ACSNANO

www.acsnano.org

ARTICLE

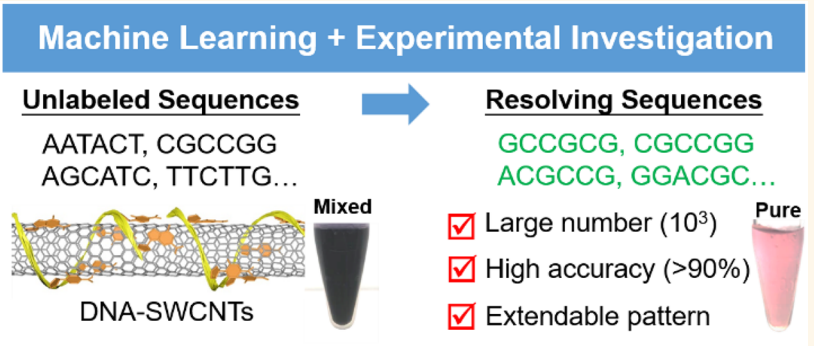
Machine Learning-Guided Systematic Search of DNA Sequences for Sorting Carbon Nanotubes

Zhiwei Lin,*,# Yoona Yang,# Anand Jagota, and Ming Zheng*

Cite This: https://doi.org/10.1021/acsnano.1c11448

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: **The prerequisite of utilizing DNA in sequence-dependent applications is to search specific sequences. Developing a strategy for efficient DNA sequence screening represents a grand challenge due to the countless possibilities of sequence combination. Herein, relying on sequence-dependent recognition between DNA and single-wall carbon nanotubes (SWCNTs), we demonstrate a method for systematic search of DNA sequences for sorting single-chirality SWCNTs. Different from previously documented empirical search, which has a low efficiency and accuracy, our approach combines machine learning and experimental investigation. The number of resolving sequences and the success rate of finding them are improved from ~10^2 to ~10^3 and from ~10% to >90%, respectively. Moreover, the resolving sequence patterns determined from 5-mer and 6-mer short sequences can be extended to sequence search in longer DNA subspaces.**

KEYWORDS: *carbon nanotubes, DNA, machine learning, sequence selection, chirality sorting*

The polymer sequence—the sequential arrangement of monomer units in a polymer chain—defines its primary structure and has a critical impact on its hierarchical structure and properties.[1] Nucleic acid polymers, such as DNA and RNA, consist of four types of nucleotide monomer units arranged in defined sequences. Such a precise sequence in DNA and RNA endows them with unparalleled functions over synthetic polymers without sequence control.[2,3] For example, the sequence of DNA in cells dictates the synthesis, folding, and function of proteins. In addition, DNA with prescribed sequences have found various applications ranging from oligonucleotide aptamers,[4−6] biomedical sensors,[7,8] to nanostructure design and fabrication.[9−11] In these sequence-dependent applications, searching targeted DNA sequences for specific purposes is a prerequisite. However, DNA sequence screening generally relies on empirical methods,[12] which are inefficient, high-cost, and system-specific. Developing a strategy for efficient DNA sequence screening is of both technological and scientific importance, while it remains an outstanding challenge due to the countless possibilities of sequence combination.

Single-wall carbon nanotubes (SWCNTs) are a family of macromolecules in tubular form defined by a roll-up vector on a graphene lattice denoted by (n, m).[13] Each (n, m) species exhibits unique optical, electrical, thermal, and mechanical properties,[14,15] and can be tailored for diverse applications.[16−25] A number of techniques and separation schemes have been developed to sort SWCNTs (n, m).[26−30] Among

Received: December 23, 2021
Accepted: February 23, 2022

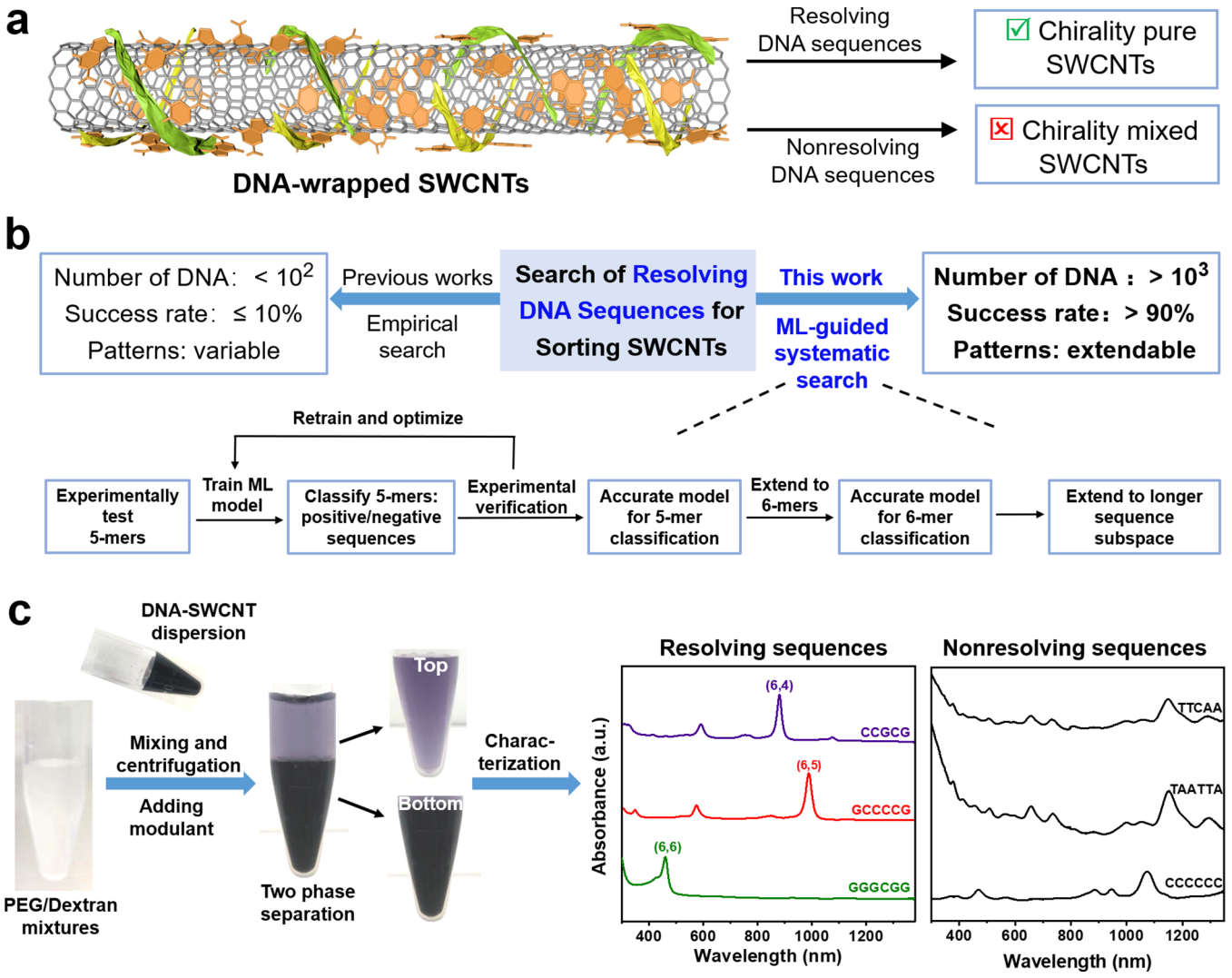


Figure 1. ML-guided systematic search of DNA sequences for sorting SWCNTs. (a) Schematic illustration of DNA-wrapped single-wall carbon nanotubes (SWCNTs). Resolving DNA sequences can purify chirality-pure SWCNTs, while nonresolving DNA sequences only give chirality-mixed SWCNTs. (b) Comparison between previous works and this work. In previous work, discovery of resolving DNA sequences was generally based on an empirical search. The number of resolving DNA sequences reported is small ($<10^2$), the success rate is quite low ($\leq$10%), and the resolving sequences have dramatically distinct patterns. In this work, the discovery of resolving DNA sequences is based on machine learning (ML)-guided systematic search. Our approach substantially increases the number ($>10^3$) and success rate (>90%) of finding resolving sequences, which can be extended to sequence search in longer DNA subspaces. (c) Schematic illustration of using an aqueous two-phase (ATP) technique to examine and test resolving/nonresolving DNA sequences. Here, one step separation is demonstrated. The bottom phase can be further extracted multiple times by adding a blank top phase (see Figure S2). If the DNA−SWCNT hybrids can give rise to at least one type of single-chirality SWCNTs with a purity of ≥90% in either the top or bottom phase, that sequence is classified as a resolving sequence; otherwise, it is classified as a nonresolving sequence. The identification relies on the UV−vis-NIR absorbance spectra of sorted fractions. Six typical examples of resolving/nonresolving sequences are illustrated.

them, sequence-dependent interactions[31] between DNA and SWCNTs are shown to provide resolution for selecting SWCNTs with defined helicity and handedness.[32,33] Previous studies have demonstrated that the aqueous two-phase (ATP) technique[34,35] can be very effective for sorting DNA-wrapped SWCNTs (DNA-SWCNTs) when they are coated by a correctly selected resolving DNA sequence (Figure 1a).[36,37] Although we have accumulated physical understanding of DNA sequence selection for sorting SWCNTs, search of resolving DNA sequences so far is still an expensive, time-consuming, and inefficient process. The empirical success rate of finding resolving sequences is rather low (≤10%),[38] which, in general, results in a relatively small number ($<10^2$) of resolving sequences discovered by previous works (Figure 1b).[36,37] In addition, because of the astronomically large DNA population, it is very challenging to conduct a comprehensive search through the DNA sequence space. By now, only a very small portion of the DNA population has been investigated, and the resolving sequences identified have dramatically distinct patterns and are unable to provide effective guidance for future sequence search.

Here, using DNA/SWCNT recognition as a model study, we demonstrate a strategy for systematic search of DNA sequences toward a targeted function. We envision that our strategy can offer an avenue for finding targeted DNA sequences in other application scenarios. Specifically, our aim is to develop a robust approach for finding resolving DNA sequences to sort single-chirality SWCNTs by addressing the

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX−XXX

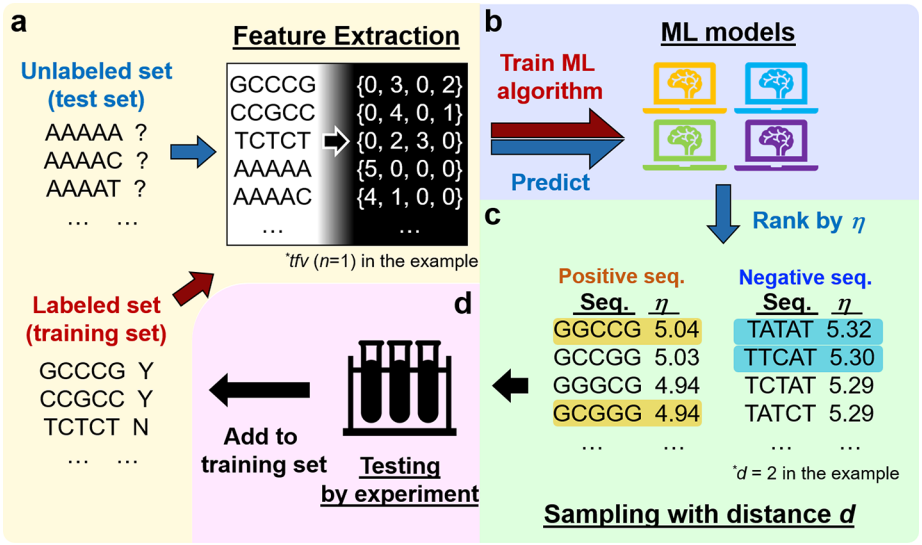


Figure 2. Schematic illustration of the ML-guided sampling process. (a) Feature extraction: DNA sequences were translated to a numeric vector. For the labeled set, the resolving sequences were labeled as Y, and nonresolving sequences were labeled as N; (b) training or prediction using ML models: ML models were trained using the labeled set with three different algorithms (Random Forest, Artificial Neural Network, and Support Vector Machine). The trained ML models were utilized to predict the label of unlabeled sequences. (c) The unlabeled sequences were ranked by the order of the ranking factor η. We then sampled the top 10 positive/negative predicted sequences, which differ from each other by at least two mutations. (d) The selected sequences were tested experimentally using ATP technique and then added to the previous training set. This process can be repeated until the models reach the desired performance.

above-mentioned challenges. Distinct from previous work relying on empirical search,[36,37] our approach effectively combines machine learning (ML) and experimental investigation. We started our exploration by experimentally examining the shortest viable DNA oligos with five nucleotides (referred to as 5-mers) using the ATP technique and collecting experimental data to serve as the first ML training set (Figure 1b). The relatively small number of distinct short sequences allows us to potentially screen the entire sequence library (4^5 = 1024). In contrast, similar screening is practically impossible for long-sequence subspaces, e.g., for a given sequence length longer than 10. We previously applied ML to predict resolving sequences in 12-mers (with $4^{12} \approx 16$ million distinct sequences).[38] However, that study was confined only to the T/C subspace due to limited availability of experimental data. The ML-model of the current study, trained on a subset of 5-mers, is employed to classify all the remaining 5-mers into either positive or negative sequence sets. Here, we denote the ML-classified resolving/nonresolving sequences as positive/negative sequences. We then tested several of the newly ML-classified sequence sets by ATP again and found the model to be highly accurate with a success rate of near 100%. This classification−verification process was further extended to 6-mers, giving a model with a success rate of >90% for classification of the 6-mers tested. The high success rate results in a large number of potential resolving sequences ($>10^3$). By iterating the procedure, we envision that our approach could be extended step-by-step to longer sequence subspaces.

RESULTS AND DISCUSSION

To simplify the problem of a large number of sequence combinations, we started our experimental investigation by testing the shortest oligos that are commercially available, containing three and four nucleotides (3-mers and 4-mers). DNA−SWCNT hybrid dispersions were prepared by a typical tip-sonication procedure reported elsewhere.[39] We found that the majority of tested 3-mers and 4-mers yielded unstable and inhomogeneous dispersions. This implies that the surface of SWCNTs may not be fully covered by DNA due to the weak binding affinity between these short DNA sequences and SWCNTs.[33] We then moved forward to DNA oligos with 5 nucleotides (5-mers). In contrast to 3-mers and 4-mers, 5-mers yielded stable and homogeneous dispersions having features similar to those made from previously reported long sequences[39] (e.g., 12-mers) (see Figure S1). This result suggests that the 5-mer is a length threshold for producing stable DNA−SWCNT hybrid dispersions.

Next, we focused on examining the performance of 5-mers in sorting SWCNTs. A total of 56 sequences were first examined by ATP experiments (Figures S3−S22). We used each sequence to make a DNA−SWCNT dispersion and then loaded it into a polyethylene glycol/dextran (PEG/DX) ATP system. Figure S2a outlines a detailed procedure used for multistep separation with ATP. Typically, DNA−SWCNT dispersions are mixed with a PEG1.5 kDa/DX250 kDa ATP system in the first separation step. After modulant polyvinylpyrrolidone (PVP) was added and centrifugation, partition of DNA−SWCNTs occurs. We collected the first top fraction (labeled 1T, or 0T if not adding PVP in the first step) and added an equal volume of blank top phase and an appropriate amount of PVP to the bottom phase. Upon mixing and centrifugation, the second top fraction (labeled 2T) was collected and again replaced with an equal amount of blank top phase and an appropriate amount of PVP. This separation process was iterated for multiple steps until the bottom phase became colorless, where the p-step top fraction is denoted as pT. We then measured the UV−vis-NIR absorbance spectra for all collected fractions.

We used GCCGC-EG150x and TGCAC-EG150x (synthetic SWCNT mixtures) dispersion as examples to illustrate the separation process and sequence classification. The absorbance spectra in Figure S2c shows that multistep separation of

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX−XXX

Table 1. Five-mer DNA Sequences Predicted by the ML-Model and Tested by the ATP Technique[a]

selected positive 5-mer sequences for test				selected negative 5-mer sequences for test			
name	sequence	class	purified CNTs	name	sequence	class	purified CNTs
P5-01	GGCGG	Y	(7,6)	N5-01	TATAA	N	N/A
P5-02	ACCGG	Y	(8,6), (6,5)	N5-02	TAATA	N	N/A
P5-03	CGGCA	Y	(7,6), (8,4), (11,1)	N5-03	TTTAT	N	N/A
P5-04	AGGCG	Y	(7,5), (8,3), (7,4)	N5-04	TATTT	N	N/A
P5-05	GCGGA	Y	(8,4), (6,5)	N5-05	AAAAA	N	N/A
P5-06	GGACC	Y	(7,6), (6,4), (7,5)	N5-06	TTCAA	N	N/A
P5-07	GGCCA	Y	(6,5)	N5-07	AATAT	N	N/A
P5-08	CACGG	Y	(6,5), (7,5)	N5-08	TTTTA	N	N/A
P5-09	CCAGG	Y	(7,5), (8,4)	N5-09	ATATA	N	N/A
P5-10	CCGGT	Y	(6,5), (10,3)	N5-10	CTTAA	N	N/A

[a]Y denotes a resolving sequence, and N denotes an nonresolving sequence.

GCCGC-EG150x enables the separation of pure (9,5) species, which can be readily collected from 0T. Although the following steps of separation do not lead to the selection of new species, we classify GCCGC as a resolving sequence as one type of pure species is sorted. It is worth mentioning that (9,5) is a low-abundance species in the parent SWCNT mixtures. The absorbance spectra in Figure S2d illustrates the multistep separation of TGCAC-EG150x dispersion. Similar absorbance profiles are observed from all fractions, and no pure species can be isolated from this dispersion. We therefore classify TGCAC as a nonresolving sequence. Table S1 summarizes the ATP separation results of 56 tested 5-mer sequences that fall in three different categories: palindrome sequences (32), C/G sequences (8), and random sequences (16). If a sequence can result in selecting at least one type of single chirality (n, m) species, we refer to it as a resolving sequence (labeled Y); otherwise, we refer to it as a nonresolving sequence (labeled N). We identified 22 resolving sequences and 34 nonresolving sequences.

To build an ML-model (Figure 2), we first translated the DNA sequences to numeric vectors. In our previous work,[38] we used two different sequence translation methods: position specific vector (*psv*) and term frequency vector (*tfv*) using overlapping n-gram, where n-gram is a substring of the length of n units from a given sequence. For example, for n = 2, "ABCD" can be translated to "AB", "BC", and "CD" by counting each character as gram. The *psv* conserves each n-gram at each position by utilizing one hot encoding for each n-gram at each position. The *tfv* counts the n-grams in the sequence and uses the count as a feature. Details of these feature extraction methods are described in our previous work.[38] In this work, we use three different training sets generated using 1-gram *psv*, 1-gram *tfv*, and 2-gram *tfv*. Once the training sets are ready, ML models are trained with three different algorithms: Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Machine (SVM). The model performance was evaluated with F-score using 10-fold cross validation. The models were built and validated using Scikit-learn.[40]

Although at this stage we have well-trained models with F-scores of 0.8−0.87 (Table S3), 56 sequences would not be sufficient to cover the entire 5-mer library (size: 4^5). So the question is, which sequences should be evaluated in the next batch of experiments to improve the model? To answer this question, we utilized the ML models to classify all the remaining sequences, sorting them according to their prediction accuracy. We sampled sequences that maximize the prediction accuracy under the additional constraint that all sequences chosen for the next round of experiments differ from each other by at least two mutations so as to better cover the entire sequence library. The schematic illustration of the ML-guided sampling is shown in Figure 2. First, we predict the unlabeled sequences using the trained ML models. Then, we rank the sequences by the ranking factor η. Here, $\eta = \sum_i F_i p_j$ where F_i is F-score of the model i and p_j is the prediction probability for sequence j. We found that the top 10 positive/negative predicted sequences are too similar to each other. We wished to sample the sequences that cover the entire library, however, the top 10 sequences are not likely to be well-distributed in sequence space. To avoid this, we set up a rule that requires the minimum distance, d, between any two selected sequences (d = 2 for 5-mers and d = 3 for 6-mers). For example, as shown in the Figure 2c, the third candidate GG**G**CG is only one base different from the first candidate GG**C**CG; thus, this sequence would not be selected as part of the next sequence set.

With the above-mentioned rule, we selected 10 positive and 10 negative sequences classified by the trained ML-model (Table 1, also Table S4 for detailed prediction probability for the selected 5-mers) for experimental verification. For simplicity, they are referred to as P5-01 to P5-10 and N5-01 to N5-10, respectively. We then experimentally tested the sorting performance of these 20 sequences by the ATP technique. The detailed ATP separation results can be found in Figures S23−S33. As summarized in Table 1, P5-01 to P5-10 are all confirmed as resolving sequences (labeled Y), while N5-01 to N5-10 are all confirmed as nonresolving ones (labeled N). The prediction is fully consistent with experimental results, indicating a highly accurate ML-model for 5-mer classification. Treating each experiment as an independent Bernoulli trial, the success rate of 100% for a sample size of 20 gives 95% confidence intervals as (0.8316, 1.0).[41] The confidence interval can also be approximated by the "rule of three"[41] to be (0.85,1.0).

Then, we used the ML models established for 5-mer to classify 6-mer library, which contains 4096 (4^6) sequences. Note that we have discarded the models using *psv* for 6-mer classification because *psv* can only handle sequences of the same length (5-mer). The trained ML models predict that 744 of 6-mers are positive sequences, and 3352 of them are negative sequences. We experimentally tested the first sequence set including 10 positive and 10 negative sequences (Table 2), which were selected following the rule with d = 3.

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX−XXX

Table 2. Six-mer DNA Sequences Predicted by the ML-Model and Tested by the ATP Technique[a]

selected positive 6-mer sequences for test				selected negative 6-mer sequences for test			
name	sequence	class	purified CNTs	name	sequence	class	purified CNTs
P6-01	GCCGCG	Y	(6,5)	N6-01	TAATTA	N	N/A
P6-02	CGCCGG	Y	(6,5), (7,6)	N6-02	TATAAA	N	N/A
P6-03	ACGCCG	Y	(7,6), (8,4)	N6-03	ATAATA	N	N/A
P6-04	GCCCGA	Y	(6,5), (7,6)	N6-04	TTAAAT	N	N/A
P6-05	GCGACC	Y	(6,4)	N6-05	AATTAG	Y	(7,6)
P6-06	ACCGGC	Y	(6,5)	N6-06	GTATAA	N	N/A
P6-07	GGACGC	Y	(6,5), (7,6), (9,4)	N6-07	AATAGT	Y	(7,6), (8,6)
P6-08	CCGGCA	Y	(6,5), (7,6)	N6-08	CAATAT	N	N/A
P6-09	CGGCCT	Y	(6,5)	N6-09	AATGTA	N	N/A
P6-10	GAGCGG	Y	(7,6), (8,5), (8,6)	N6-10	TTTATC	N	N/A

[a]Y denotes a resolving sequence, and N denotes an nonresolving sequence.

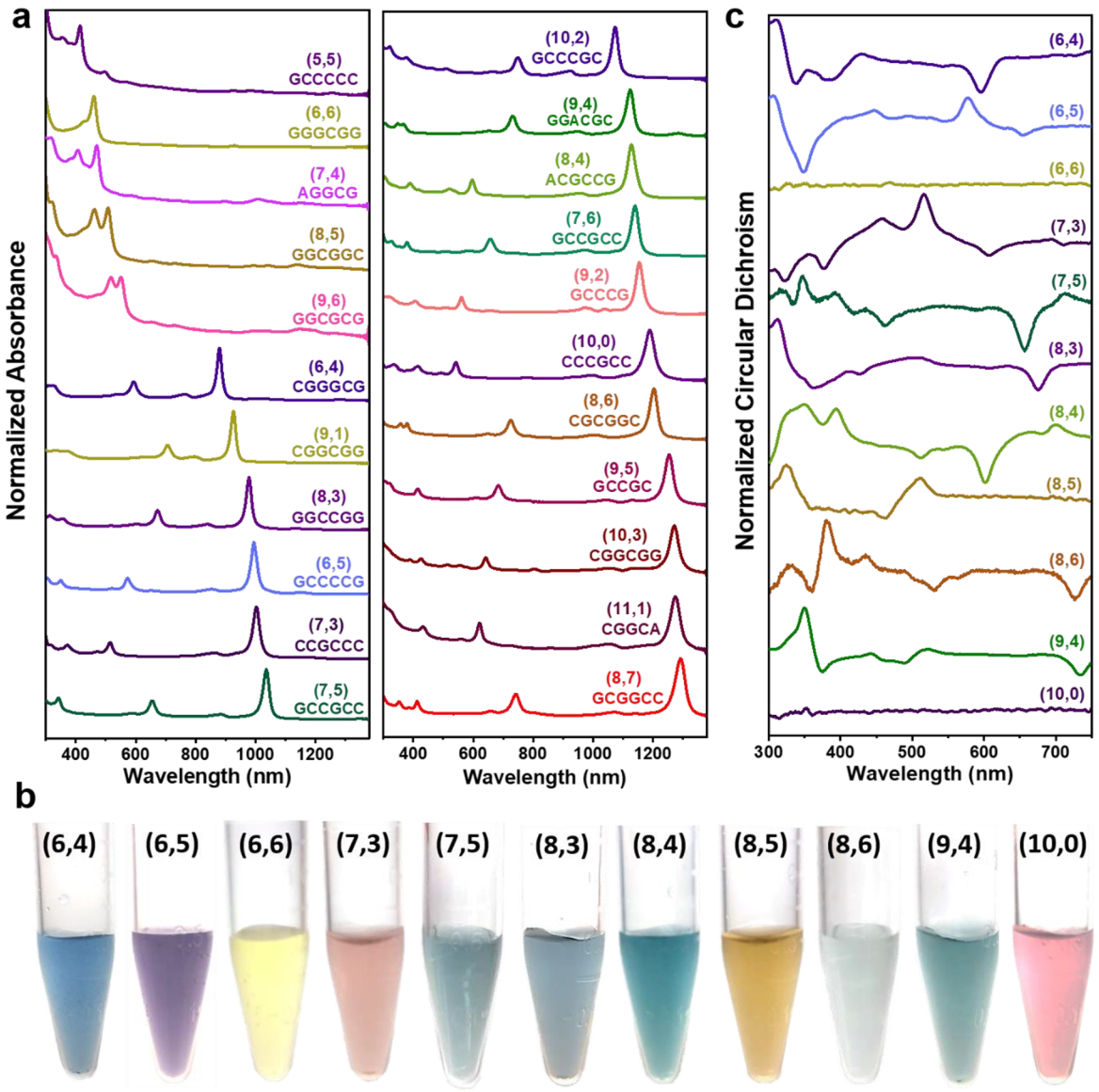


Figure 3. Optical characterization of distinct SWCNT species purified by 5-mer and 6-mer resolving sequences. (a) Absorbance spectra of the single chirality SWCNT species and corresponding resolving sequences. Each spectrum is normalized at the E_{11} peak position. (b) Photographs of a subset of single chirality SWCNT species. The photograph of each (n, m) species was taken after concentrating the original fraction by precipitation and resuspension in water. (c) CD spectra of the selected 11 SWCNTs species. They are normalized at the E_{22} peak position (M_{11} for metallic species). The original CD spectra can be seen in Figure S111.

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX−XXX

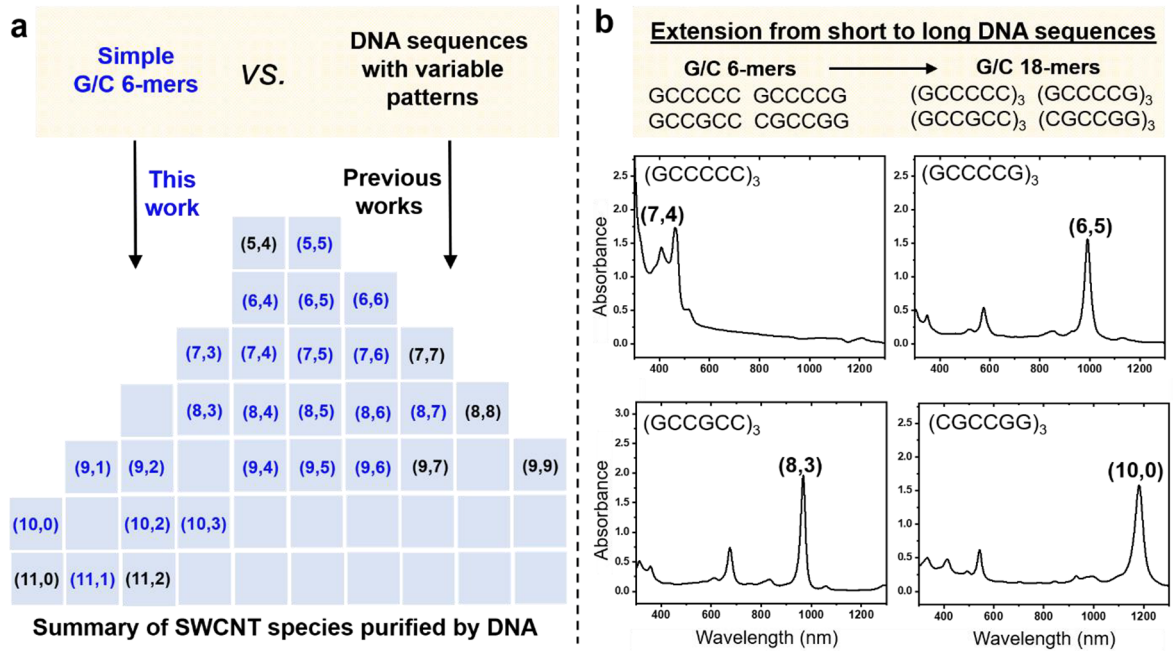


Figure 4. Discovery of super-resolving G/C patterns. (a) Summary of all SWCNT (*n*, *m*) species sorted by DNA. In the literature, one has to use resolving sequences with dramatically different patterns to get these species, while in this work, simple G/C 6-mers can purify the majority of them. (b) Extension of a G/C resolving pattern from a short to long DNA subspace. Absorbance spectra of the single chirality SWCNT species purified by four 18-mers G/C resolving sequences.

The detailed ATP separation results can be found in Figures S34−S45. We verified that the positive set are all resolving sequences. Within the negative set, we found eight non-resolving sequences and two resolving sequences. Specifically, N6-05 and N6-07 were predicted to be negative sequences but were found to be able to purify (7,6) and (8,6) species efficiently. To further verify the accuracy of 6-mer classification model, we tested a second sequence set, which included 62 positive sequences (Table S2). Notably, they included all G/C containing sequences. The experimental results show that 57 out of 62 are resolving sequences (Figures S46−S107). Taking these results together, we tested 82 predicted sequences in 6-mers. There are only seven sequences showing different results from prediction, with a prediction efficiency of 91.5%. This suggests that the model trained on 5-mers works very well for 6-mers. The success rate of 91.5% for a sample size of 82 gives 95% confidence intervals for these Bernoulli trials as (0.832, 0.965).[41] Considering the prediction accuracy in both 5-mer (100%) and 6-mer (91.5%), this is a substantial improvement over the ∼10% frequency of discovering resolving sequences by empirical search in previous works.

Figure 3a shows the UV−vis-NIR absorbance spectra of 22 single chirality SWCNT species that we have purified to date by distinct 5-mers and 6-mers. Different from previous work where distinct species were isolated in a variety of ATP systems,[36,37] all of these 22 species can be readily separated from the same PEG1.5 kDa/DX250 kDa ATP system, suggesting the high resolution of these short resolving sequences. It is worth mentioning that some sequences (e.g., CCCGCC) can purify five different SWCNTs (*n*, *m*) (Figure S49). A detailed description of the separation procedure employed for the separation of each species is presented in the Supporting Information. These species include armchair metallic tubes [(5,5) and (6,6)], nonarmchair quasi-metallic tubes [(7,4), (8,5) and (9,6)], and semiconducting tubes with a broad range of diameter and chiral angles ranging from pure zigzag tubes like (10,0) to near-armchair tubes like (8,7). Moreover, with the appropriately selected DNA sequences, even low-abundance species such as (9,5), zigzag (10,0), and near zigzag (10,2) and (11,1) can be efficiently extracted from the parent DNA−SWCNT dispersion. Figure 3b illustrates a photograph of 11 distinct (*n*, *m*) species after their original fractions were concentrated by precipitation and resuspension.

In addition to helicity, handedness is also an important structure feature of SWCNTs. We measured the circular dichroism (CD) spectra of 11 representative species shown in Figure 3c. To make a quantitative comparison of relative enantiomeric purities, we normalized CD spectra (CD_{norm}) by the absorption of the E_{22} peak (M_{11} for metallic species). Their original spectra can be seen in Figure S111. The zigzag tubes (10,0) and armchair tubes (6,6) exhibit a featureless CD profile, consistent with their achiral nature. The rest of nine chiral tubes all demonstrate the characteristic profile of enantiomerically pure species. Table S5 summarizes the CD_{norm} values of these nine species and the highest values reported in the literature. Notably, we found that the nanotube enantiomers of (6,4), (7,3), (7,5), (8,4), (8,5), (8,6), and (9,4) have even higher normalized CD intensities than the highest values reported. These results suggest that the resolving sequences we identified give rise to pure nanotubes with not only a broad range of helicities but also highly enantiomeric purities.

Figure 4a summarizes all 29 single chirality nanotubes that have been previously purified by DNA in the literature.[36,37,39,42] To get these species, hundreds of long DNA sequences were screened to find the resolving ones. Since these resolving sequences possess dramatically different patterns, it is very difficult for one to extract a rule to guide the search of new resolving sequences. Now, we are able to purify most species using G/C 6-mers. Note that there are a total of only 64 G/C

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX−XXX

6-mers. We merely miss seven low-abundance species [(5,4), (7,7), (8,8), (9,7), (9,9), (11,0), (11,2)] while isolating a new species (9,5). Given the super-resolving feature of G/C patterns, we hypothesize that they may be extended to a longer DNA sequence subspace. To test this hypothesis, we selected four 6-mer G/C sequences containing 1G, 2G and 3G, namely, GCCCCC, GCCCCG, GCCGCC, and CGCCGG, and tripled their length to 18-mers $(GCCCCC)_3$, $(GCCCCG)_3$, $(GCCGCC)_3$, and $(CGCCGG)_3$, respectively. These four 18-mers were examined by the ATP technique. Strikingly, they are all resolving sequences giving rise to four different types of single chirality species [(7,4), (6,5), (8,3), and (10,0)] (Figure 4b and Figures S108–S109). We then further tested two G/C sequences containing 36 nucleotides (36-mers), $(GCCCCG)_6$, and $(GCCGCC)_6$, and verified them as resolving sequences of (6,5) and (8,3) (Figure S110). These results suggest that the G/C resolving sequence pattern identified through a short-sequence study can be extended to longer length sequence subspaces, enabling us to discover many more resolving sequences in longer DNA subspaces.

CONCLUSIONS

In conclusion, we demonstrate in this work a ML-aided systematic search strategy to identify DNA sequences for SWCNT sorting. To address the challenge of enormous sequence combinations, we initiated our search from very short DNA sequences (e.g., 5-mers) and step-by-step extended to a longer sequence subspace. The short sequences allow us to screen the whole sequence subspaces, providing a solid foundation for systematic search. Relying on a cycle of ML-prediction and experiment-verification, we created a highly accurate model for 5-mer and 6-mer sequence classification with a success rate of near 100% and 92%, respectively, and identified hundreds of resolving sequences for sorting SWCNTs. The tested resolving sequences gave rise to 22 pure single chirality species with a broad range of chirality and highly enantiomeric purities. Moreover, we uncovered that the combination of G/C is super-resolving patterns, and they can be further extended to longer sequence subspaces.

METHODS

Preparation of DNA–SWCNT Dispersions. DNA–SWCNT dispersions were prepared according to previously published procedures.[36,37] Briefly, CoMoCAT SWCNT powders were dispersed in aqueous DNA and salt solutions by tip sonication (model VCX130, Sonics and Materials, Inc.) in an ice bath for 1.5 h with a power level of 8 W. The SWCNT concentration was kept constant at 1 mg/mL, the SWCNT/DNA mass ratios were 1:2.5 for SG65i and 1:2 for EG150x, and the salt concentration was 30 mmol/L. Unless indicated otherwise, the salts used were 1 mol/L pH = 4 sodium phosphate buffer (NaPB). For the dispersions that are specifically indicated with pH = 7, the salts used were 1 mol/L pH = 7 NaPB. After sonication, the raw dispersions were centrifuged at 17000g (g = 9.81 m/s^2) for 90 min. The supernatant of the DNA-SWCNT dispersion was collected and used for the ATP separation at room temperature.

Multistep Sorting of DNA–SWCNT by the ATP Technique. For multistep sorting of DNA–SWCNT, the following procedures were performed. Stock ATP solutions of PEG 1.5 kDa and DX 250 kDa (with two compositions 10:0 and 7:3) were prepared following the procedures identified in our previous publication.[39] The mass fractions of PEG 1.5 kDa and DX 250 kDa are 9.0% and 10.6% in 10:0 stock solution, and 12.2% and 14.4% in 7:3 stock solution, respectively.

In a typical sorting experiment (Figure S2a), 3 vol (120 μL) of SWCNT–DNA dispersion was loaded into 7 vol (280 μL) of the 7:3 stock solution, giving a 10:0 ATP system at room temperature. The first top fraction was extracted out and labeled as 0T. Then, an equal volume of blank 10:0 top phase and 2 μL of modulating agents PVP (with a mass concentration ranging from 1%, 2%, 5%, 10% to 20%) were sequentially added for the following extractions. The low concentration of PVP was added for early fractions, and the high concentration of PVP was added for later fractions. After vortex mixing and gentle centrifugation, a new top phase was formed and extracted (labeled as 1T to pT). This process can be repeated multiple times until the bottom fraction becomes colorless. To improve the purity, the fractions containing targeted DNA–SWCNT species (e.g., mT) can be incubated at 4 °C overnight or 24 h (Figure S2b), inducing a new two-phase separation (mT-T and mT-B). The pure DNA–SWCNT species were precipitated by adding 5 mol/L NaSCN to reach a final salt concentration of 0.5 mol/L and centrifuged at 17000 g (g = 9.81 m/s^2) for 10 min. The pellets were resuspended in deionized H_2O for characterization, and then free DNA was added to a final concentration of 0.1 mg/mL for SWCNT dispersion stabilization.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnano.1c11448.

Detailed sequence information, multistep ATP separation results of selected sequences, and additional figures and tables (PDF)

AUTHOR INFORMATION

Corresponding Authors

Zhiwei Lin – *Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States;* Present Address: South China Advanced Institute for Soft Matter Science and Technology, School of Emergent Soft Matter, South China University of Technology, Guangzhou 510640, China; orcid.org/0000-0001-9194-1145; Email: zhiweilin@scut.edu.cn

Ming Zheng – *Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States;* orcid.org/0000-0002-8058-1348; Email: ming.zheng@nist.gov

Authors

Yoona Yang – *Department of Chemical & Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States;* orcid.org/0000-0002-7446-8816

Anand Jagota – *Departments of Chemical & Biomolecular Engineering and of Bioengineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States;* orcid.org/0000-0002-0779-0880

Complete contact information is available at: https://pubs.acs.org/10.1021/acsnano.1c11448

Author Contributions

#Z.L. and Y.Y. contributed equally to the work.

Author Contributions

Z.L. and M.Z. conceived the idea. Z.L. designed and conducted the experiments and analyzed the data. Z.L. and M.Z. wrote the manuscript. Y.Y. and A.J. conducted machine learning study and cowrote the manuscript.

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX–XXX

All codes along with sample data files are available at the following repository (10.5281/zenodo.5702497). They are also available at https://bitbucket.org/jagotagrouplehigh/ml-guided-dna-search.

ACKNOWLEDGMENTS

We thank Dr. Jeffrey Fagan, Dr. Zeus Allen De los Santos, and Dr. Arjun Sharma for helpful discussion and suggestions. This work was supported by NIST internal funding and by a dean's fellowship to Y.Y. at Lehigh University. A.J.'s contribution is part of the NHI initiative at Lehigh University.

REFERENCES

(1) Lutz, J.-F.; Ouchi, M.; Liu, D. R.; Sawamoto, M. Sequence-Controlled Polymers. *Science* **2013**, *341*, 1238149.
(2) Lutz, J.-F. Defining the Field of Sequence-Controlled Polymers. *Macromol. Rapid Commun.* **2017**, *38*, 1700582.
(3) Aldaye, F. A.; Palmer, A. L.; Sleiman, H. F. Assembling Materials with DNA as the Guide. *Science* **2008**, *321*, 1795−1799.
(4) Ellington, A. D.; Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **1990**, *346*, 818−822.
(5) Tuerk, C.; Gold, L. Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science* **1990**, *249*, 505−510.
(6) Sun, H.; Zhu, X.; Lu, P. Y.; Rosato, R. R.; Tan, W.; Zu, Y. Oligonucleotide aptamers: new tools for targeted cancer therapy. *Mol. Ther. Nucleic Acids* **2014**, *3*, No. e182.
(7) Herzner, A.-M.; Hagmann, C. A.; Goldeck, M.; Wolter, S.; Kübler, K.; Wittmann, S.; Gramberg, T.; Andreeva, L.; Hopfner, K.-P.; Mertens, C.; Zillinger, T.; Jin, T.; Xiao, T. S.; Bartok, E.; Coch, C.; Ackermann, D.; Hornung, V.; Ludwig, J.; Barchet, W.; Hartmann, G.; Schlee, M. Sequence-specific activation of the DNA sensor cGAS by Y-form DNA structures as found in primary HIV-1 cDNA. *Nat. Immunol.* **2015**, *16*, 1025−1033.
(8) Yaari, Z.; Yang, Y.; Apfelbaum, E.; Cupo, C.; Settle, A. H.; Cullen, Q.; Cai, W.; Roche, K. L.; Levine, D. A.; Fleisher, M.; Ramanathan, L.; Zheng, M.; Jagota, A.; Heller, D. A. A perception-based nanosensor platform to detect cancer biomarkers. *Sci. Adv.* **2021**, *7*, No. eabj0852.
(9) Lee, J. Y.; Kim, Y. J.; Lee, C.; Lee, J. G.; Yagyu, H.; Tabata, O.; Kim, D. N. Investigating the sequence-dependent mechanical properties of DNA nicks for applications in twisted DNA nanostructure design. *Nucleic Acids Res.* **2019**, *47*, 93−102.
(10) Wu, J.; Tan, L. H.; Hwang, K.; Xing, H.; Wu, P.; Li, W.; Lu, Y. DNA Sequence-Dependent Morphological Evolution of Silver Nanoparticles and Their Optical and Hybridization Properties. *J. Am. Chem. Soc.* **2014**, *136*, 15195−15202.
(11) Copp, S. M.; Bogdanov, P.; Debord, M.; Singh, A.; Gwinn, E. Base Motif Recognition and Design of DNA Templates for Fluorescent Silver Clusters by Machine Learning. *Adv. Mater.* **2014**, *26*, 5839−5845.
(12) Xu, Q.; Schlabach, M. R.; Hannon, G. J.; Elledge, S. J. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 2289−2294.
(13) Yang, F.; Wang, M.; Zhang, D.; Yang, J.; Zheng, M.; Li, Y. Chirality Pure Carbon Nanotubes: Growth, Sorting, and Characterization. *Chem. Rev.* **2020**, *120*, 2693−2758.
(14) Bati, A. S. R.; Yu, L.; Batmunkh, M.; Shapter, J. G. Synthesis, purification, properties and characterization of sorted single-walled carbon nanotubes. *Nanoscale* **2018**, *10*, 22087−22139.
(15) Bachilo, S. M.; Strano, M. S.; Kittrell, C.; Hauge, R. H.; Smalley, R. E.; Weisman, R. B. Structure-Assigned Optical Spectra of Single-Walled Carbon Nanotubes. *Science* **2002**, *298*, 2361−2366.
(16) Zhang, J.; Landry, M. P.; Barone, P. W.; Kim, J.-H.; Lin, S.; Ulissi, Z. W.; Lin, D.; Mu, B.; Boghossian, A. A.; Hilmer, A. J.; Rwei, A.; Hinckley, A. C.; Kruss, S.; Shandell, M. A.; Nair, N.; Blake, S.; Şen, F.; Şen, S.; Croy, R. G.; Li, D.; Yum, K.; Ahn, J.-H.; Jin, H.; Heller, D. A.; Essigmann, J. M.; Blankschtein, D.; Strano, M. S. Molecular recognition using corona phase complexes made of synthetic polymers adsorbed on carbon nanotubes. *Nat. Nanotechnol.* **2013**, *8*, 959−968.
(17) Gillen, A. J.; Boghossian, A. A. Non-covalent Methods of Engineering Optical Sensors Based on Single-Walled Carbon Nanotubes. *Front. Chem.* **2019**, *7*, 612.
(18) Barone, P. W.; Baik, S.; Heller, D. A.; Strano, M. S. Near-infrared optical sensors based on single-walled carbon nanotubes. *Nat. Mater.* **2004**, *4*, 86−92.
(19) Heller, D. A.; Jeng, E. S.; Yeung, T.-K.; Martinez, B. M.; Moll, A. E.; Gastala, J. B.; Strano, M. S. Optical Detection of DNA Conformational Polymorphism on Single-Walled Carbon Nanotubes. *Science* **2006**, *311*, 508−511.
(20) Landry, M. P.; Ando, H.; Chen, A. Y.; Cao, J.; Kottadiel, V. I.; Chio, L.; Yang, D.; Dong, J.; Lu, T. K.; Strano, M. S. Single-molecule detection of protein efflux from microorganisms using fluorescent single-walled carbon nanotube sensor arrays. *Nat. Nanotechnol.* **2017**, *12*, 368−377.
(21) Demirer, G. S.; Zhang, H.; Goh, N. S.; Pinals, R. L.; Chang, R.; Landry, M. P. Carbon nanocarriers deliver siRNA to intact plant cells for efficient gene knockdown. *Sci. Adv.* **2020**, *6*, No. eaaz0495.
(22) Harvey, J. D.; Jena, P. V.; Baker, H. A.; Zerze, G. H.; Williams, R. M.; Galassi, T. V.; Roxbury, D.; Mittal, J.; Heller, D. A. A carbon nanotube reporter of microRNA hybridization events in vivo. *Nat. Biomed. Eng.* **2017**, *1*, 0041.
(23) Piao, Y.; Meany, B.; Powell, L. R.; Valley, N.; Kwon, H.; Schatz, G. C.; Wang, Y. Brightening of carbon nanotube photoluminescence through the incorporation of sp3 defects. *Nat. Chem.* **2013**, *5*, 840−845.
(24) Brozena, A. H.; Kim, M.; Powell, L. R.; Wang, Y. Controlling the optical properties of carbon nanotubes with organic colour-centre quantum defects. *Nat. Rew. Chem.* **2019**, *3*, 375−392.
(25) He, X.; Htoon, H.; Doorn, S. K.; Pernice, W. H. P.; Pyatkov, F.; Krupke, R.; Jeantet, A.; Chassagneux, Y.; Voisin, C. Carbon nanotubes as emerging quantum-light sources. *Nat. Mater.* **2018**, *17*, 663−670.
(26) Nish, A.; Hwang, J.-Y.; Doig, J.; Nicholas, R. J. Highly selective dispersion of single-walled carbon nanotubes using aromatic polymers. *Nat. Nanotechnol.* **2007**, *2*, 640−646.
(27) Liu, H.; Nishide, D.; Tanaka, T.; Kataura, H. Large-scale single-chirality separation of single-wall carbon nanotubes by simple gel chromatography. *Nat. Commun.* **2011**, *2*, 309.
(28) Arnold, M. S.; Green, A. A.; Hulvat, J. F.; Stupp, S. I.; Hersam, M. C. Sorting carbon nanotubes by electronic structure using density differentiation. *Nat. Nanotechnol.* **2006**, *1*, 60−65.
(29) Ghosh, S.; Bachilo, S. M.; Weisman, R. B. Advanced sorting of single-walled carbon nanotubes by nonlinear density-gradient ultracentrifugation. *Nat. Nanotechnol.* **2010**, *5*, 443−450.
(30) Zheng, M. Sorting Carbon Nanotubes. *Top. Curr. Chem.* **2017**, *375*, 13.
(31) Zheng, M.; Jagota, A.; Semke, E. D.; Diner, B. A.; McLean, R. S.; Lustig, S. R.; Richardson, R. E.; Tassi, N. G. DNA-assisted dispersion and separation of carbon nanotubes. *Nat. Mater.* **2003**, *2*, 338−342.
(32) Tu, X.; Manohar, S.; Jagota, A.; Zheng, M. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature* **2009**, *460*, 250−253.
(33) Roxbury, D.; Mittal, J.; Jagota, A. Molecular-Basis of Single-Walled Carbon Nanotube Recognition by Single-Stranded DNA. *Nano Lett.* **2012**, *12*, 1464−1469.
(34) Khripin, C. Y.; Fagan, J. A.; Zheng, M. Spontaneous partition of carbon nanotubes in polymer-modified aqueous phases. *J. Am. Chem. Soc.* **2013**, *135*, 6822−6825.
(35) Fagan, J. A. Aqueous two-polymer phase extraction of single-wall carbon nanotubes using surfactants. *Nanoscale Adv.* **2019**, *1*, 3307−3324.
(36) Ao, G.; Khripin, C. Y.; Zheng, M. DNA-controlled partition of carbon nanotubes in polymer aqueous two-phase systems. *J. Am. Chem. Soc.* **2014**, *136*, 10383−10392.

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX−XXX

(37) Ao, G.; Streit, J. K.; Fagan, J. A.; Zheng, M. Differentiating Left- and Right-Handed Carbon Nanotubes by DNA. *J. Am. Chem. Soc.* **2016**, *138*, 16677−16685.
(38) Yang, Y.; Zheng, M.; Jagota, A. Learning to predict single-wall carbon nanotube-recognition DNA sequences. *npj Comput. Mater.* **2019**, *5*, 3.
(39) Lyu, M.; Meany, B.; Yang, J.; Li, Y.; Zheng, M. Toward Complete Resolution of DNA/Carbon Nanotube Hybrids by Aqueous Two-Phase Systems. *J. Am. Chem. Soc.* **2019**, *141*, 20177−20186.
(40) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J. T.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(41) Hanley, J. A.; Lippman-Hand, A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* **1983**, *249*, 1743−1745.
(42) Tu, X.; Hight Walker, A. R.; Khripin, C. Y.; Zheng, M. Evolution of DNA Sequences Toward Recognition of Metallic Armchair Carbon Nanotubes. *J. Am. Chem. Soc.* **2011**, *133*, 12998−13001.

https://doi.org/10.1021/acsnano.1c11448
ACS Nano XXXX, XXX, XXX−XXX