Capturing sequence-dependent behavior in biomolecular materials using chemically informed machine learning models

P. Mastracco, A. Gonzàlez-Rosell, S. M. Copp (University of California, Irvine)
J. Evans (Chaffey Community College, CA) / P. Bogdanov (University of Albany, SUNY, NY)

Sequence-encoded biomolecules are promising programmable building blocks for materials, but the complexity of these biomaterials can challenge their design. To enable the design of programmable biomolecular building blocks, the UCI MRSEC team has developed a machine learning model that learns from limited experimental data sets to map biomolecule sequence onto materials properties.

- Developed for proof-of-principle material: DNA-stabilized silver nanoclusters (Ag_N-DNAs) with sequence-encoded fluorescence.
- High-throughput experiments connect DNA sequence to Ag_N-DNA color (Fig. 1), and the known atomic sizes of "magic-colored" Ag_N-DNAs inform the ML classification problem.
- Chemically informed ML classifiers combine known structural properties with training data to learn to distinguish DNA sequences classified by Ag_N-DNA color.
- These models increase success of designing Ag_N-DNAs with target colors by up to 10-fold (Fig. 2a). Moreover, the models are are *interpretable*, providing insights into the sequence-color "code" for Ag_N-DNAs (Fig. 2b).

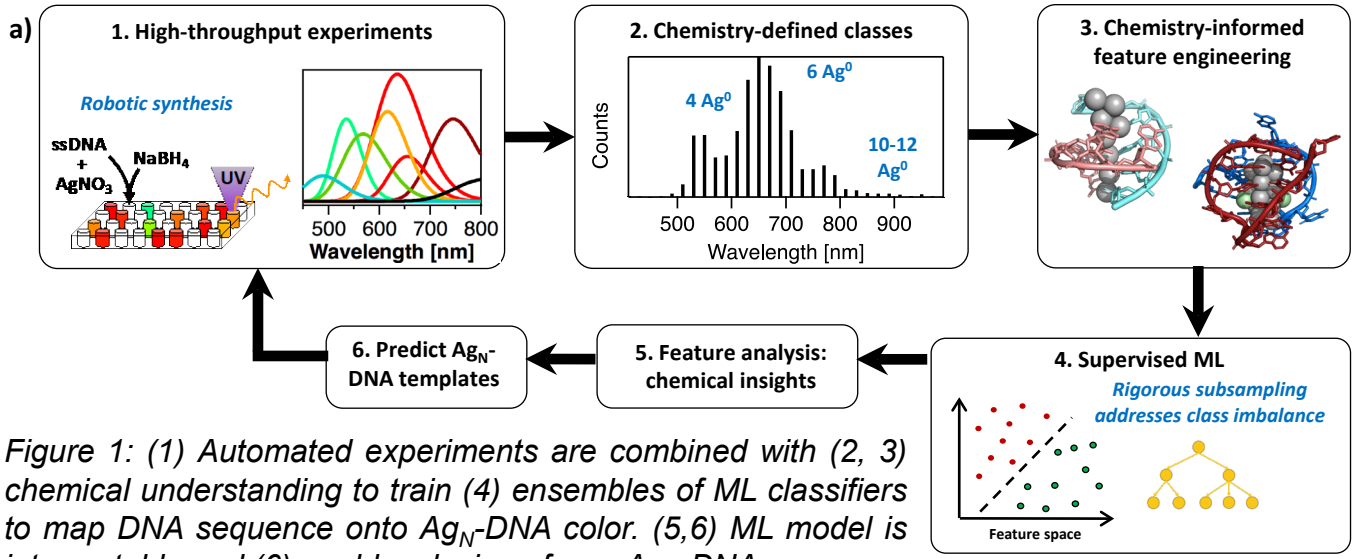

Figure 1: (1) Automated experiments are combined with (2, 3) chemical understanding to train (4) ensembles of ML classifiers to map DNA sequence onto Ag_N-DNA color. (5,6) ML model is interpretable and (6) enables design of new Ag_N-DNAs.

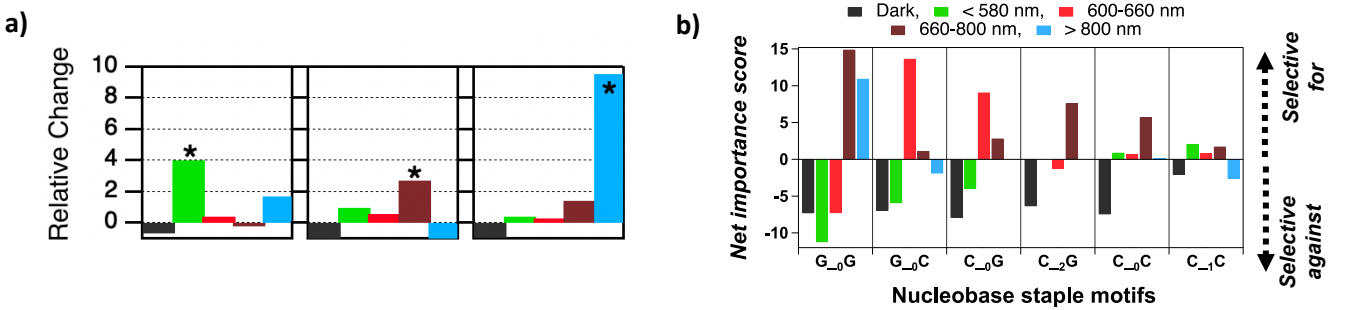

Figure 2: a) Success of ML model for increasing target color class (*). b) Feature analysis scores for the six most color-selective base patterns per Ag_N-DNA color class (see legend).

Mastracco P, Gorovits A, Gonzàlez-Rosell A, Evans J, Bogdanov P, **Copp SM**. "Chemistry-informed machine learning enhances predictive design of fluorescent DNA-stabilized silver clusters" To be submitted.