MACHINE LEARNING PREDICTION AND CLASSIFICATION OF TRANSMISSION FUNCTIONS FOR RAPID DNA SEQUENCING IN HYBRID NANOPORE

M.Sc. Thesis

By

SOUPTIK PANDIT

DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2024

MACHINE LEARNING PREDICTION AND CLASSIFICATION OF TRANSMISSION FUNCTIONS FOR RAPID DNA SEQUENCING IN HYBRID NANOPORE

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of
Master of Science

by
SOUPTIK PANDIT

DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2024

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled "**Machine Learning Prediction and Classification of Transmission Functions for Rapid DNA Sequencing in Hybrid Nanopore"** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **DEPARTMENT OF CHEMISTRY, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2022 to May 2024 under the supervision of **Dr. BISWARUP PATHAK**, Professor, Department of Chemistry, IIT Indore.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Souptik Pandit
17.05.2024

Signature of the student with date
Souptik Pandit

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

Signature of the Supervisor of
M.Sc. thesis
Prof. Biswarup Pathak

Souptik Pandit has successfully given his/her M.Sc. Oral Examination, held on May 8, 2024.

Convener, DPGC
Date:

ACKNOWLEDGEMENTS

I want to convey my sincere gratitude to my supervisor, **Prof. Biswarup Pathak**, Department of Chemistry, IIT Indore, for providing me with the opportunity to work on this M.Sc. project work under his esteemed supervision.

I am deeply indebted to **Milan Kumar Jena**, whose guidance and mentorship have been pivotal in my learning journey. His insightful feedback and unwavering support have been instrumental in helping me throughout this one-year-long M.Sc. research project. Further, I would like to extend my gratitude to other group members for the supportive environment, insightful discussions, and valuable suggestions.

Finally, I am eternally grateful to my family and friends for their continuous encouragement, and moral support.

Souptik Pandit

Dedicated to my mother

Abstract

Electrical DNA sequencing using solid-state nanopores has emerged as a promising technology due to its potential to achieve high-precision single-base resolution. However, uncontrollable nucleotide translocation, low signal-to-noise ratios, and electrical signal overlapping from nucleotide stochastic motion have been major limitations. Recent fabrication of in-plane hybrid heterostructures of 2D materials has triggered active research in sequencing applications due to their interesting electrical properties. Herein, our study explores both machine learning (ML) regression and classification framework for single DNA nucleotide identification with hybrid graphene/hexagonal boron nitride (G/h-BN) nanopore using a quantum transport approach. The optimized ML model predicted each nucleotide at their most stable configurations with the lowest root-mean-squared error of 0.07. We have also examined the impact of three locally polarised hybrid nanopore environments ($\mathrm{C}^{\delta-} - \mathrm{H}^{\delta+}$, $\mathrm{N}^{\delta-} - \mathrm{H}^{\delta+}$, and $\mathrm{B}^{\delta+} - \mathrm{H}^{\delta-}$) on ML prediction of transmission functions utilizing structural, chemical, and electrical environmental descriptors. The random forest algorithm demonstrates notable classification accuracy across quaternary (~86%), ternary (~95%), and binary (~98%) combinations of four nucleotides. Further, we checked the applicability of the hybrid nanopore device with conductance sensitivity and frontier molecular orbitals analysis. Our study showcases the potential of a hybrid nanopore with ML combined quantum transport method as a promising sequencing platform that paves the way for advancements in solid-state nanopore sequencing technologies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Deoxyribonucleic Acid (DNA)

Deoxyribonucleic acid (DNA) is the primary genetic material that stores and transmits hereditary information in nearly all living organisms. DNA is a polymer that consists of two antiparallel polynucleotide strands coiled into a double-helix structure. Each strand is composed of a sugar-phosphate backbone with attached nucleobases. In DNA, the sugar moiety is 2-deoxyribose, and the phosphate groups form the phosphodiester linkages that connect the nucleotides. The four nitrogen-containing nucleobases are adenine (A), thymine (T), cytosine (C), and guanine (G). Adenine and guanine are purine derivatives, while cytosine and thymine are pyrimidine derivatives. Based on the structural characteristics, each nucleobase can form multiple H-bonding with its complementary base pair. The adenine (A) nucleobase pairs with thymine (T) with two H-bonding interactions (A═T) whereas cytosine (C) pairs up with guanine(G) with three H-bonding interactions (C≡G). The complementary base pairing and the helical structure of DNA are critical for its ability to self-replicate during cell division. The sequence of these nucleotides encodes the genetic information required for the development and function of all known living organisms. The structure of double-stranded (ds DNA) has been elucidated in detail, as shown in **Figure 1.1**. The specific sequence of these nucleobases in DNA can carry a broad range of biological and genetic information for protein expression at the molecular level. This DNA sequence provides the blueprint for life. Decoding these DNA sequences, known as DNA sequencing, is pivotal for understanding gene and protein expression by associating genomic and proteomic data*[1]*.

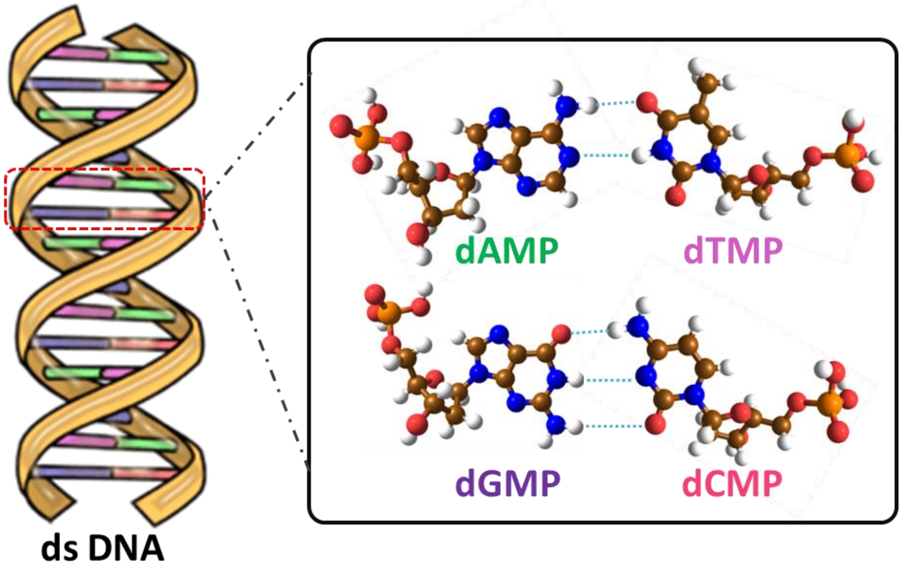


Figure 1.1: Illustration of the Watson and Crick model of dsDNA with their complementary base pair units (adenine (A), thymine (T), cytosine (C), and guanine (G)). This helical structure is stabilized by the H-bonding interaction between purine and pyrimidines ($\mathrm{A} = \mathrm{T}$ and $G \equiv C$).

DNA sequencing can help to identify genetic disorders, cancer, viral mechanisms, and antibiotic resistance. Early diagnosis, personalized treatment, and the prevention of genetic disorders become possible through the insights gained from DNA sequencing*[2]*. Ultimately, unraveling the genetic information stored in DNA sequences can lead to significant advancements in healthcare, and medicine. The mapping of the complete human genetic blueprint, achieved by the National Human Genome Research Institute (NHGRI) under the National Institutes of Health (NIH) in 2004, marked a groundbreaking accomplishment, after which the USA initiated the "$1000 Genome" project. Over the past decades, a huge number of DNA sequencing methodologies have evolved to achieve sub-$1000 genome sequencing costs. The reported DNA sequencing methods are classified into four distinct generations. In this chapter, we briefly discuss DNA sequencing using 2D solid-state nanoscale devices.

Nanopore sequencing, also known as fourth-generation DNA sequencing, is a technique that utilizes nanopores to sequence DNA nucleotides. The concept of nanopore sequencing was first proposed in the 1990s by researchers like Deamer, Church, Kasianowicz, and Branton*[3]*. In 1996, Deamer *et al.* demonstrated the translocation of DNA through a biological (α-hemolysin) nanopore/nanochannel*[4]*. When DNA or RNA oligomers are threaded through the nanopore, they produce distinct ionic current blockade signals that vary with respect to the translocation time. Researchers have been able to distinguish between purines and pyrimidines, as well as the four individual nucleobases, by analyzing these ionic current signals. In 2012, Gundlach and co-workers demonstrated DNA sequencing with single-nucleotide resolution using this approach*[5]*. In the same year, the first commercial nanopore sequencing device, the MiniIon from Oxford Nanopore Technologies, was developed and released at a price of $900. This technique has shown great potential for single DNA nucleotide sequencing and has rapidly advanced in the past decade.

In the early stages, the primary focus was on monitoring the ionic current through biological nanopores. This approach continues to be actively investigated for its advantages in nanopore sequencing. However, the ionic blockade current signals measured at any given point originate from DNA strands simultaneously residing in the nanopores/nanochannels. This requires the use of complex post-data processing algorithms to resolve the sequence information accurately. Consequently, techniques based on ionic blockade current have yet to achieve the same level of accuracy as Sanger sequencing*[6]*. A key limitation of ionic blockade current sequencing is that it requires the integration of a DNA polymerase to slow down the translocating DNA strands. This results in limited read lengths, making the technique sensitive to the nature of the translocating DNA nucleotides over the nanopores/nanochannels. In other words, the ionic blockade current sequencing approach suffers from the drawback of restricted read lengths, which can impact the overall sequencing performance and accuracy.

However, Lagerqvist *et al.* proposed measuring transverse electrical current perpendicular to the translocating DNA backbone for the identification of single nucleotides*[7]*. In quantum transport study, electrodes are embedded at the two ends of the nanopore device, and electronic current is measured for individual identification of DNA nucleotides. In 2010, a graphene-based nanogap was proposed by Postma and co-workers for DNA sequencing by measuring transverse tunneling current*[8]*. Apart from graphene, a wide range of solid-state nanopore devices have been designed and employed for DNA sequencing applications. For fabricating two-dimensional (2D) nanopore-based sequencing devices, materials like graphene, molybdenum disulfide, and boron nitride have been significantly investigated.

1.2 DNA Sequencing Using Solid-State Nanodevices

Recent years have witnessed remarkable progress in nanoscale DNA sequencing technologies, driving down the cost of genome sequencing. **Figure 1.2** depicts the schematic representation of different available solid-state nanoscale devices. Four new concepts have been developed using solid-state nanostructure for the sequencing of DNA and protein molecules. **Figure 1.2a** shows the ionic current method that can be directly measured when a ssDNA passes through the nanometer-sized solid-state nanopore. The DNA sequence is then decoded by measuring the relative changes of ion-current with time. An alternative technique is to determine the changes in transverse tunneling conductance of electrons when a ssDNA/dsDNA translocate through a solid-state nanopore or nanogap. In 2010, several independent research groups reported that the dsDNA can be translocated through atomically thin graphene nanopores *[9,10]*. However, a low signal-to-noise ratio and rapid translocation speed of the nucleotide molecules are limitations of such next-generation sequencing methodologies. Besides, in graphene-based nanopores, rapid clogging of nucleotides is observed due to strong π-π interaction that may result in nanopore blockage and irreversible nanopore-closure.

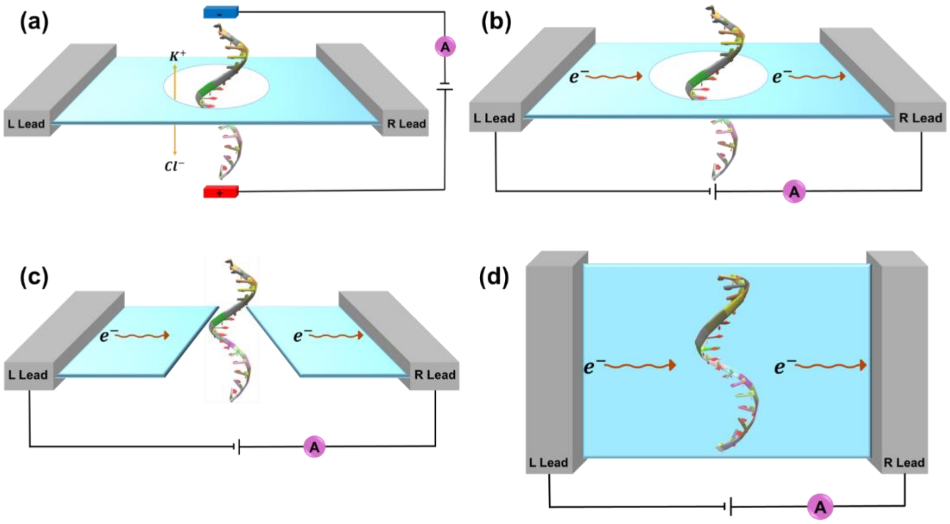


Figure 1.2: Schematics of solid-state nanostructure-based (a) nanopore to measure ionic current, (b) nanopore to measure transverse tunneling current, (c) nanogap to measure transverse tunneling current, (d) a graphene nanochannel film to measure the current due to the physical sorption of DNA bases.

Apart from the ionic current method, the quantum transport approach for DNA sequencing is well investigated technique that utilizes the principles of quantum tunneling of electrons to detect and identify individual nucleotides within a DNA molecule. Apart from the ionic current method, the quantum transport approach for DNA sequencing is a well-investigated technique that utilizes the principles of quantum tunneling of electrons to detect and identify individual nucleotides within a DNA molecule. **Figure 1.2b** describes a nanopore device that utilizes quantum transport, where the in-plane electron tunneling current is monitored as DNA strands transverse through the nanopore. **Figure 1.2c** describes a nanogap device that utilizes quantum transport, with the in-plane current, and transmission is monitored as DNA strands transverse through the nanogap. Additionally, **Figure 1.2d** also describes a nanochannel device that utilizes quantum transport. In this technique, the electronic properties of the solid-state nanostructure (nanopore, nanogap, or nanochannel) are used as direct parameters to facilitate DNA sequencing. Experimentally, the electronic conductance through these narrow solid-state nanostructures is determined when a DNA nucleobases translocate through them. Through multiple theoretical studies, Kim and co-workers proposed methodologies for identifying individual nucleotides utilizing a graphene nanoribbon-based nanochannel device*[11–15]*. This device exploits the π-π interactions and highly sensitive Fano resonance-driven conductance characteristics unique to each nucleobase, enabling the detection of single nucleotides. Moreover, reports indicate the feasibility of detecting individual nucleobases through conductance measurements utilizing narrow semiconducting nanoribbons (graphene, MoS_2, silicene, and hexagonal BN), which experience dips in the conductance curve in the presence of the nucleobases *[15]*. Similarly, Amorim *et al.* computationally proposed a 2D silicene-based device as an electrical biosensor*[16]*.

1.3 Conclusions

In summary, the quantum transport approach for DNA sequencing presents a promising methodology for single nucleotide identification. By harnessing the principles of quantum tunneling and exploiting the unique electronic properties of nanoscale structures, such as nanopores, nanogaps, and nanochannels, this technique offers the potential for highly sensitive and accurate DNA sequencing. In this study, we investigate a hybrid graphene/hexagonal boron nitride (G/h-BN) nanopore using quantum transport approach for single DNA nucleotide identification. Besides, we will also implement machine learning (ML) regression to predict the fingerprint transmission function and employ classification algorithms to distinguish single nucleotides based on their transmission characteristics.

Chapter 2

Review of Past Work and Problem Formulation

Solid-state nanopore-based next-generation sequencing (NGS) technique has garnered significant interest over the past few decades for various applications, including personalized medicine*[17,18]*. The electrical identification of nucleotides with the quantum transport approach has opened up a new horizon in genomics by rendering it an economical and resilient substitute to traditional alternatives for fast and high-precision DNA sequencing*[19,20]*. The identification of DNA nucleotides is already demonstrated at the single-molecular level by monitoring transverse conductance readouts in graphene nanogap*[8]*. Owing to its single atomic thickness and superior electronic properties, graphene nanopores have been extensively investigated both experimentally*[9,10,21]* and theoretically*[22–24]* for their potential use as electrodes compared to other two-dimensional materials. However, experimental challenges associated with fast translocation and low signal-to-noise ratios have prompted researchers to explore other nanopore devices*[25,26]*. The designing of axisymmetric nanopores and precise control of their size and geometry can be an effective way to manipulate the interaction of nucleotides with the nanopore edges*[27]*. To control the dimension of the nanopore in the sub-10 nm range, experimental techniques such as the electrochemical reaction and dielectric breakdown approach have proven to be handy tools for regulating the extent of interaction between nucleotides and nanopore edges*[28,29]*. Unlike graphene nanopores, experimental reports on boron nitride (BN) nanopore exhibited improved coupling interactions with dsDNA resulting in prolonged translocation time that can resolve sensitivity issues*[30]*.

Thanks to the minor (∼1.8%) difference in graphene and h-BN lattice parameters, the search for better nanoscale devices has led to in-plane hybrid graphene/hexagonal boron nitride (G/h-BN) heterostructures*[31,32]*. The hybrid G/h-BN material is a combination of two interesting 2D materials with impressive structural and electronic properties*[33]*. These in-plane heterostructures can address the limitations associated with graphene nanopores by exhibiting enhanced sensitivity and selectivity towards the target molecule due to the electronic confinement of local current*[34–36]*. These G/h-BN nanopores are also capable of providing distinguishable currents at lower applied voltages for each nucleotide as compared to previously reported G/h-BN nanogaps *[37,38]*. Moreover, hybrid G/h-BN devices are also reported to be experimentally synthesized using the chemical vapor deposition technique and topological substitution reactions with full control over the specificity, size, and composition of h-BN in heterostructures*[32,39–42]*. The oppositely polarised nanopore edges due to the presence of $\mathrm{B}^{\delta+}-\mathrm{H}^{\delta-}$, $\mathrm{N}^{\delta-}-\mathrm{H}^{\delta+}$and $\mathrm{C}^{\delta-}-\mathrm{H}^{\delta+}$ functionalized groups can significantly affect the conductance readouts by modulating coupling strength with the polar nucleotide molecules located inside the G/h-BN pore. Besides, the temporal hydrogen bonding interactions accompanied with dipolar coupling can also impart a combined effect resulting in a sharp change of transmission signals that offer both better sensitivity and selectivity towards nucleotides*[38]*.

Recent advancements in machine learning (ML) and neural networks have rapidly transformed the identification of biomolecules in nanopore sequencing by identifying peak positions and reducing noise in the calculated fingerprint transmission function *[43–46]*. Taniguchi *et al.* have reported an ML-aided artificially intelligent solid-state nanopore for the detection of single nanoparticles *[47]*. In addition, several other studies have thoroughly examined the application of ML algorithms on electric signals for signal identification and biomolecule detection *[48–51]*. Accurate prediction of fingerprint transmission readouts has also emphasized the relevance of using ML in the identification of single nucleotides and amino acids via the transverse quantum transport approach *[52–55]*.

Driven by the remarkable progress in ML techniques for single molecule-based DNA sequencing, we aim to investigate the meticulously crafted G/h-BN nanopore and the role of different electronic environments (C-H, B-H, and N-H) on the nucleotide transmission functions with ML integrated quantum transport approach. We have also emphasized on understanding the role of those local termination environments as descriptors in predicting the signature transmission function of the energetically favorable configuration of individual nucleotides. Precise prediction of the transmission data for the most stable rotational orientation and SHAP interpretability of ML models can provide valuable insights into the correlations between transmission function and the local electronic environment of the hybrid nanopore. The categorization of DNA nucleotides with multiclass ML classification from their overlapped transmission readouts is also investigated. The combined application of ML regression and classification on quantum transport results of hybrid G/h-BN nanopore can be interesting for the efficient identification of nucleotides.

Chapter 3

Theoretical Background

Computational modeling and calculations have wide applications in the fields of chemistry, physics, and material science. These calculations utilize first-principles density functional theory (DFT) to accomplish structure optimization and further calculation. Owing to the accuracy and computational efficiency of DFT calculations, this methodology comes with more reliable and practical choice while working with many body complex systems. In this chapter, we present an overview of the main theoretical frameworks utilized throughout this thesis. Furthermore, we will briefly discuss the key aspects of DFT as applied to atomic-scale simulations and calculations, particularly focused on electronic structure (or nuclear structure) and the determination of ground state properties. This comprehensive coverage of the underlying theoretical foundations sets the stage for the subsequent discussions.

3.1 The Many-Body Problem

The time-independent Schördinger equation is exactly solvable for one electron containing H-atom or H-like atoms (i.e., He^{+}, Li^{2+}, and Be^{2+}) which are two body problems. But practically, the materials that constitute our physical world from tiny molecules to solids, liquids, or gases are fundamentally composed of multiple electrons and atomic nuclei. To describe the ground state electronic structure of a material, we must understand the principles of quantum mechanics and comprehend the complex interactions between electrons and atomic nuclei in many-body problems. The time-independent Schördinger equation for many-body problem can be written as follows-

$$\hat{H}\Psi(r_1, r_2, \dots, r_i, R_1, R_2, \dots, R_I) = E\Psi(r_1, r_2, \dots, r_i, R_1, R_2, \dots, R_I) \qquad (3.1)$$

Where$\Psi(r_1, r_2\ , r_3, \dots, r_i, R_1, R_2, R_3, \dots, R_I)$ is the many-electron wave function. The r_i represents the position vector of i^{th} electron and R_I represents the position vector of I^{th} nuclei. E is the total energy eigenvalue of the whole system, and H is the total Hamiltonian of the system consisting of kinetic and potential energy operators (in atomic units)

$$\hat{H} = \underbrace{-\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2}_{\hat{T}_e} - \underbrace{\frac{\hbar^2}{2}\sum_I \frac{\nabla_I^2}{M_I}}_{\hat{T}_n} + \underbrace{\underbrace{\frac{1}{2}\sum_{i\neq j}\frac{e^2}{4\pi\varepsilon_0 r_{ij}}}_{\hat{V}_{ee}} - \underbrace{\sum_{i,I}\frac{e^2 Z_i}{4\pi\varepsilon_0 R_{iI}}}_{\hat{V}_{en}} + \underbrace{\frac{1}{2}\sum_{I\neq J}\frac{e^2 Z_I Z_J}{4\pi\varepsilon_0 R_{IJ}}}_{\hat{V}_{nn}}}_{\hat{V}} \qquad (3.2)$$

The Z_I represents the number of protons in the I^{th} atom. The m_e and M_I are the mass of the i^{th} electron and I^{th} nuclei, respectively. In **Equation 3.2**, the first two terms ($\hat{T}_e$ and $\hat{T}_n$) are the kinetic energy of the electron and nuclei, and the later three terms ($\hat{V}_{ee}$, $\hat{V}_{en}$, $\hat{V}_{nn}$) express the potential energy due to the Columb interaction between electrons and nuclei. Whereas, $\hat{V}_{ee}$ and $\hat{V}_{nn}$ represents the interelectronic and the internuclear repulsion respectively.

The solution of Schördinger (**Equation 3.2**) is complicated and can only be solved accurately for small systems like H-like atoms as mentioned previously. Hence to deal with such complicated many-body systems like solids consisting of huge numbers of atoms, we need to come up with some valid approximations.

The mass of one proton is approximately 1836 times higher than the mass of a single electron. Consequently, the motion of nuclei is extremely slow compared to the movement of electrons. This significant disparity in mass allows us to consider the nuclei as effectively static relative to the dynamics of electrons. This fundamental principle forms the basis of the Born-Oppenheimer approximation, which permits the multiplicative separation of degrees of freedom for the electronic and nuclei. Therefore, the total wavefunction (Ψ) in **Equation 3.1** can be expressed as a product of wavefunctions for electrons and nuclei, as given in **Equation 3.3**.

$$\Psi(r_1, r_2, \dots, r_i, R_1, R_2, \dots, R_I) = \Psi(r_1, r_2, \dots, r_i)\Psi(R_1, R_2, \dots, R_I) \qquad (3.3)$$

After simplification (**Equation 3.3**), the initial many-body problem completely reduces to an electronic problem where the coordinates of nuclei are entered only as a parameter. Considering the motion of nuclei independent of electronic motion, the last term of **Equation 3.2** can be regarded as constants. The kinetic energy of nuclei ($\hat{T}_n$) and potential energy for internuclear repulsion ($\hat{V}_{nn}$) terms can be removed from **Equation 3.2** to get the relatively simplified form of energy Hamiltonian (**Equation 3.4**).

$$\hat{H} = \underbrace{-\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2}_{\hat{T}_e} + \underbrace{\underbrace{\frac{1}{2}\sum_{i\neq j}\frac{e^2}{4\pi\varepsilon_0 r_{ij}}}_{\hat{V}_{ee}} - \underbrace{\sum_{i,I}\frac{e^2 Z_i}{4\pi\varepsilon_0 R_{iI}}}_{\hat{V}_{en}}}_{\hat{V}} \qquad (3.4)$$

Here kinetic energy of the nuclei is considered to be negligible. However, the potential energy due to inter-nuclear repulsion between nuclei contributes to the total energy. After considering the Born-Oppenheimer approximation, the overall number of degrees of freedom of the system can be reduced to only an electronic problem. However, the solution of Schördinger equation is still very difficult as the electronic repulsion term is hard to handle for a system with a huge number of electrons. As a more practical approach, it is convenient to consider the electron density instead of individual electron coordinates. In the following section, we describe the Density Functional Theory (DFT) approach, which involves the many-body energy Hamiltonian as a function of electron density, rather than the many-body wavefunctions.

3.2 Density Functional Theory

The fundamental idea behind density functional theory (DFT) is founded on the premise that the properties of interacting electron systems can be described as a functional of the ground state electron density, rather than relying on the complex many-body wavefunctions. The origins of DFT can be traced back to the early work of Thomas and Fermi, who proposed a quantum mechanical theory for illustrating the electronic structure of many-body systems based on the non-interacting homogeneous electron density, known as the Thomas-Fermi model. However, this approximation is unable to accurately describe the electrons in the many-body system due to its inherent limitations. The DFT formalism as we know it today was later developed by Kohn and Hohenberg, who established the framework of the two Hohenberg-Kohn (H-K) theorems. The H-K theorems introduced an exact theory for interacting many-body systems, providing a solid foundation for the use of electron density as the central quantity, rather than the many-body wavefunction. This revolutionary approach significantly simplified the treatment of complex quantum systems, as the electron density contains all the necessary information to determine the ground state properties of the system.

3.2.1 Hohenberg-Kohn Theorem

The Hohenberg-Kohn founded the fundamental basis of DFT showing that the properties of interacting systems can be calculated using the ground-state electron density.

> **Theorem I** *"For any system of interacting particles in an external potential V_{ext}(r), the potential V_{ext}(r) is determined uniquely, up to a constant, by ground-state electron density, n_0(r)."* **[$\boldsymbol{n_0}$(r) → $\boldsymbol{V_{ext}}$(r)]**
>
> **Theorem II** *"In any quantum state the external potential, V_{ext}(r), determine uniquely the many-body electronic wavefunction."* **[$\boldsymbol{V_{ext}}$(r) → $\boldsymbol{\Psi}$(r)]**
>
> **Theorem III** *"In any quantum state of total energy, E, is a functional of many body wavefunction."* **[$\boldsymbol{\Psi}$(r) → $\boldsymbol{E}$]**

From the above-mentioned theorems, the energy can be expressed as the functional of electronic density as shown in **Equation 3.5**

$$E_{HK}[n] = F_{HK}[n] + \int drV_{ext}(r)n(r) \tag{3.5}$$

where the internal energies consisting of kinetic and potential energies, are expressed in terms of F_{HK}. The total internal energy functional can be written as follows (**Equation 3.6**).

$$E_{HK}[n] = T[n] + E_{int}[n] \qquad (3.6)$$

Interestingly, **Equation 3.6** is independent of the external potential (V_{ext}) and solely depends upon the density of electron. The total internal energy is the sum of the kinetic energy functional ($T[n]$) and the internal energy functional ($E_{int}[n]$).

While the theorem establishes the existence of a functional form, its exact nature remains unknown. Therefore, we need relevant approximations to solve the problems. The global minima of the functional form (**Equation 3.6**) expresses the precise ground-state total energy of the system, with the corresponding electron density representing the exact ground-state electron density (n_0(r)). The variational principle, as depicted in **Equation 3.7**, can be employed to determine the ground-state electron density (n_0(r)).

However, in the theorem, there is no way out to determine the exact functional form. Therefore, it has to be approximated to apply to practical problems. The global minima of the functional form (**Equation 3.6**) is E_0 which describes the exact ground-state total energy of the system and the corresponding electron density would be the exact ground-state electron density n_0(r). The variational principle can be used to acquire electron density(n_0(r)) for the ground-state as shown below in **Equation 3.7**.

$$\frac{\delta}{\delta n} E_{HK}[n(r)]|_{n=n_0} = 0 \qquad (3.7)$$

The H-K theorems have proved to be the foundational pillars of modern DFT formalism, describing the relation between the ground-state electron density and the ground-state energy of a many-body system. Further, Kohn-Sham (K-S) came up with the idea of a non-interacting reference system constructed from a series of orbitals. This reference system enables the calculation of a significant portion of the kinetic energy with a high degree of accuracy.

3.2.2 Kohn-Sham Formalism

The Hohenberg-Kohn (H-K) theorems offer a foundational framework for addressing the many-body problem by utilizing the particle density function and the variational principle. However, for practical applications involving particles, the Density Functional Theory as realized through the Kohn-Sham (K-S) approach is more commonly used. The central concept supporting the H-K theorems is to substitute the interacting electron system with an auxiliary system of non-interacting particles that possess an identical electron density distribution. This allows the total energy functional to be expressed as shown in **Equation 3.8**.

$$E_{KS}[n] = T_S[n] + \int dr\, V_{ext}(r)n(r) + \frac{1}{2}\int dr\, dr'\, \frac{n(r)n(r')}{|r-r'|} + E_{xc}[n] \tag{3.8}$$

In **Equation 3.8**, the kinetic energy functional of a non-interacting electron gas system is represented by $T_S[n]$. The external potential is expressed as a contribution due to nuclei and another external potential ($\int dr\, V_{ext}(r)n(r)$), and the classical Coulomb potential for the interelectronic interaction is expressed as ($\frac{1}{2}\int dr\, dr'\, \frac{n(r)n(r')}{|r-r'|}$) which is called the Hartree potential. The final term, $E_{xc}[n]$,in the expression captures all the many-body effects arising from exchange and correlation interactions, which is known as the exchange-correlation functional. The exact analytical expression of these exchange-correlation functionals ($E_{xc}[n]$) is yet to be determined. In **Equation 3.8**, the Coulomb repulsion term due to internuclear repulsion is contributed directly as a constant term in the final energy expression.

According to the second H-K theorem, the solution for the Kohn-Sham (K-S) auxiliary systems can be obtained by minimizing the K-S energy functional with respect to the electron density [n(r)]. This minimization of the total energy is achieved by employing a Schrödinger-like equation, as presented in **Equation 3.9**.

$$\hat{H}_{KS}\Psi_i(r) = \left[-\frac{\hbar^2}{2m_e}\nabla^2 + V_{KS}(r)\right]\Psi_i(r) = \varepsilon_i\Psi_i(r) \qquad (3.9)$$

where $\Psi_i(r)$ corresponds to the Kohn-Sham orbital, ε_i are the eigenvalues corresponding to the energy Hamiltonian, and V_{KS} is the effective potential of the system as defined in **Equation 3.10**.

$$V_{KS}(r) = V_{ext} + \int dr' \frac{n(r)}{|r-r'|} + V_{xc} \qquad (3.10)$$

Here, V_{xc} defines the exchange-correlation potential as shown in **Equation 3.11**.

$$V_{xc} = \frac{\delta E_{xc}[n]}{\delta n(r)} \qquad (3.11)$$

The $\Psi_i(r)$ and Kohn-Sham orbitals do not represent the wave functions of electrons. These orbitals lack any direct physical significance about the system. The auxiliary functions are used to compute the electron density, as defined in **Equation 3.12**.

$$n(r) = \sum_i |\Psi_i(r)|^2 \qquad (3.12)$$

The Kohn-Sham formalism can accurately determine the ground state of a system with many interacting particles, as long as the right expression for the exchange-correlation energy ($E_{xc}[n]$) is known. It should be noted that the effective potential of the system depends on the electron density (Equation **3.10**). Hence, it is necessary to solve the K-S equations in a self-consistent manner using an iterative approach. Ultimately, the self-consistent solution guarantees the attainment of the accurate ground-state density (**Figure 3**).

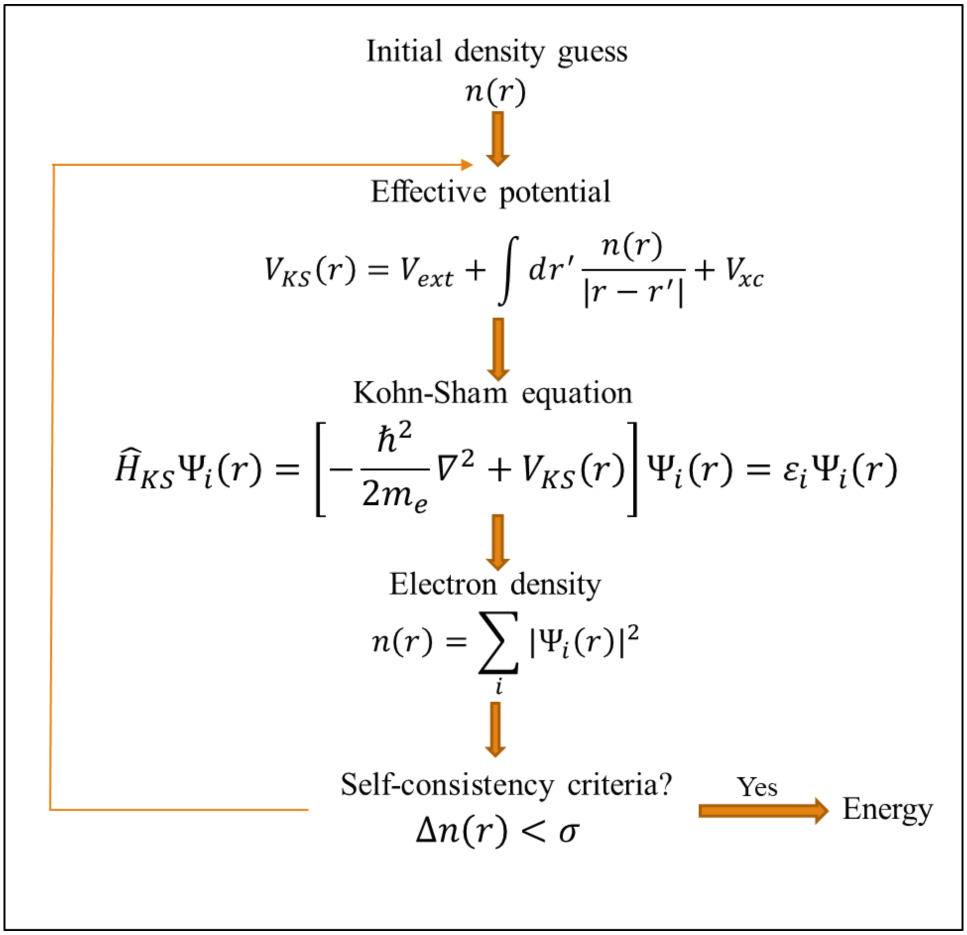

Initial density guess
$n(r)$
Effective potential
$V_{KS}(r) = V_{ext} + \int dr' \frac{n(r)}{|r - r'|} + V_{xc}$
Kohn-Sham equation
$\hat{H}_{KS}\Psi_i(r) = \left[-\frac{\hbar^2}{2m_e}\nabla^2 + V_{KS}(r)\right]\Psi_i(r) = \varepsilon_i\Psi_i(r)$
Electron density
$n(r) = \sum_i |\Psi_i(r)|^2$
Self-consistency criteria?
$\Delta n(r) < \sigma$
Yes
Energy

Figure 3: The schematic representation of the self-consistency loop method for the calculation of total energy.

3.3 Exchange-Correlation Functionals

As previously mentioned, the Kohn-Sham (K-S) formulation of Density Functional Theory (DFT) substitutes the original interacting many-electron system with an auxiliary non-interacting system that shares the same ground-state electron density. Consequently, the accuracy of DFT calculations heavily relies on the quality of the approximation employed for the exchange-correlation functional. Researchers are actively engaged in developing enhanced approximations that can more precisely account for the exchange and correlation effects, thereby improving the overall performance and reliability of DFT calculations. In essence, the approximations for the exchange-correlation functional ($E_{xc}[n]$) should be formulated with the goal of minimizing the discrepancy between the calculated total energy and the true ground-state energy. In the subsequent section, we will explore some widely adopted approximations for the exchange-correlation functional in DFT calculations, including the Local Density Approximation (LDA), Generalized Gradient Approximation (GGA), and van der Waals Density Functional (vdW-DF) approaches. These methods are designed to systematically enhance the accuracy of the exchange-correlation functionals by accounting different level of quantum mechanical interactions.

3.3.1 The Local Density Approximation (LDA)

Introduced as the pioneering approximation within the Kohn-Sham (K-S) formalism during its initial development, the Local Density Approximation (LDA) laid the foundation for subsequent refinements*[56]*. In this approach, the exchange-correlation energy density has been considered as a homogeneous electron gas. The uniform electron gas model is employed due to its incorporation of the most fundamental form of the exchange-correlation functional, which has proven to be remarkably effective for various metallic systems. The Local Density Approximation can be mathematically represented as shown in **Equation 3.13**.

$$E_{xc}^{LDA} = \int n\,(\boldsymbol{r})\varepsilon_{xc}^{uni}[n(\boldsymbol{r})]d\boldsymbol{r} \qquad (3.13)$$

where ε_{xc}^{uni} represents the exchange-correlation energy functional for a uniform electron density *n(r)* calculated at a distance **r**. This ε_{xc}^{uni} can further be sliced into two counterparts which are exchange (ε_x) and correlation (ε_c) terms respectively. The exchange (ε_x) part is obtained from an analytical methodology, but exact part of the correlation (ε_c) part is yet to be discovered. The LDA formalism reported working quite well in several model systems with slowly changing densities such as the free electrons in metallic systems*[57]*. There are some limitations associated with the LDA formalism:

(i) The calculated cohesive and binding energy values are overestimated using this correlation functional.

(ii) LDA is not suitable while working with diffused d and f orbitals, unlike s and p orbitals which are relatively localized.

(iii) The long-range interactions (i.e., van der Waals interactions) cannot be addressed due to the local nature of the LDA formalism.

3.3.2 The Generalized-Gradient Approximation (GGA)

LDA exchange-correlation formalism is not helpful for most of the systems where the electronic distribution is not uniform. Thus, GGA formalism was proposed by the H-K *[56]*, where the exchange-correlation (ε_{xc}) energy per atom is expressed not only as a function of the electron density but also the gradient of the local electronic density ($\nabla n(\boldsymbol{r})$). The GGA can be mathematically represented as shown in **Equation 3.14**

$$E_{xc}^{GGA} = \int \varepsilon_x^{uni}\, n(\boldsymbol{r}) F_{xc}[n(\boldsymbol{r}), \nabla n(\boldsymbol{r})] n(\boldsymbol{r}) d\boldsymbol{r} \tag{3.14}$$

In **Equation 3.14**, ε_x^{uni} expresses the exchange-energy density functional for a homogeneous electron gas system of electron density equal to *n(**r**)*. The F_{xc} represents a function of both electron density ($n(\boldsymbol{r})$) and the gradient of electron density ($\nabla n(\boldsymbol{r})$) which is a dimensional quantity. The F_{xc} term can further be split exchange and correlation part respectively. The exchange term in the exchange-correlation functional has been approximated in various forms, with widely used examples including the Becke (B88)*[58]*, LYP*[59]*, and Perdew-Burke-Ernzerhof (PBE)*[56]* functionals. The Generalized Gradient Approximation (GGA) formalism, which incorporates the gradient of the electron density, generally provides a smaller exchange-correlation energy compared to the Local Density Approximation (LDA). This reduction in binding energy values often improves the agreement with experimental findings, though it can also lead to under-binding in some cases. The GGA approach has therefore been successful in overcoming the shortcomings of the LDA formalism. However, the GGA formalism faces challenges in accurately capturing long-range interactions.

3.3.3 The van der Waals Density Functional (vdW-DF) Method

The accurate calculation of electronic structure and other key properties of low-dimensional or nanoscale systems presents a significant challenge. These systems are characterized by two competing types of interatomic interactions: (i) strong, localized chemical bonds between neighboring atoms, and (ii) weak, long-range van der Waals (vdW) forces between atoms separated by space. Conventional density functional theory (DFT) methods relying on the local density approximation (LDA) or generalized gradient approximation (GGA) frequently encounter challenges in accurately describing the crucial van der Waals (vdW) interactions. This limitation can result in errors in the predicted electronic structure, binding energies, and other relevant properties of the system under investigation. To address this limitation, more advanced DFT approaches have been developed that incorporate non-local vdW functionals. These so-called vdW-inclusive DFT methods aim to seamlessly account for both the short-range chemical bonds and the long-range dispersive forces, enabling a more comprehensive and accurate description of the electronic structure and physical properties of nanoscale systems*[60]*.

In this thesis, the focus is on employing the vdW-DF method to calculate the electronic structure and other vital properties of these challenging low-dimensional or nanostructured systems. The main difference between vdW-DF functional and LDA/GGA is that in vdW-DF, the energy correlation has a non-local dependence on the electron density. The full expression for the exchange and correlation energy in the vdW-DF framework is given by **Equation 3.15**

$$E_{xc}^{vdW-DF} = E_{x}^{GGA} + E_{C}^{LDA} + E_{c}^{nl} \tag{3.15}$$

In this approach, E_{x}^{GGA} represents the GGA exchange energy, E_{C}^{LDA} denotes the correlation energy within the local density approximation, and E_{c}^{nl} accounts for a non-local correlation term derived from LDA. **Equation 3.15** combines the exchange component of a GGA functional with the correlation component from LDA, further represented by a non-local correlation term. Moreover, the non-local correlation term in the vdW-DF functional takes the form of a six-dimensional integral, as expressed in **Equation 3.16**. This integral formulation captures the long-range van der Waals interactions, which are crucial for accurately describing various physical and chemical phenomena.

$$E_{C}^{nl}[n] = \frac{1}{2}\int_{r}\int_{r'} n(r)\,\Phi(r,r')\,n(r') \tag{3.16}$$

where $\Phi(r,r')$ is an interaction kernel function. In the asymptotic limit, this kernel has well-known $1/r^{6}$ behaviour characteristics for vdW interaction.

3.4 Pseudopotentials

As the dimensions of nanoscale devices grow, the computing cost of doing calculations using atomic electrons also rises, mostly because of the presence of atomic core electrons. However, these core electrons do not significantly contribute to the determination of the chemical bonding or other crucial chemical and physical characteristics of the system. These features are mostly attributed to the valence electrons. In order to address this computational difficulty, the concept of pseudopotential is proposed. The pseudopotential considers the core electrons to be chemically inert, while directly addressing the valence states that play a crucial role in chemical bonding. The consideration of pseudopotential simplifies the computing process by substituting the Coulomb potential of the nucleus and the tightly bound core electrons with an effective ionic potential. This allows the emphasis to be on the valence electrons of the device, decreasing the computational cost.

The fundamental way to lower the number of basis functions needed in electrical structure computations is to apply the PP technique. This technique comprises two practical steps: (i) Atomic core electrons are omitted from the equation, and the ionic potential is substituted with the pseudopotential. This phase efficiently minimizes the computing load by concentrating the calculation on the valence electrons, which are principally responsible for the chemical and physical characteristics of the system. (ii) The complete ionic potential, including the orthogonality of the valence WFs to the atomic core states, is substituted with a softer pseudopotential. This softer representation allows for the employment of a lesser number of basis functions since the quickly varying properties associated with the atomic core are no longer explicitly addressed.

The development of norm-conserving pseudopotentials (PPs) has been a significant advancement in the field of electronic structure calculations for nanoscale devices. In 1979, Hamann and co-workers pioneered the theory behind these energy-independent PPs, which possess several desirable properties:

(1) The eigenvalues of the all-electron wave functions correspond with those of the pseudo wave functions for the specified atomic reference configuration.

(2) Real and pseudo atomic wave functions agree beyond a chosen core radius, $r \geq r_c$.

(3) The pseudo wave functions are constrained to have the same norm as the all-electron valence wave functions within the cutoff radius, $r < r_c$. (norm-conserving)

(4) The logarithmic derivatives of the real and pseudo wave functions agree at the limit for $r \geq r_c$.

The generation of first-principles PPs has been an active area of research, with various approaches proposed to address the specific needs of electronic structure calculations. These different methods vary in terms of the functional form of the potential and the conditions used for smoothing the pseudo-wavefunctions (WFs). In the widely-used SIESTA (Spanish Initiative for Electronic Simulations with Thousands of Atoms) package, the norm-conserving PPs are typically generated according to the parameterization developed by Troullier and Martins that we have employed for our calculations.

3.5 Geometry Optimization and Force Theorem

In all the works in this thesis, geometrical optimizations have been done to search the equilibrium configuration (atoms are arranged in the ground state) before calculating the electronic and quantum transport properties. Employing the Hellman-Feynman theorem, we calculate the force acting on nuclei with ionic position (R_I) as given below in **Equation 3.17**

$$F_I = \frac{\partial \varepsilon}{\partial R_I} \tag{3.17}$$

here, ε corresponds to the total energy of the system which can be described as given below in **Equation 3.18**

$$\varepsilon = \frac{<\Psi|\hat{\mathrm{H}}|\Psi>}{<\Psi|\Psi>} \tag{3.18}$$

where Ψ is the Kohn-Sham WFs. As considered wave functions are normalized, we can write $< \Psi|\Psi >= 1$. The change in the Kohn-Sham WFs due to the variation in ionic coordinates is directly responsible for the forces acting on the ions, which is a fundamental aspect of electronic structure calculations and a key driver in the optimization of atomic structures and materials properties. By using **Equation 3.17** and **3.18**, we can write **Equation 3.19**.

$$F_I = -< \Psi \left|\frac{\partial \hat{H}}{\partial R_I}\right| \Psi > - < \frac{\partial \Psi}{\partial R_I}\left|\hat{H}\right|\Psi > - < \Psi\left|\hat{H}\right|\frac{\partial \Psi}{\partial R_I} > \tag{3.19}$$

The second and third terms represent the change in the wave function Ψ with respect to the nuclear positions, acting with the Hamiltonian $\hat{H}$, and then taking the inner product Ψ. This term captures the effect of changing the nuclear positions on the electronic structure of the system. However, it is often assumed that this effect is small compared to the direct change in the Hamiltonian due to the nuclear motion. This assumption is particularly valid in systems where the electronic structure changes relatively slowly compared to the nuclear motion (**Equation 3.19**). Hence, the explicit form of the force can be written as shown in **Equation 3.20**.

$$F_I = -< \Psi \left| \frac{\partial \hat{H}}{\partial R_I} \right| \Psi > \tag{3.20}$$

The calculation of forces on the ions is a crucial aspect of electronic structure methods, and it can be derived from the total energy of the system. Considering the system at its ground state, the partial derivative of the total energy with respect to the ionic positions can be used to describe the forces acting on the ions. The force theorem provides a framework for performing geometry optimization based on these calculated forces.

3.6 Conclusions

In this study, we have considered a hybrid graphene/h-BN nanopore that comprises two electrodes (left/right) and a central scattering region. The nanopore is sculpted in the central scattering region with diameters of 12.16 Å (y-axis) and 12.96 Å (z-axis). Out of the fully optimized unit cell of graphene and boron nitride, we have made a hybrid supercell with a size of 1 × 24.32 × 43.36 Å^3, and further structure optimization was done by employing SIESTA (Spanish Initiative for Electronic Simulations with Thousands of Atoms) *[61,62]*. Using 6-31+G* basis sets and the B3LYP correlation functional as available in Gaussian 09 code has been implemented to optimize the isolated nucleotide molecules *[63]*. The nucleotides are positioned inside the pristine nanopore in a way that ensures that the aromatic rings (purines and pyrimidines) are aligned in the same yz plane, with the phosphate group extending outward from the nanopore. This spatial configuration will ensure that the nucleotides are interacting strongly with the nanopore edges leading to enhanced interaction energy, charge transfer values, and conductance sensitivity. During structure optimization, the van der Waals density (vdW-DF) exchange-correlation functional has been employed to describe weak interactions *[16,64]*. The interaction between the inner core and the valence electrons is described using double zeta polarized (DZP) basis sets in all computations *[65,66]*. A mesh cut-off of 200 Ry is considered for integration in real space, and (1 × 3 × 2) k-point sampling has been used in the Brillouin zone for all the calculations using SIESTA. The convergence requirements for the density matrix in all optimization processes utilizing the self-consistent field (SCF) approach are established at 0.0001 eV. All four nucleotides are optimized inside the nanopore at an angle between 0^o to 180^o on an interval of 30^o, and this rotation is done around the x-axis perpendicular to the plane of the nanopore.

Chapter 4

Quantum Transport Theory

4.1 Quantum Transport Regimes

The understanding of electronic transport across nanoscale systems is a major focus of this thesis. The behavior of electron transport in these nanoscale devices can be categorized into different transport regimes, depending on the relative magnitudes of two characteristic lengths - the momentum relaxation length (L_m) and the phase relaxation length ($L_\emptyset$)*[67]*. The momentum relaxation length (L_m) is the mean free path of an electron, representing the average distance it travels before its initial momentum is lost. Similarly, the phase relaxation length ($L_\emptyset$) is the average distance it travels before its initial phase is lost. If the length of the nanoscale device (L) is much longer than L_m and $L_\emptyset$, the conductance of the nanoscale device depends on the length of the wire, as per the classical Ohm's law.

(1) **Ballistic Transport Regime ($\boldsymbol{L << L_m, L_\emptyset}$)**:

In this regime, the length of the nanoscale device (L) is much shorter than both the momentum relaxation length (Lm) and the phase relaxation length (Lϕ). This means that electrons can propagate from one electrode to the other without experiencing any scattering events that would cause a loss of their original momentum or phase. The electron transport is essentially ballistic - the electrons travel through the device without any significant interactions. The only resistance that arises is due to backscattering at the contacts between the device and the electrodes. This ballistic transport leads to a quantum conductance that is independent of the device length, as seen in materials like carbon nanotubes and graphene.

(2) **Elastic and Coherent Transport Regime ($\boldsymbol{L < L_m, L_\emptyset}$)**:

When the device length L is less than the momentum and phase relaxation lengths, the electron transport can still be considered coherent. In this regime, electrons can undergo elastic scattering events within the device, where their energy and phase are preserved, but their momentum may change. This elastic scattering reduces the transmission function compared to the ballistic case, but the electron wavefunction remains coherent throughout the transport process. The device length is short enough that the electrons maintain their quantum phase as they traverse the system.

(3) **Inelastic and Incoherent Transport Regime** ($\boldsymbol{L > L_m, L_\emptyset}$):):

In this regime, the device length L exceeds both the momentum and phase relaxation lengths. This means that the electrons experience significant inelastic scattering events, such as interactions with other electrons or phonons, as they travel through the nanoscale system. These inelastic processes lead to a change in both the electron momentum and phase. The long device length compared to the relaxation lengths results in a loss of the electron's phase coherence, leading to incoherent transport. The inelastic scattering and phase breaking events constitute an important part of the transport characteristics in this regime.

The transport problem has been extensively studied using two frequently used formalisms: the Landauer formalism and the Non-Equilibrium Green's Function (NEGF) formalism. The Landauer formalism enables the definition of non-interacting electronic transport in the ballistic or coherent transport regimes. In contrast, the NEGF formalism is a more comprehensive method that may be employed in all three transport regimes. In the following sections, we will provide a concise explanation of the Landauer and NEGF formalism used to study the transport issues in this thesis.

4.2 The Landauer Formalism

The Landauer formalism provides a way to understand electrical conductance in nanoscale systems by treating them as scattering problems *[68,69]*. It considers the system as a central scattering region connected to semi-infinite left and right electrodes, which act as electron reservoirs. The key principle is that the electrical current flowing through the system is proportional to the probability that electrons can transmit from one electrode to the other.

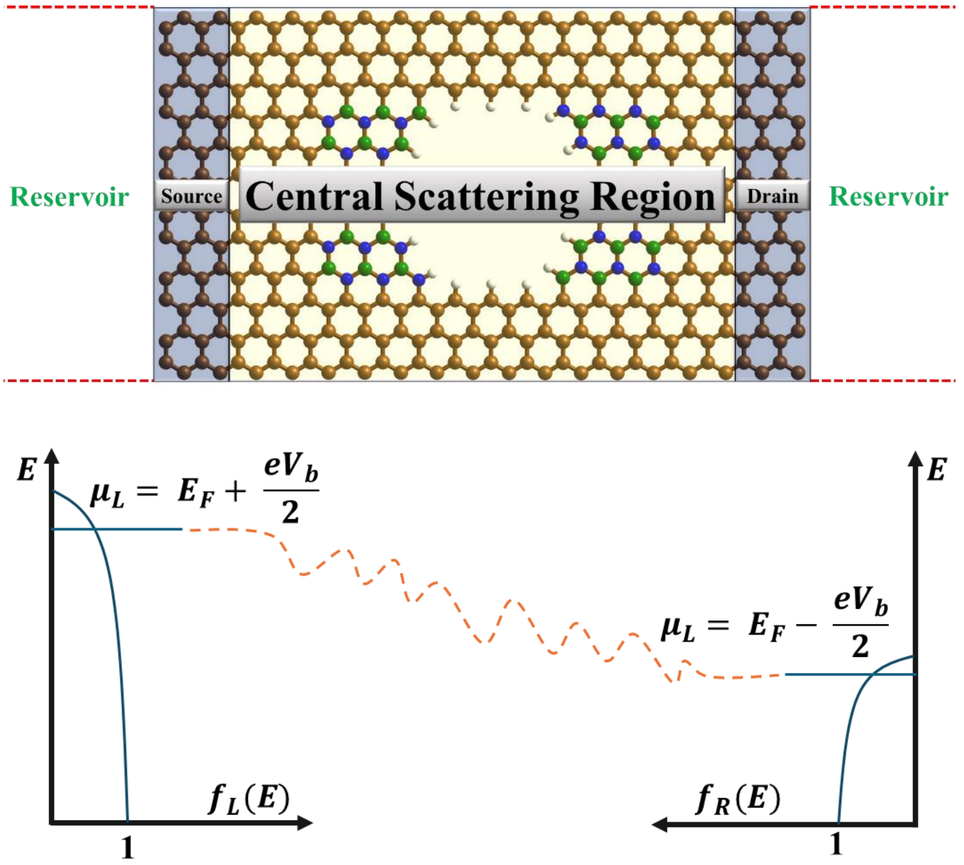

Reservoir
Source
Central Scattering Region
Drain
Reservoir
E
$\mu_L = E_F + \frac{eV_b}{2}$
E
$\mu_L = E_F - \frac{eV_b}{2}$
$f_L(E)$
1
$f_R(E)$
1

Figure 4.1: Diagram of a graphene/h-BN nanopore device connected with two electrodes described in the Landauer formalism. The μ_L and μ_R are the chemical potential at equilibrium of source and drain respectively. Both electrodes are connected to a central scattering region (CSR).

This probability is encapsulated in the transmission function T(E, V_b), defined as a sum over the transmission function of electrons flowing from source to drain at a given energy E and bias voltage (V_b). In **Figure 4.1,** V_b is considered to be the externally applied bias voltage on the left and right electrodes. Then, the chemical potentials ($\mu_{L/R}$) on the L/R electrodes can be expressed as $\mu_{L/R} = E_F \pm eV_b/2$. Here, E_F represents the Fermi energy. It is evident that the electric current would be zero if $f_L(E) = f_R(E)$, according to the Landauer formula. The net electric current is dependent on the difference between the Fermi distributions of electrons in the source and drain respectively. At a finitely applied bias voltage (V_b), the electric current simplifies to the following **Equation 4.1** at absolute temperature (T) = 0.

$$I(V_b) = \frac{2e}{h}\int_{\mu_R}^{\mu_L} T(E)\,dE \qquad (4.1)$$

If the difference between the Fermi distribution functions is significant for the two electrodes, the famous Landauer formula can expressed as the integral of transmission function ($T(E, V_b)$) as shown in **Equation 4.2**.

$$I(V_b) = \frac{2e}{h}\int T(E, V_b)\left(f_L(E) - f_R(E)\right)dE \qquad (4.2)$$

The quantum conductance can be derived by averaging the transmission over an energy window with a width centered around the Fermi level of the electrodes. Conversely, if the Fermi function exhibits minimal variation across an energy range, it can be approximated through a Taylor series expansion evaluated at the Fermi energy, as illustrated in **Equation 4.3**.

If the transmission function T(E) does not vary appreciably over the energy window corresponding to the applied bias $eV_b = \mu_L - \mu_R$, an alternative approach can be used. In this scenario, the Fermi function can be Taylor expanded around the Fermi energy $\mu_L = E_F$ of the electrodes. Performing this expansion and utilizing the properties of the delta function, one arrives at an expression for the quantum conductance G by averaging the transmission over an energy range weighted by the derivative of the Fermi distribution:

$$I(V_b) = \frac{2e}{h} \int dE\, T(E)\, (-\frac{\partial f(E)}{\partial E})(\mu_L - \mu_R) \tag{4.3}$$

Then the quantum conductance $(G) = I/V_b$ can be written as below in **Equation 4.4**

$$G = \frac{2e^2}{h} \int\, T(E)\, (-\frac{\partial f(E)}{\partial E}) dE \tag{4.4}$$

At $T = 0$ K, $-\frac{\partial f(E-\mu)}{\partial E} = \delta(\mu)$, where $\delta(\mu)$ represents a Kronecker delta.

Further, for an ideal periodic chain, where $T(E) = 1$ at $T = 0$ K, the Landauer conductance becomes as below (**Equation 4.5**).

$$G_0 = \frac{2e^2}{h} = \frac{1}{12.9}(k\Omega)^{-1} \text{ or } G_0 = (12.9\, k\Omega)^{-1} \tag{4.5}$$

This G_0 is known as the quantum conductance.

Electron scattering in nanoscale devices often deviates from ideal behavior. Consequently, the conductance of such nanoscale devices can be more accurately represented by the expression given in the upcoming **Equation 4.6**.

$$G = G_0 T(E_F) = \frac{2e^2}{h} T(E_F) \tag{4.6}$$

The **Equation 4.6** is used for two-electrode systems. Nevertheless, for devices with a gate electrode or more than two electrodes that are carrying electrons, the Landauer formula can be generalized as given below in **Equation 4.7**.

$$G = G_0 \sum_{i,j} T_{ij}(E_F) \tag{4.7}$$

where T_{ij} is the probability of electron passing from i^{th} conducting mode at the L electrode of the nanoscale device to the j^{th} conducting mode at the R electrode of the nanoscale device.

4.3 Non-Equilibrium Green’s Function (NEGF) Formalism

The study of electronic transport is simulated on the atomic level; thus, combining the NEGF formalism with DFT has a significant advantage over the other formalism. A detailed discussion of the NEGF formalism can be found in many articles and books*[70,71]*. In this section, we aim to provide a general explanation of the NEGF formalism to evaluate the current−voltage (I−V) characteristics curves of nanoscale devices. The schematic of a graphene/h-BN nanopore device, separated into three parts, is shown in **Figure 4.1**. It consists of a central scattering region (CSR) and semi-infinite left (L) and right (R) electrodes, then the Hamiltonian (H) of the device can be written as below in **Equation 4.8**.

$$H = \begin{pmatrix} H_L & \tau_L & 0 \\ \tau_L^\dagger & H_{CSR} & \tau_R^\dagger \\ 0 & \tau_R & H_R \end{pmatrix} \tag{4.8}$$

here, H_{CSR}, $H_{L/R}$ indicates the Hamiltonian matrices of the CSR and L/R electrodes, respectively. $\tau_{L/R}$ represents the matrix elements involving the interaction between the L/R electrodes and the CSR. We assume that there is no direct interaction (tunneling) between the L/R electrodes. Therefore, after writing the Hamiltonian of the nanoscale device, we aim to solve the Schrödinger equation. In this approach, the Non-Equilibrium Green's Function (NEGF) formalism is employed to solve the quantum transport equation, where the retarded Green's function (G) corresponding to the Hamiltonian matrix (H) is formulated as presented in **Equation 4.9**.

$$[E^+S - H] = I \tag{4.9}$$

here, S corresponds to an overlap matrix, $E^{+} = lim_{\eta\rightarrow 0^{+}} E + i\eta$, and I represents an ∞-dimensional matrix and G can be written as below in **Equation 4.10**.

$$G = \begin{pmatrix} G_L & G_{CSRL} & 0 \\ G_{LCSR} & G_{CSR} & G_{RCSR} \\ 0 & G_{CSRR} & G_R \end{pmatrix} \tag{4.10}$$

For our convenience, the whole system can be divided into different regions and evaluate the matrix (G) as we are not interested in electrodes. Therefore, both ends of the CSR are in surface contact with the electrode. Thus, we consider that the interaction term $(\tau_{L/R})$ would be negligibly smaller in size as compared to the Hamiltonian (H). Therefore, after solving **Equations 4.9** and **4.10,** we can obtain the following **Equation 4.11**.

$$\begin{pmatrix} E^{+}S_L - H_L & -\tau_L & 0 \\ -\tau_L^{\dagger} & E^{+}S_{CSR} - H_{CSR} & -\tau_R^{\dagger} \\ 0 & -\tau_R & E^{+}S_R - H_R \end{pmatrix} \begin{pmatrix} G_L & G_{CSRL} & 0 \\ G_{LCSR} & G_{CSR} & G_{RCSR} \\ 0 & G_{CSRR} & G_R \end{pmatrix} =$$

$$\begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} = I \tag{4.11}$$

After solving the **Equation 4.11**, the G written earlier in **Equation 4.10** becomes as given below in **Equation 4.12**

$$G = \begin{pmatrix} g_L(1 + \tau_L G_{LCSR}) & g_L \tau_L G_{CSR} & 0 \\ G_{LCSR} & G_{CSR} & G_{RCSR} \\ 0 & g_R \tau_R G_{CSR} & g_R(1 + \tau_R G_{RCSR}) \end{pmatrix} \tag{4.12}$$

here, $g_{L/R} = \frac{1}{(E^{+}S_L - H_{CSR})} = (E^{+}S_L - H_{CSR})^{-1}$ is the "surface G" of the L/R electrode uncoupled to the CSR.

Then, the final expression for retarded G of the CSR can be obtained by solving the equations as mentioned earlier and can be written as below in **Equation 4.13**

$$G_{CSR} = [E^{+}S_{CSR} - H_{CSR} - \sum_L(E) - \sum_R(E)]^{-1} \tag{4.13}$$

where $\sum_L(E) = \tau_L^\dagger g_L \tau_L$ and $\sum_R(E) = \tau_R^\dagger g_R \tau_R$ are called the "self-energies". The self-energy is associated with the energy-level shift (Δ) and the energy-level broadening (Γ), as presented in **Figure 4.2**. The Δ and Γ can be described from the real and imaginary part of the self-energy as below in **Equation 4.14** and **4.15**

$$\Delta_{L/R}(E) = Re \sum_{L/R}(E) \tag{4.14}$$

$$\mathit{\Gamma}_{L/R}(E) = i[\sum_{L/R}(E) - \sum_{L/R}^\dagger(E)] \tag{4.15}$$

Furthermore, the broadening of molecular energy levels correlates with the residence time of electrons on the molecules. When a molecule couples to electrodes, electrons can transition from the molecular states localized at the conductor-superconductor interface into either the left or right electrode. The lifetime of a given molecular state is inversely proportional to the degree of state broadening: $\tau_{L/R}\mathit{\Gamma} = \hbar$.

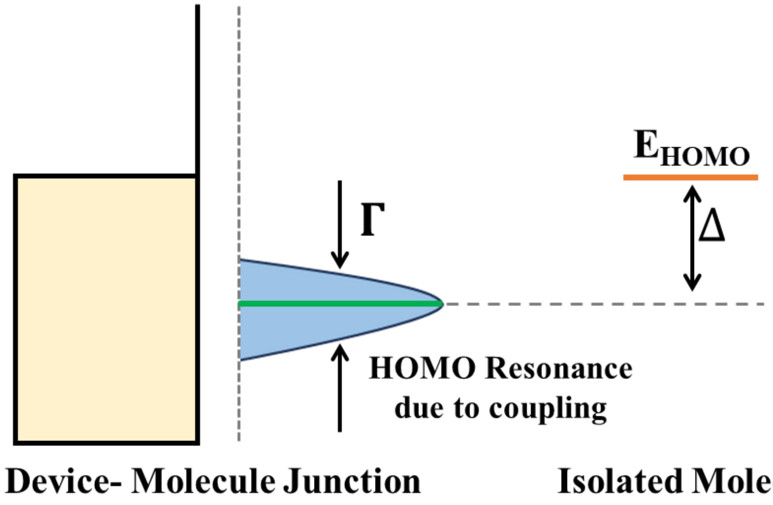


Figure 4.2: Schematic illustration of device-molecule junction. When the single molecule is in contact with semi-infinite electrodes, its energy levels are shifted (Δ). The energy level broadening due to the coupling to the contact is given by Γ.

From **Equation 4.13**, the infinite-dimensional Hamiltonian is reduced to the dimension of the CSR, where the self-energies, $\sum_{L/R}(E)$, include all information on the semi-infinite properties of the electrodes. The CSR only interacts with the surface region of the L/R electrodes. As a result, we can solely focus on the G matrix of the CSR and effective Hamiltonian (H_{eff}) can be described as given below in **Equation 4.16**.

$$H_{eff} = H_{CSR} + \sum_{L}(E) + \sum_{R}(E) \tag{4.16}$$

The effective Hamiltonian can be used to construct the scattering matrix, which describes how electron waves are scattered as they propagate through the system. Analyzing the scattering matrix provides insights into the scattering processes occurring at interfaces and within the central scattering region.

4.4 Conclusions

In this study, NEGF formalism has been utilized in combination with the DFT in the TranSIESTA code for electronic transport calculations *[62,72]*. The TranSIESTA calculations have been done with (1 × 11 × 11) k-point sampling in the Brillouin zone. For transmission function calculation using TBtrans code, we have used higher values of k- points of (1 × 45 × 45) to get more accurate results. The electronic transmission function and current can be calculated using advanced Green's function. Using this method, zero-bias (*V=0*) transmission spectra have been calculated for the pristine hybrid G/h-BN device. The transmission probability of the electron from the left electrode to the right electrode can be calculated from the zero-bias transmission function as per the Landauer−Büttiker formula.

Chapter 5

Machine Learning Details

Within this study, we applied both machine learning (ML) regression and classification techniques to identify single nucleotides from the electric readouts of the G/h-BN nanopore. Linear regression (LR), adaptive boosting regression (AdaBoost), and extreme gradient boosting regression (XGBoost) are utilized for regression tasks, while for classification purposes, K-nearest neighbor (KNN), support vector machine (SVM), decision tree classifier (DTC), and random forest classifier (RFC) are employed. These algorithms are implemented in the scikit-learn open-source library (version 1.2.2) and executed using Python code 3.10 within the Google Colab environment *[73]*. In order to assess the effectiveness of these ML algorithms, several statistical measures have been used. For regression models, parameters such as root mean square error (RMSE) and determination coefficient (R^2) have been calculated, and for classification models, the evaluation has been done based on the confusion matrix, as well as parameters like precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve.

5.1 ML Regression

In this study, we employed XGBoost to develop a regression model for transmission function prediction of single DNA nucleotides using G/h-BN nanopore. Feature engineering was done to extract valuable information as descriptors from the raw transmission data. After preprocessing, the dataset was split into training and testing sets to effectively evaluate the performance of the XGBoost regressor. XGBoost algorithm utilizes an iterative nature that constructs an ensemble of decision trees sequentially, minimizing a predefined loss function. Crucial hyperparameters like the number of trees, tree depth, and learning rate were tuned to optimize the model mitigating over or underfitting. Standard evaluation metrics, such as mean squared error (MSE), root mean squared error (RMSE), and R-squared, were used to assess the predictive performance of the ML model. Finally, hyperparameter tuning improves XGBR model performance by finding the best set of parameters for the prediction.

5.1.1 Correlation Matrices

A correlation matrix is a table that shows the correlation coefficients between each pair of chosen descriptors. It provides a simple way to visualize the patterns of relationships among the considered features. The correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

5.1.1.1 Pearson correlation coefficient

The Pearson correlation coefficient (PCC) is a statistical metric that assesses the magnitude and direction of linear connections. The coefficient varies between -1 and +1, with a value of +1 indicating a perfect positive linear correlation, -1 representing a perfect negative linear correlation, and 0 suggesting no linear correlation. The Pearson correlation coefficient (PCC) has been widely utilized to precisely measure the linear correlation between distances in the original space and the reduced space, as determined by the equation:

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}}$$

Where x_i and y_i represent the corresponding values of x and y. $\bar{x}$ and $\bar{y}$ represent the average values of x and y, respectively.

5.1.1.2 Spearman’s rank-order correlation coefficient

Spearman's rank-order coefficient of correlation (ρ) is a reliable statistical metric used to determine the magnitude and direction of the monotonic connection between two variables. Spearman's coefficient is more effective than linear-focused PCC in determining the degree of correlation between the ranks of variables. The range of the values is from -1 to +1. A value of +1 indicates a perfect positive monotonic connection, -1 indicates a perfect negative monotonic relationship, and 0 indicates no monotonic relationship. The calculation is done as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

The variable d_i represents the difference in rank assigned to the two variables, while n represents the total number of samples.

5.1.2 Root Mean Squared Error (RMSE)

RMSE is a widely used metric to evaluate the accuracy of predictive models, particularly in regression problems involving continuous data. It measures the average magnitude of errors between predicted and actual values, considering the square of individual errors. This characteristic penalizes larger deviations more heavily, providing a comprehensive assessment of model performance. A lower RMSE value indicates better prediction accuracy, making it a reliable criterion for model selection and continuous data prediction.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Where n represents the total number of data points, y_i represents the actual observed values, and $\hat{y}_i$ represents actual predicted values.

5.1.3 Coefficient of Determination (R^2):

Coefficient of Determination (R^2) is a statistical measure that evaluates how well a regression model fits the data. It quantifies the extent to which the independent variables capture the variation present in the values of the dependent variable. A higher R^2 value, ranging from 0 to 1, indicates a better fit and more predictive power of the model. R^2 provides a standardized way to assess goodness-of-fit across different regression models, making it an essential tool for model evaluation and selection

$$R^2 = 1 - \frac{sum\ squared\ residuals\ (SSR)}{total\ sum\ of\ squares\ (SST)} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where y_i represents the actual observed values, $\hat{y}_i$ represents actual predicted values, and $\bar{y}$ represents the mean value of the sample.

5.2 ML Classification

5.2.1 Classification Algorithms

5.2.1.1 K-Nearest Neighbor (KNN)

The KNN algorithm utilizes the principle of similarity based on the calculated Euclidean distance between the new point and each set of points in a class while classifying a new point in feature space.[74] The parameter ‘K’ which decides the closest data points, is declared by the user and based on the number of nearest neighbors and the new data point is assigned to a new class. The user-defined value of ‘K’ significantly impacts the overall accuracy of the model. Small values of closest points may lead to flexibility in the decision boundary that is prone to noise, while large values can lead to oversimplification of results with a smooth decision boundary. The KNN algorithm is useful for small datasets as the computational complexity increases while working with larger datasets.

5.2.1.2 Decision Tree Classifier (DTC)

This classification algorithm consists of a tree-like structure based on a series of decisions to classify new data points.[75] This decision tree is constructed based on some internal nodes that represent a feature or attribute, each branching is interpreted as a decision-making based on that feature, and finally, each leaf node represents a particular class of outcome.

As the new data comes in, it glides from the root to the leaf of the decision tree and is finally assigned to a particular class based on a series of decisions. The DTC algorithm functions recursively to partition the dataset into smaller subsets according to the features that are chosen, continuing this process until it reaches a halting criterion. The depth of the tree, minimum number of samples in a node, and other parameters can be declared by the user. DTC algorithm tends to overfit on training data set if it is allowed to grow too deep. Controlling the depth of the tree, pruning and some ensemble methods can be convenient to extenuate such problems.

5.2.1.3 Random Forest Classifier (RFC)

The Random Forest Classifier algorithm is a type of ensemble learning method that is built upon the principles of decision trees. The system functions by creating numerous decision trees and using their results to produce predictions, thus enhancing accuracy by mitigating overfitting. Every tree within the forest is separately trained using random subsets of the training data and features. For classification problems, Random Forest Classifier (RFC) combines the predictions of each tree by employing either a majority voting method or averaging the probabilities assigned by separate trees. The Random Forests are often resilient and exhibit strong performance across many complex datasets. They efficiently manage high-dimensional data, big datasets, and categorical variables. Moreover, ensemble methods have a superior ability to generalize compared to individual decision trees, resulting in reduced susceptibility to overfitting. The confusion matrix is often regarded as the most effective and fundamental evaluation matrix in machine learning classification. The table displays various combinations of predicted and actual values.

5.2.2 Machine Learning Classification Metrics

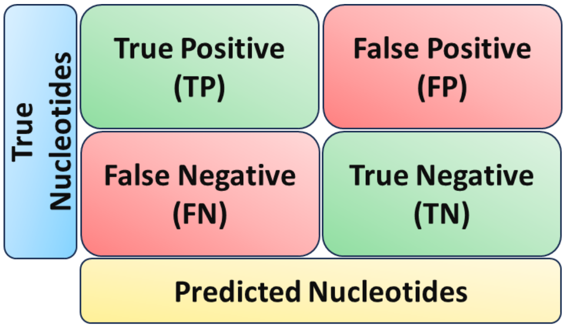


Figure 5: Illustration of confusion matrix for the evaluation of classification algorithms

A confusion matrix (**Figure 5**) is a table that allows you to visualize the performance of the ML classification model. It shows the number of correct and incorrect predictions made by the model, broken down by each class. The rows in a confusion matrix represent the actual classes, while the columns represent the predicted classes. The diagonal elements of the matrix represent the correct predictions, where the actual class matches the predicted class. The off-diagonal elements represent the incorrect predictions, where the model confused one class for another. Confusion matrices are particularly useful when you have an imbalanced dataset, where some classes have many more instances than others. In such cases, accuracy alone may not be a reliable metric to evaluate the performance of the model, as it can be skewed by the majority class. The confusion matrix provides more detailed information about the model's performance on each class, including the following predictions-

True positive (TP) = Total number of correct positive predictions made by the ML model

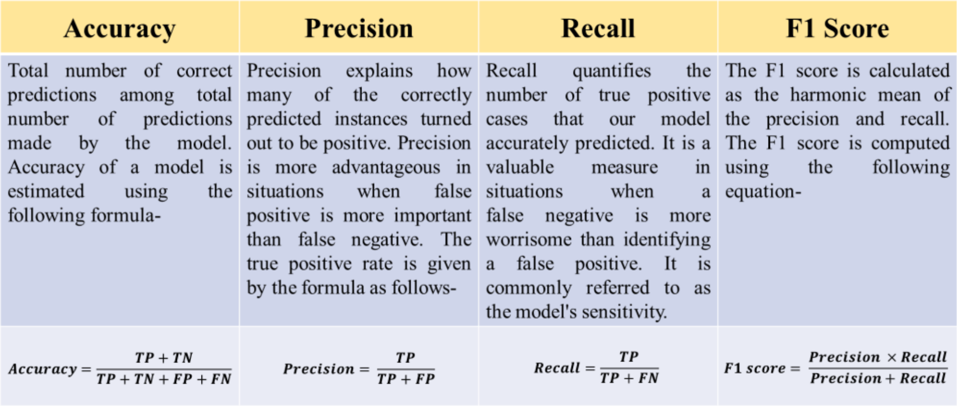
True negative (TN) = Total number of correct negative predictions made by the ML model

False positive (FP) = Total number of incorrect positive predictions made by the ML model

False negative (FN) = Total number of incorrect negative predictions made by the ML model

From the confusion matrix, various performance metrics can be calculated, such as precision, recall, and F1-score for each class, as shown in **Table 5**. These metrics can help you understand the strengths and weaknesses of the model and help us to evaluate and improve its performance.

Table 5: Tabular Representation of Accuracy, Precision, Recall, and F1 Score for the Evaluation of Classification Algorithms



Accuracy	Precision	Recall	F1 Score
Total number of correct predictions among total number of predictions made by the model. Accuracy of a model is estimated using the following formula-	Precision explains how many of the correctly predicted instances turned out to be positive. Precision is more advantageous in situations when false positive is more important than false negative. The true positive rate is given by the formula as follows-	Recall quantifies the number of true positive cases that our model accurately predicted. It is a valuable measure in situations when a false negative is more worrisome than identifying a false positive. It is commonly referred to as the model's sensitivity.	The F1 score is calculated as the harmonic mean of the precision and recall. The F1 score is computed using the following equation-
$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	$Precision = \frac{TP}{TP+FP}$	$Recall = \frac{TP}{TP+FN}$	$F1\ score = \frac{Precision \times Recall}{Precision + Recall}$

5.2.3 ROC Curve

A Receiver Operating Characteristic (ROC) curve is a visual depiction of how well a model performs at various classification criteria.

$$TPR(Sensitivity) = \frac{TP}{TP+FN} \text{ and } FPR = (1 - Specificity) = \frac{FP}{FP+TN}$$

The graph illustrates the relationship between the true positive rate (TPR) and the false positive rate (FPR) across different threshold values.

The AUC-ROC measures the whole area under the ROC curve, spanning from point (0,0) to (1,1). It provides a concise evaluation of the model's capacity to differentiate across distinct classes. A classifier with a perfect performance has an Area Under the Curve (AUC) value of 1, whereas a random classifier has an AUC value of 0.5. It facilitates the comparison between multiple classification models. A model with a higher AUC-ROC typically exhibits superior discrimination capabilities.

5.3 Machine Learning Model Explainability

SHapley Additive exPlanations (SHAPs) have been used to elucidate the lack of interpretability in the decision-making process of trained "black box" models by extrapolating the impact of input feature information on transmission prediction. Shapley's value employs cooperative game theory to break down the overall prediction into the aggregate of individual contributions from local features. This technique improves the understanding of how these input properties impact the transmission. The SHAP values may be represented by the following generic equation:

$$\emptyset_m(p) = \sum_{s \subseteq N/m} \frac{|S|!\,(n - |S| - 1)!}{n!}\left(p(S \cup m) - p(S)\right)$$

The equation defines $\emptyset_m(p)$ as the SHAP value corresponding to feature ‘m’ in instance ‘p’. N denotes the collection of all features. The summation was conducted over all subsets ‘S’ that do not include feature m. The expression $p(S \cup m)$ denotes the prediction made by the model when feature m is brought into the subset ‘S’, which already includes certain characteristics. The model's prediction, denoted as $p(S)$, is based only on the characteristics included in the subset ‘S’. The disparity between these two predictions represents the influence of feature ‘m’ on the output.

5.4 Conclusions

After extracting suitable features from our DFT-calculated transmission datasets, we have implemented ML regression to predict signature transmission for nucleotides and utilized different classification algorithms for the identification of single nucleotides. The different evaluation matrices have been used to assess the performance of ML classification and regression algorithms. Further SHAP analysis has been used to evaluate the underlying hierarchical order of feature importance in regression and classification studies.

Chapter 6

Results and Discussion

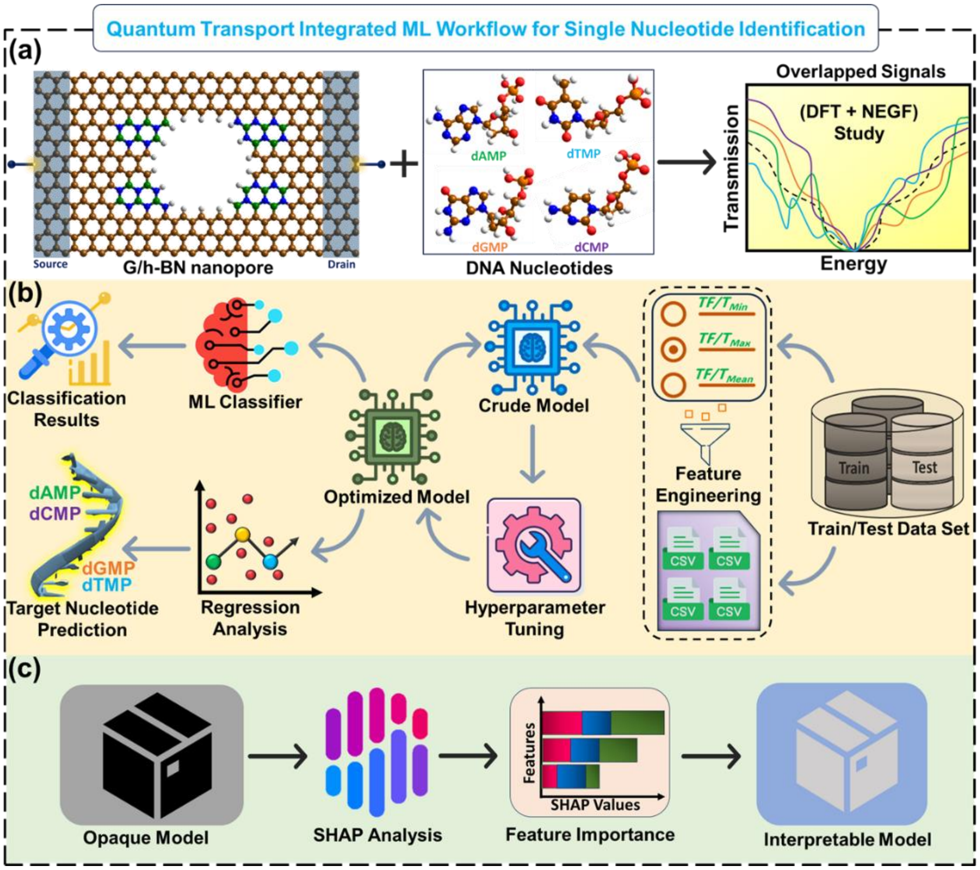


Figure 6.1: (a) Schematics of hybrid graphene/hexagonal boron nitride (G/h-BN) nanopore for single DNA nucleotides (dAMP, dCMP, dGMP, and dTMP) identification with quantum transport approach, (b) a step-by-step illustration machine learning workflow for regression and classification analysis with transmission function readouts collected from G/h-BN nanopore, and (c) machine learning interpretable analysis with SHAP to understand the descriptor-target relationships.

Figure 6.1 describes the G/h-BN nanopore that can integrate the quantum transport approach with ML algorithms for the prediction of fingerprint transmission function and classification of unlabelled DNA nucleotides. Initially, the pristine G/h-BN nanopore and isolated nucleotides are optimized using first principle DFT calculations as implemented in SIESTA*[61,62]* and Gaussian 09 code*[63]*, respectively (**Figure 6.2)**.

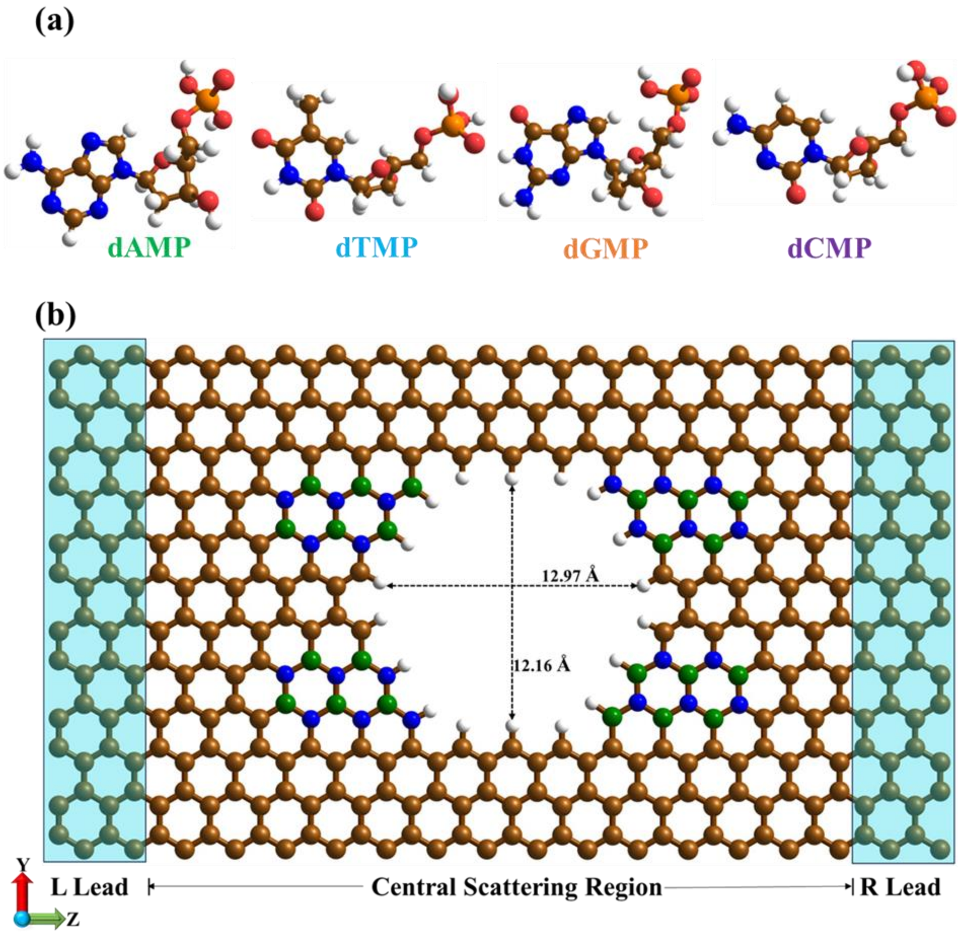


Figure 6.2: (a) Optimized structures of four single DNA nucleotide (dAMP, dGMP, dTMP, and dCMP), (b) hybrid G/h-BN pristine nanopore device. The proposed hybrid nanopore device consists of left (L) and right (R) leads and a central scattering region (device region) where the N and B atoms are depicted in blue and green spheres, respectively.

After that, G/h-BN nanopore + nucleotide geometries are relaxed by placing each nucleotide inside the hybrid nanopore by using SIESTA code *[61,62]*. In experimental conditions, the DNA strand encounters various orientation fluctuations and structural variations while moving through the nanopore. To mimic the situation, rotational dynamics are considered for nucleotides with seven different orientations (from 0º to 180º in the step of 30º) while located inside the nanopore and fully relaxed for our further calculations.

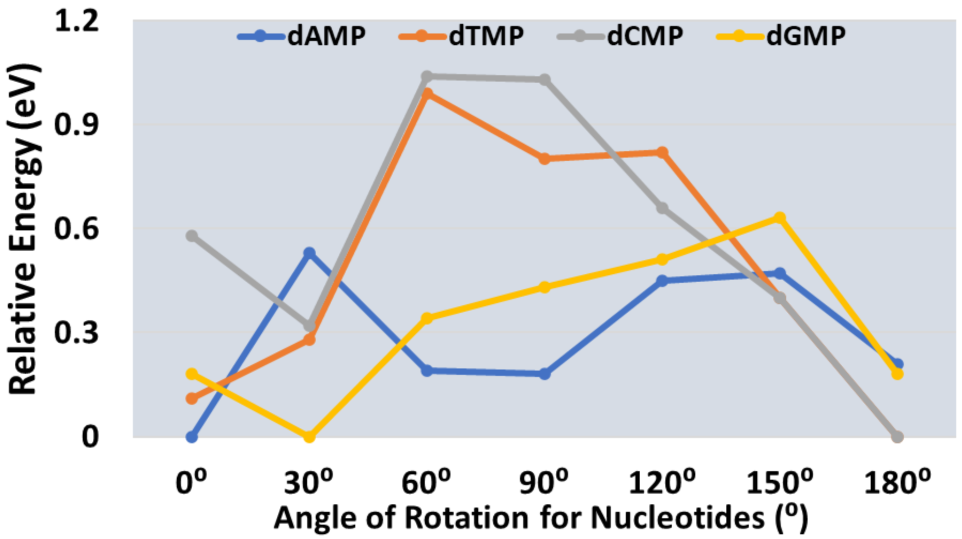


Figure 6.3: Relative energy (eV) plot of all four nucleotides at different in-plane rotational configurations inside G/h-BN nanopore with respect to the most stable configuration.

The configurations with the lowest relative energy are selected as the most favorable configuration of nucleotide inside the G/h-BN nanopore, as shown in **Figure 6.3**. By examining the relative energy plot, the 0º, 180º, 30º, and 180º orientations are identified as the energetically most stable configurations for dAMP, dTMP, dGMP, and dCMP, respectively as represented in **Figure 6.4**.

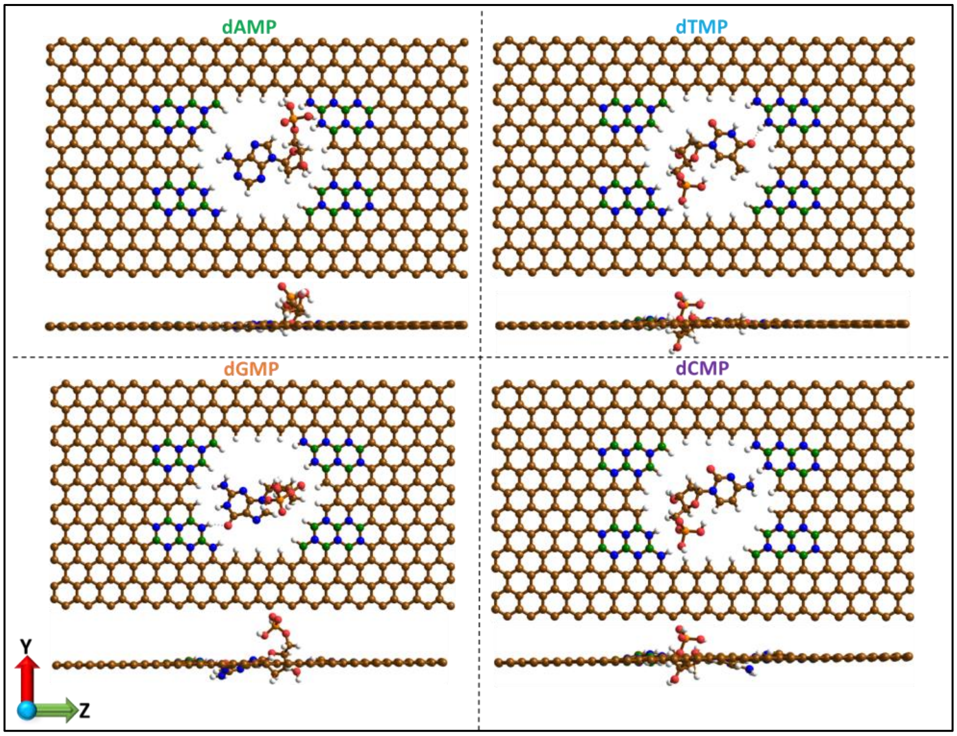


Figure 6.4: Top and side views of the fully relaxed most stable configurations of all four nucleotides inside the hybrid G/h-BN nanopore.

The transmission functions of all the structures are calculated with the quantum transport (DFT+NEGF) method as implemented in the TranSIESTA code *[62]*. The calculated signature transmission readouts for seven distinct rotational configurations of each nucleotide have been shown in **Figure 6.5**.

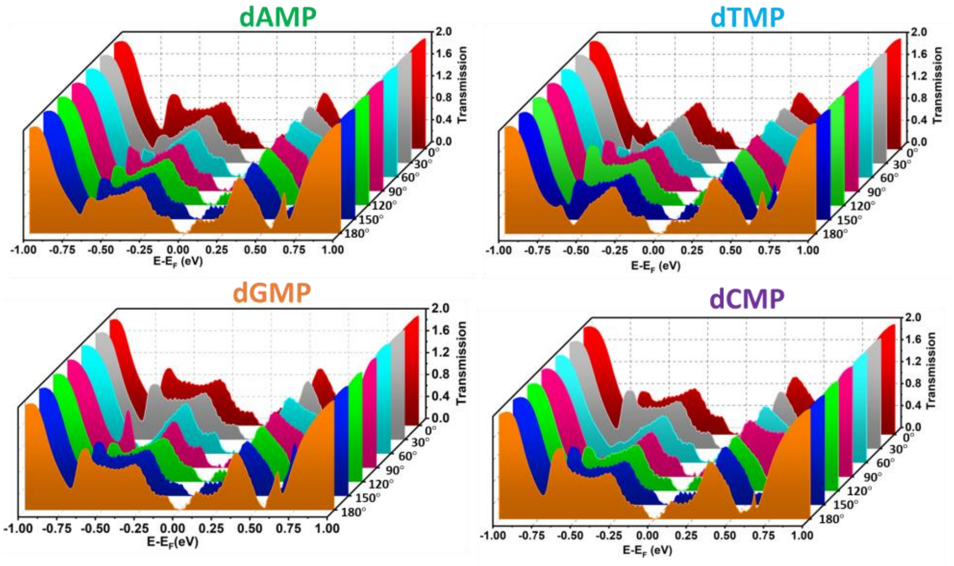


Figure 6.5: Effect of rotation on transmission function for all four nucleotides at different rotational configurations varying from 0º to 180º in an interval of 30º while located inside the G/h-BN nanopore.

For the ML study, we have first prepared four transmission databases, considering their seven rotated configuration with individual nucleotides. Each database consists of 3500 (7 × 500 data in the energy window of -2.5 to $+2.5\ eV$) transmission function data points. Selecting relevant, concise, and easily accessible descriptors that can characterize these transmission datasets is crucial for better results of ML applications. The coupling strength of nucleotides with the nanopore edges has a significant impact on transmission readouts of a particular device as reported by some previously published theoretical and experimental studies *[76,77]*. Thus, the input features are carefully extracted based on the chemical composition of individual nucleotides and their different environment present in the vicinity of C-H, B-H, and N-H substituted nanopore edges, as shown in **Table 6.1**.

Table 6.1: The Detailed Overview of Selected Features along with their Description for Machine Learning Regression

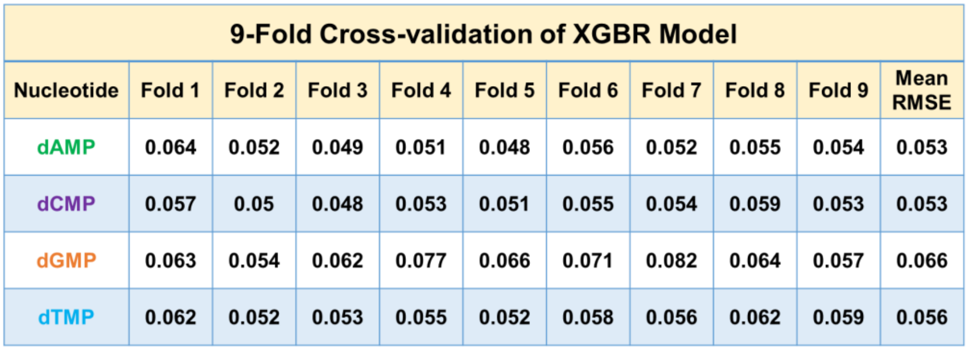
Serial No.	Type of Features	Feature Name	Description of Feature
1	Electronic Properties	Mean electronegativity (EN)	Mean electronegativity of interacting atoms in nucleotides
2		Mean electron affinity (EA)	Mean electron affinity of interacting atoms in nucleotides
3		Mean ionization energy (IE)	Mean ionization energy of interacting atoms in nucleotides
4		Mean dipole polarizability	Mean dipole polarizability of interacting atoms in nucleotides
5		Mean Z effective	Mean effective nuclear charge of interacting atoms in nucleotides
6		Energy	Total energy range of calculated transmission
7	Atomic Properties	Mean valence electrons	Mean number of valence electrons of interacting atoms in nucleotides
8		Mean molecular weight (MW)	Mean molecular weight of interacting atoms in nucleotides
9		Mean covalent radius	Mean covalent radius of interacting atoms in nucleotides
10		Mean VdW radius	Mean Van der Waals radius of interacting atoms in nucleotides
11	Structural	d_{min}(C-H & H)	C-H environment

12		d_{min}(C-H & N)	
13		d_{min}(C-H & O)	
14		d_{min}(N-H & H)	N-H environment
15		d_{min}(N-H & N)	
16		d_{min}(N-H & O)	
17		d_{min}(B-H & H)	B-H environment
18		d_{min}(B-H & N)	
19		d_{min}(B-H & O)	

The considered descriptors can be grouped into two distinct categories: electronic and chemical properties: (a) related to the target nucleotide and (b) those describing the local electronic environment within the hybrid G/h-BN nanopore. It includes electronic (mean EN, mean EA, mean IE, mean dipole polarizability, mean Z effective, and energy) and atomic (mean valence electron, mean MW, mean covalent radius and mean VdW radius) properties of respective nucleotides. However, the different electronic environments ($\mathrm{B}^{\delta+} - \mathrm{H}^{\delta-}$, $\mathrm{C}^{\delta-} - \mathrm{H}^{\delta+}$, and $\mathrm{N}^{\delta-} - \mathrm{H}^{\delta+}$) created inside the G/h-BN nanopore manifests unusual polarity differences. These three groups of descriptors determine the extent of distance-dependent dipolar coupling and H-bonding interactions of nanopore edges with the neighbouring atoms of nucleotides. These electronic, atomic, and structural descriptors can play a pivotal role in explaining the signature transmission function of nucleotides. It can be noted that the minimum interactive distances (d_{min}) between the O, N, and H atoms of nucleotides and functionalized hybrid nanopore edge atoms (H) are considered at ≤ 3.0 Å, which is reported to be well within the range of H-bonding and non-covalent interactions *[78,79]*.

After feature engineering, ML regression analysis is followed through using the transmission databases of four DNA nucleotides (dAMP, dCMP, dGMP, and dTMP). The databases are randomly rearranged and divided into two independent training and test datasets in a ratio of 80:20. These datasets are applied to several considered regression models (LR, RFR, KRR, GPR, AdaBoost, ETR, and XGBR) as available in open-sourced scikit-learn package*[73]* and assessed their performance based on the least test RMSE scores. The XGBR model has exhibited superior performance by obtaining the lowest test RMSE scores for all four nucleotides. In order to provide an impartial assessment of the XGBR model efficiency, a 9-fold cross-validation technique was utilized followed by calculation of mean test RMSE score, as summarized in **Table 6.2**.

Table 6.2: RMSE Score for Each Fold (1−9) of the 9-fold Cross-Validation along with Mean and Test RMSE Scores for the XGBR Model for All Four Nucleotides



9-Fold Cross-validation of XGBR Model										
Nucleotide	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Mean RMSE
dAMP	0.064	0.052	0.049	0.051	0.048	0.056	0.052	0.055	0.054	0.053
dCMP	0.057	0.05	0.048	0.053	0.051	0.055	0.054	0.059	0.053	0.053
dGMP	0.063	0.054	0.062	0.077	0.066	0.071	0.082	0.064	0.057	0.066
dTMP	0.062	0.052	0.053	0.055	0.052	0.058	0.056	0.062	0.059	0.056

The mean cross-validation test RMSE scores closely match the test RMSE score throughout all four datasets confirming model stability. The repeated process of cross-validation evaluates performance and helps to find out the most effective hyper-parameters for the XGBR model across all four datasets. **Figure 6.6** shows the parity plot of train-test predictive performance that compares the ML-predicted transmission with the DFT-calculated transmission function. It confirms that XGBR models are adequately trained and capable of capturing the underlying relations between the attributes and nucleotide transmission.

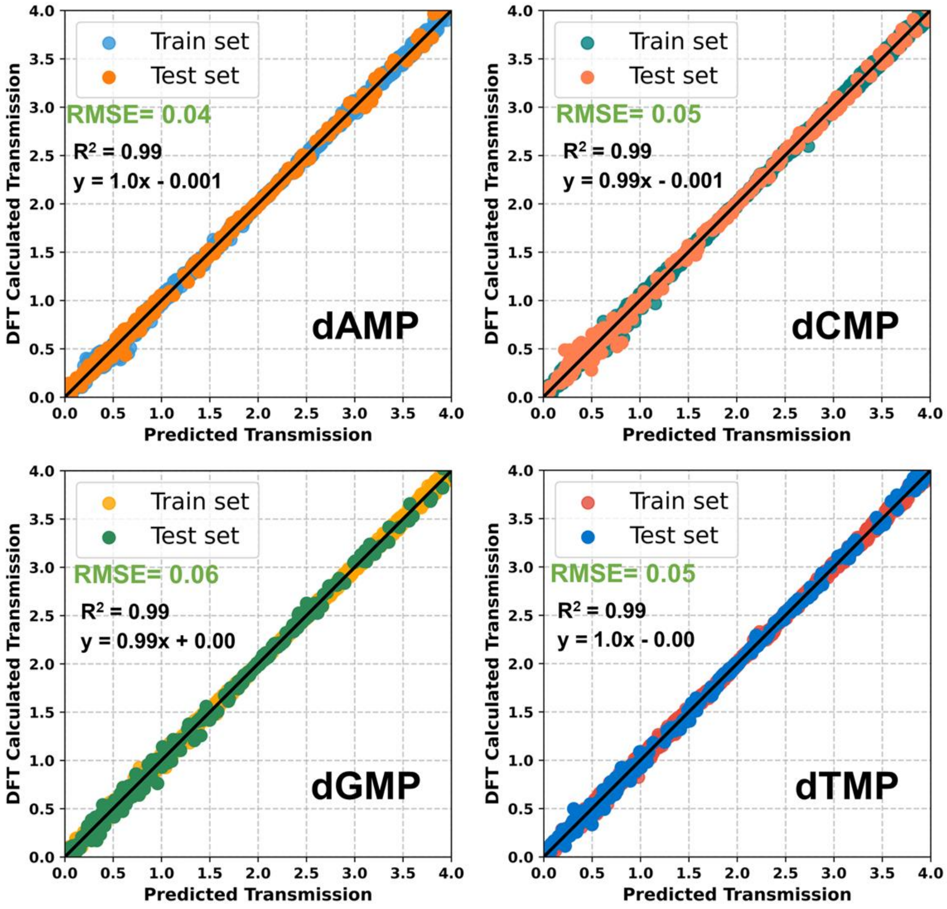


Figure 6.6: Parity plots of ML predicted versus DFT calculated transmission function along with test RMSE scores obtained from optimized XGBR model for dAMP, dCMP, dGMP, and dTMP nucleotides.

The plausible test results of these four models may be ascribed to the high quality of the datasets and also imply their robust correlation with the considered features with a lower percentage of outliers. The XGBR model operates on a gradient boosting framework which is often categorized as a black box model because of its complexity and a large number of hyper-parameters *[80]*. Therefore, it may be possible to deduce the complicated correlation between the structural, chemical, and local electrical environment-based descriptors and transmission function by incorporating decision trees driven ensemble learning method. Furthermore, we have also predicted the signature transmission of the three nucleotides with the optimized XGBR model that was trained with other single nucleotide datasets, as shown in **Figure 6.6**.

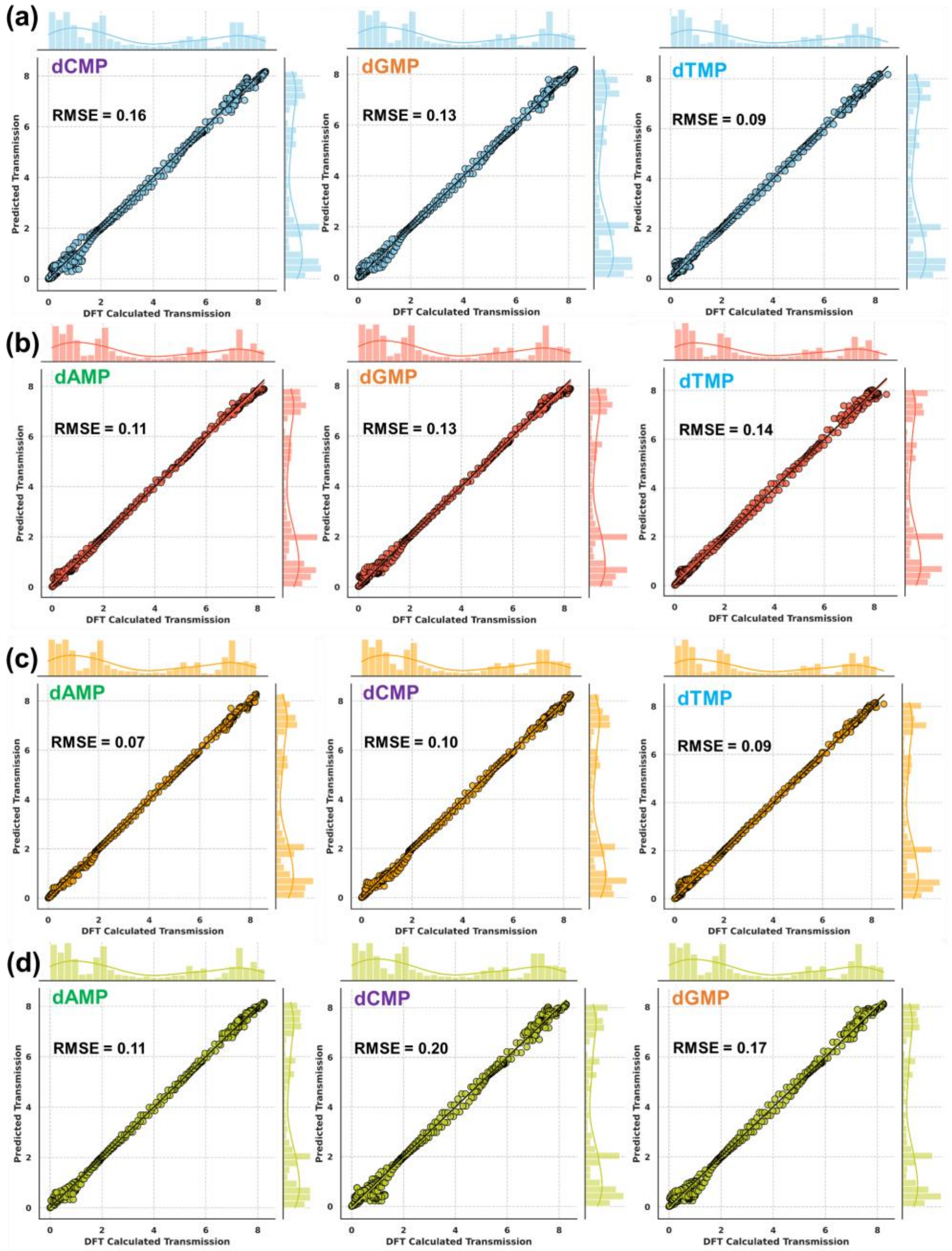


Figure 6.7: Parity plots of ML predicted versus DFT calculated transmission function along with test RMSE scores obtained from optimized XGBR model after training upon (a) dAMP, (b) dCMP, (c) dGMP, and (d) dTMP datasets.

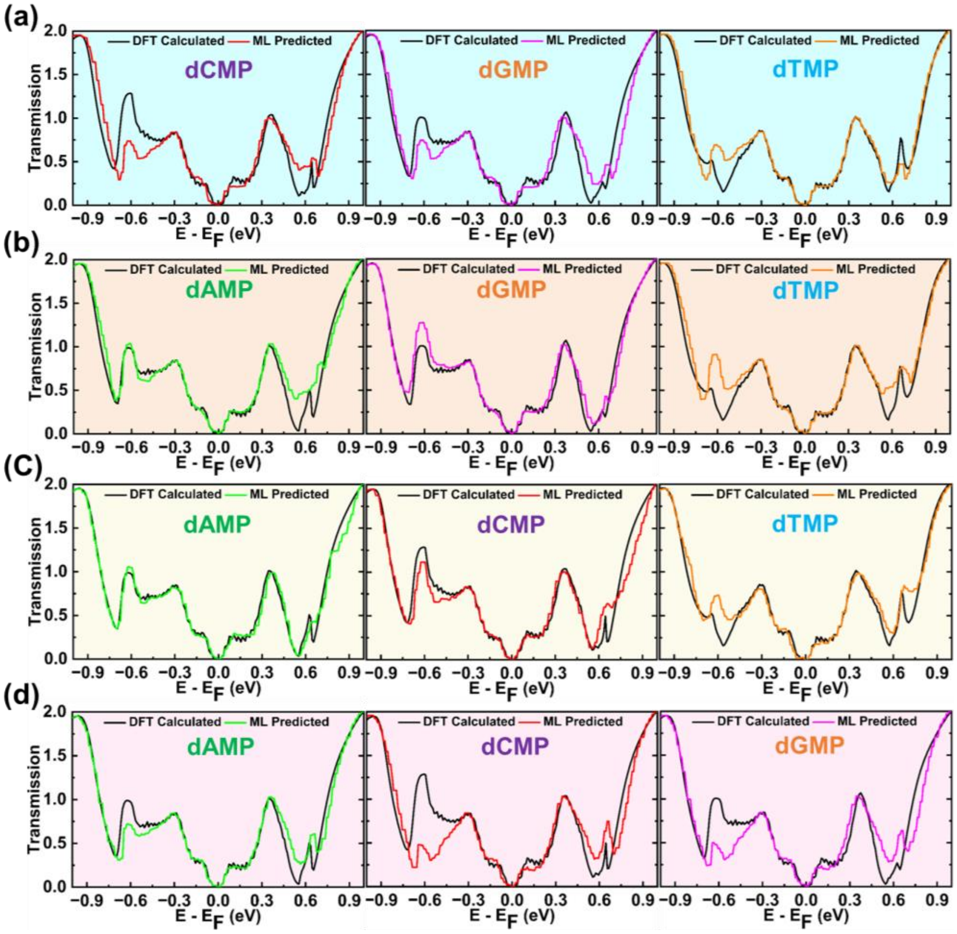
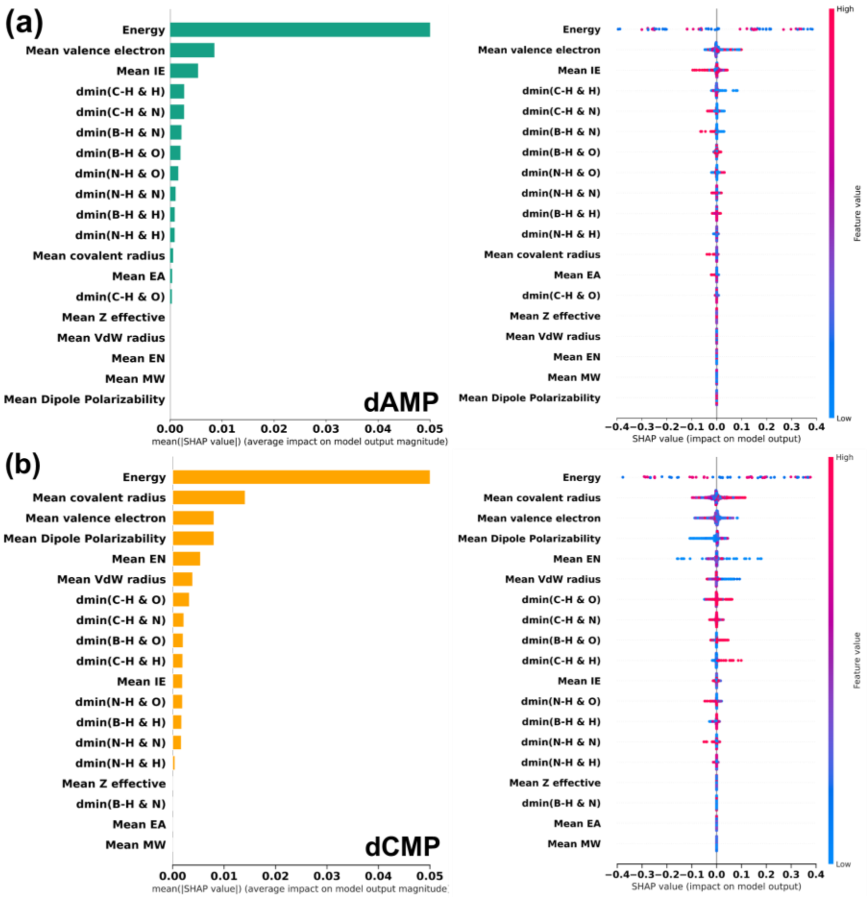


Figure 6.8: XGBR model predicted and DFT calculated transmission function for the most stable rotational configuration of nucleotides after training upon (a) dAMP, (b) dCMP, (c) dGMP, and (d) dTMP datasets.

From the parity plots of ML prediction, the ML-predicted versus the DFT-calculated transmission function can be visualized across all four nucleotide dataset pools. From the low test RMSE scores, it is evident that the optimized XGBR model can precisely predict the fingerprint transmission function of other nucleotides. Among the four nucleotides, the dGMP dataset helps to provide a more generalized prediction for dAMP, dCMP, and dTMP with excellent accuracy with test RMSE ≤ 0.1, as depicted in the parity plot in **Figure 6.7c**. The predicted transmission exhibits precise recognition of the peak locations as reported in the DFT calculated transmission readouts (**Figure 6.8a-d)**. The intrinsic electronic property of nanopore and the interaction of adjacent atoms on nucleotide and (C-H, B-H, and N-H) substituted nanopore edges are the significant contributing factors behind sharp changes in height, width, and position of the peaks in transmission function. The results highlight that the optimized XGBR algorithms, employed in this investigation, exhibit adequate competency in predicting other nucleotides. Therefore, by gathering the transmission readouts of a single nucleotide, one may accurately predict the distinctive molecular conductance shown by the remaining three nucleotides.

The XGBR model comprehends complex relations in datasets learned by an ensemble of decision trees making it difficult to understand underlying design principles, thereby impeding transparency and interpretability. Hence, we have considered a method called SHapley Additive exPlanation (SHAP) analysis, which is based on cooperative game theory, to gain insight into prediction framework the of the XGBR models, which are essentially black boxes.


(a)
Energy
Mean valence electron
Mean IE
dmin(C-H & H)
dmin(C-H & N)
dmin(B-H & N)
dmin(B-H & O)
dmin(N-H & O)
dmin(N-H & N)
dmin(B-H & H)
dmin(N-H & H)
Mean covalent radius
Mean EA
dmin(C-H & O)
Mean Z effective
Mean VdW radius
Mean EN
Mean MW
Mean Dipole Polarizability
dAMP
0.00 0.01 0.02 0.03 0.04 0.05
mean(|SHAP value|) (average impact on model output magnitude)
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
SHAP value (impact on model output)
High
Feature value
Low
(b)
Energy
Mean covalent radius
Mean valence electron
Mean Dipole Polarizability
Mean EN
Mean VdW radius
dmin(C-H & O)
dmin(C-H & N)
dmin(B-H & O)
dmin(C-H & H)
Mean IE
dmin(N-H & O)
dmin(B-H & H)
dmin(N-H & N)
dmin(N-H & H)
Mean Z effective
dmin(B-H & N)
Mean EA
Mean MW
dCMP
0.00 0.01 0.02 0.03 0.04 0.05
mean(|SHAP value|) (average impact on model output magnitude
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
SHAP value (impact on model output)
High
Feature value
Low

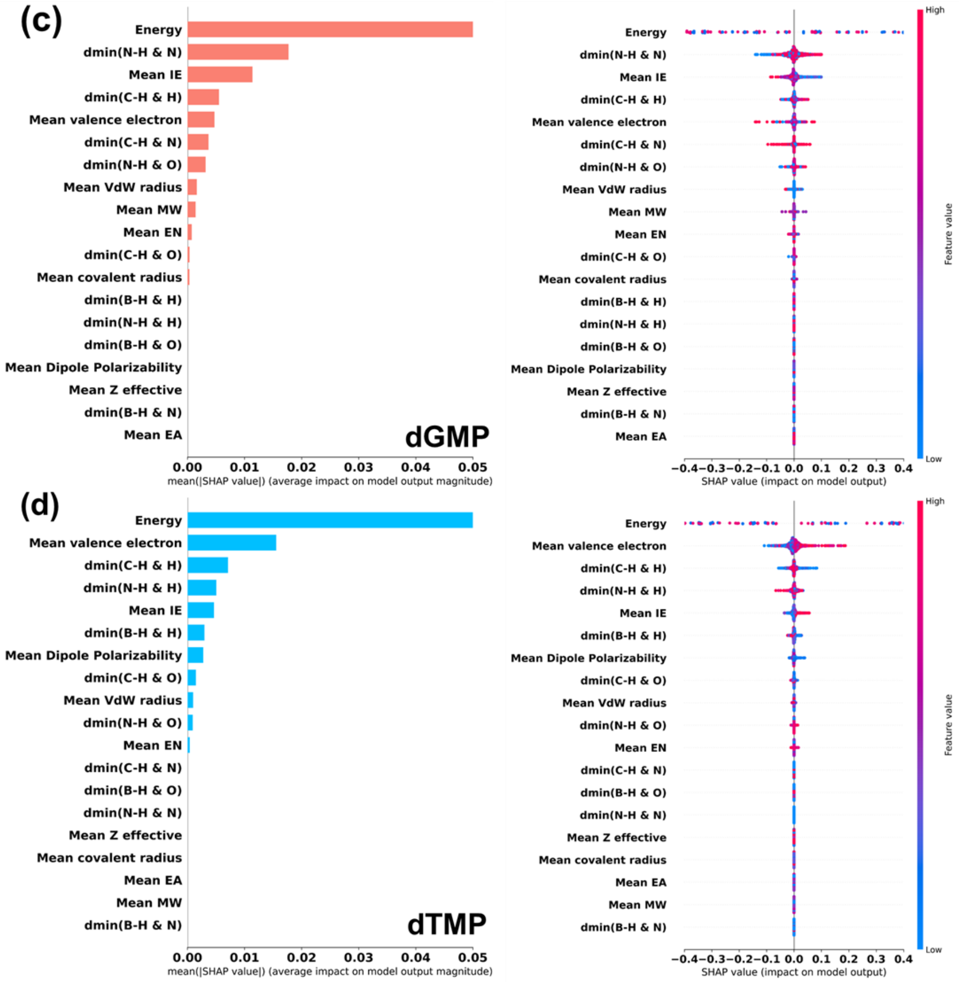


Figure 6.9: Representation of global feature importance and SHAP beeswarm (summary) plot for optimized XGBR model (a) dAMP, (b) dCMP, (c) dGMP, and (d) dTMP datasets. In the bee-swarm plot, the y-axis color gradient indicates the magnitude of the feature, with values ranging from (blue) for lower impact, whereas the red color stands for the higher impact of relevant features in output prediction, and the x-axis consists of positive and negative SHAP values.

The global feature importance and bee-swarm (summary) plots display the hierarchical ranking of the global relevance of the most important descriptors, listed in descending order of the XGBR model trained upon dAMP, dCMP, dGMP, and dTMP datasets respectively (**Figure 6.9a-d**). The summary plot provides a concise and informative overview of how the key features impact (negative/positive) the prediction of the transmission data. The ‘energy’ feature is identified as the top-performing feature in all four cases since the transmission function is calculated as a function of a specific energy range *[11]*. After that, the atomic descriptor (mean valence electrons) is the second top contributing feature in predicting the signature transmission with the XGBR models trained with each nucleotide dataset (**Figure 6.9**). Interestingly, **Figure 6.9a** shows that along with mean IE, the C-H and B-H environment descriptors have relatively higher contributions compared to N-H environment descriptors, for the dAMP nucleotide dataset. This observation could be ascribed to dipolar coupling interactions between pore edges (C-H/B-H) with their nearest electronegative heteroatoms (N and O). For the dCMP dataset, the mean covalent radius and valence electrons are observed to be more promising than structural descriptors in fingerprint transmission prediction (**Figure 6.9b**). For dGMP dataset, the stronger contribution of the N-H environment descriptor is noted, which could be due to the hydrogen bonding interaction between the N-H terminal and aromatic ring containing N atoms on dGMP (**Figure 6.9c**). Structural descriptors (C-H and N-H environments) are noted as dominant features in ML prediction with dTMP dataset, as shown in **Figure 6.9d**. The presence of distance-dependent structural features describing the local electronic environment is found to have a crucial role in the prediction of transmission function. Moreover, the contribution has a substantial impact when the distance between the pore edge and nucleotide atoms is consistently low across the considered rotational configurations.

Furthermore, we have also employed the ML classification to transmission dataset pools in order to distinguish four nucleotides from their widely scattered and significantly overlapped fingerprint signals. ML classification has already been observed to be effective in accurately categorizing nucleotides from quaternary, ternary, and binary combinations of nucleotides from their experimental electric measurements.[81] Before employing classification algorithms, we have extracted four distinct features (TF, Min, Max, and Mean) by normalizing the calculated original transmission data as summarized in **Table 6.3**.

Table 6.3: The Detailed Overview of Selected Features along with their Description for ML Classification

Serial No.	Feature Name	Description of Feature
1	**TF**	The DFT-NEGF calculated transmission dataset
2	**Min**	Each transmission dataset is normalized by dividing with the minimum transmission value (TF/TF_{min})
3	**Max**	Each transmission dataset is normalized by dividing with the maximum transmission value (TF/TF_{max})
4	**Mean**	Each transmission dataset is normalized by dividing with the mean transmission value (TF/TF_{mean})

Among the considered ML classifiers, the random forest classifier (RFC) algorithm has exhibited better performance than KNN*[74]*, DTC*[75]*, GPR, and SVM*[82]* algorithms after comparing train-test accuracy scores as elucidated in **Figure 6.10a**.

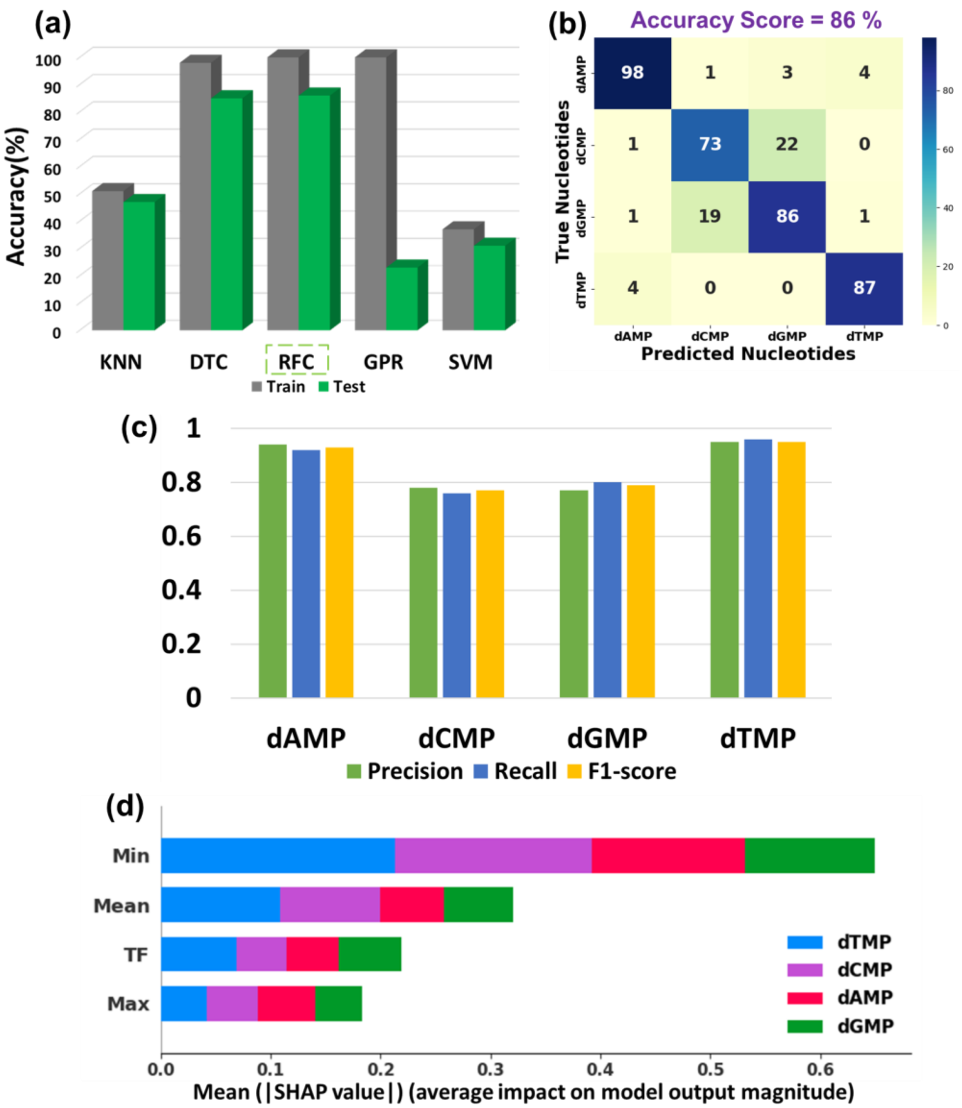


Figure 6.10: (a) Train-test accuracy histograms for the employed ML classification algorithms, (b) RFC model predicted confusion matrix for quaternary classification along with test accuracy score, (c) histogram plots of precision, recall, and F1-score, and (d) SHAP analysis of each feature contributing to quaternary classification. The considered features are defined as TF = DFT calculated transmission data, Min = TF/TF$_{min}$, Mean = TF/TF$_{mean,}$ and Max = TF/TF$_{max}$.

The RFC is an ensemble learning algorithm that incorporates multiple decision trees throughout the training process, with each tree being trained on a randomly selected part of the training data and a randomly selected subset of the features. The use of randomization serves the purpose of mitigation of overfitting. The comparison study of true nucleotide and predicted nucleotide for quaternary classification is presented in the form of a confusion matrix in **Figure 6.10b** with an overall test accuracy score of 86%. In order to evaluate the reliability of the RFC algorithm in quaternary classification, a 7-fold cross-validation technique was utilized followed by the calculation of mean test accuracy score, as summarized in **Table 6.4**.

Table 6.4: Summary of k-Fold Cross-Validation for Quaternary Classification using RFC Model along with Mean and Test Accuracy Scores for ATGC datasets below Fermi (BF) and Above Fermi (AF) Level

Data Sets	S1	S2	S3	S4	S5	S6	S7	Mean	Test
ATGC	80	84	88	86	84	78	86	84	86
ATGC-BF	83	82	79	69	77	75	77	77	76
ATGC-AF	71	76	70	67	68	74	74	71	69

The quaternary classification performance using the RFC algorithm has been assessed by analyzing the histogram plots of precision, recall, and F1 score as displayed in **Figure 6.10c**. From the SHAP analysis, it is interesting to observe that the normalized ‘Min’ (TF/T_{min}) and ‘Mean’ (TF/T_{mean}) features are found to be more effective for quaternary classification than their pure transmission data ‘TF’ of nucleotides (**Figure 6.10d** *[83]*. The quaternary classification is also performed using the collected transmission data of below and above the Fermi level, that showed a drop in the overall test accuracy score for both instances. However, the calculated transmission data at the below Fermi level classifies the nucleotides more accurately with a test accuracy of 76% as compared to ‘TF’ at the above Fermi level.

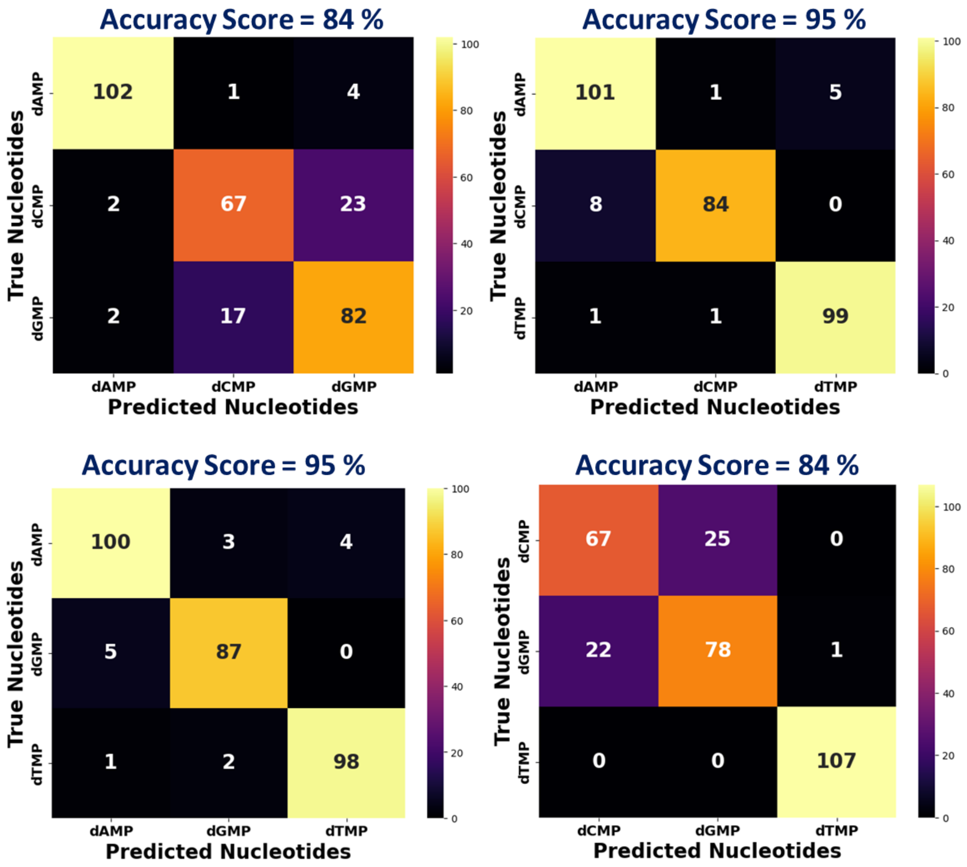


Figure 6.11: Confusion matrix for ternary classification for four possible combinations (ACG, ACT, AGT, and CGT).

Within ternary categorization, the AGT and ACT nucleotide combination achieves a higher accuracy score of 95%, while the ACG and TCG combination yields a lesser test accuracy of 84% as shown in **Figure 6.11**. The prediction performance of the optimized RFC model is visualized in the form of confusion matrices for binary classification as shown in **Figure 6.12**.

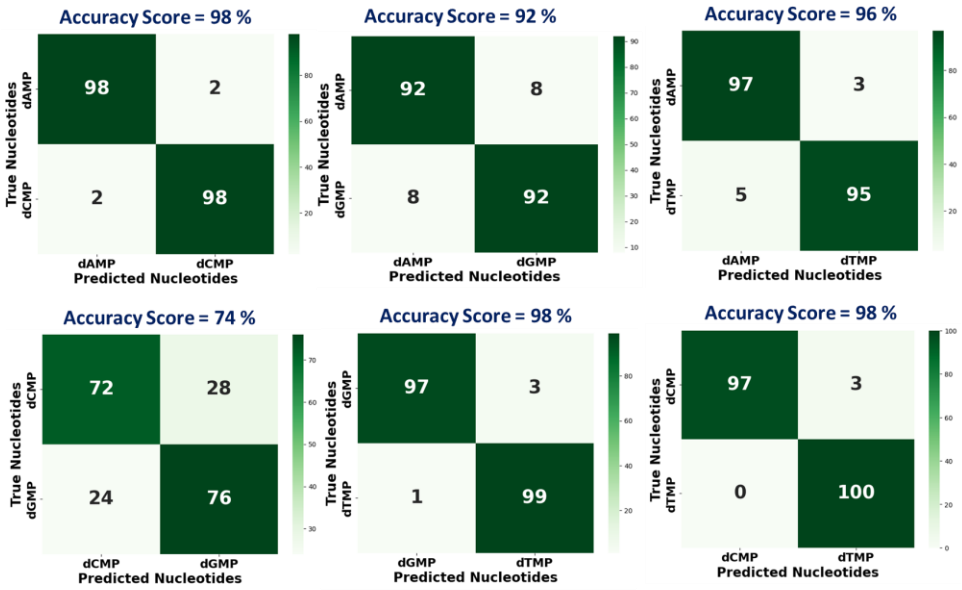


Figure 6.12: Confusion matrix and accuracy scores for binary classification of six possible combinations (AC, AG AT, CG, GT, and CT).

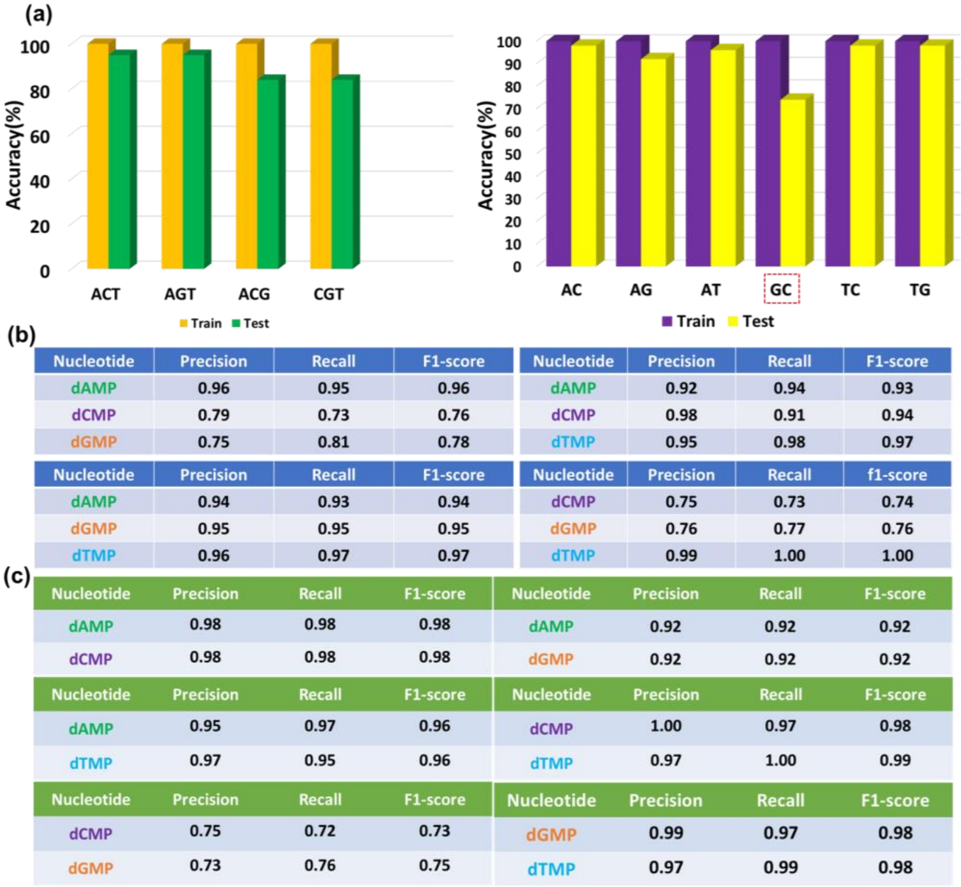
The test accuracy of binary classification was achieved as high as 98% for six possible combinations (AC, AG, AT, CG, CT, and GT), as shown in **Figure 6.13a**. The high classification accuracy is also evident from their precision, recall, and F1 scores (**Figure 6.13b-c)**. To assess the robustness of the RFC algorithm, 7-fold cross-validation is implemented in smaller subsets of the nucleotide data pool for ternary and binary classification (**Tables 6.5** and **6.6)**.

Table 6.5: Summary of k-Fold Cross-Validation for Ternary Classification using RFC Model along with Mean and Test Accuracy Score for Each Fold (S1−S7)

Data Sets	S1	S2	S3	S4	S5	S6	S7	Mean	Test
ACT	94	96	95	91	93	94	94	94	95
AGT	94	92	96	92	90	89	94	92	95
ACG	84	84	77	78	77	84	78	80	84
CGT	84	84	84	80	80	84	81	82	84

Table 6.6: Summary of k-Fold Cross-Validation for Binary Classification using RFC Model along with Mean and Test Accuracy Score for Each Fold (S1−S7)

Data Sets	S1	S2	S3	S4	S5	S6	S7	Mean	Test
AC	90	96	98	92	97	96	92	94	98
AG	89	96	92	90	96	96	91	93	92
AT	92	96	94	92	96	94	96	94	96
CG	64	76	75	68	74	80	76	73	74
CT	96	96	100	95	96	97	100	97	98
GT	96	96	97	96	95	97	97	96	98



Nucleotide	Precision	Recall	F1-score
dAMP	0.96	0.95	0.96
dCMP	0.79	0.73	0.76
dGMP	0.75	0.81	0.78

Nucleotide	Precision	Recall	F1-score
dAMP	0.92	0.94	0.93
dCMP	0.98	0.91	0.94
dTMP	0.95	0.98	0.97

Nucleotide	Precision	Recall	F1-score
dAMP	0.94	0.93	0.94
dGMP	0.95	0.95	0.95
dTMP	0.96	0.97	0.97

Nucleotide	Precision	Recall	f1-score
dCMP	0.75	0.73	0.74
dGMP	0.76	0.77	0.76
dTMP	0.99	1.00	1.00

Nucleotide	Precision	Recall	F1-score
dAMP	0.98	0.98	0.98
dCMP	0.98	0.98	0.98

Nucleotide	Precision	Recall	F1-score
dAMP	0.92	0.92	0.92
dGMP	0.92	0.92	0.92

Nucleotide	Precision	Recall	F1-score
dAMP	0.95	0.97	0.96
dTMP	0.97	0.95	0.96

Nucleotide	Precision	Recall	F1-score
dCMP	1.00	0.97	0.98
dTMP	0.97	1.00	0.99

Nucleotide	Precision	Recall	F1-score
dCMP	0.75	0.72	0.73
dGMP	0.73	0.76	0.75

Nucleotide	Precision	Recall	F1-score
dGMP	0.99	0.97	0.98
dTMP	0.97	0.99	0.98

Figure 6.13: (a) Train-test accuracy histograms for ternary and binary classification, (b) ternary, and (c) binary classification report including precision, recall, and F1-score in tabular format after implementing RFC algorithm.

Interestingly, the overall test accuracy drops for the datasets including dGMP and dCMP for both ternary and binary classification. The probable reason behind the sudden drop in test accuracy scores could be the consequence of significant overlapping of fingerprint transmission function. However, the trained model revives its performance after being trained with the ‘TF’ data of all four nucleotides in quaternary classification. By comparing the test accuracy scores for quaternary, ternary, and binary classification, it can be anticipated that the larger training dataset can minimize the classification error for dGMP and dCMP as implemented using RFC algorithm.

Figures 6.14 and **6.15** illustrate the relationship between the true positive rate (TPR) and the false positive rate (FPR) for each set of ternary and binary combinations, respectively.

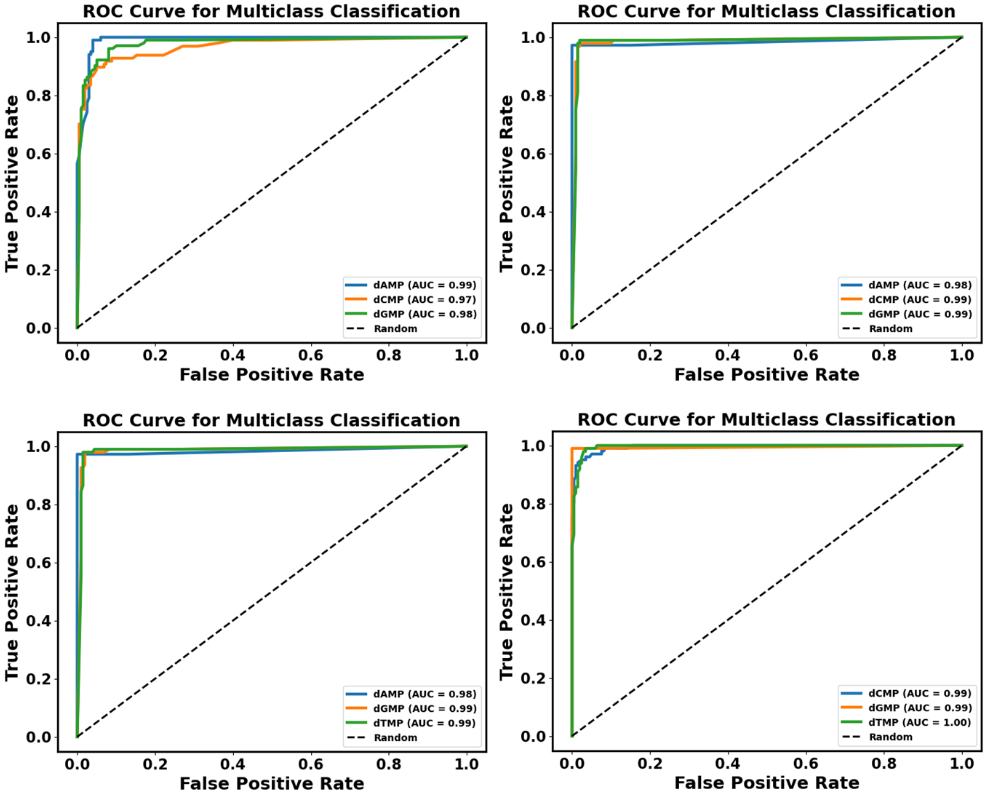


Figure 6.14: Visual representation of area under ROC curve for ternary classification utilizing RFC model.

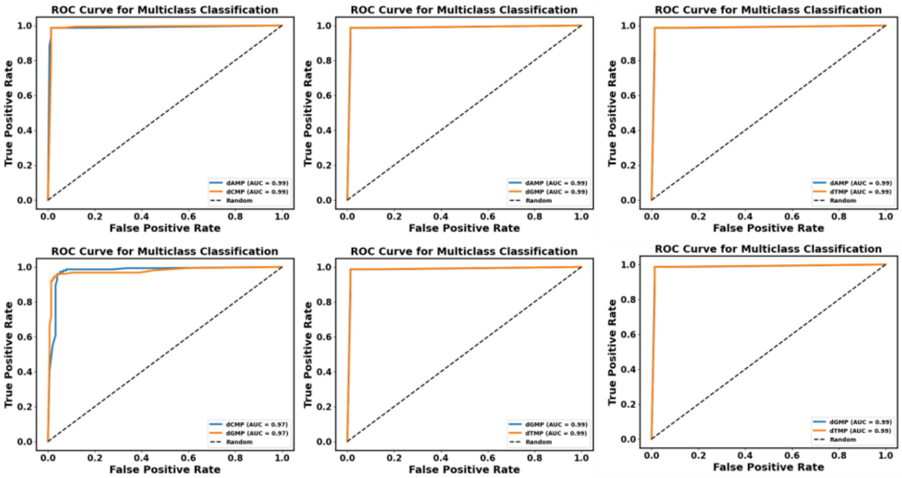


Figure 6.15: Visual representation of area under ROC curve for binary classification utilizing RFC model.

The higher value of the area under the ROC curve suggests that our RFC model emphasizes the true positive prediction of nucleotides rather than the false positive prediction. Further to understand the underlying contribution of calculated transmission data and the other three normalized features in classification, SHAP analysis was done for each set of ternary and binary combinations. **Figure 6.16** describes global feature importance as well as the hierarchical order of contribution in each class in all ternary combinations. It is worth noticing that the normalized feature 'Min' has a higher contribution as compared to the calculated transmission data for the combinations ACG and AGT. In the case of ACT and CGT combinations, all three normalized features impact ternary classification more than calculated TF.

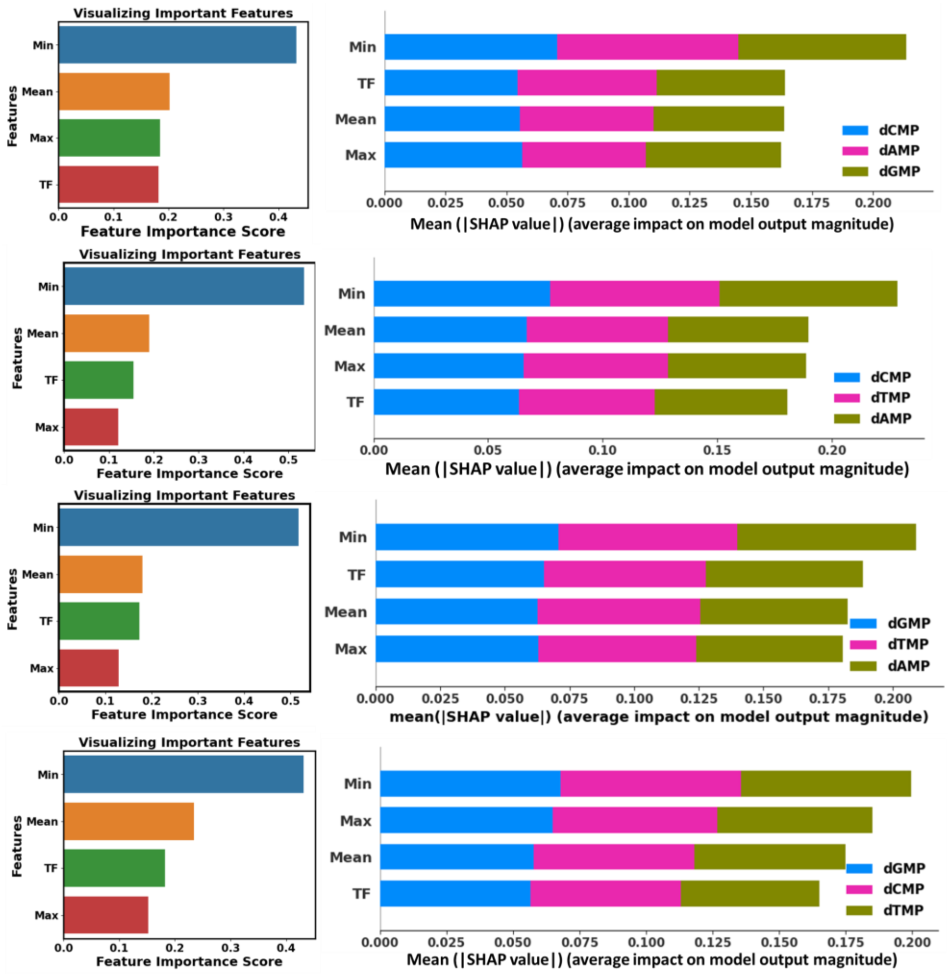


Figure 6.16: Visualization of feature importance and SHAP analysis of features contributing to ternary classification of each nucleotide by implementing RFC model.

In the case of binary combinations (**Figure 6.17**), all three normalized features are observed to contribute more as compared to our calculated TF data for AC and CG. However, the classification of dAMP and dGMP is significantly impacted by the 'Min' feature as compared to other three descriptors. For the remaining three binary combinations (CT, AT, and GT), the two normalized features 'Min' and 'Mean' secure position among the top two contributing features. The higher contribution of normalized features can be explained in the spotlight of the different scales of data after the mathematical operation on our calculated transmission data. These different scales help the RFC classifier to find patterns in data and ease the identification of single nucleotides.

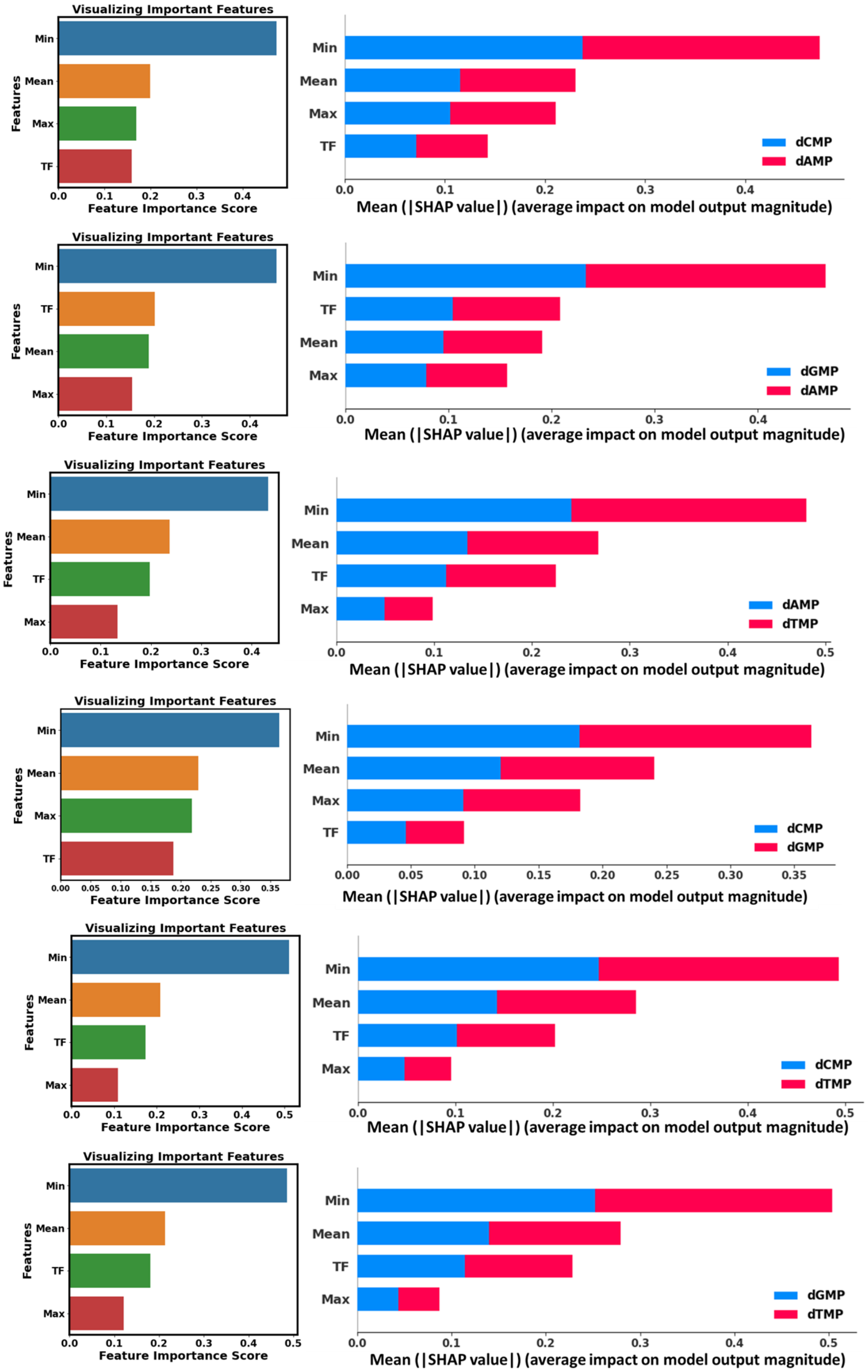


Figure 6.17: Visualization of feature importance and SHAP analysis of features contributing to binary classification of each nucleotide by implementing RFC model.

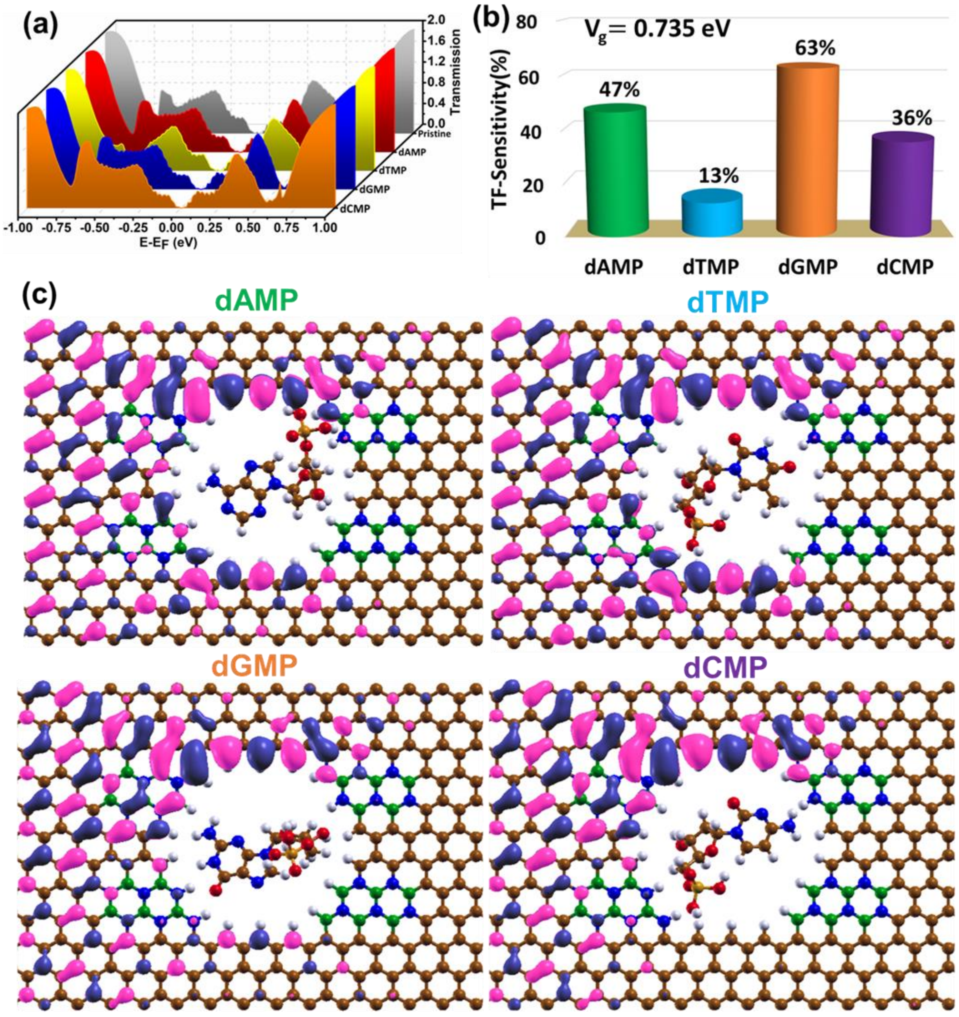


Figure 6.18: (a) Zero bias transmission function of four nucleotides (dAMP, dTMP, dGMP, and dCMP) while present inside G/h-BN nanopore, (b) conductance sensitivity (%) histogram plots with respect to the pristine nanopore, and (c) wave function analysis of G/h-BN nanopore + nucleotides at gate voltage $(V_g) = 0.735\ eV$.

Furthermore, we have also analyzed the zero bias transmission function of all four nucleotides at their most stable configurations as shown in **Figure 6.18a**. The sharp variation in the transmission signal of nucleotides below the Fermi level can be attributed to the coupling of nucleotide molecular orbitals (MOs) and (C-H, N-H, and B-H) passivated nanopore edges through non-covalent interactions as discussed earlier. Further, we have calculated conductance sensitivity (%) which is an essential property of DNA sequencing to evaluate the identification capability of the considered G/h-BN nanopore device by using the following equation:

$$Sensitivity\ (\%)\ =\ \left|\frac{G_0 - G_x}{G_x}\right| \times 100$$

Where, G_0 is the conductance of the pristine G/h-BN nanopore and G_x is the conductance of each nucleotide (dAMP, dGMP, dCMP, and dTMP) inside the nanopore. The histogram plot for conductance sensitivity is depicted in **Figure 6.18b**. In experimental conditions, the conductance sensitivity (%) can be measured by applying an external gate voltage (V_g) = 0.735V) for the G/h-BN nanopore device. From this study, the conductance sensitivity values are observed to be higher for purine and lower for pyrimidine type of nucleotide device. The conductance sensitivity analysis is found to be in the following order: dGMP > dAMP > dCMP > dTMP. Here, we deduced that the hybrid G/h-BN nanopore has enhanced the high-resolution conductance sensitivity of four nucleotides for single nucleobase identification.

Finally, to provide a better insight into the coupling interaction between nucleotide and the substituted nanopore edges the wave function analysis is performed at the same energy of $E - E_F\ =\ 0.735\ eV$ where the highest conductance sensitivity is observed as shown in **Figure 6.18c**. The wave function analysis provides insight into the probability of flowing electrons from the source to the drain corresponding to a specific energy value. At the energy ($E - E_F\ =\ 0.735\ eV$) for dAMP and dTMP, there is only a minute change in transmission compared to pristine G/h-BN pore, Whereas, there is a drop in the transmission of electrons through one of the nanopore edges which can be attributed to the current modulation effect for dGMP and dCMP *[84,85]*. The ring (O) atom in dGMP present close to the pore could be resulting in an imposition of negative potential on the nanopore edge due to excess partial negative charge on it. This negative potential blocks one of the transmission channels and compels the electrons to flow from the other alternative direction. As dCMP is in slightly skewed conformation inside the pore, the closely spaced oxygen (O) atom of the phosphate group exerts the same influence resulting in current modulation.

Chapter 7

Conclusions and Scope for Future Work

In a nutshell, we have studied a hybrid G/h-BN nanopore for single nucleotide-based DNA sequencing using both machine learning regression and classification framework combined with quantum transport approach. Initially, we prepared four databases by calculating the transmission function of all four nucleotides at their energetically favorable and six other dynamic configurations inside the nanopore with the DFT-NEGF approach. The optimized XGBR models, trained using one nucleotide database, can accurately predict the other three nucleotides with an RMSE score as low as 0.07. The ML explainability with SHAP analysis revealed that electronic descriptors (mean IE) and atomic descriptors (mean valence electrons and mean covalent radius) along with energy have a strong influence on the prediction of the transmission function. Among structural descriptors, the C-H and B-H environments are observed to have a significant impact on the prediction of dAMP and dCMP nucleotides, while N-H and C-H environment descriptors have more contribution towards the prediction of dGMP and dTMP nucleotides. The relatively lower impact of the B-H environment could be attributed to its opposite polarity ($\mathrm{B}^{\delta+} - \mathrm{H}^{\delta-}$). RFC classification of quaternary, ternary, and binary combinations of nucleotides has achieved maximum accuracy of 86%, 95%, and 98%, respectively. However, dGMP and dCMP nucleotides displayed lower accuracies across all three types of combinations, possibly due to significantly overlapped transmission readouts. Additionally, the conductance sensitivity analysis demonstrated that purine nucleotides have considerably higher sensitivity than pyrimidine nucleotides. Frontier molecular orbitals with wavefunctions analysis are also conducted to elucidate the impact of electronic coupling interactions between the nucleotides and nanopore edges on their transmission functions. This proof-of-concept ML study with hybrid nanoscale devices demonstrates a potential platform for single nucleotide-based DNA sequencing that can garner interest among researchers for further investigations.

REFERENCES

(1) Heng, H. H. Q.; Liu, G.; Stevens, J. B.; Bremer, S. W.; Ye, K. J.; Abdallah, B. Y.; Horne, S. D.; Ye, C. J. (2011), Decoding the Genome beyond Sequencing: The New Phase of Genomic Research. Genomics, 98, 242–252 (DOI: 10.1016/j.ygeno.2011.05.008)

(2) Esplin, E. D.; Oei, L.; Snyder, M. P. (2014), Personalized Sequencing and the Future of Medicine: Discovery, Diagnosis and Defeat of Disease. Pharmacogenomics, 15, 1771–1790 (DOI: 10.2217/pgs.14.117)

(3) Wang, Y.; Yang, Q.; Wang, Z. (2015), The Evolution of Nanopore Sequencing. Front. Genet., 5 (DOI: 10.3389/fgene.2014.00449)

(4) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. (1996) Characterization of Individual Polynucleotide Molecules Using a Membrane Channel. Proc. Natl. Acad. Sci., 93, 13770–13773 (DOI: 10.1073/pnas.93.24.13770)

(5) Manrao, E. A.; Derrington, I. M.; Laszlo, A. H.; Langford, K. W.; Hopper, M. K.; Gillgren, N.; Pavlenok, M.; Niederweis, M.; Gundlach, J. H. (2012), Reading DNA at Single-Nucleotide Resolution with a Mutant MspA Nanopore and Phi29 DNA Polymerase. Nat. Biotechnol., 30, 349–353 (DOI: 10.1038/nbt.2171)

(6) Hagemann, I. S. (2015), Chapter 1 - Overview of Technical Aspects and Chemistries of Next-Generation Sequencing. In Clinical Genomics; Kulkarni, S., Pfeifer, J., Eds.; Academic Press: Boston, pp 3–19 (DOI: 10.1016/B978-0-12-404748-8.00001-0)

(7) Lagerqvist, J.; Zwolak, M.; Di Ventra, M. (2006), Fast DNA Sequencing via Transverse Electronic Transport. Nano Lett., 6, 779–782 (DOI: 10.1021/nl0601076)

(8) Postma, H. W. C. (2010), Rapid Sequencing of Individual DNA Molecules in Graphene Nanogaps. Nano Lett., 10, 420–425 (DOI: 10.1021/nl9029237)

(9) Schneider, G. F.; Kowalczyk, S. W.; Calado, V. E.; Pandraud, G.; Zandbergen, H. W.; Vandersypen, L. M. K.; Dekker, C. (2010), DNA Translocation through Graphene Nanopores. Nano Lett., 10, 3163–3167 (DOI: 10.1021/nl102069z)

(10) Garaj, S.; Hubbard, W.; Reina, A.; Kong, J.; Branton, D.; Golovchenko, J. A. (2010), Graphene as a Subnanometre Trans-Electrode Membrane. Nature, 467, 190–193 (DOI: 10.1038/nature09379)

(11) Min, S. K.; Kim, W. Y.; Cho, Y.; Kim, K. S. (2011), Fast DNA Sequencing with a Graphene-Based Nanochannel Device. Nat. Nanotechnol., 6, 162–165 (DOI: 10.1038/nnano.2010.283)

(12) Cho, Y.; Min, S. K.; Kim, W. Y.; Kim, K. S. (2011), The Origin of Dips for the Graphene-Based DNA Sequencing Device. Phys. Chem. Chem. Phys., 13, 14293–14296 (DOI: 10.1039/C1CP20760A)

(13) Rezapour, M. R.; Rajan, A. C.; Kim, K. S. (2014), Molecular Sensing Using Armchair Graphene Nanoribbon. J. Comput. Chem., 35, 1916–1920 (DOI: 10.1002/jcc.23705)

(14) Rajan, A. C.; Rezapour, M. R.; Yun, J.; Cho, Y.; Cho, W. J.; Min, S. K.; Lee, G.; Kim, K. S. (2014), Two Dimensional Molecular Electronics Spectroscopy for Molecular Fingerprinting, DNA Sequencing, and Cancerous DNA Recognition. ACS Nano, 8, 1827–1833 (DOI: 10.1021/nn4062148)

(15) Thomas, S.; Rajan, A. C.; Rezapour, M. R.; Kim, K. S. (2014), In Search of a Two-Dimensional Material for DNA Sequencing. J. Phys. Chem. C, 118, 10855–10858 (DOI: 10.1021/jp501711d)

(16) Amorim, R. G.; Scheicher, R. H. (2015), Silicene as a New Potential DNA Sequencing Device. Nanotechnology, 26, 154002 (DOI: 10.1088/0957-4484/26/15/154002)

(17) Shendure, J.; Ji, H. (2008), Next-Generation DNA Sequencing. Nat. Biotechnol., 26, 1135–1145 (DOI:10.1038/nbt1486)

(18) Clarke, J.; Wu, H.-C.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H. (2009), Continuous Base Identification for Single-Molecule

Nanopore DNA Sequencing. Nat. Nanotechnol., 4, 265–270 (DOI: 10.1038/nnano.2009.12)

(19) Di Ventra, M.; Taniguchi, M. (2016), Decoding DNA, RNA and Peptides with Quantum Tunnelling. Nat. Nanotechnol., 11, 117–126. (DOI: 10.1038/nnano.2015.320)

(20) Dekker, C. (2007), Solid-State Nanopores. Nat. Nanotechnol., 2, 209–215 (DOI: 10.1038/nnano.2007.27)

(21) Garaj, S.; Liu, S.; Golovchenko, J. A.; Branton, D. (2013), Molecule-Hugging Graphene Nanopores. Proc. Natl. Acad. Sci., 110, 12192–12196 (DOI: 10.1073/pnas.1220012110)

(22) Feliciano, G. T.; Sanz-Navarro, C.; Coutinho-Neto, M. D.; Ordejón, P.; Scheicher, R. H.; Rocha, A. R. (2015), Capacitive DNA Detection Driven by Electronic Charge Fluctuations in a Graphene Nanopore. Phys. Rev. Appl., 3, 034003 (DOI: 10.1103/PhysRevApplied.3.034003)

(23) Wells, D. B.; Belkin, M.; Comer, J.; Aksimentiev, A. (2012), Assessing Graphene Nanopores for Sequencing DNA. Nano Lett., 12, 4117–4123 (DOI: 10.1021/nl301655d)

(24) Jena, M. K.; Kumawat, R. L.; Pathak, B. (2021), First-Principles Density Functional Theory Study on Graphene and Borophene Nanopores for Individual Identification of DNA Nucleotides. ACS Appl. Nano Mater., 4, 13573–13586 (DOI: 10.1021/acsanm.1c03015)

(25) Garoli, D.; Yamazaki, H.; Maccaferri, N.; Wanunu, M. (2019), Plasmonic Nanopores for Single-Molecule Detection and Manipulation: Toward Sequencing Applications. Nano Lett., 19, 7553–7562 (DOI: 10.1021/acs.nanolett.9b02759)

(26) Liang, L.; Cui, P.; Wang, Q.; Wu, T.; Ågren, H.; Tu, Y. (2013), Theoretical Study on Key Factors in DNA Sequencing with Graphene Nanopores. RSC Adv., 3, 2445–2453. (DOI:10.1039/C2RA22109H)

(27) Zhang, Z.; Shen, J.; Wang, H.; Wang, Q.; Zhang, J.; Liang, L.; Ågren, H.; Tu, Y. (2014), Effects of Graphene Nanopore Geometry

on DNA Sequencing. J. Phys. Chem. Lett., 5, 1602–1607 (DOI: 10.1021/jz500498c)

(28) Kwok, H.; Briggs, K.; Tabard-Cossa, V. (2014), Nanopore Fabrication by Controlled Dielectric Breakdown. PLOS ONE, 9, e92880 (DOI: 10.1371/journal.pone.0092880)

(29) Briggs, K.; Charron, M.; Kwok, H.; Le, T.; Chahal, S.; Bustamante, J.; Waugh, M.; Tabard-Cossa, V. (2015), Kinetics of Nanopore Fabrication during Controlled Breakdown of Dielectric Membranes in Solution. Nanotechnology, 26, 084004 (DOI: 10.1088/0957-4484/26/8/084004)

(30) Liu, S.; Lu, B.; Zhao, Q.; Li, J.; Gao, T.; Chen, Y.; Zhang, Y.; Liu, Z.; Fan, Z.; Yang, F.; You, L.; Yu, D. (2013), Boron Nitride Nanopores: Highly Sensitive DNA Single-Molecule Detectors. Adv. Mater., 25, 4549–4554 (DOI: 10.1002/adma.201301336)

(31) Dean, C. R.; Young, A. F.; Meric, I.; Lee, C.; Wang, L.; Sorgenfrei, S.; Watanabe, K.; Taniguchi, T.; Kim, P.; Shepard, K. L.; Hone, J. (2010), Boron Nitride Substrates for High-Quality Graphene Electronics. Nat. Nanotechnol.,5, 722–726 (DOI: 10.1038/nnano.2010.172)

(32) Levendorf, M. P.; Kim, C.-J.; Brown, L.; Huang, P. Y.; Havener, R. W.; Muller, D. A.; Park, J. (2012), Graphene and Boron Nitride Lateral Heterostructures for Atomically Thin Circuitry. Nature, 488, 627–632 (DOI: 10.1038/nature11408)

(33) Ci, L.; Song, L.; Jin, C.; Jariwala, D.; Wu, D.; Li, Y.; Srivastava, A.; Wang, Z. F.; Storr, K.; Balicas, L.; Liu, F.; Ajayan, P. M. (2010), Atomic Layers of Hybridized Boron Nitride and Graphene Domains. Nat. Mater., 9, 430–435 (DOI: 10.1038/nmat2711)

(34) Wasfi, A.; Awwad, F.; Ayesh, A. I. (2019), Electronic Signature of DNA Bases via Z-Shaped Graphene Nanoribbon with a Nanopore. Biosens. Bioelectron. X, *1*, 100011 (DOI: 10.1016/j.biosx.2019.100011)

(35) Souza, F. A. L. de; Sivaraman, G.; Hertkorn, J.; Amorim, R. G.; Fyta, M.; Scopel, W. L. (2019), Hybrid 2D Nanodevices (Graphene/h-BN): Selecting NOx Gas through the Device

Interface. J. Mater. Chem. A, 7, 8905–8911 (DOI: 10.1039/C9TA00674E)

(36) Nelson, T.; Zhang, B.; Prezhdo, O. V. (2010), Detection of Nucleic Acids with Graphene Nanopores: Ab Initio Characterization of a Novel Sequencing Device. Nano Lett., 10, 3237–3242 (DOI: 10.1021/nl9035934)

(37) Shukla, V.; Jena, N. K.; Grigoriev, A.; Ahuja, R. (2017), Prospects of Graphene–hBN Heterostructure Nanogap for DNA Sequencing. ACS Appl. Mater. Interfaces, 9, 39945–39952 (DOI: 10.1021/acsami.7b06827)

(38) Kiakojouri, A.; Frank, I.; Nadimi, E. (2021), In-Plane Graphene/h-BN/Graphene Heterostructures with Nanopores for Electrical Detection of DNA Nucleotides. Phys. Chem. Chem. Phys., 23, 25126–25135 (DOI: 10.1039/D1CP03597E)

(39) Liu, Z.; Ma, L.; Shi, G.; Zhou, W.; Gong, Y.; Lei, S.; Yang, X.; Zhang, J.; Yu, J.; Hackenberg, K. P.; Babakhani, A.; Idrobo, J.-C.; Vajtai, R.; Lou, J.; Ajayan, P. M. (2013), In-Plane Heterostructures of Graphene and Hexagonal Boron Nitride with Controlled Domain Sizes. Nat. Nanotechnol., 8, 119–124 (DOI: 10.1038/nnano.2012.256)

(40) Gong, Y.; Shi, G.; Zhang, Z.; Zhou, W.; Jung, J.; Gao, W.; Ma, L.; Yang, Y.; Yang, S.; You, G.; Vajtai, R.; Xu, Q.; MacDonald, A. H.; MacDonald, A. H.; Yakobson, B. I.; Lou, J.; Liu, Z.; Ajayan, P. M. (2014), Direct Chemical Conversion of Graphene to Boron- and Nitrogen- and Carbon-Containing Atomic Layers. Nat. Commun., 5, 3193 (DOI: 10.1038/ncomms4193)

(41) Meng, J. H.; Zhang, X. W.; Wang, H. L.; Ren, X. B.; Jin, C. H.; Yin, Z. G.; Liu, X.; Liu, H. (2015), Synthesis of In-Plane and Stacked Graphene/Hexagonal Boron Nitride Heterostructures by Combining with Ion Beam Sputtering Deposition and Chemical Vapor Deposition. Nanoscale, 7, 16046–16053 (DOI: 10.1039/C5NR04490A)

(42) Meng, J.; Wang, D.; Cheng, L.; Gao, M.; Zhang, X. (2018), Recent Progress in Synthesis, Properties, and Applications of Hexagonal

Boron Nitride-Based Heterostructures. Nanotechnology, 30, 074003 (DOI: 10.1088/1361-6528/aaf301)

(43) Cui, F.; Yue, Y.; Zhang, Y.; Zhang, Z.; Zhou, H. S. (2020), Advancing Biosensors with Machine Learning. ACS Sens., 5, 3346–3364 (DOI: 10.1021/acssensors.0c01424)

(44) Leong, Y. X.; Tan, E. X.; Leong, S. X.; Lin Koh, C. S.; Thanh Nguyen, L. B.; Ting Chen, J. R.; Xia, K.; Ling, X. Y. (2022), Where Nanosensors Meet Machine Learning: Prospects and Challenges in Detecting Disease X. ACS Nano, 16, 13279–13293 (DOI: 10.1021/acsnano.2c05731)

(45) Misiunas, K.; Ermann, N.; Keyser, U. F. (2018), QuipuNet: Convolutional Neural Network for Single-Molecule Nanopore Sensing. Nano Lett., 18, 4040–4045 (DOI: 10.1021/acs.nanolett.8b01709)

(46) Takashima, Y.; Komoto, Y.; Ohshiro, T.; Nakatani, K.; Taniguchi, M. (2023), Quantitative Microscopic Observation of Base-Ligand Interactions via Hydrogen Bonds by Single-Molecule Counting. J. Am. Chem. Soc., 145, 1310–1318 (DOI: 10.1021/jacs.2c11260)

(47) Taniguchi, M.; Takei, H.; Tomiyasu, K.; Sakamoto, O.; Naono, N. (2022), Sensing the Performance of Artificially Intelligent Nanopores Developed by Integrating Solid-State Nanopores with Machine Learning Methods. J. Phys. Chem. C, 126, 12197–12209 (DOI: 10.1021/acs.jpcc.2c02674)

(48) Im, J.; Sen, S.; Lindsay, S.; Zhang, P. (2018), Recognition Tunneling of Canonical and Modified RNA Nucleotides for Their Identification with the Aid of Machine Learning. ACS Nano, 12, 7067–7075 (DOI: 10.1021/acsnano.8b02819)

(49) Meyer, N.; Janot, J.-M.; Lepoitevin, M.; Smietana, M.; Vasseur, J.-J.; Torrent, J.; Balme, S. (2020), Machine Learning to Improve the Sensing of Biomolecules by Conical Track-Etched Nanopore. Biosensors, 10, 140. (DOI: 10.3390/bios10100140)

(50) Taniguchi, M.; Minami, S.; Ono, C.; Hamajima, R.; Morimura, A.; Hamaguchi, S.; Akeda, Y.; Kanai, Y.; Kobayashi, T.; Kamitani, W.; Terada, Y.; Suzuki, K.; Hatori, N.; Yamagishi, Y.; Washizu, N.;

Takei, H.; Sakamoto, O.; Naono, N.; Tatematsu, K.; Washio, T.; Matsuura, Y.; Tomono, K. (2021), Combining Machine Learning and Nanopore Construction Creates an Artificial Intelligence Nanopore for Coronavirus Detection. Nat. Commun., 12, 3726 (DOI: 10.1038/s41467-021-24001-2)

(51) Barati Farimani, A.; Heiranian, M.; Aluru, N. R. (2018), Identification of Amino Acids with Sensitive Nanoporous MoS2: Towards Machine Learning-Based Prediction. Npj 2D Mater. Appl., 2, 1–9 (DOI: 10.1038/s41699-018-0060-8)

(52) Jena, M. K.; Mittal, S.; Manna, S. S.; Pathak, B. (2023), Deciphering DNA Nucleotide Sequences and Their Rotation Dynamics with Interpretable Machine Learning Integrated C3N Nanopores. Nanoscale, 15, 18080–18092 (DOI: 10.1039/D3NR03771A)

(53) Jena, M. K.; Pathak, B. (2023), Development of an Artificially Intelligent Nanopore for High-Throughput DNA Sequencing with a Machine-Learning-Aided Quantum-Tunneling Approach. Nano Lett., 23, 2511–2521 (DOI: 10.1021/acs.nanolett.2c04062)

(54) Mittal, S.; Manna, S.; Jena, M. K.; Pathak, B. (2023), Artificial Intelligence Aided Recognition and Classification of DNA Nucleotides Using MoS2 Nanochannels. Digit. Discov., 2, 1589–1600 (DOI: 10.1039/D3DD00118K)

(55) Mittal, S.; Manna, S.; Pathak, B. (2022), Machine Learning Prediction of the Transmission Function for Protein Sequencing with Graphene Nanoslit. ACS Appl. Mater. Interfaces, 14, 51645–51655 (DOI: 10.1021/acsami.2c13405)

(56) Perdew, J. P.; Burke, K.; Ernzerhof, M. (1996), Generalized Gradient Approximation Made Simple. Phys. Rev. Lett., 77, 3865–3868 (DOI: 10.1103/PhysRevLett.77.3865)

(57) Kurth, S.; Marques, M. A. L.; Gross, E. K. U. (2005), Density-Functional Theory. In Encyclopedia of Condensed Matter Physics; Bassani, F., Liedl, G. L., Wyder, P., Eds.; Elsevier: Oxford, pp 395–402 (DOI: 10.1016/B0-12-369401-9/00445-9)

(58) Becke, A. D. (1988), Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. Phys. Rev. A, 38, 3098–3100 (DOI: 10.1103/PhysRevA.38.3098)

(59) Lee, C.; Yang, W.; Parr, R. G. (1988), Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. Phys. Rev. B, 37, 785–789 (DOI: 10.1103/PhysRevB.37.785)

(60) Stöhr, M.; Van Voorhis, T.; Tkatchenko, A. (2019), Theory and Practice of Modeling van Der Waals Interactions in Electronic-Structure Calculations. Chem. Soc. Rev., 48, 4118–4154 (DOI: 10.1039/C9CS00060G)

(61) Ordejón, P.; Artacho, E.; Soler, J. M. (1996), Self-Consistent Order-N Density-Functional Calculations for Very Large Systems. Phys. Rev. B, 53, R10441–R10444 (DOI: 10.1103/PhysRevB.53.R10441)

(62) Brandbyge, M.; Mozos, J.-L.; Ordejón, P.; Taylor, J.; Stokbro, K. (2002), Density-Functional Method for Nonequilibrium Electron Transport. Phys. Rev. B, 65, 165401 (DOI: 10.1103/PhysRevB.65.165401)

(63) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A. (2009), Gaussian, Gaussian 09, Revision A. 1. Wallingford CT Gaussian Inc.

(64) Rocha, A. R.; García-Suárez, V. M.; Bailey, S.; Lambert, C.; Ferrer, J.; Sanvito, S. (2006), Spin and Molecular Electronics in Atomically Generated Orbital Landscapes. Phys. Rev. B, 73, 085414 (DOI: 10.1103/PhysRevB.73.085414)

(65) Prasongkit, J.; Grigoriev, A.; Pathak, B.; Ahuja, R.; Scheicher, R. (2013), Theoretical Study of Electronic Transport through DNA Nucleotides in a Double-Functionalized Graphene Nanogap. J. Phys. Chem. C, 117, 15421–15428 (DOI: 10.1021/jp4048743)

(66) Pathak, B.; Löfås, H.; Prasongkit, J.; Grigoriev, A.; Ahuja, R.; Scheicher, R. (2011), Double-Functionalized Nanopore-Embedded

Gold Electrodes for Rapid DNA Sequencing. Appl. Phys. Lett., 100 (DOI: 10.1063/1.3673335)

(67) Mousavi, H.; Bamdad, M.; Jalilvand, S. (2022), Calculation of Electron Transport in Short Polyyne Nanochains. ECS J. Solid State Sci. Technol., 11. (DOI: 10.1149/2162-8777/ac8bfc)

(68) Landauer, R. (1957), Spatial Variation of Currents and Fields Due to Localized Scatterers in Metallic Conduction. IBM J. Res. Dev., 1, 223–231 (DOI: 10.1147/rd.13.0223)

(69) Büttiker, M. (1986), Four-Terminal Phase-Coherent Conductance. Phys. Rev. Lett., 57, 1761–1764 (DOI: 10.1103/PhysRevLett.57.1761)

(70) Ozaki, T.; Nishio, K.; Kino, H. (2010), Efficient Implementation of the Nonequilibrium Green Function Method for Electronic Transport Calculations. Phys. Rev. B, 81, 035116 (DOI: 10.1103/PhysRevB.81.035116)

(71) Rastkar, A.; Ghavami, B.; Jahanbin, J.; Afshari, S.; Yaghoobi, M. (2015), The Quantum Transport of Pyrene and Its Silicon-Doped Variant: A DFT-NEGF Approach. J. Comput. Electron., 14, 619–626 (DOI: 10.1007/s10825-015-0692-2)

(72) Dou, M.; Maier, F. C.; Fyta, M. (2019), The Influence of a Solvent on the Electronic Transport across Diamondoid-Functionalized Biosensing Electrodes. Nanoscale, 11, 14216–14225 (DOI: 10.1039/C9NR03235E)

(73) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. (2011), Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res., 12, 2825–2830.

(74) Guo, G.; Wang, H.; Bell, D.; Bi, Y. (2004), KNN Model-Based Approach in Classification

(75) Patel, H.; Prajapati, P. (2018), Study and Analysis of Decision Tree Based Classification Algorithms. Int. J. Comput. Sci. Eng., 6, 74–78 (DOI: 10.26438/ijcse/v6i10.7478)

(76) Chen, F.; Tao, N. J. (2009), Electron Transport in Single Molecules: From Benzene to Graphene. Acc. Chem. Res., 42, 429–438 (DOI: 10.1021/ar800199a)

(77) Furuhata, T.; Ohshiro, T.; Akimoto, G.; Ueki, R.; Taniguchi, M.; Sando, S. (2019), Highly Conductive Nucleotide Analogue Facilitates Base-Calling in Quantum-Tunneling-Based DNA Sequencing. ACS Nano, 13, 5028–5035 (DOI: 10.1021/acsnano.9b01250)

(78) Steiner, T. (2002), The Hydrogen Bond in the Solid State. Angew. Chem. Int. Ed., 41, 48–76 (DOI: 10.1002/1521-3773(20020104)41:1<48:: AID-ANIE48>3.0.CO;2-U)

(79) Grabowski, S. J. (2001), Ab Initio Calculations on Conventional and Unconventional Hydrogen BondsStudy of the Hydrogen Bond Strength. J. Phys. Chem. A, 105, 10739–10746 (DOI: 10.1021/jp011819h)

(80) Chen, T.; Guestrin, C. (2016), XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; pp 785–794 (DOI: 10.1145/2939672.2939785)

(81) Jena, M. K.; Roy, D.; Mittal, S.; Pathak, B. (2023), Artificially Intelligent Nanogap for Rapid DNA Sequencing: A Machine Learning Aided Quantum Tunneling Approach. ACS Mater. Lett., 5, 2488–2498 (DOI: 10.1021/acsmaterialslett.3c00475)

(82) Evgeniou, T.; Pontil, M. (2001), Support Vector Machines: Theory and Applications; Vol. 2049, p 257 (DOI:10.1007/3-540-44673-7_12)

(83) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. (2020), From Local Explanations to Global Understanding with Explainable AI for Trees. Nat. Mach. Intell., 2, 56–67 (DOI: 10.1038/s42256-019-0138-9)

(84) Souza, F. A. L. de; Sivaraman, G.; Fyta, M.; Scheicher, R. H.; Scopel, W. L.; Amorim, R. G. (2020), Electrically Sensing Hachimoji DNA Nucleotides through a Hybrid Graphene/h-BN

Nanopore. Nanoscale, 12, 18289–18295 (DOI: 10.1039/D0NR04363J)

(85) Souza, F. A. L. de; Amorim, R. G.; Scopel, W. L.; Scheicher, R. H. (2017), Electrical Detection of Nucleotides via Nanopores in a Hybrid Graphene/h-BN Sheet. Nanoscale, 9, 2207–2212 (DOI: 10.1039/C6NR07154F)