Electronic Supplementary Information for Predicting aggregate morphology of sequence-defined macromolecules with Recurrent Neural Networks

Debjyoti Bhattacharya,[1] Devon C. Kleeblatt,[1] Antonia Statt,[2] and Wesley F. Reinhart[1, 3, a)]

[1)] *Materials Science and Engineering, Pennsylvania State University, PA 16802*

[2)] *Materials Science and Engineering, Grainger College of Engineering, University of Illinois, Urbana-Champaign, IL 61801*

[3)] *Institute for Computational and Data Sciences, Pennsylvania State University, PA 16802*

(Dated: 6 June 2022)

[a)]email:reinhart@psu.edu

The origin of the 63 090 unique sequences used in High Throughput Screening can be explained by permutations and combinations of the fixed number of A type beads within the monomer sequence. The total number of possible sequences is $\frac{20!}{12!(8)!} = 125\,970$. However, this treats forward and reverse sequences as distinct, thereby double-counting, hence it needs to be halved to 62 985 sequences. Then, we need to add back the perfectly symmetrical sequences. In order to count the perfectly symmetrical sequences, there would be 2 possible cases wherein the 10th and 11th position would be occupied by either AA or BB and the other 2 possible cases AB or BA are not possible because then the sequence would be asymmetrical (as it would then leave behind 7 A and 11 B to arrange in either of the sides of the 10th and 11th positions). Hence, the possible symmetrical sequences for the case where AA occupies 10th and 11th positions would be $\frac{9!}{2\cdot 6!(3)!} = 42$. Similarly, for the case where BB occupies the 10th and 11th positions, the number of possible sequences would be $\frac{9!}{2\cdot 5!(4)!} = 63$. Adding these symmetrical sequences back, gives us a total of $62\,985 + 42 + 63 = 63\,090$ possible sequences.

Note that in practice, we arrived at this number by exhaustive enumeration and explicitly checking for symmetry. This explanation is only provided post hoc for the satisfaction of the interested reader.

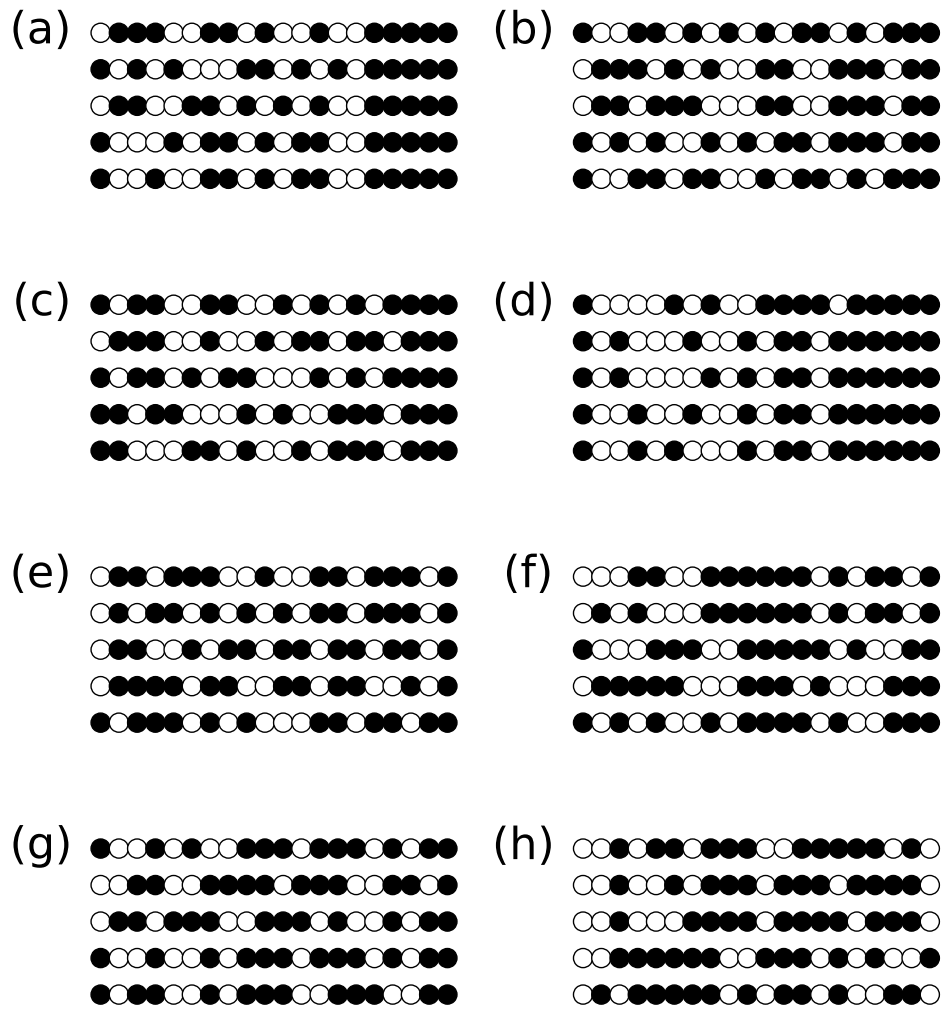

FIG. S1. Ranked list of sequences used to generate Fig. 8(b) in the main text.

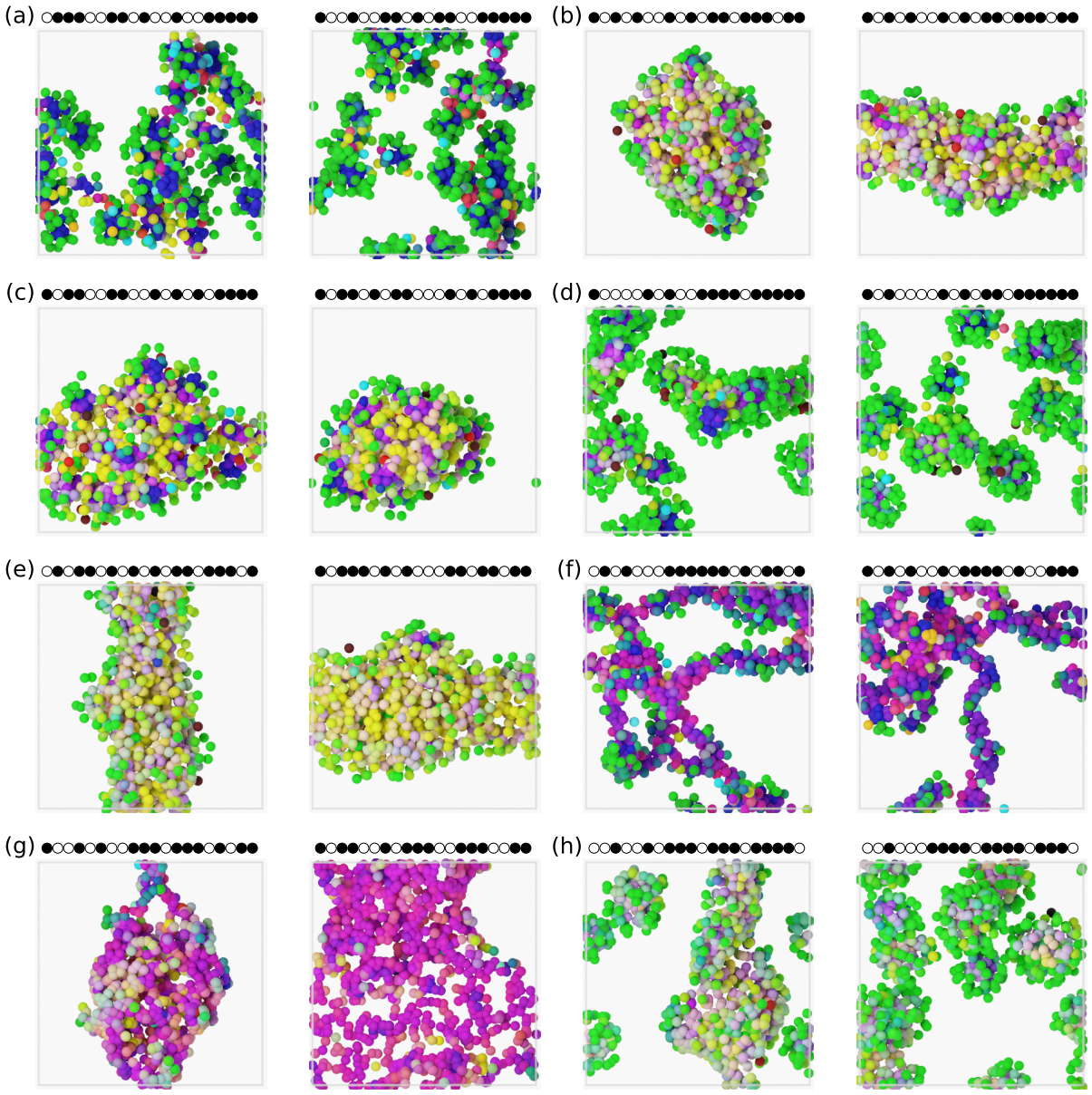

FIG. S2. Snapshots from the simulations in Fig. 8(b) in the main text, corresponding to the targets selected by K-Means clustering. Coloring is determined by the local environment around each particle, as described in Ref. 74. The left panels are closest to the target in the batch of candidates (the best result of 25 samples), right panels are farthest away (the worst result of 25 samples). Labels correspond to those in Fig. 8(b).

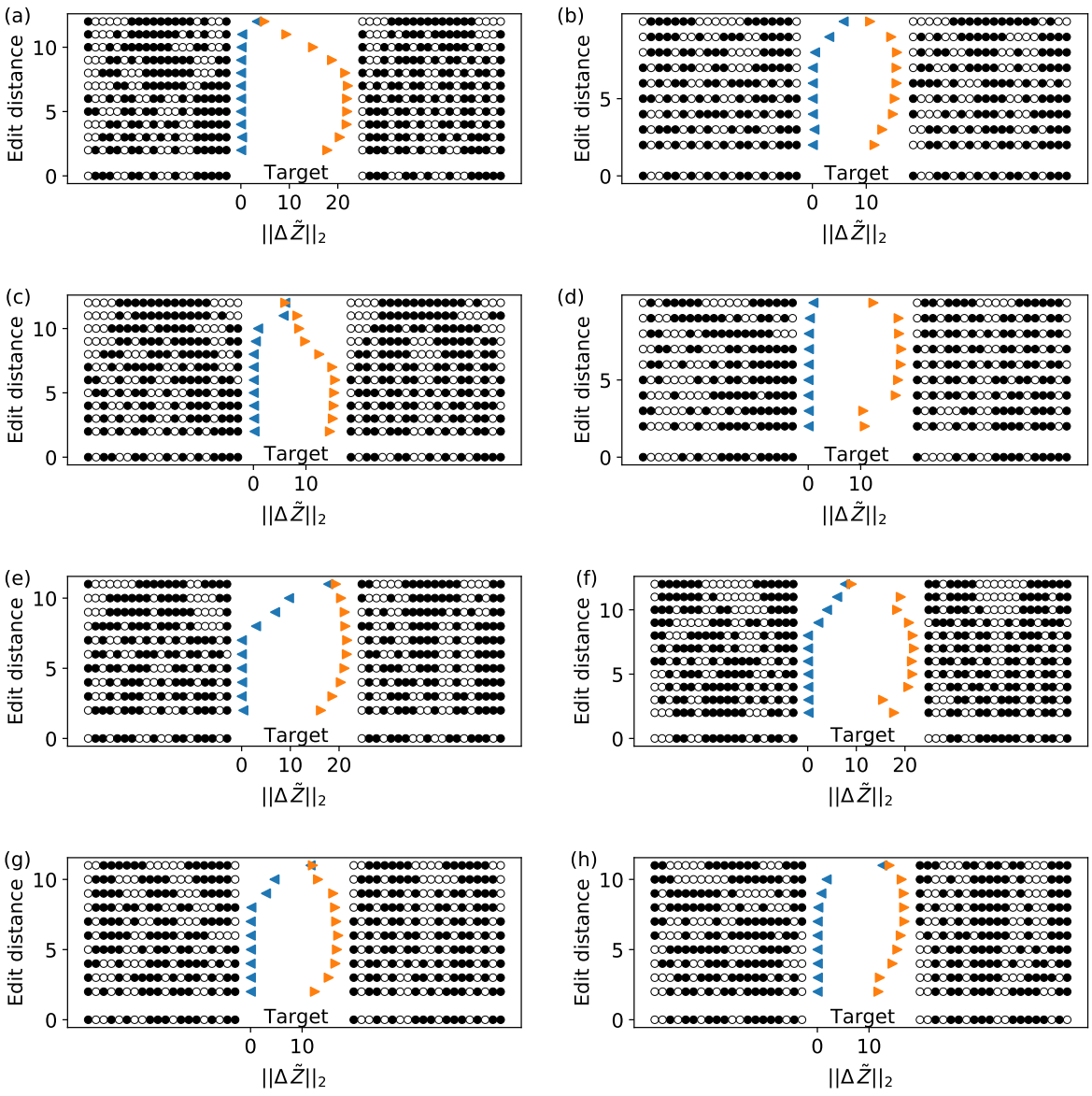

FIG. S3. Contrastive analysis of RNN inference on selected sequences from Fig. 8(b). Symbols show Levenshtein edit distance versus predicted distance from target sequence in Z space. Blue left triangles show minimum $||\Delta\tilde{Z}||_2$ at fixed edit distance from among all possible sequences (at fixed composition), while orange right triangles show maximum.