Volume 47 Spring 2024

MURJ

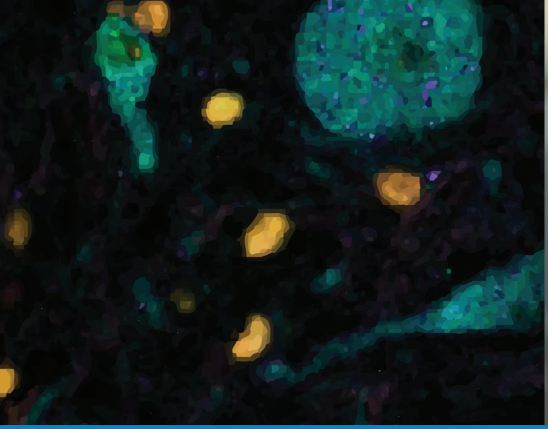
Massachusetts Institute of Technology Undergraduate Research Journal

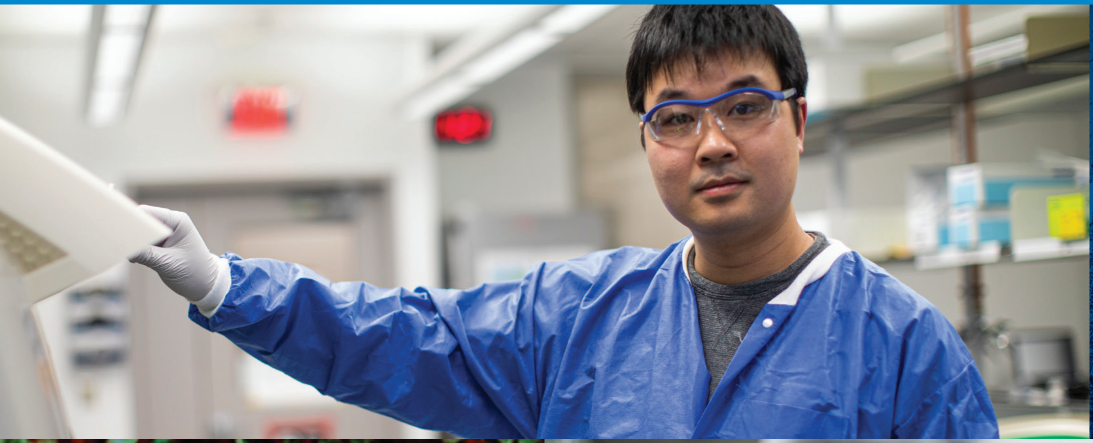
Feature p. 5

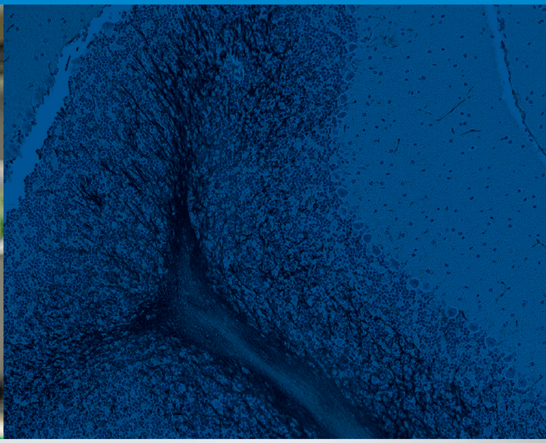
Musical Chronicles II: Unifying Artforms and Uniting People

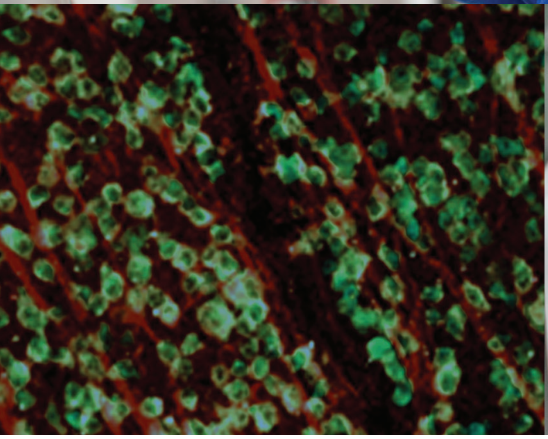
Research p. 17

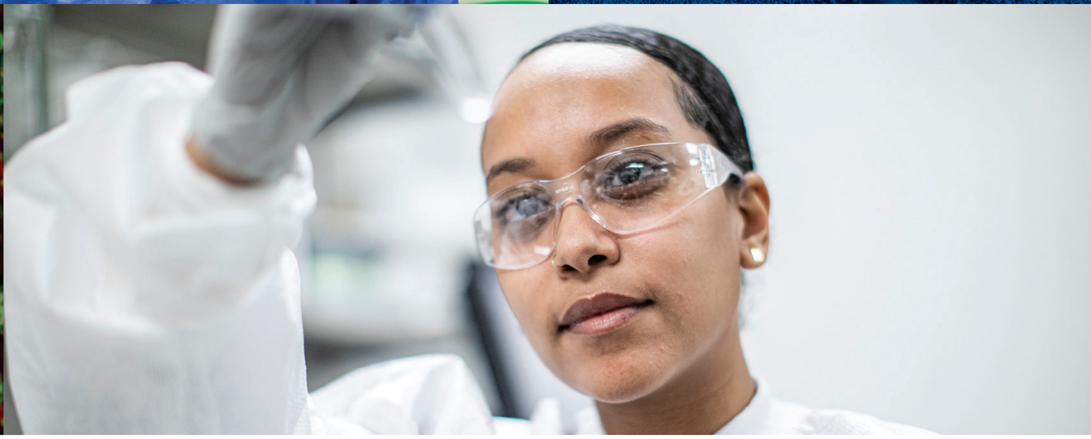
Tackling cancer cachexia with engineered macrophages

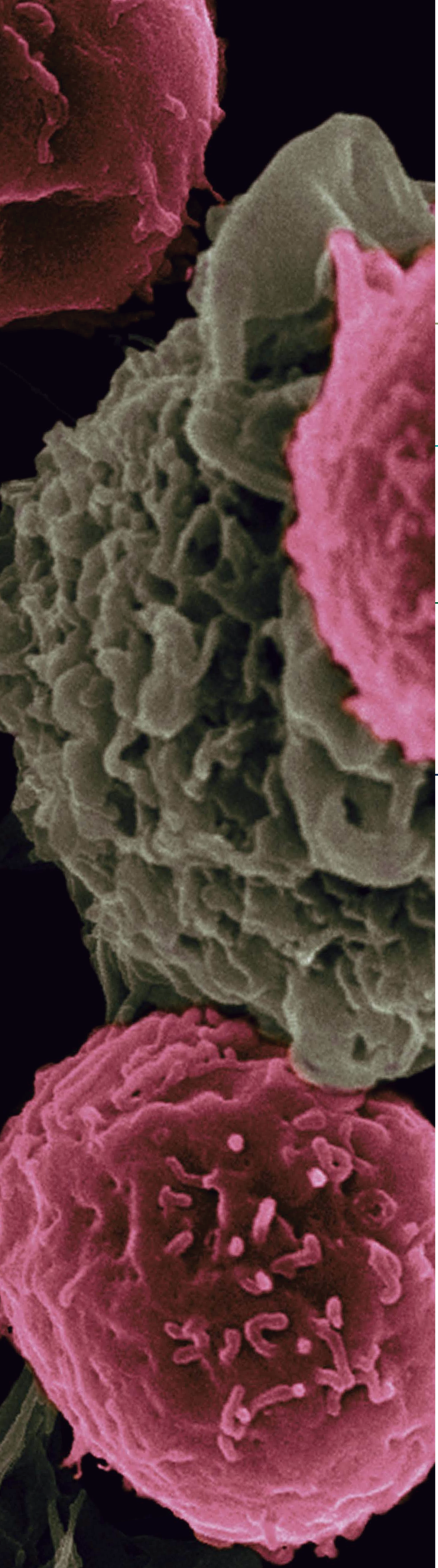












Contents

Introductory Letter

Features

UROP Summaries

Reports

Massachusetts Institute of Technology

UNDERGRADUATE
RESEARCH JOURNAL
Volume 47, Spring 2024

MURJ Staff
MIT Undergraduate Research Journal

Massachusetts
Institute of Technology

April 2024

Dear MIT Community,

We are pleased to present to you the 47th issue of the MIT Undergraduate Research Journal (MURJ). MURJ, a biannual student-run publication, consistently features innovative and exciting undergraduate research at MIT and highlights of the wonderful MIT community. In this issue, we are proud to highlight the research of students at MIT, truly driven by the university's motto of "mens et manus." This issue specifically spotlights the work conducted to improve the quality of life, demonstrating MIT students' profound abilities to contribute to the benefit of society.

Research in this issue spans various fields, including computer science, biology, and engineering. Original student work has been conducted analyzing the interpretation and impact of legal language (legalese) on decision-making in law. We also publish research regarding the identification and treatment of viral disease, with new work identifying classification models for more optimized host targeting for improved treatment. Cancer research is also highlighted in this issue, focusing on possibilities of treatment for cancer cachexia, a disease that affects up to 80% of cancer patients.

In addition to publishing the research of MIT students, we focus on reports by students highlighting MIT life. This semester, we share a small part of the musical community here at MIT, featuring Maya Cunningham, an ethnomusicologist highlighting aspects of African history, culture, and music.

Research is never an independent task, nor the publication of new research. This journal is possible only through the dedicated commitment by MURJ staff members and our undergraduate researchers. We thank the Editorial Board, contributors, and the greater community for the continued support of our journal.

For previous issues of the MIT Undergraduate Research Journal, please visit murj.mit.edu. If you are interested in contributing to future issues, we invite you to join our team of authors and editors or submit your research. If you have any questions, please contact murj-officers@mit.edu.

All the best,

Lia Bu
Editor-in-Chief

No material appearing in this publication may be reproduced without written permission of the publisher. The opinions expressed in this magazine are those of the contributors and are not necessarily shared by the editors. All editorial rights are reserved.

MURJ Features

Musical Chronicles: **Unifying Artforms and Uniting People**

Featuring Maya Cunningham

By Emily Hu

"Thousands of years before history recorded/Deep in the jungle a woman stepped on wet clay/And the print remained there/ Future reminder of ages ago."[11]

All around us, cultural footprints are being left behind. Footprints etched, footprints stamped, footprints molded. Some are immediately washed from shore as the tides of time ebb and flow. Others linger for a while, fleetingly spared from anonymity. Others still, as referenced by Wayne Shorter's jazz standard "Footprints,"[11,18] are immortalized in the clay of history, and despite a temporal obscurity, are inevitably unveiled to the larger world.

And it is precisely these footprints from the last category – footprints of rich histories and cultures from marginalized societies – that have become increasingly studied and appreciated by music lovers and musicologists alike. In the previous issue, our focus was concentrated on the preservation of musical "footprints" in Mexican folk music through the lens of non-profit musical education.[13] Through the work of Dr. Joseph Maurer, lecturer of world music at MIT, we learned about the importance of the Mexican song, or *son*, in immigrant communities. Through the establishment of ethnic music programs such as the Mexican *Son* Music School in Chicago, children of Mexican descent were given an opportunity to reconnect with their roots, cultivate closer connections with their families, and develop essential skills - confidence, self-discovery, and more - for their future adult lives.

In our follow-up inquiry, let us investigate a broader panorama still:

It is a world of African music that, despite both subconscious misconceptions and active efforts to view it as a monolith, houses a treasure trove of diverse musical, visual, and cultural arts. It is a hemisphere of ever-evolving artforms, with nations of music blending together to generate new creations, and yet somehow, each piece remaining distinct from one another.

This is the world of tribal folk music; of basket-weaving songs of the San people; of children's ethnic musical education; and of the confluence of people leading to the emergence of hollers, blues, and jazz. And to steer us through all the intricate inner workings of this immense landscape is Maya Cunningham, an ethnomusicologist, vocalist, visual artist, and activist.

Cunningham currently teaches history and cultural courses at MIT and Berklee and was awarded a Fulbright Scholarship in 2017 to study traditional music schools in Botswana. She fuses her jazz vocal singing with textile, glass, and mixed media arts and has published several book chapters on Black history and music. She currently holds the Executive Director position for Themba Arts and Culture, a non-profit education initiative and aims to promote public awareness for African and African American musical traditions.[1,14]

So, although our time is brief, let us delve into each aspect – historical, artistic, and cultural – that weave together to form the fabric of African arts and people today.

A Historical Preview of African and African American Music

Although geographic barriers like deserts and forests have catalyzed unique musical trajectories in each region of Africa, a few marked commonalities exist, such as a musical emphasis on a strong percussion beat and communal bonding. Take West Africa for instance, where rhythmic structures such as meter and form are foundational to the music. There, the Djembe ("Dje" meaning *gather* and "be" meaning *everyone*) drum has been used throughout the region for several centuries to accompany music, including singing and performances during baptisms, weddings, and funerals.[12]

Each of these drums is believed to contain three spirits that confer healing and magical powers to the instrument. Despite these supernatural associations, the craftsmanship ensures that the pitch of the instrument is tuned to the human voice, which allows the rhythmic playing to mimic the discourse of human speech (hence the alternative name of "talking drums"). This is similar to other traditional West-African instruments, such as the balafon, a xylophone-like instrument believed to have been a gift endowed by supernatural beings, which also has an emphasis on reverberating timbres (or sound quality) similar to that of human speech.[6,12,17,18]

Djembe drum. African Global News. https://africa-global news.com/djembe-history-and-sound/

Maya Cunningham. MIT. https://mta.mit.edu/person/maya-cunningham/

And it is from this foundation of communal music, with its unique instrumentation, rhythms, and tonal structure, that much of contemporary African music, and notably, African American musical forms are based. As Cunningham explains, during the trans-Atlantic slave trade, most of the African migration to the United States stemmed from the forced slave trade of West Africans, including but not limited to from the areas of Guinea, Burkina Faso, Mali, and Gambia. Over time, as the music of the West intermingled with African music traditions, now popular African American forms such as jazz and blues began to develop, harboring rhythmic elements that are heavily based upon traditional West African music.[19] A notable example of this can be seen through the musical career of the master Nigerian drummer

Olatunji, whose recording of Nigerian drum rhythms, *Drums of Passion*, garnered critical acclaim among notable jazz musicians at the time, including the African-American saxophonist John Coltrane. Through Olatunji's dedication to West African music traditions and pioneering forms, jazz music incorporated more traditional West African rhythms into the African-American jazz form, marking a beautiful combination between the two heritages.[15]

Continuing the timeline, another notable point in African American music development can be seen in the 12-bar blues, a form of the blues that is presently one of the most frequently-used chord progressions in jazz and popular music. This progression traverses four bars each of pattern A A, and then B, which is at a sharp contrast to Western European music, which often features a more symmetric 32-bar structure of A A B A. Instead, the meter and form of the modern blues are believed to have arisen from African American work songs such as field hollers sung by field slaves in plantations and chain gangs. [2,10,16]

Thus, although the precise origins of the Blues cannot be traced from solely West-African origins, the fundamental cornerstones of the genre are inseparable from the community, lifeblood, and music of Africans, first in West Africa, and later in the United States.

Field Work: African Music in Action

But, in order to truly encapsulate the importance of African musical traditions, not as a mere historical relic, but as a living and breathing art form, it is vital to examine how African traditions are being passed down, performed, and lived in action today. For this, let us turn to the work of Cunningham in her 2017 field work in Botswana, where she investigated the linkage between primary music education, traditional music, and national identity. There, she uncovered a rich heritage of intergenerational education, where cultural identity was cultivated alongside the celebration of traditional African music.[19]

In the realms of history, children were taught songs sung in the royal court, and from those melodies, educated about the history of the Tswana chiefdoms. Imageries of chiefs, of the wide expanses of territory extending into the Kalahari desert, of the former generations of Setswana people, past but not yet bygone, all coming to life. With respect to the community, classroom children were actively involved in the singing of traditional wedding songs, baby-naming songs, and game songs. From each of the songs, students were informed about how the music within their community paralleled each key event through the passage of life – from birth, to adolescence, to marriage and child-rearing. Students were reminded about the music's fundamental characteristic of community, and how their national identity was tied not entirely to geography or politics, but more so with cultural and communal identity. Even though the setting of oral tradition had shifted from solely within the tribe to into the classroom, the intrinsic element of music as a glue for common identity had not.[5,19]

"Even though the setting of oral tradition had shifted from solely within the tribe to into the classroom, the intrinsic element of music as a glue for common identity had not."

But, similar to the preservation of ethnic immigrant music, as discussed in the previous issue[13], musical tradition, and preservation are never solely restricted to the classroom. As elucidated by Cunningham's journey through Botswana, traditional folk songs were not only celebrated within the confines of primary schools, but likewise preserved by the music of the San people within the harsh wilderness of the Kalahari Desert and by the vocals of Botswana women in their daily work.[19]

For instance, Cunningham observed that within the San community, music was utilized to accompany social happenings within the community such as games, hunting, and healing rituals. As such, a strong emphasis is placed on community participation, such as by clapping, dancing, singing, or playing an instrument.[9]

As their musical tradition is entirely orally transmitted, the Sans' songs are constantly evolving, with lyrics and entire genres being replaced as popularity of songs peak or fall into disuse. For example, when certain hunting rituals became omitted during the hunting season, the songs associated with those rituals disappeared and became replaced by more popular songs, similar to the idea of "hits" within the Western music industry. What results from this natural cycle is an organic evolution of African music, unchanging in its richness but constantly developing as an art form and an integral element of maintaining community.[9]

From both of these examples, Cunningham gives us insight into the rich intergenerational heritage of African music traditions. Whether within the classroom of modern day or through the oral transmission of one generation to the next, the music of Africa is continuously inherited, treasured, and valued. In perpetuity.

Blending Artforms and Commemorating History

The San people are known not only for their diversity in music celebration, but their incorporation of music with visual art. As mentioned earlier, community songs were often used as an avenue for performing healing rituals and gathering the community. However, as witnessed by Cunningham and discussed by Dowson, author of *Rock Art and Social Change in South Africa*, rock or cave paintings were often incorporated with these musical traditions to commemorate trance dances, honor healing events, and record larger community occasions, such as rain dances or contact with Europeans.[7,8,19]

More presently, as alluded to earlier, Botswana women often weave musical traditions with tree fibers together in their basket-making art. According to Cunningham, along with nature-inspired patterns, such as "Forehead of the Zebra" or "Tears of the Giraffe," women were observed during her visit to sing songs with motifs that reflect their visual art creation.[4,19]

Thus, in order to encapsulate these discoveries of historical traditions in visual arts and music, Cunningham has continued to fuse a multitude of artistic techniques along with her vocal projects. In a musical essay, or a programmatic collection of song

Tumbuktu – Beneath the Desert Moonlight. Maya Cunning ham. https://mta.mit.edu/person/mayacunningham/

recordings, titled “All Africa,” Cunningham aligned songs that she learned from her field work and influential jazz music along with glass work, textiles, and portraits.[19]

One notable work from this project is a glasswork piece titled “Timbuktu – Beneath the Desert Moonlight,” which depicts the Sankoré Madrasah Mosque established by Mansa Musa in the Mali Empire and is inspired by the jazz trumpet legend Lee Morgan’s “Desert Moonlight.” Just as the musical piece evokes a moonlit night in an African desert, the silhouette of the Sankoré Madrasah is simply outlined against the purple moony sky; the work encapsulates a beautiful overlap between the rich historical past of African chiefdoms and the powerful African American jazz traditions of present day.[3]

But, in order to bring this junction between African and African American music to a wider audience, we must look beyond a solely artistic hemisphere and once again turn our focus back to the classroom, one of the ultimate spaces for long-term advocacy for musical preservation.

Education Outreach: Traditions in the Classroom and Future Outlook

Recalling the xylophone-like balafon, Cunningham mentions that a key advancement for the classroom and heritage music education is the incorporation of traditional instruments within the classroom. In an initiative called the Heritage Arts Institute, Cunningham and her collaborators strive to bring instruments such as the gyil, a sister instrument of the balafon originating from the Gur-speaking population of West Africa, into American classrooms. In doing so, not only are children enriched in their musical heritage, but moreover, given a glimpse into their familial and communal identity.[1,19]

Balafon. African Global News. https://africaglobal-news.com/the-balafon/

As an example, Cunningham points to a key interval (or two pitches played simultaneously) on the gyil, known as a “twin” interval. As these musical intervals are historically linked with the relationship of twins, an important family relationship within tribal culture, she mentions, students are not only informed about “musical nomenclature and pitch relation”, but also about family organization within the traditional African family. In doing so, an “experience [that is normally] reserved for' someone with “specialized training [is] offered to broad numbers of students to receive the [...] benefits of achieving what we would call a bimusicality.”[19]

While such education is culturally informative for all students, Cunningham emphasizes that this model of heritage music education is especially important for African American and Afro-descendant students. By connecting traditional instrumentation within the classroom to demonstrations of African American and African history, these heritage students are actively engaged in a history that is not only based on African American musical pioneers, but all the way back to the traditions of historical African empires. As summarized by Cunningham, “it is deeply empowering to learn [this] kind of information about our heritage through music.”[19]

From the classrooms of Botswana to the wide expanse of the Kalahari desert, and back to our own classrooms, African and African American music has been and will continue to be an immense universe inhabited by a creative community and never-ending possibilities. As our footsteps mark an end to this exploration and fade into the distance, the parallel footprints left behind by the African people in their music and art – past present, and future – will continue to be engraved in the clay of history, "a future reminder of ages ago."[11]

Botswana basket. Ndalama African Deserts Crafts. https://africandesertcrafts.com/site/wp-content/uploads/2009/12/*NXBF-10-08-1-3.jpg*

References

[1] *About Maya Cunningham*. ETHNOMUSICOLOGY IN ACTION. (n.d.). http://www.ethnomusicologyinaction.org/about-maya-cunningham---executive-director.html

[2] Appen, R. von, & Frei-Hauenschild, M. (2015). AABA, refrain, chorus, bridge, prechorus – songformen und ihre Historische Entwicklung. *German Society for Popular Music Studies*, 57–124. https://doi.org/10.1515/transcript.9783839418789.57

[3] Art. MAYA CUNNINGHAM - JAZZ VOCALIST/ETHNOMUSICOLOGIST. (n.d.). http://www.mayacunninghammusic.com/art.html

[4] Botswanacraft Marketing. (2020). *Botswana Basketry Information*. https://botswanacraft.com/botswana-baskets

[5] *Botswana*. Countries and Their Cultures. (n.d.). https://www.everyculture.com/Bo-Co/Botswana.html

[6] Colbourne, J. (n.d.). *West African talking drums and Music*. PILOT GUIDES. https://www.pilotguides.com/articles/west-african-talking-drums-and-music/

[7] de Greef, K. (2016). *In South Africa, colonialism was written on Stone*. Hakai Magazine. https://hakaimagazine.com/article-short/south-africa-colonialism-was-written-stone/

[8] Dowson, T. A. (1994). Reading art, writing history: Rock Art and social change in Southern Africa. *World Archaeology*, 25(3), 332–345. https://doi.org/10.1080/00438243.1994.9980249

[9] Emmanuelle, O. (2010, April 28). The "Success" of San Music. https://www.culturalsurvival.org/publications/cultural-survival-quarterly/success-san-music

[10] *Field hollers and work songs*. Black Music Scholar. (n.d.). https://blackmusicscholar.com/field-hollers-and-work-songs/

[11] *Footprints lyrics by Wayne Shorter*. Footprints lyrics by Wayne Shorter (2008, April 26). https://www.lyricsmode.com/lyrics/w/wayne_shorter/footprints.html

[12] *History of the Djembe*. Drum Connection World Djembe. (n.d.). https://www.drumconnection.com/africa-connections/history-of-the-djembe/

[13] Hu, E. (2023). Musical Chronicles: Preserving Music and Empowering People. *MIT Undergraduate Research Journal*, 46, 16–19. https://murj-assets.s3.amazonaws.com/assets/issues/Vol_46_Published.pdf

[14] *Maya Cunningham*. Massachusetts Institute of Technology. (2023, January 13).https://mta.mit.edu/person/maya-cunningham

[15] *Pas Hall of Fame*. Babatunde Olatunji. (n.d.). https://www.pas.org/about/hall-of-fame/babatunde-olatunji

[16] Public Broadcasting Service. (n.d.). *Understanding the 12-bar Blues*. PBS. https://www.pbs.org/theblues/classroom/essays12bar.html

[17] Sylla, C. (2019, October 24). *The story of the balafon, an ancient West African musical instrument*. The Gambia Experience. https://www.gambia.co.uk/blog/article?id=665

[18] Warner, G. (2020, September 3). *Balafon: Wood-tongue-talk*. Garland Magazine. https://garlandmag.com/loop/balafon/,%20https://garlandmag.com/loop/balafon/

[19] A special thanks is given to Maya Cunningham for her collaboration in the writing of this article.

MURJ

UROP Summaries

Analyzing the neural basis of legal reasoning

David Oluigbo[1], Eric Martinez[2], Selena She[3], Anna Ivanova[4], Ev Fedorenko[5]

1 Student Contributor, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02142

2 Student Contributor, Department of Brain and Cognitive Sciences, MIT, Cambridge MA 02139

3 Department of Brain and Cognitive Sciences, MIT, Cambridge MA 02139

4 Supervisor, School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332

5 Principal Investigator, Department of Brain and Cognitive Sciences, MIT, Cambridge MA 02139

Introduction

The debate between legal formalism and legal realism, two theories concerning how judges approach the law, has long intrigued scholars. Legal formalism states that laws are applied by a judge more or less mechanistically or deductively. In contrast, legal realism suggests that judges consider social consequences and fair outcomes when applying laws. However, little research has delved into the cognitive mechanisms behind legal interpretation and the influence of legal language, often referred to as "legalese," on these processes. Legal language, with its distinct features from everyday language, poses challenges for interpretation, especially among laypeople. This study seeks to address fundamental questions regarding legal interpretation and decision-making while also exploring the impact of the form of legal language.

Networks are groups of brain areas that exhibit functional connectivity and correlated activation during specific cognitive tasks. Imaging modalities, like functional magnetic resonance imaging (fMRI), can highlight these activity patterns in real time (Fig. 1), allowing researchers to better pinpoint which brain regions are involved in different processes. The theory of mind (ToM) network is active when an individual must consider a situation's moral and emotional aspects, which is the core tenet of legal realism. Conversely, the multiple demand (MD) network is active during logic-based tasks, such as solving math problems, which aligns with the legal formalism rooted in deductive reasoning. This study aims to elucidate whether evaluating legal scenarios activates the MD or the ToM network. Our comprehensive approach seeks to unravel the intricate interplay between cognitive processes, language, and legal interpretation, shedding light on the underlying mechanisms that shape legal decision-making and understanding.

Methods

Study participants are placed in an fMRI machine to record and examine their brain activation patterns during the

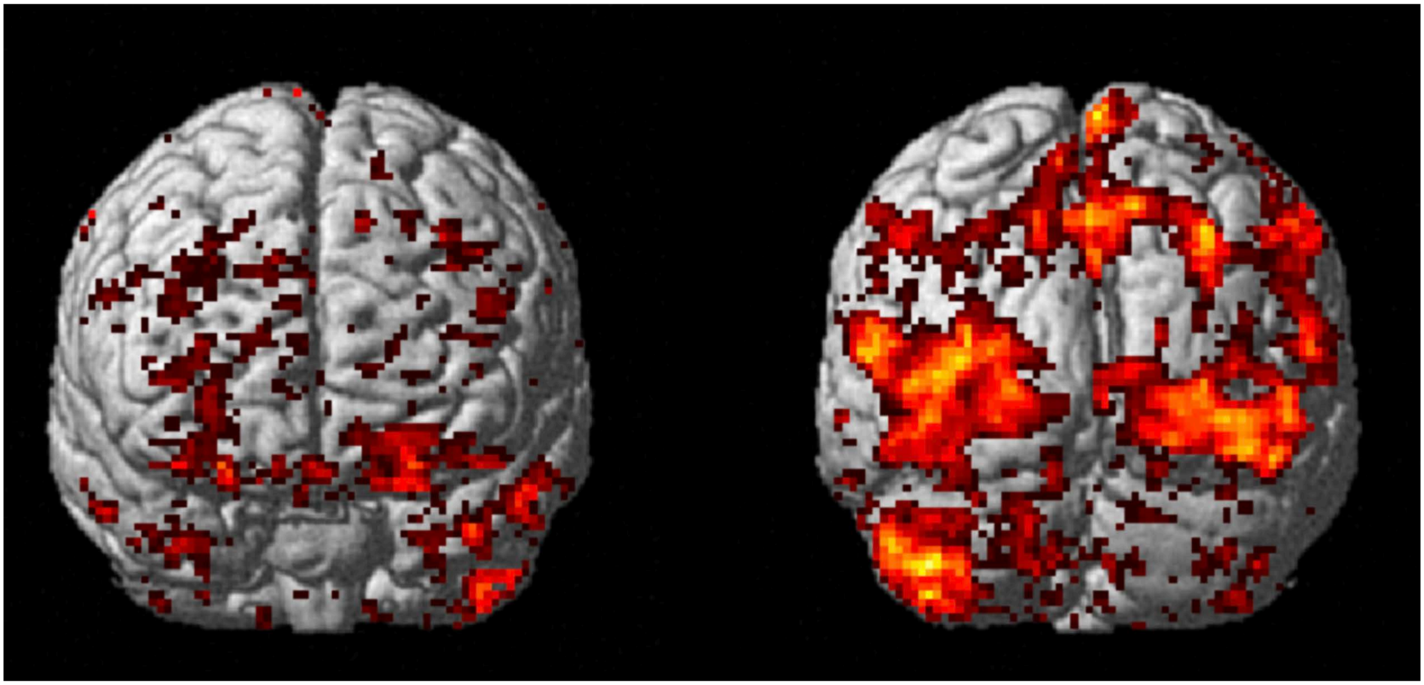
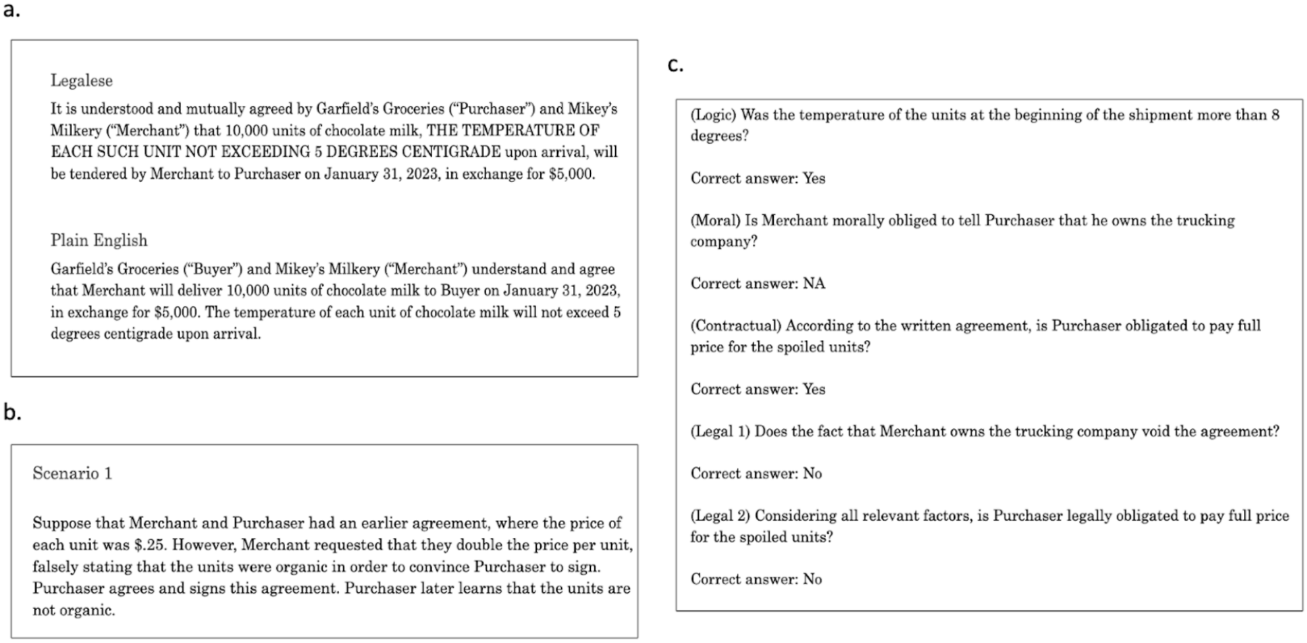


Fig. 1. Front and back views of a participant's brain featuring activation maps for the ToM network for a cognitive moral judgment task.



a.

Legalese

It is understood and mutually agreed by Garfield's Groceries ("Purchaser") and Mikey's Milkery ("Merchant") that 10,000 units of chocolate milk, THE TEMPERATURE OF EACH SUCH UNIT NOT EXCEEDING 5 DEGREES CENTIGRADE upon arrival, will be tendered by Merchant to Purchaser on January 31, 2023, in exchange for $5,000.

Plain English

Garfield's Groceries ("Buyer") and Mikey's Milkery ("Merchant") understand and agree that Merchant will deliver 10,000 units of chocolate milk to Buyer on January 31, 2023, in exchange for $5,000. The temperature of each unit of chocolate milk will not exceed 5 degrees centigrade upon arrival.

b.

Scenario 1

Suppose that Merchant and Purchaser had an earlier agreement, where the price of each unit was $.25. However, Merchant requested that they double the price per unit, falsely stating that the units were organic in order to convince Purchaser to sign. Purchaser agrees and signs this agreement. Purchaser later learns that the units are not organic.

c.

(Logic) Was the temperature of the units at the beginning of the shipment more than 8 degrees?

Correct answer: Yes

(Moral) Is Merchant morally obliged to tell Purchaser that he owns the trucking company?

Correct answer: NA

(Contractual) According to the written agreement, is Purchaser obligated to pay full price for the spoiled units?

Correct answer: Yes

(Legal 1) Does the fact that Merchant owns the trucking company void the agreement?

Correct answer: No

(Legal 2) Considering all relevant factors, is Purchaser legally obligated to pay full price for the spoiled units?

Correct answer: No

Fig. 2. Examples of stimuli materials. **a)** Example of legalese and plain English versions of a contract. **b)** Example of scenario associated with a contract. **c)** Example of set 5 yes/no questions associated with a scenario.

performance of a Legal Reasoning Task. The Legal Reasoning Task consists of 12 trials completed during fMRI scanning. Each trial features one contract, two scenarios, and ten yes/no questions in either plain English or legalese format to examine how participants interpret and evaluate the stimuli.

We created a MATLAB script to present stimuli to study participants in a highly coordinated manner. Our stimuli materials consist of a per-participant, randomized sequence of 12 contracts, each associated with two scenarios, and each scenario associated with five yes/no questions (classified as either a moral, contractual, legal, or logic question) (Fig. 2). The 120 yes/no questions are written in legalese or plain English. The moral and logic questions are mainly designed to examine activation patterns in the ToM and MD networks, respectively. Contractual questions are based on information in the contracts shown to participants, testing reading comprehension. Lastly, legal questions are administered to assess legal comprehension.

The script syncs with the fMRI machine and presents each type of stimuli for an allotted time. It shows each contract for a maximum of 50 seconds, each scenario for 30 seconds, and each question for 10 seconds. Our script incorporates a two-second pause between questions and checks when the participant is ready to move forward to address potential fatigue. Our script also records participants' responses to stimuli in real time and stores information about the exact stimuli presented, the order in which they were displayed, and the duration of their presentation. This comprehensive data allows us to assess the pace at which participants moved through the trial and analyze their responses.

Results

We ran a pilot study, as described above, on two adults. The results of our pilot study provide foundational insights into the cognitive mechanisms underlying legal interpretation. Analysis of brain fMRI activation patterns elucidated the neural correlates underlying legal interpretation. Heightened activation in the language network during plain English scenario processing may signify increased engagement of language-specific brain areas compared to legalese. The MD network demonstrated the highest activation for logic questions, whereas the ToM network demonstrated the highest activation for moral questions. These activation patterns suggest that we had correctly localized the ToM and MD network in our pilot study participants. The MD and ToM network also exhibited notable activation for legal questions, suggesting that deductive reasoning and moral dimensions are relevant in such scenarios.

Interestingly, only the ToM network showed noticeable activation when the pilot participants processed the legal scenarios, while only the MD network showed noticeable activation when reading the contract. There could be an underlying link between objective language in contracts and deductive reasoning. Participants could also have an inherent tendency to approach and evaluate scenarios like a moral dilemma. These findings shed light on the complex cognitive mechanisms underpinning legal decision-making.

To validate our pilot study, we examined the effectiveness of our stimuli by analyzing several response metrics. Mean response rates ranged from 79.2% to 94.8%, highlighting the allotted time was appropriate for processing the stimuli. In addition, mean response times for different question categories ranged from 5.06 to 6.01 seconds, demonstrating the uniformity of question processing despite variable accuracy. Our pilot study also revealed that legal and logic questions had significantly lower mean accuracies than contractual questions, which is expected given that participants can more readily answer contractual questions with information directly available in the contract stimuli. This discrepancy can confound our analysis, so we plan to ensure questions have similar difficulty levels in future studies.

Question Type	Mean Response Rate (%)	Mean Response Time (sec)	Mean Accuracy (%)
Moral	93.8	5.44	N/A
Logic	79.2	6.01	66.7
Legal	94.8	5.06	60.4
Contractual	95.8	5.11	88.3

Table 1. Question accuracy and response results. Accuracy for moral questions is marked as "N/A" because there was no definitive yes/no answer.

Conclusion

This study utilizes a multifaceted approach to investigate the cognitive mechanisms underlying legal interpretation and the impact of legal language on neural processing, offering additional insight into the long-standing debate between legal formalism and realism. We observed the activation of the MD and ToM networks during the cognitive legal reasoning task, suggesting that both legal formalism and realism can explain legal interpretation. Understanding how individuals navigate legal language and engage with legal scenarios provides nuanced perspectives that can inform discussions on the balance between mechanistic deductive reasoning, as advocated by legal formalism, and the consideration of social consequences, as emphasized in legal realism. Ultimately, this research bridges the gap between theoretical debates and empirical evidence, offering a deeper understanding of how legal decisions are made and interpreted within legal systems.

References

Dodell-Feder D, Koster-Hale J, Bedny M, Saxe R. fMRI item analysis in a theory of mind task. Neuroimage. 2011 Mar 15;55(2):705-12. doi: 10.1016/j.neuroimage.2010.12.040. Epub 2010 Dec 21. PMID: 21182967.

Fedorenko E, Behr MK, Kanwisher N. Functional specificity for high-level linguistic processing in the human brain. Proc Natl Acad Sci U S A. 2011 Sep 27;108(39):16428-33. doi: 10.1073/pnas.1112937108. Epub 2011 Sep 1. PMID: 21885736; PMCID: PMC3182706.

MURJ Reports

Tackling cancer cachexia with engineered macrophages

Jeannie She[1], Katie Spivakovsky[2], Allison Lin[3], Matthew Feng[4], Justin Buck[5]

1 Student Contributor, Department of Biological Engineering, MIT, Cambridge MA 02139

2 Student Contributor, Department of Biological Engineering and Department of Electrical Engineering and Computer Science, MIT, Cambridge MA 02139

3 Student Contributor, Department of Electrical Engineering and Computer Science, MIT, Cambridge MA 02139

4 Supervisor, Department of Electrical Engineering and Computer Science, MIT, Cambridge MA 02139

5 Principal Investigator, Department of Biological Engineering, MIT, Cambridge, MA 02139

Cancer cachexia is a severe muscle- and fat-wasting disease that is directly responsible for 30% of cancer-related deaths, yet it is strikingly understudied (Lim, 2020). No treatments in the United States exist beyond exercise and diet regimens. We propose a cancer cachexia cell therapy using macrophages equipped with chimeric antigen receptors (CARs) that recognize and engulf interleukin 6 (IL-6) cytokines. Reducing IL-6 addresses overstimulated inflammatory signaling and alleviates cachexia from its root cause. We have demonstrated that our CAR-P (-P for phagocytosis) specifically binds IL-6, suggesting our therapy may work in tandem with existing cancer treatments to help patients fight cancer. Furthermore, it may be broadly applicable for treating other inflammatory conditions, such as rheumatoid arthritis and lupus.

Introduction

Cancer cachexia debilitates up to 80% of cancer patients, compromising critical organ systems and directly causing 30% of cancer-related deaths (NIH, 2022). It remains to be known if cancer treatments like chemotherapy drive cachexia, but it is widely hypothesized that cancer can directly induce cachexia-causing inflammation. No FDA-approved treatments for cancer cachexia exist in the United States. Instead, physicians have relied on prescribing lifestyle changes to boost patients' muscle mass through exercise and diet regimens. Still, these are merely palliative measures and not targeted therapy (Cancer Cachexia, 2022). Our first step was identifying IL-6 as a promising therapeutic target. IL-6 is a cytokine released by the body in response to inflammation. The overabundance of IL-6 in cachectic patients overstimulates inflammatory signaling cascades, interrupts protein synthesis, and upregulates protein degradation (Carson & Baltgalvis, 2010). These disruptions in biological processes manifest as the body attacking its muscles and fat. **Therefore, reducing extracellular IL-6 levels can mitigate cachexia in patients.**

We were inspired by advances in chimeric antigen receptor (CAR) T-cell therapy. In this up-and-coming cancer therapy, a patient's white blood cells are isolated and a subtype of those cells (T-cells) are engineered with a surface receptor (a CAR) to recognize a specific protein on the surface of cancer cells. To reduce extracellular IL-6 levels in patients, we take a similar cell therapy approach but instead engineer macrophages, a different subtype of white blood cells that engulf small proteins and cell debris via a process called phagocytosis. The macrophages would express CARs that bind to IL-6 and induce a signaling cascade, resulting in the phagocytosis and degradation of IL-6 (Fig. 1).

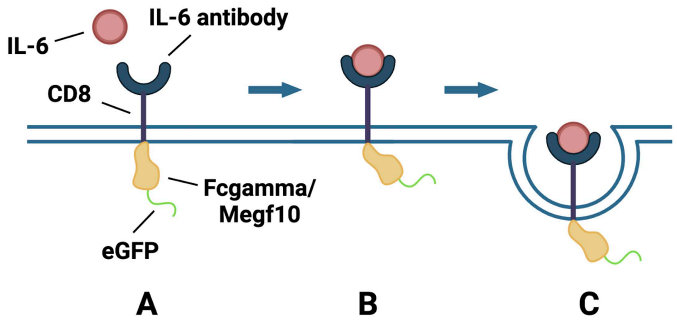


Fig. 1. A schematic representing the sense and response aspects of our design. The CAR, expressed on macrophages, uniquely senses extracellular IL-6. Once IL-6 binds, it induces a phagocytic signal to degrade the IL-6 protein. **A)** CARs are expressed on the membranes of engineered cells, and IL-6 is present extracellularly. **B)** IL-6 binding transduces a signal into the cell, initiating phagocytosis. **C)** The CAR-IL-6 complex is phagocytosed and then degraded.

Methods

Our workflow comprised two main phases: sample preparation and experimentation.

Phase 1: Sample preparation

To modify the host cell's genome, plasmids (circular components of DNA) can be introduced into cells with the help of encapsulating chemicals or electrical impulses, a process known as transfection. Successful expression of plasmids can be verified through fluorescence or other markers engineered into the plasmid. For our project, we designed three plasmids.

1.1: CAR-P protein design

CAR-T cell therapy utilizes a four-part modular CAR design. We designed our CAR-Ps similarly (Fig. 2), justified by research showing that this CAR design successfully translates from T cells to macrophages (Morrissey, 2017). Flexible glycine and serine amino acids, whose chemical properties promote the integrity of correctly folded protein domains, link the CAR components in the following order (Mazinani & Rahbarizadeh, 2022):

1. *Extracellular anti-IL-6 antibody:* This receptor is the exposed part of the protein and is responsible for specifically recognizing our target protein, IL-6. (Takeuchi, 2017).
2. *Transmembrane CD8 domain:* This protein embeds within the cell membrane, bridging components 1 and 3, enabling extracellular binding events to induce intracellular responses like phagocytosis.
3. *Intracellular Fcγ or Megf10 domain:* Fcγ and Megf10 are signaling proteins that induce phagocytosis in macrophages – our desired therapeutic response (Morrissey, 2017). We designed two CARs differing only according to these two proteins.
4. *Enhanced green fluorescent protein tag:* eGFP is a popular reporter protein, and its fluorescence allows us to verify the successful synthesis of CAR proteins in cells.

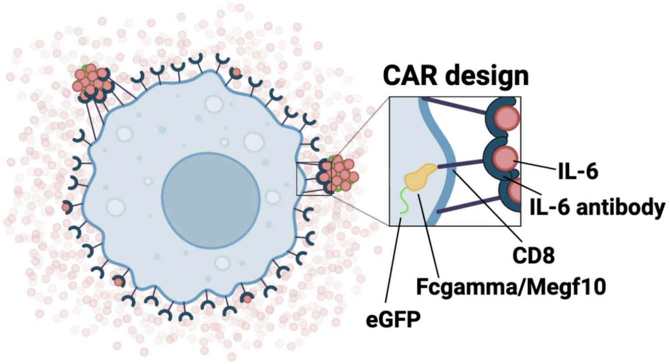


Fig. 2. A Our four-part, modular CAR-P plasmid design.

1.2: IL-6-mCherry fusion protein design

We needed a source of IL-6 molecules to validate our CAR-P's functionality. To visualize IL-6 binding to CARs, we designed IL-6-mCherry: a fusion protein of IL-6 tagged on its 5' end with 3xFLAG for protein purification and on its 3' end with mCherry (a red fluorescent protein).

1.3: Plasmid construction and transfection

To evaluate our design, we investigated the potential for extracellular IL-6 to be sequestered by CAR-engineered human embryonic kidney cells (HEK293), a robust cell line used in place of macrophages in proof-of-concept experiments. Though HEK293 cells lack phagocytic capabilities, our CAR-Ps' intracellular domains responsible for inducing phagocytosis have been previously validated in CAR-engineered macrophages (Morrissey, 2017). The priority of our experiment was to visualize the efficacy of the extracellular component of the CAR-P, so HEK293 cells suffice.

Our aforementioned plasmid sequences (two CAR-Ps and one IL-6-mCherry fusion protein) were modified to match Golden Gate DNA assembly standards (Bird, 2022) and transfected into HEK293 cells using Lipofectamine 2000, a lipid-based chemical that complexes with DNA to carry it to the cell nucleus. HEK293 cells transfected with our CAR-P plasmids efficiently expressed CAR-P proteins, as evident through their substantial green fluorescence (Fig. 3). HEK293 cells transfected with our IL-6-mCherry plasmid efficiently produced intracellular IL-6-mCherry fusion protein, as evident through their substantial red fluorescence (Fig. 4).

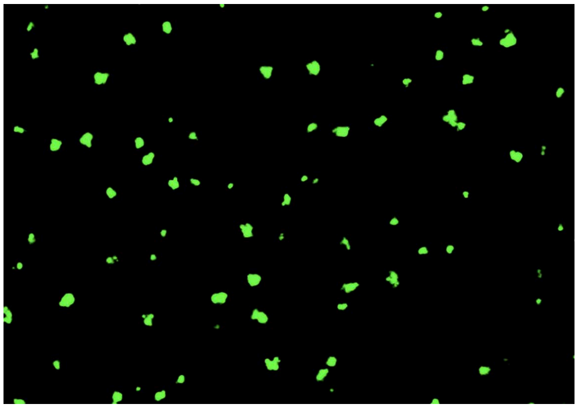


Fig. 3. HEK293 cells transfected with 600 ng Megf10 CAR-P, imaged 48 hours after transfection.

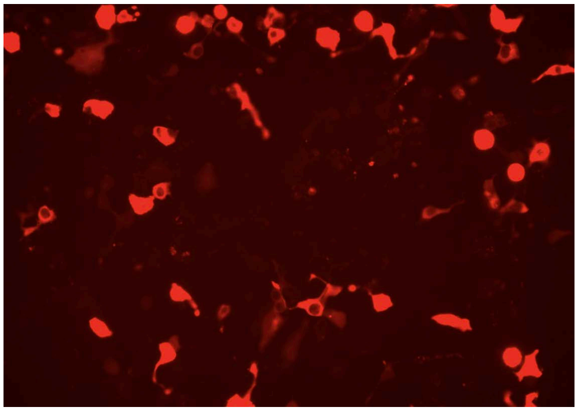


Fig. 4. HEK293 cells transfected with 600 ng IL-6-mCherry, imaged 48 hours after transfection.after transfection.

Phase 2: Experimentation

2.1: Co-culture of HEK293 cells with IL-6-mCherry fusion protein

In vivo, our CAR-P-expressing macrophages are intended to interact directly with extracellular IL-6. An in vitro representation of this cell therapy involves incubating, or co-culturing, the CAR-expressing HEK293 cells with IL-6-

mCherry fusion protein. Two cell lysis protocols were tested to extract the IL-6-mCherry produced intracellularly by HEK293 cells so that the IL-6 could then be co-cultured with CAR cells.

1. Transfected cells were lysed using a freeze-thaw protocol, splitting them open and exposing intracellular proteins like IL-6. The lysate was added to a well plate containing CAR cells and incubated for 6.5 hours (Fig. 5).
2. Transfected cells were lysed using ThermoFisher's M-PER detergent lysing agent. Lysate was purified into IgG buffer using ThermoFisher's anti-3xFLAG magnetic agarose beads. Purified IL-6 was added to a well plate containing CAR cells and incubated for 6.5 hours.

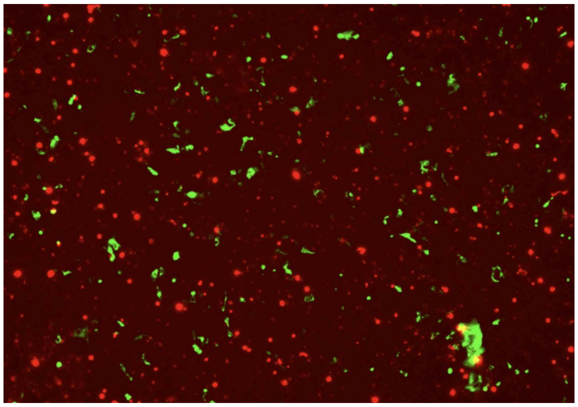


Fig. 5. CAR-P HEK293 cells (green) were incubated with IL-6-mCherry freeze-thaw lysate (red) after 6 hours and viewed using the GFP and RFP filter cubes on the Keyence BZ-X10.

2.2: Quantitative single-cell analysis of samples with flow cytometry

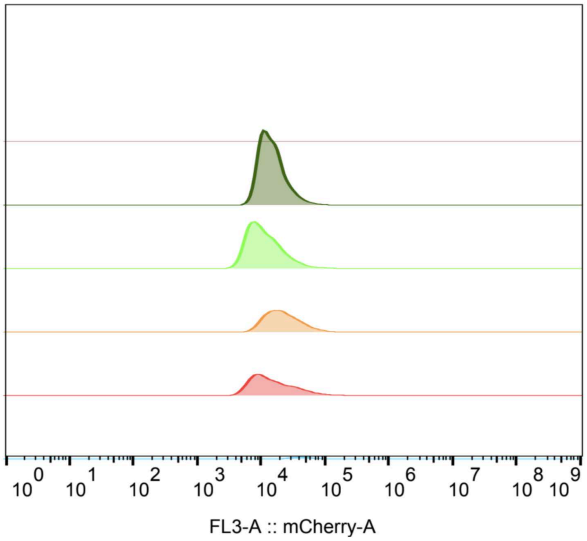
Due to the limited resolution of our microscope, it is not apparent which IL-6 molecules are bound to which HEK293 cells in Figure 5. This limitation motivated our usage of the flow cytometer, a device that precisely quantifies how many cells fluoresce any given color, to characterize levels of CAR expression and of IL-6 binding to the CAR. Since our CARs contain an intracellular GFP domain, they naturally fluoresce green. Thus, the cytometer can sort transfected CAR-P cells according to whether or not they fluoresce green. Similarly, since our IL-6-mCherry fusion protein fluoresces red, the cytometer can sort cells according to whether or not they are bound to IL-6.

Results

We examined six samples (detailed below) in the flow cytometer and gated for the populations that were double-positive: cells that fluoresced green (eGFP+) and red (mCherry+) imply CAR-P expression as well as IL6-mCherry binding. **At least 10% of each experimental sample population was double-positive** (Fig. 6).

Our six samples are analyzed as follows (Fig. 6):

1. Negative control: As expected, control HEK293 cells without CARs bound no IL-6.



	Sample Name	Subset Name	Freq. of Parent	Count
□	Negative Control.fcs	CAR+IL-6	0	0
🟩	Megf10 IgG_Data Source - 1.fcs	CAR+IL-6	10.4	1344
🟩	Megf10 FT_Data Source - 1.fcs	CAR+IL-6	12.0	1116
🟧	Fcgamma IgG_Data Source - 1.fcs	CAR+IL-6	11.0	578
🟥	Fcgamma FT_Data Source - 1.fcs	CAR+IL-6	16.2	599
🟦	EGFP.fcs	CAR+IL-6	1.15	37.0

Fig. 6. The control and double-positive cell populations for each experimental sample are plotted, showing the wavelength of light versus relative cell count. Double-positive single cells exhibit both CAR expression and IL-6 binding. The "frequency of parent" column in the legend denotes the percentage of double-positive cells.

2. Megf10 CAR-P (IgG purification): 10.4% of cells transfected with our Megf10 CAR and incubated with IgG-purified IL-6 were double positive.
3. Megf10 CAR-P (freeze-thaw purification): 12.0% of cells transfected with our Megf10 CAR and co-cultured with IL-6 freeze-thaw lysate were double positive.
4. Fcγ CAR-P (IgG purification): 11.0% of cells transfected with our Fcγ CAR and incubated with IgG-purified IL-6 were double positive.
5. Fcγ CAR-P (freeze-thaw purification): 16.2% of cells transfected with our Fcγ CAR and co-cultured with IL-6 freeze-thaw lysate were double positive.
6. eGFP positive control: As expected, cells transfected only with eGFP did not fluoresce red.

Our four experimental samples tested two different intracellular domains for our CAR-P (Megf10 CAR-P and Fcγ CAR-P) and two different methods of IL-6-mCherry purification (IgG-purified IL-6 or freeze-thaw-purified IL-6), but none of the samples notably exceeded the others in binding efficiency. Therefore, further testing is needed to identify optimal conditions. Despite this, we observed all of our experimental samples to have over 10% of the cell population read as double-positive, a **significant improvement** compared to our negative control with untransfected HEK293 cells, which read 0% double-positive.

The gating strategy used to determine the double-positive population for each sample is visible from our backgating history. In each figure below, the left graph illustrates gating for HEK293 cells and eliminating cell debris; the middle graph illustrates the subset of cells that are GFP positive (CAR+); and the right graph illustrates the subset of CAR+ cells that are mCherry positive (CAR+IL-6).

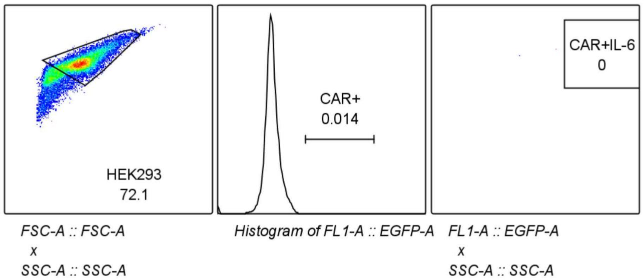


Fig. 7. Negative control backgating extracted using FlowJo

In Figure 7, we used negative control cells to separate cell debris from actual HEK293 cells, assuming that debris is significantly smaller than cells. Thus, we strictly gated for the area around the hotspot and above but discarded the area that tailed underneath. Retroactively, we applied our eGFP+ and mCherry+ gating to the negative control population to verify our gating stringency; as expected, neither eGFP+ nor mCherry+ cell populations were observed in the negative control HEK293 cells.

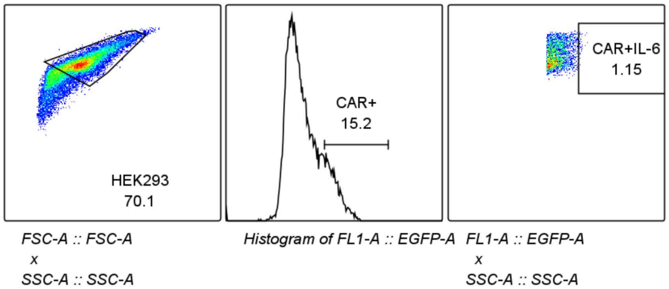


Fig. 8. eGFP control backgating extracted using FlowJo

In Figure 8, we used eGFP-expressing cells as a positive control. We intended to gate for green fluorescing cells (eGFP+) in this population, but since we lacked a distinct second peak representing the eGFP+ population, we used this population to gate for mCherry+ instead; knowing that these eGFP+ cells lacked mCherry expression, we gated for mCherry where there was no fluorescent signal. Similarly, we gated for eGFP+ using the portion of the mCherry-expressing population lacking fluorescence. Retroactively, we verified that the majority of the eGFP-expressing population was correctly gated for eGFP+ and similarly for mCherry+.

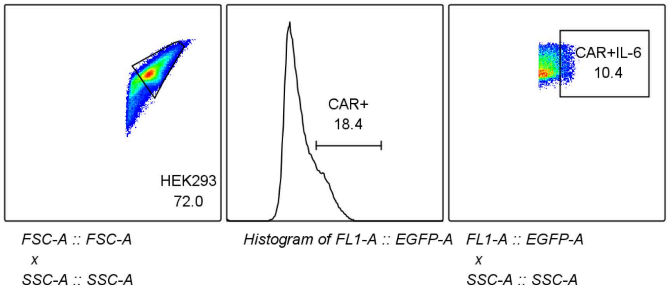


Fig. 9. Megf10 CAR-P with IgG-purified IL-6 backgating extracted using FlowJo

Having gated for green expression, red expression, and no color, we analyzed our four experimental samples. We subsetted the entire population to HEK293 cells, removing debris. Next, we gated for cells that were eGFP+, implying CAR-P expression. We further subsetted this green population to cells that fluoresced red (double-positive), implying CAR-P+IL-6 binding. Figure 9 demonstrates the backgating history for the Megf10 CAR-P co-cultured with IgG-purified IL-6-mCherry.

Conclusion

We have demonstrated that HEK293 cells expressing our CARs bind IL-6 significantly stronger than cells lacking CARs. The success of our CAR-P is supporting evidence that CARs can recognize any small protein as long as an appropriate antibody can be synthesized. Our findings suggest, in part, that it may be possible to utilize CAR therapies for a wide range of diseases caused by abnormal levels of specific proteins by simply adapting CARs for unexplored cell types that activate different responses. Not only can we utilize this protein targeting technique to induce an immune response using white blood cells, but we can also creatively apply it to any cells with behavior we wish to harness. For example, neurons could be engineered to recognize aberrant levels of neurotransmitters and then respond in a manner conducive to ameliorating neurodegenerative diseases. Additionally, platelets could be engineered to recognize aberrant levels of clotting factors and then respond by modulating blood clotting cascades to cure clotting disorders.

A key advantage of synthetic biology is that it enables tuning the sense component of a bioengineered solution and returning the appropriate response. Accordingly, our CAR could be further engineered to induce phagocytosis only when detected IL-6 concentrations are sufficiently elevated. This design would require further research characterizing thresholds for IL-6 concentrations that indicate homeostasis or disease states at various severities. We hope this synthetic circuitry could be incorporated into future experimentation, expanding on our current results, first with in vitro validation of our CAR-P in macrophage cell lines and later with in vivo validation in mouse cancer models. Moreover, if other cytokines or pathways are found to be significant in cachexia, our CAR-P could be coupled to additional CARs with their own specificities. Indeed, the modularity of CAR design and synthetic biology presents a vast set of design possibilities.

Beyond the technical undertaking, we aspired to contextualize the impact of a therapy like ours on a patient population. Specifically, we felt it crucial to understand the landscape of how cachectic patients are currently treated. Conversations with Dr. Teresa Zimmers, President of the Cancer Cachexia Society, have highlighted how the disproportionate emphasis on treating male patients is compounded by the fact that the majority of clinical data relevant to cachexia research draws from a predominantly male pool of subjects. This phenomenon is especially exacerbated since the biological factors contributing to cachexia vary significantly by sex (Zhong, 2022), so our therapeutic design is inherently biased in favor of men. While our therapy is a first step towards utilizing this data for clinical therapy, it also serves to highlight the fundamental flaws of our current understanding of cachexia. Similar consultation with

Dr. Ishan Roy, a researcher and clinician at the Shirley Ryan Ability Lab at Northwestern University, has illuminated how cachexia requires a remarkably complex, holistic treatment encompassing not only medical therapies but also tailored exercise and diet regimens adhering to an individual patient's personal cultural and lifestyle needs. We intend for the time and energy we have focused on developing our CAR-P therapy for cancer cachexia to illuminate the unmet needs of cachectic patients. In the future, we look forward to cancer cachexia receiving the research and clinical attention it deserves.

References

Bird, J. E., Marles-Wright, J., & Giachino, A. (2022). A User's Guide to Golden Gate Cloning Methods and Standards. ACS Synthetic Biology, 11(11), 3551-3563. https://doi.org/10.1021/acssynbio.2c00355

Carson, J. A., & Baltgalvis, K. A. (2010). Interleukin 6 as a key regulator of muscle mass during cachexia. Exercise and Sport Sciences Reviews, 38(4), 168–176. https://doi.org/10.1097/JES.0b013e3181f44f11

Cancer Cachexia: After Years of No Advances, Progress Looks Possible. (2022). NIH National Cancer Institute. https://www.cancer.gov/about-cancer/treatment/research/cachexia.

Lim, S., et al. (2020). Development and progression of cancer cachexia: Perspectives from bench to bedside. Sports Medicine and Health Science, 2(4), 177-185. https://doi.org/10.1016/j.smhs.2020.10.003

Mazinani, M., & Rahbarizadeh, F. (2022). CAR-T cell potency: from structural elements to vector backbone components. Biomarker Research, 10(1), 70. https://doi.org/10.1186/s40364-022-00417-w

Morrissey, Meghan A, et al. (2018). Chimeric Antigen Receptors That Trigger Phagocytosis. eLife, 7. https://doi.org/10.7554/elife.36688

Takeuchi, T., et al. (2017). Sirukumab for rheumatoid arthritis: the phase III SIRROUND-D study. Annals of the Rheumatic Diseases, 76(12), 2001–2008. https://doi.org/10.1136/annrheumdis-2017-211328

Zhong, X., et al. (2022). Sex specificity of pancreatic cancer cachexia phenotypes, mechanisms, and treatment in mice and humans: role of Activin. Journal of Cachexia, Sarcopenia and Muscle, 13, 2146–2161. https://doi.org/10.1002/jcsm.12998

Classifying Genetic Interactions Using an HIV Experimental Study

Sean C. Huckleberry[1], Mary S. Silva[2], Jeffrey A. Drocco[3]

1 Student Contributor, Department of Electrical Engineering and Computer Science, MIT, Cambridge MA 02139

2 Supervisor, Global Security-Computing Applications, Lawrence Livermore National Labs, Livermore CA 94550

3 Principal Investigator, Physical Life Science Directorate, Lawrence Livermore National Labs, Livermore CA 94550

Current methods of addressing novel viruses remain predominantly reactive and reliant on empirical strategies. To develop more proactive methodologies for the early identification and treatment of diseases caused by viruses like HIV and Sars-CoV-2, we focus on host targeting, which requires identifying and altering human genetic host factors that are crucial to the life cycle of these viruses. To this end, we present three classification models to pinpoint host genes of interest, thoroughly analyzing each one's current predictive accuracy, susceptibility to modifications of the input space, and potential for further optimization. Our methods rely on the exploration of different gene representations, including graph-based embeddings and large foundation transformer models, to establish a set of baseline classification models. Subsequently, we introduce an order-invariant Siamese neural network that exhibits more robust pattern recognition with sparse datasets while ensuring that the representation does not capture unwanted patterns, such as the directional relationship of genetic interactions. Through these models, we generate biological features that predict pairwise gene interactions, with the intention of extrapolating this proactive therapeutic approach to other virus families.

Introduction

In the ongoing pursuit of effective antiviral strategies, recent headway has been made in developing proactive therapies to improve public health by preventing or reducing the severity of infections and mitigating transmission among vulnerable groups. One main challenge, however, is identifying a subset of promising genes for proactive host targeting, as comprehensive therapy research is time- and cost-intensive. Once identified, groups of genes could provide a foundation for efficient validation studies and clinical trials. Crucially, these tasks could offer insight into common viral pathways that may be targeted to disrupt viral infection, which may be relevant to a broader group of viruses of interest. Here, we employ three predictive models and compare their efficacy in addressing this obstacle with a focus on Human Immunodeficiency Virus (HIV). Our interest in HIV stems from its well-characterized genetic mechanisms and abundant genetic data, which provide a robust framework for probing pertinent viral-host interactions.

Our models rely heavily on genetic pairwise epistasis: the interactions between pairs of genes. To model epistasis, we introduce two main featurization approaches. The first is based on graphical genetic relationships between genes, biological processes, pathways, and cellular components. These graph relationships are cached in the Scalable Precision Medicine Oriented Knowledge Engine (SPOKE), a "database of databases" comprised of approximately 20 thousand human genes and over 1 million gene expression and regulation edge types (Morris, 2023). The second method uses Geneformer, a foundation transformer model pre-trained on 30 million single-cell transcriptomes (Theodoris, 2023). We utilize the embedding outputs from this pre-trained model and perform fine-tuning with a task-specific neural network classifier to predict genetic epistasis. Using these featurization approaches, we present in the form of models a technique for identifying vital genes for host targeting, which we anticipate may be adapted for the treatment of other viruses beyond HIV.

Data Acquisition Methods

Featurization Approches

I. Gene Representations Using Graph Embeddings

Our graph-based featurization approach employs the University of San Francisco's Scalable Precision Medicine Oriented Knowledge Engine (SPOKE), a graphical network of biomedical databases that has been used for drug repurposing (Himmelstein, 2017), gene regulation in lung cells for COVID-19 patients (Huang, 2020), and numerous other link prediction tasks. This knowledge graph consists of over 20 thousand protein-coding human gene nodes from Entrez Gene (NCBI's database for gene-specific information) (Maglot, 2011), a combined 18 thousand biological process, molecular function, and cellular component node types derived from the Gene Ontology database (Dessimoz & Škunca, 2017). These nodes contain comprehensive gene information, including gene nomenclature, function and attributes, sequences, and source data. Additionally, the graph includes over 1 million gene relationships, where the knockdown or knockout of one gene, achieved either by short hairpin RNA or CRISPR-Cas9, results in the upregulation or downregulation of another gene as indicated by consensus transcriptional profiles. Using various subsets of nodes, we apply the Fast Random Projection (FastRP) graph embedding, a sparse random projection-based graph embedding algorithm (Chen, 2019). With FastRP, we represent each gene node in the SPOKE graph as a 256-dimensional numeric vector. After extracting the embeddings associated with the 356 genes associated with HIV function (which match our validation set discussed below), we use the concatenated form as inputs to model corresponding gene-pair epistasis. This allows us to explore the relationship between gene embedding pairs and their corresponding interactions.

II. Gene Representations Using Geneformer Embeddings

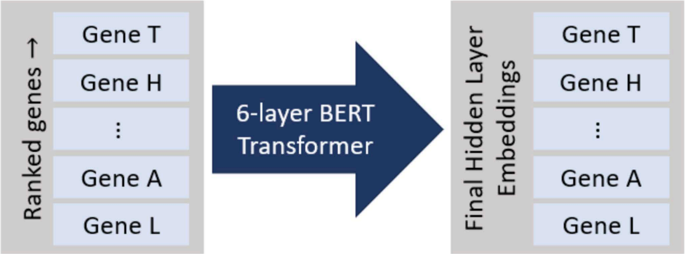


Fig. 1. Geneformer architecture and feature extraction. Each layer within the Geneformer BERT Transformer has four attention heads that individually learn to monitor distinct classes of genes and improve predictive power.

Pretrained on the massive Genecorpus-30M, which consists of 30 million single-cell transcriptomes, Geneformer contains comprehensive information in its embeddings that opens new avenues for computation. Our second featurization approach involves employing these embeddings to perform inference regarding their impact on the epistatic gene interactions that contribute to HIV.

In the standard Geneformer architecture, each single-cell transcriptome is assigned a rank value encoding: genes are ranked by their expression in each cell normalized by their expression across the entire Genecorpus. This encoding system prioritizes genes that differentiate cell state and shrinks housekeeping genes with limited distinguishing potential (Theodoris, 2023). The rank value genes are then passed into a six-layer BERT transformer encoding unit (Delvin, 2019), and the model is pre-trained using a context-aware masked learning objective (Figure 1). We utilize the pre-trained model by extracting the final hidden layers of the ranked genes, which are 256-dimensional embeddings, and using them to compare both featurization methods.

Model Validation

As our predictive models harness the HIV infection metric as the target variable, an epistasis mapping containing 63,012 pairwise interactions closely linked to HIV function serves for validation. Generated as part of a recent study concerning the quantitative mapping of genetic interactions, an HIV epistasis matrix known as a vE-MAP (viral epistatic miniarray profile) was produced by the pairwise depletion of 356 human genes closely linked with HIV. More specifically, this vE-MAP approach involves the construction of miniarrays containing HIV and uniquely perturbed culture cells, and the resulting degree of infection of each cultured cell was captured and documented in the 356 by 356 symmetric epistatic matrix (Gordon, 2020). We focus on the upper triangular entries of the symmetric epistasis matrix, which serve as the response variables for our machine learning models. In scenarios where our objective is solely to classify HIV enhancement or suppression, we establish a threshold and formulate the response as a binary outcome.

Discussion and Results of Predictive HIV Models

Contextualizing Results through Response Binarization

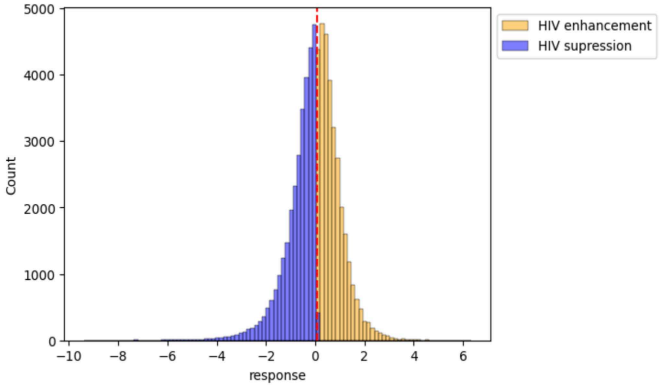


Fig. 2. A histogram representing the epistasis between each pair of genes in the 356 by 356 matrix. The tails represent HIV enhancement on the positive end and suppression at the negative end. Most pairs of genes show no epistasis response at the center of the distribution. The red line indicates the threshold used for binary classification.

For each gene pair, the results of our predictive models are formatted as a binary response indicating HIV suppression or enhancement. To this end, we determine an appropriate threshold for the HIV infection metric and categorize response into two groups. Establishing the threshold as the mean of the response, we achieve a balanced split between suppression and enhancement, mitigating concerns related to imbalance (Figure 2). Of particular interest are gene pairs associated with HIV suppression, which may be crucial to developing host-targeting therapies.

FastRP Embedding Classification

The FastRP graph-based approach enables us to capture a different feature space from the Geneformer embeddings by focusing on a subset of genes and their relationships within the larger SPOKE knowledge graph. Utilizing the SPOKE and FastRP embeddings, we explore the "gene" node types, which encompass approximately 20 thousand human genes and contain gene names, descriptions, abbreviations, and source information. Similarly, we investigate the "biological process" node types, which consist of roughly 12 thousand nodes defining biological objectives to which a gene or gene product contributes. There is a discrete, fixed number of these nodes defined within Gene ontology, so the relationships between

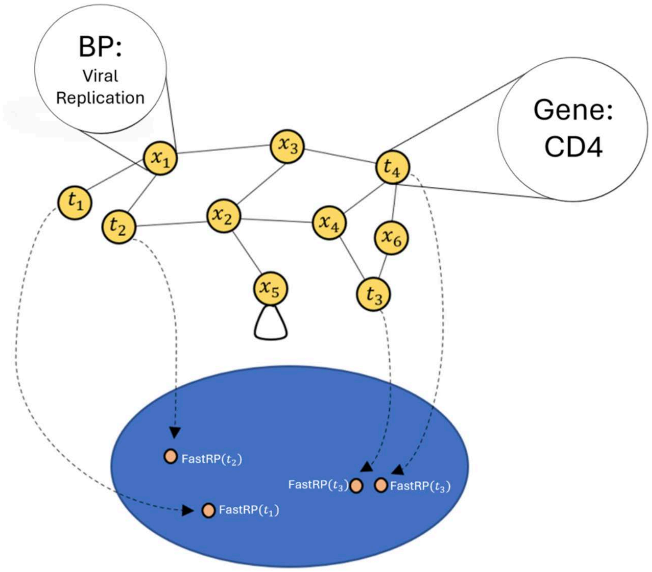


Fig. 3. Graph-like representation of the data utilized by the FastRP predictive model. Target nodes t_i represent each of the 356 "gene" nodes from our validation set. Our model seeks to capture the latent representation of these target nodes using fast random projection embeddings FastRP(ti) and information regarding the neighbors in "biological process" nodes x_j.

these node types and the genes are unique and finite. Figure 3 illustrates how we utilize each node type to predict a response.

We begin by defining the binary response threshold, then shuffling and splitting the data so that 70% is used for training and 30% is reserved for validation. Subsequently, a random forest classifier was trained on the data to establish a predictive model, given the method's exceptionally robust performance for datasets with many features (Kirasich, 2018). The model was then evaluated on the test set to assess its performance and generalization capabilities.

TABLE I
FASTRP EMBEDDINGS RANDOM FOREST MODEL

		Enhancement	Suppression
Predicted	Enhancement	36.01%	13.99%
	Suppression	15.92%	34.08%
		Actual	

Table I. Confusion matrix representing the performance of random forest classifier trained on FastRP embeddings.

This first model, which harnesses the concatenated FastRP embeddings and binarized epistatic response, obtains a prediction accuracy of approximately 70% (Table I). Results falling along the off-diagonals represent incorrect predictions about gene pair epistasis. We seek to minimize the false positives, in which a gene pair is erroneously classified as exhibiting HIV suppression despite actually promoting HIV enhancement. These data points constitute a subset of gene pairs whose presence would likely be damaging rather than beneficial to a host, and we strive to exclude them from the output of our models.

Geneformer-based Classification

We now explore the Geneformer embeddings using the same train-test splits and Random Forest parameters from the FastRP model. Notably, the Geneformer embeddings under analysis were derived from the final layer, which is recognized for capturing features more closely associated with the learning objective of the pre-trained task. Despite the lack of fine-tuning, both featurization approaches lead to fair prediction accuracy.

TABLE II
GENEFORMER EMBEDDINGS RANDOM FOREST MODEL

		Enhancement	Suppression
Predicted	Enhancement	35.98%	14.02%
	Suppression	15.87%	34.13%
		Actual	

Table II. Confusion matrix representing the performance of random forest classifier trained on Geneformer embeddings.

The prediction accuracy of the Geneformer and FastRP models is nearly identical (Table II). We conclude that both featurization approaches offer similar predictive performance as the training and testing parameters were kept constant.

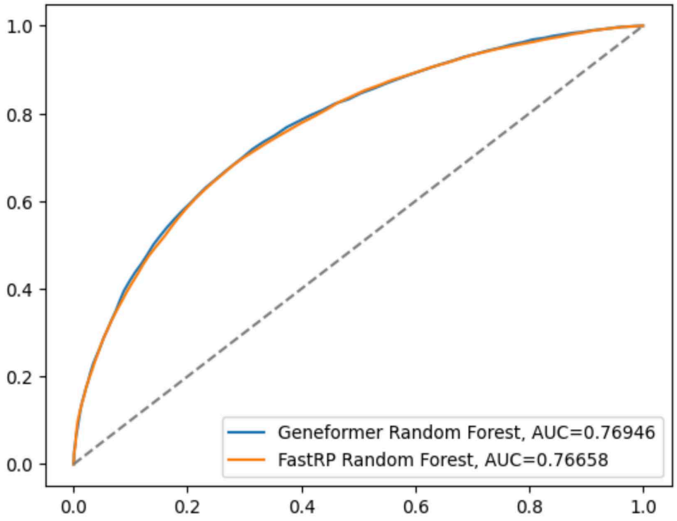


Fig. 4. ROC/AUC curve comparing the predictive performance of the Geneformer and FastRP models. The dotted line represents random chance, and the "perfect classifier" reaches the top left corner of the graph. The overall performance of each model can be summarized as a scalar: the area under the curve (AUC).

This observation is confirmed by comparing the ROC/AUC of the two models, plotting true positive against false positive rates (Figure 4). Notably, Geneformer embeddings demonstrate only marginally superior performance compared to the FastRP embeddings.

Symmetry and Order Invariance

Order invariance is vital for our epistasis modeling that utilizes a symmetric binary input of gene pairs since it

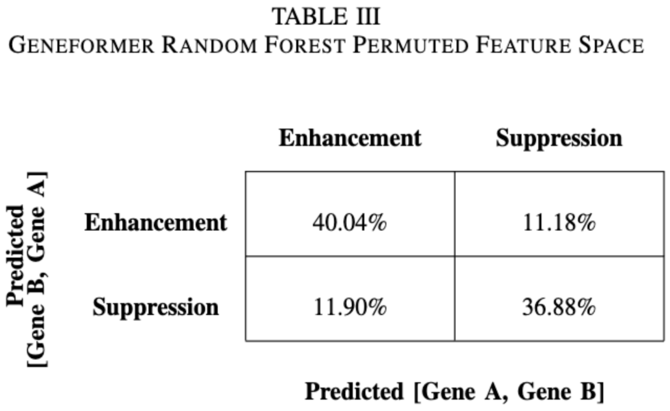


TABLE III
GENEFORMER RANDOM FOREST PERMUTED FEATURE SPACE

Predicted [Gene B, Gene A]	Enhancement	Suppression
Enhancement	40.04%	11.18%
Suppression	11.90%	36.88%

Predicted [Gene A, Gene B]

Table III. Confusion matrix comparing the binarized output of symmetric inputs for random forest classifier trained on Geneformer embeddings. For an order-invariant model, one would expect that the values of the off-diagonals are close to zero.

indicates that any trends learned by the model reflect the biological interaction between genes and are not artifacts of changes in the data representation. However, both the FastRP embedding classification and Geneformer-based classification models ignore the order of embedding concatenation. In other words, [EmbeddingGeneA, EmbeddingGeneB] and [EmbeddingGeneB, EmbeddingGeneA] may yield inconsistent results despite being pairwise symmetrical. To demonstrate this, we compare the predicted results of the Geneformer embedding with the same training inputs and random forest parameters but swapped concatenation order of the observations in the test set. Swapping the concatenation order of gene pairs resulted in a 23% inconsistency between predictions (Table III). This reveals the existence of a disagreement in the prediction of the same observations when the concatenation order is swapped, indicating that the model has captured extraneous patterns during training. We address this concern by implementing a Siamese network.

Siamese Classification Network

We implement a Siamese network to ensure cohesion between gene pairs (Figure 5). This type of network better correlates symmetric gene pairs and introduces order-invariance to our model. Siamese neural networks have been studied in the context of machine learning for classifying similarities or dissimilarities between image inputs as well as signature verification (Malhotra, 2023). They have also been used for protein-protein interaction classification problems where the input space is the embeddings extracted from protein sequence transformer models, commonly referred to as ProtBERT (Madan, 2022). Since we determined that FastRP and Geneformer produce similar results, we focus on the Geneformer embeddings to implement this network.

The Siamese network is significantly more sensitive to hyperparameter tuning than the random forest models as it consists of two identical neural networks, where the inputs are the corresponding gene embeddings. Much like a standard feed-forward neural network, the number of hidden layers and corresponding weights and biases, and the choice of the optimizer, activation functions, and regularization techniques can significantly impact the performance and accuracy of a Siamese neural network. Initially, we define the two branches of our Siamese network to have three hidden layers of size 512, 128, and 64. We select a stochastic gradient descent optimizer, which is memory-efficient and computationally inexpensive relative to other optimizers. The contrastive layer, used to measure the similarity of inputs, is defined as a concatenation of the layers of the two Siamese network branches and their Euclidean distance with an additional linear layer. Other contrastive layers to explore include a function of the Euclidean

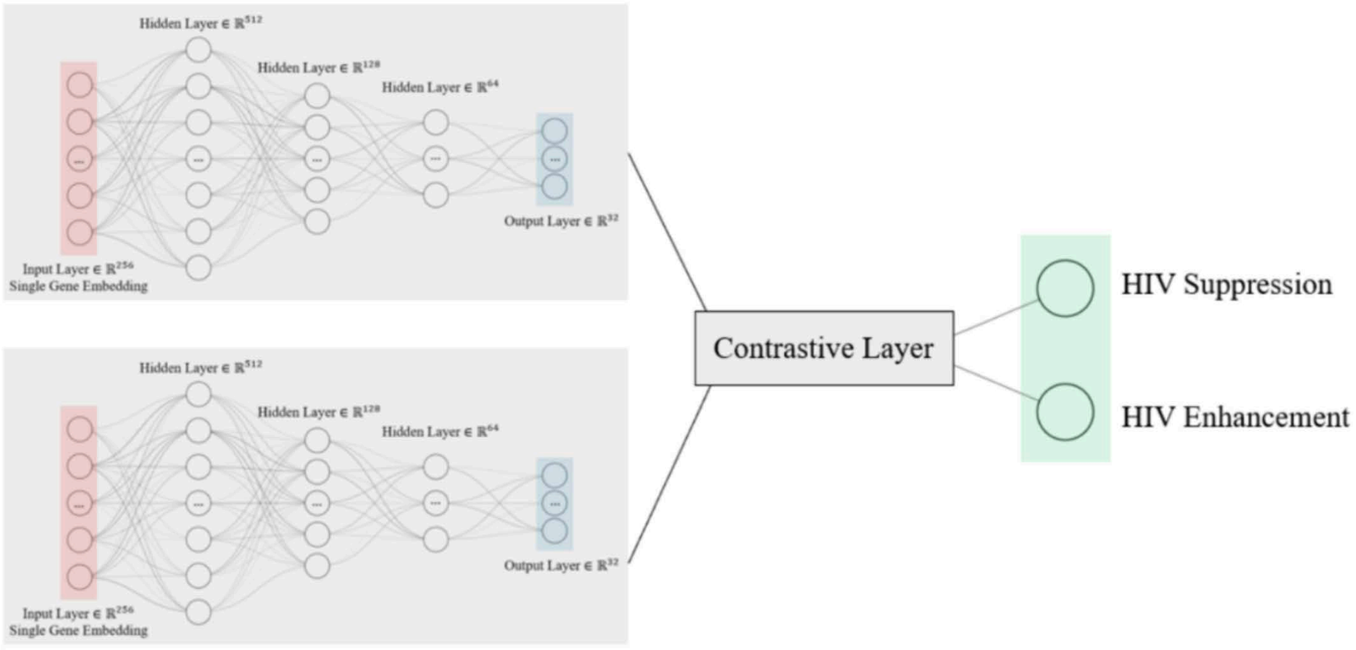

Hidden Layer ∈ $\mathbb{R}^{512}$
Hidden Layer ∈ $\mathbb{R}^{128}$
Hidden Layer ∈ $\mathbb{R}^{64}$
Output Layer ∈ $\mathbb{R}^{32}$
Input Layer ∈ $\mathbb{R}^{256}$
Single Gene Embedding
Hidden Layer ∈ $\mathbb{R}^{512}$
Hidden Layer ∈ $\mathbb{R}^{128}$
Hidden Layer ∈ $\mathbb{R}^{64}$
Output Layer ∈ $\mathbb{R}^{32}$
Input Layer ∈ $\mathbb{R}^{256}$
Single Gene Embedding
Contrastive Layer
HIV Suppression
HIV Enhancement

Fig. 5. Siamese network architecture. The model includes two identical neural networks applied independently to each gene embedding. The output layers of the two networks are then fed into a contrastive layer, which produces a single output containing information about the final similarity or dissimilarity between the input pairs.

distance or a function of the cosine similarity. Due to the large scale of the data inputs and complexity of the Siamese neural network, mini-batch training and PyTorch GPU acceleration were leveraged.

TABLE IV
Geneformer Embeddings + Siamese Neural Network

		Predicted	
		Enhancement	**Suppression**
Actual	**Enhancement**	35.41%	14.59%
	Suppression	14%	36%

Table IV. Confusion matrix comparing the binarized output of symmetric inputs for a Siamese network trained on Geneformer embeddings.

Beyond improving the model's consistency for symmetric gene pairs, the Siamese network, trained on the Geneformer embeddings, also slightly enhanced the model's predictive accuracy from approximately 70% to over 71% without additional fine-tuning (Table IV). In contrast with the previous two models, which have limited improvement potential from the current state, this Siamese network could see a vast improvement in predictive accuracy from tweaking the hyperparameters, either manually or with an optimization framework like Optuna.

TABLE V
Geneformer Embeddings + Siamese Neural Network
Permuted Feature Space

		Predicted [Gene A, Gene B]	
		Enhancement	**Suppression**
Predicted [Gene B, Gene A]	**Enhancement**	49.41%	0%
	Suppression	0%	50.59%

Table V. Confusion matrix comparing binarized output of symmetric inputs for Siamese network trained on Geneformer embeddings. The off-diagonals are zero, as expected for an order-invariant model.

The main benefit of applying a Siamese Neural Network to the Geneformer embeddings is that it solves the problem of feature input order invariance. Unlike the first model, which saw approximately 23% disagreement of the test set prediction when the order of the gene embeddings was swapped, the Siamese Neural Network now has 100% agreement (Table V). This also indicates that the model is not unintentionally learning patterns associated with the order of the inputs.

Conclusion

To better understand host-targeting viral therapies, three models were trained to classify gene pairs as inducing HIV suppression or enhancement. The first classification model explored the FastRP embeddings from UCSF's SPOKE database with a graph-based approach and achieved a prediction accuracy of 70% with a random forest algorithm. The next model, which achieved similar performance, utilized the same algorithm and parameters, but embeddings were extracted from the pre-trained Genformer foundation transformer instead. The first pair of models employ a random forest classifier, utilizing the FastRP and Geneformer embeddings as individual inputs without undergoing fine-tuning. These models perform adequately when trained on a sizable subset of data. However, the potential for significant improvement is limited due to the inherent characteristics of random forests. Random forests rely on decision trees, which independently make decisions based on individual features and do not provide much flexibility in capturing non-linear relationships and other hierarchical relationships between the input space. More importantly, with the concatenated representation of the input feature space, they fail to capture the symmetric relationship between genetic interaction.

The third model employed a Siamese network to the Geneformer embeddings, which introduced order-invariance to better handle input symmetry and demonstrated significant optimization potential through fine-tuning. By inferring gene pairs inducing HIV suppression, these models identify the most promising pairs for HIV therapies to study through more resource-intensive methods. The next steps include applying these models to different diseases and sparse datasets where they might be most informative. Regarding genetic interaction, we have only explored these models under a binary classification setting; a stricter multi-class response could include a third neutral class. The overall accuracy may be reduced, but using a multi-class response would give a more realistic identification of genetic interactions, as it would produce more nuanced results.

This experimental study underscores the critical role of various computational models in adeptly predicting HIV's associated human host factors through rigorous pattern matching of large-scale gene data. The significance of this work is manifold: it further establishes data and computation as a method to expedite and reduce the cost of traditional research methods in biology, and suggests that predictive models can enable us to take a more proactive approach to understanding viral infections.

Acknowlegement

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-JRNL-860169-DRAFT).

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

References

Morris, J. H., et al. (2023). The scalable precision medicine open knowledge engine (SPOKE): A massive knowledge graph of biomedical information. Bioinformatics, 39(2), btad080. https://doi.org/10.1093/bioinformatics/btad080.

Theodoris, C. V., Xiao, L., Chopra, A., et al. (2023). Transfer learning enables predictions in network biology. Nature, 618, 616–624. https://doi.org/10.1038/s41586-023-06139-9.

Himmelstein, D. S., et al. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife, 6, e26726. https://doi.org/10.7554/eLife.26726.

Huang, S., Kaipainen, A., Strasser, M., & Baranzini, S. (2020). Mechanical Ventilation Stimulates Expression of the SARS-Cov-2 Receptor ACE2 in the Lung and May Trigger a Vicious Cycle. Preprints, 2020050429. https://www.preprints.org/manuscript/202005.0429/v1.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research, 39, D52–D57. https://doi.org/10.1093/nar/gkq1237.

Gene Ontology Consortium. (2007). The Gene Ontology project in 2008. Nucleic Acids Research, 36 (Database issue), D440-4. https://doi.org/10.1093/nar/gkm883.

Dessimoz, C., & Škunca, N. (Eds.). (2017). The Gene Ontology Handbook. Methods in Molecular Biology, 1446.

Chen, H., Sultan, S. F., Tian, Y., Chen, M., & Skiena, S. (2019). Fast and Accurate Network Embeddings via Very Sparse Random Projection. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19). ACM, New York, NY, USA, 399–408. https://doi.org/10.48550/arXiv.1908.11512

Devlin, J. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics. Gordon, D. E. (2020). A quantitative genetic interaction map

A decade of experience in transforming processes from batch to continuous
CONTINUUS™
pharmaceuticals
continuuspharma.com
info@continuuspharma.com

of HIV infection. Mol. Cell, 78(2), 197–209.e7. https://doi.org/10.48550/arXiv.1810.04805.

Gordon, D. E. (2020). A quantitative genetic interaction map of HIV infection. Molecular Cell, 78(2), 197–209.e7. https://doi.org/10.1016/j.molcel.2020.02.004.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review, 1(3), Article 9.

Malhotra, A. (2023). Single-Shot Image Recognition Using Siamese Neural Networks. In Proceedings of the 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2550-2553.

Madan, S., Demina, V., Stapf, M., Ernst, O., & Fröhlich, H. (2022). Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings. Patterns, 3(9), 100551. https://doi.org/10.1016/j.patter.2022.100551.

Meganck, R. M., & Baric, R. S. (2021). Developing therapeutic approaches for twenty-first-century emerging infectious viral diseases. Nature Medicine, 27(3), 401-410. https://doi.org/10.1038/s41591-021-01282-0.

Better Health, Brighter Future

Takeda is a global, R&D-driven biopharmaceutical company committed to discovering and delivering life-transforming treatments and vaccines that have a lasting impact on society.

Since our founding in 1781 in a market stall in Osaka, Japan, our values endure by putting patient needs first, building trust with society, strengthening our reputation, and developing the business - in that order.

www.takeda.com

Takeda