8-1-2019

Learning about Sequence-Dependent DNA/Single-Wall Carbon Nanotube Hybrids

Yoona Yang
Lehigh University

Learning about Sequence-Dependent DNA/Single-Wall Carbon Nanotube Hybrids

by

Yoona Yang

Presented to Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Chemical Engineering

Lehigh University

August 2019

Certificate of Approval

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Date

_____
Prof. Anand Jagota,
Dissertation Director

_____
Accepted Date

Committee Members:

_____
Prof. Hugo S. Caram

_____
Prof. Javier Buceta

_____
Dr. Ming Zheng

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Anand Jagota. I consider myself fortunate to join Jagota group and to work with him. Your deep insights and broad perspectives have always inspired me. With your continuous support and confidence in me, I could motivate myself and realize my capability to overcome challenges. I sincerely appreciate everything you have invested in me.

I also would like to extend my sincere gratitude to all the members of my dissertation committee: Professor Hugo Caram and Professor Javier Buceta at Lehigh University and Dr. Ming Zheng at NIST in Gaithersburg, MD, for many valuable comments and guidance on this thesis. A special thanks to Ming for collaborating on most of the projects presented in this thesis and assistance you have given me for the aqueous two-phase separation experiment at NIST. I also would like to say thanks to Professor Roger French and other members of SDLE center at Case Western Reserve University in Cleveland, Ohio, for allowing me the opportunity to visit your team and broaden my data science knowledge including big data technologies and improved data collection techniques, as well as learning how to work with multiple team members. Additionally, thanks to Dr. Danial Heller and Dr. Zvi Yaari at the Memorial Sloan Kettering Cancer Center in New York, NY, for collaborating on *Molecular Perceptron* project and providing me the valuable advice with the biomedical perspective.

Also, I would like to thank Professor Myungsook Oh and Professor Won Sun Ryoo, Professor of Department of Chemical Engineering at Hongik University, Republic of Korea. My interest in Chemical Engineering coalesced when I was an undergraduate at Hongik University, where I was fortunate to take several excellent courses from Professor Oh. In addition, she encouraged me to M.S. under the supervision of Professor Ryoo, instead of rushing into a Ph.D. program. Professor Ryoo, my undergraduate and graduate research advisor, helped me to become a better researcher and provide me the opportunity to present academic research. Not only have they been a great teacher to me, but they have been a wonderful mentor. I am grateful for their mentorship and guidance while becoming my role model.

I wish to record my deep gratitude to all of my lab mates that I have worked with at Lehigh University: Dr. Akshaya Shankar, Dr. Nicole Lapinski, Dr. Arjun Sharma, Nichole Moyle, Thibault Aryaksama, Guillaume Noetinger, Xinyu Cui, Meng Li, and Luke Wang. Especially thanks to Akshaya, Nicole, and Nichole; we had the best time together, that is what I will miss the most. I also thank my best friend, Sukyung, for cheering me up over the years. I was again fortunate to meet friends around here: Jinny, Jookyung, Jihyun, Jae Hyung, Heuijoon, Tony, and everyone from the Korean Association. I will always cherish the memories we made together.

Last but not least, I would like to express my sincere appreciation to my family – my parents, Imseung and Myungja; my brother's family, Dongheon, Eunjin, Yoonseo, and Soobin, and my grandparents, Hwoiwan and Kyungsoon. They have supported and encouraged me throughout my entire education and finally during my Ph.D. even I have

been thousands of miles away from them for my studies. Without their selfless love and faith in me, I would never be who I am today. I dedicate this thesis to them. Also, my thanks go to my dear friend, Steve, who always stand by me to cheer me up and being there for me through thick and thin.

To everyone who has helped me along the way, thank you all for your contributions.

Yoona Yang

August 2019

Table of Contents

List of Tables

List of Figures

Abstract

Since the single-wall carbon nanotubes (SWCNTs) were discovered in 1993, they have attracted significant interest with their extraordinary electrical and optical properties in addition to their remarkable mechanical strength and thermal conductivity. Single-stranded DNA conjugated SWCNT have shown outstanding functionality in terms of dispersibility and biocompatibility. In addition, some special DNA sequences have presented an ability to recognize specific SWCNT species, called recognition sequences. Ion-exchange chromatography and aqueous two-phase (ATP) separation technique have been widely used for SWCNT separation. However, little is known about the use of ATP as an analytical technique. Furthermore, for bio-applications, DNA/SWCNT hybrids have attracted significant interest due to their high solvatochromic sensitivity to changes in the local environment, which enables their use as sensors. Recognition properties can provide good candidates for molecular detection on the assumption that the recognition DNA/SWCNT hybrids have structurally well-defined DNA wrappings. Thus, there is a growing need for discovery of new recognition sequences. In this thesis, we explore new methods to quantify difference in solvation/binding characteristics using ATP, and a new approach to predicting recognition sequences using Machine Learning techniques. Finally, a new concept for a DNA/SWCNT-based sensing system is demonstrated.

Chapter 1 : Introduction

1.1 Single-Walled Carbon Nanotube

A single-wall carbon nanotube (SWCNT) is a cylindrical nanostructure, that can be thought of as made by wrapping a single layer of graphene into a seamless tube. Its properties are sensitively determined by the precise way in which the graphene is rolled. SWCNT diameters vary from 0.4 to 2 nm, while their length is in the micrometer range.[1,2] Due to this quasi-one-dimensional character, SWCNTs have unique physical, electrical, and optical properties.[3] Since these nanostructures were discovered by Iijima et al.[4] and Bethune et al.[5], numerous studies have been conducted to understand their remarkable properties and to explore possible applications.

The structure and properties of a given SWCNT are uniquely defined by a chiral vector, $\boldsymbol{C_h}$, connecting two sites on the two-dimensional graphene sheet (Figure 1.1). By introducing basis vectors of the hexagonal honeycomb lattice, $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$, a chiral vector can be expressed by

$$\boldsymbol{C_h} = n\boldsymbol{a_1} + m\boldsymbol{a_2}$$

Thus, any SWCNT can be described by the pair of integers (n,m). For example, the chiral vector shown in Figure 1.1 describes a (6,2)-SWCNT.

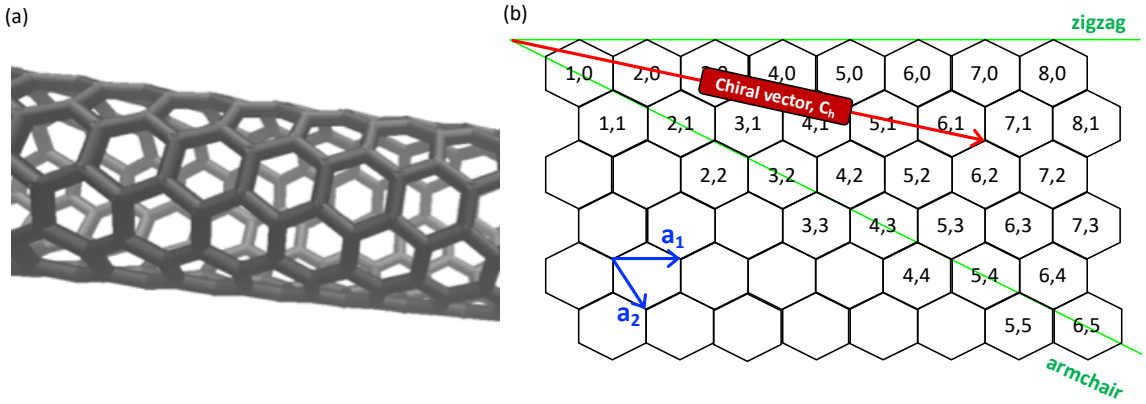

Figure 1.1. (a) A molecular representation of a semi-conducting (6,5)-SWCNT. (b) SWCNT configurations with the chiral vector $\boldsymbol{C_h}$ and basis vectors $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$. For example, the chiral vector shown by red arrow describes the (6,2)-SWCNT.

1.2 Single-Stranded Deoxyribonucleic Acid (DNA) as Aptamer

Deoxyribonucleic acid (DNA) is a biomolecule composed of nucleotides. Each nucleotide is made up of a phosphate group, a sugar group and a nucleobase.[6] There are four types of nucleobases: adenine (A), thymine (T), guanine (G) and cytosine (C). The nucleotides are attached to one another to form a strand by covalent bonding between the phosphate group of one nucleotide and the sugar group of the next. In double-stranded DNA, the nucleobases on one strand are bound to the nucleobases on another strand, following Watson-Crick base-pairing rules[6], as presented in Figure 1.2. Double-stranded DNA often exists in the celebrated two-strand, double-helix (B-DNA) structure. Other forms, including the A, Z[6], and S[7] double helices or special structures such the G-quartets formed by G-rich sequences[8] are, broadly speaking, exceptions.

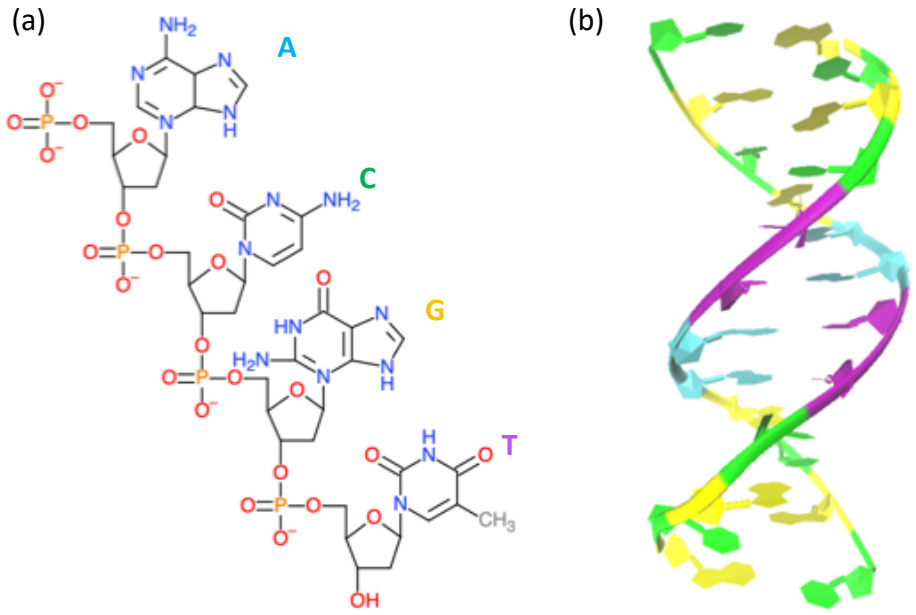

Figure 1.2. (a) Schematic representation of single-stranded DNA. (b) A molecular representation of a double-stranded DNA. The nucleobases A, C, G, and T are represented in cyan, green, yellow, and purple, respectively.

Some DNA or ribonucleic acid (RNA) sequences, called aptamers, can specifically recognize a target molecule.[9] Among the biological polymers, single-stranded DNA (ssDNA) is of great interest because of its unique and well-defined composition that can be chosen from a gigantic library, as well as bio-comparability.[10] Furthermore, it is known that ssDNA aptamers are capable of selective binding to specific target molecules from biological molecules[11,12] to single chirality SWCNT, called '*recognition sequence*'.[13,14]

1.3 Single-Stranded DNA/SWCNT Hybrid

During past few years hybrids of SWCNT and biomolecules have attracted significant interest as bio-sensors for specific molecule detection[15–17], targeted drug

delivery[18], and in-vivo imaging[19]. However, there are significant barriers to applying SWCNT in those applications. First, due to its hydrophobicity, the SWCNT needs to be functionalized, to be compatible with aqueous media. In addition, current fabrication methods always produce a mixture of different SWCNTs with varying chirality. Considerable efforts have been made to disperse individual SWCNTs and to sort them by length[20,21], diameter[22], and chirality using dispersal agents including small molecule surfactants and biological polymers[23–25]. Among these, DNA/SWCNT hybrids stand out for their remarkable colloidal stability.

It is known that the specificity and high binding affinity of aptamers to target molecules can be achieved by DNA secondary structure motifs.[26] In ssDNA-SWCNT hybrids, ssDNA sequences generally wrap SWCNT helically due to the intrinsic curvature of sugar-phosphate backbone as well as $\pi - \pi$ stacking between bases and the SWCNT surface. Since the backbone is not specific to the sequence, the selective binding characteristic of ssDNA/SWCNT are likely due to the differences in the orientation of the nucleobases. Several studies on computational molecular modeling of DNA/SWCNT[27–31] have established a number of ordered structural motifs that ssDNA can adopt when adsorbed onto an SWCNT (Figure 1.3). This indicates that the structural motifs of ssDNA in its adsorbed state are significant in identifying its specific binding characteristics, which provide a basis for SWCNT separation and molecule detection.

There are two separation methods that are primarily used for SWCNT separation: ion-exchange chromatography (IEX),[13,32] and aqueous two-phase (ATP) separation.[14,33] First, the ion-exchange chromatography method is based on differences in electrostatics

and hydrophobic interaction between the DNA/SWCNT hybrid and the positively charged column. A notable success has been made by Tu et. al.[13]; 26 recognition sequences were identified from a total of 350 sequences by systematic search of the vast DNA library. Next, more recently, the aqueous two-phase separation technique has been used with remarkable success for SWCNT separation.[14,33] The ATP separation is based on aqueous solutions of water-soluble polymers that separate into two phases.[34] The DNA/SWCNT hybrid partitions into the two phases based on small differences in solvation free energy, and this can be modulated by a modulant molecule. Thus, the ATP system potentially can offer a way to quantify and rank the solvation properties of the DNA/SWCNT hybrid. Using the ATP technique, Ao et. al[35] have designed a systematic albeit limited search of the DNA library by sequence pattern expansion, and successfully identified recognition sequences with a success rate of ~15%.

In addition to the SWCNT separation, DNA/SWCNT hybrids have attracted considerable interest due to their outstanding sensitivity to change in local environment, enabling their use as sensors.[15–17] It appears that the recognition sequences can form structurally well-defined DNA/SWCNT hybrid, which determine the pattern/size of exposed SWCNT surface. This can be the basis for molecular detection. Thus, recognition sequence discovery is essential in sensing application.

So far, sequence screening has relied on experimental work, which is costly and time-consuming. Moreover, the DNA sequence library is practically infinite in size. It is very expensive to explore the library experimentally. Several studies explored the library with the restriction of the chosen sequence expansion scheme, but the probability of finding

a recognition sequence still remains low (~15%).[35] Clearly, a different and more systematic approach to sequence prediction is needed. Furthermore, the traditional method[36] for preparation of DNA/SWCNT dispersion by direct sonication followed by centrifugation for removing impurities is highly laborious, expensive, and time-consuming. Recently, a new simple and rapid preparation method has been established by using replacement of surfactant on SWCNT by DNA aided by methanol.[37,38] However, little is known about the nature of exchange mechanism.

Further experimental studies have been conducted to understand the structural basis of sequence-specific recognition. Atomic force microscopy (AFM) based single-molecule force spectroscopy,[39,40] and solution based studies have provided quantitative information on binding free energies and activation free energy for displacement of DNA aptamers by surfactants.[41,42] Fluorescence quenching studies have been used to infer wrapping structures of recognition sequences.[43]

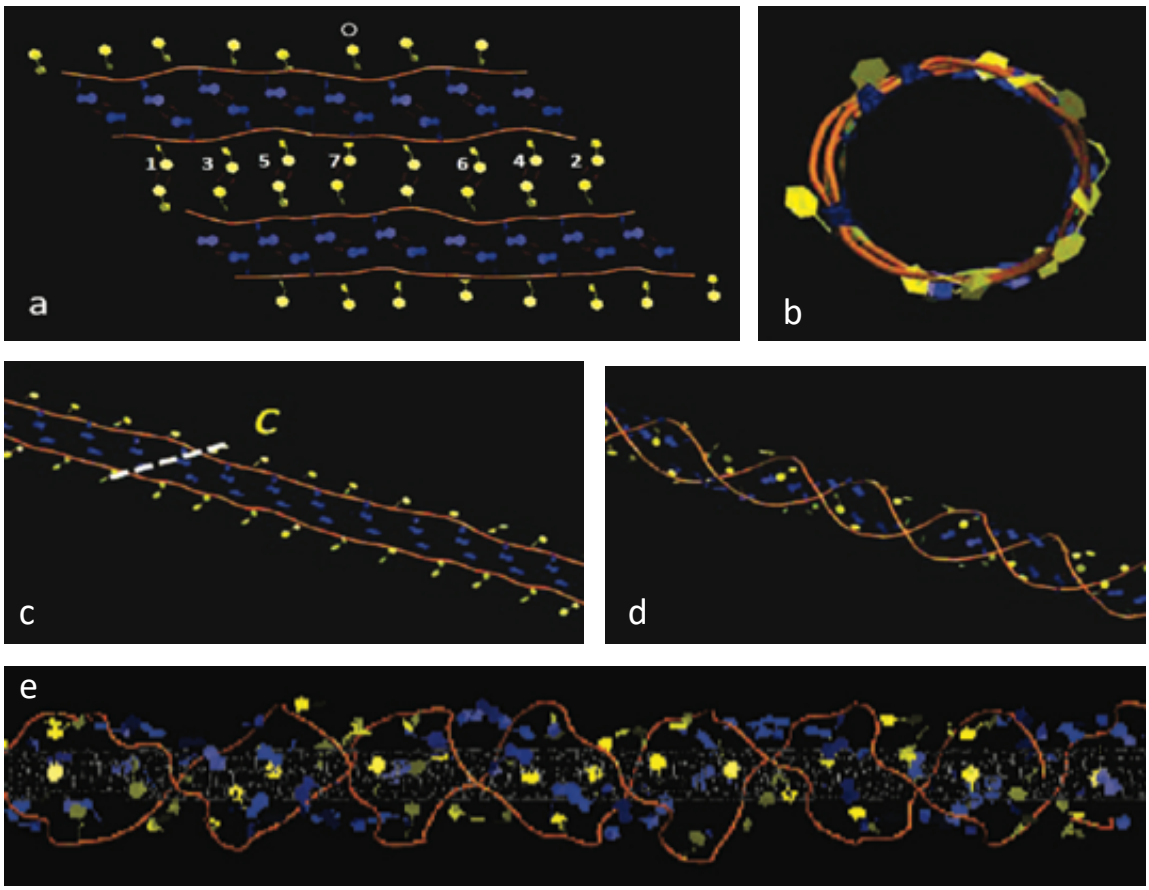

Figure 1.3.* Example of a DNA β-barrel structure based on non-Watson-Crick base pairing. (a) ssDNA strands placed adjacent and antiparallel to each other on a surface form a periodic hydrogen-bonded 2D sheet structure. (c) By following roll up vector (white dashed line), a β -barrel structure can be created (b, d). (e) ordered β-barrel structure with SWCNT.

* This image was taken from the literature:
Roxbury, D., Manohar, S. & Jagota, A. Molecular simulation of DNA β-sheet and β-barrel structures on graphite and carbon nanotubes. *J. Phys. Chem. C* **114,** 13267–13276 (2010).

1.4 Machine Learning in Bioinformatics

In recent years, Machine Learning has emerged as a powerful general methodology with the ability to create well-performing predictive models from data by recognizing unknown patterns, without requiring explicit programming instructions. In particular, machine learning techniques have become essential in bioinformatics because it is impractical to transform manually large amounts of raw sequence data into useful scientific knowledge. Many of the important bioinformatics problems are well suited for classification algorithms, including gene annotation,[44] protein function prediction,[45,46] peptide binding prediction,[47,48] and DNA binding prediction.[49]

1.5 Scope of Thesis

Although many studies have been conducted to understand DNA sequence-specific interaction on SWCNT, there is still considerable room for advancement. In this thesis, experimental studies have been performed to investigate the nature of the sequence-specific interaction between DNA and SWCNT. Furthermore, we combine the experimental and machine learning techniques to not only establish a predictive model for recognition sequence, but also to develop a new *perception*-based sensing system we call the *Molecular Perceptron*. Below is an outline of this study:

Chapter 2 presents a study of the DNA/SWCNT hybrid partition in the ATP system, specifically models for quantitatively analysis of the solvation characteristics of various DNA/SWCNT hybrids. Using these models, we extract relative solvation free energies and solubility parameters of various DNA/SWCNT hybrids. The two approaches

are found to be consistent, providing some confidence in each as a method of quantifying differences in solubility of various DNA/SWCNT hybrids.

Chapter 3 demonstrates the use of Machine Learning (ML) techniques to develop a predictive model for recognition DNA sequences. To date, even though numerous studies have reported a lot of physical understanding and a reasonable amount of data, our ability to predict recognition sequences is still absent. We built models that classify query sequences into recognition/non-recognition classes. Predictions were tested experimentally using the ATP separation technique.

Chapter 4 investigates the kinetics of a surfactant-based replacement process aided by methanol. We proposed a mechanistic model to analyze the kinetics and quantify differences in binding characteristics in terms of activation energy of the replacement process. Some recognition sequences showed significant difference in activation energy for different species of SWCNT, suggesting that the methanol-aided replacement process can be utilized as a promising low-cost and rapid way to identify recognition sequences.

Chapter 5 demonstrates a new *perception*-based sensing system for detection of the ovarian cancer serum biomarker HE4 in the presence or absence of fetal bovine serum (FBS) and bovine serum albumin (BSA) using ML techniques. The trained models successfully detect not only the target biomarker (HE4), but also other analytes (BSA and FBS). It is strongly suggestive of the idea that the perception mode of sensing can make accurate judgements in a noisy sensing environment.

1.6 References

1. Tang, Z. K., Wang, N., Li, G. D. & Chen, J. S. Materials science: Single-walled carbon nanotube arrays. *Nature* **408**, 50–51 (2000).
2. Avouris, P., Appenzeller, J., Martel, R. & Wind, S. J. Carbon nanotube electronics. *Proc. IEEE* **9**, 1772–1784 (2003).
3. Saito, R., Dresselhaus, G. & Dresselhaus, M. S. *Physical Properties of Carbon Nanotubes*. (Imperial College Press, 1998). doi:10.1142/p080
4. Iijima, S. & Ichihashi, T. Single-shell carbon nanotubes of 1-nm diameter. *Nature* **363**, 603–605 (1993).
5. Bethune, D. S. *et al.* Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls. *Nature* **363**, 605–607 (1993).
6. Saenger, W. *Principles of Nucleic Acid Structure*. (Springer New York, 1984). doi:10.1007/978-1-4612-5190-3
7. Bustamante, C., Bryant, Z. & Smith, S. B. Ten years of tension: single-molecule DNA mechanics. *Nature* **421**, 423–427 (2003).
8. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770–780 (2012).
9. Klussmann, S. *The aptamer handbook : functional oligonucleotides and their applications*. (Wiley-VCH, 2006).
10. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids. *Nature* **171**, 737–738 (1953).
11. Kim, Y. S. *et al.* Electrochemical detection of 17β-estradiol using DNA aptamer immobilized gold electrode chip. *Biosens. Bioelectron.* **22**, 2525–2531 (2007).
12. Su Jin Lee, † *et al.* ssDNA Aptamer-Based Surface Plasmon Resonance Biosensor for the Detection of Retinol Binding Protein 4 for the Early Diagnosis of Type 2 Diabetes. (2008). doi:10.1021/AC800050A
13. Tu, X., Manohar, S., Jagota, A. & Zheng, M. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature* **460**, 250–253 (2009).
14. Ao, G., Khripin, C. Y. & Zheng, M. DNA-controlled partition of carbon nanotubes in polymer aqueous two-phase systems. *J. Am. Chem. Soc.* **136**, 10383–10392 (2014).
15. Zhang, J. *et al.* Single Molecule Detection of Nitric Oxide Enabled by $d(AT)_{15}$ DNA Adsorbed to Near Infrared Fluorescent Single-Walled Carbon Nanotubes. *J. Am. Chem. Soc.* **133**, 567–581 (2011).
16. Shi, J. *et al.* Microbiosensors based on DNA modified single-walled carbon nanotube and Pt black nanocomposites. *Analyst* **136**, 4916 (2011).
17. Landry, M. P. *et al.* Single-molecule detection of protein efflux from microorganisms using fluorescent single-walled carbon nanotube sensor arrays. *Nat. Nanotechnol.* **12**, 368–377 (2017).
18. Zhang, W., Zhang, Z. & Zhang, Y. The application of carbon nanotubes in target drug delivery systems for cancer therapies. *Nanoscale Res. Lett.* **6**, 555 (2011).
19. Wen, J., Xu, Y., Li, H., Lu, A. & Sun, S. Recent applications of carbon

nanomaterials in fluorescence biosensing and bioimaging. *Chem. Commun.* **51**, 11346–11358 (2015).
20. Huang, X., Mclean, R. S. & Zheng, M. High-Resolution Length Sorting and Purification of DNA-Wrapped Carbon Nanotubes by Size-Exclusion Chromatography. *Anal. Chem.* **77**, 6225–6228 (2005).
21. Zheng, M. *et al.* Structure-Based Carbon Nanotube Sorting by Sequence-Dependent DNA Assembly. *Science (80-.).* **302**, 1545–1548 (2003).
22. Arnold, M. S., Stupp, S. I. & Hersam, M. C. Enrichment of Single-Walled Carbon Nanotubes by Diameter in Density Gradients. *Nano Lett.* **5**, 713–718 (2005).
23. Nish, A., Hwang, J.-Y., Doig, J. & Nicholas, R. J. Highly selective dispersion of single-walled carbon nanotubes using aromatic polymers. *Nat. Nanotechnol.* **2**, 640–646 (2007).
24. Liu, H., Nishide, D., Tanaka, T. & Kataura, H. Large-scale single-chirality separation of single-wall carbon nanotubes by simple gel chromatography. *Nat. Commun.* **2**, 309 (2011).
25. 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.* **1**, 60–65 (2006).
26. Lin, C. H. & Patei, D. J. Structural basis of DNA folding and recognition in an AMP-DNA aptamer complex: distinct architectures but common recognition motifs for DNA and RNA aptamers complexed to AMP. *Chem. Biol.* **4**, 817–832 (1997).
27. Johnson, R. R., A. T. Charlie Johnson & Klein, M. L. Probing the Structure of DNA−Carbon Nanotube Hybrids with Molecular Dynamics. (2007). doi:10.1021/NL071909J
28. Johnson, R. R., Kohlmeyer, A., Johnson, A. T. C. & Klein, M. L. Free Energy Landscape of a DNA−Carbon Nanotube Hybrid Using Replica Exchange Molecular Dynamics. *Nano Lett.* **9**, 537–541 (2009).
29. Roxbury, D., Manohar, S. & Jagota, A. Molecular simulation of DNA β-sheet and β-barrel structures on graphite and carbon nanotubes. *J. Phys. Chem. C* **114**, 13267–13276 (2010).
30. Roxbury, D., Jagota, A. & Mittal, J. Structural characteristics of oligomeric DNA strands adsorbed onto single-walled carbon nanotubes. *J. Phys. Chem. B* **117**, 132–140 (2013).
31. Shankar, A., Zheng, M. & Jagota, A. Energetic Basis of Single-Wall Carbon Nanotube Enantiomer Recognition by Single-Stranded DNA. *J. Phys. Chem. C* **121**, 17479–17487 (2017).
32. Zheng, M. *et al.* DNA-assisted dispersion and separation of carbon nanotubes. *Nat. Mater.* **2**, 338–342 (2003).
33. Khripin, C. Y., Fagan, J. A. & Zheng, M. Spontaneous partition of carbon nanotubes in polymer-modified aqueous phases. *J. Am. Chem. Soc.* **135**, 6822–6825 (2013).
34. Zaslavsky, B. Y. *Aqueous two-phase partitioning : physical chemistry and bioanalytical applications*. (M. Dekker, 1995).
35. Ao, G., Streit, J. K., Fagan, J. A. & Zheng, M. Differentiating Left- and Right-Handed Carbon Nanotubes by DNA. *J. Am. Chem. Soc.* **138**, 16677–16685 (2016).

36. Ao, G. & Zheng, M. Preparation and Separation of DNA-Wrapped Carbon Nanotubes. in *Current Protocols in Chemical Biology* **7**, 43–51 (John Wiley & Sons, Inc., 2015).
37. Giraldo, J. P. *et al.* A Ratiometric Sensor Using Single Chirality Near-Infrared Fluorescent Carbon Nanotubes: Application to In Vivo Monitoring. *Small* **11**, 3973–3984 (2015).
38. Streit, J. K., Fagan, J. A. & Zheng, M. A Low Energy Route to DNA-Wrapped Carbon Nanotubes via Replacement of Bile Salt Surfactants. *Anal. Chem.* **89**, 10496–10503 (2017).
39. Manohar, S. *et al.* Peeling single-stranded DNA from graphite surface to determine oligonucleotide binding energy by force spectroscopy. *Nano Lett.* **8**, 4365–72 (2008).
40. Iliafar, S., Mittal, J., Vezenov, D. & Jagota, A. Interaction of Single-Stranded DNA with Curved Carbon Nanotube Is Much Stronger Than with Flat Graphite. *J. Am. Chem. Soc.* **136**, 12947–12957 (2014).
41. Roxbury, D., Tu, X., Zheng, M. & Jagota, A. Recognition ability of DNA for carbon nanotubes correlates with their binding affinity. *Langmuir* **27**, 8282–8293 (2011).
42. Shankar, A., Mittal, J. & Jagota, A. Binding between DNA and carbon nanotubes strongly depends upon sequence and chirality. *Langmuir* **30**, 3176–3183 (2014).
43. Zheng, Y., Bachilo, S. M. & Weisman, R. B. Quenching of Single-Walled Carbon Nanotube Fluorescence by Dissolved Oxygen Reveals Selective Single-Stranded DNA Affinities. *J. Phys. Chem. Lett.* **8**, 1952–1955 (2017).
44. Gupta, R. *et al.* Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics* **11**, S65 (2010).
45. Zhao, X.-M., Wang, Y., Chen, L. & Aihara, K. Gene function prediction using labeled and unlabeled data. *BMC Bioinformatics* **9**, 57 (2008).
46. Clare, A. & King, R. D. Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics* **19**, ii42–ii49 (2003).
47. Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12**, 1007–1017 (2003).
48. Stiffler, M. A. *et al.* PDZ Domain Binding Selectivity Is Optimized Across the Mouse Proteome. *Science (80-.).* **317**, 364 LP – 369 (2007).
49. 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.* **26**, 5839–5845 (2014).

Chapter 2 : Quantification of DNA/SWCNT Solvation Differences by Aqueous Two Phase Separation[*]

Single wall carbon nanotubes (SWCNTs) coated with single-stranded DNA can be effectively separated into various chiralities using an aqueous two-phase (ATP) system. Partitioning is driven by small differences in the dissolution characteristics of the hybrid between the two phases. Thus, in addition to being a separation technique, the ATP system potentially also offers a way to quantify and rank the dissolution properties of the solute (here the DNA/SWCNT hybrids), such as the solvation free energy or solubility. In this study, we propose two different approaches to quantitatively analyze the ATP partitioning of DNA/SWCNT hybrids. First, we present a model that extracts relative solvation free energy of various DNA/SWCNT hybrids by using an expansion relative to a standard state. Second, we extract a solubility parameter by analyzing the partitioning of hybrids in the ATP system. The two approaches are found to be consistent, providing some confidence in each as a method of quantifying differences in solubility of various DNA/SWCNT hybrids.

[*] This chapter has been published in *Langmuir*:

Yang, Y., Shankar, A., Aryaksama, T., Zheng, M. & Jagota, A. Quantification of DNA/SWCNT Solvation Differences by Aqueous Two-Phase Separation. *Langmuir* **34**, 1834–1843 (2018).

2.1. Introduction

Carbon nanotubes have attracted considerable attention due to their outstanding physical, electronic, and optical properties.[1,2] Since these properties depend on it, sorting single-wall carbon nanotubes (SWCNT) by chirality is of crucial importance.[2] DNA/SWCNT hybrids have been used successfully to address this long-standing problem of structure-based separation of complex mixtures of SWCNTs. Recently, it has been shown that SWCNTs coated with single-stranded DNA can be effectively separated into various chiralities using an aqueous two-phase (ATP) system.[3] Partitioning in the ATP system is driven by small differences in the dissolution characteristics of the hybrid between the two phases.[4] In this way, the ATP system is not only a separation technique, but potentially also a way to quantify and rank the dissolution properties, such as the solvation free energy or solubility.

Solute distribution in a two-phase system depends on the solute's relative free energy of solvation in the two phases, which in part depends on the exact structure of the solute surface exposed to the phases. The single-wall carbon nanotube (SWCNT) – single stranded DNA hybrid is essentially an amphiphilic system, with the DNA backbone being the hydrophilic area and the DNA bases as well as the SWCNT surface being hydrophobic regions.[5] Various simulation studies have suggested that the DNA bases adsorb onto the SWCNT surface and the backbone is away from the SWCNT surface and solvated by the surrounding aqueous phase.[6–8] This suggests that the net solvation energy of the hybrid surface depends sensitively on the details of the DNA structure on the SWCNT surface. A polymer aqueous two-phase (ATP) system consists of two separate but permeable water

phases which vary slightly in their physical properties due to the difference in the polymer composition and concentration in the two phases.[8–10] The ATP system has been widely used for separation of biomolecules as the phases do not denature the biomolecules, the interfacial stress is much lower than in case of water – organic solvent system, and the difference in solvation energy in these systems is small, which is ideal for separating solutes with small structural differences.[4,11]

Recently, Khripin et al. and Fagan et al. showed that surfactant coated SWCNTs could be separated very effectively into various chiralities using a polymer aqueous two-phase system.[12,13] Further work by Ao et al.[3] has shown that partitioning of DNA−SWCNT hybrids in a given polymer two-phase system is strongly sequence-dependent and can be further modulated by salt and polymer additives. SWCNT partitioning in the ATP system is determined by the difference in dissolution properties of DNA−SWCNT between the two phases. Hence it is proposed that the DNA−SWCNT partitioning in the ATP is because of sensitive dependence of the dissolution energy on the spatial distribution the DNA on the SWCNT hybrid.

Here we present two different approaches for analyzing ATP partitioning of DNA/SWCNT hybrids quantitatively. In both approaches, the system we consider comprises the following elements: (a) the solvent, water, (b) two water-soluble solutes (Dextran/DX and Polyethylene Glycol/PEG) that form the basic two-phase system (DX and PEG-rich, respectively), (c) a modulant molecule (PVP) that adjusts differences in solvation free energy or solubility so as to drive the final component, (d) solute (DNA-SWCNT) from one phase to the other. The primary measurement for a given two-phase

system (fixed DX/PEG concentration) is the partitioning of the solute between the two phases, quantitatively the partition coefficient *K*, as a function of modulant concentration.

In the first approach, we propose that the solvation free energy of the solute in a particular aqueous phase is related in a certain way to its solvation free energy in pure water. We develop a method for quantitative interpretation of the partition coefficient in terms of solvation free energy of the solute being studied, and factors that affect how the modulant changes it. A discrete Taylor series is then adopted to interpret the developed model.

In the second approach, we investigate the use of solubility parameters to analyze the partitioning of the hybrid in an ATP system. The solubility parameters, introduced by Hildebrand[14], are defined as the square root of cohesive energy E_{coh} divided by the molar volume *V*.

$$\delta \equiv \left(\frac{E_{coh}}{V}\right)^{1/2} \tag{2.1}$$

The basic idea is that solutes will dissolve in solvents with solubility parameters not too different from their own, a kind of "like dissolves like".[15] The cohesive energy is the energy necessary to separate the atoms or molecules from each other, thus it is the energy required to break all interactions during vaporization. Furthermore, it has been found that the solubility parameter is strongly related to the surface free energy.[16,17] In previous research, it has been proposed that the recognition sequence pairs (DNA/SWCNT combinations that enable separation) have some special secondary structure that accounts for their difference in solvation properties.[3] Thus differences among hybrids can likely be interpreted as being due to differences in their surfaces, as defined by the arrangement of DNA on the SWCNT surface.

Formally, the Hildebrand solubility parameter considers only dispersion interactions between molecules.[14] For many polymer/solvent pairs, the cohesive energy is also affected by polar group interactions and hydrogen bonding, which led to the development of the Hansen solubility parameters.[15] According to Hansen theory, the cohesive energy can be considered as a sum of the contributions by dispersion forces E_d, polar group effects E_p, and hydrogen bonding E_h:

$$E_{coh} = E_d + E_p + E_h \tag{2.2}$$

and the corresponding solubility parameters is

$$\delta^2 = \delta_d^2 + \delta_p^2 + \delta_h^2 \tag{2.3}$$

To our knowledge, there have been no studies to quantify DNA/CNT hybrid solubilities although the solubility parameters for DNA and SWCNT themselves have been reported.[17–19] Here, we estimate solubility parameters for different species of the hybrids by relating the partition coefficient with Hildebrand solubility parameters.

2.2. Methods*

HiPco SWCNT was obtained from NanoIntegris and single-stranded DNA was purchased from Integrated DNA Technologies (IDT). Other chemicals were procured from Sigma-Aldrich.

* Some of the experimental work described in this section was performed by Dr. Akshaya Shankar of Lehigh University.

For dispersion of SWCNTs with a given DNA sequence, a total volume of 1 mL of the DNA and SWCNT mixture in phosphate buffer was sonicated in an ice bath for 90 minutes at a power level of 8 W. The SWCNT/DNA mass ratio was 1:1.5. After centrifugation at 16100g for 90 min, the supernatant of the DNA-SWCNT dispersion was collected. The dispersion was then passed through an Amicon 100 kDa filter and resuspended in DI water three times in order to remove free DNA and the phosphate salts. The concentration of the dispersion was adjusted such that at 20 times dilution, the absorbance at 990 nm was ~0.5.

Aqueous two-phase systems consisting of Dextran (70 kDa) and PEG (6 kDa) were prepared in DI water with different compositions: 10% (w/w) PEG/10% (w/w) Dextran and 14 % (w/w) PEG/14 % (w/w) Dextran. The total volume of the aqueous two-phase system including the DNA-SWCNT dispersion and PVP solution was fixed at 500 uL. The volume of DNA-SWCNT dispersion added was fixed at 25 uL. Partition of SWCNTs in the aqueous two-phase was obtained by vortex mixing of the PEG solution, Dextran solution, PVP solution and DNA-SWCNT dispersion in a microcentrifuge tube for 1 min followed by centrifugation at 16100g for 2 min. Figure 2.1 shows qualitatively the use of PVP to adjust the phase in which the DNA/SWCNT hybrids reside. With increasing PVP concentration the hybrids move from the bottom to the top phase.

Absorbance measurements were performed on a Varian Cary 50 spectrophotometer over the wavelength range of 200−1100 nm using a 10 mm path length quartz microcuvette to determine the concentration of DNA/SWCNT hybrid in each phase. Fractions of the top and bottom phases were collected using a pipette and diluted for absorbance measurements.

Blank top or bottom phases of ATP systems without SWCNTs were collected and diluted in the same way as the corresponding SWCNT fractions for baseline measurements.

For quantitative analysis we decompose the absorbance spectra using methods developed for analysis of surfactant exchange experiments.[20] Figure 2.2 shows a typical decomposition of the absorbance spectra into contributions from (9,1), (8,3), and (6,5) SWCNTs. By using this decomposition, we can separately track the change in absorbance intensity of these three SWCNT chiralities. We chose to focus on these three SWCNTs because of our UV-VIS-NIR instrument limitations and because of previous studies on surfactant exchange kinetics on the same three SWCNTs.[20]

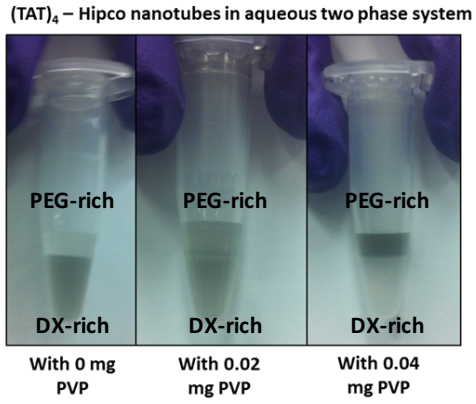

Figure 2.1. The effect of PVP in aqueous two-phase system of 10% (w/w) PEG/10% (w/w) Dextran. With the addition of PVP, the DNA-CNT hybrid moves from being mostly in the bottom phase to mostly in the top phase.

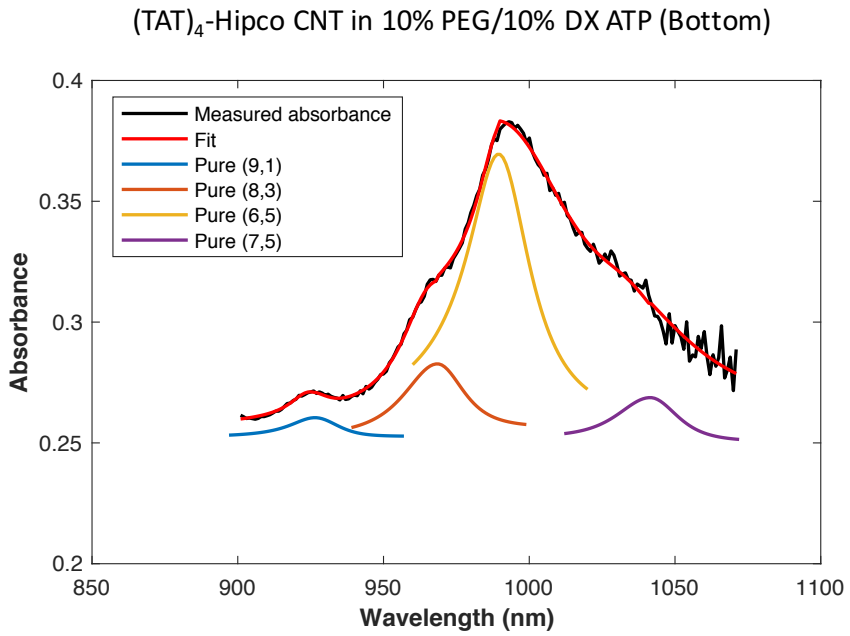

Figure 2.2. Fitted spectra showing contribution of different SWCNTs to the measured spectrum

The primary measurement, concentration of the solute in each of the two phases, is used to obtain the partition coefficient for the solute, which is defined as the ratio of its concentrations in the two phases:

$$K = \frac{C_b}{C_t}. \tag{2.4}$$

We found that an average of about 3% of the DNA/SWCNT hybrids are stuck at the interface. To minimize the error due to the interface trapping, we use only concentration in the bottom layer "*b*" and estimate the one in the top layer "*t*" by mass balance.

We assume equilibrium, that is, the chemical potential of the solute is the same in the two phases ($\mu_t = \mu_b$):

$$\mu_t = \mu_t^o + k_B T \ln a_t = \mu_b = \mu_b^o + k_B T \ln a_b \tag{2.5}$$

where μ^o is the chemical potential of the standard state; a is the solute activity; the subscript t and b refer to top and bottom phases, respectively; k_B is Boltzmann's constant; T is temperature. In the infinitely diluted condition the activity coefficient approaches unity, and then equation (2.5) can be written as

$$\mu_t^o + k_B T \ln C_t = \mu_b^o + k_B T \ln C_b \tag{2.6}$$

Combining the equations (2.4) and (2.6), the partition coefficient can be expressed in terms of the difference in standard chemical potential of the solute in the bottom and top phases,

$$K = \exp\left(-\frac{\mu_b^o - \mu_t^o}{k_B T}\right) \tag{2.7}$$

In this two-phase system, absent the modulant (PVP), most of the DNA/SWCNT hybrids partition to the bottom phase. As PVP is added to the system, the DNA/SWCNT hybrids gradually transfer to the top phase. Figure 2.3 shows the partition coefficients as a function of PVP concentration for various sequences paired with the (6,5) SWCNT. (The fit to the data will be described later.) It can be seen that there is considerable DNA sequence-dependent difference in the amount of PVP required to move the DNA/SWCNT hybrids from the bottom phase to the top phase. Thus (a) the difference in chemical potential is a function of both PVP concentration and the hybrid identity (specified by the DNA sequence and SWCNT chirality), and hence (b) the concentration of PVP can be used to probe quantitatively this difference.

The choice of DNA sequences was governed by the following considerations. We have previously studied the (TAT) family and have shown in surfactant exchange experiments that $(TAT)_4$/(6,5) has properties quite distinct from its compositional cousins $(TAT)_3T$ and $(TAT)_3TA$.[8,20] Similarly $(CCA)_{10}$ is a recognition partner for (9,1). The

remaining sequences represent a sample of repeat two-mer DNA sequences, all of the same length (30-mers).

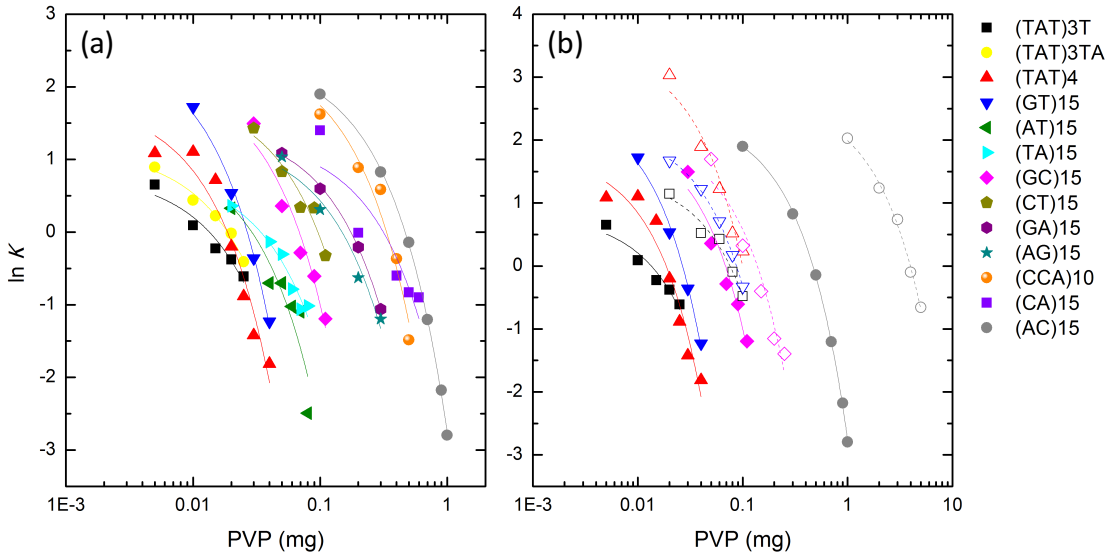

Figure 2.3. (a) Partition coefficient as a function of PVP concentration for a number of sequences paired with the (6,5) SWCNT in a 10% DX/10% PEG ATP system. (b) Partition coefficient for a reduced set of sequences with the (6,5) SWCNT in a 10% DX/10% PEG ATP (solid) and in 14% DX/14% PEG ATP system (open). The data have been fitted using a two-parameter function described in the text, equation (2.13).

2.3. Evaluation of Relative Solvation Free Energy of Hybrids

We first explore the idea that the solvation free energy of a DNA/SWCNT hybrid in either of the two ATP phases can be regarded as related to its solvation free energy in pure water, modulated both by the water-soluble polymers and by the modulant (PVP in this case). We wish to create a model that allows us to extract quantitative relative values of a property of the hybrid itself, absent the polymers and modulant. The use of such a procedure is to produce a property of the hybrid that, we propose, can be interpreted as the solvation free energy in pure water.

For a given DX/PEG system, the experimental data in Figure 2.3 show that the chemical potential μ^o depends on two factors: the concentration of the modulant (PVP) and the solute composition (DNA/SWCNT hybrid). Assume that these effects are independent and additive. Within either one of the two phases ("t" for "top phase", "b" for "bottom phase"), let the chemical potential be given by the following. (We considered several different approaches to quantification of the data based on different assumption that capture the modulation due to PVP. We present the one that is most consistent with the experimental data.)

$$\mu^o_{t/b} = \alpha_{t/b}([PVP], \xi) l \mu^o_w(n, m, d) + l\beta_{t/b}(\xi, n\ , m, d) \tag{2.8}$$

where $\mu^o_w(n, m, d)$ is the solvation free energy of DNA/SWCNT hybrid in water per unit length of the SWCNT, l is SWCNT length, $\alpha_{t/b}$ is a parameter through which the solvation free energy is modulated, and $\beta_{t/b}$ is a parameter related to adhesion between hybrid and solution that depends on the polymer solution and the composition of the ssDNA/SWCNT. Note that the chemical potential depends linearly on the length of the DNA/SWCNT. We are assuming that its distribution can be represented by its mean value. (See Appendix for a detailed discussion and justification of this assumption). Any given DNA/SWCNT hybrid can be identified by the CNT chiral indices (*n,m*) and the DNA sequence '*d*'. The polymer composition is represented by ξ. Substituting equation (2.8) into equation (2.7) we get

$$K = \exp\left(-\frac{\Delta\alpha l \mu^o_w + \Delta\beta l}{k_B T}\right) \tag{2.9}$$

or

$$lnK = -\Delta\alpha' l \mu^o_w - \Delta\beta' l \tag{2.10}$$

where $\Delta\alpha'$ *and* $\Delta\beta'$ are $(\alpha_b-\alpha_t)/k_BT$ *and* $(\beta_b-\beta_t)/k_BT$, respectively. Suppose that $\Delta\beta' = 0$, and consider the equipartition case for which $\ln K = 0$. Because the $\Delta\alpha'$ is, by assumption, independent of hybrid type, equation (2.10) would then predict that regardless of DNA/SWCNT composition, all curves on a plot of ln*K vs* [modulant] would pass through the same point when $\ln K = 0$. That is, as shown in Figure 2.4, $\Delta\alpha'$ captures differences in slope of the ln*K vs* [modulant] plot. Now imagine a fixed modulant concentration [PVP], say where $\Delta\alpha'$=0. Then, it is clear that $\Delta\beta'$ represents a vertical shift factor.

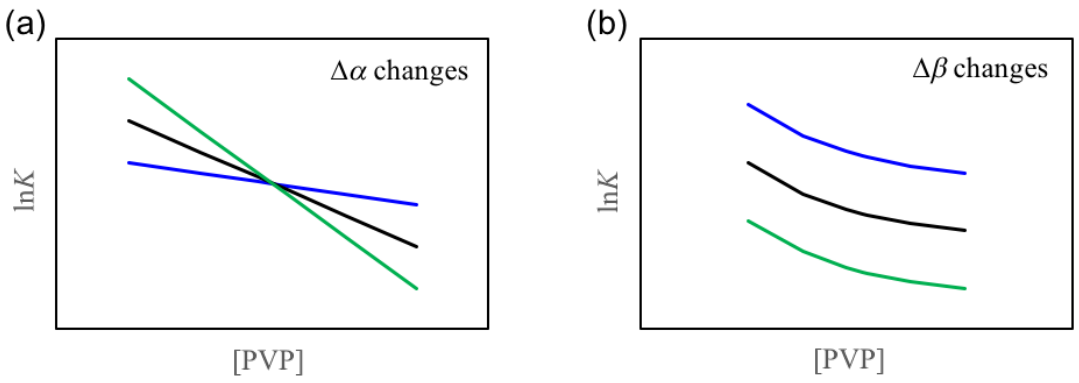

Figure 2.4. A schematic plot of ln*K vs*. PVP concentration: (a) The parameter $\Delta\alpha$ captures changes in slope whereas (b) $\Delta\beta$ represents relative vertical shifts.

For now, consider it as an empirical, minimalist representation of the experimental data using three factors, a quantity with units of energy that we provisionally term the solvation free energy μ_w^o, and two dimensionless parameters, $\Delta\alpha'$ and $\Delta\beta'$, that modulate the first factor. Later we return to discuss their physical significance. Since, by supposition, $\Delta\beta'$ is independent of [PVP], it can be eliminated by taking a derivative of equation (2.10) with respect to concentration [PVP]:

$$\frac{\mathrm{d}\ln K}{\mathrm{d}[PVP]} = -\frac{d\Delta\alpha'}{\mathrm{d}[PVP]} l\mu_w^o. \tag{2.11}$$

For any given ATP system consisting of the same polymer composition, the term $\frac{d\Delta\alpha'}{\mathrm{d}[PVP]}$ in equation (2.11) depends only on [PVP], by supposition. So, it can be eliminated by taking the ratio of the LHS of equation (2.11) for two different DNA/SWCNT concentrations at a fixed [PVP]. Thus, we can obtain the ratio of the solvation free energy in water for two different hybrid compositions *A* and *B* in terms of the ratio of derivatives of experimental data:

$$\frac{\mu_{w,A}^o}{\mu_{w,B}^o} = \frac{\frac{\mathrm{d}\ln K}{\mathrm{d}[PVP]}\Big|_A}{\frac{\mathrm{d}\ln K}{\mathrm{d}[PVP]}\Big|_B} \frac{l_B}{l_A}. \tag{2.12}$$

We assume that different DNA/SWCNT hybrids have the same length distribution so the factor l_B/l_A equals one. The right-hand side of equation (2.12) can be obtained in a model-independent way by derivatives calculated from experimental data at the same PVP concentration. Therefore, it offers a method to quantify the ranking of solvation free energy of various DNA sequence and SWCNT chirality combinations. It is clear that in this formulation the ratio of solvation free energy by itself does not indicate separability, in which the factor $\Delta\beta'$ plays a critical role. Since comparison has to be done at the same [PVP], it is important to have a series of experimental plots in which each one overlaps at least another one, or all overlap some shared standard.

Here we use a different method which is simpler and yields results very similar to those following the procedure just described. We find that the experimental data (Figure 2.3) can be fitted well by:

$$\ln K = C_1[PVP] + C_2 \tag{2.13}$$

where C_1 represents the "strength" of the influence of [PVP] on change of partition coefficient and C_2 is the logarithm of partition coefficient in the absence of PVP. Using eq. (2.13) in (2.12) then gives a single number for the ratio of solvation energy, independent of [PVP]:

$$\frac{\mu^o_{w,A}}{\mu^o_{w,B}} = \frac{C_1|_A}{C_1|_B}\frac{l_B}{l_A}. \tag{2.14}$$

Each calculated ratio was then standardized using a single reference, in this case the sequence $(TAT)_3T$, as shown in Figure 2.5. (The choice of reference hybrid is arbitrary – we picked the one that required the least amount of PVP for equipartition.)

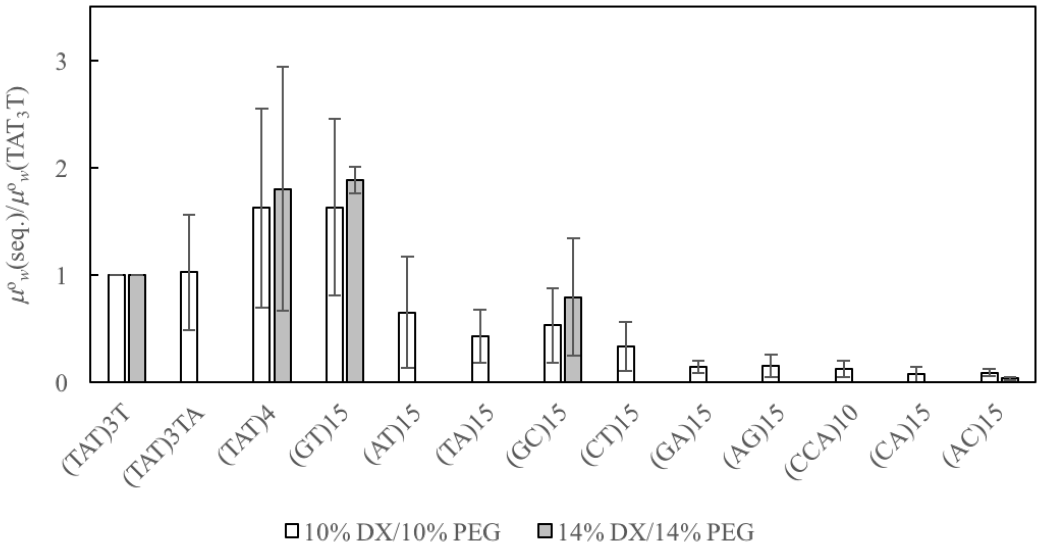

Figure 2.5. The relative solvation free energy for different DNA/(6,5)SWCNT hybrids in ATP systems with two different polymer compositions. Each free energy is standardized to $(TAT)_3T$ paired with the (6,5) SWCNT by equation (2.14). Error bars show 90% confidence intervals of the ratio of the fit parameter C_1 in equation (2.13).

To test if the relative solvation free energy so extracted is independent of polymer combination (as assumed in equation (2.8)), we repeated the experiment for a different polymer composition, 14% DX and 14% PEG (for a reduced set of sequences). Figure 2.3b shows the difference in the partition coefficients compared to the ATP system of 10% DX and 10% PEG. In the new system, all the partition coefficient curves are shifted to the right; more PVP is required to move the hybrids from bottom to top phase. The solvation free energy was then obtained and compared to the results of the 10% DX and 10% PEG system, presented in Figure 2.5. Evidently, the solvation free energy ratio extracted using equation (2.14) is consistent between the two polymer compositions used.

We return now to extract the remaining two parameters, $\Delta\alpha'$ and $\Delta\beta'$. For a given hybrid, *K* is measured as a function of [PVP]. So, the data can be stitched into an underlying master curve for $\Delta\alpha'$ and jumps in $\Delta\beta'$ that depend only on the particular hybrid. The details of the procedure are presented in the Appendix. $\Delta\alpha'$ and $\Delta\beta'$were estimated for different chiralities and are presented in Figure 2.6. As expected, measurements of a number of DNA/SWCNT sequences within the same polymer composition ATP system can be collapsed into a single $\Delta\alpha'([PVP])$ plot. We also found that the polymer composition of ATP system significantly affects $\Delta\alpha'$ and $\Delta\beta'$ unlike the relative solvation free energy (Figure 2.5). This suggests that the basic functional form chosen is appropriate. The same analysis was implemented to obtain the relative solvation free energy and $\Delta\beta'$ for different DNA sequences paired with the (8,3) and (9,1) SWCNT. The results are consistent with those presented in the main text for the (6,5) SWCNT and can be found in the Appendix.

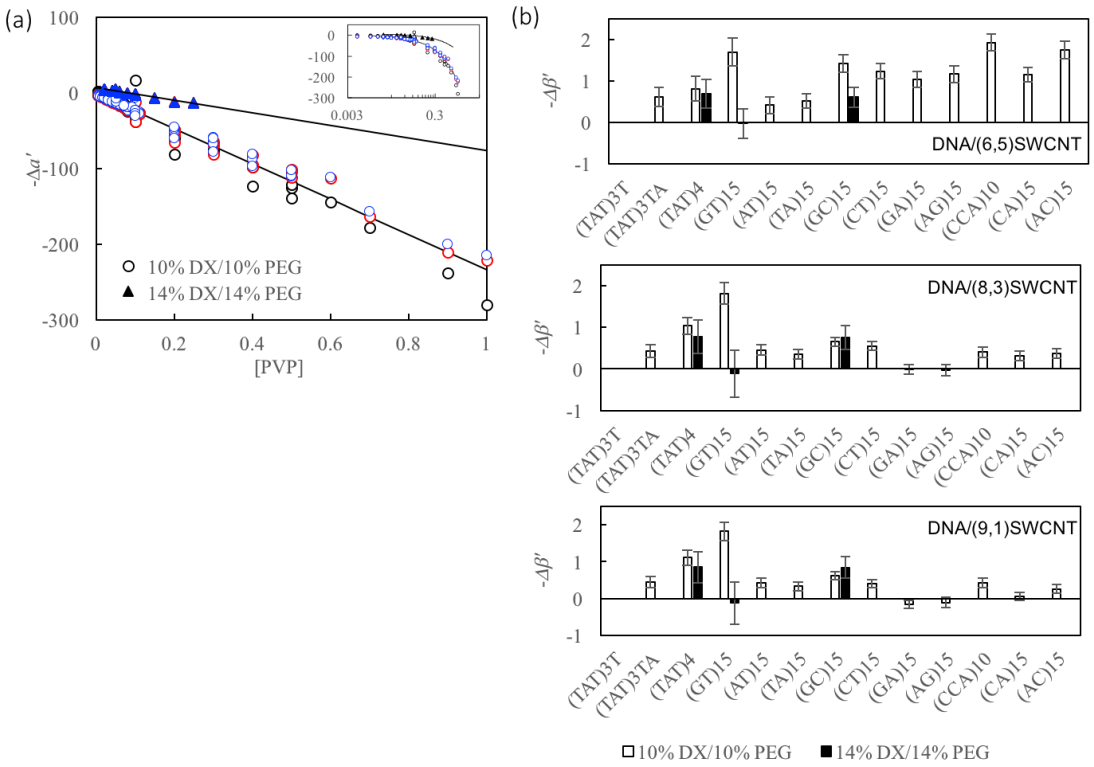

Figure 2.6. (a) Master curve for $\Delta\alpha'$ for a number of sequences paired with (6,5) (black), (8,3) (red) and (9,1) (blue) SWCNT in 10% DX/10% PEG ATP system (○) and with (6,5) SWCNT in 14% DX/14% PEG ATP system (▲). The data have been fitted by a linear equation (solid line). (b) $\Delta\beta'$ for a number of sequences paired with the (6,5) (8,3) or (9,1) SWCNT in different polymer composition ATP systems. Sequences $(TAT)_3T$, $(TAT)_4$, $(GT)_{15}$, and $(GC)_{15}$ paired with (6,5) SWCNT were examined in 14% DX/14% PEG ATP system. The error bars on $\Delta\beta'$ show 90% confidence intervals for the parameter $\Delta\beta'$ in equation (2.27).

The basic forms assumed for the difference in chemical potential (eqs. 2.8 & 2.10) can be viewed as an expansion relative to solvation free energy in pure water. The chemical potential differences (or the partition coefficients) are continuous functions of [PVP] but are also functions of discrete variables: the sequences and chiralities (n,m,d). Thus, we propose an interpretation of eqs. 2.8 & 2.10 as a continuous/discrete Taylor series expansion:

$$\Delta\mu([PVP],\xi,n,m,d) = \Delta\mu^{o} + \frac{d\Delta\mu([PVP],\xi,n,m,d)}{d[PVP]}[PVP] + \omega(n,m,d) \qquad (2.15)$$

where $\Delta\mu^{o}$ is the chemical potential difference at the reference state for a chosen particular sequence of DNA on SWCNT at [PVP] = 0; $\omega(n,m,d)$ is a parameter describing the discrete difference in the chemical potential in terms of the hybrid species. Let $\Delta\mu^{*}(\xi,n,m,d)$ be the value of $\Delta\mu$ at [PVP] = 0. Then,

$$\Delta\mu([PVP],\xi,n,m,d) = \frac{d\Delta\mu([PVP],\xi,n,m,d)}{d[PVP]}[PVP] + \Delta\mu^{*}(\xi,n,m,d) \qquad (2.16)$$

According to the experimental observation shown in Figure 2.3 and Figure 2.6a, both $\Delta\mu$ and $\Delta\alpha'$ are found to be linear in [PVP]; specifically, let $\Delta\alpha' = c_1 + c_2[PVP]$ (see Appendix). Substituting this into equation (2.10) and comparing with equation (2.16) establishes the correspondence that

$$\frac{d\Delta\mu([PVP],\xi,n,m,d)}{d[PVP]} = -c_2 l\mu_w^{o}k_B T \qquad (2.16a)$$

and

$$\Delta\mu^{*}(\xi,n,m,d) = -(\Delta\beta' + c_1\mu_w^{o})lk_B T \qquad (2.16b)$$

Thus, the term c_2 is directly related to how the chemical potential difference changes with [PVP]. Moreover, because in our system the intercept c_1 is small, equation (2.16b) suggests a meaning for $\Delta\beta'$; it essentially corresponds to $\Delta\mu^{*}$.

We see, for example, that $(GT)_{15}$ and $(CCA)_{10}$ behave very differently even though the lengths are same (30mers). This clearly confirms that the solvation free energy difference depends strongly on the DNA sequence. It is interesting to compare the solvation free energy with binding affinity of DNA sequences on SWCNT. Our group previously reported the activation energy for removal of several DNA sequences from different

species of SWCNT by a surfactant molecule. We found that the rate at which a surfactant removes $(CCA)_{10}$ is about ten times slower than removal of $(TAT)_4$ (from (6,5)).[20] This appears to correlate with the fact that it takes considerably larger amounts of PVP to move the $(CCA)_{10}$/(6,5) hybrid to the top phase. It is also reported that $(GT)_{15}$/(6,5) is removed about 7 times slower than $(TAT)_4$/(6,5).[8] However, our results from ATP partitioning show that $(GT)_{15}$/(6,5) differs only slightly from $(TAT)_4$/(6,5). This suggests that the order of hydrophilicity may not be directly correlated with the order of binding activation energy. For the surfactant exchange experiment, we previously proposed that a defect in the DNA coating admits adsorption of the surfactant, which is the activated state, following which the surfactant can replace DNA from the entire SWCNT.[8] The defects can be created due to local disorder in the DNA strand arrangement, or be thermally activated. Thus, defects in DNA coverage on SWCNT is a very important factor to determine the activation energy. On the other hand, the solvation free energy is related to the hydration interaction which is dependent more on the average spatial distribution of hydrophilic groups rather than the rare few defects.[3]

2.4. Evaluation of Solubility Parameters

We now consider an alternative analysis of the ATP experiment. Ao et al.[3] suggested that the difference in partitioning arises from its surface functionalities, i.e., the surface free energy of the hybrids. It is well known that the surface free energy is directly related to the intermolecular forces in liquid, and Hildebrand and Scott[21] reported that this

force can be measured in terms of the heat of the vaporization, or its square root, the solubility parameter δ.

The expression for the partition coefficient in terms of solubility parameters was derived by Srebrenik and Cohen[22]:

$$\ln K = \frac{V_i}{RT}[(\delta_a - \delta_i)^2 - (\delta_b - \delta_i)^2] + \ln\frac{V_a}{V_b} \tag{2.17}$$

where the subscripts a, b and i denote the solvents a, b and solute i, and V is the molar volume. In this study, a and b are top and bottom phases, respectively, and i is a species of hybrid. There are several assumptions involved in the derivation of equation (2.17). The solutions are assumed to be regular so that the total volume remaining unchanged. The geometric mean approximation is used to estimate the cohesive energy density between dissimilar molecules (i.e., a and i or b and i), i.e., it can be given by the geometric mean of the homogeneous cohesive energy density of two molecules: $\delta_{ai}^2 = \delta_{aa}\delta_{ii}$ and$\delta_{bi}^2 = \delta_{bb}\delta_{ii}$. This assumption is based on an analogy with the semi-quantitative relation for intermolecular forces called the *combining relation*.[23]

In solubility parameter theory, the solubility parameter for mixtures, $\bar{\delta}$, depends directly on the relative amount present, and it can thus be related to the volume fraction ϕ_j and solubility parameters δ_j of the components by the expression

$$\bar{\delta} = \sum \phi_j \delta_j \tag{2.18}$$

where the summation extends over all components, i.e., PEG, Dextran, water, and PVP in our system. Although the use of such an expression for mixtures is not always

quantitatively accurate,[24] it has been widely used successfully. For any given ATP system, the amounts of PEG, Dextran and water are constant. Thus, within the same polymer composition, the equation (2.18) can be rewritten as a function of PVP concentration:

$$\bar{\delta}_{t/b} = A_{t/b} + \phi_{PVP}\delta_{PVP} \tag{2.19}$$

where $A_{t/b} = \phi_{t/b,PEG}\delta_{PEG} + \phi_{t/b,DX}\delta_{DX} + \phi_{t/b,w}\delta_w$ and the subscript t, b for top and bottom phase, respectively. Assume that the concentration of PVP is equal in both phases so that $\phi_{PVP} = \phi_{t,PVP} = \phi_{b,PVP}$. Substituting equation (2.19) into (2.17), the partition coefficient becomes

$$\ln K = \frac{V_i}{RT}\left[A_t^2 - A_b^2 + 2(A_t - A_b)\left(\phi_{PVP}\delta_{PVP} - \delta_i([PVP])\right)\right] + \ln\frac{V_t}{V_b} \tag{2.20}$$

Here, by incorporating the PVP term in equation (2.19), we have assumed that PVP is in sufficient excess such that it affects the solubility parameter of the solution. It can also affect the solubility parameter of the hybrid in equation (2.20), say by adsorbing to its surface. For small PVP conditions, the PVP term in equation (2.19) can be ignored. Then, the partition coefficient can be rewritten as

$$\ln K = \frac{V_i}{RT}\left[A_t^2 - A_b^2 - 2(A_t - A_b)\delta_i([PVP])\right] + \ln\frac{V_t}{V_b} \tag{2.20a}$$

Let us use equation (2.20a) to see, approximately, how the solubility parameter of the hybrid compares to that of the two aqueous phases. Neglect the logarithm term $\ln\frac{V_t}{V_b}$ since $V_t \approx V_b$. By rearranging the remaining equation for $\delta_i([PVP])$, equation (2.20a) becomes $\delta_i([PVP]) = -\frac{RT\ln K}{V_i(A_t - A_b)} + \frac{A_t + A_b}{2}$. Here, $\bar{\delta}_{t/b} = A_{t/b}$ and the second term is numerically

dominant. Thus, approximately, the solubility parameter of hybrid δ_i can be estimated to be an average value of the solubility parameters of the top and bottom phases: $\delta_i([PVP]) \approx \frac{\overline{\delta}_t+\overline{\delta}_b}{2}$, that is, its value lies between those of the two phases. Because the solubility parameters of the top and bottom phases themselves differ only slightly, the ATP system is able to discriminate between solutes with small difference in solubility.

To extract the solubility parameter of the hybrid, δ_i, the remaining parameters need to be determined. First, the solubility parameter and molar volume for each component were obtained from the literature and the values are listed in Table 2.1.[25–29] The volume fractions of PEG and Dextran are estimated from the phase diagram of PEG and Dextran system.[30] Next, molar volume of the hybrid is given by

$$V_i = \pi r_{hybrid}^2 L_{cnt} N_{Avo} \tag{2.21}$$

where the length of SWCNT, L_{cnt}, is set to be 200 nm,[31] N_{Avo} is Avogadro number, and the radius of hybrid, r_{hybrid}, is determined as $r_{hybrid} = r_{cnt} + d_{cb}$; the radius of SWCNT, r_{cnt}, is calculated from its (n, m) indices in Å:[32]

$$r_{cnt} = 0.783\sqrt{(n+m)^2 - nm} \tag{2.22}$$

and the distance from SWCNT to the backbone, d_{cb}, is determined to be 0.592 nm.[7] We recognize that the molar volume could be sequence-dependent even for the same species of the SWCNT because DNA structure on its surface may be different, but neglect these differences.

Finally, the solubility parameters of each hybrid are calculated using equation (2.20) and shown in Figure 2.8. It is not surprising that the values of the solubility parameters of all hybrids are close to those of water because the hybrid is well dispersed in water. The solubility parameter is an intrinsic property; we presume that it changes as a function of PVP concentration because PVP absorbs on the hybrid surface. As shown in Figure 2.8, the solubility parameter changes linearly with [PVP], which is consistent with concentration-dependent adsorption of PVP on the hybrid.

Fitting the dependence of solubility parameter by a linear function provides two quantities: the slope and intercept. The intercept (Figure 2.7) is significant because by excluding the effect of PVP it reflects an intrinsic property of the hybrid. However, for the compositions studied here, we find that the differences between intercept values (Figure 2.7) are not statistically significant. As shown in Figure 2.8, the solubility parameters decrease slightly as PVP concentration increases, with different slopes characteristic of each hybrid. The larger the slope of the solubility parameter variation with respect to [PVP], the lower the value of [PVP] required to move that hybrid from bottom to top phase. Therefore, the difference in the effectiveness of PVP can be seen by plotting the slopes (Figure 2.9). It is not surprising that the slope is smaller at high polymer concentration conditions, which can be interpreted in terms of entropy or adsorption theory. It also shows that the solubility of $(TAT)_4$/(6,5) is most easily modulated by the PVP concentration. Among the sequences of DNA with the same length, $(GT)_{15}$/(6,5) has significantly smaller values than others.

Table 2.1. Hildebrand (δ) and Hansen ($\delta_d, \delta_p, \delta_h$) solubility parameters and molar volume (V) of water and polymers used in this study. The solubility parameters of SWCNT are presented as reference. Units are $MPa^{0.5}$ for solubility parameter and m^3/mol for molar volume.

	δ [a]	δ_d [b]	δ_p [b]	δ_h [b]	$V \times 10^3$ [c]
PEG	19.71	17.56	3.22	8.31	4.992
DX	38.6	24.3	19.9	22.5	43.82
water	47.8	15.5	16.0	42.3	0.018
PVP	24.3	18.8	13.4	7.5	8.330
SWCNT	20.8	18.0	7.8	6.9	52.76 – 56.24

[a] Hildebrand solubility parameters were calculated by geometric mean of the corresponding Hansen solubility parameters. [b]Hansen solubility parameters were taken from the literature.[17,25–27] [c]Molar volumes of polymers and water were calculated from corresponding specific volume, which were taken from the literature.[28,29] Molar volume of SWCNT can be calculated by equation (2.21) using the radius of SWCNT (r_{cnt}), given by equation (2.22), instead of that of hybrid (r_{hybrid}).

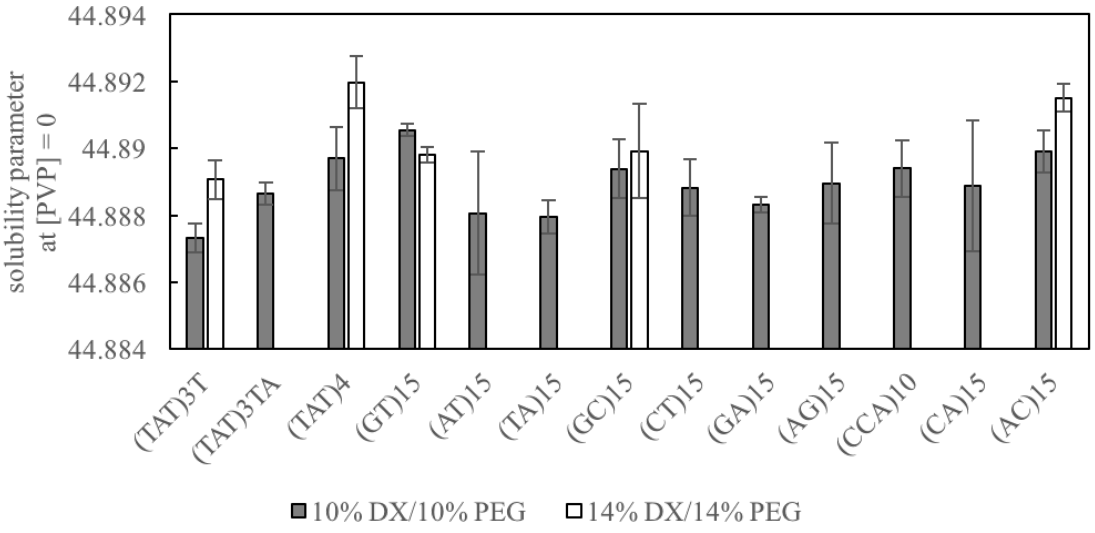

Figure 2.7. Hildebrand solubility parameters at [PVP] = 0 for a number of sequences paired with the (6,5) SWCNT in the DX/PEG ATP system. Error bars show 90% confidence intervals for the intercept given by a two parameter linear fit.

Previously in Figure 2.3b, we found that more PVP is required to change the partition at higher polymer concentrations. In terms of solubility parameters, hybrids move from bottom to top phase as the solubility parameter of hybrid approaches that of top phase due to the PVP absorption on the hybrid. Therefore, as shown in Figure 2.10, if the solubility parameter of hybrid is independent of the polymer combination, the difference in the solubility parameter between the hybrid and the top phase is greater at high polymer concentration conditions. As expected, the solubility parameters at both the bottom and top phases at high polymer concentration are smaller than at lower polymer concentration, for example the values of the solubility parameter in the $(AC)_{15}$ hybrid system are $\delta_t = 43.95, \delta_b = 45.80$ in 10% PEG/10% DX system and $\delta_t = 43.90, \delta_b = 45.74$ in 14% PEG/14% DX system, respectively.

To better understand the solubility parameter of the hybrids, we compared the values of the hybrid with those of free DNA and the SWCNT (Table 2.1). To the best of our knowledge, no studies have been reported the solubility parameter of the ssDNA, but dsDNA is $\delta = 29.7$ MPa$^{0.5}$ ($\delta_d = 19.0$, $\delta_p = 20.0$, and $\delta_h = 11.0$)[15], much less than that of the hybrid as we report here. Since dsDNA is stabilized by forming a helical structure in which the hydrophobic bases hydrogen-bond and stack with each other, we might expect that the solubility characteristics of ssDNA to be very different from that of dsDNA. Therefore, we have investigated the solubility parameter of ssDNA using same method presented here. Poly(T) of four different length (10, 30 and 60 and 90-mers), poly(A) of two different length (30 and 50-mers), and poly(C) (30-mers) in 10% PEG/10% DX ATP system were examined, shown in the Appendix. The solubility parameter of ssDNA was

determined to be 44.91 $MPa^{0.5}$ by an average of the intercepts. It is noteworthy that this value is quite close to those of the hybrids. This is consistent with our structural models for the hybrid in which the SWCNT is well-covered by ssDNA with its bases and the sugar-phosphate backbone both exposed to the solvent. Small differences in (rare) wrapping structure defects are captured by the differences in hybrid solubility parameters. The high sensitivity of the ATP technique to these small differences makes it well-suited as a method to quantitatively discriminate between different hybrid species, something that it is currently not possible to predict.

Note that we have considered the Hildebrand parameter as a single parameter that represents solubility characteristics, not dispersion force parameters. The concept of the Hildebrand parameter has been used not only to interpret the partitioning of hybrid in the ATP system, but also to successfully extract the solubility parameter as an intrinsic property of the hybrid. Multiple-component concepts, such as Hansen solubility parameter, can be used to extract partial parameters from different contributions, which can help to better understand the intermolecular interactions of hybrids. Further work can be implemented by applying Hansen solubility parameter. This requires three solubility parameters, so more experimental work is required to calculate them uniquely. We suggest this for future work; see Appendix for a discussion of how to extract Hansen solubility parameters from ATP experiments.

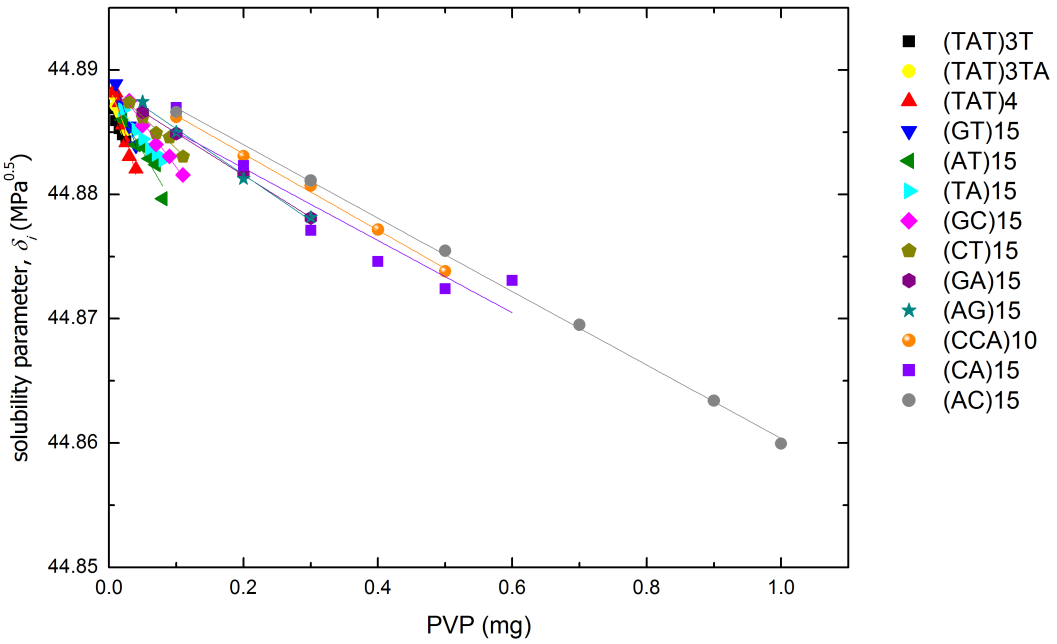

Figure 2.8. Hildebrand solubility parameters as a function of PVP concentration for a number of sequences paired with (6,5) SWCNT in 10% DX/10% PEG ATP system.

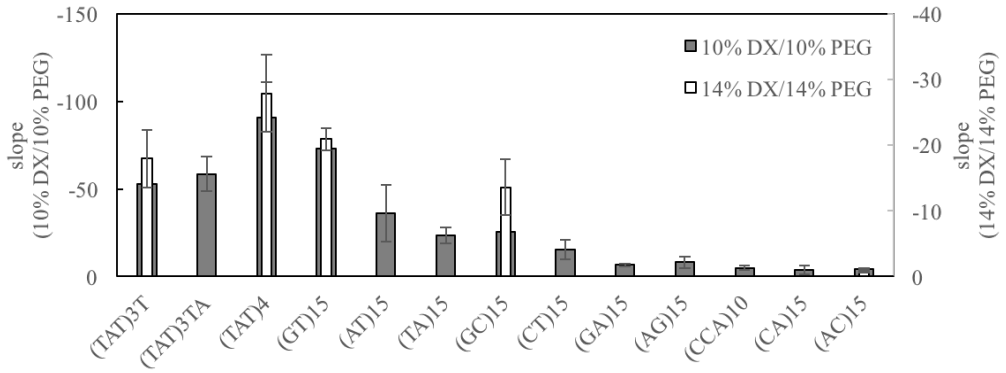

Figure 2.9. Slopes of Hildebrand solubility parameters for a number of sequences paired with (6,5) SWCNT in DX/PEG ATP system. Error bars show 90% confidence intervals for the slope.

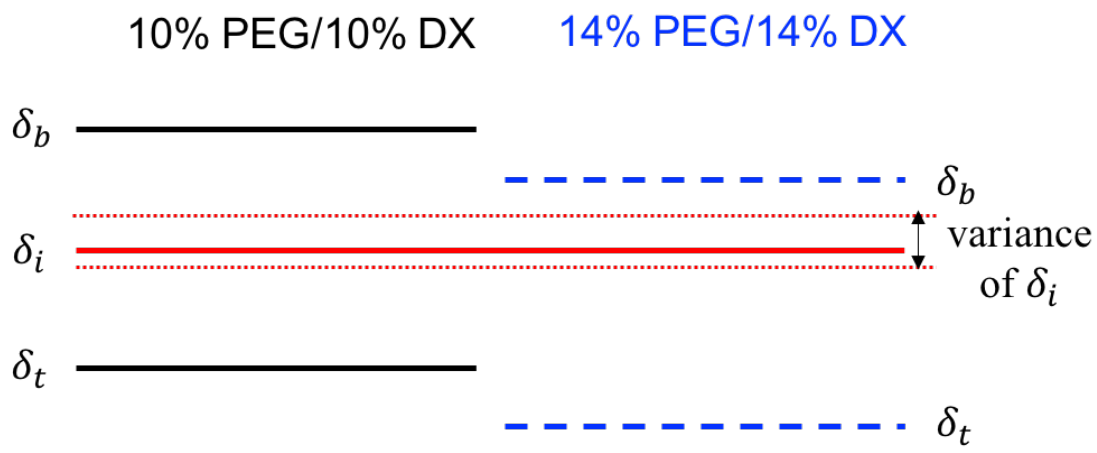

Figure 2.10. Schematic diagram showing the Hildebrand solubility parameters of a DNA/SWCNT hybrid relative to the top and bottom phases for two different compositions. The diagram illustrates why it takes a greater amount of PVP to shift hybrids from the bottom to the top phase as the concentration of PEG and DX is increased.

2.5. Conclusions

In this study, we have measured the partition coefficient as a function of PVP concentration for a number of sequences paired with three different species of SWCNT in ATP system. To analyze the partitioning of DNA/SWCNT hybrids, we proposed two different approaches which relate the measurements of partition coefficient to the solvation free energy or solubility.

First, based on an expansion of free energy difference from a reference state and composition, the relative solvation free energies have been extracted for a number of DNA/SWCNT hybrid combinations. The solvation free energies obtained from two different solvent compositions shows consistent values. Furthermore, the extracted values of $\Delta\alpha'$, representing the effect of modulant, were obtained as master curve from all hybrid combination within the same polymer composition, as expected. Thus, we suggest that the model can reasonably quantify our experimental observation.

Next, a method has been proposed for estimating the Hildebrand solubility parameters of the hybrids. The value of solubility parameter at [PVP] = 0, which could be interpreted as an in intrinsic property of each hybrid, do not exhibit statistically significant differences. The solubility parameter decreases with increasing [PVP], presumably because of its adsorption on the DNA/SWCNT. This shows that PVP can modulate the solubility of the hybrid by modifying its surface. The sensitivity to [PVP], characterized by the rate of reduction of the Hildebrand parameter with [PVP], varies with hybrid composition, indicating that the interaction between PVP and the hybrids is sequence-dependent. To compare this difference in the interaction, we compare the slope of solubility

parameter. This comparison clearly shows that the sequence $(TAT)_4$, known as the recognition sequence for (6,5) SWCNT, and $(GT)_{15}$, known for its strong affinity to SWCNT, have stronger sensitivity to [PVP] than other hybrids. These results suggest that the measurement of the slope and intercept of the solubility parameter for various hybrid combination can provide quantitative insight into aspects of DNA structure on the nanotube underlying the sequence-specific interaction. Note that the particular ATP system we have studied here (DX/PEG) is not ideal for successful separation of SWCNT species. Generally, the partition coefficient can be determined accurately only when it is on the order of 1; otherwise, experimental errors become large. Ideally, the partition coefficient should be approximately unity at [PVP] ~ 0; this would allow more accurate extraction of the intrinsic solubility parameter (absent the modulant) as the intercept of the [PVP] vs. δ_i plot. Although our system (DX/PEG) is not ideal from this point of view, it serves well the purpose of extracting and comparing quantitatively the solubility characteristics of DNA/SWCNT hybrids.

To compare the two approaches offered in this work, the values of the relative solvation free energies and the slope of the solubility parameters has been transformed to compare them in the same manner. The slope of the solubility parameter (Figure 2.9) is normalized to $(TAT)_4$/(6,5) SWCNT by $\widehat{s_i} = \frac{s_i - s_{(CA)15}}{s_{(TAT)4} - s_{(CA)15}}$. Here, the sequences of $(TAT)_4$ and $(CA)_{15}$ were chosen because they have minimum and maximum values in range of all the sequences paired with the (6,5) SWCNT as shown in Figure 2.9. Then the values of $\widehat{s_i}$ were compared with the relative solvation free energy which, here, is standardized to $(TAT)_4$ paired with the (6,5) SWCNT instead of $(TAT)_3T$. The results are presented in

Figure 2.11. It is interesting that the transformed values from two different approaches are in good agreement with each other.

Both approaches developed in this study have some limitations. While both provide quantitative data, neither contains direct information about underlying structural differences. Also, the solvation free energy method only provides values relative to a reference hybrid. (To the best of our knowledge, no study has yet reported absolute values of the solvation free energy of DNA/SWCNT hybrids.) For the solubility parameter approach, the (DX/PEG) system turned out to have some limitations because of small measured difference in the values of ln$K$ (or the corresponding δ_i) when [PVP] = 0. This limitation can likely be alleviated by choosing other ATP systems. We leave this as future work.

In summary, the ATP system can be used for not only separation technique but also a method by which to quantify and rank the dissolution properties such as the solvation free energy or the solubility. We expect that such quantification can provide a basis for data-analytic searches for new sequences.

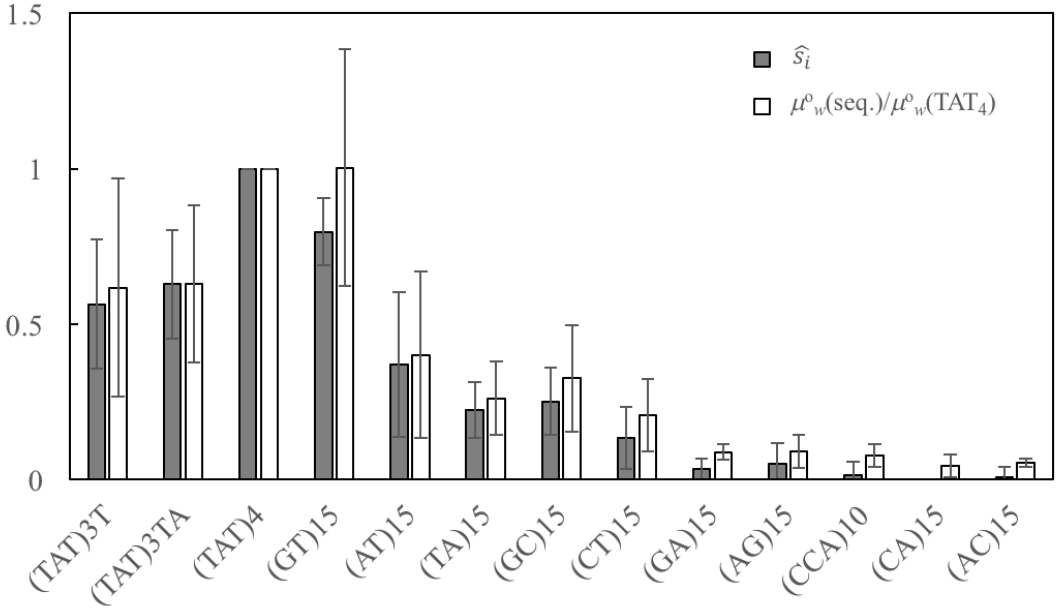

Figure 2.11. Comparison between the normalized solvation free energy and the slope of the solubility parameters. Error bars on $\widehat{s_i}$ show 90% confidence intervals for the normalized slope given by a two parameters linear function. Error bars on $\mu_w^o(seq.)/\mu_w^o(TAT_4)$ show 90% confidence intervals for the ratio of the fit parameter C_1 in equation (2.13).

2.6. Acknowledgement

This work was performed in collaboration with Dr. Akshaya Shankar of Lehigh University and Thibault Aryaksama of ESPCI in Paris, France. I would like to thank Prof. Anand Jagota and Dr. Ming Zheng who have served as mentors in this work.

2.7. References

1. Ajayan, P. M. Nanotubes from Carbon. (1999). doi:10.1021/CR970102G
2. Baughman, R. H., Zakhidov, A. A. & de Heer, W. A. Carbon Nanotubes--the Route Toward Applications. *Science (80-.).* **297,** (2002).
3. Ao, G., Khripin, C. Y. & Zheng, M. DNA-controlled partition of carbon nanotubes in polymer aqueous two-phase systems. *J. Am. Chem. Soc.* **136,** 10383–10392 (2014).
4. Zaslavsky, B. Y. *Aqueous two-phase partitioning : physical chemistry and bioanalytical applications*. (M. Dekker, 1995).
5. Zheng, M. *et al.* DNA-assisted dispersion and separation of carbon nanotubes. *Nat. Mater.* **2,** 338–342 (2003).
6. Roxbury, D., Jagota, A. & Mittal, J. Structural characteristics of oligomeric DNA strands adsorbed onto single-walled carbon nanotubes. *J. Phys. Chem. B* **117,** 132–140 (2013).
7. Roxbury, D., Jagota, A. & Mittal, J. Sequence-Specific Self-Stitching Motif of Short Single-Stranded DNA on a Single-Walled Carbon Nanotube. *J. Am. Chem. Soc.* **133,** 13545–13550 (2011).
8. Roxbury, D., Tu, X., Zheng, M. & Jagota, A. Recognition ability of DNA for carbon nanotubes correlates with their binding affinity. *Langmuir* **27,** 8282–8293 (2011).
9. Roxbury, D. Sequence Specific Interactions Between DNA and Single-Walled Carbon Nanotubes. (2012).
10. Roxbury, D., Zhang, S. Q., Mittal, J., Degrado, W. F. & Jagota, A. Structural stability and binding strength of a designed peptide-carbon nanotube hybrid. *J. Phys. Chem. C* **117,** 26255–26261 (2013).
11. Albertsson, P. a. Partition of cell particles and macromolecules in polymer two-phase systems. *Adv. Protein Chem.* **24,** 309–341 (1970).
12. Khripin, C. Y., Fagan, J. A. & Zheng, M. Spontaneous partition of carbon nanotubes in polymer-modified aqueous phases. *J. Am. Chem. Soc.* **135,** 6822–6825 (2013).
13. Fagan, J. A. *et al.* Isolation of specific small-diameter single-wall carbon nanotube species via aqueous two-phase extraction. *Adv. Mater.* **26,** 2800–2804 (2014).
14. Hildebrand, J. H., Prausnitz, J. M. & Scott, R. L. *Regular and related solutions; the solubility of gases, liquids, and solids*. (Van Nostrand Reinhold Co., 1970).
15. Hansen, C. M. *Hansen Solubility Parameters A User's Handbook*. *Journal of Chemical Information and Modeling* **53,** (2013).
16. Samaha, M. W. & Naggar, V. F. Relationship Between the Solubility Parameter and the Surface Free Energy of Some Solids. *Drug Dev. Ind. Pharm.* **16,** 1135–1151 (1990).
17. Bergin, S. D. *et al.* Multicomponent Solubility Parameters for Single-Walled Carbon Nanotube−Solvent Mixtures. *ACS Nano* **3,** 2340–2350 (2009).
18. Usrey, M. L., Chaffee, A., Jeng, E. S. & Strano, M. S. Application of Polymer Solubility Theory to Solution Phase Dispersion of Single-Walled Carbon Nanotubes.

J. Phys. Chem. C **113,** 9532–9540 (2009).

19. Hansen, C. M. *Hansen solubility parameters : a user's handbook.* (CRC Press, 2007).
20. Shankar, A., Mittal, J. & Jagota, A. Binding between DNA and carbon nanotubes strongly depends upon sequence and chirality. *Langmuir* **30,** 3176–3183 (2014).
21. Henry Hildebrand, J. & L Scott, R. *The solubility of nonelectrolytes*. (Reinhold Pub. Corp., 1950). doi:10.1021/ed042pA318.1
22. Srebrenik, S. & Cohen, S. Theoretical derivation of partition coefficient from solubility parameters. *J. Phys. Chem.* **80,** 996–999 (1976).
23. Israelachvili, J. N. *Intermolecular and surface forces*. (Academic Press, 2011).
24. Purkayastha, A. & Walkley, J. Studies in Solubility Parameter Theory for Mixed Solvent Systems. *Can. J. Chem.* **50,** 834–838 (1972).
25. Kitak, T., Dumičić, A., Planinšek, O., Šibanc, R. & Srčič, S. Determination of Solubility Parameters of Ibuprofen and Ibuprofen Lysinate. *Molecules* **20,** 21549–21568 (2015).
26. Antoniou, E., Tsianou, M. & Alexandridis, P. *Solvent Modulation of Polysaccharide Conformation*. *AIChE Annual Meeting, Conference Proceedings* (2008).
27. Li, L., Jiang, Z., Xu, J. & Fang, T. Predicting poly(vinyl pyrrolidone)'s solubility parameter and systematic investigation of the parameters of electrospinning with response surface methodology. *J. Appl. Polym. Sci.* **131,** n/a-n/a (2014).
28. Kang, H. & Sandler, S. I. Effects of Polydispersivity on the Phase Behavior of Aqueous Two-Phase Polymer Systems. *Macromolecules* **21,** 3088–3095 (1988).
29. Liu, Y., Lipowsky, R. & Dimova, R. Concentration Dependence of the Interfacial Tension for Aqueous Two-Phase Polymer Solutions of Dextran and Polyethylene Glycol. *Langmuir* **28,** 3831–3839 (2012).
30. Madeira, P. P., Teixeira, J. A., Macedo, E. A., Mikheeva, L. M. & Zaslavsky, B. Y. Correlations between distribution coefficients of various biomolecules in different polymer/polymer aqueous two-phase systems. *Fluid Phase Equilib.* **267,** 150–157 (2008).
31. Pease, L. F. *et al.* Length Distribution of Single-Walled Carbon Nanotubes in Aqueous Suspension Measured by Electrospray Differential Mobility Analysis. *Small* **5,** 2894–2901 (2009).
32. Dresselhaus, M. S., Dresselhaus, G. & Saito, R. Physics of carbon nanotubes. *Carbon N. Y.* **33,** 883–891 (1995).

2.8. Appendix

2.8.1. Additional data analysis for (8,3) and (9,1) chirality of SWCNT

The relative solvation free energy for different DNA sequences paired with (8,3) and (9,1) SWCNT hybrids were obtained in the same way as in the main text. The results are consistent with those presented in the main text for the (6,5) SWCNT.

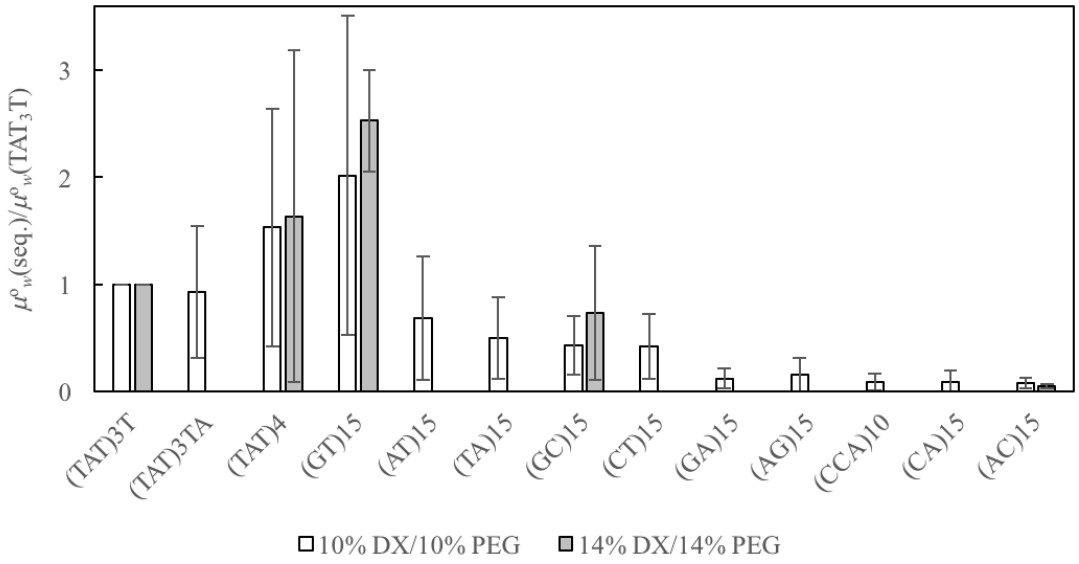

Figure 2.12. The solvation free energy for ATP systems with two different polymer compositions. Each free energy is standardized to $(TAT)_3T$ paired with the (8,3) SWCNT by equation (2.14). Error bars show 90% confidence intervals of the ratio of the fit parameter C_1 in equation (2.13).

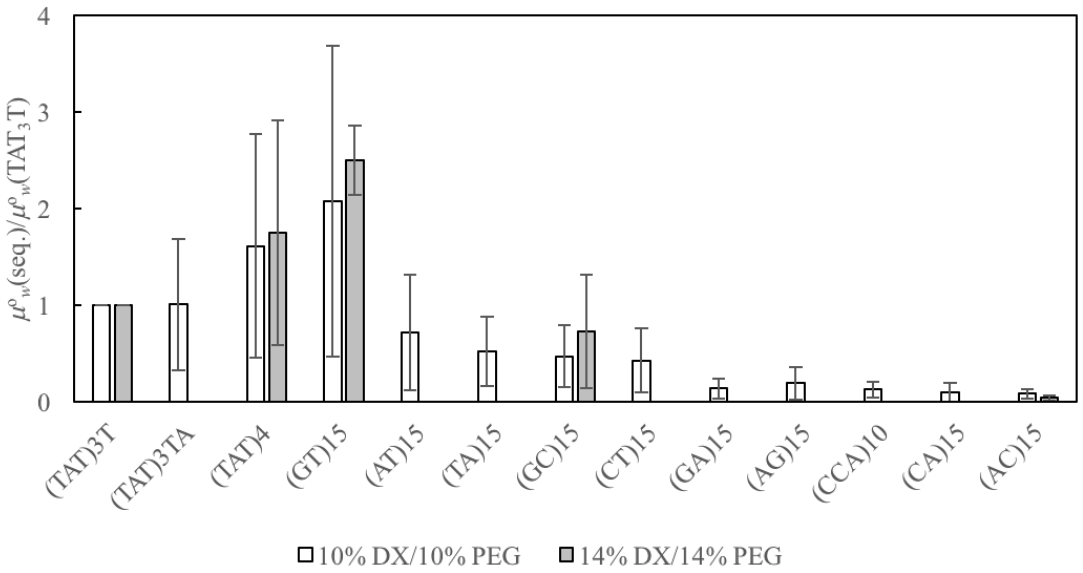

Figure 2.13. The solvation free energy for ATP systems with two different polymer compositions. Each free energy is standardized to $(TAT)_3T$ paired with the (9,1) SWCNT by equation (2.14). Error bars show 90% confidence intervals of the ratio of the fit parameter C_1 in equation (2.13).

2.8.2. Determination of $\Delta\alpha$' and $\Delta\beta$'

First, the relative solvation free energy can be estimated using equation (2.14).

Using eq. (2.10) $\Delta\alpha'$ can be written as

$$\Delta\alpha' = -\left(\frac{lnK+\Delta\beta'}{l\mu_w^o}\right) \tag{2.23}$$

Since both $\Delta\alpha'$ and $\Delta\beta'$ are relative properties, they can be set to zero at an arbitrarily chosen point. We assumed that $\Delta\beta'$ is zero for a sequence A of hybrid so the equation (2.23) for the sequence A becomes

$$\Delta\alpha'|_A = -\frac{lnK}{l\mu_w^o}\Big|_A \tag{2.24a}$$

Then $\Delta\alpha'$ for other sequences of DNA hybrid can be described as

$$\Delta\alpha'|_B = -\left(\frac{lnK}{l\mu_w^o}\Big|_B + \frac{\Delta\beta'|_{B/A}}{l\mu_w^o|_B}\right) \tag{2.24b}$$

Here, the subscript B/A means that $\Delta\beta'|_{B/A}$ is relative to $\Delta\beta'|_A$ which is set to be zero. Because the experimental data shows that ln *K* is linear in [PVP], we assume that $\Delta\alpha'$ also is linear to [PVP]:

$$\Delta\alpha' = c_1 + c_2[PVP] \tag{2.25}$$

where c_1 and c_2 are unknown parameters. Let $y = \frac{lnK}{l\mu_w^o}$, which can be obtained from the experimental data. Here, l is set to be 200 nm.[31] As defined, $\Delta\alpha'$ has a master curve for various hybrid combinations, so c_1 and c_2 are constant for the entire data set. Then, equations (2.24a) and (2.24b) can be rewritten in terms of y, c_1 and c_2:

$$y|_A = c_1 + c_2[PVP] \tag{2.26a}$$

and

$$y|_B = c_1 + c_2[PVP] + \frac{\Delta\beta'|_{B/A}}{l\mu_w^o|_B} \tag{2.26b}$$

Suppose that we have experimental data at *n* different concentration for *m* hybrid combinations. Then we have *n* x *m* equations of type (2.23) which can be expressed in the similar form in (2.26b). Then c_1, c_2 and $\Delta\beta'|_{seq./A}$ can be estimated by solving the following system of equation (in a least-squares sense).

$$\begin{pmatrix} 1 & [PVP]_{A1} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & [PVP]_{An} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & [PVP]_{B1} & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & [PVP]_{Bn} & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ & & & \vdots & & & & & \\ 1 & [PVP]_{M1} & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & [PVP]_{Mn} & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ -\Delta\beta'_{BA}/l\mu_w^o\big|_B \\ -\Delta\beta'_{CA}/l\mu_w^o\big|_C \\ \vdots \\ -\Delta\beta'_{LA}/l\mu_w^o\big|_L \\ -\Delta\beta'_{MA}/l\mu_w^o\big|_M \end{pmatrix} = \begin{pmatrix} y_{A1} \\ \vdots \\ y_{An} \\ y_{B1} \\ \vdots \\ y_{Bn} \\ \vdots \\ y_{M1} \\ \vdots \\ y_{Mn} \end{pmatrix} \tag{2.27}$$

This procedure was repeated for different chiralities and the obtained results for $\Delta\beta'$ are shown in Figure 2.6b.

2.8.3. The effect of the hybrid length distribution

In the main text, we assumed that the SWCNT length distribution in each phase is narrow enough to be represented by its mean value, and that this mean value is the same in the two phases so that their ratio is unity. Here we explore further the conditions under which the assumption of replacing length by its mean value is an acceptable approximation. To do so, we compute the expected value of partition coefficient when the solute (DNA/SWCNT hybrid) has a distribution of lengths.

First, we start with the probability of a nanotube of certain length *l* placed in bottom phase, p_b, and top phase, p_t:

$$p_{t/b} = \frac{\exp(-\mu'_{t/b} l/k_B T)}{Z} \tag{2.28}$$

where $Z \equiv \exp\left(-\frac{\mu'_b l}{k_B T}\right) + \exp\left(-\frac{\mu'_t l}{k_B T}\right)$ and $\mu'_{t/b}$ is the solvation free energy of top or bottom phase per unit length.

For a given SWCNT, its optical absorption is proportional to the total length of that SWCNT in the sample.* That is, the partition coefficient *K* can be written as the ratio of the total length of that SWCNT in each phase. The total length of nanotubes in the bottom or top phase can be obtained in terms of the product of the probability of finding it in either phase and the number probability of finding SWCNT of a certain length:

$$L_{t/b} = N \int_0^\infty lp(l) p_{t/b} dl \tag{2.29}$$

where *N* is the total number of the nanotubes and $p(l)$ is the number probability of finding SWCNT of certain length *l*. The corresponding partition coefficient is

$$K = \frac{L_b}{L_t} = \frac{\int_0^\infty lp(l) p_b dl}{\int_0^\infty lp(l) p_t dl} \tag{2.30}$$

Here, $lp(l)$ gives the length distribution $p_l(l)$. Let us set $p_l(l) \equiv lp(l)/\bar{l}$ where $\bar{l} = \int_0^\infty lp(l) dl$ and assume for the sake of concreteness that the probability has Gaussian distribution

$$p_l(l) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(l-\bar{l})^2}{2\sigma^2}\right) \tag{2.31}$$

Finally, the partition coefficient can be reduced to

$$K = \frac{\int_0^\infty \frac{p_l(l)}{1+\exp(\Delta\mu' l/k_B T)} dl}{\int_0^\infty \frac{p_l(l)}{1+\exp(-\Delta\mu' l/k_B T)} dl} \tag{2.32}$$

where $\Delta\mu' \equiv \mu'_b - \mu'_t$.

* Fagan, J. A.; et al. Length-Dependent Optical Effects in Single-Wall Carbon Nanotubes. *J. Am. Chem. Soc.* **2007**, *129*, 10607−10612.

First, we note that in the case of very narrow distribution when $p_l(l)$ approaches the delta function $\delta(\bar{l})$, K takes the form given by equation (2.7), as expected. Also, if $\Delta\mu = 0$, then K=1 identically, regardless of the length distribution.

In most experiments, the CNTs have a reasonably broad distribution but measurements are necessarily made for relatively small $\Delta\mu$, because otherwise K would either be immeasurably small or large. We wish to establish whether equation (2.7) is still accurate under such conditions. When $\Delta\mu$ is small enough, the equation (2.32) can be simplified by series expansion to

$$K = \frac{\int_0^\infty p_l(l)(1-\Delta\mu' l/2k_BT)dl}{\int_0^\infty p_l(l)(1+\Delta\mu' l/2k_BT)dl} \tag{2.33}$$

For the special case where $p_l(l)$ is given by the Normal distribution as defined above, these integrals can be evaluated if the distribution is narrow enough that we can ignore contributions from fictitious negative lengths. The partition coefficient then becomes

$$K = \frac{1-\Delta\mu'\bar{l}/2k_BT}{1+\Delta\mu'\bar{l}/2k_BT} \approx \left(1-\frac{\Delta\mu'\bar{l}}{2k_BT}\right)^2 \approx 1-\frac{\Delta\mu'\bar{l}}{k_BT} \tag{2.34}$$

This is identical to the form given by equation (7) in the appropriate limit of $\Delta\mu \to 0$:

$$K = \exp\left(-\frac{\Delta\mu' l}{k_BT}\right) \approx 1-\frac{\Delta\mu' l}{k_BT} \tag{2.35}$$

This indicates that for the interpretation of partition coefficient, for small departures from the equipartition state, i.e., when $\Delta\mu'$ is small, the length distribution can be regarded as narrow-enough to be represented simply by its mean value.

Further to demonstrate that the assumption of narrow length distribution or small $\Delta\mu'$ is valid for our experimental results, equation (2.32) was numerically solved as a function of $\Delta\mu'$ when $p_l(l)$ has Normal distribution with given value of $\bar{l}$ and σ. The mean

length $\bar{l}$ was roughly estimated as 200 nm* and the standard deviation σ was varied in the range from 20 to 200 nm.

Our experimental data have shown ln K to be a linear function of [PVP]. Since $\Delta\mu'$ is linear in ln K by the definition of the partition coefficient, $\Delta\mu'$ also can be presented as a linear function of [PVP] in our experimental range. That is, it is known by experiment that ln K is linear in $\Delta\mu'$. Based on the results represented in Figure 2.14, ln K is linear to $\Delta\mu'$ only if either $\Delta\mu'$ or σ is small enough. Because our experimental results show linear behavior, they are likely to be in such range. Thus, the comparison justifies that the length distribution can be represented by its mean value.

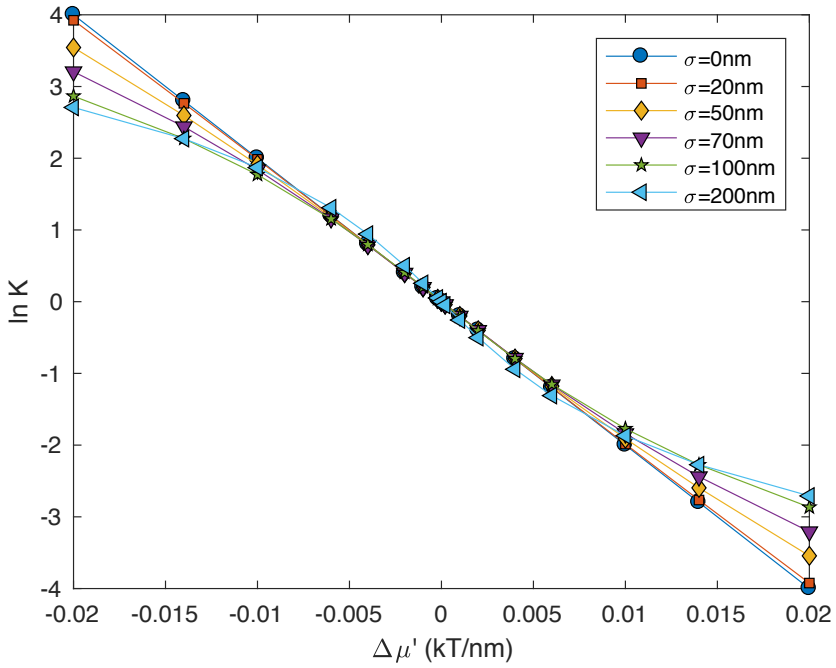

Figure 2.14. ln K as a function of $\Delta\mu'$ with $\bar{l} = 200\ nm$ for various σ, which is evaluated by solving equation (2.32) numerically.

* Hearst, J. E. The specific volume of various cationic forms of deoxyribonucleic acid. *J. Mol. Biol.* **1962**, *4*, 415−7.

2.8.4. Estimating the solubility parameter of single-stranded DNA*

We conducted ATP experiments using poly(T) ssDNA of four different lengths (T_{10}, T_{30}, T_{60}, and T_{90}), poly(A) ssDNA of two different lengths (A_{30}, A_{50}), and poly(C) ssDNA (C_{30}). The absorbance was monitored from 200 to 400 nm using a UV/vis/NIR spectrophotometer. A prominent peak was observed at 266 nm for the ssDNA. The partition coefficients were then obtained using equation (2.4), shown in Figure 2.15a. The solubility parameters of each ssDNA type were calculated using equation (2.20) in the same way as in the main text. The molar volumes for each type of ssDNA were calculated as the product of the specific volume (0.55 ml/g [2]) and their molecular weights (3059, 9075.8, 18204.5, 26714.3, 9329.3, 15574.3, and 8611.0 Da for T_{10}, T_{30}, T_{60}, T_{90}, A_{30}, A_{50}, and C_{30}, respectively).

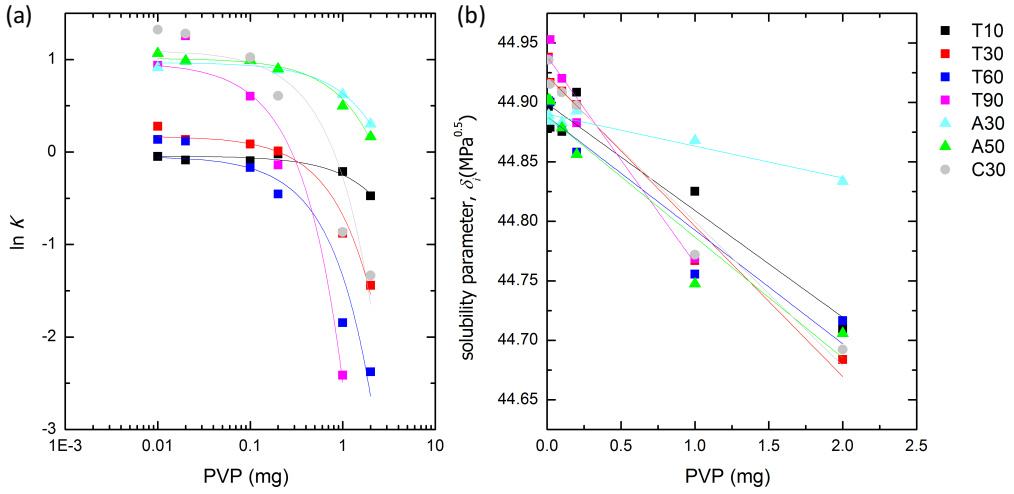

Figure 2.15. (a) Partition coefficients and (b) Hildebrand solubility parameters as a function of PVP concentration for poly(T) of four different lengths (T_{10}, T_{30}, T_{60}, and T_{90}), poly(A) ssDNA of two different lengths (A_{30}, A_{50}), and poly(C) ssDNA (C_{30}) in 10% DX/10% PEG ATP system. The solid lines are fits using a two-parameter linear function.

* The experimental work in this section has been performed by Thibault Aryaksama (ESPCI, Paris, France) at Lehigh University.

2.8.5. Hansen Solubility Parameter

From the equation (2.3), the expression of the partition coefficient shown in equation (2.20) can be replaced by the following

$$\ln K = \frac{V_i}{RT}\sum\left[\left(\delta_{k,t} - \delta_{k,i}\right)^2 - \left(\delta_{k,b} - \delta_{k,i}\right)^2\right] + \ln\frac{V_t}{V_b} \tag{2.36}$$

where the summation extends over all the partial contribution terms ($k = d, p, h$) and d, p, h denotes dispersion, polar, and hydrogen bonding contribution, respectively. The solubility parameters for top and bottom phase $\delta_{k,(t/b)}$ are then replaced by their approximate expression in terms of the volume fraction of each component:

$$\bar{\delta}_{k,(t/b)} = B_{k,(t/b)} + \phi_{PVP}\delta_{k,PVP} \tag{2.37}$$

Here, $B_{k,(t/b)} = \phi_{k,(t/b),PEG}\delta_{(t/b),PEG} + \phi_{k,(t/b),DX}\delta_{(t/b),DX} + \phi_{k,(t/b),w}\delta_{(t/b),w}$ and is constant for a given polymer composition. Substituting equation (2.37) to (2.36), the partition coefficient is then

$$\ln K = \frac{V_i}{RT}\left[\sum\left({B_{k,t}}^2 - {B_{k,b}}^2\right) + 2\sum\left(B_{k,t} - B_{k,b}\right)\left(\phi_{PVP}\delta_{k,PVP} - \delta_{k,i}\right)\right] + \ln\frac{V_t}{V_b} \tag{2.38}$$

To calculate a set of solubility parameters for a given hybrid, $\delta_{k,i}$, at least three different sets of experimental data sets are required. We recognized that the change of $B_{k,(t/b)}$ by PVP concentration is too small to make valid sets to solve equation (2.38). Thus, we recommend the use of experimental data set from different polymer composition.

2.8.6. Link to Public Repository:

The following link is to a public repository where we provide a collection of scripts for spectra decomposition and solvation free energy and solubility parameter analysis.

https://bitbucket.org/jagotagrouplehigh/dna_swcnt_atp/

Chapter 3 : Learning to Predict Single-Wall Carbon Nanotube-Recognition DNA Sequences[*]

DNA/SWCNT hybrids have enabled many applications because of their special ability to disperse and sort SWCNTs by their chirality and handedness. Much work has been done to discover recognition sequences which recognize specific chiralities of SWCNT, and significant progress has been made in understanding the underlying structure and thermodynamics of these hybrids. Nevertheless, de novo prediction of recognition sequences remains essentially impossible and the success rate for their discovery by search of the vast ssDNA library is very low. Here, we report an effective way of predicting recognition sequences based on machine learning analysis of existing experimental sequence data sets. Multiple input feature construction methods (position-specific, term-frequency, combined or segmented term frequency vector, and motif-based feature) were used and compared. The transformed features were used to train several classifier algorithms (logistic regression, support vector machine and artificial neural network). Trained models were used to predict new sets of recognition sequences, and consensus among a number of models was used successfully to counteract the limited size of the data set. Predictions were tested using aqueous two-phase separation. New data thus acquired

[*] This chapter has been published in *npj Computational Materials*:

> Y Yang, M. Zheng, A. Jagota. "Learning to predict single-wall carbon nanotube-recognition DNA sequences" *npj Computational Materials*. **2019**, 5:3; https://doi.org/10.1038/s41524-018-0142-3

was used to retrain the models by adding an experimentally tested new set of predicted sequences to the original set. The frequency of finding correct recognition sequences by the trained model increased to >50% from the ~10% success rate in the original training data set.

3.1. Introduction

In recent years, machine learning has emerged as a powerful general methodology with the ability to create well-performing predictive models from data. In particular, these techniques have become essential in bioinformatics because it is impractical to transform manually large amounts of raw sequence data into useful scientific knowledge, without requiring explicit programming instruction. Many of the important bioinformatics problems are well suited for classification algorithms, including gene annotation,[1] protein function prediction,[2,3] peptide binding prediction,[4,5] and DNA binding prediction.[6]

Single-wall carbon nanotubes (SWCNTs) comprise a family of nanomaterials with remarkable electronic, optical, and mechanical properties.[7] The structure of SWCNTs can be viewed as a cylinder obtained by rolling a hexagonal graphene sheet. The properties of SWCNTs are highly dependent on exactly how the graphene sheet is rolled, which is identifiable by chiral indices (*n*,*m*); all synthetic methods result in mixtures of different chiralities. Especially for electronic and optical applications, chirality control of the SWCNTs is of critical importance.[8,9] A number of strategies for SWCNT separation by their chirality have been developed,[10–12] and notable success has been achieved using special short DNA sequences called *recognition sequences*.[13,14] These recognize specific corresponding partner SWCNTs by forming special hybrids with sufficiently different physical and chemical properties to enable their separation from mixtures.[15] Furthermore, there is evidence that special recognition DNA/SWCNT hybrids are also effective as biosensors for specific molecular detection.[16–18]

Several studies have contributed to our understanding of the structural basis for sequence-specific recognition. Computational molecular modeling[19–23] has established a number of ordered structural motifs that single-stranded DNA (ssDNA) can adopt when adsorbed onto an SWCNT. Single-molecule force spectroscopy,[24,25] and solution based studies have provided quantitative information on strength of association between ssDNA and SWCNTs.[26,27] Aqueous two-phase (ATP) separations have been analyzed to quantify solubility of DNA-SWCNTs,[13,28,29] and fluorescence quenching studies have been used to infer wrapping structures of recognition sequences.[30]

Despite all this knowledge and understanding, we have essentially no ability to predict ssDNA sequences that will form recognition pairs with SWCNTs. Discovery of new recognition sequences has relied upon systematic searches through the vast sequence space of ssDNA. For example, Tu et. al.[31] designed a systematic search of the DNA library by sequence pattern expansion, and achieved a success rate of ~7%. In another recent study[28] some sequence patterns were found in a directed and limited search of a reduced (12-mer, T/C bases only) DNA library, achieving somewhat better performance (success rate of ~10%). We may surmise that the probability of finding a recognition sequence, conditioned upon this sequence expansion scheme, is no better than about 10%. Thus, although we have a lot of physical understanding and a reasonable amount of data, our ability to predict recognition sequences is still absent, and the search process remains time-consuming and inefficient – the number of distinct sequences in the sequence space is enormous. (For the typical sequence lengths l in the range 10 – 30, there are 10^6 – 10^{18}

distinct sequences.) Clearly, a different and more systematic approach to sequence prediction is needed.

Here, we investigate a new approach to prediction of recognition sequences using machine learning techniques. The aim is to create models to classify query sequences as either recognition or non-recognition. Multiple input feature construction methods including *n*-gram position-specific vector (*psv*), *n*-gram term-frequency vector (*tfv*), combined or segmented *tfv*, and motif-based features[6,32] were used. The models were built using a machine learning tool (WEKA).[33] As an initial study for the work presented in this manuscript, we manually tried *all* the algorithms that the WEKA package provides for binary classification using unigram and trigram position specific vector (*psv*) features. This preliminary study showed that ANN and random-forest methods worked best. However, both are of similar complexity. We decided to try three different algorithms, each algorithm representing a different level of complexity. Specifically, we used three different algorithms: logistic regression (LR, simplest),[34] support vector machine (SVM, moderately complex),[35] and artificial neural network (ANN, most complex)[35].

After training and validation using labeled data, they were used to predict new recognition sequences. The relatively small data-set size, a common issue in applying machine learning techniques to problems in materials science,[36] was mitigated by choosing consensus sequences from a number of models, i.e., we combined multiple models by cross-validation and selected the sequences only from the intersection of each set of classifier results. Predictions were tested experimentally using the ATP separation technique.[37] We retrained the model using the updated data set. This cycle of prediction,

testing, and retraining was repeated twice. Models were built on DNA sequence information only. To interpret the results in the context of previous computational[19–23] and experimental work,[13,24–29] we examined discovered motif using saliency measures within the ANN models. This study is the first attempt to predict recognition sequences for SWCNTs by adopting machine learning techniques.

3.2. Materials and Methods

3.2.1. Data Collection*

The available data on ssDNA sequences that form recognition pairs with specific SWCNTs has been obtained under varying conditions (e.g., solution conditions), sequence lengths (~8-30), and classification methods (ion-exchange chromatography, ATP, etc.). Here, we chose a recently reported set of sequences[28] that were all handled under identical conditions. To reduce complexity, in this set the DNA base type was restricted to the 2-letter (Thymine;T/Cytosine;C) alphabet and DNA length was fixed to be 12 bases. This set initially contained 9 recognition sequences (labeled as 'Y') and 73 non-recognition sequences (labeled as 'N').

To test our predicted sequences experimentally, we utilized the ATP separation technique. Preparation of DNA/SWCNT hybrids and ATP separation followed the protocols described in ref 37. Briefly, CoMoCAT SWCNTs (1 mg, SG65i grade and EG150X grade; Southwest Nanotechnologies) were suspended in 1 mL of deionized water

* The experimental part was performed at National Institute of Standards and Technology (NIST) in Gaithersburg, MD, in direct collaboration with Dr. Ming Zheng of NIST.

with 0.1 M NaCl (Sigma-Aldrich) and 2 mg ssDNA (Integrated DNA Technologies). The DNA/SWCNT mixture was dispersed using tip sonication with a power output of 8 W for 1.5 h in an ice bath. The dispersion was then centrifuged at 16,000 g for 1.5 h and the supernatant was collected. Typically, an ATP system comprising 7.76% PEG (MW 6 kDa, Alfa Aesar) and 15% polyacrylamide (PAM, 10 kDa, Sigma-Aldrich), denoted as PEG/PAM, was used for SWCNT separation, but 16 % poly(vinylpyrrolidone) (PVP, MW 10 kDa, Sigma-Aldrich) and 11 % Dextran 70 (DX, MW ~70 kDa, TCI) ATP system, denoted as PVP/DX, was used for some of the DNA/SWCNT hybrids. Both DX and PVP were used as DNA−SWCNT partition modulators. UV−vis−NIR absorbance measurements were performed on a Varian Cary 5000 spectrophotometer over the wavelength range of 200−1400 nm.

3.2.2. Feature Construction

We wish to build models that predict the class to which a sequence belongs (i.e., recognition or non-recognition). Choice of sequence representation by features is important for classifier algorithms to function well. We investigated several input feature construction (or sequence encoding) methods: position-specific vector (*psv*), term frequency vector (*tfv*), combined *tfv*, segmented *tfv*, and motif-based feature vector (*mfv*), described schematically in Figure 3.1.

A common input feature construction technique in bioinformatics is fixed-length overlapping *n*-gram analysis, which breaks sequences into subsequences using various types of vocabulary, in the case of DNA the nucleotides or the codon types.[38] Using the

method, sequences can be represented by overlapping *n*-gram patterns.

The *Position-Specific Vector* encoding method uses an indicator vector to represent each *n*-gram word at each position. Thus, a given sequence *S* can be represented by $psv(S) = \{w_{i1}, w_{i2}, \ldots, w_{il}\}$, where $w_{ij} \in$ *n*-gram vocabulary; *l* is the number of positions that is given by $(L - n + 1)$; *L* is sequence length. For example, for the sequence *A* = TTCTCC, with *n* = 2, $w_{ij} \in \{TT, TC, CT, CC\}$ and $psv(A) = \{TT, TC, CT, TC, CC\}$. To enter into the ML models, the *psv* is converted into binary features using the one-attribute-per-value approach (i.e., {TT, TC, CT, CC} ~ {(1,0,0,0), (0,1,0,0), ..., (0,0,0,1)}) by a built-in function in WEKA.[33] The *psv* represents the entire base position information but is not suitable for long sequences as the size of the feature vector becomes large. In addition, sequences with different lengths cannot be compared easily, because they result in feature vectors of different sizes.

The *term frequency vector* (*tfv*) defines the feature vector using the frequency of the *n*-gram in the sequence. For sequence *A*, $tfv(A) = \left\{1/5, 2/5, 1/5, 1/5\right\}$. The *tfv* method loses global positional sequence information – several different sequences correspond to the same *tfv* – unless the word length approaches that of the sequence itself. The *psv* method, on the other hand, contains the complete sequence information in that there is a 1-1 mapping between *psv* and the original sequence, but by treating each base as a feature it does not capture more complex features very efficiently. The *tfv* method is computational inexpensive, and can accommodate different sequence lengths.[39] However, it has a limitation that many sequences give the same *tfv*, e.g, $tfv(T_{12}) = tfv(T_{13}) = \{1,0\}$, especially for small *n*.

Previous work[28] suggests that both frequency and position information could be important for sequence prediction, and so we considered a new encoding scheme that combines features of *psv* and *tfv*. The basic idea of the method is to divide a sequence into m ($m \in [1,L]$) smaller segments of roughly equal length l_s ($l_s = L/m$). We construct a *tfv* for each segment, and then *tfv* for the entire sequence S in the following way to include position information of each segment: $tfv_{m,n}(S) = \{tfv_n(seg_1), tfv_n(seg_2), \dots, tfv_n(seg_m)\}$. Contribution to the *tfv* from terms that straddle segment boundaries are made according to a weighted average of their occupancy in either segment. For example, for sequence A, where $m = 2$ and $n = 2$, segment 1 = TTC, segment 2 = TCC, and overlapped segment = CT, so $tfv_{2,2}(A) = \left[\left\{{}^{1}/_{2.5}, {}^{1}/_{2.5}, {}^{0.5}/_{2.5}, 0\right\}, \left\{0, {}^{1}/_{2.5}, {}^{0.5}/_{2.5}, {}^{1}/_{2.5}\right\}\right]$.

With a similar purpose in mind, but in a simpler way, a combined *tfv* method was also investigated. Using n-grams with different n, different properties can be captured. For example, unigram is based only on the base frequency, while trigram captures some of the location information as well as their frequency. Thus, by combining different n-gram features, one can capture more information. The combined *tfv* can be formed as following: $tfv_{1-2-\cdots-k}(S) = \{tfv_1(S), tfv_2(S), \dots, tfv_k(S)\}$.

We next considered features based on motifs. The basic hypothesis of this method is that there are recurring patterns or motifs in the DNA sequence which recognize a special type of SWCNT. We employed a motif-discovery tool called MERCI[32] to search for motif patterns. In order to systematically select discriminative motif features, we ranked the motifs based on their conditional probabilities that a sequence is labeled 'Y', given motif: $P(Y|motif)$. The top ten recognition and non-recognition motifs were chosen for use as

features. Maximum motif lengths were limited to 5 - 7 bases for recognition motifs and 5 bases for non-recognition motifs. The extracted motifs were coded as a 20-dimensional binary feature vector, *mfv*. Entry *m* is set to '1' if motif *m* occurs in a given sequence and '0' otherwise.

Note that the range of all feature vectors were rescaled to the range in [−1, 1] to weigh all features equally.

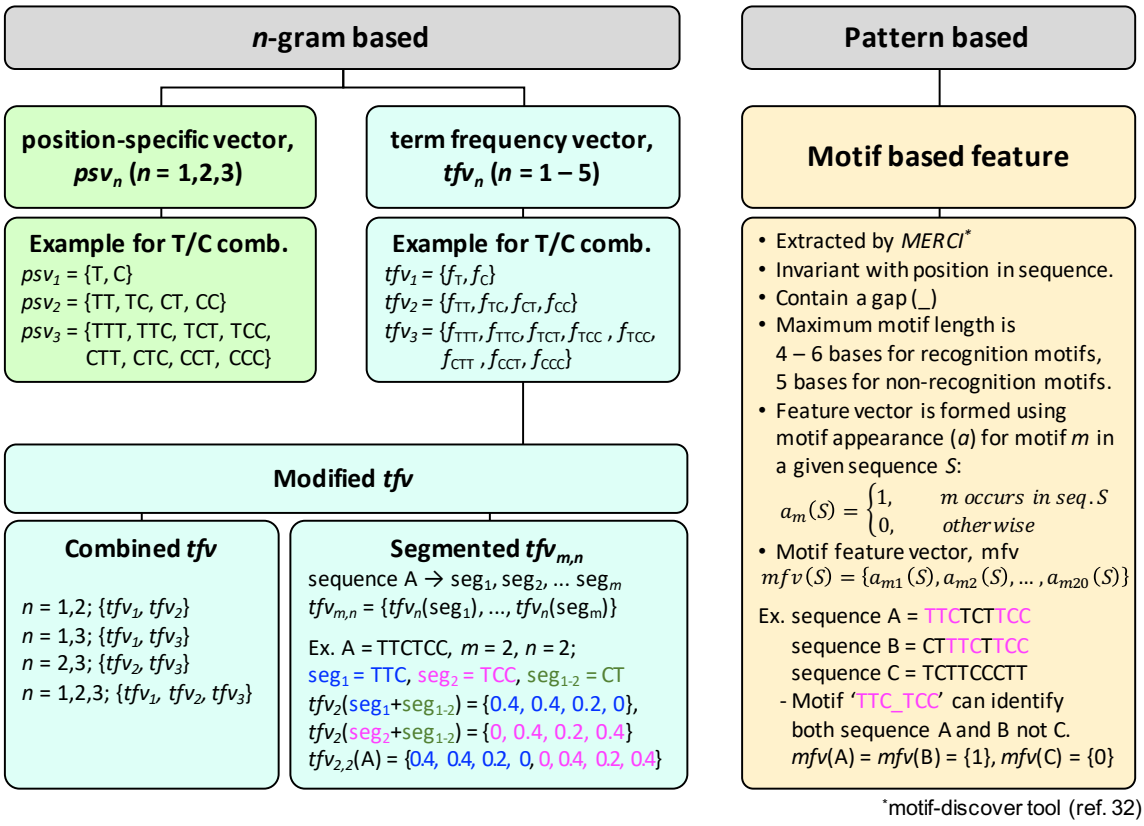

Figure 3.1. Overview of input feature construction methods explored. Feature types can be broadly categorized into two types: *n*-gram-based and pattern-based. The *n*-gram feature vectors represent DNA sequences as a collection of n-gram entities in a position-specific manner (*psv*), in terms appearance frequency (term frequency vector, *tfv*), or some combination of these two. In the pattern-based feature vector, following discovery of motifs in the training set, the DNA sequences are represented by the occurrence or absence of a given motif in that sequence.

3.2.3. Learning, Validation, and Evaluation

We began by evaluating a number of common learning algorithms for binary classification: logistic regression (LR) with ridge estimator,[40] support vector machine (SVM) using sequential minimal optimization (SMO),[41] and feedforward artificial neural network (ANN). To build and validate the classification models, we employed the open-source machine learning tool WEKA[33].

To optimize the artificial neural network models, we trained them with different numbers of hidden layers (N_l) and hidden nodes (N_h). Additionally, we optimized the cost factor γ, the ratio of false positive to false negative "cost" to vary from '1'. By maximizing γ, we reduce the chance of failure in follow-up experiments.

We also tried automated ML packages to explore all models and adjust the hyperparameters automatically using the Auto-WEKA[42] and "h2o"[43] AutoML packages. Both packages return choices for algorithms and hyperparameters – examples are provided in SI. However, because of lack of transparency, we decided to focus on the three chosen algorithms along with "manual" optimization of hyperparameters.

The performance of each the classifier was evaluated using a standard 10-forld cross-validation. Because the sample set is relatively small, and examples with the 'Y' label smaller still, we chose not to use strategies that include training, test, and validation subsets. Instead of so splitting the training set, we tested our models by using them to predict new sets of sequences that were tested experimentally. Evaluation results can be examined by the *confusion matrix*, which reports the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) predictions. To measure

prediction quality, we computed the conventional evaluation parameters such as precision ($Prc = \frac{TP}{TP+FP}$), recall ($R = \frac{TP}{TP+FN}$), or F^1–score ($F^1 = \frac{2Prc \cdot Recall}{Prc+Recall}$).

In addition, the performance was evaluated using the area under the receiver operating characteristic (ROC) curve, known as AUC.

To validate the models with newly identified sequences, normalized prediction error *E* is calculated by

$$E = \sum|t - t_c| / 2n \quad (3.1)$$

Here, t_c is the prediction probability for each instance calculated by the classifier and *t* is the experimentally determined truth value, '1' for recognition sequences, '–1' for non-recognition sequences, and '0' for marginal sequences, and *n* is the number of instances.

3.3. Results and Discussion

3.3.1. Initial models – Training, validation, prediction, and evaluation

The overall scheme of our approach is shown in

Figure 3.2. During the first round of learning, the models were trained by using three types of algorithms (LR, SVM, and ANN) with n-gram *psv* and *tfv* (*n* = 1–3) using the data set described in data collection section (listed in Table 3.1). The final models that gave the highest precision were chosen. This is because precision is directly related to the ability to find new recognition sequence (TP) correctly in the experiment, which is the most labor-intensive and time-consuming part of the entire process. The performance of models is shown in Table 3.2. Once a model was built, we generated a query sequence set, including all possible sequences (~2^{12}). These were then classified as recognition or non-

recognition sequence using each of our previously trained models. Each model typically predicted hundreds of recognition sequences, still far too many to test. Furthermore, because our training set is small relative to the size of the query sequence set (i.e. 82 vs. 4014), one needs to be wary of overfitting. To resolve these issues, we combined multiple models by cross validation; sequences for experimental testing were selected only from the intersection of each set of classifier results.

We experimentally tested the 10 most frequently occurring sequences among the sequences predicted to be recognition by our classifiers (Table 3.3). We identified 5 sequences (labelled 'Y') that lead to partitioning of only one particular (*n*,*m*) SWCNT species with high yield. Figure 3.3 shows the absorbance spectra of the purified SWCNT species by the five sequences and the starting material. In each spectrum of the purified species, the observed sharp peaks correspond to the characteristic optical transitions of a particular (*n*,*m*) species. Considering the prediction efficiency, this is a remarkable result, with prediction efficiency of 50%, a significant improvement over the ~10% frequency of recognition sequences in the training set.[28] We also found two marginal sequences that could not safely be classified as recognition sequence because they had insufficient yield or selectivity although they did show enrichment of a particular (*n*,*m*) SWCNT species in a given phase. These sequences were labeled as non-recognition sequence in order to maximize stringency of 'Y' labels in the training set.

The previously trained models were then evaluated based on their prediction errors on the newly tested sequences using eq. (3.1) (depicted as a heat map in Figure 3.4a). The total prediction errors among the models using *psv* are not significantly different from each

other, while the models using *tfv* showed considerable difference. Compared within the same input feature construction method, the trigram ANNs are better on both, showing the normalized prediction error of 0.38 and 0.423 for *psv* and *tfv,* respectively.

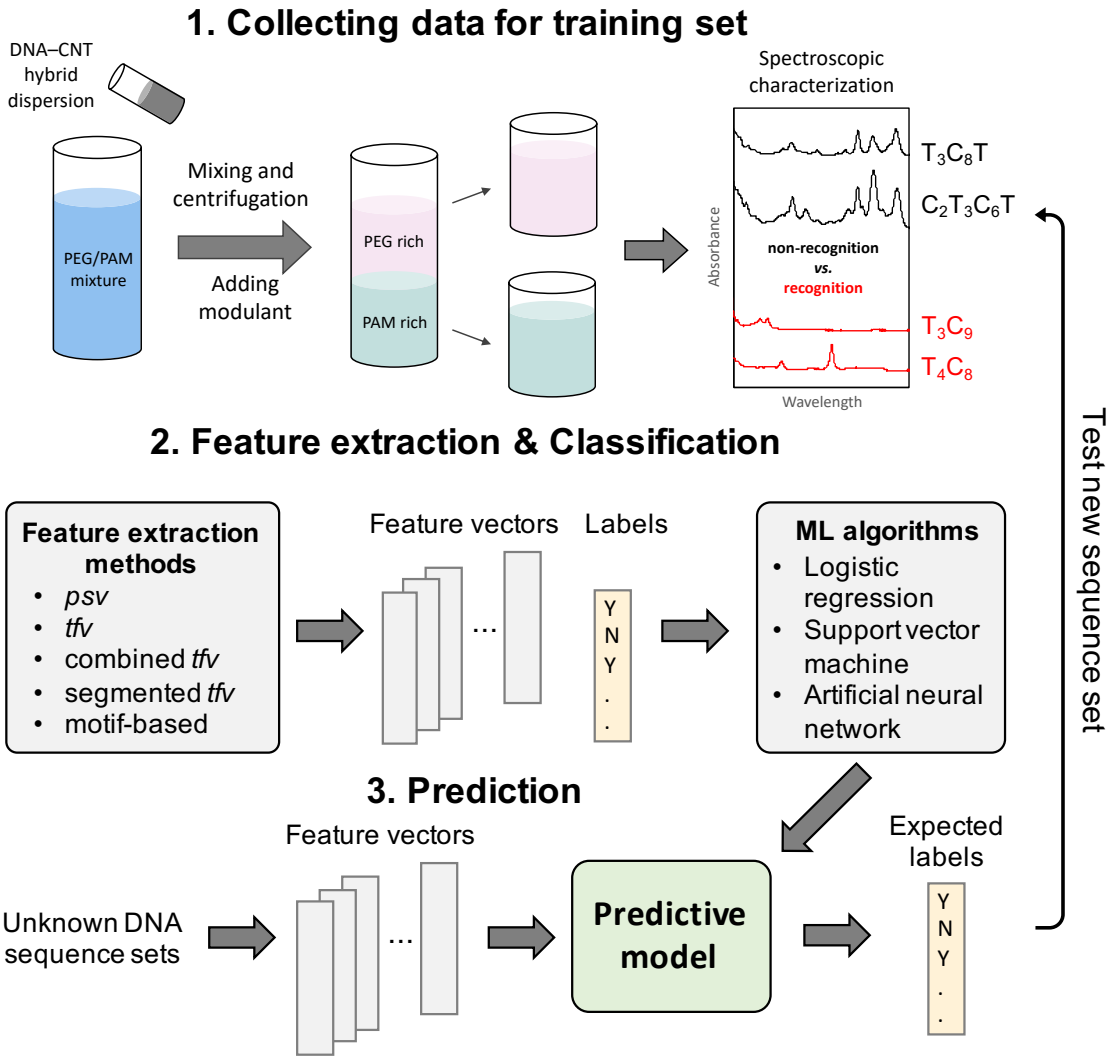

Figure 3.2. Overall scheme to develop a model to predict and test DNA recognition sequences. First, the training data set is collected using the ATP technique. If the DNA/CNT hybrid can allow partitioning one type of SWCNT in either the top or the bottom phase, that sequence is labeled as a recognition sequence ('Y'). This is done via the NIR absorbance spectra of sorted fractions. Once the data are collected, the DNA sequences and their labels are encoded to a numeric vector, which is called input feature construction. Then, the models with three different types of classification algorithms are trained using the training set feature vectors. A generated query sequence set including all possible sequences (~2^{12}) in the 12 mer C/T library are then classified using the trained models. Limitations due to small data set size are mitigated by choosing the consensus of a number of models. The predicted recognition sequences are tested using the ATP technique again. The new data are added to the existing labeled sequence data and the models are retrained. This procedure was repeated twice.

Table 3.1. Training data set of DNA sequences and partner SWCNT species with their corresponding labels, used to develop the predictive models in the first round of learning. The data set was identified by Ao et al. (2014)[13] using the ATP technique. If the DNA/CNT hybrid can allow partitioning one type of SWCNT in either the top or the bottom phase, that sequence is labeled as a recognition sequence ('Y'). To reduce complexity, the DNA base type was restricted to the 2-letter (T/C) alphabet and DNA length was fixed to be 12 bases.

#	Sequence	CNT species	Class
1	TTTCTCCTCTCT		N
2	TCTTTCCTCTCT		N
3	TCTCTTCTCTCT		N
4	TCTCTCTTCTCT		N
5	TCTCTCCTTTCT		N
6	TCTCTCCTCTTT		N
7	CCTCTCCTCTCT		N
8	TCCCTCCTCTCT		N
9	TCTCCCCTCTCT		N
10	TCTCTCCCCTCT		N
11	TCTCTCCTCCCT	(10,2)	Y
12	TCTCTCCTCTCC		N
13	TCCCCCCCCCCC		N
14	CCCCTCCCCCCC		N
15	CCCCCTCCCCCC	(10,3)	Y
16	CCCCCCTCCCCC		N
17	CCCCCCCTCCCC		N
18	CCCCCCCCCTCC		N
19	TTTTTTTTTTTT		N
20	TTTTTCCTTTTT		N
21	TTTTCTTCTTTT		N
22	TTTTCCCCTTTT	(11,1)	Y
23	TTTCTTTTCTTT		N
24	TTTCTCCTCTTT		N
25	TTTCCTTCCTTT		N
26	TTTCCCCCCTTT	(8,5), (9,6), (9,9)	Y
27	TTCTTTTTTCTT		N
28	TTCTTCCTTCTT		N
29	TTCTCTTCTCTT		N
30	TTCTCCCCTCTT		N
31	TTCCTTTTCCTT		N
32	TTCCTCCTCCTT		N
33	TTCCCTTCCCTT		N
34	TTCCCCCCCCTT		N
35	TCTTTTTTTTCT		N
36	TCTTTCCTTTCT		N
37	TCTTCTTCTTCT		N
38	TCTTCCCCTTCT		N
39	TCTCTTTTCTCT		N
40	TCTCTCCTCTCT	(7,3)	Y
41	TCTCCTTCCTCT		N
42	TCTCCCCCCTCT		N
43	TCCTTTTTTCCT		N
44	TCCTTCCTTCCT		N
45	TCCTCTTCTCCT		N
46	TCCTCCCCTCCT		N
47	TCCCTTTTCCCT		N
48	TCCCTCCTCCCT		N
49	TCCCCTTCCCCT		N
50	TCCCCCCCCCCT	(8,5)	Y

#	Sequence	CNT species	Class
51	CTTTTTTTTTTC		N
52	CTTTTCCTTTTC		N
53	CTTTCTTCTTTC		N
54	CTTTCCCCTTTC		N
55	CTTCTTTTCTTC		N
56	CTTCTCCTCTTC		N
57	CTTCCTTCCTTC		N
58	CTTCCCCCCTTC	(11,2)	Y
59	CTCTTTTTTCTC		N
60	CTCTTCCTTCTC		N
61	CTCTCTTCTCTC		N
62	CTCTCCCCTCTC		N
63	CTCCTTTTCCTC		N
64	CTCCTCCTCCTC		N
65	CTCCCTTCCCTC		N
66	CTCCCCCCCCTC		N

#	Sequence	CNT species	Class
67	CCTTTTTTTTCC		N
68	CCTTTCCTTTCC		N
69	CCTTCTTCTTCC		N
70	CCTTCCCCTTCC		N
71	CCTCTTTTCTCC		N
72	CCTCTCCTCTCC		N
73	CCTCCTTCCTCC		N
74	CCTCCCCCCTCC		N
75	CCCTTTTTTCCC		N
76	CCCTTCCTTCCC		N
77	CCCTCTTCTCCC		N
78	CCCTCCCCTCCC		N
79	CCCCTTTTCCCC		N
80	CCCCTCCTCCCC		N
81	CCCCCTTCCCCC	(8,8), (9,7)	Y
82	CCCCCCCCCCCC	(11,0)	Y

Table 3.2. Validation results of the initial models using *n*-gram position-specific vector (*psv*) and term frequency vector (*tfv*). Note that an empty cell indicates there were no true positives. The evaluation factors were obtained by 10-fold cross-validation in the training set. In general, SVM showed poor performance (especially with *tfv*). Notice that none of models with unigram ($n = 1$) *tfv* can classify any recognition sequences in the training set correctly. The best performances in terms of F^1-score or *S'* are highlighted.

	psv ($n = 1$)			*psv* ($n = 2$)			*psv* ($n = 3$)		
	LR	ANN	SVM	LR	ANN	SVM	LR	ANN	SVM
Optimization	$\gamma = 25$			$\gamma = 2$	$\gamma = 5$ N_l=1, N_h=30		$\gamma = 2$	$\gamma = 1$ N_l=1, N_h=38	$\gamma = 1$
Precision	0.286			0.357	0.222		0.278	0.200	0.143
Recall	0.222			0.556	0.222		0.556	0.222	0.111
F^1-score	0.250			0.435	0.222		0.370	0.211	0.125
S'	0.577			0.717	0.564		0.690	0.557	0.515

	tfv ($n = 1$)			*tfv* ($n = 2$)			*tfv* ($n = 3$)		
	LR	ANN	SVM	LR	ANN	SVM	LR	ANN	SVM
Optimization				$\gamma = 1$			$\gamma = 1.5$	$\gamma = 1$ N_l=2, N_h=4	
Precision				0.200			0.500	0.500	
Recall				0.111			0.111	0.333	
F^1-score				0.143			0.182	0.400	
S'				0.529			0.549	0.646	

Table 3.3. DNA sequences predicted by our classifiers and tested using ATP separation. 'Y' denotes recognition sequence and 'N' denotes non-recognition sequence. The superscript '*' denotes marginal sequence due to its low yield or selectivity.

Initial models				1st retrained models			
Name	Sequence	CNT species	Class	Name	Sequence	CNT species	Class
S01	CTT CCC CCC CCT	(7,3)	Y	S11	TTT TCC CCC CTC		N
S02	CTT CCC CCC CCC		N	S12	TTT CCC CCC CTC	*(7,5)*	N*
S03	TTT CCC CCC CCC	(6,4)	Y	S13	TTT TTC CCC CCT	*(9,6)*	N*
S04	TTT CCC CCC CCT		N	S14	TTT TTT CCC CCT	(10,2)	Y
S05	TTT TCC CCC CCT	(10,4)	Y	S15	CCC CCC CCC CTC	*(8,5)*	N*
S06	TTT TCC CCC CCC	(8,5)	Y	S16	TTT CTC CCC CCT	(7,6), (6,5)	Y
S07	CTC CCT CCC CCT	*(7,6)*	N*	S17	CCC CCC CCC CCT	*(8,5)*	N*
S08	CCT TTC CCC CCT		N	S18	CCC CCC CCC TTC	(11,0)	Y
S09	CCT TCC CCC CCT	*(9,7)*	N*	S19	TTT TTC CCC CCC	(8,5)	Y
S10	CCC CCT CCC CCT	(7,5)	Y	S20	TTC TCC CCC CCT	(8,5)	Y

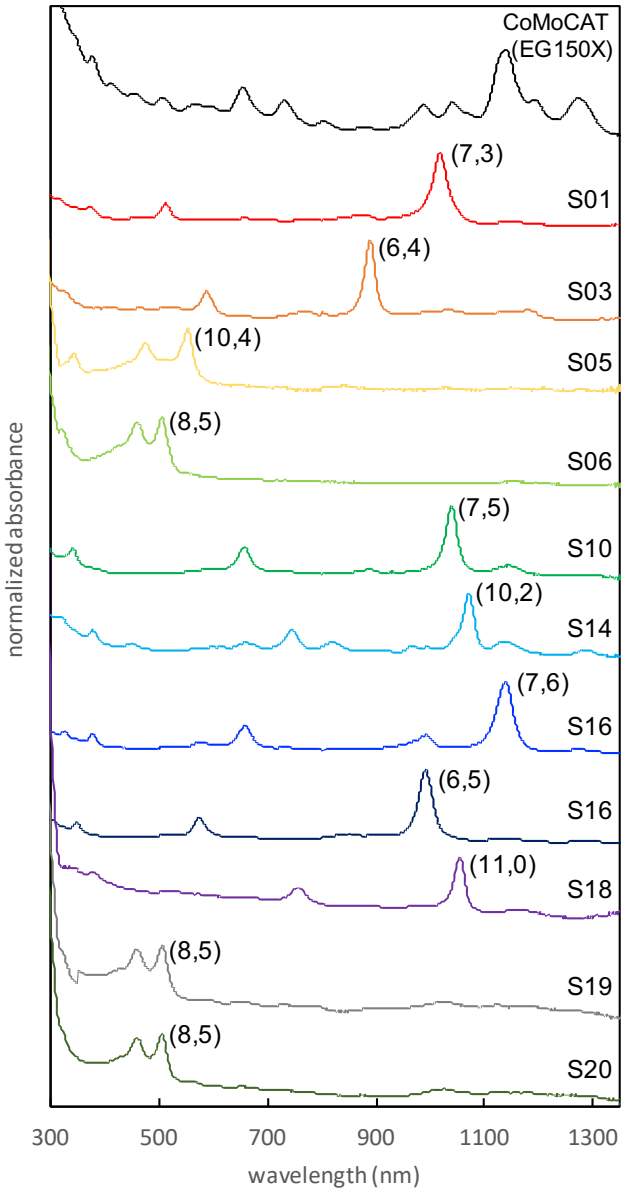

Figure 3.3. Absorbance spectra of SWCNT species purified by ATP using new sequences and the starting CoMoCAT (EG150X) mixture. The SWCNT species have been identified by their E_{11} and E_{22} peak positions (M_{11} for metallic species). Each spectrum is normalized at the E_{11} peak position (M_{11} for metallic species) and the baseline level of each spectrum was manually offset for visual clarity.

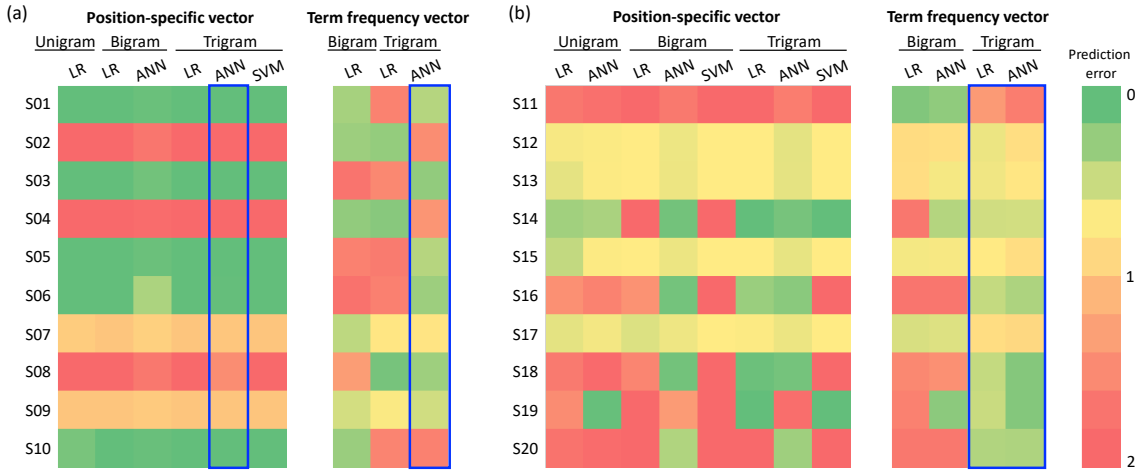

Figure 3.4. Absolute prediction error heat map of (a) initial models and (b) 1st retrained models *vs*. experimentally identified sequences. Total prediction errors were calculated using eq (3.1) then normalized by twice the number of instances (= 2 ×10): For the initial models with *psv*, values are 0.4, 0.4, 0.395, 0.4, 0.38, and 0.4 for unigram LR, bigram LR and ANN, trigram LR, ANN, and SVM, respectively. For the initial models with *tfv*, values are 0.479, 0.564, and 0.423 for bigram LR, trigram LR and ANN, respectively. For the 1st retrained models with *psv*, values are 0.638, 0.605, 0.763, 0.394, 0.8, 0.42, 0.405, and 0.6 for unigram LR and ANN, bigram LR, ANN, and SVM, trigram LR, ANN, and SVM, respectively. For the 1st retrained models with *tfv*, values are 0.6, 0.434, 0.317, and 0.324 for bigram LR and ANN and trigram LR and ANN, respectively. Overall, ANN for both trigram *psv* and *tfv* of initial models and LR and ANN with trigram *tfv* of 1st retrained models showed the best performance (highlighted with blue box).

3.3.2. Retrained models – Training, validation, and prediction

In the second round of learning, the training set was updated by including newly determined sequences by ATP separation, and the models were retrained. Ten new recognition sequences (S11–S20) were predicted and tested experimentally.

Although most retrained models showed improved validation performance (Table 3.4), the actual prediction performance of 50% remained the same as that of the initial models (Table 3.3, Figure 3.3). Note that only one sequence was determined as non-recognition sequence

and the remaining 4 were deemed marginal (Table 3.3). This indicates that the retrained models performed somewhat better than initial models but not well enough to drastically increase the prediction efficiency. Four of ten predicted sequences interestingly have an ability to purify (8,5) species. Evidently, our retrained models are likely to predict recognition sequences for (8,5) species.

Table 3.4. Validation results of 1st retrained models using *n*-gram position-specific vector (*psv*) and term frequency vector (*tfv*)

	psv (*n* = 1)			*psv* (*n* = 2)			*psv* (*n* = 3)		
	LR	ANN	SVM	LR	ANN	SVM	LR	ANN	SVM
Optimization	$\gamma = 1.1$	$\gamma = 1$ N_l=2, N_h=6		$\gamma = 3$	γ=1.7 N_l=3, N_h=23	$\gamma = 1$	$\gamma = 7$	$\gamma = 2.2$ N_l=3, N_h=43	$\gamma = 2$
Precision	0.200	0.304		0.308	0.467	0.462	0.300	0.400	0.364
Recall	0.143	0.500		0.286	0.500	0.429	0.429	0.429	0.286
F^1-score	0.167	0.378		0.296	0.483	0.444	0.353	0.414	0.320
S'	0.520	0.647		0.585	0.699	0.669	0.625	0.657	0.598

	tfv (*n* = 1)			*tfv* (*n* = 2)			*tfv* (*n* = 3)		
	LR	ANN	SVM	LR	ANN	SVM	LR	ANN	SVM
Optimization				$\gamma = 1.5$	$\gamma = 1$ N_l=1, N_h=6		$\gamma = 1$	$\gamma = 1$ N_l=1, N_h=6	
Precision				0.500	0.500		0.714	0.400	
Recall				0.071	0.214		0.357	0.286	
F^1-score				0.125	0.300		0.476	0.333	
S'				0.529	0.588		0.666	0.604	

3.3.3. Design of improved models

In the first round, to find optimal models, we used cross-validation. Although cross-validation is designed to minimize overfitting, there is still some concern because the validation set is not independent of the training set. In the second phase, we estimate the model performance based on the prediction errors calculated using a newly tested sequence set that is independent of training sets.

Figure 3.4b shows the prediction errors of the retrained models. In general, the models with *tfv* gave smaller error than models with *psv*. For the models with *psv*, bigram ANN and trigram ANN and LR perform much better (with the error of 0.394, 0.405, and 0.42, respectively) than others. Among the models with *tfv*, trigram LR and ANN showed smaller error of 0.317 and 0.324.

Although the prior models already showed very good performance, we explored improved training methods to further enhance the prediction accuracy in next round of experiments. First, we selected and focused on *tfv*, for its ability to handle sequences of different lengths. Next, we dropped the use of SVM since validation results revealed that SVM models are generally poor (Table 3.2 & Table 3.4 and Figure 3.4). We also found that the models with small *n*-gram of *psv* and *tfv* showed poor performance (Table 3.2 & Table 3.4), so higher *n*-gram (n = 3-5) *tfv* were examined in 2nd retrained models. For ANN models, in most cases we found best performance with a single layer. Additionally, given the size of our training set, we restricted N_l to be single and N_h to be no larger than twice the size of the feature vector to avoid overfitting.

The overall optimization was previously performed on the precision because it is more important to classify actual non-recognition sequence incorrectly (i.e., low FP) than to classify actual recognition sequence incorrectly (i.e., high FN) when we test the predicted sequences in the lab. However, a better indicator of model quality should account for both FP and FN, so F^1 and S scores were subsequently used for optimization. Furthermore, additional feature construction methods were examined as described in the section 3.2.2.

The motif were searched by the motif-mining tool, MERCI,[32] with the minimal occurrence frequency for positive sequences f_P and the maximal occurrence frequency for negative sequences f_N. It is worth noting that the motifs identified when f_N is set to zero (motifs are absent from the negative set) contained at least 8 bases (Table 3.5). This may indicate that the minimum length of the recognition sequence pattern is 8 bases, which is consistent with the fact that most recognition sequences found so far contain 8 bases or more. Furthermore, earlier computation studies have shown that the ssDNA/SWCNT free energy can be minimized by maximizing base/SWCNT stacking and inter-base hydrogen bonding, which requires ssDNA of sufficient length that can wrap a nanotube at least one round.[21] The motif discovery results would suggest that at least 8 bases are required to wrap around SWCNT tightly. Longer motifs may play a stronger role in distinguishing recognition sequences, but considering that our DNA length is 12 bases, the motif should not be too long. Using 8 bases motifs allows only four degrees of freedom which can result in too limited a search. Thus, we set the maximum motif lengths to be 5 to 7 for recognition motifs. For non-recognition motifs, there are more short motifs. When the maximum motif

length is 6 or more, the number of motifs with $P(N|motif) = 1$ is larger than 10, and the top 10 motifs cannot be selected. Thus, the maximum motif lengths were limited to 5 bases for non-recognition motif. The motifs were ranked based on their conditional probabilities of a recognition or non-recognition sequence given motif, $P(Y\ or\ N|motif)$, shown in Figure 3.5. Top 10 motifs for both recognition and non-recognition motifs were chosen for use as features, listed in Table 3.5.

Table 3.5. Base motifs used for recognition and non-recognition classes.

Recognition motifs			**Non-recognition motifs**
$L_{rec} \leq 5$	$L_{rec} \leq 6$	$L_{rec} \leq 7$	$L_{non\text{-}rec} \leq 5$
CCCCC	TCCCCC	CCCCCTT	CTCTT
TTCCC	CCCCC	TCCCCCT	CTTCT
CCCC	CCCCTT	TTCCCCC	TCCTT
TCCCC	TTTCCC	CCCCTTC	TCTT
C_CCC	TTCCCC	CCCCTTT	TCTTC
CCC	CCCCCT	CCTCCCT	TCTTT
CCCTT	CCCCCC	TTCTCCC	TTCCT
CCCCT	TT_CCC	TTTCCCC	C_TCT
TCCC	TTCCC	TTTTCCC	CCTCT
T_CCC	CCCC	TCCCCC	CTTTT

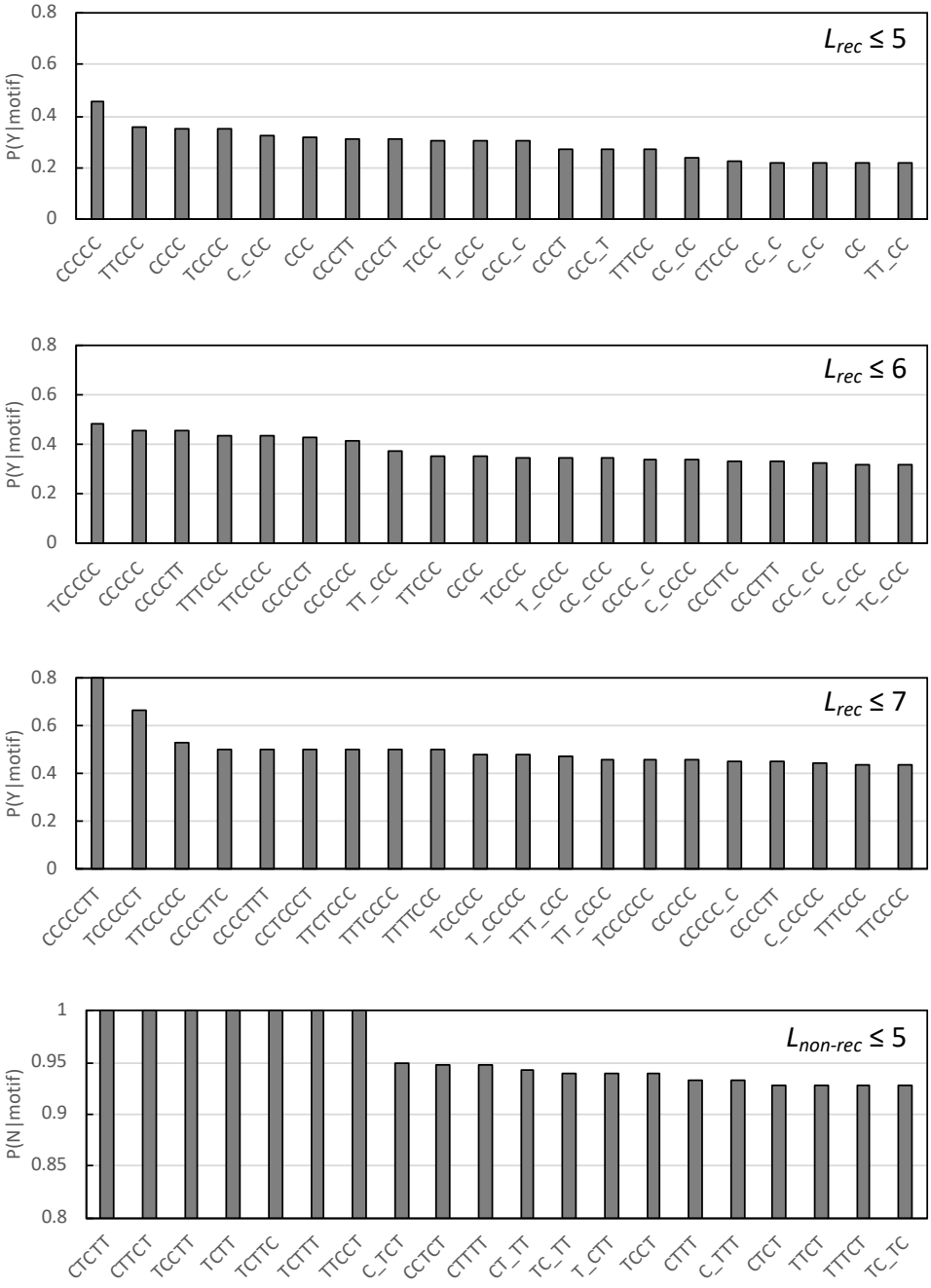

Figure 3.5. Top ranking base motifs with varying maximum length of motifs from 5 to 7 for recognition sets and maximum length of 5 for non-recognition set.

Finally, we retrained the models using LR and ANN with simple *tfv*, combined or segmented *tfv*, and motif-based features using the updated training set (Table 3.6 and Figure 3.6).

Top five models that gave the highest F^1–scores are listed in Table 3.7. In general, ANN showed better performance than LR, and trigram *tfv* and motif-based features showed high performance. ANN with simple trigram *tfv* (tfv_3) shows the best performance, while the combined bigram & trigram *tfv* ($tfv_{2\text{-}3}$) and bi-segmented trigram *tfv* ($tfv_{2,3}$) show third best performances. It is interesting that combined or segmented trigram *tfv* do not perform better than simple *tfv*, even though they already contain simple *tfv* inside. This implies that irrelevant features can cause poor performance, which leads to the need for a saliency analysis.

Table 3.6. Validation results of the 2nd retrained models with different input feature construction methods. Top two models with each input feature construction methods were listed. The evaluation factors were obtained by 10-fold cross-validation in the training set.

Algorithms	Feature	Optimization	Precision	Recall	F^1-score	S'
LR	tfv_4	$\gamma = 1.3$	0.474	0.474	0.474	0.677
	tfv_5	$\gamma = 1$	0.385	0.526	0.444	0.667
ANN	tfv_3	$N_h = 11$, $\gamma = 1$	0.600	0.632	0.615	0.768
	tfv_4	$N_h = 6$	0.444	0.421	0.432	0.650
LR	combined $tfv_{1\text{-}2\text{-}3}$		0.533	0.421	0.471	0.668
	combined $tfv_{2\text{-}3}$		0.533	0.421	0.471	0.668
ANN	combined $tfv_{2\text{-}3}$	$N_h = 4$	0.556	0.526	0.541	0.710
	combined $tfv_{1\text{-}2\text{-}3}$	$N_h = 9$	0.529	0.474	0.500	0.689
LR	segmented $tfv_{2,3}$		0.471	0.421	0.444	0.646
	segmented $tfv_{2,4}$		0.385	0.526	0.444	0.704
ANN	segmented $tfv_{2,3}$	$N_h = 9$	0.556	0.526	0.541	0.710
	segmented $tfv_{2,2}$	$N_h = 2$	0.450	0.474	0.462	0.676
LR	motif ($L_{rec} \leq 6$)		0.480	0.632	0.545	0.770
	motif ($L_{rec} \leq 5$)		0.429	0.474	0.450	0.675
ANN	motif ($L_{rec} \leq 7$)	$N_h = 4$	0.529	0.474	0.500	0.678
	motif ($L_{rec} \leq 6$)	$N_h = 3$	0.500	0.421	0.457	0.647

Table 3.7. Top five 2nd retrained models showing best performance

Algorithms	Feature	Optimization	Precision	Recall	F^1–score
ANN	tfv_3	$N_h = 11$, $\gamma = 1$	0.600	0.632	0.615
LR	motif ($L_{rec} \leq 6$)		0.480	0.632	0.545
ANN	combined $tfv_{2\text{-}3}$	$N_h = 4$	0.556	0.526	0.541
ANN	segmented $tfv_{2,3}$	$N_h = 9$	0.556	0.526	0.541
ANN	combined $tfv_{1\text{-}2\text{-}3}$	$N_h = 9$	0.529	0.474	0.500

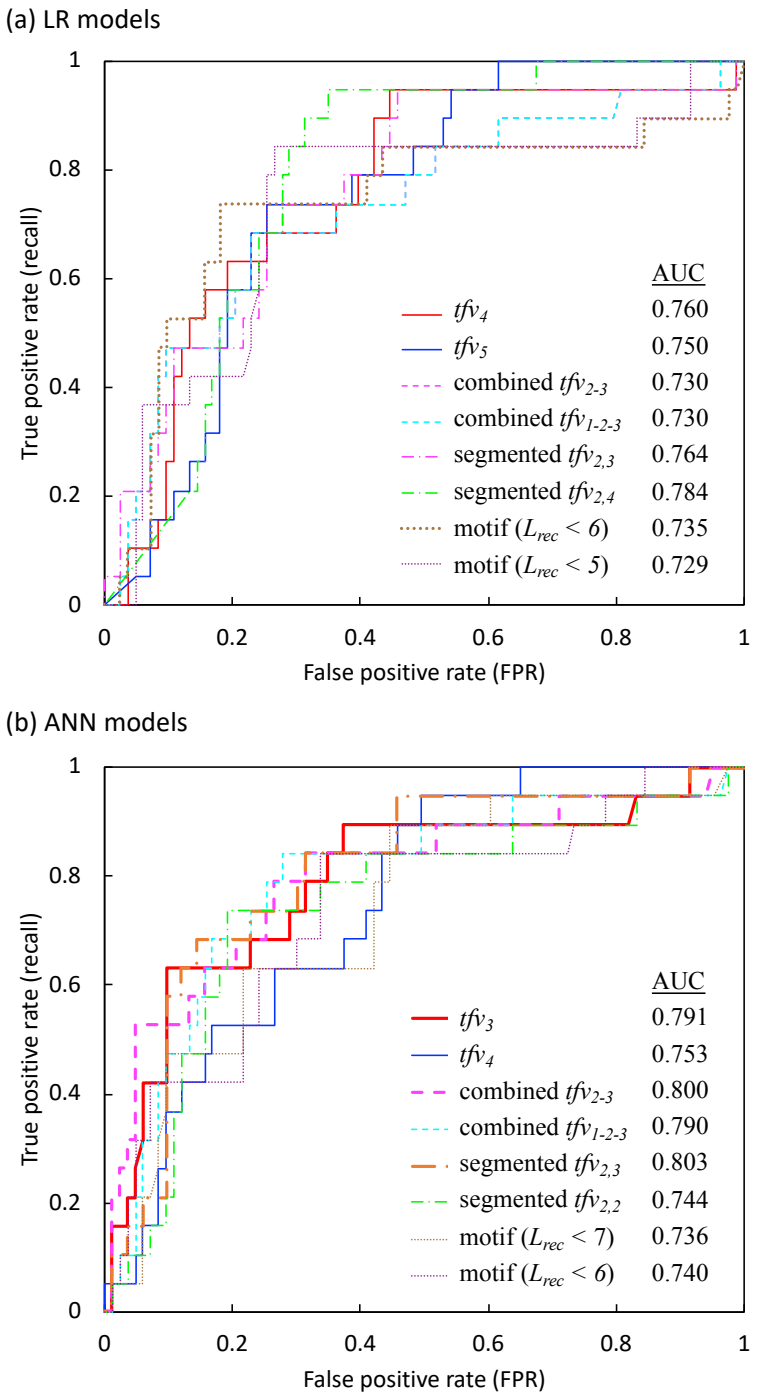

Figure 3.6. ROC curve performances of top two (a) LR models and (b) ANN models with each input feature construction methods. In general, ANN models show higher AUC than LR models.

3.3.4. Saliency analysis and overall observations

The saliency measures can be used to identify important input features.[44] Figure 3.7 shows that the saliency of segmented $tfv_{4,3}$ ANN models is high in the features of the first and last segment (i.e., at the ends of the sequences). Previous studies on the displacement of ssDNA by surfactants[26,27] suggest that the difference between recognition and non-recognition sequences is due to structural differences at sequence ends. Saliency results support that experimental finding.

Saliency also can be used to study model performance by examining the number of irrelevant features, defined by when the standard deviation is larger than the mean value. We rank models by the ratio of the irrelevant to total features. The top 4 models with lowest irrelevant feature ratio are tfv_3, motif-based feature with $L_{rec} = 7$, the combined $tfv_{2\text{-}3}$, and $tfv_{1\text{-}2\text{-}3}$. These four are also the top 4 ANN models based on the validation results.

Figure 3.8 shows the *n*-gram frequency of the final training set. Recognition sequences evidently contain higher frequency of 'CCC', especially in the newly discovered sequences (red box). This is consistent with a previous experimental finding.[28]

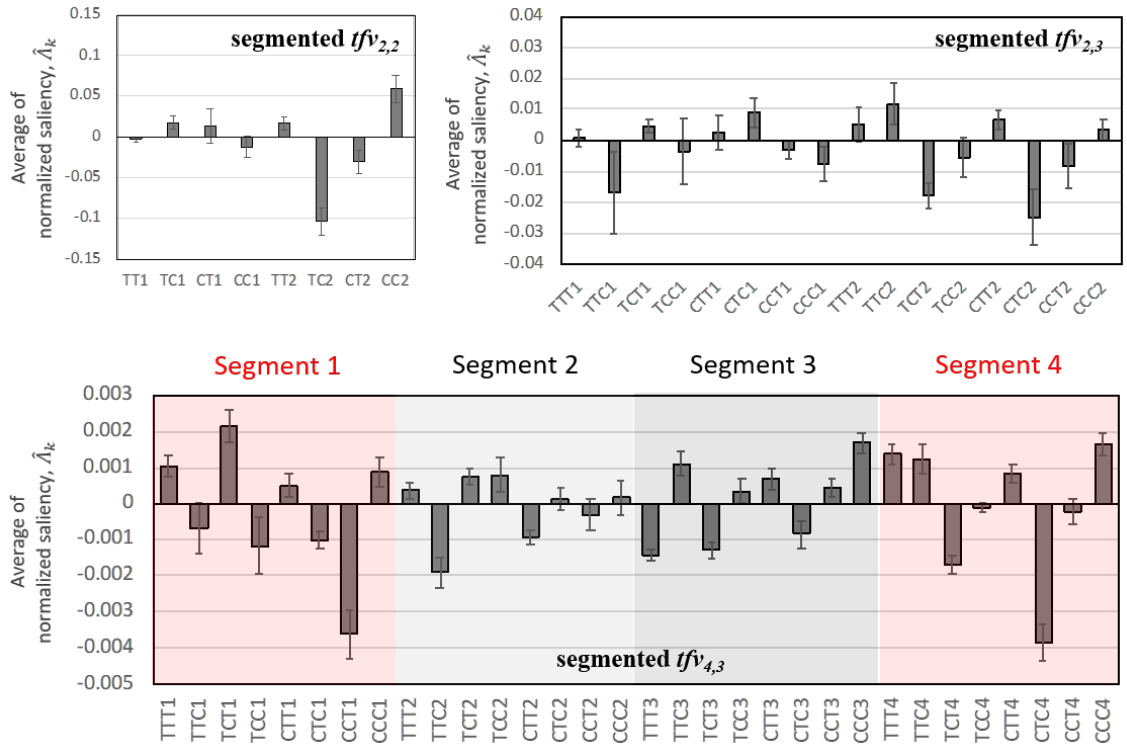

Figure 3.7. Average of normalized saliency of the feedforward ANN models with segmented $tfv_{2,2}$ (top left), $tfv_{2,3}$ (top right), and $tfv_{4,3}$ (bottom) calculated by eq. (3.5b) in Appendix. Error bar represents standard deviation.

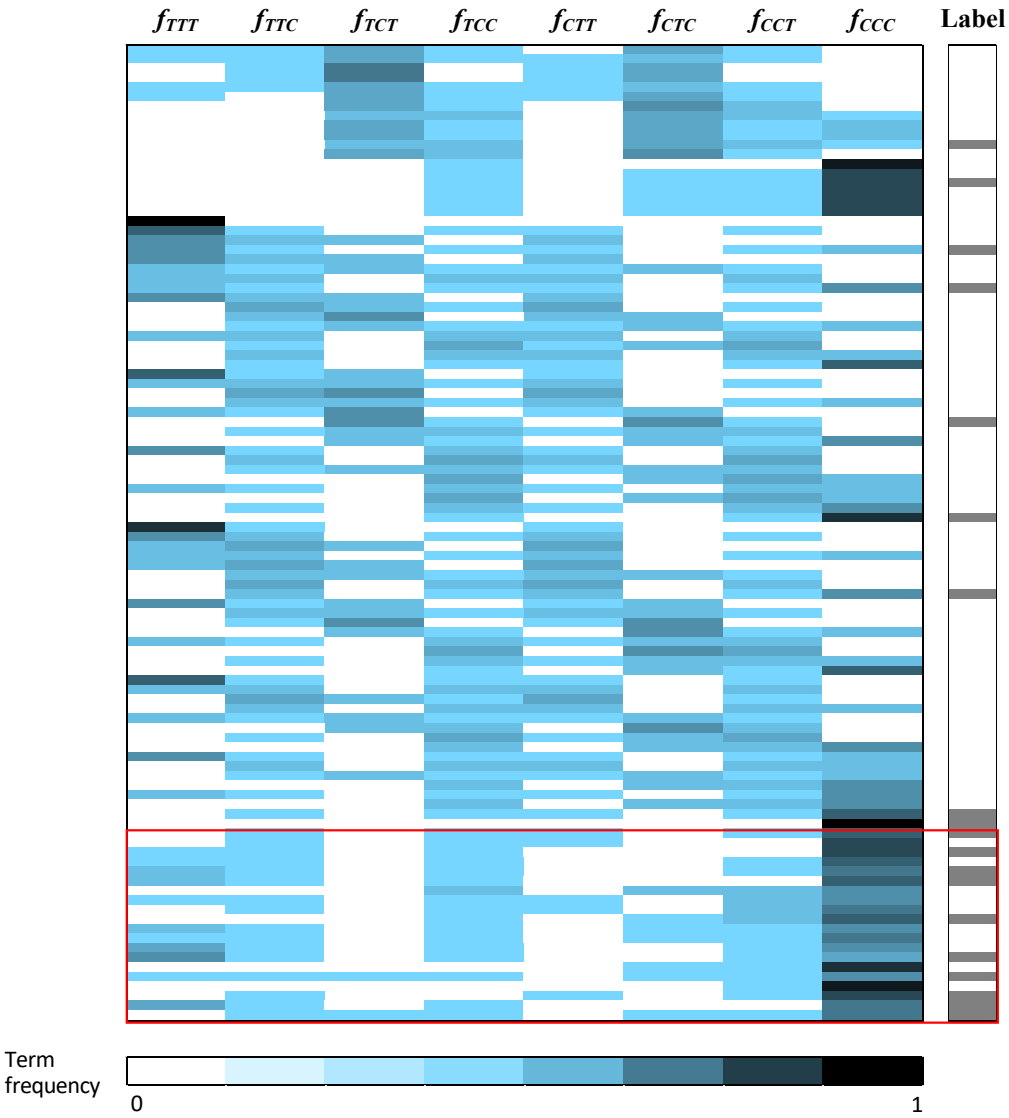

Figure 3.8. Heatmap analysis of relative enrichment of trigram terms in recognition sequences. The horizontal axis represents each of the eight 3-gram words. The vertical axis represents individual experimentally tested sequences. Those inside the red rectangle are new sequences discovered in this work. Grey in the rightmost column represents a label of 'Y', i.e., that of a recognition sequence. The heatmap itself displays the frequency of occurrence of that word in that particular sequence.

3.4. Conclusions

The DNA/SWCNT hybrid system comprises a vast set of sequence/(*n*,*m*) combinations. A small fraction of these form recognition pairs that allow separation of individual (*n*,*m*) SWCNT from a mixture. Our considerable knowledge about their structure and thermodynamics has not previously translated into an ability to predict recognition sequences. Here, we systematically applied machine learning techniques to predict recognition sequences. For simplicity and illustrative purposes, we restricted ourselves to a 12-mer sequences with a 2-letter alphabet (C & T). ML models were trained on available data, and re-trained twice based on new experimental data. We showed a remarkable increase in the frequency of recognition sequences from 10% in the original training set to >50% in the model-predicted sequence sets.

To design an improved model, detailed analyses were carried out. Performance was measured in terms of evaluation parameters (F^1–score) by cross-validation and prediction errors on the newly tested sets. Often model performance depends strongly on choice of sequence representation by input features. We chose a number of feature representation methods including *tfv, psv,* and mixed models. These methods have competing advantages when it comes to capturing information embedded in a set of sequences. When predicting new sequences to be tested experimentally, we chose on the basis of consensus of a number of methods, on the notion that the intersection of predictions made by different models would mitigate the limitations of our data set size and feature encoding schemes.

Among individual models, prediction performance of the *tfv* models was generally better than *psv*; trigram *tfv* models showed smaller prediction error. Based on these

analyses, we directed attention to ANN and LR using *tfv*. We also explored new input feature construction methods such as combined or segmented *tfv*, and motif-based features. We obtained highly encouraging models that showed an improved F^1–score of ~ 27% when compared to the best previous model. In general, the ANN algorithm in combination with trigram *tfv* showed the best performance.

As aids to model interpretation, we investigated the discovered motif and feature saliency. We found that the top ranked motifs found with no motif-length limitation contained at least 8 bases. This result may suggest that at least 8 bases are needed to tightly wrap around SWCNT to exhibit a specific binding characteristic. According to the saliency analysis, the sequence at the ends contributes more to the classification, consistent with experiment.[26,27]

One may question the representation of recognition DNA sequence prediction as a binary classification problem, since each pairs with a different SWCNT. Success despite this assumption, indicates that recognition sequences may share common features although individual recognition sequences recognize a particular (*n*,*m*) species. Although our model is promising, we believe that there is considerable room for improvement. For example, recognition sequences differ in terms of selectivity, represented by purification yield. Some special sequences are known to be capable of separating enantiomers[28]. Yet, in the current model, these are all assigned the same label/score.

These considerations suggest future research in two major directions: one is to develop resolution-based multi-level classification. For example, multi-level classification would allow us to capture improvement in the model between the first and second rounds

of experiment by allowing cases labeled as N* to be accounted for as their own level of classification. The other is the study of methods for the interpretability of ML models such as saliency analysis. More broadly, bio/nano hybrid materials made of inorganic nanostructures and sequence-defined polymers such as DNA and peptides represent an emerging class of materials that have many promising applications. Design of this new class of material inevitably has to solve the challenging problem of efficient exploration of a vast sequence space. The learnings we obtained in this work should provide some insight to the more general sequence selection problem.

3.5. Acknowledgement

This work was performed with the collaboration of Dr. Ming Zheng at National Institute of Standards and Technology in Gaithersburg, MD. I would also like to acknowledge Prof. Anand Jagota and Dr. Arun Jagota with helpful discussion and suggestions during this work. Additionally, this work was supported by a Dean's Fellowship. This work is part of the NHI Initiative at Lehigh University.

3.6. References

1. Gupta, R. *et al.* Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics* **11,** S65 (2010).
2. Zhao, X.-M., Wang, Y., Chen, L. & Aihara, K. Gene function prediction using labeled and unlabeled data. *BMC Bioinformatics* **9,** 57 (2008).
3. Clare, A. & King, R. D. Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics* **19,** ii42-ii49 (2003).
4. Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12,** 1007–1017 (2003).
5. Stiffler, M. A. *et al.* PDZ Domain Binding Selectivity Is Optimized Across the Mouse Proteome. *Science (80-.).* **317,** 364 LP-369 (2007).
6. 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.* **26,** 5839–5845 (2014).
7. Baughman, R. H., Zakhidov, A. A. & de Heer, W. A. Carbon Nanotubes--the Route Toward Applications. *Science (80-.).* **297,** (2002).
8. Eatemadi, A. *et al.* Carbon nanotubes: properties, synthesis, purification, and medical applications. *Nanoscale Res. Lett.* **9,** 393 (2014).
9. Yang, N., Chen, X., Ren, T., Zhang, P. & Yang, D. Carbon nanotube based biosensors. *Sensors Actuators B Chem.* **207,** 690–715 (2015).
10. Nish, A., Hwang, J.-Y., Doig, J. & Nicholas, R. J. Highly selective dispersion of single-walled carbon nanotubes using aromatic polymers. *Nat. Nanotechnol.* **2,** 640–646 (2007).
11. Liu, H., Nishide, D., Tanaka, T. & Kataura, H. Large-scale single-chirality separation of single-wall carbon nanotubes by simple gel chromatography. *Nat. Commun.* **2,** 309 (2011).
12. 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.* **1,** 60–65 (2006).
13. Ao, G., Khripin, C. Y. & Zheng, M. DNA-controlled partition of carbon nanotubes in polymer aqueous two-phase systems. *J. Am. Chem. Soc.* **136,** 10383–10392 (2014).
14. Tu, X., Manohar, S., Jagota, A. & Zheng, M. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature* **460,** 250–253 (2009).
15. Zheng, M. Sorting Carbon Nanotubes. *Top. Curr. Chem.* **375,** 13 (2017).
16. Zhang, J. *et al.* Single Molecule Detection of Nitric Oxide Enabled by d(AT) $_{15}$ DNA Adsorbed to Near Infrared Fluorescent Single-Walled Carbon Nanotubes. *J. Am. Chem. Soc.* **133,** 567–581 (2011).
17. Shi, J. *et al.* Microbiosensors based on DNA modified single-walled carbon nanotube and Pt black nanocomposites. *Analyst* **136,** 4916 (2011).
18. Landry, M. P. *et al.* Single-molecule detection of protein efflux from microorganisms using fluorescent single-walled carbon nanotube sensor arrays. *Nat. Nanotechnol.* **12,** 368–377 (2017).

19. Roxbury, D., Jagota, A. & Mittal, J. Structural characteristics of oligomeric DNA strands adsorbed onto single-walled carbon nanotubes. *J. Phys. Chem. B* **117,** 132–140 (2013).
20. Johnson, R. R., A. T. Charlie Johnson & Klein, M. L. Probing the Structure of DNA−Carbon Nanotube Hybrids with Molecular Dynamics. (2007). doi:10.1021/NL071909J
21. Johnson, R. R., Kohlmeyer, A., Johnson, A. T. C. & Klein, M. L. Free Energy Landscape of a DNA−Carbon Nanotube Hybrid Using Replica Exchange Molecular Dynamics. *Nano Lett.* **9,** 537–541 (2009).
22. Roxbury, D., Manohar, S. & Jagota, A. Molecular simulation of DNA β-sheet and β-barrel structures on graphite and carbon nanotubes. *J. Phys. Chem. C* **114,** 13267–13276 (2010).
23. Shankar, A., Zheng, M. & Jagota, A. Energetic Basis of Single-Wall Carbon Nanotube Enantiomer Recognition by Single-Stranded DNA. *J. Phys. Chem. C* **121,** 17479–17487 (2017).
24. Manohar, S. *et al.* Peeling single-stranded DNA from graphite surface to determine oligonucleotide binding energy by force spectroscopy. *Nano Lett.* **8,** 4365–72 (2008).
25. Iliafar, S., Mittal, J., Vezenov, D. & Jagota, A. Interaction of Single-Stranded DNA with Curved Carbon Nanotube Is Much Stronger Than with Flat Graphite. *J. Am. Chem. Soc.* **136,** 12947–12957 (2014).
26. Roxbury, D., Tu, X., Zheng, M. & Jagota, A. Recognition ability of DNA for carbon nanotubes correlates with their binding affinity. *Langmuir* **27,** 8282–8293 (2011).
27. Shankar, A., Mittal, J. & Jagota, A. Binding between DNA and carbon nanotubes strongly depends upon sequence and chirality. *Langmuir* **30,** 3176–3183 (2014).
28. Ao, G., Streit, J. K., Fagan, J. A. & Zheng, M. Differentiating Left- and Right-Handed Carbon Nanotubes by DNA. *J. Am. Chem. Soc.* **138,** 16677–16685 (2016).
29. Yang, Y., Shankar, A., Aryaksama, T., Zheng, M. & Jagota, A. Quantification of DNA/SWCNT Solvation Differences by Aqueous Two-Phase Separation. *Langmuir* **34,** 1834–1843 (2018).
30. Zheng, Y., Bachilo, S. M. & Weisman, R. B. Quenching of Single-Walled Carbon Nanotube Fluorescence by Dissolved Oxygen Reveals Selective Single-Stranded DNA Affinities. *J. Phys. Chem. Lett.* **8,** 1952–1955 (2017).
31. Tu, X., Manohar, S., Jagota, A. & Zheng, M. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature* **460,** 250–253 (2009).
32. Vens, C., Rosso, M.-N. & Danchin, E. G. J. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **27,** 1231–1238 (2011).
33. Witten, I. H., Frank, E. & Hall, M. a. *Data Mining: Practical Machine Learning Tools and Techniques*. *Annals of Physics* **54,** (2011).
34. Cox, D. R. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* **20,** 215–242 (1958).
35. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **2,** 183–197 (1991).

36. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *npj Comput. Mater.* **4,** 25 (2018).
37. Ao, G. & Zheng, M. in *Current Protocols in Chemical Biology* **7,** 43–51 (John Wiley & Sons, Inc., 2015).
38. Srinivasan, S. M., Vural, S., King, B. R. & Guda, C. Mining for class-specific motifs in protein sequence classification. *BMC Bioinformatics* **14,** 96 (2013).
39. Vinga, S. & Almeida, J. Alignment-free sequence comparison--a review. *Bioinformatics* **19,** 513–523 (2003).
40. Cessie, S. Le & Houwelingen, J. C. Van. Ridge Estimators in Logistic Regression. *Appl. Stat.* **41,** 191 (1992).
41. Platt, J. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. (1998).
42. Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F. & Leyton-Brown, K. *Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA*. *Journal of Machine Learning Research* **17,** (2016).
43. Aiello, S., Eckstrand, E., Fu, A., Landry, M., and Aboyoun, P. *Machine Learning with R and H2O*. (http://h2o.ai/resources/, 2018).
44. W. Ruck, D., Rogers, S. & Kabrisky, M. *Feature Selection Using a Multilayer Perceptron*. *Journal of Neural Network Computing* **2,** (1993).

3.7. Appendix

3.7.1. Scoring Factor, S

We have introduced a new scoring factor, *S*:

$$S = \frac{TP-FN}{p} + \frac{TN-FP}{1-p} \tag{3.2}$$

where *p* is the frequency (estimate of probability) of the positive instances (i.e., recognition sequence) in the original data set. This is based on the consideration that TP and FN are drawn from positive instances with probability *p*, and TN and FP from negative instances with probability 1–*p*, and a correct (incorrect) classification should gain ±1 point, respectively. The scaling by probability is important, because it gives more credit (punishment) to the correct (incorrect) prediction of less frequent and more interesting events. Here, the probability p is approximated by $\frac{TP+FN}{N_T}$ where N_T is the number of total instances ($N_T = TP + FP + FN + TN$). Then, eq (3.2) can be rewritten in terms of recall and false positive rate ($FPR = \frac{FP}{FP+TN}$) as $S = 2N_T(R - FPR)$. Finally, the factor S is normalized as eq (S2), so that it can be compared directly with the F^1 score.

$$S' = \frac{S/2N_T+1}{2} = \frac{R-FPR+1}{2} \tag{3.3}$$

3.7.2. Saliency Analysis

It is well known that saliency measures in feedforward ANN can be used to assess a feature's relative importance.[*†‡] The higher a saliency value, the stronger the relationship between input feature and output.

Derivative-based saliency measures were developed by Ruck et. al.[*] in which the sensitivity of the network output to its input feature can be expressed by the derivative of an output with respect to a given input. When the sigmoid function is used for a feedforward single hidden layer ANN, the derivative is

$$\frac{\partial z_i}{\partial x_k} = z_i(1 - z_i) \sum_{m=1}^{N_h} w_{mi}^2 x_m^1 (1 - x_m^1)\, w_{km}^1 \tag{3.4}$$

where z_i is the output node i in the output layer, x_k is the input feature k, x_m^1 is the output of node m in hidden layer, w_{km}^1 is the weight connecting node k in the input layer to node m in the hidden layer, w_{mi}^2 is the weight connecting node m in the hidden layer to node i in the output layer, and N_h is the number of nodes in the hidden layer.

The integrated saliency is defined by

$$\Lambda_k = \sum_{\mathbf{x} \in \mathbf{S}} \sum_i \sum_{x_k \in D_k} \frac{\partial z_i}{\partial x_k}(\mathbf{x}, \mathbf{w}) \tag{3.5a}$$

where **x** is the n-dimensional input features, **S** is the set of p training vectors, **w** is the weights in the network, and D_k is a set of R points for input x_k which will be sampled.[*]

---

[*] W. Ruck, D., Rogers, S. & Kabrisky, M. *Feature Selection Using a Multilayer Perceptron*. *Journal of Neural Network Computing* **2**, (1993).

[†] Belue, L. M. & Bauer, K. W. Determining input features for multilayer perceptrons. *Neurocomputing* **7**, 111–121 (1995).

[‡] Steppe, J. M. & Bauer, K. W. Feature saliency measures. *Comput. Math. with Appl.* **33**, 109–126 (1997).

Here, $\frac{\partial z_i}{\partial x_k}(\mathbf{x}, \mathbf{w})$ can be obtained by eq (3.4). To compare the errors in different models, eq (3.5a) is normalized by the total number of the derivatives,

$$\overline{\Lambda_k} = \frac{\Lambda_k}{npR} \tag{3.5b}$$

The feedforward single layer ANN was trained at least 101 times with randomly initialized weights, from which the average of normalized saliency $\hat{\Lambda}_k$ was obtained.

The averaged saliency of top two ANN models for each feature are shown in Figure 3.7, 3.9, and 3.11. (Positive saliency indicates that the input feature is related to the recognition class; negative saliency indicates it is related to the non-recognition class.)

Figure 3.11 shows the normalized saliency of ANN with motif-based features models. Generally, saliencies of the recognition/non-recognition motifs are positive/negative. Also, longer motifs tend to show higher absolute values of saliency, suggesting that the longer motifs have more significant information. In general, the non-recognition motifs show low saliency values compared to the recognition motifs even though the conditional probabilities of non-recognition motifs are much higher during motif selection (Figure 3.5). We suggest that there are more of the non-recognition motifs than the recognition motifs in entire motif space; or, the recognition motifs are highly relevant in determining the recognition sequences as compared to non-recognition motifs. This can explain why few recognition sequences are found in random searches of DNA library.

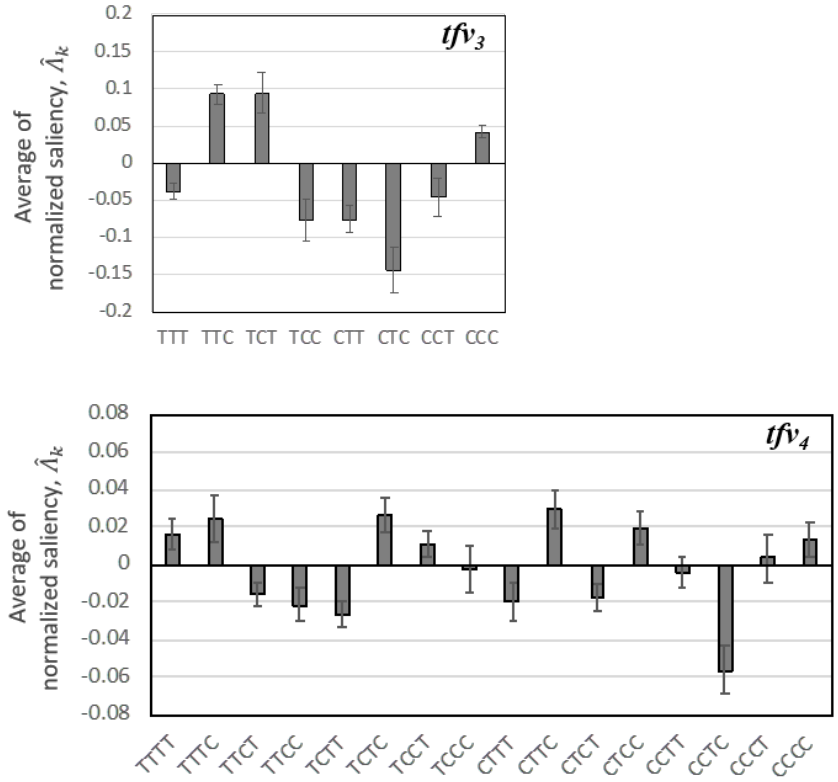

Figure 3.9. Average of normalized saliency of the feedforward ANN models with trigram *tfv* (top) and 4-gram *tfv* (bottom) calculated by eq. (3.5b). Error bar represents standard deviation.

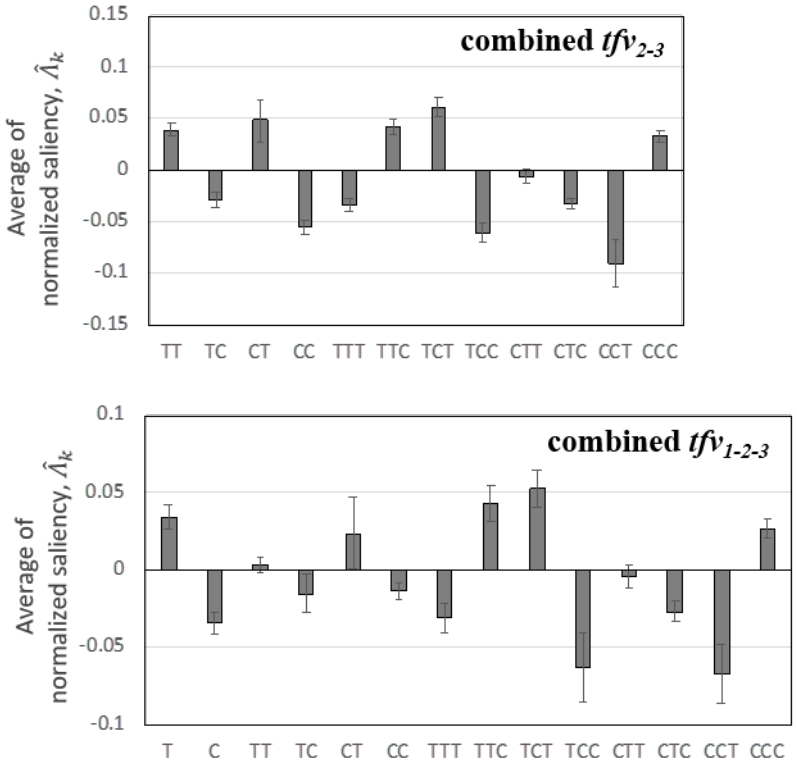

Figure 3.10. Average of normalized saliency of the feedforward ANN models with combined tfv_{2-3} (top) and tfv_{1-2-3} (bottom) calculated by eq. (3.5b). Error bar represents standard deviation.

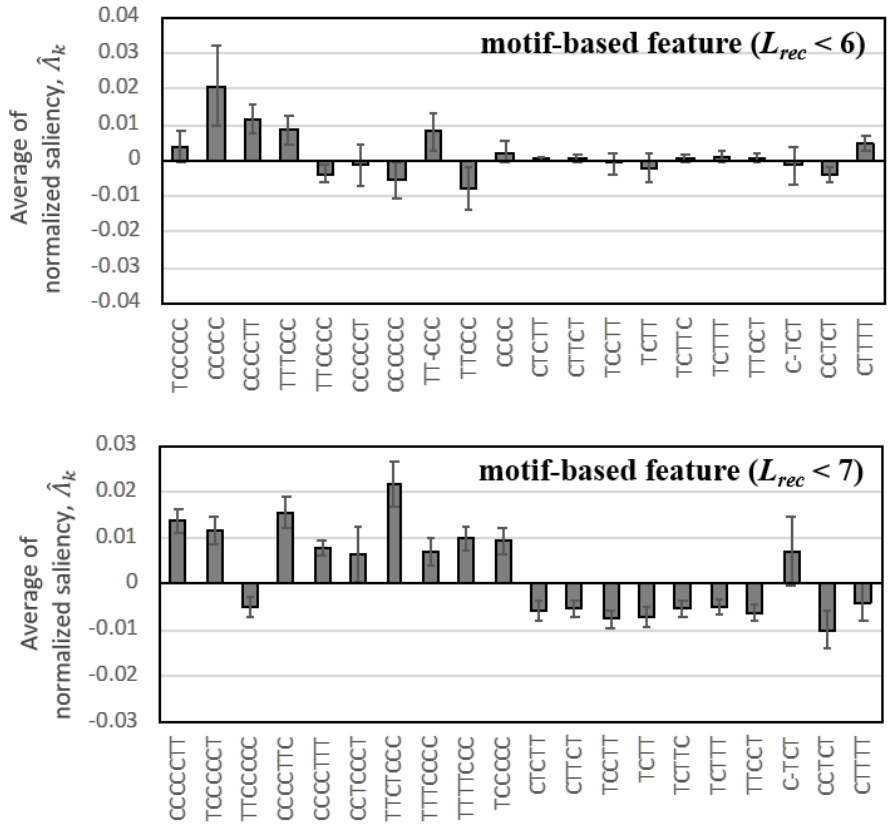

Figure 3.11. Average of normalized saliency of the feedforward ANN models with motif-based feature (L_{rec} < 6) (top) and (L_{rec} < 7) (bottom) calculated by eq. (3.5b). Error bar represents standard deviation. The first ten sequences are recognition motifs while the second ten are non-recognition motifs.

3.7.3. Probability of Finding Recognition Sequences

Based on our initial training set, the population of recognition sequence is ~ 0.1 (9 recognition sequences and 73 non-recognition sequences). Suppose that the population is a normal distribution with standard deviation, σ: $N(\mu, \sigma)$ where μ is 0.1 and σ is assumed to be 0.1 (black dotted line in Figure 3.12). Note that the standard deviation might be much smaller than 0.1 because we set a high standard when defining the recognition

sequence, however, to show the extreme case, we set the standard deviation to a relatively large value (0.1). Similarly, the population of the recognition sequence in the predicted set is 0.5 (5 recognition sequences and 5 non-recognition sequence for each set predicted by initial and 1st retrained model), and the population can be presented as $N(0.5, 0.1)$, shown as red solid line in Figure 3.12.

Suppose we randomly sample 10 test sequences in the entire sequence space for 12 mer C/T library (2^{12}). According to the central limit theorem, the mean is the same and the standard deviation of the sampling distribution is the value obtained by dividing the standard deviation of the population, σ, by the sample size, n, so the sampling distribution is $\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ (blue dashed line in Figure 3.12). As shown in Figure 3.12, the random sampling distribution is much narrower than that of the training set and is far from the population in the predicted set by our models. This demonstrates that it is very unlikely that 50% success rate could be achieved by random sampling.

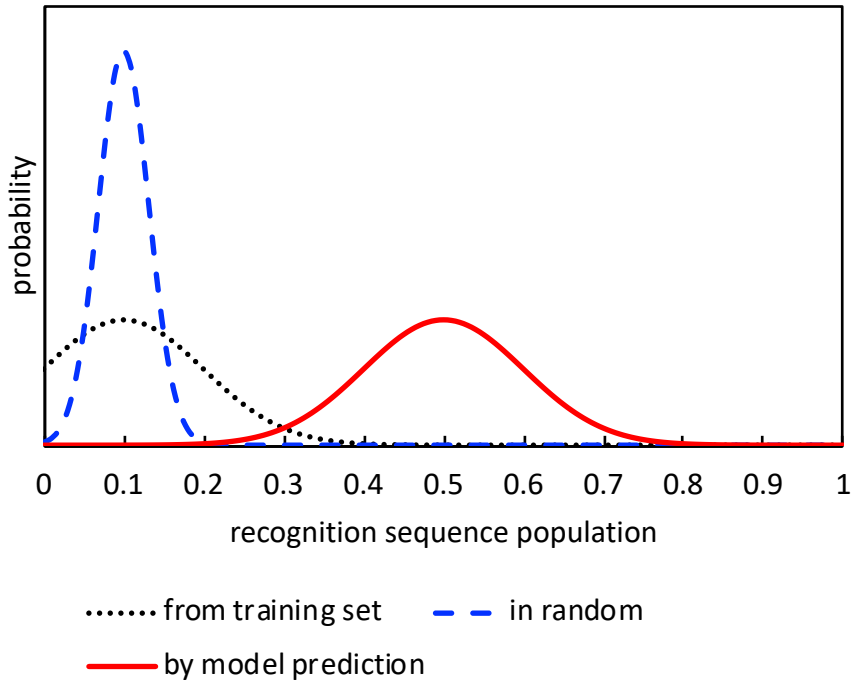

Figure 3.12. Probability density of the recognition sequence population in training set (black dotted line) and when ten sequences are randomly drawn from the population of training set (blue dashed line), and the population in predicted set (red solid line)

3.7.4. MERCI* Command Example

```
perl MERCI.pl -p TC12_pos.txt
-n TC12_neg.txt
-g 1
-gl 12
-k ALL
-o TC12_rec_motif_l7
-fp 1
-fn 83
-l 7
```

Here, '*-p filename*' and '*-n filename*' set the file with the positive (recognition) and negative (non-recognition) sequences. The file should be in Fasta format. '*-k value*' sets the number of motifs requested. '*ALL*' means all the possible motifs requested. '*-g value*' and '*-gl value*' set the maximal number of gaps and maximal gap length. '*-o file*' set the output file where motifs will be saved. '*-fp value*' and '*-fn value*' set the minimal frequency for the positive sequences f_P and the maximal frequency for the negative sequences f_N, respectively. The frequency should be an absolute number between 0 and the total number of positive (negative) sequences. We set $f_P = 1$ and $f_N = 83$ (the number of non-recognition sequences in the training set) when we discover recognition motifs, so that all the possible recognition motif can be searched. When we discover non-recognition motifs, we use $f_P = 10$ and $f_N = 10$. This is because it is seen that we have more non-recognition motifs based on the portion on non-recognition sequences in the training set, so limited f_P and f_N could

* Vens, C., Rosso, M.-N. & Danchin, E. G. J. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **27**, 1231–1238 (2011).

provide more robust motifs. '*-l*' sets the maximal length of motif. In this study, we vary the maximal length of motif, *l*, from 5 to 7 for recognition sequences and set to 6 for non-recognition sequences.

3.7.5. WEKA* Command Example - Artificial Neural Network

```
java weka.classifiers.functions.MultilayerPerceptron
-L 0.3
-M 0.2
-N 500
-E 20
-H 11
-S 0
-t ./trainingSet/TC_tfv_3gram.arff
-threshold-file ./WEKAresults/roc/tfv3/TC_tfv_3gram_seed0.arff
> ./WEKAresults/tfv3/ TC_tfv_3gram_seed0
```

First, specify the learning algorithm ('weka.classifiers.functions.MultilayerPerceptron' for ANN / 'weka.classifiers.functions.Logistic' for LR / 'weka.classifiers.functions.SMO' for SVM). '*-L value*', '*-M value*', '*-N value*', '*-E value*' set the learning rate, momentum, the number of epochs, and the number of consecutive increases of error allowed for validation testing, respectively. '*-H value*' sets the number of hidden nodes in a single hidden layer that was optimized manually. '*-S value*' sets the value used to seed the random number

* Witten, I. H., Frank, E. & Hall, M. a. *Data Mining: Practical Machine Learning Tools and Techniques*. *Annals of Physics* **54**, (2011)

generator which was varied to get randomly initialized models required for saliency analysis. '*-t filename*' sets the training file. '*-threshold-file filename*' set the file to save the threshold data required to obtain ROC curve. '> *filename*' in the last sentence saves the model parameters and cross-validation results as the filename.

The models using LR and SVM can be implemented in a similar way.

3.7.6. Automated machine learning packages (Auto-WEKA [*] and h2o [†] AutoML)

Even though we presented results based on our choice of three algorithms and manual optimization of hyperparameters in the main manuscript, we also tried automated ML packages to explore all models and adjust the hyperparameters. To be consistent with the models used throughout this work, we used the Auto-WEKA package. In addition, to compare with the WEKA package, we used the "h2o" AutoML package. Both packages return choices for algorithms and hyperparameters. We tested both packages with the trigram *tfv* of the training set that was used for the 2nd retrained model.

First, the Auto-WEKA searched more than 3,000 different models through the joint space of WEKA's learning algorithms and their hyperparameters to maximize F^1-score. Each model was evaluated by 10-fold cross-validation. The results show that the locally weighted learning (LWL) model using repeated incremental pruning to produce error

[*] Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F. & Leyton-Brown, K. *Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research* **17**, (2016).

[†] Aiello, S., Eckstrand, E., Fu, A., Landry, M., and Aboyoun, P. *Machine Learning with R and H2O*. (http://h2o.ai/resources/, 2018).

reduction (RIPPER) with reduced feature set by feature selection showed the best performance (F^1-score = 0.649). This value is comparable to the best values we obtained using manually optimized ANN models. This provides some extra assurance that our manual search has resulted in a near-optimal choice of model and hyperparameters. However, Auto-WEKA results could not be reproduced on multiple executions, nor even by using the optimized hyperparameter setting for the chosen "best" model. Given this finding and the lack of transparency, we much prefer to present results based on our choice of three algorithms and "manual" optimization.

Next, the h2o AutoML package was used to compare with the Auto-WEKA results. The 10-fold cross-validation results shows that the best model is ANN (F^1-score = 0.688). However, the ANN model has too many hidden nodes (200), which strongly implies it is overfitting, given the limited size of our dataset. (The default value of the number of hidden nodes is 200 and there is no option to set the hyperparameter range in the h2o AutoML package.)

Overall, we believe that our model in the main manuscript is well-optimized and more reliable than the models produced by the automated ML packages.

3.7.7. Link to Public Repository:

The following link is to a public repository where we provide a collection of scripts for translation from sequence data to features. We have also included details and typical example scripts for WEKA, MERCI, and code for saliency analysis.

https://bitbucket.org/jagotagrouplehigh/dna_swcnt_ml/

Chapter 4 : DNA-Wrapped Carbon Nanotubes via Methanol-Aided Replacement of Surfactants[*]

DNA/SWCNT hybrids have attracted recent interest due to their ability for SWCNT separation and their use as promising agents in biosensing and bioimaging applications. In order to perform such SWCNT separation and application development, special DNA sequences are needed that have an ability to recognize specific chiral SWCNT, called recognition sequences. So far, sequence screening has relied on various sorting methods which are costly and time-consuming. Recently, a new simple, rapid way to produce DNA/SWCNT hybrids was reported using replacement of strong surfactant on SWCNT by DNA, aided by methanol. However, little is known about the nature of the exchange mechanism. Here, we investigated the kinetics of the replacement process aided by methanol. A mechanistic model was proposed to analyze and extract the activation energy of the exchange process. We found that some recognition sequences have significantly different activation energy for different species of SWCNT, while some sequences do not. The results suggest that the replacement process does not always produce the same structure of DNA/SWCNT that was produced by traditional direct sonication. In addition, for some sequences, it takes too long to reach the equilibrium wrapping configuration, suggesting that the replacement method might not be suitable to make DNA/SWCNT

[*] This work has been performed in collaboration with Dr. Arjun Sharma of Lehigh University, Guillaume Noetinger of ESPCI (Paris, France), and Dr. Ming Zheng of National Institute of Standards and Technology.

hybrids for these sequences. Nevertheless, our results suggest that the methanol-aided replacement process not only can reduce preparation time for DNA/SWCNT dispersion, but also could be useful as a promising low-cost, rapid way to identify recognition sequences.

4.1 Introduction

Single-wall carbon nanotubes (SWCNTs) are tubular nanostructures obtained conceptually by rolling a two-dimensional graphene sheet.[1] The structure of a given SWCNT is uniquely defined by a chiral vector that specifies how the graphene sheet is rolled into a tube. The chiral vector is typically expressed by chiral indices (n,m), representing the number of steps taken along each of two unit cell vectors.[1] Since the electrical and optical properties of SWCNTs are highly dependent on their chirality, its control is significant in many applications.[2,3] Previous studies have reported that certain special short single-stranded DNA (ssDNA) sequences have recognition ability toward partner SWCNT species.[4,5] Furthermore, ssDNA-wrapped SWCNT (DNA/SWCNT) have attracted considerable interest not only for their ability to be dispersed and sorted in aqueous solution, but also for the sensitivity of their optical properties to molecular analytes, which enables their use as promising agents in biosensing and bioimaging applications.[6–9]

In order to utilize DNA/SWCNT in applications, it is necessary to build a well-identified DNA/SWCNT library. Since the DNA sequence library is practically infinite in size, a search strategy is required. Significant effort has been expended to discover recognition sequences – SWCNT pairs, and significant progress has been made in understanding the underlying structure and thermodynamics of these hybrids.[10–17] Furthermore, a recent study has reported remarkable improvement in success rate of finding recognition sequences with a new approach using machine learning techniques.[18] Nevertheless, the pace of the process is still severely limited by experimental generation of

DNA/SWCNT samples. In particular, preparation of DNA/SWCNT via direct sonication of DNA and SWCNT mixtures[19] followed by a centrifugation for removing impurities is highly laborious and time-consuming. Recent studies have established a new technique for preparing DNA/SWCNTs by replacing surfactants in methanol/water solution (Figure 4.1).[20,21] It sped up the preparation with high SWCNT recovery by carrying out the sonication and centrifugation once with a surfactant followed by the replacement of the surfactant by any chosen DNA sequence. However, the nature of exchange mechanism is not fully understood at this time. For example, the DNA/SWCNT hybrids prepared via direct sonication are likely to form equilibrium structures unless the constituents are damaged. However, it is not clear whether the methanol aided surfactant replacement produces equilibrium structures immediately or over a period of time. Furthermore, the final structure of DNA/SWCNT can be dependent on the reaction conditions such as the concentration of methanol or DNA and temperature, which is directly related to the reaction rate.

In this study, we investigated the kinetics of the replacement process of several DNA sequences and SWCNTs. We used a surfactant molecule, sodium deoxycholate (SDC), for initial nanotube dispersion, then replaced it by DNA, aided by the addition of MeOH. By monitoring the kinetics using optical spectroscopy, our aim was to determine the relative binding characteristics for ssDNA on SWCNTs. A mechanistic model was developed for desorption of SDC and adsorption of DNA on the nanotube surface. The activation energy for the replacement process was extracted to quantitatively compare the binding characteristics for each DNA on SWCNT. We found that the replacement process

is highly sequence- and chirality-dependent. This finding shows that DNA/SWCNT preparation aided by MeOH could be a low-cost and rapid way to identify recognition/resolving sequences for sorting and sensing applications.

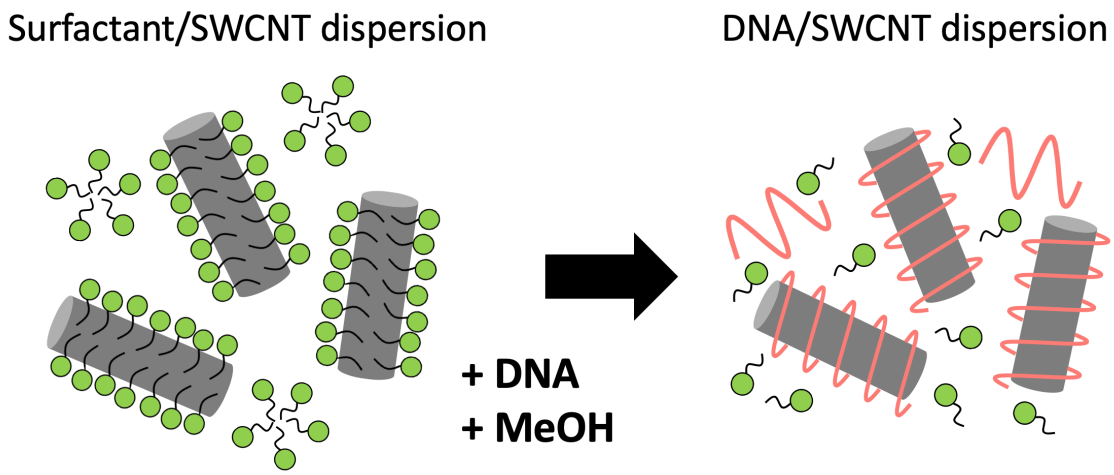

Figure 4.1. Schematic illustration for the methanol-aided exchange process of surfactant on the surface of SWCNTs by DNA

4.2 Materials and Methods

4.2.1 Materials

Cobalt−molybdenum catalyst (CoMoCAT, SG65i grade) SWCNT was purchased from Southwest Nanotechnologies. Single-stranded DNA (ssDNA) was obtained from Integrated DNA Technologies (IDT). Sodium deoxycholate (SDC) (BioXtra, >98%), and methanol (MeOH) were acquired from Sigma-Aldrich and used without modification. The DNA sequences were chosen were among those reported as recognition sequences. Specifically, $(TAT)_4$, $(TTA)_4TT$, and $(CCA)_{10}$ are recognition sequences for (6,5), (8,3),

and (9,1), respectively; they have also been studied in the exchange reaction of DNA/SWCNT by sodium dodecylbenzenesulfonate (SDBS). Other sequences, $(TTA)_2(ATT)_2$ and $(TG)_2T_4(GT)_2$, were chosen as the 12mer recognition sequences for (8,3) and (9,1), respectively. $(CCA)_4$ and $(TAT)_{10}$ were additionally examined to investigate the effect of sequence length.

4.2.2 Sample Preparation*

The CoMoCAT SWCNTs suspended in a 10 g/L of SDC solution were sonicated using tip sonicator for 1 h at a power of 1 W/mL. The dispersion was then centrifuged for 1 h at 1885 rad/s (Beckman J-2 centrifuge, JA-20 rotor) to remove SWCNT aggregates and residual impurities, and the supernatant was recovered.[21]

4.2.3 Exchange Procedure

SDC/SWCNT dispersion was diluted such that the final concentration of SWCNT was 5 mg/L. DNA was then added into the SDC/SWCNT dispersion in a 40:1 weight ratio of DNA to SWCNT. The DNA and SDC/SWCNT mixture was placed in a water bath at a chosen temperature. The volume of methanol was chosen to end up with various concentration (20 to 50 v/v %). The desired amount of methanol was held in a quartz cuvette at the same temperature. Once the temperature of the two solutions stabilized, the DNA and SDC/SWCNT mixture was added to the cuvette and mixed gently using a pipette.

* The sample of SDC dispersed SWCNT was prepared by Dr. Ming Zheng at National Institutes of Standards and Technology in Gaithersburg, MD.

4.2.4 Spectroscopy Measurements*

It has been reported that peak position in absorbance and intensity in fluorescence are very sensitive to change in local environment of SWCNT. This allows SWCNT to sense changes in the material wrapping it.[22] In this manner, the peak shift in absorbance and intensity changes in fluorescence can be used as a measure of DNA or SDC coverage, which is directly correlated to the progress of the replacement reaction.

The absorbance spectrum of the sample was monitored immediately after the addition of the mixture of DNA, MeOH, and SDC/SWCNT dispersion and then at 2 min intervals in the range from 200 to 1100 nm using a UV/vis/NIP spectrophotometer (Varian Cary 50) at various temperatures, 15°–35 °C. A prominent peak was observed at 983 nm and 990 nm for the (6,5)-SWCNT covered by SDC and ssDNA, respectively, indicative of the E11 bandgap transition. The absorbance spectra of the SDC/SWCNT dispersion in the presence of DNA and MeOH were measured at 5 °C and after overnight incubation at 40 °C as controls. In addition, the SDC/SWCNT dispersion with DNA in absence of MeOH was measured as another control.

Fluorescence spectra were recorded on a Fluorolog-3 fluorometer (HORIBA Jobin Yvon) in conjunction with a near-NIR sensitive PMT. The excitation source was a 450-W Xenon lamp. Emission wavelengths from 900-1200 nm with increments of 1 nm and slit widths of 8 nm were used. The excitation were performed at 569 nm and 572 nm for SDC/(6,5)-SWCNT and DNA/(6,5)-SWCNT hybrids, 668 nm and 672 nm for SDC/(8,3)-

* The fluorescence spectroscopic measurements were performed by Dr. Arjun Sharma of Lehigh University.

SWCNT and DNA/(8,3)-SWCNT, and 695 nm and 698 nm for SDC/(9,1)-SWCNT and DNA/(9,1)-SWCNT, respectively. Additionally, dark count correction factors were applied with integration time of 1 s. For controls without methanol and with or without free DNA, excitations were done at 569 nm. Note that the SWCNT mixture is rich in (6,5)-SWCNT, therefore, we measured (6,5)-SWCNT peak intensity in most cases, but (8,3) and (9,1)-SWCNT were also measured for their recognition sequences.

4.2.5 Spectra decomposition

The absorbance spectra were decomposed into individual Voigt profiles, convolution of Gaussian and Lorentzian profiles (see Supporting Information for details). Figure 4.2 shows a typical decomposition of the absorbance spectra into contributions from species in the mixture. We assume that the peak location can be used as a linear measure of how far the reaction has proceeded based on the fact that the peak shift is caused by the replacement of wrapping molecules[22]. (See also Supporting Information for a discussion of the correlation between peak shift and the reaction progress.) In fluorescence spectra, the intensity change is dominant as the replacement proceeds, so the intensity change of each SWCNT species was used as a measure of reaction progress.

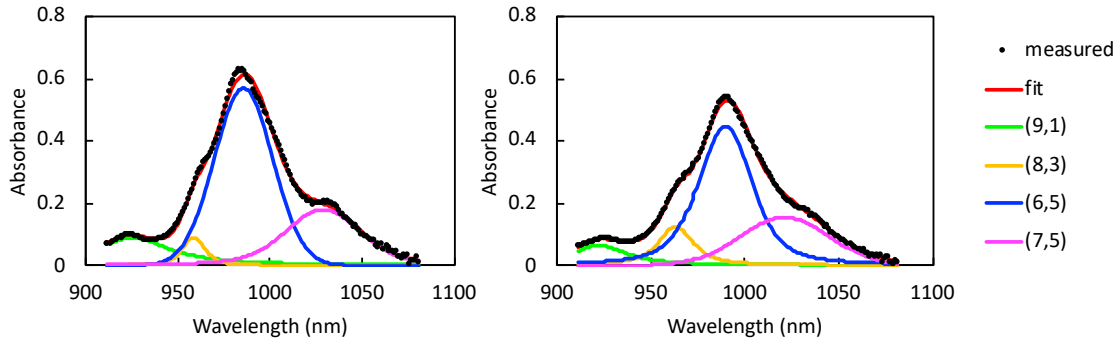

Figure 4.2. Decomposition of measured absorbance of SDC-coated SWCNTs and $(TAT)_4$-coated SWCNTs into contributions from four CNT species.

4.3 Results and discussions

We investigated the kinetics of SDC replacement by DNA on SWCNT facilitated by methanol in aqueous solution using optical peak shift and intensity changes. To find the optimal conditions that allow kinetic observations in a reasonable time (2 h) at mild temperatures, preliminary experiments were carried out by monitoring the kinetics of absorbance while varying the concentrations of MeOH and DNA. It was found that the overall reaction is very sensitive to changes of the MeOH concentration. For example, the reaction was completed quickly at room temperature (25 °C) at high concentrations of methanol (60%) while replacement hardly progressed at low concentration of methanol (20%) (Figure 4.3a). To examine the effect of the DNA concentration, we compared the absorbance spectra of the SDC/SWCNT with different concentration of $(TAT)_4$ incubated at room temperature for 30 min with the control sample which was incubated overnight at 40 °C. We assumed that the control sample was entirely converted to DNA-covered

SWCNT. Figure 4.4a shows that 30x mass excess of DNA to SWCNT is sufficient to reach the reaction completion. Based on these results, we chose for our study the experimental conditions of 40 v/v% MeOH concentration with 40x mass excess of DNA to SWCNT.

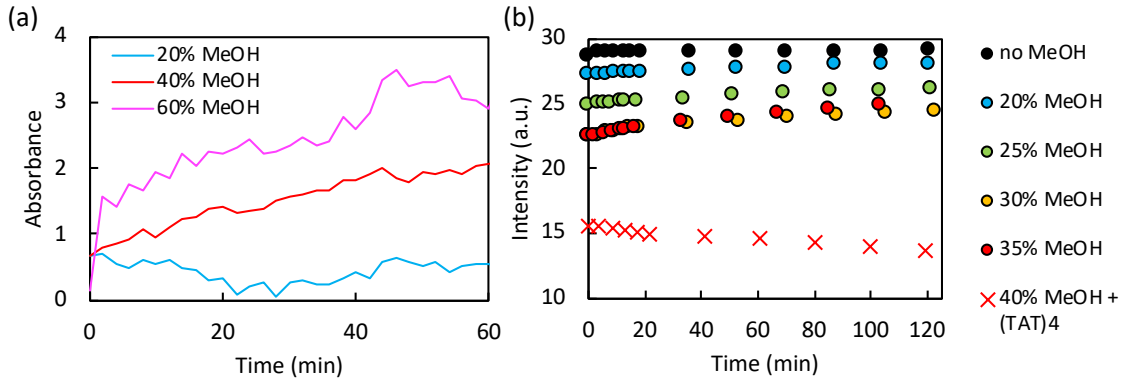

Figure 4.3. MeOH concentration-dependent kinetics measured, (a) by absorbance, and (b) by fluorescence. (a) The absorbance was measured by the peak shift for the $(TAT)_4$/(6,5)-SWCNT. The peak locations of (6,5)-SWCNT were obtained by decomposing the spectra (b) the fluorescence was measured by decay of the peak for (6,5)-SWCNT with the excitation wavelength of 569 nm. Note that the SDC/SWCNT sample with 40% MeOH sample immediately aggregates in absence of DNA.

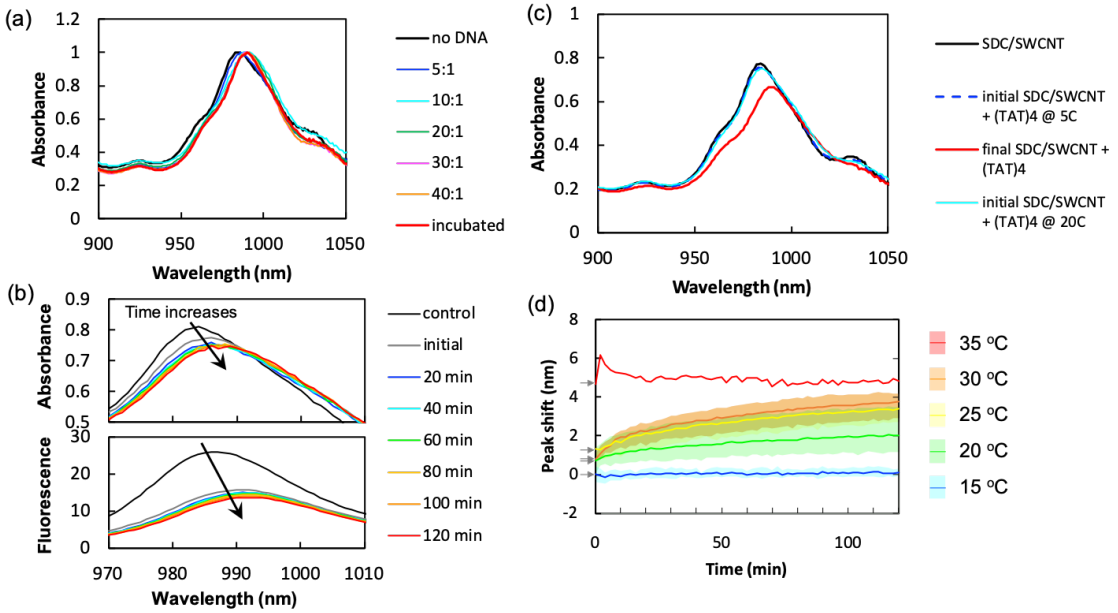

Figure 4.4. (a) SDC/SWCNT replaced by $(TAT)_4$ under incubation for 30 min at 40 °C with different DNA concentration (5:1 to 40:1 mass excess of DNA to SWCNT) and the control sample which we assume to be entirely DNA-covered SWCNT (i.e., incubated overnight at 40 °C). Note that all samples were prepared with 40 v/v% MeOH concentration. The 30x excess or higher concentration of DNA results in a complete peak shift. (b) Time evolution of absorbance and fluorescence spectra for SDC/SWCNT replaced by $(TAT)_4$, showing an instantaneous peak shift in absorbance and intensity changes in fluorescence, followed by slow changes. (c) Absorbance spectra for $(TAT)_4$/SWCNT species show peak shift as SDC is replaced by DNA on SWCNT. The spectra of the sample without DNA (black) and the immediate scan after DNA addition at 5 °C (blue) are seen to be identical. The spectrum of the immediate scan after DNA addition at 20 °C (cyan) shows a slight redshift. (d) Temperature-dependent kinetics in the absorbance measured by the peak shift for the (6,5)-SWCNT. The peak location of each species of SWCNT was obtained by decomposing the spectra. Data show an instantaneous initial increase (indicated by arrows along y-axis), followed by a further gradual increase.

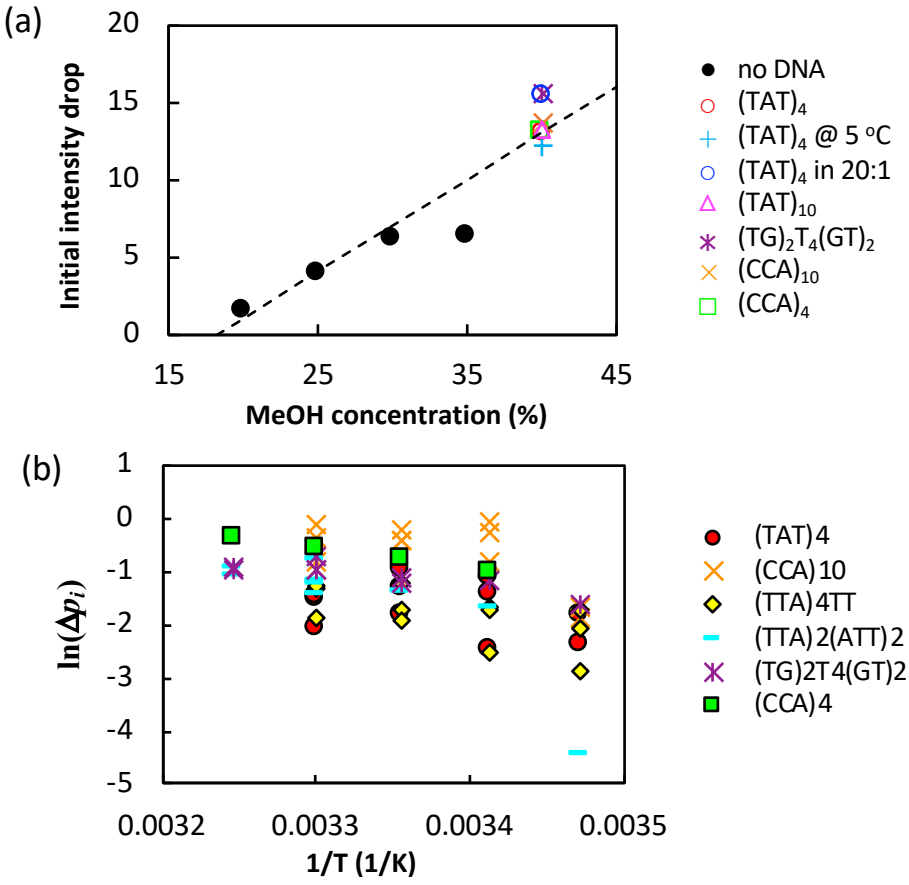

Figure 4.5. (a) The effect of MeOH concentration, DNA sequence, and the temperature on the initial fluorescence intensity drop for (6,5)-SWCNT. The results show that the initial intensity drop does not change with DNA sequence or temperature at fixed MeOH concentration, but depends linearly on the MeOH concentration. Note that all data were measured at room temperature except for the sample with temperature notation (+); all data were measure in a 40:1 weight ratio of DNA to SWCNT except for the sample with notation (o in blue). (b) The effect of the temperature on the initial peak shift in the absorbance.

The kinetics in absorbance and fluorescence were observed. Figure 4.4b shows time-evolution of absorbance and fluorescence spectra of SDC/SWCNT for the (6,5)-SWCNT peak. By comparing the spectrum taken immediately after addition to MeOH to that of the control sample prepared in the absence of MeOH, we see clearly that there is an

instantaneous peak shift in absorbance and change in fluorescence intensity. We also found subsequent time-dependent slow changes in absorbance peak location. Note that the fluorescence showed no major changes over time after the initial drop occurred.

Next, we measured the kinetics of SDC/(6,5)-SWCNT with varying MeOH concentration in the absence of DNA. As shown in Figure 4.3b, the fluorescence peak intensity values for (6,5)-SWCNT show an instantaneous drop that increases as methanol concentration increases. Interestingly, addition of methanol to a 40% total sample volume immediately aggregates SWCNTs in absence of DNA. It is worth noting that SDC/SWCNT in presence of DNA at even higher MeOH concentrations (60%) is stable.[21] This suggests the hypothesis that the initial drop is observed due to the (likely partial) desorption of SDC by MeOH. Figure 4.5a presents the initial drop as a function of the concentration of MeOH. Note that the initial drop was obtained by subtracting the intensity of a given sample at t = 0 from the intensity of the sample without MeOH. The data suggest that the initial intensity drop does not change with DNA sequence or temperature, but it depends linearly on the MeOH concentration, which is only associated with SDC, not DNA. Figure 4.6 also shows that the absorbance spectra of DNA/SWCNT does not change in the presence or absence of MeOH. Therefore, we suggest that the interaction between DNA and MeOH is negligible, and we propose that the initial drop can be used as a measure of SDC desorption.

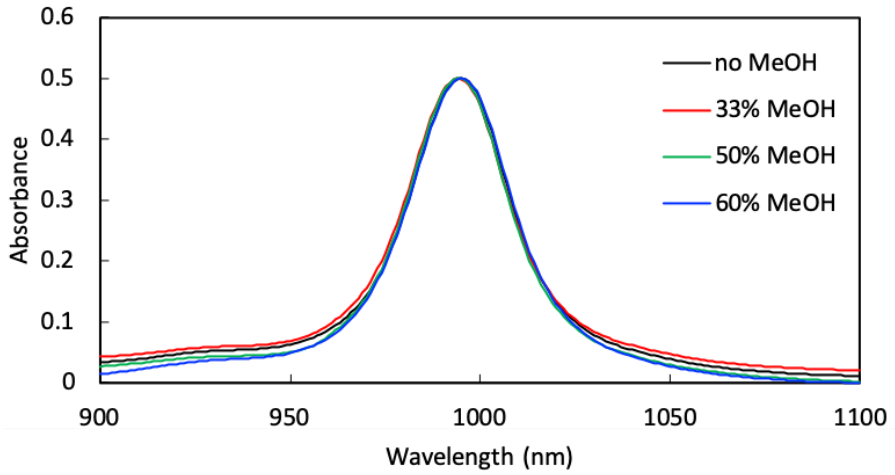

Figure 4.6. Absorbance spectra of TTA(TAT)$_2$ATT/(6,5)-SWCNT in various concentration of MeOH.

4.3.1 SDC replacement process by DNA

The SDC exchange process by different DNA sequences was monitored using the absorbance and fluorescence spectra for the (6,5) and (8,3)–SWCNT. Figure 4.4c depicts the absorbance spectra of the pure SDC-covered SWCNT, the initial mixture measured immediately at 5 °C and 20 °C, and the final mixture obtained after incubation overnight at 40 °C. The spectra of pure SDC-covered SWCNT and the initial mixture at 5 °C are seen to be identical while a slight difference is observed in the initial mixture at 20 °C. This indicates that the replacement reaction has not started at 5 °C; it starts at higher temperature. On the contrary, the uniform instantaneous intensity drop in fluorescence is observed regardless of the temperature (Figure 4.5a). Especially at high MeOH concentration (>40%), most SDCs are destabilized as soon as MeOH is added. This is supported by the previous experiments in which, in the absence of DNA, 40% MeOH shows instantaneous SWCNT aggregation. We interpret this difference to mean that the effect of MeOH on

partially destabilizing SDC adsorption is nearly instantaneous. The temperature-dependent kinetics were then measured by the peak shift of each chirality (Figure 4.4d for $(TAT)_4$/(6,5)-SWCNT, other sequences are shown in Appendix). We observed a rapid initial peak shift followed by a gradual increase over longer periods of time. We interpret the initial increase as a nearly instantaneous, partial and disordered, adsorption of DNA bases on regions of SWCNT exposed by desorption of SDC. We then interpret the gradual increase over the next two hours as conformational changes of DNA/SWCNT structures as they approach equilibrium.

Based on these experimental findings, we therefore propose a three-step model for the interaction of DNA with the SDC/SWCNT hybrid, sketched in Figure 4.7. The initial state consists of an aqueous solution containing SDC/SWCNT and dissolved excess DNA. Prior to addition of MeOH, this state is indefinitely stable. Upon addition of MeOH, the steps are as follows: (1) instantaneous partial SDC desorption by MeOH (desorption increases with MeOH concentration); (2) rapid partial adsorption of DNA to SWCNT surface exposed by SDC desorption; (3) time-dependent DNA rearrangement (and adsorption) which we will model as a simple pseudo-first-order reaction.

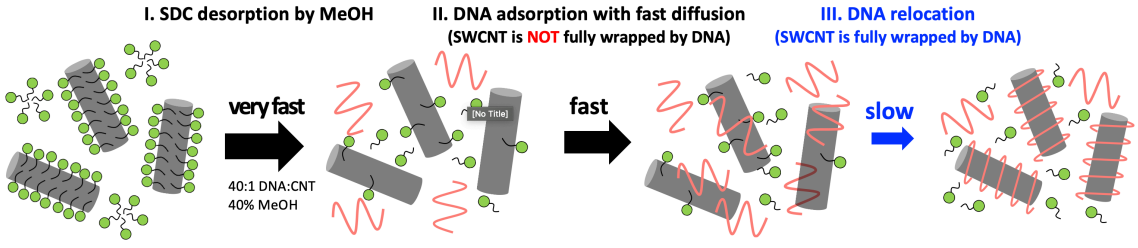

Figure 4.7. Schematic illustration of the mechanistic model for the exchange process of SDC on the surface of SWCNTs by DNA.

The first SDC desorption step is too fast for to see its temperature dependence, as depicted in Figure 4.5a. At the second step, we suggest that the rapid initial peak shift can be interpreted by a competitive adsorption isotherm. The initial peak shift (Δp_i) is calculated by the difference between the peak location at 5 °C and the peak location intermediately measured at a given temperature. Figure 4.5b reveals the linear dependence of $\ln(\Delta \overline{p_i})$ on $1/T$ where $\Delta \overline{p_i}$ was scaled to be in range of [0,1]. Furthermore, we investigated the sequence-dependence of the exchange process. We found a significant difference in initial DNA adsorption rate between T/A or T/G rich sequences versus C/A rich sequences. The C/A rich sequences, $(CCA)_4$ and $(CCA)_{10}$, are initially replaced faster than other T/A or T/G rich sequences, as presented in Figure 4.5b.

We hypothesize that most SDCs are desorbed and DNAs are partially adsorbed rapidly in stages 1 and 2. This is supported by the prior experimental data that the instantaneous SWCNT aggregation occurred in 40% MeOH in absence of DNA. The rate of stage 3, we propose, is limited not by DNA adsorption but by DNA rearrangement on SWCNTs.

4.3.2 Kinetics of DNA Rearrangement

The kinetics of the gradual rearrangement of DNA is analyzed as an activated process to extract the activation energy. The DNA rearrangement can be modeled as a simple first-order reaction where a disordered DNA on SWCNT (A) forms an ordered structure (B) as shown below:

$$A \xrightarrow{k} B \tag{4.1}$$

Then, the rate of production of *B*, DNA rearrangement on the SWCNT surface, can be expressed by

$$\frac{d[B]}{dt} = k[A] \tag{4.2}$$

By assuming the transition state theory and quasi-equilibrium states, the concentration of DNA on the SWCNT in ordered structure can be found:

$$[B] = [B]_0 + [A]_0\{1 - \exp(-kt)\} \tag{4.3}$$

where $[A]_0$ and $[B]_0$ are initial concentration of disordered and ordered structure of DNA/SWCNT. Note that the initial quick DNA adsorption (step 2) was observed, thus we cannot apply the initial condition such that $[B]_0 = 0$ at $t = 0$.

We previously assumed that the absorbance peak location (p) is linearly dependent on how far the reaction has proceeded which is directly related to the concentration of each SWCNT so that following the peak location is the same as following the reaction. Therefore, eq (4.3) can be represented as a function of peak shifts:

$$\Delta p = \Delta p_i - C\exp(-kt) \tag{4.4}$$

Here, Δp is calculated by subtracting the peak location of initial mixture at 5 °C from the peak location at a given temperature and time, Δp_i is the initial peak shift, and C is constant related to $[A]_0$.

Based on Eyring reaction rate theory, the overall reaction constant *k* is

$$k = k_2 \exp\left(-\frac{\Delta H^{\ddagger}}{k_B T} + \frac{\Delta S^{\ddagger}}{k_B}\right) \tag{4.5a}$$

$$\ln(k) = \ln(k_2) - \frac{\Delta H^{\ddagger}}{k_B T} + \frac{\Delta S^{\ddagger}}{k_B} \tag{4.5b}$$

Therefore, the activation enthalpy can be obtained from the slope of ln(k) versus $1/T$ in the Eyring plot. Details are described in the Appendix.

Figure 4.8 shows the Eyring plot for various DNA on (6,5)- and (8,3)-SWCNT and the slope for each DNA/SWCNT hybrid that represents the relative enthalpy ($\Delta H^{\ddagger}/k_B$). Previously, our group investigated the activation energies for the removal of DNA from a given SWCNT species by a surfactant molecule, SDBS, and reported that DNA sequence has noticeably higher activation energy on its recognition-partner species of SWCNT than on non-partner species[15]. Interestingly, it has been found that the activation energy of the reverse reaction in the presence of MeOH is chirality dependent for some DNA sequences. For example, $(TAT)_4$ and $(TTA)_2(ATT)_2$ have lower activation energy on their partner species of SWCNT (i.e., $(TAT)_4$/(6,5)–SWCNT and $(TTA)_2(ATT)_2$/(8,3)–SWCNT). Note that $(TG)_2T_4(GT)_2$, known as the (9,1) recognition sequence,[16] also showed considerable difference in the activation energy between (6,5) and (8,3)-SWCNT. Although other sequences such as $(CCA)_{10}$ and $(TTA)_4TT$ have shown the ability to sort SWCNT species,[4] they showed no differentiation between (6,5) and (8,3)-SWCNT. This suggests that, in the surfactant replacement route, not all DNA sequences are forming the same structure of DNA/SWCNT as those prepared via direct sonication.

In general, T/A rich sequences are seen to have lower activation enthalpy than T/G rich sequences. In addition, we found that $(CCA)_{10}$ shows similar activation energy to that of T/A rich ~12-mer sequences even though $(CCA)_{10}$ is much longer.

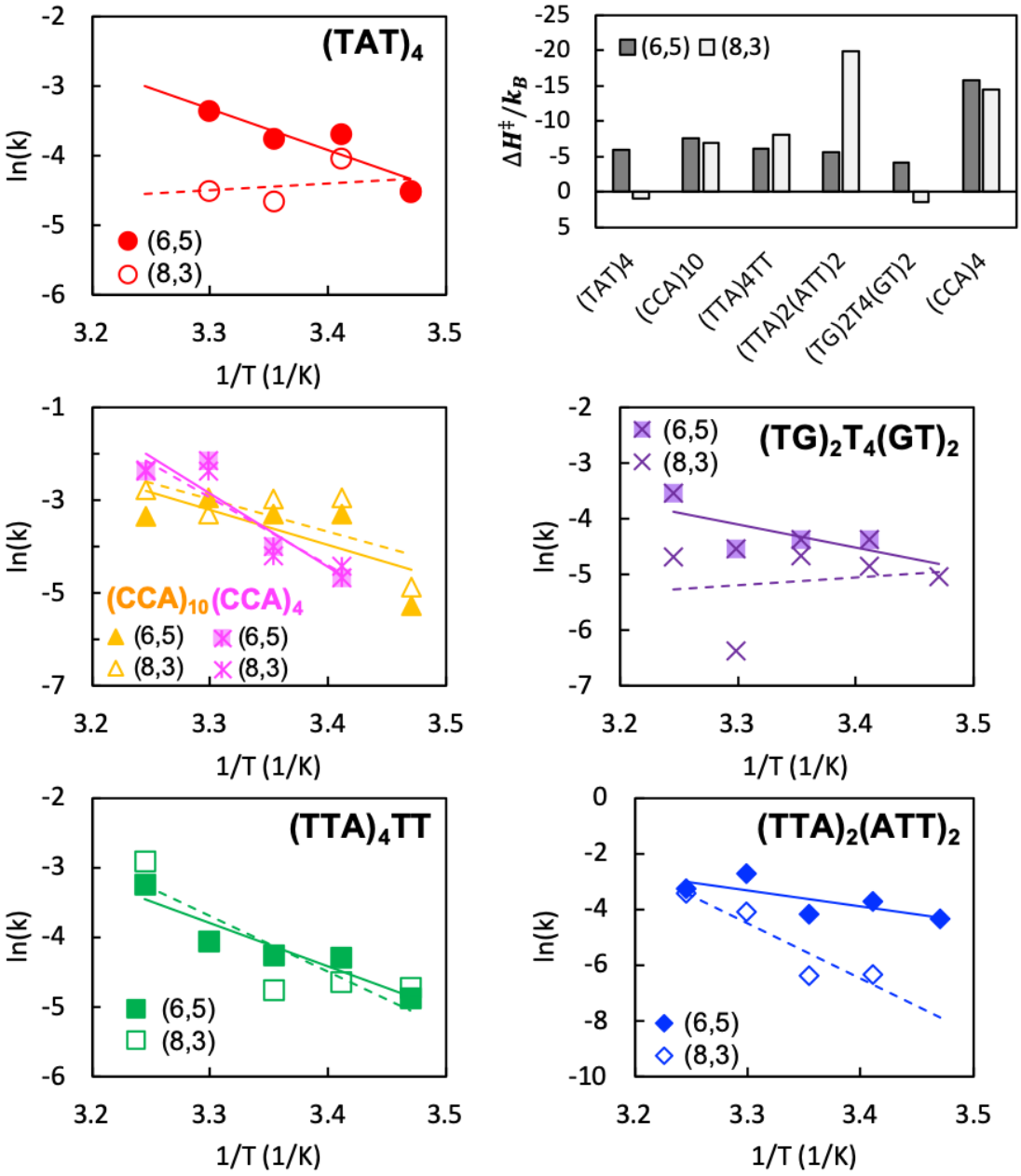

Figure 4.8. Eyring plots for the exchange process of SDC on (6,5) or (8,3)-SWCNT by different DNA sequences and the relative activation enthalpy ($\Delta H^{\ddagger}/k_B$) estimated from the slope of Eyring plot.

Although $(CCA)_{10}$ is a considerably longer sequence than other sequences we tested, no significant difference in activation enthalpy was observed. Since the kinetics are seen to be sequence dependent, it could not be appropriate to examine the length effect by comparing different lengths in different sequence combinations. Thus, we additionally tested 12-mer C/A rich sequence, $(CCA)_4$, and 30-mer T/A rich sequence, $(TAT)_{10}$. It is interesting to note that $(CCA)_4$ shows a significantly faster reaction than T/A or T/G rich sequences of the same length (Figure 4.8). It is worth noting that C/A rich sequences, including $(CCA)_{10}$, exhibited abnormal behavior in aqueous two-phase solution that required more PVP to move the SWCNT into the top phase, indicating higher hydrophilicity of the hybrid surface compared to others.[23]

Furthermore, $(TAT)_{10}$ showed an extremely slow reaction. For example, the peak location did not reach its final value (i.e., the peak location of DNA-covered SWCNT) even after 20 h at 50 °C, suggesting that the reaction was not completed (Figure 4.10). A similar instantaneous drop in fluorescence was still observed (Figure 4.5a) and no aggregation was detected. This experimental finding indicates that although some DNA covered the free sites on SWCNT as SDC was desorbed (i.e., the first and second stage in Figure 4.7), it still takes a long time to complete the rearrangement of DNA required to completely cover the SWCNT with DNA.

4.4 Conclusions

We have investigated the kinetics of the SDC exchange process by DNA, aided by the addition of MeOH. Observations suggest a three-stage process in which partial SDC is immediately desorbed by the addition of MeOH, DNA then (also rapidly) adsorbs to SWCNT surface exposed by SDC desorption, followed by a slower process interpreted as DNA rearrangement on the SWCNT surface as the system moves towards its equilibrium state. The activation energies for each DNA sequence on (6,5) and (8,3)-SWCNT were extracted to quantitatively compare their binding characteristics. We found two classes of behavior in Figure 4.8. Class 1 shows significant different activation energies between (6,5) and (8,3)-SWCNT. This class includes some recognition sequences, e.g., $(TAT)_4$, $(TTA)_2(ATT)_2$, and $(TG)_2T_4(GT)_2$. In contrast, Class 2 presents no chirality-dependence on the activation energy. The Class 2 includes the non-recognition sequence, $(CCA)_4$, also some recognition sequences, e.g., $(TTA)_4TT$ and $(CCA)_{10}$. This suggests that the newly developed MeOH aided replacement process does not always produce the same structure of DNA/SWCNT as does direct sonication. For example, the Class 1 sequences presumably have same structure as that prepared by direct sonication, so they exhibit chirality-dependence. However, the Class 2 sequences are formed in slightly different DNA/SWCNT structure, so their recognition ability to its partner species of SWCNT is lost. This finding strongly suggest that the recognition ability is based on the secondary structure of DNA on SWCNT. To demonstrate this interpretation, further experiments are required.

4.5 Acknowledgements

This work was performed in direct collaboration with Dr. Arjun Sharma of Lehigh University, Guillaume Noetinger of ESPCI (Paris, France), and Dr. Ming Zheng of National Institute of Standards and Technology (Gaithersburg, MD). I thank to Dr. Jeffrey Fagan of National Institute of Standards and Technology (Gaithersburg, MD) for providing the SDC/SWCNT dispersion. I also would like to thank Prof. Anand Jagota for his mentoring activities.

4.6 References

1. Iijima, S. & Ichihashi, T. Single-shell carbon nanotubes of 1-nm diameter. *Nature* **363**, 603–605 (1993).
2. Eatemadi, A. *et al.* Carbon nanotubes: properties, synthesis, purification, and medical applications. *Nanoscale Res. Lett.* **9**, 393 (2014).
3. Yang, N., Chen, X., Ren, T., Zhang, P. & Yang, D. Carbon nanotube based biosensors. *Sensors Actuators B Chem.* **207**, 690–715 (2015).
4. Tu, X., Manohar, S., Jagota, A. & Zheng, M. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature* **460**, 250–253 (2009).
5. Ao, G., Khripin, C. Y. & Zheng, M. DNA-controlled partition of carbon nanotubes in polymer aqueous two-phase systems. *J. Am. Chem. Soc.* **136**, 10383–10392 (2014).
6. Zhang, J. *et al.* Single Molecule Detection of Nitric Oxide Enabled by d(AT) $_{15}$ DNA Adsorbed to Near Infrared Fluorescent Single-Walled Carbon Nanotubes. *J. Am. Chem. Soc.* **133**, 567–581 (2011).
7. Landry, M. P. *et al.* Single-molecule detection of protein efflux from microorganisms using fluorescent single-walled carbon nanotube sensor arrays. *Nat. Nanotechnol.* **12**, 368–377 (2017).
8. Jena, P. V. *et al.* A Carbon Nanotube Optical Reporter Maps Endolysosomal Lipid Flux. *ACS Nano* **11**, 10689–10703 (2017).
9. Williams, R. M. *et al.* Noninvasive ovarian cancer biomarker detection via an optical nanosensor implant. *Sci. Adv.* **4**, eaaq1090 (2018).
10. Johnson, R. R., Kohlmeyer, A., Johnson, A. T. C. & Klein, M. L. Free Energy Landscape of a DNA−Carbon Nanotube Hybrid Using Replica Exchange Molecular Dynamics. *Nano Lett.* **9**, 537–541 (2009).
11. Roxbury, D., Jagota, A. & Mittal, J. Structural characteristics of oligomeric DNA strands adsorbed onto single-walled carbon nanotubes. *J. Phys. Chem. B* **117**, 132–140 (2013).
12. Shankar, A., Zheng, M. & Jagota, A. Energetic Basis of Single-Wall Carbon Nanotube Enantiomer Recognition by Single-Stranded DNA. *J. Phys. Chem. C* **121**, 17479–17487 (2017).
13. Iliafar, S., Mittal, J., Vezenov, D. & Jagota, A. Interaction of Single-Stranded DNA with Curved Carbon Nanotube Is Much Stronger Than with Flat Graphite. *J. Am. Chem. Soc.* **136**, 12947–12957 (2014).
14. Iliafar, S., Mittal, J., Vezenov, D. & Jagota, A. Interaction of Single-Stranded DNA with Curved Carbon Nanotube Is Much Stronger Than with Flat Graphite. *J. Am. Chem. Soc.* **136**, 12947–12957 (2014).
15. Shankar, A., Mittal, J. & Jagota, A. Binding between DNA and carbon nanotubes strongly depends upon sequence and chirality. *Langmuir* **30**, 3176–3183 (2014).
16. Ao, G., Streit, J. K., Fagan, J. A. & Zheng, M. Differentiating Left- and Right-Handed Carbon Nanotubes by DNA. *J. Am. Chem. Soc.* **138**, 16677–16685 (2016).
17. Zheng, Y., Bachilo, S. M. & Weisman, R. B. Quenching of Single-Walled Carbon

Nanotube Fluorescence by Dissolved Oxygen Reveals Selective Single-Stranded DNA Affinities. *J. Phys. Chem. Lett.* **8**, 1952–1955 (2017).
18. Yang, Y., Zheng, M. & Jagota, A. Learning to predict single-wall carbon nanotube-recognition DNA sequences. *npj Comput. Mater.* **5**, 3 (2019).
19. Ao, G. & Zheng, M. Preparation and Separation of DNA-Wrapped Carbon Nanotubes. in *Current Protocols in Chemical Biology* **7**, 43–51 (John Wiley & Sons, Inc., 2015).
20. Giraldo, J. P. *et al.* A Ratiometric Sensor Using Single Chirality Near-Infrared Fluorescent Carbon Nanotubes: Application to In Vivo Monitoring. *Small* **11**, 3973–3984 (2015).
21. Streit, J. K., Fagan, J. A. & Zheng, M. A Low Energy Route to DNA-Wrapped Carbon Nanotubes via Replacement of Bile Salt Surfactants. *Anal. Chem.* **89**, 10496–10503 (2017).
22. Choi, J. H. & Strano, M. S. Solvatochromism in single-walled carbon nanotubes. *Appl. Phys. Lett.* **90**, 1–4 (2007).
23. Yang, Y., Shankar, A., Aryaksama, T., Zheng, M. & Jagota, A. Quantification of DNA/SWCNT Solvation Differences by Aqueous Two-Phase Separation. *Langmuir* **34**, 1834–1843 (2018).
24. Roxbury, D., Manohar, S. & Jagota, A. Molecular simulation of DNA β-sheet and β-barrel structures on graphite and carbon nanotubes. *J. Phys. Chem. C* **114**, 13267–13276 (2010).

4.7 Appendix

4.7.1 Spectral Decomposition based on Voigt Profile

An individual spectrum of pure SWCNT species is well-approximated by a Voigt profile which is the convolution of a Gaussian, $G(x';\sigma)$, and a Lorentzian, $L(x-x';\gamma)$.

$$V(x;\sigma,\gamma) \equiv \int_{-\infty}^{\infty} G(x';\sigma)L(x-x';\gamma)dx' \tag{4.7}$$

where $G(x;\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$ and $L(x;\gamma) = \frac{\gamma}{\pi(x^2+\gamma^2)}$.

Here, σ is the standard deviation of the Gaussian profile, γ is the half-width at half-maximum (HWHM) of the Lorentzian profile, and x is the shift from the line center. The Voigt profile can be evaluated using the real part of the Faddeeva function $w(z)$:

$$V(x;\sigma,\gamma) = \frac{Re[w(z)]}{\sigma\sqrt{2\pi}} \tag{4.8}$$

where $z = \frac{x+i\gamma}{\sigma\sqrt{2}}$ and $w(z) = \mathrm{e}^{-z^2}\mathrm{erfc}(-\mathrm{i}z)$.

The spectra of a mixture of SWCNT is considered as a sum of the spectra of pure species. Thus, fit the spectra of the mixture of SWCNT as a sum of contributions from each of the purified species by adjusting the fitting coefficients: peak position, peak height, σ, and γ. Figure 4.2 shows a typical decomposition of the absorbance spectra into contributions from species in the mixture.

4.7.2 Correlation between Peak Shift and Reaction Progress

Let us consider a dominant peak like (6,5) that is not really influenced by its neighbors (although it strongly affects them). For this peak (let's assume it is Gaussian), let

$$f_S(\lambda) = a_S \exp\left(-\frac{(\lambda-\lambda_S)^2}{2\sigma_S^2}\right) \tag{4.9a}$$

$$f_D(\lambda) = a_D \exp\left(-\frac{(\lambda-\lambda_D)^2}{2\sigma_D^2}\right) \tag{4.9b}$$

Suppose that the spectra sum as $f = \alpha f_S + (1-\alpha) f_D$, where α indicates the extent to which the reaction has proceeded with value of '1' denoting fully surfactant coated and '0' representing fully DNA coated SWCNTs. So,

$$f = \alpha a_s \exp\left[-(\lambda-\lambda_s)^2/\left(2\sigma_s^2\right)\right] + (1-\alpha) a_D \exp\left[-(\lambda-\lambda_D)^2/\left(2\sigma_D^2\right)\right] \tag{4.10}$$

The peak location, λ^*, is given by the solution of

$$\frac{\partial f}{\partial \lambda} = 0 \Rightarrow$$

$$\alpha a_S \frac{(\lambda^*-\lambda_S)}{\sigma_S^2} \exp\left[-\frac{(\lambda^*-\lambda_S)^2}{2\sigma_S^2}\right] + (1-\alpha) a_D \frac{(\lambda^*-\lambda_D)}{\sigma_D^2} \exp\left[-\frac{(\lambda^*-\lambda_D)^2}{2\sigma_D^2}\right] = 0 \tag{4.11}$$

It is not obvious that the solution of this equation goes from one limit to the other according to α. Let's look at some special cases. Firstly, we know that

$$\frac{(\lambda^*-\lambda_S)^2}{2\sigma_S^2} \ll 1 \tag{4.12a}$$

$$\frac{(\lambda^*-\lambda_D)^2}{2\sigma_D^2} \ll 1 \tag{4.12b}$$

This is because our shifts (~ 5 nm) are small compared to the peak width (~20 nm). So, we approximate the exponentials in the previous equation by '1' and our equation becomes:

$$\alpha a_s \frac{(\lambda^* - \lambda_s)}{\sigma_s^2} + (1-\alpha) a_D \frac{(\lambda^* - \lambda_D)}{\sigma_D^2} = 0 \tag{4.13}$$

This can be solved for λ^*,

$$\lambda^* = \frac{\alpha a_s \sigma_D^2 \lambda_s + (1-\alpha) a_D \sigma_s^2 \lambda_D}{\alpha a_s \sigma_D^2 + (1-\alpha) a_D \sigma_s^2} \tag{4.14}$$

Suppose that the two peaks are such that $a_s \sigma_D^2 = a_D \sigma_s^2$. This could be because they have the same standard deviation and amplitude, for example. In that case,

$$\lambda^* = \alpha \lambda_s + (1-\alpha) \lambda_D \tag{4.15}$$

That is, under the assumption that the peak shift is small compared to standard deviation and that the amplitude and standard deviation of the DNA-coated and surfactant-coated cases are very similar, then one can say that just following the peak location is the same as following the reaction.

Creating counter examples where following the peak position is not a good idea is not so difficult. For example, take the case, as occurs in fluorescence spectroscopy where, say, $a_S >> a_D$. Then it is clear from looking at the solution for the peak position λ^* that for most of the reaction the peak position will not change from λ_S; it would be much better to follow the peak amplitude to follow the reaction.

4.7.3 Effect of Oxidant on SDC/SWCNT and DNA/SWCNT*

To demonstrate the stability of SDC surface adsorption on SWCNT in solution we treated SDC/SWCNT to a one electron oxidant reagent, $K_2Ir(Cl)_6$ (Potassium Hexachloroiridate). Fluorescence emission peak intensity measurements show that SDC/SWNCT are resistant to quenching even at increasing concentrations (1 μM - 10 μM) of $K_2Ir(Cl)_6$ (Figure 4.9a). On the other hand, DNA/SWCNT at the same $K_2Ir(Cl)_6$ concentration range show an increased fluorescence quenching (Figure 4.9b). It is well known that surfactants like SDC tightly pack around SWCNT, thereby limiting solvent access to the SWCNT surface. Based on our prior work[24] DNA oligomers form secondary structures on SWCNT that are specific to nucleotide sequence and SWCNT chirality. It is hypothesized that only some DNA oligomers form highly ordered conformations on specific SWCNTs, also known as recognition sequences. Hence DNA/SWCNT are generally very much susceptible to changes in solvent dielectric environment which manifests as large changes in fluorescence emission intensity.

* This work has been performed by Dr. Arjun Sharma of Lehigh University.

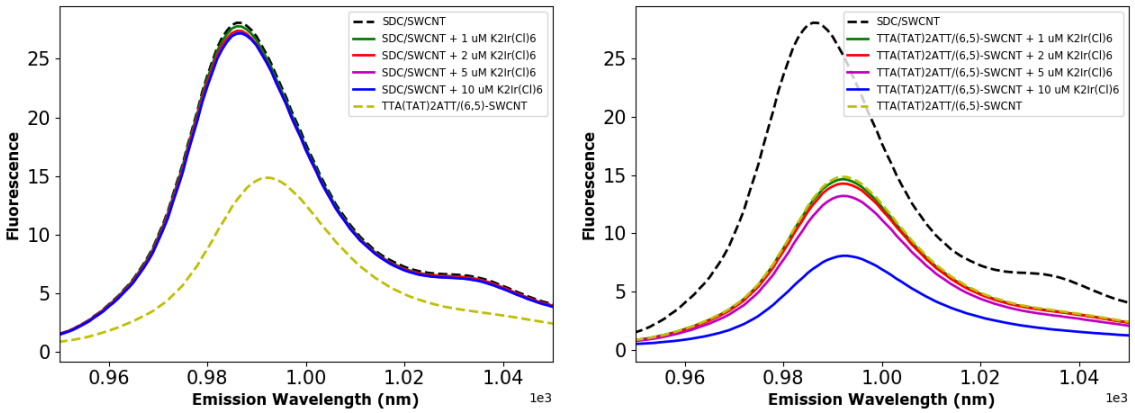

Figure 4.9. Comparison of fluorescence emission spectra of (a) SDC/SWCNT and (b) TTA(TAT)$_2$ATT/(6,5)-SWCNT in presence of $K_2Ir(Cl)_6$, a strong oxidizing reagent in solution (1 μM – 10 μM) . The fluorescence intensity values of SDC/SWCNT barely change while the emission peak of TTA(TAT)$_2$ATT/(6,5)-SWCNT shows substantial fluorescence quenching with increasing concentrations of $K_2Ir(Cl)_6$. The controls include SDC/SWCNT (solid black dashed line) and TTA(TAT)$_2$ATT/(6,5)-SWCNT (solid yellow dashed line) in absence of $K_2Ir(Cl)_6$.

4.7.4 DNA Rearrangement Kinetics and Transition State Theory

Consider a reaction of DNA rearrangement on SWCNT where disordered structure of DNA on SWCNT (*A*) forms an ordered structure DNA/SWCNT (*B*). The rate of production of *B* is then expressed by eq (4.2) in the main text of this chapter. Similarly, the rate of consumption of *A* can be found

$$\frac{d[A]}{dt} = -k[A] \tag{4.16}$$

Integrating $[A]$ with respect to *t*, we can obtain

$$[A] = [A]_0 \exp(-kt) \tag{4.17}$$

Plugging the expression into eq (4.2),

$$\frac{d[B]}{dt} = k[A] = k[A]_0 \exp(-kt) \tag{4.18}$$

Transition state theory assumes that there is an intermediate step in which a transition complex, $B^{\ddagger}$, is formed.

$$A \overset{k_1}{\Leftrightarrow} B^{\ddagger} \overset{k_2}{\rightarrow} B \tag{4.19}$$

The rate of change in the concentration of intermediate complex ($B^{\ddagger}$) can be found:

$$\frac{d[B^{\ddagger}]}{dt} = k_1[A] - k_{-1}[B^{\ddagger}] - k_2[B^{\ddagger}] \tag{4.20}$$

It is assumed that the reactants and the transition state are in a quasi-equilibrium:

$$k_1[A] = k_{-1}[B^{\ddagger}] \tag{4.21}$$

Therefore, the rate equation (4.20) can be simplified to

$$\frac{d[B^{\ddagger}]}{dt} = -k_2[B^{\ddagger}] = -\frac{k_1}{k_{-1}}k_2[A] = -\frac{d[B]}{dt} \tag{4.22}$$

where *k* is overall reaction rate constant $\left(k = K^{\ddagger}k_2,\ K^{\ddagger} = \frac{k_1}{k_{-1}}\right)$.

Then, integrating the right-hand side of equation (4.22) with respect to t and applying an initial condition,

$$[A] = [A]_0 \; and \; [B] = [B]_0 \; at \; t = 0$$

Finally, the concentration of ordered structure of DNA on the SWCNT $[B]$ can be found by eq (4.3) in the main text.

Using thermodynamic relationships, activation enthalpy, entropy, and free energies can be found:

$$\Delta G^{\ddagger} = -k_B T \ln K^{\ddagger} = \Delta H^{\ddagger} - T\Delta S^{\ddagger} \tag{4.23}$$

Thus,

$$\ln K^{\ddagger} = -\frac{\Delta H^{\ddagger}}{k_B T} + \frac{\Delta S^{\ddagger}}{k_B} \tag{4.24}$$

Then, the overall reaction constant k is

$$k = k_2 \exp\left(-\frac{\Delta H^{\ddagger}}{k_B T} + \frac{\Delta S^{\ddagger}}{k_B}\right) \tag{4.25a}$$

$$\ln(k) = \ln(k_2) - \frac{\Delta H^{\ddagger}}{k_B T} + \frac{\Delta S^{\ddagger}}{k_B} \tag{4.25b}$$

Therefore, the activation enthalpy can be obtained from the slope of $\ln(k)$ versus $1/T$ in the Eyring plot.

4.7.5 Raw Data of Temperature-dependent Kinetics in Absorbance for Various DNA on (6,5)-SWCNT

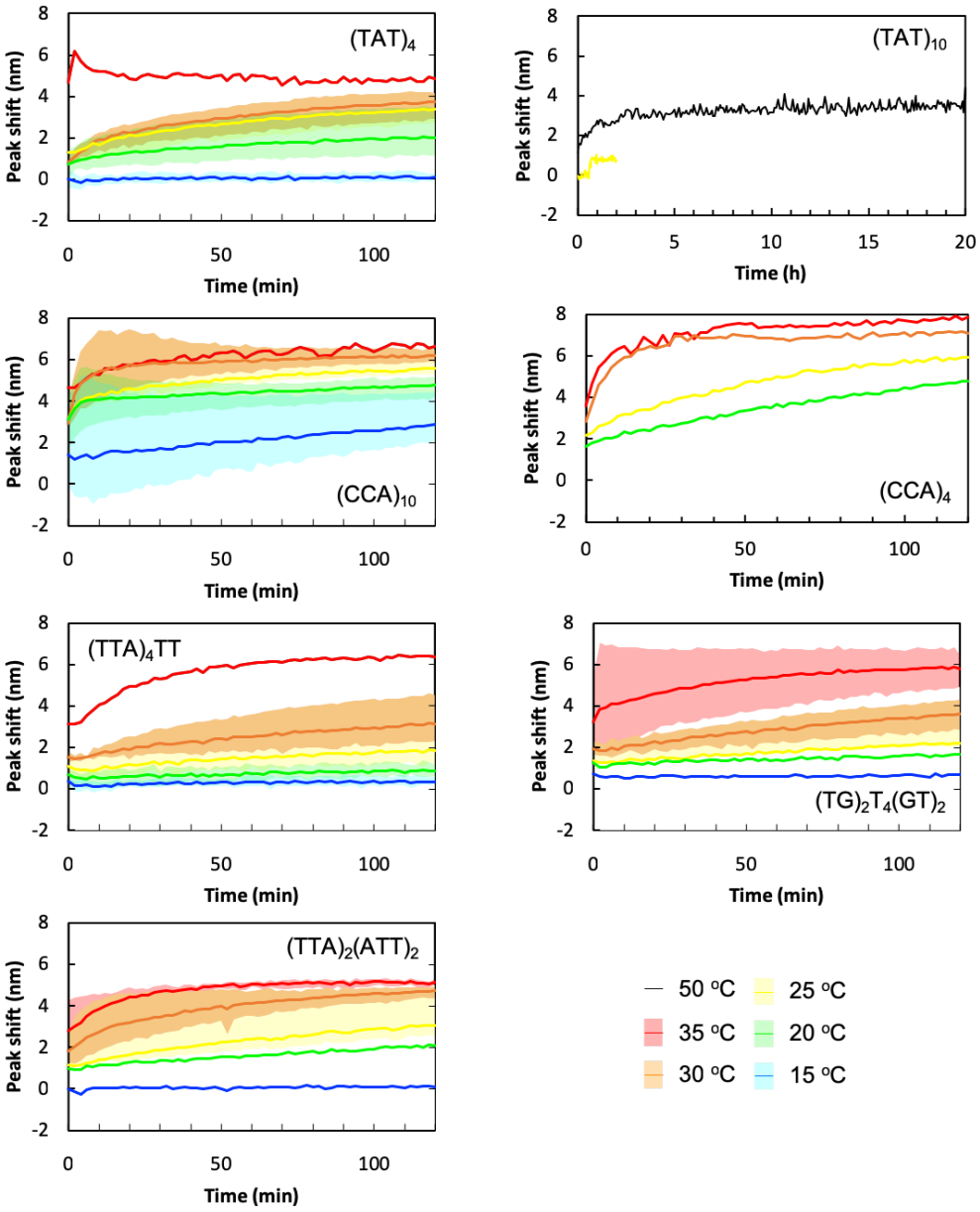

Figure 4.10. Temperature-dependent kinetics measured by the peak shift in the absorbance spectrum for the (6,5)-SWCNT. The peak location of each species of SWCNT was obtained by decomposing the spectra. Note that all the kinetic experiments shown here were performed at 40 v/v% MeOH concentration. Note that for some sequences, e.g., $(TAT)_{10}$, it takes too long to reach equilibrium wrapping configuration, and the exchange method might not be a good way to make DNA-CNTs for these sequences.

4.7.6 Link to Public Repository:

The following link is to a public repository where we provide a collection of scripts for spectral decomposition based on Voigt profile and kinetic analysis.

https://bitbucket.org/jagotagrouplehigh/dna_swcnt_SDCexchange/

Chapter 5 : Molecular Perceptron: A New Perception-based Sensor to Detect Ovarian Cancer Biomarker using Machine Learning*

Ovarian cancer is the fifth-leading cause of cancer-related deaths among females in the United States. Early and accurate detection of cancer can significantly improve the five-year survival rate. One of two FDA-approved serum biomarkers for ovarian cancer, Human epididymis protein 4 (HE4), provides noticeable sensitivity and specificity for ovarian cancer diagnosis. Current research on sensing applications has largely been based on one-to-one recognition. However, this is an inefficient way to detect various molecules since it requires the same number of receptors as the number of molecules one wishes to detect. To detect a combination of various analytes simultaneously, an effective and automatic data processing system is essential. In this study, we propose a new perception-based sensing system using weakly-specific sensor arrays that can be analyzed by an artificial perception model, we call the Molecular Perceptron. We demonstrate that the Molecular Perceptron can detect HE4 in the presence or absence of other analytes. The Molecular Perceptron is based on the DNA/SWCNT system, which has attracted considerable interest due to its unique optical properties and their strong sensitivity to

* Portions of this work have been submitted as grant proposal to Designing Materials to Revolutionize and Engineer our Future program of National Science Foundation. This work has been performed in direct collaboration with Dr. Ming Zheng at National Institute of Standards and Technology (Gaithersburg, MD) and Dr. Daniel Heller, Dr. Zvi Yaari, and Alex Settle at Memorial Sloan Kettering Cancer Center (New York, NY).

changes in the local environment. DNA/SWCNT hybrids were utilized to optically detect the analytes by observing changes in the fluorescence spectra of each SWCNT. Using the experimental data, machine learning models were trained using three different algorithms: Support Vector Machine, Random Forest, and Artificial Neural Network. Overall, the machine learning models achieved remarkable trainability giving F^1-scores of ~0.95. It is strongly suggestive of the idea that the perception mode of sensing can make accurate judgements in a noisy sensing environment.

5.1 Introduction

Cancer patient prognosis and quality of life are significantly affected by the failure to accurately diagnose disease at an early stage. One such example is ovarian cancer, the fifth-leading cause of cancer-related deaths among females in the United States and first among gynecologic malignancies,[1] resulting in 22,000 new cases and 14,000 deaths per year.[1] There is currently no method to achieve early, accurate diagnosis, nor are there strategies to rapidly determine patient response to treatment in order to inform the choice of therapy. The five-year relative survival rate for all patients diagnosed with ovarian cancer is 44%.[2] If detected at stage one, five-year survival rates are approximately 91%.[3]

Conventionally, serum CA-125 measurements and ultrasonography have been used to detect ovarian carcinoma but these methods do not result in early stage detection and convey little survival benefit.[4,5] Human epididymis protein 4 (HE4) is one of two FDA-approved serum biomarkers for ovarian cancer, along with CA-125, and plays a factor in ovarian tumorigenesis.[6] This protein is overexpressed by malignant epithelial cells[7] and found in increased levels in patient serum,[8,9] ascites,[10] and uterine fluid.[11] Serum-based HE4 provides similar sensitivity and specificity for ovarian cancer diagnosis as CA-125, although it may be more useful in differentiating benign from malignant disease.[8]

The DNA/SWCNT system has been widely used in biosensing applications due to its strong optical absorption in the near-infrared (NIR) region and high sensitivity to the local environment.[12–16] Current research on sensing applications is primarily based on specific one-to-one *recognition*-based sensing. This is conceptually simple and easier to design and interpret. However, it imposes strong requirements on needing to find one-to-

one recognition pairs since it requires the same number of receptors as the number of molecules to detect. Thus, it is an inefficient way by which to detect various molecules. For example, in a real system, biofluids contain a plethora of molecules that together can accurately define the physiological state of a person. A grand challenge is to sense this entirety by a single device. We believe that rather than attempting to accomplish this on a one-to-one recognition basis, it is far more effective to grasp it by perception. This leads us the need of a *perception*-based system with multiple receptors; each one captures certain features of the target molecule and the overall ensemble response is then analyzed by an artificial model resulting in perception. In past, Staii et. al.[17] have shown the feasibility of *perception*-based sensing utilizing an *Electronic Nose*-based system[18] by a DNA-decorated field-effect transistor, but exhibited limited success.

Here, we propose a new *perception*-based system using DNA/SWCNT hybrids along with a machine learning (ML) framework to construct an artificial perception system that we call a *Molecular Perceptron*. We demonstrate a *Molecular Perceptron* for detection of the ovarian cancer serum biomarker HE4 in the presence or absence of fetal bovine serum (FBS) and bovine serum albumin (BSA) using machine learning techniques. Optical responses induced by analytes, represented by both fluorescence peak position and intensity changes, were obtained. We then built an artificial perception model using machine learning methods. To find the best model, three different algorithms (support vector machine (SVM), random forest (RF), and artificial neural network (ANN)) were examined, with multiple input feature vector representations including different missing value/outlier treatments, DNA sequence encoding method, and creating feature vectors by

Principal Components Analysis (PCA). In addition to varying input feature vectors, we examined three different ways to define target variables (bi-class, multi-class, and multi-label classification). The classification algorithms, performance validation, and optimization were implemented using the Scikit-learn machine learning library.[19] The models successfully detect not only the target biomarker (HE4), but also other analytes (BSA and FBS). Furthermore, the feature importance within the RF models and saliency within the ANN models were analyzed to extract physical meaning, as well as to help in the feature selection.

5.2 Materials and Methods

5.2.1 Data Collection*

To develop training sets for the machine learning models, we began with a small set of previously identified recognition sequences,[14,20] showing the ability to recognize specific molecules such as certain SWCNT species, protein, or their relatives, as listed in Table 5.1. Each sequence was used to disperse a synthetic mixture of SWCNTs containing ~ 10 semiconducting chiral species (e.g. SG65i from Sigma Aldrich). The analyte, a specific cancer biomarker HE4, was introduced to the DNA-SWCNTs in presence or absence of bovine serum albumin (BSA) or fetal bovine serum (FBS). Near-infrared photoluminescence spectra were acquired on these samples. The spectroscopy was conducted using a custom-built high-throughput setup that allows for measurements

* This work has been performed by Dr. Daniel A Heller, Dr. Zvi Yaari, and Alex Settle at the Memorial Sloan Kettering Cancer Center, New York, NY.

directly within 96 or 384 well plates, with each well containing the SWCNT synthetic mixture dispersed by one type of sequence, followed by the addition of different analytes. The instrumentation consists of an epifluorescence microscope with automated translation stage programmed for high-throughput data collection. The samples were excited at 660 nm and 730 nm using a supercontinuum light source. Spectra was acquired using a Princeton Instruments IsoPlane spectrometer coupled to a NIRvana 640x512 InGaAs array detector.

Table 5.1. Initial DNA sequence set. The sequences were chosen from previously identified recognition sequences or their relatives.

DNA sequence	**Specialty**
$(GT)_{12}$	Relative of $(GT)_{20}$ (8,4), (7,4), and (5,5)-recognition[20]
$(ATT)_5$	Relative of $(ATT)_4$ (7,5)-recognition[20]
$(AT)_{11}$	Protein-recognition[14]
$(AT)_{15}$	Protein-recognition[14]
$(AT)_{20}$	Protein-recognition[14]
$(AC)_{15}$	Non-recognition, but shows special characteristic in ATP system[21]
$(TCT)_5$	(6,5) and (6,6)-recognition
$T_3C_3T_3C_3T_3$	(6,5)-recognition
$C_3T_9C_3$	(8,3)-recognition
$C_3T_3C_9$	(6,4), (9,1), (7,3), and (9,2)-recognition

5.2.2 Data Preprocessing

The dataset comprises the photoluminescence spectra of each combination of DNA-SWCNT hybrid exposed to different combinations of a small number of analytes (HE4, BSA, FBS). That is, we have total $N{\cdot}M{\cdot}L$ combinations where N is the number of DNA sequences, M is the number of SWCNT chiralites, and L is the number of analyte combinations. The spectra are analyzed to yield two parameters for each SWCNT type: the relative peak position and intensity:

$$dwl_i = wl_i - wl_0 \tag{5.1a}$$

or

$$dint_i = \frac{int_i}{int_0} \tag{5.1b}$$

where wl_0 and int_0 are the wavelength and intensity of a control sample (DNA/SWCNT without analyte); wl_i and int_i are the wavelength and intensity of DNA/SWCNT with analyte combination, i. We have considered five different analyte combinations for each DNA/CNT pair (i.e., $L = 5$): HE4, BSA, BSA+HE4, FBS, and FBS+HE4. In this way, each DNA/SWCNT/analyte combination has two relative spectroscopic measurement values $(dwl_i, dint_i)$.

Next, we identify input and output (target) variables for the machine learning algorithm. The input variables include DNA sequence, SWCNT chirality, DNA modification, and the two spectroscopically measured parameters $(dwl_i, dint_i)$. The output variable could be analyte type or concentration of each analyte. The learning problem is defined by the target variable. If it is continuous (e.g., concentration) the problem would be considered as regression, and if it is a discrete number of values, the

problem would be considered as classification. In a classification problem, if there are only two classes (e.g., yes/no or 1/0), it is a binary classification problem. If there are three or more classes and each example is assigned to only one class, this is called 'multi-class' classification; if each example can be assigned multiple classes, this is called 'multi-label' classification. Our problem is originally of multi-label classification, but we can consider the output class as the presence or absence of HE4 (bi-class) or consider the output class as every analyte combination present in training set (multi-class).

For learning models, categorical data (such as SWCNT chirality and analyte type) must be transformed to numeric values, for which we use the common *one-hot encoding* technique.[22] We encode the DNA sequence by taking two or three bases as a *term*, then encoding each sequence as its *term-frequency vector*.[23] We have considered several ways to construct feature vectors and present the experimental results for the best-performing ones. Figure 5.1 depicts the overall scheme for the input feature construction.

There are several ways to construct feature vectors by defining examples and features differently. We aim to form a proper feature space to enhance machine learning model performance. When considering each DNA/SWCNT/analyte hybrid as a single example (*fvtype1* in Table 5.2), the number of examples can be maximized. However, from a practical point of view, single DNA/SWCNT hybrid seems to lack sufficient information to predict the presence of the analyte. It is also diametrically opposed to the main idea of the *Molecular Perceptron.* Another extreme case (*fvtype4* in Table 5.2) considers each DNA/SWCNT/analyte combination as a single example and each set of measurements ($dwl_i, dint_i$) with different DNA sequence and SWCNT chirality as features. A single

example of this feature vector contains much more information, but the number of examples is equal to the number of combinations of analytes (five in our dataset). Both *fvtype2* and *fvtype3* have plenty of information in a single example. Considering feature vector expandability for new data, DNA library is much larger than SWCNT library, so it is more likely to add new DNA sequences rather than introduce new species of SWCNT. Since *fvtype3* can be easily expandable with more DNA sequence data, we decided to construct feature vector using *fvtype3*.

The feature vector involves some missing values that occur due to unobserved data or outliers that are abnormally large. In the case of unobserved data, we eliminate an example (complete-example set) or feature (complete-feature set) that includes missing values. For outliers, we eliminate an example if $dwl_i > 20\ nm$.

Note that the range of all feature vectors were rescaled to the range in [0, 1] to weigh all features equally.

5.2.3 Feature Space Reduction using Principal Component Analysis (PCA)

Feature selection is also very important to achieve better model performance. It is particularly a concern if the feature space is large, because there is otherwise greater chance of overfitting. In order to reduce the feature space, we used principal component analysis (PCA) to find the directions of greatest variation in the dataset. This allows us to define fewer but more relevant features of which we used the first five principal components.

The preprocessing including PCA was implemented in R.

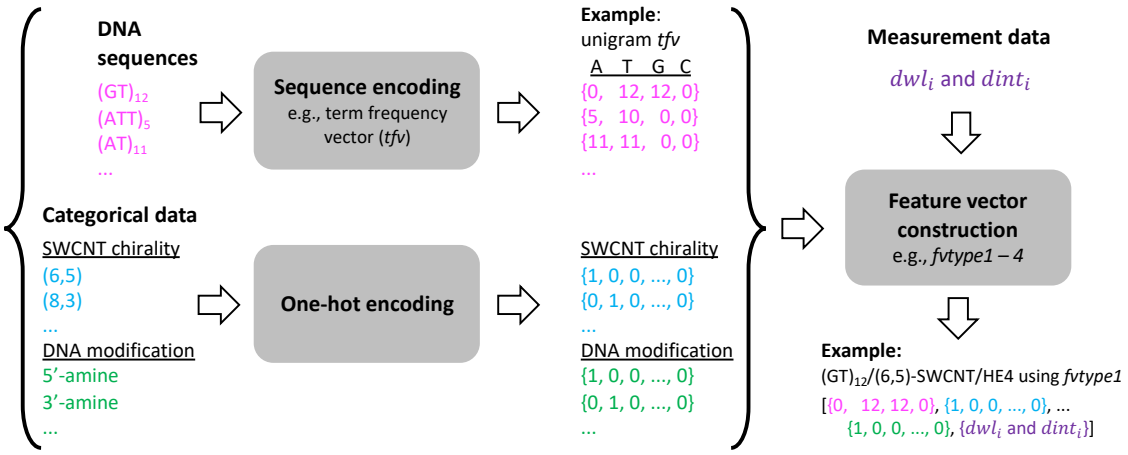

Figure 5.1. Overall scheme for input feature construction. First, DNA sequence is encoded to numeric vector using *term-frequency* method.[23] All the categorical data such as SWCNT chirality and DNA modification are also transformed to numeric vectors using *one-hot encoding* technique. The numeric vectors can be then combined in various ways as shown in Table 5.2. This figure presents an example of a feature vector for a single example obtained by the *fvtype1* method.

Table 5.2. Feature vector construction types: DNA sequence, DNA modification, SWCNT chirality, and the analyte combination can be transformed to a numeric vector, as shown in Figure 5.1. The numeric vectors representing different classes (DNA sequence, SWCNT chirality, etc.) can be combined in various ways as presented in this table.

Name	**Feature vector**	**Single example representation**	**# of example**	**# of features**
fvtype1	[{DNA sequence}, {Modification type}, {Chirality}, {Measurements set}] e.g., [{0, 12, 12, 0}, {0,0,0,0,1}, {1,0,0, … ,0}, $\{dwl, dInt\}$]	DNA/modification/ SWCNT/analyte e.g., $(GT)_{12}$/none/ (6,5)-SWCNT/HE4	$N{\cdot}M{\cdot}L$	$4^k{+}M{+}P{+}S$
fvtype2	[{Chirality}, {Measurements set for each DNA sequence}] e.g., [{1,0,0, … ,0}, $\{dwl_{seq1}, dwl_{seq2}, \dots, dwl_{seqN}, dInt_{seq1}, dInt_{seq2}, \dots, dInt_{seqN}\}$]	SWCNT/analyte e.g., (6,5)-SWCNT/HE4	$M{\cdot}L$	$M{+}S{\cdot}Q$
fvtype3	[{DNA sequence}, {Modification type}, {Measurements set for each chirality}] e.g., [{0, 12, 12, 0}, {0,0,0,0,1}, $\{dwl_{(6,5)}, \dots, dwl_{(8,7)}, dInt_{(6,5)}, \dots, dInt_{(8,7)}\}$]	DNA/analyte e.g., $(GT)_{12}$/none/HE4	$Q{\cdot}L$	$4^k{+}P{+}S{\cdot}M$
fvtype4	{Measurements set for each chirality with different DNA sequence} e.g., $\{dwl_{seq1,(6,5)}, \dots, dwl_{seq1,(8,7)}, \dots, dwl_{seqN,(6,5)}, \dots, dwl_{seqN,(8,7)}, dInt_{seq1,(6,5)}, \dots, dInt_{seq1,(8,7)}, \dots, dInt_{seqN,(6,5)}, \dots, dInt_{seqN,(8,7)}\}$	Analyte e.g., HE4	L	$S{\cdot}Q{\cdot}M$

Note that N is the number of sequences in training set; M is the number of chirality in training set; L is the number of analyte combination; P is the number of modification type; Q is number of DNA/modification combination (Q < N+P); k is size of n-gram in term frequency vector; S is the number or measurement type.

5.2.4 Learning, Validation, and Evaluation

In order to find the best ML model, we examined three different algorithms (support vector machine, SVM, random forest, RF, and artificial neural network, ANN), each with multiple input feature vectors including bi/trigram *term-frequency vector* for DNA sequence encoding, different missing value treatment (elimination of example or feature), DNA sequence feature existence, and feature vectors by PCA. In addition, we examined bi-class, multi-class, and multi-label classification algorithms. Each model was evaluated by 10-fold cross-validation. Classification algorithms and cross-validation approaches were implemented using the Scikit-learn machine learning library.[19] Since hyperparameters, such as learning rate and activation function, have a significant impact on model performance, we used Bayesian hyperparameter optimization to find optimal hyperparameters to maximize F^1-score, implemented in HyperOpt.

5.3 Results and Discussion

5.3.1 Spectroscopic Data Analysis*

First, we demonstrated that the response of DNA/SWCNT intensity to HE4 can vary according to DNA sequences. Figure 5.2a clearly shows the sequence dependence of the spectroscopic characteristic on HE4. In particular, the peak shift (dwl_i) not only shows strong dependence on DNA sequence, but also can be either positive or negative. (In general, the intensity decreases as HE4 is added.) It is interesting to note that the peak shift

* The experimental work was carried out by Alex Settle, Dr. Zvi Yaari, and Dr. Danial Heller at Memorial Sloan Kettering Cancer Center, New York, NY.

and intensity changes appear not to be correlated with each other. This indicates that there is not simple rule such as the peak is blue/red shifted as a response to HE4 for single species of SWCNT. Also, the sequence dependence can be evidence that each DNA/SWCNT structure is sequence-dependent, which affects the binding ability of HE4 on DNA/SWCNT.

In addition to the sequence dependence, we examined the response of DNA/SWCNT to HE4 in the presence or absence of other interferents (e.g., BSA and FBS). We found that some sequences show significant difference in peak shift and intensity changes. Note that most sequences generally showed the differences in the intensity in presence of HE4 (see Figure 5.6 in Appendix). Figure 5.2b depicts the special sequence, $(ATT)_5$, which shows significant difference in both peak shift and intensity changes, on (7,5)-SWCNT in different analyte conditions. Both peak shift and intensity is significantly decreased in presence of HE4 regardless of the presence of BSA and FBS.

We also found that most DNA sequences showed specific response to HE4. This was expected by the fact that the initial sequences set was derived from structurally well-defined DNA/SWCNTs that have previously enabled separation[20] or molecular sensing[14] (Table 5.1).

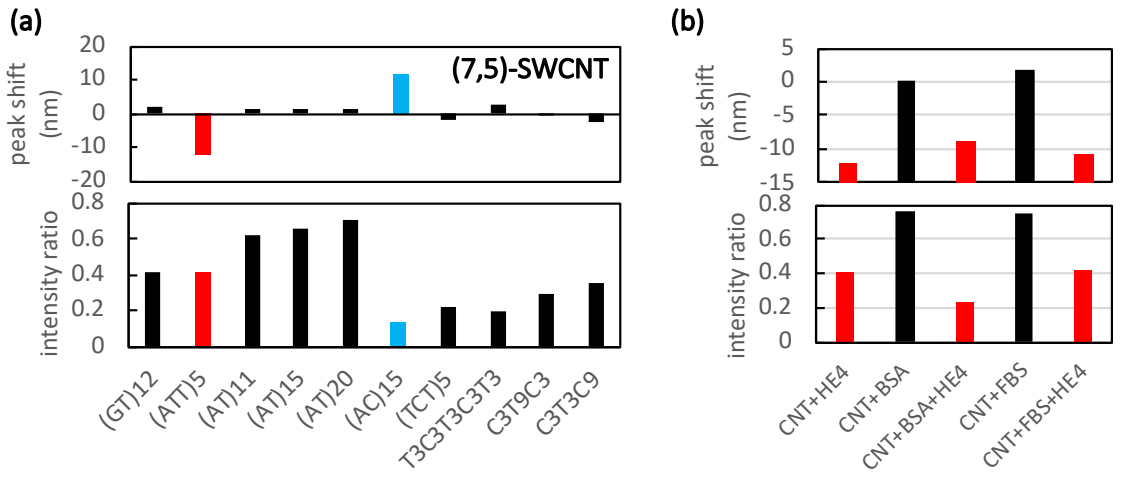

Figure 5.2. (a) DNA sequence dependence of (7,5)-SWCNT response to HE4; (b) Response of the $(ATT)_5$-(7,5) SWCNT to HE4 vs. interferents. Note that $(ATT)_5$-(7,5) SWCNT is depicted by the red bar in panel (a).

5.3.2 Machine Learning Model Development

The overall scheme of our approach is presented in Figure 5.3. Each model was optimized, and the model performance was estimated by the F^1–score using 10-fold cross-validation. Note that the target label from the lower right model in Figure 5.3 was labeled as the target label in the multi-label classification, but in practice, all three classification types were used.

The best models that gave the highest F^1–scores are listed in Table 5.3. In general, RF showed better performance than ANN and SVM. In particular, SVM showed poor performance in multi-class/multi-label classification. The performance of bi-class classifiers is slightly better than multi-class and multi-label classifiers. In terms of the missing value treatment, *complete-example* set showed better performance than *complete-feature* set. For multi-class/multi-label classification, it is seen that DNA sequence

information rarely has an impact on the model performance. However, this is unlikely to be correct, as will be discussed later. Overall, the validation results are highly encouraging given that the F^1–scores achieved approach the maximum value of 1.0.

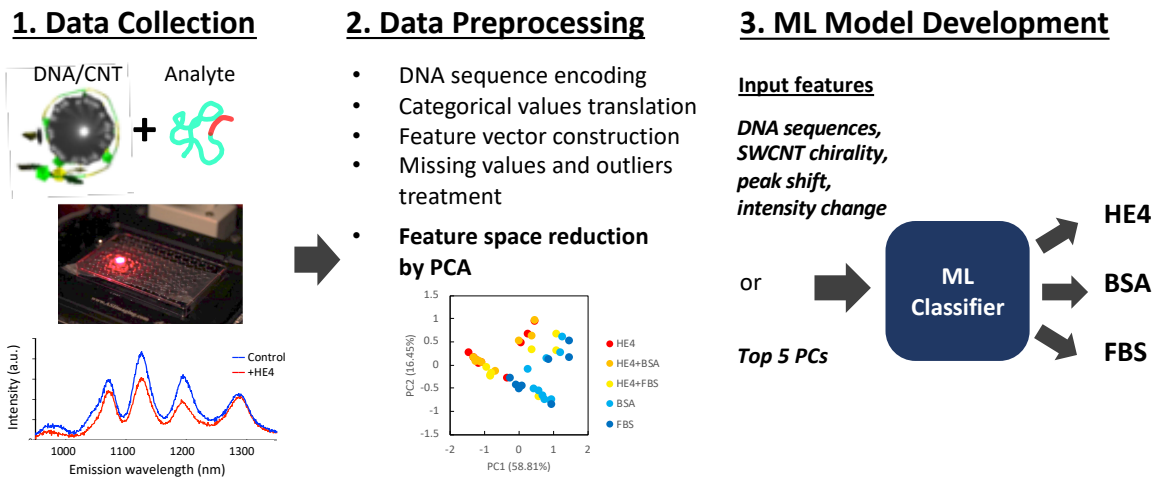

Figure 5.3. Overall scheme to develop a Molecular Perceptron to detect HE4. First, to collect data for the training set, near-IR photoluminescence spectra were measured for various DNA/SWCNT/analyte combinations. The picture presents high-throughput near-IR fluorescence spectroscopy. Each well contains a single DNA sequence and the entire complement of SWCNTs in presence or absence of analytes.* The corresponding target label depends on a classification type. For example, if a bi-class classification is considered, the target labels would be the presence or absence of HE4. For multi-class classification, the target labels will be different analyte combinations. For multi-label classification, the target labels will be each analyte type. Note that the target label from the ML classifier, depicted in lower right, is presented as the target label in multi-label classification, but in practice, all three classification types were used. Data preprocessing step involves translation of all categorical values to a numeric vector, missing values and outlier treatment, feature vector construction, and feature space reduction. Once the feature vector is created, the models with three different types of classification algorithms and feature extraction method are trained using the training set feature vectors.

* The picture and spectra on the left in Figure 5.3 were created by Dr. Daniel Heller at Memorial Sloan Kettering Cancer Center, New York, NY.

Table 5.3. Top five bi-class and multi-class/multi-label algorithms (ALG) showing excellent trainability

Bi-class classifier			Multi-class/multi-label classifier			
Feature vector	**ALG**	**F^1-score**	**Output type**	**Feature vector**	**ALG**	**F^1-score**
PCA Bigram complete-example	RF	0.947	Multi-class	Bigram complete-example	RF	0.930
Bigram complete-feature	RF	0.942	Multi-class	noDNA complete-example	RF	0.930
Trigram complete-feature	RF	0.942	Multi-label	noDNA complete-example	RF	0.930
Bigram complete-example	SVM	0.927	Multi-label	noDNA complete-example	ANN	0.927
Trigram complete-example	ANN	0.927	Multi-class	Trigram complete-example	RF	0.923

5.3.3 Feature Importance and Saliency Analysis

Although PCA can combine correlated features and reduce the dimensionality of feature space, it is difficult to understand what exactly the principal components mean, because they are combinations of the original features. Furthermore, it is important to note that the criteria for PCA are completely unsupervised. For example, PC1 can be correlated with other factors such as DNA length and PC2 can be correlated with an analyte concentration. While, the feature importance and saliency analysis depend on the output class, so the analysis selects dimensions which lead to an improved classifier performance. In addition, the analysis can provide a physical meaning, for example, which SWCNT chirality is more sensitive to detect an analyte or which nucleobase is more important. This analysis may reduce future experimental work.

Figure 5.4 shows the feature importance of top two bi-class and multi-class models. The results show that DNA sequence feature has negligible importance for determining the

output class. However, this is unlikely to be correct based on the prior knowledge that DNA sequences do show selective recognition on SWCNT with different electrostatic, solvation, and energetic characteristics[20,21,24,25]. In addition, the DNA sequence-dependence on the SWCNT emission response to HE4 has been observed, as depicted in Figure 5.2a and Figure 5.6. This implies that the current choice of feature vector likely does not capture salient information from the DNA sequences, which might be caused by the feature vector type. For example, the feature vector type we used (*fvtype3*) combines the chirality and spectroscopic measurement values (dwl_i, $dint_i$). Since the measurement values have more direct information on the HE4 detection, it can be considered overwhelming compared to other features such as DNA sequences. On the other hands, the *fvtype2* combines the DNA sequence and spectroscopic measurement values. It is likely that the trained model by fvtype2 feature vector will help to reveal sequence dependence in HE4 detection. As the previous model by the input feature vector of *fvtype3* showed, it is likely that the model trained by the input feature vector of *fvtype2* helps to demonstrate the sequence-dependence in HE4 detection. This leads to further model development and feature importance analysis with the feature vector type, *fvtype2*, to see the sequence dependence.

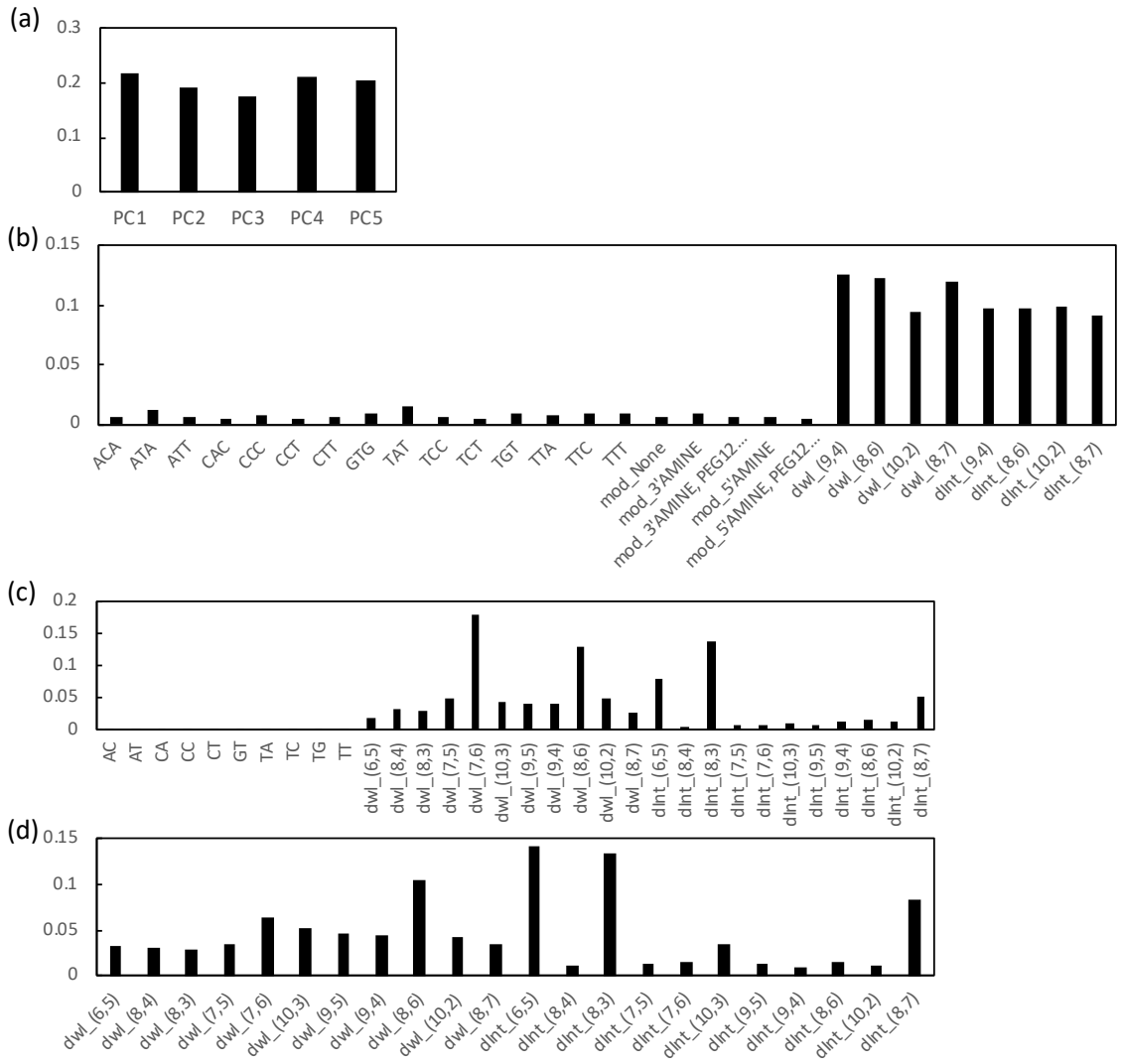

Figure 5.4. Feature importance of top two bi-class and multi-class models. (a) PCA/Bigram/*complete-example* feature vector using bi-class RF; (b) Trigram/*complete-feature* feature vector using bi-class RF; (c) Bigram/*complete-example* feature vector using multi-class RF; (d) noDNA/*complete-example* feature vector using multi-class RF.

5.4 Conclusions

We demonstrated a new *perception*-based sensing system using machine learning techniques to detect the ovarian cancer biomarker HE4 in the presence or absence of other analytes (BSA and FBS). The DNA/SWCNT hybrids were utilized to optically detect the analytes by observing changes in the peak intensity and peak location of each SWCNT. Using the experimental data, the machine learning models were trained. Overall, the models achieved remarkable trainability giving F^1-scores of ~0.95. This is strongly suggestive of the idea that the perception mode of sensing can make accurate predictions.

The feature importance and saliency analysis of the trained models imply that the current feature vector may not capture relevant information in DNA sequences. This may be because the feature vector type *fvtype3* emphasizes the effect of SWCNT chirality by combining chirality and measurement values. Further work is required to resolve the problem of insufficient DNA information in current feature vectors; we suggest the use of the feature vector constructed by *fvtype2* that combines DNA sequence and measurement values. We expect that the feature importance results of two different feature construction type (*fvtype2* and *fvtype3*) can complement each other, which can improve model performance by feature selection based on the results.

Furthermore, we found that most DNA sequences in the initial set showed specific response to HE4. Note that our initial sequence set was derived from previously identified recognition sequences. By the fact that recognition sequences can form structurally well-defined DNA/SWCNTs,[26,27] it can be interpreted that there is more chance of finding a special sequence for analytes detection in a set of pre-identified recognition sequences. We

can connect this study to our previous work on the recognition sequences predictive model using ML technique. The predictive model can provide good candidates for the *Molecular Perceptron*.

So far, we have considered the sensing model as the classification problem. However, from an early detection perspective, it is advisable to monitor the analyte concentration, even if the concentrations of target analytes are not in the range to determine the class. This can be resolved by utilizing a multi-target regression model that takes the output as the analyte concentration. To build the regression model, additional experimental in difference concentration of analytes is required.

Finally, we would like to note that the methods developed herein to detect a given analyte using molecular perceptron can be expanded to detect any number of important bioanalytes, and allow multiplexed detection via the use of multiple SWCNT chiralities and/or separately-addressable sensors. Eventually, these ideas both have the potential to change clinical practice to screen and detect disease at early stages.

5.5 Acknowledgement

This work was performed in direct collaboration with Dr. Daniel Heller, Dr. Zvi Yaari, and Alex Settle at Memorial Sloan Kettering Cancer Center (New York, NY) and Dr. Ming Zheng at National Institute of Standards and Technology (Gaithersburg, MD). In particular, the experimental work, described in Data Collection (section 5.2.1), was performed by Dr. Daniel Heller, Dr. Zvi Yaari, and Alex Settle at Memorial Sloan Kettering Cancer Center, New York, NY. I would also like to acknowledge Prof. Anand Jagota, Dr. Daniel Heller, and Dr. Ming Zheng for serving as mentors during this work.

5.6 References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA. Cancer J. Clin.* **66**, 7–30 (2016).
2. Vaughan, S. *et al.* Rethinking ovarian cancer: recommendations for improving outcomes. *Nat. Rev. Cancer* **11**, 719–725 (2011).
3. Maringe, C. *et al.* Stage at diagnosis and ovarian cancer survival: Evidence from the International Cancer Benchmarking Partnership. *Gynecol. Oncol.* **127**, 75–82 (2012).
4. Brown, P. O. & Palmer, C. The Preclinical Natural History of Serous Ovarian Cancer: Defining the Target for Early Detection. *PLoS Med.* **6**, e1000114 (2009).
5. Buys, S. S. *et al.* Effect of Screening on Ovarian Cancer Mortality. *JAMA* **305**, 2295 (2011).
6. Moore, R. G. *et al.* HE4 (WFDC2) gene overexpression promotes ovarian tumor growth. *Sci. Rep.* **4**, 3574 (2015).
7. Drapkin, R. *et al.* Human Epididymis Protein 4 (HE4) Is a Secreted Glycoprotein that Is Overexpressed by Serous and Endometrioid Ovarian Carcinomas. *Cancer Res.* **65**, 2162–2169 (2005).
8. Hellström, I. *et al.* The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma. *Cancer Res.* **63**, 3695–700 (2003).
9. Wu, L. *et al.* Diagnostic Value of Serum Human Epididymis Protein 4 (HE4) in Ovarian Carcinoma: A Systematic Review and Meta-Analysis. *Int. J. Gynecol. Cancer* **22**, 1106–1112 (2012).
10. Shender, V. O. *et al.* Proteome-metabolome profiling of ovarian cancer ascites reveals novel components involved in intercellular communication. *Mol. Cell. Proteomics* **13**, 3558–71 (2014).
11. Levine, D. A. Detection of ovarian cancer. (2013).
12. Zhang, J. *et al.* Single Molecule Detection of Nitric Oxide Enabled by $d(AT)_{15}$ DNA Adsorbed to Near Infrared Fluorescent Single-Walled Carbon Nanotubes. *J. Am. Chem. Soc.* **133**, 567–581 (2011).
13. Shi, J. *et al.* Microbiosensors based on DNA modified single-walled carbon nanotube and Pt black nanocomposites. *Analyst* **136**, 4916 (2011).
14. Landry, M. P. *et al.* Single-molecule detection of protein efflux from microorganisms using fluorescent single-walled carbon nanotube sensor arrays. *Nat. Nanotechnol.* **12**, 368–377 (2017).
15. Jena, P. V. *et al.* A Carbon Nanotube Optical Reporter Maps Endolysosomal Lipid Flux. *ACS Nano* **11**, 10689–10703 (2017).
16. Galassi, T. V *et al.* An optical nanoreporter of endolysosomal lipid accumulation reveals enduring effects of diet on hepatic macrophages in vivo. *Sci. Transl. Med.* **10**, eaar2680 (2018).
17. Staii, C., Johnson, A. T., Chen, M. & Gelperin, A. DNA-Decorated Carbon Nanotubes for Chemical Sensing. *Nano Lett.* **5**, 1774–1778 (2005).
18. Persaud, K. & Dodd, G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* **299**, 352–355 (1982).
19. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*

12, 2825–2830 (2011).
20. Ao, G., Streit, J. K., Fagan, J. A. & Zheng, M. Differentiating Left- and Right-Handed Carbon Nanotubes by DNA. *J. Am. Chem. Soc.* **138**, 16677–16685 (2016).
21. Yang, Y., Shankar, A., Aryaksama, T., Zheng, M. & Jagota, A. Quantification of DNA/SWCNT Solvation Differences by Aqueous Two-Phase Separation. *Langmuir* **34**, 1834–1843 (2018).
22. Harris, D. & Harris, S. *Digital Design and Computer Architecture, Second Edition*. (Morgan Kaufmann Publishers Inc., 2012).
23. Yang, Y., Zheng, M. & Jagota, A. Learning to predict single-wall carbon nanotube-recognition DNA sequences. *npj Comput. Mater.* **5**, 3 (2019).
24. Zheng, M. *et al.* Structure-Based Carbon Nanotube Sorting by Sequence-Dependent DNA Assembly. *Science (80-.).* **302**, 1545–1548 (2003).
25. Shankar, A., Zheng, M. & Jagota, A. Energetic Basis of Single-Wall Carbon Nanotube Enantiomer Recognition by Single-Stranded DNA. *J. Phys. Chem. C* **121**, 17479–17487 (2017).
26. Tu, X., Manohar, S., Jagota, A. & Zheng, M. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature* **460**, 250–253 (2009).
27. Ao, G., Khripin, C. Y. & Zheng, M. DNA-controlled partition of carbon nanotubes in polymer aqueous two-phase systems. *J. Am. Chem. Soc.* **136**, 10383–10392 (2014).

5.7 Appendix

5.7.1 Dimensionality Reduction by Principal Component Analysis (PCA)

In order to reduce the feature space, we used principal component analysis (PCA) to find the directions of most variation in the dataset. As shown in Figure 5.5a, there is an overlap between the examples for the *complete-feature* set that is labelled as analytes. On the other hands, the examples of the *complete-example* set are well spread along the first and second principal component axes (Figure 5.5b). To train ML model, we used five of the most contributed principal components.

It is worth noting that the direction of the eigenvector of features of DNA sequence is seen to be orthogonal to the direction in which the output data (i.e., HE4 presence) are the most spread out. Based on the observation, we created the feature vector without DNA sequence features.

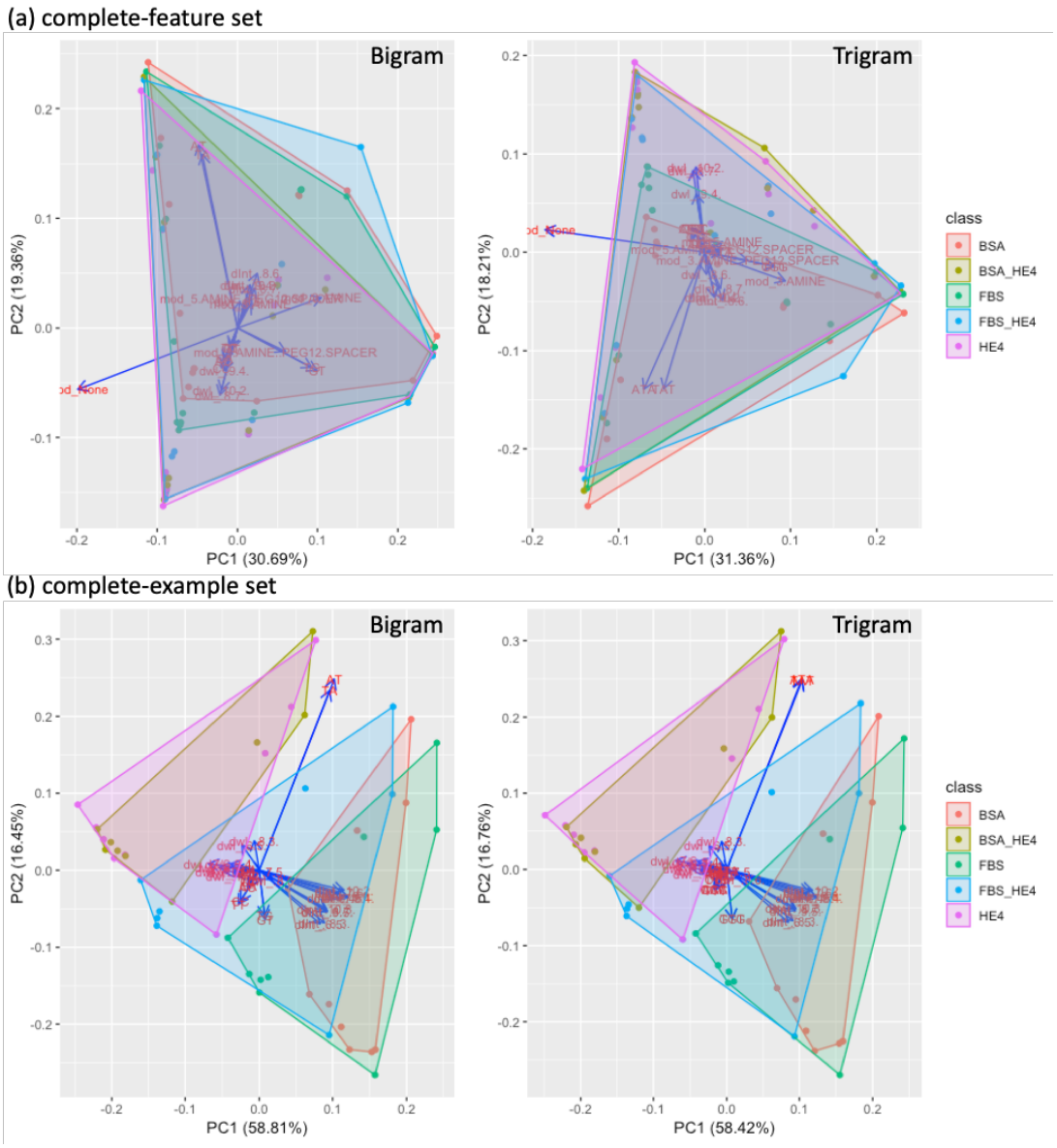

Figure 5.5. Principal component analysis (PCA) of (a) *complete-feature* and (b) *complete-example* set. In *complete-feature* set, the analyte-labelled examples are overlapped, however, in *complete-example* set, the labelled examples are well spread out. Vectors (blue arrow) represent eigenvectors and are scaled to the square root of their eigenvalue.

5.7.2 Raw data of peak shift (dwl_i) and peak intensity change ($dint_i$)

In general, most sequences showed differences in the intensity in presence of HE4 regardless of the presence of BSA and FBS.

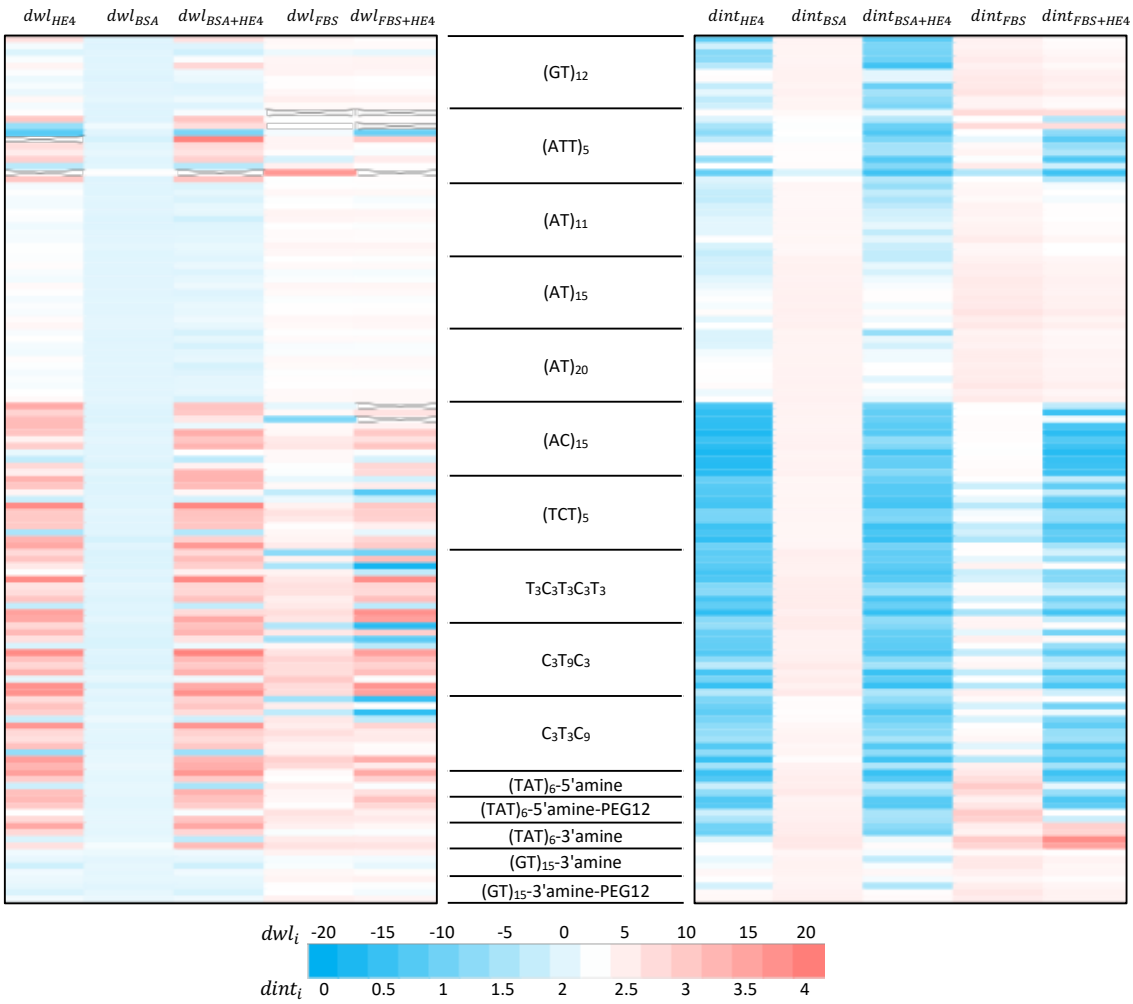

Figure 5.6. Heatmap analysis of peak shift (dwl_i) and intensity changes ($dint_i$) in presence/absence of the analyte. The horizontal axis represents the peak shift (left) and intensity changes (right) for each analyte combination (HE4, BSA, BSA+HE4, FBS, and FBS+HE4). The vertical axis represents each DNA/SWCNT combination. A total of 11 chiral SWCNTs with each DNA sequences were measured except for the modified DNA sequences; (6,5), (8,4), (8,3), (7,5), (7,6), (10,3), (9,5), (9,4), (8,6), (10,2), and (8,7)-SWCNT are shown in each row. Four of chiral SWCNT were measured for the modified sequences; (10,2), (9,4), (8,6), and (8,7)-SWCNT are shown in each row for modified sequences. The colorbar represents the peak shift and intensity change values. Different range was for dwl_i and $dint_i$.

5.7.3 Link to Public Repository:

The following link is to a public repository where we provide a collection of scripts for translation from DNA sequence and optical spectra data to features and machine learning models.

https://bitbucket.org/jagotagrouplehigh/dna_swcnt_mp/

Chapter 6 : Conclusion

6.1 Experimental Characterization of DNA/SWCNT Hybrid

It is well-known that some special DNA/SWCNT hybrids have different ordered structures enabling their use for separation of SWCNT species. Aqueous two-phase (ATP) systems have been used to separate SWCNTs using the difference in partitioning of DNA/SWCNT hybrid in the ATP. The partitioning is determined by the difference in solvation properties which can result from the secondary structure of DNA on SWCNT. In chapter 2, we proposed two ways to extract the solvation properties of DNA/SWCNT using the ATP system: relative solvation free energy in water, $\mu_{w,A}^{o}/\mu_{w,B}^{o}$, and the Hildebrand solubility parameter, δ_i. The results from two different approaches are found to be consistent with each other, providing some confidence in each as a method of quantifying differences in solubility of various DNA/SWCNT hybrids. In chapter 4, we measure the binding characteristics of DNA on SWCNT by observing the kinetics of the SDC exchange process by DNA, aided by the addition of methanol. The activation energies for the DNA rearrangement process were estimated using Eyring kinetics. We found that the activation energies of some recognition sequences present significant difference between (6,5) and (8,3)-SWCNT, while the non-recognition sequence presents no chirality-dependence on the activation energy. We expect that such quantification can provide a basis for data-analytic searches for new sequences.

6.2 SWCNT-Recognition DNA Sequence Prediction

For many years, much effort has been expended on finding SWCNT-recognition sequences. However, this has not lead to the ability to predict recognition sequences. Recently, machine learning applications in bioinformatics have attracted considerable interest because of their ability to transform large amounts of raw sequence data into useful scientific knowledge, without requiring explicit programming instructions. As such, we utilized machine learning techniques to discover new recognition sequences in the vast ssDNA library. In chapter 3, we built machine learning models based on the DNA sequence information. We found a remarkable increase in the frequency of recognition sequences from 10% in the original training set to >50% in the model-predicted sequence sets.

6.3 New Perception-based Sensing System using Machine Learning

Current research on sensing applications has been mostly based on one-to-one recognition. However, it is impossible and inefficient to find receptors for each of the molecule in a complex sample. To detect various molecules simultaneously, an effective and automatic data processing system is essential. In chapter 5, a new perception-based sensing system named *Molecular Perceptron* is proposed. We demonstrated the perception-based sensing system using machine learning techniques to detect the ovarian cancer biomarker HE4 in the presence or absence of other analytes (BSA and FBS). The DNA/SWCNT hybrids were utilized to optically detect the analytes. Encouragingly, the models achieved remarkable trainability with the F^1-score of ~0.95.

6.4 Future work

In chapter 2, we introduced a way to estimate the Hildebrand solubility parameters for DNA/SWCNT hybrid. The Hildebrand solubility parameter considers only dispersion interactions between molecules.[1] For many polymer/solvent pairs, the cohesive energy is also affected by polar group interactions and hydrogen bonding, thus multiple-component concepts, such as Hansen solubility parameters,[2] can be used to extract partial parameters from different contributions, which can help to better understand the intermolecular interactions of hybrids. In order to extract Hansen solubility parameters from ATP system, three solubility parameters are required, so more experimental work is needed to calculate them uniquely.

In chapter 3, we considered the recognition DNA sequence prediction as a binary classification problem despite each pair with a different SWCNT. In addition, it is known that some special sequences have the separability of enantiomers,[3] which means that the recognition sequences can be differed in terms of selectivity/yield. In this respect, the predictive model can be expanded to a resolution-based multi-level classification. More broadly, our approach will provide some insight to the sequence selection problem for bio/nano hybrid materials made of inorganic nanostructures and sequence-defined polymers such as DNA and peptides.

In chapter 4, the preliminary findings suggest that the structure of DNA on SWCNT prepared by the exchange process can be different compared to that prepared by direct sonication, which is related to the recognition ability. The ATP separation technique can be utilized to test the recognition ability of the DNA/SWCNT hybrid. Furthermore,

additional treatment after the exchange process such as mild sonication or incubation at high temperature can be studied systematically to obtain the same structure as that prepared by direct sonication.

In chapter 5, the sensing model was considered as a classification problem. However, from a diagnostic perspective, it is advisable to monitor the concentration of analytes. A multi-target regression model can be considered to predict the analyte concentration. Additional experimental in difference concentration of analytes is required to build the regression model.

Overall, all studies presented in this dissertation can be connected (Figure 6.1). The properties of DNA/SWCNT obtained by experiments in Chapter 2 and 4 could be utilized to build a model to predict new recognition sequences. The predictive model can provide good candidates for the *Molecular Perceptron* system by its high probability of having well-defined secondary structure.

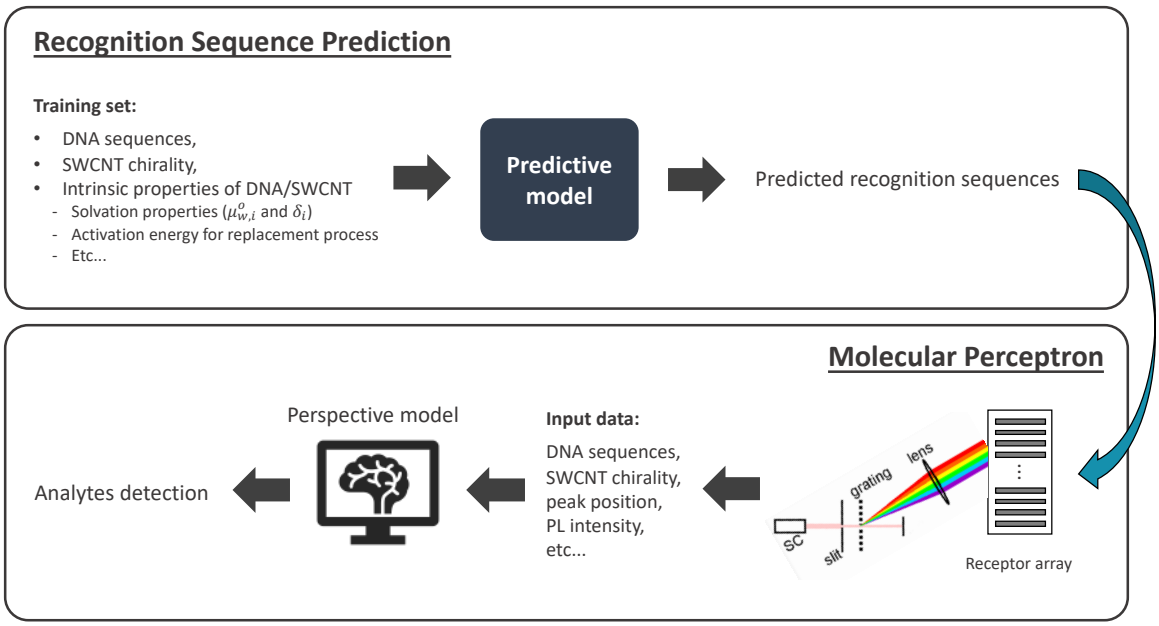

Figure 6.1. Schematic pipeline for future work. It mainly comprises two systems: (top) Recognition sequence prediction and (bottom) Molecular Perceptron. The sequence predictive model can generate well-chosen sequence candidates for Molecular Perceptron.

YOONA YANG

4 Duh Dr. APT 114 yoy214@lehigh.edu
Bethlehem, PA 18015 (201) 281-9044

Education

Lehigh University, Bethlehem, Pennsylvania **August 2019**
Ph.D. Chemical and Biomolecular Engineering (GPA 3.7/4.0)
Thesis Title: Learning about Sequence-Dependent DNA/Single-Wall Carbon Nanotube

Hongik University, Seoul, Republic of Korea August 2013
M.S. Chemical Engineering (GPA 4.5/4.5)
Thesis Title: Numerical Analysis of Reverse Electro-Dialysis (RED) for Electrical Power Generation

Hongik University, Seoul, Republic of Korea August 2011
B.S. Chemical Engineering (GPA 4.1/4.5)

Research Experience

Lehigh University, Bethlehem, Pennsylvania **Aug. 2014 – present**
Graduate research assistant (Advisor: Prof. Anand Jagota)
- Investigate the sequence-specific interaction between DNA and Single Wall Carbon Nanotubes (SWCNTs).
- Discover new recognition sequences based on experimental or physical observation using machine learning techniques.
- Develop an automatic data-processing system to detect various analytes simultaneously using machine learning.
- Quantify DNA/SWCNT hybrid properties by aqueous two-phase (ATP) sorting.
- Conduct molecular dynamic simulation to study on the sequence-specific structure at the molecular level using Gromacs.
- Develop a physical model for the energetics behind DNA based separation using Monte Carlo simulation coupled with Bayesian optimization.

Case Western Reserve University, Cleveland, Ohio June 2018 – Aug. 2018
Visiting research assistant (Advisor: Prof. Roger French)
- Developed data analysis framework for spectra decomposition using principal component analysis (PCA) and partial least square regression (PLSR) on time resolved optical spectra showing displacement by surfactant of DNA on different SWCNT species.
- Developed an automated technique to analyze micro indentation load-displacement curve.

National Institute of Standards and Technology, Gaithersburg, Maryland
Visiting research assistant (Advisor: Dr. Ming Zheng) Sep. 2017 – Nov. 2017
- Conducted single chirality SWCNT separation by an aqueous two-phase (ATP) system.

Korea Institute of Science and Technology (KIST) – Republic of Korea
Research assistant – (Supervisor: Dr. Myung-Suk Chun) Aug. 2013 – June 2014
- Developed model to predict solute sorption behavior in polymer gel.
- Investigated size-dependent cell separation based on hydrodynamic filtration utilizing microfluidic chip.

Hongik University – Republic of Korea Sep. 2011 – July 2013
Graduate research assistant – (Advisor: Prof. Won Sun Ryoo)
- Studied salinity gradient energy for electric power generation utilizing Reverse Electro-Dialysis (RED) and Pressure Retarded Osmosis (PRO).
- Developed inorganic ion exchange membrane by modifying surface of porous alumina membranes.
- Conducted numerical simulation on ion transport characteristics in nano-channel with surface charge.
- Developed an effective method for bitumen production using CO_2.

Hongik University – Republic of Korea Mar. 2010 – Aug. 2011
Undergraduate research assistant – (Advisor: Prof. Won Sun Ryoo)
- Synthesized and characterized self-doped conducting polymers for resistive random-access memory (RRAM) application.

Hongik University – Republic of Korea June 2009 – May 2010
Undergraduate research assistant – (Advisor: Prof. Young Sik Kim)
- Simulated and synthesized novel iridium(III) complexes for organic light emitting devices (OLEDs).

Publications (*: corresponding author)

1. N. Senanayake, **Y. Yang**, R. H. French, A. Jagota, J. L.W. Carter, An Automated Technique to Analyze Micro Indentation Load - Displacement Curve, *Society of Experimental Mechanics*. (*submitted*)
2. **Y. Yang,** M. Zheng, and A. Jagota*, Learning to Predict Single-Wall Carbon Nanotube-Recognition DNA Sequences, *npj Computational Materials*. *5* (*3*), **2019.**
3. **Y. Yang**, A. Shankar, T. Aryaksama, M. Zheng, and A. Jagota*, Quantification of DNA/SWCNT Solvation Differences by Aqueous Two-Phase Separation*, Langmuir, 34* (5), 1834-1843, **2018**.
4. **Y. Yang** and M.-S. Chun*, The effect of chain stiffness on moisture diffusion in polymer hydrogel by applying obstruction-scaling model, *Korea-Australia Rheology Journal*, 25 (4), 267-271, **2013**.
5. H. W. Ham, **Y. A. Yang**, and Y. S. Kim*, Blue Phosphorescent Mono-cyclometalated Iridium(III) Complexes, *Journal of Korean Physical Society, 57* (6), 1695-1698, **2010**.

Patent

1. M.-S. Chun, **Y. Yang**, and D. H. Woo, Method and apparatus for sensing the flow properties of solution in microfluidic channel having multiple branches, *Korea Patent*, May 14, 2015 (*registered #10-1521879*).
2. G.Y. Chung, W. Ryoo, and **Y. Yang**, Power Generating System from Salinity Gradient Energy Utilizing Automatic Self-Reciprocating Pressure Exchange, , *Korea Patent*, April 24, 2014 (*registered #10-1388694*)
3. G.Y. Chung, S.O. Lee, W. Ryoo, T. Lee, and **Y. Yang**, Power-Generating System using Salinity Difference Energy Between Salty Water and Fresh Water, Capable of Recovering Pressure using a Pressure Exchange Device which Includes Two Interlocked Pistons, *Korea Patent*, February 26, 2013 (*registered #10-1239440*).

Awards

1. Best Student Poster Award, "Quantitative Analysis of Aqueous Two Phase Separation of DNA-SWCNT Hybrids", *231st ECS meeting*, Nanocarbon (NANO) Division of the Electrochemical Society, June 2017.
2. Best Poster Presentation Award, "Development of Eco-friendly Bitumen Production using CO2: Swelling Factors and Diffusion Coefficients of CO2 in Bitumen", *KSCT Fall Conference 2012*, The Korean Society of Clean Technology (KSCT), November 2012.
3. Best Poster Presentation Award, "Numerical analysis of ion transport in stacked cells of reverse electro-dialysis for generating electricity", *KIChE Fall Meeting 2011*, The Korean Institute of Chemical Engineers (KIChE), October 2011.

Selected Conference Presentations (*: corresponding author)

1. **Y. Yang***, M. Zheng and A. Jagota, Learning How to Predict SWCNT-Recognition DNA Sequences, *2018 AIChE Annual Meeting*, Pittsburgh, PA, October 2018.
2. **Y. Yang***, A. Cruz, D. Gordon, R. French, M. Zheng, and A. Jagota, Applying Machine Learning Techniques to Prediction and Data Analysis of DNA/SWCNT Sequence-dependent Interaction, *CWRU/Kyocera Materials Data Science Symposium 2018*, Cleveland, OH, August 2018.
3. **Y. Yang***, M. Zheng and A. Jagota, Learning DNA/SWCNT Recognition Sequences, *233rd ECS Meeting*, Seattle, WA, May 2018
4. **Y. Yang**, A. Shankar, T. Aryaksama, M. Zheng, and A. Jagota*, Quantitative Analysis of Aqueous Two Phase Separation of DNA-SWCNT, *231st ECS Meeting*, New Orleans, LA, May 2017
5. M.-S. Chun*, **Y. Yang**, Unsteady electrokinetic microfluidics with hydrodynamic slippage effect, *March Meeting of The American Physical Society*, Denver, CO, March 2014.
6. **Y. Yang**, D. Y. Lee, and M.-S. Chun*, Chain Properties and Hindered Moisture Diffusion in Polymer Hydrogel Based on Obstruction-Scaling Model, *The 9th*

International Workshop for East Asian Young Rheologists, Seoul, Republic of Korea, February 2014.

7. K. Yoon, **Y. Yang**, M.-S. Chun*, and H. W. Jung*, Electrolytic non-Newtonian Fluids in a Curved Microchannel with Charged Wall, *The 9th International Workshop for East Asian Young Rheologists*, Seoul, Republic of Korea, February 2014.
8. I. Yim, **Y. Yang**, and W. Ryoo*, Performance Analysis of Reverse Electrodialysis (RED) Cell Stacks Utilizing Equivalent Circuit Model (ECM), *KIChE Fall Meeting 2013*, Daegu, Korea, October 2013.
9. **Y. Yang** and W. Ryoo*, Numerical analysis of selective ion transport in nano-channels with charged surfaces, *87th ACS Colloid & Surface Science Symposium*, Riverside, USA, June 2013.
10. **Y. Yang**, S. Oh, W. Bae, and W. Ryoo*, Development of Eco-friendly Bitumen Production using CO_2: Swelling Factors and Diffusion Coefficients of CO_2 in Bitumen, *KSCT Fall Conference 2012*, Daegu, Korea, November 2012.
11. Y. Kim, **Y. Yang**, and W. Ryoo*, Lumped parameter analysis of Reverse Electro-Dialysis (RED) cell performance based on equivalent circuit model, *KSCT Fall Conference 2012*, Daegu, Korea, November 2012.
12. **Y. Yang**, G. Y. Chung, M.-S. Chun, and W. Ryoo*, Preparation and characterization of functionalized alumina membranes for ion exchange applications, *12th International Conference on Inorganic Membranes*, Enschede, Netherlands, July 2012.
13. **Y. Yang**, B. Oh, G.Y. Chung, and W. Ryoo*, Mechanical Power generation utilizing Pressure Retarded Osmosis (PRO), *KIChE Fall Meeting 2011*, Incheon, Korea, October 2011.
14. **Y. Yang**, S. Kim, S.O. Lee, G.Y. Chung, and W. Ryoo*, Numerical Simulation of Ion Transport in Reverse Electrodialysis (RED) Cell for Electrical Power Generation, *Tech Connect World Conference & Expo 2011*, Boston, USA, June 2011.
15. **Y. Yang** and W. Ryoo*, Synthesis and Characterization of Self-doped Conductive Polymers, *KIChE Spring Meeting 2011*, Changwon, Korea, April 2011.
16. **Y. Yang**, S. Kim, S.O. Lee, G.Y. Chung, and W. Ryoo*, Numerical Analysis of Ion Transport in Stacked Cells of Reverse Electro-dialysis for Generating Electricity, *KIChE Spring Meeting 2011*, Changwon, Korea, April 2011.

Research Technique Proficiency

Experimental Skills

- **Spectroscopic analysis:** UV-vis-NIR spectroscopy, Photoluminescence (PL) spectroscopy, Circular Dichroism (CD) spectroscopy, Fourier Transform Infrared (FT-IR) spectroscopy, and Nuclear magnetic resonance (NMR) spectroscopy
- **Electrical/Electrochemical characterization:** Cyclic voltammetry (Potentiostat), Electrochemical Impedance Spectroscopy (EIS), pH/Conductivity meter, Multimeter, Power Supply, and Electric Load
- **Microscopy:** Scanning Electron Microscopy (SEM) and Atomic Force Microscopy (AFM)

- Sorting single species of SWCNT using aqueous two-phase (ATP) separation
- Design and Fabrication of microfluidic chip by using photolithography process
- Synthesis (Suzuki-Miyaura reaction) and surface functionalization (silanization and plasma treatment) of polymer, membrane, or glass
- Fabrication and electrical characterization of lab-scale Random Access Memory (RAM) and Field Effect Transistor (FET)
- Deposition of metal using magnetron sputter and thermal evaporator
- Chemical analysis using Gas Chromatography (GC)
- Interfacial energy analysis using Contact Angle Meter

Computational Skills

- Experience with **machine learning/bioinformatics** techniques (SciPy, keras, and variable libraries in R) and **big data** technology (Hadoop)
- **Molecular simulation:** Molecular modeling by calculating with Density Functional Theory (DFT) utilizing Gaussian 98; molecular dynamics modeling using Gromacs; coarse-grained model using Monte Carlo simulation coupled with Bayesian optimization
- **Numerical analysis:** Finite-difference Method (FDM) and Finite-element Method (FEM) utilizing Matlab and COMSOL Multiphysics
- **Programming languages:** R, Python, and Matlab
- **Other software applications:** Git, Microsoft Office (including Visio and Front Page), Chembio Office, LabVIEW, Origin, SigmaPlot, AutoCAD, Adobe Photoshop, and Adobe Illustrator

Teaching Experience

- Teaching assistant for *[ChE 415] Transport processes* for PhD students, Lehigh University, Spring 2019
- Teaching assistant for *[BioE 210] Introduction to Engineering Physiology* (including laboratory work) for undergraduate students, Lehigh University, Spring 2017
- Teaching assistant for *[ChE 400] Chemical Engineering Thermodynamics* for PhD students, Lehigh University, Fall 2016
- Teaching assistant for *[BioE 110] Elements of Bioengineering* for undergraduate students, Lehigh University, Fall 2015
- Teaching assistant for *[ChE 201] Methods of Analysis in Chemical Engineering* for undergraduate students, Lehigh University, Fall 2014
- Teaching assistant for *Chemical Engineering Laboratory. II* for undergraduate students, Hongik University, Spring 2012
- Teaching assistant for *Chemical Engineering Laboratory. I* for undergraduate students, Hongik University, Fall 2011