ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Beyond Nature's base pairs: machine learning-enabled design of DNA-stabilized silver nanoclusters†

Peter Mastracco[a] and Stacy M. Copp‡[a,b,c]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

ABSTRACT: Sequence-encoded biomolecules such as DNA and peptides are powerful programmable building blocks for nanomaterials. This paradigm is enabled by decades of prior research into how nucleic acid and amino acid sequences dictate biomolecular interactions. The properties of biomolecular materials can be significantly expanded with non-natural interactions, including metal ion coordination of nucleic acids and amino acids. However, these approaches present design challenges because it is often not well-understood how biomolecular sequence dictates such non-natural interactions. This Feature Article presents a case study in overcoming challenges in biomolecular materials with emerging approaches in data mining and machine learning for chemical design. We review progress in this area for a specific class of DNA-templated metal nanomaterials with complex sequence-to-property relationships: DNA-stabilized silver nanoclusters (Ag_N-DNAs) with bright, sequence-tuned fluorescence colors and promise for biophotonics applications. A brief overview of machine learning concepts is presented, and high-throughput experimental synthesis and characterization of Ag_N-DNAs are discussed. Then, recent progress in machine learning-guided design of DNA sequences that select for specific Ag_N-DNA fluorescence properties is reviewed. We conclude with emerging opportunities in machine learning-guided design and discovery of Ag_N-DNAs and other sequence-encoded biomolecular nanomaterials.

1 Introduction

Nature's nucleic acids are powerful molecular tools for nanotechnologies. Since Ned Seeman's seminal paper in 1982 presented the concept that oligomeric DNA and RNA can be used to build static junctions and networks,[1] researchers have become increasingly adept at harnessing the sequence-programmed rules of Watson-Crick-Franklin base pairs to assemble DNA nanostructures,[2] organize colloids into nanoscale architectures,[3] and create dynamic machines and computers.[4,5] (This article refers to canonical hydrogen-bonded base pairing of the natural nucleobases adenine (A), cytosine (C), guanine (G), and thymine (T) as Watson-Crick-Franklin base pairs,[6] as is becoming commonly adopted by the scientific community.[7,8]) The diverse functionalities of DNA nanotechnologies are enabled by the degree to which scientists now understand the nature of Watson-Crick-Franklin base pairing, an understanding that has been built by decades of intense prior research on the structure and formation of the DNA duplex within the biochemistry community.

More recently, nucleic acid nanotechnologies are expanding beyond the confines of the Watson-Crick-Franklin A-T and C-G base pairs. One particular area of significant development is metal-nucleic acid interactions, including metal-mediated DNA base pairing[9–11] and DNA-templated metallic nanostructures.[12,13] Metal-nucleic acid interactions have been well-studied for decades.[14] While initially of interest for their biomedical relevance, recent studies are demonstrating the promise of metal-nucleic acid interactions for both expanding the self-assembly rules of DNA and realizing new properties with DNA nanomaterials.[15] DNA-organized metal atom arrays and DNA-templated metal nanostructures have been explored for their applications in magnetic materials,[16,17] nanoelectronics,[18,19] photonics,[20] and catalysis.[21] This field is now expanding even more rapidly due to the last two decades of rapid advances in "traditional" DNA nanotechnologies based on Watson-Crick-Franklin base pairs.[22]

Because the scientific understanding of metal-mediated base pairs remains far more nascent than of Watson-Crick-Franklin base pairs, it is a major challenge to develop the same degree of predictive power over the molecular conformations adopted by nucleic acids due to metal-mediated interactions. This is especially challenging due to the combinatorially large DNA sequence space (for L-base sequences of the four canonical nucleobases A, C, G, T, there exist 4^L unique DNA oligomer sequences) and

[a] *Department of Materials Science and Engineering, University of California, Irvine, California 92697, United States*
[b] *Department of Physics and Astronomy, University of California, Irvine, California 92697, United States*
[c] *Department of Chemical and Biomolecular Engineering, University of California, Irvine, California 92697, United States*
‡ stacy.copp@uci.edu

to the computational challenges of simulating DNA-metal complexes, which are very large for *ab initio* approaches but cannot be fully described by classical techniques. While rapid advances in molecular simulations are beginning to address this challenge, it is important to also develop experimental approaches to interrogate the full complexity of nucleic acid-metal interactions. Crystallographic studies continue to advance understanding of metal-mediated base pairing.[9,23] A recent comprehensive study by Vecchione, *et al.*, of crystallographic structures of 32 different base pairs mediated by Ag^+, Hg^{2+}, and Au^+ suggests a rich diversity of possible metal-mediated nucleic acid interactions that remain to be fully understood and harnessed.[23] It is important to note that crystallographic studies are relatively low-throughput and require crystallization, motivating the development of additional experimental approaches to understand when and how nucleic acid sequence influences metal-mediated base pairing.

This Feature Article discusses how combinatorial experiments and computational tools including data mining and machine learning (ML) have been used to advance the state-of-the-art for a particular class of nucleic acid-templated metal nanomaterials: DNA-stabilized silver nanoclusters (Ag_N-DNAs). These tiny clusters of silver atoms stabilized by DNA oligomers were first discovered in 2004 by Petty and coauthors[12] and quickly attracted attention for their bright fluorescence colors that range from visible to near-infrared (NIR) wavelengths[24,25] and for the unique sequence tunability of emission wavelength with nucleic acid sequence.[26] Ag_N-DNAs have shown significant promise in a diversity of chemical and biomolecular sensing schemes[27] and are promising emitters for NIR bioimaging[28–30] and emerging low-background, anti-Stokes shift fluorescence imaging modalities.[31,32] Since 2012, studies of chromatographically purified Ag_N-DNA species have significantly advanced understanding of the compositions and optical properties of Ag_N-DNA emitters.[20] However, it has remained a major challenge to determine how DNA sequence selects Ag_N-DNA composition and optical properties. This challenge hinders the design of Ag_N-DNAs that are engineered for specific applications.

Here, we review how the development of high-throughput experimentation platforms for Ag_N-DNA synthesis and characterization has enabled curation of sufficiently large, consistent data sets to enable informatics approaches to determine the sequence-to-color rules for Ag_N-DNAs (Figure 1). To make this review accessible to the broad community of materials chemistry, we begin with an overview of what is known about Ag_N-DNAs and other nucleic acid-stabilized metal nanoclusters in **Section 2**. We then provide a short introduction to ML and associated terminology in **Section 3**. **Section 4** discusses the experimental strategies employed to generate large, well-controlled data libraries for Ag_N-DNAs. Then, we review how data mining tools and ML classifiers have been developed for predictive design of DNA oligomers that template brightly emissive Ag_N-DNAs (**Section 5**) and Ag_N-DNAs with specific emission colors (**Section 6**). Finally, **Section 7** discusses how interpretable ML models have been used not only to design Ag_N-DNAs but also to understand the sequence-to-color rules for these nucleic acid-templated nanomaterials.

Through this case study, this Feature Article intends to provide researchers within the field of materials chemistry with insights into how ML and data mining can be successfully harnessed for advancing both discovery and fundamental chemical understanding of materials systems. In particular, the approaches and models discussed are highly promising for nucleic acid-based materials[33–36] and other systems whose properties are governed by biomolecular sequence, such as peptide- and protein-based materials.[37–40]

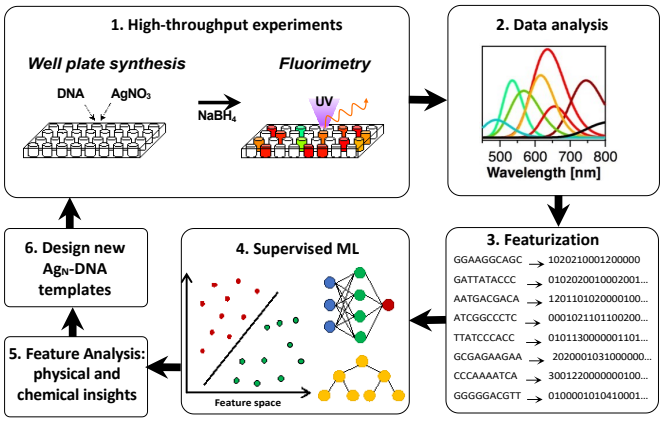

Fig. 1 General workflow for high throughput synthesis, characterization, and design of Ag_N-DNAs.

2 Fundamentals of nucleic acid-stabilized metal nanoclusters

It has been known that Ag^+ has affinity for the nucleobases but not the phosphate backbone of natural DNA at neutral pH.[14] This affinity allows DNA to act as a stabilizing ligand for silver nanoclusters (Ag_N), as was first reported by the groups of Petty and Dickson in 2004.[12] A crystal structure of a 16-atom Ag_N-DNA[46,47] is shown in Figure 2a. Ag_N-DNAs are synthesized by $NaBH_4$ reduction of a neutral pH aqueous solution of Ag^+ and single-stranded DNA.[12] Because A, C, G, and T have varying affinities for Ag^+, primarily through nucleobase ring nitrogens,[14,48,49] nucleobase sequence selects Ag_N size, shape, and optical properties, with Ag_N-DNA emission peaks from 400 nm up to 1,200 nm.[20,50] In some cases, altering a single nucleobase can dramatically shift Ag_N-DNA color.[51] This sequence tunability of a metal nanocluster by biomolecular sequence is one of the most unique properties of Ag_N-DNAs. RNA can also be used to stabilize Ag_N with bright emission.[52]

Recently, the fundamental understanding of Ag_N-DNA structure and properties has been significantly advanced by detailed investigations of atomically precise samples of Ag_N-DNA species that are prepared by high performance liquid chromatography (HPLC). We summarize the current understanding here; readers can find more details in a recent comprehensive review.[20] The compositions of Ag_N-DNAs can be determined using electrospray ionization mass spectrometry (ESI-MS) to count the total numbers of silver atoms N and DNA strands n_s per nanocluster[53] and to determine the charge of Ag_N-DNA species.[42] Knowledge of

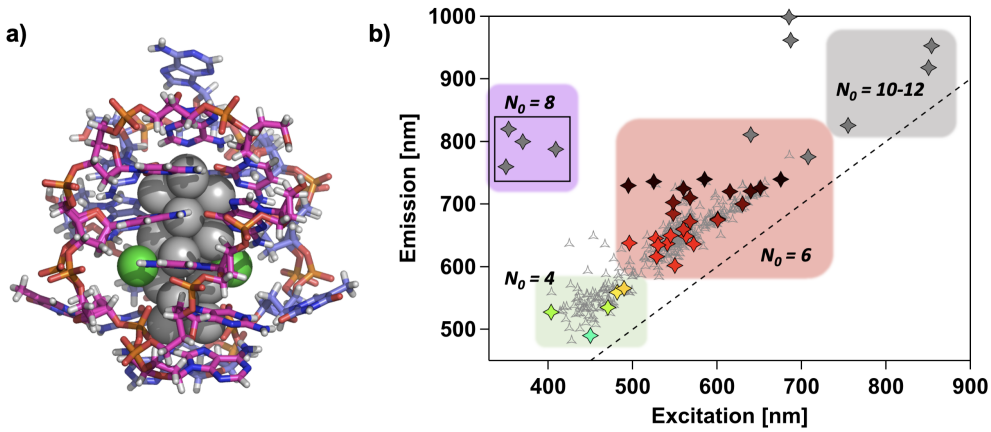

Fig. 2 a) Crystal structure of a NIR-emissive Ag_N-DNA with 16 Ag atoms (grey) and two chlorido ligands (green). Adapted with permission from Gonzàlez-Rosell, et al.,[41] licensed under CC BY-NC-ND. Copyright 2023. b) Excitation and emission wavelengths of Ag_N-DNAs, including those reported previously[42–45] and current work in preparation in the Copp lab. Empty triangles represent Ag_N-DNAs characterized by high-throughput experiments, and filled diamonds represent Ag_N-DNAs whose sizes have been determined by ESI-MS. Marker color represents the emission wavelength converted to RGB values. NIR wavelengths are shown in grey. Note that N_0 = 8 Ag_N-DNAs have multiple emission peaks;[45] here the most intense excitation peak beyond the DNA absorption peak is shown.

nanocluster charge then allows one to determine the effective valence electron content, N_0, of Ag_N-DNAs,[43] which is an important property according to the superatomic model for ligand-stabilized metal nanoclusters.[54] By these methods, researchers have identified Ag_N-DNA species ranging in size from N = 10 to N = 30 Ag atoms stabilized by n_s = 1 to n_s = 3 copies of the templating DNA oligomer. Recent results have also shown that some Ag_N-DNAs possess additional stabilizing chlorido ligands.[47] Ag_N-DNAs are partially oxidized,[43,55] with effective valence electron counts of N_0 = 4, 6, 8, and 10 – 12 identified by mass spectral analysis. The peak excitation and peak emission wavelength, λ_p, scale strongly with N_0 (Figure 2b).[43,44] Thus, studies that show that DNA sequence selects for Ag_N-DNA excitation and emission wavelengths provide evidence that DNA sequence is selecting for the atomic size and shape of the Ag_N-DNA.

Not only does DNA sequence select the size and optical properties of individual Ag_N-DNAs, but DNA's supramolecular chemistry also enables programmable supracluster assembly of Ag_N-DNAs. Watson-Crick-Franklin base pairing can be used to organize Ag_N-DNA into pairs of atomically precise nanoclusters with interesting nanophotonic behavior[56,57] and higher-order architectures on DNA nanostructures.[58–60] Stimulus-induced "color-switching"[61] and cytotoxicity[62] also vary with DNA ligand sequence. These properties make Ag_N-DNA promising and uniquely tunable fluorophores for bioimaging and biochemical sensing.

To realize the full potential of Ag_N-DNA emitters for such applications, it is critical to develop a clear understanding of how nucleobase sequence selects the optical and chemical properties of Ag_N-DNAs. Because DNA oligomer templates for Ag_N-DNAs are typically 10-40 nucleobases in length, the design of Ag_N-DNA template sequence is inherently challenging due to the combinatorially large space of DNA sequences. Early designs for DNA template strands for Ag_N-DNAs often assumed that canonical secondary DNA structures were conducive for forming fluorescent Ag_N-DNA, such as i-motif,[63] G-quadruplex,[64] and hairpin structures.[65] While many of these designs appeared successful, it should be noted that more recent studies of silver-mediated DNA base pairing call into question the stability of Watson-Crick-Franklin and other canonical nucleic acid structures in the presence of Ag^+.[11,66] Thus, the apparent success of some of these designs may not have been correlated with natural DNA secondary structure before Ag_N-DNA synthesis but rather with abundance of C and/or G nucleobases, which have much higher affinities for Ag^+ than A and T, as discussed below.[66,67] One study of four Ag_N-DNA emitters used UV circular dichroism (CD) spectroscopy, which is sensitive to DNA secondary structure, to show that the CD signatures of Ag_N-DNA emitters are much more similar to the CD signatures of Ag^+-DNA mixtures pre-reduction than to the signatures of the bare DNA template strands.[68]

Other groups have used single base mutations to understand how sequence influences Ag_N-DNA fluorescence properties.[51,69] For instance, one such study identified the role of guanine in the formation of a specific near-infrared emissive Ag_N-DNA.[70] While these early studies showed the strong dependence of Ag_N-DNA fluorescence spectrum on DNA sequence, the results of single base mutation studies of Ag_N-DNAs are difficult to generalize as design rules for Ag_N-DNAs.

Ab initio calculations have also been used to understand the silver-nucleobase interactions involved in the stabilization of Ag_N-DNAs. Density functional theory (DFT) studies have investigated these interactions in the context of silver-mediated base pairing, showing that Ag^+ has much higher affinities for cytosines and guanines than for adenines and thymines (Figure 3).[66,67] Because Ag^+-DNA complexes are the precursors for Ag_N-DNA chemical synthesis, such studies do provide insights into general trends in DNA sequence-to-color relationships.[71] Computational studies have also given important insights into the ability of silver cations to pair together DNA duplexes in noncanonical orientations[72,73] and into the formation and optical properties of Ag_N-DNAs.[74,75] However, full *ab initio* calculations of Ag_N-DNA structures are less

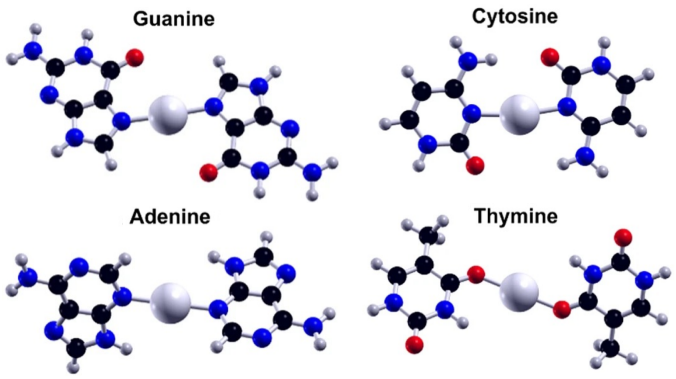

Fig. 3 Computational modeling of Ag^+-nucleobase interactions for homobase pairs[66] Adapted from Swasey, et al.,[66] licensed under CC BY. Copyright 2015.

mature than for other types of ligand-protected nanoclusters.[76] Thanks to breakthrough X-ray crystal structures of the first Ag_N-DNAs,[46,77] *e.g.* Figure 2a, realistic models of Ag_N-DNAs are now emerging.[47,78] At present, such models cannot yet be used to predict how DNA sequence selects Ag_N-DNA properties.

Due to the complexity of DNA sequence-to-color relationships, a data-driven approach is a promising alternative solution, whereby the correlations between DNA sequence and Ag_N-DNA properties is learned from many experimental observations. Screening on DNA microarrays was used by the Dickson group to identify cytosine-rich DNA templates for Ag_N-DNAs.[79] This first demonstration of Ag_N-DNA synthesis in "high-throughput" suggested the power of combinatorial experiments on hundreds of DNA sequences to discern the sequence-to-color rules that govern Ag_N-DNAs. However, this study only reported a few of the successfully designed sequences on the microarray, prohibiting informatics-based approaches using this dataset. Copp, *et al.*, developed the high-throughput experimental approach described in Section 4,[43] and the Yeh group has developed a platform for rapid screening of "light-up" Ag_N-DNA probes called NanoCluster Beacons (NCBs)[80] based on Illumina MiSeq chips.[81]

An alternate approach is to mine the existing literature for reports of Ag_N-DNAs and their stabilizing DNA sequences. New, *et al.* compiled a detailed list of 133 different DNA template sequences and the properties of their associated Ag_N-DNAs.[82] This significant task provides a view into the diversity and complexity of the Ag_N-DNA sequence-to-color relationship. However, different laboratories use differing synthesis methods, such as $[Ag^+]$:[DNA] ratio and solution or buffer conditions, as well as different spectroscopic methods, such as excitation wavelength(s) and spectral range. These discrepancies complicate attempts to use this library for ML or other investigations into general trends in how DNA sequence selects the properties of Ag_N-DNAs.

3 Fundamentals of machine learning

In this section, we provide an overview of concepts related to machine learning (ML), which is a branch of computer science that develops algorithms that learn from data without being explicitly programmed. Because ML is useful for identifying trends inside complex datasets, it has attracted significant and increasing attention from scientists in other fields in recent years, as both computational and experimental studies across fields have begun to produce and curate ever-expanding data sets. In the context of chemistry and materials science, ML can be employed to eluci-date structure-property relationships and to determine the phys-ical and chemical effects that govern materials properties.[83–85] We briefly review key ML terminology and concepts here. For an excellent detailed review of ML in the context of soft materials, the reader is directed to a topical review by Ferguson.[86]

ML for chemical and materials discovery requires several key "ingredients." First, one must define the question one seeks to answer about a chemical or materials system, *e.g.*, how does DNA sequence dictate Ag_N-DNA fluorescence emission color? Second, a large experimental or computational "training data" set is needed for ML algorithms to learn to answer the question posed. Third, one must choose a ML algorithm or class of algorithms for this tasks. There are a wide variety of potential ML algorithms to choose from, and it is best to choose a ML model based on the goals and the data available for the task. Figure 4 outlines key steps in the decision-making process for ML algorithm selection, as explained here. The first thing to consider when choosing a ML algorithm is whether one wishes to perform supervised or unsupervised ML. Supervised learning involves training a ML model with labeled data, *i.e.* a data set where input data is correlated to or "labeled by" output data, where the goal is to learn a mapping between inputs and their corresponding output labels. (This could, for example, be a data set that correlates DNA sequence to Ag_N-DNA fluorescence emission color.) Commonly used supervised ML algorithms include random forests, support vector machines (SVMs) and neural networks. In contrast, unsupervised learning involves training a ML model on unlabeled data, where there are no predefined output labels, with the goal of discovering patterns or relationships among input data. Common unsupervised ML algorithms include principle component analysis (PCA) and k-means clustering. When choosing between supervised and unsupervised ML, one should consider both the time required to label input data (Is it impossible or prohibitively difficult to perform this labeling?) and the information one seeks to learn from the ML process.

Another issue to consider when selecting a ML algorithm is the complexity of the model and the characteristics of the available training data set. Simple ML models are easier to interpret, less prone to overfitting, and require less computational resources for training but may underperform when the underlying relationship between input and output data is complex. Models such as random forests and SVMs are easy to train and have a higher level of interpretability than more complex learning algorithms, making these algorithms better suited for for applications involving small experimental datasets and/or when the researcher needs to understand what the ML algorithm has learned to capture trends within the training data. More complex models, including deep learning architectures, can capture much more complex trends in the data and can potentially achieve better accuracies, but these models may also suffer from overfitting (which can cause artificially high accuracies), require much greater computational re-

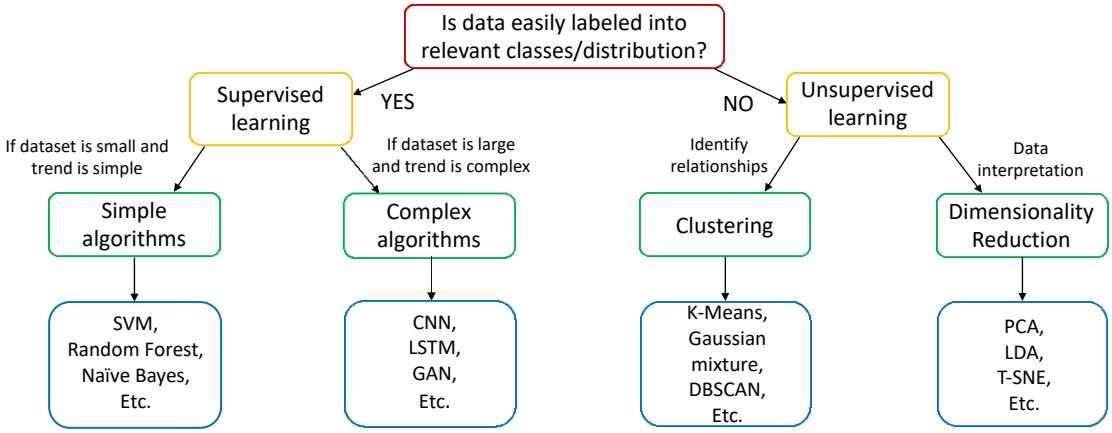

Fig. 4 Flow chart for selecting a ML algorithm for a learning task.

sources, and are more difficult or even impossible to interpret. Large datasets like those generated by high-throughput quantum chemical calculations are often suited for more complex models, but small experimental datasets collected in a single lab are not likely to benefit from using deep learning without caution.

A fourth equally important and often challenging step in developing ML models for chemical and materials research is feature engineering. Also referred to as featurization or "choice of descriptors," feature engineering is the process of choosing how to represent data in the form of feature vectors, which are the inputs to the ML algorithm. The choice of feature vectors is especially important because ML models will perform best when feature vectors represent the properties of the input data that are most correlated to the trend(s) one seeks to learn. It is ideal to choose feature vectors that provide the relevant physical and chemical information that are necessary for the ML algorithm to perform the learning task posed. However, one often does not know these physical and chemical principles in their entirety or, in some cases, at all, which is often the motivating factor for the use of ML in the first place. Thus, feature engineering can be both a highly challenging *and* scientifically enlightening step in the ML process. It is also important to ensure that the dimensions of the feature vectors ("feature space") are reduced to decrease the chance of overfitting.[87] Feature selection approaches can be used to reduce feature space and to gain new fundamental insights into the materials system.[88]

Once a ML model and features have been selected, it is critical to assess model performance by *in silico* testing using training data before employing the model for prediction and design tasks. An intuitive metric for model performance is accuracy, which is the fraction of predictions made correctly by the model (*i.e.*, the number of correctly made predictions divided by the total number of predictions). Accuracy and other performance metrics are commonly assessed by k-fold cross-validation, whereby (1) training data is randomly divided into k folds, (2) the model is trained using data in k − 1 folds, (3) the unseen fold of data is used to calculate accuracy, and (4) the process is iterated over all k folds to determine an average k-fold cross-validation metric. k-fold cross-validation accuracy provides an important estimate of how the ML model may perform on unseen data, but on its own is often insufficient to assess ML model performance. Accuracy should be provided in conjunction with other metrics, such as F_1-score (the harmonic mean of precision and recall) and analysis of the receiver operating characteristic (ROC) curve. Both metrics better assess over- or under-fitting. *In silico* assessment is critical to ensure a ML model's fitness for making predictions and should always be applied when selecting ML models and tuning model parameters. Of course, the best method to assess a ML model is by experiment, which is far more costly and time-consuming. Thus, *in silico* validation should always precede experimental testing.

A key issue for ML newcomers to note is the problem of learn-ing from imbalanced data. Data imbalance refers to the uneven distribution of input training data instances across the set of possible outputs one seeks to learn. Consider a simple example of a data set of 500 possible catalysts, 50 of which are effective at catalyzing a desired reaction and 450 of which are poor catalysts for the desired reaction. This data is "imbalanced" in its distribution of effective *vs.* ineffective catalysts. Without correction, a simple ML classifier trained to discriminate between effective and ineffective catalysts using this data set could achieve 90% accuracy by assigning all inputs as ineffective catalysts, since 450/500 = 0.9 of the training input instances are ineffective. However, the usefulness of a model trained on this imbalanced data set, without any additional steps, would clearly be very low if one seeks to search for effective catalysts. Learning from imbalanced data is a well-studied problem with multiple strategies developed to address this issue, such as data subsampling or supersampling to balance data as well as algorithmic changes such as uneven misclassification costs during the training process.[89,90] Because data imbalance is inherent to most chemical and materials data sets, it is critical that researchers ensure best practices are used for effective ML in chemical design.

Most significant progress in the emerging field of materials informatics has been in the areas of crystalline materials and small molecules.[91–93] This is largely because of the dozens of materials databases that have been curated for these systems, such as the Materials Project,[94] Open Quantum Materials Database,[95] and Organic Materials Database.[96] These databases compile the

results of hundreds of thousands of quantum chemical calculations, as well as crystallographic studies. Such large databases have allowed researchers to develop and implement ML and other data informatics approaches to expedite the study and design of new molecules and solid-state materials. Progress in developing ML approaches for soft and biological materials, such as polymers[97–99] and peptides,[86] has been slower by comparison. Soft and biological materials tend to be more complex and less crystalline than solid-state materials systems; thus, simplifications by applying periodic boundary conditions may not be appropriate. This complexity makes the use of computational techniques like DFT much more difficult and time-consuming to perform. Moreover, molecular dynamics (MD) approaches often require course-graining to feasibly model large systems, which has also been a major challege for scalability of methods to multiple systems. Together, this has limited the use of computational simulations of soft and biological materials for generating large databases and thus slowed the development of ML techniques for their discovery.

There is a key advantage of advancing ML-based approaches to soft and biological materials in the context of sequence-based biomolecules - nucleic acids and peptides - as model systems. These molecules are parameterized by a sequence of nucleic acids or amino acids, providing an approach to feature engineering for ML models that are trained to map biomolecule sequence onto the physical and chemical properties of materials systems derived from these molecules. By establishing these sequence-to-property connections through ML, we can not only better understand the complex sequence-to-property relationships but can also design new biomolecules that yield desired materials properties.

4 High-throughput experimental synthesis and characterization

ML-based approaches to Ag_N-DNA design require high-quality experimental data sets that connect DNA sequence to Ag_N-DNA properties. Because computational models for these complex systems are still in development,[75,100,101] experimental training data is required for ML. In order to determine how DNA sequence selects Ag_N-DNA emission brightness and color, this data set should ideally hold the following factors constant:

- Synthesis stoichiometry, including $[Ag^+]$:[DNA] and $[Ag^+]$:$[NaBH_4]$ ratios
- Solution conditions: ionic strength and/or buffer system, pH, mixing, *etc.*
- Time between chemical reduction and Ag_N-DNA spectral characterization
- Spectroscopic details: excitation wavelength(s), emission scan window, gain, *etc.*
- Length of DNA oligomer template

Data sets collected with these factors held constant enable direct comparisons of Ag_N-DNA peak emission and emission brightness among a large set of DNA sequences. Additionally, it is favorable to sufficiently sample DNA sequence space in order to capture a representative set of examples of how nucleobase sequence selects emission properties.

Since 2014, Copp and coauthors have curated a data library of nearly 4,000 10-base DNA sequences and the peak emission wavelength(s) and brightnesses (*i.e.* intensities) of the Ag_N-DNAs that these DNA oligomers stabilize (illustrated schematically in Figure 5). Ag_N-DNA synthesis is performed in 384 well plates using robotic liquid handing, with settings including pipetting speed and the number of pipet mixing cycles that are optimized for Ag_N-DNA brightness and reproducibility.[43] Well plates are centrifuged at slow speed immediately after synthesis to remove any small bubbles that would impact subsequent fluorimetry. $NaBH_4$ solutions for chemical reduction are prepared freshly before each experiment. All experiments have been performed at constant synthesis stoichiometry: 20 μM DNA, 100 μM $AgNO_3$, 50 μM $NaBH_4$ in 10 mM NH_4OAc aqueous solution, pH 7. This stoichiometry was chosen to maximize the yield of brightly fluorescent products across the visible spectral range.[43] NH_4OAc solutions have routinely been used for Ag_N-DNA synthesis because these solutions are directly compatible with ESI-MS analysis of nanocluster products.[20,102] Finally, universal UV excitation of all Ag_N-DNA products at the 260 nm DNA absorbance band, as demonstrated by O'Neill *et al.*,[103] is used to screen for all emissive Ag_N-DNA species with a single excitation wavelength in a commercial plate reader equipped with a monochromator, with emission spectra collected from 400 to 850 nm. Swasey, Nicholson, and coauthors recently developed custom instrumentation that extends high-throughput screening of emission in well plate format into the NIR using low-cost InGaAs photodetectors.[104]

To enable direct comparison of *both* Ag_N-DNA peak emission wavelength(s) *and* emission intensity, or "brightness," among all DNA sequences that were screened over the last decade, a well-studied Ag_N-DNA template sequence, 5'-TTCCCACCCACCCCGGCCCGTT-3', has been included as a control sample in about 10 wells per experiment. This control Ag_N-DNA is well-known to produce bright red emission with 93% quantum yield[42] and to evolve over time into a green emissive product.[105] The relative abundance of the green and red products produced by this control strand is used to screen for any discrepancies in the efficiency of the chemical reduction process in the well plates, and the emission intensity of this control Ag_N-DNA is used to normalize emission brightness across all experiments since 2014.[50]

Ag_N-DNA emission spectra are fitted to a series of Gaussian functions using an automated fitting code, thereby extracting the peak wavelength and brightness of each emissive product.[43] Peak brightness is correlated with the area of each Gaussian peak fit. Sequences are reported with up to three associated spectral peaks and peak brightnesses. (Ag_N-DNA emitters have emission peaks with 50 to 100 nm full-width-at-half-maximum.[53,106] For this reason, peak fitting uncertainty becomes significant in the rare case where more than three peaks are present in the 400 to 850 nm window for which spectra are collected, prohibiting accurate assignment of peak wavelengths and areas.) The presence of multiple different emission spectral peaks for a single DNA sequence indicates that the DNA strand can stabilize multiple dif-

ferent species of silver nanoclusters, with different cluster sizes and/or geometries.[50,71,107]

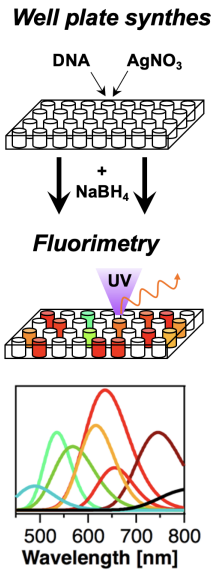

Fig. 5 Schematic of high-throughput synthesis and fluorimetry approach.

Other groups have used other experimental approaches to screen hundreds of Ag_N-DNAs in a single experiment. As stated earlier, the Dickson group used DNA microarrays to screen for DNA strands that hosted brightly fluorescent Ag_N-DNA products.[79] Recently, Yeh and coauthors have used next-generation-sequencing chips to screen nearly 40,000 activator strands for NanoCluster Beacon (NCB) sensing schemes based on Ag_N-DNAs.[108] NCBs are composed of a nonfluorescent or "dark" Ag_N-DNA that becomes brightly emissive when bound to a target DNA strand, typically rich in guanines.[109] These NCBs have utility in biomedical sensing, such as single nucleotide polymorphism (SNP) detection,[80] nucleobase methylation detection,[110] and enzymatic activity detection.[111,112] Systematic studies have shown that the activator strand sequence affects the emission spectrum of the activated NCB,[61] but given the large combinatorial space of possible activator sequences, this problem is challenging to solve by intuition-based design. To rapidly screen 10^4 activator sequences in a single experiment, Yeh and coauthors adapted Illumina MiSeq chips for NCB activation characterization. Activator oligomers were immobilized on the MiSeq chip, and the "dark" Ag_N-DNA was then flowed into the chip. DNA activator sequences that produced significant fluorescence enhancement were identified and used for ML-guided design of activator sequences for NCBs, as discussed later.[108]

5 ML-guided design of DNA sequences for bright Ag_N-DNA emission

The first demonstration of ML for Ag_N-DNA design focused on the selection of DNA template sequences for time-stable and brightly fluorescent Ag_N-DNAs, without selectivity for peak emission wavelength, λ_p. This simpler problem is an important starting point for Ag_N-DNA design because studies have shown that ≤ 25% of DNA sequences yield brightly fluorescent Ag_N-DNA products. Moreover, while it is well-known that cytosines and guanines are critical for formation of bright Ag_N-DNAs,[53,79] the incorporation of multiple nucleobases appears important for Ag_N-DNA time stability.[102] Thus, it is important to determine which DNA sequences yield Ag_N-DNAs with bright fluorescence that persists for multiple days.

To address this design challenge, Copp, *et al.* used a training data set of 684 random 10-base DNA sequences of all four nucleobases (A, C, G, T) and the UV-excited fluorescence emission spectra of the Ag_N-DNA products stabilized by these DNA strands (Figure 6a), which were collected in a previous study.[43] Training sequences were completely random, other than the criterion that at least three nucleobases in each sequence must be C and/or G. (684 sequences represents 0.065% of all possible 10-base DNA sequences.) To quantify fluorescence brightness, the integrated fluorescence emission spectrum from 450-850 nm, I_{int}, was used. I_{int} represents emission from all UV-excited products for a given DNA sequence and combines chemical yield, quantum yield, and extinction coefficient among all emitters in a given solution. Sequences corresponding to the top 30% of I_{int} values in the data set were defined as "bright." Sequences corresponding to the bottom 30% of I_{int} values were defined as "dark." Then, supervised ML was performed using a SVM classifier trained to distinguish between bright and dark sequences (Figure 6a), omitting the middle 40% of sequences.[41]

Feature engineering was critical to the success of this ML model. It was found that simple parameterization of DNA sequences without feature selection resulted in low 50-60% SVM accuracies as quantified by cross-validation (cross-validation is a process involving training on the majority of data in the training library while reserving a smaller portion of the data library for testing how well the trained SVM assigns the correct class to input sequences). To more effectively select features that are predictive of Ag_N-DNA brightness, the bioinformatics tool MERCI[113] was used to extract short DNA base motifs (contiguous sets of nucleobases, with or without a single "wildcard" nucleobase) that were correlated to either bright or dark I_{int} values (Figure 6a). By constructing feature vectors that represent the presence or absence of these motifs, a SVM accuracy of 82% was achieved. This accuracy was slightly boosted to 86% by also including features that quantified the positions of bright-correlated motifs within the 10-base sequence . Based on the greater importance of position-invariant base motifs as compared to position-specific sequence information, it was suggested that select base patterns are more important for brightly emissive Ag_N-DNA formation, rather that specific positions of base motifs in a DNA strand.

New DNA sequences for brightly emissive Ag_N-DNAs were designed by sampling MERCI-identified bright DNA base motifs from an I_{int}-weighted distribution to construct candidate sequences. The trained SVM was then used to assign a "brightness" probability to each of the constructed sequences, and the 374 sequences predicted with most certainty to be "bright" were selected for experimental testing. Experimental synthesis showed that 78% of the designed sequences produced Ag_N-DNAs with I_{int} values above the brightness threshold established in the training

data (Figure 6b).

This study was the first successful demonstration of ML for Ag_N-DNA design.[41] Notably, the MERCI-identified DNA base motifs confirmed existing understanding that C- and G-rich DNA base sequences are highly likely to stabilize brightly emissive Ag_N-DNAs, while ^{T}-rich sequences strongly disfavor emissive Ag_N-DNA formation. The method did have limitations, however. Notably, the SVM was only predictive of overall brightness and not of peak emission wavelength, λ_p. Also, the ML-designed Ag_N-DNA template sequences favored Ag_N-DNAs with redder λ_p wavelengths and disfavored shorter-wavelength green Ag_N-DNAs (Figure 6c). This motivates the development of ML models that can discriminate among Ag_N-DNA emission colors.

6 Multi-class ML models for prediction of Ag_N-DNA color

It is ideal to develop ML models to design Ag_N-DNAs with λ_p in specific spectral windows. Because Ag_N-DNA atomic size and structure correlates to fluorescence spectral properties,[106] this design problem involves selecting a DNA strand of specific nucleobase sequence that sculpts a silver nanocluster of the appropriate size and shape to yield a desired λ_p value. The relationship between DNA sequence and nanocluster size/shape has been poorly understood. Thus, data mining and ML techniques provide a promising way to map DNA sequence onto Ag_N-DNA λ_p.

Several key challenges have been overcome to train accurate ML models that predict Ag_N-DNA color given an input sequence. First, there is the challenge of feature engineering. Because training data sets correlating Ag_N-DNA color to sequence remain limited compared to the space of all possible DNA sequences, it has not generally been feasible to use more complex deep learning approaches that perform featurization (recent work is now making advances in this area[114]). Instead, a combination of data mining and known chemical information has been used to successfully engineer features, together with feature selection to avoid overfitting and achieve reasonably high cross-validation accuracies. Second, there is the challenge of data imbalance. Ag_N-DNA emission color is inherently unevenly distributed as a function of wavelength because of the enhanced stabilities of certain "magic" atomic sizes of these nanoclusters as compared to others.[43,44] Moreover, training data for NIR-emissive Ag_N-DNAs is far more limited than for the visible spectral regions.[115] This imbalanced training data challenges simple regression approaches to map sequence onto λ_p. A series of recent studies have overcome these challenges to achieve predictive design of DNA templates for Ag_N-DNAs across the spectral range of these emitters.

Copp, *et al.*, first presented a ML approach to design DNA sequences that select λ_p by using an ensemble of SVMs.[71] This study harnessed a training data set of 1,432 10-base DNA sequences from prior studies, including both random sequences[43] and ML-designed sequences.[41] The distribution of λ_p values for this training dataset, determined by spectral fitting described in Section 4, is approximately bimodal, a result of the "magic" colors of Ag_N-DNAs (Figure 7a).[43] Specifically, past ESI-MS studies showed that green-emissive Ag_N-DNAs have N_0 = 4 valence electrons, while red-emissive Ag_N-DNAs have N_0 = 6 valence electrons. Motivated by the distinct chemical differences between green and red Ag_N-DNAs, which likely correlate to distinct base sequence differences in their DNA templates, a supervised ML classification problem was defined based on the following color classes:

- Green: λ_p < 580 nm, I_{int} above the previously defined "bright" threshold[41]
- Red: 600 nm < λ_p < 660 nm, I_{int} above the "bright" threshold
- Very Red λ_p > 660 nm, I_{int} above the "bright" threshold
- Dark: I_{int} below the previously defined "dark" threshold

Red and Very Red classes were defined based on the hypothesis that two distinct distributions that appear above 600 nm in the Ag_N-DNA histogram correlate to two distinct classes of Ag_N-DNA structures, leading to different DNA sequence signatures for these λ_p windows (Figure 7a).

Training sequences were sorted into the four defined color classes. Notably, sequences with multiple bright peaks in more than one color class were excluded from the training data, as these sequences may contain patterns correlated with multiple

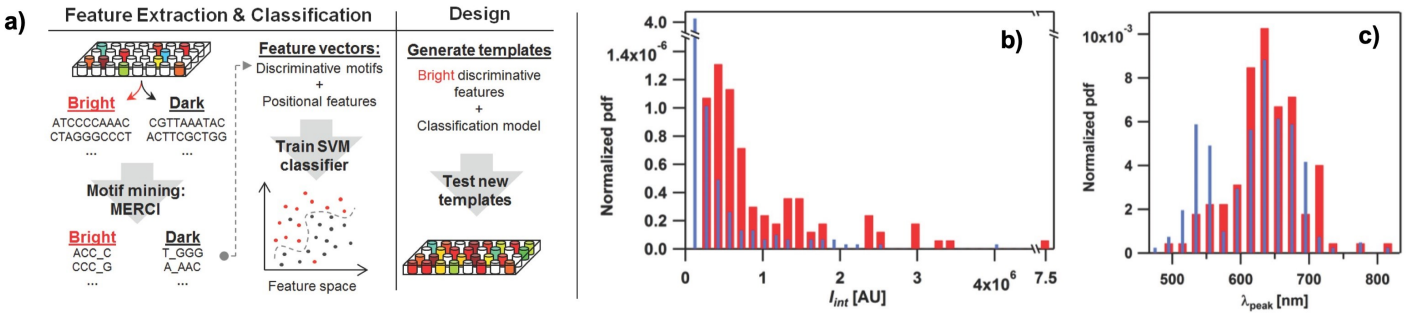

Fig. 6 ML-guided design of brightly fluorescent Ag_N-DNAs.[41] a) Schematic of the experimental, data processing, feature engineering, and ML steps to design DNA templates for brightly fluorescent Ag_N-DNAs. b,c) Normalized probability distribution functions (pdfs) of b) integrated intensities, I_{int}, c) and peak emission wavelengths in training data (thin blue bars) and for ML-designed sequences. ML-guided design increased I_{int} on average and tended to favor formation of red-emissive Ag_N-DNAs over green-emissive Ag_N-DNAs. Adapted from Copp, et al.,[41] with permission from John Wiley and Sons. Copyright 2014.

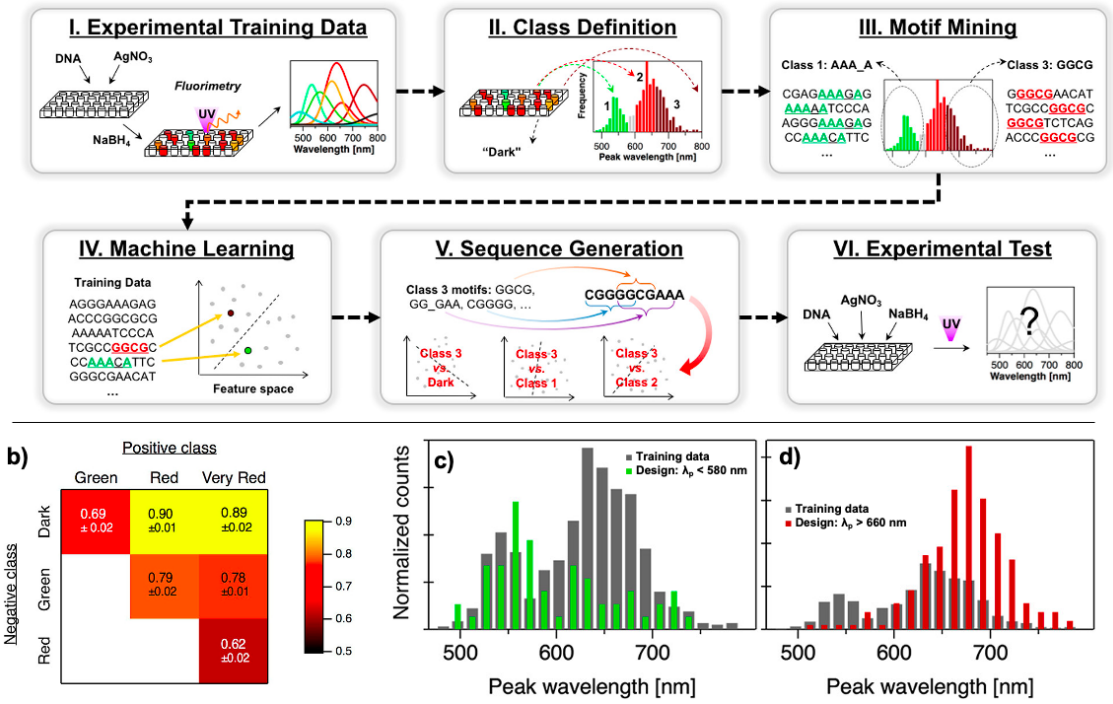

Fig. 7 ML-guided design of Ag_N-DNAs with fluorescence emission in specific spectral windows.[71] a) Schematic of ML workflow. b) Cross-validation scores of one-versus-one SVMs for each color class pair. c,d) Adapted from Copp, et al.,[71] with permission from the American Chemical Society. Copyright 2018.

color classes and therefore complicate ML. Sequences with I_{int} values between the "bright" and "dark" thresholds were also excluded to better learn to design brightly fluorescent Ag_N-DNAs.

Unlike the simpler problem of discriminating between "bright" and "dark" sequences only, learning to discriminate among sequences in the four color classes is a *multi-class* ML problem. Thus, a *one-versus-one* approach was used, *i.e.* a set of six SVMs was trained to discriminate between each pair of color classes, with one SVM per class pair. Because of significant training data imbalance - far more Dark and Red sequences than Green sequences - random subsampling of the more abundant class was used to balance the training data for each SVM (see Section 3 for training on imbalanced data).

Feature engineering was performed by first using MERCI[113] to identify > 1,600 DNA base motifs correlated with a specific Ag_N-DNA color, and then using "greedy" feature selection[116] to reduce the large set of MERCI-identified motifs to a more succinct set of about 200 motifs. This feature selection process is essential for reducing ML model overfitting and was found in this case to be essential to achieve sufficiently high cross-validation scores for the SVM ensemble (Figure 7b).

The trained SVM ensemble was used to design 10-base DNA sequences for Green and Very Red, the two least abundant color classes. Sequences were constructed by sampling color-correlated motifs, as done previously,[41] and then using all three SVMs associated with the target color class to assign the probability of falling within the desired color class. The minimum probabil-ity from the three SVMs was then used as a conservative metric to rank sequences. The top 180 designed Green and Very Red sequences were experimentally tested, showing that the SVM ensemble increased selection of bright Very Red Ag_N-DNAs by 330% and of bright Green Ag_N-DNAs by 70% (Figure 7c,d.) The lower relative success for Green as compared to Very Red was suspected to be related to (1) lower cross-validation scores of Green-associated SVMs, (2) the apparent similarity between Green and Dark sequences, and (3) a higher degree of variability of Green Ag_N-DNA emission spectra between experiments, which was investigated in the study.[71]

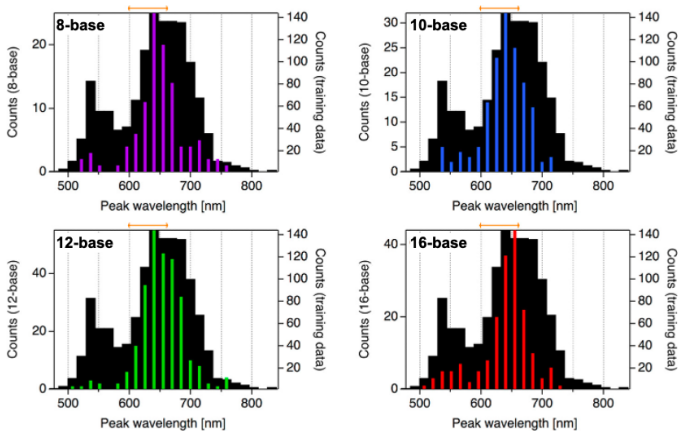

Fig. 8 Length-invariant design of DNA templates for Ag_N-DNAs. DNA templates of multiple lengths designed for Ag_N-DNAs with 600 nm < λ_p < 660 nm (orange brackets above graphs). Training data shown in black. Adapted from Copp, et al.,[107] with permission from the American Chemical Society. Copyright 2020.

The *one-versus-one* ML approach was later expanded to predict

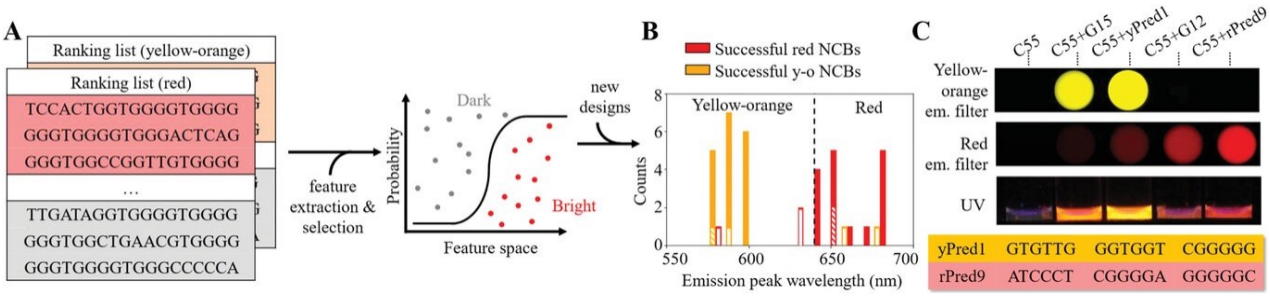

Fig. 9 ML-guided design of activator sequences for NCBs.[108] a) Schematic of ML workflow to train logistic regression classifiers to discriminate between "bright" NCBs (top 30% of emission intensities) and "dark" NCBs (bottom 30% of emission intensities). b) Success metrics for 40 activator sequences designed using the trained ML classifiers. Hatched bars represent unsuccessful designs. c) Plate reader images of lit-up NCBs using successfully designed activator sequences. Adapted from Kuo, et al.,[108] with permission from John Wiley and Sons. Copyright 2022.

Ag_N-DNA template sequences beyond 10-base oligomers without any additional training data for sequences of other lengths.[107] Using the growing library of 2,161 10-base DNA sequences with their associated Ag_N-DNA λ_p values and peak brightnesses, a new set of *one-versus-one* SVMs was trained, using the previous feature engineering and feature selection methods.[71] Because these feature vectors quantify the instances of select DNA base motifs rather than explicitly encoding DNA sequence, the feature vectors are general for sequences of any length, and the ML model could in principle be used to predict color for sequences of any length. To test the hypothesis that DNA base motifs selective for color are invariant with DNA sequence length, the method for constructing new DNA sequences by sampling color-correlated motifs[71] was generalized for DNA sequences of any length. Then, 8-base, 10-base, 12-base, and 16-base Red DNA template sequences were designed. (Red was chosen to test the hypothesis because the training data already included Green and Very Red designed sequences from past work,[71] which might make subsequent design of Green and Very Red easier than Red.) For all lengths, ML-guided design increased the prevalence of Red sequences by 100 - 150 % (Figure 8), supporting that the color-correlated base motifs in 10-base sequences also select Ag_N-DNA size/color at other DNA template lengths, as well.

Kuo, *et al.*, recently adapted the approach of Copp, *et al.*, using simple ML classifiers together with MERCI-based motif identification to design the activator sequences for NanoCluster Beacon Ag_N-DNAs[80] that "light up" either in the "yellow-orange" or red spectral windows, as defined based on filter cubes designed for the conventional fluorophores TRITC (Ex/Em: 535/50, 605/70 nm) and Cy5 (Ex/Em: 620/60, 700/75 nm) (Figure 9).[108] Using a next-generation sequencing chip platform to screen nearly 40,000 activator strands on a fluorescence microscope equipped with TRITC and Cy5 filter cubes,[81] they identified thousands of new activator sequences that yield bright NanoCluster Beacon fluorescence. This dataset was then used to train logistic regression classifiers to distinguish between DNA activator sequences that light-up brightly fluorescent "yellow-orange" and red NanoCluster Beacons and DNA activator sequences associated with low fluorescence ("dark"). The ML model was used to design 20 new "yellow-orange" and red activator sequences, which were 8.5 and 2.9 times more likely to be bright yellow–orange and red than randomly generated sequences, respectively. Feature analysis tools as discussed in Section 7 below could be used to interpret the sequence features learned by ML classifiers that select for suitable NanoCluster Beacon activator sequences, providing mechanistic insights into the sequence patterns that cause "light-up" effects in these Ag_N-DNA sensors.

In 2018, Swasey, *et al.*, dramatically expanded the number of known Ag_N-DNA emitters in the NIR window with their discovery of 161 new NIR-emissive Ag_N-DNAs.[115] This study has prompted significant interest in the design of NIR-emissive Ag_N-DNAs, which have exciting potential for applications in bioimaging.[29,30,117] Mastracco, *et al.*, achieved this goal by exploiting the limited data set of newly reported NIR-emissive Ag_N-DNAs together with feature engineering inspired by the first X-ray crystal structures of Ag_N-DNAs.[46,77] Namely, it was noted that pairs of both adjacent *and* nonadjacent nucleobases appear important for describing how a DNA template interacts with its encapsulated nanocluster in the crystal structure of a particular Ag_{16}-DNA reported by Cerretani, *et al.*[46] This crystal structure shows that adjacent C's and G's protect the long sides of the rod-shaped Ag_{16}, while two A's separated by three other nucleobases cap the nanocluster's ends (Figure 10). Such pairs of nucleobases can be referred to as nucleobase "staple" features $X\_{m}Y$, representing two distinct nucleobase ligands X and Y separated by m arbitrary nucleobases, m = 0, 1, ..., 8, which can coordinate the Ag_N at zero, one, or two sites. To capture such nucleobase patterns, length-144 feature vectors were constructed to enumerate the occurrences of all possible $X\_{m}Y$ motifs within a 10-base sequence. Then, a series of L1-norm SVMs were trained. L1-norm regularization, which uses the sum of the magnitude of vector components as the vector normalization, was chosen because L1-norm SVMs encourage sparse solutions and naturally perform feature selection.[118]

In addition to chemically-informed feature engineering, the same chemically informed color classes were used as previously defined based on the "magic" N_0 values of Ag_N-DNAs of distinct λ_p color classes.[71] The inclusion of Ag_N-DNA training data up to nearly λ_p = 1,000 nm motivated the definition of an additional color class at high wavelengths. Very little information about N_0 values is available for Ag_N-DNAs in the spectral window near the far red-NIR border. Instead, using a combination of unsuper-

vised k-means clustering and supervised ML, the previous Very Red class (λ_p > 660 nm, I_{int} above the "bright" threshold) was divided into two new color classes: Far Red (660 nm < λ_p < 800 nm) and NIR (λ_p > 800 nm) (Figure 10a). With this data distribution, there is even more significant training data imbalance between the largest classes, Dark and Far Red, and the smallest class, NIR, with only 55 training instances (Figure 10b). This data imbalance was addressed through the ML model architecture design. Specifically, a *one-versus-one* approach was used together with a consensus approach: for each pair of color classes, 10 different SVMs were trained on 10 different training data sets consisting of all data instances of the smaller class together with a random data subsampling of the larger class (Figure 11a). In this manner, variances in the subsampled data set are averaged over, providing a more accurate representation of the prediction of color class. 10-fold cross-validation shows that the ensemble SVM model performs best on average for pairs of color classes with the lowest data imbalance, as would be expected (Figure 11b).

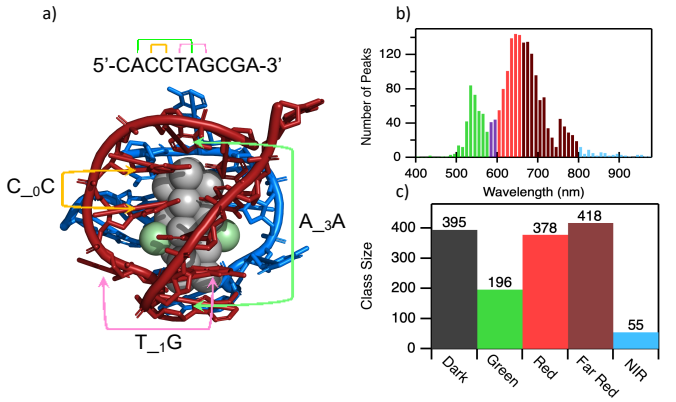

Fig. 10 Chemistry-informed ML for design of Ag_N-DNAs. a) Illustration of staple features as inspired by crystal structure.[46] b) Distribution of bright emission peaks identified in high-throughput experiments. c) Sizes of the five color classes are significantly imbalanced. Adapted with permission from Mastracco, et al.,[50] licensed under CC BY. Copyright 2022.

The ensemble of 100 SVMs was trained on all training data and then used to screen all 4^{10} 10-base DNA sequences for those with the highest likelihood of falling in the Green, Far Red, and NIR color classes. Experimental testing of these predictions showed high success rates for all three design cases, including significant improvement for Green prediction compared to past work[71,107] (Figure 11c). The success rate of NIR Ag_N-DNA design (defined as λ_p > 800 nm in this study) increased more than 12-fold compared to the training data(Figure 11e); only about 2% of sequences in the training data were in the NIR class, while 34 of the 124 sequences designed for NIR emission were experimentally found to yield NIR emission. Due to this success, this study nearly doubled the number of Ag_N-DNAs with bright near-infrared emission above 800 nm.[50]

7 Interpretable ML: feature analysis uncovers the sequence-to-color connection.

A fundamental tenet of ML is that certain information about the training data, contained in input features to the ML model, contains the information needed to learn the specific trend or classification task that one seeks to determine. This is why ML works best when input features contain only the information needed for a given learning tasks. The field of interpretable ML is focused on determining exactly what a ML model is learning and how important that information is to the predictions made by the model.[119] Interpretable ML opens up exciting possibilities in the chemical sciences to not only design and discover new molecules and materials but also to learn new chemistry about the structures and properties of these systems.

Interpretable ML is a promising approach to begin to understand the complex "sequence-structure-property" relationships that connect DNA sequence to the resulting size and color of Ag_N-DNAs. For each of the models discussed in Sections 5 and 6, feature analysis has yielded new insights into how DNA sequence selects for Ag_N-DNA fluorescence brightness and emission color. In this section, we summarize these insights.

The first ML-guided study of Ag_N-DNAs (Section 5) focused on simple discrimination between 10-base DNA sequences that stabilize brightly emissive Ag_N-DNAs *versus* "dark" sequences that do not stabilize detectably emissive Ag_N-DNAs.[41] This first large-scale studies of 684 random DNA sequences and their associated Ag_N-DNA emission spectra enabled examination of how DNA base motifs correlated to Ag_N-DNA brightness for the first time. Figure 12a shows the ratio of average 2-base motif counts, $R_{B/D}$, per strand in bright to dark templates. C- and G- motifs are both clearly important for stabilizing brightly emissive Ag_N-DNAs, an important insight at the time given that cytsoine was primarily associated with stabilizing Ag_N-DNAs.It is also clear that T is generally selective for "dark" while A appears to play a more nuanced role. This study also listed the top 10 most frequently occurring discriminative motifs for both bright and dark sequences, as identified by MERCI.[113] These motifs were 3-5 bases long and corroborated the observations from Figure 12a that C and G both play a role in selection of brightly emissive Ag_N-DNAs,[41] while T selects for dark DNA sequences.

A similar approach was later applied to determine how DNA sequence selects for Ag_N-DNA emission color classes, which were defined based on Ag_N-DNA magic number properties.[71] In this model, feature selection was used to reduce the number of MERCI-identified DNA sequence motifs to a succinct set of about 200. Then, the base composition of these selected motifs was investigated, showing that certain short 2-base and 3-base motifs were highly selective for Ag_N-DNA color class. (Figure 12b) shows that CC is selective against Dark but not selective for Ag_N-DNA color. G-rich motifs, and in particular those containing "GG", are selective of larger, longer wavelength Ag_N-DNAs, while A-rich motifs select for smaller, Green-emissive Ag_N-DNAs. It is also notable that the sequence signatures for Green and Dark are similar, which may explain the lower success rate for designing Green Ag_N-DNAs as compared to Red and Far Red Ag_N-DNAs.[71]

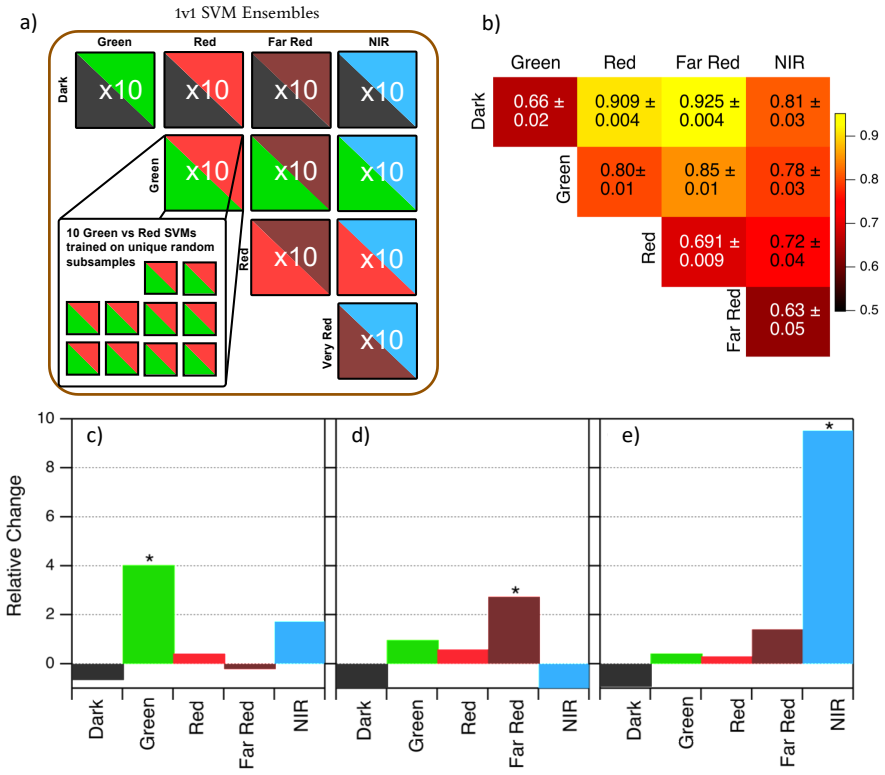

Fig. 11 a) One-versus-one SVM ensemble used to design Ag_N-DNAs with targeted fluorescence in the Green, Far Red, and NIR classes. b) Heatmap of average 10-fold cross-validation accuracies of the model in (a). c-e) Relative change of each class size for DNA sequences designed for c) Green, d) Far Red, and e) NIR, showing significant increases in the target color classes in all cases. Adapted from Mastracco, et al.,[50] licensed under CC BY. Copyright 2022.

Our recent work used the feature analysis tool BorutaShap to enable model interpretability.[50] BorutaShap combines feature selection with the Boruta algorithm[120] with Shapley additive explanations (SHAP), which uses methods from coalitional game theory to interpret model predictions.[121] Most feature analysis methods use a leave-one-out approach whereby a ML model's cross-validation accuracy is determined when each feature in the input feature vector is left out one-by-one. The feature whose removal results in the largest drop in cross-validation accuracy is scored as most "important" for the ML model's performance, and features are typically ranked by their importance scores. While this method is often useful to interpreting the weights placed on each feature by a ML model, it does not determine how "relevant" each feature is for the learning task; to quantify "relevance," it is important to compare the importance scores of features to randomly generated "shadow features" that, by nature of their randomness, are not relevant to the ML learning task.[122] We used BorutaShap to score the relevance of all nucleobase staple motifs and then calculate a "net importance score" for each staple feature's selectivity per color class. Figure 12c shows the top 15-scored nucleobase staple motifs. This analysis further confirms the critical importance of consecutive G's for formation of larger, long-wavelength Ag_N-DNAs. The complexity of these sequence-to-color relationships also underscores the importance of ML-based design strategies for Ag_N-DNAs. As X-ray crystallographic studies of Ag_N-DNAs expand, these studies may confirm whether the sequence motifs in Figure 12 are, indeed, relevant for selecting Ag_N-DNA size and emission color.

8 Conclusions

This Feature Article has summarized recent progress in ML-enabled design of Ag_N-DNAs. Progress has been enabled by several key advances: (1) training data libraries developed with well-controlled experimental synthesis and fluorescence emission spectroscopy of 10^3 Ag_N-DNAs with uniform synthesis parameters and universal UV excitation of emission, (2) simple ML classifiers that are well-suited for learning on limited training data sets, (3) chemically motivated ML classification based on known Ag_N-DNA size-to-color correlations from ESI-MS studies, (4) statistical sampling to address training data imbalance, (5) chemistry-informed feature engineering together with feature selection to reduce overfitting and gain new chemical insights into the mechanisms behind the sequence-to-property relationships for Ag_N-DNAs. These strategies have dramatically advanced the state-of-the-art in the design of this class of nanomaterials, with exciting potential to design new Ag_N-DNA-based emitters for bioimaging and biosensing.

Very recent advances illustrate the potential of ML-designed Ag_N-DNAs for bioimaging applications. The fundamental understanding of Ag_N-DNAs has advanced significantly since 2019 thanks to one Ag_{16}-DNA that was first identified in ML studies.[71] Cerretani, *et al.* solved the crystal structure of this Ag_{16}-DNA,

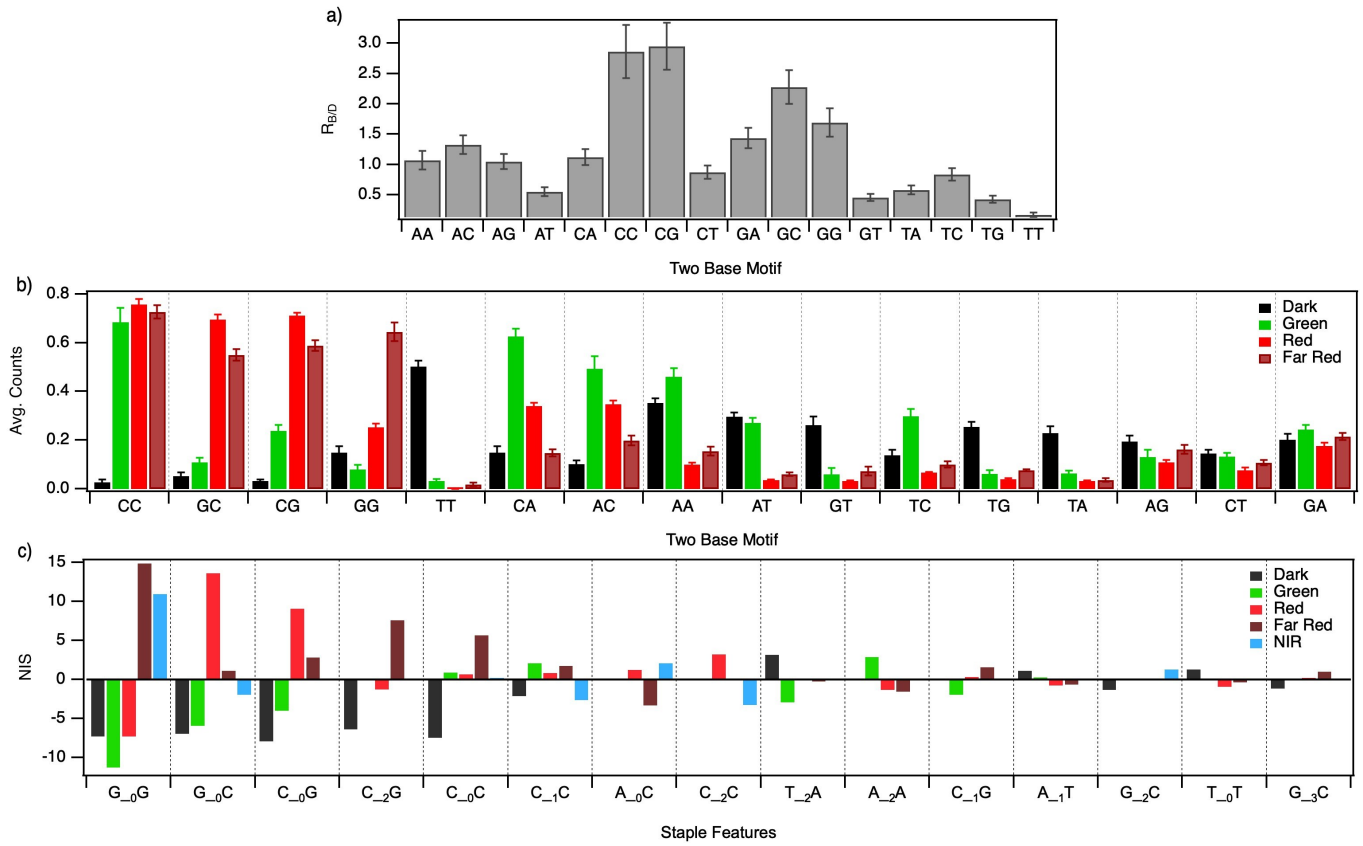

Fig. 12 Feature analysis provides fundamental insights about sequence-to-property relationships for Ag_N-DNAs. a) Ratio of average 2-base motif counts, $R_{B/D}$, per strand in bright to dark templates. Adapted from Copp, et al.,[41] with permission from John Wiley and Sons. Copyright 2014. b) Prevalence of 2-base patterns in motifs identifed as color-selective by feature selection.[71] Adapted from Copp, et al.,[71] with permission from the American Chemical Society. Copyright 2018. c) Net importance score (NIS) for staple motifs as scored by Boruta Adapted from Mastracco, et al.,[50] with permission from the American Chemical Society. Copyright 2022.

and Gonzàlez-Rosell, *et al.*, later determined its electron count by ESI-MS.[47] This combination of structural information and molecular formula has now enabled key theoretical breakthroughs in electronic structure modeling of Ag_N-DNAs.[47,78] Enabled by the degree of structural understanding of this Ag_{16}-DNA, Vosch and coauthors recently developed and demonstrated functionalized NIR-emissive biological labels for targeted cell labeling and *in vivo* imaging, providing a framework for developing generalizable biolabels based on these emitters.[123,124] New chemical synthesis strategies to significantly increase the yields of NIR-emissive Ag_N-DNAs that are discovered by ML-guided experiments may rapidly expand this class of potential biolabels.[125] We anticipate that ML-designed Ag_N-DNAs will continue to advance both theory and applications. Promising areas for progress include first-principles simulations that exploit nucleobase features identified as important by feature analysis (Figure 12) and an expanded palette of NIR-emitting Ag_N-DNAs.

Moreover, the development of ML-based models for Ag_N-DNA design provides a roadmap for the design of other DNA- and peptide-based nanomaterials that move beyond well-understood natural sequence-dependent interactions, such as Watson-Crick-Franklin base pairing. Data mining and ML approaches developed in the field of computer science have significant potential for the advancement of these and other chemical and materials systems. Researchers seeking to develop these approaches for a specific system may find success in adapting the approaches and models presented here. We hope that the ML tutorial and case studies in this Feature Article will inspire new innovations in the field of ML and data mining for chemical discovery.

Author Contributions

Both authors have discussed the organization and presentation of the contents and participated in the writing and editing of the Feature Article.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors acknowledge primary support by the National Science Foundation Materials Research Science and Engineering Center program through the UC Irvine Center for Complex and Active Materials (DMR-2011967) and partial support from the NSF Biophotonics program, CBET-2025790, and from the UC multicampus research programs and initiatives (MRPI) under Award No. M23PL5990.

Notes and references

1 N. C. Seeman, *Journal of Theoretical Biology*, 1982, **99**, 237–247.
2 N. C. Seeman and H. F. Sleiman, *Nature Reviews Materials*, 2017, **3**, 1–23.
3 R. J. Macfarlane, B. Lee, M. R. Jones, N. Harris, G. C. Schatz and C. A. Mirkin, *Spherical Nucleic Acids*, Jenny Stanford Publishing, 2020, pp. 539–553.
4 M. DeLuca, Z. Shi, C. E. Castro and G. Arya, *Nanoscale Horizons*, 2020, **5**, 182–201.
5 E. Del Grosso, E. Franco, L. J. Prins and F. Ricci, *Nature Chemistry*, 2022, **14**, 600–613.
6 B. Maddox, *NOVA, Science programming on air and online, April*, 2003.
7 B. Maddox, *Nature*, 2003, **421**, 407–408.
8 S. Rauch and B. C. Dickinson, *Programmable RNA binding proteins for imaging and therapeutics*, 2018.
9 Y. Takezawa, J. Müller and M. Shionoya, *Chemistry Letters*, 2017, **46**, 622–633.
10 S. Naskar, R. Guha and J. Mueller, *Angewandte Chemie International Edition*, 2020, **59**, 1397–1406.
11 T. Atsugi, A. Ono, M. Tasaka, N. Eguchi, S. Fujiwara and J. Kondo, *Angewandte Chemie International Edition*, 2022, **61**, e202204798.
12 J. T. Petty, J. Zheng, N. V. Hud and R. M. Dickson, *Journal of the American Chemical Society*, 2004, **126**, 5207–5212.
13 G. Shemer, O. Krichevski, G. Markovich, T. Molotsky, I. Lubitz and A. B. Kotlyar, *Journal of the American Chemical Society*, 2006, **128**, 11006–11007.
14 R. M. Izatt, J. J. Christensen and J. H. Rytting, *Chemical reviews*, 1971, **71**, 439–481.
15 J. Müller and B. Lippert, *Modern Avenues in Metal-Nucleic Acid Chemistry*, CRC Press, 2023.
16 K. Tanaka, A. Tengeiji, T. Kato, N. Toyama and M. Shionoya, *science*, 2003, **299**, 1212–1213.
17 H. D. A. Mohamed, S. M. Watson, B. R. Horrocks and A. Houlton, *Nanoscale*, 2012, **4**, 5936–5945.
18 N. Fardian-Melamed, G. Eidelshtein, D. Rotem, A. Kotlyar and D. Porath, *Advanced Materials*, 2019, **31**, 1902816.
19 N. Fardian-Melamed, L. Katrivas, G. Eidelshtein, D. Rotem, A. Kotlyar and D. Porath, *Nano Letters*, 2020, **20**, 4505–4511.
20 A. Gonzàlez-Rosell, C. Cerretani, P. Mastracco, T. Vosch and S. M. Copp, *Nanoscale Advances*, 2021, **3**, 1230–1260.
21 H. D. A. Mohamed, S. M. Watson, B. R. Horrocks and A. Houlton, *Journal of Materials Chemistry C*, 2015, **3**, 438–446.
22 Y. Hu and C. M. Niemeyer, *Advanced Materials*, 2019, **31**, 1806294.
23 S. Vecchioni, B. Lu, W. Livernois, Y. P. Ohayon, J. B. Yoder, C.-f. Yang, K. Woloszyn, W. Bernfeld, M. Anantram, J. W. Canary *et al.*, *Advanced Materials*, 2023, 2201938.
24 T. Vosch, Y. Antoku, J.-C. Hsiang, C. I. Richards, J. I. Gonzalez and R. M. Dickson, *Proceedings of the National Academy of Sciences*, 2007, **104**, 12616–12621.
25 J. Sharma, H.-C. Yeh, H. Yoo, J. H. Werner and J. S. Martinez, *Chemical Communications*, 2010, **46**, 3280–3282.
26 E. G. Gwinn, P. O'Neill, A. J. Guerrero, D. Bouwmeester and D. K. Fygenson, *Advanced Materials*, 2008, **20**, 279–283.
27 Y. Chen, M. L. Phipps, J. H. Werner, S. Chakraborty and J. S. Martinez, *Accounts of Chemical Research*, 2018, **51**, 2756–2763.
28 S. A. Bogh, M. R. Carro-Temboury, C. Cerretani, S. M. Swasey, S. M. Copp, E. G. Gwinn and T. Vosch, *Methods and applications in fluorescence*, 2018, **6**, 024004.
29 V. A. Neacşu, C. Cerretani, M. B. Liisberg, S. M. Swasey, E. G. Gwinn, S. M. Copp and T. Vosch, *Chemical Communications*, 2020, **56**, 6384–6387.
30 M. B. Liisberg, Z. Shakeri Kardar, S. M. Copp, C. Cerretani and T. Vosch, *The Journal of Physical Chemistry Letters*, 2021, **12**, 1150–1154.
31 B. C. Fleischer, J. T. Petty, J.-C. Hsiang and R. M. Dickson, *The Journal of Physical Chemistry Letters*, 2017, **8**, 3536–3543.
32 S. Krause, C. Cerretani and T. Vosch, *Chemical Science*, 2019, **10**, 5326–5331.
33 X. Tu, S. Manohar, A. Jagota and M. Zheng, *Nature*, 2009, **460**, 250–253.
34 Z. Lin, Y. Yang, A. Jagota and M. Zheng, *ACS nano*, 2022, **16**, 4705–4713.
35 C. Lachance-Brais, M. Rammal, J. Asohan, A. Katolik, X. Luo, D. Saliba, A. Jonderian, M. J. Damha, M. J. Harrington and H. F. Sleiman, *Advanced Science*, 2023, 2205713.
36 M. G. Rafique, J. M. Remington, F. Clark, H. Bai, V. Toader, D. F. Perepichka, J. Li and H. F. Sleiman, *Angewandte Chemie*, 2023, e202217814.
37 N. L. Ing, R. K. Spencer, S. H. Luong, H. D. Nguyen and A. I. Hochbaum, *ACS nano*, 2018, **12**, 2652–2661.
38 M. Kumar, N. L. Ing, V. Narang, N. K. Wijerathne, A. I. Hochbaum and R. V. Ulijn, *Nature chemistry*, 2018, **10**, 696–703.
39 N. J. Sinha, M. G. Langenstein, D. J. Pochan, C. J. Kloxin and J. G. Saven, *Chemical Reviews*, 2021, **121**, 13915–13935.
40 S. Kumar, A. Pearse, Y. Liu and R. E. Taylor, *Nature communications*, 2020, **11**, 2960.
41 S. M. Copp, P. Bogdanov, M. Debord, A. Singh and E. Gwinn, *Advanced Materials*, 2014, **26**, 5839–5845.
42 D. Schultz, K. Gardner, S. S. Oemrawsingh, N. Markeševiċ, K. Olsson, M. Debord, D. Bouwmeester and E. Gwinn, *Advanced materials*, 2013, **25**, 2797–2803.
43 S. M. Copp, D. Schultz, S. Swasey, J. Pavlovich, M. Debord, A. Chiu, K. Olsson and E. Gwinn, *The journal of physical chemistry letters*, 2014, **5**, 959–963.
44 S. M. Copp and A. Gonzàlez-Rosell, *Nanoscale*, 2021, **13**, 4602–4613.
45 A. Gonzàlez-Rosell, R. Guha, C. Cerretani, V. Rück, M. B. Liisberg, B. B. Katz, T. Vosch and S. M. Copp, *The journal of physical chemistry letters*, 2022, **13**, 8305–8311.
46 C. Cerretani, H. Kanazawa, T. Vosch and J. Kondo, *Angewandte Chemie International Edition*, 2019, **58**, 17153–17157.
47 A. Gonzàlez-Rosell, S. Malola, R. Guha, N. R. Arevalos, M. F. Matus, M. E. Goulet, E. Haapaniemi, B. B. Katz, T. Vosch, J. Kondo, H. Häkkinen and S. M. Copp, *Journal of the American Chemical Society*, 2023, **145**, 10721—-10729.
48 A. Ono, S. Cao, H. Togashi, M. Tashiro, T. Fujimoto, T. Machinami, S. Oda, Y. Miyake, I. Okamoto and Y. Tanaka, *Chemical communications*, 2008, 4825–4827.
49 P. Scharf and J. Müller, *ChemPlusChem*, 2013, **78**, 20–34.
50 P. Mastracco, A. Gonzàlez-Rosell, J. Evans, P. Bogdanov and S. M. Copp, *ACS nano*, 2022, **16**, 16322–16331.
51 J. T. Petty, C. Fan, S. P. Story, B. Sengupta, M. Sartin, J.-C. Hsiang, J. W. Perry and R. M. Dickson, *The Journal of Physical Chemistry B*, 2011, **115**, 7996–8003.
52 D. Schultz and E. Gwinn, *Chemical communications*, 2011, **47**, 4715–4717.
53 D. Schultz and E. G. Gwinn, *Chemical Communications*, 2012, **48**, 5748–5750.
54 M. Walter, J. Akola, O. Lopez-Acevedo, P. D. Jadzinsky, G. Calero, C. J. Ackerson, R. L. Whetten, H. Grönbeck and H. Häkkinen, *Proceedings of the National Academy of Sciences*, 2008, **105**, 9157–9162.
55 J. T. Petty, O. O. Sergev, M. Ganguly, I. J. Rankine, D. M. Chevrier and P. Zhang, *Journal of the American Chemical Society*, 2016, **138**, 3469–3477.
56 D. Schultz, S. M. Copp, N. Markeševiċ, K. Gardner, S. S. Oemrawsingh, D. Bouwmeester and E. Gwinn, *ACS nano*, 2013, **7**, 9798–9807.
57 Q. Wu, C. Liu, C. Cui, L. Li, L. Yang, Y. Liu, H. Safari Yazd, S. Xu, X. Li, Z. Chen *et al.*, *Journal of the American Chemical Society*, 2021, **143**, 14573–14580.
58 R. Orbach, W. Guo, F. Wang, O. Lioubashevski and I. Willner, *Langmuir*, 2013, **29**, 13066–13071.
59 S. M. Copp, D. E. Schultz, S. Swasey and E. G. Gwinn, *ACS nano*, 2015, **9**, 2303–2310.
60 L. Yourston, L. Rolband, C. West, A. Lushnikov, K. A. Afonin and A. V. Krasnoslobodtsev, *Nanoscale*, 2020, **12**, 16189–16200.
61 J. M. Obliosca, M. C. Babin, C. Liu, Y.-L. Liu, Y.-A. Chen, R. A. Batson, M. Ganguly, J. T. Petty and H.-C. Yeh, *ACS nano*, 2014, **8**, 10150–10160.
62 N. Bossert, D. de Bruin, M. Götz, D. Bouwmeester and D. Heinrich, *Scientific Reports*, 2016, **6**, 37897.
63 B. Sengupta, K. Springer, J. G. Buckman, S. P. Story, O. H. Abe, Z. W. Hasan, Z. D. Prudowsky, S. E. Rudisill, N. N. Degtyareva and J. T. Petty, *Journal of Physical Chemistry C*, 2009, **113**, 19518–19524.
64 J. Ai, W. Guo, B. Li, T. Li, D. Li and E. Wang, *Talanta*, 2012, **88**, 450–455.
65 P. R. O'Neill, L. R. Velazquez, D. G. Dunn, E. G. Gwinn and D. K. Fygenson, *The Journal of Physical Chemistry C*, 2009, **113**, 4229–4233.
66 S. M. Swasey, L. E. Leal, O. Lopez-Acevedo, J. Pavlovich and E. G. Gwinn, *Scientific Reports*, 2015, **5**, 10163.
67 E. Makkonen, P. Rinke, O. Lopez-Acevedo and X. Chen, *International Journal of Molecular Sciences*, 2018, **19**, 2346.
68 S. M. Swasey, N. Karimova, C. M. Aikens, D. E. Schultz, A. J. Simon and E. G. Gwinn, *ACS nano*, 2014, **8**, 6883–6892.
69 W. Guo, J. Yuan, Q. Dong and E. Wang, *Journal of the American Chemical Society*, 2010, **132**, 932–934.
70 Y. Teng, X. Yang, L. Han and E. Wang, *Chem. Eur. J.*, 2014, **2-**, 1111–1115.
71 S. M. Copp, A. Gorovits, S. M. Swasey, S. Gudibandi, P. Bogdanov and E. G. Gwinn, *ACS nano*, 2018, **12**, 8240–8247.
72 X. Chen, A. Karpenko and O. Lopez-Acevedo, *ACS Omega*, 2017, **2**, 7343–7348.
73 X. Chen, E. Makkonen, D. Golze and O. Lopez-Acevedo, *Journal of Physical Chemistry Letters*, 2018, **9**, 4789–4794.
74 J. Wu, Y. Fu, Z. He, Y. Han, L. Zheng, J. Zhang and W. Li, *The Journal of Physical Chemistry B*, 2012, **116**, 1655–1665.
75 X. Chen, M. Boero and O. Lopez-Acevedo, *Physical Review Materials*, 2020, **4**, 065601.
76 M. F. Matus and H. Häkkinen, *Nature Reviews Materials*, 2023, 1–18.
77 D. J. Huard, A. Demissie, D. Kim, D. Lewis, R. M. Dickson, J. T. Petty and R. L. Lieberman, *Journal of the American Chemical Society*, 2018, **141**, 11465–11470.
78 H. Häkkinen, S. Malola and M. F. Matus, *ChemRxiv*, 2023, DOI: 10.26434/chemrxiv–2023–x8899.
79 C. I. Richards, S. Choi, J.-C. Hsiang, Y. Antoku, T. Vosch, A. Bongiorno, Y.-L. Tzeng and R. M. Dickson, *Journal of the American Chemical Society*, 2008, **130**, 5038–5039.
80 H.-C. Yeh, J. Sharma, I.-M. Shih, D. M. Vu, J. S. Martinez and J. H. Werner, *J. Am. Chem. Soc*, 2012, **134**, 37.
81 Y.-A. Kuo, C. Jung, Y.-A. Chen, J. R. Rybarski, T. D. Nguyen, Y.-A. Chen, H.-C. Kuo, O. S. Zhao, V. A. Madrid, Y.-I. Chen *et al.*, Reporters, Markers, Dyes, Nanoparticles, and Molecular Probes for Biomedical Applications XI, 2019, pp. 31–42.
82 S. New, S. Lee and X. Su, *Nanoscale*, 2016, **8**, 17729–17746.
83 A. Seko, H. Hayashi, K. Nakayama, A. Takahashi and I. Tanaka, *Physical Review B*, 2017, **95**, 144110.

84 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, *Matter*, 2021, **4**, 1578–1597.
85 Y. T. Shih, Y. Shi and L. Huang, *Journal of Non-Crystalline Solids*, 2022, **584**, 121511.
86 A. L. Ferguson, *Journal of Physics: Condensed Matter*, 2017, **30**, 043002.
87 M. Köppen, 5th online world conference on soft computing in industrial applications (WSC5), 2000, pp. 4–8.
88 J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang and H. Liu, *ACM computing surveys (CSUR)*, 2017, **50**, 1–45.
89 H. He and E. A. Garcia, *IEEE Transactions on knowledge and data engineering*, 2009, **21**, 1263–1284.
90 B. Krawczyk, *Progress in Artificial Intelligence*, 2016, **5**, 221–232.
91 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Machine learning for molecular and materials science*, 2018.
92 G. R. Schleder, A. C. Padilha, C. M. Acosta, M. Costa and A. Fazzio, *Journal of Physics: Materials*, 2019, **2**, 032001.
93 J. Wei, X. Chu, X. Y. Sun, K. Xu, H. X. Deng, J. Chen, Z. Wei and M. Lei, *InfoMat*, 2019, **1**, 338–358.
94 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Materials*, 2013, **1**, 011002.
95 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, *npj Computational Materials 2015 1:1*, 2015, **1**, 1–15.
96 S. S. Borysov, R. M. Geilhufe and A. V. Balatsky, *PLOS ONE*, 2017, **12**, e0171501.
97 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Computational Materials 2017 3:1*, 2017, **3**, 1–13.
98 W. Sha, Y. Li, S. Tang, J. Tian, Y. Zhao, Y. Guo, W. Zhang, X. Zhang, S. Lu, Y. C. Cao and S. Cheng, *InfoMat*, 2021, **3**, 353–361.
99 C. M. Heil, A. Patil, A. Dhinojwala and A. Jayaraman, *ACS Central Science*, 2022, **8**, 996–1007.
100 M. Berdakin, M. I. Taccone, G. A. Pino and C. G. Sánchez, *Physical Chemistry Chemical Physics*, 2017, **19**, 5721–5726.
101 M. A. Jabed, N. Dandu, S. Tretiak and S. Kilina, *The Journal of Physical Chemistry A*, 2020, **124**, 8931–8942.
102 E. Gwinn, D. Schultz, S. M. Copp and S. Swasey, *Nanomaterials*, 2015, **5**, 180–207.
103 P. R. O'Neill, E. G. Gwinn and D. K. Fygenson, *The Journal of Physical Chemistry C*, 2011, **115**, 24061–24066.
104 S. M. Swasey, H. C. Nicholson, S. M. Copp, P. Bogdanov, A. Gorovits and E. G. Gwinn, *Review of Scientific Instruments*, 2018, **89**, 095111.
105 C. Cerretani and T. Vosch, *ACS omega*, 2019, **4**, 7895–7902.
106 S. M. Copp, D. Schultz, S. M. Swasey, A. Faris and E. G. Gwinn, *Nano Letters*, 2016, **16**, 3594–3599.
107 S. M. Copp, S. M. Swasey, A. Gorovits, P. Bogdanov and E. G. Gwinn, *Chemistry of Materials*, 2019, **32**, 430–437.
108 Y.-A. Kuo, C. Jung, Y.-A. Chen, H.-C. Kuo, O. S. Zhao, T. D. Nguyen, J. R. Rybarski, S. Hong, Y.-I. Chen, D. C. Wylie *et al.*, *Advanced Materials*, 2022, **34**, 2204957.
109 H.-C. Yeh, J. Sharma, J. J. Han, J. S. Martinez and J. H. Werner, *Nano letters*, 2010, **10**, 3106–3110.
110 Y.-A. Chen, J. M. Obliosca, Y.-L. Liu, C. Liu, M. L. Gwozdz and H.-C. Yeh, *Journal of the American Chemical Society*, 2015, **137**, 10476–10479.
111 S. Juul, J. M. Obliosca, C. Liu, Y.-L. Liu, Y.-A. Chen, D. M. Imphean, B. R. Knudsen, Y.-P. Ho, K. W. Leong and H.-C. Yeh, *Nanoscale*, 2015, **7**, 8332–8337.
112 M. Peng, N. Na and J. Ouyang, *Chemistry–A European Journal*, 2019, **25**, 3598–3605.
113 C. Vens, M.-N. Rosso and E. G. Danchin, *Bioinformatics*, 2011, **27**, 1231–1238.
114 F. Moomtaheen, M. Killeen, J. Oswald, A. Gonzàlez-Rosell, P. Mastracco, A. Gorovits, S. M. Copp and P. Bogdanov, Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 3593–3602.
115 S. M. Swasey, S. M. Copp, H. C. Nicholson, A. Gorovits, P. Bogdanov and E. G. Gwinn, *Nanoscale*, 2018, **10**, 19701–19705.
116 M. A. Hall and L. A. Smith, Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98, 1998, pp. 181–191.
117 V. Rück, C. Cerretani, V. A. Neacşu, M. B. Liisberg and T. Vosch, *Physical Chemistry Chemical Physics*, 2021, **23**, 13483–13489.
118 P. S. Bradley and O. L. Mangasarian, ICML, 1998, pp. 82–90.
119 W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu, *Proceedings of the National Academy of Sciences*, 2019, **116**, 22071–22080.
120 M. B. Kursa and W. R. Rudnicki, *Journal of statistical software*, 2010, **36**, 1–13.
121 S. M. Lundberg and S.-I. Lee, *Advances in neural information processing systems*, 2017, **30**, 1–10.
122 A. Costine, P. Delsa, T. Li, P. Reinke and P. V. Balachandran, *Journal of Applied Physics*, 2020, **128**, 235303.
123 V. Rück, N. K. Mishra, K. K. Sørensen, M. B. Liisberg, A. B. Sloth, C. Cerretani, C. B. Mollerup, A. Kjaer, C. Lou, K. J. Jensen *et al.*, *Journal of the American Chemical Society*, 2023.
124 X. Wang, M. B. Liisberg, G. L. Vonlehmden, X. Fu, C. Cerretani, L. Li, L. A. Johnson, T. Vosch and C. I. Richards, *ACS nano*, 2023.
125 R. Guha, M. Rafik, A. Gonzalez-Rosell and S. M. Copp, *Chemical Communications*, 2023, DOI: 10.1039/D3CC02896H.