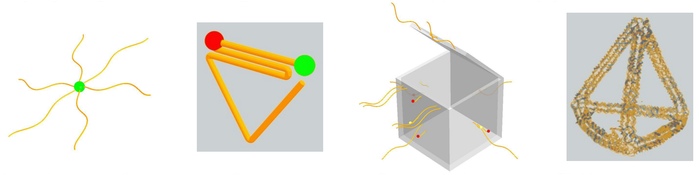
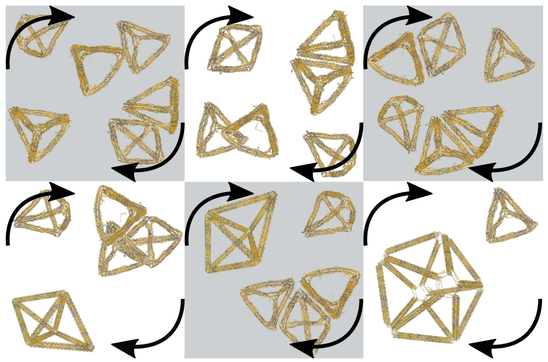
Predicting Nanorobot Shapes via Generative Models

Emma Benjaminson
Rebecca E. Taylor
Matthew Travers

Background on DNA Nanotechnology

DNA nanotechnology can be used to construct a variety of components that are capable of moving through fluids [1], sensing pH and metal ions [2] and carrying cargo, such as drug loads for medical applications [3].





Conceptual illustration of self-assembly process for DNA components (visualizations obtained using [6])

Challenges in Studying Nanoscale Self-Assembly

Studying nanoscale self-assembly processes can be challenging for several reasons:

- Nanoscale components cannot be imaged by conventional fluorescence microscopy; more time-consuming characterization methods are required
- Assembly processes for multicomponent structures often have low yields

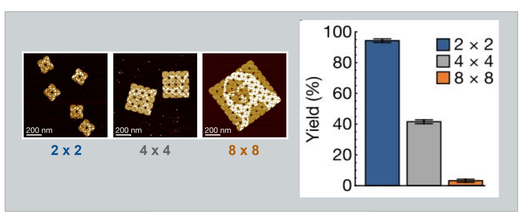


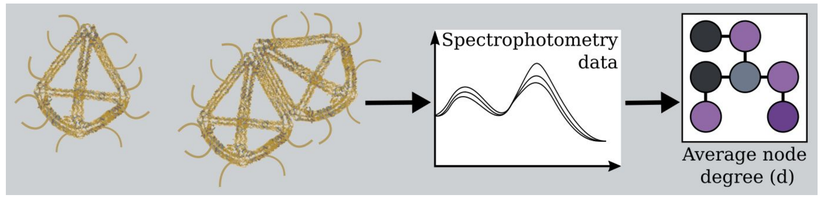
Figure source [4]

Research Question

How can we use artificial intelligence to predict the outcome of a nanoscale self-assembly process with limited data so that we can optimize the manufacturing process to maximize yield?

Low- vs. High-Fidelity Data

We characterize DNA nanotechnology assemblies using low- and high-fidelity data, and we propose that a combination of both can be used to train a model to quickly and accurately predict the structures produced during a manufacturing process. We use graphs to represent the assemblies as combinations of subcomponents.



We hypothesize that low-fidelity spectrophotometry data can be correlated to the number of bonds that each component forms with its neighbors, conveying information about average node degree.

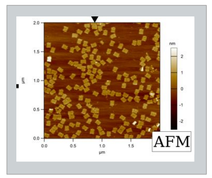


Image credit to Ying Liu

High-fidelity data includes atomic force microscopy (AFM) and cryo-electron microscopy (cryo-EM) images of individual structures.

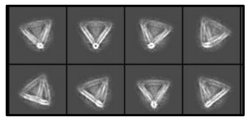
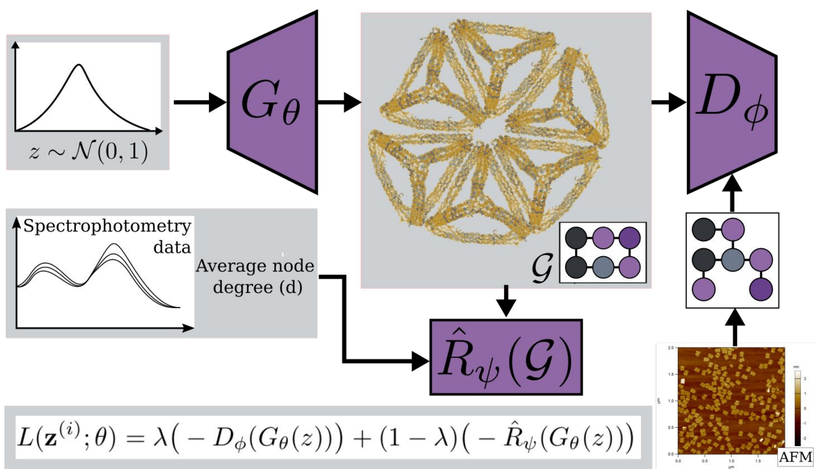


Figure source [5]

Using a GAN to Predict Nanoscale Assembly Shapes

This preliminary work uses a GAN (inspired by MolGAN [7]) to generate predictions of DNA nanotechnology structures (represented as graphs) using low-fidelity data to inform the prediction, and high-fidelity data as a training dataset. For this project we used the QM9 dataset [9-10] as a surrogate for DNA nanotechnology assemblies.

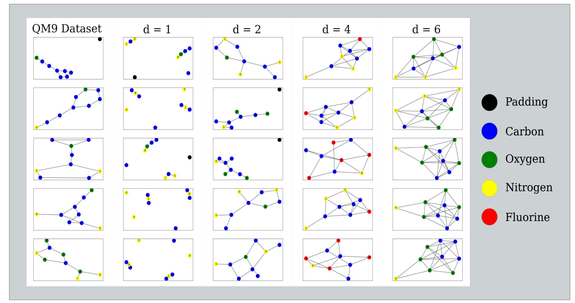


We conducted two experiments with our generative model:

1. How varying the desired average node degree affects the output
2. How varying λ affects the model's performance

Results

We found that the model was able to generate sample graphs that followed the desired node degree, which we varied over several experiments.



However, the number of unique samples produced by the generator with the reward network decreased significantly as compared to tests conducted with no reward network.

Our Results		
Objective	% Unique	Average d
$d = 1$	0.1	1.1 (average over 3 trials)
$d = 2$	0.8	2.0
$d = 4$	0.7	4.0
$d = 6$	0.7	6.0 (average over 3 trials)
Baseline ($\lambda = 1.0$)	12.8	1.4

	MolGAN [7]		**MolGAN** [8]		**Our Results**	
λ	% Unique	Solubility	% Unique	QED	% Unique	Average d
0.0	2.3	0.86	3.16	0.61	0.8	2.0
0.05	2.5	0.67	-	-	2.3	2.0
0.25	1.9	0.65	10.16	0.61	6.3	2.0
0.5	1.8	0.48	31.29	0.56	8.9	2.0
0.75	2.5	0.57	64.35	0.51	9.8	2.1
1.0	2.5	0.54	63.91	0.50	12.8	1.4

As we decreased λ we also saw a decrease in the number of unique samples produced by the generator. Smaller values of λ correspond to favoring the reward more heavily, which could lead to mode collapse.

Conclusions

This approach is able to tune the output of a generative model towards graphs with specific desired characteristics. However, our approach of using a reward network also causes significant mode collapse as compared to samples generated in the absence of a reward network. We will investigate alternative methods of tuning our generator's output in future work, and proceed to train with a dataset obtained from DNA nanotechnology experiments.

References

[1] R. Dreyfus et al. (2005). Nature, 437(6):2–5.
[2] Y. Liu et al. (2018). Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology, pp. e1518.
[3] S. M. Douglas et al. (2012). Science, 335(6070):831–834.
[4] G. Tikhomirov et al. (2017). Nature, 552(7683):67–71.
[5] H. Jun et al. (2019). ACS Nano, 13(2), 2083–2093.
[6] E . Poppleton et al. (2020). Nucleic Acids Research, 48(12), e72.
[7] N. De Cao et al. (2018). ArXiv Preprint ArXiv: 1805.11973. http://arxiv.org/abs/1805.11973
[8] N. De Cao (2019). Machine Learning for Physics and the Physics of Learning 2019, Workshop I: From Passive to Active: Generative and Reinforcement Learning with Physics, hosted by the Institute for Pure & Applied Mathematics, UCLA.
[9] L. Ruddigkeit et al. (2012). Journal of Chemical Information and Modeling 52(11):2864–2875.
[10] R. Ramakrishnan et al. (2014). Scientific Data 1, 140022.