

Self-assembling 2D arrays with *de novo* protein building blocks

Zibo Chen^{1,2}, Matthew C. Johnson¹, Jiajun Chen^{3,4}, Matthew J. Bick^{1,2}, Scott E. Boyken^{1,2}, Baihan Lin², James J. De Yoreo^{3,4}, Justin M. Kollman¹, David Baker^{1,2,5}, Frank DiMaio^{1,2,*}

*Email: dimaio@u.washington.edu

Contents

Program Code	1
Data Availability	3
Supplemental Materials	4
Computational design methods	4
1. Connecting the homodimer into monomer	4
2. Systematic sampling of lattice parameters	4
3. Design calculations	4
4. Selection criteria and metrics used to evaluate designs	5
Visualization and figures	5
Experimental methods	5
Buffer and media recipe	5
Construction of synthetic genes	5
Protein expression	5
Protein purification	5
Circular dichroism (CD) measurements	6
X-ray crystallography and structure determination	6
In Situ Atomic Force Microscopy Imaging	6
Negative stain EM	6
Figure S1-S4.	8
Table 1. Design Sequences	13
Table 2. Data collection and refinement statistics	14

Program Code

All of the computational tools described in the manuscript are available with Rosetta, and can be downloaded at rosettacommons.org and <https://github.com/uagaug/2DLattice>. Rosetta design xmls are attached as supplemental files. Command line options for the design calculations can be found below.

Step 1. Rapid generation of connecting and non-clashing 2D lattices from protein building blocks

```
~/Rosetta/main/source/bin/flatland.static.linuxgccrelease
-in:file:s [input pdb model]
-database [path to Rosetta database]
-ignore_unrecognized_res
-mh:path:scores_BB_BB
/gscratch/baker/zibochen/utilities/aa_count_ACDEFHIKLMNQRSTVWY_resl1_ang15_
msc0.2_smooth1.3_ROSETTA/aa_count_ACDEFHIKLMNQRSTVWY_resl1_ang15_ms
c0.2_smooth1.3_ROSETTA -mh:score:use_ss1 true -mh:score:use_ss2 true -
mh:score:use_aa1 false -mh:score:use_aa2 false #motif score specific options
-symmetry_definition dummy
-output_virtual
-tag [user defined name tag for the job]
-rot_step [search step size for the self-rotation of the building block, takes a real number]
-Cn [internal cyclic symmetry of the building block, 2]
-wallpaper [layer symmetry of the final 2D lattice, C211]
-dump_silent [dump a silent file containing all the lattices, boolean]
-C211_B [lattice parameter B for the C 1 2 layer group, takes a real number]
-cell_upper [upper limit for the cell dimensions, takes a real number]
-single_chain_version [if the input model is monomerized, the code accomodates for this
psudeo-symmetry. Boolean]
-cell_step [search step size for the lattice cell dimensions, takes a real number]
```

Step 2. HBNet search at the interfaces of extracted adjacent building blocks

```
~/Rosetta/main/source/bin/rosetta_scripts.static.linuxgccrelease
-in:file:s [input pdb model]
-out::file::pdb_comments
-run:preserve_header
-use_input_sc
-out:prefix HBNet_
-beta
-missing_density_to_jump true
-parser:protocol 2D_HBNet.xml
-database [path to Rosetta database]
-chemical:exclude_patches LowerDNA UpperDNA Cterm_amidation VirtualBB ShoveBB
VirtualDNAPhosphate VirtualNTerm CTermConnect sc_orbitals pro_hydroxylated_case1
pro_hydroxylated_case2 ser_phosphorylated thr_phosphorylated tyr_phos phorylated
tyr_sulfated lys_dimethylated lys_monomethylated lys_trimethylated lys_acetylated
glu_carboxylated cys_acetylated tyr_diiodinated N_acetylated C_methylamidated
MethylatedProteinCterm
-in:file:fullatom
-multi_cool_annealer 10
-no_optH false
-optH_MCA true
-flip_HNQ
```

Step 3. Regenerate the complete 2D lattice and map newly designed interfaces to all symmetric copies

```
~/Rosetta/main/source/bin/symm_seq_gen_2D.default.linuxgccrelease
-database [path to Rosetta database]
-s [input pdb model]
-cn [symmetry of the building block, 2]
```

Step 4. Symmetric design of the 2D lattice in the context of its symmetry

```
~/Rosetta/main/source/bin/symm_seq_gen_2D.default.linuxgccrelease  
-database [path to Rosetta database] \ -in:file:silent [input Rosetta silent file containing  
the 2D lattice]  
-parser:script_vars resfile=[input resfile to enforce newly designed interfaces stay intact]  
-out::file::pdb_comments  
-run:preserve_header  
-multi_cool_annealer 10  
-use_input_sc  
-symmetry_definition dummy  
-out:prefix packed_  
-beta -missing_density_to_jump true  
-symmetry:detect_bonds false  
-parser:protocol 2D_final_design.xml
```

Data Availability

Coordinates and structure files have been deposited to the Protein Data Bank with accession code: 6EGC (SC_2L4HC2_23).

Supplemental Materials

Computational design methods

1. Connecting the homodimer into a monomer

The two monomers from the homodimer 2L4HC2_23 are connected into a single chain monomer with a 5-residue loop using the method described previously¹. Briefly, a database of backbone samples composed of short fragments connecting two helices via a loop of less than six residues was generated from high resolution crystal structures. Loop regions in this database were then structurally aligned to terminal residues on the helices of the design structure, and those that aligned within 0.35 Å RMSD were carried forward with full Rosetta design. The lowest-scoring candidate was selected as the final loop design.

2. Systematic sampling of lattice parameters

A custom Rosetta protocol was developed to dock the building block into pseudo-C 1 2 layer group and systematically sample the three parameters that control lattice geometry: two parameters describing the lattice dimensions, and one parameter controlling rotation of the building block around its central axis (Fig. 1C). Taking into account the dimension of the building block, lattice parameter “*a*” was sampled from 60 Å to 100 Å, with a step size of 0.5 Å; lattice parameter “*b*” was sampled from 30 Å to 50 Å, with a step size of 0.5 Å; rotation of the building block around its central axis, θ , was sampled from 0° to 180° with a step size of 1°, resulting in 576,000 possible docked conformations. A rapid evaluation protocol in Rosetta was applied to remove lattices that have either clashes of building blocks or inter building block distance greater than 10 Å, resulting in 4,139 candidate lattices for further design. Two adjacent building blocks were extracted from the lattice for interface design calculations.

3. Design calculations

RosettaDesign² calculations were carried out on the interfaces between adjacent building blocks, while keeping the rest of the sequences fixed. To enhance the binding specificity among subunits, we optionally used the Rosetta HBNets algorithm¹ to design buried hydrogen bond networks at the interface between subunits, selecting for networks that involve at least 3 side chain residues with all heavy-atom donors and acceptors participating in hydrogen bonds. Low energy sequences were identified using RosettaDesign calculations in which the hydrogen bond networks were held fixed. A final step of minimization and side chain repacking without atom pair constraints was applied to identify the movement of HBNets, filtering out designs with significantly moved hydrogen bond networks. The complete 2D lattice was then regenerated using the adjacent building blocks (now with designed interfaces), with the newly designed sequences applied to all building blocks. A final round of Rosetta design was carried out in the context of the C 1 2 layer group symmetry with the newly designed sequences fixed, to resolve potential side chain clashes in the final lattice.

4. Selection criteria and metrics used to evaluate designs

Fully designed models were selected based on the shape-complementarity of the designed interface ($SC > 0.6$), size of the designed interfaces ($dSASA > 500 \text{ \AA}^2$), average binding energy

($\text{ddG/dSASA} < -0.02$ Rosetta Energy Unit/Å²) and no buried unsatisfied hydrogen bonds introduced at the new interfaces. Selected designs were then visually inspected for good packing of hydrophobic side chains at the interfaces.

Visualization and figures

All structural images for figures were generated using PyMOL ³.

Experimental methods

Buffer and media recipe

TBM-5052

1.2% [wt/vol] tryptone, 2.4% [wt/vol] yeast extract, 0.5% [wt/vol] glycerol, 0.05% [wt/vol] D-glucose, 0.2% [wt/vol] D-lactose, 25 mM Na₂HPO₄, 25 mM KH₂PO₄, 50 mM NH₄Cl, 5 mM Na₂SO₄, 2 mM MgSO₄, 10 μM FeCl₃, 4 μM CaCl₂, 2 μM MnCl₂, 2 μM ZnSO₄, 400 nM CoCl₂, 400 nM NiCl₂, 400 nM CuCl₂, 400 nM Na₂MoO₄, 400 nM Na₂SeO₃, 400 nM H₃BO₃

TBS buffer

20 mM Tris pH 8.0, 100 mM NaCl

Construction of synthetic genes

Synthetic genes were ordered from Genscript Inc. (Piscataway, N.J., USA) and delivered in pET28b(+) *E. coli* expression vector, inserted between the NdeI and XhoI sites.

Protein expression

Plasmids were transformed into chemically competent *E. coli* expression strains BL21(DE3)Star (Invitrogen) for protein expression. Single colonies were picked from agar plates following transformation and overnight growth, and 5 ml starter cultures were grown at 37°C in Luria-Bertani (LB) medium containing 100 μg/mL kanamycin with shaking at 225 rpm for 18 hours at 37°C. Starter cultures were diluted into 500 ml TBM-5052 containing 100 μg/mL kanamycin, and incubated with shaking at 225 rpm for 24 hours at 37°C.

Protein purification

Cells were harvested by centrifugation for 15 minutes at 5000 rcf at 4°C and resuspended in 20 ml lysis buffer. Lysozyme, DNase, and EDTA-free cocktail protease inhibitor (Roche) were added to the resuspended cell pellet before sonication at 70% power for 5 minutes. Designs precipitated into cell pellet after clearing the cell lysate at 12,000g for 1 hour. Pellets were twice resuspended in 10 ml TBS followed by centrifugation at 12,000g for 20 min. The resulting pellet was resuspended in 1 M GdmHCl followed by centrifugation at 12,000g for 20 min. The supernatant was dialyzed overnight into TBS buffer.

Circular dichroism (CD) measurements

CD wavelength scans (260 to 195 nm) and temperature melts (25 to 95°C) were performed using an AVIV model 420 CD spectrometer. Temperature melts were carried out at a heating rate of

4°C/min when recording the change in ellipticity at 222 nm; protein samples were diluted to 0.25 mg/mL in PBS pH 7.4 in a 1 mm cuvette.

X-ray crystallography and structure determination

Crystals of SC_2L4HC2_23 were grown by mixing 0.1ul of protein at 20 mg/ml plus 0.1ul of crystallization condition Morpheus H9 (Molecular Dimensions, 0.1M Amino acids, 0.1M Buffer System 3 pH 8.5, 50% (v/v) Precipitant Mix 1). As this solution is already a suitable cryoprotectant, crystals were flash-frozen directly in liquid nitrogen prior to data collection. Diffraction data was collected at the Advanced Light Source, Lawrence Berkeley National Laboratory, beamline 8.2.1. Diffraction data was indexed and scaled using HKL2000⁴. Initial models were generated by the molecular-replacement method using PHASER⁵ within the Phenix software suite⁶, with the computational design model as the search model. Model bias were reduced by using prime-and-switch phasing and simulated annealing within Phenix.autobuild⁷. Iterative rounds of manual building in COOT⁸ and refinement in Phenix were used to produce the final model. Due to the high degree of self-similarity in coiled coils, the dataset for the reported structure suffered from a high degree of pseudo translational non-crystallographic symmetry, as report by Phenix.Xtriage, which complicated structure refinement and may explain the higher than expected R values reported. Phenix⁶ was used to calculate RMSDs of bond lengths, angles and dihedrals from ideal geometries. The overall quality of the final model was assessed using MOLPROBITY⁹. Summaries of diffraction data and refinement statistics are provided in Supplementary Table 2.

In Situ Atomic Force Microscopy Imaging

40 µL of protein solution (5 or 10 µg/mL protein with 21mM Tris-HCl, 400 mM KCl, pH 8) was added on top of a freshly cleaved mica surface (15 mm; Ted Pella). In situ images were captured using silicon nitride probes (OTR4 and OTR8, k: 0.08 N/m, 0.15 N/m, tip radius: 15 nm; Bruker) under tapping with a Cypher ES AFM (Asylum Research) at room temperature. Images were analyzed using Gwyddion SPM data analysis software.

Negative stain EM

Samples were applied to glow-discharged EM grids and stained with either uranyl acetate (UA), uranyl formate (UF) or NanoW (Nanoprobes, Inc, Yaphank, NY, USA) for screening or analysis. Data was collected using a Tecnai T12 equipped with a Gatan Orius CCD. CTF estimation was performed using GCTF¹⁰, and all other image processing steps were completed via Relion 2.1¹¹. For the analysis in Figure 2G, 421 2D array segments were picked manually from 42 micrographs, and the resulting 2D class average used as a template for Relion autopicking, which yielded 5,823 2D array segments. After subsequent 2D classification and alignment, the dominant 2D class consisted of 1,893 array segments (each approximately 20 nm diameter).

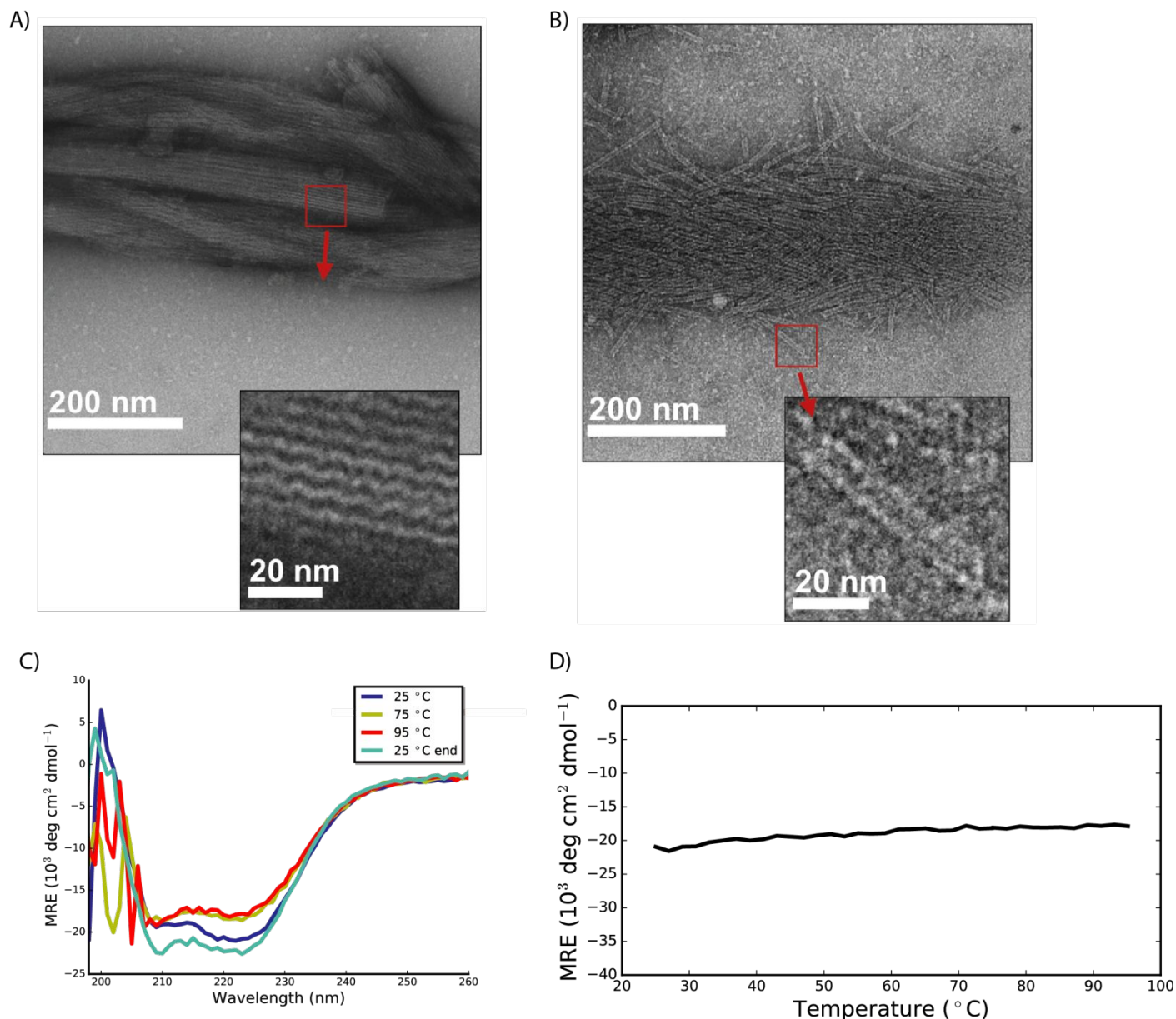
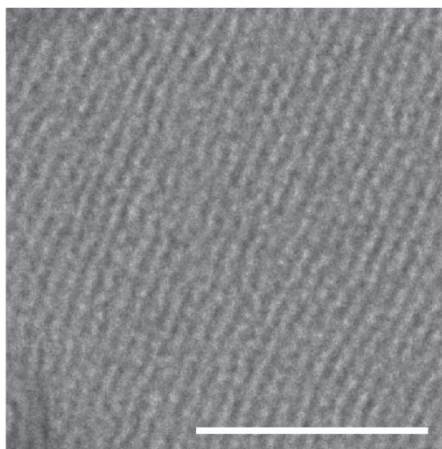


Figure S1. Design 2D-HP displayed two distinct morphologies under different staining solutions. (A) Thick, bundle-like structures formed in the uranyl acetate staining solution, likely due to the overall flexibility of designed 2D assemblies. (B) Individual fiber-like structures can be seen in the nanoW staining solution. (C) Circular dichroism (CD) spectra for the thermal denaturation of 2D-HP. Wavelength scans were performed at 25°C, 75°C, 95°C, and final 25°C. Design was alpha helical and stable up to 95°C. (D) CD temperature melt indicating high thermal stability of 2D-HP.

A)



B)

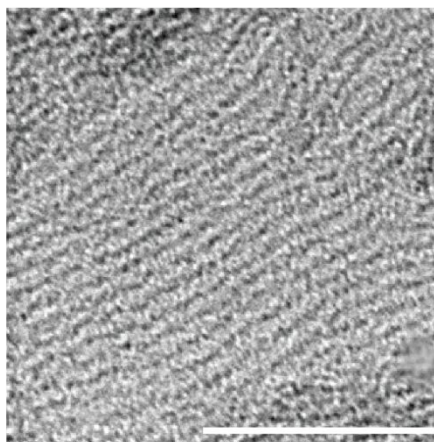
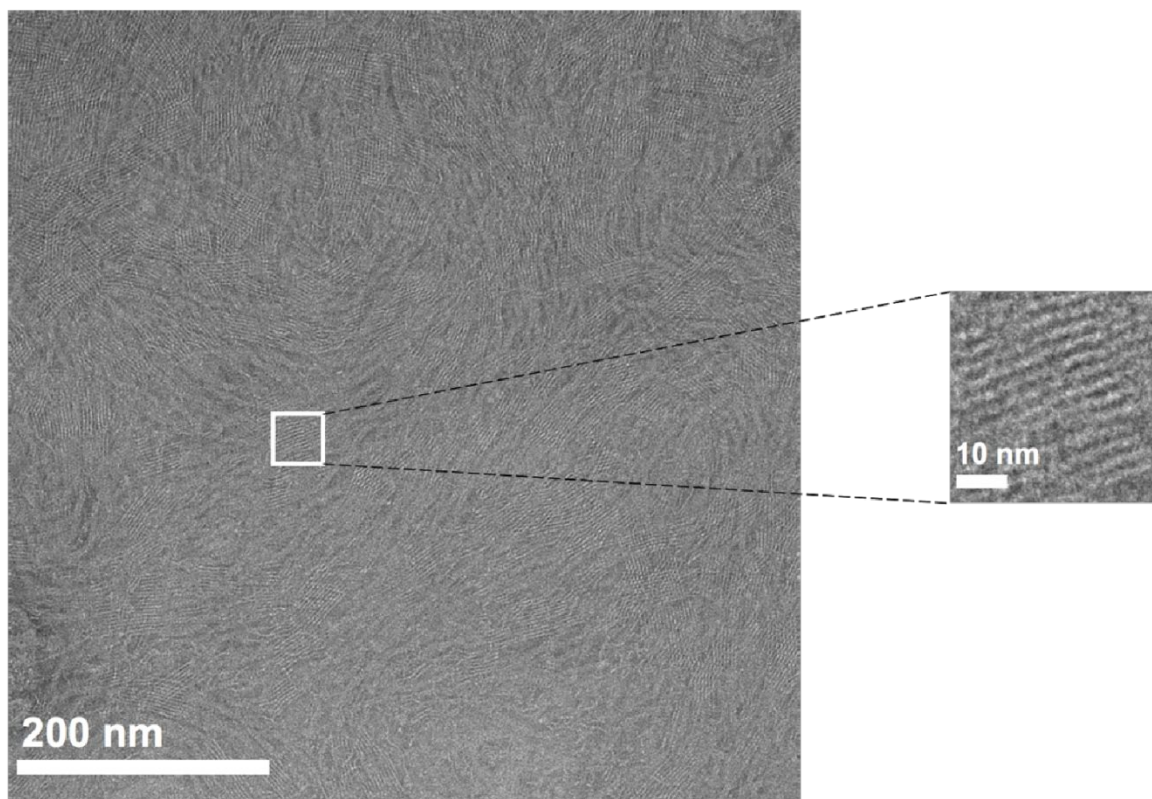
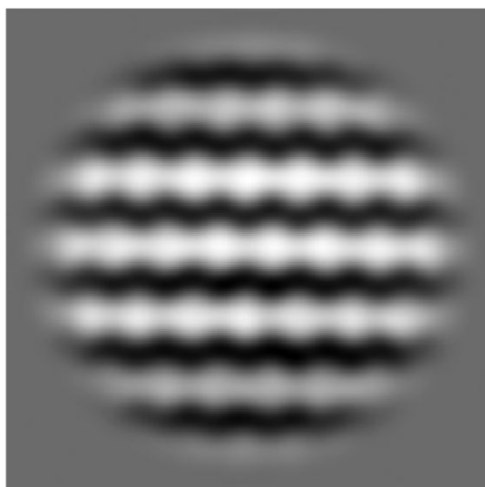


Figure S2. Comparison of 2D-HBN lattices using the monomer SC_2L4HC2_23 (A) or the homodimer 2L4HC2_23 (B) as building blocks. Negative stain EM shows similar patterns of 2D lattice formation in both cases. Scale bar: 50 nm.

A)



B)



C)

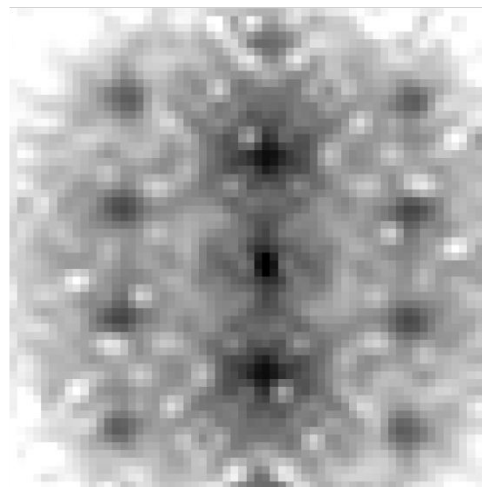


Figure S3. 2D class average of 2D-HBN under EM. (A) Representative image of 2D-HBN used for 2D class averaging. Inset on the right represents one of the 1,893 boxed sections picked for 2D averaging. (B) 2D class average of 2D-HBN assembly with homodimer building blocks. (C) Fourier transform of the 2D class average in (B).

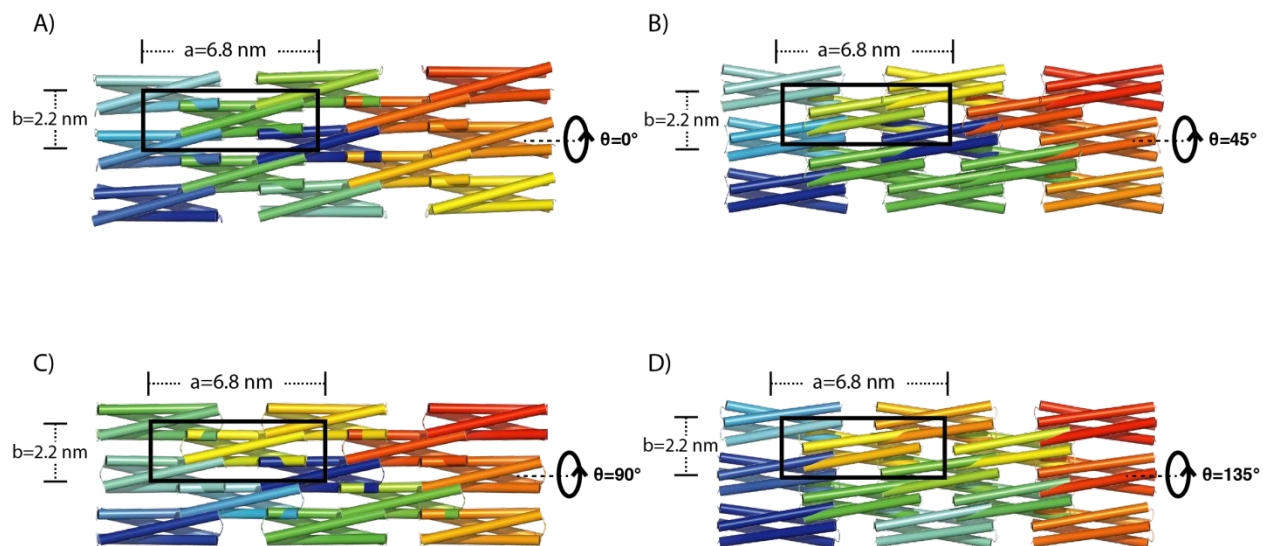


Figure S4. Docking the building block into observed lattice dimensions of 68 and 22 Å results in clashes. Black box marks the unit cell, with dimensions shown outside the box. A) - D): 0° , 45° , 90° , and 135° rotation of the building block around its central axis.

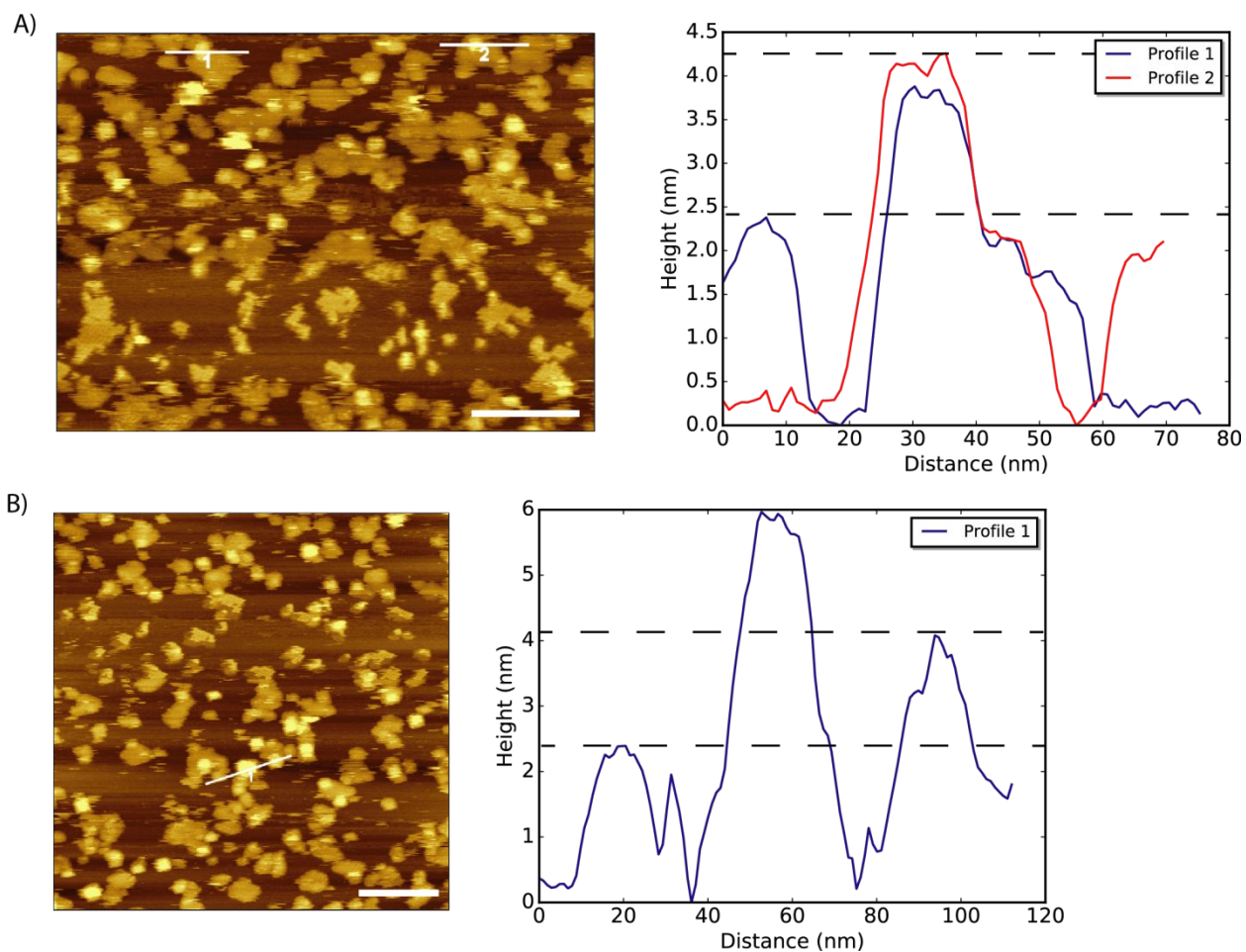


Table 1. Design Sequences

>SC_2L4HC2_23

TRTEIIRELERSLREQEELAKRLKELLRELERLQREGSSDEDVRELLREIKELVEEIEKLAREQKY
LVEELKRQQGPPGNEIIRELERSLREQEELAKRLKELLRELERLQREGSSDEDVRELLREIKELV
EEIEKLAREQKYLVEELKRQD

>2D-HP

SRTMYIRALEQSLREQEELAKRLKELLRELERLQREGSSDRDVKVLLWEIEALVEEIEKLARLQK
ELVEKLKRQGGSGNMYIRALEQSLREQEELAKRLKELLRELERLQREGSSDRDVKVLLWEIEALV
EEIEKLARLQKELVEKLKRQD

>2D-HBN

GELTDIILKLIKSLQTQKLLAERLKTLLKVLEISQDSGADDKQVKLLDEIRKLVEKIEKLARKQTKL
VEKLLKKGPGNDIILKLIKSLQTQKLLAERLKTLLKVLEISQDSGADDKQVKLLDEIRKLVEKIEKL
ARKQTKLVEKLLKGD

>2D-HBN_Homo (using the homodimer as the building block)

GELTDIILKLIKSLQTQKLLAERLKTLLKVLEISQDSGADDKQVKLLDEIRKLVEKIEKLARKQTKL
VEKLLKGD

Table 2. X-Ray Data collection and refinement statistics

	SC_2L4HC2_23
Wavelength	0.9998
Resolution range	21 - 1.74 (1.802 - 1.74)
Space group	P 1 21 1
Unit cell	41.253 49.36 41.239 90 104.303 90
Total reflections	59303 (5381)
Unique reflections	16336 (1241)
Multiplicity	3.6 (3.4)
Completeness (%)	92.01 (76.04)
Mean I/sigma(I)	9.11 (1.26)
Wilson B-factor	36.92
R-merge	0.05331 (0.8799)
R-meas	0.06269 (1.034)
R-pim	0.03271 (0.5389)
CC1/2	0.998 (0.669)
CC*	1 (0.895)
Reflections used in refinement	15268 (1241)
Reflections used for R-free	1461 (121)
R-work	0.2266 (0.3887)
R-free	0.2657 (0.4216)
CC(work)	0.939 (0.792)

CC(free)	0.913 (0.658)
Number of non-hydrogen atoms	1185
macromolecules	1134
solvent	51
Protein residues	147
RMS(bonds)	0.019
RMS(angles)	1.45
Ramachandran favored (%)	98.6
Ramachandran allowed (%)	1.4
Ramachandran outliers (%)	0
Rotamer outliers (%)	0
Clashscore	3.62
Average B-factor	54.64
macromolecules	54.26
solvent	63.1
Number of TLS groups	4

References

- (1) Boyken, S. E.; Chen, Z.; Groves, B.; Langan, R. A.; Oberdorfer, G.; Ford, A.; Gilmore, J. M.; Xu, C.; DiMaio, F.; Pereira, J. H.; Sankaran, B.; Seelig, G.; Zwart, P. H.; Baker, D. De Novo Design of Protein Homo-Oligomers with Modular Hydrogen-Bond Network-Mediated Specificity. *Science* **2016**, *352* (6286), 680–687.
- (2) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y. A.; Fleishman, S. J.; Corn, L. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* **2011**, *487*, 545–574.
- (3) Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. November 2015.
- (4) Otwinowski, Z.; Minor, W. Processing of X-Ray Diffraction Data Collected in Oscillation Mode. *Methods Enzymol.* **1997**, *276*, 307–326.
- (5) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser Crystallographic Software. *J. Appl. Crystallogr.* **2007**, *40* (Pt 4), 658–674.
- (6) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (Pt 2), 213–221.
- (7) Terwilliger, T. C.; Grosse-Kunstleve, R. W.; Afonine, P. V.; Moriarty, N. W.; Zwart, P. H.; Hung, L.-W.; Read, R. J.; Adams, P. D. Iterative Model Building, Structure Refinement and Density Modification with the PHENIX AutoBuild Wizard. *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64* (1), 61–69.
- (8) Emsley, P.; Cowtan, K. Coot: Model-Building Tools for Molecular Graphics. *Acta Crystallogr. D Biol. Crystallogr.* **2004**, *60* (Pt 12 Pt 1), 2126–2132.
- (9) Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B., 3rd; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. MolProbity: All-Atom Contacts and Structure Validation for Proteins and Nucleic Acids. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W375–W383.
- (10) Zhang, K. Gctf: Real-Time CTF Determination and Correction. *J. Struct. Biol.* **2016**, *193* (1), 1–12.
- (11) Scheres, S. H. W. RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination. *J. Struct. Biol.* **2012**, *180* (3), 519–530.