BAR-ILAN UNIVERSITY–YESHIVA UNIVERSITY

Summer Science Research Internship Program 2023

The Bar-Ilan University–Yeshiva University Summer Science Research Internship Program is an amazing research opportunity for undergraduate men and women, allowing them to contribute to the forefront of science research taking place in Israel. Generously supported by former chairman of Bar-Ilan's Global Board of Trustees, Dr. Mordecai D. Katz z"l and his wife Dr. Monique Katz, the Irving I. Stone Foundation and the Zoltan Erenyi Fund, students gain invaluable laboratory skills, along with an unforgettable summer experience.

Program Directors: Prof. Arlene Wilson-Gordon and Prof. Ari Zivotofsky
Av and Em Bayit: Rabbi Azi and Ellie Fine

TABLE OF CONTENTS

Editors: Hallie Gordon and Esther Karman
Contributing Editors: Bracha Weinberger, Akiva Lipshitz, and Jordan Levovitz

BRAIN SCIENCES

Esther Karman, Sara Stein, Talia Simpson, Cayla Muschel, Ma'ayan Tzur
Ezra Cooper, Jordan Levovitz

Entrainment of Tidal Rhythms in Aplysia by Artificial Tidal Cycles

Esther Karman and Jordan Levovitz
Advised under Prof. Avy Susswein and Dr. Yisrael Schnytzer

Almost all organisms have endogenous biological clocks that entrain to the natural cycles. It has been shown in many animals that adopting rhythms that follow these cycles confers an advantage [1]. The most commonly studied rhythm is the circadian clock, which follows the rising and setting of the sun. There are, however, additional cycles, such as the rotation of the moon around the earth. The lunar cycle is the dominant force responsible for the rising and falling of the tides. The solar and lunar cycles interact and together drive tidal rhythms.

Aplysia are marine gastropod mollusks, commonly known as sea hares. *Aplysia* have been under investigation for decades, and their neural circuits are well-mapped. Eric Kandel was awarded the Nobel prize in the year 2000 for his neurobiology research in Aplysia. *Aplysia* have been demonstrated to be malleable to feeding training paradigms [2] and are found in tide

zones. Previous clock work in *Aplysia* has focused on the circadian clock [3]; however, due to their occurrence in the tidal zone, as well as observations of their prolonged exposure to air, we theorized that they might possess a tidal clock as well. In this exploratory study, we set out to determine whether *Aplysia* can be entrained to a tidal rhythm, and whether they may be beneficial for research about the underpinnings of circatidal clocks.

We attempted to entrain several species of *Aplysia*, including 2 *depilans*, 6 *punctata*, and 1 *fasciata*. During the entrainment phase, eight cycles of high/low tide were simulated using a draining and filling system. During the free run phase, the *Aplysia* were transferred to individual tanks at constant water level. Their behavior was monitored over approximately 72 hours. To remove any obscuring factors, the experiment was conducted first in dark-only (DD) and then light-only (LL) conditions to account for the possibility of circadian behavior in the animals. After activity data was collected and recorded, the data was analyzed using charts in Excel and the Lomb-Scargle Periodogram (LSP) method to search for trends in behavior.

Entrainment **Free Run DD** **Free Run LL**

Figure 1. Setup during entrainment and free run phases.

We first conducted the DD entrainment. During free run, animal movement was recorded at 10-minute intervals, where movement was defined as displacement of the animal by one body length from its original position. Small head movements were disregarded. Initial data analysis revealed some patterns in the animals, with 55% (5 of 9) showing moderate tidal rhythm. Of the remainder, 22% (2 of 9) were excluded from the experiment, 11% (1 of 9) showed no tidal rhythm, and 11% (1 of 9) showed strong circadian rhythm.

A trend in data was noted in 80% (4 of 5) of the animals showing some form of tidal entrainment. Specifically, there were abnormal activity patterns and additional noise during low tide of the 3rd-5th tidal cycles. However, it was discovered that there was, in error, an additional low tide during the tidal simulation of the first entrainment phase which correlated with the abnormal behavior patterns displayed during the free run. We concluded that *Aplysia* are susceptible to tidal entrainment and that a repeat experiment could yield conclusive results.

We then conducted the entrainment in LL conditions. During the second free run, further data analysis revealed that activity moderately correlated with tidal rhythms in 37% (3 of 8) of the animals. An additional 37% (3 of 8) did not exhibit tidal behavior, and 25% (2 of 8) were excluded from the experiment. LSP analysis revealed unusual patterns in the animals showing some form of circatidal activity, specifically that the animals exhibited 8-hour or 15-hour rhythms, where 12.4-hour or 24.8-hour cycles were expected.

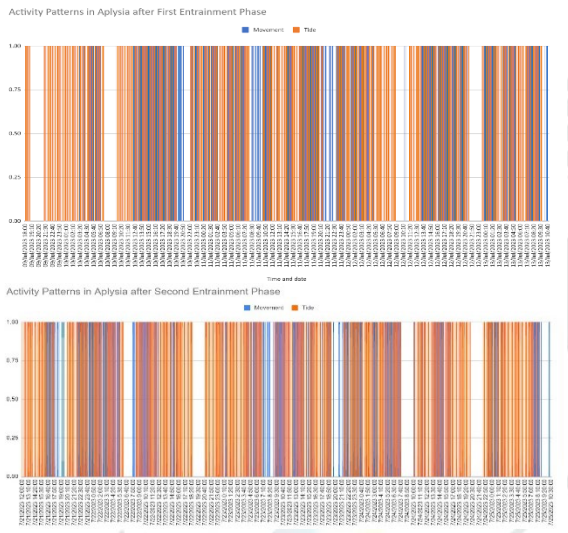


Figure 2. Activity patterns in *Aplysia* following the first and second entrainment phase.

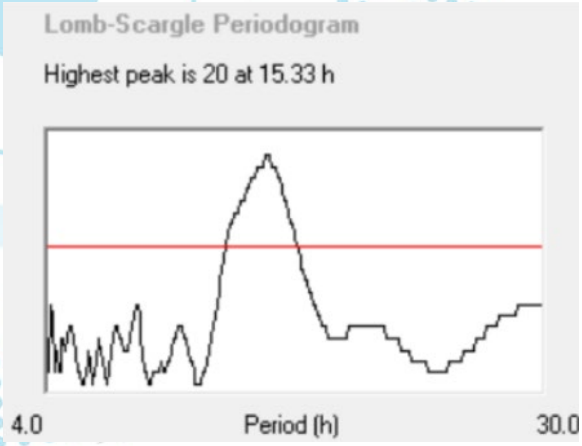


Figure 3. LSP analysis shows a significant peak at ~15 hours.

Several factors could account for the observed trend in behaviors. It may be that a longer entrainment period is necessary for the animals to exhibit a tidal rhythm. Other factors to consider are sample size and species. It's possible that our sample size was too small to draw conclusions about behavior at the population level, or that the inconsistency observed was due to variations among the different species used in entrainment. We concluded that *Aplysia* may be susceptible to tidal entrainment, but that a longer entrainment period and larger sample size are necessary to obtain accurate results. Moreover, it's important to consider differences between natural and experimental conditions. In the wild, *Aplysia* are found among seaweed; therefore, it makes sense that their activity patterns would change in the absence of algae.

Over the course of data analysis, several points were raised to bear in mind for repeat experiments. First, it was noted that the animals tend to remain on walls at low tide and are rarely found on the surface during low tide. Thus, classifying movement based on position may grant greater insight into tidal rhythm than analysis of movement alone. Additionally, it was observed that the animals occasionally allow themselves to be carried passively by the water's motion, an activity that should be distinguished from movement or non-movement. It was hypothesized that this behavior may be a replacement for the passive clinging to algae observed in *Aplysia* in their native habitats. Lastly, it is interesting to note that one species under investigation, the *Aplysia fasciata*, retained strong circadian rhythms despite attempts at tidal entrainment. This species is known to be a nocturnal animal that is primarily active at night. Their precise biological clock may be an area of interest for future research.

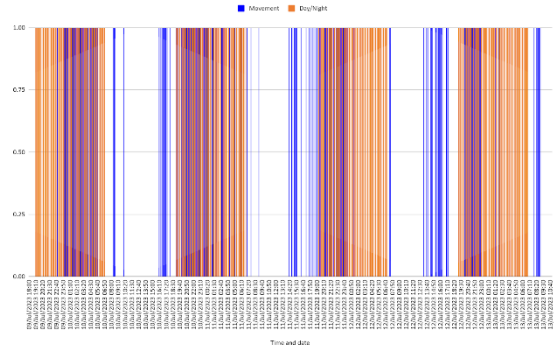


Figure 4. Aplysia fasciata exhibits strong circadian rhythm despite tidal entrainment.

[1] Häfker NS, et. al., Annual Review of Marine Science, **15**, 509-38 (2023).
[2] McManus J, et. al., Learning and Memory, **26**, 151-65 (2019).
[3] Lickey ME. Journal of Comparative and Physiological Psychology, **68**, 9-17 (1969).

The Depressometer: Using AI Technology in the World of Psychology

Ma'ayan Tzur, Sara Stein, Ezra Cooper
Advised under Prof. Eva Gilboa-Schechtman

The amount of people with depression today unfortunately is growing rapidly, and there are not enough therapists to properly treat and diagnose everyone. Additionally, therapy is expensive and sometimes can be subjective. If it were possible to use artificial intelligence to diagnose depression, therapy could become

more standardized, accessible, and cost effective.

One of the key aspects of the research at hand is to utilize computer programming to possibly aid in the diagnosis of depression. Additionally, this research is part of the process of trying to use AI and advances in computer programming as an aid in the field of psychology.

Using AI, the research seeks to explore the relationship between emotional congruence and depression. Thus, computer programming was used to note the emotional congruence across vocal, facial, and verbal channels. It is possible that both positive and negative emotional congruence are correlated with depression, and this too was relevant to the research.

Congruence levels were explored during the initial intake videos of individuals who were subsequently diagnosed with depression and treated for it. The research attempts to see if more severe initial depression is correlated with more or less congruence during intake, and if congruency can predict treatment outcomes. This is in the hope that some sort of computer algorithm which measures vocal, facial, and verbal congruence may be developed in order to help diagnose depression.

In order to do this, first, assessors diagnosed each participant with depression, rating its severity. Those diagnosed with depression then underwent sixteen sessions of psychodynamic therapy and their depression was evaluated upon completion. Then, emotional congruence during intake videos was assessed by measuring arousal and valence of voice, face, and content respectively [1]. Arousal describes how strong an emotion is, with a higher rating denoting intensely felt emotions such as anger and excitement and a lower rating denoting less intensely felt emotions such as tiredness (see Figure 1). Valence describes an emotion on a spectrum from positive to negative, with a higher rating denoting a more positive emotion such as happiness, and a lower rating denoting a more negative emotion such as sadness (Figure 1).

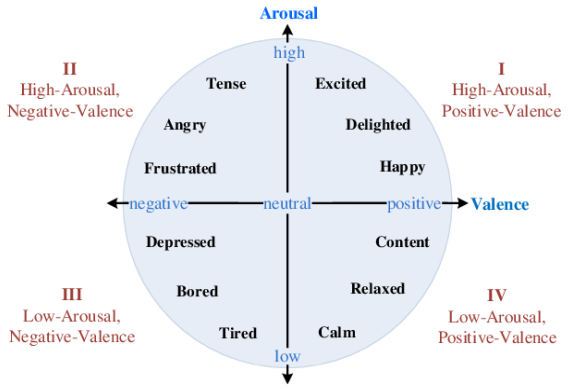


Figure 1. Valence and Arousal (Lung-Hao Lee).

To analyze the participants' speech, the segments of the therapist speaking in the intake videos were muted so that the videos could be run through The Geneva Minimalistic Acoustic Parameter Set (GeMAPS), a software that measures voice arousal and valence levels. Then, the videos were run through Facereader, a program that measures the valence and arousal of facial expressions (see Figure 2). The participants' speech was also transcribed and rated based on arousal and valence to assess content.

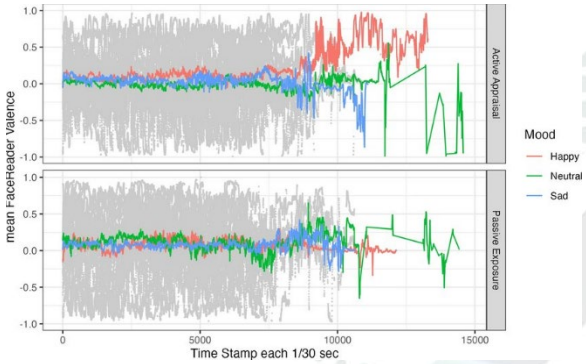


Figure 2. Example of FaceReader Analysis for Valence and Arousal (not actual data, just an example).

The data will be compared to determine congruence levels between voice and vocal channels, voice and content channels, and vocal

and content channels. The congruence levels will be used to determine if the severity of depression is correlated with congruence levels. Congruence levels between the channels were also evaluated by three researchers in order to compare the levels of congruence determined by software to those determined by individuals. In addition, the initial and final evaluation of depression levels will be compared with the level of congruence in order to determine how congruence may be correlated with treatment outcomes.

Currently, the research is still ongoing. That being said, the hope is that this research will not only merge computer programs with the world of psychology, but will even enhance therapeutic treatment for psychopathologies like depression, and perhaps other psycho-maladaptive disorders as well. While there is certainly more work to be done, we are excited about the future prospects stemming from the current research, as well as the other ways in which the newest computer programs will possibly enhance the world of psychology.

[1] L.-C. Yu et al., Proc. NAACL-HTL, 540-545 (2016).

A Meta-Analysis of Social Cognition in Turner Syndrome

Talia Simpson and Cayla Muschel
Advised under Dr. David Anaki

Turner Syndrome (TS) is a genetic disorder affecting between 1/2500 and 1/3000 live female births where one of the X chromosomes is completely or partially deleted. TS is characterized by unique physical, cognitive, and psycho-social features. Physical features include short stature, webbed neck, and ovarian dysfunction which leads to estrogen and androgen deficiency. Cognitive features include difficulties in visual-spatial tasks, visual-motor tasks, mathematics tasks, executive functioning tasks, and social cognition tasks. Psycho-social features include difficulties creating new relationships and maintaining those relationships, as well as being more withdrawn socially [1].

While factors such as insecurities, social stigma, and the struggles of living with TS play a role in psycho-social difficulties experienced by people with TS, the fact that social cognition has been shown to be impaired demonstrates that the chromosomal abnormality and its resulting physical differences themselves most likely contribute to those difficulties.

Two aspects of social cognition are focused on in the existing literature on TS. The first is emotion perception, which is the ability to use information on another person's facial expressions, voice, and body language to determine what emotions are being conveyed (Figure 1). The second is Theory of Mind (ToM), which is the ability to attribute thoughts, emotions, and intentions to oneself and others, thus enabling a person to predict others' behavior based on their perceived mental states (Figure 2). Studies have shown that patients with TS perform significantly worse on both emotion perception tasks and ToM tasks when compared

to healthy controls, demonstrating that there is a measurable impairment in social cognition.

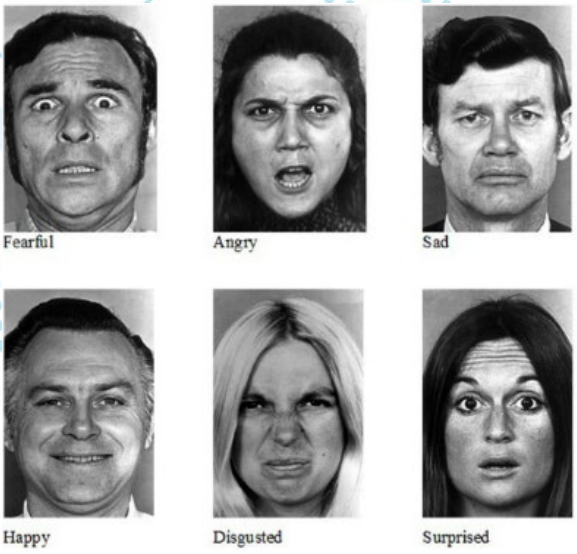


Figure 1. Emotion perception task example.

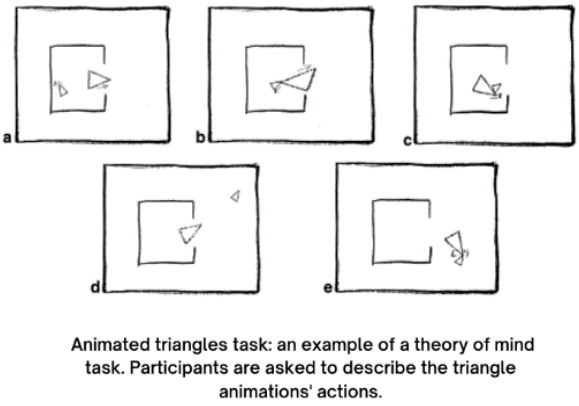


Figure 2. ToM task example.

Although many individual studies have been published examining social cognition in TS, to our knowledge a meta-analysis analyzing and combining those results into a comprehensive paper has not yet been published. Meta-analyses are important because individual studies on their own can be difficult to properly interpret when attempting to reach an evidence-based conclusion on the topic as a whole (Figure 3). In a meta-analysis, all relevant studies meeting the criteria for high quality data are analyzed and weighted appropriately, and the data is combined to create a pooled estimate.

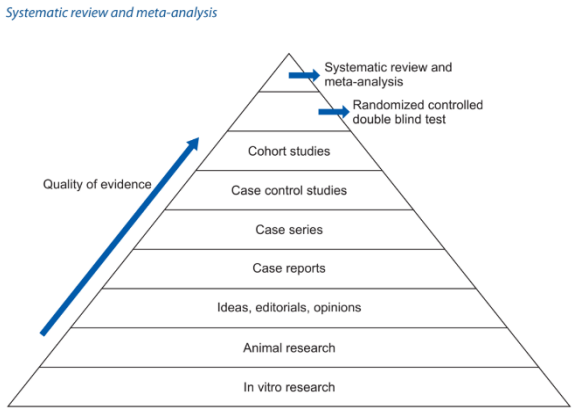


Figure 3. Levels of evidence.

The primary goals of the current meta-analysis were to measure the extent of social-cognitive deficits in TS and explore whether the relationship between TS and impairments in social cognition is moderated by other variables, such as the specific aspect of social cognition being tested, sensory modality, and age. To measure the extent of social-cognitive deficits in TS, studies about social cognition in TS that met the criteria were gathered (Table 1). Data was extracted, weighted, and pooled using a random effects model, measuring effect size with Cohen's *d* (Figure 4). To measure whether social cognition impairment in TS is moderated by other variables, subgroup analyses were performed. The first analyzed the impact of the type of social cognition task performed, the second analyzed the impact of the sensory modality in which the task was performed, the third analyzed the impact of the type of emotion being measured, and the fourth analyzed the impact of the participants' age when the task was performed.

Author	Year	TS Age	TD Age	TS N	TD N	M1/TS	M2/TD	S1/ TS	S2/ TD	Cohen's d	Task	HRT/pubertal onset
Anaki, D. et al.	2016-2018	30.58	29.07	26	26					0.38	AR and MSA	unknown
Elgar, K.	2001	24.91	25.3	23	23	7.643	8.717	2.071	1.253		AR	unknown
Good, C. et al.	2003	25.1	24.1	51	56	-5.164	-1.066	-4.39	-3.067		AR	Yes
Hong D.S. et al.	2011	7.89	7.56	25	22	9.49	10.49	2.89	2.92		AR	No
Klabunde, M., et al.	2020	14.41	13.41	14	12	0.836	0.934	0.287	0.341		MSA	unknown
Lawrence K., et al.	2003	24.6	24.4	42	56	7.835	8.835	0.847	0.445		AR	Yes
Mazzola, F., et al.	2006	32.6	31.1	18	17	7.338	8.782	0.751	0.888		AR	unknown
McCauley, E., et al.	1987	13.11	12.91	17	16	24.71	27.75	2.7	2.4		AR	unknown
Romans S.M., et al.	1998a	14.2	14.3	64	64					1.015437	AR	Yes
Romans S.M., et al.	1998b	17.3	17.5	35	25					0.34	AR	Yes
Ross, J. L., et al.	2004	31.2	33.5	94	96	78.33	84.17	5.83	4.984		AR	Yes
Ross, J.L., et al.	1995a	8.8	8.7	35	50	13.8	15	2.4	1.1		AR	No
Ross, J.L., et al.	1995b	12.3	12.2	21	50	15.2	15.7	0.9	0.9		AR	No

Table 1. Studies included in the meta-analysis, where AR is affect recognition and MSA is mental state attribution.

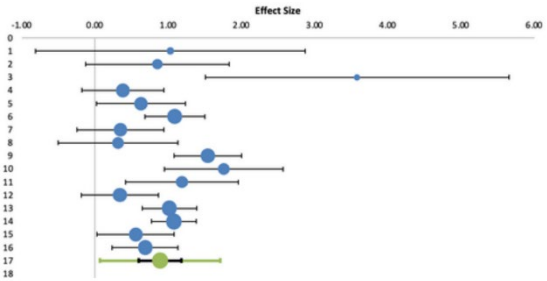


Figure 4. Forest plot of results.

The reason that type of social cognition task was analyzed as a potential moderating variable was because it is not yet clear from the existing literature whether TS specifically impacts emotion perception or ToM, or if both types of social cognition are impaired to a similar extent.

Sensory modality was analyzed as a potential moderating variable because while existing literature demonstrates impairments in visual-motor skills, visual-spatial skills, and gaze-tracking abilities, the connection between those visual perceptual deficits and impairments in social cognition is unclear. It is known that vision perception is an important contributor to social cognition skills, but we also wanted to explore the effect of other sensory modalities such as auditory and verbal tasks to determine whether the mode of perception has an impact.

Types of emotions were analyzed as a potential moderating variable since research has shown that TS patients have trouble discerning fear and anger specifically [2]. Therefore, we wanted to determine whether there was a difference in the social cognition skills of people with TS in certain emotions compared to others. For this we compared the six basic emotions of happiness, sadness, fear, anger, surprise, and disgust that were used in the existing literature.

Lastly, age at the time of testing was analyzed as a potential moderating variable because pubertal development may have an impact on social cognition skills [3]. Since TS is a chromosomal disorder that impacts the reproductive system as well as the development of secondary sex characteristics, many TS patients do not experience puberty

spontaneously. Rather, TS patients often receive hormone-replacement therapy (HRT) to induce development. To account for this, in the sub-group analysis we grouped studies where the participants were children (expected to be pre-pubescent), teenagers (expected to be in the midst of puberty), and adults (expected to be post puberty) separately. Some studies specified whether the TS participants were receiving HRT/ experienced spontaneous puberty or not, while others had mixed groups or did not specify. For the studies that specified, we analyzed ERT/post pubertal groups versus no ERT/pre-pubertal groups so that the results would be as accurate as possible.

The results of the meta-analysis yielded a large and significant effect size (Cohen's d= 0.89; p< 0.001; 95% CI [0.6 1.8]), which means the pooled data supports the conclusion that social cognition in TS is impaired to a statistically significant degree. For the sub-analyses the effect sizes were generally moderate, suggesting that the moderating variables influenced the heterogeneity of the original effect size analysis. Future research can examine how the types of social cognition tasks, sensory modalities, specific emotions, and age brackets impact social cognition in TS.

[1] D. Anaki, T. Zadikov Mor, & Z. Hochberg, Neuropsychologia, **90**, 274-285, (2016).
[2] K. L. Elgar, University College London, **146**, (2002).
[3] D. S. Hong, B. Dunkin, A. L. Reiss, J. Dev. Behav. Peditr., **32 (7)**, 512-520, (2011).

ENGINEERING

Jesse Lerner, Aaron Meoded, Effie Bluestone, Jeremy Zarge, Akiva Lipshitz
Hallie Gordon, Emily Haller, Emuna Rouhani, Nicole Haller

Voltage Modulated Charge Screening in Monolayer Tungsten Disulfide by Graphene

Hallie Gordon and Akiva Lipshitz
Advised under Prof. Doron Naveh and PhD student Aviv Schwarz

Background

Advances in materials research, microelectronics, and photonics achieved over the past few decades have led to promising new computing and sensing technologies. The basic science fueling these systems deals with the physics of semiconductors, materials whose electric conductivity falls between that of a conductor and an insulator, and when treated properly can adopt properties of either. Semiconductors have proven integral to the advancement of technologies including diodes and transistors, integrated circuits, and even solar cells. Our research focuses on a type of semiconductor known as two-dimensional transition metal dichalcogenides (TMDCs), thin lattice materials commonly fabricated via chemical vapor deposition (CVD) or exfoliation. Here, we investigate optical properties of Tungsten Disulfide (WS_2), and how they change

in response to charge screening by an applied layer of graphene.

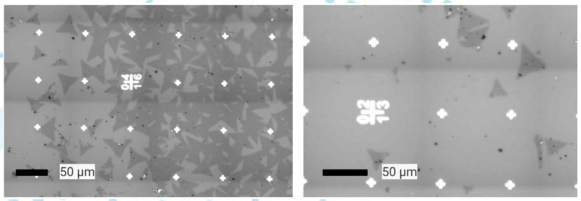


Figure 1. High resolution photo of the WS_2 flakes on our chip, with high magnification, taken under microscope. Flakes are about 25 microns wide.

Attempts to characterize the optoelectronic properties of new materials, such as WS_2, inevitably face a source of measurement bias when lasers in optical instruments excite charged quasiparticles in a material. Quasiparticles form when optically excited electrons become bound to electron holes, forming either electrically neutral or charged particles. Physically, electron holes are vacancies in the solution of states to the wavefunction. They behave as positively charged fermions with nonzero mass and have coulombic interactions with electrons and other holes.

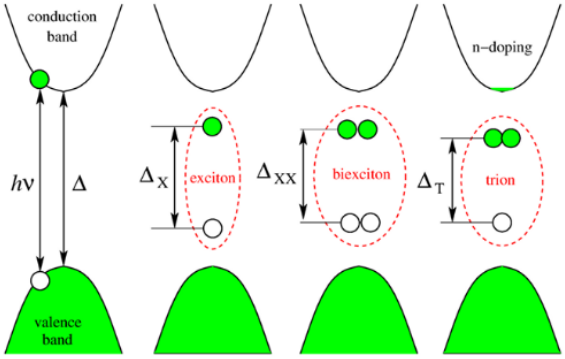


Figure 2. Formation of quasiparticles in semiconductors, such as electron holes, excitons (X), biexcitons (XX) and negative trions (X-). The positive trion (X+) is not shown. Image Source: Rodin, et al. (2021). Collective Excitations in 2D Materials.

The displacement of charged trions creates unwanted electromagnetic fields which bias measurements. Free-electron rich conductors functioning as electric screens have been proposed to correct this error. Electrons within the conductor rearrange themselves to counteract charges within the material, moving away from negative changes and towards positive charges. We select graphene for our electric screening system, as graphene has long been employed as a versatile conductor in the context of two-dimensional electrical materials.

Methodology

Set Up: In this experiment, we conduct measurements of WS_2 flakes deposited on a silicon support before and after the addition of graphene, and note their differences. To begin, we take CVD-grown WS_2 monolayer flakes and transfer them atop a SiO_2/Si substrate used for voltage biasing. Raman and PL spectroscopy measurements of the sample are taken before the addition of graphene.

Raman Spectroscopy: Raman spectroscopy measures the quantized photonic energy losses which correspond to phononic resonance modes of electrons within a material. Results of this measurement indicate what materials are present in our sample.

Photoluminescence (PL) Spectroscopy: PL involves the absorption of photons by a material's particles, which excites them to higher energy levels within their quantized orbitals. As these excited particles return to lower energy levels, excess energy corresponding to the difference between the excited and relaxed energy levels [band gaps] is released in the form of electromagnetic radiation. PL Spectroscopy measures the intensity of different light energies emitted during recombination. Results from this measurement help identify the various particles excited within a material.

We conduct our PL measurements at cryogenic temperatures (-173 °C) to sharpen the resolution of our measurements, using an automated liquid nitrogen cooling system. At higher temperatures, the Doppler effect on the interaction between the thermal energy of the particles and light causes a broadening of peaks

on a PL graph, which is a form of uncertainty. By lowering the temperature and decreasing excess thermal energy, low temperature PL graphs have narrower peaks, making it possible to identify quasiparticles in the data with higher confidence.

Voltage Modulated Graphene Charge Screen: Using standard techniques, CVD-grown graphene is layered across the entire chip. We etch the graphene off of the edgeplates of the chip to prevent electrical shorting, enabling us to create a potential difference between the graphene and the silicon support. We then connect a voltage source to electrical contacts on our chip and conduct PL and Raman scattering measurements. We expect a shifted and narrowed PL spectrum due to charge screening, such that the shift magnitudes correspond to the voltage strength. Producing a voltage difference between the graphene and the silicon support creates a charge-storing dielectric capacitor. Modulation of the gate voltage modifies the fermi level of the graphene. The graphene voltage dynamically controls the screening of photoexcited trions and the charge storage capacity of the capacitor. We expect that negative charge buildup causing a heightened Fermi level in the graphene will screen positively charged trions. Conversely, a lowered fermi level in the graphene created by positive charge buildup in graphene screens negatively charged trions.

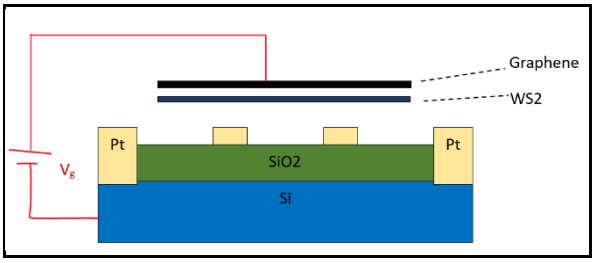


Figure 3. Cross-section diagram of our device.

Analysis and Results

Raman: Before the addition of the graphene, our Raman measurements showed two peaks, corresponding to the WS_2 (~350) and the silicon support (~520). Post graphene transfer, we see the addition of known graphene signatures.

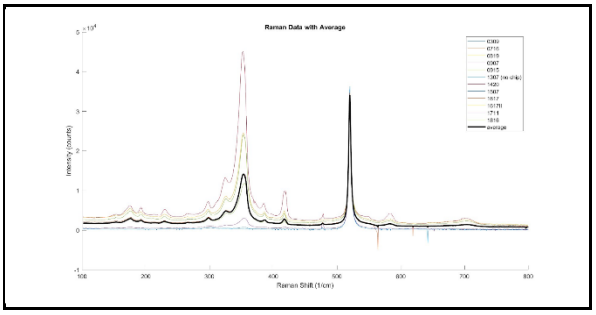


Figure 4. Raman data.

PL: The bare WS_2 PL spectrum contains multiple peaks which indicates that multiple quasiparticles are excited in the material during measurement. After analysis, we attribute each peak to a different species of quasiparticle excited. Post graphene transfer, we observe the decrease in intensities corresponding to trion recombination, as anticipated.

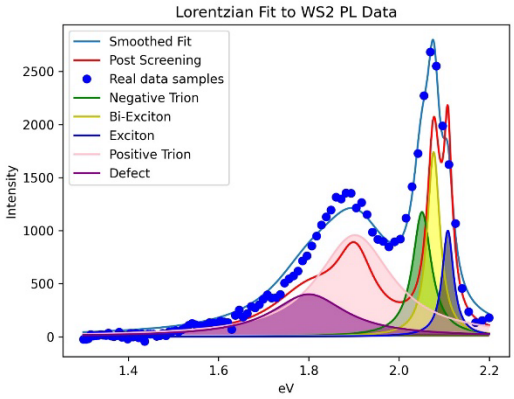


Figure 5. PL of WS_2 flakes at -173 °C with deconvolution of spectrum showing peaks of individual quasiparticles.

Our data shows a correspondence between the voltage applied to the silicon and the intensity of the PL spectrum. Notably, the positive trion peak height is diminished, consonant with the predictions made by screening theory. We expect to see that height of the peak varies symmetrically with the sign of the voltage – when the silicon voltage is positive, the positive trions will be screened, and the height of the positive trion peak is diminished, and the same is

to be observed for the negative trion peak when the silicon voltage is negative, but these measurements are yet to be taken.

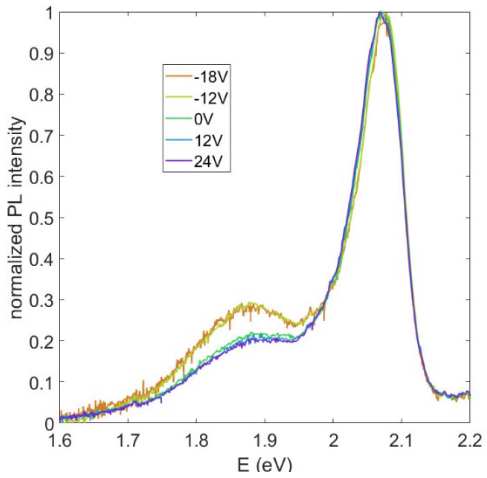


Figure 6. PL data from voltage modulated device.

Conclusion

In this work, we studied electron-quasiparticle interactions between monolayer WS_2 and conductive graphene. We used cryogenic photoluminescence spectroscopy to characterize optical emissions of WS_2 before and after a graphene transfer and observed changes in the PL spectrum consonant with what we would expect for screening of charged trions, proportional to the gate voltage in the silicon-graphene MOS capacitor. Charge screening is an important phenomenon to investigate as it will open up the possibility for new varieties of transistors and other electrical components.

[1] Lorchat, E., et al. Nat. Nanotechnol. **15**, 283–288 (2020).
[2] Ugeda, M., et al. Nature Mater. **13**, 1091–1095 (2014).
[3] Fábio J R Costa, et al. Nanotechnology **34**, 385703 (2023).
[4] Rahul Kesarwani et al. Sci. Adv. **8**, eabm0100 (2022).

Optimization in Conformance Checking

Emuna Rouhani and Aaron Meoded
Advised under Prof. Izack Cohen, PhD student Eli Bogdanov, and undergraduate students Liri Benzinou and Yuval Gerber

Process mining is a technique used to analyze, visualize, and improve business processes based on the data generated during their execution. It involves extracting information from event logs and other data sources, then using that information to gain insights into how processes are actually performed in practice. The main goal of process mining is to discover, monitor, and enhance process efficiency, effectiveness, and compliance.

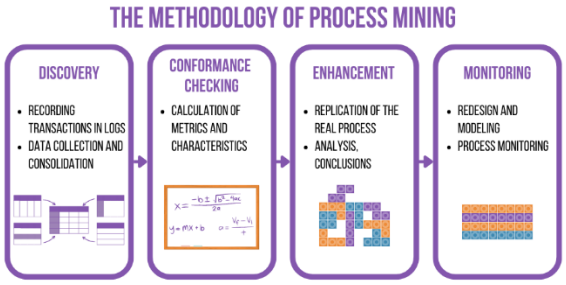


Figure 1. Process Mining Stages.

Conformance checking in process mining is a technique used to compare observed behavior, which is extracted from event logs, with a pre-defined process model or process specification. The goal is to assess the extent to which the observed behavior adheres to the expected or intended process. Alignment-based conformance checking is a specialized technique used to compare the observed behavior in event logs with the expected behavior represented by process models. It involves transforming event log traces into sequences of activities and systematically aligning them with the corresponding paths in the process model, in a model called the synchronous product.

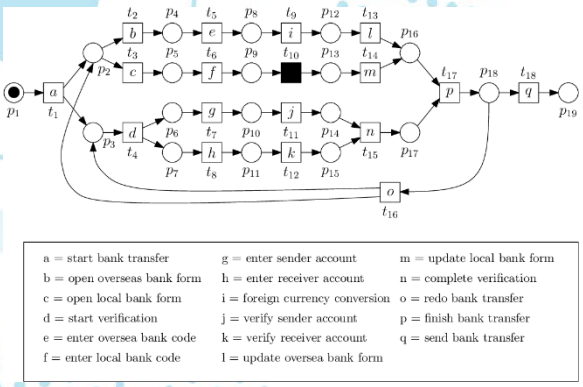


Figure 2. Synchronous Product Example of Money Transfer.

The alignment process identifies deviations, such as skipped or extra activities, and quantifies the conformance level. This approach provides a quantitative measure of how well the event log aligns with the process model, enabling organizations to diagnose process compliance, uncover inefficiencies, and focus on targeted process improvements. Finding optimal alignments means identifying the cheapest execution sequence of a synchronous product model according to some cost function. We use incremental shortest path algorithms for constructively computing optimal alignments through only building the relevant parts of the reachability graph in memory to conserve time and storage.

Enhanced Marking Equation Approach: The basis of this approach involved the A* shortest path search algorithm, which uses a function that underestimates the remaining costs of the optimal alignment. One option for the heuristic function is an ILP-based underestimation function, which uses the marking equation to estimate the remaining costs of the alignment. Process models are represented by Petri nets, and the marking equations of a Petri net encapsulate its behavior and structure through mathematical equations. The marking equation, denoted as $\mathbf{m} + C \cdot \mathbf{x} = \mathbf{m'}$, encompasses both the initial marking vector **m** and the target marking vector **m'**, along with an incidence matrix, C, that characterizes the changes in marking. Within these vectors, each entry corresponds to the quantity of tokens present at a specific place. A column in the incidence matrix defines a transition and its effect on the places that are represented by the rows. A transition removes or adds a token into a place that is represented with 1 or -1 respectively in the incidence matrix.

To enhance the efficiency of the marking equation heuristic, we aim to reduce the number of linear problems to be solved by minimizing the incidence matrix, which forms the foundation for solving linear problems. The incidence matrix, comprising |P| rows and |T| columns representing places (P) and transitions (T) in the Petri net model, can be reduced by converting rows and columns of irrelevant places and transitions to zero. This reduction streamlines the marking equation solving process, as a linear problem with all-zero coefficients possesses infinite solutions and no constraints, requiring relatively little processing time to solve. To identify irrelevant transitions, we employ Tarjan's algorithm to find strongly connected components (SCCs) in the reachability graph. We store the SCCs in a directed graph, perform topological sorting, and subsequently eliminate transitions within preceding SCCs, as cyclic and backward arcs are absent in the sorted structure. This targeted approach optimizes the incidence matrix, significantly reducing computational time and memory requirements during the heuristic's execution. The time complexity of the marking equation heuristic is about $O(n \cdot \log(n))$ time, where n represents the number of nodes in the reachability graph. This new heuristic is estimated to take $O(n \cdot \log(\log(n)))$ time.

Reinforcement Learning Approach: In Reinforcement Learning (RL), an agent uses its current state and information about the possible actions around it (including the rewards associated with that action) as inputs in deciding which action it should take next in the environment. It receives rewards from states that it transitions to, and gradually learns which

decisions yield higher rewards. Deep Reinforcement Learning (RL) is a machine learning approach that relies on modeling a problem in a way that conforms to the RL structure, while using neural networks to estimate the cumulative future rewards from each action taken. We utilized a value-based Deep Q-learning Neural Network (DQNN) that incrementally explores the synchronous product reachability graph. The RL agent explores the synchronous product graph from the initial marking (node) until the final marking, gradually learning to do so in the fastest way. The RL agent is trained on different synchronous product graphs of the same model and differing traces, to be generalized to new trace instances that are applied to the same model that the agent was trained on. This means that after training, the agent is able to traverse new trace instances very quickly, though at the cost of being suboptimal.

Sequence Alignment Approach: Sequence alignment is a method used in bioinformatics and other fields to compare and find similarities between two sequences of characters, such as DNA, RNA, or protein sequences. The goal is to arrange the characters in the sequences to maximize the number of matched or aligned characters and minimize the number of insertions, deletions, or substitutions required. The Needleman-Wunsch algorithm computes the alignment of two n-length sequences in $O(n^2)$ time. We propose reducing alignment-based conformance checking by splitting up the reachability graph of the process model into each of the unique paths it is composed of, and then performing a sequence alignment using the Needleman-Wunsch algorithm between each one of the unique paths and the given trace. The lowest-cost alignment is saved and returned as the optimal alignment. The main obstacles present in making this reduction include encountering a potentially exponential number of unique paths in the model (relative to the number of nodes in the graph) as well as the potential presence of cycles. Both are issues that are either solved directly or tackled with heuristic solutions.

DNA Genomic Diagnostics Optimized Using a Hamming Distance Tolerant Content-Addressable Memory (HD-CAM) as an Accelerator

Effie Bluestone
Advised under Prof. Adam Teman and Master's student Victor Galindo

In this groundbreaking abstract, we presented a novel and revolutionary approach that expedited the DNA Genomic Diagnostics process, with a specific focus on bacterial DNA classification. Our method harnessed the advantage of a Hamming Distance tolerant Content-Addressable Memory (HD-CAM) as a powerful hardware accelerator, leading to enhanced efficiency and reliability. The inner workings of the circuitry in the HD-CAM relied heavily on NOR-type logic gates to properly implement associative memory bitcells, as shown in Figure 1. Specifically, this design excelled at approximate matching operations and comparisons with programmable tolerance while also being efficient with memory allocation and energy consumption.

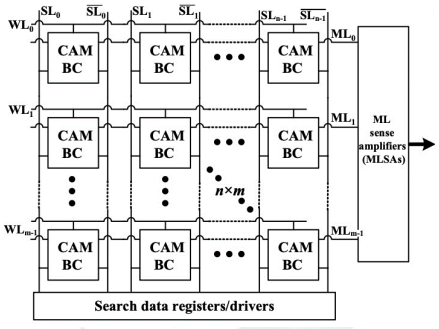


Figure 1. The inner workings of a HD-CAM.

The optimization of the DNA Genomic Diagnostics process revolved around three key areas:

Firstly, the computation time that each operation took directly depended on what clock frequency the HD-CAM was chosen to operate under. This implementation allowed the approximate matching operations to use the matchline charge held in the NOR gates, making this solution differ from the classic rise or fall time typically employed in other scientific solutions. This NOR gate structure can be seen in Figure 2 below. The matchline charge depended largely on the frequency of the clock. We addressed the critical issue of computation time by elevating the sequencing clock frequency to an optimal level. This advancement enabled us to achieve better data processing speed without compromising the accuracy of exact match sequencing. As a result, we obtained results much faster and streamlined analyses. The clock frequency was improved from a mere 25MHz to an astonishing 142MHz. This was just above an 800 percent increase in frequency, reducing the period eight times the original time it took.

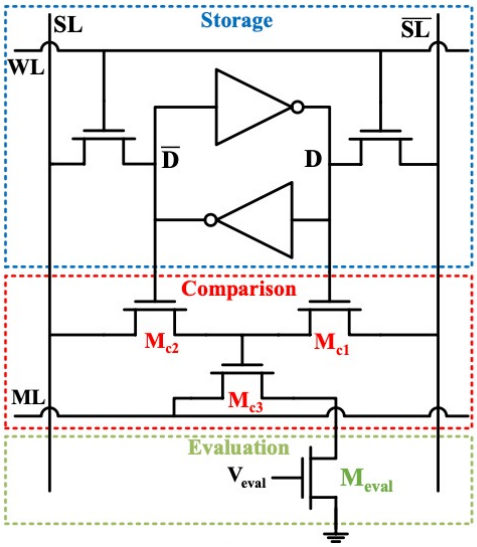


Figure 2. The NOR-type based static associative memory bitcells.

Secondly, the process of DNA encoding was extremely unorganized and inaccurate. Letters were constantly generated in wrong locations throughout the string, disrupting the major parts of the full computation DNA genetic read process. To get around this major drawback, the entire DNA encoding algorithm was modified to introduce a highly robust DNA letter encoding method that significantly improved resilience to errors during data retrieval and analysis. This innovative encoding technique reduced the likelihood of inaccuracies, resulting in elevated data integrity levels previously unattainable. Additionally, we developed a specialized Python script to validate the binary encoding of genetic reads, ensuring efficient verification of genomic diagnostics results and instilling confidence in the entire process. The DNA encoding process was completely refined and became superior to what preceded it, making it extremely reliable with zero errors in the encoding process. The genetic reads were able to be verified, and the process is now flawless.

Thirdly, our current algorithm used the brute force approach, which was extremely inefficient in interfacing with the HD-CAM. This could be thought of as a bottleneck in the overall computation time in the DNA genomic diagnostics process. To enhance and speed up the communication between the algorithm and the HD-CAM in this part of the analysis process, the integration of two new cutting-edge algorithms known as *Jellyfish and Bowtie2* was researched. In order to complement the HD-CAM hardware accelerator, both are beginning to be integrated into our genomic diagnostics approach. This would seemingly boost the computational efficiency of DNA genomic diagnostics, pushing the boundaries of speed and performance even further. The power of isolating a computation done entirely on software and using hardware could vastly improve the overall time. The integration of these two advanced algorithms with HD-CAM

could create a formidable and unparalleled computational powerhouse, revolutionizing the field of DNA genomic diagnostics.

Moreover, the integration of *Jellyfish* and *Bowtie2* algorithms with HD-CAM can not only improve computational efficiency but also enable a more comprehensive analysis of complex bacterial DNA. The combination of hardware and software-based algorithms should provide a seamless and synergistic approach, allowing for more accurate identification and classification of bacterial strains, even in the presence of genetic variations and mutations. This breakthrough could have significant implications in fields such as epidemiology, where rapid and precise bacterial identification is crucial for disease control and outbreak management. These advancements paved the way for wider adoption of this transformative technology across diverse scientific domains. Our novel approach opens up new possibilities for studying the human microbiome, offering insights into the diverse microbial communities that reside in and on the human body. Understanding the intricacies of these microbial ecosystems has profound implications for personalized medicine, as it can aid in the development of targeted therapies and interventions based on an individual's unique microbiota composition. The full Genomic Diagnostics process is shown below in Figure 3.

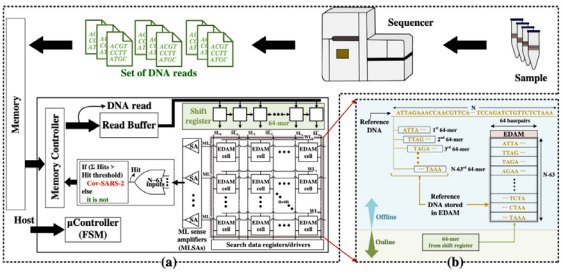


Figure 3. The full DNA Genomic Diagnostics process.

Furthermore, the application of HD-CAM in DNA genomic diagnostics goes beyond bacterial classification. This hardware accelerator can be adapted for various genomic analysis tasks, including gene expression profiling, structural variant detection, and metagenomics. As a result, the impact of this research extends far beyond the realm of bacterial DNA classification, unlocking new avenues for exploring and understanding the complexities of the genome. By leveraging HD-CAM and incorporating innovative algorithms, we redefined the boundaries of achievable speed, accuracy, and computational efficiency. Our research unlocked new opportunities for scientific discovery and laid the groundwork for more advanced applications in the future. This is an important turning point in revolutionizing DNA genomic diagnostics, shaping the landscape of modern genomics, and its potential impact on various scientific disciplines. As DNA genomic diagnostics continues to play a pivotal role in advancing various scientific disciplines, our research marks a significant milestone in propelling the frontiers of genomics, ushering in a new era of understanding and innovation.

Acceleration of Gene Interaction Analysis with Abstract Boolean Networks

Nicole Haller and Emily Haller
Advised under Prof. Hillel Kugler and PhD student Eitan Tannenbaum

Biologists have long studied gene interactions in order to better understand the dynamic behavior of biological systems. Studying these interactions gives scientists insight into an organism's characteristics, inheritance, possible mutations, etc. These observations also give scientists insight into the change in system behavior under different inputs or under genetic manipulations. This information enables scientists to reverse engineer gene networks, a process in which the regulatory system is

reasoned from observational behavior [1]. Yet, the large number of gene interactions in most organisms has made it extremely complex for humans to study them efficiently by hand. To study these interactions concretely and more effectively, scientists have utilized computational biology. For instance, scientists commonly utilize Boolean Network diagrams (Figure 1) to represent these gene interactions. In Boolean Network diagrams nodes represent genes that can be either on (1) or off (0) and gene interactions are represented by directed edges between nodes. These edges are either positive, representing an activating relationship, or negative, representing an inhibiting relationship.

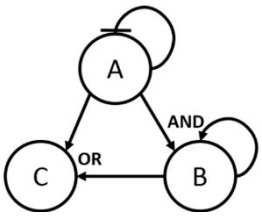


Figure 1. Boolean Network with components A, B, and C. Arrows indicate a positive/activating interaction, while flat-end symbols indicate a negative/inhibiting interaction. A, B, and C can be set to 0 (off) or 1 (on).

Many Boolean Networks contain optional gene interactions, whose existence is uncertain because they lack sufficient experimental evidence, as well as definite interactions, which are proven to exist by well founded experimental evidence [2]. These networks can be represented utilizing an Abstract Boolean Network (ABN). In an ABN, optional interactions are represented by a dotted-lined edge, and definite interactions are represented by a solid-lined edge. When experimental constraints, in which component states are defined at specific times, are added to the ABN, it is called a Constrained Abstract Boolean Network (cABN) (Figure 3) [1].

```
//Components
Otx2 (0..17); Esrrb(0..17); Otx2(0..17);

//Definite interactions
Esrrb Esrrb positive;
Dnmt3a Dnmt3a positive;
Otx2 Otx2 positive;

//Optional interactions
Dnmt3a Esrrb negative optional;
Otx2 Esrrb positive optional;
Esrrb Dnmt3a negative optional;
Esrrb Otx2 positive optional;
Dnmt3a Otx2 positive optional;
Otx2 Dnmt3a positive optional;

//Experiments
#ExperimentOne[0] |=  $rgb "Exp1 initial expression pattern";
$rgb :=
{
 Esrrb = 1 and
 Dnmt3a = 1 and
 Otx2 = 1
};
```

Figure 2. Original unfiltered RE:IN file which contains all definite and optional gene interactions, and constraints defined as individual experiments.

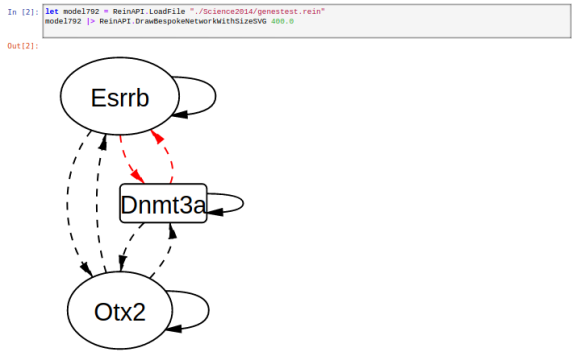


Figure 3. Constrained Abstract Boolean Network (cABN) which is output after the RE:IN file (Figure 1) is entered into RE:IN. Dotted lines indicate an optional interaction, while solid lines indicate definite interactions.

Since manually building these models is time-consuming, tedious, and introduces bias, automated formal reasoning is applied to make the process of analysis and testing more efficient [2]. The Reasoning Engine for Interaction Networks (RE:IN), which is executed within the SMT Solver Z3 [2], is a tool that uses automated formal reasoning to automate this analysis and testing. RE:IN works by processing RE:IN files (Figure 2) which contain components, gene interactions, and experimental constraints. After processing this information, a cABN diagram of the model (Figure 3) is created that is consistent with the experimental observations. In this way,

we only consider models which define the biological behavior we are studying. RE:IN also outputs whether a network has solutions or not. A network has solutions if there is a model which is satisfiable, meaning it can function given the experimental constraints. If solutions are found to exist, we can create minimal networks by eliminating some optional interactions, allowing for a better and simpler understanding of that organism's biological processes [2].

Currently, RE:IN files (Figure 2) that are inputted into RE:IN include all optional interactions, even if there is a very low likelihood of that interaction occurring. Since RE:IN considers all of these optional interactions, there is potential for the creation of inaccurate models in RE:IN. To reduce this inaccuracy, before the file is inputted into RE:IN, scientists reference a corresponding Excel file (Figure 4) in an attempt to manually filter out optional interactions with lower confidence levels (those less likely to interact). In the Excel file, a number represents the confidence level, meaning how likely the two genes are to interact. For example, the closer the number is to 0, the less likely the genes are to interact. This manual process is extremely time-consuming, rigorous, and susceptible to errors. Therefore, we developed a tool that automatically eliminates some optional interactions from the RE:IN file. Our tool works by storing a threshold number which is requested from the user. When this threshold number is applied, any optional interactions whose confidence level is below the user threshold are eliminated from the output RE:IN file. As a result, our tool successfully eliminates a number of less likely optional interactions, based on the number the user chooses. This is a much more systematic and efficient approach than performing this extensive search and elimination manually, and it also creates more accurate models of gene interaction in RE:IN.

	Esrrb	Dnmt3a	Otx2
Esrrb	1	-0.3	0.5
Dnmt3a	-0.3	1	0.8
Otx2	0.5	0.8	1

Figure 4. Excel File containing the confidence levels of specific gene interactions.

After some optional interactions are eliminated from a RE:IN file, the user inputs the file into RE:IN to test it for satisfiability. Often, during this process, scientists are interested in what the lowest threshold number is that produces a satisfiable model in RE:IN. Therefore, they manually calculate thresholds based off of the Excel file, which is extremely time-consuming and difficult. It also results in the creation of multiple new RE:IN files, often with the same or similar amounts of optional interactions. This is inefficient because the same files are then tested multiple times. To address this inefficiency, one feature of our tool allows a user to choose how many output RE:IN files to create. Based on the number of total optional interactions in the original file, our tool automatically calculates the thresholds which will output the desired files. It then distributes a linearly decreasing amount of optional interactions into each respective file using the automated thresholds. This provides a more uniform distribution of optional interactions and allows for faster testing and analysis by ensuring that each file is unique before testing.

Additional research in this area could further accelerate the process of studying gene interaction networks, such as: 1) a tool that automates the process of inputting the new filtered files into RE:IN 2) an algorithm which halts the creation of RE:IN files when a threshold has been found that creates a model that is satisfiable in RE:IN. In this way, less time is wasted from continuing to create unnecessary files.

Tools of computational biology are necessary to gain an accurate understanding of gene

interaction networks. Therefore, our tool and its subsequent development, are necessary for working towards both the acceleration and advancement of scientists' study of genes and their interactions in the future.

[1] Liu, Zhi-Ping. *Current Genomics*, (2015).
[2] Yordanov, B. et al. *npj Systems Biology and Applications* **2** 16010, (2016).

Creating a Conversational Chatbot Using Google Software API

Jeremy Zarge
Advised under Prof. Sharon Gannot and Mr. Pinchas Tandeitnik

Amidst the challenge of severe hospital understaffing, effective patient communication is paramount. ARI, our groundbreaking robot, was crafted to tackle this need head-on. In addition to using its cameras to detect people close by, ARI uses advanced audio processing to actively listen to and comprehend nearby conversations. Our vision for ARI extends to fostering thoughtful, human-like interactions, creating a more empathetic and personalized communication experience, especially tailored for older patients.

To create a conversational chatbot, we had to use three main APIs: Google Cloud Text-to-Speech (TTS), Google Cloud Speech-to-Text (STT), and Google Bard (an AI chatbot). First, the user speaks into the device's microphone, and the audio is recorded. Once this audio is recorded, it is sent to the STT API and converted into text. The text is sent as input to Google Bard, which outputs its response as text. This output text is sent to the TTS API, where it is converted to audio and then played by the device's loudspeakers. Thus, through this process, a real-time conversation is simulated. This process can be visualized below in Figure 1.

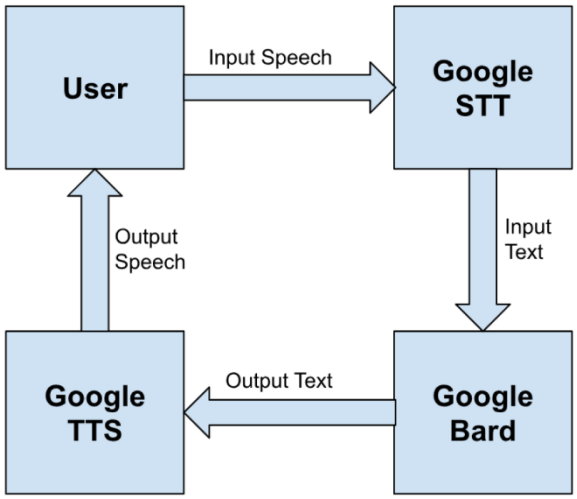


Figure 1. General Flow Diagram for Chatbot.

Although Google Bard is not as popular as GPT, it is among the more powerful LLMs (Large Language Models). While we would have liked to use GPT, it would have been very expensive to use, whereas Bard had a free API. Unlike many other LLMs we tested, Bard has a strong memory, meaning it can remember previous parts of the conversation such as a user's name and interests, and answer their questions keeping those in mind.

This process went through many iterations during optimization of the program. For example, we initially used audio files to process audio data on both the input and output ends. On the input end, this meant that after each time the user spoke, an audio file was created, which was subsequently sent to the STT API. On the output end, this meant that the TTS created an audio file, which was then played on the computer's speaker. This process was effective, but took more time than necessary. Additionally, creating additional files could lead to storage issues on some devices. To confront these challenges, we found a way to stream the data directly on both ends. Audio input data was streamed directly to TTS, and audio output data was streamed directly from STT to the speaker. Doing so made the whole process much cleaner and quicker.

Another problem encountered was managing the length of Bard's response. Often, when answering questions and explaining concepts, Bard would give extremely long responses. While very informative, such answers are not conducive to a normal conversation and may exhaust potential users. We had to brainstorm ways to get shorter answers. First, we tried asking Bard to give shorter answers in the future. Despite its capability to recall other pieces of information from earlier in the conversation, it continued to give long responses. Next, we tried asking Bard to provide a three-sentence summary after its response and put that summary between certain character symbols, like "%!%". Doing so allowed us to extract the summary from Bard's response. This strategy worked well but was not consistent. Sometimes, it would forget the symbols or put each individual sentence between the symbols. This made it very difficult to consistently extract what we wanted. Our current strategy is to calculate how large Bard's response is, and if it is above a certain threshold (600 characters), we ask Bard to give a shorter (1-3 sentence) version of its previous response. So far, this has worked well.

One other modification we made (but have yet to apply to the process) is incorporating WebRTC's VAD (Voice Activity Detector). VAD allows us to monitor audio data and determine which parts contain audio and which are silent. Because using the speech-to-text API costs money, limiting the amount of data we send by cutting out the silent parts is financially beneficial. An example of audio data run through VAD is shown below in Figure 2.

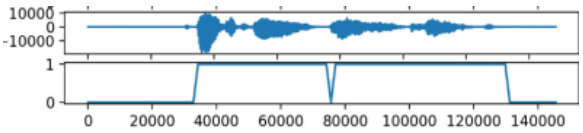


Figure 2. Top Graph: Audio File; Bottom Graph: Audio Detection (1 for Audio and 0 for No Audio).

While the current approach shows promise, practical implementation demands a more human-like chatbot such as GPT-4. The Bard chatbot provided us with an incredible start, but its frequently lengthy answers make it difficult to optimize both authenticity and efficiency. Furthermore, adhering to European regulations, such as GDPR (General Data Protection Regulation), restricts the use of Google Cloud's Speech-to-Text and Text-to-Speech software in hospitals. Such a setup would breach GDPR guidelines, as patient data cannot be transmitted outside the hospital premises.

Despite these limitations, this project served as an important proof of concept for such a chatbot. There is still much work to be done, but we've provided a stepping stone for future work in this field. Some of this future work includes generating images from text explanations, creating moving (and speaking) avatars from images, and training models to generate speech in anyone's voice. We are only beginning to explore AI's role in our future.

Circuit Level Design and Simulation Tool of DNA Strand Displacement Computational Circuits

Jesse Lerner
Advised under Prof. Adi Teman and PhD students Noa Edri and Roman Golman

Computer researchers are hitting physical limits when it comes to fitting more transistors onto smaller chips, which means Moore's Law is starting to fail. We have been reliant on Moore's Law for the advancement of computer speeds over time for the past 40-50 years, but as it fails we need to start looking towards other avenues to advance computer speeds. One such avenue of research is the use of a form of molecular computation, where computation is done using

reactions between physical particles rather than through electrical signals. The particular path of molecular computation that we are focusing on is DNA Strand Displacement (DSD) reactions.

The custom DNA used in DSD is prohibitively expensive, and therefore makes research into the subject difficult. The DNA is not so expensive that it is impossible to obtain, but doing consistent experiments becomes a large burden on a lab. This, along with the fact that it is easy to make small mistakes on experiments of this nature and that the DNA is not reusable, makes it hard to do research on strand displacement. The goal of our research group is to build a tool, currently called the DNA Simulator, to make this process easier and less expensive.

The goal of the DNA Simulator is to enable researchers to design and test large circuits in a free, virtual environment before spending money to test on real DNA. This will provide them with more experience, thus minimizing mistakes in procedure because they know which steps are necessary, and will also enable them to predict the outcome of the experiment according to current hypotheses. This makes testing and updating knowledge about new hypotheses easier as it can be compared to a virtual experiment run on a tool built by using past results. This tool is also open-source so that any researcher can easily download the source code and modify it for their own uses, or even suggest and implement changes based on their own experiments and experience with DSD.

Our tool is coded in Python and uses something called a "Bucket Model" (Figure 1). In the bucket model, we view concentrations of our reactants (individual and paired strands of DNA) as buckets, calculate the rate the buckets are flowing into each other, and record the amounts at each step. For example, if we know A and B are reacting to form C and D at a certain rate X, then we would remove timestep*X from A and B and add timestep*X to C and D. Most DSD reactions can be simulated like this, but this only covers one single reaction. Our goal is to scale this so that researchers working with complicated networks don't have to look at every single individual reaction, as that would be intractable to calculate by hand.

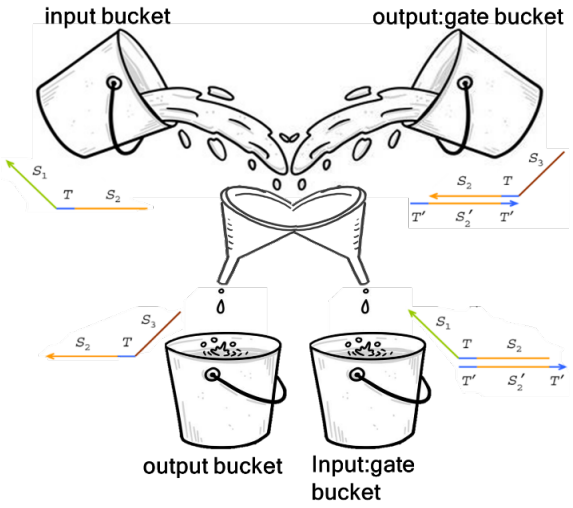


Figure 1. Visualization of the Bucket Model.

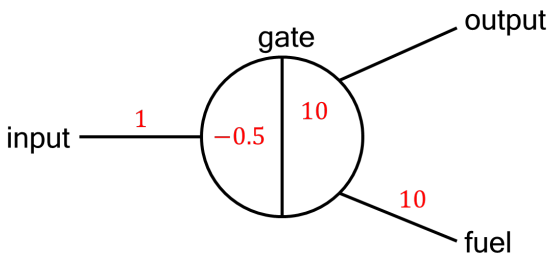


Figure 2. The circuit symbol for a simple Thresholding Seesaw.

We use the Bucket Model to generalize the reactions, so that we can use these reactions in gates. Gates are logical elements that perform basic operations. Commonly known gates that are the building blocks of modern computers are gates such as the AND gate and the OR gate, but in DSD there are simpler gates that make up those gates called Seesaws. A Thresholding seesaw (see Figure 2) takes an input and only produces an output if the input is above a certain amount. This is necessary in DSD environments because signals are not digital and reactants do

not react completely, instead reaching dynamic equilibrium and staying there. A threshold can be used to re-digitize the signal (that is, to turn it back into a 0 or 1 depending on whether it is above the threshold or not). It does this via a thresholding strand which consumes the input at a faster rate than the gate does, preventing it from activating the gate and producing an output. If the threshold gets consumed, then the remaining input will activate the gate and produce the output. The fuel present in the system will make sure the reaction continues to completion by making the input reusable. It does this by removing it from the input-gate combined strand and taking its place, allowing the input to react with a new output-gate compound and forcing it to release its output (Figure 3).

In addition to simple logic gates like the AND gate and OR gate, and seesaws like the Thresholding Seesaw and Integrating (adding) Seesaw, our tool now also has a Delay gate and NOT gate, both of which are newer additions. The lack of a NOT gate was one of the main questions and roadblocks in DSD research, and we are still developing this tool. We hope that it can become a very useful resource in DSD research, and that we will see many advances as DSD research become more approachable and easier to understand.

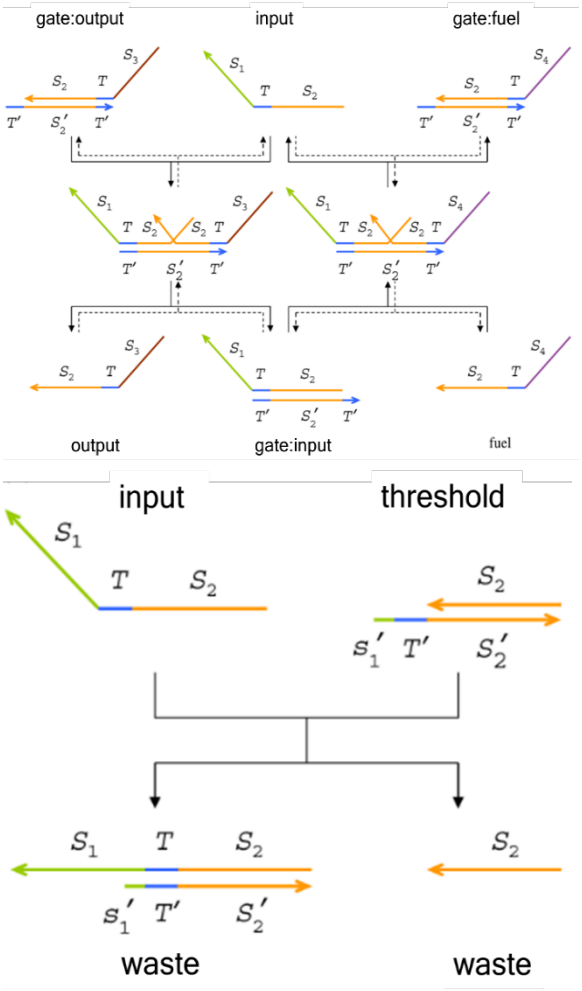


Figure 3. Reaction map of a simple Thresholding Seesaw.

LIFE SCIENCES

Gittel Levin, Meira Steiner, Joshua Brafman, Noa de Louya, Sivan Mussaffi

The Gene GPS: Guide RNA Competition for a CRISPR Gene Therapy

Gittel Levin
Advised under Dr. Ayal Hendel and PhD student Charles Krul

Severe combined immunodeficiency (SCID), colloquially known as Bubble Boy Disease, describes a set of rare genetic diseases where the immune system cannot fight off infections most people beat with a simple cold. My research focused on a type of SCID called Bare Lymphocyte Syndrome (BLS), where a mutation prevents the immune system from recognizing infections. BLS is prevalent in some ethnic communities, like the Bedouin community in Israel. BLS occurs when Antigen Presenting cells (APCs) are missing a receptor on their outer membrane called Major Histocompatibility Complex Class II (MHC II). When there is an infection, MHC II presents an antigen (a protein particle from a foreign object) to a T cell that alerts the immune system. Since MHC II is missing in BLS patients, their immune system does not recognize infection. The gene coding for MHC II is on chromosome six in humans. The gene contains a promoter region, the X Box, that needs activation for gene expression. The X Box is activated when an enzyme complex called RFX binds to it. RFX contains three components, RX5,

RFXAP, and RFXANK. A mutation on any of the genes for the subunits causes the RFX complex not to form. Therefore, X Box activation does not occur, MHC II is not transcribed, and the receptor is missing from APCs, resulting in BLS.

The current treatment for BLS and other SCID diseases is a bone marrow transplant. However, finding a match donor can be difficult. Furthermore, a transplant from an incomplete donor match might result in adverse effects like Graft v. Host disease and transplant rejection. With genetic engineering, patient stem cells are edited with the correct gene outside the patient and placed back into the patient to reconstitute a new healthy immune system. Current gene editing uses a recognition complex to find the target gene and a nuclease to create a double-stranded break (DSB) at the site. Once the complex induces a breakage, there are two repair pathways:

1. Non-homologous end joining (NHEJ) occurs when the cell repair mechanism deletes a few nucleotides and reconnects the two ends.
2. Homology-directed recombination (HDR) occurs when a provided donor gene is inserted into the break site.

Since 2012, when Jennifer Doudna and Emmanuelle Charpentier proposed CRISPR Cas9 for genetic engineering, for which they won the Nobel Prize in Chemistry in 2020, laboratories worldwide have implemented CRISPR for creating novel gene therapies. CRISPR uses a complementary guide RNA (gRNA) to locate the target site. Once the gRNA locates the target, Cas9 induces a DSB allowing NHEJ or HDR to occur. This experiment tested five different guide RNAs to find the most efficient one to target mutations in the RFX5 gene.

Methodology

After extensive research about the best locations to place the gRNA, five gRNAs were inserted into a px330 plasmid. The positive control was a plasmid called NP108, containing a guide for the gene IL2RG, which also causes SCID. This guide previously displayed high levels of CRISPR activity. The plasmids were introduced into human cells using electroporation technology. Electroporation is a process where cells receive an electric shock, weakening the cell membrane, and allowing the plasmid to enter the cells. Once the cells recovered, PCR was conducted to isolate and amplify the RFX5 gene in the cells. Gel electrophoresis confirmed PCR success. The DNA products from the PCR were sequenced and underwent TIDE analysis.

When the guides work, an insertion or deletion (indels) appear at the break site. TIDE measures the frequency of indels compared to the control and translates it into an efficiency percentage. Efficiency refers to variance in the sequence attributable to Cas9-guide activity.

Results & Discussion

Four rounds of electroporation were performed on the five guides. TIDE analysis for Round 1 (R1) guides 2, 3, and 4 displayed efficiencies of 29%, 36 and 40 (Figure 1). Normally, those percentages are a good indicator of activity. However, R1 was performed without NP108 and therefore there is no standard to judge the efficiencies against. Additionally, TIDE was not performed on R1 guide 1 (G1) due to poor sequencing. Compared to R1, R2 displayed less activity. However, most of the guides' efficiencies were in close proximity to NP108's efficiency at 18%, with G4 at 16% and G1 surpassing with an efficiency of 25%. Results from R3 and R4 could help determine whether G1 or G4 is a more efficient guide, but there was inadequate time to collect R3 and R4 data.

Conclusion

More data from R3 and R4 is needed to confirm which guide RNA is the most efficient. The next steps are designing homology arms for the donor DNA, and deciding on a vector to introduce the donor DNA to the cells.

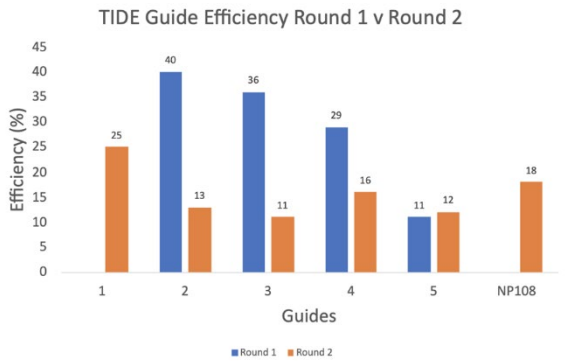


Figure 1. TIDE Analysis performed on Round 1 and Round 2 guide activity in HEK 293 cells.

[1] Reith, W., & Mach, B. Annual review of immunology, **19**, 331–373. (2001).
[2] Porteus M. H. The New England journal of medicine, **380** (**10**), 947–959. (2019).

Train-test Leakage in CRISPR Off-target Prediction Models

Joshua Brafman
Advised under Prof. Yaron Orenstein

CRISPR/Cas9 is a widely used gene-editing technique, but the occurrence of unexpected off-target editing presents a significant limitation in its application within biological and clinical contexts. Consequently, researchers have developed and trained machine-learning models to predict when and where off-target editing occurs. However, existing studies have largely overlooked the potential impact of train-test leakage on the evaluation of these models.

Train-test leakage occurs when a machine-learning model is tested on data that closely resembles the data on which it was trained. In such cases, the model may achieve high performance during testing because it has simply "memorized" the characteristics that these data points share. For instance, imagine training a model to recognize cats by showing it images of felines that happen to have dark birthmarks near their eyes. If the model is then tested with a picture of a cat exhibiting a similar birthmark, it might accurately classify the image as depicting a cat. However, we cannot be certain whether the model has genuinely learned what a cat looks like; there's a possibility that it has merely "memorized" an association between dark birthmarks and cats. Therefore, if confronted with an image of a cat lacking a birthmark, the model may fail. In a similar vein, if there was leakage between the data used to train and test off-target prediction models, it is possible that they performed better than they should have during testing because they memorized the "leaked" characteristics. Addressing train-test leakage is therefore critical for evaluating model performance in real-world scenarios.

The purpose of this study is to determine whether there is correlation between train-test leakage and the performance of off-target prediction models, and to offer strategies to minimize this risk so a more realistic evaluation of these models can be performed. To analyze this issue, I partitioned a popular off-target dataset into training and tested these sets in 680 different ways. For each of these train-test splits, I calculated the leakage between the two sets using seven different metrics. These metrics included various ways of evaluating the similarity between the guide RNA sequences across the split, as well as calculating the percentage of target locations (regions in the genome where off-target editing may have occurred) in the test set that physically overlapped with a target location in the training set.

In addition to calculating the leakage, for each train-test split, I trained a Random-Forest,

Logistic-Regression, and XGBoost model (machine-learning types). I evaluated these models using AUROC, which is a common metric to evaluate binary classification, because the models were trained to classify between experimental off-targets (positives) and potential off-targets (negatives) that were found in the genome. Finally, I calculated the correlation between the train-test leakage and model performance for each leakage metric and machine learning model across the 680 different partitions.

The two metrics for calculating leakage over the training guide RNAs, which were used in a previous study, did not yield a significant correlation to model performance. The new leakage metrics, which I defined in this study, resulted in a statistically significant correlation between train-test leakage and model performance (Figure 1). Moreover, calculating leakage in terms of overlapping target coordinates yielded the highest correlation for all machine-learning models evaluated in this study.

Figure 1. Correlation with Performance by Leakage and Model Type.

This analysis indicates that eliminating target overlap is the most effective approach to reduce train-test leakage. Therefore, to address this issue I propose implementing chromosomal partitioning of the off-target sites. The off-target sites should be partitioned in a way that no target in the training set shares a chromosome with any target in the test set. This approach eliminates target overlap—if no pair of targets from the two sets appear on the same chromosome, then no pair of targets can overlap. My initial results from applying this method suggests that it is an effective way to reduce train-test leakage because both the average AUROC score of the models and the correlation between leakage and model performance decreased significantly. However, further analysis of these results is required and is still ongoing.

Congenital Dyserythropoietic Anemia type 1: Unraveling the Enigma of Codanin-1

Noa De Louya
Advised under Prof. Benny Motro

Background

Congenital Dyserythropoietic Anemia type 1 is a rare autosomal recessive disorder with macrocytic anemia characterized by pathognomonic abnormalities. The erythroid precursors in the bone marrow appear as spongy chromatin and occasionally chromatin bridges appear as well. CDA type 1 mutation was initially found in a Bedouin tribe which intermarried, resulting in this disease. Unfortunately, the mutated gene responsible for this phenomenon was yet to be found. Professor Hannah Tamary, from Schneider Hospital, quickly spiked an interest in this research, leading to the discovery of this intricate gene. Subsequent to multiple years of hard work, Prof. Tamary successfully pinpointed the gene, naming it Codanin-1. Following this major breakthrough, Prof. Tamary reached out to Professor Benny Motro, a highly experienced biologist with expertise in knocking out genes in mice, to collaborate in the hopes of finding the mechanism, location, and purpose of

this gene. Currently, extraordinary advancements have been made by Prof. Motro. CDAN1 was found to be lethal as a homozygous null type mutation, indicating that this gene may be essential for human biological function. His previous research proves that the gene is indispensable for mouse embryonic development even before erythrocyte formation through cell Knockout experiments. Additionally, the experiments demonstrated the widespread presence of Codanin-1 in different mouse cell lines. These findings opened a broader avenue for investigation that is not limited to CDA type 1 disease, but for the infinite biological processes Codanin-1 may be involved with. Focused research on Codanin-1 has suggested a possible relationship between the encoded protein and DNA damage repair system. This conserved protein also serves as a scaffold for C15Orf41 (a nuclease) and ASF1 (chaperone protein), which are involved in DNA damage control, which may reveal other crucial functions of the gene [1].

This research project aimed to elucidate the possible processes or behaviors of Codanin-1 by understanding its localization in the cell and its involvement in DNA damage repair. To achieve this objective, it is beneficial to leverage existing knowledge about ASF1's function. This chaperone protein is involved in DNA damage response as well as chromatin stability and organization, which can prompt inquiries about Codanin-1's influence on these processes. The experiment was planned as follows using the Western-blot technique. Firstly, the HeLa cells were exposed at different time intervals to UV light in order to induce DNA damage. (Table 1).

0h	2h	4h
8h	10h	12h

Table 1. Cell plates exposed to UV light at respective times.

To knock out Codanin-1 in 7 out of the 14 plates, the auxin signaling system was activated. Essentially, using the CRISPR/Cas9 system, mAID (degron) was inserted into the end of the target endogenous gene, CDAN1. Once codanin-mAID is produced, and the hormone auxin is bound to its receptor, it attaches a small molecule called ubiquitin to the degron molecule. The ubiquitin is recognized by E3 ligase complex and is sent to the proteasome for rapid degradation. Codanin-1 stabilizes C15Orf41, thus Codanin-1 degradation results in C15Orf41 elimination as well. Subsequent to protein extraction, the Western blot experiment was performed. The results are shown in Figure 1 below.

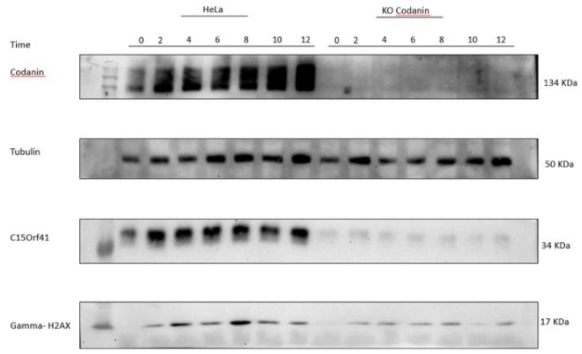


Figure 1. The influence of Codanin-1 degradation on DNA damage response following UV irradiation.

When Codanin-1 is not present in the nucleus as seen in the section titled "KO Codanin" (Figure 1), there is less DNA damage over time compared to untreated cells. Both Codanin-1 and C15Orf41 do not appear after the auxin signaling system was activated. Tubulin is used as a control and Gamma-H2AX is a DNA damage marker. Codanin-1's presence in the nucleus correlates with higher DNA damage, suggesting faster repair rates, which can more likely induce mutations and cellular apoptosis. In the absence of Codanin-1, DNA damage repair is slower, hinting at its regulatory role in this process.

After understanding that Codanin-1 degradation affects the DNA damage response system, investigating the localization of Codanin-1 and C15Orf41 subsequence to UV treatment seemed interesting. Furthermore, we wanted to test if the positive correlation between the two proteins is reflected as similar localization. To

test for the effects of UV on the location of the proteins, the immunofluorescence technique was used. After preparing the slides, two with UV treatment and one with no treatment, the results were visualized using a fluorescence microscope as seen in Figure 2 below.

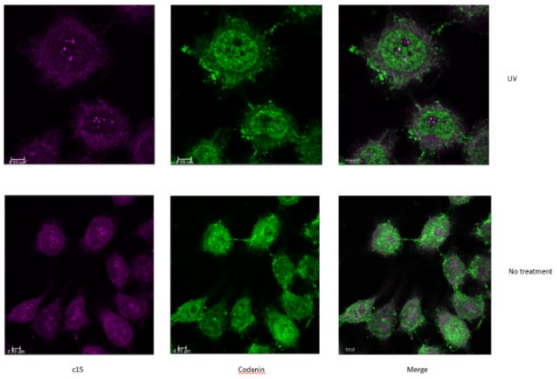


Figure 2. Transducing cells with C15Orf41 and Codanin-1 determining their location after exposure to UV light.

C15Orf41, in green, is primarily found in the nucleolus. After DNA damage is induced by exposure to UV light, C15Orf41 aggregates together, while the untreated cells seem to have the protein dispersed throughout the nucleus and the cell. Codanin-1 is located throughout the cells and exposure to UV does not cause a significant change in its location. It can be concluded that DNA damage affects the localization of C15Orf41 in the cell. Its aggregation in the center may suggest that it participates in DNA damage repair. However, since Codanin-1 does not seem affected, C15 and Codanin-1 may also have distinct functions. Further investigations should be performed to confirm these results.

Another investigation, currently in progress, explored the phenotypic effects of Codanin-1. Codanin-1 might play a crucial role in maintaining testes function, particularly in DNA damage response. Our interest stems from prior studies involving knockout mice suggesting that heterozygous Codanin-1 knockouts have smaller testes, further hinting at the gene's significance in testicular biology. To execute the experiment, the gene needs to be knocked out using the Cre-lox system. A PCR gel was run to find out which mice could be candidates. Firstly, these mice should be carriers for AMH-cre (Anti-Mullerian Hormone), and it should be expressed only in Sertoli cells. Mice homozygotes for Codanin-1-Floxed and carriers for AMH-cre recombinase will recognise the loxP leading to elimination of Codanin in Sertoli cells. It's evident from Figure 3 that mice 574, 589, 590, 591, 592, and 593 are carriers of AMH-cre on the right and mouse 591 is the only one homozygote for CodF on the left.

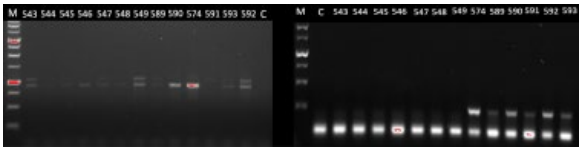


Figure 3. PCR genotyping of mice for AMH-Cre and for Codanin-floxed.

In conclusion, this research project brought us one step closer to unraveling the intricate functions of Codanin-1. These findings highlight Codanin-1's multifaceted nature, ranging from DNA damage repair to potential involvement in testicular biology. This research not only sheds light on the mechanisms underlying specific diseases but also opens up avenues for investigating broader biological processes where Codanin-1 may play a pivotal role. Further research and exploration are required to fully comprehend the complexities of Codanin-1 and its contributions to diverse physiological processes.

[1] Swickley, G. et al. BMC Molecular and Cell Biology, **21**, 18 (2020).

Truncated SIRT6 and the Effects of Mutated CBS in Mice

Sivan Mussafi and Michelle Steiner
Advised under Prof. Haim Cohen and PhD student Noga Touitou

The science of aging and longevity has recently become a growing area of study. As old age becomes an increasingly prevalent cause of mortality, it has become more obvious that the metabolic processes and genomics that relate to aging are largely unknown. A genetically conserved family of proteins known as Sirtuins are known to have a profound effect on metabolic regulation. SIRT6 is one of the seven proteins (SIRT1-SIRT7) in this family found to prevent metabolic effects of aging. An ADP-ribosyl transferase and NAD^+-dependent deacetylase specific for H3K9 (histone 3, lysine 9) and H3K56 (histone 3, lysine 56), SIRT6 regulates genome stability, DNA double strand break repair, telomere integrity, and gene transcription through the use of the transcription factor Sp1. The overexpression of SIRT6 in B6 male and female mice both preserved glucose homeostasis and hepatic glucose output in their old age, repressed glycolysis, and activated pathways such as AMP kinase [1]. All of these effects helped maintain homeostasis in the aging mice by promoting normoglycemia. In the liver specifically, SIRT6 was found to repress glycolysis, triglyceride synthesis, and increase beta oxidation. Structurally, SIRT6 is composed of a core domain of 276 amino acids. The C terminal is mainly responsible for nuclear localizations where the N terminal is the primary site of the catalytic activity. While the overall location of catalytic activity in SIRT6 is known, little research has gone into pinpointing the Sp1 binding site of SIRT6 [2].

Gaseous H_2S has also been found to have a positive effect on health and aging by preventing hypertension and protecting against neurodegeneration associated with Huntington's disease, atherosclerosis, and type 1 diabetes [3]. A major product of the transsulfuration (TSS) pathway, H_2S is generated by CBS and CGL, enzymes in the TSS pathway. This pathway is present in both humans and mice. Our lab demonstrated that the mouse CBS protein is acetylated on K386 (lysine 386). Further investigation revealed that this acetylation regulates CBS's H_2S production activity. However, in human CBS, this lysine residue is substituted with an arginine, which is traditionally used in molecular biology for mimicking constant deacetylation. Thus, this region in human CBS resembles a constant state of deacetylation, whereas in mice this region is variable. Little is known about the effects of the differing residues in human and mice CBS, and consequently how the acetylation/deacetylation of CBS affects metabolic activity.

Our goal was twofold. First, to determine the effects of acetylation on CBS activity, H3K9 acetylation, and CGL concentration; second, to successfully truncate SIRT6 in order to determine the location of the Sp1 binding site in the future. In order to test the effects of acetylated CBS, we used mutant BQ mice, in which K386, normally the variable region of acetylation, was replaced with glutamine to mimic constant acetylation. The liver samples of 5 male and 5 female transgenic mice were extracted along with a corresponding number of wild type mice of

each gender. The livers were homogenized and lysed. The samples were then standardized using Ponceau and a western blot was repeatedly used to visualize the protein levels of CBS, CGL, and acetylated H3K9 in each mouse. ImageJ was used to assess the relative concentrations of each sample, and T-tests were performed to determine the significance of the results.

In order to create the truncated SIRT6 along with the proper vectors, primers specific for the N-Terminal, C-Terminal, and core region of SIRT6 were used to create the proper inserts. A PCR was performed in order to truncate and amplify the regions of SIRT6. DNA cloning was used to fuse a pcDNA vector with the SIRT6 segments, using restriction enzymes BamHI and EcoRI to create a plasmid which can be used in the future to determine the Sp1 binding site. The truncations were sequenced and analyzed using BLAST Sequencing in order to obtain the most viable sequences.

To determine the concentrations of the proteins analyzed in each western blot, the protein levels of each sample, obtained through assessing the image of each nitrocellulose membrane using ImageJ, was divided by the general protein levels indicated by the ponceau solution and obtained through imageJ. The results of the western blots for the various proteins are indicated in Figure 1 for males and Figure 2 for females. It is noteworthy that the western blot used to determine H3K9 acetylation in the male mice (Figure 1) did not yield results. The western blot was performed twice, but in both instances the nitrocellulose membrane appeared randomly stained and splotchy. It is possible there may have been an issue with the antibodies themselves. The averages and standard deviations of each of these values was also calculated. T-tests were used to assess the significance of the results by comparing between each of the groups. However, no statistically significant results were obtained. These results indicate that the acetylation of CBS does not change its expression and stability, meaning that future phenotypes seen in those mice are a result of the mutation alone. These results indicated the effect of the Q mutant of CBS in mice. The next step would be to engineer mice that possess an arginine (R) residue in place of lysine, mimicking constant deacetylation, and to then compare the physiological and molecular differences between the two mutants (Q and R).

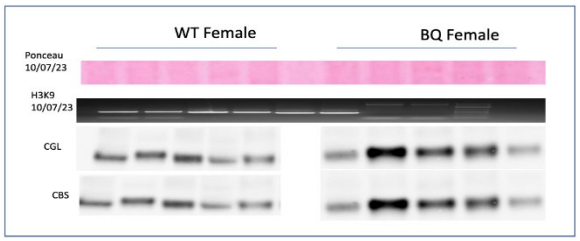


Figure 1. Western Blot Results for WT and BQ Female Mice.

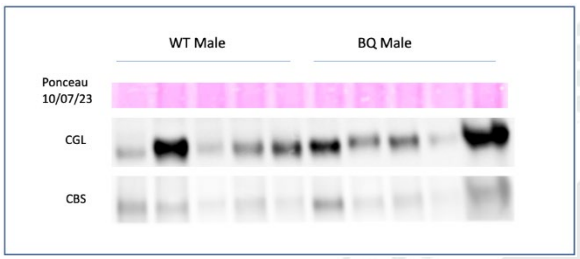


Figure 2. Western Blot Results for WT and BQ Male Mice.

In our second project, the cloning of the truncated forms of the human SIRT6 to a pcDNA3.1+ expression vector, three separate PCRs were performed at slightly different annealing temperatures (55, 57, 59

degrees celsius). The results of the initial truncation are depicted in Figure 3. All the reactions showed an amplification of the aforementioned desired regions. Once the cloning was performed, the samples were again assessed to determine the success of the vector insertion. All samples (Figure 4) indicated success. Separate samples of each of the truncated segments of SIRT6 were isolated and sequenced in order to obtain the sample with the correct sequence for future use. DC4, DN3, and Core1 had the best sequences and were set aside for future use, namely to determine the Sp1 binding site.

The relative concentrations were obtained for each of the tested values for the transgenic female mice, but results were inconclusive for H3K9 in male transgenic mice. Ultimately, acetylated CBS does not have a significant effect on CGL production of H3K9 acetylation. However, these results will be used in future procedures comparing mutant BQ mice (acetylated) to transgenic R mice (deacetylated). The cloning for truncated SIRT6 was successful and the plasmids will later be used in order to further the objective of determining the Sp1 binding site.

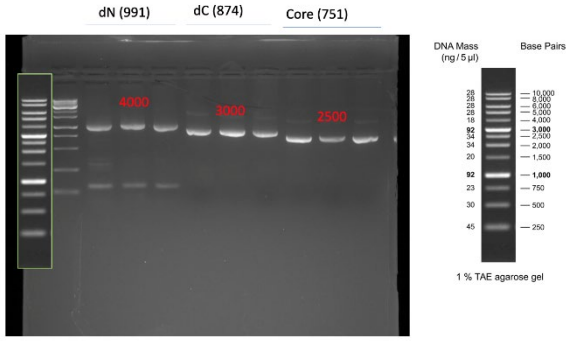


Figure 3. PCR Results for Truncated SIRT6.

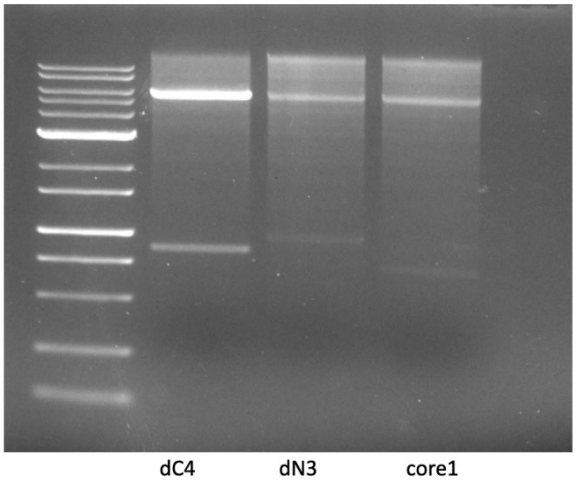


Figure 4. PCR Results for Truncated SIRT6 with Vector.

[1] Roichman A et al., Nature Communications, **12**, 3208 (2021).
[2] Tennen RI et al., Mechanisms of Ageing and Development, **131(3),** 185-192 (2010).
[3] Hine C and Mitchell JR, Experimental Gerontology, **68**, 26-32 (2015).

PHYSICS, CHEMISTRY, AND MATHEMATICS

Yisrael Weiner, Ned Krasnopolsky, Ezra Goldfarb
Bracha Weinberger, Leora Kronenberg, Maya Rubenstein, Leah Baron (not in photo)

Stealthy Hyperuniform Lasers

Bracha Weinberger
Advised under Prof. Patrick Sebbah and Post-Doc Aswathy Sundaresan

Hyperuniformity refers to a set of points or objects in a specific form that exhibit a high degree of order and regularity. Disordered hyperuniform systems exhibit both disorder and hyperuniformity simultaneously. In such systems, on average, the density fluctuations within a certain window size are suppressed, leading to a more uniform distribution of points or objects, despite the inherent disorder in their arrangement. A defining feature of a disordered hyperuniform system is that the structure factor $S(k)$ is precisely zero as the wave number $|k|$ approaches zero [1]. An important subclass is the stealthy disordered hyperuniform system in which $S(k) = 0$ for a finite range of wavenumbers $0 < k \leq K$; this range defines the exclusion region in Fourier space. The degree of stealthiness χ is the ratio of the number of the constrained wavevectors in the reciprocal space to the total number of degrees of freedom. i.e., in one dimension, $\chi = K/(2\pi\rho)$, where ρ is the number density of particles [1].

The objective is to understand stealthy hyperuniform systems, design and fabricate

stealthy hyperuniform lasers in one dimension and investigate their spectral response for different values of stealthiness χ. The stealthy hyperuniform lasing structures were developed by generating a point pattern with 100 points for a given stealthy parameter. This was performed algorithmically by starting with a Poissonian point-distribution, then adding a small random displacement to a point chosen at random; the change is kept so long as it increases the stealthy parameter, and the process is repeated until the desired χ is reached. Each point is replaced with rectangles of 350 nm width and 50 µm height using the software Lumerical FDTD.

The Elionix ELS-G100 e-beam lithography system is used to carve the pattern of the structure. A fused silica wafer (15 mm x 15 mm) is used as the substrate with 5% DCM doped PMMA A6 495 as the resist. The resultant Stealthy hyperuniform lasers consist of specifically positioned air grooves of 350 nm width and 50 µm height.

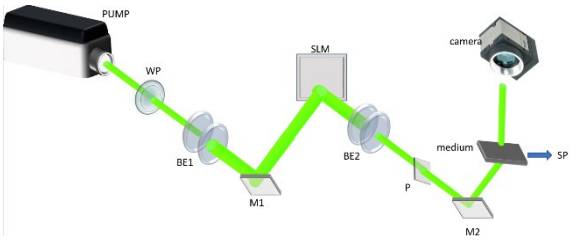


Figure 1. Experimental Set-up [2].

As seen in Figure 1, the input laser beam from a picosecond laser (10 Hz, 100 ps) is passed through a half-wave plate (WP) which rotates the polarization direction of the light. Then it passes through a beam expander (BE1) and then reflected off the first mirror (M1) to a spatial light modulator (SLM) which modulates the intensity of the pump based on the given profile. Then it passes through a second beam expander which adjusts the magnification of the pump profile (BE2) reflected off the SLM and a polarizer (P) before it is finally reflected off the second mirror (M2) and travels normal to the sample. A fiber coupled spectrometer (SP) collects the in-plane emission from the sample. A CCD camera is employed to image the structure from above.

Stealthy hyperuniform structures are designed in one dimension with varying stealthiness, and the structures are fabricated via e-beam lithography. It is understood that with the increment in the stealthiness, χ, from 0.1-0.6, the system approaches a more uniform arrangement. It is seen that when χ reaches above 0.4 the system is no longer disordered, rather, it reverts back to being periodic. Hence χ = 0.4 is the transitional value in one-dimensional stealthy hyperuniform systems from disordered to ordered arrangement.

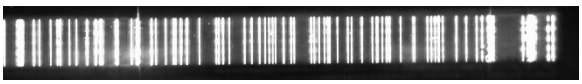


Figure 2. χ = 0.1

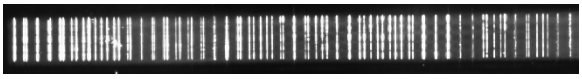


Figure 3. χ = 0.4

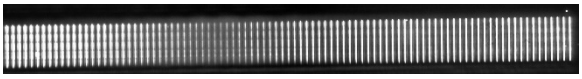


Figure 4. χ = 0.6

In conclusion, we designed and fabricated the stealthy hyperuniform systems in one dimension, and found that with an increment in stealthiness (χ), the system becomes more ordered. An experimental investigation on the stealthy hyperuniform lasers with different values of stealthiness is performed to analyze the spectral response of the systems. Further exploration of one dimensional hyperuniform systems can be performed regarding the mode selection of the stealthy hyperuniform systems.

[1] J. Kim, and S. Torquato, Optica, **10**, 965 (2023).
[2] B. Kumar et al., Optica, **8**, 1034 (2021).

Synthesis and Performance of Platinum Nanoparticles on Carbon Black for Oxygen Reduction Reaction in Fuel Cells

Ezra Goldfarb
Advised under Prof. Lior Elbaz and MSc student Yeela Persky

The rapid growth of the global population and advances in civilization have resulted in an exponential growth in energy demand, and unsustainable fuels like fossil fuels account for about 70% of total energy consumption. Hydrogen Fuel Cells, a viable alternative fuel source that is both renewable and efficient, are a promising alternative energy source. A significant limitation of this technology is the sluggish Oxygen Reduction Reaction (ORR), which needs to be catalyzed by precious metals. This project focuses on incorporating Platinum Nanoparticles (Pt-NPs) set in Carbon Black to act as a catalyst for the Oxygen Reduction Reaction and to increase the efficiency of the fuel cell system.

The synthesis of the catalyst used an Incipient Wetness Impregnation method. 72XC carbon and chloroplatinic acid were dissolved in acetone. The solution was dried in an oven at 80°C with a 150 ml/min flow rate for twelve hours. This step used H_2 to reduce the salt and leaves behind Pt-NPs in the carbon. This synthesis method and proportions had an expected yield of 50mg and a 20%wt of Pt. With synthesis finished, the catalyst is hereinafter examined for quality. Multiple methods are used to confirm different qualities of the synthesized catalyst. Commercial Pt-NPs in Carbon is then used in a fuel cell to examine catalyst performance in ORR.

XRD: XRD is an analytical technique that can measure the crystallographic structure of a sample. Incident X-rays are beamed towards the sample at various angles and resulting intensities and scattering angles are observed according to Bragg's Law.

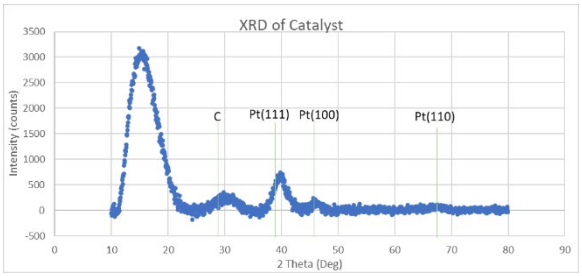


Figure 1. XRD of synthesized Pt-NPs catalyst.

The above graph shows the results from an XRD of the sample. The three peaks at 40°, 46°, and 68° match the Pt(111), Pt(100), and Pt(110) peaks, respectively. The peak at 30° matches the carbon support (XC-72) peak. The peak at 15° may match Pt-NP scattering. This resembles literature Pt on C with a Pt FCC structure.

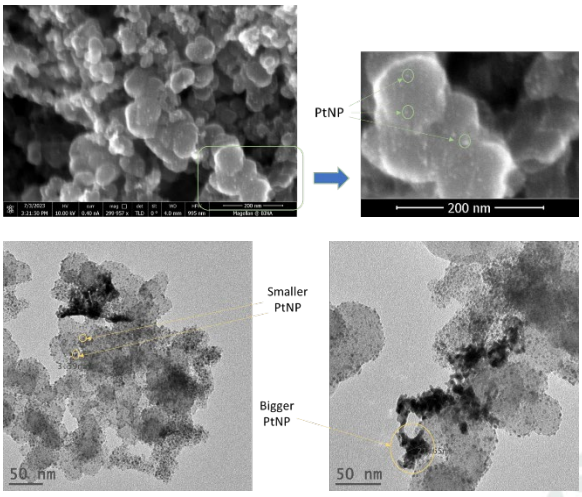


Figure 2. Images obtained from HR-SEM (top) and TEM (bottom).

Microscopy: The top images show results from a High Resolution – Scanning Electron Microscopy imaging of the sample. Incident electrons are focused onto the sample, and X-rays and backscattered electrons are examined to create an image. Small dots in the HR-SEM image show Pt-NPs. A small section is enlarged on the right to more clearly show the Pt-NPs on carbon. Pt-NPs ranged from 4 to 12 nm. EDAX indicated a 40%wt Pt composition. The lower image shows results

from a Transmission Electron Microscopy imaging of the sample. In TEM, incident electrons pass through an ultrathin sample to create an image. Both images are taken at a higher magnification than those taken by HR-SEM. Black dots in the TEM image show Pt-NPs. Both images show many distributed Pt-NPs of various sizes. In the image on the left, two specific Pt-NP clusters are measured as having lengths of ~2 and ~4 nm. In the right image, one Pt-NP cluster is measured with a length of ~22 nm. In both HR-SEM and TEM, the Pt-NPs are visible as specks on bigger Carbon structures, as expected. As shown in the TEM, the Pt-NPs clumped in areas, thereby reducing Pt-NPs' electrochemical surface area and resulting catalyst efficiency.

Electrochemistry: The following electrochemical experiments were done to verify the electron transfer of the desired catalyzed reaction and to determine to Electrochemical Surface Area (ECSA) of the catalyst.

Half Electrochemical Cell Measurements: The system was purged with Ar to remove dissolved O_2 from the solution and headspace. 40 scans were done at a scan rate of 100 mV/s.

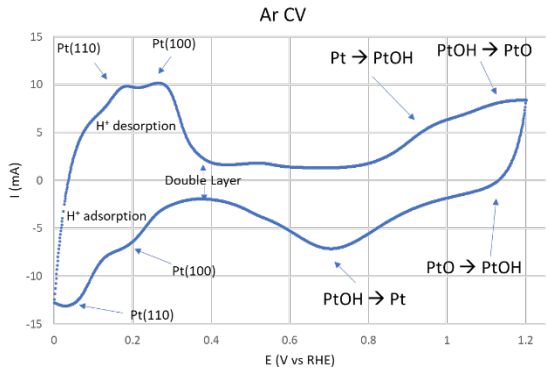


Figure 3. Voltammogram of CV of Ar.

Above is the voltammogram of the CV of Pt-NPs on C catalyst in a 0.5 M H_2SO_4 solution. On the left, the different facets of the Pt-NPs show their different hydrogen adsorption and desorption peaks. Multiple Pt oxidation and reduction peaks can be seen on the right.

RDE was done on the H_2SO_4 solution from the previous experiment. Runs were done with Ar to establish a baseline for O_2 runs. The data from the final cycle of O_2 was split into its oxidation and reduction sections and adjusted for the Ar baseline. The limiting Current is obtained from the average between the stable phase of either voltammogram.

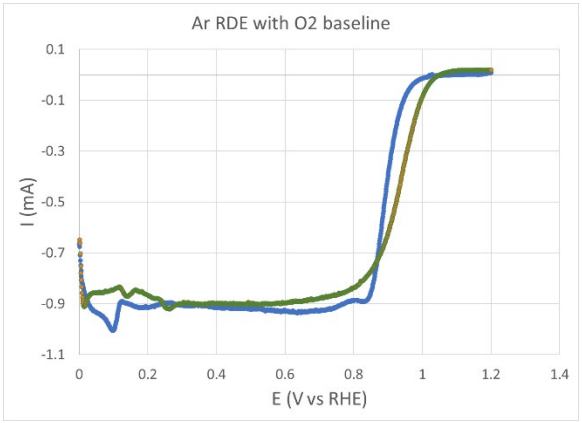


Figure 4. Voltammogram of RDE of O_2 with Ar baseline.

e⁻ and ECSA calculations: Various parameters and results from the aforementioned voltammograms are used in the following two equations-- ECSA and Levich equations, respectively-- to determine ECSA and electron transfer.

$$ECSA = \frac{|Q_H|}{210\mu C}(cm^2)$$

$$I_L = 0.620nFAD^{2/3}w^{1/2}v^{-1/6}C$$

Equation 1 and 2. ECSA and Levich equations.

The limiting current is used in the Levich equation to calculate the e^- transfer as ~3.7. This is within error of the identifying 4 e^- transfer of hydrolysis. Using integrated currents from the CV's, the ECSA can be calculated as ~21 m^2/g and ~31 m^2/g for the synthesized and commercial catalysts, respectively.

Fuel Cells Measurements: A fuel cell was prepared using similar commercial grade Pt-NPs

on XC72. O_2 and H_2 were purged through the system and CV was done with a range of 0.05 to 1.1 V and a scan rate of 100 mV/s. As shown below, peaks appear in the Fuel Cell CV at similar places to the Half Cell CV.

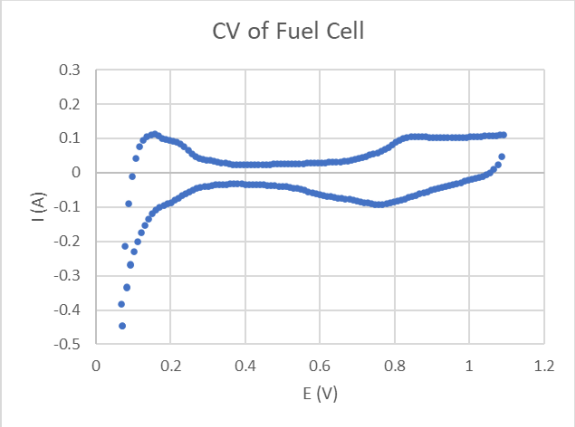


Figure 5. Voltammogram of CV of Fuel Cell.

Conclusion: The catalyst was prepared by an incipient wetness impregnation method and examined with the following methods to verify its identity and quality. The XRD showed each expected peak of the Pt-NP on Carbon catalyst. Additionally, HR-SEM and TEM both confirmed Pt-NP on the Carbon. From the EDAX at the HR-SEM, it seems that the synthesis produced a higher Pt % composition than intended. The CV further verified the individual H^+ adsorption and desorption peaks as well as the Pt oxidation and reduction peaks. The calculated e^- transfer of the reaction indicates H_2O formation as opposed to unwanted peroxides. The electrochemical surface area was also found to be within the expected range. Finally, a commercial version of the Pt-NP on XC72 catalyst was used in a fuel cell and examined for electrochemical surface area, H_2 crossover, and resulting power and voltage from current density. The electrochemical surface area was found to be ~150% compared to the synthesized catalyst. H_2 crossover was at a minimal level. Power and voltage versus current density graphs resembled the expected results. These results indicate the success of the synthesis and the effectiveness of the catalyst.

Pathways to Disordered Protein Dynamics: Biomolecular-NMR Analysis of WIP

Leora Kronenberg
Advised under Dr. Inbal Sher and Prof. Jordan Chill

Wiskott-Aldrich syndrome protein (WASp) is a 502-residue polypeptide that is expressed in hematopoietic cells and is responsible for cytoskeleton rearrangement when activated. WASp has an activity regulator, the WASp interacting protein (WIP), which is a 503-residue polypeptide and a member of the verprolin family of actin binding proteins [1]. While WIP is most known for the phosphorylation-dependent process that prevents WASp's degradation, there are multiple proteins within the cell that WIP is likely to interact with (Figure 1). However, because WIP is an intrinsically disordered protein (IDP), the reaction processes that WIP carries out with each protein are difficult to track and characterize.

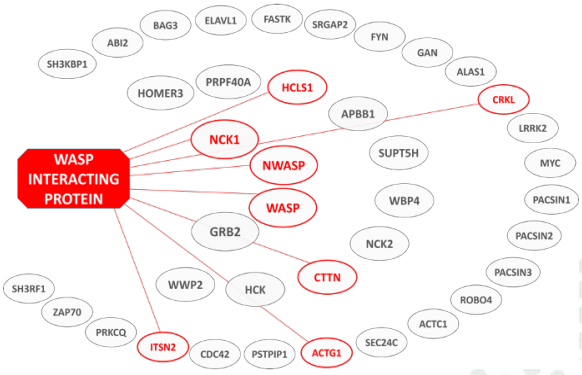
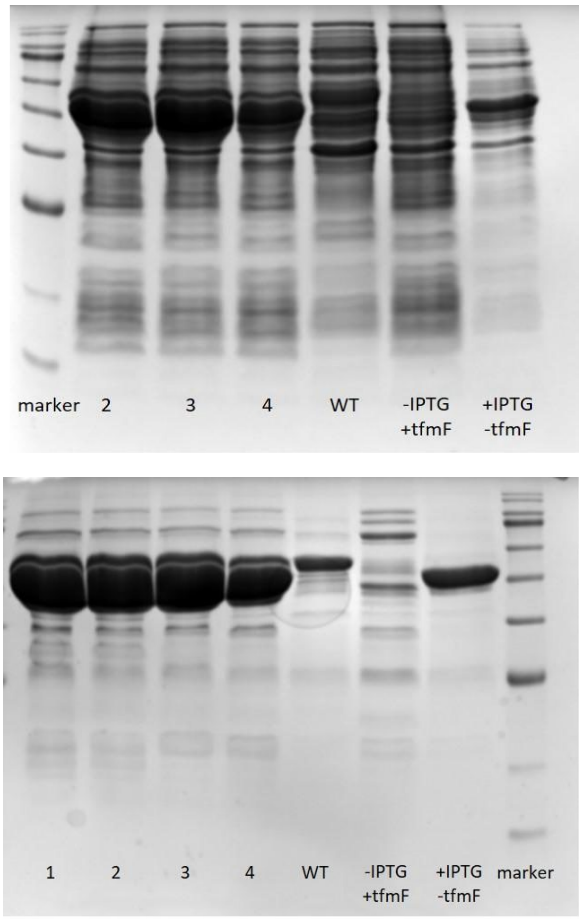


Figure 1. Interaction partners of WIP. The Human Integrated Protein-Protein Interaction rEference (HIPPIE) database indicates proteins with a good probability ($p \geq 0.5$) of interacting with WIP. Inner circle—$0.94 \leq p \leq 0.99$, middle circle—$0.68 \leq p \leq 0.86$, outer circle—$0.52 \leq p \leq 0.63$. [2]

The Chill laboratory aims to follow WIP behavior within its natural cellular habitat using nuclear magnetic resonance spectroscopy (NMR). This can be done by replacing an amino acid in WIP with a mutated version containing a fluorine nucleus and acquiring ^{19}F-NMR spectra. This

ensures the NMR is detecting WIP only and nothing else within the cell. The laboratory aims at understanding the multiple reactions of the protein itself and establishing methods for in-vivo protein analysis to better understand the complex interactions of IDPs in general. For this to succeed, a protocol to prepare ^{19}F-containing mutant WIP must be generated.

DNA that corresponds to the active site of WIP that typically interacts with WASp (residues 440-503) was isolated. The DNA was mutated using standard oligonucleotide-directed mutagenesis to create a plasmid that would code for a fluorinated phenylalanine instead of tyrosine at residue 455, valine at residue 469, or isoleucine at residue 501, respectively. The plasmid was transformed into DH5α E-Coli bacterial cells. A transformation was carried out in tandem with DNA that codes for modified tRNA that can interact with the fluorine-substituted phenylalanine (tfmF). The newly formed plasmids were isolated using a standard Miniprep kit and then co-transformed in BL21 *E. coli* cells. The cells were incubated and isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to induce over-expression. The cells were then lysed by sonication and the mutant protein was isolated using a selectively binding nickel column. To analyze purity and concentration of the isolated protein, SDS-PAGE gels were run of the supernatant before (Figure 2) and after (Figure 3) nickel column filtration. Four samples of lysed cells were used, as well as a cell containing a WIP wild type (WT) and control samples, one without IPTG and one without tfmF. The gels clearly show that (i) expression of our desired protein is induced by IPTG, and (ii) full-length expression is tfmF-dependent, since the mutated site is interpreted as a stop-codon without the fluorinated amino acid. Overall, these results prove that our efforts were successful and that ^{19}F-labeled WIP can be produced.



Figures 2 and 3. Supernatant (top) and elution (bottom) SDS-Page gel.

In preparation for mutant protein analysis, we also ordered a synthetic ^{19}F-labeled polypeptide (residues 440-503 of WIP, with residues 455 and 501 replaced by tfmF) and acquired its spectrum on the Bar Ilan 400 MHz NMR spectrometer (Figure 4). As can be seen on the spectra, sufficient signal from the two ^{19}F nuclei was obtained within a few minutes of measurement. Notably, the signal from both ^{19}F-containing amino acids is overlapping. This is likely due to both ^{19}F nuclei being in a similar chemical environment since the polypeptide is disordered and not folded. Upon interaction with cellular binding partners, we predict that the environments surrounding the ^{19}F nuclei would be different enough to exhibit different chemical shifts, allowing them to report on their interactions with other proteins.

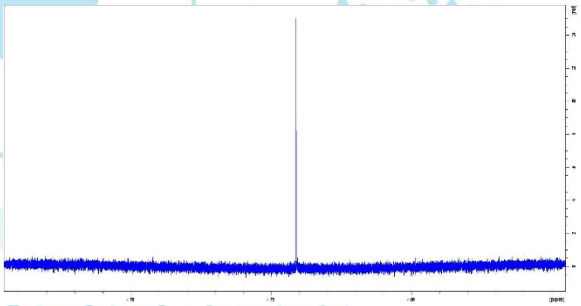


Figure 4. NMR Spectra of synthetic WIP.

With the successful growth and isolation of the fluorinated WIP mutant and the ^{19}F-NMR analysis of the synthetic version, the Chill Lab will attempt ^{19}F-NMR analysis on isolated WIP mutant, and then within a cell, to truly understand the dynamics and interactions of WIP and IDPs in general.

[1] Halle-Bikovsi, A. et al., ACS Chem. Biol., **13**, 100-109 (2017).
[2] Sokolik, C. G. et al., Biomolecules, **10** (7), 1084 (2020).

An Atomistic Model to Predict Raman Spectra

Maya Rubenstein
Advised under Prof. Ilya Grinberg and Dr. Atanu Paul

Raman spectroscopy is an essential tool in materials science used to understand the vibrational and structural properties of various systems. Raman spectra result from changes in vibrational frequency. When an incident photon collides with an electron, the energy of the electron may temporarily increase. The electron will then emit a photon to return to its ground state energy. While the electron will return to its original electronic energy level, it may transition to a different vibrational frequency. This change in vibration results in a difference in energy between the absorbed and emitted photons. Various energy differences from various vibrational transitions are displayed in a Raman spectrum [1].

Only vibrational modes in which the molecule's polarizability changes are Raman active. Polarizability is the tendency of a molecule's electron cloud to be distorted by an external electric field. Electric fields exert opposite charges on the positively charged nuclei and negatively charged electrons, inducing a dipole. The ease in which this dipole is induced is the molecule's polarizability. For a vibration to be Raman active, the molecule's polarizability must change as the spatial coordinates of the atoms move throughout the vibration [1]. The intensity of a Raman peak is proportional to the Fourier transform of the autocorrelation function of the time derivative of electronic polarizability.

$$I_{||}(\omega) \propto \frac{(\omega_{in}-\omega)^4}{\omega}\frac{1}{1-exp(-\frac{\hbar\omega}{k_B T})}\int \langle \dot{\alpha_{xx}}(\tau)\dot{\alpha_{xx}}(t+\tau)\rangle_\tau \, e^{-iwt}\,dt$$

Equation 1. Raman intensity is proportional to the Fourier Transform of the autocorrelation function of the time derivative of polarizability. $I_{||}$ is the parallel component of the intensity, a_{xx} is the xx element of the polarizability tensor, ω_{in} is the frequency of the incoming light, and ω is the frequency of the outgoing light. The perpendicular component of the intensity is analogous with a_{xy} instead of a_{xx} [2].

Because calculating Raman spectra requires the time derivative of polarizability, we must obtain the evolution of polarizability over a long time with small time steps. Density functional theory (DFT) provides a robust framework for modeling molecular vibration and polarizability. However, DFT is limited to small numbers of molecules (<40 molecules) over short time scales (<0.1 nanosec) because calculation of polarizability using DFT is computationally expensive. Currently, it is impossible to use DFT to calculate polarizability for large systems [3].

To address this issue, our group has developed an atomistic model to calculate the polarizability for predicting Raman spectra. In the model, we assumed that the total polarizability of a given system is the sum of polarizability from each

bond. The bond polarizability can be expressed using only a few Lorentzian functions. The Lorentzian functions are functions of the atomic coordinates, not the individual electrons. The bond polarizability term contains 6 to 10 constant parameters which depend on the system. These unknown parameters in the model are determined using an optimization procedure with respect to the polarizability calculated from DFT. This atomistic model is significantly more computationally efficient than DFT, and thus can handle larger systems and time scales (millions of atoms over several nanoseconds) [3].

We applied this model to 4 molecules: Cl_2, H_2S, NH_3, and SO_2. We first calculated the time dependent trajectory of each molecule using molecular dynamics at 300 K from DFT. This provided the position of each atom in the molecule at time steps of 0.0015 picoseconds for 30 picoseconds. We then calculated polarizability using DFT for 50 data points spread over regular intervals along the molecular dynamics trajectory. In order to extract the unknown parameters in the polarizability model, we optimized the model polarizability with respect to the DFT polarizability for those 50 data points. We then used the extracted parameters and positions to calculate the model polarizability for each time step along the trajectory. This allowed for the calculation of the derivative of polarizability with respect to time. We produced a theoretical Raman spectrum from the time derivative.

Overall, the polarizability trajectories predicted by DFT and by the atomistic model showed great agreement. Figure 1 visually displays the evolution of polarizability over time for NH_3. As can be seen, the model's peaks closely match that of DFT. To more rigorously assess the correspondence, linear regression was performed, as displayed in Table 1. A regression line with slope 1 and intercept 0 indicates perfect correspondence between the 2 models. In all cases, each value deviated by less than 10%.

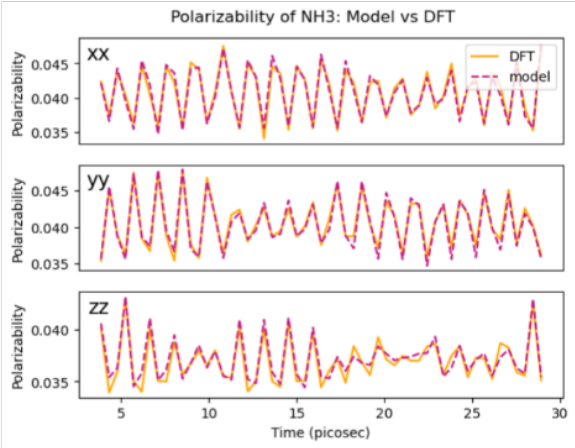


Figure 1. Polarizability over time for ammonia as predicted by DFT (orange) and the novel atomistic model (pink). From top to bottom, graphs show the xx, yy, and zz components of the tensor. The two trajectories show great agreements.

Model vs DFT: Polarizability Predictions

	xx	yy	zz
Cl_2	1.005 0.000	1.028 -0.001	1.004 0.001
H_2S	0.932 0.006	0.941 0.006	0.980 0.001
NH_3	0.967 0.001	0.918 0.003	0.962 0.001
SO_2	0.962 0.006	0.951 -0.001	0.950 0.006

Table 1. This table demonstrates the degree of correspondence between the DFT and atomistic model's predictions of each molecule's polarizability trajectory. The values give the slope and intercept (indented) of the best-fit line of the graph of the model vs DFT. Slope of 1 and intercept of 0 indicates perfect fit.

The theoretical Raman spectra showed more deviation with experimental data. However, it is unclear whether this is due to failure of the model to recreate the DFT trajectory or to the approximate nature of DFT. In each case, the theoretical spectrum displayed peaks in similar regions to the experimental spectra. However, the peaks were shifted by up to 250 cm^{-1}, and varied in shape and relative height. Figure 2 demonstrates these results for NH_3. Both the theoretical and experimental spectra exhibit two peaks in the 900-1000 wavenumber region; however, the peaks of the theoretical model are

shifted ~40 cm^{-1} to the right. The proximity of the two peaks is also altered. Nevertheless, the ability of the model to reproduce the Raman spectra with low computational cost is powerful and holds potential as a theoretical tool. The success in reproducing Raman peaks in similar ranges is also remarkable.

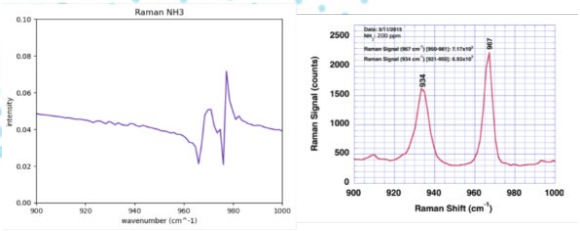


Figure 2. Theoretical (left) vs experimental (right) Raman spectrum for NH_3. Both spectra exhibit 2 peaks in the 900-1000 wavenumber range. The spectra also agree in the shorter height of the lower-energy peak. The peaks are shifted ~40 cm^{-1} to the right in the theoretical spectrum [4].

[1] "Raman Spectroscopy," University of Michigan, Adv. Phys. Lab. (2006).
[2] Thomas, M. et al., Phys. Chem. Chem. Phys., **15**, 6608 (2013).
[3] Paul, A. et al., arXiv, 2304.07526 (2023).
[4] Aggarwal, R.L. et al., Advances, **6**, 025319 (2016).

9-Loci Multi-Race Graph Imputation and Matching for HLA Genotypes

Leah Baron
Advised under Prof. Yoram Louzoun and PhD student Sapir Israeli

Human leukocyte antigen (HLA) matching plays a crucial role in the success of allogeneic hematopoietic stem cell (HSC) transplantation, as it significantly impacts the outcome of the procedure. Identifying suitable donors with matching HLA variants from unrelated donor registries is of paramount importance to improve patient outcomes. In September of 2019, researchers from Professor Yoram Louzoun's lab at Bar Ilan University and colleagues made significant strides in approaching this issue, developing a system of algorithms coined GRIMM: GRaph IMputation and Matching for HLA genotypes [1]. In it, HLA-matching algorithms predict the most likely genotype matches between patients and potential donors, considering ambiguous HLA typing data and the population HLA frequency distribution of haplotypes, including complex population substructure.

The initial algorithm used (HapLogic) performs HLA imputation by directly enumerating all possible haplotype pairs consistent with 5-loci HLA typing data, and subsequently calculates the probability of each haplotype pair based on population haplotype frequency distributions. GRIMM improved that process and ensured efficient matching by storing the imputation results in a graph database, which enables rapid and flexible query processing to identify potential donors that match a patient's HLA profile. GRIMM provides real-time results with a cost sublinear to the size of the donor registry, overcoming the scalability issues faced by traditional methods. In addition, using the World Marrow Donor Association's cross-validation dataset, it has been demonstrated that GRIMM's results are in concordance with the consensus results from the registry. In the years following GRIMM's development, a number of improvements have been made to it. Most notably, one improvement introduces multi-race imputation (MR-GRIMM) to further enhance the accuracy of HLA type matching by reducing the dependency on self-identified race, leading to a 20% decline in matching error [2].

The original GRIMM was developed in the context of five-locus matching (A,B,C,DR,DQB), which is the current standard. However, modern centers require nine-locus matching (adding DQA,DPA,DPB, DRB3/4/5). There are significant computational challenges to be surmounted in the realm of efficient HLA imputation and matching to impute with so many loci. The more loci available to be processed, the more accurate

the typing result, but MR-GRIMM's graph construction software cannot handle more than a six-locus input. I was tasked with figuring out how an expansion to nine loci– the amount that typically constitutes a complete HLA typing– could be executed by building on top of MR-GRIMM.

I worked on developing logic for this improvement, which is outlined in several stages (see Figure 1). The most crucial aspects of the nine-loci MR-GRIMM involve preserving a speed similar to the one used for five loci. The first step to performing this higher-level imputation involves cutting down the imputation input to three loci, which makes running the process even more efficient. Then, the results produced could each be split up by chromosome into two sets, which would allow specific loci in the desired typing to be searched for and compared. This would produce a trimmed version of the results list, which could then be fed into a high-efficiency string-searching algorithm, such as Rabin-Karp, to produce even better typing options. If there were still too many valid options at this point, a high-resolution, more time-consuming search similar in nature to the Smith-Waterman algorithm could be performed.

I began to enhance the code in the py-graph-imputation Python package, which runs the MR-GRIMM version of imputation and was developed by researchers in Professor Louzoun's lab. After spending multiple days acclimating myself to the complexity of the code and its nuances, as well as running a few simulation examples, I worked on the first two steps of the logic, cutting down GRIMM to three loci and splitting the results by chromosome.

The graph-based approach to HLA imputation and matching, as demonstrated by the GRIMM tool, represents a practical and efficient solution for bone marrow registry applications. The tool allows for bulk imputation, making it valuable for both clinical and research purposes, including virtual crossmatch for organ allocation and disease association studies. To fully leverage GRIMM's capabilities in the clinic, it requires seamless integration with operational registry software for donor selection. The open-source nature of the implementation enables customization and facilitates its adoption by other registries worldwide. The use of graph database applications in genetics and bioinformatics holds great promise, providing a versatile and powerful platform for mapping complex relationships between genotypes and phenotypes, and GRIMM is a compelling example of the potential of graph-based algorithms in this domain. The expansion to nine loci should only improve matching and survival of HSCT patients, and I highly anticipate those outcomes in the future.

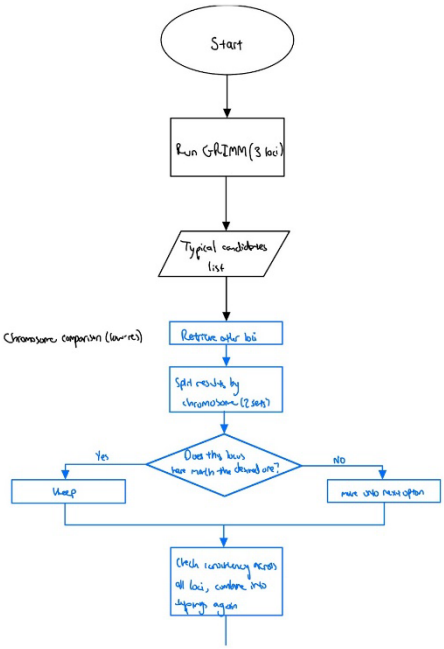


Figure 1. Flow chart of nine-loci logic.

[1] Maiers, M et al., Bioinformatics, 35(18), 3520–3523 (2019).
[2] Israeli, S et al., Combined Imputation of HLA Genotype and Race Leads to Better Donor-Recipient Matching (Unpublished).

Investigation of Non-covalent Complexes via Electrospray Ionization Mass Spectrometry

Ned Krasnopolsky
Advised under Professor Yoni Toker, PhD student Ori Licht, and MSc student Mirit Anaby

Amino acids, the building blocks of proteins, are known to engage in various non-covalent interactions. These interactions play a crucial role in the formation of peptide chains, in a process known as Peptide Bond Formation (PBF). The study of non-covalent amino acid clusters is highly relevant in questions pertaining to biogenesis, or the origin of life. Molecular clusters are aggregates containing two or more entities which are held together non-covalently, via weakly bonding interactions such as hydrogen bonds. Meng and Finn (1989) demonstrated the clustering of arginine, leucine, and histidine. Later studies showed the viable, stable formation of the serine octamer [1]. More comprehensive studies of all amino acids were carried out with an eye towards the stability of the different clusters [2].

The early emergence of proteins likely involved the self-assembly of simple organic molecules into more complex structures. Amino acids, as the constituents of proteins, play a vital role in this assembly process. Indeed, a more recent study argues for the possibility of the gas-phase transformation of clusters into dipeptide bonds following energetic excitation in the form of collision, in a process known as low-energy collision-induced dissociation (LE-CID) [3]. Further analysis of amino acid clusters is a critical step in discovering more about the early stages of protein formation. In recent years, works have shown that different excitation methods yield different dynamics in the cluster, with two main competing channels: PBF and cluster evaporation [4,5,6].

Electrospray Ionization Mass Spectrometry (ESI-MS) has emerged as a powerful tool for the study of non-covalent interactions in amino acid clusters, enabling the exploration of their structural characteristics and stability. ESI-MS allows for the investigation of these cluster interactions, thereby facilitating the identification of preferred binding partners and a better understanding of the relevant thermodynamic processes.

Variations in ESI-MS operation settings lead to a better understanding of how temperature and other conditions impact the formation of these clusters. These insights are critical in developing a theory of biogenesis, which is highly dependent on environmental conditions. ESI-MS experiments have successfully been paired with computer models to provide a more comprehensive understanding of the energetics of non-covalent interactions within amino acid clusters. Developments in ESI-MS technology promise to further improve our understanding of non-covalent amino acid interactions and their role in biogenesis on Earth and beyond.

In our study, we investigated the formation of protonated valine and serine clusters, in addition to heterogenous serine-valine clusters. Capillary, sample core, and extraction core voltages were held at 3200 V, 50-70 V, and 0 V, respectively. The desolvation temperature was kept at around 160 °C. Standard clusters were observed for valine, as shown below (Figure 1). Isotopes of carbon were also observed (Figure 2). The dimer, trimer, and tetramer were all present in the expected abundances, as were all possible heterogeneous clusters. MS-MS runs confirmed that the tetramer was constructed from two dimers. Further study is needed to determine the necessary binding energies of these interactions.

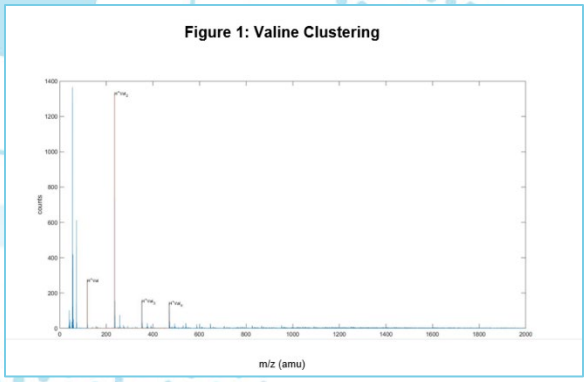


Figure 1. Valine clustering.

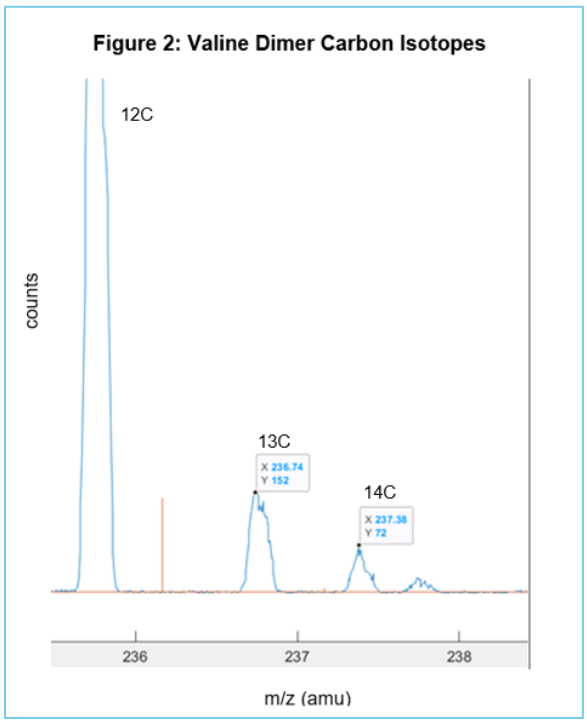


Figure 2. Valine dimer carbon isotopes.

[1] Nanita, SC and Cooks, RG, Angewandte Chemie (International ed. in English), **45**, 554-69 (2006).
[2] Nemes, P et al., Journal of Mass Spectrometry, **40**, 43-9 (2005).
[3] Singh, A et al., Rapid Communications in Mass Spectrometry, 28, 2019-23 (2014).
[4] Rousseau, P et al., Nature Communications, 11, 3818 (2020).
[5] Nihamkin, M et al., The Journal of Physical Chemistry Letters, 11, 10100-10105 (2020).
[6] Licht, O et al., Angewandte Chemie (International ed. in English) 62, e202218770 (2023).

Microbiomic Pathways Data Analysis Pipeline

Yisrael Wiener
Advised under Prof. Yoram Louzon and PhD student Oshrit Shtossel

The intricate interplay between the human microbiome and its host's physiological processes has far-reaching implications for health. Microbial activity often translates into discernible changes in metabolite concentrations, requiring the development of methods to predict metabolic profiles from microbial taxonomic frequencies. These methods assume a direct connection between gut microbiome composition and blood metabolite levels. While this direct correspondence is essential to note, the potential insights embedded within the complex metabolic pathways intrinsic to these interactions warrant further exploration. In the pursuit of predicting host phenotypes, specific metabolic pathways emerge as an essential factor. These pathways have the potential to yield valuable insight into various facets of host health, as the expression levels of certain pathways may correlate with certain conditions.

Consequently, the need arises for a systematic approach to convert genomic and metabolic data into relevant pathway information. In approaching the question of pathway analysis, two initial approaches regarding the output of the data processing pipeline were developed. The first involved taking the relative abundance of a species within a sample and deriving pathway information based on already existing databases containing the genomic content of that species. However, this approach had two limitations. First, the relative abundance of a species in the microbiome does not correlate directly with its impact, as often a minority of species within a sample will dominate the metabolic production [1]. Secondly, current genomic databases offer little insight into actual

pathway expression, as they do not often contain information regarding gene duplication or expression levels.

To effectively navigate the challenge of pathway data analysis, a second approach was developed that involved deriving genetic information from gut samples and using gene counts to derive pathway information. Granted, the nature of pathway expression is multifaceted, being influenced by an interplay of genetic, environmental, and physiological factors. However, amidst this complexity, the relative abundance of specific genes constituting a given pathway does serve as a reliable indicator of pathway expression levels. As such, a meticulously designed pipeline was developed.

This pipeline systematically derives genetic and pathway data from raw sequencing data unveiling pathway-related information encoded within genetic material. The pipeline encompasses a series of sequential steps. Initially, raw genomic data in fastq format, which includes both sequence data and each base pair's accuracy score, is subjected to processing using tools like Metaphlan, aiming to detect the species present within the data. The output of this initial step contains the species (and their taxonomic levels) present within a sample and the relative abundance of each species. This information has to be filtered using a newly developed process that detects species within the sample with accessible pathway data. Once species are identified, species-specific reference files are retrieved from the NCBI database using a locally developed algorithm to serve as a basis for sequence alignment.

Following this, sequences are aligned against these references, facilitating the identification of significant gene expression within the genomic data. Utilizing specialized R packages, feature count matrices are generated for each species, offering a structured representation of gene expression. Notably, the analysis progresses to the identification of pathways by utilizing accumulated genetic data, revealing clusters of functionally related genes. In order to produce such data, information from the KEGG pathway databases has to be extracted, translated, and applied to each species. Thus, relevant data regarding hundreds of bacteria and the hundreds of pathways they share can be derived for each sample.

To bolster the analytical rigor, a logarithmic statistical test was employed to assign scores to pathways based on gene abundance, thereby providing a quantitative perspective on the dynamics of these pathways. The average expression of certain genes within a pathway was determined across a sample space, and using this as a basis, the ratios of actual

	Fructose Metabolism					Galactose Metabolism		
	K02313	K03217	K06348	K03501	K03498	K03497	K03722	K02990
Adlercreutzia_equolifaciens	6	3	12	9	6	9	4	9
Alistipes_communis					2	3		
Alistipes_dispar	5	2		3	10	2		
Alistipes_putredinis								
Anaerobutyricum_hallii	2	20	12	9	23	63		14
Anaerostipes_hadrus	266	224	181	121	37	187		78
Bacteroides_caccae	11	18		12	5	6		10
Bacteroides_cellulosilyticus	6	17		3	4	7		7
Bacteroides_faecis	7	2			6	2		5
Bacteroides_intestinalis	2	2		3	2	3		
Bacteroides_ovatus	9	40		20	13	20		5
Bacteroides_salyersiae	2	4			2			
Bacteroides_stercoris	148	86		49	47	66		32
Bacteroides_uniformis	128	84		18	65	74		40
Bacteroides_xylanisolvens	12				5	25		6

Table 1. Example of gene and pathway output in heatmap format.

expression versus the average expression could be calculated, thus suggesting when certain pathways or genes are enriched within a sample. However, the log of the ratio was used in developing the pathway score, so as not to give too much power to any gene in particular. Finally, these values, often numbering in the hundreds, were summed together for a particular pathway and divided by the total number of genes to keep the scores consistent across the samples.

Thus, this meticulous scoring methodology provides a comprehensive and quantitative framework for pathway analysis, contributing to a more profound understanding of the intricate molecular mechanisms at play. By considering both the complexities of microbial interactions and the detailed genetic expressions, this approach sheds light on the intricate relationship between the microbiome and host health, potentially paving the way for new avenues of therapeutic intervention and personalized healthcare strategies.

[1] Shtossel, O. et al., Microbiome-metabolomeinteractions predict host phenotype. Research Square (2022).