage, Y. Zhao, N. I. Edman, C. Li, D. Wang, T. M. Schlaeger, L. I. Zon, T. E. North, U. Lendahl, R. G. Rowe, D. Baker, S. C. Blacklow, G. Q. Daley, Design of soluble Notch agonists that drive T cell development and boost immunity. *Cell* **188**, 5980-5994.e28 (2025).

Acknowledgements: We thank S. Woodbury for helping set up RFDiffusion2 scripts. LTY thanks Dr. J. Chen and the rest of Baker lab members for general discussions. We thank IPD Core Research members X. Li and C. Möller for mass spectrometry data collection and processing. We thank HHMI Janelia Research Campus for cryoEM data collection; J. Quispe, F. Zhang and W. Jiang for the management of Arnold and Mabel Beckman CryoEM Center at University of Washington. We thank IPD Yeast Production/Processing Center members for providing us with LM670 vectors. We thank the IPD lab management team led by K. VanWormer, IPD IT team led by L. Goldschmidt, IPD Electron Microscopy Research Core (EMRC) led by A. Borst for technical support. Part of the nanotubes and 2D lattices designs (HP) as well as the Liquid-phase AFM and associated mechanical mapping (XL, SZ) were based upon work supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, as part of the Energy Frontier Research Centers (EFRC) program: CSSAS – The Center for the Science of Synthesis Across Scales under Award Number DE-SC0019288 and at Pacific Northwest National Laboratory under contract FWP 72448. PNNL is a multiprogram national laboratory operated for DOE by Battelle under Contract No. DE-AC05-76RL01830. STEM/EDS characterization performed at PNNL was supported by the U.S. DOE BES, Material Sciences and Engineering Division, Synthesis and Processing Science program, FWP 67554 (D. Li). Part of the TEM experiments were conducted at the Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure site at the University of Washington which is supported in part by the National Science Foundation (grant NNCI-1542101), the University of Washington, the Molecular Engineering & Sciences Institute, the Clean Energy Institute, and the National Institutes of Health.

Funding:

This work was funded by the following: The Breakthrough Fund Mineralization at the Institute for Protein Design (to LTY), The Audacious Project at the Institute for Protein Design (to PSK, SM, RDK); The Howard Hughes Medical Institute (to DB); The Bill and Melinda Gates Foundation INV-043758 (to AJB, CW, KDC); HHMI Hanna H. Gray Fellowship, Award No. 17718 (to NPB), The DOE BES Center for the Science of Synthesis EFRC (to JDY, DB, SZ). This material is based on work supported by the Air Force Office of Scientific Research under award number FA9550-22-1-0506 (to YL).

Author contributions:

Conceptualization: LTY, HP, NPB, DB

Methodology: LTY, HP, XL, NPB, AJB, CW, KDC, APP, PK

Investigation: LTY, HP, XL, AJB, YL, PK, RDK, APP, DL, SM

Data processing: LTY, XL, APP, AJB, CW, KDC

Visualization: LTY, XL, APP, AJB

Funding acquisition: HR-B, SZ, JDY, DB

Supervision: HR-B, SZ, JDY, DB

Writing – original draft: LTY, HP, XL, AJB, JDY, DB

Writing – review & editing: All authors

Competing interests: Authors declare that they have no competing interests.

Data, code, and materials availability: All data are available in the main text or the supplementary materials. CryoEM structure of the HAp1_Nt was deposited in the Protein Data Bank under accession number 9ZNK.
Scripts for RFdiffusion are available at https://github.com/RosettaCommons/RFdiffusion. Scripts for ProteinMPNN and LigandMPNN are available at https://github.com/dauparas/ProteinMPNN and https://github.com/dauparas/LigandMPNN, respectively. Scripts for AlphaFold2 and AlphaFold3 are available at https://github.com/google-deepmind/alphafold and https://golgi.sandbox.google.com/ respectively. All detailed scripts, designed structures in pdb format and designed sequences in fasta files in this work are publicly available at Zenodo(*50*). All raw data of the plots presented in this work are available in the Data S1 file.

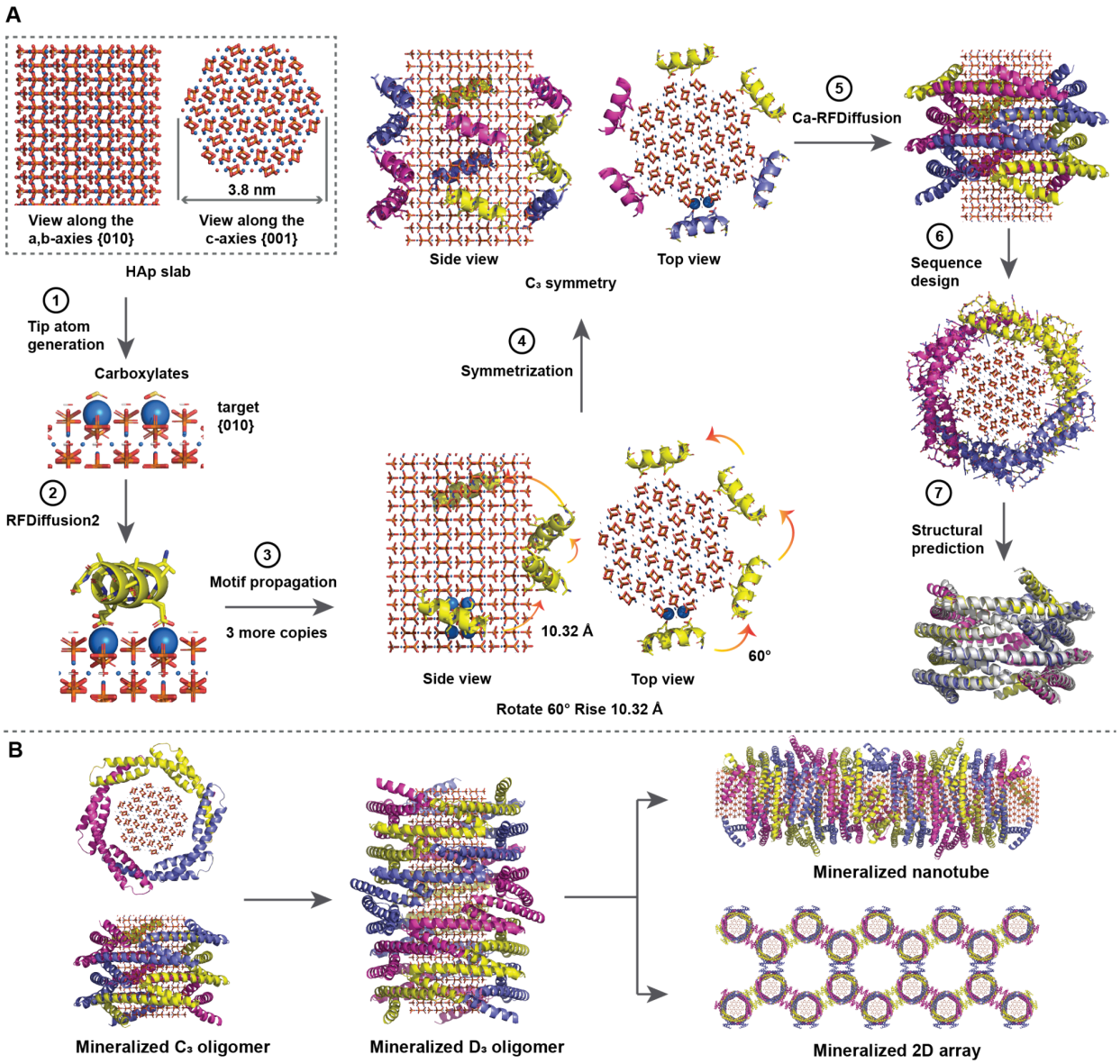

A
3.8 nm
View along the a,b-axies {010}
View along the c-axies {001}
HAp slab
1
Tip atom generation
Carboxylates
target {010}
2
RFDiffusion2
3
Motif propagation
3 more copies
Side view
Top view
C₃ symmetry
4
Symmetrization
10.32 Å
60°
Side view
Top view
Rotate 60° Rise 10.32 Å
5
Ca-RFDiffusion
6
Sequence design
7
Structural prediction
B
Mineralized C₃ oligomer
Mineralized D₃ oligomer
Mineralized nanotube
Mineralized 2D array

Fig. 1. Design strategy. (**A**) Computational design work flow for generating C_3 oligomers to stabilize HAp precursor and promote nucleation. The design starts with a hexagonal model of HAp that terminates at {010} facets at a and b directions. Carboxylates (yellow residues) were placed to interact with the calcium ions (blue balls) from HAp. The yellow alpha helix was generated with RFD2 and copies were propagated with a helical rise and twist to “wrap” the HAp model. The four helices together form an asymmetric unit which was transformed into a C_3 oligomer (the two additional chains are in magenta and purple). Gaps between the helices were filled using RFdiffusion(*32*). In step 7, the AF2 prediction is in gray and the design model is colored. (**B**) Hierarchical design plan. The first step of the hierarchy is the mineralized C_3 oligomer; the feature in the middle of the ring represents HAp. C_3 oligomers were then designed into D_3 oligomers which were then designed into one-dimensional protein nanotubes or two-dimensional protein arrays.

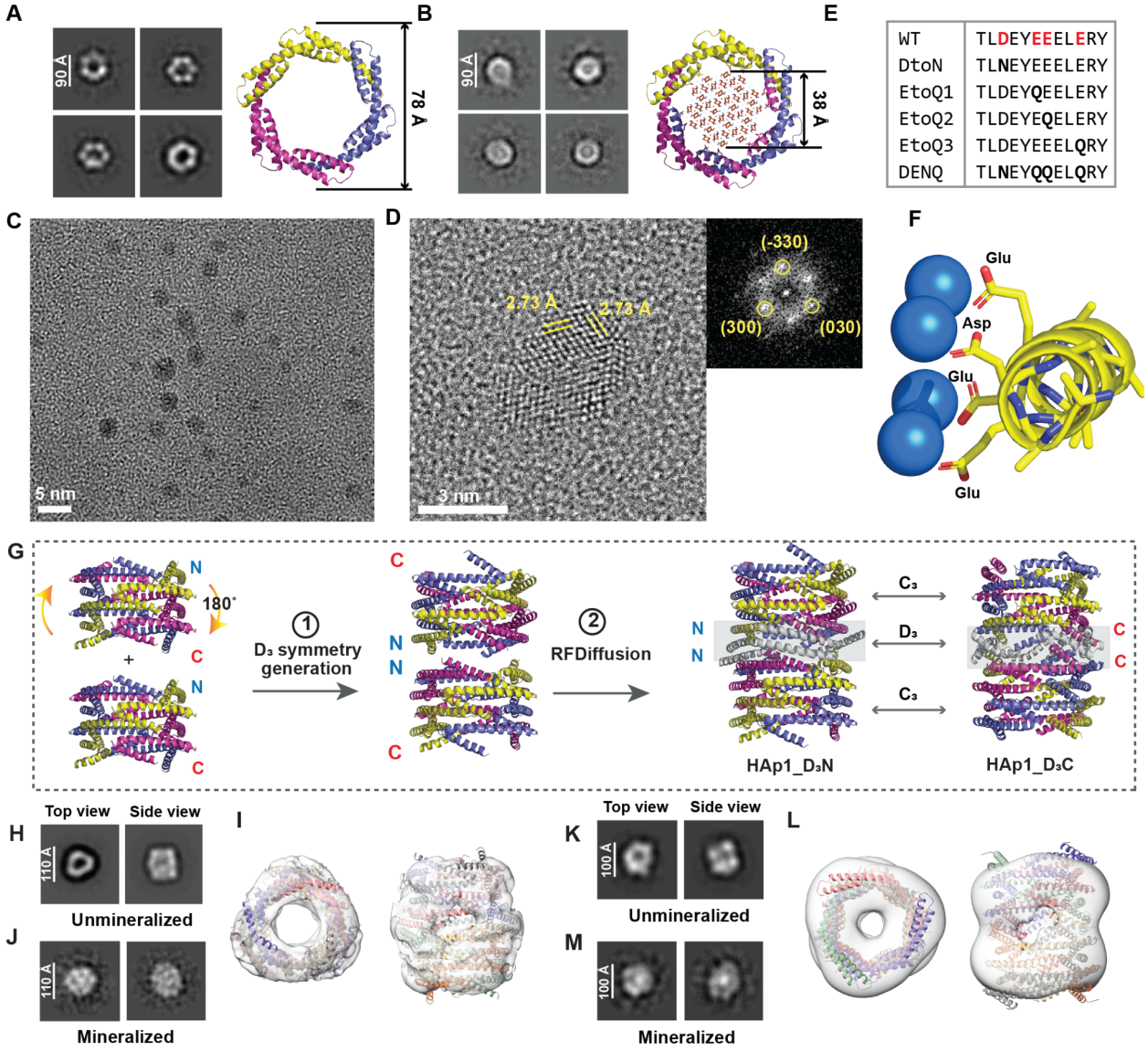

A
90 Å
B
78 Å
90 Å
38 Å
E
WT TLDEYEEELERY
DtoN TLNEYEEELERY
EtoQ1 TLDEYQEELERY
EtoQ2 TLDEYEQELERY
EtoQ3 TLDEYEEELQRY
DENQ TLNEYQQELQRY
C
5 nm
D
2.73 Å
2.73 Å
3 nm
(-330)
(300)
(030)
F
Glu
Asp
Glu
Glu
G
N
180˚
C
+
N
C
① D₃ symmetry generation
C
N
N
C
② RFDiffusion
N
N
C₃
D₃
C₃
C
C
HAp1_D₃N
HAp1_D₃C
H
Top view
Side view
110 Å
Unmineralized
J
110 Å
Mineralized
I
K
Top view
Side view
100 Å
Unmineralized
M
100 Å
Mineralized
L

Fig. 2. HAp crystal templating by the HAp1_C_3 and HAp1_D_3 designs. (**A**) Unmineralized HAp1_C_3 nsEM 2D classes. (**B**) Mineralized HAp1_C_3 cryoEM 2D classes (**C**) Mineralized HAp1_C_3 TEM image. (**D**) Mineralized HAp1_C_3 HRTEM. The top right panel shows the FFT of the nanocrystal. (**E**) Protein sequence of the repeated interfacial helix of HAp1_C_3 and knockout mutants. Residues in bold indicate mutations. (**F**) Model illustration of designed carboxylates to interact with calcium from HAp. (**G**) HAp1_D_3 oligomer design. The D_3 interface does not incorporate lattice-matching sequences. (**H**) and (**K**) nsEM 2D class averages of HAp1_D_3N and HAp1_D_3C, respectively. (**I**) and (**L**) Design models fitted into a volume density map generated from nsEM results of HAp1_D_3N and HAp1_D_3C, respectively. The “pores” visible in the side views are at the D_3. (**J**) and (**M**) nsEM 2D class of mineralized HAp1_D_3N and HAp1_D_3C, respectively. Raw micrographs are in Fig. S22.

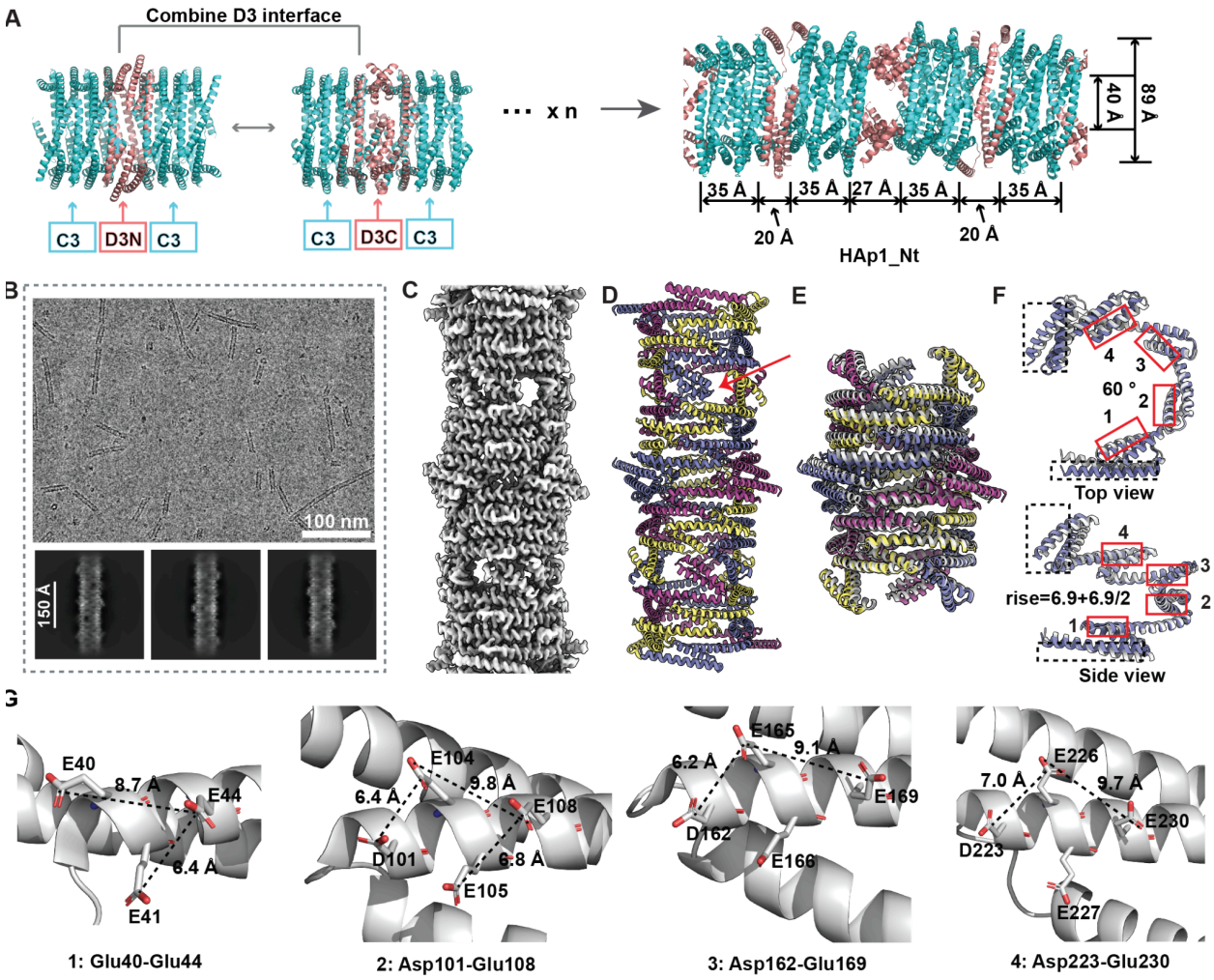


Fig. 3. Design and cryoEM structure of the HAp1_Nt nanotube design. (**A**) Schematic of nanotube design showing the stacking of the lattice-matching region derived from the original C_3 oligomer (cyan) and the non-lattice-matching D_3 polymerization region (salmon). The resulting nanotubes have alternating lattice matching (3.5 nm) and non-lattice matching (2-2.7 nm) regions along the extended dimension. (**B**) cryoEM raw micrograph (top) and 2D class averages (bottom) of HAp1_Nt. (**C**) CryoEM density map. (**D**) CryoEM derived 3D structure. Arrow indicates periodic pores designed at D_3 interfaces that may facilitate molecular transport. (**E**) Design model of the nanotube (grey) overlaps with the cryoEM structure (colored). (**F**)

Monomer C_α RMSD ~1.1 Å. CryoEM structure is in purple and the design model in grey part. Red boxes indicate the initial mineral templating regions (see Fig 1A); black-dash boxes indicate non-nucleating D_3 polymerization interfaces near the N- (1-39) and C-termini (276-332). (**G)** distance measurements between pairs in Glu40-Glu44, Asp101-Glu108, Asp16240-Glu169, and Asp223-Glu230.

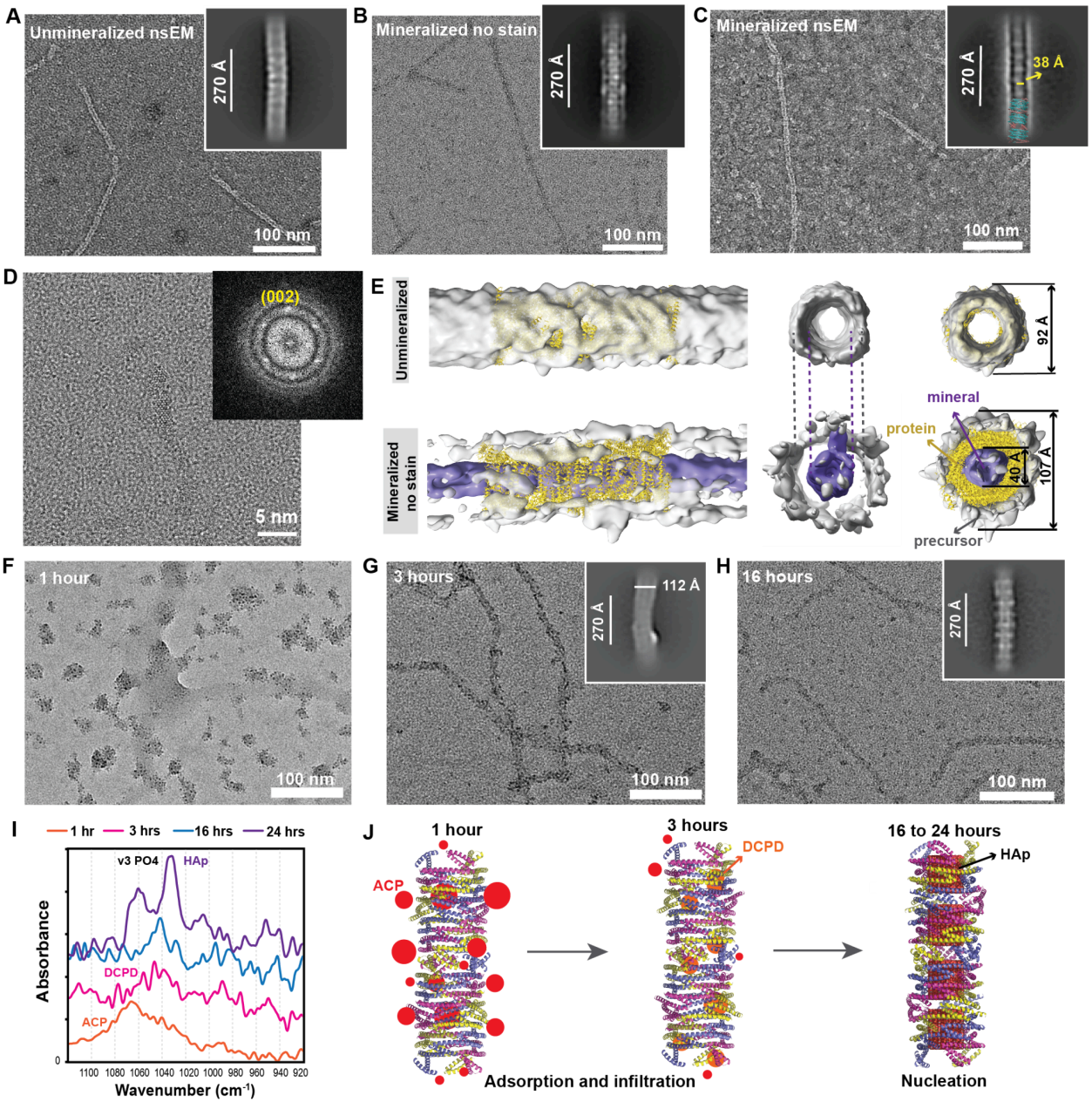


Fig. 4. Time-course of HAp

mineralization within HAp1_Nt nanotubes. (A) nsEM of unmineralized nanotubes, with 2D class average in inset. (**B**) TEM of HAp1_Nt after 24 hours of mineralization in presence of 180 µg/mL 27kda pAsp *(assay 2*). (**C**) nsEM of the sample in panel B. The inset shows the 2D class average superimposed on the HAp1_Nt cryoEM structure. (**D**) HRTEM of mineralized HAp1_Nt sample in panel B, with FFT in inset. (**E**) The cryo-EM structure of apo HAp1_Nt fitted into nsEM-derived density map of unmineralized nanotubes in panel A and TEM-derived density map of mineralized nanotubes with pAsp in panel C. (**F-H**) TEM of HAp1_Nt mineralized after 1 hour, 3 hours, and 16 hours mineralization, respectively. The insets are 2D class averaging. (**I**) ATR-FTIR characterization of HAp1_Nt mineralized after different mineralization times. (**J**) schematic illustration of the nanotube mineralization process.

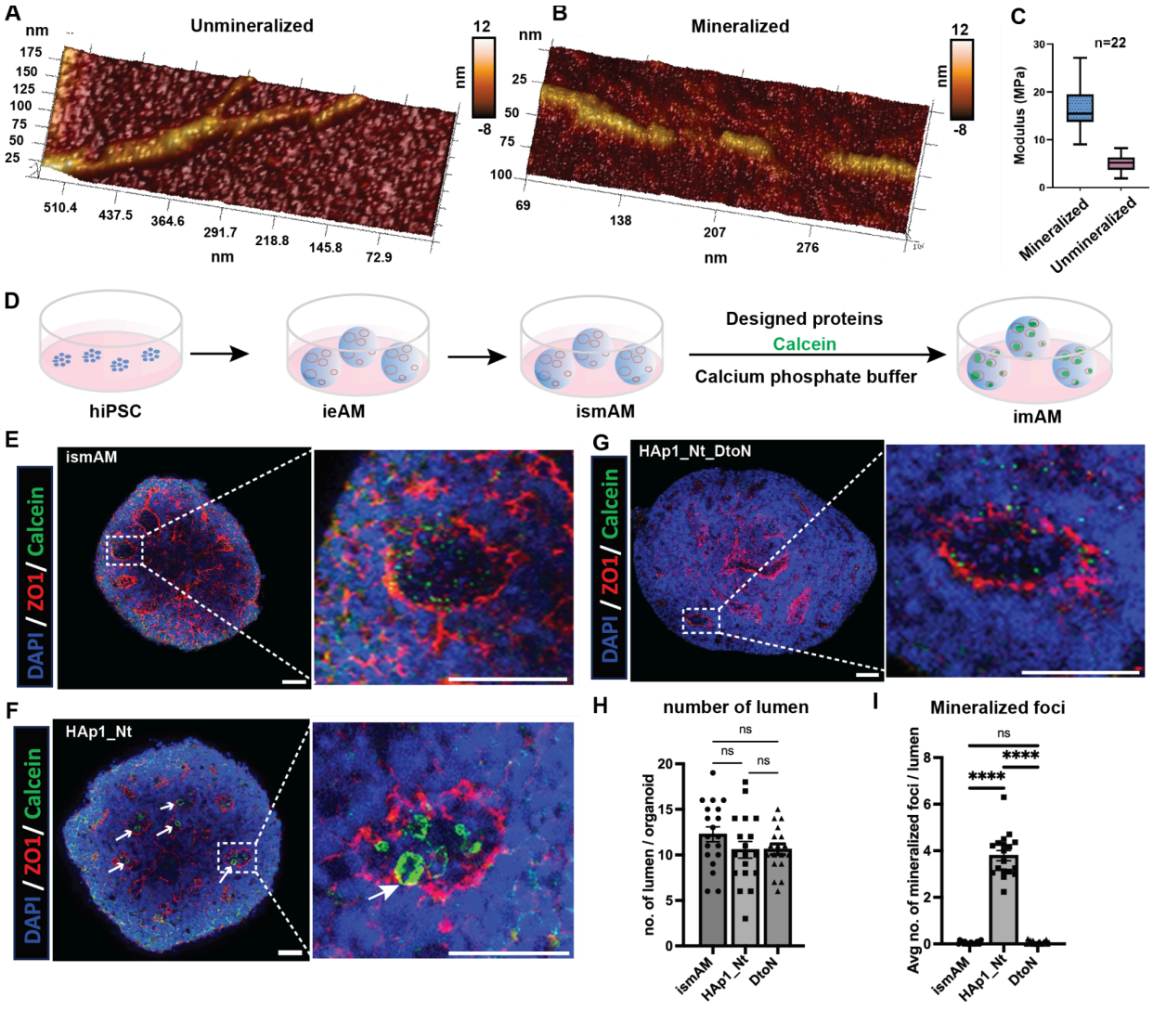

A
Unmineralized
nm
175
150
125
100
75
50
25
510.4
437.5
364.6
291.7
218.8
145.8
72.9
nm
12
nm
-8
B
Mineralized
nm
25
50
75
100
69
138
207
276
nm
12
nm
-8
C
n=22
Modulus (MPa)
30
20
10
0
Mineralized
Unmineralized
D
hiPSC
ieAM
ismAM
Designed proteins
Calcein
Calcium phosphate buffer
imAM
E
ismAM
DAPI / ZO1 / Calcein
F
HAp1_Nt
DAPI / ZO1 / Calcein
G
HAp1_Nt_DtoN
DAPI / ZO1 / Calcein
H
number of lumen
no. of lumen / organoid
20
15
10
5
0
ns
ns
ns
ismAM
HAp1_Nt
DtoN
I
Mineralized foci
Avg no. of mineralized foci / lumen
8
6
4
2
0
ns
****
****
ismAM
HAp1_Nt
DtoN

Fig. 5. HAp1_Nt mineralization increases nanotube stiffness and promotes mineral deposition in dental organoids. (**A, B**) LP-AFM topography images of unmineralized and mineralized HAp1_Nt nanotubes, respectively. (**C**) Elastic modulus distributions calculated from Fig. S31A, B. (**D**) Schematic of the ismAM (induced secretory-stage maturation ameloblast) dental organoid mineralization assay showing the human induced pluripotent stem cell (hiPSC) differentiation trajectory from ieAM (induced early ameloblast) to ismAM and, following HAp1_Nt-driven mineralization, to imAM (induced mature ameloblast). (**E–G**) Confocal images of ismAM organoids cultured without added protein (**E**), with HAp1_Nt (**F**), or with HAp1_Nt_DtoN (**G**). Nuclei (DAPI, blue), tight junctions and lumen boundaries (ZO-1, red), and mineralized foci (Calcein, green) are shown. Scale bars: 50 μm. (**H**) Quantification of lumen number per organoid across treatment groups. (**I**) Quantification of average number of mineralized foci per lumen across treatment groups ($p < 0.0001$).

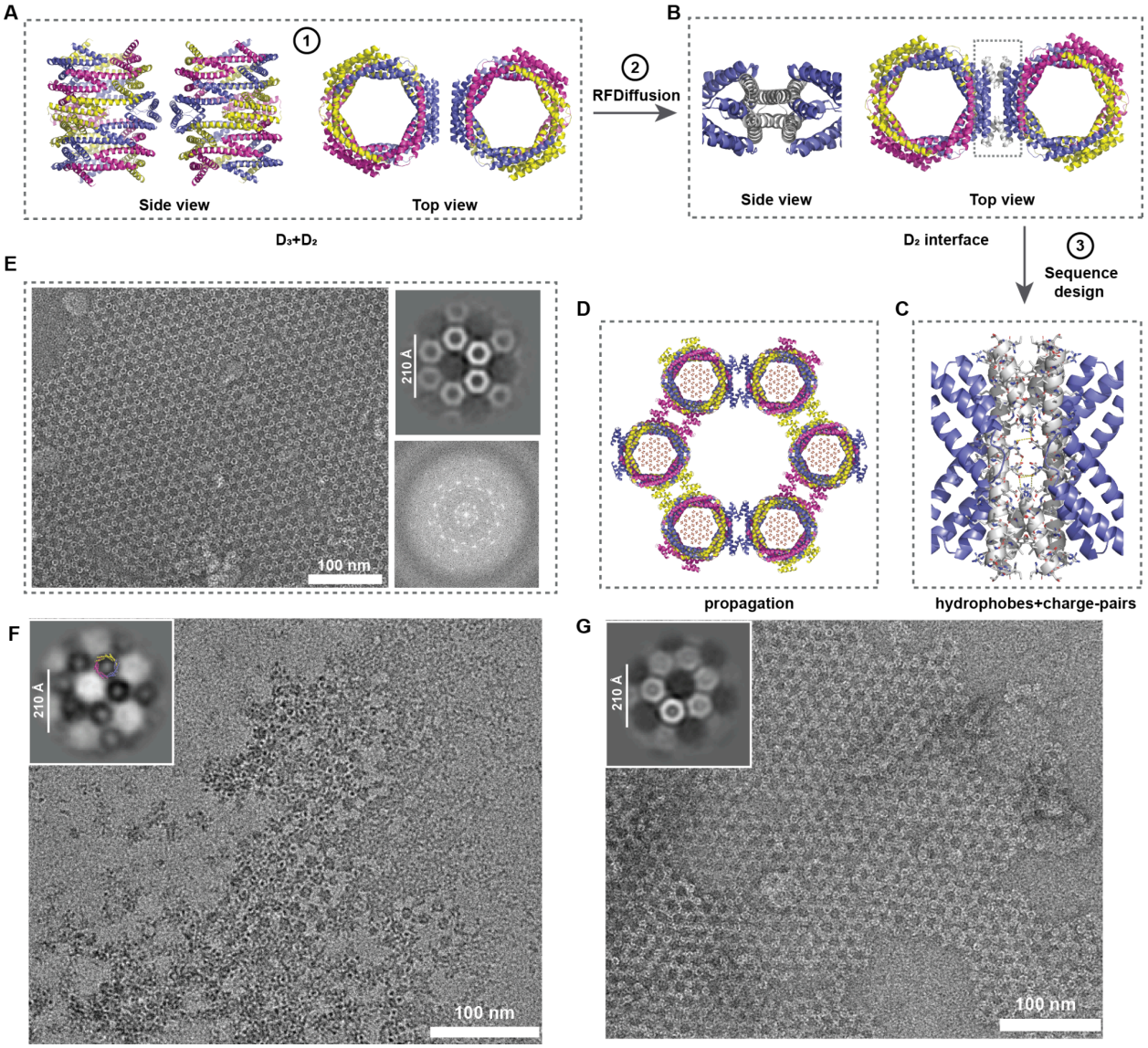

A
1
Side view
Top view
D3+D2
2
RFDiffusion
B
Side view
Top view
D2 interface
3
Sequence design
E
210 Å
100 nm
D
propagation
C
hydrophobes+charge-pairs
F
210 Å
100 nm
G
210 Å
100 nm

Fig. 6. Two-dimensional array HAp1_Ar design, characterization and mineralization. (**A**) Two HAp1_D_3N oligomers were arranged to generate D_2 symmetry. (**B**) D_2 interface generation using RFDiffusion (grey part). (**C**) Side chain interactions of D_2 interface. (**D**) Propagation models with templated HAp nanocrystals generated in PyMOL(*51*). (**E**) and (**G**) nsEM results of the unmineralized and mineralized array, respectively. (**F**) TEM of mineralized array. The inset is 2D class averaging overlapping with the D_3 design model.

Appendix 1: De novo design of RNA and nucleoprotein complexes

This Appendix contains additional background for (and limited reproduction of) content submitted as: Andrew Favor, Riley Quijano, Elizaveta Chernova, Andrew Kubaney, Connor Weidle, Morgan A. Esler, Lilian McHugh, Ann Carr, Yang Hsia, David Juergens, Kenneth D. Carr, Paul T. Kim, Yuliya Politanska, Enisha Sehgal, Paul S. Kwon, Robert J. Pecoraro, Cameron Glasscock, Andrew J. Borst, Frank DiMaio, Barry L. Stoddard, David Baker. *Nature Methods* (in review).

Abstract

Nucleic acids fold into sequence-dependent tertiary structures and carry out diverse biological functions, much like proteins. However, while considerable advances have been made in the *de novo* design of protein structure and function, the same has not yet been achieved for RNA tertiary structures of similar intricacy. Here, we describe a generative diffusion framework, *RFdiffusion-polymer (RFDpoly)*, for generalized *de novo* biopolymer (RNA, DNA and protein) design, and use it to create diverse and designable RNA structures. We design RNA structures with novel folds and experimentally validate them using a combination of chemical footprinting (SHAPE-seq) and electron microscopy. We further use this approach to design protein-nucleic acid assemblies; the crystal structure of one such design is nearly identical to the design model. This work demonstrates that principles of structure-based *de novo* protein design can be extended to nucleic acids, providing a route toward the structural design of RNA architectures and protein–nucleic acid complexes.

Introduction

Noncoding RNAs adopt complex three-dimensional structures that underpin a wide range of biological functions, including genetic regulation, catalysis, and scaffolding of biochemical machinery. In the field of nucleic acid design, most strategies have focused on three main areas: aptamer generation via randomized sequence selection[1,2], the design of small RNAs with defined secondary structures[3,4], and the construction of large, geometrically regular RNA/DNA origami[5,6,7,8]. Aptamer generation has been largely structure-independent, while secondary structure-based approaches focus on two-dimensional base pair templates and do not account for three-dimensional geometry, which limits their utility in applications requiring precise spatial control. Three-dimensional nucleic-acid design approaches — such as origami, hierarchical alignment of known motifs[9,10], and parametric methods that propagate idealized geometry[11,12] — often depend on rigid geometric assumptions (e.g., uniform helical axes) and on reuse of pre-existing structural fragments to simplify placement of features such as crossovers and tertiary contacts. While these approaches can generate sophisticated and functional nucleic-acid structures, parametric methods often fail to capture the complexity observed in native RNA folds (which feature multi-angled junctions, bulges, bent helices, and non-canonical contacts), and fragment-alignment methods remain constrained to the design space covered by existing motif libraries, which limits the range of geometric problems they can solve. Thus, there remains a critical need for structure-oriented nucleic acid design approaches that enable exploration of

the space of diverse three-dimensional structures exemplified by natural RNAs.

We reasoned that recent advances in generative deep learning for protein design could be leveraged to address the current limitations in nucleic acid design. Generative models such as RFdiffusion have shown considerable success in structure-oriented protein design[13,14,15], and the RoseTTAfold framework has been adapted for nucleic acid structure prediction[16,17]. Although denoising diffusion probabilistic models (DDPMs) have been applied to generate short RNA monomers[18,19,20,21,22], to date, these efforts lack hierarchical design capabilities, lack multi-polymer structure generation, and none have yet been experimentally validated. We set out to generalize the RFdiffusion *de novo* protein design approach to nucleic acids, and to explore the use of this method by designing novel RNA structures and protein-nucleic acid assemblies.

Computational approach:

Our design approach proceeds in two steps. First, we generate RNA or DNA backbone structures using an extended version of RFdiffusion, *RFdiffusion-polymer* (*RFDpoly*), described in the following paragraphs. Second, we design base sequences on the generated backbones using NA-MPNN[23], which generalizes the widely used ProteinMPNN[24,25] to nucleic acids.

RFDpoly extends RFdiffusion to enable the generation of nucleic acid structures by building on RF2-AllAtom[16,17], which models both proteins and nucleic acids and accepts conditional information through 1D, 2D, and 3D channels. Previous RFdiffusion implementations could only denoise protein structures around fixed nucleic acid ligands; RFDpoly generates nucleic acid structures as well. Unlike proteins, where all non-frame backbone atoms (O, C_β) can be deterministically placed given frame coordinates, nucleic acid backbones have increased conformational freedom due to variable sugar and phosphate geometries. To address this, RFDpoly still denoises frame rotations and translations as in previous RFdiffusion models, but goes further by using rigid body parameters *and* torsion angles predicted from RoseTTAfold to construct complete nucleic acid backbones (and side chains, when sequence or motif geometry is provided). To increase the extent to which the denoised frames determine nucleobase positioning, we switched from the phosphate-centered frame atoms used in RF2-AllAtom (OP1-P-OP2) to O4'-C1'-C2' frame atoms, which are closer to bases.

To condition the diffusion process, RFDpoly uses one-hot encoded molecule-class features, provided through RoseTTAfold's 1D track, to guide the generation of appropriate backbone geometry for each polymer type – producing chemically accurate RNA and DNA without compromising protein structure quality. With RFDpoly trained to denoise these different biopolymers, we can specify which polymer class is generated in each contiguous chain during inference (**Fig. 1c**).

To enable RFDpoly to condition on base-pairing networks when generating 3D RNA structures, we expanded RoseTTAfold to accept base pair partner labels through its 2D-track (**Fig. 1b**). These conditional features allow RFDpoly to generate backbones that possess specified secondary structure patterns. To enable user guidance over structure generation in ways that are relevant to nucleic acid design, we provide several routes for controlling secondary structure features at inference time (**Fig. 1d**). Secondary structure strings (in the form of dot bracket notation) can be provided to generate structures with specific pseudoknot topologies, allowing users to generate 3D models for RNA puzzles[3,26,27] or other sources of secondary structure templates. Beyond generating canonical single-partner base pairs, users can specify whether pairs are parallel or antiparallel, giving precise control over tertiary structure features such as triple helices or G-quadruplexes. Lists of paired sequence regions can be specified, with cyclic- or repeat-symmetry, to reflect design patterns often seen in nucleic acid origami.

We train RFDpoly using a randomized masking scheme in which training examples are divided into *motif* regions (containing ground truth structure information) and *diffused* regions – to emulate design tasks such as motif-bridging, docking, or generating new structure around central motifs. To enable RFDpoly to generate both protein and nucleic acid structures, and assemblies containing both polymer types, we train the model using datasets consisting of proteins, RNA, and DNA either as monomers or as mixed polymer complexes. While previous versions of RFdiffusion only generate structure within protein chains, keeping nucleic acids as fixed motifs, the mask generators for RFDpoly treat all polymer classes equally, denoising noised regions within nucleic acid chains as well as proteins.

Once backbone structures have been generated, sequence design is performed using two methods. In the first approach, NA-MPNN is used to assign sequences to complete diffusion-generated backbones[23], and PyRosetta is then used to build and repack sidechains. In the second approach, sequence-structure codesign, RoseTTAfold sequence predictions are autoregressively decoded along with rigid-body and torsion parameters to produce full atom sidechain representations. Sequences designed with NA-MPNN were predicted to fold with better self-consistency metrics, but the presence of side chains from codesign trajectories allowed immediate filtering of structures for target base pair pattern satisfaction, without going through the process of MPNN-design and PyRosetta repacking.

Designed sequences were filtered based on similarity between design models and predicted structures at multiple levels. Initial filtering was based on predicted secondary structure self-consistency using RibonanzaNet[28]; similarity between designed and predicted secondary structure was evaluated using F1-scores[29]. Tertiary structures of designs with high F1 scores were predicted using Chai-1[30] or AlphaFold3[31], and compared to design models, using lDDT[29,32,33] or TM-score[34] as metrics for structure similarity. We use the term “designable’’ below to describe designs for which the predicted structure matched the design model.

In unconditional structure generation calculations, the generated structures maintained high designability across increasing lengths, with little drop-off in secondary structure self-consistency metrics, and minimal steric clashing even in large-scale generation tasks (**Extended Data Fig. 1a**). Tertiary structure *diversity* was evaluated by comparing inter- and intra-group distributions of 3-dimensional fold similarities for diffusion-generated structures against native RNAs from structural databases, or against other diffusion-generated structures. Pairwise alignment of RNA structures (within 20% sequence length) from both generated and native sets (~1,700 RCSB RNA structures) using USalign[34], yielded distributions of tertiary structure similarity (given by TM-score) across various length scales (**Extended Data Fig. 1b**). While the unconditionally generated structures of short RNAs had folds resembling those in the PDB (e.g., tRNAs), diversity increased rapidly for structures longer than 120 bases with the majority of RFDpoly-generated structures dissimilar from native RNAs. This short-length enrichment for common native-like folds likely reflects biases in available RNA structural training data; in practice, base-pair conditioning provides a way to guide generation away from these unconditional sampling biases toward user-specified topologies.

A standard test of protein designs is the similarity of the designed backbone with structure predictions based on the designed sequence using a method such as AF3 (if the structures match closely, the designed sequence and backbone are internally consistent, and the backbone is said to be “designable”). A challenge for assessing RNA designs in silico is that AF3 and other prediction methods are substantially less accurate for RNA than for DNA. Nevertheless, just as the Rosetta ab initio structure prediction method was very useful for evaluating protein designs prior to the development of the much more accurate deep learning based prediction methods, we reasoned that the extent of similarity between our designed RNA backbones and AF3 predictions of the structures of the corresponding MPNN designed sequences could still provide a useful albeit imperfect in silico design metric. A partial justification of this approach is the well documented result that designed protein structures are far easier to predict than native protein structures, not requiring a multiple sequence alignment to extract co-evolution constraints–designed structures generally have fewer irregular and suboptimal sequence and structure features than native proteins. Furthermore, to reduce reliance on any single oracle, we assess self-consistency using multiple independent predictors (AlphaFold3 and RibonanzaNet) trained on distinct datasets; the concordance between these methods (Extended Data Fig. 1a) provides partial support for the assessment. Consistent across two independent predictors (AlphaFold3 and RiboNanzaNet), self-consistency F1 decreases only gradually with length, whereas cross-design redundancy (mean top-5 design–design F1 to other backbones of the same length) drops much more steeply (**Extended Data Fig. 1c**), indicating that RFDpoly expands structural diversity at longer lengths while preserving designability according to current *in silico* filters.

Experimental evaluation in the Eterna OpenKnot challenges:
To benchmark the ability of our method to generate structures with specific base-pair patterns, we participated in the Eterna OpenKnot challenges[27]**.** Base-pairing templates from a set of 57 pseudoknot puzzles were used to condition denoising trajectories

(**Fig. 2a**), and sequences were assigned using NA-MPNN. For all puzzles attempted, reference wild-type sequences were provided for comparison, and many puzzles had experimentally determined 3D structures, which were also redesigned with NA-MPNN. To experimentally characterize the secondary structures of designed sequences, the OpenKnot organizers used SHAPE-seq, which generates reactivity profiles that distinguish between base pairing and non-base pairing regions[28,35,36,37]**.**

For clarity, we first describe the design approach and the comparison to experimental data in more depth for a specific case, puzzle Q19 (SV_r7_240_4), which consists of five hairpins (α_1–α_5). In the target structure, hairpin pairs (α_1 : α_3) and (α_4 : α_5) form kissing-loop interactions, while hairpin α_2 remains unpaired (**Fig. 2a**, left). Starting from random noise, the base pair-conditioned diffusion denoising trajectory progressively builds an RNA structure adhering closely to these predefined contacts and hairpin geometries (**Fig. 2a**, right). Hosts of the OpenKnot competition experimentally assessed this design using selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq), which reports on nucleotide flexibility and accessibility by chemically modifying unpaired or flexible nucleotides. Modified nucleotides impede reverse transcription, allowing sequencing readouts to quantify site-specific reactivities. Regions displaying high SHAPE reactivity correspond to unpaired or loop nucleotides, whereas paired bases are protected from modification and thus exhibit low reactivity. To show how reactivity data map back to structural features, we color secondary and tertiary structures by reactivity profile for a selected design (*) (**Fig. 2b**). Consistent with expectation, we observe low reactivity in the paired kissing loops, but high reactivity in the unpaired stem loop bases (α_2). Thus, the SHAPE-seq reactivity profiles indicate that the design secondary structure closely matches the target secondary structure (horizontal axes, with desired paired and unpaired positions colored in blue and red, respectively). To systematically evaluate performance across multiple design strategies, SHAPE-reactivity profiles were compiled into two-dimensional heatmaps (**Fig. 2c**), comparing our diffusion-based method with alternative backbone redesigns and wild-type RNA sequences. The selected design (*) displayed optimal agreement with the target structure, as quantitatively captured by the highest achieved target OpenKnot score among designs for puzzle Q19 (**Fig. 2d**).

Overall, SHAPE-seq experimental reactivity profiles for targeted puzzles were close to the input base pairing profiles (**Fig. 2c**). *Target OpenKnot Scores*[38] were used as an aggregate metric to summarize how well each design folded into the target secondary structure (**Fig. 2d-e**)**.** For most puzzles attempted, at least one sequence from MPNN-design of RFDpoly-backbones outperformed the wild-type sequences and MPNN-redesign of experimental reference structures (**Fig. 2f**, **Extended Data Fig. 2a-c**).

These results show that RFDpoly can generate 3D models across a wide range of target secondary structures. User-input enables the generation of backbones for a diverse set of pseudoknot topologies without loss in designability (**Fig. 2e**). Many of the designs characterized by SHAPE-seq display highly accurate formation of intricate non-standard tertiary elements. For example, P15-07 and P20-10 exhibit perpendicular helix–helix groove docking, Q17-07 and Q16-09 form high-order multi-way junction hubs, and W08-146

and P05-14 use loop–groove clasping interactions to stabilize packing (**Extended Data Fig. 3**). The diversity and complexity of secondary structures designed indicate that RFDpoly can explore a complex fold space beyond that which is accessible to origami approaches or even observed in native systems.

Design of pseudocycles:

To further assess RFDpoly's design capabilities, we sought to create larger novel RNA architectures. We focused on designing *RNA pseudocycles*: large, monomeric RNA structures composed of nearly cyclically repeating subunits connected in a single continuous chain. These folds were chosen as design targets because their large size, relative to our previously designed RNA pseudoknots, would enable characterization via electron microscopy, and their nearly symmetric morphologies would be easily identifiable during screening. We chose single-chain pseudocycles rather than cyclic assemblies to avoid the complications of multimeric assembly, enabling direct application of *in silico* screening metrics optimized for single-chain RNAs.

To generate large pseudocycles with numerous intra-chain contacts to reduce structure flexibility, we developed a secondary structure template generation protocol that accepts graph-based definitions of regional pairing and efficiently searches for fitting base-pair patterns, diversified using random sampling of subregion lengths (**Fig. 3a**). Once candidate base pair templates were generated, they were provided to RFDpoly as 2D templates to guide the denoising process. Pseudocyclic RNA backbones were generated using symmetric noise propagation during denoising, while maintaining backbone connectivity between subunits in encoded bond features. After backbone generation and sequence design, *in silico* predictions of tertiary structure with AlphaFold3 showed conformational variability between models, while having matching secondary structures between designed and predicted structures (**Fig. 3b-c**). Synthetic genes were obtained and *in vitro*-transcribed for eight 372-base designs, all sharing a common twofold pseudocyclic symmetry separated by hinge-like loops.

Characterization with negative stain electron microscopy (nsEM) revealed well-formed C2 pseudosymmetric structures for six designs (**Extended Data Fig. 4a-c**). Observed morphologies were polydisperse, with the butterfly-like folds sampling multiple conformational states, and the two rigid side-domains folding about their central hinge loops. For two of the four designs (pC2-372-07, pC2-372-11), class averages for one of the primary observed particles (denoted *set-1*) matched the design model, while others (*set-2*) more closely matched the AlphaFold3 predicted models (**Fig. 3d-f**). For the remaining designs, reconstructed volumes did not closely resemble either the design or predicted structures, while nsEM revealed alternative pseudosymmetric folds (**Extended Data Fig. 4d**). All of the nsEM-characterized designs had similar secondary structure features (defined by four characteristic cloverleaf folds: σ_1, σ_2, σ_1', σ_2'), and for each design with nsEM 3D reconstructions, one of the primary observed states had the target double-triangle fold (*set-1*). Fitting of design models into 3D-reconstructed volumes for particle *sets-1* demonstrated

structural consistency within this subpopulation (**Fig. 3f**). To further characterize the hinge-motion between the open and closed conformations, we classified particles into multiple 3D volumes, which revealed connectivities between the designed conformation and alternative states (**Fig. 3g**).

We next designed smaller (240 nt) compact RNA pseudocycles, aiming for well-defined tertiary structures by enriching for supportive base pair contacts and iteratively refining the overall structure by recycling the most well-ordered regions of the models back into RFDpoly as base pair templates, over multiple rounds of design and selection. These shorter-length sequences could be characterized with SHAPE-seq, allowing us to compare experimental results between chemical probing and electron microscopy. Fifty designs with two- or three-fold pseudocyclic symmetry were selected for SHAPE-seq characterization based on high *in silico* self-consistency metrics. Overall, there was strong agreement between experimental reactivity profiles and target secondary structures (**Fig 4a**), with 42 designs achieving *Target OpenKnot Scores* above a target threshold of 0.8 (**Extended Data Fig. 5a**). Of these, 19 were selected for further characterization by electron microscopy, based on whether Chai-1[30] or AlphaFold3[31] predictions resembled the design models. RNA transcribed from DNA templates was screened by nsEM, and the most structured candidates were further characterized by cryo-EM. Three designs (pC2-240-02, pC2-240-03, pC2-240-04) had good SHAPE-reactivity profiles (**Extended Data Fig. 5b**) and well-defined 2D classes from cryo-EM (**Fig. 4b,** top; **Extended Data Fig. 5c-d**).

For pC2-240-04, cryo-EM characterization led to a 6.6 Å reconstructed volume, with clearly observable helical pitch and a global fold matching the design model (**Fig. 4b**, middle and bottom). The experimental SHAPE-reactivity profile matches the designed secondary structure (**Fig. 4c**, left), and the design model could be fit into the cryo-EM reconstructed volume (**Fig. 4c**, right). The locations of paired and loop regions observed in the cryo-EM reconstruction are consistent with the SHAPE-reactivity profile. Of particular interest are three central helical domains, δ_1 , δ_2 and δ_3 (**Fig. 4c**, middle), with crossover regions in the secondary structure that make this fold a complex high-order pseudoknot (since these central crossover domains can occupy separate locations in 2D secondary structure diagrams, we label paired strands as δ_i and δ_i', denoting the upstream and downstream components, respectively). The interweaving of these three helical domains between protruding helical hairpins in the design model is confirmed in the 3D tertiary structure. To assess the novelty of this fold, we aligned the experimentally refined model against all representative RNAs from structural databases (see Methods) and found no close matches: the best hit was to a small region of the plant mitochondrial ribosome (PDB 6XYW, chain A) with a TM-score of 0.33 at 0.55 coverage, which is below the threshold for RNA fold similarity. The combined validation using SHAPE-seq and cryo-EM shows that RFDpoly can accurately create intricate new RNA tertiary structures.

Motif scaffolding and hierarchical design:

We next set out to use RFDpoly to design protein–nucleic acid assemblies. We began by incorporating cognate protein and DNA motif fragments designed and characterized in previous work[39] into larger assemblies; simultaneously fusing interacting protein (C-to-N) and DNA (3′-to-5′) motifs with newly generated rigid connective structures. We used 2D motif-template conditioning to embed three sequence-orthogonal DNA-binding proteins (DBPs) and their respective target DNAs (**Fig. 5a**) into larger connected assemblies, allowing their global placement and the connections between them to be inferred during the denoising process (**Fig. 5b**). Base pair templates that specified DNA strand exchange were used to condition the denoising process, and control long-range connectivity (**Fig. 5c,d**), enabling the creation of intricate nucleoprotein complexes with distributed, multi-contact binding surfaces that would be challenging to achieve with purely parametric approaches.

We first applied this hierarchical strategy to design protein–DNA complexes in which protein homodimers bind to strand-exchanging DNA heterodimers to form hybrid complexes reminiscent of classical Holliday junctions[40] (**Fig. 5e**). These designs were created using *inter-helical fusions* (**Fig. 5b, top**) which connect two DBPs across different bound helices, acting as rigid "staples" that lock the relative orientation of the DNA helices, reflecting prior strategies in protein-DNA hybrid nanostructures[41]. An initial set of designs was generated in which DBP fusions form bridges across DNA helical extensions; upon characterization by nsEM, these designs showed high conformational variability, likely due to limited buttressing interactions, which complicated structural characterization. Despite uncertainty in the geometries of these complexes, we did observe on-target trimer assembly for one design, ECC09 (**Fig. 5e-g**, top row). Difficulties in characterizing the flexible complexes motivated a second design campaign, in which the generated protein structure not only connects DBPs but also forms a homooligomeric interface with the other protein chains in the complex, to create a more rigid protein-DNA assembly. This second round of protein-protein interface-forming complexes was expressed and again characterized by nsEM, revealing successful assembly for two designs, ECC36 and ECC38 (**Fig. 5e-g**, bottom two rows). The pseudo-symmetric structure of these complexes is shown clearly in 2D class averages (**Fig. 5f**) and 3D reconstructed volumes (**Fig. 5g**), consistent with the bent configuration of the DNA in these strand-exchanging complexes. Thus, RFDpoly can carry out motif scaffolding while simultaneously generating designs with complex secondary structures, providing a route to bridge the worlds of protein design and DNA origami.

We next used the same flexible scaffolding approach to generate new proteins predicted to bind extended contiguous DNA sequences (**Fig. 5h**). This second hierarchical design campaign utilizes *intra-helix DNA fusions* (**Fig. 5b**, bottom) in which DBPs are concatenated along a shared helical axis to create a longer sequence-specific recognition domain, expanding on existing techniques for the modular assembly of DNA-binding proteins[42,43]. By sampling the geometry of the connecting DNA during DBP fusion, the designs can bind DNA in a wide range of conformations, in contrast to a previous approach limited to linear, parametrically generated B-form DNA[39]. The newly designed intra-helix DBP fusions were evaluated by yeast-display flow cytometry experiments (**Fig. 5h**,

bottom) in which biotinylated dsDNA targets were titrated to generate binding curves, yielding apparent K_D values from ~2 μM (RQ04) to ~10 nM (RQ02).

We assessed model accuracy by X-ray crystallography, obtaining a 3.0 Å resolution crystal structure of RQ01 bound to its DNA target (**Fig. 5i**). The crystal structure closely matches the design model over the full assembly, including the rigid linkers; notably, the bent DNA conforms to the design, indicating that RFDpoly can accurately design non-ideal nucleic-acid geometries. Together, these results showcase the utility of RFDpoly for generating new DNA-binding proteins specific for extended DNA sequences and the ability to scaffold multiple biopolymer types simultaneously. Overall, our hierarchical design approach allows for the reuse of previously characterized structures to accelerate further design efforts and enables the construction of increasingly complex classes of biomolecular architectures.

Discussion

The *de novo* design of nucleic acid tertiary structures has been a longstanding challenge due to their intrinsic flexibility and the relatively small number of structures available for training deep learning models compared to proteins. Previous efforts, including adaptations of diffusion models for RNA, have been restricted to short sequences, exhibit limited structural diversity, and lack robust motif scaffolding or experimental validation, leaving the physical plausibility of their outputs uncertain[18,19,20,21,22]. In direct *in silico* benchmarks against alternative RNA tertiary structure generation models, RFDpoly yielded higher secondary-structure self-consistency, improved AF3-based tertiary-structure agreement, and higher backbone quality scores (**Extended Data Fig. 7**). RFDpoly enables user-guided generation of structured RNA with intricate, native-like folds, going beyond the regular origami-style topologies that have long defined the field of three-dimensional nucleic acid design. The use of randomly-sampled motif masks and conditional feature labels during training allows the model to learn local features of molecular structure, and to generalize these globally at inference time. Explicit hydrogen bond conditioning during the RFDpoly denoising process enables control over the topologies of the structures generated through the installation and enrichment of base-interaction networks. The successful design and cryo-EM validation of a 240-nucleotide RNA pseudocycle demonstrates that RFDpoly can produce novel RNAs with intricate folds. Despite being trained on a relatively small dataset of experimentally determined RNA structures, RFDpoly generalizes beyond known topologies (**Extended Data Fig. 1a, Extended Data Fig. 3**), in some cases producing designs more stable than their natural counterparts (**Extended Data Fig. 2**). The ability of RFDpoly to generate both base pair-conditioned and unconditional structures with high accuracy and diversity suggests that it does not simply memorize motifs from the training data, but instead learns fundamental structural principles, enabling the exploration and creation of physically viable, previously unobserved RNA folds.

The ability of RFDpoly to generate protein-nucleic acid assemblies enables the design of increasingly complex structures by reusing existing components. Combining the capabilities of RFDpoly to carry out polymer-class conditioning, 2D-templated scaffolding of preexisting motifs, and base-pair templating to guide nucleic acid secondary structures enabled the design of the protein-DNA assemblies in **Fig. 5**. The close agreement of the crystal structure of a designed protein-DNA assembly, RQ01, with the design model showcases the ability of RFDpoly to generate atomically accurate designs and to stabilize irregular (bent) DNA backbone structures. The relative placements of the input motif sets were not pre-defined within a global coordinate frame as in previous work: instead, the model inferred a physically plausible arrangement while generating the connecting backbones, eliminating the need for manual motif positioning, and enabling joint optimization of motif placement and backbone generation.

Despite the many new design capabilities of RFDpoly, there are still many avenues to explore to expand our ability to design RNA and nucleoprotein complexes. Designed compact and specific protein-RNA complexes would be of considerable interest in themselves, and would enable integration into higher order RNA-protein assemblies by hierarchical assembly, analogous to the DBP-DNA fusions presented here. The development of more reliable tertiary structure prediction methods for protein-nucleic acid assemblies would considerably improve *in silico* evaluation of designs, and in combination with experimental design results, help uncover nucleic acid folding principles[44]. Broadening RFDpoly training to include greater chemical diversity (small-molecules, noncanonical nucleotides, and post-synthetic modifications) would further extend generative capabilities, and enable the design of aptamers and riboswitches programmed to bind more generalized targets; future work may also benefit from training on high-quality synthetic datasets containing diverse fold representation to overcome the redundancies of current RNA structural databases. Finally, much of the utility of RNA in nature arises from conformational flexibility and stimulus-dependent behavior; fully realizing this potential in *de novo* design will require designing for dynamics to achieve *functions* of interest and pairing these efforts with experimental characterization of activity, responsiveness, and dynamics. Ultimately, the continued development of generative frameworks like RFDpoly will likely transform the design of structured nucleic acids and proteins, unite protein design and DNA origami methods, and enable rapid, precise engineering of biomolecular assemblies for synthetic biology, therapeutic control, and beyond.

Data Availability

Structure files for RNA pseudocycle *pC2-240-04*, and DBP-fusion *RQ01* have been deposited in the PDB, and will be publicly released at the time of publication.
Structure files for computational design models are included with this manuscript as supplementary data.

Experimental SHAPE-seq data and *in silico* scoring metrics for compact (240 nt) *de novo* RNA pseudocycles are provided as supplementary materials. Data for OpenKnot designs are available at: https://github.com/eternagame/OpenKnotAIDesignData.

Code Availability

The code for running the RFDpoly has been released as a GitHub repository (https://github.com/RosettaCommons/RFDpoly), which includes links to download model weights. Interactive tutorials highlighting the design techniques used in this paper are available on the RFDpoly GitHub page, and are also included as Supplementary Material files in this manuscript: https://github.com/RosettaCommons/RFDpoly/tree/colab_tutorial/tutorials

Acknowledgements

We thank the Eterna project and the Das lab (Stanford, HHMI) for coordination and experimental evaluation of the Eterna OpenKnot challenge.
The authors would like to thank Lance Stewart and Lynda Stuart for their work in maintaining operations at the Institute for Protein Design; Luki Goldschmidt and Kandise VanWormer and for maintaining the computational and wet lab resources that were used in this research; Madison Kennedy and Bulat Faezov for technical help in preparation of this manuscript; Han Altae-Tran, Pooja D Bandawane, Andrew C. Hunt, Angela M. Yu, and Florian Praetorius for experimental advice and insightful discussions over the course of this project.

Funding

This work was supported by Bill and Melinda Gates Foundation INV-043758 (C.W., D.J., K.C., Y.P., A.B., A.C., D.B.); Gift from Microsoft (D.B.); Howard Hughes Medical Institute (D.B.); The Audacious Project at the Institute for Protein Design (A.F., P.K., D.B.); The Open Philanthropy Project Improving Protein Design Fund (E.S., Y.H., D.B.); the National Science Foundation Grant No. CHE-2226466 (L.M., R.J.P., C.G., B.L.S., D.B.); the National Institutes of Health (NIH) NIGMS GM148166 (B.L.S.); Washington Research Foundation (C.G.).

Crystals were screened at the Fred Hutchinson Cancer Center X-ray diffraction facility (supported by NIH grant S10OD028581). Crystallographic data were collected at the Advanced Light Source (ALS), which is supported by the Director, Office of Science, Office of 20 Basic Energy Sciences, and US Department of Energy under contract number DE-AC02-05CH11231, and the ALS-ENABLE grant P30 GM124169, which supports a consortium of structural biology beamlines.

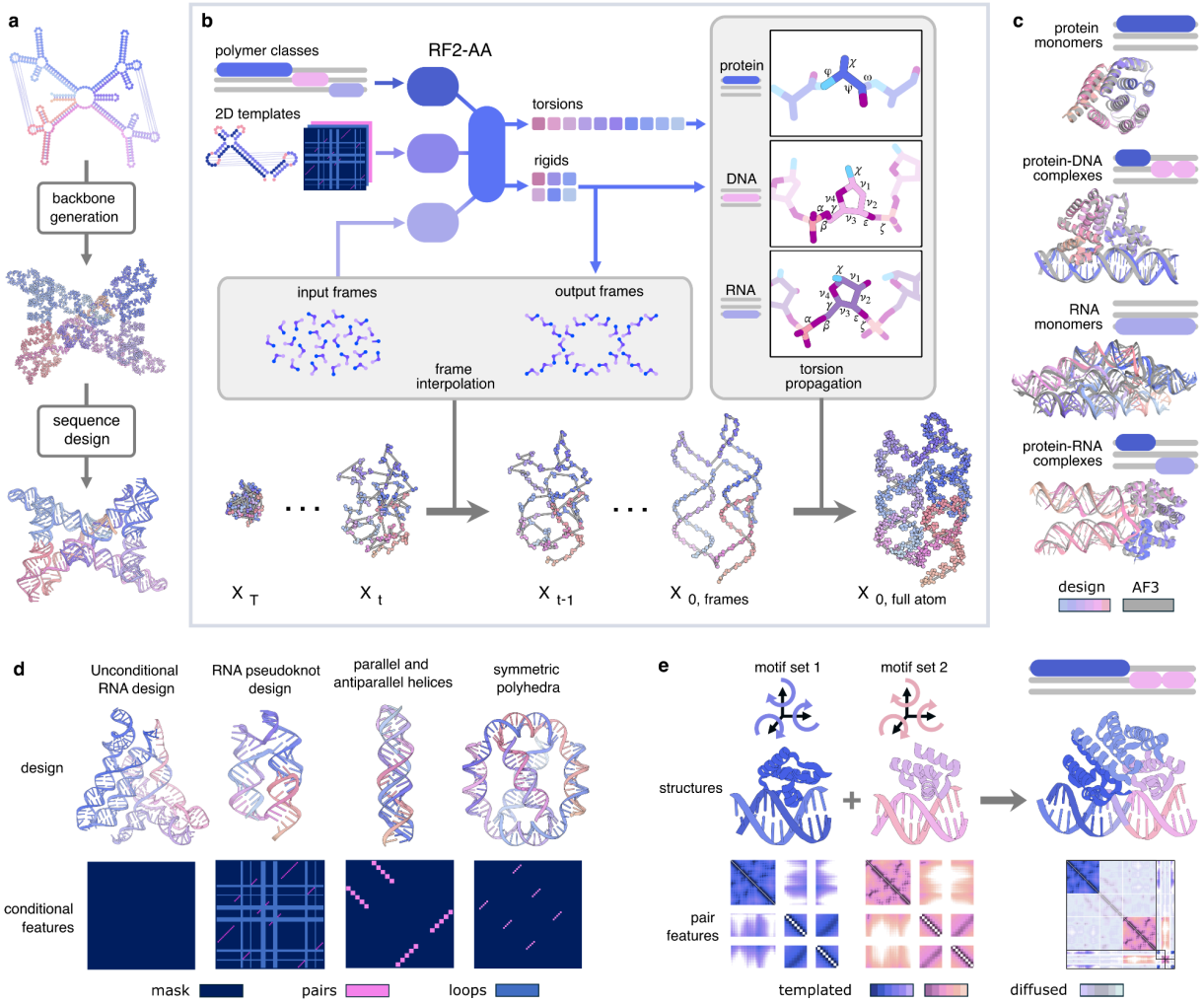
Author contributions

A.H.F. conceived the study, wrote the manuscript, trained RFDpoly, and developed the RNA design pipeline. A.H.F., R.Q., and E.C. generated experimentally characterized designs. R.Q., E.C., E.S., Y.P., A.C., and Y.H. purified proteins. A.H.F purified RNA. A.J.K., L.M., P.T.K., and D.J. contributed additional code. C.G., A.J.B., P.S.K., R.P., F.D., and D.B. offered supervision throughout the

project. B.L.S. and M.A.E. characterized crystal structures. C.W., A.J.B., K.C., A.H.F., A.C., and Y.H. performed electron microscopy and data processing. A.J.K. developed and trained NA-MPNN. L.M. and F.D. trained RF-NA.

Figures

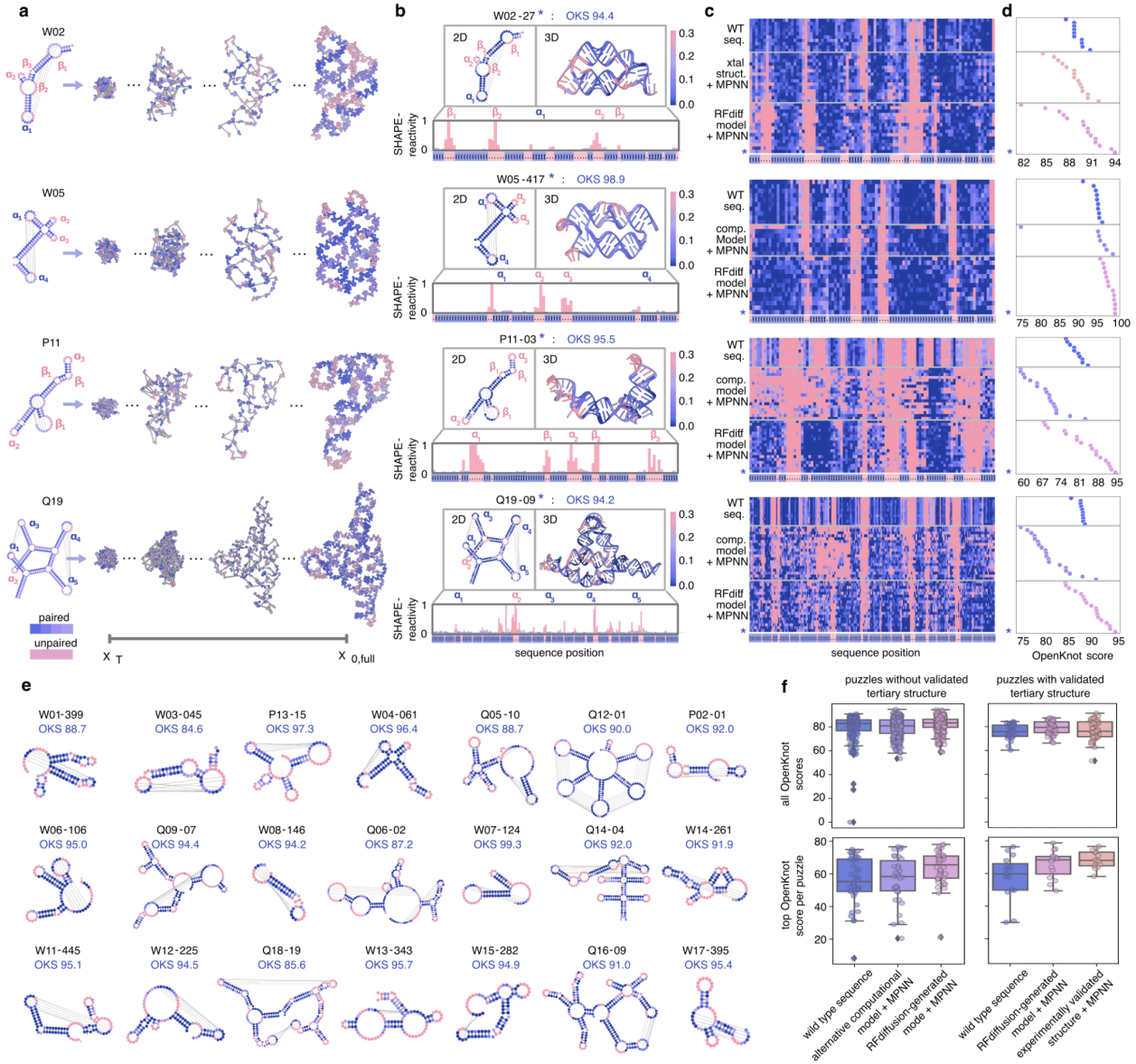
Figure 1. Generalized biopolymer structure generation with RFDpoly



a, Standard *de novo* structure design pipeline, consisting of backbone generation using RFDpoly, followed by sequence design using NA-MPNN. **b**, RFDpoly denoises backbone frames using RoseTTAfold's 3D track, while conditional information from the 1D and 2D

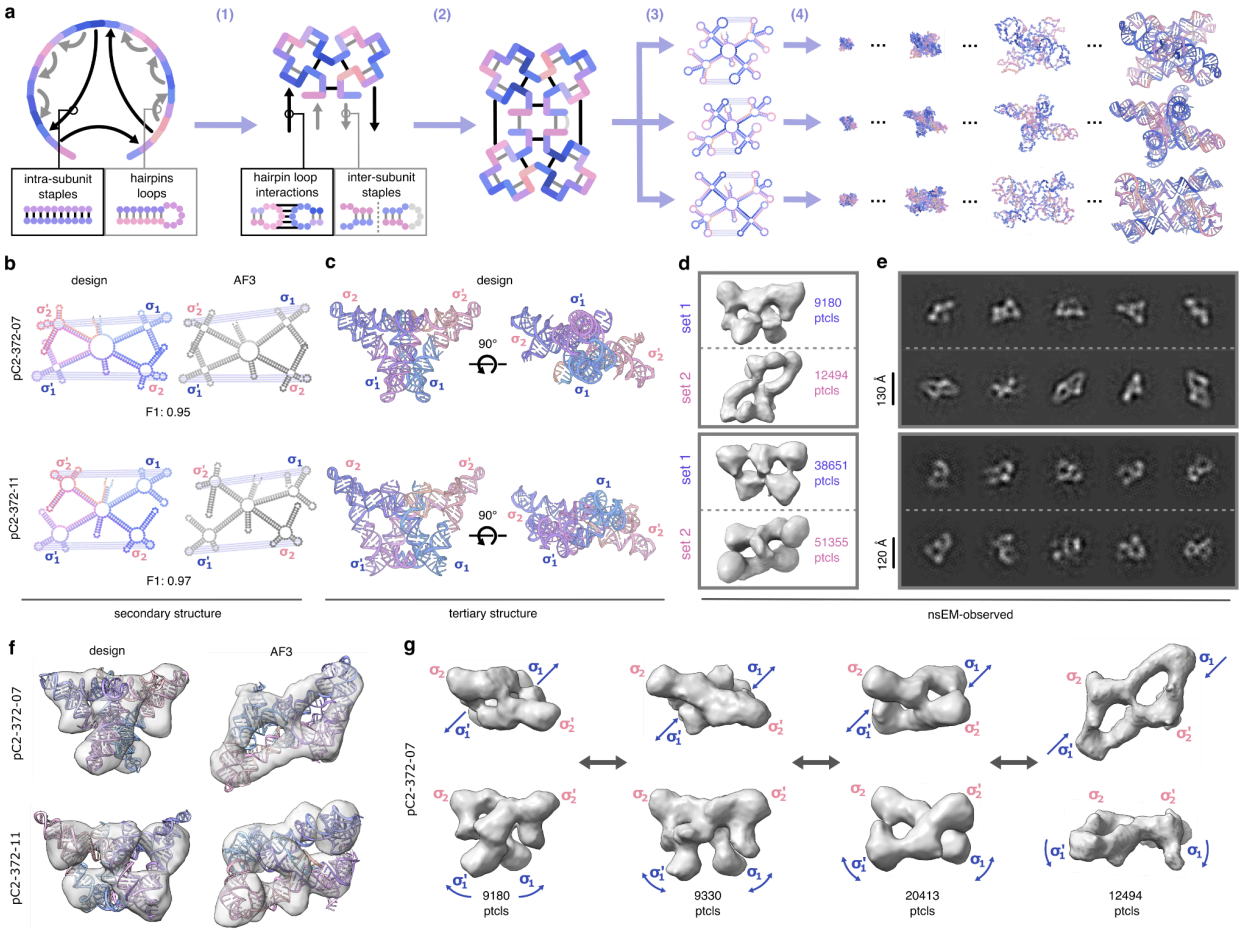
tracks (such as polymer class and base-pair patterns, respectively) guides the generation toward desired structures. Full sugar–phosphate backbone coordinates are constructed by propagating torsion angles about frames, using connectivity graphs defined by sequence predictions. **c**, Polymer-class labels provided at inference instruct the model to generate chemically accurate structures appropriate for each molecule type. *In silico* predictions with AlphaFold3 show high self-consistency for proteins, DNA, and RNA. **d**, Structural control via base-pair templating enables RFDpoly to generate nucleic acid structures with diverse topologies. User-specified controls define templates for RoseTTAfold's 2D track. **e**, Hierarchical design in RFDpoly uses 2D templates for motif structure, allowing global placement to be inferred during the denoising process.

Figure 2. Experimental validation with SHAPE-seq



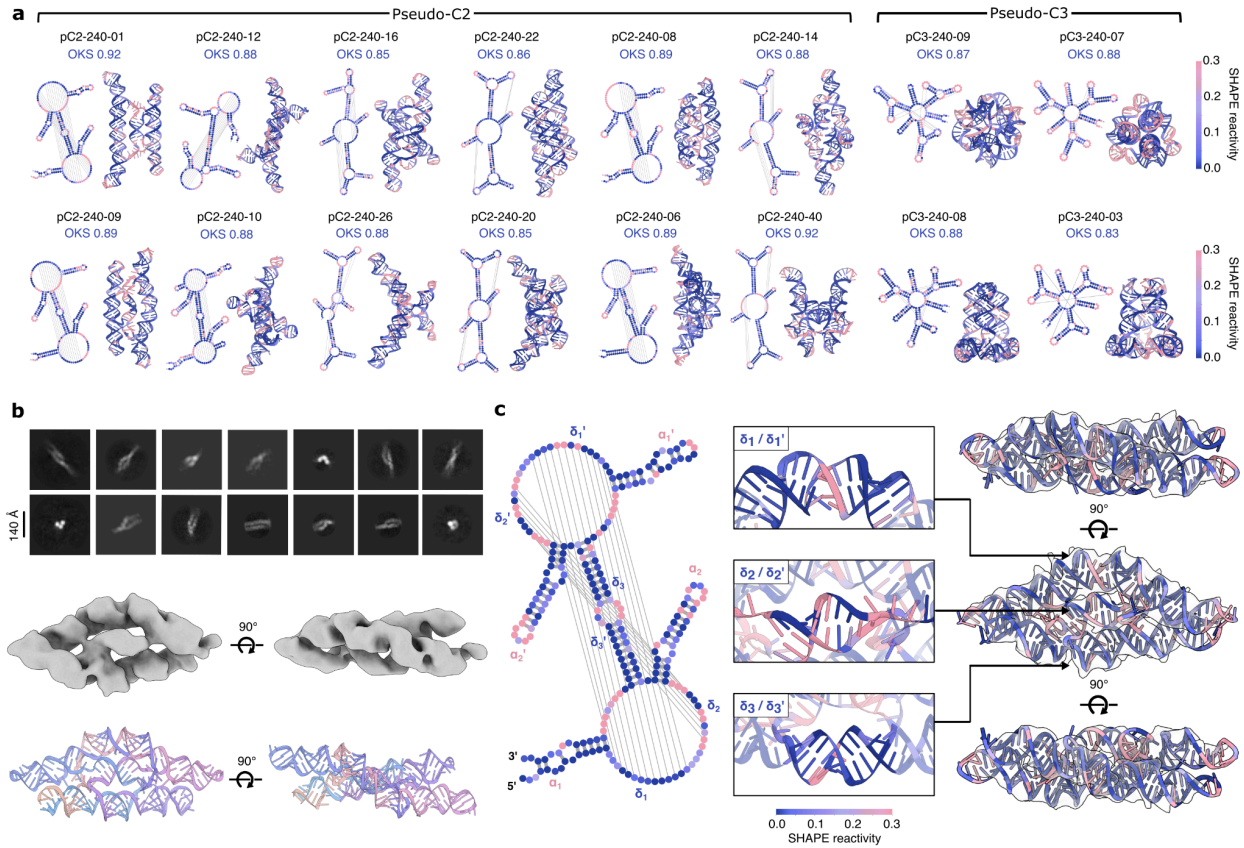
a, Denoising trajectories for representative RNA puzzles, where puzzle templates specify base-pair patterns and unpaired loop regions. **b**, SHAPE profiles for selected designs show high reactivity in unpaired regions (structures colored by reactivity), consistent with target secondary structure templates. Annotated structural features include hairpins (α) and bulges (β). **c**, Experimental SHAPE profiles compiled (rows correspond to experimentally tested sequences) for: wild-type sequences, MPNN-redesigns of alternative backbones (experimental or computational), and MPNN-designs of diffusion outputs. Selected designs (*) best matched target structures. **d**, Target OpenKnot scores derived from SHAPE profiles of experimentally tested sequences. **e**, SHAPE profiles mapped onto puzzle secondary structures with corresponding OpenKnot scores, for selected puzzles. **f**, Distribution of all Target OpenKnot scores (top) and best scores achieved per puzzle (bottom), separated by puzzles with experimentally determined 3D structures (left) and those with only computational models (right).

Figure 3. Design and characterization of 372-base pseudocycles



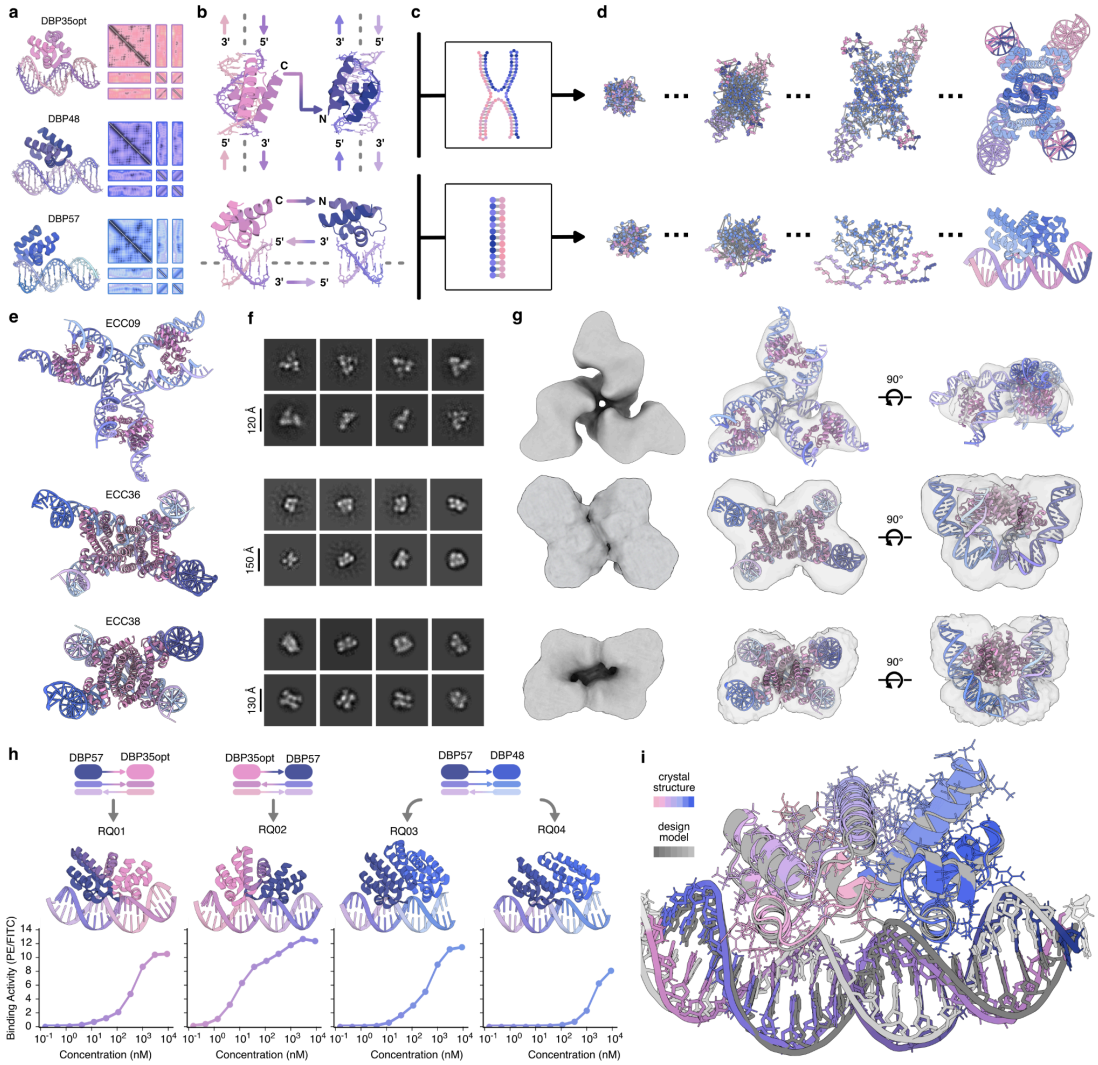
a, Workflow for generating base pair templates and RNA pseudocycles: (1) initialize subunit segments and assign *intra*-subunit partner regions; (2) propagate subunits symmetrically and assign *inter*-subunit pairs; (3) randomly vary segment lengths to diversify structures; (4) run RFDpoly denoising with secondary structure conditioning. **b**, Comparison of secondary structures and F1-similarity scores between design models and AlphaFold3 predictions for selected designs (top: pC2-372-07, bottom: pC2-372-11). **c**, Tertiary structures of selected designs. **d**, 3D reconstructed nsEM volumes showing butterfly-like folds consistent with design models and AF3 models. **e**, 2D class averages from nsEM data. **f**, Design models and AF3 models fit into reconstructed volumes. **g**, Conformational variability in pC2-372-07 visualized across multiple nsEM volumes.

Figure 4. Characterization of small, compact pseudocycles



a, Representative compact pseudocycle designs; secondary (left) and tertiary (right) structures colored by SHAPE-seq reactivity. High reactivity values in helical regions is indicative of local misfolding, which provides a course-grained estimation of which designs formed the target structure, and which did not. **b**, Cryo-EM 2D classes (top), and 3D reconstructed volume (middle) of designed pseudocycle, pC2-240-04, revealing the overall fold of the design model (bottom). **c**, SHAPE reactivity profile of pC2-240-04, mapped onto the secondary structure (left) and tertiary structure (right), refined to fit into the cryo-EM volume. Three central helical domains (δ_1, δ_2, δ_3) are highlighted (middle), showing crossover base pairs in the secondary structure and corresponding strand interweaving in the tertiary fold.

Figure 5. Multi-polymer fusions and motif scaffolding



a, DNA-binding proteins (DBPs) and their 2D templates, provided to RFDpoly for motif scaffolding. **b,** Multi-chain protein-DNA fusions: either *inter*-helical fusions bridging two DBPs across different helices (top) or *intra*-helical fusions parallel to the DNA axis (bottom). **c,** Base-pair pattern specification enables topological control over generated DNA interfaces. **d,** Denoising trajectories run

with or without symmetry (top and bottom, respectively) to generate complex assemblies. **e,** Design models of three nucleoprotein junctions, with protein domains bound to strand-exchanging DNA hetero-oligomers. **f,** nsEM 2D class averages of selected designs. **g,** nsEM 3D reconstructed volumes shown alone (left), or superimposed over design models (middle and right); final maps refined using symmetry constraints. **h,** Intra-helical fusion of different DBPs in various binding orientations (top) generated larger sequence-specific DNA-binding proteins (middle), which bind to longer stretches of target DNA, as evaluated by yeast-display flow cytometry assays (bottom). Binding quantified as phycoerythrin/FITC ratios. **i,** Crystal structure of DBP-fusion RQ01 superimposed on the design model using USalign[34] (RMSD = 1.76 Å, TM-score = 0.90).

Table 1: Cryo-EM Data Collection, Processing, and Modeling Statistics

	Design: pC2-240-04
Data Collection	
Microscope	Titan Krios
Voltage (kV)	300
Detector	Gatan K3
Energy filter	Gatan BioQuantum Gif
Recording mode	Counting
Magnification (kX)	105,000 X
Movie micrograph pixel size (Å)	0.83
Dose rate ($e^-/Å^2/s$)	10
No. of frames per movie micrograph	99
Frame exposure time (s)	0.0505
Movie micrograph exposure time (s)	5.0
Total dose ($e^-/Å^2$)	50
Defocus range (μm)	0.8 - 1.8
Number of movie micrographs	8,030
Map Processing	
Symmetry imposed	C1
Extraction box size (pix)	400
Initial particle images (no.)	5,482,343
Final particle images (no.)	140,174
Map resolution (Å)	6.63
FCS threshold	0.143
Map resolution range (Å)	14.96 - 5.63
Refinement	
PDB ID	9YN7
EMDB ID	EMD-73141
Initial model used	Computationally designed model
Map resolution (Å)	6.63
FCS threshold	0.143
Model resolution range (Å)	14.96 - 5.63

Map sharpening B factor	546.8
Model composition	
Non-hydrogen atoms	5,145
Protein Residues	0
Nucleotides	242
Ligands	—
R.M.S. deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.349
Validation	
MolProbity score	2.09
Clashscore	1.80
RNA Backbone	0.73

Table 2: Crystal structure data collection and refinement statistics

	Design: RQ01
Wavelength	
Resolution range	43.63 - 2.116 (2.25 - 2.12)
Space group	C 1 2 1
Unit cell	113.148 33.745 104.306 90 107.23 90
Total reflections	115915 (5311)
Unique reflections	36743 (2847)
Multiplicity	3.2 (1.9)
Completeness (%)	82.33 (20.91)
Mean I/sigma(I)	12.02 (0.68)
Wilson B-factor	34.95
R-merge	0.04013 (1.089)
R-meas	0.04781 (1.473)
R-pim	0.02569 (0.9807)
CC1/2	0.999 (0.432)
CC*	1 (0.777)

Reflections used in refinement	18082 (751)
Reflections used for R-free	892 (28)
R-work	0.2192 (0.2972)
R-free	0.2473 (0.2885)
Number of non-hydrogen atoms	2418
macromolecules	2274
ligands	0
solvent	144
Protein residues	166
RMS(bonds)	0.004
RMS(angles)	0.55
Ramachandran favored (%)	98.15
Ramachandran allowed (%)	1.85
Ramachandran outliers (%)	0
Rotamer outliers (%)	1.55
Clashscore	3.14
Average B-factor	48.86
macromolecules	39.69
solvent	45.94

References

[1] Sefah, Kwame, Dihua Shangguan, Xiangling Xiong, Meghan B. O'Donoghue, and Weihong Tan. 2010. "Development of DNA Aptamers Using Cell-SELEX." *Nature Protocols* 5 (6): 1169–85.

[2] Kyung-Nam Kang and Yoon-Sik Lee. RNA aptamers: a review of recent trends and applications. Future Trends in Biotechnology, pages 153– 169, 2012.

[3] Lee, Jeehyung, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, et al. 2014. "RNA Design Rules from a Massive Open Laboratory." *Proceedings of the National Academy of Sciences* 111 (6): 2122–27.

[4] Zadeh, Joseph N., Conrad D. Steenberg, Justin S. Bois, Brian R. Wolfe, Marshall B. Pierce, Asif R. Khan, Robert M. Dirks, and Niles A. Pierce. "NUPACK: Analysis and design of nucleic acid systems." *Journal of computational chemistry* 32, no. 1 (2011): 170-173.

[5] Poppleton, Erik, Niklas Urbanek, Taniya Chakraborty, Alessandra Griffo, Luca Monari, and Kerstin Göpfrich. 2023. "RNA Origami: Design, Simulation and Application." *RNA Biology*, July, 510–24.

[6] Zadegan, Reza M., and Michael L. Norton. 2012. "Structural DNA Nanotechnology: From Design to Applications." *International Journal of Molecular Sciences* 13 (6): 7149–62.

[7] Rothemund, Paul W. K. 2006. "Folding DNA to Create Nanoscale Shapes and Patterns." *Nature* 440 (7082): 297–302.

[8] Seeman, N. C. 1982. "Nucleic Acid Junctions and Lattices." *Journal of Theoretical Biology* 99 (2): 237–47.

[9] Geary, Cody, Guido Grossi, Ewan K. S. McRae, Paul W. K. Rothemund, and Ebbe S. Andersen. 2021. "RNA Origami Design Tools Enable Cotranscriptional Folding of Kilobase-Sized Nanoscaffolds." *Nature Chemistry* 13 (6): 549–58.

[10] Yesselman, Joseph D., Daniel Eiler, Erik D. Carlson, Michael R. Gotrik, Anne E. d'Aquino, Alexandra N. Ooms, Wipapat Kladwang et al. "Computational design of three-dimensional RNA structure and function." *Nature nanotechnology* 14, no. 9 (2019): 866-873.

[11] Lu, Xiang‐jun, and Wilma K. Olson. 2003. “3DNA: A Software Package for the Analysis, Rebuilding and Visualization of Three‐dimensional Nucleic Acid Structures.” *Nucleic Acids Research* 31 (17): 5108–21.

[12] Li, Shuxiang, Wilma K. Olson, and Xiang-Jun Lu. 2019. “Web 3DNA 2.0 for the Analysis, Visualization, and Modeling of 3D Nucleic Acid Structures.” *Nucleic Acids Research* 47 (W1): W26–34.

[13] Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang et al. "Accurate prediction of protein structures and interactions using a three-track neural network." *Science* 373, no. 6557 (2021): 871-876.

[14] Wang, Jue, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L. Watson, Karla M. Castro, Robert Ragotte et al. "Scaffolding protein functional sites using deep learning." *Science* 377, no. 6604 (2022): 387-394.

[15] Watson, Joseph L., David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, et al. 2023. “De Novo Design of Protein Structure and Function with RFdiffusion.” *Nature* 620 (7976): 1089–1100.

[16] Krishna, Rohith, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee et al. "Generalized biomolecular modeling and design with RoseTTAFold All-Atom." *Science* 384, no. 6693 (2024): eadl2528.

[17] Baek, Minkyung, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker, and Frank DiMaio. "Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA." *Nature methods* 21, no. 1 (2024): 117-121.

[18] Ma, Runze, Zhongyue Zhang, Zichen Wang, Chenqing Hua, Zhuomin Zhou, Fenglei Cao, Jiahua Rao, and Shuangjia Zheng. "RiboFlow: Conditional De Novo RNA Sequence-Structure Co-Design via Synergistic Flow Matching." *arXiv preprint arXiv:2503.17007* (2025).

[19] Nori, Divya, and Wengong Jin. "Rnaflow: Rna structure & sequence design via inverse folding-based flow matching." *arXiv preprint arXiv:2405.18768* (2024).

[20] Morehead, Alex, Jeffrey Ruffolo, Aadyot Bhatnagar, and Ali Madani. "Towards joint sequence-structure generation of nucleic acid and protein complexes with SE (3)-discrete diffusion." *arXiv preprint arXiv:2401.06151* (2023).

[21] Anand, Rishabh, Chaitanya K. Joshi, Alex Morehead, Arian R. Jamasb, Charles Harris, Simon V. Mathis, Kieran Didi, Rex Ying, Bryan Hooi, and Pietro Liò. "Rna-frameflow: Flow matching for de novo 3d rna backbone design." *ArXiv* (2025): arXiv-2406.

[22] Tarafder, Sumit, and Debswapna Bhattacharya. "RNAbpFlow: Base pair-augmented SE (3)-flow matching for conditional RNA 3D structure generation." *bioRxiv* (2025).

[23] Kubaney, Andrew, Andrew Favor, Lilian McHugh, Raktim Mitra, Robert Pecoraro, Justas Dauparas, Cameron Glasscock, David Baker. "RNA sequence design and Protein–DNA specificity prediction with NA-MPNN." *bioRxiv* (2025).

[24] Dauparas, Justas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile IM Wicky et al. "Robust deep learning–based protein sequence design using ProteinMPNN." *Science* 378, no. 6615 (2022): 49-56.

[25] Dauparas, Justas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. "Atomic context-conditioned protein sequence design using LigandMPNN." *Nature Methods* (2025): 1-7.

[26] Cruz, José Almeida, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das et al. "RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction." *Rna* 18, no. 4 (2012): 610-625.

[27] Eterna. OpenKnot. https://eternagame.org/challenges/11843006, 2025. Eterna challenge.

[28] He, Shujun, Rui Huang, Jill Townley, Rachael C. Kretsch, Thomas G. Karagianes, David BT Cox, Hamish Blair et al. "Ribonanza: deep learning of RNA structure through dual crowdsourcing." *bioRxiv* (2024).

[29] Kretsch, Rachael C., Alissa M. Hummer, Shujun He, Rongqing Yuan, Jing Zhang, Thomas Karagianes, Qian Cong, Andriy Kryshtafovych, and Rhiju Das. "Assessment of nucleic acid structure prediction in CASP16." *bioRxiv* (2025): 2025-05.

[30] Boitreaud, Jacques, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. "Chai-1: Decoding the molecular interactions of life." *BioRxiv* (2024).

[31] Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3." *Nature* 630, no. 8016 (2024): 493-500.

[32] Das, Rhiju, Rachael C. Kretsch, Adam J. Simpkin, Thomas Mulvaney, Phillip Pham, Ramya Rangan, Fan Bu et al. "Assessment of three‐dimensional RNA structure prediction in CASP15." *Proteins: Structure, Function, and Bioinformatics* 91, no. 12 (2023): 1747-1770.

[33] Zemla, Adam. "LGA: a method for finding 3D similarities in protein structures." *Nucleic acids research* 31, no. 13 (2003): 3370-3374.

[34] Zhang, Chengxin, Morgan Shine, Anna Marie Pyle, and Yang Zhang. "US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes." *Nature methods* 19, no. 9 (2022): 1109-1115.

[35] Marinus, Tycho, Adam B. Fessler, Craig A. Ogle, and Danny Incarnato. "A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy." Nucleic acids research 49, no. 6 (2021): e34-e34.

[36] Wayment-Steele, Hannah K., Wipapat Kladwang, Alexandra I. Strom, Jeehyung Lee, Adrien Treuille, Alex Becka, Eterna Participants, and Rhiju Das. "RNA secondary structure packages evaluated and improved by high-throughput experiments." *Nature methods* 19, no. 10 (2022): 1234-1242.

[37] Mortimer, Stefanie A., Cole Trapnell, Sharon Aviran, Lior Pachter, and Julius B. Lucks. "SHAPE–Seq: high‐throughput RNA structure analysis." *Current protocols in chemical biology* 4, no. 4 (2012): 275-297.

[38] Rhiju Das. Openknotscorematlab. https://github.com/eternagame/OpenKnotScoreMATLAB, 2025. GitHub repository. 1.

[39] Glasscock, Cameron J., Robert J. Pecoraro, Ryan McHugh, Lindsey A. Doyle, Wei Chen, Olivier Boivin, Beau Lonnquist et al. "Computational design of sequence-specific DNA-binding proteins." *Nature Structural & Molecular Biology* (2025): 1-10.

[40] Holliday, Robin. "A mechanism for gene conversion in fungi." *Genetics Research* 5, no. 2 (1964): 282-304.

[41] Praetorius, Florian, and Hendrik Dietz. "Self-assembly of genetically encoded DNA-protein hybrid nanoscale shapes." *Science* 355, no. 6331 (2017): eaam5488.

[42] Joung, J. Keith, and Jeffry D. Sander. "TALENs: a widely applicable technology for targeted genome editing." *Nature reviews Molecular cell biology* 14, no. 1 (2013): 49-55.

[43] Bhakta, Mital S., and David J. Segal. "The generation of zinc finger proteins by modular assembly." In *Engineered Zinc Finger Proteins: Methods and Protocols*, pp. 3-30. Totowa, NJ: Humana Press, 2010.

[44] Kwon, Diana. "RNA function follows form-why is it so hard to predict?." *Nature* 639, no. 8056 (2025): 1106-1108.

[45] Croll, Tristan Ian. "ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps." *Biological Crystallography* 74, no. 6 (2018): 519-530.

[47] Wang, Wenkai, et al. "trRosettaRNA: automated prediction of RNA 3D structure with transformer network." *Nature communications* 14, no. 1 (2023): 7266.