Developing an AI-Enhanced Protocol for Simulating Abiogenesis: Novel Nano-Structured Carbon Materials as a Templatizing Surface for Prebiotic Nucleic Acid Oligomer Formation

Anna Du

Phillips Academy

180 Main St, Andover, MA 01810

Table of Contents:

Abstract

This project explores novel strategies for investigating the origins of life and enhancing the field of molecular dynamics (MD) simulations. A novel concept is put forth that takes advantage of the inherent irregularities within carbonaceous crystal lattices such as diamond like carbon (DLC), as well as its superlative thermal diffusivity and conductivity along all planes due to its anisotropism, as the ideal templates for the abiotic formation of oligomers. Variability within the lattice structures plays a crucial role in facilitating the molecular interactions that paved the way for life to emerge. Increased entropy is a driving force that provides conditions necessary for oligomer formation. Moreover, I introduce a new artificial intelligence-driven methodology to inform optimal input parameters with the goal of generating key endpoints (such as forming a Van der Waals polymer-template attachment, translation/rotation leading to double strandedness, elongation and bridging, as well as detachment). As MD simulations are extremely computationally expensive due to large quantities of electromagnetic and force field interactions for trajectory calculations, Polychrom/OpenMM simulations were used as a proxy for biological/chemical interactions for training transformer models that predict optimal parameters resulting in collision events. Given the emergence of the 4D Nucleome project, which studies the 3D geometric properties of oligomers through time, I utilized time reversal with a transformer model to form more accurate predictions. These innovations furnish a robust framework for probing fundamental questions regarding the origins of life and have the potential to revolutionize synthesized RNA based drug discovery and delivery platforms.

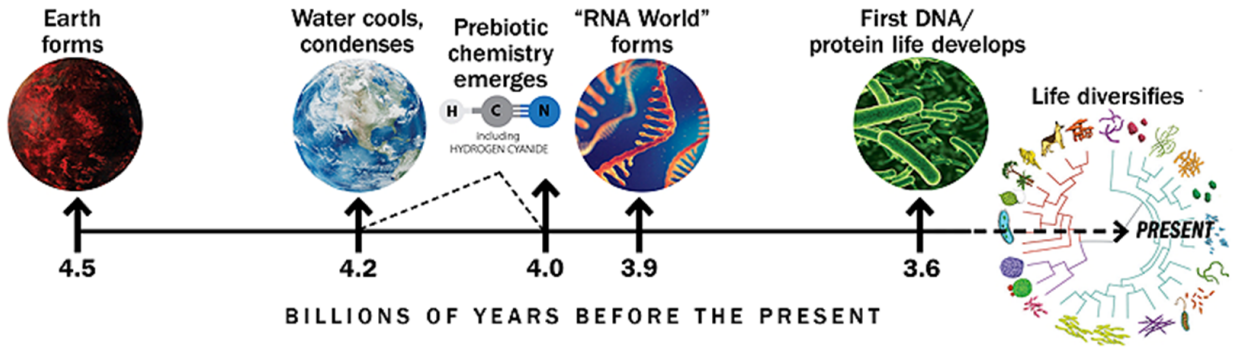
Background

“Where Do We Come From? What Are We? Where Are We Going?”

~ Paul Gauguin

Humanity, from all walks of life, whether artists or scientists, have wondered about this since the dawn of conscious thought. While scientists have studied this field for over half a century, significant progress has been made in the last few decades due to greater availability of analytical chemistry resources, historical geochemical knowledge, and data about the constitution of the universe itself.

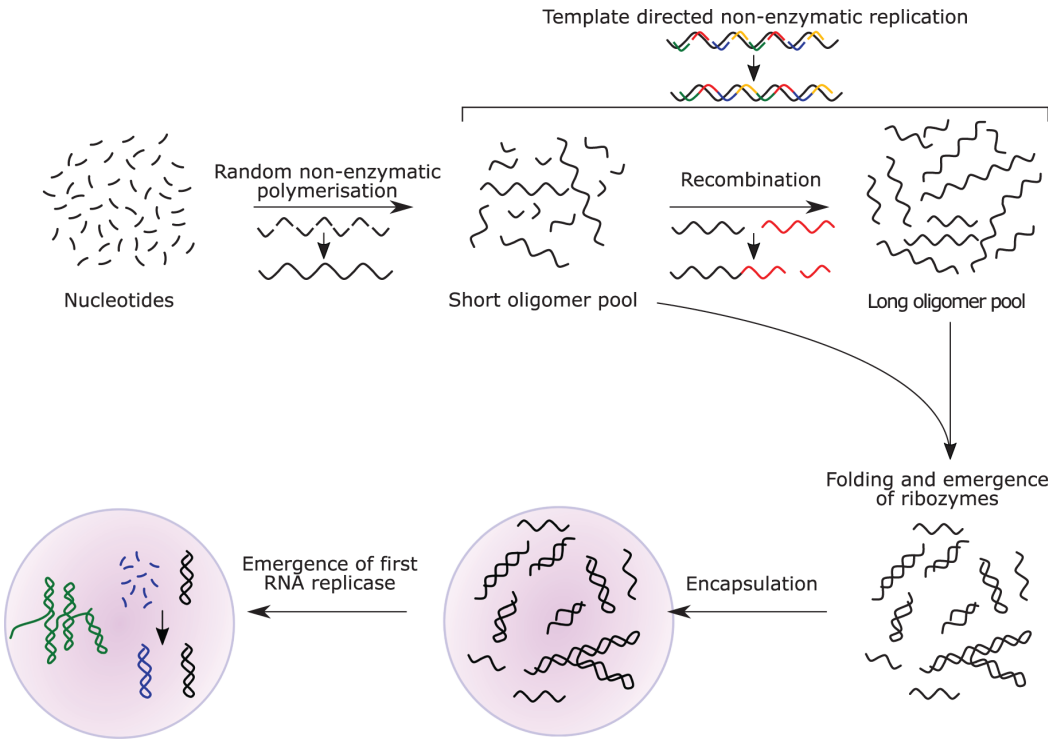
The question of life's origin has driven researchers to unravel the mysteries underlying the transition from non-life to life. This pursuit is not merely an intellectual endeavor but has profound implications for our understanding of the universe, our place within it, and the potential for extraterrestrial life. It also casts a light on how humans can deal with future problems, such as diseases or environmental disasters. While numerous theories and experimental approaches have been proposed, a comprehensive understanding of how life might emerge from simple chemical compounds remains elusive.



Visual representation of Earth's history across billions of years for the formation of life (Walker, 2017).

Among the emergent theories, is the concept of the RNA world, suggesting that RNA played a pivotal role in the transition from simple molecules to complex, self-replicating life forms. Efforts to simulate and recreate prebiotic conditions have led to significant breakthroughs. The Miller-Urey experiment of the 1950s demonstrated that amino acids could be synthesized from simple gases under conditions mimicking those of the early Earth (Miller & Urey, 1959). This experiment marked a foundational step towards understanding how the basic building blocks of life could have arisen from inorganic compounds.

This is further proven by recent discoveries of organic compounds from extraterrestrial sources. Uracil, one of the key nucleobases of RNA, was discovered on the surface of the near-Earth asteroid Ryugu. The presence of uracil on an extraterrestrial body holds profound implications for the plausibility of abiogenic nucleotide synthesis. The identification of uracil on an asteroid surface, far removed from Earth's biosphere, suggests that the formation of these vital molecules might have occurred through natural processes prevalent in space. Thus, organic compounds are proven to both have the potential to have formed on Earth, or arrived via extraterrestrial sources.

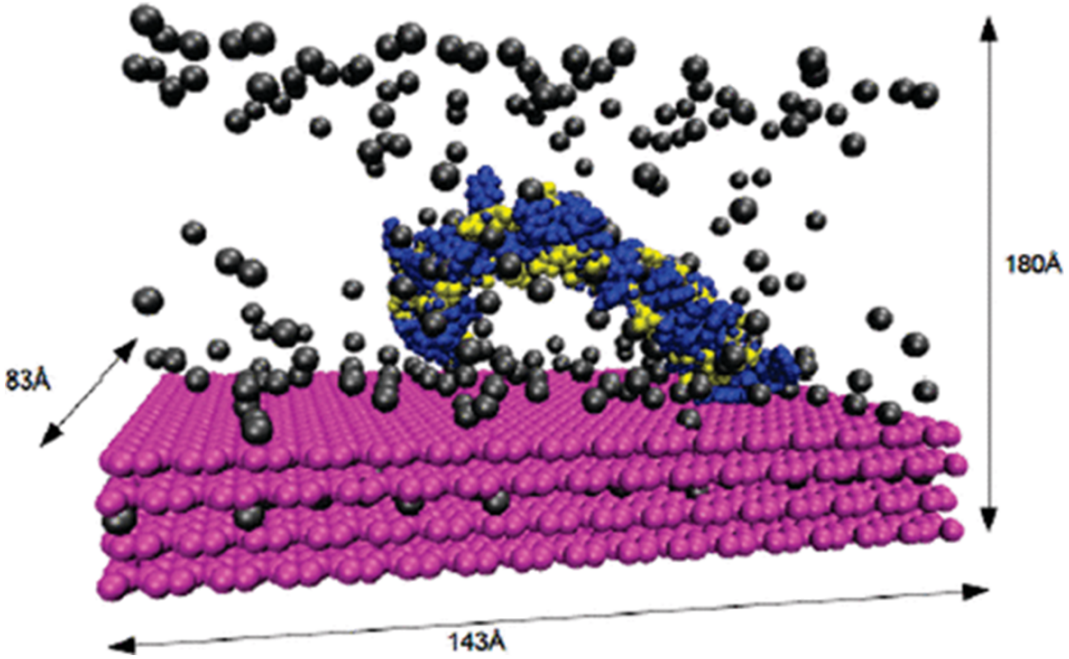


The current prevailing theory for abiogenesis is the RNA World Hypothesis. According to that hypothesis, RNA is a foundational biopolymer and of significant value for the ongoing transfer and evolution of genetic data (Le Vay, 2019).

Floating ionic catalysts have also garnered attention as potential contributors to prebiotic chemistry. Clays, with their layered structure and affinity for organic molecules, have been investigated for their role in concentrating and facilitating reactions that could lead to the formation of complex molecules. Similarly, in-solution catalysts and solid substrate templates have been explored as plausible mechanisms for organizing and aligning reactive molecules, enabling the creation of more intricate molecular structures. The role of calcium and other metals in prebiotic chemistry has been a subject of considerable investigation. These metal ions have demonstrated the ability to mediate the phosphorylation of nucleotides, a crucial step in the

synthesis of nucleic acids. In this context, the interplay between metal catalysts and the dynamic chemistry of early Earth environments has been an area of exploration, aiming to unravel the potential pathways that could have led to the formation of biomolecular precursors.

A significant body of research has also focused on the potential of solid crystalline catalysts, particularly clays, to support prebiotic reactions. It is commonly accepted that out of the 300+ clays that exist today, which likely existed towards the beginning of Earth, one of the most promising has been and continues to be montmorillonite. While on the one hand, this demonstrates a possible scenario for what we observe in today's living systems, the mechanism of attachment of long chain biopolymers to inorganic templatizing surfaces inhibits longer chains from forming.



Molecular dynamics visualization of RNA that has formed a Van der Waals attachment to the montmorillonite clay in a parallel manner, preventing further elongation.

There have been other confirmatory experiments, which have validated the idea that abiotic development of life is plausible, purely through a combination of entropy and random stochastic chemical interactions (Prigogine 1975, England, 2013). Recent advances in force-field based molecular dynamics (in silico) and economically plausible large scale studies of organic chemical formation (in vitro), have led to a number of newer theories proposed for the progression of life between the early inorganic formation of basic organic molecules and the last universal common ancestor (LUCA) during a period commonly referred to as the RNA World period in prevailing abiogenesis hypotheses. Current theories consider plausible prebiotic origins in hydrothermal vents, intertidal pools, and even extraterrestrial sources. The existence of abiogenetically formed informational molecule precursors outside of Earth are validated by the presence of confirmed extraterrestrial amino acids and nucleobases within a number of meteorites.

Hypothesis

My project presents a novel theory for abiogenesis, departing from the assumptions that homogenous materials like montmorillonite or diamonds were the only relatively abundant likely templates for organic polymer growth. I propose that heterogenous carbon lattice structures, like amorphous diamond / DLC would have been far more ideal, both structurally, and chemically, for the temporary, yet robust, attachment and detachment of organic compounds. Structures that contain predominantly sp3 hybridized and sp2 bonded carbon atoms could have served as

plausible binding and nucleation sites for stable biopolymer growth, leading to a localized, saturated condition where adjacent molecules/strands could interact, form helices and other more complex, functional 3D structures, and become incorporated as informational (genetic) molecules into early precellular/protocellular organisms. These seeding points allowed for the maximum high density growth of stable biopolymers, at lengths that are more biologically relevant (100's of mers versus a few dozen as has been typically observed in lab experiments thus far), as well as subsequent release into the local aqueous environment. This could have plausibly achieved a saturation point of required biopolymers to make the various subsequent abiogenesis theories possible, if not likely.

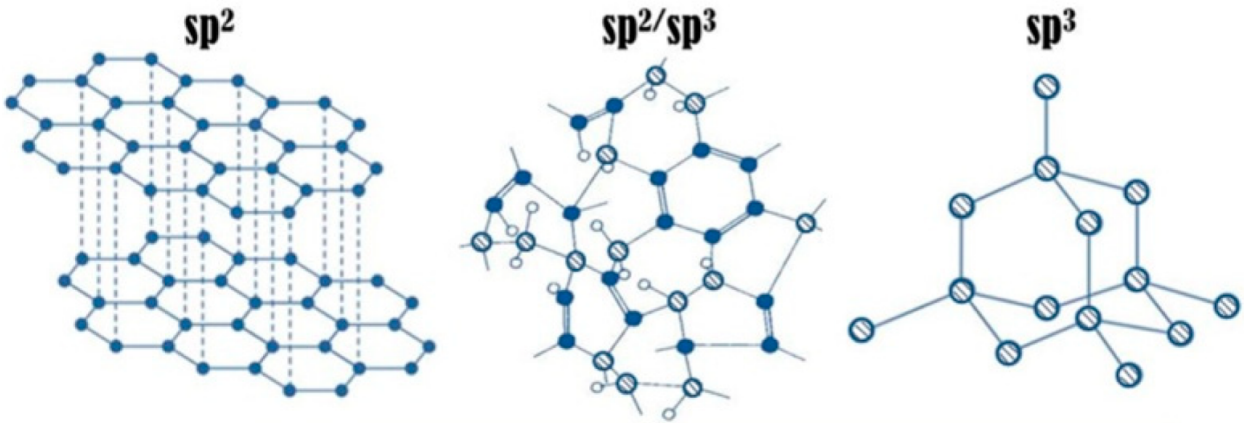
By incorporating insights from diverse fields such as materials science, catalysis, biology, and physics, this project aims to explore potential mechanisms that could have driven the early stages of chemical evolution. In particular, diamond-like carbon (DLC) is viewed as a potential catalyst for biopolymer growth, due to the implications of stochastic resonance, anisotropism, and its superlative thermal conductivity and diffusivity. Among the diverse array of materials investigated for their role in prebiotic chemistry, diamond-like carbon (DLC) emerges as a unique candidate with many benefits. DLC, composed of carbon atoms bonded in a mixture of sp2 and sp3 hybridization states, possesses a distinct nanocrystalline structure characterized by its variability and heterogeneity. This distinctive feature gives rise to a range of potential binding sites, making DLC an attractive template for seeding and catalyzing the growth of biopolymers.

DLC's presence and formation conditions are compelling factors that make it an appealing candidate for playing a role in prebiotic chemistry. The prevalence of methane in outer

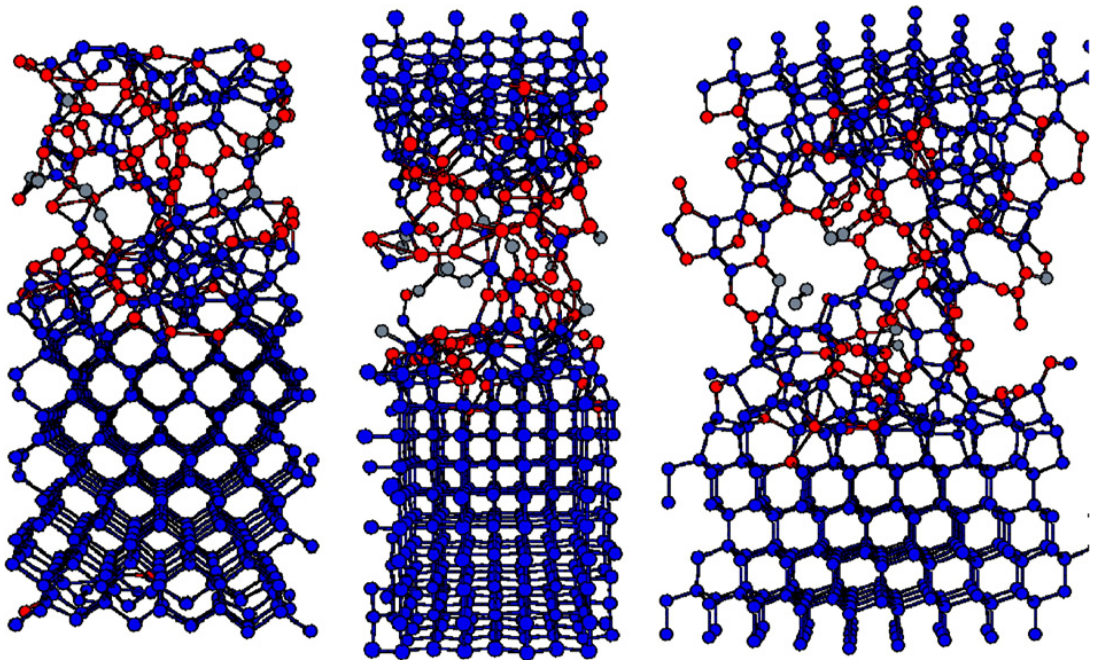
space and the high vacuum environment, coupled with radiation and energy sources, closely mimic the conditions that can lead to the formation of DLC. These are the same conditions used to produce synthetic DLC in modern science labs. There is also evidence that instantiated high pressure environments, which were plentiful within the early Earth crust, could have formed both amorphous diamond/DLC and fully sp3 hybridized diamond crystals, which were subsequently exposed on the surface through constant upheaval of tectonic/volcanic activity, which was prevalent on early Earth. Whether or not aluminosilicate clays/minerals were abundant, the likely presence of heterogeneous crystal lattice structures as well as mixtures of homogeneous carbonaceous materials would have been highly efficient template for seeding and biopolymerization of long chain informational molecules.

Synthetic production methods for DLC involve similar conditions, making the material versatile and applicable across various scientific domains, including prebiotic chemistry. It is reasonable to infer that DLC must have existed on a prebiotic Earth, possibly through kimberlite pipes or other mechanisms, but also potentially originating from extraterrestrial sources. Notably, telescopes including James Webb and Chandra have recently provided confirmatory evidence of methane and acetylene within distant-universe galaxies, both of which are precursor gases for DLC formation, in the vacuum of space. James Webb has also detected carbon methyl cations, which aids the formation of complex carbon molecules, in June 2023. Furthermore, the discovery of asteroids containing nanodiamonds on Earth provides additional evidence of the potential widespread distribution of diamond-like carbon in the cosmos, fostering intriguing connections between space chemistry and the emergence of life on our planet. Serpentinite xenoliths have been found in Sicily containing carbonaceous materials, demonstrating the presence of organic compounds during the Mesozoic era.

Unlike many crystalline structures that maintain relatively stable and repeating lattices, DLC's lattice structure can be modulated to have a range of sp2 / sp3 bonds based on tuned deposition conditions. This modulation creates a heterogeneous structure comprising an amalgamation of closely packed intermingled nanostructures, including diamond, graphene, single-walled carbon nanotubes, fullerene, and graphite. This inherent variability provides a wide spectrum of potential binding sites for biomolecules. The ability to control the ratio of sp2 and sp3 hybridization bonds allows for tunability in the material's properties, ranging from more graphitic to more diamond-like. This morphological extensibility extends to surface features, including convex and concave structures, which could accommodate the growth of biopolymers of varying lengths, enhancing the plausibility of polymerization on DLC templates. In light of its multifaceted advantages, DLC emerges as a compelling candidate for the templatizing material of prebiotic oligonucleotides.



A comparison of the molecular structures of graphite, DLC, and diamond. DLC is a heterogeneous mix of various carbon species, and thus contains a range of sp2 to sp3 bonds (Voytech, 2014).



Lattice structure of DLC / amorphous diamond (Kopidakis, 2007).

This representation shows a mismatch between two crystal lattices, and provides an opportunistic attachment point for the Van der Waals bonding of a nucleotide base / functional group for a covalent bond. In the case of this project, I'm primarily identifying potential VDW sites for variable polymer types, initially looking at RNA but also aggregation of various forms of RNA based structure and early polypeptides and proteins, as those aggregate structures are extremely crucial functional groups. This overall creates more functional biopolymers.

In prebiotic chemistry, diffusion enables the movement and interaction of molecules, fostering chemical reactions necessary for life's emergence. DLC's heterogeneity, with its varied binding sites and morphological features, aligns with the diverse diffusion paths molecules can take on its surface.

The field of molecular simulations has progressed rapidly in recent years. Among these, transformers—initially introduced for natural language processing tasks—have emerged as a compelling choice for analyzing complex molecular interactions. The sequence-to-sequence nature of transformers aligns with molecular simulations. One of the advantages to using transformers in MD is their capability to capture long-range dependencies. Furthermore, transformers have superior parallel processing capabilities. Unlike their sequential counterparts, such as recurrent neural networks (RNNs), transformers process sequences in parallel—an attribute conducive to leveraging distributed computing environments, such as GPUs or TPUs. The resultant shortening of simulation runtimes allows researchers to conduct larger and more comprehensive studies.

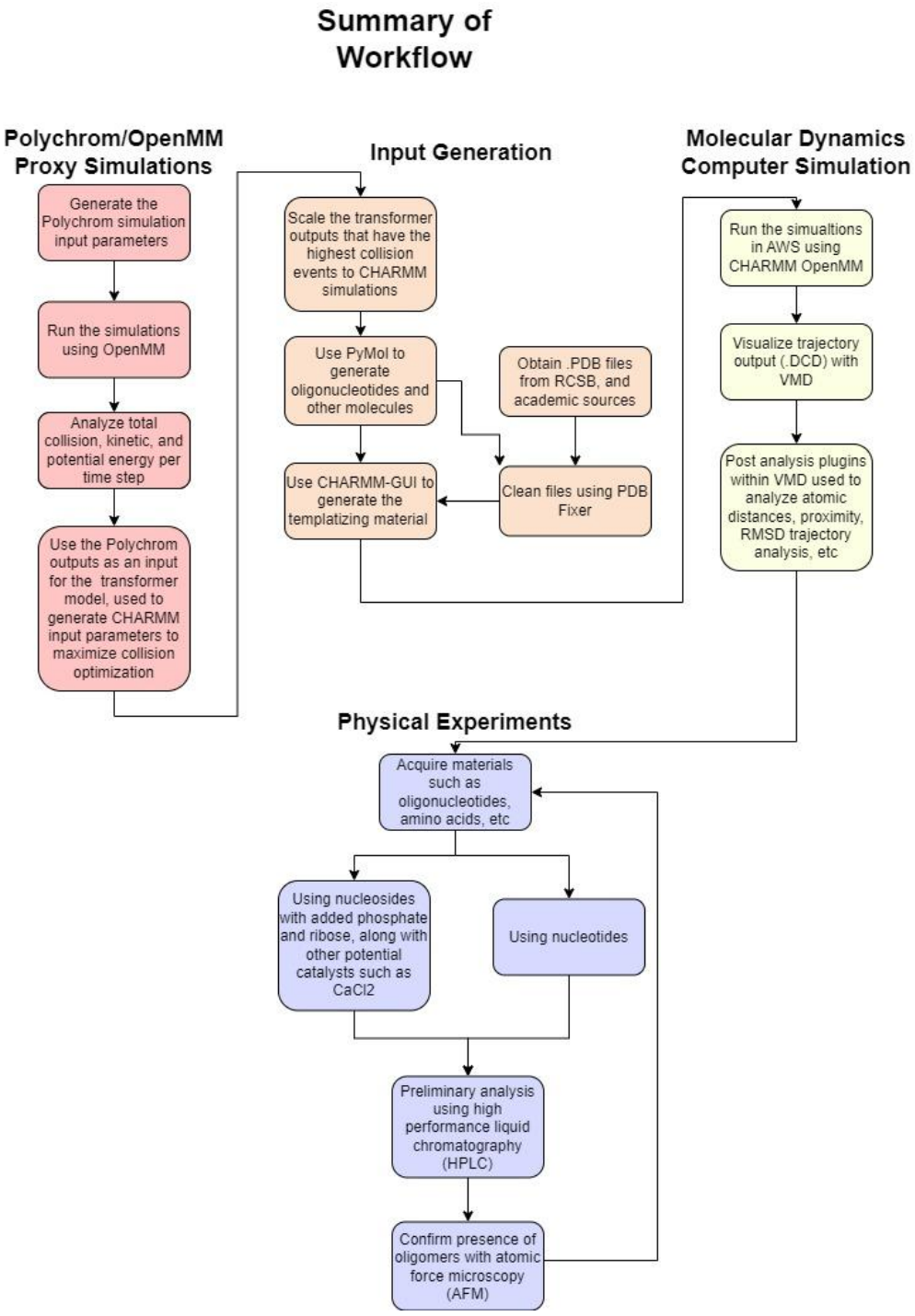
Flexibility is an attribute of transformers, that makes them advantageous for the inherent variability of molecular structures. Traditional methodologies have rigid input size requirements, a limitation that is circumvented by transformers' innate capacity to handle variable-length sequences. This adaptive ability aligns with the diverse nature of molecular simulations. Unlike traditional methods reliant on manual feature engineering, transformers' attention mechanism calculates the spatial relationships and trajectories of molecular structures. In the past few decades, researchers have used RNN for time sequence based data, and the 4D Nucleome project at MIT primarily uses CNN and time reversal, however, it is my belief that a transformer, is among the most versatile and efficient networks used today.

Methods

Polychrom/OpenMM Simulations

Achieving accurate results for molecular dynamics simulations necessitates calculating the locations of hundreds of thousands or more molecules. These interactions, which are dictated by force fields, electromagnetism, and other factors involving the molecules, inherently introduce complexity, which poses challenges to simulation speed, accuracy, and computational efficiency. This section delves into the concept of complexity in molecular dynamics simulations, proposes strategies to mitigate it, and outlines the pursuit of parameter optimization through AI-driven techniques to enhance simulation efficiency.

It is imperative to explore avenues for reducing complexity to expedite simulations. This can happen in two ways. The first, is through methods such as encoding, convolutions, or compression. The second, is in the form of parametric optimization. This strategy tailors simulation parameters to achieve a balance between accuracy and computational efficiency. AI-powered algorithms systematically explore parameter space, identifying configurations that are most likely to create a collision event (representative of a key endpoint) while preserving the accuracy of the results. The ramifications of enhanced simulation efficiency extend beyond speed. By mitigating complexity, simulations become more cost-effective and accessible, allowing for a larger volume of molecular dynamics runs.



The rationale driving the proposed workflow is deeply rooted in the pursuit of efficient and accurate molecular dynamics simulations. This workflow aims to reduce the overall volume of information while simultaneously enhancing the quality and interpretability of the data.

Solving the issue of complexity can be done in three ways: improving the quality of input data, refining encoding techniques, and parametric optimization. This project is an attempt at addressing the third. The initial step involves mass generating the simulations through Polychrom/OpenMM. Transformer models are trained based on this mass-generated data, to identify the parameters which will result in the highest amount of collision events. The proposed workflow builds upon my current work at the Mirny lab in the MIT PRIMES program.

The complexity of a transformer model is represented by the following equation, in which n = sample size, and d = depth of the model (Farsani, 2020).

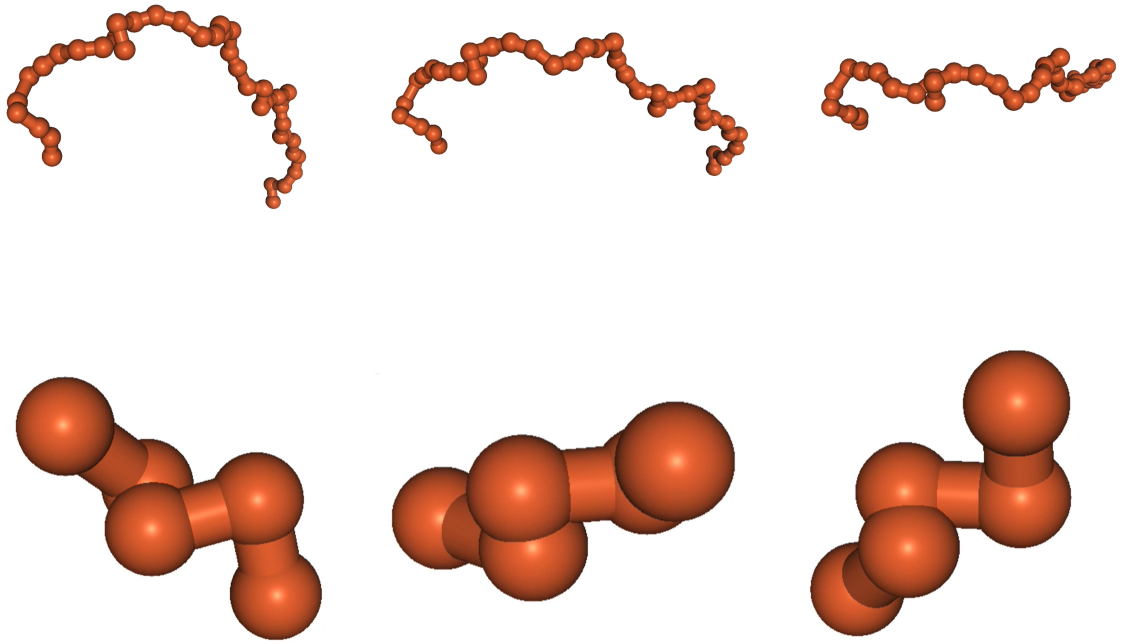
$$Complexity = O(n^2 \bullet d)$$

In this particular model, the input size is determined by the product of the number of conformations and the corresponding blocks. The complexity of sequential operations stands at O(1), indicating that the prediction subsequent to the training phase is notably expedited. Thus, predictions generated by the transformer are provided to CHARMM as initial configurations.

The computational complexity intrinsic to CHARMM is described as $O(n^2)$. However, it is imperative to note that the associated coefficient is substantially higher, often by several orders of magnitude. Consequently, when employed in tandem, the transformer and CHARMM framework is significantly more efficient than utilizing CHARMM in isolation.

I used the Polychrom/OpenMM simulation, a molecular dynamics modeling tool, to explore key parameters including polymer density, simulation waterbox temperature, polymer chain amount, and length of polymer. These parameters were systematically manipulated en masse to generate an extensive dataset, which would serve as the foundation for training our

transformer-based model for parameter optimization. Polychrom simulations, which use representative geometric structures such as beads and chains, to emulate biological structures such as nucleotides and phosphate or pentose groups, acted as a proxy for molecular simulations. This approach delivers an accelerated pathway to obtaining predictive outcomes, mitigating the resource-intensive and time-consuming nature of traditional biological experimentation involving MD.



3D representations of RNA were generated by using the 'ngutils' visualization tool, from data collected through the Polychrom simulations. The upper set of images demonstrates the bead and spring depiction of a polymer of 40 nucleotides, while the lower set shows a closer view of a smaller polymer of 6 nucleotides. In this visualization, each individual bead corresponds to a single nucleotide.

This simulation process involved configuring Polychrom to model polymer systems under varying conditions of polymer density, length, and temperature. By using these parameter settings as input settings, a multitude of simulated data points that emulated the behavior of polymer systems was generated. This synthetic dataset served as the training data for the transformer model, for subsequent parameter optimization. The adoption of Polychrom simulations enabled the exploration of a wide parameter space, enabling a more comprehensive investigation than physical or molecular dynamics based experimental approaches would allow. The rapid generation of simulated data allowed the transformer model to have the capacity to identify optimal parameter settings for enhanced polymer behavior. This approach presents a novel avenue for advancing polymer research, where data-driven insights from simulations can potentially revolutionize experimental design and optimization strategies.

Polychrom is a polymer physics engine that utilizes a coarse-grained bead-spring model to simulate the dynamics and interactions of polymer chains. In the Polychrom model, each monomer unit along a polymer chain is represented by a single spherical bead. Connections between beads are modeled using springs to approximate the physical forces and bond lengths between monomers. Polychrom utilizes Newtonian physics to simulate the motions and interactions of the polymer models over time. This fundamental principle dictates that the forces acting on particles can be derived from potential energy functions, leading to equations of motion that govern their trajectories through space. These forces can arise from both bonded interactions within the molecule's structure and non-bonded interactions with other nearby particles.

Polychrom integrates the equations of motion for the bead positions and momenta using an algorithm known as the Langevin integrator. This approach introduces friction and random noise terms to implicitly represent the effects of water collisions on the polymer beads, avoiding the need to explicitly model individual water molecules.

The Polychrom simulations are generated via the following steps:

1. SET simulation storage parameters:
   a. Define the path to the folder where simulations are to be stored.
   b. Create the folder if it doesn't exist.
2. SET bead and other necessary parameters for simulation:
   a. Set block size.
   b. Set the number of blocks to be simulated.
   c. Set polymer length.
   d. Set the max distance between beads.
   e. Set the number of polymer chains.
   f. Calculate the box length based on the polymer length, number of chains, and desired density.
3. GENERATE the initial conformation of the polymer:
   a. Create a compact initial structure.
4. SETUP reporters to save the simulation data:
   a. For each initial configuration, create an HDF5 reporter.
   b. Store the reporters in a list.
5. INITIATE the simulation:
   a. Set the platform, integrator, error tolerance, temperature, GPU number, and other necessary parameters.
   b. Initialize the Simulation object with the set parameters and reporters.
6. ADD forces to make the beads behave like a polymer:
   a. Define bond forces, angle forces, and non-bonded forces.
7. FOR EACH polymer chain in the initial configurations:
   a. Set the initial configuration for the simulation.
   b. Perform an energy minimization to prevent explosions at the start.
   c. FOR the set number of blocks:
      i. Execute the simulation block.
      ii. Get the simulation data.
      iii. Set the modified data for the next block.

Polychrom's ability to modulate the density of oligonucleotides, the length of oligo chains, and the temperature of the system was instrumental in these investigations. Polychrom-generated data, including collision information, kinetic and potential energy, temperature, polymer length, and density were processed using transformer models. The attention mechanisms within transformers make it suitable for the complexity of time reversal. Various parameter changes over time emphasize the significance of temporal dynamics. An approach that combines forward and backward simulations can enhance the credibility of simulation outcomes.

		Transformer Output Collision Probabilities	
		Predicted Collision Events >547	Predicted Collision Events ≤547
Polychrom Simulation Generated Collision Count	Collision Events >547	274	12
	Collision Events ≤547	85	109

This confusion matrix compares the collision count generated by the Polychrom simulations to those predicted by the transformer based on the input parameters that were given.

Higher amounts of nucleotides and polymer chains in the Polychrom simulation than would be modeled with molecular dynamics, and because these collisions include contacts between a single polymer. Thus, a threshold was set at 547 collision events, due to the diminishing returns. A simulation that had over 547 collision events was considered to be a simulation likely to generate a key endpoint when using molecular dynamics simulations.

Input Generation

The next phase involved converting the Polychrom parameters into a format compatible with CHARMM GUI. This required attention to several specific details: Polychrom operates using an internal unit based on the length of a single nucleotide, and a single voxel unit corresponds to the volume of a nucleotide if condensed into a cube. This had to be accounted for in the conversion process. Polychrom also assumes that the nucleotides are fully submerged within the waterbox, and that the nucleotides share the same density as the water in the waterbox.

The computational expense of CHARMM necessitated a scaling down of the Polychrom parameters. To maintain feasibility while ensuring meaningful output, the polymer chain number and length were limited proportionally. This was due to the hardware constraints of running CHARMM, given its extensive time and resource requirements. The challenge of translating volumetric representations of beads obtained from Polychrom simulations to real nucleotides was addressed by transposing the simulation's volumetric data into real-world parameters, accounting for Van der Waals volume of nucleotides. This transformation enabled me to bridge the gap between simulation and real-world scenarios, facilitating more accurate predictions.

The transformer models yielded a series of probabilities corresponding to each parameter, facilitating the identification of the most collision-inducing conditions. The density was derived from the formula below, with d = density coefficient based on percentage of volume of the waterbox taken up by the nucleotides, p = polymer length, s = distance between polymer chains in nucleotide units:

$$density = d * (\frac{p}{p+s})$$

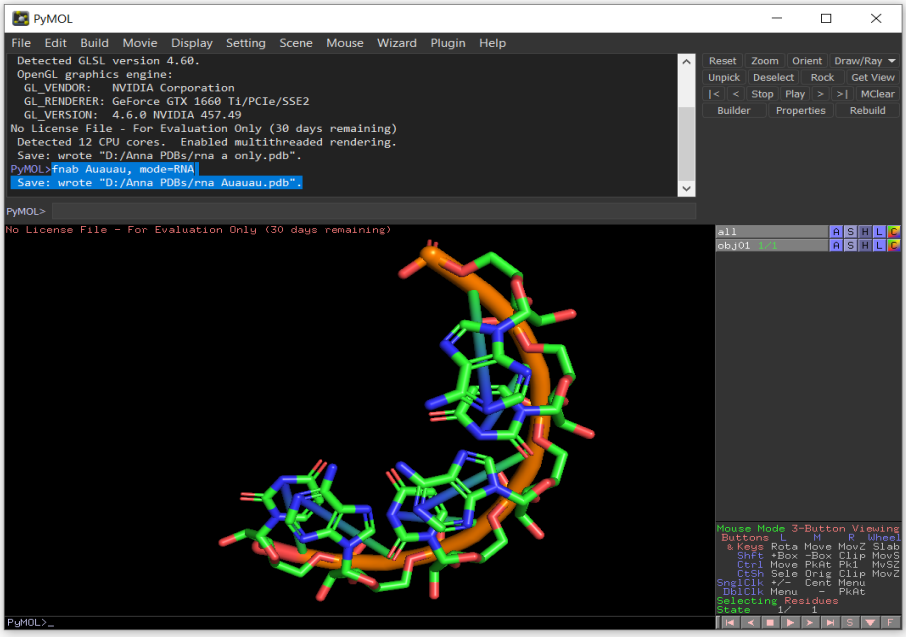
The box length was calculated using the formula, with n = amount of polymer chains present, d = density coefficient based on percentage of volume of the waterbox taken up by the nucleotides, p = polymer length, s = distance between polymer chains in nucleotide units:

$$BoxLength = \frac{n(p+s)}{d}^{\frac{1}{3}}$$

The scaling was based on density and volume. The volume was derived from the Van der Waals volume of nucleotides and carbonaceous materials, using the known Van der Waals volumes of Cytosine (18.97 Å), Guanine (23.77 Å), Adenine (22.25 Å), and Uracil (17.74 Å). By applying the previous formulas to the optimal parameters, a box length that was calculated based on the internal Polychrom scale could be translated into angstroms, the unit used in CHARMM.

PyMol, an open source molecular visualization tool, was utilized to generate the .PDB (protein data bank) files of the oligonucleotides based on the scaled down parameters from the transformer output. The resultant .PDB files categorized each individual atom as a ‘HETATM’

(heterogeneous atom), and furthermore, PDB manipulation software packages tended to alter the formats and atom definitions, causing simulations to fail. To address these compatibility issues and ensure accurate representation, the PDBFixer application was employed. PDBFixer redesignates the molecules as separate RNA strands, thereby standardizing the positions of all atoms. Furthermore, other necessary conditions were normalized, rendering the resultant .PDB file fully compatible with the input generators in CHARMM-GUI.

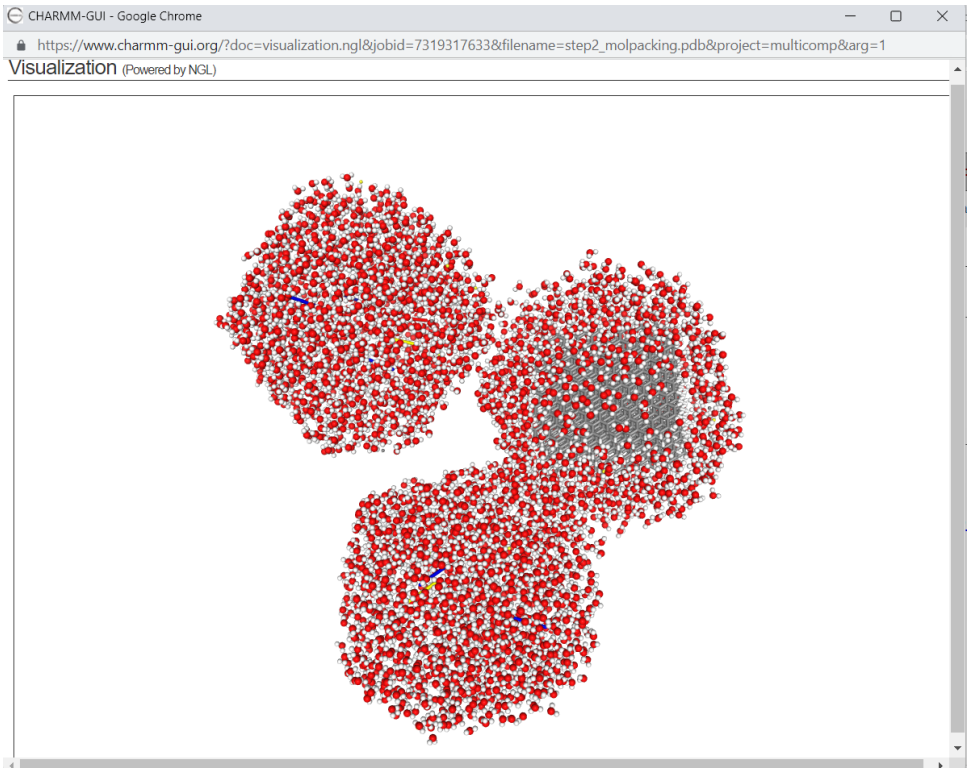


Using PyMol to generate oligonucleotide strands with specific sequences to use in CHARMM simulations

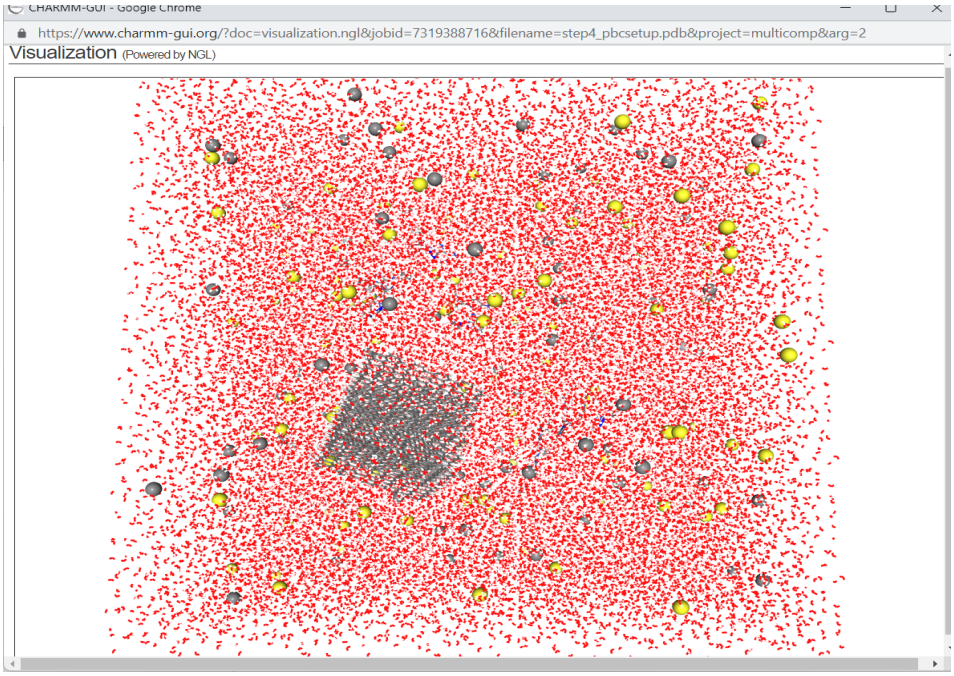
The generated and refined .PDB files were then fed into CHARMM-GUI, a versatile suite of tools for molecular dynamics simulations. Specifically, the Solution Builder module within CHARMM-GUI was used. This step involved the creation of a waterbox surrounding the

oligonucleotides, effectively providing the aqueous environment for simulations. For simulations that tested the parameters derived from the transformer output, the setup involving the waterbox around the oligonucleotides was used. However, for simulations that incorporated carbonaceous templates, additional steps were used.

The outputs from PDBFixer were used as inputs in CHARMM-GUI's Nanomaterial Modeler. This facilitated the generation of files for different carbonaceous species such as graphene, carbon nanotubes (CNT), and diamond. Subsequently, the Multicomponent Assembler module was utilized to combine these various components along with the refined oligonucleotide .PDB files. This process yielded a composite system encompassing all components. Furthermore, the system was ionized using sodium chloride (NaCl) to simulate an ionic environment.



A preview of a simulation with a sample of graphite and two oligonucleotide strands on CHARMM-GUI prior to the encasement of a waterbox



After the water box is constructed around the graphite, the two oligonucleotides, and 0.15 M NaCl ions are scattered using the Monte-Carlo method.

CHARMM Simulations

CHARMM (Chemistry at HARvard Molecular Mechanics) is an open source software to run molecular dynamics. CHARMM utilizes sophisticated force fields, which encode the potential energy surfaces governing molecular interactions. These force fields include parameters

that define bond lengths, angles, dihedral angles, non-bonded interactions, and more. They are created based on experimental data and quantum mechanical calculations, effectively translating complex molecular interactions into computationally tractable models.

The resulting files were input into CHARMM C47B1 OpenMM 7.6 to create trajectory files (.DCD). The following parameters were used:

- Fortran 90
- CUDA 9.2
- CMake 2.8.12.2
- C++ 7.3.1

This was done on AWS cloud-based on demand processing computing services, as well as a small cluster of 2 test laptops with NVIDIA GPU and Ubuntu Linux.

1. Large Server n=1
    - 64 GB GPU
    - 16 core CPU, 64 RAM
2. Small Server n=3
    - 16 GB GPU
    - Quad core CPU, 16 RAM

Optimal Configuration Testing			
GPU Type	GPU	Swap for CPU Mem	Average time to finish 1 iteration of production step (hours)
None	0	Yes	> 50
NVIDIA T4	1	Yes	49
NVIDIA T4	4	Yes	23.2
NVIDIA T4	24	Yes	21.3
NVIDIA T4	1	No	13.8
NVIDIA T4	4	No	8.1
NVIDIA T4	24	No	7.8

The resulting .DCD trajectory files were then visualized and analyzed using Visual Molecular Dynamics (VMD). Normalizing simulation models were first run to gain a preliminary understanding of how long it would take for a single simulation to run, given a certain amount of parameters. This section lays the groundwork for the subsequent analysis and results, providing a comprehensive understanding of the methodologies employed to address complexity in molecular dynamics simulations.

Physical Experimentation

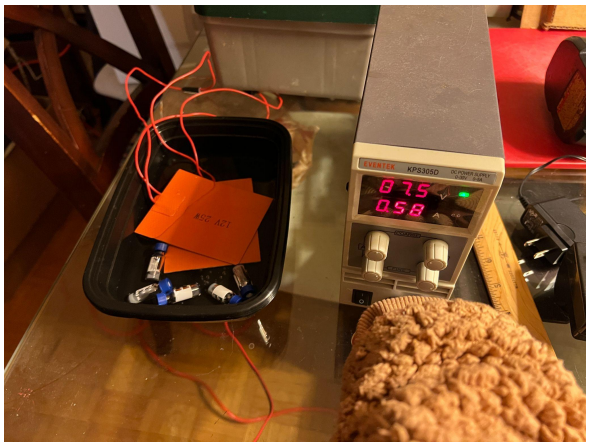
The physical experiments were designed with the purpose of verifying if DLC would serve as an ideal templatizing surface, and would facilitate the formation of hyper-saturated oligonucleotide solutions. The majority of the research has been conducted under the assumption that these reactions are taking place in an aqueous environment, perhaps in a saline solution. However, experimental reactions observed have indicated a need for some form of catalysis to justify the plausibility of arguments that the precursor organics would achieve a saturation point mathematically sufficient to justify a possible association and self-assemblage of such biopolymeric molecules by sheer chance and entropy. Catalysts such as Ca^{2+} ions have been proposed, and certainly, these have been demonstrated to assist in the phosphorylation of nucleobases, as well as facilitate the continued growth of nucleic acid chains, as well as peptides. In recent decades, scientists have further expanded this theory to include the possibility of sheet-like material templatizing surfaces as a means of seeding the growth of stable, longer biopolymers. It has been suggested by numerous researchers that the co-occurence of a reasonable saturation of relatively short chain biopolymers of different types (proteinaceous and informational) as a prerequisite for most abiogenesis theories, yet the greatest enzymatic and informational value of these biopolymers exists only when the polymerization process is stable, reliable, repeatable, and yields a variety of polymer lengths, particularly longer chains.

Due to equipment availability constraints it was not possible to emulate the high pressures of a hydrothermal vent environment. However, several key steps were taken to emulate the conditions of a deeper, warm intertidal pool (greater than atmospheric pressures, with geothermal heat influence, yet not as high as the temperatures or pressures of a hydrothermal vent). Experiments placed in the hydrothermal chamber were exposed to temperatures of roughly 20 - 90 C, controlled by resistive DC-powered 12V silicon heating pads and redundant K-type thermocouples. Nucleosides (cytosine, guanine, adenine, and uracil) were obtained from ThermoFisher were used, along with phosphate and ribose were utilized to best replicate early Earth conditions, with the goal of forming nucleotides/oligonucleotides from primary source inorganic chemicals. In some experiments, catalyzing ions such as calcium were also included, in the form of calcium chloride.

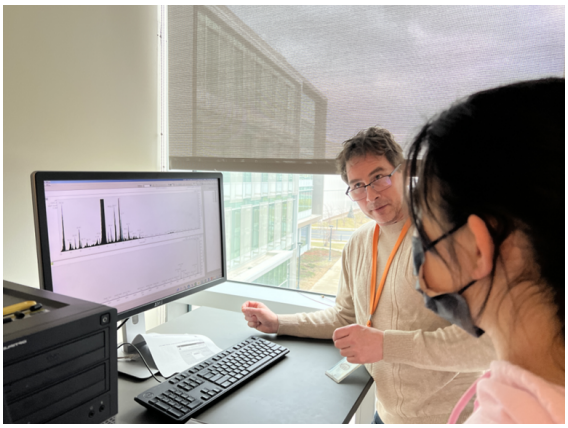
Custom made "Mini Hydrothermal Vent".

Furthermore, a regular rocking motion was applied every twelve hours, at a rate of 120 cycles / hour, based on the wave period in the middle of a large pond. Based on my assumption, abiogenesis could not have occurred on the surface of any body of water due to increased radioactivity. Therefore, it can be assumed that the polymerization process was likely to have occurred below the significant influence of UV radiation. Samples were periodically rotated to a specific slope pre-calculated based on the geometry of the sample chamber, which allowed for cycles of settling and resuspension of nanoparticles.

To emulate the conditions of an isolated aqueous environment at relatively low pressure, such as an intertidal pool (or a “warm little pond/mini intertidal pool”), samples were placed in waterproof 2 mL borosilicate containers with water, a microcontroller-automated variable temperature flux between 40 and 60 C, modulated by the resistive heating pads and a variable power supply to modulate the wattage the samples are exposed to, thus managing the temperature. With the release of some vapor in a sterile chamber, evaporite formed on the periphery of the sample container.



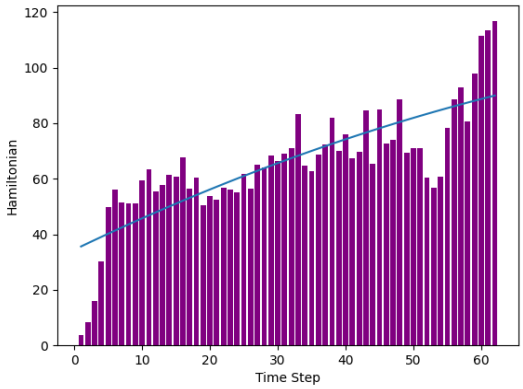
Homemade "Mini Intertidal Pool".

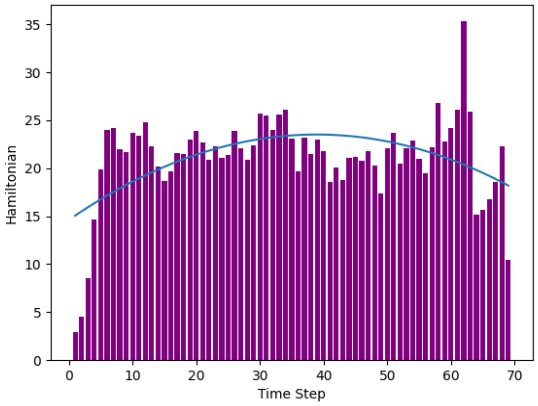


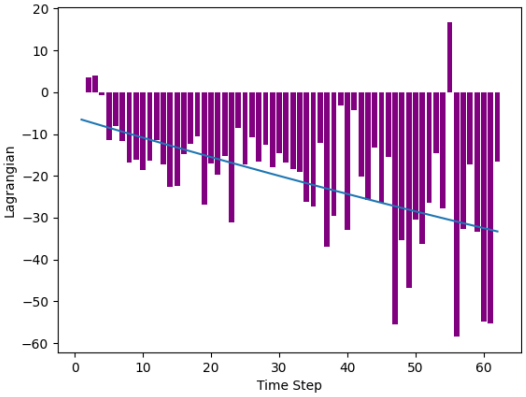
Further data analysis was conducted at UMASS Boston, with the assistance of Professor Evans, using a LC-MS/MS.

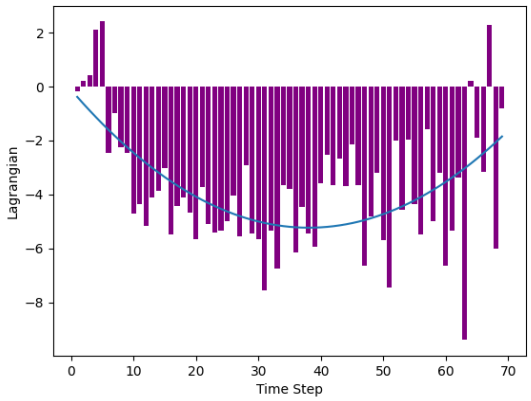
Results

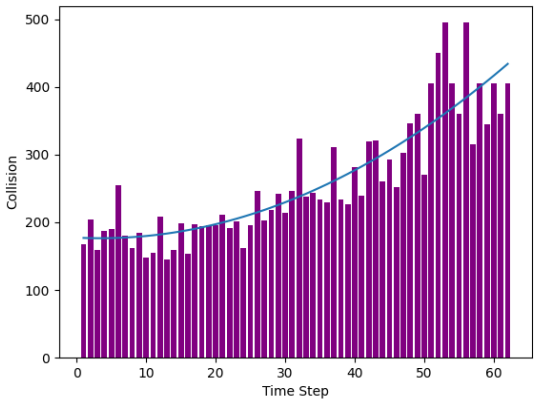
The graphs of the simulation data provided valuable insights into molecular dynamics results. The Hamiltonian and Lagrangian representations of energy showed spikes that correlated with increased collision counts in the early time steps. This alignment indicates that rises in energy levels under the optimized parameters were associated with more frequent molecular collisions.

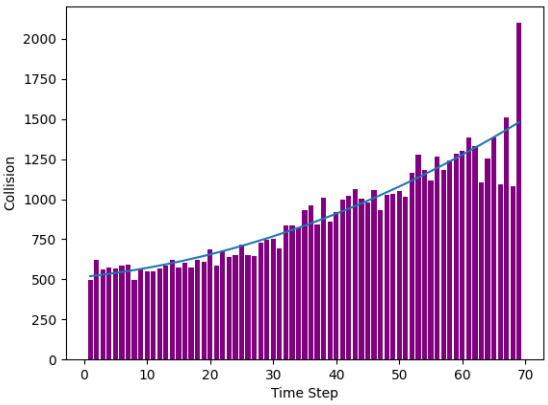

Hamiltonian
Time Step


Hamiltonian
Time Step


Lagrangian
Time Step

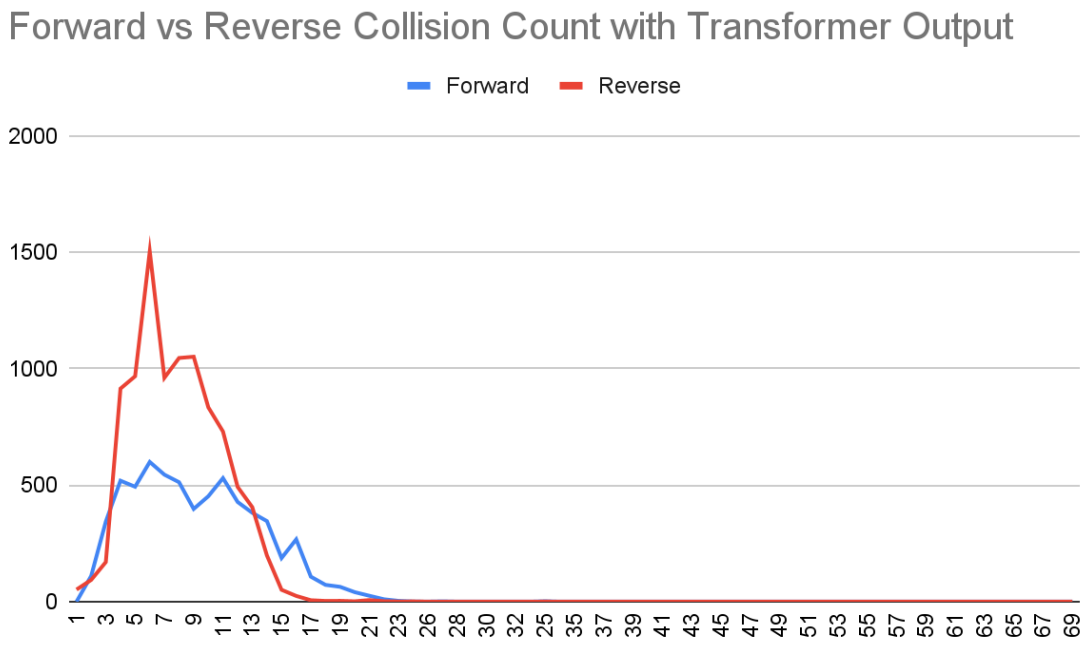

Lagrangian
Time Step

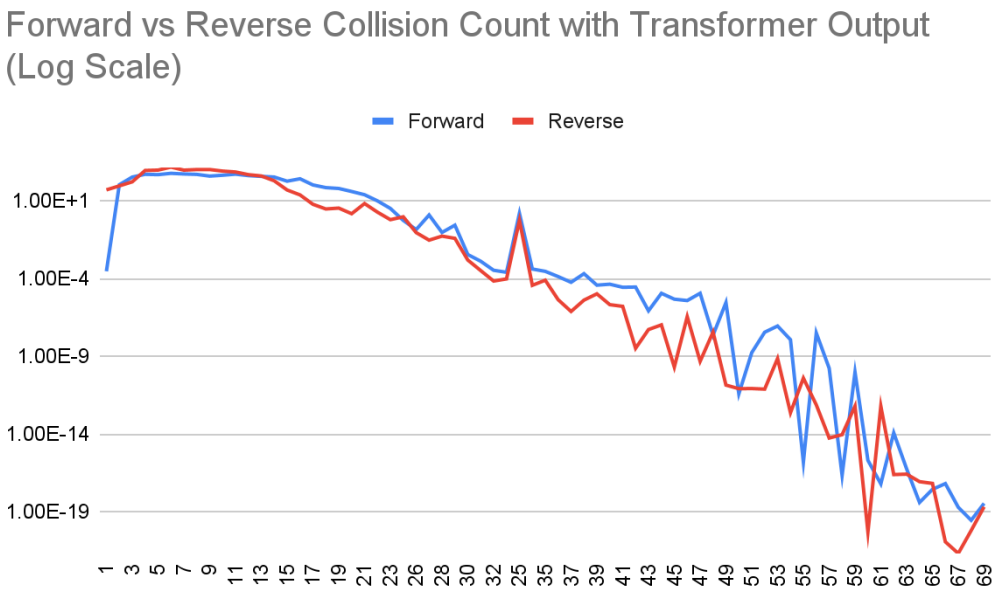

Collision
Time Step

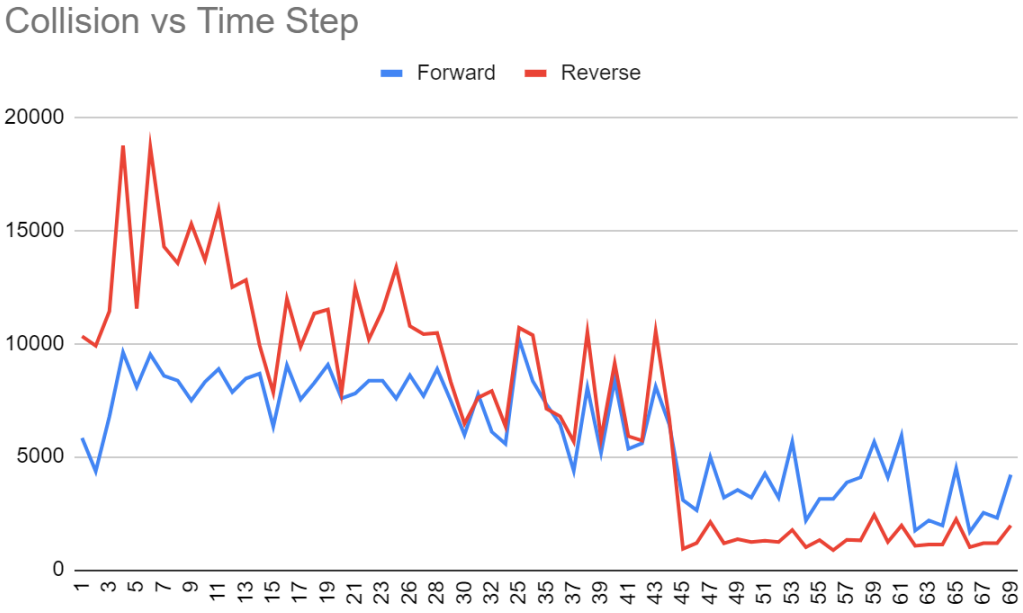

Collision
Time Step

For all simulations that take place at 304K, Top Left) the Hamiltonian representation (KE + PE), Middle Left) Lagrangian representation (KE - PE), and Bottom Left) the collision count vs time step for all simulations.

For simulations that take place ato 300K, Top Right) the collision count vs time step for all simulations, Middle Right) the Hamiltonian representation (KE + PE), and Bottom Right) Lagrangian representation (KE - PE)




Forward vs Reverse Collision Count with Transformer Output (Log Scale)
Forward
Reverse
1.00E+1
1.00E-4
1.00E-9
1.00E-14
1.00E-19
1 3 5 7 9 11 13 15 17 19 21 23 26 28 30 32 25 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69


Collision vs Time Step
Forward
Reverse
20000
15000
10000
5000
0
1 3 5 7 9 11 13 15 17 19 21 23 26 28 30 32 25 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69

While the majority of collisions took place within the first twenty time steps, the graphs demonstrated a secondary increase in collisions toward the end of the simulations. As the polymers continue to interact, their configurations and bindings gradually shift in ways that promote renewed interactions. The transient dissociations and reconfigurations of the molecules would allow for fresh collision potentials.

t-Test: Paired Two Sample for Means

	Forward	*Reverse*
Mean	167.7654	142.1219
Variance	52546.1	51314.64
Observations	480	480
Pearson Correlation	0.029505	
Hypothesized Mean Difference	0	
df	479	
t Stat	1.769599	
P(T<=t) one-tail	0.038715	
t Critical one-tail	1.648041	
P(T<=t) two-tail	0.07743	
t Critical two-tail	1.964929	

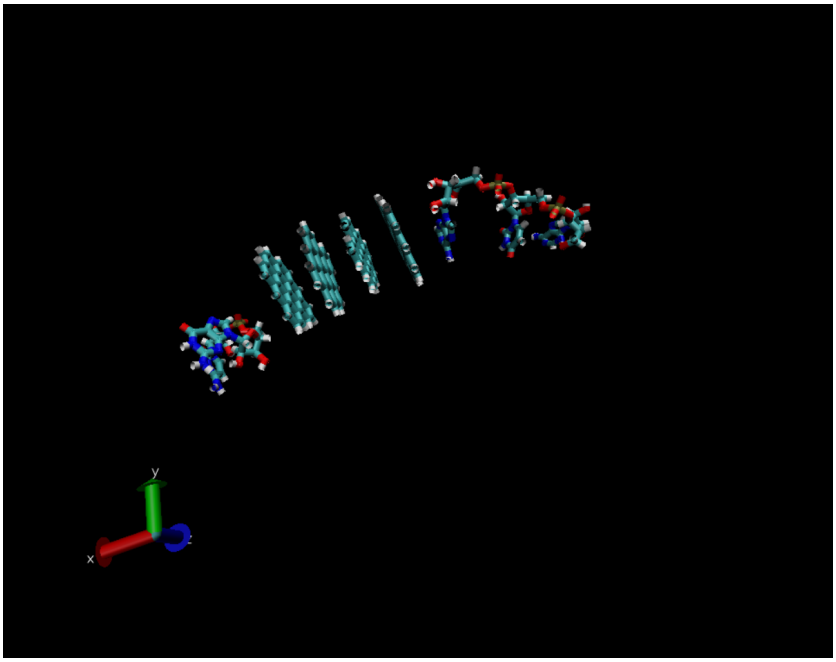
The statistical analysis of the forward versus reverse collision data yielded an extremely significant p-value of 0.039. This p-value indicates a negligible probability that the observed difference in collisions between the forward and reverse simulations could have occurred by chance alone, and validates that the identified parameters specifically caused the increase in collision counts. The transformer model yielded crucial insights into the parameters that allow for maximal molecular collisions. The model's identification of temperature, chain lengths, and density combinations highlights the interdependence of these variables in influencing interaction potentials. At lower temperatures, thermal energy would be insufficient to drive collisions.

Shorter chain lengths would reduce interaction cross sections, and excessive density could hinder polymer motion.

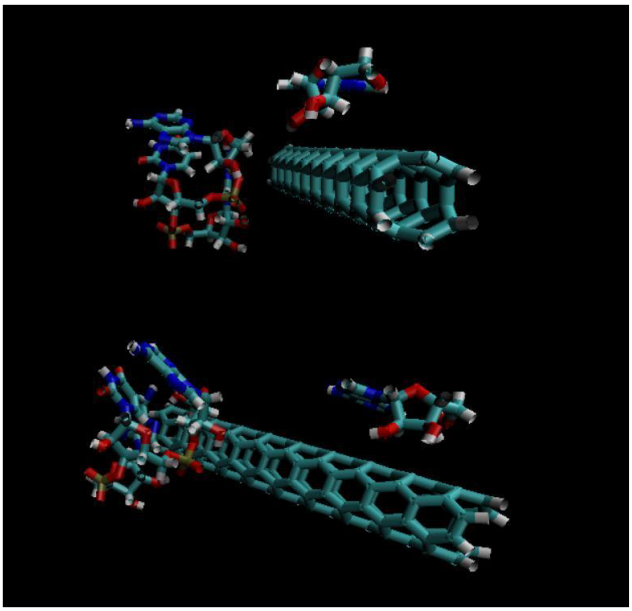
The transformer results align with molecular motion and reaction kinetics. Higher temperatures, to a degree, provide greater kinetic energy to overcome activation barriers (so long as it remains within ranges viable for life to sustain). Longer chains increase reaction probabilities via greater surface contact. But densities must be optimized to allow chain movements. The model encapsulated these dynamics within its parameterized framework. By uncovering the optimal conditions, this demonstrates transformers models' utility in scientific applications involving multifaceted time-series data. This application paves the way for advanced simulations that leverage AI to uncover key insights and guide experimental designs.

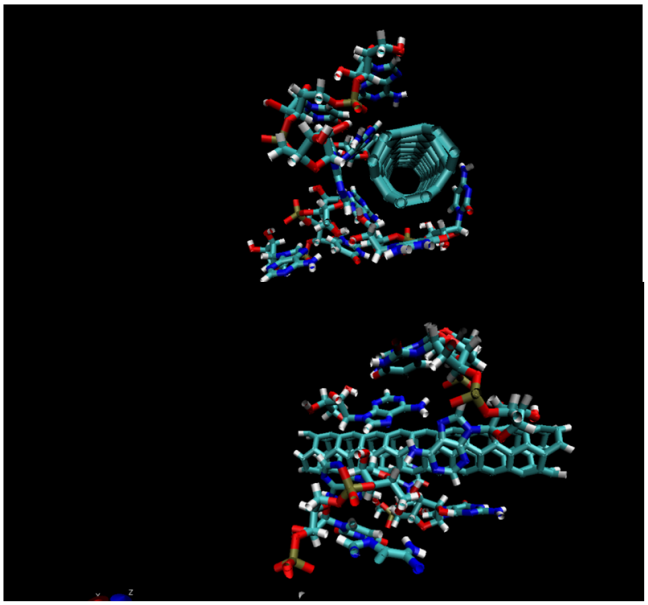
From the extensive CHARMM simulations conducted, a myriad of significant key endpoints were successfully generated, unveiling intriguing molecular interactions and behaviors. Among the notable outcomes were:

- Oligomer Bridging via Carbon Sheet Structures: The simulations revealed instances where oligomers exhibited bridging behavior, connecting on both sides of the carbonaceous templatizing material.

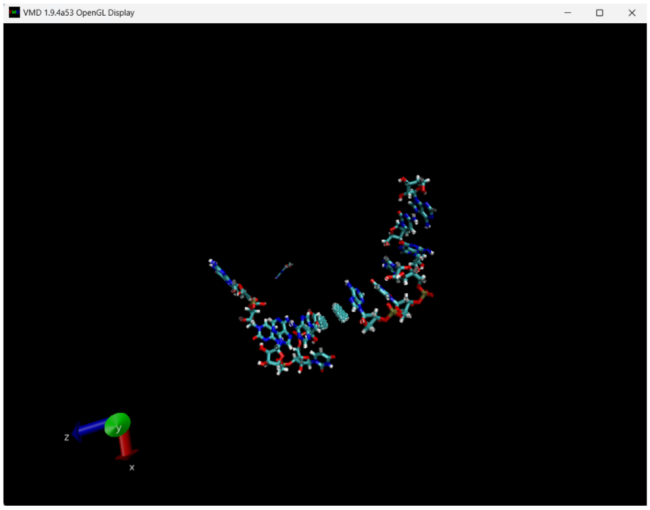


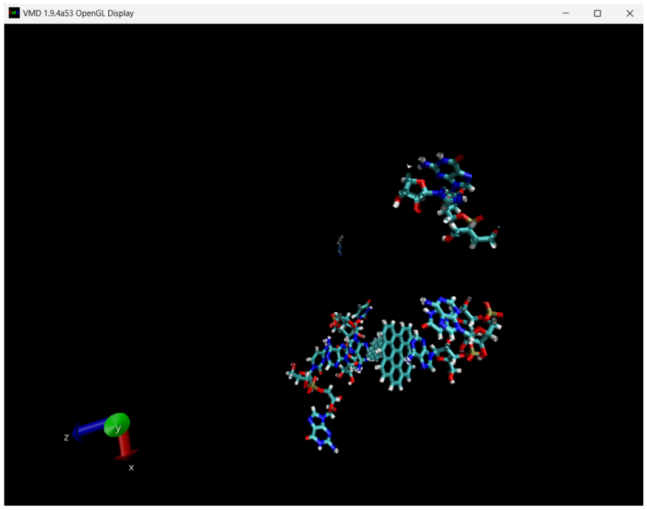
- Oligomer Attachment to Carbon Nanotube: The simulations captured the attachment of oligonucleotides to the surface of carbon nanotubes. The oligonucleotides form in a parallel manner to the surface of the CNT, thus showing that, like montmorillonite, CNT on its own, with a large, flat surface, is not an ideal templatizing material to facilitate polymer growth.



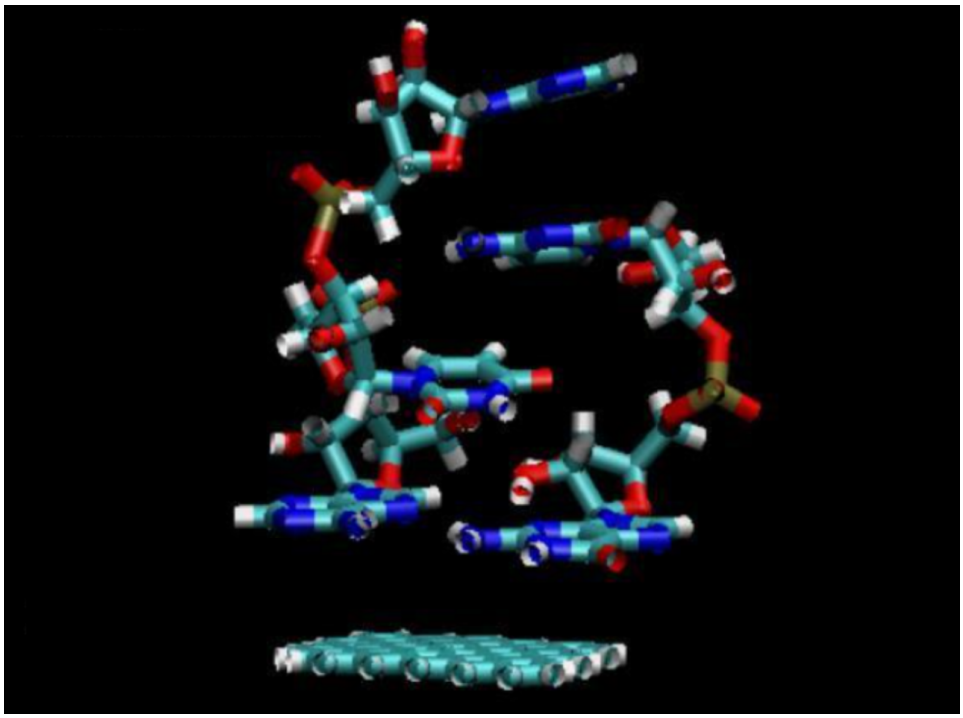


- Detachment of Dimer from Oligonucleotide Complex: Notably, a dimer dissociated from a complex involving a 5-mer oligonucleotide. This detachment event occurred due to an increase of temperature in the CHARMM simulations, emulating temperature cycles in the natural environment. This demonstrates that a hyper-saturated solution of oligonucleotides could have formed in an environment such as a warm little pond.

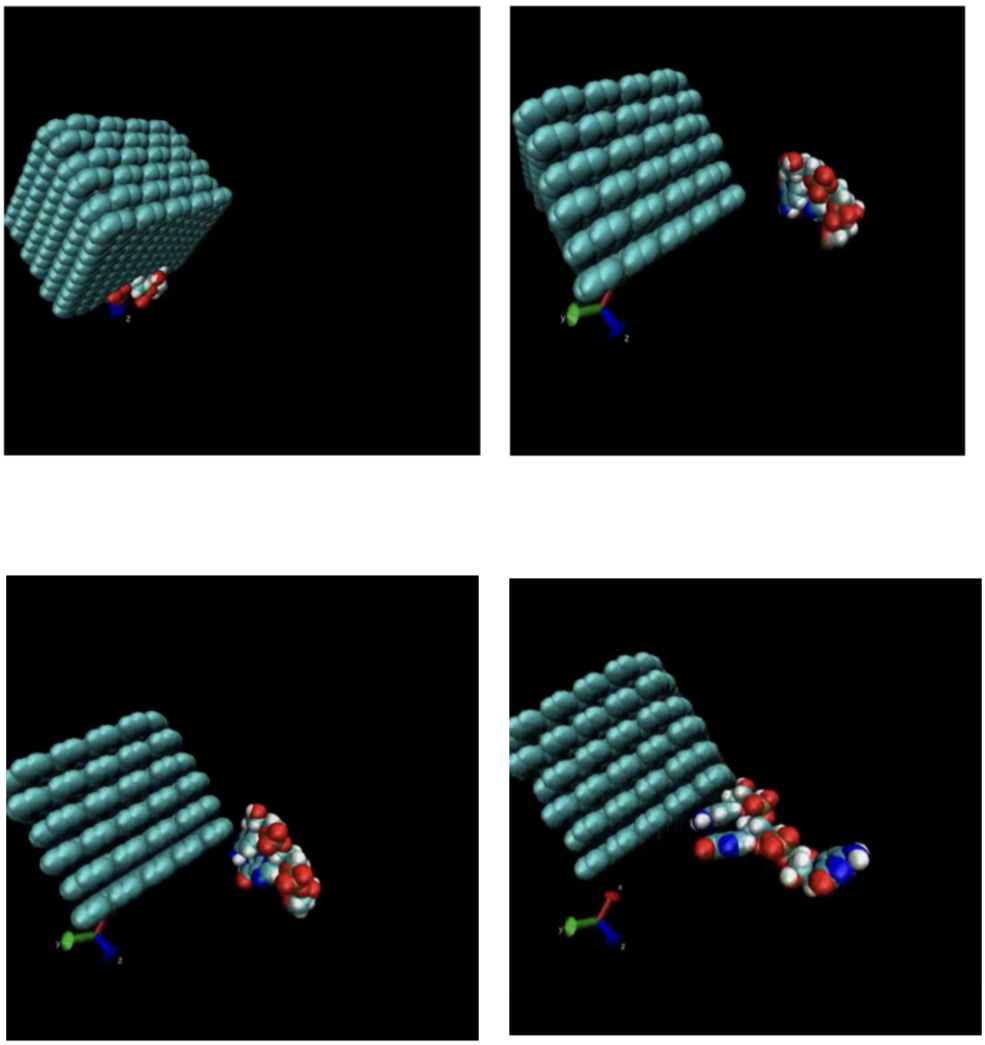




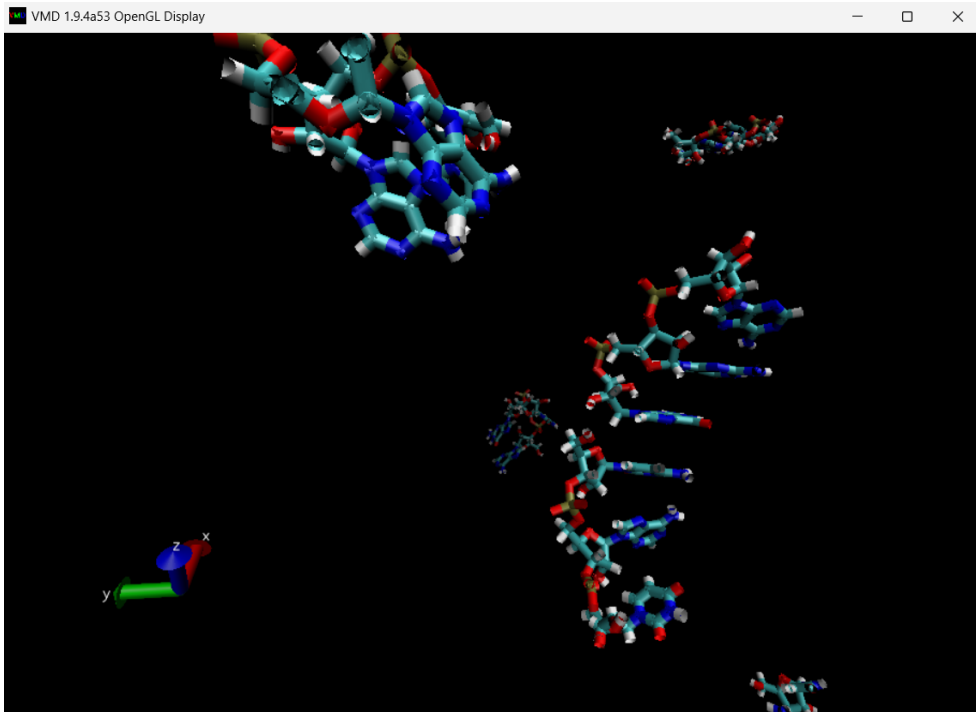
- Double-Stranded Formation Potential: This scenario demonstrates two oligonucleotides in close proximity on a graphene sheet. This arrangement shows potential for these dimers to initiate the formation of double-stranded structures in subsequent frames, or given slightly different input parameters.



- Surface Attachment and Translational Dynamics: The simulations unveiled intricate dynamics involving oligonucleotides bonding to the surface of a graphite sheet. Moreover, the movement of oligonucleotides, including translation across the surface of graphite, demonstrated the adaptability and mobility of these molecules in response to the carbonaceous templatizing material.



- Elongation Events: The simulations captured instances of elongation, where multiple oligonucleotides formed a Van der Waals bond along the sugar phosphate backbone, showing the possibility for the formation of long chain oligonucleotides. For the most basic lifeforms to exist, oligonucleotide chains must exceed 300 nucleotides, therefore, elongation was a crucial key endpoint to have been able to achieve.



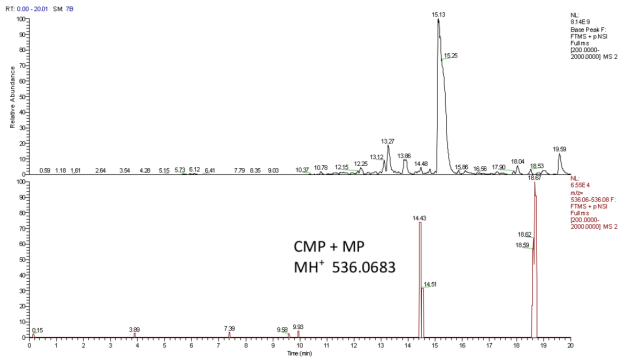
These multifaceted observations underscore the complexity of molecular interactions within carbonaceous systems and underscore the potential for diverse molecular configurations. The transformer outputs not only enabled the identification of these key endpoints but also provided a comprehensive view of how various parameters and conditions influenced the observed behaviors.

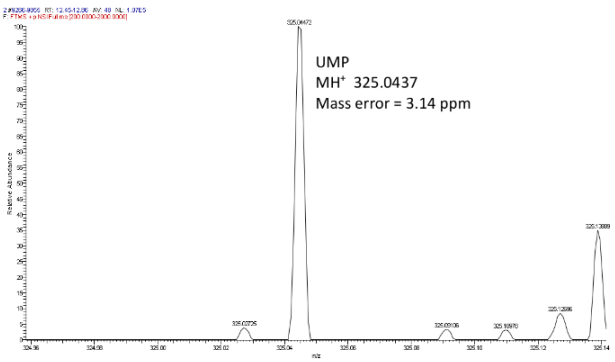
Further data analysis was conducted at UMASS Boston, using high performance liquid chromatography (HPLC) and liquid chromatography, with tandem mass spectroscopy (LC-MS-MS). The following experimental parameters were used:

- Gradient: 0-2 min 0% B, 2-12 min 0-50% B, 12-15 min 50-90% B, 15-16 min at 90 % B, 16-17 min 90-0% B, 17-20 min 0% B
- MS1: Orbitrap detector, 120000 resolution, 200-2000 m/z scan range, 30% RF lens, Custom AGC target 50%, custom injection time 100 ms, MIPS filter = small molecule,

Intensity filter = 1.0e3, dynamic exclusion filter: exclude after 1 time for 10 s with +/- 10 ppm excluding isotopes,

- MS2: quadrupole isolation, isolation window 1.2 m/z, activation type CID, fixed 35% collision energy, activation time 10ms, activation Q 0.25, detector type Ion Trap, San rate rapid, mass range normal, custom AGC target 200%, custom injection time 50 ms





In the results, there were chromatographic peaks for MH+ ions of UMP, CMP+MP, 2CMP, CMP+AMP and 2 CMP, demonstrating the presence of these, especially in conditions that contain Ca^{2+} ions.

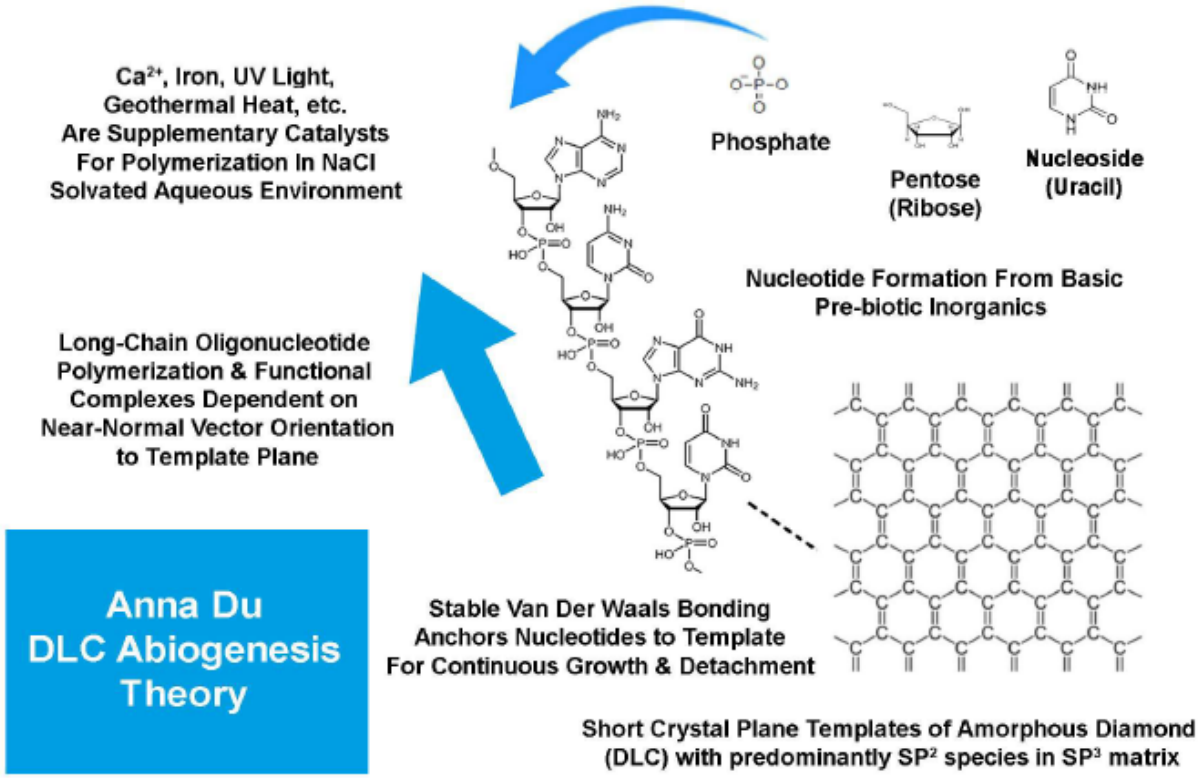
Conclusions

The first innovation presented in this research introduces a novel concept that challenges traditional theories surrounding the origin of life, and shows that DLC is the ideal templatizing material for oligonucleotide formation. 11 of the 38 experiments which were run contained DLC as the templatizing material. 9 of those, when analyzed with an LC-MS/MS, contained dimer and

trimer growth. 6 out of the 8 experiments which contained nanodiamond, and 4 out of 10 experiments which contained graphite, and 4 out of 9 for the experiments that contained graphene.

The anisotropism of the material allows for a continuous gradient of molecular spacings, creating a dynamic environment that allows for the key endpoints necessary for the formation of life. This variability in the structure allows for unique configurations of carbon rings and their respective distances, contributing to the diversity of organic polymers. Unlike other inorganic templates, the inherent stochastic inconsistencies present in carbonaceous crystal lattices are integral to this process. The notion of variable interspacing, whereby carbon rings and molecular constituents are not uniformly distributed, introduces a level of diversity that is critical for life's emergence.

This project has implications for the future of drug discovery. The current use of CPG (controlled pore glass), and to a lesser extent, polystyrene as a material for modern solid phase RNA synthesis only allows for oligonucleotides of ~100 -mers to build up. However, even simple proteins require double stranded nucleic acids of no less than ~150 base pairs in length. Given that DLC is an ideal environmental friendly templatizing surface for RNA to form, as it is not only also a stable surface resistant to swelling and shrinkage (unlike polystyrene), it is also more compatible with the carbon ring structures that are present in organic molecules, thus, longer oligonucleotides could form on the surface of DLC.



The second innovation advanced by this research introduces a transformative approach to parameter optimization by utilizing transformer models. The core concept behind this innovation is to utilize advanced transformer models, which are advantageous in handling complex and diverse data patterns. By using transformer models for parameter optimization, this innovation enables researchers to efficiently explore a vast parameter space, significantly reducing the need for exhaustive experimental runs. The transformer-based AI approach rapidly identifies the optimal parameter configurations that yield desired outcomes, having a 94% accuracy in predicting collision events when compared to the output generated by the Polychrom simulations.

By automating and enhancing the parameter optimization process, researchers can rapidly generate high-quality simulation outputs without compromising on statistical significance. This is especially critical in fields like materials science, drug discovery, and molecular biology,

where accuracy and efficiency are crucial, and has the potential to revolutionize the field of drug discovery. By carefully studying and understanding the beginning of life, we may very well find much-needed solutions to ensure the future of life.

…

"The cosmos is within us. We are made of star-stuff. We are a way for the universe to know itself."

~Carl Sagan

References

1. England, J. L. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics*, *139*(12), 121923. https://doi.org/10.1063/1.4818538

2. Ilya Prigogine (1975). Dissipative structures, dynamics and entropy. *The Quantum Chemistry, 19/25 January 1975* https://onlinelibrary.wiley.com/doi/10.1002/qua.560090854

3. Rothstein, D. M. (2012, May 1). Solid-Phase supports for oligo synthesis. *GEN - Genetic Engineering and Biotechnology News*. https://www.genengnews.com/magazine/solid-phase-supports-for-oligo-synthesis/

4. Swadling, J. B. (2010, September 15). *Clay minerals mediate folding and regioselective interactions of RNA: A large-scale atomistic simulation study*. ACS Publications. https://pubs.acs.org/doi/abs/10.1021/ja104106y

5. S. Jo, T. Kim, V.G. Iyer, and W. Im (2008) CHARMM-GUI: A Web-based Graphical User Interface for CHARMM. J. Comput. Chem. 29:1859-1865

6. B.R. Brooks, C.L. Brooks III, A.D. MacKerell, Jr., L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D.M. York, and M. Karplus (2009). CHARMM: The Biomolecular Simulation Program. J. Comput. Chem. 30:1545-1614

7. J. Lee, X. Cheng, J.M. Swails, M.S. Yeom, P.K. Eastman, J.A. Lemkul, S. Wei, J. Buckner, J.C. Jeong, Y. Qi, S. Jo, V.S. Pande, D.A. Case, C.L. Brooks III, A.D. MacKerell Jr, J.B. Klauda, and W. Im (2016). CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations using the CHARMM36 Additive Force Field. J. Chem. Theory Comput. 12:405-413

8. S. Jo, X. Cheng, S.M. Islam, L. Huang, H. Rui, A. Zhu, H.S. Lee, Y. Qi, W. Han, K. Vanommeslaeghe, A.D. MacKerell, Jr., B. Roux, and W. Im (2014) CHARMM-GUI PDB Manipulator for Advanced Modeling and Simulations of Proteins Containing Non-standard Residues. Adv. Protein Chem. Struct. Biol. 96:235-265

9. Y.K. Choi, N.R. Kern, S. Kim, K. Kanhaiya, S.H. Jeon, Y. Afshar, S. Jo, B.R. Brooks, J. Lee, E.B. Tadmor, H. Heinz, and W. Im (2022) CHARMM-GUI Nanomaterial Modeler for Modeling and Simulation of Nanomaterial Systems. J. Chem. Theory Comput. in press

10. crystallography365. (2014, September 22). *Tetrahedral amorphous carbon: Rough in the diamond*. Crystallography365.

https://crystallography365.wordpress.com/2014/09/22/tetrahedral-amorphous-carbon-rough-in-the-diamond/
11. Dhindsa, G. K., Bhowmik, D., Goswami, M., O'Neill, H., Mamontov, E., Sumpter, B. G., Hong, L., Ganesh, P., & Chu, X. (2016). Enhanced Dynamics of Hydrated tRNA on Nanodiamond Surfaces: A Combined Neutron Scattering and MD Simulation Study. *The Journal of Physical Chemistry B*, *120*(38), 10059–10068. https://doi.org/10.1021/acs.jpcb.6b07511
12. Ferris, J. P. (2006). Montmorillonite-catalysed formation of RNA oligomers: The possible role of catalysis in the origins of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1474), 1777–1786. https://doi.org/10.1098/rstb.2006.1903
13. Kopidakis, G., Remediakis, I. N., Fyta, M. G., & Kelires, P. C. (2007). Atomic and electronic structure of crystalline–amorphous carbon interfaces. *Diamond and Related Materials*, *16*(10), 1875–1881. https://doi.org/10.1016/j.diamond.2007.07.013
14. Olawanle, J. (2022, October 5). Big O cheat sheet – time complexity chart. *freeCodeCamp.Org*. https://www.freecodecamp.org/news/big-o-cheat-sheet-time-complexity-chart/
15. Ren, H., Wang, J., Zhao, W. X., & Wu, N. (2021, August 14). RAPT: Pre-training of time-aware transformer for learning robust healthcare representation. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. http://dx.doi.org/10.1145/3447548.3467069
16. Seif, A., Hafezi, M., & Jarzynski, C. (2020). Machine learning the thermodynamic arrow of time. *Nature Physics*, *17*(1), 105–113. https://doi.org/10.1038/s41567-020-1018-2
17. Sitapure, N., & Kwon, J. S.-I. (2023). CrystalGPT: Enhancing system-to-system transferability in crystallization prediction and control using time-series-transformers. *Computers & Chemical Engineering*, *177*, 108339. https://doi.org/10.1016/j.compchemeng.2023.108339
18. Yu, H., Zhong, Y., Ji, J., Gong, X., & Xiang, H. (2022). *Time-reversal equivariant neural network potential and Hamiltonian for magnetic materials*. American Chemical Society (ACS). http://dx.doi.org/10.26434/chemrxiv-2022-h6f69
19. Zhang, S., Murray, N., Wang, L., & Koniusz, P. (2022). Time-rEversed DiffusioN tEnsor Transformer: A New TENET of Few-Shot Object Detection. In *Lecture Notes in Computer Science* (pp. 310–328). Springer Nature Switzerland. http://dx.doi.org/10.1007/978-3-031-20044-1_18