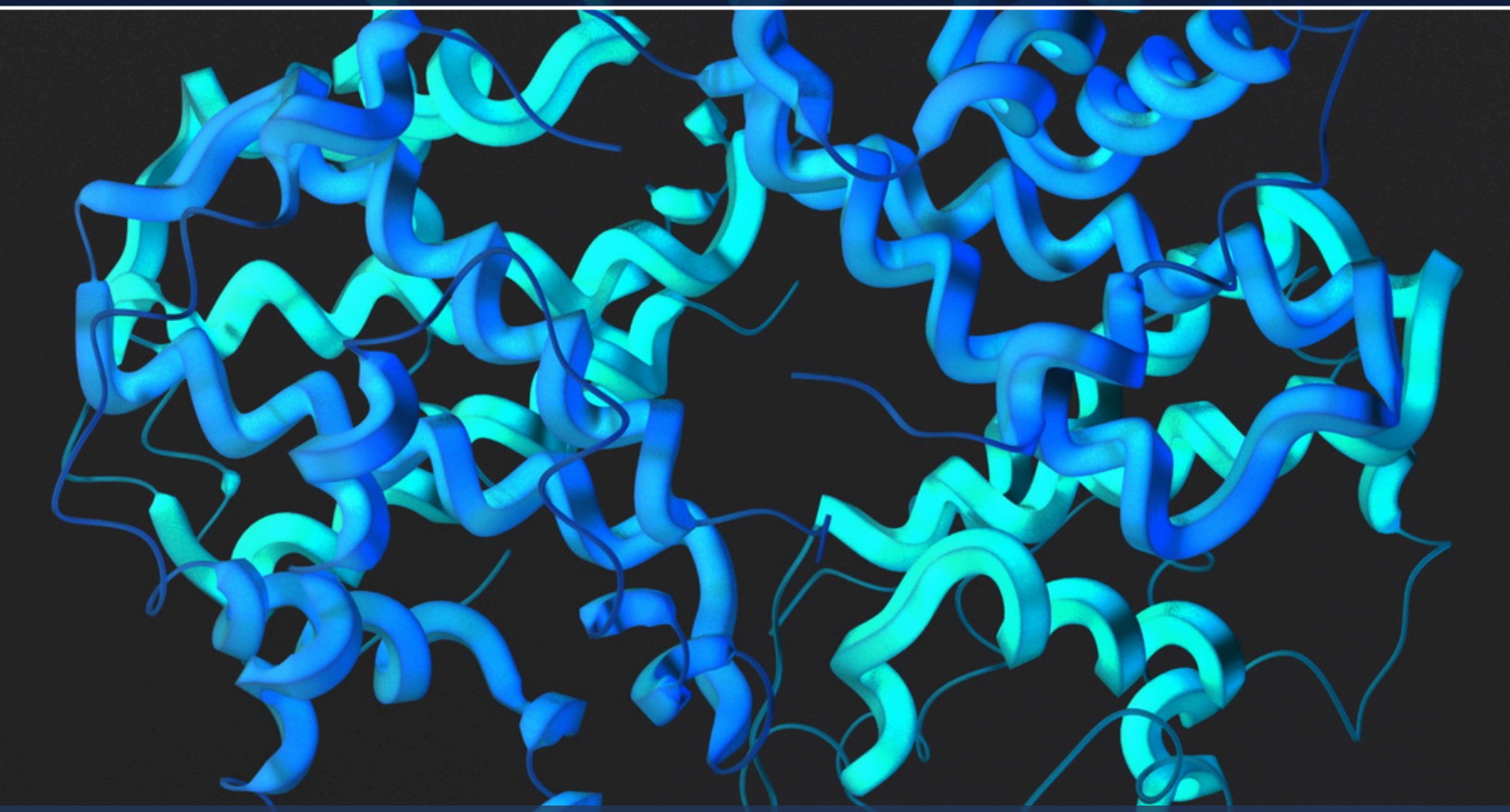
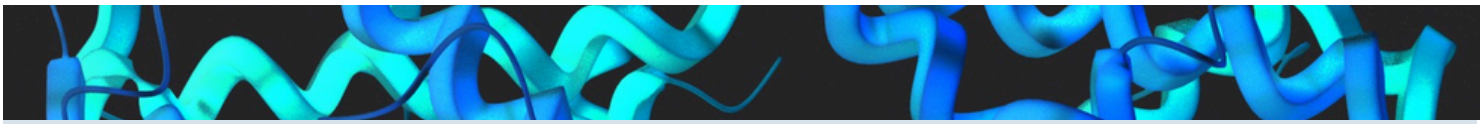
VOL. 5 | NO. 1 | 2026

OSURJ

UNIVERSITY OF OTTAWA SCIENCE UNDERGRADUATE RESEARCH JOURNAL
JOURNAL D'ÉTUDIANT DE RECHERCHE SCIENTIFIQUE DE L'UNIVERSITÉ D'OTTAWA





JRSUO
JOURNAL D'ÉTUDIANT DE RECHERCHE SCIENTIFIQUE DE L'UNIVERSITÉ D'OTTAWA

OSURJ
UNIVERSITY OF OTTAWA SCIENCE UNDERGRADUATE RESEARCH JOURNAL

À PROPOS DE NOUS

Le Journal d'étudiant de recherche scientifique de l'Université d'Ottawa est une revue bilingue, évaluée par les pairs et en libre accès, consacrée à la promotion de la recherche scientifique au premier cycle universitaire. Publiée et dirigée par des étudiants de l'Université d'Ottawa, la revue présente des travaux de recherche originaux, des articles de synthèse et des perspectives couvrant les sciences naturelles, de la santé et quantitatives. Le JÉRSUO conjugue un processus rigoureux d'évaluation par les pairs à une approche de publication académique axée sur le mentorat afin d'offrir une plateforme accessible aux chercheurs émergents du Canada et d'ailleurs.

ÉNONCÉ DE LA MISSION

Le JRSUO cherche à enrichir l'expérience scientifique au premier cycle universitaire en offrant aux étudiants une plateforme leur permettant de rédiger, d'évaluer et de publier des travaux de recherche académique. En impliquant les étudiants dans toutes les étapes du processus de publication scientifique, le JRSUO favorise la rigueur intellectuelle, l'esprit critique et une communication scientifique efficace. Par cette initiative, la revue vise à cultiver un engagement durable envers la recherche, la collaboration et la découverte au sein de la communauté scientifique.

SUR LA COUVERTURE

Couverture adaptée de S*tructure cristalline de l'hémoglobine* par Anirudh sur Unsplash, sous licence Unsplash ; couleurs modifiées.

ABOUT US

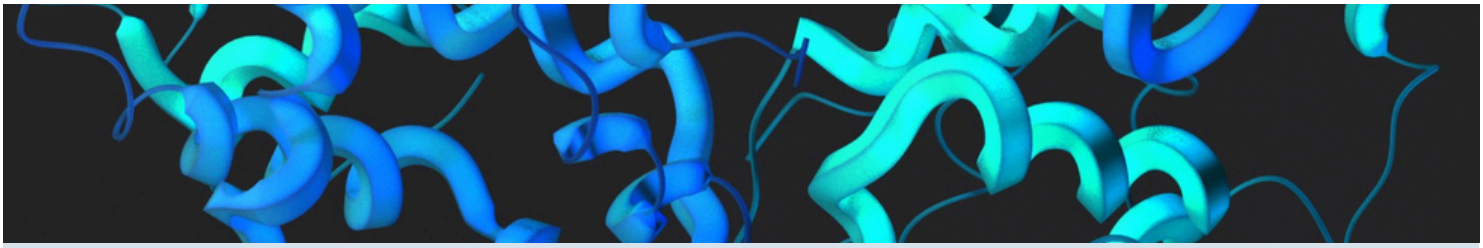
The University of Ottawa Science Undergraduate Research Journal is a bilingual, peer-reviewed, open-access journal dedicated to advancing undergraduate scientific scholarship. Published and managed by students at the University of Ottawa, the journal features original research, reviews, and perspectives spanning the natural, health, and quantitative sciences. OSURJ combines rigorous peer review with mentorship-driven academic publishing to provide an accessible platform for emerging researchers across Canada and beyond.

MISSION STATEMENT

OSURJ seeks to enrich the undergraduate science experience by providing students with a platform to write, review, and publish scholarly research. By engaging students in all stages of the scientific publishing process, OSURJ fosters rigorous inquiry, critical thinking, and effective scientific communication. Through this initiative, the journal aims to cultivate a lasting commitment to research, collaboration, and discovery within the scientific community.

ON THE COVER

Cover adapted from *Crystal hemoglobin structure* by Anirudh on Unsplash, licensed under the Unsplash license; Colours modified.



JRSUO OSURJ

JOURNAL D'ÉTUDIANT DE RECHERCHE SCIENTIFIQUE DE L'UNIVERSITÉ D'OTTAWA

UNIVERSITY OF OTTAWA SCIENCE UNDERGRADUATE RESEARCH JOURNAL

University of Ottawa Science Undergraduate Research Journal 2025 - 2026 Executive Team

Ishaan S. Goswami
Co-Editor-in-Chief

Sivany Kathir
Co-Editor-in-Chief

Ayman Assaaoudi
Lead Translator

Isra F. Omar
Lead Copy Editor

Mars Wichmann-Young
Lead Layout Editor

Hafsa Ahmed
Associate Reviewer

Ahona Deb
Associate Reviewer

Zoha Fatima
Associate Reviewer

Seb Parmasad
Associate Reviewer

Bilal Siddiqi
Associate Reviewer

Faiz Nameer Ahmed
Copy Editor

Haider Ikram
Copy Editor

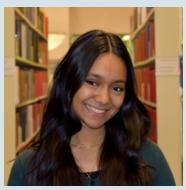
Daksh Maini
Copy Editor

Abraxas Petit
Copy Editor

Shreya Pal
VP External

Marie Babineau
Translator

Chloé Hajjar
Translator



Tabassum Howlader
VP Social Media

Jenna Abu-Dieh
Managing Editor

Varna Prapakaran
Managing Editor

Editorial Team | Comité de rédaction

Editorial Team | Comité de rédaction

Contributors | Contributeurs

Ayman Assaaoudi, University of Ottawa, Ottawa, ON, Canada
Maximiliano Araneda Suárez, University of Ottawa, Ottawa, ON, Canada
Sharon Barden, University of Ottawa, Ottawa, ON, Canada
Olivia Chen, Merivale High School, Ottawa, ON, Canada
Joseph Costello, University of Ottawa, Ottawa, ON, Canada
Ahona Deb, University of Ottawa, Ottawa, ON, Canada
Zoha Fatima, University of Ottawa, Ottawa, ON, Canada
Parastoo Golzarian, Toronto Metropolitan University, Toronto, ON, Canada
Ishaan S. Goswami, University of Ottawa, Ottawa, ON, Canada
Maïka Harvey, University of Ottawa, Ottawa, ON, Canada
Tabassum Howlader, University of Ottawa, Ottawa, ON, Canada
Victor Lee, University of Ottawa, Ottawa, ON, Canada
Sum Ki Kelsie Ling, University of Ottawa, Ottawa, ON, Canada
Jack Madden, University of Ottawa, Ottawa, ON, Canada
Paul M. Mayer, University of Ottawa, Ottawa, ON, Canada
Isra F. Omar, University of Ottawa, Ottawa, ON, Canada
Shreya Pal, University of Ottawa, Ottawa, ON, Canada
Natasha Saltarelli, University of Ottawa, Ottawa, ON, Canada
Zoya Sharma, Liberal Arts and Science Academy, Austin, TX, United States of America
Josiah A. W. Smith, University of Ottawa, Ottawa, ON, Canada
Cindy Yao, University of Ottawa, Ottawa, ON, Canada
Annie Xiang, University of Ottawa, Ottawa, ON, Canada
Pegah Yousefirad, University of Ottawa, Ottawa, ON, Canada
Kamron Yunusov, University of Ottawa, Ottawa, ON, Canada

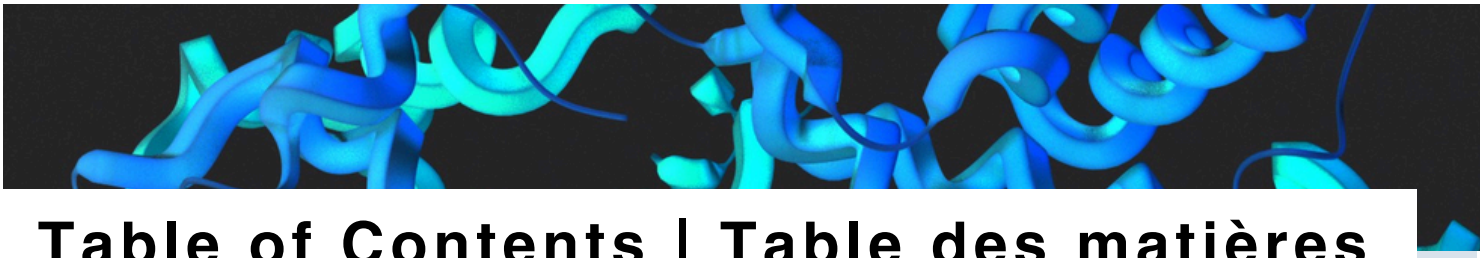
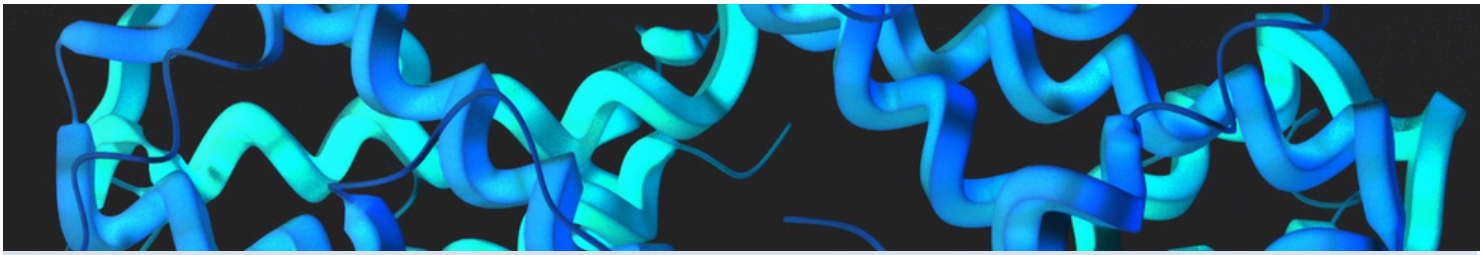
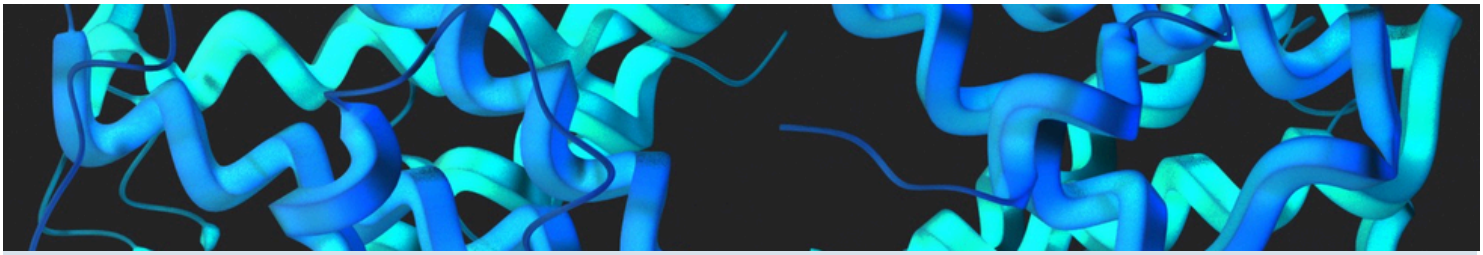


Table of Contents | Table des matières

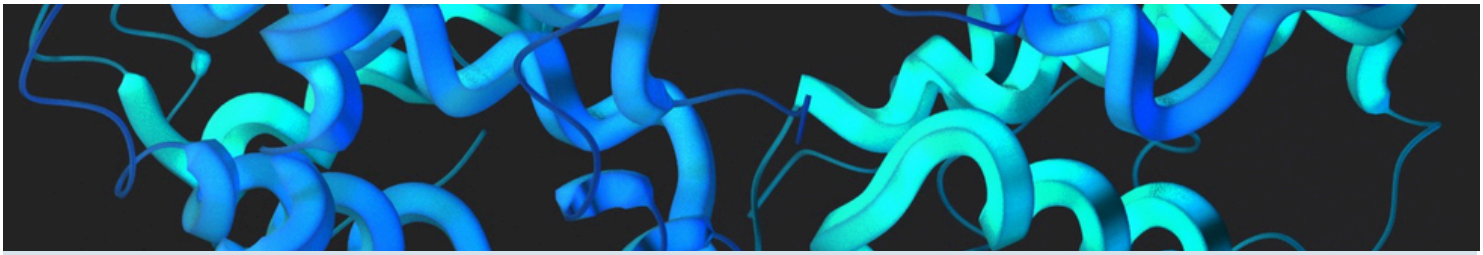
Original Research | Recherche originale



Reviews | Articles de revue



Commentaries | Commentaires



Foreword

Dear Reader,

Scientific research is driven by an instinct for discovery and exploration. Science is a field that is constantly evolving. Axioms once held as the truth are constantly scrutinized as new evidence is discovered. Undergraduate study allows for this curiosity to become discipline, where questions about the world are refined into methods. Research allows students to become contributors to knowledge, rather than just consumers of it.

This issue of the University of Ottawa Science Undergraduate Research Journal reflects this transition. The works collected here span disciplines, methodologies, and perspectives, but share a common thread: a commitment to rigorous inquiry and intellectual growth. From experimental sciences to interdisciplinary analyses, each article represents the persistence, creativity, and critical thinking that define undergraduate scholarship at the University of Ottawa and beyond.
As a student-run journal, OSURJ exists because of a collective effort. We are deeply grateful to our authors for trusting us with their work, to our reviewers and editors for their care and precision, and to our faculty mentors who continue to support undergraduate research.

Over the past year, OSURJ has evolved into a steadily growing platform for undergraduate scholarship. We have seen a marked increase in submissions, broader disciplinary representation, and an expansion of our editorial team, reflecting a growing confidence in student-led academic publishing. Alongside this expansion, we formalized a double-blind peer-review process, refined our editorial workflows, and strengthened our review team to ensure that growth is matched by rigour. These changes mark an important transition for the journal: from a developing initiative to a more structured and sustainable venue for undergraduate research.

Growth does not simply mean expansion; it demands refinement. As we look ahead, our focus is not only on continuing to expand the reach of OSURJ but on deepening quality and accessibility. We aim to further strengthen our review standards, broaden disciplinary representation, and build a culture in which undergraduate researchers feel supported in bringing their work into the public academic sphere. With a stronger editorial foundation in place, OSURJ is well positioned to continue evolving as a collaborative space for thoughtful, student-driven scholarship.

We hope this volume not only showcases the breadth of undergraduate research conducted at the University of Ottawa but also encourages more students to see their questions as worthy of investigation and publication. Inquiry begins with curiosity, and journals like OSURJ exist to give that curiosity a home. Cover article selections were conducted through a blinded editorial review process. Editorial board members were recused from decisions involving their own submissions.
On behalf of the editorial team, we are proud to present to you Volume V issue i and thank you for reading and for supporting undergraduate research.

Ishaan Goswami & Sivany Kathir
Co-Editors-in-Chief
University of Ottawa Science Undergraduate Research Journal

Avant-propos

Chère lectrice, cher lecteur,

La recherche scientifique est guidée par le désir de découvrir, de comprendre et d'explorer. La science est un domaine qui évolue constamment. Des idées autrefois considérées comme des vérités peuvent être remises en question lorsque de nouvelles preuves apparaissent. Les études de premier cycle permettent à cette curiosité de devenir plus structurée : les questions que les étudiants se posent sur le monde se transforment peu à peu en démarches de recherche. Grâce à la recherche, les étudiants ne font pas seulement qu'apprendre des connaissances déjà existantes; ils peuvent aussi contribuer à en créer de nouvelles.

Ce numéro du Journal de recherche de premier cycle en sciences de l'Université d'Ottawa reflète bien cette transition. Les travaux présentés ici touchent à plusieurs disciplines, méthodes et perspectives, mais ils ont tous un point commun : ils montrent un engagement envers une recherche rigoureuse et un développement intellectuel. Des sciences expérimentales aux analyses interdisciplinaires, chaque article témoigne de la persévérance, de la créativité et de l'esprit critique qui caractérisent la recherche de premier cycle à l'Université d'Ottawa et ailleurs.

En tant que revue dirigée par des étudiants, le JRSUO existe grâce à un effort collectif. Nous sommes profondément reconnaissants envers les auteurs, qui nous ont confié leurs travaux, envers les réviseurs et les éditeurs, qui ont fait preuve de soin et de précision, ainsi qu'envers les mentors professoraux, qui continuent de soutenir la recherche de premier cycle.

Au cours de la dernière année, le JRSUO est devenu une plateforme de plus en plus importante pour la recherche de premier cycle. Nous avons observé une augmentation du nombre de soumissions, une plus grande diversité de disciplines représentées et un élargissement de notre équipe éditoriale. Ces changements montrent une confiance grandissante envers la publication académique dirigée par des étudiants. En parallèle, nous avons officialisé un processus d'évaluation par les pairs en double aveugle, amélioré nos méthodes de travail éditoriales et renforcé notre équipe de révision afin que cette croissance s'accompagne aussi d'un haut niveau de rigueur. Ces progrès marquent une étape importante pour la revue, qui passe d'une initiative en développement à un espace plus structuré et durable pour la recherche de premier cycle.

La croissance ne signifie pas seulement s'agrandir; elle demande aussi de s'améliorer. Pour l'avenir, notre objectif n'est pas seulement d'élargir la portée du JRSUO, mais aussi d'en renforcer la qualité et l'accessibilité. Nous souhaitons continuer à améliorer nos normes de révision, représenter un plus grand nombre de disciplines et créer un environnement où les chercheurs de premier cycle se sentent soutenus lorsqu'ils souhaitent partager leurs travaux dans le milieu académique. Avec une base éditoriale plus solide, le JRSUO est bien placé pour continuer à évoluer comme un espace collaboratif consacré à une recherche réfléchie et menée par des étudiants.

Nous espérons que ce volume mettra en valeur la diversité des recherches de premier cycle réalisées à l'Université d'Ottawa, tout en encourageant davantage d'étudiants à considérer leurs questions comme dignes d'être étudiées et publiées. Toute recherche commence par la curiosité, et des revues comme le JRSUO existent pour offrir un espace à cette curiosité. Les articles sélectionnés pour la couverture ont été choisis dans le cadre d'un processus d'évaluation éditoriale à l'aveugle. Les membres du comité éditorial se sont retirés des décisions concernant leurs propres soumissions.

Au nom de toute l'équipe éditoriale, nous sommes fiers de vous présenter le volume V, numéro i. Nous vous remercions de votre lecture et de votre soutien à la recherche de premier cycle.

Ishaan Goswami et Sivany Kathir
Co-rédacteurs en chef
Journal d'étudiant de recherche scientifique de l'Université d'Ottawa

Benzylium Versus Tropylium-like Ion Formation: The Dissociation of Ionized 1-Methylnaphthalene

Formation des ions de type benzylium versus ions de type tropylium : la dissociation du 1-méthylnaphtalène ionisé

Joseph Costello[1], Paul M Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

The formation of benzylium versus tropylium ions in the dissociation of gas-phase ions has a long and rich history. In this brief report, the possible formation of these two structurally isomeric ions in the dissociation of ionized 1-methylnaphthalene was explored. Hydrogen (H)-loss from 1-methylnaphthalene gave a distinct ion species compared to those generated by bromine (Br) atom-loss from 2-(bromomethyl)naphthalene, 1-methyl-4-bromonaphthalene, and 1-bromo-2-methylnaphthalene. The primary dissociation pathway for the tropylium ion structure [molecule (M)-H]$^+$ ion from cycloheptatriene was H-atom loss. This was also the primary reaction for [M-H]$^+$ from 1-methylnaphthalene, a result consistent with the ion-molecule reaction chemistry of Gotkis and Lifshitz, which suggested this population of ions was largely benzotropylium. The three brominated species must thus lose Br to make the naphthylmethyl cation structure as the common reacting configuration.

La formation des ions benzylium par rapport aux ions tropylium lors de la dissociation d'ions en phase gazeuse possède une longue et riche histoire. Dans ce court rapport, la formation possible de ces deux ions isomères de structure lors de la dissociation du 1-méthylnaphtalène ionisé a été étudiée. La perte d'un atome d'hydrogène (H) à partir du 1-méthylnaphtalène a produit une espèce ionique distincte de celles générées par la perte d'un atome de brome (Br) à partir du 2-(bromométhyl)naphtalène, du 1-méthyl-4-bromonaphtalène et du 1-bromo-2-méthylnaphtalène. La principale voie de dissociation menant à la structure de l'ion tropylium, c'est-à-dire l'ion [molécule (M) − H]$^+$ issu du cycloheptatriène, correspond à la perte d'un atome d'hydrogène. Cette réaction est également la voie principale pour l'ion [M − H]$^+$ issu du 1-méthylnaphtalène, un résultat cohérent avec la chimie des réactions ion–molécule décrite par Gotkis et Lifshitz, ce qui suggère que cette population d'ions est majoritairement de type benzotropylium. Ainsi, les trois composés bromés doivent perdre un atome de brome pour former la structure de cation naphtylméthyle, qui constitue la configuration réactive commune.

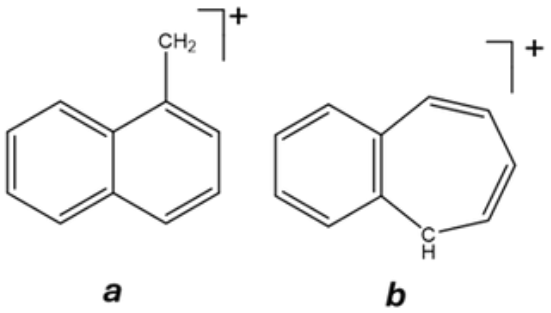
Keywords: benzylium ion; tropylium ion; benzotropylium; 1-methylnaphthalene; MIKE spectrometry; gas-phase ion chemistry; ion dissociation; naphthylmethyl cation; density functional theory; mass spectrometry

Introduction

One of the more famous stories in gas phase ion chemistry is the formation of the benzylium versus tropylium ion structure ($C_7H_7^+$, mass-to-charge ratio (m/z) 91, where C is carbon). Hydrogen-loss from toluene forms a mixture of the two structures, with tropylium being the dominant one (1-4). Tropylium is formed uniquely via the H-loss reaction of cycloheptatriene radical cations. This story bled over into astrochemical interest as researchers probe the dissociative photoionization of methyl-substituted polycyclic aromatic hydrocarbon molecules, which are expected to be carriers of some of the unidentified infrared bands (5). For example, the H-loss reaction in photoionized 1-methylpyrene demonstrated the formation of only the benzylium-like structure due to the ring strain in the key transition state leading to the tropylium-like product (6). Jusko et al. concluded that the tropylium-analogue, if present, is below the detection limit of their infrared multi-photon dissociation experiment. They computationally modeled both isomers and found C_2H_2-loss was the dominant reaction from both, making the distinction of the two structures even more difficult (7). Recently, Rossi et al. demonstrated that the two ions can be distinguished based on vacuum ultraviolet photodissociation action spectra (8). Benzylium and tropylium ions also factor in the dissociation of ions with multiple rings such as substituted coumarins (9).

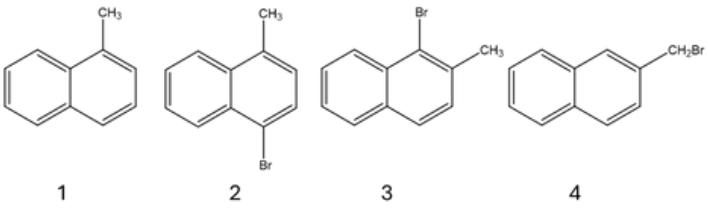
Taking a step back from this more complex system, this study focuses on the ion formed by H-loss from 1-methylnaphthalene. Gotkis and Lifshitz explored the nature of the m/z 141 ion ($C_{11}H_9^+$) formed upon photoionization of 1-methylnaphthalene, and using time-resolved measurements, they monitored the dissociative photoionization of the above-mentioned molecules on timescales from microseconds to seconds. Importantly, they observed a side reaction of a percentage of the m/z 141 ions ($C_{11}H_9^+$) with their

precursor molecule to form $C_{12}H_{11}^+$ (m/z 155) and $C_{10}H_8$, which is analogous to the reaction observed between the benzylium ion and toluene (10, 11). This reaction showed that H-loss from the 1-methylnaphthalene ions resulted in both 1-naphthylmethyl (a) and benzotropylium (b) ion structures (Scheme 1). The ion-molecule reactions suggested the content of 1-naphthylmethyl was about 20% across a photon energy range of 10–13 eV, corresponding to a molecular ion internal energy of up to 5 eV (10).



Scheme 1. Chemical structures of 1-naphthylmethyl and benzotropylium ions. The 1-naphthylmethyl ion (a) and the benzotropylium ion (b).

To probe this further, this study examines the $C_{11}H_9^+$ (m/z 141) ions formed by electron ionization of 1-methylnaphthalene and three brominated isomers (Scheme 2) with mass-analyzed ion kinetic energy (MIKE) spectrometry and theory to probe the dissociation over a narrow internal energy window for these ions.



Scheme 2. Precursor compounds of 1-methylnapthalene and brominated isomers. 1-methylnaphthalene (1), 1-methyl-4-bromonaphthalene (2), 1-bromo-2-methylnaphthalene (3), and 2-bromomethylnaphthalene (4).

Methods

Compounds 1-4 (Scheme 2) were purchased from Millipore Sigma and used without further purification. MIKE spectrometry was performed on a VG ZAB-R mass spectrometer (Manchester) (12, 13). Samples were introduced into the ion source of the instrument by thermal desorption from a solids inlet insertion probe. The ion source pressure (read above the ion source turbopump using an ion gauge) was kept below 1.0 x 10^{-6} Torr. The radical cations were generated by electron ionization with 80 eV electrons. The resulting ions were accelerated to 8 kV toward the magnetic sector where m/z 141 was momentum-selected and transmitted to the second field-free region. Ions with internal energies such that their unimolecular dissociation rate constant is between 10^4 and 10^6 s^{-1} can dissociate in this region, and they are coined “metastable ions.” Due to the conservation of energy and momentum, product ions from unimolecular dissociation (F^+) have a fraction of the precursor ion’s (P^+) translational energy, T_{P+}, according to the equation:

$$T_{F^+} = T_{P^+}\left(\frac{M_{F^+}}{M_{P^+}}\right) \quad \text{(Eq. 1)}$$

Where M_{F+} and M_{P+} are the m/z ratios of the product and precursor ions, respectively. The electrostatic analyzer separates the ions according to their kinetic energies, and the ions are then detected by an off-axis conversion dynode/scintillator/photomultiplier assembly. By measuring T_{F+} and T_{P+}, and knowing M_{P+}, M_{F+} can be deduced for a singly-charged ion.

All structures were optimized using density functional theory with the B3LYP/6-311+G(d,p) level of theory using the Gaussian 16 software suite (14). Transition states were confirmed with the intrinsic reaction coordinate protocol.

Results and Disucssion

Figure 1 shows the MIKE spectra for the source-generated m/z 141 ion from Scheme 2 compounds 1-4. For 2-4, m/z 141 resulted from the loss of Br in the ion source upon dissociative ionization. Three fragment ions were observed. The main peaks are due to either H-loss (m/z 140) or C_2H_2-loss (m/z 115), with some evidence for 2H-loss in some of the spectra. The spectra appear to cluster into two types: one in which C_2H_2-loss predominates (for m/z 141 from 2-4), and one where H-loss dominates (for m/z 141 from 1). Interestingly, Br-loss from 4, which one would assume produced the 2-naphthylmethyl cation (c), produces an ion with the same MIKE spectrum as those formed from Br-loss from 2 and 3, suggesting that the three initially formed $[M\text{-}Br]^+$ ions may isomerize to a common reacting configuration before metastable dissociation (it would be highly improbable for multiple reacting configurations to be formed in the same relative abundance from three precursors) (12). If this is the case, then the data suggests a minimum of two predominant reacting configurations, at least one formed from 1 and the other formed from 2-4.

Computational investigation of the interconversion of a and b

At the presently employed level of theory, ion b is the global minimum, lying 0.3 eV below a (Figure 2). Ion a first isomerizes to d by bending the CH_2 moiety over to the adjacent carbon. The CH_2 group then inserts into the ring to form e, which undergoes a 1,2-H shift to form b. The reverse barriers from d to a, and from e to d are tiny with the slightest geometry change leading back to a or e,respectively. The high relative energy for e and the subsequent transition state is due to the breaking of the delocalization in a and b and the incorporation of both a carbene carbon and the CH_2 moiety into the seven-member ring.

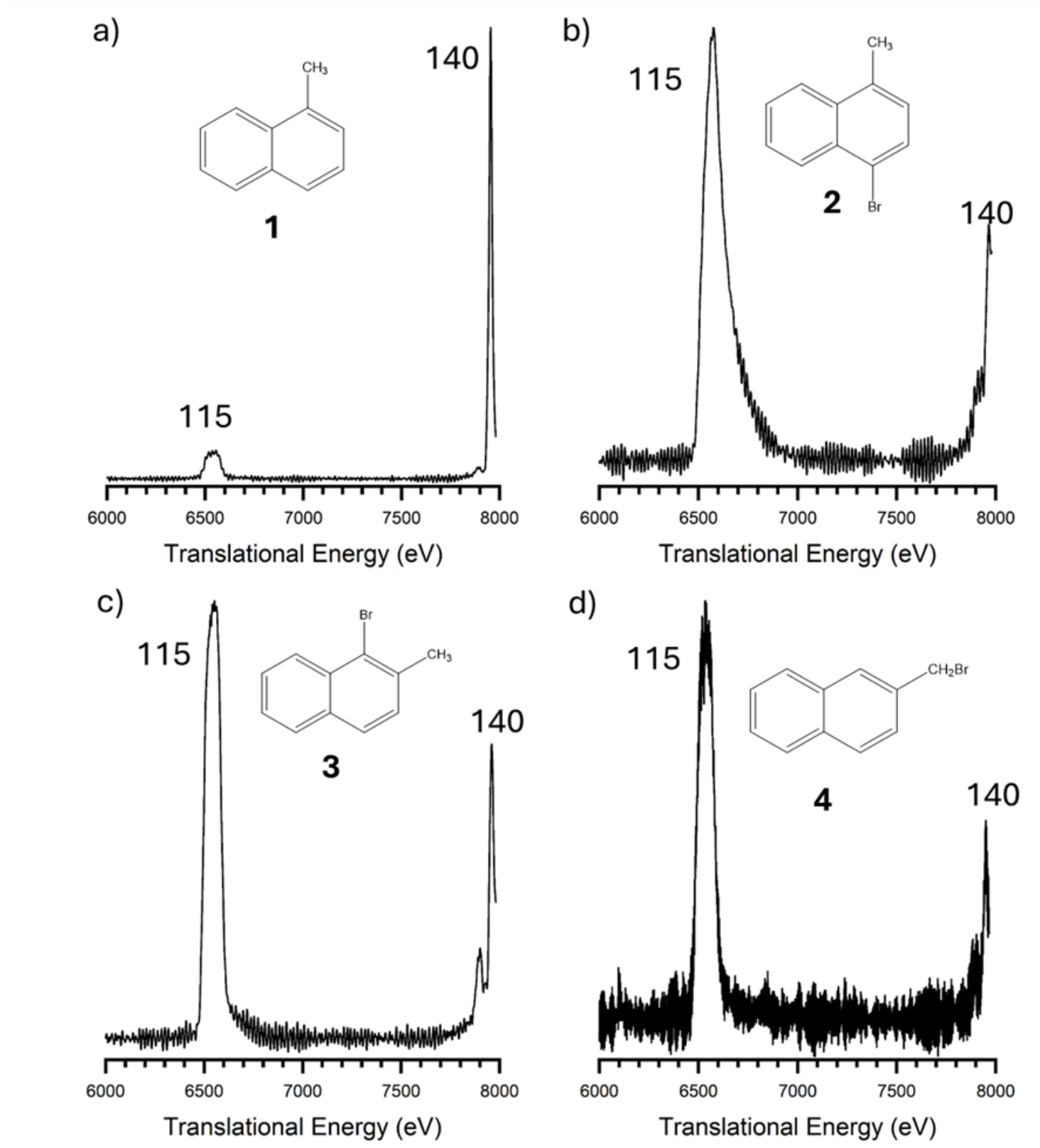


Figure 1. MIKE mass spectra for compounds 1-4. In (b), the skewed appearance of the peak at 6500 eV is due to the lag of the amplifier at high gain. Peaks are labelled with their m/z as calculated from Equation 1.

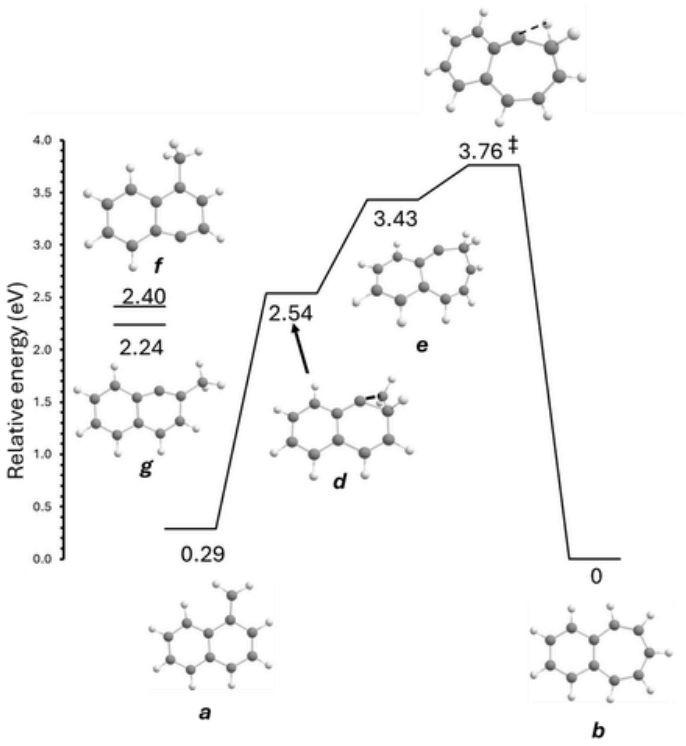


Figure 2. The interconversion of a and b at the B3LYP/6-311+G(d,p) level of theory. Structures f and g are included for comparison.

Br-loss from 2 and 3 will initially form distorted structures with the charge on a bare ring carbon, structures f and g (Figure 2). These lie at high relative energy (2.40 and 2.24 eV, respectively). Hydrogen migration along the ring would result in their isomerization to a (fromf) and c (from g), the 2-naphthylmethyl cation. For the naphthalene radical cation, such hydrogen migration occurs below 3.5 eV (15). Combined with the fact that Br-loss from 4 should initially make c, these combined experimental and theoretical results suggest that the reactive configuration(s) for the m/z 141 ions from 2-4 are a form of naphthylmethyl cation (a and/or c). This would be consistent with the results of Gotkis and Lifshitz that showed m/z 141 from 1 is primarily b (10). The results also imply that H-loss would be the primary fragmentation pathway for b. The MIKE spectrum was obtained for m/z 91 from cycloheptatriene, which should be 100% tropylium ion, and the main peak is indeed H-loss (Figure 3). As the ring size of the polycyclic aromatic hydrocarbon increases, it seems either the predominant loss from b-type structures switches to C_2H_2 (according to Jusko et al.) (7) or the H-loss energy was overestimated in their theoretical model.

Conclusion

This study employed MIKE spectra to explore the structure of the $C_{11}H_9^+$ ions (m/z 141) formed by H-loss from ionized 1-methylnaphthalene. The spectrum of these m/z 141 ions was dominated by the loss of an H atom together with a small amount

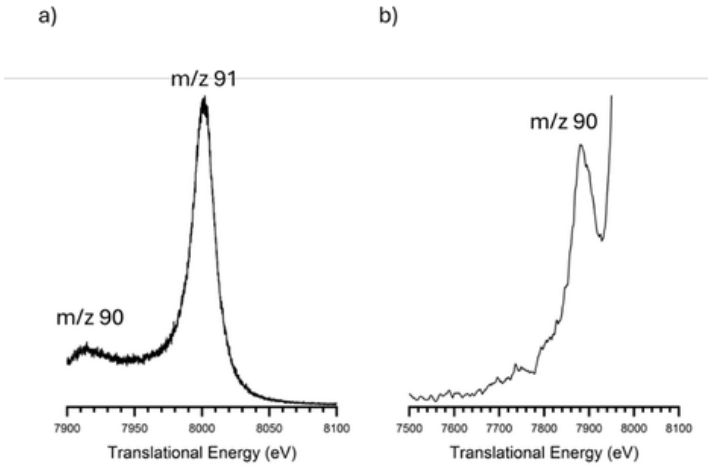


Figure 3. MIKE mass spectra of ion source generated m/z 91 ([M-H]$^+$) from cycloheptatriene. Panel (a) illustrates the relative abundance of the precursor and fragment ion, while (b) shows the entire H-loss peak.

of C_2H_2-loss. By comparison, the source-generated m/z 91 ion from cycloheptatriene, which should have the tropylium structure, also exhibited primarily H-loss in its MIKE spectrum. Thus, in conclusion, ionized 1-methylnaphthalene loses H to form primarily the benzotropylium ion (b), in agreement with the ion-molecule results of Gotkis and Lifshitz (10). Three isomeric brominated ions also generated m/z 141 ions (by Br-loss) that had virtually identical MIKE spectra consistent with a naphthylmethyl cation, such as a, which exhibited a dominant reaction leading to loss of C_2H_2.

Acknowledgements

P.M.M. thanks the Natural Sciences and Engineering Research Council of Canada for continuing financial support and the Digital Research Alliance of Canada for computational resources.

References

1. J.H. Moon, J.C. Choe, M.S. Kim. Kinetic energy release distribution in the dissociation of toluene molecular ion. The tropylium vs benzylium story continues. J. Phys. Chem. A. 104, 458-463 (2000).
2. F.-S. Huang, R.C. Dunbar. Time-resolved photodissociation of toluene ion. Int. J. Mass Spectrom. Ion Process. 109, 151-170 (1991).
3. N. Ohmichi, I. Gotkis, L. Steens, C. Lifshitz. Time-dependent mass spectra and breakdown graphs: 15—toluene-d8. Org. Mass Spectrom. 27, 383-389 (1992). 10.1002/oms.1210270407
4. T. Baer, W.A. Brand, T.L. Bunn, J.J. Butler. Isomerization of internal energy selected ions. Faraday Discuss. Chem . Soc. 75, 45-55 (1983).
5. H.W. Jochims, H. Baumgartel, S. Leach. Structure-dependent photostability of polycyclic aromatic hydrocarbon cations: Laboratory studies and astrophysical implications. Astrophys. J. 512, 500-510 (1999). 10.1086/306752

6. B. West, B. Lowe, P.M. Mayer. Unimolecular dissociation of 1-methylpyrene cations: Why are 1-methylenepyrene cations formed and not a tropylium-containing ion? J. Phys. Chem. A. 122, 4730-4735 (2018).
7. P. Jusko, A. Simon, G. Wenzel, S. Brünken, S. Schlemmer, C. Joblin. Identification of the fragment of the 1-methylpyrene cation by mid-ir spectroscopy. Chem. Phys. Lett. 698, 206-210 (2018)
8. C. Rossi, B. Gans, A. Giuliani, U. Jacovella. Vacuum ultraviolet fingerprints as a new way of disentangling tropylium/benzylium isomers. J. Phys. Chem. Lett. 14, 8444-8447 (2023).
9. H.J. Dias, W.H. Santos, L.C.S. Filho, E.J. Crevelin, J.S. McIndoe, R. Vessecchi, A.E.M. Crotti. Electrospray ionization tandem mass spectrometry of 4-aryl-3,4-dihydrocoumarins. J. Mass Spectrom. 59, e5062 (2024). 10.1002/jms.5062
10. I. Gotkis, C. Lifshitz. Time-dependent mass spectra and breakdown graphs. 16–the methylnaphthalenes. Org. Mass Spectrom. 28, 372-377 (1993). 10.1002/oms.1210280418
11. J. Shen, R.C. Dunbar, G.A. Olah. Gas phase benzyl cations from toluene precursors. J. Am. Chem. Soc. 96, 6227-6229 (1974).
12. J.L. Holmes, C. Aubry, P.M. Mayer. Assigning structures to ions in mass spectrometry (CRC Press, Boca Raton, ed. 1, 2007).
13. R.G. Cooks, J.H. Beynon, R.M. Caprioli, G.R. Lester. Metastable ions (Elsevier Sci. Pub. Co. Amsterdam, 1973).
14. M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A.V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox. Gaussian 16, Revision B.01. Gaussian, Inc., Wallingford, CT (2016).
15. E.A. Solano, P.M. Mayer. A complete map of the ion chemistry of the naphthalene radical cation? Dft and rrkm modeling of a complex potential energy surface. J. Chem. Phys. 143, e104305 (2015). 10.1063/1.4930000

A Comparison of Bioactive Molecules in Three Sage Varieties

Une comparaison de molécules bioactives de trois variétés de sauge

Cindy Yao[1], Sharon Barden[1], Paul M. Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

Sage belongs to the genus *Salvia* in the family Lamiaceae and is a large, globally distributed aromatic plant with a wide variety of species. This work explores and compares the chemical composition of raw plant material from common sage (*Salvia officinalis*), clary sage (*Salvia sclarea*) and white sage (*Salvia apiana*) using microwave-distilled hydrosol extraction, supercritical fluid carbon dioxide (sc-CO_2) extraction, and alcohol extraction. All extracts were analyzed by gas chromatography-mass spectrometry (GC-MS). The results show that the sc-CO_2 and alcohol extraction methods yield more chemical components than hydrosol extraction, such as epimanool, humulene, sclareol and linalyl acetate. The chemical composition of common sage closely resembles that of a commercially obtained dried white sage, while clary sage varies greatly depending on the extraction method. The powdered white sage does not show any components other than camphor and eucalyptol, likely due to its age and how it is dried or powdered.

———

La sauge appartient au genre *Salvia* de la famille des Lamiacées et est une plante aromatique distribuée globalement ayant une vaste variété d'espèces. Ce travail explore et compare la composition chimique de matière première végétale provenant de la sauge commune (*Salvia officinalis*), la sauge sclarée (*Salvia sclarea*) et la sauge blanche (*Salvia apiana*) en utilisant l'extraction d'hydrolat par distillation assistée par micro-ondes, l'extraction au dioxyde de carbone en fluide supercritique (CO2 supercritique) et l'extraction par alcool. Tous les extraits ont été analysés par chromatographie en phase gazeuse couplée à la spectrométrie de masse (CPG-SM). Les résultats ont démontré que les méthodes d'extraction au CO2 supercritique (sc-CO2) et à l'alcool permettent d'obtenir un plus grand nombre de composants chimiques que l'extraction par hydrolat, particulièrement l'épimanool, l'humulène, la sclaréol et l'acétate de linalyle. La composition chimique de la sauge commune ressemblait étroitement celle de la sauge blanche séchée commerciales, alors que celle de la sauge sclarée variait grandement selon la méthode d'extraction. La sauge blanche en poudre ne démontre pas des composantes autres que le camphre et l'eucalyptol, probablement dû à son âge et à sa méthode de séchage ou de poudrage.

Keywords: Salvia officinalis; Salvia sclarea; Salvia apiana; GC-MS; supercritical CO_2 extraction; phytochemical analysis; bioactive compounds; essential oils; epimanool; α-humulene

Introduction

The earliest medicines came from natural products and were mostly derived from plants. For example, salicylic acid, the foundation of the aspirin we rely on today for a pain killer, was first discovered in willow bark, while the cancer-fighting agent taxol originates from the Pacific yew tree. Some drugs are extracted directly from plants, while others are created by modifying chemical compounds found in them. Though a few are synthesized from inorganic materials, many have their origins in research on the active compounds discovered in plants (1).

The genus *Salvia L.* is one of the best-known medicinal and aromatic plants of the Lamiaceae family, comprising 900 species, and found throughout Europe, Asia, and the Americas (2). Sage has a long history of being used as a seasoning and for health purposes. It is promoted for treating conditions such as sore throat, memory loss, diabetes and high cholesterol. *Salvia officinalis*, known as common sage, is a valuable industrial plant widely used in the food and pharmaceutical industries. It has been reported to be a potential therapeutic option for Alzheimer's disease based on its *in vitro* cholinergic binding properties and its role in regulating mood and improving cognitive performance in humans (3). Indigenous peoples across much of California incorporate both the seeds and green parts of white sage into their diets, medicines, and ceremonies (4). Although there is no historical record of its use in religious ritual during the pre-colonial period, the ritual smudging of white sage (*Salvia apiana*) is widely used in the Chumash religion for its sedative, diuretic, and common-cold healing properties (5).

The objective of this study is to explore and analyze the chemical composition of the hydrosol, supercritical-CO2, and alcohol extracts of common sage (*Salvia officinalis*), clary sage (*Salvia sclarea*) and white sage (*Salvia apiana*).

Figure 1. (a) Common sage (*Salvia officinalis*), (b) Clary sage (*Salvia sclarea*), (c) White sage powder (*Salvia apiana*) and (d) hydroponic white sage (*Salvia apiana*).

Figure 2. Microwave-distilled hydrosol apparatus and a bottle of common sage hydrosol.

Methods

Common sage (*Salvia officinalis*) was collected from a garden patch of sage which is 3 years old and thriving in a sunny exposure in a 5-acre gardening zone, south of Ottawa, Canada (Fig. 1a). Clary sage (*Salvia sclarea*) was harvested from the same area from plants that were two years old and in full flower (Fig. 1b). White sage (*Salvia apiana*) was obtained in dried powder from Mountain Herb (USA) (Fig. 1c) and from a hydroponics system in Calgary, Canada (NuLeaf Inc.) (Fig. 1d).

Microwave-Distilled Hydrosol Extraction

The roots and stems of the fresh common and clary sage were removed. The leaves were washed, torn into small pieces, weighed to 100 g and 615 g respectively (due to availability), and placed at the bottom of the jar in a circle around the beaker along with some water to have prevented over-drying during microwaving. The common sage was then microwaved four times for a total of 32 minutes (Fig. 2). The clary sage was microwaved seven times for a total of 50 minutes. Previous experience in the lab has demonstrated that further cycling does not result in increased analyte concentrations. The powdered white sage was weighed to 100 g and soaked in 500 mL of boiled water overnight. The slurry was then microwaved five times for a total of 39 minutes.

Solid Phase Extraction (SPE)

Chemical constituents from water extracts were removed via SPE to prepare them for instrumental analysis. The C-18 SPE cartridge was wetted with 1 mL of methanol and 1 mL of deionized water to activate the solid phase before loading the hydrosol. 10 mL and 3 mL of common sage and white sage hydrosol, respectively, were loaded and passed through the cartridge. 100 mL of clary sage hydrosol was used due to the low concentration of the chemical species present. After the water and other unwanted materials were removed, four aliquots of 250 µL 50:50 MeOH:CH_3CN were used to elute the analytes retained on the solid phase, resulting in 1 mL of eluted product to be analyzed by GC-MS.

Supercritical Fluid Carbon Dioxide (sc-CO_2) Extraction

To extract both polar and non-polar compounds and thus produce a more comprehensive overview of the plant material, Sc-CO_2 was used. Washed common sage and clary sage leaves were freeze-dried to minimize moisture content and ground into powder. The ground common sage, clary sage, and both white sages (Powder Mountain Herb & Nuleaf) were placed in the extraction chamber of the Supercritical Fluid Technologies Inc. SFT-250 SFE/SFR System. The flow rate of CO_2, pressure (300 bar), and extraction temperature (45°C) were controlled to ensure successful extraction and maximize the yield. The dissolved compounds were carried by sc-CO_2 into a separator vessel where the pressure was lowered, causing the sc-CO_2 to return to its gaseous state and the compounds were collected at the end. 5-minute intervals of soaking the plant material in CO_2 and then extracting were repeated between 4 and 6 times. The process was stopped when there was no extract entering the collection vial. The collected common sage, clary sage, Nuleaf and Mountain Herb white sage extracts were each diluted in 1 mL of ethyl acetate. 35 µL of diluted freeze-dried common sage, 0.036 µL of diluted clarysage, 10 µL of diluted white sage (Nuleaf Dried), 5 µL of diluted white sage powder (Mountain Herb) extracts were diluted to 1 mL with ethyl acetate for GC-MS analysis.

Alcohol Extraction

9 mL of methanol was added to each of 0.3 g of common sage, Mountain Herb white sage powder and NuLeaf white sage, and each was vortexed for 10 minutes. The supernatants from all the extracts were filtered several times using syringe filters before being analyzed by GC-MS.

Gas Chromatography-Mass Spectrometry

An Agilent 7820A Gas Chromatography coupled to a 5975C Mass

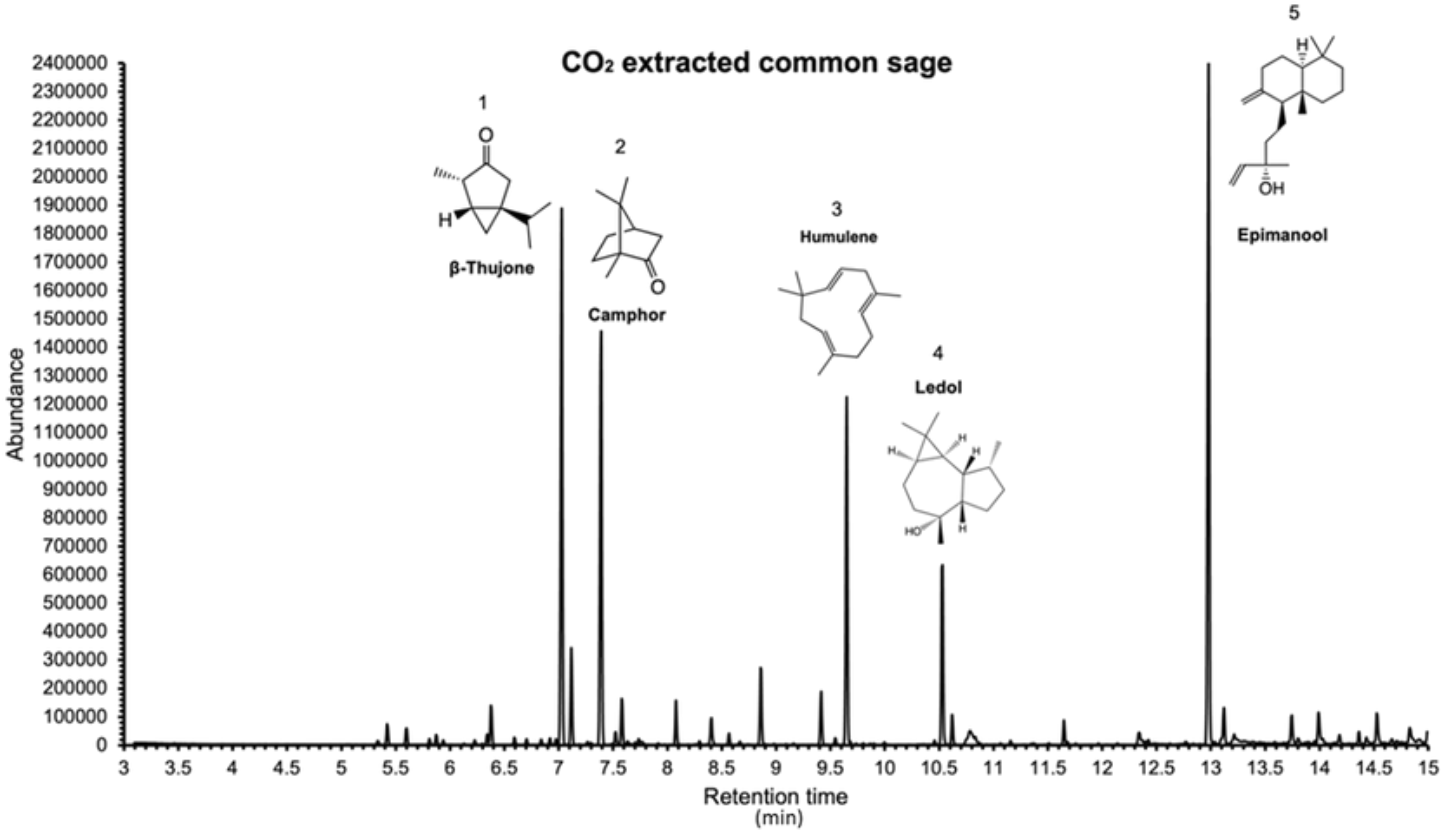


Figure 3. Major chemical components of the CO_2 extracted common sage and their molecular structures.

Spectrometer was used to identify the major compounds in all the extracts of the sages. The inlet was 200°C and the oven ramped from 40-250°C at a rate of 15°C per minute for hydrosol extraction and essential oil, with a 2-minute hold time at the beginning. The oven ramped from 40-300°C at a rate of 15°C per minute for sc-CO_2 and alcohol extraction, with a 1-minute hold time at the beginning and end of the run. Helium gas ran at a constant flow of 1.3 ml min^{-1}. The column was a Rx 5 sil MS 30m x 0.25mm x 0.025µm. The mass spectrometer was calibrated weekly using PFTBA, (Sigma Millipore 77299).

Retention times were calibrated using a C7-C40 saturated alkane standard (Supelco 49452-U) to create the Kovats Retention Indices (RI) using the equation:

$$RI = 100 * \left[n + (N - n)\frac{(tr(unknown) - tr(n))}{tr(N) - tr(n)}\right] \quad \text{(Eq. 1)}$$

where n is the number of the alkane below the unknown, N is the number of the alkane above the unknown, tr (unknown) is the retention time of the analyte, tr(N) is the retention time for the alkane above the unknown, and tr(n) is the retention of the alkane below the unknown.Compounds were identified by the NIST 2 database. Retention Indices were calculated and compared to the NIST Chemistry WebBook for the MS 5 Column. Percentage abundances were used to compare the amounts of compounds present in different extracts.

Results and Discussion

sc-CO_2 Common Sage Extracts

An example of the results obtained is shown in Fig. 3 for common sage extracted by sc-CO_2. The main compounds present are determined to be β-thujone (RT = 7.029 min, 1); camphor (RT = 7.391 min, 2); α-humulene (RT = 9.651 min, 3); ledol (RT = 10.531 min, 4); and 13-epimanool (RT = 12.981 min, 5).

Comparison of Common Sage Using Different Extraction Methods

The results show that the main bioactive molecules in all common sage extracts are α/β-thujone, camphor, ledol, humulene and epimanool (Fig. 4). However, the hydrosol lacks humulene but contains a small amount of carvacrol (2.2%). The chemical components of both CO_2 and alcohol extracts appear similar, while the hydrosol shows fewer peaks due to the steam-distilled method, which extracts only more volatile and water-soluble compounds; some heat-sensitive compounds might also be lost during the processing.

Studies typically showed a variety of cyclic monoterpenes such as eucalyptol, α- and β-pinene, isothujone, camphene, sabinene,

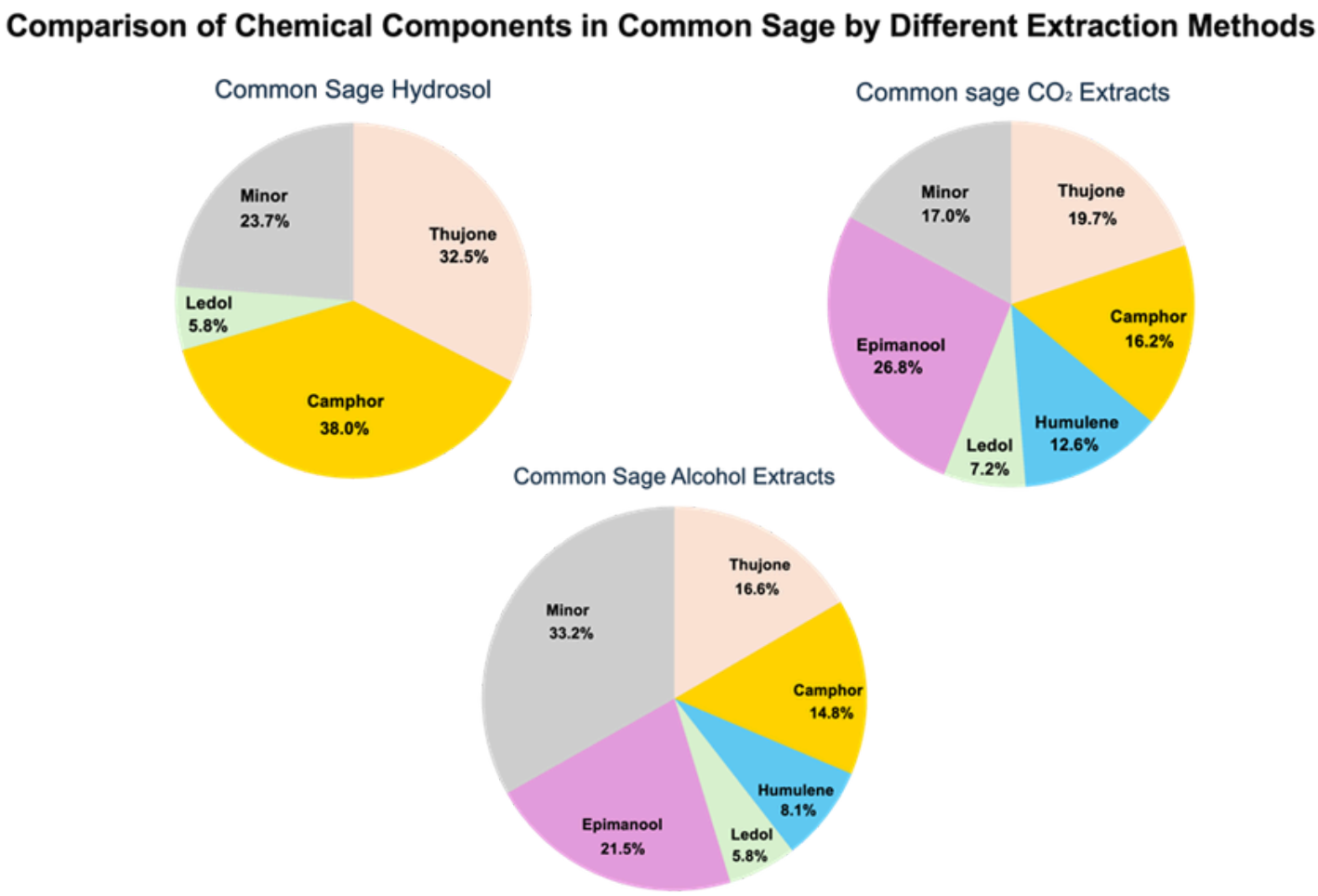


Figure 4. Major chemical components of common sage obtained using different extraction

limonene, myrcene, camphor and (+)-bornyl diphosphate. The wide variety of acyclic, monocyclic, and bicyclic monoterpenes are attributed to the activity of various synthases, which converted the common substrate geranyl diphosphate into multiple products (6). Among these chemical constituents, the most dominant was eucalyptol (62.0%), followed by camphor (8.0%), β-pinene (6.0%), borneol (5.0%), α-pinene (3.7%), β-myrcene (3.0%) and (–)-camphene (2.6%) (7). The chemical composition of common sage leaf essential oil from three different global regions suggested that eucalyptol (26.9%) is again the major component, followed by α-thujone (17.2%), camphor (12.8%), camphene (5.2%), α-pinene (5.0%), β-caryophyllene (4.9%) and β-pinene (4.1%). Notably, α-humulene (5.7% in a sample from Mexico) was detected in the essential oil, as does manool, a stereoisomer of epimanool (8). Eucalyptol was only observed as a minor component (~ 1-2 %) in the current study. Although ledol is not typically reported in common sage, it was detected in all three extracts analyzed in this experiment. It is a sesquiterpenoid alcohol commonly found in plants like wild rosemary, sage and wormwood.

Comparison of Clary Sage Components Using Different Extraction Methods

The chemical composition of clary sage extracts varied greatly depending on the extraction method (Fig. 5). The hydrosol contains β-thujone, camphor, carvacrol, germacrene D and caryophyllene oxide. In contrast, the sc-CO_2 extract only showed two main peaks, linalyl acetate (13.3%) and sclareol (86.1%). This is due to the use of high pressure and temperature during the CO_2 extraction process, which separated the less volatile compounds. Other main components present in the hydrosol were found in minor proportions in the CO_2 extracts.

Among various parts and regional sources of clary sage, the essential oil extracted from leaves of Slovak Republic and flowers of clary sage was found to contain linalool (18.9%), sclareol (15.7%), linalyl acetate (13.7%), α-terpineol (6.5%), germacrene D (5.0%), and geranyl acetate (4.3%) as the most abundant components (9). According to another study, the most significant difference between steam distillation and sc-CO_2 extraction was the amount of sclareol in the extract. Steam distillation yielded nearly no sclareol, whereas sc-CO_2 extraction concentrated it up to 50%, consistent with our current results. It is found that changing the extraction conditions from 90 bar/50 °C to 100 bar/40 °C increased the sclareol concentration from 0 to 25.3% while linalyl acetate decreased from 13.4% to 2.5%, selectively reducing other components while increasing the amounts of sclareol (10). In this

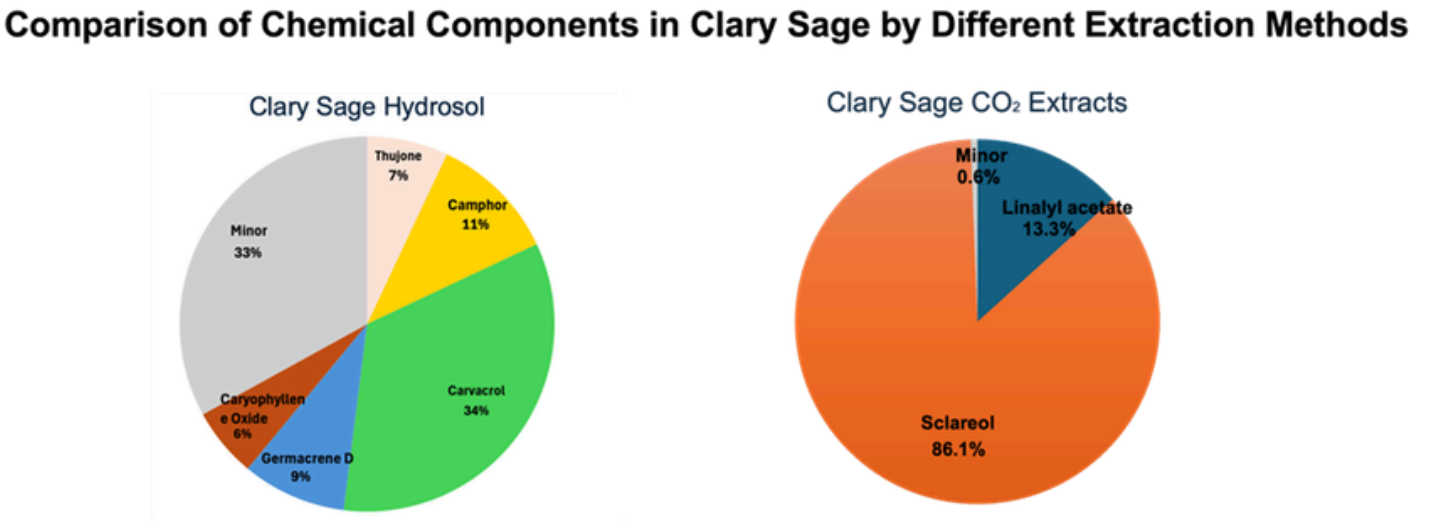


Figure 5. Major chemical components of clary sage obtained using different extraction methods.

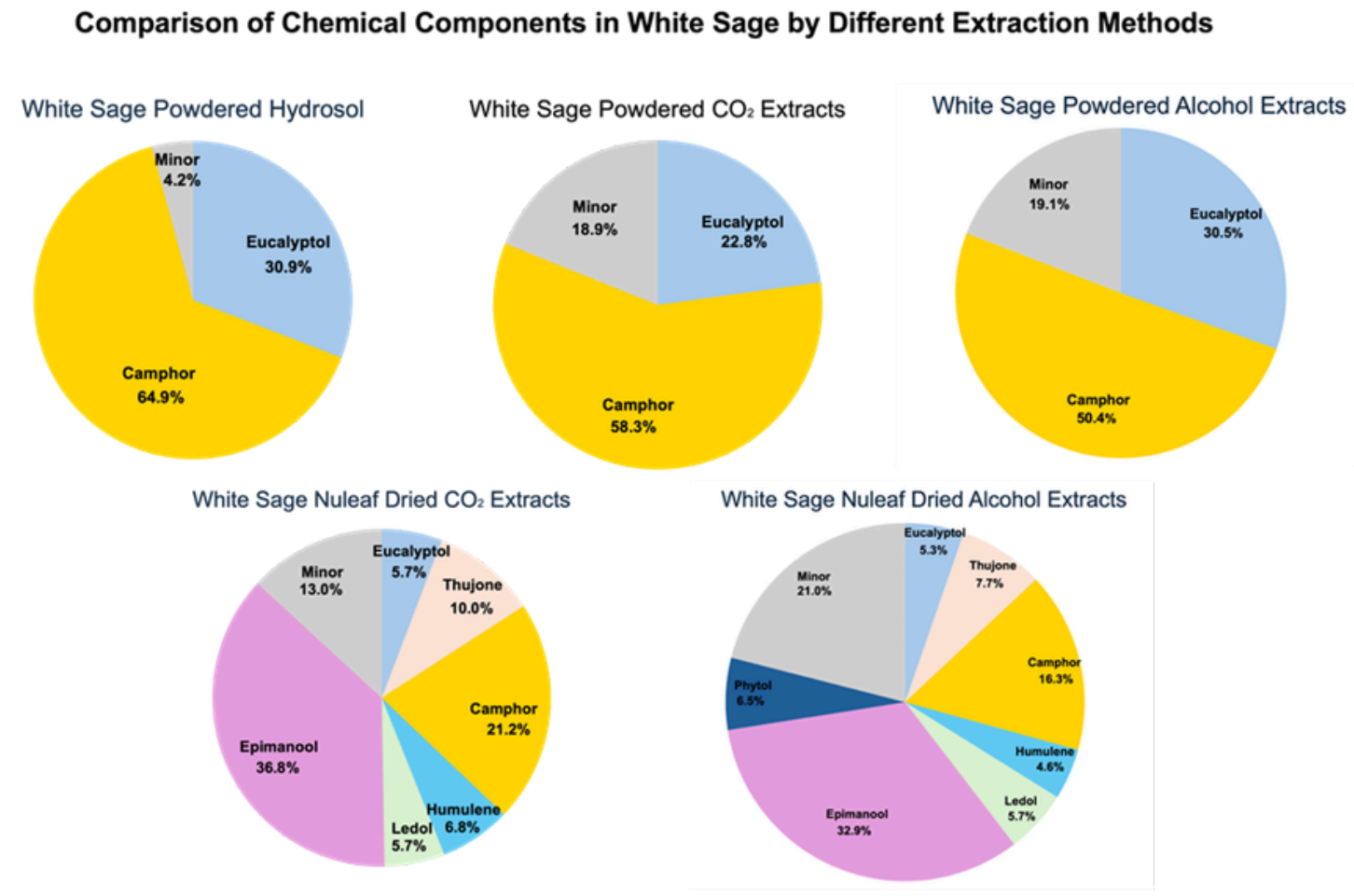


Figure 6. Major chemical components of white sage obtained using different extraction methods.

study, by conducting the sc-CO_2 extraction at 300 bar/45 °C, a yield of 86.1% sclareol was obtained.

The hydrosol contains a variety of compounds, including carvacrol, which is not typically seen in clary sage extracts. It is a monoterpenoid alcohol commonly found in plants such as thyme and oregano. Studies have shown that the use of this compound can effectively inhibit the growth of pathogenic bacteria, including Staphylococcus aureus, Escherichia coli, Salmonella enterica subsp. enterica serovar Typhimurium, Listeria monocytogenes, and Shigella sonnei (11). Germacrene D and caryophyllene oxide were otherwise spotted in small amounts in clary sage. Germacrene D is a sesquiterpene commonly found in plants such as cannabis, rosemary, ginger, lavender and black pepper. It has been found to isomerize under acidic conditions to produce a variety of naturally occurring sesquiterpenes, including cadinane, muurolane and amorphane (12).

The white sage powder extracts contained a significant amount of eucalyptol and camphor (Fig. 6). The powdered white sage was a year old when extracted, and it is unknown how it was dried or powdered. This could be responsible for the lack of chemical diversity. In contrast, the dried white sage derived from the Nuleaf hydroponic system displayed significantly more chemical variety, including eucalyptol, α/β-thujone, camphor, humulene, ledol and epimanool, in the alcohol and CO_2 extracts. Nuleaf white sage is chemically very similar to common sage, with the main difference being the presence of eucalyptol, which is unique to white sage.

Previous research on the aerial parts of white sage have demonstrated that the major chemical components are eucalyptol (34.5%), camphor (21.7%), β-pinene (7.4%), α-pinene (6.4%), camphene (3.9%), limonene (3.5%) and myrcene (3.2%) (13). In another study on white sage, α-humulene was detected only in trace amounts (0.1%) (14). In contrast, the current analysis found α-humulene present in both CO_2 and alcohol extracts of Nuleaf Dried white sage. α-humulene is a sesquiterpene commonly found in plants like hops, cannabis, and some spices like sage, ginseng, and basil. Epimanool, while not observed in white sage essential oil (15), was identified as a common compound in the other extracts. Based on available knowledge, this is the first time that epimanool has been identified in sage extracts and opens the door to further in-depth research, from isolation and purification to exploring its exact mechanism of action (16). Epimanool also has significant antifungal and anticancer properties, as demonstrated in studies involving human leukemia cells and human neuroblast cells (17).

Conclusions

In conclusion, this study focused on hydrosol, alcohol, and sc-CO_2 extracts of common, white and clary sages. One notable difference was the presence of eucalyptol, which was found only in both types of white sage as one of the major components. Contrarily, other studies have shown that eucalyptol was found in large quantities in common sage, whereas in our results it only accounts for less than 3% in minor parts. Despite its rare occurrence in the literature, this study identified ledol in both common sage and white sage. α-Humulene is commonly detected in common sage, though in minor amounts, but isnot consistently present in white sage. Epimanool has been detected in specific conditions in common sage but is rarely reported in white sage. In this study, both epimanool and α-humulene are found in the CO_2 extracts of both common and Nuleaf Dried white sage. Epimanool, being a less well-known diterpenoid alcohol, demonstrates significant biological activity.

A limitation of the present study was the diversity of the samples collected. From fresh to store-bought dried samples, this range of plant sources could almost certainly have affected the results. Future research could focus on further biological testing to validate the medicinal properties and applications of these identified compounds. Comparing and identifying the chemical components of different sage species provides valuable insights into their bioactive diversity, pharmacological and commercial potential.

Acknowledgements

PMM and SB thank the JLH Mass Spectrometry Core Facility at the University of Ottawa for instrument time and extraction supplies used in the course of this study.

References

1. Taneja, N.; Alam, A.; Patnaik, R. S.; Taneja, T. Unmasking the Potential Role of Plant-Based Medicine “Plumbagin” in Oral Cancer—A Novel Paradigm. Oral Sci Int 2022, 19, 3–18. https://doi.org/10.1002/OSI2.1107.
2. Jash K. Shyamal; Gorai Dilip; Roy Rajiv. Salvia Genus and Triterpenoids. Int. J. Pharm. Sci. Res., 2016, 7, 4710-4732. https://doi.org/10.13040/IJPSR.0975-8232.7(12).4710-32.
3. Ertas, A.; Yigitkan, S.; Orhan, I. E. A Focused Review on Cognitive Improvement by the Genus Salvia L. (Sage)—From Ethnopharmacology to Clinical Evidence. Pharmaceuticals 2023, 16, 171. https://doi.org/10.3390/PH16020171.
4. Timbrook, J. Chia and the Chumash: A Reconsideration of Sage Seeds in Southern California. J Calif Gt Basin Anthropol 1986, 8 (1), 50–64.
5. Krol, A.; Kokotkiewicz, A.; Luczkiewicz, M. White Sage (Salvia Apiana)-a Ritual and Medicinal Plant of the Chaparral: Plant Characteristics in Comparison with Other Salvia Species. Planta Med 2021, 88, 604–627. https://doi.org/10.1055/A-1453-0964/BIB.
6. Wise, M. L.; Savage, T. J.; Katahira, E.; Croteau, R. Monoterpene Synthases from Common Sage (Salvia Officinalis). CDna Isolation, Characterization, and Functional Expression of (+)-Sabinene Synthase, 1,8-Cineole Synthase, and (+)-Bornyl Diphosphate Synthase. J. Bio. Chem. 1998, 273, 14891–14899. https://doi.org/10.1074/JBC.273.24.14891/ASSET/3230215A-E5F1-4F42-9A10-5EECB9A621E3/MAIN.ASSETS/GR4.JPG.

7. Hamidpour, M.; Hamidpour, R.; Hamidpour, S.; Shahlari, M. Chemistry, Pharmacology, and Medicinal Property of Sage (Salvia) to Prevent and Cure Illnesses Such as Obesity, Diabetes, Depression, Dementia, Lupus, Autism, Heart Disease, and Cancer. J Tradit Complement Med 2014, 4, 82–88. https://doi.org/10.4103/2225-4110.130373.
8. Craft, J. D.; Satyal, P.; Setzer, W. N. The Chemotaxonomy of Common Sage (Salvia Officinalis) Based on the Volatile Constituents. Medicines 2017, 4, 47. https://doi.org/10.3390/MEDICINES4030047.
9. Hans, M.; Deeksha; Nayik, G. A.; Salaria, A. Clary Sage Essential Oil. Essential Oils: Extraction, Characterization and Applications 2023, 459–478. https://doi.org/10.1016/B978-0-323-91740-7.00001-3.
10. Zanotti, A.; Baldino, L.; Scognamiglio, M.; Reverchon, E. Supercritical Fluid Extraction of Essential Oil and Sclareol from a Clary Sage Concrete. Molecules 2023, 28, 3903. https://doi.org/10.3390/MOLECULES28093903.
11. Kachur, K.; Suntres, Z. The Antibacterial Properties of Phenolic Isomers, Carvacrol and Thymol. Crit Rev Food Sci Nutr 2020, 60, 3042–3053. https://doi.org/10.1080/10408398.2019.1675585.
12. Bülow, N.; König, W. A. The Role of Germacrene D as a Precursor in Sesquiterpene Biosynthesis: Investigations of Acid Catalyzed, Photochemically and Thermally Induced Rearrangements. Phytochemistry 2000, 55, 141–168. https://doi.org/10.1016/S0031-9422(00)00266-1.
13. Takeoka, G. R.; Hobbs, C.; Park, B. S. Volatile Constituents of the Aerial Parts of Salvia Apiana Jepson. J. Essen. Oil Res. 2010, 22, 241–244. https://doi.org/10.1080/10412905.2010.9700314.
14. Krol, A.; Kokotkiewicz, A.; Luczkiewicz, M. White Sage (Salvia Apiana)-a Ritual and Medicinal Plant of the Chaparral: Plant Characteristics in Comparison with Other Salvia Species. Planta Med 2021, 88, 604–627. https://doi.org/10.1055/A-1453-0964/BIB.
15. Bozzini, M. F.; Pieracci, Y.; Ascrizzi, R.; Najar, B.; D'Antraccoli, M.; Ciampi, L.; Peruzzi, L.; Turchi, B.; Pedonese, F.; Alleva, A.; Flamini, G.; Fratini, F. Chemical Composition and Antimicrobial Activity against the Listeria Monocytogenes of Essential Oils from Seven Salvia Species. Foods 2023, 12, 4235. https://doi.org/10.3390/FOODS12234235.
16. Letaief, T.; Garzoli, S.; Laghezza Masci, V.; Mejri, J.; Abderrabba, M.; Tiezzi, A.; Ovidi, E. Chemical Composition and Biological Activities of Tunisian Ziziphus Lotus Extracts: Evaluation of Drying Effect, Solvent Extraction, and Extracted Plant Parts. Plants 2021, 10, 2651. https://doi.org/10.3390/PLANTS10122651.
17. Sandulovici, R. C.; Gălăţanu, M. L.; Cima, L. M.; Panus, E.; Truţă, E.; Mihăilescu, C. M.; Sârbu, I.; Cord, D.; Rîmbu, M. C.; Anghelache, Ş. A.; Panţuroiu, M. Phytochemical Characterization, Antioxidant, and Antimicrobial Activity of the Vegetative Buds from Romanian Spruce, Picea Abies (L.) H. Karst. Molecules 2024, 29, 2128. https://doi.org/10.3390/MOLECULES29092128/S1.

Derivatization of Rosemary is Integral to its Analysis

La dérivatisation du romarin est intégrale à son analyse

Ahona Deb[1], Sharon Barden[1], Paul M Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

Plants of the Lamiaceae family, such as rosemary (*Salvia rosmarinus*), common sage (*Salvia officinalis*), and white sage (*Salvia apiana*), are known for their bold flavours, fragrant leaves, and rich sources of bioactive compounds. A key characteristic shared by these plants is their high content of carnosic acid, a prominent phenolic diterpene, and its derivatives. However, the analysis of these compounds via gas chromatography–mass spectrometry (GC-MS) is challenging and not frequently explored due to their high polarity and low volatility. Derivatization techniques like trimethylsilylation and appropriate solvent selection can address these limitations by enhancing compound volatility, allowing for effective GC-MS identification through characteristic fragmentation patterns. We employed trimethylsilyl derivatization of acetonitrile extracts of rosemary, white and common sage, and quantified carnosic acid, rosmanol, and 12-O-methylcarnosic acid. Rosemary contained the highest amount of carnosic acid at 3.4 mg/mL, followed by common sage at 2.0 mg/mL and white sage at 0.7 mg/mL. Rosemary contained the most rosmanol at 2.2 mg/mL followed by common sage (0.3 mg/mL), and white sage contained minimal amounts (0.003 mg/mL). Additionally, both common sage and rosemary were found to contain 12-O-methylcarnosic acid, with concentrations of 1.3 mg/mL and 0.4 mg/mL, respectively. Thus, we demonstrated that simple solvent extraction and TMS derivatization is an effective approach to quantifying these bioactive compounds in plants.

———

Les plantes de la famille des Lamiacées, tel le romarin (*Salvia rosmarinus*), la sauge commune (*Salvia officinalis*) et la sauge blanche (*Salvia apiana*) sont connues pour leurs saveurs prononcées, leurs feuilles parfumées et leurs sources riches de composés bioactifs. Une caractéristique clé partagée par ces plantes est leur teneur élevée en acide carnosique, un diterpène phénolique important, ainsi que ses dérivées. Cependant, l'analyse de ces composés via la chromatographie en phase gazeuse couplée à la spectrométrie de masse (CPG-SM) pose un défi et n'est pas fréquemment explorée dû à leur polarité élevée et leur faible volatilité. Les techniques de dérivatisation telles la triméthylsilylation (TMS) et la sélection de solvant approprié peuvent remédier à ces limitations en améliorant la volatilité des composés, ainsi permettant à une identification par CPG-SM efficace par des modèles de fragmentation caractéristiques. Nous avons utilisé la dérivatisation triméthylsilylée d'extraits d'acétonitrile de romarin, de sauge blanche et de sauge commune, et avons quantifié l'acide carnosique, le rosmanol et l'acide 12-O-méthylcarnosique. Le romarin contenait le montant le plus élevé d'acide carnosique à 3,4 mg/mL, suivi par la sauge commune à 2,0 mg/mL et la sauge blanche à 0,7 mg/mL. Le romarin contenait le plus de rosmanol à 2,2 mg/mL suivi par la sauge commune (0,3 mg/mL) et la sauge blanche qui contenait des montants minimes (0,003 mg/mL). De plus, la sauge commune et le romarin se sont révélés contenir de l'acide 12-O-méthylcarnosique, avec des concentrations de 1,3 mg/mL et de 0,4 mg/mL respectivement. Ainsi, nous avons démontré que la simple extraction de solvant et la dérivation TMS sont des approches efficaces pour quantifier ces composés bioactifs dans les plantes.

Keywords: rosemary; common sage; white sage; carnosic acid; rosmanol; phenolic diterpenes; trimethylsilylation; GC-MS derivatization; GC-MS; phytochemical analysis

Introduction

The Lamiaceae family is recognized worldwide for its diversity, distinctive flavours, and high versatility. While many of these plants are widely used in culinary applications, they also hold an important place in traditional medicine worldwide. In ancient Greece and Rome, rosemary (*Salvia rosmarinus*) have been traditionally used to alleviate migraines, muscle pain, respiratory conditions, and insomnia (Figure 1a) (1). In North America, common sage (*Salvia officinalis*) and white sage (*Salvia apiana*) have been used to relieve respiratory and digestive issues (Figure 1b,c). A key characteristic shared by these plants is their high content of carnosic acid, a prominent phenolic diterpene, and its derivatives. These compounds demonstrate a wide range of therapeutic properties, including anti-inflammatory, antimicrobial, and antioxidant effects, Nieto et al. explored the inhibitory effects of rosemary, finding the primary bioactive species to be carnosic acid, rosmanol, and related molecules (1).

These compounds interact with the cell membrane, leading to disruptions in electron transport and leakage of cellular components. Similarly, Vegara et al. reported that carnosic acid demonstrates greater efficacy against pathogenic bacteria compared to any other major extract components (2). Nonetheless, its effectiveness and related applications are still under active investigation. Traditionally, high-performance liquid chromatography (HPLC) has been the primary method for analyzing carnosic acid. However, very few studies have explored its analysis using gas chromatography–mass spectrometry (GC-MS). A 2016 study by Islamčević Razboršek, and Ivanović investigated several extraction techniques and adapted GC-MS methods for diterpene analysis (3). Despite this, there remain limitations such as prolonged wait times and extra sample preparation steps. The present study aims to develop a unique method to analyze carnosic acid and related molecules in a rapid and efficient manner and quantify the amounts of each molecule across various natural sources using GC-MS.

Methods

Dried plant material was purchased from Mountain Rose Herbs Inc. (Eugene, Oregon) for rosemary, common sage, and white sage. One gram (1 g) of dried plant material was ground into a fine powder and placed in 10 mL of acetonitrile and left to sit at room temperature for two days. The extract underwent two rounds of filtration: first using a standard vacuum filtration apparatus and then using a microneedle filter to ensure no solid material entered the GC–MS. Notably, no further evaporation or drying steps were required before derivatization with the trimethylsilyl (TMS) reagent, Regisit-1%TMCS (bis(trimethylsilyl)trifluoroacetamide + 1% trimethylchlorosilane). The role of TMS in this experiment was to convert polar hydroxyl groups into non-polar derivatives. This transformation occurs through the substitution of each acidic proton with a trimethylsilyl, $Si(CH_3)_3$, group (Figure 2). Excess TMS reagent handled any residual moisture in the samples.

The reactivity of functional groups during derivatization is influenced by heating: carboxylic acid groups typically react first, followed by alcohol groups. For this experiment, 100 µL of acetonitrile extract was mixed with 100 µL of TMS reagent, then heated for 40 minutes at 70 °C, with shaking every 10 minutes to ensure complete derivatization of all acidic groups. The GC oven temperature was programmed from 80 °C to 300 °C, with a total run time of 18 minutes. The instrument used was an Agilent 7820A Gas Chromatography coupled to a 5975C Mass Spectrometer. The inlet was 200°C and the oven ramped from 40-250°C at a rate of 15°C per with a 2-minute hold time at the beginning. Helium gas ran at a constant flow of 1.3 ml min^{-1}. The column was a Rx 5 sil MS 30m x 0.25mm x 0.025µm.

Each extract was analyzed in triplicate. Quantification was performed using abietic acid as an external standard, with a calibration curve constructed from concentrations of 0.125 mg/mL, 0.25 mg/mL, and 0.5 mg/mL (Figure 3). Three points in the calibration curve necessarily limit the accuracy of this approach.

The identification of specific molecules was done through the testing of pure standards and library matches through NIST databases. Values are quoted for concentration in the derived hydrosol (mg/mL) and per gram of plant material (mg/g).

Figure 1. a) Rosemary (*Salvia rosmarinus*), b) common sage (*Salvia officinalis*), and c) white sage (*Salvia apiana*).

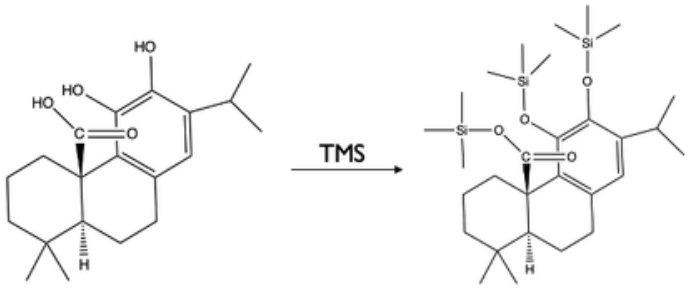


Figure 2. Carnosic acid before and after complete TMS derivatization. Derivatization increases the molecular weight from 332.4 g/mol to 548.4 g/mol and lowers the boiling point, allowing GC–MS detection.

Results and Discussion

Prior to derivatization, the chromatograms for all three plants resemble that in Figure 4, in which there is a large, undefined area of molecules from 11 to 13 minutes that cannot be analyzed directly.

The peaks become much more defined post-derivatization. The chromatograms obtained from the rosemary, common sage, and white sage derivatized acetonitrile extracts are shown in Figure 5.

From these results, it was observed that carnosic acid and rosmanol were present in varying concentrations across all products. Rosemary contained the highest amount of carnosic acid at 3.4 ± 0.15 mg/mL (34 ± 1.5 mg/g), followed by common sage at 2.0 ± 0.15 mg/mL (20 ± 1.5 mg/g), and white sage at 0.7 ± 0.15 mg/mL (7 ± 1.5 mg/g). A similar trend was seen with rosmanol: rosemary contained the most at 2.20 ± 0.15 mg/mL (22 ± 1.5 mg/g), common sage had 0.30 ± 0.15 mg/mL (3 ± 1.5 mg/g), and white sage contained only 0.003 ± 0.15 mg/mL (0.03 ± 1.5 mg/g). Additionally, both common sage and rosemary were found to contain 12-O-methylcarnosic acid, with concentrations of 1.3 ± 0.15 mg/mL (13 ± 1.5 mg/g) and 0.4 ± 0.15 mg/mL (4 ± 1.5 mg/g), respectively. This aligns with previous studies, as common sage tends to have higher concentrations of 12-O-methylcarnosic acid than rosemary (4). It should be noted that several unknown peaks appeared in each chromatogram within the 11 to 13-minute region, which are most likely phenolic diterpenes that could not be identified.

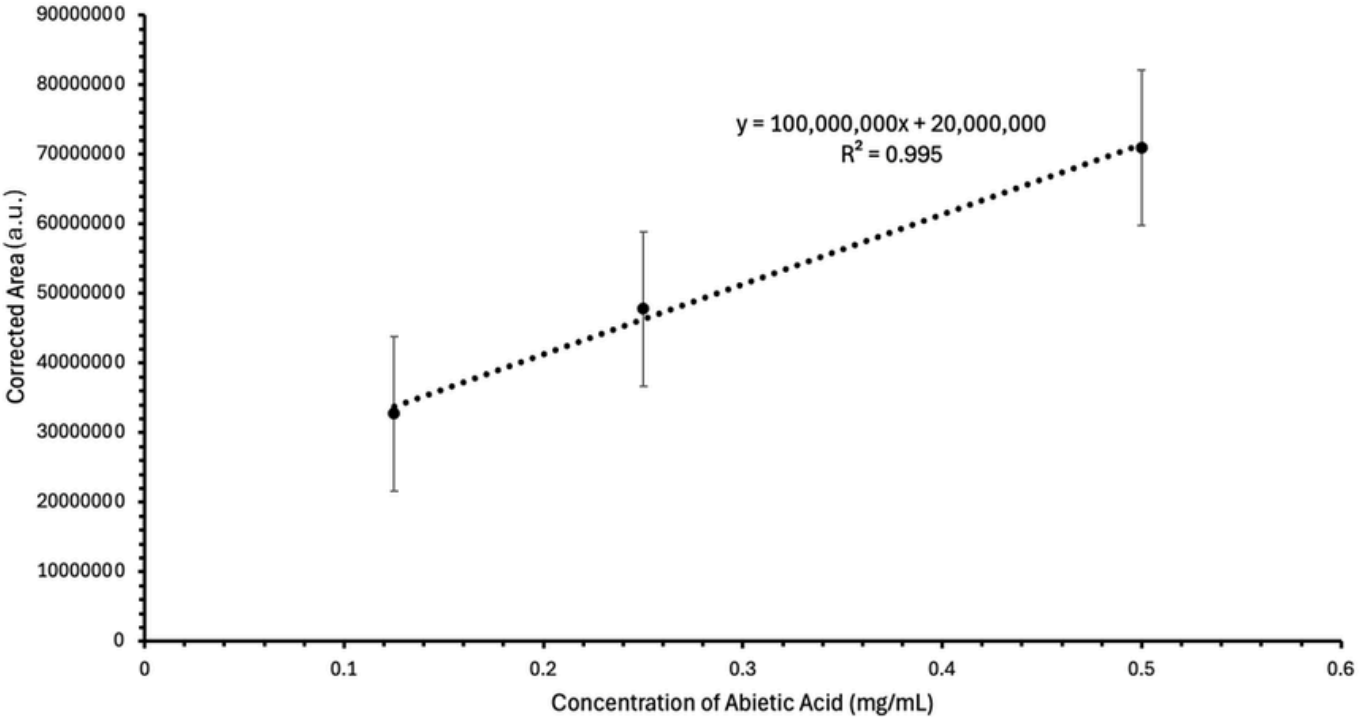


Figure 3. Calibration curve of the external standard, abietic acid, from concentrations of 0.125 mg/mL, 0.25 mg/mL, and 0.5 mg/mL.

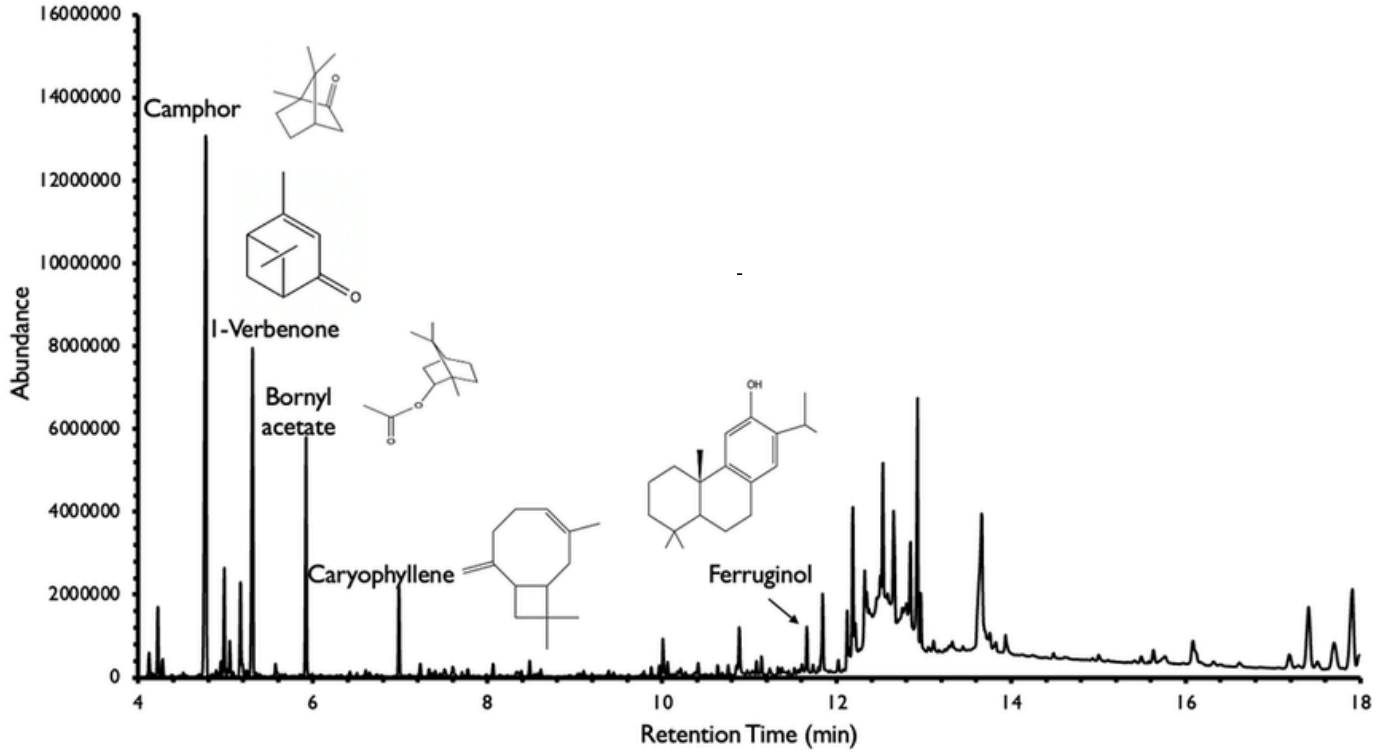


Figure 4. The chromatogram obtained from the GC-MS of an underivatized rosemary acetonitrile extract.

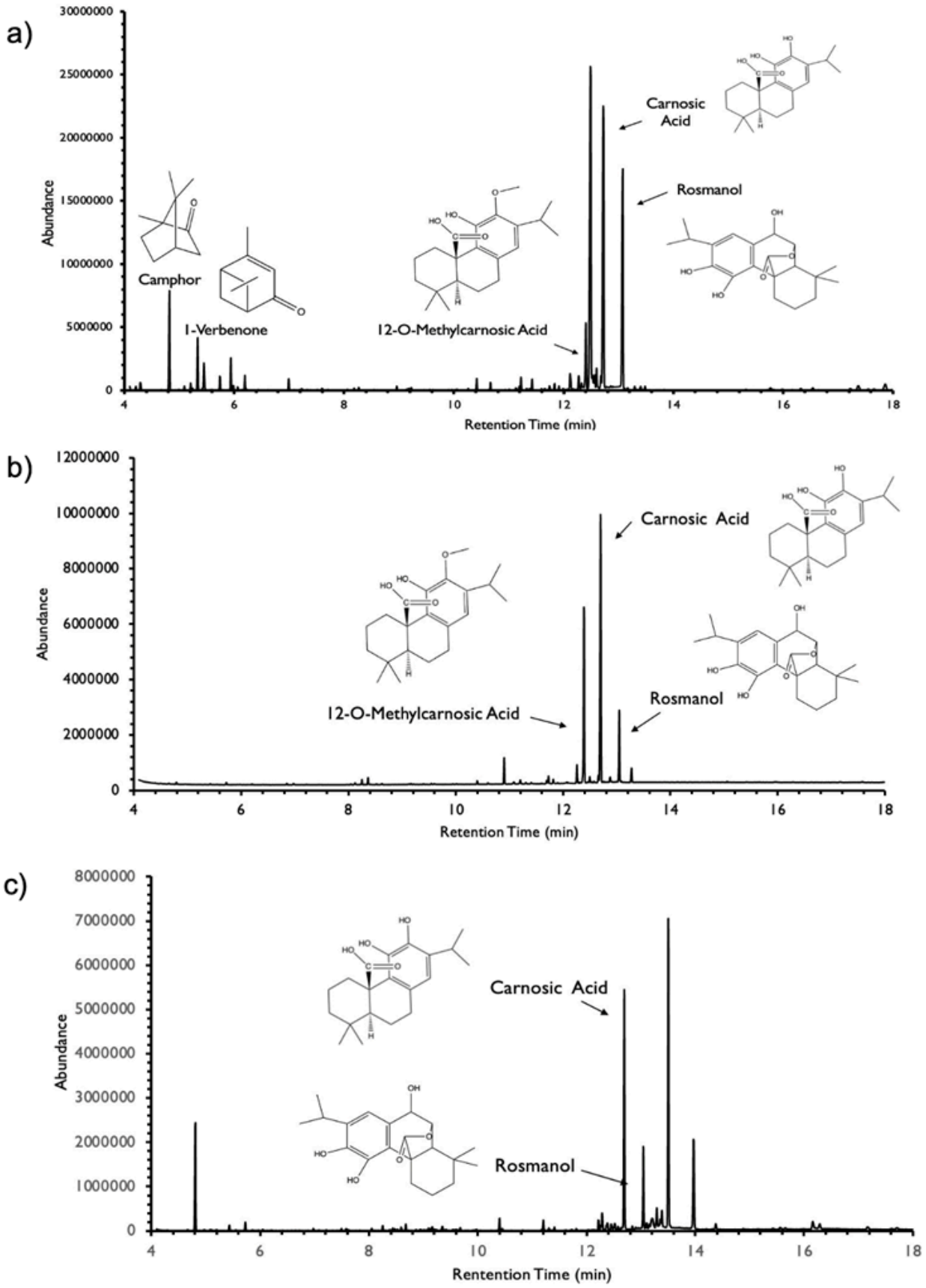


Figure 5. The chromatogram obtained from the GC-MS of derivatized rosemary, b) common sage, and c) white sage acetonitrile extracts.

Conclusion

A reliable, reproducible method has been developed to analyze carnosic acid in a quick and effective manner. The choice of solvent, acetonitrile, is optimal for direct derivatization without the need for an intermediate drying/evaporation step. While maceration occurs over a period of two days, the GC-MS run time is only 18 minutes and is still able to elucidate multiple phenolic diterpenes and other major constituents. Derivatization is an important step needed for the analysis of phenolic diterpenes with GC-MS instrumentation, as it increases their volatility, decreases polarity, and allows prominent, well-defined peaks to emerge. As expected, rosemary contained the highest amount of carnosic acid and rosmanol among the various plant samples, while white sage contained the lowest. Likewise, both common sage and rosemary contained 12-O-methylcarnosic acid, with common sage having the higher concentration.

Acknowledgments

I would like to thank the JLH Mass Spectrometry Core Facility of the University of Ottawa for providing instrument access and consumables related to this project.

References

1. G. Nieto, G. Ros, J. Castillo, Antioxidant and Antimicrobial Properties of Rosemary (Rosmarinus officinalis, L.): A Review. *Medicines*. 5, 98 (2018).
2. S. Vegara, L. Funes, N. Martí, D. Saura, V. Micol, M. Valero, Bactericidal activities against pathogenic bacteria by selected constituents of plant extracts in carrot broth. *Food Chem.* 128, 872–877 (2011).
3. M. I. Razboršek, M. Ivanović, Stability studies and determination of carnosic acid and its oxidative degradation products by gas chromatography–mass spectrometry. *Int. J. Mass Spectrom.* 407, 29–39 (2016).
4. T. Tounekti, S. Munné-Bosch, Enhanced Phenolic Diterpenes Antioxidant Levels Through Non-transgenic Approaches. *Crit. Rev. Plant Sci.* 31, 505–519 (2012).

Effect of Sustainable Aviation Fuels on Contrail Formation Across Flight Routes: A Thermodynamic Analysis

Effet des carburants aéronautiques durables sur la formation de traînées de condensation à travers les routes de vol: une analyse thermodynamique

Zoya Sharma[1*]

1. Liberal Arts & Science Academy, Austin, Texas, USA
*Corresponding author. Email: zoya.sharma@g.austincc.edu

Abstract | Résumé

Aviation-induced contrail cirrus is known to be responsible for significant climate impact, yet fuel and engine-based mitigation strategies receive less emphasis than flight path and altitude optimization research. Flight path optimization studies require an enormous amount of multiparty coordination, real-time data collection, and resource-rich research partners. This study assesses the impact of sustainable aviation fuels (SAFs) on contrail formation, which is a research avenue more amenable to modeling. The usage of three SAF types, Hydroprocessed Esters and Fatty Acids Synthetic Paraffinic Kerosene (HEFA-SPK), Fischer-Tropsch Synthetic Paraffinic Kerosene (FT-SPK), and Alcohol-to-Jet Synthetic Paraffinic Kerosene (ATJ-SPK), was modeled on four representative flight routes. A mathematical framework based on the Schmidt-Appleman criterion (SAC) was developed. This framework was applied to atmospheric data at each waypoint along great circle routes for the four flight paths, enabling spatially resolved contrail formation probabilities to be evaluated for each fuel type. All three SAFs produced a marginal increase in mean contrail formation probability relative to baseline kerosene across all routes. This result is attributed to the higher water vapor emission indices and specific combustion heat of SAFs, which increase the slope parameter G and shift the threshold temperatures toward warmer values. However, this thermodynamic effect operates independently of the primary mitigatory mechanism of SAFs: the reduction of non-volatile particulate matter (nvPM) emissions. The reduced nvPM emissions suppress ice crystal nucleation through a distinct physical pathway not captured within the thermodynamic framework. The results therefore represent a worst-case estimate of contrail formation probability, and the full climate benefit of SAF deployment is expected to be realized when nvPM effects are incorporated alongside the thermodynamic assessment presented here.

———

Le cirrus de traînée induit par l'aviation est reconnu pour être responsable d'un impact climatique important, pourtant les stratégies d'atténuation basées sur le carburant et le moteur reçoivent moins d'attention que la recherche sur l'optimisation de la trajectoire de vol et de l'altitude. Les études d'optimisation de trajectoire de vol nécessitent une énorme coordination multipartite, une collecte de données en temps réel et des partenaires de recherche riches en ressources. Cette étude évalue l'impact des carburants aéronautiques durables (SAF) sur la formation de traînées de condensation, qui constitue une voie de recherche plus propice à la modélisation. L'utilisation de trois types de SAF, esters hydrotraités et acides gras (kérosène paraffinique synthétique HEFA-SPK), kérosène paraffinique synthétique Fischer-Tropsch (FT-SPK) et kérosène paraffinique synthétique alcool-à-jet (ATJ-SPK), a été modélisée sur quatre routes de vol représentatives. Un cadre mathématique basé sur le critère de Schmidt-Appleman (SAC) a été développé. Ce cadre a été appliqué aux données atmosphériques à chaque point de passage le long de chacune des quatre trajectoires de vol orthodromiques, permettant d'évaluer spatialement les probabilités de formation de traînées de condensation pour chaque type de carburant. Les trois SAF ont produit une augmentation marginale de la probabilité moyenne de formation de traînées par rapport au kérosène de base sur toutes les voies. Ce résultat est attribué aux indices d'émission de vapeur d'eau plus élevés et à la chaleur de combustion spécifique des SAF, qui augmentent le paramètre de pente G et déplacent les températures seuils vers des valeurs plus chaudes. Cependant, cet effet thermodynamique fonctionne indépendamment du principal mécanisme d'atténuation des SAF : la réduction des émissions de particules non volatiles (nvPM). Les émissions réduites de nvPM suppriment la nucléation des cristaux de glace par une voie physique distincte non capturée dans le cadre thermodynamique. Les résultats représentent donc une estimation dans le pire des cas de la probabilité de formation de traînées de condensation, et on s'attend à ce que le bénéfice climatique complet du déploiement du SAF soit réalisé lorsque les effets nvPM sont intégrés à l'évaluation thermodynamique présentée ici.

Keywords: Sustainable aviation fuels (SAFs), contrail formation, Schmidt-Appleman criterion, aviation climate impact, contrail cirrus, non-volatile particulate matter (nvPM), radiative forcing, flight route modelling.

Introduction

While carbon dioxide (CO_2) emissions from aircrafts create long-term effects, non-CO_2 components such as water vapor, nitrogen oxides (NOx), and soot also contribute significantly to atmospheric warming through the formation of contrails. Contrails are thin, cirrus-like ice clouds that form when water vapor from aircraft engine exhausts mixes with extremely cold ambient air at high altitudes and rapidly condenses around particulate matter like soot, forming ice crystals. Contrails are responsible for approximately two-thirds of the net radiative forcing currently produced by the aviation industry (1). Radiative forcing is the difference between the amount of energy entering Earth's atmosphere from the sun and amount of energy leaving the atmosphere. An energy imbalance is created when heat gets trapped in the earth's atmosphere. Although contrails can produce a cooling effect during the daytime by reflecting incoming shortwave radiation, they exert a net warming effect by absorbing and re-emitting outgoing longwave radiation, particularly at night, creating an energy imbalance and contributing to global warming (2). As a result, contrails represent a major non-CO_2 climate impact of aviation. Figure 1 explains the various factors impacting the formation of persistent contrail cirrus clouds in the upper troposphere region of the atmosphere.

Contrail mitigation research tends to focus on flight path and altitude optimization as potential pathways, yet targeted deployment of SAFs is often seen as a supplementary solution in the field (4, 5). However, SAFs have the potential to significantly lower CO_2 emissions in the aviation industry. SAFs are renewable jet fuels that can be blended with conventional jet fuel without requiring new aircraft or fueling infrastructure. Increased adoption of SAFs could help the aviation industry progress toward its goal of net-zero emissions by 2050; for example, a sustained annual SAF growth rate of 1–2% could reduce aviation CO_2 emissions by 5.5–9.5% over 15 years (6). The effect of SAFs on contrail formation is complex and this research aims to address both positive and negative impacts of SAFs that can contribute to persistent contrails in the atmosphere.

Contrail formation begins when activated soot particles first form liquid water droplets through condensation. At temperatures near 235 K, the droplets freeze via homogeneous ice nucleation, forming ice crystals (7). Through these processes, both soot particle and water vapor emissions affect contrail formation frequency and contrail radiative forcing. SAFs can reduce the non-volatile particulate matter (nvPM) emissions index by up to 70%, while fleetwide adoption of 100% SAF could lower annual mean contrail net radiative forcing by 44% (8, 9). On the other hand, SAFs have higher water vapor emissions and can therefore also increase the frequency of contrail formation (10). Contrail formation is governed by both engine emissions and atmospheric conditions and occurs only when the Schmidt-Appleman criterion (SAC) is satisfied. The SAC defines the threshold at which the mixing of hot, moist aircraft exhaust with cold, ambient air leads to saturation with respect to liquid water (11, 12). The SAC depends on many parameters that vary with different fuels, including combustion heat, emission indices, and engine thrust. Different values for these will result in different temperature thresholds for contrail formation. Differences among fuels arise primarily from variations in the hydrogen-to-carbon ratio, combustion efficiency, and water vapor production per unit energy released. For example, hydrogen combustion produces substantially more water vapor per unit heat released than conventional kerosene, resulting in higher likeliness for contrail formation (13). Conversely, kerosene exhibits higher particle mass and number emissions, which enhance ice crystal number concentrations once contrails form (14).

Whether this threshold is met depends on ambient temperature and pressure, atmospheric humidity, engine efficiency, and the amount of water vapor emitted during fuel combustion (15). If ambient temperatures are insufficiently low, contrails will not form (13). Consequently, atmospheric conditions are critical in determining contrail formation and their associated radiative forcing. Atmospheric humidity and ambient temperature and pressure vary at different points of the atmosphere and on different flight routes. Contrails form in regions with high humidity and low temperatures. When relative humidity with respect to ice exceeds 100%, the atmospheric zone is called an ice-supersaturated region. These regions are ideal for the formation of

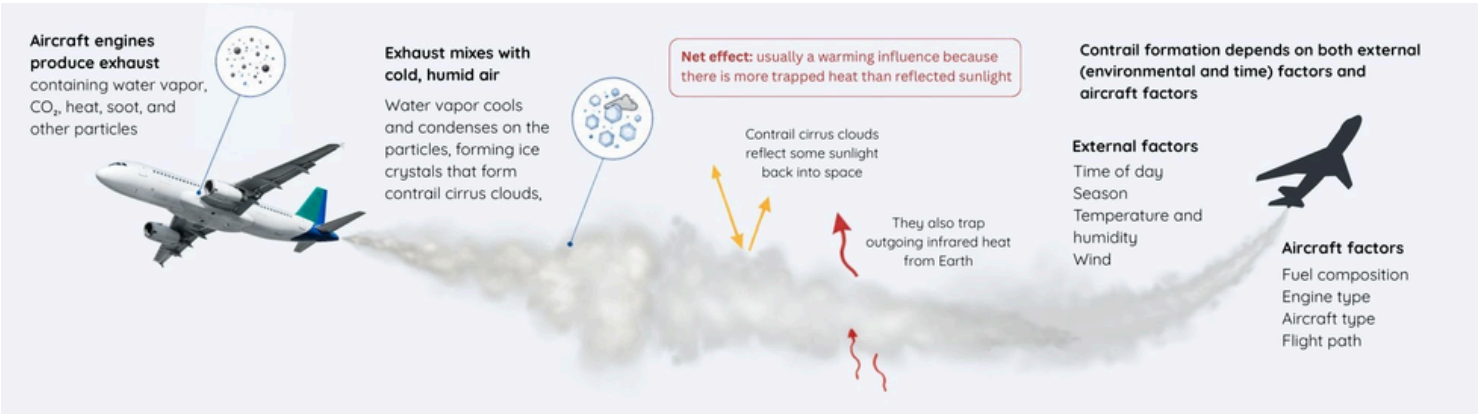


Figure 1. Contrail Cirrus Formation. Factors leading to contrail formation and positive radiative forcing.

cirrus clouds and the long-lasting persistence of contrails. It is important to examine the effects of SAFs on contrails, not at specific points, but on actual flight routes.

This paper aims to utilize the SAC to determine the effect of various SAF types on the temperature threshold for contrail formation, as well as apply this threshold on multiple routes to find the impact on actual contrail formation probability.

Methods

This research study combined a mathematical framework with a flight path modeling approach. The mathematical framework was utilized to have a rigorous handle on the threshold temperature calculation for different kinds of aviation fuels. These include kerosene (the baseline fuel), compared to Hydroprocessed Esters and Fatty Acids Synthetic Paraffinic Kerosene (HEFA-SPK), Fischer-Tropsch Synthetic Paraffinic Kerosene (FT-SPK), and Alcohol-to-Jet Synthetic Paraffinic Kerosene (ATJ-SPK), which are the most common types of SAFs currently being used. The flight modeling builds from the calculated threshold temperature and provides a way to understand relative impact of SAFs on flight routes of varying lengths.

Mathematical Framework

This framework was based on the SAC for contrail formation, as formulated by Schumann (13). Contrail formation is assumed to occur when the exhaust plume from an aircraft engine, mixing adiabatically with ambient air, becomes supersaturated with respect to ice. The threshold condition is defined by the critical ambient temperature, below which contrails form. The various factors that must be considered to determine threshold temperature include propulsion efficiency, slope parameter G, and the maximum temperature thresholds for liquid (T_{LM}) and ice-saturated formation (T_{IM}).

η is the propulsion efficiency of the aircraft which is dependent on engine type. It considers thrust (F), true air speed (V), specific combustion heat (Q), and rate of fuel flow (m_F).

$$\eta = \frac{FV}{Qm_F} \quad \text{(Eq. 1)}$$

The threshold temperature can be expressed in terms of slope parameter G, which characterizes the ratio of heat release to water vapor emission in the exhaust plume. This slope parameter is defined by:

$$G = \frac{EI_{H2O}c_p p}{\varepsilon Q(1-\eta)} \quad \text{(Eq. 2)}$$

ε is the ratio of gas constants or molar masses of dry air and water vapor and the value is taken as 0.622. c_p is the specific heat capacity and represents the exhaust gas temperature of modern jet engines. The value used in this study is 1050 J kg^{-1} K^{-1} at T = 600K., Other variables include the emission index of water vapor EI_{H2O} (the amount of water produced per unit fuel), the ambient air pressure at cruise altitude (p), and overall propulsion efficiency of the aircraft engine (η).

For the SAC, G is the thermodynamic parameter that describes how temperature decreases relative to water vapor increase when hot aircraft exhaust mixes with cold ambient air. Physically, it represents the slope of the exhaust-air mixing line in temperature-humidity space. This slope determines whether the mixture will reach saturation with respect to water during mixing. If the mixing line intersects with the saturation curve, condensation occurs, and a contrail can form. Thus, G directly controls the critical ambient temperature threshold for contrail formation.

A larger G means the mixture reaches saturation more easily, allowing contrails to form at relatively warmer temperatures. A smaller G requires colder ambient conditions for contrail formation. Therefore, G links aircraft engine thermodynamics with atmospheric conditions in determining when persistent contrails can develop.

Using G, T_{LM} and T_{IM} can be approximated using Equations 3 and 4. If the ambient temperature is higher than T_{LM}, no condensation occurs. If it is lower than T_{LM}, water droplets form, which generally freeze into ice crystals. If the temperature is lower than T_{IM} and the atmosphere is sufficiently humid, the contrail will persist and spread, which is the key factor in aircraft-induced radiative forcing and climate warming.

$$T_{LM} = -46.46 + 9.43\ln(G - 0.053) + 0.720[\ln(G - 0.053)]2 \quad \text{(Eq. 3)}$$

$$T_{IM} = -43.36 + 9.08\ln(G - 0.02) + 0.49[\ln(G - 0.02)]2 \quad \text{(Eq. 4)}$$

When relative humidity U = 1, $T_{(LM,\ IM)} = T_{(LC,\ IC)}$. When assuming relative humidity is at 100%, T_{LM} and T_{IM} can then be used as the threshold temperatures, where persistent contrails form at temperatures below T_{LM}. T_{LM}, T_{IM}, and G are used for calculating T_{LC} (liquid-condensation temperature).

$$T_{LC} = T_{LM} - x, x = -A + (A^2 + 2B)^{\frac{1}{2}}, \quad \text{(Eq. 5)}$$

$$with\ A = \frac{(1-U)G}{U^2 e_L^{\cdot}(T_{LM})},\ B = \frac{e_L C - e_E}{U^2 e_L^{\cdot}(T_{LM})},\ e_E = U_{e_L}(T_{LM})$$

e" is calculated from the Clausius-Clapeyron relationship, which estimates vapor pressure at different temperatures.
T_{LC} is the temperature at which the exhaust plume from an aircraft first becomes saturated with respect to liquid water as it mixes with the surrounding air. When the plume cools below T_{LC}, the water vapor emitted by the engine can condense into tiny liquid droplets.

For a visible contrail to form and persist, the ambient air temperature must be below T_{LC} so that condensation can begin. Contrails form when the surrounding atmosphere is cold enough that the mixing line of the exhaust plume crosses the ice saturation curve. Thus, T_{LC} marks the threshold for initial condensation, while

Table 1. Engine Parameters. CFM56-7B26 engine input parameters used to calculate propulsion efficiency, η.

	Unit	Value
Thrust per engine (*F*)	kN	32.295
True air speed (*V*)	m s^{-1}	233.89
Rate of fuel flow (*mF*)	kg s^{-1}	0.575
Relative humidity (*U*)	%	0.42

Table 2. Fuel Parameters. Combustion properties and values of kerosene and SAFs used in the analysis.

	Specific Combustion Heat (*Q*)	EI_{h2o}
kerosene (Baseline)	43.2	1.25
HEFA-SPK	44.1	1.37
FT-SPK	43.9	1.37
ATJ-SPK	44	1.38

T_{IC} determines whether those droplets freeze and create the ice crystals that make contrails visible and long-lasting. The T_{IC} can be calculated using the same formula as T_{LC} by replacing T_{LM} with T_{IM}. From this mathematical framework, key parameter values to be utilized during the flight modeling phase of the research were identified. The engine parameter values are based on one of the most commonly used commercial aircraft: a B737 equipped with a CFM56-7B26 engine (16,17).

Fuel parameters were calculated for kerosene, which is used as baseline fuel, and three commonly used sustainable aviation fuels. HEFA-SPK is produced by hydroprocessing biological feedstocks such as waste cooking oils, animal fats, or plant oils. It is currently the most commercially mature and widely available SAF and can be blended with conventional jet fuel at up to 50% without modification to existing engines. Experimental evidence from the Emission and Climate Impact of Alternative Fuels (ECLIF3) campaign has confirmed that 100% HEFA-SPK combustion reduces contrail ice crystal numbers by 56% compared to conventional Jet-A1 fuel, with modeled reductions in contrail radiative forcing of 26% (18). FT-SPK is produced via the Fischer-Tropsch process. The resulting fuel is highly refined with very low sulfur and aromatic content, contributing to notably reduced nvPM emissions relative to conventional kerosene. ATJ-SPK, produced by converting fermentation-derived alcohols, offers a flexible feedstock pathway, however, it currently has higher production costs than HEFA (19, 20).

Engine and Fuel parameters presented in Table 1 and Table 2 are consistent throughout all the routes modeled during this research. Table 3 presents an example calculation based on a flight at a pressure of 216.63 hPa. This calculation will change at different points in the flight based on flight altitude and atmospheric conditions.

Flight Modeling

The threshold temperature formula from the previous section was applied to specific routes to observe potential impact on contrail formation in an applied setting. Four representative routes were chosen to understand mitigatory effects of SAFs under varying atmospheric and fuel usage conditions. These include a long-haul, high latitude route (London–New York); a very long-haul, mid-to-high latitude route (Los Angeles–Tokyo); a medium-haul, subtropical latitude route (Paris–Dubai); and a short-haul, mid-latitude route (Berlin–Rome).

Flight paths for the four selected routes were constructed using the great circle method, which defines the shortest geodesic distance between two points on the surface of a sphere and serves as the standard geometric basis for long-haul aviation route planning. Since a great circle route requires continuous heading adjustment as it traverses the curvature of the Earth, direct implementation in flight simulations is not computationally feasible. Therefore, each route was discretized into a series of intermediate waypoints at regular intervals along the great circle arc, with each segment between consecutive waypoints approximated as a rhumb line of constant bearing (21). This waypoint-based discretization approach is well established in the aviation literature as a method for accurately representing geodesic flight trajectories and it helps maintain computational efficiency (22). The geographic coordinates of each waypoint, defined by latitude and longitude, were derived trigonometrically from the origin-destination airport pair using the haversine formula. This accounts for the spherical geometry of the Earth and has been validated for route distance calculations in global aviation fuel and emissions modelling (22). For each of the four routes, origin and destination coordinates were used from OurAirports database as shown in Table 4.

Atmospheric data (pressure, temperature, and relative humidity) for each waypoint was implemented through accessing the ERA5 atmospheric reanalysis data which is the fifth generation of climate data produced by the European Center for Medium range weather forecasts.

Table 3. Example calculations for contrail formation parameters for a flight at pressure of 216.63 hPa. Overall efficiency (η), thermodynamic factor (G), and critical threshold temperatures for kerosene and SAF alternatives at cruise pressure.

	η	G (Pa.K^{-1})	T_{LM} (°C)	T_{IM} (°C)	T_{LC} (°C)	T_{IC} (°C)
kerosene (Baseline)	0.3040852476	1.452438883	-43.21	-40.03	-43.95	-40.70
HEFA-SPK	0.2978794262	1.545602895	-42.57	-39.44	-43.27	-40.08
FT-SPK	0.2992365079	1.555651179	-42.50	-39.37	-43.19	-40.01
ATJ-SPK	0.2985564249	1.561929085	-42.46	-39.34	-43.15	-39.97

Table 4. Origin and destination Latitude and Longitude of 8 airports on 4 flight routes from Our Airports database (23). Latitude and Longitude for origin and destination airport used to calculate waypoints along the route.

IATA Code	Airport Name	Latitude	Longitude
LHR	London Heathrow	51.47001	-0.4543
JFK	New York JFK	40.6413	-73.7781
LAX	Los Angeles	33.9416	-118.4085
NRT	Tokyo Narita	35.7720	140.3929
CDG	Paris CDG	49.0097	2.5479
DXB	Dubai	25.2532	55.3657
BER	Berlin Brandenburg	52.3667	13.5033
FCO	Rome Fiumicino	41.8003	12.2389

This data was accessed by using the Copernicus Climate Data Store Application programming interface or API. Data averaged over three months (December, January, February) was used, as there is a higher likelihood of contrail formation during winter. The data was finally converted into a probability function that compares probability of contrail formation ($T_{IC} < T_{ambient} \leq T_{LC}$) and the impact of SAF type on each flight route. The probability was calculated from 0 to 1 based on threshold temperature and relative humidity across all waypoints on the flight route.

Results and Discussion

As a first step, the overall probability of contrail formation along each flight path was calculated. As contrail formation is governed by atmospheric temperature and humidity, which change with latitude and altitude, the four routes in different latitudinal bands could capture different atmospheric conditions (1).

Figure 2 presents the distribution of contrail formation probability along each of the four representative flight routes using baseline kerosene fuel. The color scale represents the likelihood of contrail formation at the waypoints along each flight route.

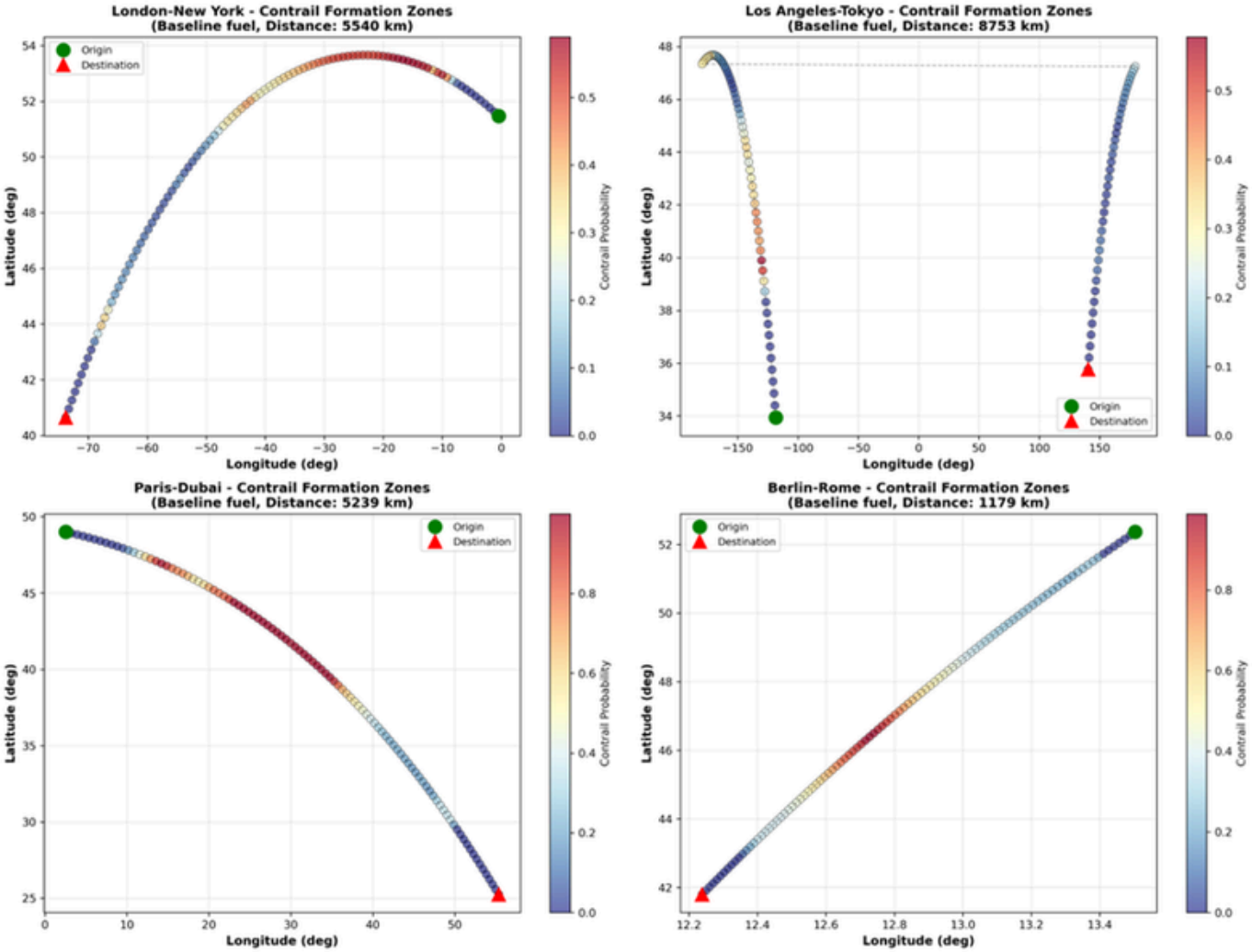


Figure 2. Probability of Contrail formation along the flight path on each route. Probability of 0.0 to 1.0 indicated at each waypoint along the flight path.

The London–New York route (5,540 km) exhibited a characteristic arc of elevated contrail probability concentrated along the high-latitude mid-route section, reaching peak values of approximately 0.5–0.6 near 54°N. This reflects the lower ambient temperatures encountered at high latitudes during transatlantic cruise, which more frequently satisfy the SAC threshold temperature condition. Probability falls toward zero near both endpoints as the route descends to lower latitudes where ambient temperatures are warmer relative to the threshold.

The Los Angeles–Tokyo route (8,753 km) demonstrated that the contrail formation probability is highest in the polar arc section of the route, concentrated in the first half of the flight where the great circle trajectory reaches its highest latitudes. This part of the flight path represents the extremely cold temperature where the SAC threshold is almost always met. As the route curved southward, probability decreased, reflecting the transition to warmer, drier subtropical conditions in the western Pacific. This asymmetry has implications for contrail avoidance strategies, as mitigation efforts on this route would be most effective during the departure phase.

The Paris–Dubai route (5,239 km) displayed the most pronounced latitudinal gradient, with high contrail formation probabilities near the origin at approximately 47°N, declining continuously toward the destination. Peak probabilities approached 1.0 in the northern section of the route over central Europe, where cold upper-tropospheric conditions and higher relative humidity consistently satisfy the SAC. This contrasts with near-zero probabilities closer to the Arabian Peninsula, where the warm and dry subtropical atmosphere provides strongly unfavorable conditions for contrail formation. This route therefore illustrates the strongest dependence on latitude of all routes studied.

The Berlin–Rome route (1,179 km) showed low contrail formation probability along most of its length. The short cruise phase limited exposure to cold upper-tropospheric air, and the mid-latitude trajectory over central Europe did not reach the high-altitude, cold conditions required to consistently satisfy the SAC. Even though peak contrail formation probability was high, it was a smaller proportion of the total route length when compared to contrail-forming conditions for the long-haul routes.

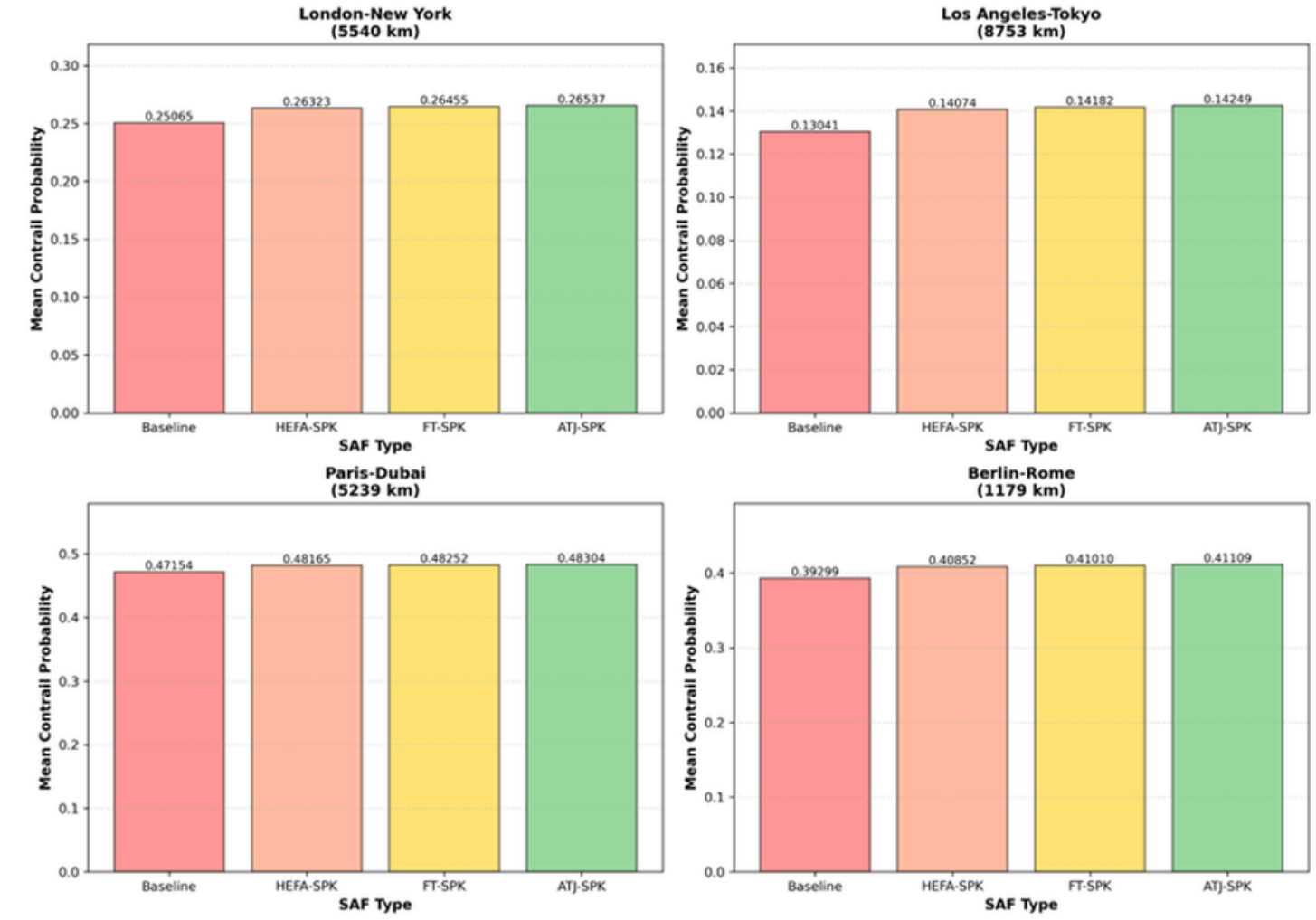


Figure 3. Mean Contrail Probability for kerosene and SAFs on flight routes. Mean Contrail Probability along each of the flight paths used to compare overall impact of SAFs, instead of along each waypoint on the routes.

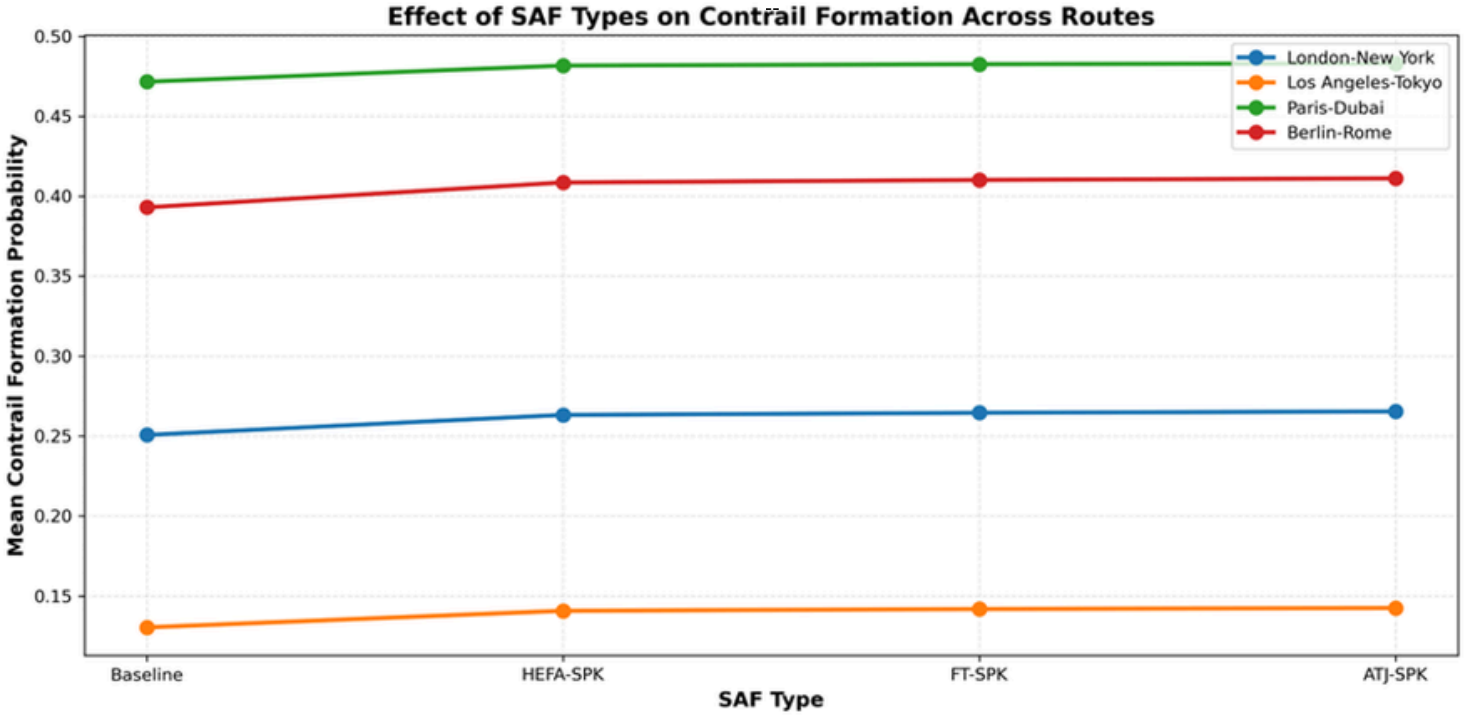


Figure 4. Effect of SAFs on mean contrail formation probability. Fuel types have similar impact on contrail formation across flight routes.

Taken together, the four routes supported that contrail formation probability is governed by ambient temperature, which, in turn, is driven by latitude and cruise altitude.

When studying the impact of SAFs on each of the routes, mean contrail probability for each route was utilized as an evaluative measurement. Figure 3 presents the mean contrail formation probability for each SAF type across the four representative flight routes, shown as bar charts per route. Figure 4 presents a combined line plot for cross-route comparison. Across all routes, mean contrail formation probability increased marginally from baseline kerosene to each SAF blend, with ATJ-SPK consistently recording the highest values, and HEFA-SPK the lowest among the three alternative fuels.

In Figure 3, for the London–New York route, mean probability rose from 0.251 for baseline kerosene to 0.265 for ATJ-SPK, representing an increase of 1.4 percentage points. A comparable pattern was observed on the Los Angeles–Tokyo route, where probabilities increased from 0.130 to 0.142 across the same fuel range. The Paris–Dubai route recorded the highest absolute probabilities of all four routes, rising from 0.472 for baseline to 0.483 for ATJ-SPK, consistent with the high-probability in the northern section of the route identified in Figure 1. The Berlin–Rome route similarly showed an increase from 0.393 to 0.411, which, while numerically larger than the London–New York shift, must be interpreted in the context of the short route length and limited total contrail exposure discussed previously.

This counterintuitive increase in contrail formation probability with SAFs is attributable to the thermodynamic properties of the fuels rather than a failure of mitigation. As shown in Table 3, SAFs exhibit higher water vapor emission indices and higher specific combustion heat relative to baseline kerosene. These properties combine to increase the slope parameter G, which shifts the SAC threshold temperatures T_{LC} and T_{IC} to slightly warmer values. This expands the range of atmospheric conditions under which the criterion is satisfied. This thermodynamic effect is well established in literature and has been noted as a trade-off inherent to the fuel chemistry of low-aromatic, synthetically derived aviation fuels (13, 24).

It is important to note that this thermodynamic effect acts in isolation within the current model framework. The primary mitigatory mechanism of SAFs is the reduction of nvPM emissions, which limits the availability of ice nucleation sites in the exhaust plume. This operates through a separate physical pathway not captured by the SAC threshold comparison alone. The results presented here therefore represent a conservative upper bound of all contrail formation and does not separate the persistent contrail formation that SAFs tend to mitigate. The net climatic impact of SAF deployment is expected to be beneficial when nvPM suppression of ice crystal nucleation is accounted for. This is discussed further in the conclusion section.

Figure 4 consolidates the cross-route comparison, and it supports that the fuel-related trend is consistent in direction across all routes but differs in magnitude. Paris–Dubai maintained the

highest mean probability throughout, followed by Berlin–Rome, London–New York, and Los Angeles–Tokyo. The parallel trajectories of all four routes across the fuel axis indicate that the thermodynamic response to SAF blend properties is route-independent. The shift in G affects threshold temperatures uniformly regardless of the atmospheric conditions encountered along a given trajectory.

Conclusion

This study assessed the mitigatory effects of three sustainable aviation fuels, HEFA-SPK, FT-SPK, and ATJ-SPK, on contrail formation across four representative flight routes. Contrail formation probability at each waypoint was determined by evaluating whether ambient temperature satisfied the SAC threshold condition. The threshold temperatures T_{LC} and T_{IC} were computed from fuel-specific thermodynamic parameters including propulsion efficiency, slope parameter G, and ambient relative humidity. The four routes were selected to span a representative range of operational conditions encountered in global commercial aviation, from high-latitude, long-haul corridors to short-haul, intra-European services.

Across all four routes, the thermodynamic properties of SAFs, specifically their higher water vapor emission indices and specific combustion heat relative to baseline kerosene, resulted in a marginal increase in mean contrail formation probability. The scope of this research was kept narrow to focus on the thermodynamic effect which shifts the SAC threshold temperatures T_{LC} and T_{IC} to slightly warmer values, broadening the range of atmospheric conditions under which the criterion is satisfied. This thermodynamic effect is well established in the literature and does not contradict the mitigatory potential of SAFs. The results presented here should therefore be interpreted as reflecting the thermodynamic component of contrail formation in isolation, in the absence of nvPM effects.

Several directions for future research emerge directly from the findings and limitations of this work. The immediate extension is the integration of nvPM emission indices into the probability framework. Combining the temperature threshold computation with a nvPM activation model to produce a net contrail formation probability will account for both thermodynamic and particle-based effects. This will also give probability for more persistent contrails rather than all contrails including transient ones. Additionally, current jet engines operate in what is known as the soot-rich regime, where a high concentration of nvPM particles are available to act as ice nucleation sites. In this regime, reducing nvPM emissions directly reduces the number of ice crystals that form, which produces a clear mitigatory benefit. However, as SAF blend ratios increase and nvPM emissions fall substantially, engines may transition into the soot-poor regime (25). Here, the relationship between nvPM concentration and ice crystal formation breaks down, and smaller volatile particles in the exhaust plume begin to take over as the main nucleation sites. Under these conditions, further reductions in nvPM no longer guarantee fewer contrail ice crystals. Understanding where this transition occurs and key factors affecting it is essential for confidently predicting the climate benefit of high-blend SAF usage.

A second area in need of further investigation is the sensitivity of route-level contrail formation probability to atmospheric data resolution. ERA5 reanalysis data, while widely used and validated, operates at approximately 31 km horizontal resolution. This is wider than the spatial scale of many ice-supersaturated regions, which can have horizontal extents of only a few kilometers (26). This limitation likely causes the model to underestimate both the frequency and variability of persistent contrail conditions, particularly on routes traversing complex regions such as the North Atlantic storm track. Higher resolution atmospheric data would better capture the variability of ice-supersaturated regions and reduce uncertainty regarding when and where contrail formation conditions are met along each route.

Finally, translating the threshold temperature-based formation probability metric developed here into a direct measure of climate impact presents a valuable extension of this work. The current framework weighs all contrail-forming waypoints equally, irrespective of the time of day, surface albedo, or season. These factors can substantially change the net radiative forcing of any given contrail (1,3). Incorporating radiative forcing weighting into the waypoint probability calculation would convert the metric from a formation indicator into a climate impact index. This will enable route-specific and fuel-specific recommendations to be expressed in terms of warming potential rather than formation frequency alone. Combined with targeted SAF deployment strategies, such a framework could directly inform airline fuel allocation decisions that maximize climate benefit of SAF used.

Acknowledgements

The author thanks Dr. Nafiz Chowdhary from Oxford Thermofluids Institute, University of Oxford for his guidance during this research.

References

1. D.S. Lee, D. W. Fahey, A. Skowron, M. R. Allen, U. Burkhardt, Q. Chen, S. J. Doherty, S. Freeman, P. M. Forster, J. Fuglestvedt, A. Gettelman, R. R. De León, L. L. Lim, M. T. Lund, R. J. Millar, B. Owen, J. E. Penner, G. Pitari, M. J. Prather, R. Sausen, L. J. Wilcox. The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. Atmos Environ 244, e117834 (2020). 10.1016/j.atmosenv.2020.117834
2. R. Meerkötter, U. Schumann, D. R. Doelling, P. Minnis, T. Nakajima, Y. Tsushima. Radiative forcing by contrails. Ann. Geophys. 17, 1080–1094 (1999).
3. J. Cathcart, A. Chen, J. Majholm, A. Jardine Wall. “Aviation Contrails: What We Know—and What We Don’t—about This Warming Phenomenon” (RMI, 2024); https://rmi.org/aviation-contrails-what-we-know-and-what-we-dont-about-this-warming-phenomenon/

4. E. Roosenbrand, J. Sun, J. Hoekstra. Contrail minimization through altitude diversions: A feasibility study leveraging global data. Transp. Res. Interdiscip. Perspect. 22, e100953. 10.1016/j.trip.2023.100953
5. J. Cathcart, A. Chen. “Contrail Mitigation: A Collaborative Approach in the Face of Uncertainty” (RMI, 2022); https://rmi.org/contrail-mitigation-a-collaborative-approach-in-the-face-of-uncertainty/
6. S. Sgouridis, P. A. Bonnefoy, R. J. Hansman. Air transportation in a carbon constrained world: Long-term dynamics of policies and strategies for mitigating the carbon footprint of commercial aviation. Transp. Res. A: Policy Pract. 45, 1077–1091 (2011).
7. B. Kärcher, U. Burkhardt, A. Bier, L. Bock, I. J. Ford. The microphysical pathway to contrail formation. J. Geophys. Res. Atmos. 120, 7893–7927 (2015).
8. L. Durdina, B. T. Brem, M. Elser, D. Schönenberger, F. Siegerist, J. G. Anet. Reduction of nonvolatile particulate matter emissions of a commercial turbofan engine at the ground level from the use of a sustainable aviation fuel blend. Environ. Sci. Technol. 55, 14576–14585 (2021).
9. R. Teoh, U. Schumann, A. Majumdar, M. E. J. Stettler, M. E. J. Targeted use of sustainable aviation fuel to maximize climate benefits. Environ. Sci. Technol. 56, 17246–17255 (20220.
10. M. Narciso, J. M. Melo de Sousa. Influence of sustainable aviation fuels on the formation of contrails and their properties. Energies, 14, e5557 (2021). 10.3390/en14175557
11. E. Schmidt. (1941). Die Entstehung von Eisnebel aus den Auspuffgasen von Flugmotoren [The formation of ice fog from the exhaust gases of aircraft engines]. Schriften der Deutschen Akademie der Luftfahrtforschung 44, 1–15 (1941).
12. H. Appleman. The formation of exhaust condensation trails by jet aircraft. Bull. Am. Metero. Soc. 34, 14–20 (1953).
13. U. Schumann. On conditions for contrail formation from aircraft exhausts. Meteorologische Zeitschrift 5, 4–23 (1996).
14. G. Quante, C. Voigt, M. Kaltschmitt. Targeted use of paraffinic kerosene: Potentials and implications. Atmos. Environ.: X 23, e100279 (2024). 10.1016/j.aeaoa.2024.100279
15. B. Kärcher, C. Voigt. Susceptibility of contrail ice crystal numbers to aircraft soot particle emissions. Geophys. Res. Lett. 44, 8037–8046 (2017).
16. R. Balas, C. Danilet, M. Stefan. “Boeing aircraft overview 2025: Complete fleet guide. The Flying Engineer (2025).” https://theflyingengineer.com/boeing-aircraft-overview/
17. “AM International CFM56 — engine specs & aircraft. PlaneFYI.” https://planefyi.com/engines/cfm56/
18. S. Märkl, C. Voigt, D. Sauer, R. K. Dischl, S. Kaufmann, T. Harlaß, V. Hahn, A. Roiger, C. Weiß-Rehm, U. Burkhardt, U. Schumann, A. Marsing, M. Scheibe, A. Dörnbrack, C. Renard, M. Gauthier, P. Swann, P. Madden, D. Luff, P. Le Clercq. Powering aircraft with 100% sustainable aviation fuel reduces ice crystals in contrails. Atmos. Chem. and Phys. 24, 3813–3837 (2024).
19. E. Cabrera, J. M. Melo de Sousa. Use of Sustainable Fuels in Aviation—A Review. Energies 15, e2440 (2022). 10.3390/en15072440
20. ICAO Committee on Aviation Environmental Protection. “Assessment report on fuel composition effects on non-volatile particulate matter (nvPM) emissions” (International Civil Aviation Organization, 2022); https://www.icao.int/sites/default/files/sp-files/environmental-protection/Documents/ScientificUnderstanding/ICAO-CAEP12-Assessment-Report-on-fuel-composition-effects-on-nvPM-emissions.pdf
21. E. Williams (2024). Aviation Formulary V1.47
22. K. Seymour, M. Held, G. Georges, K. Boulouchos. Fuel estimation in air transportation: Modeling global fuel consumption for commercial aviation. Transp. Res. D: Transp. Environ.88, e102528 (2020). 10.1016/j.trd.2020.102528
23. Our Airports. "OurAirports Data" (OurAirports, 2024); https://ourairports.com/data/
24. J. Ponsonby, L. King, B. J. Murray, M. E. J. Stettler. Jet aircraft lubrication oil droplets as contrail ice-forming particles. Atmos. Chem. and Phys. 24, 2045–2058 (2024).
25. B. Kärcher, F. Yu. Role of aircraft soot emissions in contrail formation. Geophys. Res. Lett. 36, e01804 (2009). 10.1029/2008GL036649
26. K. Gierens, P. Spichtinger. "On the size distribution of ice-supersaturated regions in the upper troposphere and lowermost stratosphere." Ann. Geophys. 18, 499–504 (2000).

Forest-Fire Intensity, and Age-Standardized Heart-Attack Hospitalization Rates in Canada, 2014–2022: An Ecological Observational Time-Series Study

Intensité des feux de forêt et taux d'hospitalisation standardisée par âge des crises cardiaques au Canada, 2014–2022 : une étude écologique observationnelle en série temporelle

Jack Madden[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: jackmadden1078@gmail.com

Abstract | Résumé

Wildfires have become more prevalent and intense with climate change. The inhalation of wildfire smoke is known to increase systemic inflammation and oxidative stress, increasing rates of respiratory disease and potentially some cardiovascular diseases. while some studies link exposure to particulate matter of size 2.5 microns ($PM_{2.5}$) with respiratory and cardiovascular outcomes, the effect on heart-attack hospitalizations, particularly at the national scale in Canada, remains unclear. This ecological observational time-series study used 108 province-year observations between 2014 and 2022 from the CIHI (Canadian Institute for Health Information) age-standardized heart-attack hospitalization rates database and the National Forestry Database for all 10 Canadian provinces and three territories. The outcome variable was age-standardized heart-attack hospitalization rates per 100,000, and the main exposure was a wildfire exposure index aggregating wildfire intensity metrics into a single variable. Using multiple linear regression with province-fixed effects adjusting for year ($\alpha = 0.05$). Wildfire exposure was not significantly associated with heart-attack hospitalization rates ($B = -0.0017$, $p = 0.338$). The year showed significant negative association with heart-attack hospitalization rates ($B = -5.25$ per 100,000 per year, $p < 0.001$). The model explained a large proportion of the variance in heart-attack hospitalization rates (adjusted $R^2 = 0.82$). At the province-year scale, wildfire intensity does not appear to be a major driver of annual heart-attack hospitalization rates. Variation in heart-attack rates is more driven by provincial differences and secular declines over time. More individual-level studies using high-resolution wildfire-smoke exposure data are required to clarify potential cardiovascular impacts, especially at the short-term scale.

———

Les incendies de forêt sont devenus plus fréquents et intenses avec le changement climatique. L'inhalation de fumée des feux de forêt est connue pour augmenter l'inflammation systémique et le stress oxydatif, augmentant les taux de maladies respiratoires et potentiellement certaines maladies cardiovasculaires. Bien que certaines études associent l'exposition aux particules d'une taille de 2,5 microns (PM2,5) aux résultats respiratoires et cardiovasculaires, l'effet sur les hospitalisations pour crise cardiaque, en particulier à l'échelle nationale au Canada, reste incertain. Cette étude écologique observationnelle en série temporelle a utilisé 108 observations provinciales-années entre 2014 et 2022 issues de la base de données standardisée des taux d'hospitalisation des crises cardiaques de l'Institut canadien de la santé (CIHI) et de la base de données nationale des forêts pour les 10 provinces canadiennes et trois territoires. La variable de résultat était les taux d'hospitalisation standardisés par âge pour 100 000 habitants, et l'exposition principale était un indice d'exposition aux feux de forêt regroupant les métriques d'intensité des feux en une seule variable. En utilisant une régression linéaire multiple avec des effets fixés par province ajustant pour l'année ($\alpha = 0,05$). L'exposition aux feux de forêt n'était pas significativement associée aux taux d'hospitalisation pour crise cardiaque ($B = -0,0017$, $p = 0,338$). L'année a montré une association négative significative avec les taux d'hospitalisation par crise cardiaque ($B = -5,25$ pour 100 000 par an, $p < 0,001$). Le modèle expliquait une grande partie de la variance des taux d'hospitalisation pour crise cardiaque ($R2$ ajusté = 0,82). À l'échelle provinciale et annuelle, l'intensité des feux de forêt ne semble pas être un facteur majeur des taux annuels d'hospitalisation par crise cardiaque. La variation des taux d'infarctus est davantage due aux différences provinciales et à une réduction séculaire. Davantage d'études individuelles utilisant des données d'exposition à haute résolution à la fumée d'incendie de forêt sont nécessaires pour clarifier les impacts cardiovasculaires potentiels, en particulier à court terme.

Keywords: Wildfire; Myocardial Infarction; Ecological Study; Canada

Introduction

There is no doubt that climate change has many deleterious effects on the environment which have been detracting from quality of life over the last century. The global surface temperature has been increasing rapidly, with rates tripling over the last 40 years (1). Simultaneously, rates of wildfires have been steadily rising for centuries, with rates in 2023/2024 doubling those from 2000–2022 (2). Canada has seen a dramatic increase in the frequency of wildfires in recent years. By the end of 2023, Canada saw record-breaking numbers with more than 6,000 fires and over 2.5 million hectares burned (3). A study conducted by the World Weather Attribution found that the likelihood of wildfires in Quebec was at least doubled as a direct result of climate change (4). Climate change has increased the frequency of wildfires through multiple mechanisms: higher temperatures lead to low humidity, drier plants, and higher winds to spread the fire. Additionally, climate-driven atmospheric circulation shifts have prolonged fire seasons, while also altering large-scale weather patterns, trapping some regions under heat waves. Independent of climate change, fuel accumulation plays a lesser role. Areas which have suppressed natural fires for decades, especially urban regions of Canada, have allowed dense and crowded forests to grow, increasing their risk of ignition and propagation (5).

As wildfire activity increases, public health concerns are not limited to the immediate danger of burns, evacuations, and property loss; wildfires also generate smoke that can travel long distances and expose large populations to fine particulate matter. Wildfires are linked to several pathologies due to the release of particulate matter with size of particle 2.5 microns ($PM_{2.5}$). In Canada about 16% of $PM_{2.5}$ comes from wildfire smoke, which is linked to increased mortality, especially from respiratory and cardiovascular diseases (6). $PM_{2.5}$ also spreads rapidly over long distances far from the source, posing a risk to surrounding communities (6). $PM_{2.5}$ raises mortality by triggering systemic inflammation and oxidative stress. (Pei et. al. (7)) demonstrated that in ApoE-deficient mice (a model prone to heart disease), oxidative stress in the form of reactive oxygen species and lipid peroxidation, as well as systemic inflammation, inflammatory cytokines were elevated in response to real ambient $PM_{2.5}$ exposure. These processes damage blood vessels and promote atherosclerosis, increasing the risk of acute and chronic cardiovascular pathologies, including heart-attack, stroke, arrhythmias, and heart failure (8). The theoretical link between wildfires and cardiovascular pathologies in humans raises concerns considering the acceleration of climate change and the pervasive nature of air pollution.

Several earlier studies have examined the link between wildfire-smoke and cardiovascular events. (Delfino et. al. (9)) conducted a time-series/panel analysis which focused on the effect of wildfire-related $PM_{2.5}$ exposure on daily hospital admissions for cardiovascular and respiratory disease in seven million Southern California residents during the 2003 Southern California wildfires. They found a strong increase in various types of respiratory admissions, but weak evidence supporting an effect on cardiovascular incidents. (Johnston et. al. (10)) explored the effects of bushfire-related particulate matter with size of particle 10 microns (PM_{10}) emissions on respiratory and cardiovascular mortality in Sydney, Australia. They ran a 13-year time-series study on roughly four million residents from 1994–2007 and again found strong evidence linking the air pollutant to respiratory mortality, but no consistent association for cardiovascular-related mortality. Since PM_{10} is known to pose fewer health risks than $PM_{2.5}$, it is largely unsurprising that the cardiovascular findings were consistent with that of (9).

While time-series/panel analysis studies are useful epidemiological study designs, they are limited in their ability to identify weak associations due to their lack of control for between-person confounders. (Haikerwal et. al. (11)) was the first group to conduct a case-crossover study in examining the link between wildfire $PM_{2.5}$ and acute cardiac pathologies. The study examined all adult OHCA (Out-of-Hospital Cardiac Arrest) and ischemic events reported in Victoria, Australia from December 2006 to January 2007. The case-crossover design allows for individuals to serve as their own control, comparing their exposure to $PM_{2.5}$ around the day they had a heart-attack to exposure the weeks to months before they had it. This improvement in study design allowed them to identify a statistically significant increase in cardiac arrest and ischemic heart disease on high-smoke days, one of the first pieces of evidence that wildfire smoke may be a factor in triggering acute cardiac events. However, this effect was only found in a narrow time window during one fire season, which may restrict the generalizability of their findings to other regions or less severe fire seasons.

(Liu et. al. (12)) conducted a large multi-state cohort with exposure modelling on over five million Medicare enrollees (65+) from 2004–2009. They compared wildfire-specific $PM_{2.5}$ exposure to hospital admissions for cardiovascular and respiratory disease. They found no significant evidence of wildfire exposure affecting cardiovascular admissions, further obscuring the effects of wildfire exposure on cardiovascular pathologies.

(Hao et. al. (13)) followed about 22 million Americans (65+) from 2007–2018 using detailed air-pollution models to differentiate wildfire smoke from other ambient sources of $PM_{2.5}$. Using a long-term cohort analysis, they analyzed who developed heart failure subsequently. Uniquely, this study measures the long-term effects of wildfire exposure on heart failure, rather than short-term effects during the days and weeks of a fire. They found a 1.4% increase in risk of heart failure for every extra 1 $\mu g/m^3$ of average wildfire $PM_{2.5}$ exposure over the two-year period. This is among the strongest evidence linking wildfire exposure to an increased risk of heart failure, making the positive association between wildfire exposure and heart-attack rates seem more probable, but the effect is still unknown.

Most studies show clear respiratory effects from wildfire smoke but the effects on myocardial infarction are mixed. The focus tends to be on short-term effects of wildfire smoke on overall

cardiovascular disease, but there is no Canada-wide, multi-year analysis explicitly linking provincial fire activity to age-standardized heart-attack hospitalization rates. Therefore, this study seeks to quantify the association between wildfire intensity and age-standardized heart-attack hospitalization rates across Canadian provinces/territories from 2014–2022. This study hypothesizes that higher wildfire activity is associated with an increase in heart-attack hospitalization rates, even after adjusting for year and province.

Methods

This study used an observational ecological panel design, with 108 province-year observations from 2014–2022 in Canada. The sample population represents all Canadian provinces and territories. This analysis used de-identified, aggregate, and publicly available data and thus did not require individual consent or research ethics board approval.

This analysis compared the rates of heart-attack hospitalization and intensity of Canadian wildfires using the hospitalized heart-attack dataset from the CIHI (14) and the National Forestry Database from the Council of Canadian Forest Ministers (15). The datasets overlapped between 2014 and 2022; data in these years were used in the analysis. The sample size was 108 province-year observations for each fire class size.

Outcome

The outcome variable is age-standardized heart-attack hospitalization rate per 100,000 Canadians (continuous). Rates were already age standardized by CIHI to the 2011 standard population. This variable includes all patients age 18 or older who were admitted to an acute care hospital. It only measures events which are either the first ever hospitalization for acute myocardial infarction (AMI), or that happen more than 28 days after the admission date for the previous AMI admission. Therefore, the outcome captures hospitalized AMI events but does not include any out-of-hospital cardiac arrests, fatal myocardial infarctions occurring before hospitalization, nonhospitalized ischemic events, unstable angina, or arrhythmias.

Exposure

Wildfire intensity (continuous) was measured using the annual counts for each fire class size per province in the National Forestry Database. These size categories were grouped by hectare (ha) range (10,000 m²). These ranges were: up to 0.1 ha; 0.11–10 ha; 10.1–100 ha; 100.1–1 000 ha; 1 000.1–10 000 ha; 10 000.1–100 000 ha, and over 100 000 ha. Since larger fires have the capacity to produce more harm, an exposure index combines all individual size categories into a single level statistic, weighing the smallest category by , the second smallest by , the third smallest by , etc. The combination of these ranges into one statistic reduces multicollinearity, while maintaining the relative fire-size effects on human health. The weights were chosen as a heuristic ordinal scoring approach to preserve the increasing severity of larger fire-size categories while avoiding the inclusion of multiple highly correlated fire-size variables in the regression model. These weights were not intended to estimate actual emissions, burned area, or $PM_{2.5}$ concentrations. Because the weighting scheme is arbitrary, the exposure index should be interpreted only as a relative indicator of annual wildfire activity, not as a validated measure of smoke exposure.

This index functions as an ecological proxy rather than a direct measure of population-level smoke exposure. The index is based on fire size, frequency, and does not incorporate population density, smoke dispersion, exposure duration, or any direct measure of $PM_{2.5}$. It is assumed larger fires have a greater potential to generate harmful smoke, but it should not be interpreted as equivalent to population-level $PM_{2.5}$ exposure.

Covariates

Year (continuous) accounted for national temporal trends, while province (categorical) was used to account for unmeasured province-level differences such as geography, demographics, healthcare access, structural differences, and chronic disease burden, which typically do not vary meaningfully over time. Alberta was used as the reference category.

Statistical *analysis*

Descriptive statistics were used to summarize the two primary study variables: age-standardized heart-attack hospitalization rates and exposure index. The means, standard deviations and ranges were used to characterize the overall variability across provinces and years. These results are presented in Table 1. Two figures were used to visualize province-fixed effects, yearly trends, and outliers. A third figure was used to visualize the adjusted prediction plot of the linear model, with 95% CI.

The main analysis used multiple linear regression with province fixed effects to estimate the association between heart-attack hospitalization rates, and exposure index. These results are shown in Table 2. Model assumption checks indicated no multicollinearity concerns (VIF < 1.5). A Q-Q plot was used to assess normality of residuals, which found approximate normality with slight deviation in the upper tail. Statistical significance was defined as $\alpha = 0.05$ (two-sided). All analyses were performed using jamovi version 2.3 (16).

Results

Across the 108 province-year observations, the mean heart-attack hospitalization rate across provinces between 2014–2022 was 322.70 per 100,000 (SD = 61.65), with rates ranging between 212.82–479.95. The average exposure index was 1,942.30 (SD = 2,095.50), with a range between 0.00–10,387.00, reflecting high variation in yearly wildfire activity from province to province (Table 1).

Figure 1 shows that there is substantial variability in heart-attack hospitalization rates across provinces and territories. British Columbia, Saskatchewan, and Alberta had the lowest rates, while Quebec, Newfoundland and Labrador, and New Brunswick had the

Table 1. Descriptive statistics for age-standardized heart-attack hospitalization rates, and wildfire-intensity categories across 108 province-year observations, Canada (2014–2022). Statistics are based on 108 province–year observations. Heart-attack hospitalization rates are age-standardized per 100,000 population; the exposure index reflects annual wildfire intensity

Statistic	Age-Standardized Heart-Attack Rate per 100,000	Exposure Index
N	108	108
Mean	322.70	1942.30
Standard Deviation	61.65	2095.50
Minimum	212.82	0
Maximum	479.95	10387

highest. Many provinces had large interquartile ranges, and scattered outlier values, reflecting high year-to-year variability.

Rates of heart-attack hospitalization generally declined between 2014, and 2022. The interquartile ranges for each year had some area of overlap with other years, reflecting province-level differences. Quebec in 2019 was marked as an outlier for its significantly higher rate of heart-attack hospitalizations in that year compared to other provinces/territories (Figure 2).

The regression model explains approximately 83% of the variance in heart-attack hospitalization rates. The overall F test ($F(13, 94) = 34.878$, $p < 0.001$) was significant, indicating that the predictors collectively accounted for substantial variation (Table 2). The year fit showed that national rates of heart-attack hospitalization significantly decreased over time ($p < 0.001$), with each calendar year associated with five fewer hospitalizations per 100,000, controlling for fires, and province differences.

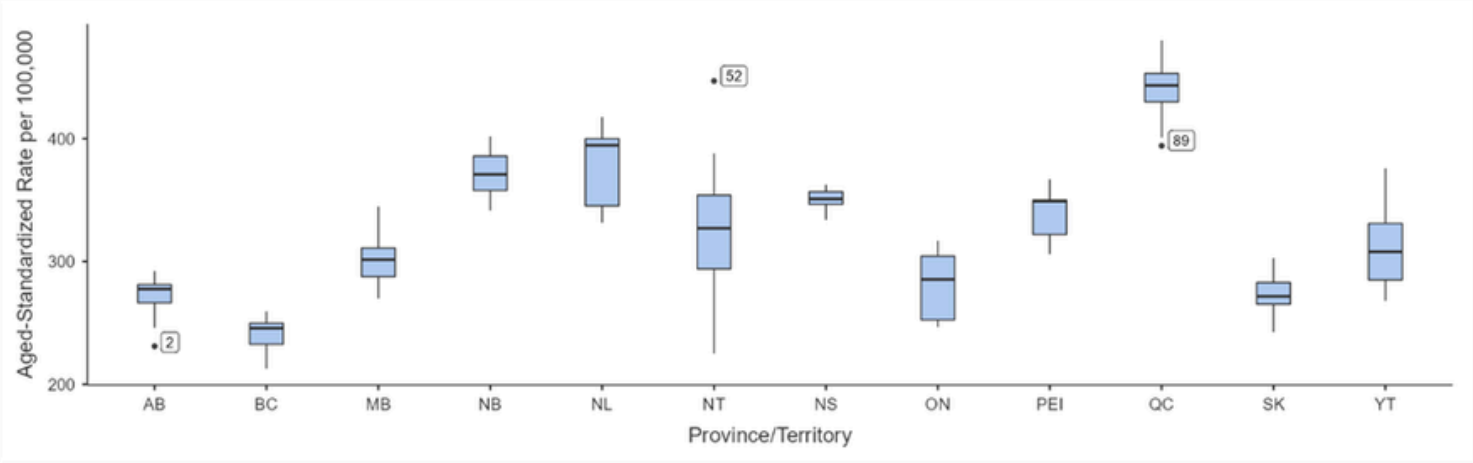


Figure 1. Distribution of age-standardized heart-attack hospitalization rates by province/territory, Canada, 2014-2022. Boxplots show the median, interquartile range, and range of values by province/territory; points indicate outliers.

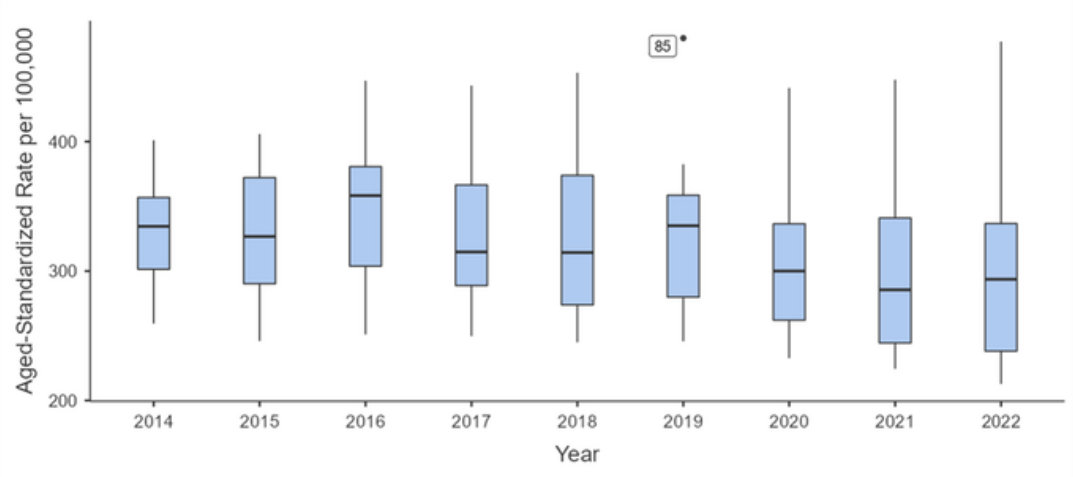


Figure 2. Distribution of age-standardized heart-attack hospitalization rates by year across Canadian provinces/territories (2014-2022). Boxplots show the median, interquartile range, and range of age-standardized rates per 100,000 by year; points indicate outliers.

Table 2. Multiple linear regression predicting age-standardized heart-attack rates from wildfire exposure index, province/territory, and year (Canada, 2014-2022). Coefficients represent the change in the age-standardized heart-attack rate (per 100 000 population) associated with a one-unit increase in each predictor, controlling for all other variables. The reference category for Province is Alberta.

Predictor	Estimate (B)	SE	t	p
Intercept	10874.42	2062.17	5.27	< .001
Year	−5.25	1.02	−5.14	< .001
Exposure Index	−0.00174	0.00181	−0.96	.338
Province/Territory (Ref: Alberta)				
British Columbia	−21.38	13.11	−1.63	.106
Manitoba	37.93	12.73	2.98	.004
New Brunswick	99.70	14.34	6.95	< .001
Newfoundland & Labrador	107.81	14.14	7.62	< .001
Northwest Territories	60.26	12.87	4.68	< .001
Nova Scotia	81.06	14.01	5.79	< .001
Ontario	15.92	12.93	1.23	.221
Prince Edward Island	68.54	14.30	4.79	< .001
Quebec	172.56	13.42	12.86	< .001
Saskatchewan	6.81	12.86	0.53	.598
Yukon	42.39	13.53	3.13	.002

The exposure index was insignificant, showing no association between yearly wildfire incidence, and rate of hospitalized heart-attacks.

To show provincial fixed effects, Alberta was used as the reference. British Columbia, Ontario, and Saskatchewan had no significant difference in their adjusted rate of heart-attack hospitalizations, while every other province had significantly higher rates ($p < 0.05$). These fixed effects indicate that there remains substantial between-province/territory variation after adjusting for year, and exposure.

Model assumption checks indicated no major multicollinearity concerns. VIF values were low for year (VIF = 1.0058), exposure index (VIF = 1.4379), and province/territory (VIF = 1.0337).

Figure 3 represents the adjusted relationship between wildfire exposure index, and age-standardized heart-attack hospitalization rates. After controlling for year, and province fixed effects, the association was weak. Lower exposure seemed to have a slightly negative association, while confidence intervals widened substantially with increasing exposure, indicating considerable uncertainty in the adjusted association at higher exposure values.

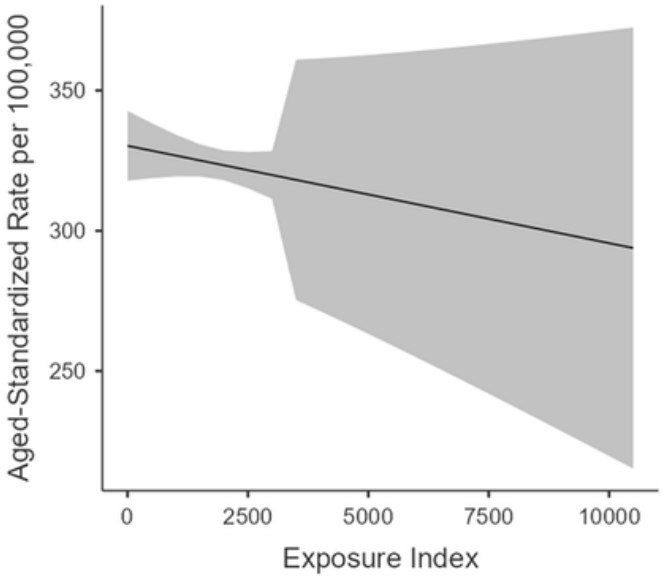


Figure 3. Adjusted association between wildfire exposure index, and age-standardized heart-attack hospitalization rates (Canada, 2014-2022). Predictions adjusted for year (continuous), and province (fixed effects). The shaded region represents the model-based 95% confidence interval.

Discussion

This ecological panel study examined whether annual variation in wildfire intensity was associated with age-standardized heart-attack hospitalization rates across Canadian provinces and territories from 2014–2022. In a multiple linear regression model with province fixed effects, and adjustment for year, there was no significant association between wildfire exposure index, and heart-attack hospitalization (B = -0.0017, p =.338). The year showed significant negative association with heart-attack hospitalization rates (B = -5.25 per 100,000 per year, p < 0.001). This model explained a large proportion of the variance in heart-attack hospitalization rates (adjusted R^2 = 0.82), suggesting most of the difference was accounted for by province-level differences, and a stable decline over time, rather than wildfire activity. The large province/territory fixed effects likely reflect persistent regional differences in cardiovascular risk, healthcare access, coding or hospitalization practices, socioeconomic conditions, comorbidity burden, and demographic structure not fully captured by age standardization. For example, Quebec, Newfoundland and Labrador, New Brunswick, and Nova Scotia had substantially higher adjusted rates than Alberta, suggesting that stable regional factors explained more variation in annual heart-attack hospitalization rates than wildfire activity. Consistent with national cardiovascular trends, we observed a significant decrease in heart-attack hospitalization rates from 2014–2022, even after accounting for wildfires and province differences. This reflects the consistent year-to-year improvements in cardiovascular prevention, and treatment (17). These findings suggest that if heart-attack risk is affected by wildfire exposure, it is small relative to the other determinants or not well captured by the aggregate exposure metric.

Figure 3 shows a weak negative adjusted slope between wildfire exposure index, and age-standardized heart-attack hospitalization rates; however, this pattern should not be interpreted as evidence of a protective effect. The confidence intervals widen substantially at higher exposure values, indicating sparse data, and greater uncertainty at the upper end of the wildfire exposure distribution. Therefore, the model is least precise for the highest wildfire-intensity province-years, limiting interpretation of the association in extreme fire years.

Our null association for heart-attack hospitalization is broadly consistent with systematic reviews showing that wildfire exposure has clear respiratory effects but less evidence that it affects cardiovascular morbidity, and myocardial infarction rates. The Annual Review of Medicine reported there is a clear effect on respiratory hospitalization but less consistent evidence on cardiovascular morbidity (18). Earlier short-term time series studies during the major wildfire episodes in Southern California, and Sydney (9, 10) reported strong associations between respiratory admissions, but null or weak evidence supporting associations of cardiovascular admissions, slikethe absence of a clear myocardial infarction signal in our study.

Newer research has identified a short-term increase in acute cardiac arrests on high-smoke days, suggesting that wildfire smoke may trigger an effect in susceptible individuals. In Victoria, Australia, (Haikerwal et. al. (11)) found that short-term increases in $PM_{2.5}$ exposure was associated with an increase in cardiac arrest, and ischemic heart disease, supporting the role of $PM_{2.5}$ as a trigger for acute cardiac events. Conversely, large population-based studies focusing on wildfire-$PM_{2.5}$ exposure show mixed results on cardiovascular effects. For example, (Liu et. al. (12)) observed no major difference in cardiovascular admissions between normal days, and “smoke wave” days, while (Hao et. al. (13)) found a modest but significant increase in the incidence of heart failure with increased average exposure of $PM_{2.5}$ over a 2-year period in Americans 65 years or older. Taken together, these studies suggest that wildfire exposure can affect cardiovascular health, but the magnitude and detectability depend on the outcome type, exposure resolution, and study design. Our findings complement this literature by demonstrating that at the province-year level in Canada, variation in heart-attack hospitalization rates is not predicted by changes in wildfire exposure, even though individual risks may be elevated during high-smoke periods.

This study has several strengths. First, its national, multi-year dataset covering all Canadian provinces, and territories allows us to examine wildfire-cardiovascular relationships across diverse geographic, climatic, and demographic contexts. Secondly, our study used a panel design with province fixed effect and control for time-invariant provincial characteristics, including healthcare, socioeconomic conditions, and baseline disease burden, that would otherwise confound the relationship between wildfire exposure and heart-attack rates. Lastly, the wildfire exposure index integrated various fire class sizes, capturing the overall severity of each province-year in one statistic, reducing multicollinearity between correlated fire metrics.

However, several limitations should be considered when interpreting these results. Firstly, the ecological panel design measures the exposure, and outcome at the province-year level, so the results cannot be interpreted as individual-risk estimates and can be susceptible to the ecological fallacy.

Secondly, a major limitation of this study is potential nondifferential exposure misclassification. The exposure index was based on fire size, and frequency, not individual or population-level $PM_{2.5}$ exposure. Larger fires in remote regions may have contributed heavily to the exposure index while potentially having a significantly lower relative exposure to humans than a smaller fire in a densely populated area. Because of this misclassification, the effect of the exposure index on heart-attack hospitalization rate may have been biased towards the null. Therefore, the absence of statistical significance should not be interpreted as evidence that wildfire smoke has no association with myocardial infraction (MI).

Thirdly, comparing our results to the finding of (Haikerwal et. al. (11)), we find that the annual time scale was likely to mask the

short-term effects of smoke exposure. (Haikerwal et. al. (11)) identified an association between $PM_{2.5}$ exposure, and acute cardiovascular effects, precisely because they measured exposure over short exposure windows. The temporal averaging of our study likely reduced the sensitivity to acute cardiovascular effects. For example, a province with a few dangerous smoke days would have been unlikely to produce any statistical significance as those days would be diluted in the data from that whole year, biasing results towards the null. It should be noted that there was a positive association, identified by (Hao et. al. (13)), between long-term smoke exposure, and the incidence of heart failure, but not MI.

We also lacked data on individual risk factors (e.g., smoking, hypertension, diabetes, socioeconomic status) and co-pollutants (e.g., non-wildfire $PM_{2.5}$, NO_2, ozone), which could confound any wildfire-MI relationship. Finally, our outcome was limited to hospitalized heart-attacks, so we could not capture out-of-hospital cardiac arrests, fatal MIs occurring before hospitalization, nonhospitalized ischemic events, unstable angina, or arrhythmias. Because wildfire smoke may precipitate sudden cardiac events or rhythm disturbances, these excluded outcomes may be among the cardiovascular events most sensitive to acute $PM_{2.5}$ exposure.

Taken together, these findings suggest that wildfire exposure at the province-year level within Canada does not have a clear effect on heart-attack hospitalization rates compared with secular declines over time, and provincial differences in cardiovascular burden. From a public-health perspective, while wildfire exposure remains an important factor in individual health, its contribution to overall annual heart-attack rates is modest at the scale of provincial health-system planning. Future research that links individual-level MI with high-resolution wildfire-specific $PM_{2.5}$ exposure, focusing on susceptible subgroups (e.g., 65+, preexisting cardiovascular disease) will be important to identify individual-level associations masked in aggregate provincial data.

Conclusion

This ecological observational time-series study examined the effects of wildfire intensity on age-standardized heart-attack hospitalization rates across Canadian provinces and territories from 2014–2022. There was no significant association between wildfire intensity and heart-attack hospitalization rates after controlling for year-to-year declines and province-fixed effects. Wildfire activity varied substantially across province-years, however most of the variation in heart-attack rates was explained by secular declines and provincial differences. At this scale, wildfires do not appear to be a main driver of myocardial infarction rates, even in the context of worsening fire seasons. Policy, and prevention should focus on traditional cardiovascular risks, and social determinants to reduce MI, while continuing to treat wildfires as an important environmental health issue, especially for respiratory disease.

References

1. L. Wang, L. Wang, Y. Li, J. Wang, A century-long analysis of global warming and earth temperature using a random walk with drift approach. Decision Anal. J. 7, 100237 (2023).
2. S. S. Sayedi, et al., Assessing changes in global fire regimes. Fire Ecol. 20, 18 (2024).
3. Government of Canada, "Canada's record-breaking wildfires in 2023: A fiery wake-up call" (Natural Resources Canada, 2023).
4. C. Barnes, et al., "Climate change more than doubled the likelihood of extreme fire weather conditions in Eastern Canada" (Natural Resources Canada, 2023).
5. J. T. Abatzoglou, et al., Climate change has increased the odds of extreme regional forest fire years globally. Nat. Commun. 16, 6390 (2025).
6. Z. Jin, et al., Fire smoke elevated the carbonaceous PM2.5 concentration and mortality burden in the contiguous U.S. and southern Canada. Res. Sq. DOI (2024).
7. Y. Pei, et al., Effects of fine particulate matter (PM2.5) on systemic oxidative stress and cardiac function in ApoE−/− mice. Int. J. Environ. Res. Public Health 13, 484 (2016).
8. C. Krittanawong, et al., PM2.5 and cardiovascular diseases: State-of-the-art review. Int. J. Cardiol. Cardiovasc. Risk Prev.19, 200217 (2023).
9. R. J. Delfino, et al., The relationship of respiratory and cardiovascular hospital admissions to the southern California wildfires of 2003. Occup. Environ. Med. 66, 189–197 (2009).
10. F. Johnston, et al., Extreme air pollution events from bushfires and dust storms and their association with mortality in Sydney, Australia 1994–2007. Environ. Res. 111, 811–816 (2011).
11. A. Haikerwal, et al., Impact of fine particulate matter (PM2.5) exposure during wildfires on cardiovascular health outcomes. J. Am. Heart Assoc. 4, e001653 (2015).
12. J. C. Liu, et al., Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. Epidemiology 28, 77–85 (2017).
13. H. Hao, et al., Long-term wildfire smoke exposure and increased risk of heart failure in older adults. J. Am. Coll. Cardiol. 85, 2439–2451 (2025).
14. Canadian Institute for Health Information, "Hospitalized Heart Attacks" (CIHI, 2025).
15. Canadian Council of Forest Ministers, "National Forestry Database: Wildland Fire Statistics" (CCFM, 2024).
16. The jamovi project, jamovi version 2.3 (2022); https://www.jamovi.org.
17. L. C. P. Botly, et al., Recent trends in hospitalizations for cardiovascular disease, stroke, and vascular cognitive impairment in Canada. Can. J. Cardiol. 36, 1081–1090 (2020).
18. C. F. Gould, et al., Health effects of wildfire smoke exposure. Annu. Rev. Med.75, 277–292 (2024).

Original Research

Fragrant Fakery: Sniffing Out the Truth in "Pure" Green Coffee Oil

Faux parfums: Flairer la vérité dans l'huile de café vert « pure »

Maximiliano Araneda Suárez[1], Sharon Barden[1], Paul M Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

The natural extracts industry is plagued by imitation products that pose health risks, despite oversight by agencies such as the FDA and Health Canada. Coffee oils are a popular extract and are particularly susceptible to adulteration. The study compares the store-bought oil to a pure cold-pressed extract of green coffee beans using gas chromatography–mass spectrometry to find signs of adulteration. The store-bought oil exhibited markers of adulteration such as cyclamen aldehyde (cyclamal) and isopropyl myristate, while lacking several natural compounds such as palmitic acid and vitamin E found in cold-pressed green coffee oil (GCO). In comparison, through examination, the results show that an unregulated product, a "pre-workout" powder, is reported to contain methylhexanamine (DMAA) and its analogue DMHA, both of which were indeed identified, supporting the authenticity of the supplement. These results highlight the need for enforcement of regulations for consumer safety.

———

L'industrie des extraits naturels est rongée par des produits d'imitation qui présentent des risques pour la santé, malgré la surveillance d'agences telles que l' Agence américaine des produits alimentaires et médicamenteux (FDA) et Santé Canada. Les huiles de café sont un extrait populaire et sont particulièrement susceptibles d'être adultérées. L'étude compare l'huile achetée en magasin à un extrait pur pressé à froid de grains de café vert en utilisant la chromatographie en phase gazeuse couplée à la spectrométrie de masse pour détecter des signes d'adultération. L'huile achetée en magasin présentait des marqueurs d'adultération tels que l'aldéhyde de cyclamen (cyclamal) et le myristate d'isopropyle, tout en manquant de plusieurs composés naturels tels que l'acide palmitique et la vitamine E présents dans l'huile de café vert pressée à froid. En comparaison, à l'examen, les résultats montrent qu'un produit non réglementé, une poudre « pré-entraînement », contient de la méthylhexanamine (DMAA) et son analogue DMHA, tous deux identifiés, ce qui confirme l'authenticité du complément. Ces résultats soulignent la nécessité de faire respecter les réglementations pour la sécurité des consommateurs.

Keywords: green coffee oil; adulteration; GC-MS; natural product authentication; DMAA; DMHA; cosmetic chemistry; phytochemical analysis; consumer safety; synthetic additives

Introduction

Natural extracts are widely used commodities for scenting, skin hydration and flavouring. Valued at $18.6 billion USD in 2020, it stands as a growing global industry (1). However, studies analyzing various natural oils have found that many commercially available "pure" products were adulterated with cheaper substitutes by exploiting legal loopholes, despite oversight enforced by the FDA and Health Canada (2,3). These adulterants pose documented potential health risks, including dermatological reactions such as rashes and psoriasis (4,5).

Green coffee oil (GCO) is challenging to extract as the raw beans have a tough, dense exterior and low oil yield. These physical characteristics make GCO particularly prone to adulteration as the process is labour-intensive and expensive (6, 7). GCO is used for skincare and fragrance because of its distinct aroma and natural caffeine content. This study investigates the authenticity of commercially available GCO by comparing its chemical profile to that of pure, cold-pressed GCO using gas chromatography-mass spectrometry (GC-MS). As a parallel case, an unregulated athletic "pre-workout" supplement marketed as containing methylhexanamine (DMAA) is examined via acid-base extraction and GC-MS to determine its authenticity. DMAA is a prohibited stimulant for athletes in Canada and by World Anti-Doping Agency, but it is available in supplements (8). The purpose was to compare the ingredients of a "regulated" natural product (GCO) and an unregulated one (workout powder). Together, these analyses highlight the extent of adulteration in the natural products industry and the potential risks posed to consumers associated with long-term exposure to synthetic additives.

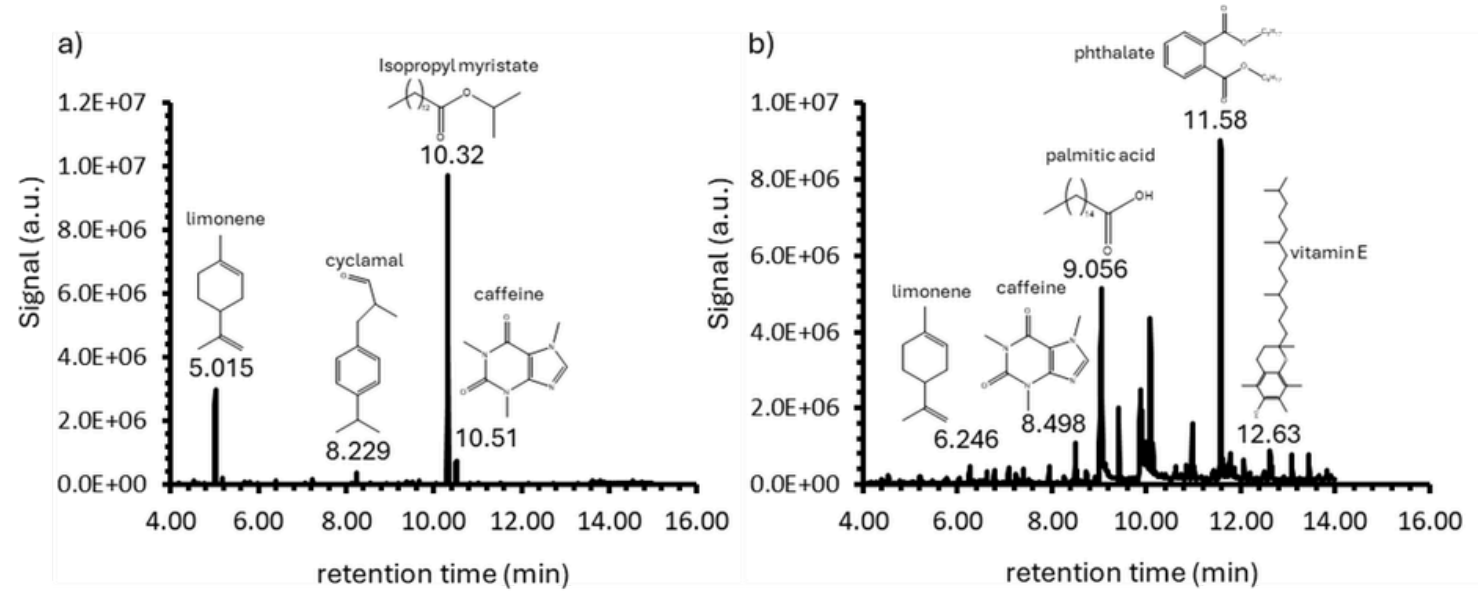


Figure 1. a) GC-MS results of store-bought GCO diluted in ethyl acetate. Peaks are labelled with respective compounds and their retention times. b) GC-MS results of cold-pressed GCO.

Experimental Procedures

Cold-Press

40 grams of green coffee beans were placed into the top of the cold-press. The entire nozzle was heated with a heat gun on the medium/low setting. The press was operated for 10 minutes or until oil is visible at the nozzle tip. Oil was extracted from the tip of the cold-press with a 10 µL syringe and transferred to a vial.

Oil Analysis

1 µL of store-bought or cold-pressed GCO was diluted to 100µL with ethyl acetate. 1 µL of this solution was injected into the GC-MS using a DB-5 capillary column (split ratio 10:1, split flow: 12.8 mL/min, inlet temperature: 200°C, oven program: 80°C (2 min), ramp to 300-320°C at 15-20°C/min, carrier gas: Helium, detector: MS (electron ionization, 70 eV)). A solvent delay of four minutes was employed throughout. Compounds were identified by their retention index and mass spectrum.

Workout Powder Analysis

10 g of the pre-workout powder was dissolved in 250 mL of water and stirred for five minutes. The solution was acidified with 1 mol/L HCl until the pH was less than two. Liquid-liquid extraction was performed with hexane, retaining the aqueous layer. The aqueous layer was basified with six mol/L NaOH until the pH was greater than 11. Liquid-liquid extraction on this solution was performed with diethyl ether, retaining the top organic layer. One µL of diethyl ether solution was diluted to 100µL in ethyl acetate. One µL of this sample was injected into the GC-MS.

Results and Discussion

The chromatograms of commercial and cold-pressed GCO are shown in Figure 1. A visual comparison between Figures 1a and 1b reveals substantial differences in chemical composition. For example, the cold-pressed GCO contains peaks of several documented compounds in GCO, such as palmitic acid and vitamin E, both of which are very common in nature and plant oils in general (9-12). Limonene and caffeine were the only overlapping compounds, though only caffeine appeared at comparable abundance. The high abundance of phthalates comes from the packaging material used to store the green coffee beans, and can be a commonly found impurity in such products (13). In Figure 1a., the significant abundance of isopropyl myristate and cyclamal demonstrates that the store-bought oil was heavily adulterated, as these compounds are common emulsifiers and fixatives, respectively. Neither compound is naturally found in GCO.[9,10] Cyclamal has been documented to cause dermatological reactions and is known to contribute to aquatic degradation (14). Isopropyl myristate causes rashes in prolonged usage and has various adverse symptoms when ingested (15, 16). The store-bought GCO also lacks palmitic acid and vitamin E found in cold-pressed GCO. Figure 2 confirms that DMAA and DMHA were both present in the pre-workout supplement, as reported on the ingredient list, supporting the authenticity of the product, despite the lack of regulation in the industry. This presents a stark contrast between the regulated natural extracts industry, hampered by imitation products and synthetic substitutions, and the unregulated pre-workout supplement industry. The adulterants found in the store-

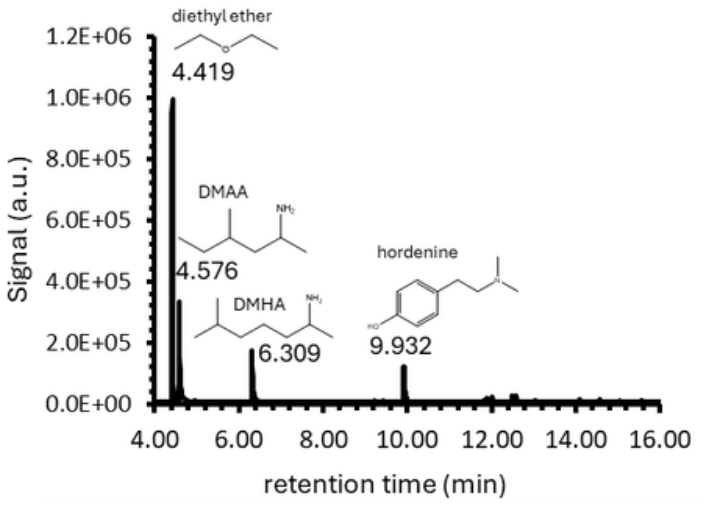


Figure 2. GC-MS results of extracted DMAA. Peaks are labelled with respective compounds and their retention times.

bought GCO, though relatively low in toxicity to humans, highlight the broader risk of substituting natural compounds with unregulated synthetic ones. Many deaths have been caused by this, as seen in the 2002 Slim10 Diet Pill incident and the 2009 Hydroxycut recall (17-19). As the industry grows rapidly, these findings demonstrate the need for tighter regulatory oversight and research into potential contraindications between adulterants and commonly used pharmaceuticals.

Conclusion

In this study, results show that a store-bought GCO was heavily adulterated compared to a cold-pressed GCO, despite current regulations and its marketing as a "pure" natural extract. The adulterants found may cause several different types of irritation. The pre-workout powder examined contains DMAA and its analogue, DMHA, as advertised, even without regulatory bodies overseeing the synthesis process. These results show that great strides in enforcement are required by regulatory bodies. The possibility of companies using toxic substituents, as has been seen in the past, alongside the growth of the industry, further emphasizes the need for regulation. Further research should be done into common adulterants and their contraindications with widespread medications to fully evaluate their safety.

Acknowledgements

The authors thank the JLH Mass Spectrometry Core Facility of the University of Ottawa for providing instrument access and consumables related to this project.

References

1. D.C. Baudoin, E. Bush, T. Gauthier, A.B. Hernandez, H. Kirk-Ballard, Effects of Irrigation and Drought on Growth and Essential Oil Production in O. Vulgare and R. Officinalis. Am. J. Plant Sci. 13, 659–667 (2022).
2. About Natural Health Product Regulation in Canada. (Health Canada 2025); https://www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/regulation.html.
3. Dietary Supplements. (U.S. Food and Drug Administration, 2025) https://www.fda.gov/food/dietary-supplements.
4. S.E. Anderson, B.J. Meade, Potential Health Effects Associated with Dermal Exposure to Occupational Chemicals. Environ. Health Insights 8, EHI.S15258 (2014).
5. S. Nicolai, M. Wegrecki, T.-Y. Cheng, E.A. Bourgeois, R.N. Cotton, J.A. Mayfield, G.C. Monnot, J.L. Nours, I.V. Rhijn, J. Rossjohn, D.B. Moody, A. de Jong, Human T Cell Response to CD1a and Contact Dermatitis Allergens in Botanical Extracts and Commercial Skin Care Products. Science Immunology 5, eaax5430 (2020).
6. L.D. Dias, F.M. Carbinatto, I. Almeida, K.C. Blanco, F. Marquele-Oliveira, C.C. Munari, V.S. Bagnato, Eco-Friendly Extraction of Green Coffee Oil for Industrial Applications: Its Antioxidant, Cytotoxic, Clonogenic, and Wound Healing Properties. Fermentation 9, 370 (2023).
7. F.J. Almeida; F. Gosuen, A. Paula, J. Maria, Scaling up the Two-Stage Countercurrent Extraction of Oil and Protein from Green Coffee Beans: Impact of Proteolysis on Extractability, Protein Functionality, and Oil Recovery. Food Bioproc. Tech. 15, 1794–1809 (2022).
8. P.A. Cohen, A. Wen, R. Gerona, Prohibited Stimulants in Dietary Supplements After Enforcement Action by the US Food and Drug Administration. JAMA Intern Med. 178, 1721–1723 (2018).
9. L.F. Echeverri-Giraldo, M.I. Pinzón Fandiño, L.M. González Cadavid, N.D. Rodriguez Marín, D.A. Moreno Ríos, V. Osorio Pérez, Determination of Lipids and Fatty Acids in Green Coffee Beans (Coffea Arabica L.) Harvested in Different Agroclimatic Zones of the Department of Quindío, Colombia. Agronomy 13, 2560 (2023).
10. A. Bobková, K. Poláková, A. Demianová, L' Belej, M. Bobko, L. Jurčaga, B. Gálik, I. Novotná, A. Iriondo-DeHond, M.D. Castillo, Comparative Analysis of Selected Chemical Parameters of Coffea Arabica, from Cascara to Silverskin. Foods 11, 1082 (2022).
11. A.V. Zhukov, Palmitic Acid and Its Role in the Structure and Functions of Plant Cell Membranes. Russ. J. Plant Phys. 62, 706–713 (2015).
12. Y. Niu, Q. Zhang, J. Wang, Y. Li, X. Wang, Y. Bao, Vitamin E Synthesis and Response in Plants. Front. Plant Sci. 13, 994058 (2022).
13. J. Wang, X. Weng, S. Liu, H. Zhang, Q. Zhu, Q. Fu, T. Wang, C. Liao, G. Jiang, Occurrence, fate, and reduction measures of phthalates in the cooking process: A review. Environment & Health 1, 300-314 (2023).

14. European Chemical Agency, 3-p-cumenyl-2-methylpropionaldehyde (ECHACHEM; 2025) https://chem.echa.europa.eu/100.002.874/overview?searchText=3-p-cumenyl-2-methylpropionaldehyde.
15. European Chemical Agency, Isopropyl myristate. (ECHACHEM; 2025) https://echa.europa.eu/registration-dossier/-/registered-dossier/16077/7/5/1.
16. M.A. Liebert, Final Report on the Safety Assessment of Myristyl Myristate and Isopropyl Myristate. J. Am. Coll. Toxic. 1, 55-80 (1982).
17. Chinese diet pills linked to deaths, illness in Asian women. (CBC News; 2002) https://www.cbc.ca/news/canada/chinese-diet-pills-linked-to-deaths-illness-in-asian-women-1.303776.
18. Hydroxycut 24 Hour Caplets (May 01, 2009) - Recalls and Safety Alerts. (Health Canada; 2009), https://healthycanadians.gc.ca/recall-alert-rappel-avis/hc-sc/2009/9948r-eng.php.
19. N. Singer, Hydroxycut Supplements Recalled after Warning. (The New York Times; 2009), https://www.nytimes.com/2009/05/02/business/02fda.html.

Global Patterns of Skilled Birth Attendance, Socioeconomic Factors, and Maternal Mortality

Modèles mondiaux de l'assistance qualifiée à l'accouchement, des facteurs socioéconomiques et de la mortalité maternelle

Ayman Assaaoudi[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: aassa087@uottawa.ca

Abstract | Résumé

Skilled birth attendance (SBA) is an important proxy for quality of maternal care, and is negatively correlated with maternal death. The present cross-national ecological study aimed to examine the association between SBA and maternal death considering socioeconomic predictors of MMR across a broad range of countries, and World Bank World Development Indicators were extracted for 180 countries between 2010 and 2023. Hierarchical linear regressions were performed to examine the association between SBA and MMR adjusting for GDP per capita, life expectancy, female literacy, sanitation, adolescent fertility and health expenditures.

SBA ranges from 58.5% (low-income) to 98.8% (high-income) while the MMRs are 538.3 (low-income) to 19.3 (high-income). The correlations of SBA with MMR show a high degree of negative association ($r = -0.775$, $p < 0.001$). Once fully adjusted for wealth, social and demographic factors, the association between SBA and log(MMR) turned out to be non-significant ($\beta = -0.004$, $p = 0.539$). GDP per capita, life expectancy ($\beta = -0.049$, $p = 0.010$) and adolescent fertility ($\beta = +0.006$, $p = 0.032$) have remained significant independent predictors of log(MMR). The full adjusted model explained 81.4% of the variance in log(MMR).

While SBA was not an independent predictor of maternal death at the ecological level controlling for level of development in the country, its beneficial effects were included within country level social determinants. This must be interpreted with caution due to ecological fallacy and a cross-section study design. At the individual level, there are direct clinical benefits for mothers receiving skilled birth attendance. Sustained reduction in maternal mortality requires the combination of development inputs related to education, sanitation, reproductive health and access to health care.

———

L'assistance qualifiée à l'accouchement est un indicateur important de la qualité des soins maternels et est généralement associée à une diminution de la mortalité maternelle. La présente étude écologique menée à l'échelle internationale avait pour objectif d'examiner l'association entre l'assistance qualifiée à l'accouchement et la mortalité maternelle, tout en tenant compte de plusieurs facteurs socioéconomiques pouvant influencer le ratio de mortalité maternelle. Pour cela, des indicateurs de développement de la Banque mondiale ont été recueillis pour 180 pays entre 2010 et 2023. Des régressions linéaires hiérarchiques ont ensuite été réalisées afin d'analyser l'association entre l'assistance qualifiée à l'accouchement et le ratio de mortalité maternelle, en ajustant les résultats selon le PIB par habitant, l'espérance de vie, le taux d'alphabétisation des femmes, l'accès à l'assainissement, la fécondité chez les adolescentes et les dépenses en santé.

Les résultats montrent que la proportion d'accouchements assistés par du personnel qualifié varie de 58,5 % dans les pays à faible revenu à 98,8 % dans les pays à revenu élevé. De leur côté, les ratios de mortalité maternelle varient de 538,3 décès maternels pour 100 000 naissances vivantes dans les pays à faible revenu à 19,3 dans les pays à revenu élevé. L'assistance qualifiée à l'accouchement présente une forte corrélation négative avec la mortalité maternelle ($r = -0,775$, $p < 0,001$), ce qui signifie que les pays où l'accès à une assistance qualifiée est plus élevé tendent à avoir une mortalité maternelle plus faible. Cependant, après ajustement complet pour la richesse, les facteurs sociaux et les facteurs démographiques, cette association entre l'assistance qualifiée à l'accouchement et le logarithme du ratio de mortalité maternelle n'était plus statistiquement significative ($\beta = -0,004$, $p = 0,539$). Le PIB par habitant, l'espérance de vie ($\beta = -0,049$, $p = 0,010$) et la fécondité chez les adolescentes ($\beta = +0,006$, $p = 0,032$) sont demeurés des prédicteurs indépendants significatifs du logarithme du ratio de mortalité maternelle. Le modèle entièrement ajusté expliquait 81,4 % de la variance du logarithme du ratio de mortalité maternelle.

Ainsi, même si l'assistance qualifiée à l'accouchement n'était pas un prédicteur indépendant de la mortalité maternelle au niveau écologique après ajustement pour le niveau de développement des pays, ses effets bénéfiques semblent être intégrés dans des déterminants sociaux plus larges, comme l'éducation, les conditions sanitaires et l'accès général aux soins. Ces résultats doivent toutefois être interprétés avec prudence, notamment en raison du risque d'erreur écologique et du devis transversal de l'étude. À l'échelle individuelle, l'assistance qualifiée à l'accouchement demeure cliniquement bénéfique pour les mères. Une réduction durable de la mortalité maternelle nécessite donc une combinaison d'interventions liées au développement, incluant l'éducation, l'assainissement, la santé reproductive et l'accès aux services de santé.

Keywords: Maternal mortality; Skilled birth attendance; Healthcare accessibility; Socioeconomic development; Ecological study; World Bank WDI

Introduction

Globally, maternal mortality is the leading cause of death among women of reproductive age. There are estimated to be 287,000 maternal deaths per year from pregnancy and childbirth causes (1), nearly 95% of which occur in low- and middle-income countries, with Sub-Saharan Africa comprising approximately 70% of the global burden (2). The maternal mortality ratio (MMR, maternal death per 100,000 live births) varies by country, from less than 5 to greater than 1000 among the highest burden countries. With the exception of the smallest numbers, the most common causes of maternal death — postpartum hemorrhage, eclampsia, sepsis, and obstructed labor — can be managed effectively by skilled providers with timely obstetric care (3).

Skilled attendance at birth (i.e., a birth attended by doctors, midwives, or other trained healthcare personnel) is one of the formal SDGs indicators (3.1.2) which measures the country-level availability of delivery services (4) and which shows this association with individual-level clinical benefit for maternal outcome (5). Despite the fact that global skilled birth attendance (SBA) coverage has increased, a huge gap remains at the country level to achieve SDG Target 3.1 among the high burden countries.

At the population level, the ecological association between skilled birth attendant and maternal mortality might be affected by groups of factors together. The most skilled attendant countries are those with more social economic developing conditions (e.g., countries with higher gross domestic product (GDP) per capita, higher female literacy, better sanitation conditions, and low adolescent fertility rates) compared to those with less developed conditions (6). High literacy may relate to healthcare-seeking behaviors and good patient-provider relationships (7), while poor sanitation relates to the risk of infection and it is independent of attendance skill (8). Low adolescent fertility rate relates to reproductive health situation and higher obstetric risks (9). These development indices often cluster together rather than change independently; thus, determining the ecological effect of SBA alone becomes challenging.

This ecological study, conducted across 180 countries between the years 2010–2023, had two main purposes: I) to provide an overview of the world distribution of skilled birth attendant and maternal mortality, and to assess the correlation between SBA and MMR and socioeconomic indicators; II) to examine the SBA-MMR association and its changes over time, adjusted for the effects of socioeconomic confounders. Finally, the interaction of female literacy and sanitation with SBA was also investigated. The main goal of the study was to determine the standalone ecological contribution of SBA toward reducing maternal mortality.

Methods and Materials

Study design and data sources
This ecological, cross-national study used publicly available country-level indicators for the period of 2010 to 2023, sourced from the World Bank World Development Indicators (WDI). Country eligibility was based on data being available for SBA and MMR for at least one year. If multiple years of data were available for a country, then a mean value for the period of 2010–2023 was used. Institutional review board approval was not required due to the use of aggregated, publicly available data.

Variables
The dependent variable of interest was the MMR (WDI indicator: SH.STA.MMRT). Data was generated by the World Health Organization/United Nations International Children's Emergency Fund/United Nations Population Fund/World Bank Maternal Mortality Estimation Inter-Agency Group and was modeled using the Bayesian BMat model. Logarithm (log) of MMR was used in the regression analyses as MMR is right-skewed.

The independent variable of interest was the percentage of births attended by skilled health personnel (WDI indicator: SH.STA.BRTC.ZS), derived from DHS and UNICEF MICS surveys. Skilled health personnel included doctor, nurse, or midwife in attendance at birth.

A set of socioeconomic variables were pre-selected a priori including GDP per capita (current 2015 USD), adult female literacy (%), life expectancy at birth (years), access to basic sanitation (%), adolescent fertility rate (per 1,000 women aged 15-19), and health expenditure per capita (USD). GDP per capita and health expenditure per capita were log-transformed and all continuous variables were mean-centered to decrease multicollinearity. Countries were categorized by the World Bank income groupings.

Statistical analysis
Descriptive analyses provided summary statistics for each variable stratified by country income category. The bivariate association of each variable with the dependent variable was evaluated using Pearson correlations. These were represented using color-coded heat maps using R software (version 4.6.0).

Using Hierarchical Ordinary Least Squares regression, the change in the SBA-log(MMR) association was examined after sequential adjustment of the primary exposure. Model I included only the exposure and the primary outcome, the SBA-log(MMR) association. Model II added log(GDP per capita) to the model. Model III added life expectancy at birth and adult female literacy to the model. Model IV (full adjustment) included log(health expenditure per capita), sanitation, and adolescent fertility. Model fit was analyzed using R and adjusted R statistics. Multicollinearity was evaluated using variance inflation factors (VIF); VIF greater than 10 suggests high collinearity. Model coefficient plots were generated showing the beta coefficients with 95% confidence intervals (CI) at each step and plotted using standard functions within R.

An analysis was conducted using interaction terms between the exposure (SBA) and two covariates to examine effect modification between the exposure and I) adult female literacy and II) access to

Table 1. Descriptive analysis of study variables by World Bank income group. Summary statistics for maternal mortality ratio (deaths per 100,000 live births) and skilled birth attendance coverage (percentage of births) stratified by income classification. MMR = maternal mortality ratio. Data: World Bank WDI, country-level means 2010–2023.

Income Group	n	MMR Mean	MMR SD	MMR Min	MMR Max	Skilled % Mean	Skilled % SD
Low income	24	538.3	298.4	70.0	1,440.0	58.5%	20.4%
Lower middle	48	250.9	215.2	21.9	1,106.0	76.5%	17.4%
Upper middle	53	67.7	58.6	2.0	254.0	96.2%	7.1%
High income	53	19.3	24.2	2.2	116.0	98.8%	1.5%
Total	**178**	**189.2**	**248.1**	**2.0**	**1,440.0**	**82.4%**	**23.8%**

basic sanitation. Visualization of geographical trends was conducted using income-stratified scatter plots showing linear regression trendlines. All analyses were performed in R using WDI data accessed through the WDI R package. All tests were two-tailed with =0.05.

Results

Descriptive characteristics

180 countries were included in the analyses of SBA and MMR for the period of 2010–2023. The two countries that had a "Not classified" status were removed from income strata analyses, resulting in 178 countries. Significant variation was found across countries for both MMR (ranging from 2.0-1440.0 deaths/100,000 live births) and SBA (<60% to near universal) and distinct gradients were evident for each indicator according to country income category (Table 1). High income countries had a mean SBA of 98.8% (standard deviation, SD: 1.5%) while low-income countries had a mean of 58.5% (SD: 20.4%). The high-income group had a mean MMR of 19.3 (SD: 24.2), whereas the low-income group had a mean of 538.3 (SD: 298.4).

Geographic distribution

Figure 1 graphs SBA coverage and MMR by income group. There appears to be gradients for SBA coverage (from low coverage in low-income countries to high coverage in high-income countries) and for MMR (from high MMR in low-income countries to low MMR in high-income countries). However, particularly in low-income countries, the large standard deviations demonstrate a significant heterogeneity among the income categories.

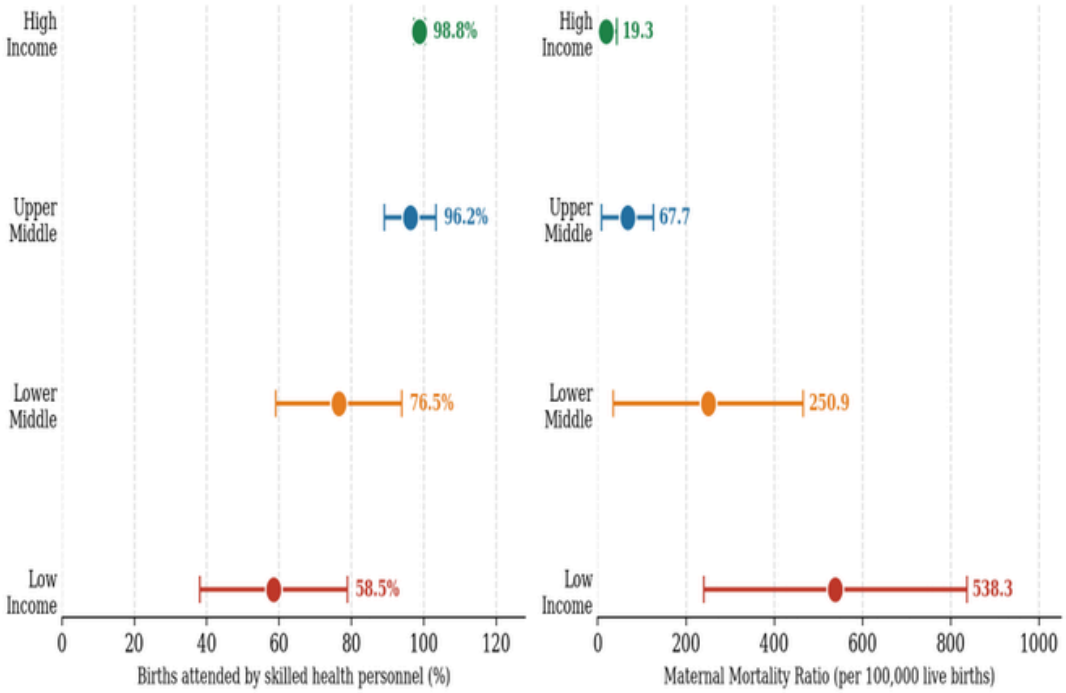


Figure 1. Skilled birth attendance and maternal mortality ratio by World Bank income group. Dot plots with error bars showing mean skilled birth attendance as percentage of births (left panel) and mean maternal mortality ratio per 100,000 live births (right panel) across income classifications. Error bars represent standard deviations. Country-level means are from 2010–2022; n = 178. Data: World Bank WDI.

Correlation structure

From the bivariate correlations, large relationships between variables were observed (Figure 2). A strong correlation was found between SBA and MMR (r = -0.775; p < 0.001; n = 180). The largest correlations with MMR were with life expectancy (r = -0.852). Both access to sanitation and female literacy had strong correlations with MMR (r = -0.807 and -0.793, respectively). Adolescent fertility had a strong positive correlation with MMR (r = +0.721). The correlation between SBA and log(GDP per capita) was also large (r = 0.73), showing high degrees of clustering of development indicators.

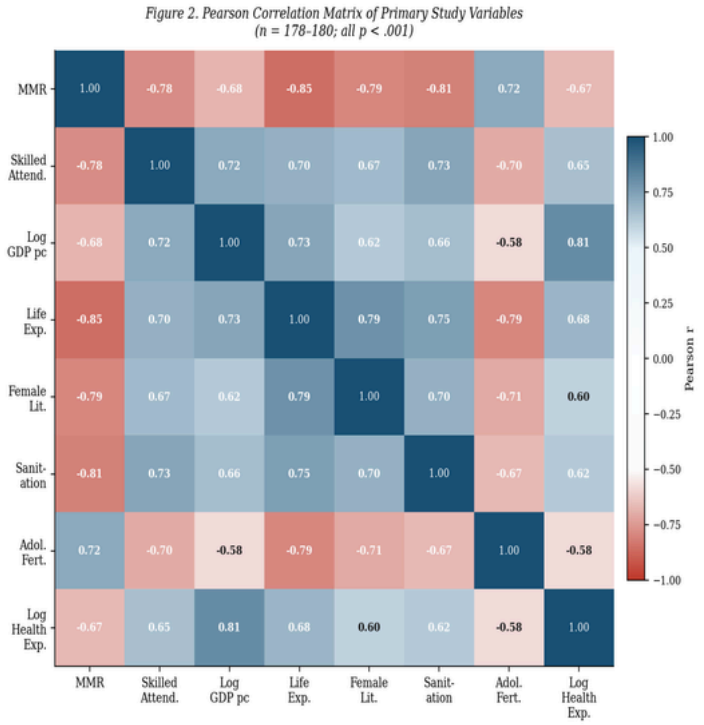


Figure 2. Pearson Correlation Matrix of primary study variables. Heat map displaying correlation coefficients between all study variables (n = 178–180). Color intensity indicates correlation strength: darker shades represent stronger correlations (positive in teal/green, negative in pink/red). All correlations significant at p < 0.001. Data: World Bank WDI 2010–2022, analyzed in R v4.6.0.

Figure 3 displays a scatter plot of SBA coverage against MMR by income group. There appears to be a strong negative relationship. There is a lot of overlap between income groups in the middle ranges of SBA coverage. On the other hand, high-income countries are situated mainly in the top-right (high SBA coverage, low MMR), while low-income countries are more dispersed.

Hierarchical regression models

Hierarchical regression analyses were used to test changes in the SBA-log(MMR) relationship after successive adjustment (Table 2). In Model I, a significant negative relationship (of -0.061, 95% CI: -0.069, -0.052, p < 0.001, R = 0.530) was present between SBA and log(MMR). Following inclusion of log(GDP per capita) (Model II), the SBA coefficient became negative but weaker (-0.027, 95% CI:

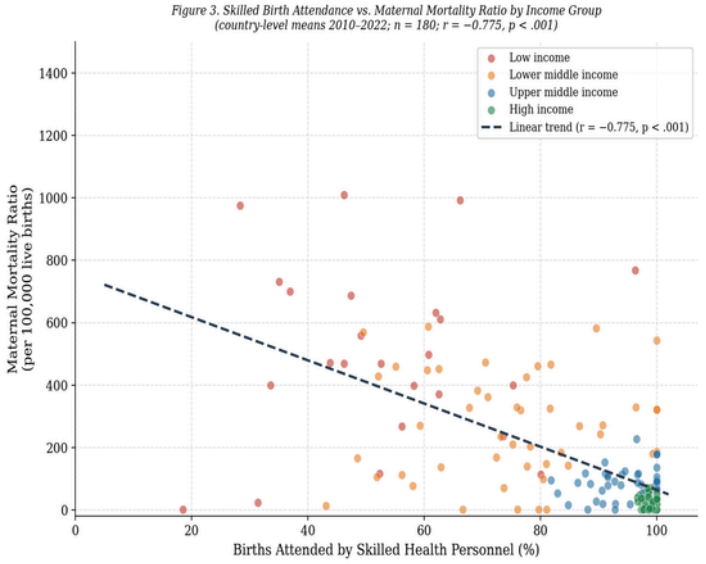


Figure 3. Scatter plot of skilled birth attendance and maternal mortality ratio by income group. Each point represents one country, color-coded by World Bank income classification (high-income = blue, upper-middle-income = green, lower-middle-income = yellow, low-income = red). Dashed line shows overall linear trend (r = -0.775, p < .001). Country-level means from 2010–2022 are used; n = 180. Data: World Bank WDI.

-0.036, -0.017, p < .001, R = 0.717). Adding life expectancy and female literacy to the regression (Model III) further attenuated the SBA relationship (-0.013, p = 0.052, R = 0.792). Finally, the addition of sanitation, adolescent fertility rate, and health expenditure to the full regression model (Model IV) made the SBA term insignificant (-0.004, 95% CI: -0.018, +0.010, p = 0.539, R = 0.814).

In the full regression model (Model IV), life expectancy (-0.049, p = 0.010) and adolescent fertility rate (+0.006, p = 0.032) had significant relationships with log(MMR). High levels of collinearity for log(GDP) (VIF = 14.26) and log(health expenditure) (VIF = 14.27) probably explain their non-significant relationships with log(MMR) despite their significant bivariate relationships, as indicated by the VIF values.

The stepwise decrease in the SBA coefficient and the increase in explained variance are shown in Figure 4 (left and right panels, respectively). The SBA coefficient decreased from -0.061 in Model I to -0.004 in Model IV and, in the fully adjusted model, crossing zero and gaining a negligible confidence interval. The R^2 value grew from 0.530 to 0.814, showing that socioeconomic characteristics account for most of the variation in cross-national MMR.

Effect modification

Significant effect modification was also noted. Interaction of SBA female literacy was significant (-0.0006, 95% CI: -0.0009 to -0.0002, p = 0.002) and SBA sanitation interaction was also significant (-0.0006, 95% CI: -0.0009 to -0.0003, p < 0.001). Negative coefficients indicated a stronger ecological association of increased SBA with decreasing log(MMR) in settings with higher female literacy and increased access to sanitation facilities.

Table 2. Hierarchical linear regression models predicting log(MMR). Unstandardized beta coefficients with 95% confidence intervals from four sequential models. MMR = maternal mortality ratio; GDP = gross domestic product; pc = per capita; VIF = variance inflation factor; SBA/skilled = skilled birth attendance. *** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.10; ns not significant. VIF in Model IV: skilled = 5.02, log(GDP) = 14.26, life expectancy = 5.44, literacy = 4.09, sanitation = 7.10, adolescent fertility = 3.08, log(health expenditure) = 14.27. High collinearity flagged for GDP and health expenditure. Sample size varies due to missing data, particularly female literacy. Data: World Bank WDI 2010–2022.

Predictor	Model I β [95% CI]	Model II β [95% CI]	Model III β [95% CI]	Model IV β [95% CI]
Skilled attend.	-0.061 [-0.069, -0.052]***	-0.027 [-0.036, -0.017]***	-0.013 [-0.026, +0.000]†	-0.004 [-0.018, +0.010]ns
Log(GDP pc)	—	-0.720 [-0.854, -0.587]***	-0.309 [-0.520, -0.098]**	-0.023 [-0.403, +0.358]ns
Life expectancy	—	—	-0.080 [-0.113, -0.047]***	-0.049 [-0.086, -0.012]*
Female literacy	—	—	-0.007 [-0.019, +0.004]ns	-0.004 [-0.015, +0.007]ns
Sanitation	—	—	—	-0.011 [-0.023, +0.001]†
Adol. fertility	—	—	—	+0.006 [+0.001, +0.011]*
Log health exp.	—	—	—	-0.245 [-0.590, +0.100]ns
R^2	**0.530**	**0.717**	**0.792**	**0.814**
Adjusted R^2	**0.528**	**0.714**	**0.784**	**0.802**
n	**180**	**178**	**113**	**112**

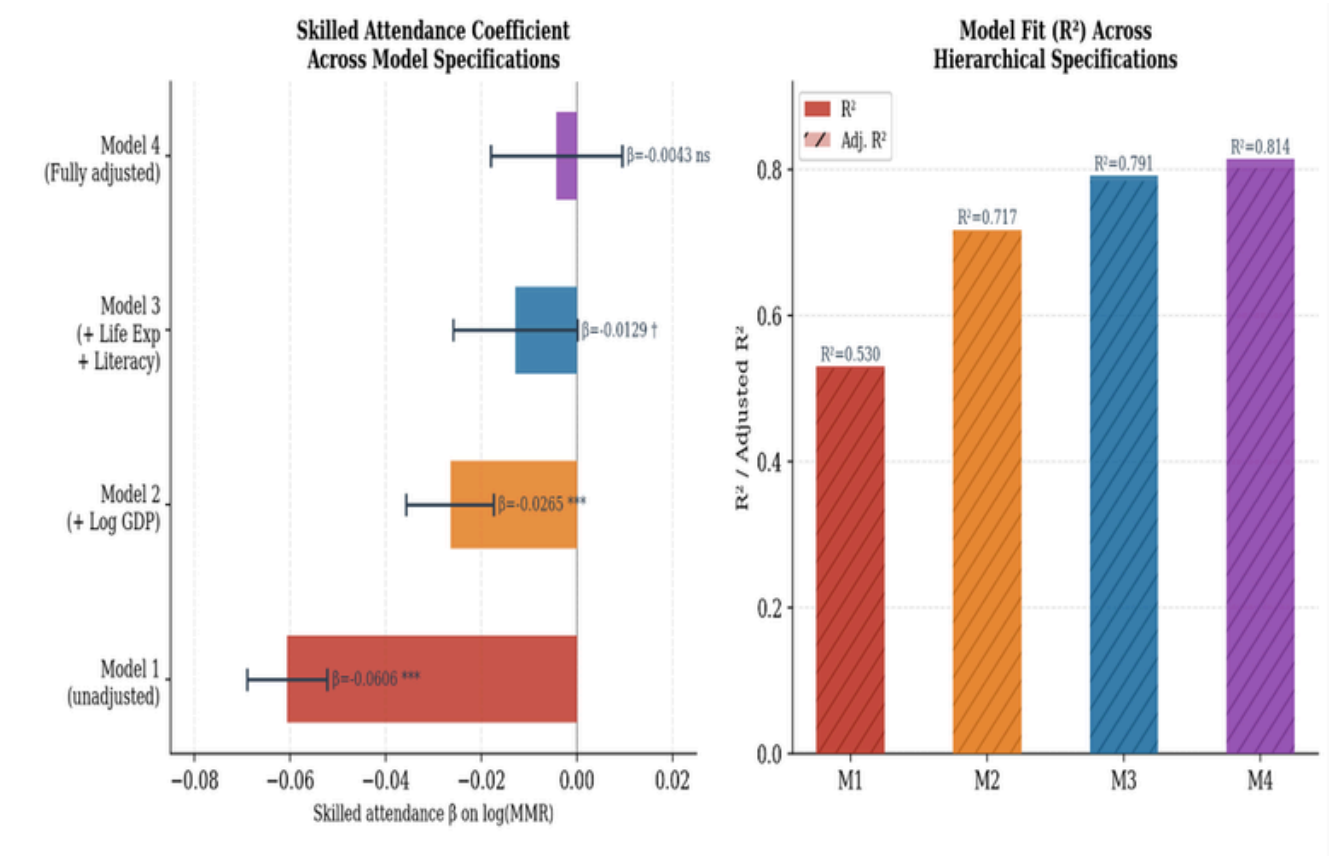


Figure 4. Evolution of skilled birth attendance coefficient and model fit across hierarchical models. Left panel: SBA beta coefficient on log(MMR) with 95% confidence intervals across four model specifications. Right panel: R^2 and adjusted R^2 by model. Sample sizes: n = 180, 178, 113, 112 for Models I-IV. Data: World Bank WDI, analyzed in R v4.6.0.

Discussion

This study conducted an ecological investigation into the associations between SBA and economic development and maternal mortality in 180 countries from 2010 to 2023. At a crudely ecological level, SBA showed a robust negative association with MMR. However, with hierarchical regression modeling, most of this relationship was mediated by broader socioeconomic determinants. After adjusting for all the identified covariates (i.e., GDP, life expectancy, female literacy, sanitation, adolescent fertility, and health expenditure), additional variation in SBA coverage explained no additional variance in maternal mortality.

It is vital to state that the relationships examined in this study are ecological and may not apply at the individual level; this concept is known as ecological fallacy (10). A country with high SBA coverage and low MMR does not necessarily translate to individuals in these settings having a lower mortality risk compared to individuals receiving SBA from a low SBA coverage country. The processes occurring at the individual level may not necessarily be at the population level. Clinical evidence at the individual level overwhelmingly supports that skilled attendance provides essential interventions during childbirth which lower the risk of maternal mortality (Graham et al., 2001; Say et al., 2014). The population-level finding that any additional variability in SBA coverage, beyond the socio-economic correlates, had no additional impact on maternal mortality does not invalidate this clinical evidence.

The stepwise increase in the coefficient reduction shown through the hierarchical regression is notable. In the unadjusted model, the negative association was large (-0.061, R = 0.530). After the addition of GDP per capita, the reduction was more than half (-0.027, R = 0.717), suggesting that income at the national level is a significant confounder. Following addition of life expectancy and female literacy, the association continued to decrease until it was no longer distinguishable from zero in the fully adjusted model. The interpretation of this finding is that in contexts with similar levels of socio-economic development, residual differences in SBA coverage are not related to subsequent differences in maternal mortality. This finding is not suggestive of the ineffectiveness of SBA scaling up, but rather that SBA coverage is intertwined with many of the determinants that are collectively associated with maternal mortality.

This is consistent with many previous ecological studies which emphasize the role of socioeconomic conditions in health status (6). Countries are generally not developing along individual parameters; improvements in health services access and health outcomes are highly correlated with improvements in the educational system, health, and economic infrastructure. The significant inter-correlation of SBA with its associated development determinants (e.g., R = 0.73 between SBA and GPD per capita) reflects this clustering. At the individual level, the benefits of skilled attendance are documented and the context within which it is delivered also contributes to maternal mortality (Karlsen et al., 2011; Prüss-Ustün et al., 2016). At the population level, it appears these components are tightly intertwined and it is difficult to separate their impact on maternal mortality patterns.

Life expectancy, in this study, remained the most robust predictor across all adjusted models, as found in other studies which identified life expectancy as an overall indicator that accounts for changes in nutrition, disease, health system performance, and age structure (2). Adolescent fertility rate remained positively associated with maternal mortality in the fully adjusted model, suggesting that the broader context surrounding reproduction is significant, as well as that pregnant adolescents have an increased risk of adverse obstetric events (9). In terms of interaction effects, higher female literacy and improved access to sanitation were associated with a larger ecological negative relationship between SBA and maternal mortality, suggesting that these enabling characteristics are important for healthcare access interventions.

Significant high correlation between GDP and health expenditures made interpretation difficult as this likely implies that individual parameter estimates may be unstable due to the strong correlation between these variables (VIF > 14 for health expenditure on GDP). While both variables are individually known to have a bivariate relationship with maternal mortality, it is not possible to distinguish between them individually when the model is adjusted for both, but further investigation would be necessary to fully account for this possibility.

Limitations

This analysis has significant limitations due to both its design and the nature of the data structure. Firstly, and perhaps most critically, this is an ecological analysis; all variables are measured at the country level. In ecological designs, inferences cannot be made about individual-level associations—this phenomenon is known as the ecological fallacy (10). If high coverage of SBA was not associated with low MMR after adjusting for the covariates of development, it does not mean that SBA does not work at the individual level. Clinical studies demonstrate clear reductions in mortality with SBA (5), and this ecological study is asking a different question (i.e., “do residual cross-country variations in SBA coverage explain variations in MMR over and above variations accounted for by overall development?”).

Secondly, the cross-sectional nature of the analysis does not allow for inferring cause and effect. It is possible that increasing SBA coverage leads to reductions in MMR; it is also possible that in countries where MMR is falling (whether due to better access to SBA, to other variables, or both) countries invest more in the infrastructure of SBA; or it is possible that underlying societal factors are driving both developments independently. Temporal precedence is inherently impossible to assess with cross-sectional data. Averaging data across the 2010-2023 timeframe mitigated any spuriousness or unusual fluctuations caused by single year data but did not allow any examination of changes within countries over time.

Thirdly, country-specific estimates of MMR vary in completeness and accuracy. It is possible that low-income countries' MMR is underreported due to weak civil registration systems and misclassification, though the MMEIG attempts to address this using Bayesian models. Despite these adjustments, there is likely still some level of residual error which may differ between countries and could bias estimates of the relation between SBA coverage and MMR.

Fourthly, this study's measure of SBA coverage is broad, capturing presence of a skilled attendant but not necessarily the quality of care, existence of essential equipment, or adequate referral facilities. Different forms of skilled attendance may have differing impact on mortality outcomes, which are masked by aggregate country-level measures. Like other covariates used, this measure is of national level and does not capture the situation at a facility level.

Fifthly, due to widespread lack of female literacy data, particularly in sub-Saharan Africa, there was a large drop-off in sample size for model III (n = 113) and model IV (n = 112). States without this data might systematically differ from those with it, meaning that the results presented for the fully adjusted models may not be generalizable.

Finally, although the study controlled for several factors, unmeasured confounders may remain. Unspecified cultural factors, varying quality of emergency obstetric care, availability of blood transfusion, and myriad other country-level phenomena might explain some portion of the observed variation. The tight correlations among the predictor variables also resulted in multicollinearity, meaning it is difficult to make definite statements regarding the relative influence of individual predictor variables.

Future directions

Based on the ecological findings, there are several promising avenues for future research. First, studies of individual-level data linking SBA to facility factors and maternal outcome would go beyond population-level correlations to assess the mechanisms implicated by the population-level data. A prospective cohort study following women from pre-conception to after childbirth would allow for stronger causal inference regarding the relationship between SBA and maternal survival.

More refined measures of quality of skilled birth attendance are also needed. Future research would benefit from separating the effects of doctors, nurses, and midwives, and should also incorporate data on emergency care capacity and facility resources, as well as data on quality of intrapartum care provided. Tracking SBA coverage and MMR trends within countries over time would help to establish temporal relations at the population level. Natural experiments that document the effects of interventions that dramatically increase SBA coverage on subsequent mortality rates would provide more robust quasi-experimental estimates of SBA's impact but require reliable civil registration data which remains elusive for the majority of countries that carry the highest burden.

Thirdly, research must continue to build comprehensive civil registration and maternal death surveillance systems for both middle- and high-income countries, in particular, high-burden countries where the stakes are highest. Improvement of civil registration infrastructure is not just "data gathering;" it is a critical public health intervention that enables all other health improvements.

Conclusion

This ecological study sought to explore associations between SBA coverage, socioeconomic development and MMR among 180 countries between 2010 and 2023. Significant ecological associations between SBA coverage and low MMR were found at a crude ecological level, but these associations became largely non-significant when controlling for other factors related to the general level of development of the country. After adjusting for GDP per capita, life expectancy, female literacy, sanitation, adolescent fertility, and health expenditure, residual variations in SBA coverage were not statistically associated with variations in MMR.

The limitations of the analysis restrict interpretation to some extent. Crucially, ecological relations cannot be generalized to individual-level associations (i.e., ecological fallacy). In addition, the cross-sectional design of the study prevents from inferring causality, and variations in measurement quality of MMR across countries may introduce bias. However, this does not imply that SBA may not be important for maternal survival, as clinical individual-level studies show considerable impact (5).

The results indicate that, at the population level, SBA clustered with other development factors that together predict maternal mortality. SBA coverage did not act as a unique predictor of MMR after taking other indicators of development into account. Future interventions intended to decrease MMR should therefore reflect multi-faceted approaches that consider development from a range of perspectives.

Future research needs to move beyond broad ecological analyses to investigate individual-level pathways and different dimensions of SBA and establish temporal precedence. Nevertheless, this study provides supporting evidence that while access to healthcare remains an important factor, it cannot be disentangled from other facets of development at the population level, highlighting the necessity of multi-dimensional policy approaches to reach SDG targets for maternal mortality reduction.

References

1. World Health Organization, Maternal mortality (2023); https://www.who.int/news-room/fact-sheets/detail/maternal-mortality.

2. L. Alkema, D. Chou, D. Hogan, S. Zhang, A. B. Moller, A. Gemmill, D. M. Fat, T. Boerma, M. Temmerman, C. Mathers, L. Say, Global, regional, and national levels and trends in maternal mortality between 1990 and 2015. Lancet 387, 462–474 (2016).
3. L. Say, D. Chou, A. Gemmill, Ö. Tunçalp, A. B. Moller, J. Daniels, A. M. Gülmezoglu, M. Temmerman, L. Alkema, Global causes of maternal death. Lancet Glob. Health 2, e323–e333 (2014).
4. United Nations, Transforming our world: The 2030 agenda for sustainable development (United Nations General Assembly, 2015).
5. W. J. Graham, J. S. Bell, C. H. W. Bullough, Can skilled attendance at delivery reduce maternal mortality in developing countries? Stud. Health Serv. Organ. Policy 17, 97–130 (2001).
6. C. G. Victora, A. Wagstaff, J. A. Schellenberg, D. Gwatkin, M. Claeson, J. P. Habicht, Applying an equity lens to child health and mortality. Lancet 362, 233–241 (2003).
7. S. Karlsen, L. Say, J. P. Souza, C. J. Hogue, D. L. Calles, A. M. Gülmezoglu, M. Rani, The relationship between maternal education and mortality among women giving birth in health care institutions. BMC Public Health 11, 606 (2011).
8. A. Prüss-Ustün, J. Wolf, C. Corvalán, R. Bos, M. Neira, Preventing disease through healthy environments (World Health Organization, 2016).
9. T. Ganchimeg, E. Ota, N. Morisaki, M. Laopaiboon, P. Lumbiganon, J. Zhang, B. Yamdamsuren, M. Temmerman, L. Say, Ö. Tunçalp, J. P. Vogel, J. P. Souza, R. Mori, Pregnancy and childbirth outcomes among adolescent mothers. BJOG 121, 40–48 (2014).
10. W. S. Robinson, Ecological correlations and the behavior of individuals. Am. Sociol. Rev. 15, 351–357 (1950).
11. R Core Team, R: A language and environment for statistical computing (Version 4.6.0) (R Foundation for Statistical Computing, 2026); https://www.R-project.org/.
12. World Bank, World Development Indicators (The World Bank Group, 2023); https://databank.worldbank.org/source/world-development-indicators.

Internet Use, Socioeconomic Indicators, and Suicide Mortality: An Ecological Analysis

Utilisation d'Internet, indicateurs socioéconomiques et mortalité par suicide : analyse écologique

Ayman Assaaoudi[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: aassa087@uottawa.ca

Abstract | Résumé

This study examined the associations between internet use, socioeconomic factors, and suicide mortality from 2015 to 2023 using World Bank data. The objective was to evaluate how technological access correlates with social and economic development indicators in relation to population-level mental health outcomes. The assessment analyzed Country-level data on internet penetration, GDP per capita, literacy, suicide mortality rates, life expectancy, school enrollment, unemployment, electricity access, and population growth. Correlation analyses, descriptive statistics, and global mapping summarized patterns, while hierarchical linear regression models adjusted for demographic and socioeconomic factors in stages. Results showed significant variation in internet use (2.2–100%) and suicide mortality (0.63–29.53 per 100,000). High-income countries had higher internet penetration (≅ 85%) and elevated suicide rates (≅ 9–11 per 100,000) compared to low-income countries (≅ 11% internet use; ≅ 6.5 per 100,000 suicide). Internet use was strongly correlated with GDP (r ≅ 0.70), literacy (r ≅ 0.60), and life expectancy (r ≅ 0.65). In regression models, internet use showed a weak positive association in unadjusted analyses that reversed to a negative association after accounting for socioeconomic factors (β = -0.30, 95% CI: -0.50, -0.10). Life expectancy and school enrollment were also associated with suicide mortality.

These findings indicate that the ecological, cross-sectional association between access to the internet and deaths from suicide is influenced by national socioeconomic circumstances. While more developed countries that had higher levels of internet access were more likely to have lower levels of suicide, the nature of the cross-sectional data and ecological design does not allow causal inference. These results are also important because they show the need to take national socioeconomic status into account when studying relationships between digital technology and the population's health status.

———

Cette étude a examiné les associations entre l'utilisation d'Internet, les facteurs socioéconomiques et la mortalité par suicide de 2015 à 2023, à partir des données de la Banque mondiale. L'objectif était d'évaluer comment l'accès aux technologies est lié aux indicateurs de développement social et économique, en lien avec les résultats de santé mentale à l'échelle des populations. L'analyse portait sur des données nationales concernant la pénétration d'Internet, le PIB par habitant, l'alphabétisation, les taux de mortalité par suicide, l'espérance de vie, la scolarisation, le chômage, l'accès à l'électricité et la croissance démographique. Des analyses de corrélation, des statistiques descriptives et des cartes mondiales ont permis de résumer les tendances observées, tandis que des modèles de régression linéaire hiérarchique ont ajusté progressivement les résultats pour des facteurs démographiques et socioéconomiques. Les résultats ont montré une variation importante de l'utilisation d'Internet (2,2 à 100 %) et de la mortalité par suicide (0,63 à 29,53 décès par 100 000 habitants). Les pays à revenu élevé présentaient une pénétration d'Internet plus élevée (environ 85 %) et des taux de suicide plus élevés (environ 9 à 11 décès par 100 000 habitants) que les pays à faible revenu (environ 11 % d'utilisation d'Internet; environ 6,5 décès par 100 000 habitants). L'utilisation d'Internet était fortement corrélée avec le PIB (r ≅ 0,70), l'alphabétisation (r ≅ 0,60) et l'espérance de vie (r ≅ 0,65). Dans les modèles de régression, l'utilisation d'Internet montrait une faible association positive dans les analyses non ajustées, mais cette association devenait négative après la prise en compte des facteurs socioéconomiques (β = -0,30; IC à 95 % : -0,50 à -0,10). L'espérance de vie et la scolarisation étaient aussi associées à la mortalité par suicide.

Ces résultats indiquent que l'association écologique et transversale entre l'accès à Internet et les décès par suicide est influencée par les conditions socioéconomiques nationales. Même si les pays plus développés, qui avaient généralement un meilleur accès à Internet, étaient plus susceptibles de présenter des niveaux plus faibles de suicide après ajustement, la nature transversale des données et le devis écologique ne permettent pas d'établir un lien de causalité. Ces résultats sont également importants, car ils montrent qu'il faut tenir compte du statut socioéconomique national lorsqu'on étudie les relations entre les technologies numériques et l'état de santé des populations.

Keywords: Suicide mortality, ecological study, digital access, Internet use, socioeconomic development.

Introduction

Suicide is a major global public health concern and one of the leading causes of death worldwide, with more than 700,000 deaths each year (1). In 2021, it was the third leading cause of death in the 15–29 age group, and nearly 75% of all suicides occur in low- and middle-income countries (1-2). Suicide mortality varies significantly by country-from an estimated 6 per 100,000 in low-income countries to about 11 per 100,000 in high-income countries. Conversely, global internet penetration has rapidly expanded the availability of information and modes of communication and interaction, from an estimated 11% of users in low-income countries to nearly 85% in high-income countries.

The relationship between internet use and mental health outcomes, including suicide-related behaviors, has been the subject of considerable debate in the literature. A systematic review by Marchant et al. (2017) examining internet use, self-harm, and suicidal behavior in young people identified a complex and bidirectional pattern (3). Their findings indicated that while problematic internet use and exposure to harmful online content are associated with increased self-harm and suicidal ideation, internet access could also facilitate help-seeking behaviors and connection to supportive communities. Similarly, other research has documented both the potential risks of cyberbullying, social comparison, and access to pro-suicide content, as well as the benefits of online mental health resources, crisis support services, and reduced social isolation (4-5). These mixed results imply that the relationship between access and suicide results may be complex and highly conditional on a number of environmental factors: types of internet usage, the type of available information, and the socio-economic milieu.

Among indicators of socioeconomic development, at the population level, GDP per capita, literacy, life expectancy, electrical power, and school enrollment are known predictors of mental health conditions. Prior studies suggest that increased levels of education and income in certain populations may be linked to decreased risk of suicide in other studies, although such relations can vary considerably among locations and populations. (6-7). For instance, the well-documented paradox wherein wealthier nations often report higher suicide rates than lower-income countries has been attributed to differences in reporting accuracy, age structure, cultural factors, and access to lethal means (8). The interaction between technological infrastructure and these traditional socioeconomic determinants remains poorly understood, particularly at the global level.

Few studies have investigated the relationship between global internet penetration and suicide mortality at the ecological level while controlling for the effects of socioeconomic development. Such an ecological investigation is important for several reasons. First, as digital infrastructure continues to spread globally, it is critical that policy decisions surrounding digital infrastructure investment are informed by an understanding of technology's impact on population mental health. Second, determining if an association between internet access and suicide is modified by the level of national development will help inform appropriate strategies for targeting mental health interventions to prevent suicide. Third, ecological studies can help identify patterns that exist at the population level that are not observable at the individual level and which can serve as a source of hypotheses for future in-depth study.

This study sought to examine ecological associations between internet penetration, socioeconomic indicators, and suicide mortality from 2015 to 2023 using country-level data. Specifically, this study aimed to: (1) describe global variation in internet access and suicide mortality across income groups, (2) assess the crude association between internet penetration and suicide rates, (3) evaluate how this association changes after accounting for key socioeconomic confounders, and (4) explore whether the relationship between internet access and suicide mortality varies by level of national development. By addressing these aims through hierarchical modeling approaches that account for the complex interdependencies among development indicators, this research contributes to understanding the broader context in which digital infrastructure expansion occurs and its relationship with population mental health outcomes. As internet access continues to expand rapidly across low- and middle-income countries, understanding how digital connectivity intersects with suicide mortality within different socioeconomic contexts is increasingly urgent for global health policy. These findings can help inform evidence-based decisions about digital infrastructure investments, identify settings where internet expansion may need to be coupled with mental health supports, and guide the development of targeted suicide prevention strategies that account for both technological and socioeconomic realities across diverse national contexts.

Methods and Materials

An ecological, cross-national study design was employed using publicly available country-level indicators retrieved from the World Bank World Development Indicators database spanning 2015 to 2023. Correlations were explored between national internet penetration and age-standardized suicide mortality while accounting for key socioeconomic development indicators. Countries were eligible for inclusion if data were available for both the primary exposure and outcome variables for at least one year during the study window. For countries with multiple years of available data, mean values from 2015 to 2023 were calculated to represent overall country-level conditions and to reduce the effect of interannual fluctuations.

The primary outcome variable was the age-standardized suicide mortality rate per 100,000 population, as reported in the World Bank World Development Indicators. Meaningful international comparisons were facilitated by age standardization to account for cross-national differences in age distribution. The primary exposure variable was internet penetration, measured as the percentage of the population reporting internet use. This indicator

captures national access to digital infrastructures and information technologies. Socioeconomic covariates were selected a priori based on theoretical relevance and prior literature on the social determinants of health. These variables included adult literacy rate (percentage of the population aged fifteen and older who can read and write), gross domestic product (GDP) per capita (constant US dollars), gross school enrollment rate, life expectancy at birth (years), unemployment rate (percentage of the labor force), percentage of the population with access to electricity, and annual population growth rate. GDP per capita was log-transformed before regression analyses to reduce skewness and to account for proportional changes across national income groups. Continuous variables were mean-centered to aid interpretation and reduce multicollinearity in multivariable and interaction models.

Descriptive analyses were used to describe the country-level distribution of suicide mortality, internet penetration and socioeconomic indicators. Summary statistics including means, standard deviations, and ranges were calculated for the full sample and stratified by World Bank income group. Geographic variation in suicide mortality and internet penetration was examined using global choropleth maps created using standard geographic visualization packages. Bivariate associations among internet penetration, suicide mortality, and socioeconomic indicators were evaluated using Pearson correlation coefficients. These analyses provided an initial assessment of the direction and magnitude of associations between variables.

Multivariable associations were evaluated using hierarchical ordinary least squares regression models. The modeling strategy was designed to assess how the estimated association between internet penetration and age-standardized suicide mortality changed with progressive adjustment for national socioeconomic factors. The initial model specification included internet penetration as the sole predictor. Subsequent models sequentially introduced covariate blocks: demographic and health indicators (life expectancy and population growth), economic indicators (log GDP per capita), infrastructure and labor market indicators (electricity access and unemployment), and education-related measures (literacy rate and school enrollment). This sequential modeling strategy allowed assessment of potential confounding and the extent to which development indicators accounted for the unadjusted association between internet penetration and suicide mortality.

To evaluate effect modification by development stage, interaction terms were included between internet penetration and key socioeconomic indicators (e.g., internet × log GDP per capita and internet × literacy rate). Variable importance plots were generated using standardized regression coefficients from models predicting GDP per capita, with importance quantified as the absolute magnitude of standardized beta coefficients. The relative contribution of each predictor in explaining variation in national income was illustrated by these plots, both overall and stratified by World Bank income classification. Population-weighted models were used in sensitivity analyses to evaluate the influence of country population size on aggregate associations.

Coefficients are reported with 95% confidence intervals; two-sided tests were performed, and statistical significance was set at $\alpha = 0.05$. Model diagnostics and robustness analyses were conducted to evaluate assumptions and the consistency of findings. Multicollinearity was assessed using variance inflation factors and addressed by re-estimating alternative specifications when collinearity was high. Influential observations were identified using leverage statistics and Cook's distance, and sensitivity analyses excluding high-leverage countries were conducted. Alternative specifications explored included transformation of the outcome when appropriate, and quantile regression was used to examine heterogeneity across the distribution of suicide mortality. For countries with adequate temporal coverage, fixed-effects panel models were estimated as an additional sensitivity analysis. Patterns of missing data were summarized, and primary analyses were conducted using complete-case data when missingness was minimal, with multiple imputation by chained equations used as a sensitivity analysis when missingness exceeded predefined thresholds. All data management, statistical analyses, and visualizations were performed using the World Bank World Development Indicators dataset and the R statistical environment. Since the study relied exclusively on aggregated, publicly available country-level data, institutional ethics review and informed consent were not required.

Results

Descriptive characteristics of the study sample

The analytic sample included countries with available data on internet penetration, suicide mortality, and key socioeconomic indicators for at least one year from 2015–2023. Significant cross-national variation was observed across all study variables. Internet penetration ranged from approximately 2.2%–100%, reflecting substantial disparities in access to digital infrastructure worldwide. Age-standardized suicide mortality rates varied from approximately 0.63–29.53 deaths per 100,000 population. When stratified by World Bank income classification, clear gradients were evident (Table 1). High-income countries had substantially higher average internet penetration (mean ≅ 85%) compared to approximately 11% in low-income nations. Suicide mortality rates were also higher on average in high-income countries (≅ 9–11 per 100,000) compared to low-income countries (≅ 6.5 per 100,000), with middle-income countries showing intermediate patterns. Socioeconomic indicators including GDP per capita, life expectancy, electricity access, school enrollment, and literacy rates increased consistently across income categories, while unemployment and population growth showed more heterogeneous patterns without a monotonic relationship to income group. These descriptive results demonstrate substantial clustering of development-related variables, underscoring the need for multivariable modeling to account for confounding and collinearity.

Table 1. Descriptive statistics of study variables by income group. Summary statistics showing mean values and standard deviations for internet penetration (% of population), suicide mortality (age-standardized deaths per 100,000), and socioeconomic indicators stratified by World Bank income classification (low-income, lower-middle-income, upper-middle-income, high-income). GDP = gross domestic product. Data obtained from World Bank World Development Indicators, 2015–2023.

income	variable	mean	sd	min	max
High income	electricity	99.94	0.11	99.70	100.00
	gdp	31891.64	18962.85	13433.92	80056.13
	internet	84.54	9.71	64.57	100.00
	life_exp	80.37	2.79	72.81	83.60
	literacy	97.55	1.44	93.04	99.70
	pop_growth	0.73	1.79	-4.17	4.83
	school_enroll	107.70	9.16	86.20	124.85
	suicide	9.15	7.16	1.03	23.40
	unemployment	7.43	5.28	1.80	22.06
Low income	electricity	26.37	17.44	8.40	71.50
	gdp	670.09	228.25	210.01	996.38
	internet	11.29	7.16	2.20	37.62
	life_exp	60.37	3.54	51.60	66.35
	literacy	53.08	17.38	22.31	76.72
	pop_growth	2.64	0.36	1.72	3.22
	school_enroll	43.46	16.10	20.19	107.03
	suicide	6.85	2.12	3.46	12.48
	unemployment	3.86	2.43	1.03	10.76
Lower middle income	electricity	87.37	17.40	26.20	100.00
	gdp	2867.53	1472.33	877.26	9174.54
	internet	40.59	19.03	10.00	82.18
	life_exp	69.58	5.10	52.19	78.26
	literacy	81.75	15.51	43.59	100.00
	pop_growth	1.41	1.02	-3.22	3.26
	school_enroll	71.67	19.02	26.09	98.33
	suicide	6.79	5.30	0.63	26.73
	unemployment	7.02	6.05	0.12	26.39
Upper middle income	electricity	97.25	7.91	49.70	100.00
	gdp	7360.53	2168.90	3880.69	12425.03
	internet	64.89	13.12	28.81	88.21
	life_exp	73.90	3.65	60.81	79.71
	literacy	94.81	3.63	80.20	99.97
	pop_growth	1.04	0.82	-0.54	3.48
	school_enroll	95.04	13.69	54.08	124.33
	suicide	7.22	5.41	0.76	29.53
	unemployment	8.90	6.51	0.60	34.01

Geographic distribution of internet penetration and suicide mortality

Figures 1 and 2 represent the spatial distribution of internet penetration and suicide mortality globally and their respective distributions by income level. The data showcase considerable geographic clustering of the suicide mortality data, where the highest reported rates appear in Central Asia, Eastern Europe, East Asia, and parts of South America and Africa. On the other hand, much of Sub-Saharan Africa, along with parts of South America and the Middle East, have very low suicide mortality data. The geographic distribution of Internet penetration data is equally stark, with the majority of countries in North America, East Asia, Western Europe, and Oceania considered to have universal Internet access, whilst a large portion of countries in South Asia and Sub-Saharan Africa have penetration rates of under 20%. Visual observation clearly shows some geographical overlap between locations with high internet penetration and those with higher rates of suicide mortality, such as wealthier areas of East Asia and Europe. However, this is not always true, there are some highly connected countries with moderate suicide rates and some with very low internet penetration rates with high rates. This indicates that internet penetration cannot simply be interpreted without regard to the overarching issues such as geography and socioeconomics.

Correlation structure and bivariate associations

Pearson correlation analyses (Figure 2) revealed that internet penetration was strongly associated with socioeconomic development indicators. Internet use showed strong positive correlations with GDP per capita ($r \cong 0.70$), literacy rate ($r \cong 0.60$), and life expectancy ($r \cong 0.65$), as well as with school enrollment and electricity access. In contrast, the unadjusted association between internet penetration and suicide mortality was weak and positive, suggesting that countries with higher internet usage tended to report slightly higher rates of suicide mortality at the ecological level. The relationship among the development indicators had moderate to strong correlation, particularly between GDP per capita, life expectancy, literacy, and electricity (implying that the variables were indeed very correlated). Therefore, hierarchical multiple regression evaluated the independent effects of internet use on suicide mortality, while controlling for socioeconomic development.

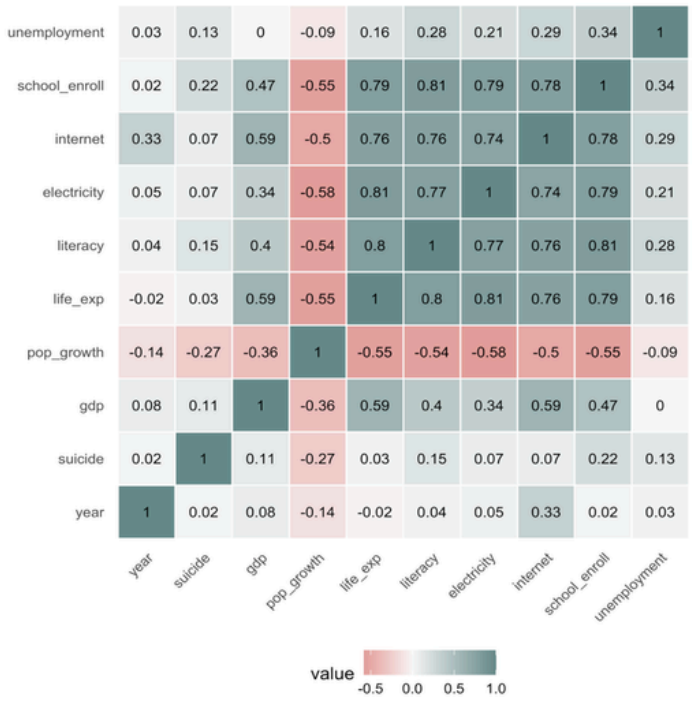


Figure 2. Correlation matrix of study variables. Pearson correlation coefficients between internet penetration, suicide mortality, and socioeconomic indicators. Coefficient magnitude shown by color intensity and circle size. GDP = gross domestic product. Data obtained from World Bank World Development Indicators, 2015–2023.

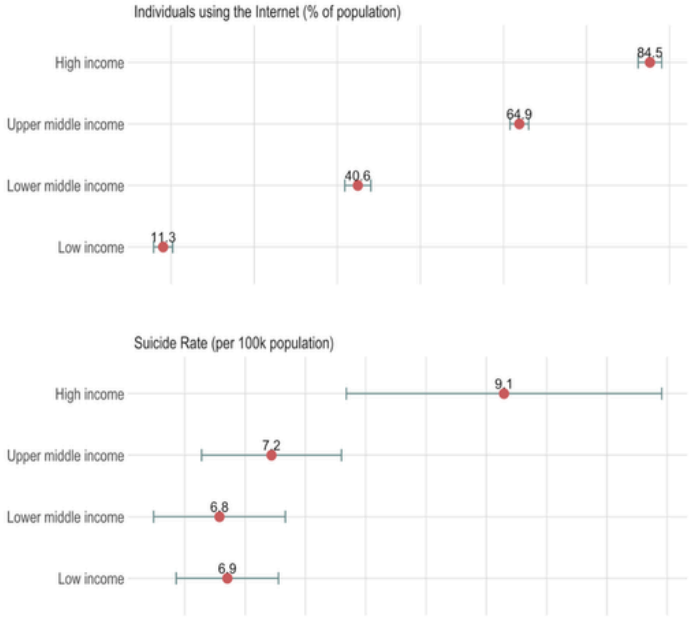


Figure 1. Internet use and suicide rates by income groups. Box plots showing distribution of internet penetration (% of population) and age-standardized suicide mortality rates (per 100,000) across World Bank income groups (low-income, lower-middle-income, upper-middle-income, high-income). Boxes show interquartile range, horizontal lines show medians. Data obtained from World Bank World Development Indicators, 2015–2023.

Hierarchical regression models

Hierarchical regression models assessed how the association between internet penetration and suicide mortality changed after sequential adjustment for socioeconomic development indicators (Table 2). In the unadjusted model, internet penetration showed a small positive association with suicide mortality that did not reach statistical significance. After adjustment for GDP per capita and life expectancy, the magnitude of the internet coefficient was substantially attenuated, indicating that the crude positive association was largely explained by differences in demographic

Table 2. Hierarchical regression models predicting suicide mortality rate. Unstandardized beta coefficients with 95% confidence intervals from four sequential hierarchical regression models. Model I: internet penetration only. Model II: adds GDP per capita and life expectancy. Model III: adds electricity access, unemployment, literacy, and school enrollment. Model IV: adds population growth. GDP = gross domestic product. ***p < 0.001; **p < 0.01; *p < 0.05. Data obtained from World Bank World Development Indicators, 2015–2023.

	Model 1		Model 2		Model 3		Model 4		Model 5	
Characteristic	**Beta**	**95% CI** [1]	**Beta**	**95% CI** [1]	**Beta**	**95% CI** [1]	**Beta**	**95% CI** [1]	**Beta**	**95% CI**
internet	0.07	-0.04, 0.18	0.01	-0.13, 0.14	-0.16	-0.34, 0.03	-0.28	-0.47, -0.08	-0.30	-0.50, -0.10
gdp			0.11	-0.03, 0.24	0.20	0.07, 0.34	0.20	0.05, 0.34	0.21	0.07, 0.36
pop_growth					-0.32	-0.45, -0.19	-0.29	-0.42, -0.17	-0.30	-0.43, -0.17
life_exp					-0.46	-0.66, -0.26	-0.55	-0.78, -0.33	-0.54	-0.77, -0.31
literacy					0.38	0.19, 0.58	0.24	0.04, 0.44	0.23	0.03, 0.43
electricity							-0.08	-0.29, 0.13	-0.07	-0.28, 0.14
school_enroll							0.49	0.28, 0.69	0.46	0.25, 0.67
unemployment									0.07	-0.04, 0.18
R^2	0.005		0.013		0.145		0.202		0.206	
Adjusted R^2	0.002		0.006		0.131		0.184		0.185	

[1] CI = Confidence Interval

and economic development. With the inclusion of education (literacy and school enrollment), unemployment, infrastructure (electricity access), and population growth, the association reversed direction. In the fully adjusted model, higher internet penetration was associated with lower suicide mortality (β = -0.30; 95% CI: -0.50, -0.10), suggesting that among countries at comparable levels of socioeconomic development, greater internet access was correlated with lower population-level suicide mortality. This ecological association corresponds to approximately 0.3 fewer deaths per 100,000 people for each 10 percentage-point increase in national internet penetration, holding other factors constant. Several socioeconomic variables remained independently associated with suicide mortality. Life expectancy showed a positive association, consistent with demographic composition effects in countries with older age distributions. School enrollment demonstrated a strong inverse association at the ecological level. GDP per capita and electricity access showed weaker associations after adjustment, while population growth and unemployment did not remain statistically significant predictors in the fully adjusted model.

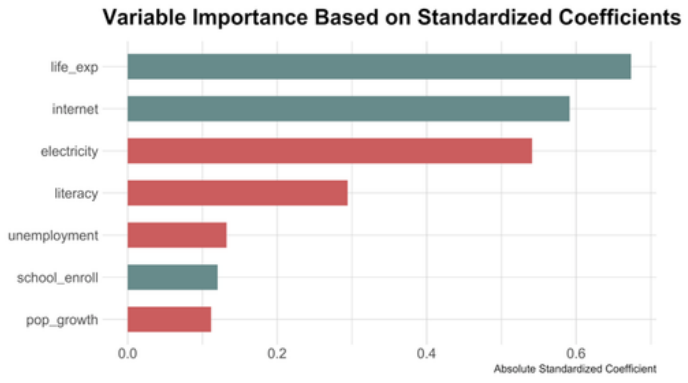


Figure 3. Variable importance for predictors of GDP per capita. Variable importance quantified as absolute magnitude of standardized regression coefficients from models predicting GDP per capita. Positive bars indicate positive associations; negative bars indicate inverse associations. GDP = gross domestic product. Data obtained from World Bank World Development Indicators, 2015–2023.

Variable importance analysis

The variable importance analysis used standardized regression coefficients from models predicting GDP per capita. Variable importance was measured as the magnitude (absolute value) of standardized coefficients in models explaining national income; higher standardized coefficients indicate greater explanatory contribution. The variable with the highest importance was life expectancy, followed by internet access and electricity access; literacy and schooling were moderately correlated with national income. Unemployment and population growth were inversely associated with national income. This illustrates the background context for the observed changes in the association of internet access and suicide mortality following adjustment: it is not an isolated variable but is associated with other indicators of development.

Effect modification by socioeconomic development

Interaction analyses showed that the association between internet penetration and suicide mortality differed by national level of development. Internet penetration interacted with GDP per capita and literacy rate; countries with greater levels of economic development and literacy demonstrated a greater inverse

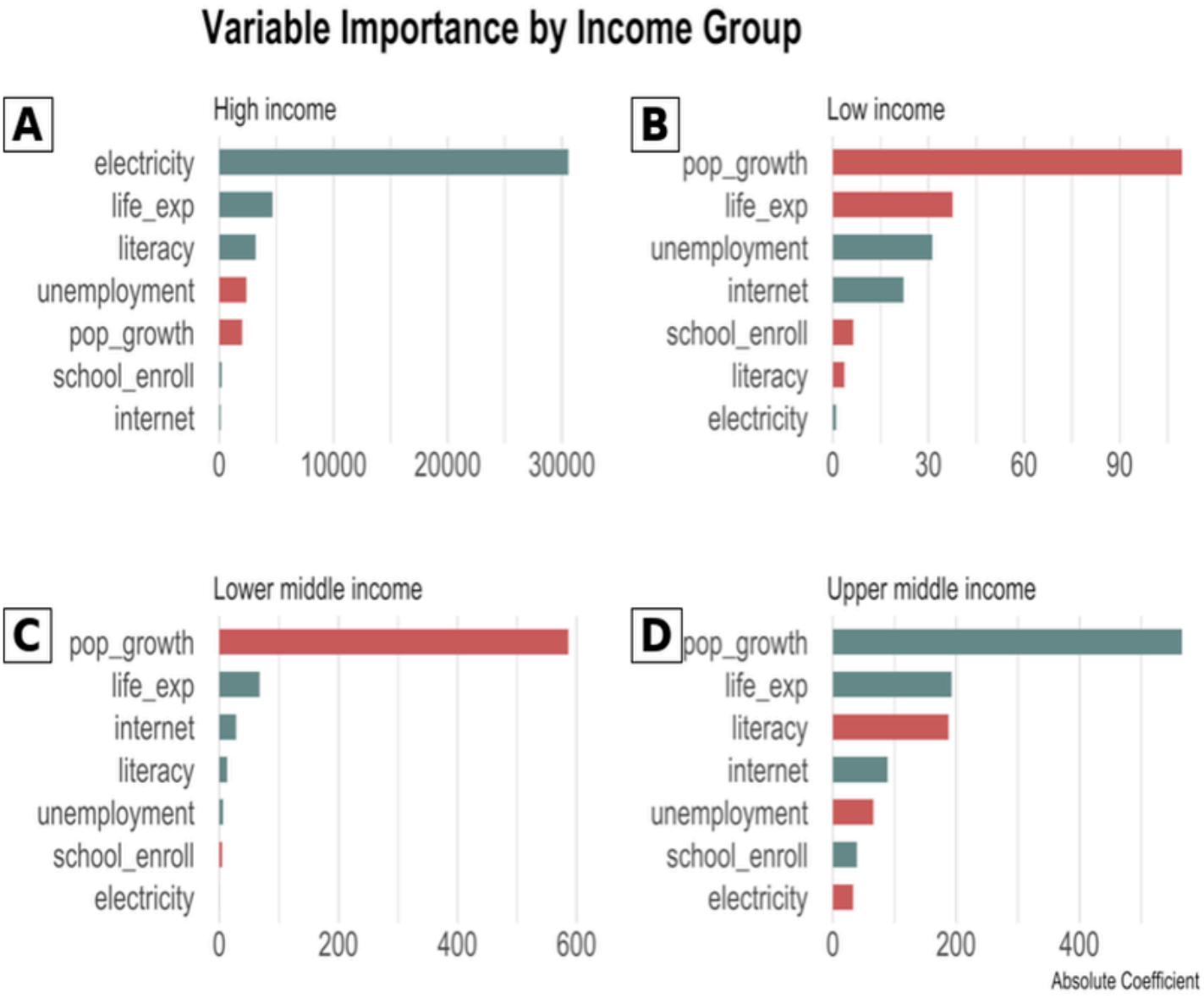


Figure 4. Variable importance stratified by income group. Panel A reports results for high-income countries, Panel B reports results for low-income countries, Panel C reports results for lower-middle-income countries, and Panel D reports results for upper-middle-income countries. Variable importance stratified by World Bank income classification shows the difference in variables that are predictive of GDP per capita at different levels of development. Positive bars correspond to positive associations while negative bars correspond to negative associations.GDP is gross domestic product. Data are from the World Bank World Development Indicators (2015-2023).

relationship between internet penetration and suicide mortality. For countries with lower literacy or lower income, the association was weaker and in some cases, non-significantly distinguishable from null. These analyses suggest that the ecological association between internet penetration and suicide mortality is more significant in environments with greater levels of socioeconomic development. Figure 4 reveals results of variable importance analysis by World Bank income levels; although the analysis results vary considerably across levels of development, associations between other development variables and GDP per capita also appear to differ by economic level of country development. Among high-income countries, life expectancy was the most significant determinant of variation in GDP, followed by internet use and school enrollment. Internet use was the most important predictor of GDP for upper-middle-income countries, followed by literacy rate and life expectancy. For lower-middle-income countries, the analysis showcased a pattern whereby unemployment had strong negative importance for GDP per capita, but life expectancy and internet use were positively associated with GDP per capita. In low-income countries, population growth was strongly negatively associated with GDP per capita, and access to electricity was among the strongest positively associated variables in the model.

Sensitivity analyses and robustness checks

A series of sensitivity analyses confirmed the robustness of the main findings. Centering predictors and alternative model specifications addressed moderate collinearity among development indicators. Exclusion of high-leverage countries did not substantially alter the direction or magnitude of the internet coefficient. Alternative outcome specifications, including transformations of suicide mortality, yielded consistent results. Quantile regression showed that the inverse association between internet penetration and suicide mortality was strongest near the median of the distribution and weaker at the extremes. Fixed-effects panel models for countries with sufficient temporal coverage produced directionally consistent but less precise estimates, likely due to limited within-country variation over the relatively short study period.

Discussion

This ecological study examined associations between internet penetration, socioeconomic development, and suicide mortality across countries from 2015 to 2023. At the crude ecological level, internet use showed a positive correlation with suicide mortality, a pattern largely explained by the tendency for higher-income countries to have both greater internet access and higher recorded suicide rates. However, after accounting for socioeconomic development through hierarchical modeling, a different pattern emerged. Among countries at comparable levels of development, higher internet penetration was associated with lower rates of suicide mortality. This reversal demonstrates that the crude positive association observed globally reflects confounding by national development rather than a direct relationship between internet access and suicide risk.

It is critical to emphasize that these findings are ecological in nature and describe associations at the country level. Ecological associations cannot be used to infer individual-level relationships —a limitation known as the ecological fallacy (9). A country with high internet penetration and low suicide mortality does not necessarily mean that individuals who use the internet are less likely to die by suicide. Individual-level mechanisms may differ substantially from population-level patterns. For example, while the ecological analysis suggests an inverse association after accounting for development, individual-level studies have documented both beneficial and harmful effects of internet use on mental health, with outcomes depending heavily on how individuals engage with digital technologies (3).

The hierarchical regression results illustrate how sequential adjustment for confounders alters the estimated association. In the unadjusted model, internet penetration showed a small positive but non-significant association with suicide mortality. The introduction of GDP per capita and life expectancy moved the internet coefficient toward zero, indicating that wealthier and healthier countries largely explained the crude association. The addition of electricity access, education, population growth, and unemployment reversed the direction of the association and became statistically significant, with higher internet penetration associated with lower suicide rates. This pattern suggests that when comparing countries at similar developmental stages, internet access is correlated with reduced population-level suicide mortality. However, this association should not be interpreted causally given the cross-sectional design and potential for unmeasured confounding. Moreover, as an ecological study, the risk of ecological fallacy prevents the extrapolation of these country-level associations to individual-level relationships.

These findings both align with and diverge from prior research in important ways. The systematic review by Marchant et al. (2017) highlighted the complexity of internet-mental health relationships at the individual level, documenting both risks (cyberbullying, exposure to harmful content) and benefits (access to support, reduced isolation). The ecological findings suggest that at the population level, in contexts with strong socioeconomic infrastructure, internet access may co-occur with lower suicide mortality. However, the results contrast with concerns that increased internet access might uniformly increase suicide risk through mechanisms such as social media-related distress or access to information about suicide methods. The reversal of the association after adjustment for development indicators suggests that the apparent positive correlation between internet use and suicide in crude comparisons is attributable to confounding by national development level rather than internet access itself being harmful.

Several socioeconomic indicators were independently associated with suicide mortality. Life expectancy shows a positive association, consistent with the demographic reality that countries with older populations have higher proportions of individuals in age groups with elevated suicide risk. School enrollment

demonstrates a strong inverse association, suggesting that broader access to education may be associated with population-level suicide prevention, potentially through economic opportunity, social integration, or mental health literacy. These associations align with prior literature linking education and economic opportunity to mental health outcomes (6). The correlation matrix reveals why unadjusted analyses can be misleading in ecological research. Internet penetration was strongly correlated with life expectancy, literacy, GDP, and electricity access, reflecting the tendency for development indicators to co-occur rather than evolve independently. Without controlling for this clustering, internet use appears associated with suicide mortality simply because both correlate with economic development and modernization. This underscores a fundamental challenge in ecological research: separating the independent contribution of specific exposures from the broader context in which they occur.

The interaction analyses reveal that the association between internet penetration and suicide mortality varied by development context, being strongest in settings with higher literacy rates and GDP per capita. This heterogeneity suggests that internet access may represent different realities across developmental stages. In more developed countries with robust healthcare systems, mental health services, and educational infrastructure, greater connectivity may facilitate access to supportive resources. In less developed settings, limitations in infrastructure, affordability, or content quality may constrain potential benefits. However, these interpretations are speculative and would require individual-level data to test mechanistically. The variable importance analyses across income groups revealed substantial heterogeneity in which development indicators best predicted national income. In high-income countries, life expectancy and internet use were most strongly associated with GDP variation. In upper-middle-income nations, internet use emerged as the strongest predictor, highlighting the role of digital infrastructure during economic transitions. In lower-middle-income countries, unemployment played a larger role, while in low-income nations, electricity access and population growth were most influential. These patterns suggest that technological infrastructure interacts differently with development depending on a country's developmental stage, which may partly explain why the internet-suicide association varied across contexts.

Limitations

This study has several important limitations inherent to its design and data structure. First and most critically, the analysis is ecological, with all variables measured at the country level. This design precludes inference about individual-level relationships, a limitation known as ecological fallacy (9). Associations observed at the population level cannot be assumed to hold at the individual level. For example, while higher national internet penetration associates with lower suicide mortality after adjustment for development, this pattern does not mean that individual internet users have lower suicide risk. Individual-level mechanisms may operate differently, and the relationship between personal internet use and suicide risk likely depends on numerous factors that ecological data fail to capture, including patterns of use, content accessed, pre-existing mental health status, and availability of offline support.

Second, the cross-sectional nature of these data prevents any causal interpretation. The design makes it impossible to determine whether internet penetration influences suicide mortality, whether countries experiencing declining suicide mortality invest more in internet infrastructure, or whether unmeasured common causes drive both. Temporal precedence cannot be established from cross-sectional ecological data. While sensitivity analyses using fixed-effects models for countries with adequate temporal coverage provide some support for consistency over time, the limited time window (2015–2023) and substantial missing data constrain these analyses.

Third, suicide reporting accuracy and completeness vary substantially across countries. Underreporting and misclassification are common, particularly in low-income countries and settings where suicide carries social stigma (7). The geographic maps revealed regions with missing or incomplete data, especially in parts of sub-Saharan Africa and small island states. If suicide is systematically underestimated in certain regions, biases in observed differences between income groups may arise. This measurement error could produce spurious associations or mask true patterns.

Fourth, the study aggregated data across multiple years to form a single cross-section, which limits observation of temporal trends within individual countries. While this approach reduces the impact of year-to-year fluctuations, it prevents examination of how changes in internet access over time relate to changes in suicide mortality within the same country. Such within-country analyses would provide stronger evidence about temporal relationships but remain unfeasible with the available data. Fifth, internet penetration is a crude measure of digital access. It captures the proportion of the population with internet access but does not measure how the internet is used, what content users access, the quality of connectivity, or whether usage patterns are beneficial or harmful for mental health. Different forms of online engagement may have divergent mental health implications, which aggregate penetration rates cannot differentiate. Finally, despite extensive adjustment for socioeconomic confounders, unmeasured confounding remains possible. Cultural factors, mental health service availability, availability of lethal means, substance use patterns, and other country-level characteristics that correlate with both internet access and suicide mortality could account for some or all of the observed associations. The strong correlations among development indicators also introduce multicollinearity, which can increase uncertainty in coefficient estimates despite efforts to address this through centering and alternative specifications.

Future Directions

Future research should build on these ecological findings in

several ways. Most importantly, individual-level data linking internet use patterns, mental health status, and suicide outcomes would enable direct evaluation of the mechanisms that these population-level results suggest. Longitudinal cohort studies following individuals over time as their digital engagement varies would help establish temporal precedence and support stronger causal inference than is possible with cross-sectional ecological data. More detailed measures of digital engagement are needed. Differentiating between types of internet use—social media, entertainment, education, health information seeking—could reveal which aspects of connectivity are most relevant for mental health. Research should also examine potentially harmful online exposures (cyberbullying, pro-suicide content) alongside potentially beneficial ones (crisis resources, social support) to understand the overall association between internet access and mental health outcomes.

Longitudinal country-level studies tracking nations as internet access expands would strengthen understanding of temporal relationships. Natural experiments—such as policy changes that rapidly expand internet access in specific regions—could provide quasi-experimental leverage for assessing whether changes in internet penetration precede changes in suicide mortality. Including indicators of mental health service availability, quality of online mental health resources, cultural attitudes toward suicide and help-seeking, and social support systems would allow more comprehensive modeling of the context in which internet access operates. This would help explain why associations between internet penetration and suicide mortality vary across income groups in this study. Finally, improving suicide mortality surveillance in low-income countries is essential. Strengthening vital registration systems and reducing underreporting would reduce measurement bias and enable more accurate assessment of global mental health patterns. Without improved data quality in these settings, differential measurement error across regions will continue to limit ecological studies.

Conclusion

This ecological study examined associations between internet access, socioeconomic development, and suicide mortality at the country level from 2015 to 2023. The analyses reveal that the apparent relationship between digital access and suicide outcomes varies substantially depending on how the study addresses socioeconomic confounding. Crude analyses suggested a positive association between internet penetration and suicide mortality, but this pattern was largely explained by the tendency for wealthier countries to have both higher internet access and higher recorded suicide rates. After hierarchical adjustment for educational, demographic, economic, and infrastructural factors, higher internet penetration was associated with lower suicide mortality among countries at comparable developmental stages.

However, several critical limitations constrain interpretation of these findings. Most importantly, ecological associations cannot be used to infer individual-level relationships. The ecological fallacy means that population-level patterns may not reflect individual-level mechanisms. Additionally, the cross-sectional design precludes causal inference, preventing determination of whether internet access influences suicide mortality or whether both are influenced by common underlying factors. Variations in suicide reporting accuracy across countries, particularly systematic underreporting in low-income settings, introduce measurement error that may bias estimates. The interaction and stratified analyses revealed that associations between internet penetration and both economic development and suicide mortality varied substantially across income groups. This heterogeneity suggests that digital infrastructure operates within different socioeconomic contexts at different stages of development. In high-income countries with established healthcare and educational systems, internet access may co-occur with features that support mental health. In lower-income settings, basic infrastructure needs and demographic pressures may be more salient.

These findings suggest that simplistic narratives about internet access being uniformly harmful or beneficial for mental health are likely inadequate. The relationship between technology and population mental health appears to be embedded within broader patterns of socioeconomic development, and the direction and magnitude of associations depend critically on the developmental context and the presence of confounding factors. From a policy perspective, these results underscore that expanding internet access alone is unlikely to reduce suicide mortality without concurrent investments in healthcare, education, and social infrastructure. Future research should move beyond ecological associations to examine individual-level mechanisms through which digital engagement relates to mental health across diverse socioeconomic settings. Longitudinal designs that can establish temporal precedence, more granular measures of internet use quality and content, and improved suicide surveillance globally are all necessary to advance understanding in this area. Nonetheless, this study provides evidence that the relationship between digital development and population mental health is complex, context-dependent, and inseparable from broader socioeconomic development processes. Mental health policies for the population should thus take technological infrastructure into account within a broad context, including education, opportunity, health care services, and support.

References

1. World Health Organization, “Suicide” https://www.who.int/news-room/fact-sheets/detail/suicide (2025).
2. H. Ritchie, “Suicide rates are higher in men than women” https://ourworldindata.org/data-insights/suicide-rates-are-higher-in-men-than-women (2025).
3. A. Marchant, K. Hawton, A. Stewart, P. Montgomery, V. Singaravelu, K. Lloyd, N. Purdy, K. Daine, A. John, A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: The good, the bad and the unknown. PLOS ONE 12, 1-16 (2017).

4. S. M. Dunlop, E. More, D. Romer, Where do youth learn about suicides on the Internet, and what influence does this have on suicidal ideation? Journal of Child Psychology and Psychiatry, 52, 1073-1080 (2011).
5. J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, H. Herrman, Social media and suicide prevention: a systematic review. Early Intervention in Psychiatry, 10, 103-121. (2015)
6. M.K Nock, G. Borges, E. J. Bromet, J. Alonso, M. Angermeyer, A. Beautrais, R. Bruffaerts, W. T. Chiu, G. de Girolamo, S. Gluzman, R. de Graaf, O. Gureje, J. M. Haro, Y. Huang, E. Karam, R. C. Kessler, J. P. Lepine, D. Levinson, M. E. Medina-Mora, Y. Ono, J. Posada-Villa, D. Williams, Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. British Journal of Psychiatry, 192, 98-105 (2018).
7. World Health Organization, “Preventing suicide: A global imperative” https://www.who.int/publications/i/item/9789241564779 (2014).
8. P. Värnik, Suicide in the world. International Journal of Environmental Research and Public Health, 9, 760-771 (2012).
9. W. S. Robinson, Ecological correlations and the behavior of individuals. American Sociological Review, 15, 351-357 (1950).

Original Research

Liquid Gold: Quantification of Vitamin E in Mustard Seed Oil

Or liquide: quantification de la vitamine E dans l'huile de graine de moutarde

Pegah Yousefirad[1], Sharon Barden[1], Paul M Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

In addition to its popularity in the culinary world, the mustard plant, a member of the *Brassicaceae* family, has a variety of medicinal applications. The plant is well known for its potent antioxidant and anti-inflammatory properties, often derived from its golden seed oil. Existing studies acknowledge the presence of vitamin E in mustard oil; however, extensive research has not been done to quantify vitamin E in mustard seed oil. This study aimed to examine the vitamin E content in yellow (*Brassica alba*) and black (*Brassica nigra*) mustard seed to determine the best source. Analysis and quantification, carried out by gas chromatography-mass spectrometry (GC-MS), revealed that a sample of black mustard oil contained a total vitamin E content of 8.0 µL/mL; 6.8 µL of γ-tocopherol and 1.2 µL of α-tocopherol. Alternatively, yellow mustard oil contained 0.43 µL/mL in only the g-tocopherol form. These results suggest that black mustard seed oil is a better source of vitamin E, as it contains an overall higher concentration and various forms of tocopherol, in contrast to the singular form found in yellow mustard at a lower concentration.

En plus de sa popularité dans le monde culinaire, la plante de moutarde, membre de la famille des Brassicacées, a une variété d'applications médicales. Cette plante est bien connue pour ses puissantes propriétés antioxydantes et anti-inflammatoires, souvent issues de l'huile dorée de ses graines. Les études existantes reconnaissent la présence de la vitamine E dans l'huile de moutarde, cependant, des recherches approfondies n'ont pas encore été effectuées pour quantifier la vitamine E dans l'huile de graine de moutarde. Cette étude visait à examiner la quantité de vitamine E dans la graine de moutarde jaune (*Brassica alba*) et noire (*Brassica nigra*) pour en déterminer la meilleure source. L'analyse et la quantification, réalisées par la chromatographie en phase gazeuse couplée à la spectrométrie de masse (CPG-SM), ont révélé que le contenu en vitamine E d'un échantillon d'huile de graine de moutarde noire était de 8.0 µL/mL; 6.8 µL étant du γ-tocophérol et 1.2 µL étant du α-tocophérol. D'une autre part, l'huile de graine de moutarde jaune contenait 0,43 µL/mL uniquement sous forme de g-tocophérol. Ces résultats suggèrent que l'huile de graine de moutarde noire est une meilleure source de vitamine E puisqu'elle possède une concentration globalement plus élevée et des formes variées de tocophérol, en contraste avec l'huile de graine de moutarde jaune qui en possédait seulement une forme et ce, à une concentration plus faible.

Keywords: mustard seed oil; vitamin E; tocopherols; Brassica nigra; Brassica alba; GC-MS; antioxidant compounds; γ-tocopherol; α-tocopherol; phytochemical analysis

Introduction

The mustard plant, a member of the Brassicaceae family, is cultivated worldwide for its oil-rich seeds. Human use of mustard dates to 3000 BCE, with ancient civilizations such as those in India, Greece, and Rome using it as both a spice and a medicine (1). Modern research has confirmed that the golden seed oil has potent antioxidant and anti-inflammatory effects stemming from metabolic compounds found within it. Various phenolic compounds have been linked to antioxidant benefits while anti-inflammatory properties are attributed to glucosinolates, sulfur-containing secondary metabolites responsible for mustards distinctive bitter taste and pungent smell (2). When hydrolyzed by the enzyme myrosinase, glucosinolates are converted into isothiocyanates, their biologically active form (3). This transformation is triggered by damage to the plant tissues, which occurs when seeds are crushed during the oil extraction process (4). As a result, mustard oil is enriched with isothiocyanates. Different mustard species can offer slightly different degrees of therapeutic benefits due to their varying chemical compositions (1).

Yellow mustard (*Brassica alba*) is a species native to the Mediterranean (Figure 1). Its predominant glucosinolate is sinalbin, which is converted to 4-hydroxybenzyl isothiocyanate. Its phenolic compounds include sinapine, sinapic acid, and vitamin E (3,5). Yellow mustard seeds are especially acknowledged for their high content of omega-3 fatty acids such as alpha-linolenic acid (2).

Figure 1. Seeds of black mustard (left) (Brassica nigra) and yellow mustard (right) (Brassica alba). Original image created by Veganbaking.net, licensed under CC-BY-SA-2.0 and cropped.

Black mustard (*Brassica nigra*) is native to regions of Africa, Asia, and Europe (Figure 1). These seeds contain a high level of anthocyanins, water-soluble pigments that give them their characteristic dark color. Like yellow mustard, black mustard is rich in sinapine, sinapic acid, and vitamin E, but also contains p-coumaric acid (3,5). The predominant glucosinolate in black mustard is sinigrin, which is converted to allyl-isothiocyanate (2). Studies suggest that sinigrin is the glucosinolate most strongly associated with anti-inflammatory response, making black mustard especially valuable for medicinal applications related to treating inflammation (6). With their higher oil content, black mustard seeds also offer a stronger flavor profile.

Despite the many beneficial components of these seed oils, mustard is also recognized for a less highly praised counterpart: a monounsaturated fatty acid known as erucic acid. Studies conducted in animal models have suggested that prolonged consumption of this fatty acid may result in the development of myocardial lipidosis, a condition involving abnormal accumulation of fats in the heart muscle (7). Erucic acid can potentially make up 50% of mustard oil's fatty acid content (8). Given this high percentage, North American regulatory bodies such as the Food and Drug Administration and Health Canada have issued warnings and placed restrictions on its use in cooking to prevent negative impacts on cardiac health. Regardless of its culinary limitations, mustard seed oil remains widely available for external use and is instead commonly incorporated into skincare routines as a result. In the cosmetic industry, many brands have begun incorporating vitamin E into products to obtain benefits mirroring those provided by mustard seed oil. This potent and well-studied fat-soluble antioxidant has been used in dermatology for over 50 years (9). Vitamin E's mechanism of action involves quenching reactive oxygen species, reducing oxidative stress (10). The human body cannot produce vitamin E on its own, and as a result, it must be obtained from an external source, such as food, or in the case of cosmetics, applied directly on the skin.

It is important to note that vitamin E exists in eight chemical forms referred to as tocopherols and tocotrienols (11). These molecules are differentiated based on 1) the arrangement and number of methyl groups on the main chromanol ring and 2) the saturation of the carbon side chain (Figure 2).

Although all forms are proven to be antioxidants, they have varying degrees of biological activity and potency due to their slight structural differences. While gamma (γ) and delta (δ) forms are more potent than antioxidants *in vitro* (as they do not have as many methyl groups blocking the OH (hydroxyl group) on the main ring), they are less bioavailable in the human body than alpha (α) tocopherol (12-16). Thus, the α form is considered the most potent and is usually used in supplements and skincare products. Previous studies have confirmed the natural presence of vitamin E in its tocopherol forms in both yellow and black mustard oil (1, 4-5, 9). These tocopherols are recognized as the main contributors to the antioxidant effects provided by the oils. Mustard plants produce Vitamin E as a protective mechanism against UV and oxidative stress induced by the environment (4).

With mustard seed oil being restricted to topical applications in Canada, it is important to further assess its effectiveness in delivering benefits through this route of administration. Although the presence of vitamin E in mustard oil has been confirmed, extensive research has not been done specifically on the tocopherol content of the seed oil from various species. This study aims to examine and quantify vitamin E tocopherols in different species of mustard seed and determine the best source.

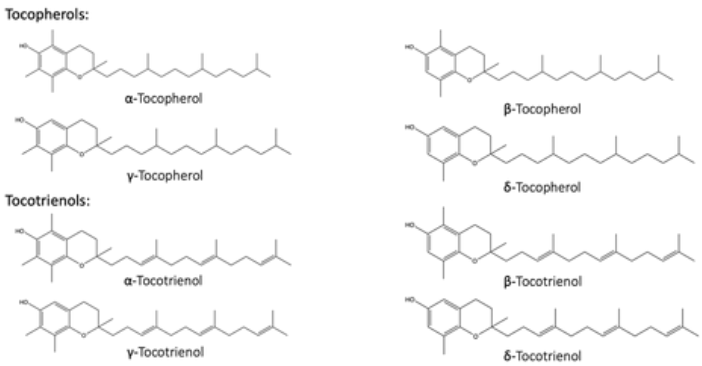


Figure 2. Eight different structural isomers of vitamin E

Methods

Yellow and black mustard seeds from the brand Suraj were obtained from Real Canadian Superstore. Approximately 50 grams of yellow and black mustard seeds were separately weighed out and crushed using a hand-operated seed oil press. Before operation, the hand press was preheated to approximately 40ºC with a heat gun for 10 minutes. 5 mL of oil from each species was kept for testing. 10 µL of black mustard seed oil was then diluted to 1 mL in ethyl acetate solvent and filtered using a chromatography syringe filter. The same was done for 70 µL of yellow mustard seed oil.

1 µL of each of the diluted oil samples was injected into the gas chromatography-mass spectrometry (GC-MS) for analysis. The GC used was the Agilent 6890 with RXI-5SIL MS column from ResTech (30 m in length, 0.25 mm internal diameter, 0.25 micrometer coating - DB-5-MS 25 column) with a Mass Spec 5975C detector and 7683B Series injector autosampler. Helium gas was used as the mobile phase, and a (5%-phenyl)-methylpolysiloxane crosslinked column was used as the stationary phase. 1 µL of vitamin E oil standard containing mixed tocopherols (New Directions Aromatics) was diluted in 1 mL ethyl acetate and 1 µL of this solution was injected into the GC-MS for analysis. Reference mass spectra for the four tocopherol types were obtained from this standard for the four different tocopherol types. An external calibration curve was constructed using 100% pure vitamin E oil from the brand Cliganic as a standard (Figure 3). Compounds were identified using the National Institute of Standards and Technology's mass spectral library.

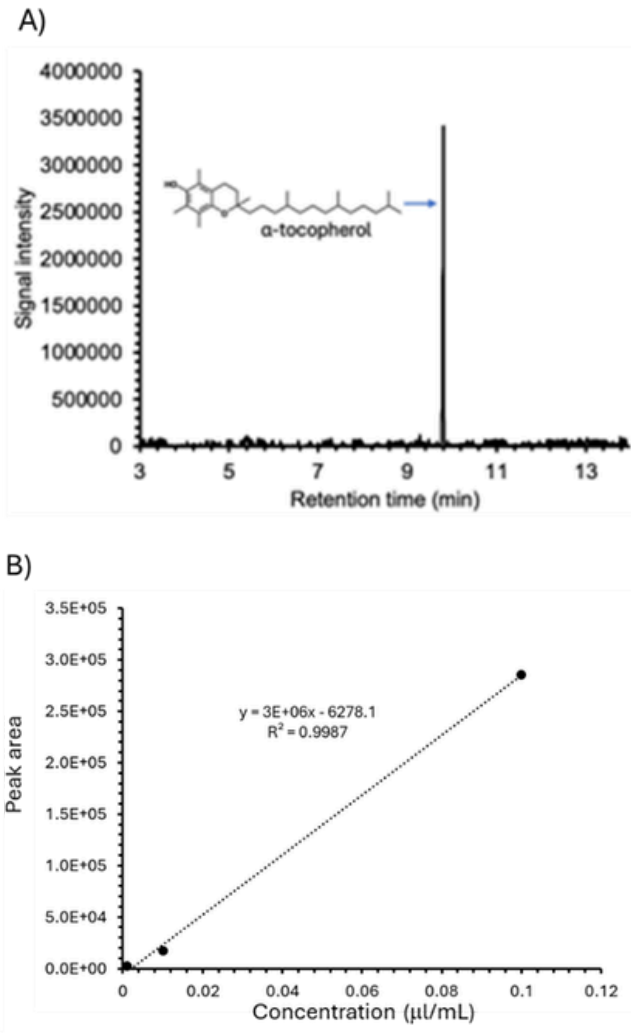


Figure 3. A) Gas chromatogram of pure a-tocopherol standard used for calibration curve (1 µL/mL) and B) Calibration curve generated using pure a-tocopherol standard.

Results and Discussion

Mixed Tocopherols in Vitamin E Oil Standard

Naturally sourced vitamin E oil contains the four types of tocopherols. As tocopherols all have the same structural backbone, the number of methyl groups on the rings serves as their differentiating factor. The varying numbers of these side chains give each tocopherol a slightly different retention time on the gas chromatogram (Figure 4) and a characteristic mass spectrum (Figure 5).

Yellow Mustard Oil

The chromatogram and subsequent compound identifications can be found in Figure 6 and Table 1. The chromatogram identified peaks 1–3 as fatty acids (Table 1). A single γ-tocopherol peak appeared at a retention time of 12.81 minutes. Peaks 4 and 5 were identified as long-chain non-polar metabolites, and peaks 6–8 were identified as sterols (Table 1). The calibration curve yielded a vitamin E concentration of 0.03 µl/mL for the injected yellow mustard sample. These values correspond to a vitamin E concentration of 0.43 µl per mL of pure yellow mustard oil. This concentration is attributed completely to γ-tocopherol.

Black Mustard Oil

The chromatogram and subsequent compound identifications can be found in Figure 7 and Table 2. The chromatogram revealed two forms of vitamin E in the black mustard sample: a peak at 9.48 minutes corresponding to γ-tocopherol and one at 9.88 minutes matched with α-tocopherol. Peaks 1–3 correspond to sterols (Table 2).

The black mustard oil exhibited no detectable fatty acids. While the literature identifies yellow mustard as being more abundant in fatty acids compared to black mustard, black mustard is still confirmed to contain several fatty acids. Their lack of appearance on the chromatogram indicates that they may not have been completely removed from the cell membrane of the seed shells during extraction and mixed with the oil, or they exist in concentrations that are too low for the GC to detect. Some of these compounds may need to be derivatized to be properly vaporized and visualized by a GC, especially at lower concentrations. The calibration curve yielded a total tocopherol concentration of 0.08 µl/mL for the injected black mustard sample; 0.068 µl/mL belonging to γ-tocopherol and 0.012 µl/mL belonging to α-tocopherol. These values correspond to a total vitamin E concentration of 8.0 µl per mL of pure black mustard oil pherol and 1.2 µL of α-tocopherol.

Conclusions

This study successfully utilized GC-MS to analyze and quantify various vitamin E tocopherols in two distinct species of mustard. Results revealed measurable concentrations of γ-tocopherol in both species, as well as α-tocopherol in black mustard alone. Noticeable variations in concentrations between the seed species indicated the possible influence of genetic and environmental factors on tocopherol content. Black mustard's higher overall

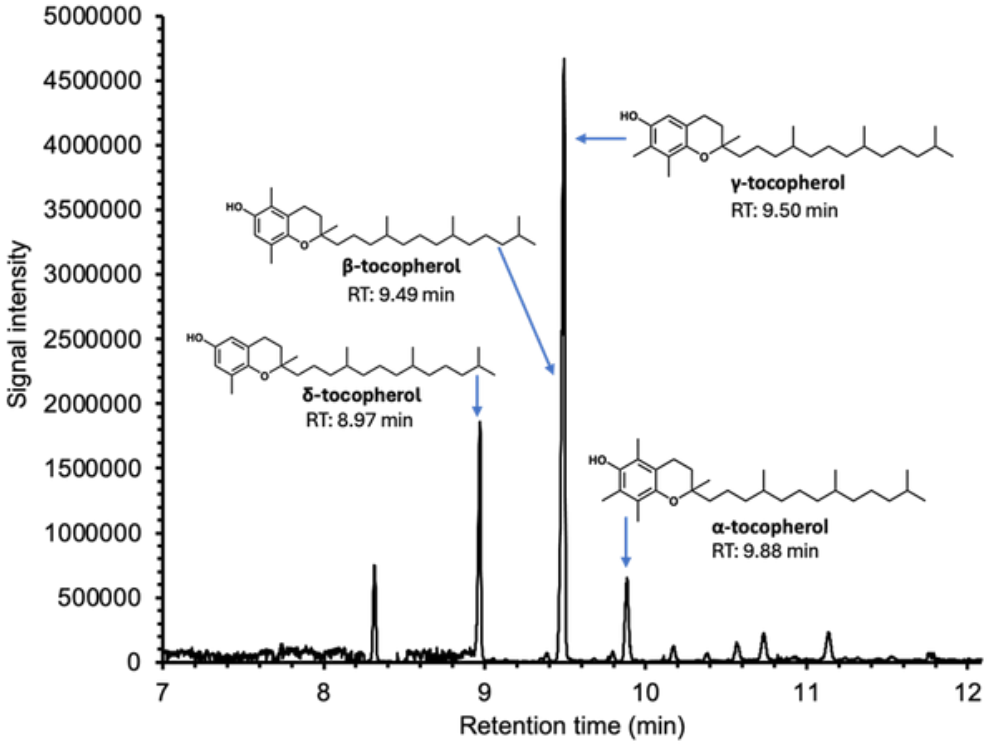


Figure 4. Gas chromatogram of vitamin E oil standards dissolved in ethyl acetate (1 µl/mL) containing all four forms of tocopherol.

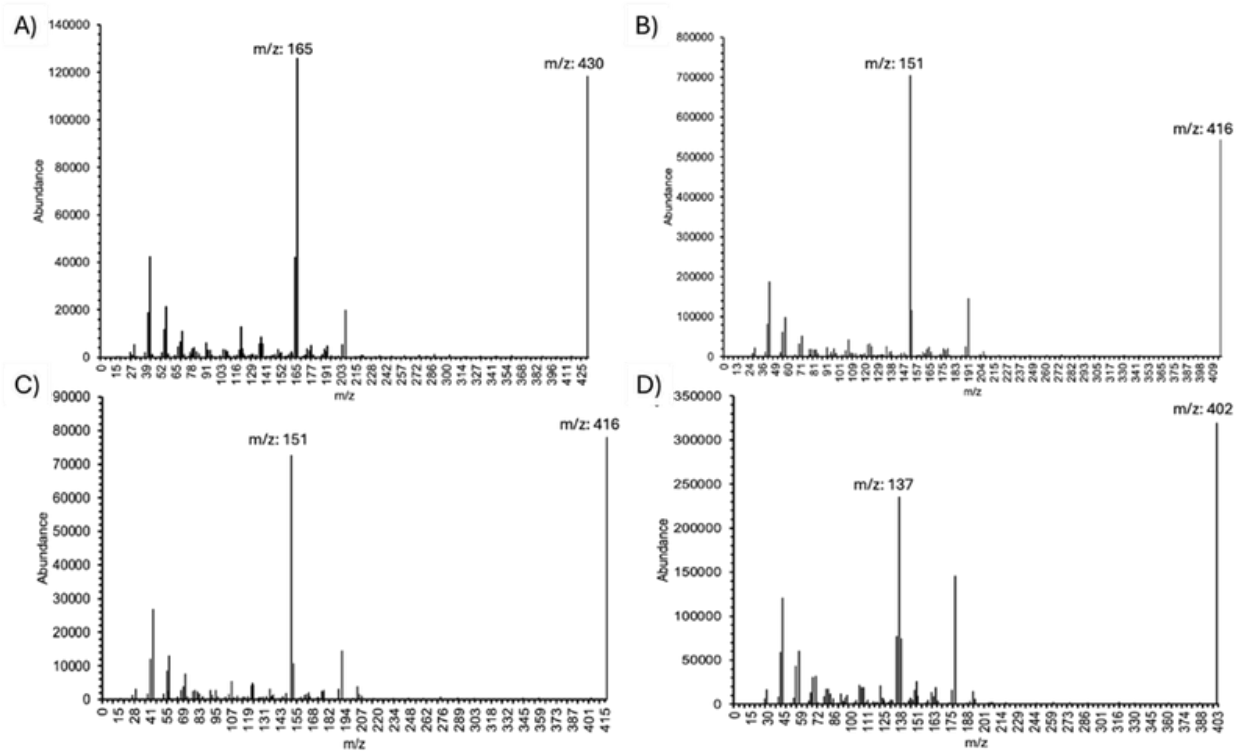


Figure 5. Characteristic mass spectra of A) α-tocopherol B) β-tocopherol C) γ-tocopherol and D) δ-tocopherol.

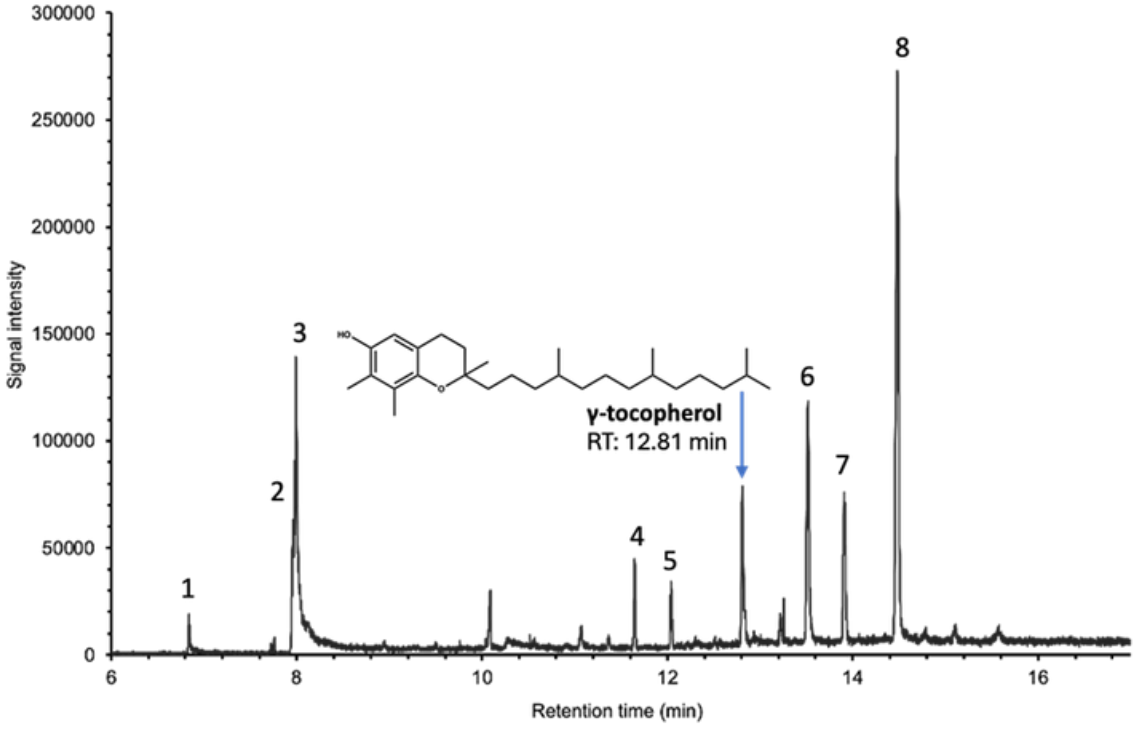
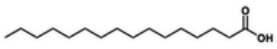
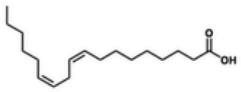
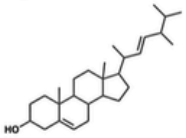
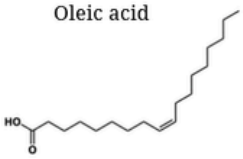
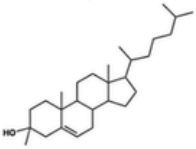
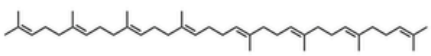
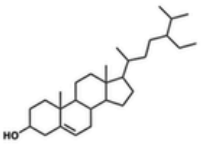


Figure 6. Gas chromatogram of yellow mustard seed oil in ethyl acetate solvent (70 µl/mL).

Table 1. Peak characterization of yellow mustard oil sample

Peak	Characterization	Peak	Characterization
1	Palmitic acid 	5	Tricosane
2	Linoleic acid 	6	Epibrassicasterol 
3	Oleic acid 	7	23-R-Methylcholesterol 
4	Lycopersene 	8	Clionasterol 

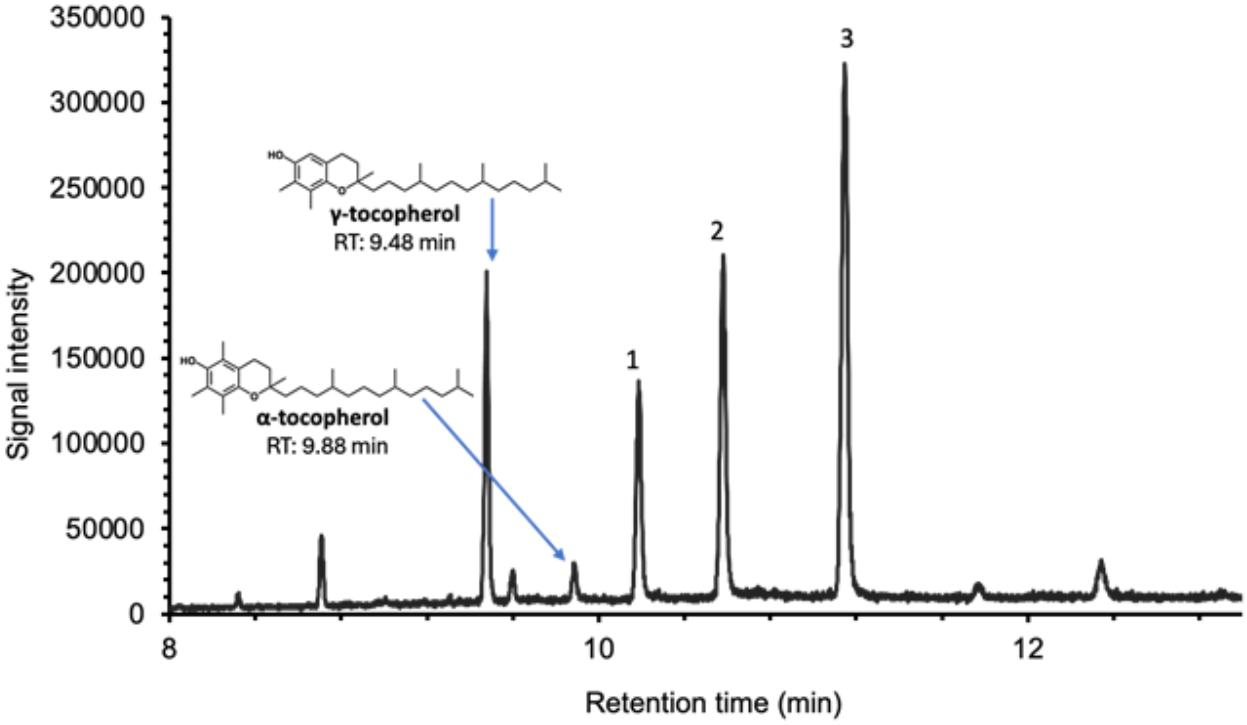
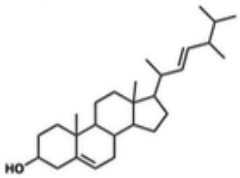
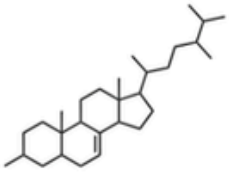
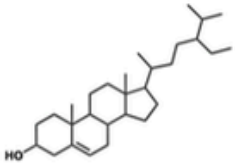


Figure 7. Gas chromatogram of black mustard seed oil in ethyl acetate solvent.

Table 2. Peak characterization of black mustard oil sample

Peak	Characterization
1	Epibrassicasterol 
2	Ergost-7-en-3-ol 
3	Clionasterol 

vitamin E levels indicate it as the better source of antioxidants. The presence of the more biologically active α-tocopherol in black mustard also highlights it as a candidate of interest for future investigations involving topical benefits. This study contributes to a better understanding of mustard oil composition and tocopherol concentrations found in these exact species. As research in this area is limited, further studies are essential for validating reproducibility and optimizing results. Future investments in similar studies could help determine more cost-effective and viable natural sources of beneficial compounds such as vitamin E, unlocking new opportunities for both cosmetic and medicinal applications.

Acknowledgements

The authors thank the JLH Mass Spectrometry Core Facility for facilitating this study.

References

1. Theertha, P.; Prasad, R.; Jyoti, S.; Sawinder, K.; Jaspreet, K.; Mahendra, G.; Harshal, A.; Nidhi, D.; Dipendra Singh, M., Bioactive compounds of mustard, its role in consumer health and in the development of potential functional foods. Current Nutrition & Food Science 19, 950-960 (2023).

2. Das, G.; Tantengco, O. A. G.; Tundis, R.; Robles, J. A. H.; Loizzo, M. R.; Shin, H. S.; Patra, J. K., Glucosinolates and omega-3 fatty acids from mustard seeds: Phytochemistry and pharmacology. Plants 11, 2290 (2022).
3. Grygier, A., Mustard seeds as a bioactive component of food. Food Rev. Int. 39, 4088-4101 (2023).
4. Lietzow, J., Biologically active compounds in mustard seeds: A toxicological perspective. Foods 10, 2089 (2021).
5. Nguyen, T.; Nandasiri, R.; Fadairo, O.; Eskin, N. A. M., Phenolics of mustard seeds: A review on composition, processing effect and their bioactvities. J. Am. Oil Chem. Soc. 101, 5-21 (2024).
6. Lee, H.-W.; Lee, C. G.; Rhee, D.-K.; Um, S. H.; Pyo, S., Sinigrin inhibits production of inflammatory mediators by suppressing nf-κb/mapk pathways or nlrp3 inflammasome activation in macrophages. Int. Immunopharm. 45, 163-173 (2017).
7. Schwarzinger, B.; Feichtinger, M.; Blank-Landeshammer, B.; Weghuber, J.; Schwarzinger, C., Quick determination of erucic acid in mustard oils and seeds. J. Anal. Appl. Pyro. 164, 105523 (2022).
8. Wendlinger, C.; Hammann, S.; Vetter, W., Various concentrations of erucic acid in mustard oil and mustard. Food Chem. 153, 393-397 (2014).
9. Thiele, J. J.; Ekanayake-Mudiyanselage, S., Vitamin e in human skin: Organ-specific physiology and considerations for its use in dermatology. Molec. Aspects Med. 28, 646-667 (2007).
10. Pinto, C. A. S. d. O.; Baby, A. R.; Velasco, M. V. R.; Batello Freire, T.; Miliani Martinez, R.; Azevedo Martins, T. E. A., Vitamin e in human skin: Functionality and topical products. In Vitamin E in Health and Disease - Interactions, Diseases and Health Aspects, Erkekoğlu, P.; Scherer Santos, J., Eds. IntechOpen: Rijeka, (2021).
11. Keen, M. A.; Hassan, I., Vitamin E in dermatology. Indian Derm. Online J. 7, 311-5 (2016).
12. Kamal-Eldin, A, and L A Appelqvist. The chemistry and antioxidant properties of tocopherols and tocotrienols. Lipids31, 671-701 (1996).
13. Ohkatsu, Yasukazu, Tetsuto Kajiyama, and Yuji Arai. Antioxidant Activities of Tocopherols. Polymer Degradation and Stability, 72, 303–311 (2001).
14. Mathur, Pankaj et al. Tocopherols in the Prevention and Treatment of Atherosclerosis and Related Cardiovascular Disease. Clinical Cardiology 38, 570-6 (2015).
15. Tucker, J M, and D M Townsend. Alpha-tocopherol: roles in prevention and therapy of human disease. Biomedicine & Pharmacotherapy 59, 380-7 (2005).
16. Pahrudin Arrozi, Aslina et al. Comparative Effects of Alpha- and Gamma-Tocopherol on Mitochondrial Functions in Alzheimer's Disease In Vitro Model. Scientific Reports 10, 8962 (2020).

Original Research

Probing Metal Ion Interactions in Stacked Polycyclic Aromatic Hydrocarbons as a Molecular Model for Superconductivity

Exploration des interactions métallique-ion dans les hydrocarbures aromatiques polycycliques empilés comme modèle moléculaire de la supraconductivité

Victor Lee[1], Olivia Chen[2], Natasha Saltarelli[1], Paul Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
2. Merivale High School, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

Metal ion intercalated graphene has been shown to display superconductive properties. Most metals chosen for intercalation are alkali and alkaline earth metals. This study explores the intercalation of graphene with nickel as a more environmentally-friendly alternative. As a first step, the matrix-assisted laser desorption ionization was used to study the structure of the dimer of coronene intercalated with a Ni atom. The ions were formed by laser desorption from a target containing a mixture of coronene and Ni salt, the dimer ion with m/z 657 was selected, and the collision-induced dissociation mass spectrum was obtained with the LIFT technology on our Bruker MALDI-TOF/TOF. The results showed that the dimer dissociated by H-atom transfer from a coronene molecule to Ni. This was supported by density functional calculations of the reaction energetics. The results suggest that the laser desorption process leads to a cluster ion better described as an ionic complex between CorNiH and Cor-H, rather than a formal [$(Cor)_2Ni$] ion.

———

Il a été démontré que le graphène intercalé aux ions métalliques présente des propriétés supraconductrices. La plupart des métaux choisis pour l'intercalation sont des métaux alcalins et alcalino-terreux. Cette étude explore l'intercalation du graphène avec le nickel comme alternative plus respectueuse de l'environnement. En première étape, l'ionisation laser assistée par matrice a été utilisée pour étudier la structure du dimère de coronène intercalé avec un atome de Ni. Les ions ont été formés par désorption laser à partir d'une cible contenant un mélange de coronène et de sel de Ni, l'ion dimère avec m/z 657 a été sélectionné, et le spectre de masse de dissociation induit par collision a été obtenu avec la technologie LIFT sur notre Bruker MALDI-TOF/TOF. Les résultats ont montré que le dimère se dissocie par transfert d'atome d'H d'une molécule coronène vers Ni. Cela était soutenu par des calculs de la fonction de densité de l'énergétique de réaction. Les résultats suggèrent que le processus de désorption laser conduit à un ion cluster mieux décrit comme un complexe ionique entre CorNiH et Cor-H, plutôt qu'un ion formel [$(Cor)_2Ni$].

Keywords: superconductivity; graphene intercalation; coronene; polycyclic aromatic hydrocarbons; nickel intercalation; MALDI-TOF/TOF; collision-induced dissociation; density functional theory; graphite intercalation compounds; host–guest interactions

Introduction

Superconductivity is a quantum mechanical phenomenon that allows materials to conduct electricity without resistance and expel magnetic fields when cooled to below a specific critical temperature (T_c) (1). It has been found that the superconducting state was not only related to temperature but also to the current in the superconductor and the applied external magnetic field. These three effects have limited the practical application of many superconducting materials (2). There is an ongoing quest for higher-temperature superconductors from practical materials that challenge these limitations. Recent advances include carbon-based superconductors that have been developed through intercalation chemistry and carrier doping techniques (3). Intercalation refers to the insertion of guest species (atoms, ions, molecules) into the spaces between layers of a host material framework without majorly disrupting its molecular framework. This occurs in highly anisotropic structures where intraplanar binding forces are larger than interplanar binding forces (4). Doping refers to the addition or substitution of impurity guest species into a material, altering its lattice structure in the process. These techniques provide the host molecule a means for varying several physical properties over wide ranges, depending on the type and degree of intercalation or doping.

Graphene (Figure 1a) can present superconductivity upon metal ion intercalation and doping (5). Intercalation is particularly interesting as it presents the opportunity to tune material electronic properties (6). Graphite intercalation compounds (GICs) are materials that are formed by inserting guest species between graphitic layers without disrupting the planar hexagonal carbon framework. In 1965, superconductivity was observed in

K-intercalated graphite (KC_8) (7). Interestingly, no superconductivity was observed in KC_{24} or KC_{36}, which suggests that the alternating arrangement of metal and graphene layers is critical for superconductivity to be observed in GICs. This highlights the fundamental importance of the structure-property relationship of these materials.

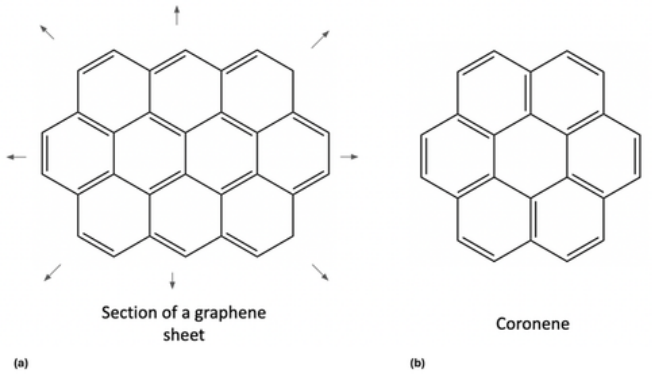


Figure 1. (a) Structure of a section of a graphene sheet, and (b) Structure of coronene

After this discovery, Ca intercalated graphite was discovered, with a T_c as high as 11.5K in CaC_6 (8). The tunability of GICs, through control over staging (number of graphene layers between intercalant layers), type of intercalation, and concentration of intercalant, makes them ideal platforms for probing electronic effects in layered carbon systems.

Since the discovery of graphene-based superconductors, experimental measurements of their binding patterns have relied on assumptions and often yielded contradictory results (9). While the average structures of the bulk material have been studied, investigating the intercalant binding patterns at specific sites remains challenging. Metal binding to graphitic sheets can occur in various forms, including direct intercalation between layers, surface functionalization, or edge binding between sheets. These distinct binding patterns can significantly influence the material's electronic properties. Previous studies of GICs have employed techniques such as X-ray diffraction and electron diffraction; however, these methods have limited spatial resolution, which affects characterization at the atomic level (10). This creates challenges in studying GIC's binding systems and the properties that arise from them. A potential solution to the study of graphene's molecular properties can be found using polycyclic aromatic hydrocarbons (PAHs) as molecular models of graphene, since they can be viewed as fragments of graphene sheets. Their extensive π-molecular orbitals allow electrons to become delocalized, leading to characteristics of conductivity (11). Although it is very different in size than graphene, heteroatom-intercalated PAHs, such as coronene (Figure 1b) could yield valuable information on the graphene system (12,13).

The objective of the study is to determine the type of metal binding that can occur on coronene and in coronene dimers (14,15), using matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) mass spectrometry and theory. While Li has been employed extensively in GICs due to the mobility of Li in the lattice, it is a rare metal whose extraction has significant environmental cost. Thus, we decided to investigate a more common metal, nickel, that also has economic advantages for the Canadian economy, to see if it has desirable intercalation properties.

Methods

MALDI-TOF Mass Spectrometry

All chemicals were obtained from MilliporeSigma (Oakville, ON) and used without further purification. For MALDI-TOF, 3 mg of nickel chloride salt was dissolved in 1 ml of a solvent comprised of 50:50 (v/v) acetonitrile and water. 10 mg of coronene was dissolved separately in 1 ml of chloroform. 25µl of each solution was deposited into a glass vial and sonicated until homogeneous. The dried-droplet spotting method was employed, which involves depositing 1 µl spots of the prepared solution onto the surface wells of the Bruker MTP 384 ground steel plates and drying under ambient conditions. Mass spectra were acquired using a Bruker ultrafleXtreme MALDI-TOF/TOF reflecting mass spectrometer in positive-ion mode, covering a mass range of m/z 0-1200. Laser desorption was achieved via a frequency-tripled Nd:YAG laser at a wavelength of 337 nm and a pulse rate of 1000 Hz. Tandem mass spectrometry was performed using the LIFT functionality on the instrument. High-resolution MS peaks corresponding to the precursor ion of interest were selected with a PCIS (precursor ion selection) window of 2 Da.

Computational Procedures

Geometry optimization and vibrational frequency calculations employing the M06/6-31+G(d) level of theory were performed with the GAUSSIAN 16 suite of programs (16).

Results and Discussion

Coronene-nickel clusters were observed using MALDI-TOF. Figure 2 showcases the coronene-nickel monomer and dimer clusters at their respective m/z range. The clusters were confirmed using isotope distribution models, which accounted for the natural abundances of carbon and nickel isotopes.

The dissociation of the coronene-nickel complex ion with m/z 358 showed one fragment ion at m/z 299, corresponding to a coronene fragment with hydrogen loss ($[Cor\text{-}H]^+$), Figure 3a. The only possible pathway that explains this peak is the loss of a neutral 59 Da fragment corresponding to NiH. The hydrogen transfer was also seen for the sodium cluster on ESI-MS. The absence of further fragmentation products implies that the coronene skeleton is intact, reinforcing its stability under LIFT conditions and the lack of sequential fragmentation events. The LIFT mass spectrum for the $(Cor)_2Ni^+$, m/z 658, is shown in Figure 4b. There are three notable fragment peaks at m/z 299, 357, and 359. The peak at m/z 308 may be the result of a C_4H_2 loss to form a stable corannulene intermediate ($C_{20}H_{10}^+$) with Ni coordination. However, this is highly

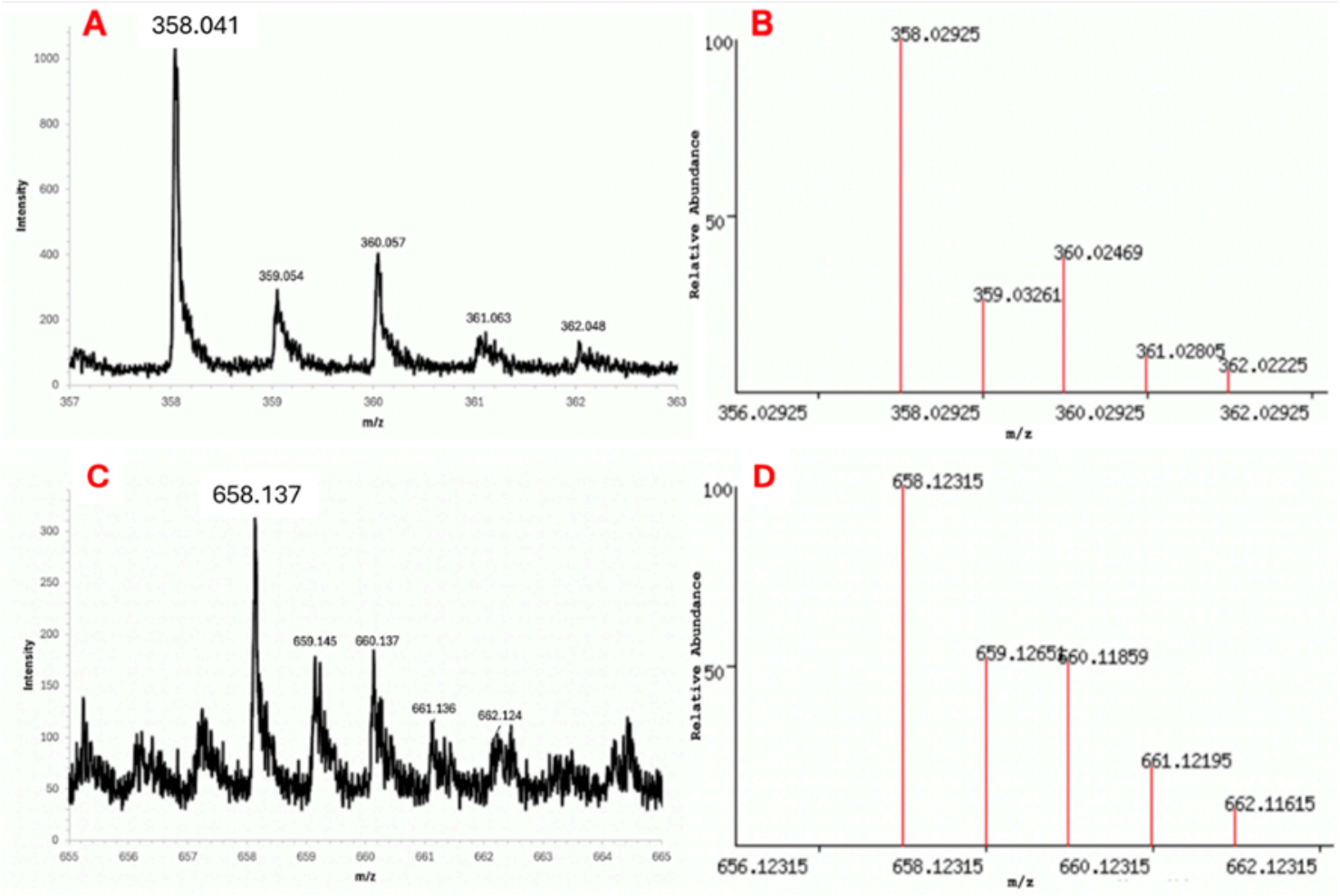


Figure 2. A) MALDI-TOF mass spectrum of coronene-nickel monomer cluster $[C_{24}H_{12}Ni]^+$ B) isotope distribution model of the monomer, C) Mass spectrum of the coronene-nickel dimer cluster $[(C_{24}H_{12})_2Ni]^+$, and D) isotope distribution model of the dimer cluster.

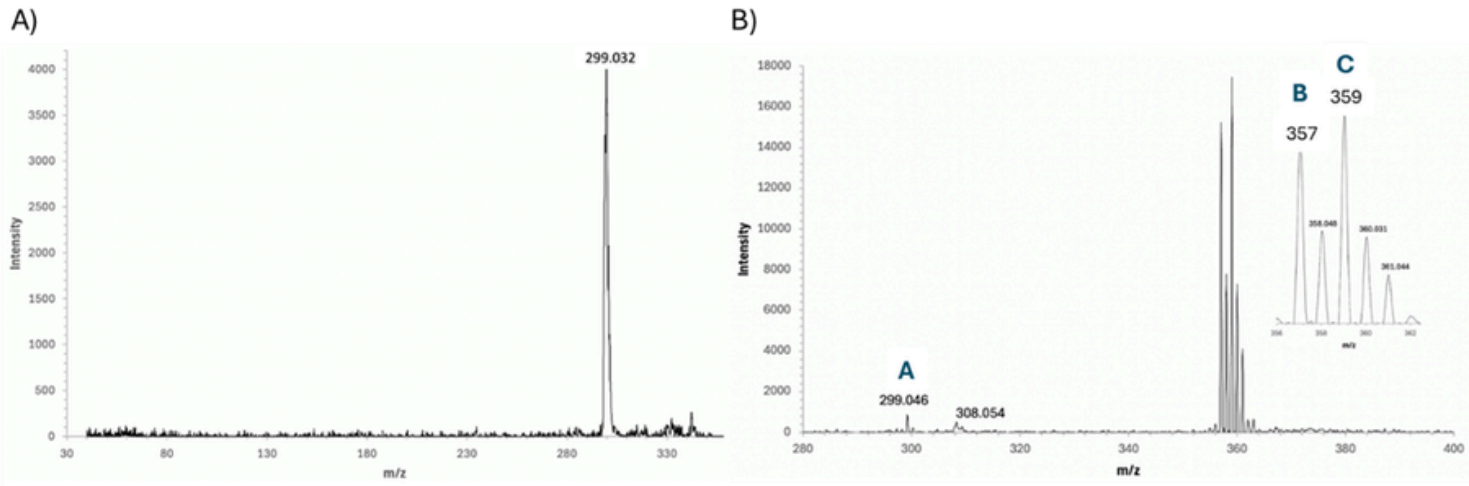


Figure 3. MALDI LIFT TOF/TOF mass spectrum A) [(Cor)Ni]+ and B) of the dimer cluster ($[(Cor)_2Ni]^+$), with an expansion of the m/z 357-361 region in the inset. Blue labels are referred to in the text.

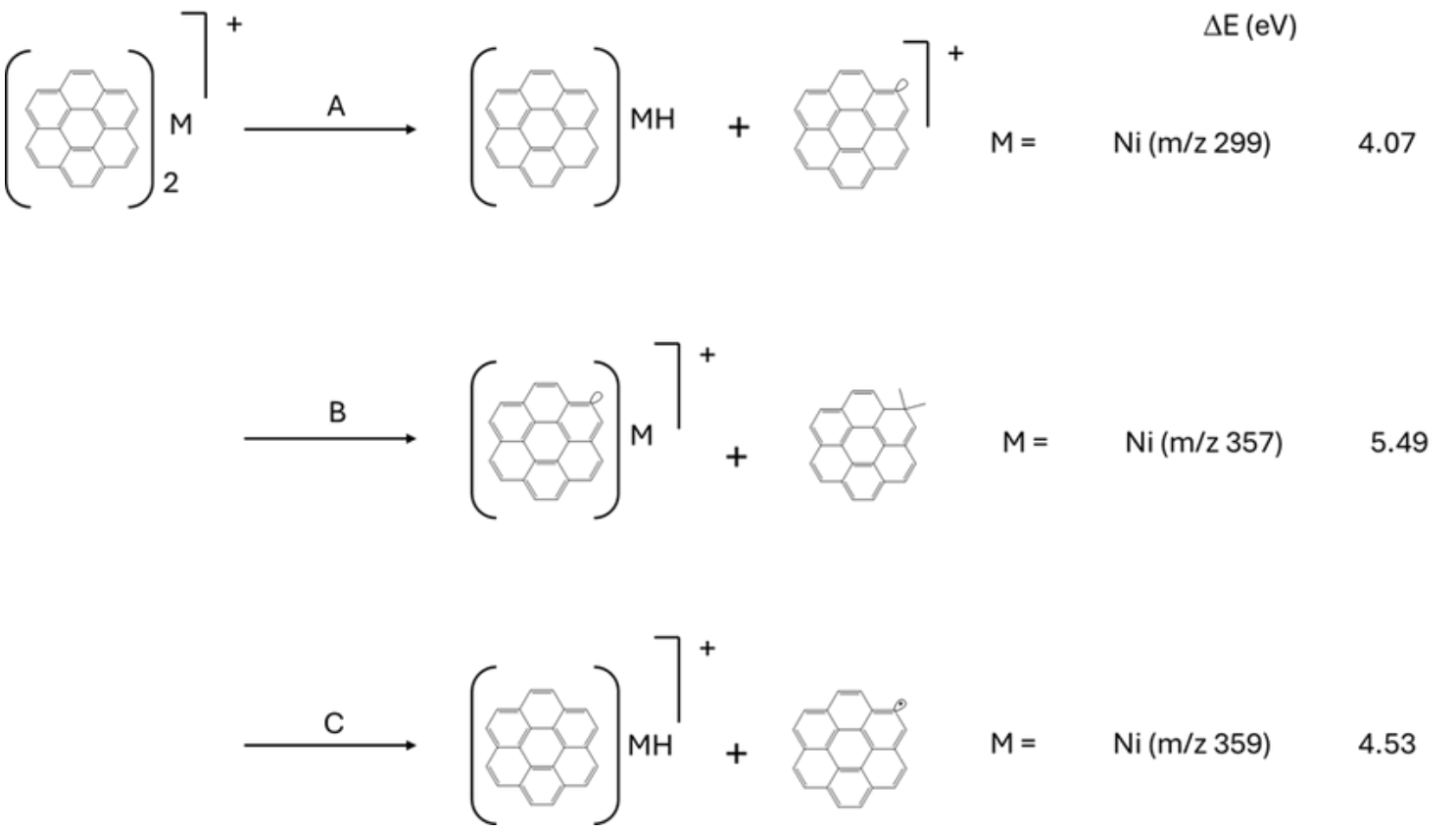


Figure 4. Summary of the experimentally observed dissociation reactions of the $[(Cor)_2M]^+$ system, where M = Na and/or Ni. Calculated M06/6-31+G(d) energies (relative to the respective dimer ions) are shown for the two metals.

speculative as C_4H_2 loss is an uncommon loss channel in PAH ions (17).

The isotope distribution of the series of peaks from m/z 357-361 does not match the isotope distribution of the obvious $C_{24}H_{11}Ni^+$ fragment; the intensity of the peak at 359 is too high. Therefore, it is likely that two fragments contribute to this series of peaks: $[C_{24}H_{11}Ni]^+$ and $[C_{24}H_{12}NiH]^+$. This suggests the presence of three competing dissociation pathways labelled A-C in Figure 4b. Reaction A has the Ni abstracting an H atom from a coronene ring before the complex splits to form m/z 299. Reaction C is the complementary reaction in which the ion charge is retained on the coronene/NiH fragment. Reaction B is the splitting of the dimer complex with H-atom transfer to the departing neutral coronene ($C_{24}H_{13}$). The observed reactions are summarized in Figure 4.

Computational Study

The most stable dimer ion was found to be the sandwich structure with the Ni atom between two coronene rings (Figure 5). For $[(Cor)_2Ni]^+$,Mulliken partial charges show the charge to be split between the organic ring and Ni atom due to the similarity in the two ionization energies (IE), 7.68 eV (Ni) (18) vs 7.29 eV for coronene (19). The M06/6-31+G(d) relative product energies are summarized in Figure 4. Structures for each species in Figure 4 are shown in Figure 5.

Interestingly, the calculated binding energy for the dimer (relative to $[CorNi]^+$ + Cor) is a scant 0.1 eV. This suggests that these ions are not made during the laser desorption stage of the experiment, as they would not survive to the LIFT process for CID, and there is definitely no m/z 300 or 358 peak in the LIFT spectra. Insight into the structure of these ions comes from the observed mass spectra.

For the ions formed in pathways B ($[(Cor\text{-}H)Ni]^+$) and C ($[CorNiH]^+$)

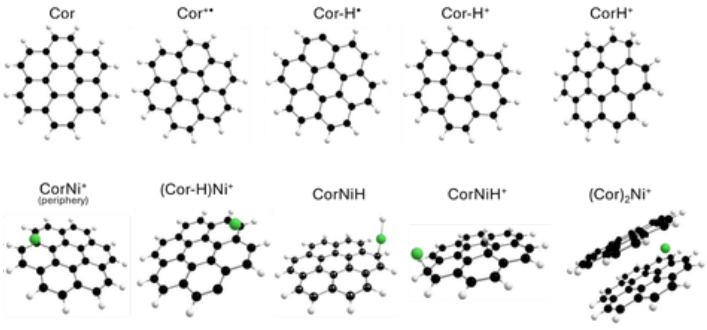


Figure 5. Structures of species in Figure 4, calculated at the M06/6-31+G(d) level of theory.

of the Ni system, the Mulliken analysis reveals that the positive charge is primarily localized on the nickel atom. Upon metal coordination, nickel is the electron-deficient atom that preferentially carries the positive charge in these systems. Through this interpretation, pathway A leading to the metal-absent ion is energetically higher because there is an unfavourable charge transfer from the nickel to the coronene. A similar conclusion was reached by Pozniak and Dunbar, as a similar unfavourable charge transfer process was thought to compete with coronene-metal ion cluster formation for the reaction of the bare metal ion onto the monomer cluster (14).

The formation of a Ni–H bond in pathway C leads to a slightly delocalized positive charge on the nickel atom. The hydride is an electron-rich ligand that can donate electron density towards the metal center, leading to a lower partial positive charge. In pathway B, the partial charge of the metal center is higher due to the absence of any electron-donating ligands. This interpretation is supported by the Mulliken charge data, which shows a higher partial charge on the nickel atom in pathway B (+0.92) compared to pathway C (+0.90). The lower partial positive charge in the ion produced from pathway C due to the hydride ligand may be a factor contributing to its higher reaction energy. It is also possible that hydrogen transfer between the coronene complexes in pathway B presents a lower energy transition state.
The results suggest that the laser desorption process leads to a cluster ion better described as an ionic complex between CorNiH and Cor-H, rather than a formal $[(Cor)_2Ni]$ ion.

Conclusion

H-abstraction by the metal ion is a common observation in the CID of $[(Cor)_2Ni]^+$. Density functional theory calculations of the energetics support the observation of the major kinetic pathways. The formation of the kinetic product ions strongly suggests that the nickel atom is intercalated between the coronene molecules, as the formation of these ions is energetically improbable in the dissociation of other conformations (surface or edge bound). This conclusion is supported by Mulliken charge analysis, which reveals that the positive charge is preferentially located on the nickel atoms in both the parent and fragment ions. In addition, the calculated trivial binding energy of the $[(Cor)_2Ni]^+$ ion indicates that laser desorption leads to a cluster ion better described as an ionic complex between CorNiH and Cor-H.

The investigation of other types of intercalant species is also encouraged, as it will expand the scope of this project and shed light on the fundamental host-guest interactions and their underlying mechanisms. Upon ion intercalation, different host-guest interactions usually induce different intercalation chemistries. New approaches to intercalation include binary or ternary element intercalation, multivalent ion intercalation, and complexed-ion intercalation (20). Further research will deepen the understanding of intercalation chemistry and will guide the exploration of new guests and hosts.

Acknowledgement

PMM thanks the Natural Sciences and Engineering Research Council of Canada for continuing financial support, and the authors thank the JLH Mass Spectrometry Core Facility of the University of Ottawa for the use of the MALDI-TOF instrument.

References

1. Rahman, Md. A., Rahaman, Md. Z, A Review on High- T_c Superconductors and Their Principle Applications. J. Adv. Phys., 4, 87–100 (2015). DOI:10.1166/jap.2015.1175
2. Yao, C., Ma, Y, Superconducting materials: Challenges and opportunities for large-scale applications. IScience, 24, 102541 (2021). doi.org/10.1016/j.isci.2021.102541
3. Kubozono, Y., Eguchi, R., Goto, H., Hamao, S., Kambe, T., Terao, T., Nishiyama, S., Zheng, L., Miao, X., & Okamoto, H. Recent progress on carbon-based superconductors. J. Phys.: Condens. Matt. 28, 334001 (2016). DOI: 10.1088/0953-8984/28/33/334001
4. Dresselhaus, M. S., Dresselhaus, G. Intercalation compounds of graphite. Advances in Physics, 51, 1–186 (2002). DOI:10.1080/00018730110113644
5. Zamani, M., Abbasnejad, M. Optical properties of superconductor-graphene-superconductor junction. Physica C: Superconductivity and Its Applications, 554, 19–26. (2018). doi.org/10.1016/j.physc.2018.09.001
6. Huynh, T. M. D., Hung, G.-S., Gumbs, G., Tran, N. T. T. Fundamental properties of alkali-intercalated bilayer graphene nanoribbons. Phys. Chem. Chem. Phys., 25, 18284–18296 (2023). doi.org/10.1039/D3CP02266H
7. Hannay, N. B., Geballe, T. H., Matthias, B. T., Andres, K., Schmidt, P., MacNair, D. Superconductivity in Graphitic Compounds. Phys. Rev. Lett., 14, 225–226 (1965). DOI: https://doi.org/10.1103/PhysRevLett.14.225
8. Gauzzi, A., Takashima, S., Takeshita, N., Terakura, C., Takagi, H., Emery, N., Hérold, C., Lagrange, P., Loupias, G. Enhancement of Superconductivity and Evidence of Structural Instability in Intercalated Graphite under High Pressure. Phys. Rev. Lett., 98, 067002 (2007). DOI: https://doi.org/10.1103/PhysRevLett.98.067002.
9. Feng, C., Lin, C. S., Fan, W., Zhang, R. Q., van Hove, M. A. Stacking of polycyclic aromatic hydrocarbons as prototype for graphene multilayers, studied using density functional theory augmented with a dispersion term. J. Chem. Phys.131, 194702 (2009). DOI: 10.1063/1.3251785.
10. Lin, Y.-C., Matsumoto, R., Liu, Q., Solís-Fernández, P., Siao, M.-D., Chiu, P.-W., Ago, H., Suenaga, K. Alkali metal bilayer intercalation in graphene. Nature Commun., 15, 425. (2024). https://doi.org/10.1038/s41467-023-44602-3.
11. Wang, X. F., Liu, R. H., Gui, Z., Xie, Y. L., Yan, Y. J., Ying, J. J., Luo, X. G., Chen, X. H. Superconductivity at 5 K in alkali-metal-doped phenanthrene. Nature Commun., 2, 507 (2011). https://doi.org/10.1038/ncomms1513.
12. Wang, X.-Y., Yao, X., Müllen, K. Polycyclic aromatic hydrocarbons in the graphene era. Science China Chemistry, 62, 1099–1144 (2019). https://doi.org/10.1007/s11426-019-9491-2.

13. Cristadoro, A., Räder, H. J., Müllen, K. Clustering of polycyclic aromatic hydrocarbons in matrix-assisted laser desorption/ionization and laser desorption mass spectrometry. Rapid Commun. Mass Spectrom., 21, 2621–2628 (2007). DOI: 10.1002/rcm.3134
14. Pozniak, B. P., Dunbar, R. C. Monomer and Dimer Complexes of Coronene with Atomic Ions. J. Am. Chem. Soc., 119, 10439–10445 (1997). doi.org/10.1021/ja9716259
15. Dunbar, R. C. Binding of Transition-Metal Ions to Curved π Surfaces: Corannulene and Coronene. J. Phys. Chem. A, 106, 9809–9819. (2002). doi.org/10.1021/jp020313b
16. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 rev. C.01, Wallingford, CT, (2016).
17. West, B., Rodriguez Castillo, S., Sit, A., Mohamad, S., Lowe, B., Joblin, C., Bodi, A., Mayer, P. M. Unimolecular reaction energies for polycyclic aromatic hydrocarbon ions. Phys. Chem. Chem. Phys., 20, 7195–7205 (2018). doi.org/10.1039/C7CP07369K
18. NIST Chemistry Webbook, NIST Standard Reference Database Number 69. National Institute of Standards and Technology: Gaithersburg, MD (2023).
19. Schröder, D., Loos, J., Schwarz, H., Thissen, R., Preda, D. V., Scott, L. T., Caraiman, D., Frach, M. V., Böhme, D. K. Single and Double Ionization of Corannulene and Coronene. Helvetica Chimica Acta, 84, 1625–1634 (2001). DOI:10.1002/1522-2675(20010613)84:6<1625::AID-HLCA1625>3.0.CO;2-0
20. Li, Y., Lu, Y., Adelhelm, P., Titirici, M.-M., Hu, Y.-S. Intercalation chemistry of graphite: alkali metal ions and beyond. Chem. Soc. Rev., 48, 4655–4687 (2019). doi.org/10.1039/C9CS00162J.

Sand Ginger versus Real Ginger: Investigating the composition of *Kaempferia Galanga*

Gingembre de sable vs gingembre véritable : étude de la composition de *Kaempferia galanga*

Sum Ki Kelsie Ling[1], Sharon Barden[1], Paul M Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

The aromatic rhizome *Kaempferia Galanga*, also known as sand ginger, has traditionally been used to relieve cough, inflammation and high blood pressure. This study aims to determine an optimal extraction method and ascertain the composition of commercially available *Kaempferia Galanga* powder. The extraction efficiencies for acetonitrile maceration extraction, microwave hydrodistillation, and supercritical-CO_2 (sc-CO_2) extraction are compared, with a focus on the target bioactive compounds ethyl cinnamate and ethyl p-methoxycinnamate. The results showed that the main essential oil components ethyl cinnamate, n-pentadecane, and ethyl p-methoxy cinnamate are present in the extracts of store-bought sand ginger powder. Acetonitrile maceration and sc-CO_2 extractions have similar composition, with the three main compounds extracted alongside long-chain fatty acids and terpenoids such as cyperene and germacrene–D. On the other hand, microwave hydrodistillation, followed by solid-phase extraction, produced a profile containing mainly ethyl cinnamate and ethyl p-methoxycinnamate with minimal co-extracted impurities. Processing 500 ml of hydrosol through solid-phase extraction yielded 48.3 mg of extract containing the two main bioactive compounds. Final derived concentrations for ethyl cinnamate, n-pentadecane, and ethyl p-methoxy cinnamate were 0.43-0.45, 0.24-0.26, and 8.7-8.9 mg/g (acetonitrile extract); 7.1, nil, and 16.3–16.4 mg/L (hydrosol); and 2.9%, 1.5%, and 92.9% (sc-CO_2).

———

Le rhizome aromatique *Kaempferia galanga*, également connu sous le nom de gingembre de sable, est traditionnellement utilisé pour soulager la toux, l'inflammation et l'hypertension. Cette étude vise à déterminer une méthode d'extraction optimale et à établir la composition de la poudre de *Kaempferia galanga* disponible dans le commerce. Les efficacités d'extraction par macération à l'acétonitrile, hydrodistillation assistée par micro-ondes et au CO_2 supercritique (sc-CO_2) sont comparées, en mettant l'accent sur les composés bioactifs cibles : le cinnamate d'éthyle et le p-méthoxycinnamate d'éthyle. Les résultats montrent que les principaux constituants de l'huile essentielle — le cinnamate d'éthyle, le n-pentadécane et le p-méthoxycinnamate d'éthyle — sont présents dans les extraits de poudre de gingembre de sable achetée en magasin. Les extractions par macération à l'acétonitrile et par sc-CO_2 présentent une composition similaire, avec ces trois composés principaux extraits en plus d'acides gras à longue chaîne et de terpènoïdes tels que le cyperène et le germacrène-D.

Keywords: Kaempferia galanga; sand ginger; GC-MS; ethyl cinnamate; ethyl p-methoxycinnamate; supercritical CO_2 extraction; microwave hydrodistillation; phytochemical analysis; bioactive compounds; medicinal plants

Introduction

Kaempferia Galanga, commonly known as sand ginger (沙姜), is a frequently used spice for traditional cuisines and for medicinal purposes to treat ailments such as inflammation, cough, and high blood pressure in South-East Asia. However, *K. Galanga* is not botanically related to common ginger, *Zingiber officinale*. Although they belong in the same family, Zingiberaceae, they differ in genus and sand ginger lacks the distinct volatile compounds geraniol, borneol, terpineol, and zingiberene (1). Morphologically, sand ginger has a small, rounded appearance in reddish-brown colour, while common ginger has a long-branched structure with a yellowish-brown colour as observed in Figure 1.

Literature has demonstrated that *K. Galanga* contains 19 identified compounds as determined by gas chromatography-mass spectrometry (GC-MS), (2) with ethyl cinnamate, ethyl p-methoxycinnamate, and n-pentadecane making up the highest abundance. Ethyl cinnamate is commonly used as a fragrance and food additive. A recent study on ethyl cinnamate proves its ability to block tumor growth by directly interfering with VEGF (Vascular Endothelial Growth Factor)/VEGFR2 signaling (3). On the other hand, ethyl p-methoxycinnamate is shown to have the ability to inhibit the activity of the cancer survival protein, transcription factor NF-κB (4). Thus, when used in combination with paclitaxel, which itself induces NFκB activation, it restores and enhances the chemotherapy compound's interference with the growth and division of cancer cells (4).

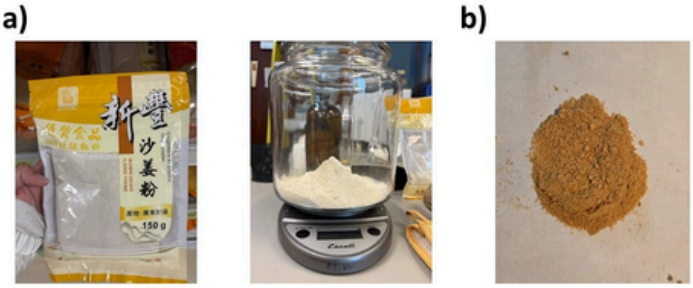


Figure 1. a) *Kaempferia Galanga* (sand ginger) vs b) *Zingiber officinale* (ginger).

This study seeks to ascertain the volatile composition of store-bought *K. Galanga* powder. Additionally, maceration extraction, microwave hydrodistillation, and supercritical-CO_2 (sc-CO_2) extraction will be utilized to identify the optimal extraction method for the ethyl cinnamate and ethyl p-methoxy cinnamate for future applications.

Methods

Acetonitrile Maceration Extract

Commercially acquired *K. Galanga* powder was sourced from Guangdong, China. Approximately one g of the sample was extracted with acetonitrile using the cold maceration method. After 24 hours, the solution was filtered and used for GC-MS analysis.

Supercritical-CO_2 Extraction

Approximately 10 g of *K. Galanga* powder was placed in the sample holder of an SFT-250 SFE System from Supercritical Fluid Technologies Inc. operating at 300 bar and 45 °C. It was subjected to a five-minute soak, followed by five minutes of extraction and collection. This cycle was repeated at least six times for a total of one hour.

Microwave Hydrodistillation

Approximately 108 g of *K. Galanga* powder was evenly placed along the inner walls of a large microwave-safe jar with a collection beaker positioned in the center. The sample was then hydrated with 300 ml of boiling water, and the opening of the jar was covered with an ice cone of distilled water. This setup was then placed into the 1,000-Watt microwave for seven minutes, and the heating cycle was repeated six times. Rising vapour in the container condenses on the ice cone and subsequently drips into the collection beaker. The collected hydrosol was passed over a C18 solid-phase extraction (SPE) cartridge, and the organics were eluted with one:one (v/v) acetonitrile: methanol for GC-MS analysis.

Instrument Method

One µl of sample was injected into the Agilent 6890 GC with mass spectrometer detector (MSD). The column was a 30 m Agilent 19091J-433 HP-5 column, with an internal diameter of 250 mm, a stationary phase thickness of 0.25 mm, He carrier gas with a flow rate of 30 cm/s. The oven temperature was programmed from 80 °C for two minutes, then increased by 15 °C/min to 300 °C. Chemical identification was carried out by comparing the mass spectra of each chromatographic peak with the National Institute of Standards and Technology (NIST) database (5). The hydrosol and acetonitrile maceration extract results were quantified using an external calibration curve of cinnamaldehyde standard, Figure 2.

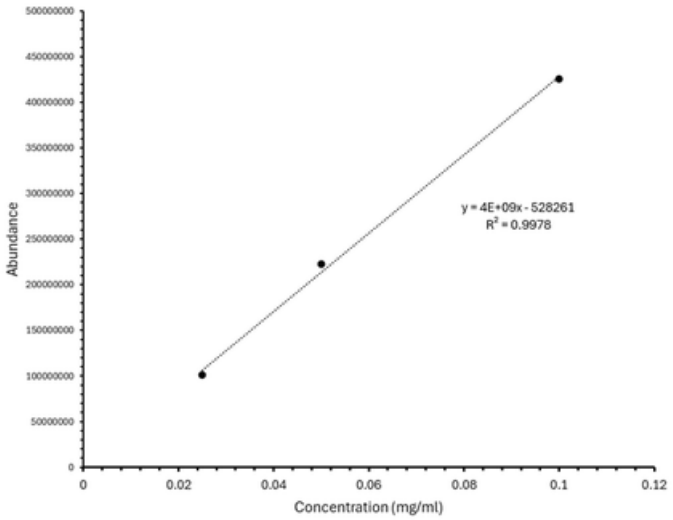


Figure 2. External calibration curve of cinnamaldehyde standard.

Results and Discussion

The three major compounds of *K. Galanga* essential oil: ethyl cinnamate, n-pentadecane, and ethyl p-methoxy cinnamate, were present in the samples as expected; however, other terpenoids and terpenes were present only in trace amounts and thus were not quantified, Figures 3 and 4 (2). The low abundance could be explained by the unknown length of time the commercially acquired powder was dried and exposed to the atmosphere, potentially losing many compounds due to oxidation. The compounds in the *K. Galanga* hydrosol and acetonitrile extracts were quantified using an external calibrant of cinnamaldehyde, whereas only a relative percent composition was established for the sc-CO_2 extract. The calculated retention indices (RI) for ethyl cinnamate (RI =1476), n-pentadecane (RI = 1499), and ethyl p-methoxy cinnamate (RI = 1762) and mass spectra were consistent with NIST values, validating their identity (5). The derived quantities are shown in Table 1.

Table 1. Quantified Target Constituents of *K. Galanga*

	Compounds	Acetonitrile (mg/g)	Hydrosol (mg/L)	sc-CO_2 (% of Total)
1	ethyl cinnamate	0.43-0.45	7.1	2.9%
2	n-pentadecane	0.24-0.26	-	1.5%
3	ethyl p-methoxy cinnamate	8.7-8.9	16.3–16.4	92.9%

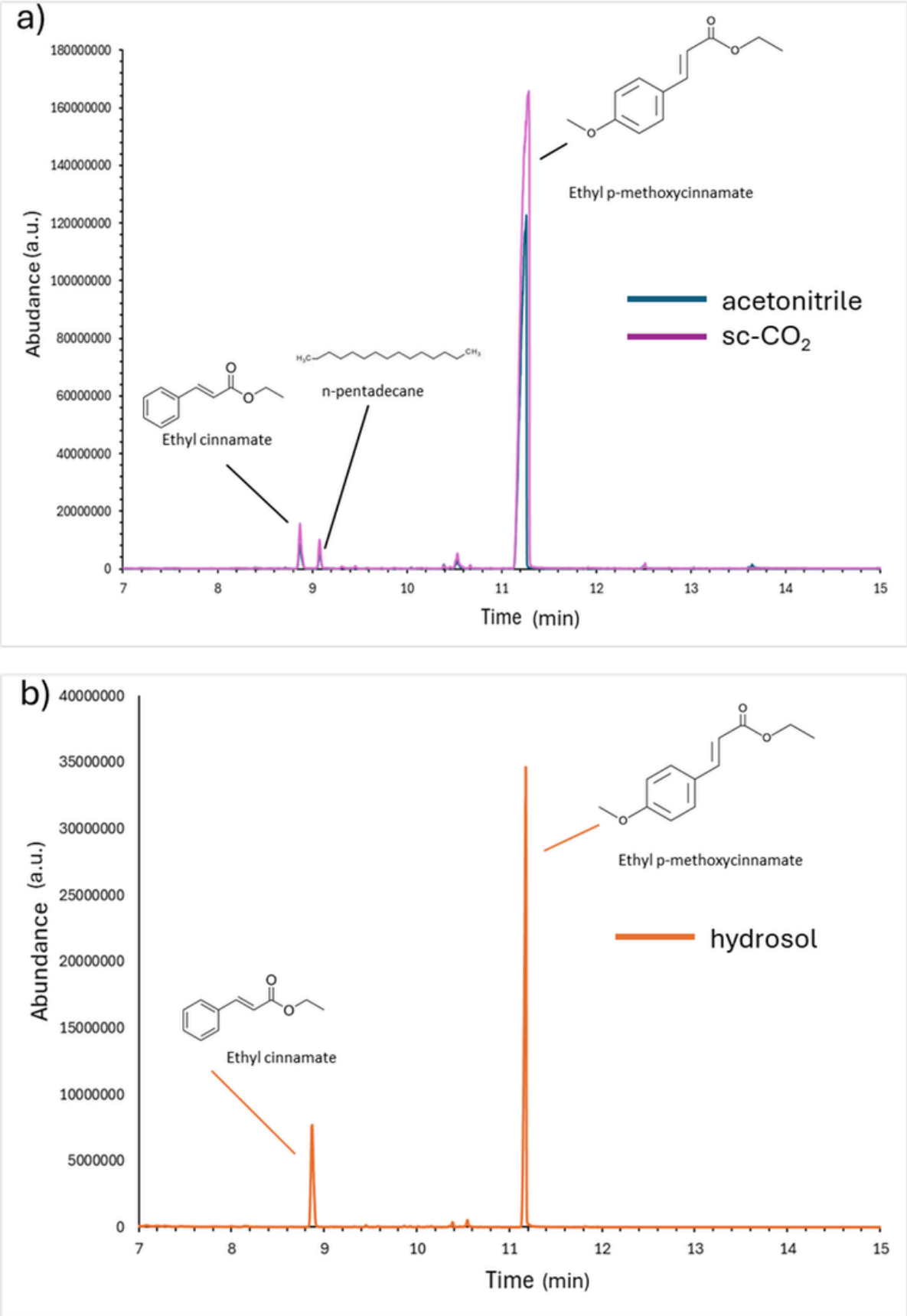


Figure 3. Chromatograms and compound identification for a) acetonitrile and sc-CO_2 extracts, and b) hydrosol.

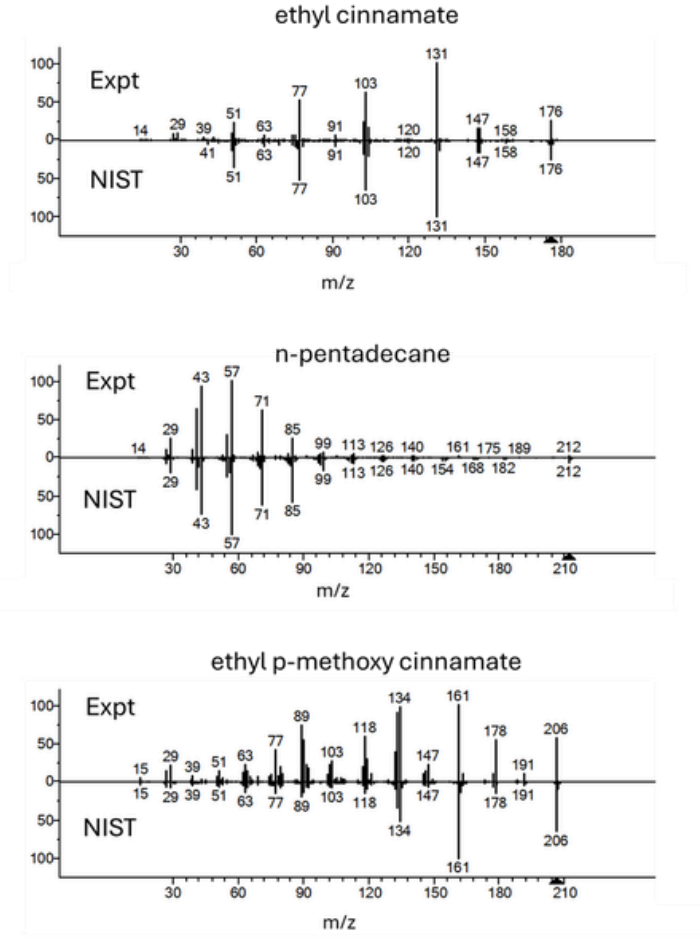


Figure 4. Observed mass spectra (Expt) of ethyl cinnamate, n-pentadecane, and ethyl p-methoxy cinnamate compared to the NIST database.

In the acetonitrile and sc-CO_2 extractions, multiple minor peaks could be observed in the chromatograms due to the low polarity solvent compared to water. This leads to the creation of extracts with unwanted impurities, which subsequently necessitate more purification, time and cost. However, these methods are selective in extracting ethyl p-methoxycinnamate in approximately the same abundance as ethyl cinnamate, and there is ten times more ethyl p-methoxycinnamate in these two extracts compared to the hydrosol, which may be due to its lower volatility. On the other hand, n-pentadecane was not present in the hydrosol extract due to the solvent's polarity, demonstrating selectivity for the two cinnamates. The hydrosol also does not contain significant amounts of the fatty acids present in the other two extracts, and thus, potentially, is more suitable for direct isolation of the target compounds.

It should also be noted that fresh *K. Galanga* material could not be obtained. This may limit the abundance of the target compounds. Additionally, SPE to remove the target compounds from water is not foolproof; the wastewater from preparing the SPE cartridges did have a similar smell to the sample, indicating that some volatile compounds were lost during processing and not caught by the SPE, decreasing their final concentration.

Due to high concentrations of both ethyl cinnamate and ethyl p-methoxycinnamate in this rhizome and their proven properties, this easily acquired, commercially dried plant material could be extracted in large amounts to develop new drugs. However, the optimal extraction method is dependent on the purpose. On one hand, natural product chemists would prefer a natural solvent for research, therapeutic and cosmetic purposes. The hydrosol extraction provides a water-based and cleaner extract, which is a more suitable choice for compound isolation. On the other hand, despite acetonitrile and sc-CO_2 extraction delivering a broader yield of volatiles, they have a larger total yield of the two cinnamates and may be more suited and efficient for industrial-scale operations.

Conclusions

While sc-CO_2 and acetonitrile extraction yielded a broader range of compounds and higher cinnamate abundance, they may demand more elaborate purification. Microwave hydrodistillation displayed a superior selectivity for the targeted volatile polar compounds. Distillation also produced a cleaner extract without the long-chain fatty acids, facilitating easier purification. The hydrosol component is often discarded in essential oil creation, but could now be repurposed, improving sustainability.

Acknowledgements

The authors thank the JLH Mass Spectrometry Core Facility of the University of Ottawa for providing instrument access and consumables related to this project.

References

1. M.N. Shaukat, A. Nazir, B. Fallico, Ginger Bioactives: A Comprehensive Review of Health Benefits and Potential Food Applications. Antioxidants 12, 2015 (2023).
2. Etsy (2026) https://www.etsy.com/ca/listing/1016957058/250g-kencur-sand-ginger-kaempferia. accessed 02/03/2026.
3. AM Produce (2026) www.amproduce.ca/product/ginger-organic-peru. accessed 02/03/2026.
4. S.-Y. Want, H. Zhao, H-T. Xu, X.-D. Han, Y.-S. Wu, F.-F. Xu, X.-B. Yang, U. Göransson, B. Liu, Kaempferia Galanga L.: Progresses in Phytochemistry, Pharmacology, Toxicology and Ethnomedicinal Uses. Front. Pharmacol. 12, 675350 (2021).
5. S. Wang, J. Yang, X. Kuang, H. Li, H. Du. Y. Wu, F. Xu, B. Liu, Ethyl Cinnamate Suppresses Tumor Growth through Anti-Angiogenesis by Attenuating VEGFR2 Signal Pathway in Colorectal Cancer. J. Ethnopharmacol. 326, 117913 (2024).
6. S. Lallo, B. Hardianti, S. Sartini, I. Ismail, D. Laela, Y. Hayakawa, Ethyl P-Methoxycinnamate: An Active Anti-Metastasis Agent and Chemosensitizer Targeting NFκB from Kaempferia Galanga for Melanoma Cells. Life 12, 337 (2022).
7. NIST Chemistry Webbook, NIST Standard Reference Database Number 69 (2023).

Original Research

Sourcing the Antifeedant Properties of Pineapple Weed (*Matricaria discoidea*)

L'approvisionnement des propriétés antialimentaires de la matricaire odorante (*Matricaria discoidea*)

Josiah A. W. Smith[1], Sharon Barden[1], Paul M Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

This study investigated the chemical composition of *Matricaria discoidea*, commonly known as pineapple weed, for the purpose of analyzing potential medicinal applications. A gas chromatograph mass spectrometer was utilized to identify the compounds in acetonitrile, CO_2, and hydrosol extracts of the plant. It was found that the acetonitrile and CO_2 extracts contained various expected molecules, including terpenes and terpenoids, commonly present in plant concentrates. However, the most significant finding was the presence of the two isomers of tonghaosu (1,6-dioxaspiro[4.4]non-3-ene, 2-(2,4-hexadiynylidene)). The hydrosol was found to have a composition of solely tonghaosu, promoting further applications due to its isolation. The presence of this molecule reinforces pineapple weed's insect-repelling activity. Consequently, the sizeable quantities of tonghaosu in pineapple weed hydrosol extract establish a promising source of the bioactive compound, supporting a potential for use in natural insect repellent and additional medicinal applications.

———

Cette étude a examiné la composition chimique de *Matricaria discoidea*, communément appelée la matricaire odorante, dans le but d'analyser les applications médicinales potentielles. Un chromatographe en phase gazeuse couplé à un spectromètre de masse a été utilisé pour identifier les composés dans les extraits d'acétonitrile, de CO_2 et d'hydrolat de la plante. Il a été constaté que les extraits d'acétonitrile et de CO_2 contenaient diverses molécules attendues, notamment des terpènes et terpénoïdes, couramment présents dans les concentrés végétaux. Cependant, la découverte la plus significative a été la présence des deux isomères de tonghaosu (2-(2,4-hexadiynylidène)-1,6-dioxaspiro[4.4]non-3-ène). Il a été constaté que l'hydrosol avait une composition uniquement de tonghaosu, favorisant de nouvelles applications en raison de son isolement. La présence de cette molécule renforce les propriétés anti-insectes de la matricaire odorante. Par conséquent, les quantités importantes de tonghaosu dans l'extrait d'hydrosol de la matricaire odorante établissent une source prometteuse de ce composé bioactif, soutenant un potentiel d'utilisation dans les répulsifs naturels et des applications médicinales supplémentaires.

Keywords: Matricaria discoidea; pineapple weed; tonghaosu; antifeedant compounds; GC-MS; hydrosol extraction; botanical insecticides; phytochemical analysis; natural insect repellents; supercritical CO_2 extraction

Introduction

Matricaria discoidea, commonly referred to as pineapple weed, is a small plant belonging to the Asteraceae family (1). Originating in northwest North America and northeastern Asia, this weed has spread widely across America and Canada, growing wild in fields as well as throughout urban settings (Figure 1), able to reach 12 inches in height (1). This common plant is widely recognized for its fragrant and edible properties (1,2). When crushed, it emits a tropical and fruity aroma with an undertone of chamomile, a closely related organism (1,2). Pineapple weed can be eaten cooked or raw, with the leaves giving a fresh and bitter taste, whereas the flowers are recognized for their herbaceous sweetness (3). Shared with chamomile, pineapple weed also demonstrates medicinal behaviour, helping with digestion and stress in addition to its physical analgesic application to treat sores, bites, and irritated skin (1,2,4). In terms of consumption forms, it is commonly steeped into a tea, used as a garnish, or possibly concentrated into a syrup for flavouring (2). Beyond its ingestion benefits, pineapple weed's aerial components can be burned, rubbed on skin, or sprayed as an extract to repel mosquitoes (2). A distinctive aroma, taste, and biological activity like the one demonstrated by pineapple weed suggest a rich chemical composition, encouraging further analysis to connect structure to its properties.

The chemical composition of *Matricaria discoidea* has been revealed to be similar to chamomile (1,2). A previous study on the essential oil of pineapple weed has identified several terpenes as primary constituents, including myrcene, [E]-beta-farnesene, and germacrene D, with the three compounds making up a

Figure 1. Pineapple weed growing next to a sidewalk

approximately sixty percent of the oil (1). These molecules are responsible for the distinct aroma and additional medicinal properties found throughout terpenes and terpenoids. However, the essential oil of a plant mainly focuses on the composition of volatile and water-soluble components; thus, the chemical background for pineapple weeds' various properties remains an area open to further investigation. In particular, we are looking into the origin of the plant's antifeedant properties, 1,6-dioxaspiro[4.4]non-3-ene, 2-(2,4-hexadiynylidene). Also known as tonghaosu, this compound is recognized as an unsaturated spiroketal enol ether with two isomers (Figure 2) (5).

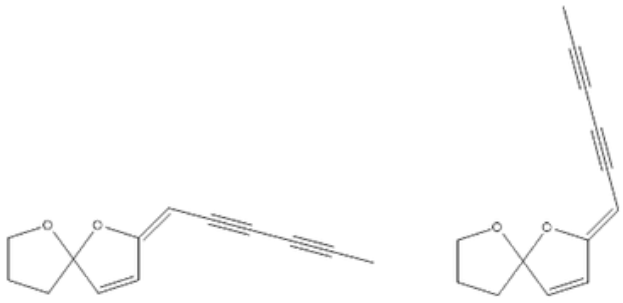


Figure 2. Chemical structure of E-Tonghaosu (left) and Z-Tonghaosu (right)

The history of these molecules originates with a very common vegetable in southern China, the tonghao plant, recognized for its specific odor and immunity to insects (5). Tonghaosu has been discovered in various plants such as Chrysanthemum coronarium L., Dendranthema indicum, and others from the Asteraceae family, including the Chrysanthemum and Matricaria genuses (5). In addition, the cis- or Z-isomer of tonghaosu has been found in German chamomile, a plant with a very close relation to pineapple weed, being a part of the same matricaria genus (2). With such a wide range of plants containing this very particular compound, it suggests that the antifeedant properties exhibited by tonghaosu might have had a survival impact on the evolution and diversification of these plants. It has been found that the antifeedant properties of the Z-isomer of tonghaosu often surpass those of the E-isomer when tested on large white butterflies; however, specific quantified data is not available (6). While no specified data relating to the particular mechanism of tonghaosu is accessible, antifeedant species often work by interacting with sensory cell receptors in insects as a deterrent instead of toxicity (7).

Given the apparent bioactivity of *M. discoidea*'s chemical composition and the significant potential impacts of specific compounds such as tonghaosu present in closely related plant species (8), an in-depth study of pineapple weeds' chemical profile is beneficial. Thus, the purpose of this study is to identify and distinguish chemical constituents of *Matricaria discoidea* using gas chromatography-mass spectrometry, solution extracts, hydrosols, solid phase extraction, and supercritical-CO_2 extractions. During this procedure and study, an emphasis was placed on the presence and abundance of tonghaosu isomers, aiming to provide more information on its chemical availability for traditional applications of the plant and further natural product development.

Experimental Procedures

Acetonitrile Extracts

Pineapple weed was harvested from a suburban sidewalk in Ottawa, Ontario, Canada. Some of the material was then dehydrated and ground into a powder using an electronic grinder. Using an electronic scale, 1 g of the powdered material was weighed and placed into a 10 mL vial. The remaining fresh material was then separated into its head (flower) and leaf (not including stem) components, placing 1 g of each into two separate 10 mL vials. 9 mL of acetonitrile was then pipetted into each vial, shaken, and allowed to sit at room temperature for 4 hours with agitation every 30 minutes.

Hydrosol Extract

250 g of fresh pineapple weed aerial segments were harvested from a relatively low-traffic area and placed in a 250 mL beaker that is part of the microwave distillation apparatus. A distilled water ice cone was screwed onto the inner surface of the apparatus's lid, and a funnel was used to direct the ice water into the beaker. The material was subjected to four 7-minute heating cycles, replacing the ice cone each time. The collected solution was

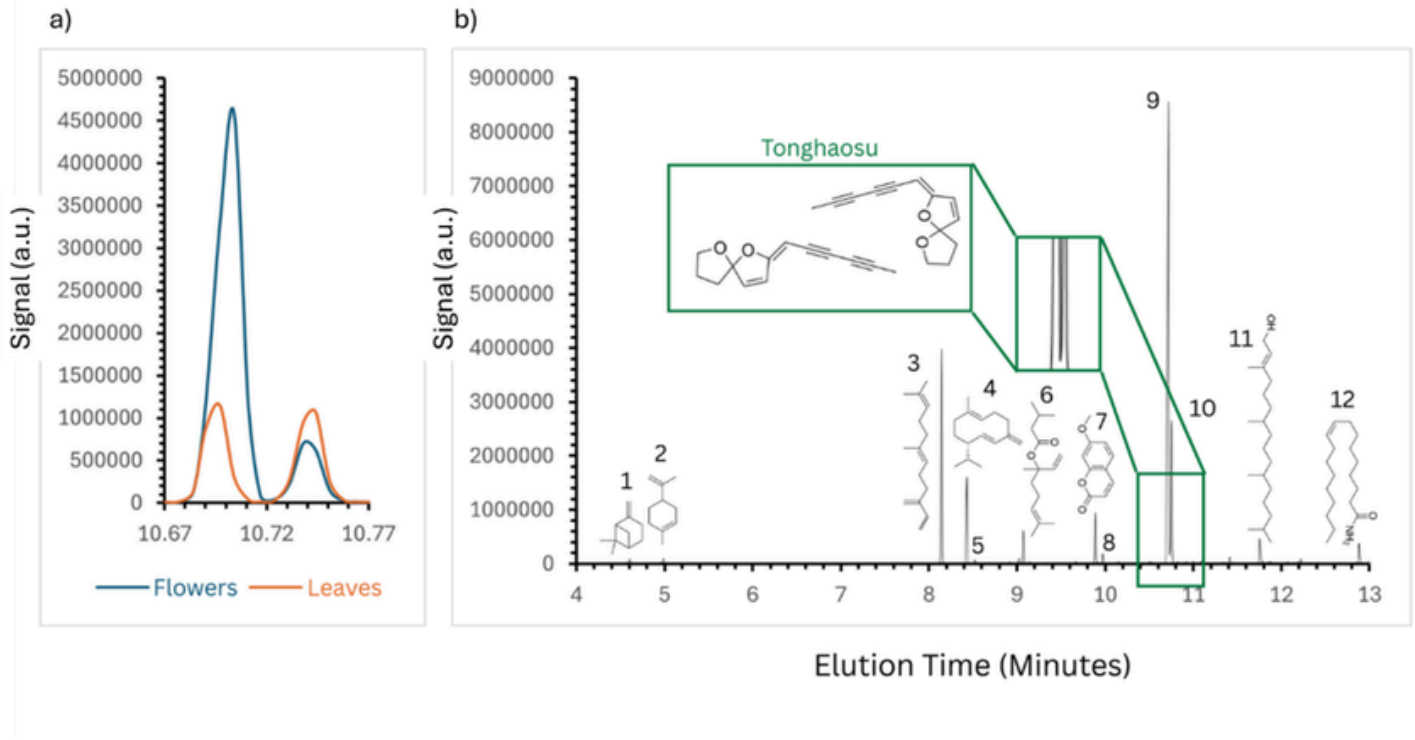


Figure 3. GC-MS chromatogram for *Matricaria discoidea* acetonitrile extracts a) zoom-in on the Tonghaosu region and comparison of separate leaves and flower extracts and b) total extract.

Table 1. Compounds Found in Pineapple Weed Acetonitrile Extract.

Peak Number	Name (Common)	Area %	Elution Time (min)	Retention Index
1	b-Pinene	0.68	4.605	
2	Limonene	0.79	4.988	10.34
3	Beta-Farnesene	16.22	8.142	14.58
4	Germacrene-D*	6.72	8.431	15.02
5	Elixene*	Traces	8.520	15.17
6	Linalyl iso-valerate*	2.47	9.070	16.07
7	Herniarin*	4.28	9.890	17.49
8	Unidentified	0.80	9.971	17.64
9	**Tonghaosu Isomer 1 (Z)**	**50.85**	**10.717**	**19.04**
10	**Tonghaosu Isomer 2 (E)**	**12.73**	**10.751**	**19.11**
11	Phytol	2.41	11.753	21.16
12	Oleamide*	2.03	12.883	23.71

*Unable to definitively confirm identity; however, high percentage quality reports suggest promising specifications.

run through solid phase extraction (Chromosep C18 500mg/6ml, PK50 SPE Column) Compounds were eluted with a 1:1 solution of acetonitrile and methanol prior to GC-MS analysis.

Supercritical-CO_2 Extraction

15 g of fresh pineapple weed was harvested from a low-traffic area, and the aerial segments (flowers, leaves, stems, not roots) were dried. The material was then run through a supercritical CO_2 extractor with an amber vial attached to the output. 1 mL of ethyl acetate was then pipetted into a 2 mL vial, and using a new pipette tip, the CO_2 extract was scraped off the amber vial and mixed into the ethyl acetate. The vial was then agitated until the material was fully dissolved. The procedure was repeated using 10 g of fresh pineapple weed flowers.

GC-MS Analysis

Solutions were filtered, and 1 µL injected onto the GC-MS. The GC-MS conditions were: temperature ramp 40 - 300 °C at a rate of 20 °C per minute; inlet temperature: 200 °C; inlet pressure: 13.96 psi; DB-5 column. Compounds were identified by comparing their mass spectra to those in the NIST mass spectral database. The retention indices were calculated using a C8-C40 standard.The hydrosol was quantified using an internal standard (4-ethylguaiacol).

Results and Discussion

As shown in Figure 3, the pineapple weed extract contains various chemicals (Table 1), including terpenes and terpenoids. However, the most abundant molecules are the two isomers of tonghaosu. The separated flower and leaf extracts demonstrate a large difference in isomer quantities with the flowers having a 6:1 peak area ratio, while the leaves have an even ratio between the two isomers. Overall, tonghaosu represents about 60 % of the derived extract.

The use of gas chromatography-mass spectrometry allowed for a comprehensive analysis of the volatile compounds in *Matricaria discoidea* (pineapple weed). Notably, tonghaosu, a molecule found in related species, was identified as a major component of the plant's aerial segments. As observed, this compound's characteristic double peak on the chromatogram indicates the existence of two isomers, identified to be Z and E structures. To identify each peak, their retention index was calculated by comparison to the C8-C40 standard mixture, and then to the NIST Chemistry Webbook. It was found that the Z-isomer elutes first (9). The Z-isomer comprises approximately 51% of the volatile extract, and the E-isomer represents 13%. However, when analyzing the

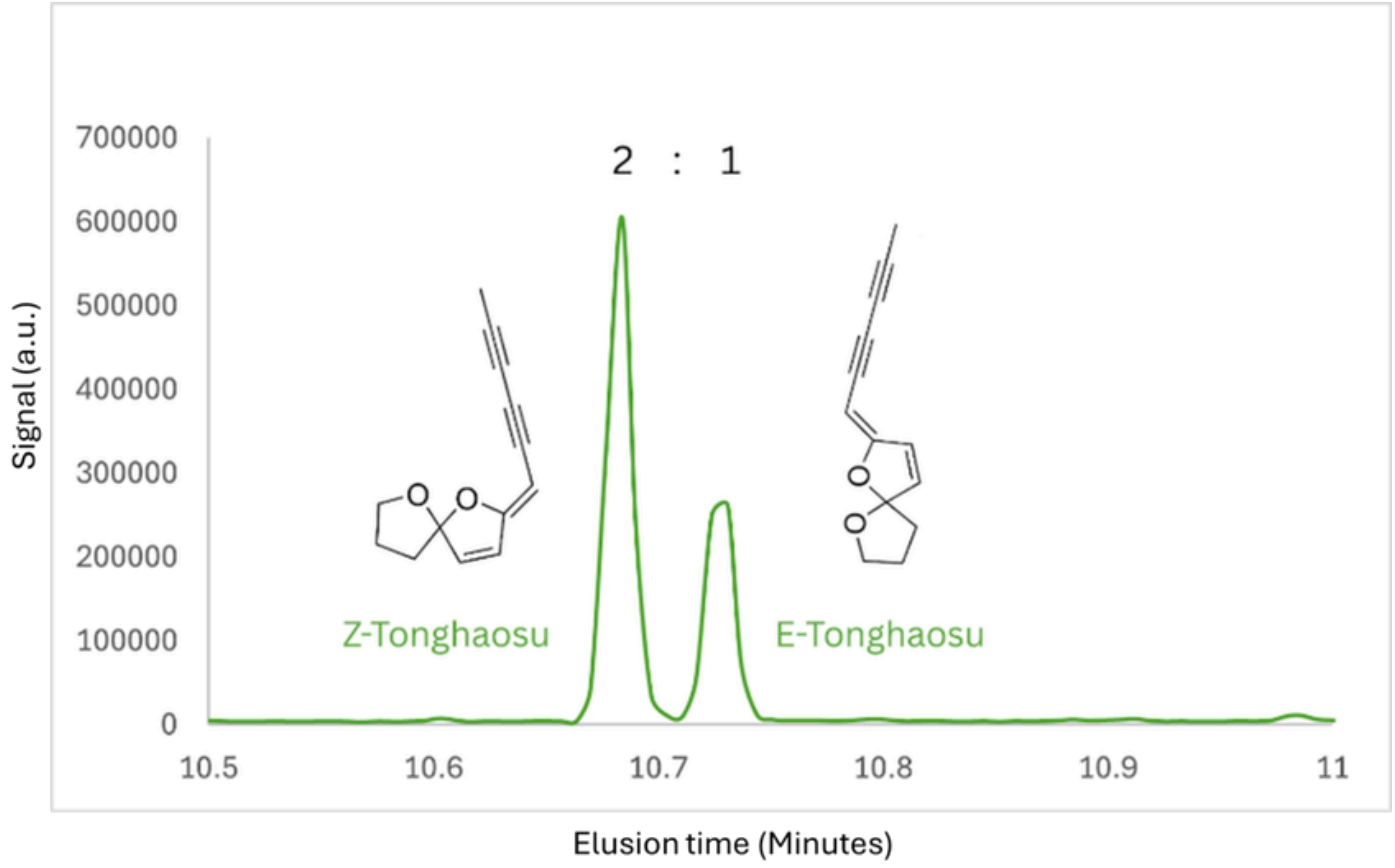


Figure 4. GC-MS chromatograms for *Matricaria discoidea* hydrosol extract, zooming in to the two Tonghaosu isomers.

leaves and flowers individually, it is evident that the flowers have a greater relative abundance of Z-tonghaosu. In contrast, the leaf extract had an even ratio of around 14% per isomer with a greater percentage of E-tonghaosu than found in the flowers.

In contrast to the acetonitrile extract, the hydrosol presented only tonghaosu as observable compounds, Figure 4 and supercritical CO_2 extractions, tonghaosu still appears as the most prevalent compound.

Antifeedant Properties

Originally identified to be the major component responsible for the tanghao plants' antifeedant capabilities, deterring insects while avoiding toxins, tanghaosu presents a potential development of natural or non-toxic insect repellents and insecticides. Thus, its large chemical presence in a plant as common as pineapple weed provides the potential for future development. As previously mentioned, the Z isomer of tonghaosu has been found to have greater antifeedant activity. Thus, as shown through all 6 extracts, Z-tonghaosu is the more prominent compound in pineapple weed, supporting insect-repelling applications through pineapple weed specifically. This apparent abundance also potentially supports recognized applications such as the previously mentioned rubbing, burning, and spraying to repel insects alongside the predator-repelling properties of aromatic terpenes.

Looking specifically at the results of the pineapple weed hydrosol, the GC-MS only revealed tonghaosu as a significant compound and thus represents a promising opportunity for isolated experimentation. The tonhaosu was quantified 3.6 g of Z-tonghaosu and 1.8 g of E-tonghaosu in the initial 250 g sample of pineapple weed. The use of water for the extraction permits a broad potential for applications. The potential insect-repelling properties of tonghaosu may be easily utilized through at-home distillation, only requiring simple kitchenware to extract a hydrosol from the plant.

Conclusion

This study was able to successfully investigate the chemical composition of the common *Matricaria discoidea* under different extraction conditions by GC-MS analysis. Tonghaosu was the major component in its Z and E isomeric forms, with Z-tonghaosu having an earlier elution and greater abundance in all analyzed segments of the plant. This provides support for traditional uses of pineapple weed as an insect repellent and for potential further application as an insecticide, with tonghaosu being a recognized antifeedant. Particularly, the isolation and abundance of tonghaosu, with emphasis on the Z-isomer through a hydrosol, present a promising source for further investigation. In summary, *Matricaria discoidea*'s rich chemical composition provides a favorable resource for the insecticide industry and therapeutic applications.

Acknowledgements

The authors thank the JLH Mass Spectrometry Core Facility of the University of Ottawa for providing instrument access and consumables related to this project.

A partial version of this article, focused only on the hydrosol,originally appeared in the NAHA Aromatherapy Journal 2025, vol. 3, p. 39, and is re-published here according to the NAHA Writers Guidelines 2022 copyright statement.

References

1. Lopes D, Kolodziejczyk PP. Essential oil composition of pineapple-weed (Matricaria discoidea dc.) grown in Canada. J. Essent. Oil-Bearing Plants. 8, 178–182 (2005). https://doi.org/10.1080/0972060X.2005.10643440.
2. Cantrell CL, Ali A, Jones AMP, Isolation and identification of mosquito biting deterrents from the North American mosquito repelling folk remedy plant, Matricaria discoidea DC. PLoS ONE 13, e0206594 (2018) . https://doi.org/10.1371/journal.pone.0206594
3. Pineapple Weed Information and Facts. Specialty Produce. 2025. https://specialtyproduce.com/produce/Pineapple_Weed_10433.php
4. Formisano C et al. Correlation among environmental factors, chemical composition, and antioxidative properties of essential oil and extracts of chamomile (Matricaria chamomilla L.) collected in Molise (South-central Italy). Industrial Crops and Products. 63, 256–263 (2015). https://doi.org/10.1016/J.INDCROP.2014.09.042
5. Yin B-L, Fan J-F, Gao Y, Wu Y-L. Progress in molecular diversity of tonghaosu and its analogs. Arkivoc. ii, 70–83 (2003). https://www.arkat-usa.org/get-file/20120/
6. Chen L et al. Synthesis and antifeeding activities of tonghaosu analogues. Journal of Agricultural and Food Chemistry. 52, 6719–6723 (2004). https://doi.org/10.1021/JF049479V,
7. Pavela R, Kovaříková K, Novák M. Botanical Antifeedants: An Alternative Approach to Pest Control. Insects. 16, 136 (2025). https://doi.org/10.3390/INSECTS16020136/S1
8. Zhang F et al. Peroxisome proliferator-activated receptor-γ agonistic effect of Chrysanthemum indicum capitulum and its active ingredients. Pharmacognosy Magazine. 14, 461–464 (2018). https://doi.org/10.4103/pm.pm_607_17
9. Senatore F, Rigano D, De Fusco R, Bruno M. Composition of the essential oil from flowerheads of Chrysanthemum coronarium L. (Asteraceae) growing wild in Southern Italy. Flavour Fragr. J. 19, 149–152 (2004). https://doi.org/10.1002/ffj.1285

Original Research

Too Hot to Handle: Chemical Profiling of Six Global Peppercorn Varieties

Trop chaud pour être manipulé : Profil chimique de six variétés de grains de poivre à travers le monde

Tabeeb Howlader[1], Sharon Barden[1], Paul M Mayer[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: pmmayer@uottawa.ca

Abstract | Résumé

For centuries, peppercorns have been valued for their aroma, spice, and distinctive flavour profiles across many cultures. While piperine is widely recognized as the primary compound responsible for their characteristic spicy taste, less is known about the broader chemical composition that contributes to their flavour, especially across different varieties of peppercorns. The goal of this study is to identify and compare the volatile compounds present in six types of peppercorns: black pepper (*Piper nigrum*), white pepper (*Piper nigrum*), red Pondicherry pepper (*Piper nigrum*), Indian long pepper (*Piper longum var. longum*), pink peppercorn (*Schinus terebinthifolius*), and Szechuan pepper (*Zanthoxylum bungeanum*). Particular attention is given to assessing the distribution of piperine within the *Piper* species. Ground pepper samples were subjected to supercritical carbon dioxide (sc-CO_2) extraction, and the resulting extracts were diluted in ethyl acetate before analysis using gas chromatography-mass spectrometry (GC-MS). The analysis enabled comparisons between *Piper* and non-*Piper* peppercorns regarding their dominant chemical constituents, revealing clear differences between true peppercorns and botanically distinct species. Additionally, the pepper samples were subjected to headspace analysis using solid phase microextraction (SPME) before GC-MS analysis. It was found that while all peppercorn varieties shared many of the same volatile compounds in varying amounts, the false peppercorns contained additional unique compounds not observed in the samples belonging to the *Piper* genus. These findings underscore the potential role of distinct volatile constituents in shaping aromatic differences among peppercorn varieties.

———

Depuis des siècles, les grains de poivre sont appréciés pour leur arôme, leur piquant et leurs profils de saveur distinctifs dans de nombreuses cultures. Bien que la pipérine soit largement reconnue comme le principal composé responsable de leur goût piquant caractéristique, on en sait peu sur la composition chimique globale qui contribue à leur saveur, en particulier entre les différentes variétés de poivre. L'objectif de cette étude est d'identifier et de comparer les composés volatils présents dans six types de poivres : le poivre noir (*Piper nigrum*), le poivre blanc (*Piper nigrum*), le poivre rouge de Pondichéry (*Piper nigrum*), le poivre long indien (*Piper longum var. longum*), le poivre rose (*Schinus terebinthifolius*) et le poivre de Sichuan (*Zanthoxylum bungeanum*). Une attention particulière a été portée à la répartition de la pipérine au sein des espèces du genre *Piper*. Les échantillons de poivre moulu ont été soumis à une extraction au dioxyde de carbone supercritique (CO_2 sc), puis les extraits obtenus ont été dilués dans de l'acétate d'éthyle avant d'être analysés par chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS). L'analyse a permis de comparer les poivres du genre *Piper* et ceux qui n'en font pas partie en fonction de leurs constituants chimiques dominants, révélant des différences nettes entre les « vrais » poivres et les espèces botaniquement distinctes. De plus, les échantillons ont été analysés en phase gazeuse (headspace) à l'aide d'une microextraction en phase solide (SPME) avant l'analyse GC-MS. Il a été observé que, bien que toutes les variétés de poivre partagent de nombreux composés volatils en proportions variables, les « faux » poivres contiennent des composés supplémentaires uniques qui ne sont pas présents dans les échantillons appartenant au genre *Piper*. Ces résultats mettent en évidence le rôle potentiel de composés volatils distincts dans la formation des différences aromatiques entre les variétés de poivre.

Keywords: peppercorns; Piper nigrum; piperine; GC-MS; SPME; supercritical CO_2 extraction; volatile compounds; terpenes; pepper aroma; flavor chemistry

Introduction

Peppercorns, the dried fruits of various *Piperaceae* species, play a central role in global cuisines and have reported medicinal properties. Their use in traditional medicine is due to their anti-inflammatory and antibacterial effects (1). The pungency of *Piper nigrum* is primarily attributed to piperine, an alkaloid that activates the Transient Receptor Potential Vanilloid 1 (TRPV1) receptor, contributing to the characteristic sensation of heat. While piperine has been widely studied and is considered the principal bioactive component in *P. nigrum* varieties, peppercorns from other species, such as *Schinus terebinthifolius* (pink peppercorns)

and *Zanthoxylum bungeanum* (Szechuan pepper), are not considered "true peppercorns" and display distinct sensory properties that suggest the presence of additional and/or alternative compounds contributing to their spiciness. All peppercorns, whether true or false, possess a similar smell. The olfactory sense is commonly said to contribute to about 75-95% of the flavours that are tasted (2), a fact that motivated an investigation into the similarity of false peppercorns to true peppercorns.

Previous studies largely focused on obtaining essential oils using traditional extraction methods such as steam distillation, hydrodistillation, and organic solvent extraction to identify volatile compounds (3). When analyzing the chemical composition of *P. nigrum* extracted essential oils, the compounds found were 3-carene, limonene, β-caryophyllene, and α-and β-pinene (4). For *Piper longum var. longum* (Indian long pepper), β-caryophyllene, pentadecane, and β-bisabolene were the main chemical constituents identified (5). For the *S. terebinthifolius* sample, the main compounds identified were 3-carene, limonene, α-phellandrene, and α-pinene (6). Lastly, for the *Z. bungeanum*, the major compounds detected were β-pinene, 1,8-terpinene, cis-piperitol acetate, oleic acid, palmitic acid, and 4-terpineol (7).

To expand the study of these peppercorns, this study employed both supercritical carbon dioxide (sc-CO_2) extraction and solid-phase microextraction (SPME) to elucidate their chemical profile. The former will extract non-volatiles from the ground peppercorns, while SPME will capture volatile components in the headspace above a ground sample. When using SPME as a method of extraction, the same variety of molecules, which are largely terpenes, are expected to be observed (8-10).

Methods

Whole peppercorns were purchased from various commercial sources. 10 g of each peppercorn was ground up in an electric grinder and placed into extraction thimbles and covered with cotton wool. The extraction thimble was placed in the sample chamber of the SFT-250 supercritical CO_2 System (Supercritical Fluid Technologies Inc.). The CO_2 was pressurized to 300 bar at a temperature of 40ºC. 1 µL of the extract was diluted in 2 mL of ethyl acetate. 1 µL of this solution was injected into the gas chromatography-mass spectrometry (GC-MS, Agilent 6890 GC with MSD) having a 30 m Agilent 19091J-433 HP-5 column, with an internal diameter of 250 mm, a stationary phase thickness of 0.25 mm, and helium carrier gas having a flow rate of 30 cm/s. The oven temperature was programmed from 150ºC to 320ºC at 20ºC/min, with a hold time of 5 minutes. The front inlet temperature was set at 250ºC. Compound identification was by mass spectral comparisons to the NIST database and retention times of standards, although a limitation of the study was the absence of Kováts retention indices (11).

Solid phase microextraction (SPME) was implemented to sample the headspace of the ground peppercorns. A small amount of each pepper, 100 mg, was ground and placed into a sealed vial. The vial was heated on a hot plate at 40ºC for 5 minutes, allowing for volatile components to enter the gas phase and be absorbed by the coated fibre of the SPME syringe. The syringe was then injected into the GC-MS, running the previously described program.

Results and Discussion

GC-MS analysis of peppercorn extracts obtained via sc-CO_2 extraction revealed that piperine was the major component in samples belonging to the *P. nigrum* species (white pepper is shown in Figure 1A, others in Figure 2). This was expected of the black, white, and red Pondicherry Peppercorns based on the literature outlined previously. However, not all of these species were *P. nigrum* peppercorns, and thus, the other peppercorns had different chemical compositions. While *P. longum* is from the same genus as *P. nigrum*, it contains piperine, isobutyl-2,4-decadienamide (which gives a spicy, herbal odour, complementing piperine) and simple hydrocarbons. There was an unidentified peak at 9 min. This unidentified compound was also apparent in *Z. bungeanum*, the extract of which also contained palmitic (C16:0) and oleic (C18:1) fatty acids (Figure 2E).

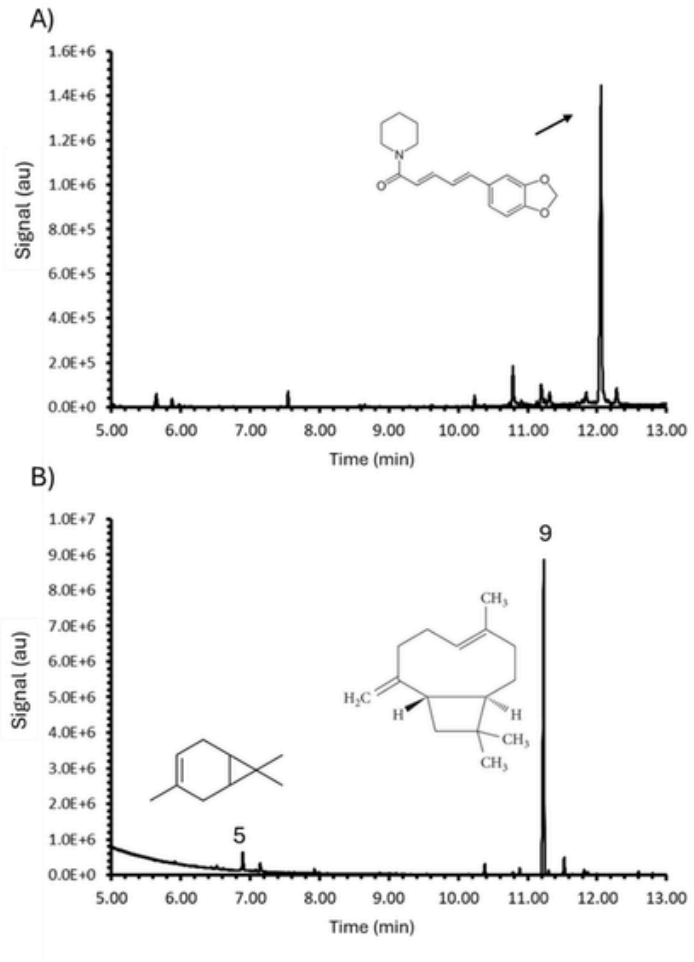


Figure 1. GC-MS results for the analysis of White Peppercorn (*P. nigrum*). A) sc-CO_2 extraction results. B) Solid phase microextraction results. The chromatograms show piperidine (12.1 min), 3-carene (6.9 min) and β-caryophyllene (11.2 min) as major compounds.

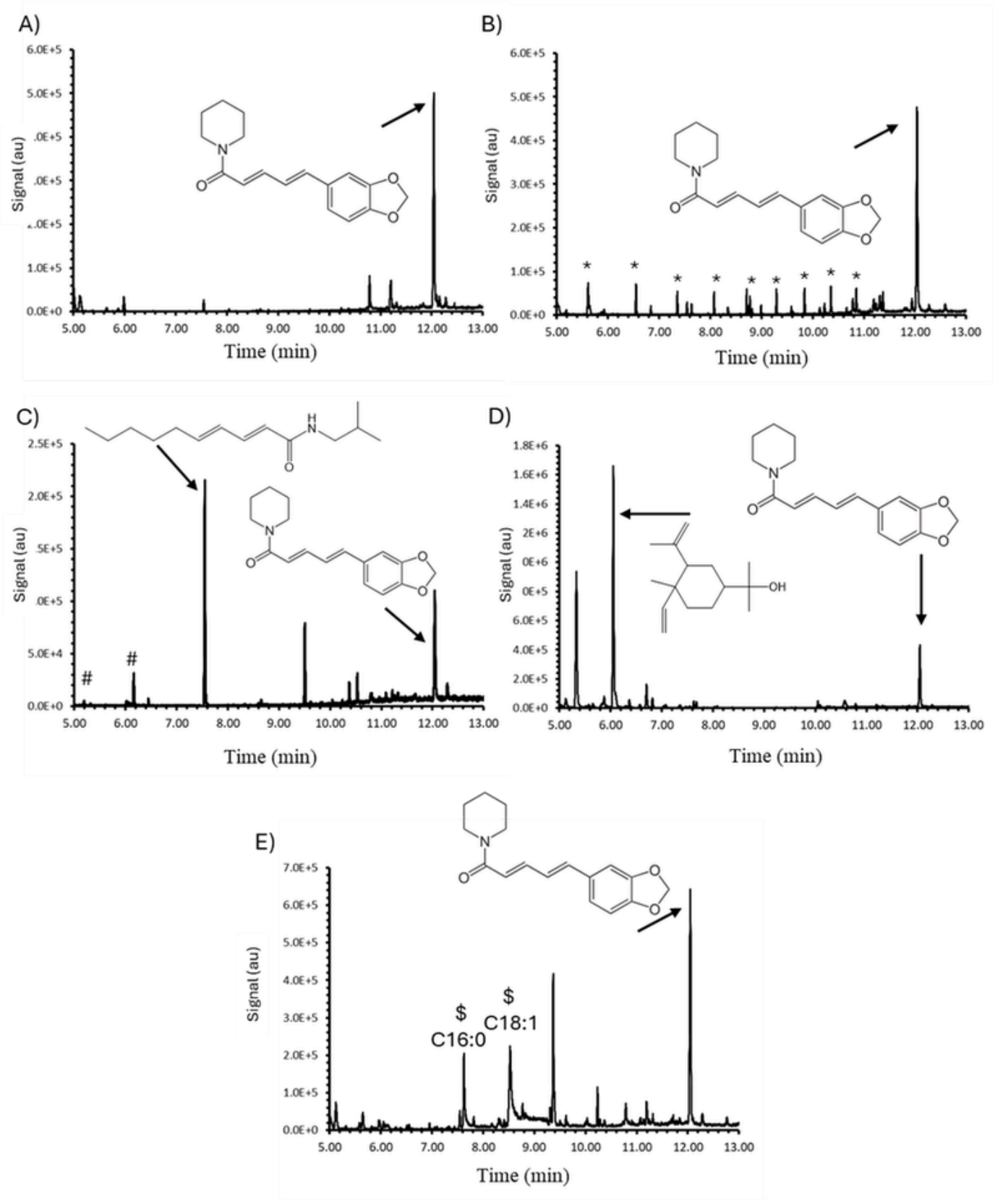


Figure 2. GC-MS results for the analysis of sc-CO_2 extracts of A) Black Peppercorn (*P. nigrum*), B) Red Pondicherry Pepper (*P. nigrum*), C) Indian Long Pepper (*P. longum*), D) Pink Peppercorn (*S. terebinthifolius*), and E) Szechuan Pepper (*Z. bungeanum*). * represents column bleed. # represents hydrocarbon and $ represents fatty acid.

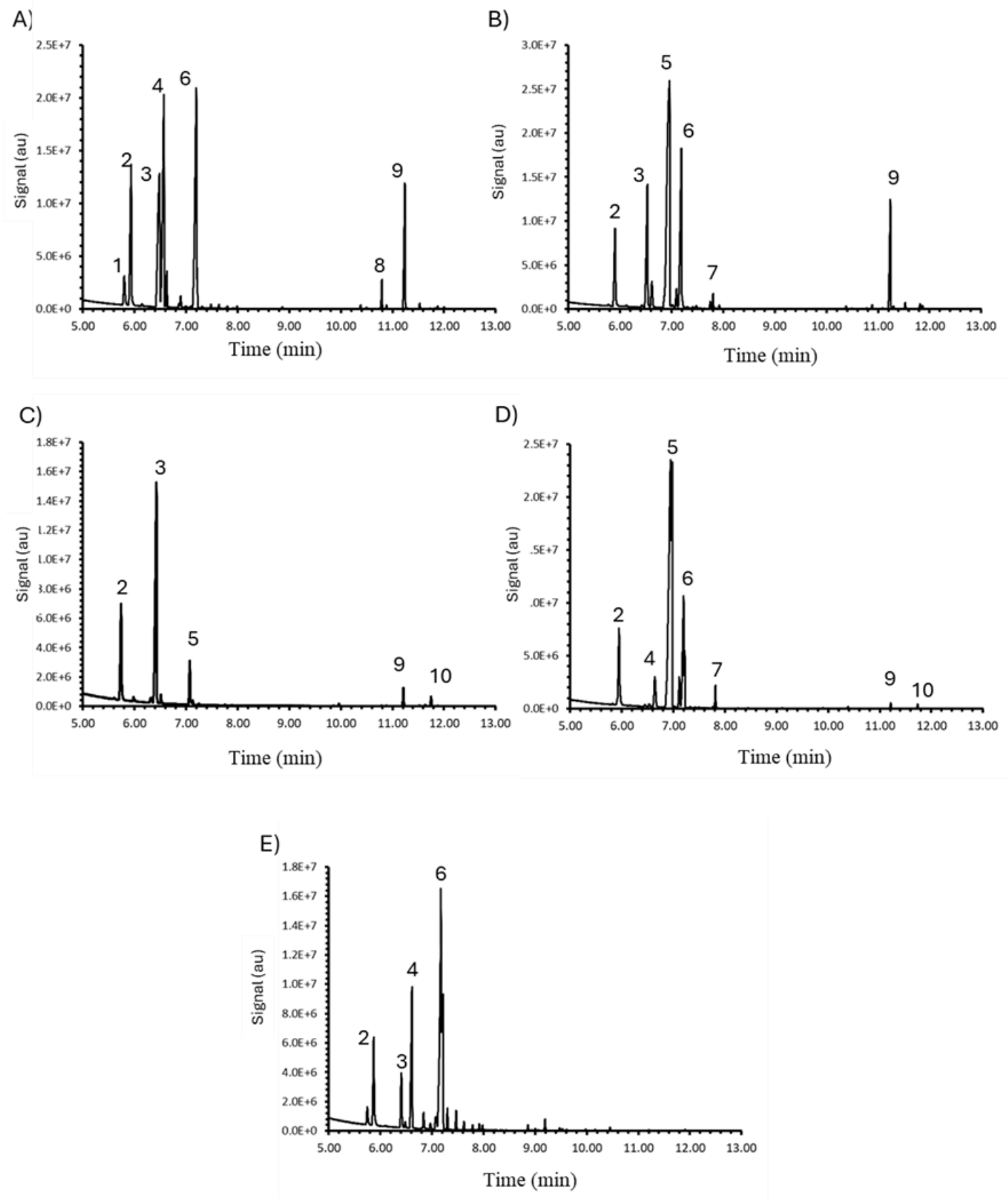


Figure 3. GC-MS results for the analysis of SPME extracts of A) Black Peppercorn (*P. nigrum*), B) Red Pondicherry Pepper (*P. nigrum*), C) Indian Long Pepper (*P. longum*), D) Pink Peppercorn (*S. terebinthifolius*), and E) Szechuan Pepper (*Z. bungeanum*). Compound assignments are in Table 1.

Table 1. Major compounds detected by SPME across the six peppercorn varieties. Numbers refer to the peak labels in Figure 3.

Identified Compound \ Peppercorn	Black Pepper (*P. nigrum*)	White Pepper (*P. nigrum*)	Red Pondicherry Pepper (*P. nigrum*)	Indian Long Pepper (*P. longum*)	Pink Peppercorn (*S. terebinthifolius*)	Szechuan Pepper (*Z. bungeanum*)
α-phellandrene	1	–	–	–	–	–
α-pinene	2	–	2	2	2	2
β-pinene	3	–	3	3	–	3
myrcene	4	–	–	–	4	4
3-carene	–	5	5	–	5	–
limonene	6	–	6	6	6	6
1-methyl-4-(1-methyl)ethylidene	–	–	–	–	7	–
Germacrene D	8	–	–	–	–	–
β-caryophyllene	9	9	9	9	9	–
pentadecane	–	–	–	10	10	–

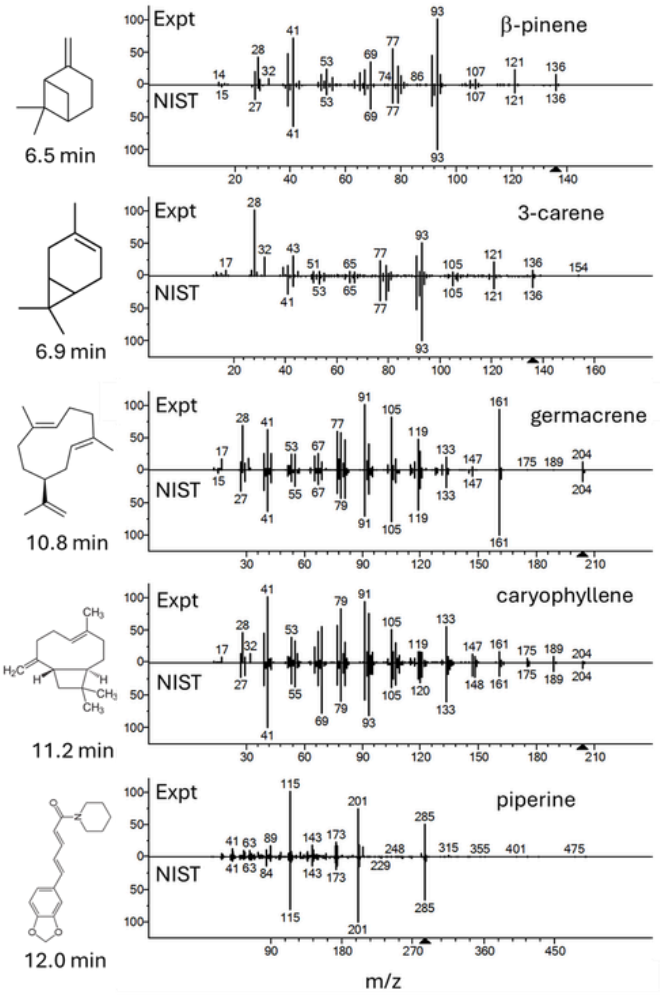


Figure 4. Comparison of selected observed and NIST mass spectra. Included compounds are β-pinene, 3-carene, germacrene, caryophyllene, and piperine. "Expt" represents observed mass spectra.

Across all peppercorn samples, the SPME results revealed a largely overlapping profile dominated by monoterpenes and sesquiterpenes, including α-pinene, β-pinene, β-caryophyllene, and limonene (see Figure 1B for White Pepper and Figure 3 for the other peppers studied, and Figure 4 for representative mass spectral identifications). Table 1 summarizes the results. It should be noted that most of the monoterpenes exhibit very similar mass spectra, and thus identification is based on both mass spectra and retention time. These compounds are commonly associated with woody and citrusy aromas and are consistent with previously reported volatile profiles of peppercorn essential oils. Despite these shared components, differences in relative abundance and the presence of additional minor compounds were observed among the false peppercorns. *S. terebinthifolius* and *Z. bungeanum* displayed unique volatile constituents not detected in the *Piper* species, which may contribute to subtle differences in perceived aroma despite their overall similarity. For instance, *Z. bungeanum* is dominated by the monoterpenes, contributing to its pungent and citrus aroma. These findings support the hypothesis that while true or false peppercorns share a common aromatic framework of terpenes, the species-specific volatile compounds may underlie their distinct sensory characteristics.

Conclusion

This study compared the volatile chemical composition of six peppercorn varieties using sc-CO_2 extraction and SPME coupled with GC-MS analysis. As predicted, piperine was identified as the main compound in the *P. nigrum* samples, confirming its presence and relative dominance with true peppercorns. In contrast, *P. longum*, *S. terebinthifolius*, and *Z. bungeanum* exhibited non-piperine dominant chemical profiles, highlighting the compositional diversity among peppercorn species and the difference between true and false peppercorns. SPME analysis revealed that both true and false peppercorns share a similar

volatile framework composed of terpenes, specifically monoterpenes and sesquiterpenes, explaining their comparable aromas. However, the false peppercorns contained additional unique volatile compounds not observed in the *Piper* species, suggesting that species-specific terpenes contribute to subtle aromatic differences. Future studies could investigate interactions between piperine identified by GC-MS and the endocannabinoid system, particularly its reported inhibition of fatty acid amide hydrolase. Such work may help elucidate biochemical mechanisms underlying the anti-inflammatory properties traditionally attributed to peppercorns.

Acknowledgements

Thank you to the JLH Mass Spectrometry Core Facility of the University of Ottawa for providing instrument access and consumables related to this project. A special thank you to Dr. Paul Mayer, who has granted me this wonderful opportunity, and to Dr. Sharon Barden for helping me along the way.

References

1. R. Balakrishnan, S. Azam, I.-S. Kim, D.-K. Choi. Neuroprotective Effects of Black Pepper and Its Bioactive Compounds in Age-Related Neurological Disorders. Aging Dis. 14, 750–777 (2023).
2. C. Spence. Just how much of what we taste derives from the sense of smell? Flavour 4, e30 (2015). 10.1186/s13411-015-0040
3. B. S. Feitosa, O. O. Ferreira, C. J. P. Franco, H. Karakoti, R. Kumar, M. M. Cascaes, R. D. Jawarkar, S. N. Mali, J. N. Cruz, I. C. de Menezes, M. S. de Oliveira, E. H. de Aguiar Andrade. Chemical Composition of Piper Nigrum L. Cultivar Guajarina Essential Oils and Their Biological Activity. Molecules 29, 947 (2024). 10.3390/molecules29050947
4. T. H. Tran, L. Ke Ha, D. C. Nguyen, T. P. Dao, L. Thi Hong Nhan, D. H. Nguyen, T. D. Nguyen, N. D.-V. Vo, Q. T. Tran, L. G. Bach. The Study on Extraction Process and Analysis of Components in Essential Oils of Black Pepper (Piper nigrum L.) Seeds Harvested in Gia Lai Province, Vietnam. Processes 7, 56 (2019).
5. N. B. Shankaracharya, L. Jaganmohan Rao, J. Pura Naik, S. Nagalakshmi. Characterisation of Chemical Constituents of Indian Long Pepper (Piper longum L.). J. Food Sci. Technol. (Mysore) 14, 73–75 (1997).
6. E. R. Cole, R. B. dos Santos, V. Lacerda Júnior, J. D. L. Martins, S. J. Greco, A. Cunha Neto. Chemical Composition of Essential Oil from Ripe Fruit of Schinus Terebinthifolius Raddi and Evaluation of Its Activity against Wild Strains of Hospital Origin. Braz. J. Microbiol. 45, 821–828 (2014).
7. Y. Li, J. Zeng, L. Liu, X. Jin. GC-MS analysis of supercritical carbon dioxide extraction products from pericarp of Zanthoxylum bungeanum. Zhong Yao Cai 24, 572–573 (2001).
8. H. H. Jeleń, A. Gracka. Analysis of black pepper volatiles by solid phase microextraction–gas chromatography: A comparison of terpenes profiles with hydrodistillation. J. Chromatogr. A 1418, 200-209 (2015).
9. M. Zhao, T. Li, F. Yang, X. Cui, T. Zou, H. Song, Y. Liu. Characterization of key aroma-active compounds in Hanyuan Zanthoxylum bungeanum by GC-O-MS and switchable GC × GC-O-MS. Food Chem. 385, e132659 (2022). 10.1016/j.foodchem.2022.132659
10. Y. G. Figueiredo, F. C. Bueno, A. H. de Oliveira Júnior, A. C. do C. Mazzinghy, H. de O. P. Mendonça, A. F. de Oliveira, A. C. de Melo, L. D. C. B. Reina, R. Augusti, J. O. F. Melo. Profile of the volatile organic compounds of pink pepper and black pepper. Sci. Electron. Arch. 14, 39-46 (2021).
11. Andriamaharavo, N.R., Retention Data. NIST 23 GC Method / Retention Index Database NIST Mass Spectrometry Data Center, National Institute of Standards and Technology (2023).

Review

After the Revolution: Where X-Ray Crystallography Stands in Context with Cryo-Electron Microscopy

Après la Révolution : où la cristallographie aux rayons X se situe dans le contexte de la cryomicroscopie électronique

Isra F. Omar[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: iomar056@uottawa.ca

Abstract | Résumé

Despite the surge of cryo-electron microscopy (cryo-EM) as a leading method for structural determination in recent years after its "resolution revolution," X-ray crystallography continues to play a key and crucial role in the field of structural biology. Although cryo-EM makes it possible to resolve non-crystallizable macromolecules, X-ray crystallography remains more effective in achieving atomic resolution and has the ability to define side chain and ligand orientations and interactions, making it useful in the mechanistic and enzymological study of proteins. This review highlights X-ray crystallography and emphasizes that X-ray crystallography and cryo-EM remain complementary methods in determining macromolecular structures instead of the former being fully replaced by the latter. In this context, together with other methods for structural analysis, X-ray crystallography is shown to adapt to the evolving structural biology landscape.

———

Malgré l'essor de la cryomicroscopie électronique (cryo-ME) comme méthode de premier plan pour la détermination des structures ces dernières années après sa « révolution de la résolution », la cristallographie aux rayons X continue de jouer un rôle clé et crucial dans le domaine de la biologie structurale. Bien que la cryo-ME permette de résoudre des macromolécules non cristallisables, la cristallographie aux rayons X reste plus efficace pour atteindre la résolution atomique et a la capacité de définir les orientations et interactions des chaînes latérales et des ligands, ce qui la rend utile dans l'étude mécanistique et enzymologique des protéines. Cette revue met en lumière la cristallographie aux rayons X et souligne que la cristallographie aux rayons X et la cryo-ME restent des méthodes complémentaires pour déterminer les structures macromoléculaires, au lieu que la première soit entièrement remplacée par la seconde. Dans ce contexte, avec d'autres méthodes d'analyse structurale, la cristallographie aux rayons X s'est révélée s'adapter au paysage évolutif de la biologie structurale.

Keywords: X-ray crystallography; Cryo-electron microscopy; Cryo-EM; Structural biology; Crystallization; Resolution; Resolution revolution

Introduction

Until the twentieth century, much of the molecular world was invisible because it could not be seen with the light microscope (1). With the emergence of the field of structural biology, however, the structure, function, and behaviour of biological macromolecules, particularly proteins, could be predicted based on physical properties and sequences (2). New protein structures continue to be determined routinely, enhancing the understanding of molecular processes and contributing to biotechnological and medical breakthroughs (1, 3). This was made possible through imaging techniques such as X-ray crystallography, cryo-electron microscopy (cryo-EM), or nuclear magnetic resonance spectroscopy.

As of today, X-ray crystallography is the most productive technique, with over 200,000 total contributed structures in the Protein Data Bank (PDB), relative to cryo-EM, which contributed only 33,000 structures (4, 5). However, recent improvements in cryo-EM-related software and hardware have allowed the technique to go through a "resolution revolution," surpassing X-ray crystallography in the determination of certain types of protein structures, such as large or membrane proteins (5, 6). As a whole, the determination of cryo-EM structures also continues to increase exponentially, with over 7,000 cryo-EM submissions to the PDB compared to 10,000 X-ray crystallography submissions in 2025 (7, 8). The field of structural biology is thus changing dynamically, with some arguing that X-ray crystallography's relevance is becoming insignificant (5). This review contests this claim, demonstrating how X-ray crystallography not only remains a key technique within the field, but also has the ability to adapt to and complement the changing landscape.

X-Ray Crystallography

Preceding electron microscopy techniques, X-ray crystallography was invented in 1912 and continues to be used to produce most of the protein structures deposited in the PDB. The technique uses X-rays and takes advantage of their having wavelengths that match the interatomic distance of crystals (5). X-rays can thus be diffracted by crystals and resultantly depict the crystal's arrangement of atoms. As such, X-ray crystallography requires crystallized protein and a diffraction pattern to build a structural model (9).

Protein crystallization

X-ray crystallography requires the formation of crystals (i.e., a well-ordered 3D packing of homogenous molecules), for structural determination. This is the bottleneck of the process, with many proteins failing to be resolved using crystallography due to their inability to crystallize. The principle of crystallization is to induce the protein to slowly precipitate out of a supersaturated solution (10). In a supersaturated solution, the amount of protein present exceeds its solubility limit, so protein is pushed out of solution (11). However, if this happens too quickly, then amorphous precipitation occurs instead. Other factors, such as the protein concentration, buffer contents, pH, temperature, and crystallization method also affect the crystallization process, with initial experiments requiring significant trial and error to establish the optimal set of conditions to encourage crystallization for a given macromolecule (10).

The most frequently used method to achieve supersaturation and crystallization of a protein is vapor diffusion. This method includes two different techniques: hanging-drop vapor diffusion and sitting-drop vapor diffusion. In both techniques, a drop of protein solution is mixed with an equal volume of precipitating solution, with the drop/mixture either hanging over a reservoir solution for hanging-drop (11) or, if the mixture has low surface tension (12), seated above the reservoir solution on a platform in sitting-drop. The concentration difference between the drop and the reservoir solution drives the system towards equilibrium via water vapor diffusion from the drop to the reservoir, thus concentrating the protein and promoting crystallization (11).

Diffraction data collection and analysis

In X-ray crystallography, a monochromatic X-ray beam collides with the crystal that is rotated, and a transducer behind the sample detects and counts the number of photons that collide into it, displaying spots of different intensities and arranged in a particular pattern. This is the diffraction pattern of the crystal of interest, and it is specifically related to the crystal morphology and protein structure. Diffraction occurs when an X-ray encounters an atom's electron cloud (instead of passing through the crystal) and bends around said atom (12, 13).

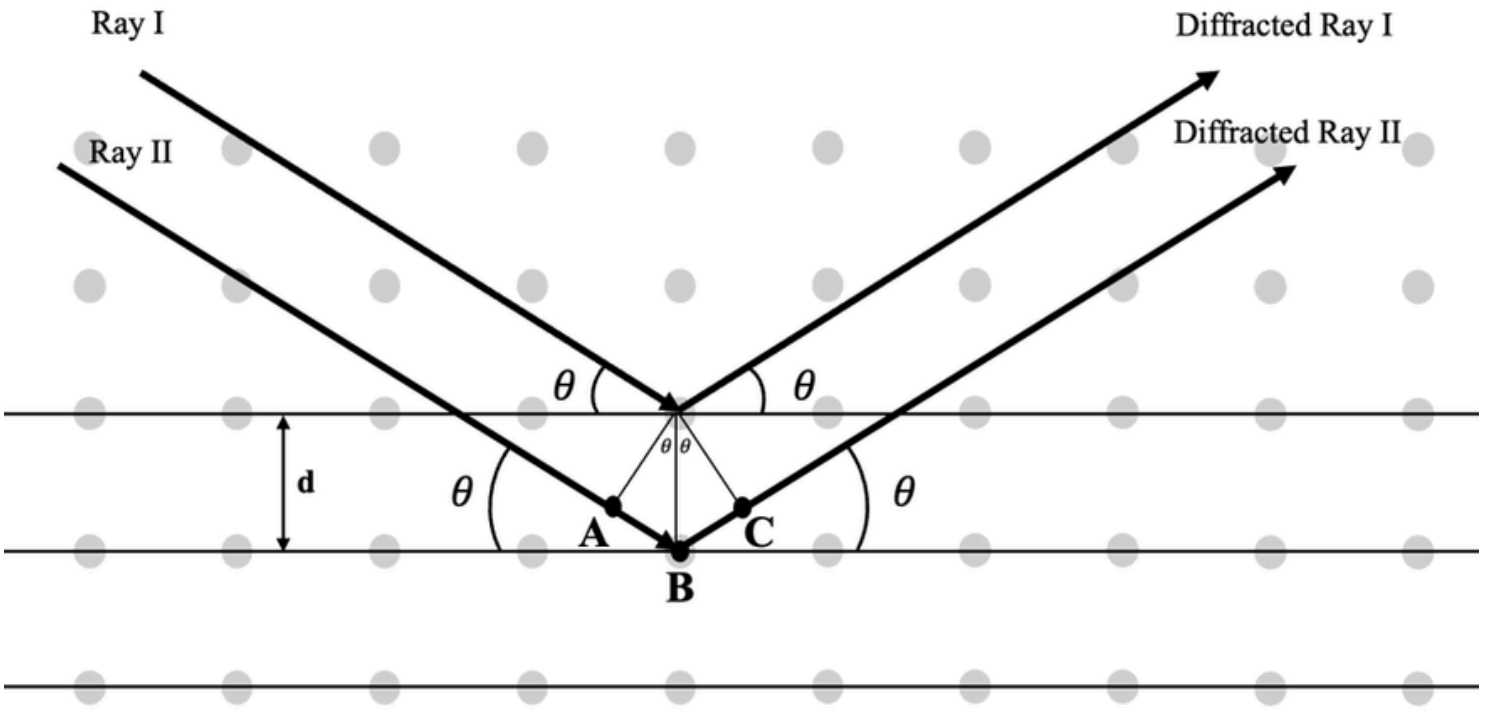


Figure 1. Bragg's Law. In a crystal lattice, atoms (grey circles) are regularly distanced, forming parallel equidistant (distance d) planes. Ray I and Ray II will diffract from identical atoms in adjacent unit cells and interfere. For constructive interference that results in a diffraction spot, Ray II must diffract in phase with Ray I. This will only occur if the extra pathlength traveled by Ray II (i.e., AB + BC) is equivalent to an integer of the wavelength of Ray I. Geometrically, this is equivalent to $2d \sin\theta = n\lambda$, which is known as Bragg's Law.

The resulting diffracted waves can also interfere constructively or destructively. In the former, two waves interact in phase and combine to form a wave of larger amplitude, while in the latter, two waves interact out of phase and cancel each other out to form a wave of zero or lower amplitude. Interestingly, because of the repeating ordered nature of atoms within a crystal, an X-ray interacts with equivalent atoms in adjacent unit cells (repeating structural units), on equidistant parallel planes (Figure 1). This means that diffracted rays from adjacent planes will interact with each other (14).

Bragg's Law predicts whether these diffracted rays will interfere constructively or destructively. Since the rays on parallel planes are diffracted at the same angle that they initially interact with the crystal's atom, successive rays will travel different pathlengths before interacting. For example, in Figure 1, diffracted Ray II travels a pathlength of AB + BC more than diffracted Ray I. Bragg's Law states that these diffracted rays will only interact constructively if the difference in their pathlength is equal to an integer of their wavelength (14):

$$AB + BC = n\lambda \quad \text{(Eq. 1)}$$

Where n represents an integer value, and λ the wavelength of the X-ray. Geometrically, this is equivalent to:

$$AB + BC = d \sin\theta \quad \text{(Eq. 2)}$$

Where d is the interplanar distance between the atoms of interaction, and θ the angle of incidence (which is the same as the angle of diffraction).
These two equations can therefore be combined to give the following equation, commonly referred to as Bragg's Law (14):

$$2d \sin\theta = n\lambda \quad \text{(Eq. 3)}$$

Where $2d\sin\theta$ is the total extra pathlength traveled by the diffracted ray (Ray II), and its equivalence to $n\lambda$ being the condition for constructive interference (with Ray I).

Additionally, if X-rays diffracted in a given direction from two identical atoms in adjacent unit cells interfere constructively, then all X-rays diffracted in that direction from identical atoms in adjacent unit cells will interfere constructively, too. The diffraction pattern's spot positions are thus characteristic of the ordered arrangement of atoms within a crystal, and they can be used to determine the unit cell's type, size, and dimensions. Each diffracted X-ray is also a manifestation of a structure factor, which describes the wavelength, intensity, and phase of the combined diffracted ray from a given angle, from every atom in a 3D reciprocal lattice (13). If all structure factor values are known, then the data can be Fourier transformed into a 3D electron density map (13, 14).

However, the phase of the diffracted rays cannot be directly obtained from the diffraction dataset (9). This is known as the "phase problem," but it can be solved indirectly. The two most common methods for this are isomorphous replacement and molecular replacement (10).

In molecular replacement, the structure factors from a closely related protein structure (homologous amino acid sequence) are used and their phases are applied to the protein of interest's dataset to calculate the new structure factor (10). This method is based on the observation that proteins with homologous amino acid sequences undergo very similar polypeptide chain folding and therefore have similar 3D protein structures. Interestingly, molecular replacement can also be used to determine relative positions of subunits or protein molecules within structural complexes (12). Nevertheless, although the reference structure must be placed in the unit cell in the same orientation and position as the protein of interest, there will always be bias towards the existing structure factor calculations (10).

On the other hand, isomorphous replacement is used when no closely related structure is available. Instead, after collecting the initial diffraction dataset, the crystal of interest is soaked in a heavy atom salt solution (e.g., mercury, platinum, gold) to incorporate heavy atoms into the protein molecule without changing its confirmation or the crystal's unit cell dimensions. By comparing the changes in diffraction intensities between this altered crystal's dataset and the original crystal of interest's dataset, the locations of the heavy atoms can be determined and phases can be estimated (10). Notably, while perfect isomorphisms rarely occur, a change in cell dimensions of ¼ of the resolution limit tend to be tolerated (12).

Building a structural model

Finally, the collected X-ray diffraction data and structure factor coordinates are converted into an electron density map to visualise real space atomic positions of the protein via inverted Fourier transformation. Fourier transformation is a mathematical operation that can decompose a signal or function in the form of signal intensities and phases (usually in its frequencies), converting complex data into its more basic components and making the data more understandable (15). The resulting map therefore depicts the 3D contours into which the protein sequence will be fitted, and the structure will be built and refined (10).

Advantages and considerations of X-ray crystallography

With highly developed robust experimental and computational methods, X-ray crystallography can solve protein structures at atomic-level resolution, excelling especially with small- to medium-sized macromolecules. Individual atoms and ligand-receptor interactions can also be resolved within crystallographic structures, allowing X-ray crystallography to be used to deduce exact atomic positions, precise mechanisms of enzymes, the binding of a drug to its target, and the structure of small and stable proteins (16).

However, X-ray crystallography is limited by its need to have highly purified samples of crystallizable protein. Not only does this mean that (should crystallization occur) more flexible proteins, like membrane proteins, may result in a less-resolved structures, but it also means that the protein in a crystal lattice may not be captured in its native-state (16, 17). Similarly, although X-ray crystallography and cryo-EM both provide snapshots of a protein's structure, because crystal structures have lattice-constrained orientations, X-ray crystallography may provide less structures for a given protein compared to cryo-EM (16).

Cryo-EM

General concepts and methodology

Although cryo-EM, primarily developed in the 1980s (5), can refer to different related techniques, the type that is attributed with revolutionizing structural biology is single particle cryo-EM, or single-particle analysis (18). This is a type of transmission electron microscopy, where the interactions of scattered electrons are used to generate contrast in an image. For purified proteins, samples are applied onto a carbon-film-covered copper grid and either embedded in a heavy-element stain (i.e., negative-stained) or flash frozen (i.e., vitrified) for regular transmission electron microscopy or cryo-EM imaging, respectively. By imaging samples dispersed on the thin and continuous carbon film, negative staining assesses the sample's homogeneity and purity and estimates optimal concentrations by revealing the protein samples in good contrast. This, however, introduces additional background noise and reduces the information that can be obtained of the internal protein structure, so it is not a process suitable for high-resolution structural determination. On the other hand, vitrified samples are prepared by depositing protein preparations on a thin and holey carbon film and quickly plunging the sample into a cryogen, like liquid ethane. This avoids the formation of ice crystals that can block imaging of single-particle samples. The protein samples are also frozen within the thin-layered ice in the individual holes on the carbon film, which, in principle, reserves the proteins' native architecture. Additionally, due to the sensitivity of biological specimens, low-dose electron beams are necessary for cryo-EM to image the sample while minimizing radiation damage (18–21).

The electron microscopy data is then collected as micrographs (i.e., a series of two-dimensional (2D) projections of the sample), which, in theory, show images of the sample in different orientations. Particles that show identical views of the same projection are then aligned two-dimensionally and averaged into a "2D classification" to reduce noise, providing a single micrograph of the protein at a given orientation. These signal-enhanced 2D images are then processed in their reciprocal space by Fourier transformation and mathematically combined and convoluted by three-dimensional (3D) image reconstruction. The 3D picture is then plotted as an electron density map into which the protein's amino acid residues can be fitted to generate an atomic model (19, 21).

Advantages and considerations of cryo-EM

Cryo-EM can consistently be used to obtain structures at intermediate to high resolution, without the need for crystals. Additionally, because macromolecules remain in their soluble state before vitrification, this technique can be used to solve structures of proteins with post-translational modifications or flexible domains in more near-native states (9, 22). Although it is true that flexible regions may still be unresolved in a cryo-EM structure, the entire protein can still be used for data acquisition and structure determination. On the other hand, in X-ray crystallography, protein truncation to remove flexible residues is common since it can reduce conformational heterogeneity and enhance lattice formation (11). Similarly, cryo-EM is particularly useful for resolving the structure of membrane proteins, which are more difficult to resolve using crystallography due to their flexibility, instability after extraction from the membrane, and difficulty in purifying the large quantities needed for crystallization (18, 23). Most notably, recent developments within the field now allow cryo-EM to be used for resolving smaller macromolecular structures as well, instead of only being limited to larger structures or supra-assemblies (24, 25). As such, cryo-EM offers great potential in structural determination, and as analysis software continues to improve, it is not surprising that the technique is becoming increasingly popular and possibly preferred.

However, limitations still exist. Because of the thermal motion of proteins before vitrification, heterogeneity in a cryo-EM sample is unavoidable (9), potentially causing regions of a cryo-EM map or structure to have different resolutions or be biased in favour of predominant structure states/conformations over rarer, intermediate ones (26). The sample is also always exposed to radiation damage. High-dose electrons destroy the sample before sufficient images can be obtained for high-resolution structural determination, while lose-dose electrons enhance background noise and reduce resolution. In either situation, the radiation limits the resolution that can be achieved within structures (19). Similarly, grid quality and sample uniformity also remain limiting factors in the quality of cryo-EM-resolved structures, often requiring significant time to achieve optimal conditions for a well-resolved structure. While trial-and-error is still present in crystallization experiments, X-ray crystallography is more well-established, with some facilities offering increased automation for crystal selection. This makes it possible to simultaneously screen multiple crystallization conditions to select for the best diffracting crystal (27).

Furthermore, cryo-EM instrumentation (e.g., electron microscopes and grid plunge-freezing devices) are also expensive, often costing millions of dollars, and requiring high operating and maintenance costs along with specific conditions for optimal performance. This contributes to the instrumentation being less accessible for use by non-expert scientists. Computational requirements are also significant; large datasets must be handled, and many computational resources are required to process the cryo-EM data. Resultantly, labs often need to invest in additional graphics processing unit workstations and storage space (27, 28).

X-Ray Crystallography's Role in a Cryo-EM-Dominant Period

Although X-ray crystallography was established earlier than cryo-EM, recent technological advancements have contributed to remarkable success for single-particle cryo-EM as a leading method for macromolecular structure determination. With cryo-EM no longer limited to resolving structures larger than 200 kilodaltons (24, 25) and achieving near-atomic resolutions, the continued relevance of X-ray crystallography is called into question.

However, to say that the future of structural biology will only be X-ray crystallography or cryo-EM, with the other becoming obsolete, is an overstatement. The future use of either technique not only depends on their own advantages and limitations, but also the biological questions that need to be answered and the goals that must be achieved. Notably, cryo-EM can skip the crystallization bottleneck entirely, offer high resolution for larger, more flexible, and/or disordered macromolecules, and it can provide information on more macromolecule conformations in a single experiment. In contrast, crystallography yields better resolution for macromolecules smaller than a few hundred kilodaltons, and it can provide dynamic information as a function of time, temperature, and pressure, even highlighting protein-ligand interactions and side chain orientations. Most importantly in pharmaceutical discovery, it allows high-throughput screening of drug candidates. X-ray crystallography therefore still plays an essential role in structural biology, as the field's overarching goal is to relate structure to biological function (5).

How X-ray and electron beams interact with samples is also to be considered. X-rays interact with an atom's electrons, so a structure solved by X-ray crystallography reflects the electron cloud distribution (i.e., the electron density) within the macromolecule. Electron beams instead interact with the Coulomb/electric potential of atoms, with electrons being diffracted based on the charge distribution within the sample (9, 29). These two resultant densities thereby represent different physical properties of a macromolecule depending on the resolution quality. At resolutions less than two angstroms, electron and Coulomb potential densities give very similar molecular shapes and features, but this is no longer the case when the resolution is higher than one angstrom. In such a case, the densities will reflect different physical properties of the macromolecule and result in different structures (9). As such, for the time being, cryo-EM cannot simply replace X-ray crystallography. X-ray crystallography is however being repositioned within the field of structural biology, especially in the context of how the two techniques may be used to complement one another to improve the understanding of biological mechanisms.

Complementarity Between X-Ray Crystallography and Cryo-EM

X-ray crystallography and cryo-EM data can complement each other in the determination of accurate structures in many ways. Some method examples include I) docking crystallographic structures within cryo-EM maps, II) solving phases of crystallographic data using cryo-EM maps, and III) crystallizing domains or subunits to fill in non-resolved sections from cryo-EM maps (9).

I) Docking crystallographic structures within cryo-EM maps

There are two key considerations that birthed this complementary technique. The first is that due to the static conformations that are required for crystalline structures, X-ray crystallography does not always reveal interaction modes or full mechanistic insights for a complex. The second is that cryo-EM maps tend to be less resolved than electron density maps from X-ray crystallography. As a result, by docking a higher-resolution crystal structure into a lower-resolution cryo-EM map, the crystal structure can be used as a template to solve the cryo-EM structure at a higher resolution than possible from using the cryo-EM map alone. With the possibility of rigid-body fitting, which aligns domains as fixed units to find optimal positions and orientations, or flexible fitting, which allows the crystal structure to undergo conformational changes (while maintaining stereo-chemistry) to fit the cryo-EM map more accurately, conformational differences between the crystallographic and cryo-EM structures can be identified, unlocking mechanistic insights that would not have been identified from crystallography or cryo-EM, in isolation (9, 30).

II) Solving phases of crystallographic data using cryo-EM maps

In X-ray crystallography, the phase problem largely occurs because there are no lenses that can collect and focus X-rays. On the other hand, electron microscopes like cryo-EM use electromagnetic lenses to collect and recombine scattered waves. As such, both the amplitude and phase information are retained within cryo-EM structure factors, thereby avoiding the same phase problem seen in X-ray crystallography (31). With cryo-EM reconstructions also increasing in resolution with the improvement of processing software, there can now be sufficient resolution overlap between X-ray crystallographic and cryo-EM data (for the same macromolecule) to use the cryo-EM map as an initial phasing model in molecular replacement for determining the crystallographic phases. Additionally, with cryo-EM now able to be used for smaller-sized macromolecules, even at low resolution this method is more applicable and may be used to solve the crystallographic phases in smaller molecules than what was previously possible (9). This method is thus particularly useful to solve the phase problem when no other existing structure exists for molecular replacement, or when isomorphous replacement becomes challenging.

III) Crystallized domains/subunits to fill non-resolved cryo-EM map sections

Although cryo-EM does not require crystallization and can resolve larger macromolecules/complexes, due to the flexible nature of a protein or its domains and the inevitable conformational heterogeneity within a sample, cryo-EM reconstructions may not have evenly distributed resolutions at each point in the density

map. This commonly results in the peripheral structure of the protein being less resolved. However, crystallization of either the entire protein or of individual domains or subunits locks the structure in one rigid conformation, so these lower- or non-resolved regions from a cryo-EM reconstitution can be solved using X-ray crystallography, and its atomic model can be fitted back into the cryo-EM map to improve resolution in these lacking areas. Because samples may also be prepared under different conditions in crystallization versus cryo-EM, the individually solved structures may also represent distinct biological states, which can contribute to a more complete structural representation of the macromolecule (9).

Conclusion

Currently, X-ray crystallography and cryo-EM are both at the forefront of structural biology. Despite recent increased favour for cryo-EM analyses, X-ray crystallography remains a fundamental technique. The distinct advantages for crystallography thus remain relevant, with it being unlikely that cryo-EM may fully replace X-ray crystallography. One method does not need to diminish the other's capabilities, nor be used entirely alone. Instead, the current era of structural biology and increased reach towards cryo-EM is reshaping the role of X-ray crystallography within a complementary framework, where better structures and more complete mechanistic profiles can be obtained in combination than when either technique is used alone. All in all, in structural biology, it is not only a question of resolving a protein structure, but also a question of how the structure elucidates function. Structural biologists must therefore use the method that will most effectively answer the biological question at hand, regardless of a technique's changing relative popularity.

References

1. S. Curry, Structural biology: A century-long journey into an unseen world. Interdisciplinary Science Reviews 40, 308–328 (2015).
2. National Research Council (US) Committee on Research Opportunities in Biology, "Molecular Structure and Function" in Opportunities in Biology (National Academies Press (US), 1989), pp. 39–41.
3. G. N. Phillips, E. E. Lattman, Happy 100th, structural biology. Struct. Dyn. 12, e061001 (2025). 10.1063/4.0000788
4. "PDB Statistics: PDB Data Distribution by Experimental Method and Molecular Type" from H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank. Nucleic Acids Research 28, 235–242 (2000). 10.1093/nar/28.1.235.
5. S. C. Shoemaker, N. Ando, X-rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. Biochemistry 57, 277–285 (2018).
6. E. Callaway, Revolutionary cryo-EM is taking over structural biology. Nature 578, 201 (2020). 10.1038/d41586-020-00341-9
7. "PDB Statistics: Growth of Structures from 3DEM Experiments Released per Year" from H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank. Nucleic Acids Research 28, 235–242 (2000). 10.1093/nar/28.1.235.
8. "PDB Statistics: Growth of Structures from X-ray Crystallography Experiments Released per Year" from H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank. Nucleic Acids Research 28, 235–242 (2000). 10.1093/nar/28.1.235.
9. H. W. Wang, J. W. Wang, How cryo-electron microscopy and X-ray crystallography complement each other. Protein Science 26, 32–39 (2017).
10. M. S. Smyth, J. H. J. Martin, x Ray crystallography. Journal of Clinical Pathology - Molecular Pathology 53, 8–14 (2000). 10.1136/mp.53.1.8
11. J. Holcomb, N. Spellmon, Y. Zhang, M. Doughan, C. Li, Z. Yang, Protein crystallization: Eluding the bottleneck of X-ray crystallography. AIMS Biophys. 4, 557–575 (2017).
12. J. Drenth, J. Mesters, Principles of Protein X-Ray Crystallography (Springer New York, ed. 3, 2007), pp. 1–44, 64–108, 123–171, 210–230, 241–247.
13. G. Rhodes, Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models, Third Edition (Academic Press, ed. 3, 2006), pp. 20–30, 49–107.
14. D. E. Sands, Introduction to Crystallography (Dover Publications, Revised ed., 1994), pp. 85-127.
15. W. van Drongelen, "Continuous, Discrete, and Fast Fourier Transform" in Signal Processing for Neuroscientists (Academic Press, 2007), pp. 91–105.
16. K. R. Acharya, M. D. Lloyd, The advantages and limitations of protein crystal structures. Trends Pharmacol. Sci. 26, 10–14 (2005).
17. C. Juhl, A. G. Beck-Sickinger, "Molecular Tools to Characterize Adiponectin Activity" in Vitamins and Hormones, Gerald Litwack, Ed. (Academic Press, 2012) vol. 90, pp. 31–56.
18. Y. Cheng, Single-particle cryo-EM-How did it get here and where will it go. Science (1979). 361, 876–880 (2018).
19. E. Nwanochie, V. N. Uversky, Structure determination by single-particle cryo-electron microscopy: Only the sky (and intrinsic disorder) is the limit. Int. J. Mol. Sci. 20, e4186 (2019).
20. V. Cabra, M. Samsó, Do's and don'ts of cryo-electron microscopy: A primer on sample preparation and high quality data collection for macromolecular 3D reconstruction. Journal of Visualized Experiments, e52311 (2015). 10.3791/52311
21. D. Lyumkis, Challenges and opportunities in cryo-EM single-particle analysis. Journal of Biological Chemistry 294, 5181–5197 (2019).
22. L. A. Earl, V. Falconieri, J. L. Milne, S. Subramaniam, Cryo-EM: beyond the microscope. Curr. Opin. Struct. Biol. 46, 71–78 (2017).
23. E. P. Carpenter, K. Beis, A. D. Cameron, S. Iwata, Overcoming the challenges of membrane protein crystallography. Curr. Opin. Struct. Biol. 18, 581–586 (2008).

24. K. Zhang, T. Grant, N. Grigorieff, Improved cryo-EM reconstruction of sub-50 kDa complexes using 2D template matching. eLife RP109790 [Preprint] (2026). 10.7554/eLife.109790.1
25. R. Castells-Graells, K. Meador, M. A. Arbing, M. R. Sawaya, M. Gee, D. Cascio, E. Gleave, J. É. Debreczeni, J. Breed, K. Leopold, A. Patel, D. Jahagirdar, B. Lyons, S. Subramaniam, C. Phillips, T. O. Yeates, Cryo-EM structure determination of small therapeutic protein targets at 3 Å-resolution using a rigid imaging scaffold. Proc. Natl. Acad. Sci. USA 120, e2305494120 (2023). 10.1073/pnas.2305494120
26. S. Zheng, Exploring the Bottleneck in Cryo-EM Dynamic Disorder Feature and Advanced Hybrid Prediction Model. Biophysica 5, e39 (2025). 10.3390/biophysica5030039
27. S. P. Muench, S. V. Antonyuk, S. S. Hasnain, The expanding toolkit for structural biology: synchrotrons, X-ray lasers and cryoEM. IUCrJ 6, 167-177 (2019).
28. H. Zafar, K. L. Malone, A. K. Singh, M. A. Cianfrocco, K. C. Glass, Breaking barriers: transitioning from X-ray crystallography to cryo-EM for structural studies. Acta Crystallogr D Struct Biol 19, 253-273 (2026).
29. M. A. Marques, M. D. Purdy, M. Yeager, CryoEM maps are full of potential. Curr. Opin. Struct. Biol. 58, 214–223 (2019).
30. A. Ahmed, P. C. Whitford, K. Y. Sanbonmatsu, F. Tama, Consensus among flexible fitting approaches improves the interpretation of cryo-EM data. J. Struct. Biol. 177, 561–570 (2012).
31. W. A. Hendrickson, Facing the phase problem. IUCrJ 10, 521-543 (2023).

Review

Beyond the Native State: From Protein Structure to Function Through Energy Landscapes and Conformational Ensembles

Au-delà de l'état natif : de la structure protéique à la fonction à travers les paysages énergétiques et les ensembles conformationnels

Ishaan S. Goswami[1†*], Isra F. Omar[1†]

1. University of Ottawa, Ottawa, ON, Canada
†Co-Authors
*Corresponding author. Email: igosw085@uottawa.ca

Abstract | Résumé

Rather than viewing protein folding as the formation of a single native structure, modern biophysics describes proteins as statistical ensembles of interconverting conformations whose populations are determined by thermodynamics, kinetics, and the energy landscape imposed by molecular interactions. Within this framework, folding is not considered a discrete event, but a dynamic process. When placed in a cellular context, these dynamic conformational ensembles are coupled to protein function and are directly influenced by the intracellular environment and interacting proteins. Protein folding and function still follow biophysical laws, albeit with varying significances, and can be described using kinetic models of intra- and intermolecular interactions. This review explores the biophysical, biochemical, and cellular determinants of protein folding, specifically highlighting that protein behaviour is a property of ensemble dynamics and the intracellular environment.

———

Plutôt que de considérer le repliement des protéines comme la formation d'une structure native unique, la biophysique moderne décrit les protéines comme des ensembles statistiques de conformations interconvertissantes dont les populations sont déterminées par la thermodynamique, la cinétique et le paysage énergétique imposé par les interactions moléculaires. Dans ce cadre, le pliage n'est pas considéré comme un événement discret, mais comme un processus dynamique. Lorsqu'ils sont placés dans un contexte cellulaire, ces ensembles conformationnels dynamiques sont couplés à la fonction des protéines et sont directement influencés par l'environnement intracellulaire et les protéines interagissant. Le repliement et la fonction des protéines suivent toujours les lois biophysiques, bien que leurs significations varient, et peuvent être décrits à l'aide de modèles cinétiques d'interactions intra- et intermoléculaires. Cette revue explore les déterminants biophysiques, biochimiques et cellulaires du repliement des protéines, en soulignant spécifiquement que le comportement des protéines est une propriété de la dynamique d'ensemble et de l'environnement intracellulaire.

Keywords: Protein folding; Energy landscapes; Conformational ensembles; Statistical mechanics; Allostery; Intrinsically disordered proteins; Molecular chaperones; Conformational selection

Introduction

Protein folding is a dynamic process governed by the principles of thermodynamics, kinetics, energetics, and statistical mechanics that is undertaken in unique biochemical and cellular contexts. This allows for the emergence of a dynamic conformation ensemble that is statistically distributed across an underlying energetic landscape, contributing to diverse biochemical functionality. This paradigm is opposed to the notion of protein folding as the deterministic acquisition of a single native structure.

The Biophysical Determinants of Protein Folding

In this section, the determinants of protein folding are highlighted from the angles of thermodynamics, molecular biophysics, and statistical mechanics (1–6).

Classical thermodynamics and the folding problem

The classical thermodynamic hypothesis of protein folding, first articulated by Anfinsen, postulates that the native structure corresponds to the global free-energy minimum under physiological conditions (2, 4–7). Evidence suggests that the "instructions" for folding a polypeptide into a native globular protein structure are contained within its primary sequence. This is known as Anfinsen's dogma, or the potential for self-assembly. However, this is not the full picture, as it portrays protein folding as deterministic, wherein one sequence will invariably produce only one structure through the same folding pathway.

Proteins are not a simple dichotomy between their amino acid sequence and the final native state. Rather, in folding, proteins populate several states, some of which are more favourable than others (2, 4–6).

As proteins can fold spontaneously in physiological conditions, protein folding must be a favourable energetic reaction. As such, it would have a negative Gibbs Free Energy (as defined by the following equation:

$$\Delta G_{folding} = \Delta H_{folding} - T\Delta S_{folding} \quad \text{(Eq. 1)}$$

Where:

$$\Delta G_{folding} < 0 \quad \text{(Eq. 2)}$$

And where enthalpy $(\Delta H_{folding})$ is defined as:

$$\Delta H_{folding} = \sum H_{folded} - \sum H_{unfolded} \quad \text{(Eq. 3)}$$

Gibbs Free Energy is a function of enthalpy $(\Delta H_{folding})$, entropy $(\Delta S_{folding})$, and temperature (T). The enthalpy of folding is due to contributions from intrachain non-covalent bonds, which is an exothermic process. This folding takes place through three mechanisms: charge-charge interactions, internal hydrogen (H)-bonding, and van der Waals interactions. Charge-charge interactions occur between cationic and anionic side chains at physiological pH. Internal H-bonding occurs between H-bond donors, the amide nitrogen of the peptide backbone, and acceptors, the carbonyl oxygen of the backbone. Van der Waals interactions are the result of the dense packing of non-polar groups in the protein core, which increases the strength of dipole-induced dipole interactions. All of these individual interactions provide low contributions, however when all intrachain bonds are summed, a large value of exothermic enthalpy is found (2, 4–6).

The entropy of folding is defined as its conformational entropy. Folding reduces the mobility, or degrees of freedom, of the polypeptide, resulting in negative entropy. However, this is overcome by contributions from solvation, or the entropy of solvation. When a hydrophobic solute is exposed to water, H-bonds in the solvent are broken. To restore the energy lost, water forms clathrates during the dissolution of said hydrophobic solute, which decreases the entropy of the solvent. With protein folding, hydrophobic residues tend to pack to the core of the protein, which minimizes the ordering of water molecules and the disruption of their H-bonds and thus restores degrees of freedom in the solvent. This is known as the hydrophobic collapse, and it can overcome the loss of peptide entropy and favour binding. The hydrophobic effect can harness entropy to create an apparent increase in order by coupling it to a greater increase in disorder among a class of smaller, more numerous objects like water molecules (2, 4–6). Therefore,

$$\Delta S_{folding} = \Delta S_{conformational} + \Delta S_{solvent} \quad \text{(Eq. 4)}$$

Combining the enthalpic and entropic contributions gives:

$$\Delta G_{folding} = \sum H_{folded} - \sum H_{unfolded} - T(\Delta S_{conformational} + \Delta S_{solvent}) \quad \text{(Eq. 5)}$$

This means that protein folding is the balance of competing energetic terms; it is a thermodynamic balance that allows for an understanding of which potential protein states are favourable. Protein folding is thus not simply proteins seeking the lowest energy, but rather, balancing competing energetic contributions (2, 4–6).

It is important to note that contributions from enthalpy and entropy also result in a small net stability of a protein. This marginal stability allows proteins to be on “the edge” of unfolding, such that they can remain flexible for function (2, 4–6).

Kinetics and folding pathways

Thermodynamics gives an idea of how favourable folding may be. However, even if a reaction or process is favourable, it may be kinetically limited. Therefore, the kinetics of protein folding must also be examined in order to diagnose which states are reachable and on what timescales (2, 4–6).

The most famous thought experiment, or paradox, of protein folding is Levinthal’s Paradox (8). The Phi-Psi (φ-ψ) dihedral angles that dictate the steric orientation between subsequent amino acids have preferred states, as visualized via the Ramachandran plot, to minimize any steric clashes between the side chains of the amino acids. As such, each residue samples a limited number of sterically permitted regions of conformational space. Assuming 10^{15} conformations can be sampled per second, the amount of time to sample all possible conformations of any protein of an arbitrary size would be absurdly large —larger than the lifetime of the universe, let alone the time biological life has existed on Earth. This is because Levinthal’s Paradox makes a false assumption that folding is a random process that requires the sampling of all possible conformations before reaching the final native state. Rather, the working kinetic “pathway” model of protein folding outlines that the primary sequence dictates the native fold which undergoes a hydrophobic collapse into molten globule intermediates in many proteins. The molten globule is a compact, native-like structure with secondary structures defined by H-bonding and backbone topology, but without defined-tertiary native structure. This pathway of protein folding necessarily eliminates the need for the primary structure to exhaustively sample all possible conformations. Instead, it follows a biased stochastic search that will lead to the native fold. The classical pathway model entails an overall decrease in ΔG as the protein folds to its native fold. Each step in the pathway is temporarily “trapped” in its conformational state as the final favourable ΔH is not yet achieved but is overcome via the loss of ΔS with folding (2, 4–6).

This challenges the notion of the two-state folding model as a universal model for protein folding, wherein two states of the protein can exist: the denatured primary sequence and the fully

folded native structure. Thus, even if the native state is thermodynamically favourable, the path taken to it matters. Proteins must traverse an energy landscape, "sampling" different conformations to find the optimal fold (2, 4–6).

Energy landscapes and force fields

The classical thermodynamic and kinetic models of protein folding paint a deterministic model of folding. However, modern biophysics portrays protein folding as a more complex and dynamic process, wherein the "sampling funnel" of classical kinetics is maintained but is not a singular path of gradient descent towards a universal ΔG minimum. Rather, it is a rugged energy landscape with constantly competing interactions dictated by energy force fields. The implication of this is that a sequence does not correspond to a single structure but to an ensemble of conformations with different probabilities (1, 3, 9–11).

Moving from macroscopic thermodynamic descriptions to microscopic conformational probabilities requires turning to statistical mechanics. Force fields are a model of molecular interactions that determine the microscopic energy of a particular protein conformation. These force fields thus determine the shape of the energetic landscape that is sampled during protein folding. Here, bonded interactions such as bond stretching, angle bending, and torsion, non-bonded terms such as van der Waals/Lennard-Jones interactions and electrostatics, and solvent-mediated effects are used to describe the energy of a certain conformation. The bonded terms constrain local geometry, torsion governs backbone and side-chain geometry and flexibility, Lennard-Jones accounts for packing and steric interactions, electrostatics captures charge interactions, and solvation accounts for water exposure and burial. All of these interactions together determine which conformations are low or high in energy (1, 3, 9–11). The following equations describe the molecular mechanics, statistical thermodynamics, and free-energy aspects of protein folding, with Table 1 containing a summary of the symbols and variables used.

$$E(R) = E_{bonded} + E_{nonbonded} + E_{solvation} \quad \text{(Eq. 6)}$$

Where is the vector, or full set of atomic coordinates of the protein.

Expanded out, the energy equation is:

$$E(R) = E_{Bond\ Stretching} + E_{Angle\ Bending} + E_{Dihedral\ rotation} + E_{Lennard-Jones} + E_{electrostatics} + E_{solvation} \quad \text{(Eq. 7)}$$

Or:

$$E(R) = \sum_{bonds\ i} k_{b,i}(b_i - b_{0,i})^2 + \sum_{angles\ i} k_{\theta,i}(\theta_i - \theta_{0,i})^2 + \sum_{dihedrals\ i} k_{\phi,i}[1 + \cos(n_i\phi_i - \delta_i)] + \sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + E_{solvation} \quad \text{(Eq. 8)}$$

Where:

$$E_{Bond\ Stretching} = \sum_{bonds\ i} k_{b,i}(b_i - b_{0,i})^2 \quad \text{(Eq. 9)}$$

Which is a harmonic spring approximation, wherein the bonds resist being stretched away from their equilibrium length b_0.

$$E_{Angle\ Bending} = \sum_{angles\ i} k_{\theta,i}(\theta_i - \theta_{0,i})^2 \quad \text{(Eq. 10)}$$

Which is another harmonic restoring force, which enforces the local geometry of angles.

Table 1. Definitions of symbols and variables used in the molecular mechanics, statistical thermodynamics, and free-energy descriptions of protein folding. The notation includes atomic coordinates, force-field parameters, thermodynamic quantities, and reaction-coordinate variables used throughout the derivation

Symbol	Meaning
R	Vector of all atomic coordinates
P	Vector of all atomic momenta
r_{ij}	Distance between atoms i and j
b_i	Length of bond i
$b_{0,i}$	Equilibrium bond length
θ_i	Bond angle
$\theta_{0,i}$	Equilibrium bond angle
ϕ_i	Dihedral angle
$k_{b,i}$	Bond force constant
$k_{\theta,i}$	Angular force constant
$k_{\phi,i}$	Torsional force constant
ε_{ij}	Lennard-Jones well depth
σ_{ij}	Distance at which Lennard-Jones potential is zero
q_i	Charge on atom i
m_i	Mass of atom i
k_B	Boltzmann constant
T	Absolute temperature (K)
Z	Partition function
Q	Folding reaction coordinate
γ	Solvent friction coefficient
$\Gamma(t)$	Random thermal force

$$E_{Dihedral\ rotation} = \sum_{dihedrals\ i} k_{\phi,i}[1 + \cos(n_i\phi_i - \delta_i)] \quad \text{(Eq. 11)}$$

Which determines the allowed backbone φ and ψ conformations and therefore secondary structure. Dihedral rotation describes the energetic cost of rotation around a bond. Rotation is periodic, so the energy repeats in a cosine pattern rather than as a harmonic operation.

$$E_{Lennard-Jones} = \sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad \text{(Eq. 12)}$$

Where the r^{-12} term is the short-range steric repulsion from Pauli's exclusion, and the r^{-6} term is the attractive London dispersion forces, which allow for tight core packing and exclude impossible conformations.

$$E_{electrostatics} = \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad \text{(Eq. 13)}$$

Which accounts for electrostatic attractive and repulsive forces (1, 3, 9–11).

The force field gives the equation for one exact conformation. The force field itself does not create the "folding funnel." Rather, the set of all force fields for all conformations creates an energy landscape for every point in conformational space:

$$R \mapsto E(R) \quad \text{(Eq. 14)}$$

This mapping creates a remarkably high-dimensional landscape which includes all possible contributions of bond angles, torsions, sidechain rotamers, backbone motions, and so on (1, 3, 9–11).

Hamiltonian and microscopic states

To formally connect these energetic descriptions to observable conformational populations, the framework of classical statistical mechanics is required. In classical mechanics, a complete microscopic state of the protein needs the position of all atoms, R, and the momenta of all atoms, P (1, 3). Thus, in order to represent the total energy of that microstate, the Hamiltonian is required, which is defined as:

$$H(R,P) = T(P) + E(R) \quad \text{(Eq. 15)}$$

Where:

$T(P) = the\ kinetic\ energy\ term$

$E(R) = the\ potential\ energy\ term$

T(P) is defined as:

$$T(P) = \sum_i \frac{p_i^2}{2m_i} \quad \text{(Eq. 16)}$$

Which is the sum of all kinetic energies of all microstates in terms of momentum.

Hence, the expanded Hamiltonian is:

$$H(R,P) = \sum_i \frac{p_i^2}{2m_i} + \sum_{bonds\ i} k_{b,i}(b_i - b_{0,i})^2 + \sum_{angles\ i} k_{\theta,i}(\theta_i - \theta_{0,i})^2 + \sum_{dihedrals\ i} k_{\phi,i}[1 + \cos(n_i\phi_i - \delta_i)] + \sum_{i<j} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + E_{solvation} \quad \text{(Eq. 17)}$$

Thus, force fields are not simply a description of protein energetics. Rather, they define the probability distribution over conformational space and determine the ensemble from which function emerges (1, 3).

Langevin dynamics

While the Hamiltonian defines the total energy of a microstate, the actual trajectory of a protein across this landscape is a stochastic process. In a cellular environment, the protein is not an isolated system but is in constant contact with a solvent "heat bath." This motion is described by the Langevin Equation (2):

$$m\frac{d^2R}{dt^2} = -\nabla E(R) - \gamma\frac{dR}{dt} + \Gamma(t) \quad \text{(Eq. 18)}$$

Here, $-\nabla E(R)$ represents the force derived from the molecular force field, $\gamma\frac{dR}{dt}$ accounts for the viscous drag of the solvent, and $\Gamma(t)$ represents the random thermal "kicks" from water molecules. Instead of simple gradient descent down the energetic funnel, the protein undergoes Brownian motion. This stochastic term is mathematically vital; it provides the energy for the protein to escape the local kinetic traps and "rugged" minima of the energy landscape, eventually allowing the system to sample the full conformational ensemble. Without this thermal noise, a protein would get stuck in the first local minimum it reached and would never fold into the native state (2).

This stochastic sampling is temperature-dependent. Below the dynamical "glass" transition (~200 K), the thermal noise, , is insufficient to overcome potential barriers, effectively "freezing" the ensemble into a single microstate and severely restricting conformational dynamics requisite for biological function (9–11).

The Ergodic Hypothesis posits that, given enough time, a single molecule governed by stochastic Langevin dynamics is assumed to visit every microstate in its conformational ensemble with a frequency proportional to that state's Boltzmann weight over sufficiently long timescales. This is the basis of the assumption that a time average of a single protein's trajectory is equivalent to an ensemble average. This allows the partition function to become a physical prediction of how a single protein may structurally fluctuate (9–11).

Boltzmann distribution and partition function

At thermal equilibrium, the probability of a microstate is proportional to its Boltzmann weight (1, 3):

$$P(R,P) \propto e^{\frac{-H(R,P)}{k_B T}} \quad \text{(Eq. 19)}$$

In order to become a true probability, the Boltzmann must be normalized by the partition function:

$$P(R,P) = \frac{e^{\frac{-H(R,P)}{k_B T}}}{Z} \quad \text{(Eq. 20)}$$

Where:

$$Z = \int e^{-\frac{H(R,P)}{k_B T}} dR dP \quad \text{(Eq. 21)}$$

Now, the Hamiltonian may be substituted into the partition function:

$$Z = \int e^{-\frac{[T(P)+E(R)]}{k_B T}} dR dP \quad \text{(Eq. 22)}$$

Via the exponential identity:

$$e^{-\frac{[T(P)+E(R)]}{k_B T}} = e^{\frac{-T(P)}{k_B T}} e^{\frac{-E(R)}{k_B T}} \quad \text{(Eq. 23)}$$

Thus,

$$Z = \int e^{\frac{-T(P)}{k_B T}} e^{\frac{-E(R)}{k_B T}} dR dP \quad \text{(Eq. 24)}$$

$$Z = \left[\int e^{\frac{-T(P)}{k_B T}} dP\right]\left[\int e^{\frac{-E(R)}{k_B T}} dR\right] \quad \text{(Eq. 25)}$$

The integral may be separated since one factor depends only on P and the other only on R.

When integrating over momenta, the Gaussian integral over all momenta is obtained which depends only on masses and temperature, not on the conformation R. This means that when comparing conformations, the whole momentum contribution is just a constant pre-factor. This allows for the simplification wherein P(R, P) may be treated as P(R) for the purposes of protein folding. This separation thus implies that conformational probabilities only depend on the potential energy surface E(R), allowing for reduction from phase space to configurational space (1, 3).

These conformations are populated statistically, according to a Boltzmann distribution, with the force fields giving E(R) (1, 3).

$$P(R) = \frac{e^{\frac{-E(R)}{k_B T}}}{Z_{conf}} \quad \text{(Eq. 26)}$$

With the partition function specific for configurational probability:

$$Z_{conf} = \int e^{\frac{-E(R)}{k_B T}} dR \quad \text{(Eq. 27)}$$

Which allow for the normalization over all possible conformations.

As this follows a Boltzmann distribution, lower-energy conformations are more greatly populated, but are not uniquely occupied. Native-like structures tend to satisfy multiple favorable interactions simultaneously, including hydrophobic burial, hydrogen-bond stabilization, electrostatic complementarity, steric packing, and reduced solvent penalty. Thus, these conformations occupy a lower free energy than most unfolded states and are thus more stable. However, there are still many local minima, so the landscape is rugged, not a smooth absolute global minimum. Consequently, the native state is better understood as a basin of closely related low free-energy conformations instead of a single rigid structure (1, 3).

Projection onto a reaction coordinate

The distribution of conformations is then projected onto a low-dimensional reaction coordinate that may have different ways of measuring folding progress, such as the fraction of native contacts, root-mean-squared distance from native state, or radius of hydration. This creates a distribution of all states on this simplified coordinate diagram (1, 3):

$$F(Q) = -k_B T \ln P(Q) \quad \text{(Eq. 28)}$$

Where:

$Q = a\ folding\ coordinate$

$P(Q) = the\ probability\ of\ finding\ the\ protein\ at\ that\ value\ of\ Q$

$F(Q) = the\ free - energy\ profile\ at\ that\ coodinate$

Because the full configurational space is too high-dimensional to visualize directly, the probability distribution is often projected onto a reduced reaction coordinate (1, 3).

And *P(Q)* is defined as:

$$P(Q) = \sum_{R \in Q} P(R) \quad \text{(Eq. 29)}$$

As moves toward more native-like conformations, tends to decrease on average, accounting for “bumps” from intermediates and kinetic conformational traps. defines the free-energy profile whose topology gives rise to the folding funnel, which is a visualization of the statistical distribution of all conformations. The resulting free-energy landscape can be visualized as a rugged

folding funnel (Figure 1) in which the top of the folding funnel is broad, not because all unfolded states are necessarily of similar energies, but due to the high degeneracy of these structures. The bottom of the funnel has many closely related native-like conformations which are dynamic, not a perfectly rigid structure (1, 3).

Force-field terms generate the microscopic energetic surface over conformational space. When this surface is viewed through the lens of statistical mechanics and projected onto suitable coordinates, it yields these rugged free-energy funnels to describe protein folding (1, 3).

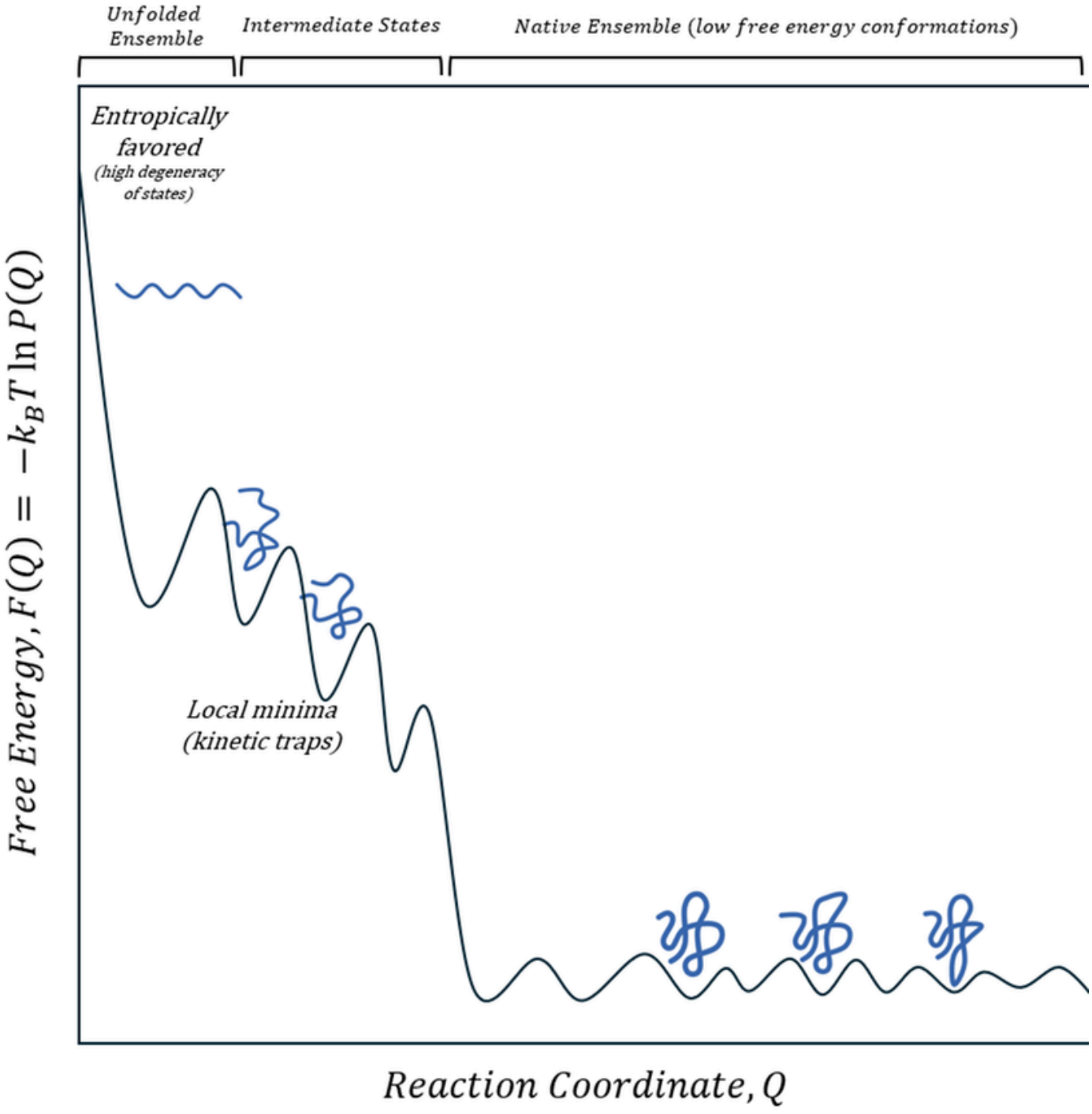


Figure 1. Rugged free-energy landscape of protein folding. The free energy F(Q) is plotted as a function of a reaction coordinate Q describing folding progress. The unfolded ensemble occupies a high-entropy region with a large degeneracy of accessible conformations. Intermediate states correspond to local minima and kinetic traps arising from competing interactions. The native state is represented as a basin of low free-energy conformations, reflecting a dynamic conformational ensemble rather than a single structure. The overall funnel shape indicates a thermodynamic bias toward the native basin, while the ruggedness reflects the complexity of the underlying energy landscape.

Critically, the folding funnel represents a free-energy landscape rather than a pure potential-energy landscape. Although unfolded conformations often possess higher potential energy, they are entropically favored because of their enormous number (1, 3).

$$F(Q) = E(Q) - TS(Q) \quad \text{(Eq. 30)}$$

Thus, the logic of this derivation may be understood as follows:

$$E(R), T(P) \rightarrow H(R,P) = T(P) + E(R) \rightarrow P(R,P) = \frac{e^{\frac{-H(R,P)}{k_B T}}}{Z} \rightarrow P(R) = \frac{e^{-\frac{E(R)}{k_B T}}}{Z_{conf}}$$
$$\rightarrow P(Q) = \sum_{R \in Q} P(R) \rightarrow F(Q) = -k_B T \ln P(Q) \quad \text{(Eq. 31)}$$

This has the profound implication that a sequence does not dictate one singular protein conformation; it dictates a dynamic ensemble of conformations (1, 3).

Proteins as Dynamic Ensembles

Protein function is not encoded in a single idealized native structure, but in the statistical distribution of conformational states and their relative populations within the native landscape, as shown above (1, 3). This may be modeled as:

$$\langle f \rangle = \sum_I P_i f_i \quad \text{(Eq. 32)}$$

Where:

$P_i = the\ probability\ of\ state\ i$

$f_i = the\ value\ of\ a\ measurable\ property\ in\ the\ state\ i$

Thus, if one conformation binds a ligand strongly, another binds weakly, and another is completely inactive, the observed behaviour of the protein will be the weighted average of all three. This directly upsets the traditional structure-function dogma, wherein the function is the weighted sum of all constituent components of the ensemble, rather than a clean one-to-one mapping (1, 3).

The Biochemical Determinants of Protein Folding

The biophysical frameworks outlined above provide a quantitative framework for understanding how proteins navigate energy landscapes. However, proteins exist in a cellular context, where folding does not occur in isolation, nor where it acts as an endpoint to ensure biological function. In a biochemical context, protein folding is coupled to function, with conformations modulated by ligands, the intracellular environment, the intrinsic disorder within proteins, and chaperones.

Conformational selection within ensembles

Of the many ways protein activity can be regulated, allostery is particularly relevant in the context of conformational selection. An allosteric effect occurs when the binding of a molecule to a non-orthosteric site (i.e., a site distinct from the primary, active site) triggers a structural shift in the protein and alters the protein's function. Allostery is thus the ability of a protein to transduce signals from an allosteric site to its (often distant) orthosteric site and influence its activity and thus the biological outcome. This can occur in both enzymes and proteins without catalytic ability (12). Two "textbook" models of allostery are frequently referenced: the Monod-Wyman-Changeux (MWC) or concerted model, and the Koshland-Nemethy-Filmer (KNF) or sequential model (Figure 2). The former describes allostery as a cooperative conformational transition of protein oligomers, while the latter describes allostery as progressive conformational transitions of individual/distinct domains within a protein (12–14).

Taking hemoglobin (which has four subunits) binding oxygen as an example, the MWC model assumes there is an equilibrium between two protein states, relaxed (R) and tense (T), where all subunits are simultaneously shifting between the R-state (which has higher affinity for oxygen) or the T-state (which has lower affinity for oxygen). Oxygen preferentially binds R-state subunits, and when binding of oxygen to all four subunits occurs, a shift in the R-T equilibrium causes a concerted conversion of all subunits from the T- to the R-state, creating more favourable binding sites for subsequent oxygen molecules. Allosteric effects are thus a result of an equilibrium-shift between distinct states in the MWC model, where the conformation of each subunit is constrained by its association with the other three subunits (12–14).

On the other hand, the KNF model describes an induced-fit mechanism; ligand binding induces structural changes. It assumes that binding of oxygen to one T-state subunit induces a conformational change only in this subunit, which then shifts the conformation and affinity of its neighbouring subunits. The KNF model therefore includes intermediate states for hemoglobin's structure, with the conformation and binding to each subunit being distinct, yet cooperative (12, 15).

Taking into account dynamics and thermodynamics, however, both the MWC and KNF models for allostery are oversimplified. In 1992, Parsegian et al. identified that the transition from hemoglobin's deoxygenated T-state to its oxygenated R-state was also characterized by the binding of 60 additional water molecules (16). Moreover, within the same year, Arnone et al. determined a third quaternary structure for hemoglobin (i.e., the "R2" state) using X-ray crystallography (17). Although the MWC model, especially, may already hint at conformational selection, both MWC and KNF models assume that protein states are well-defined structures (instead of continuous conformational distributions) and they fail to consider the protein's environment.

Parsegian et al.'s discovery showed that hemoglobin's transition between states is not only a result of the protein's structural rearrangement but also coupled to solvent changes. Of course, the activity of only one solution component cannot be varied with all

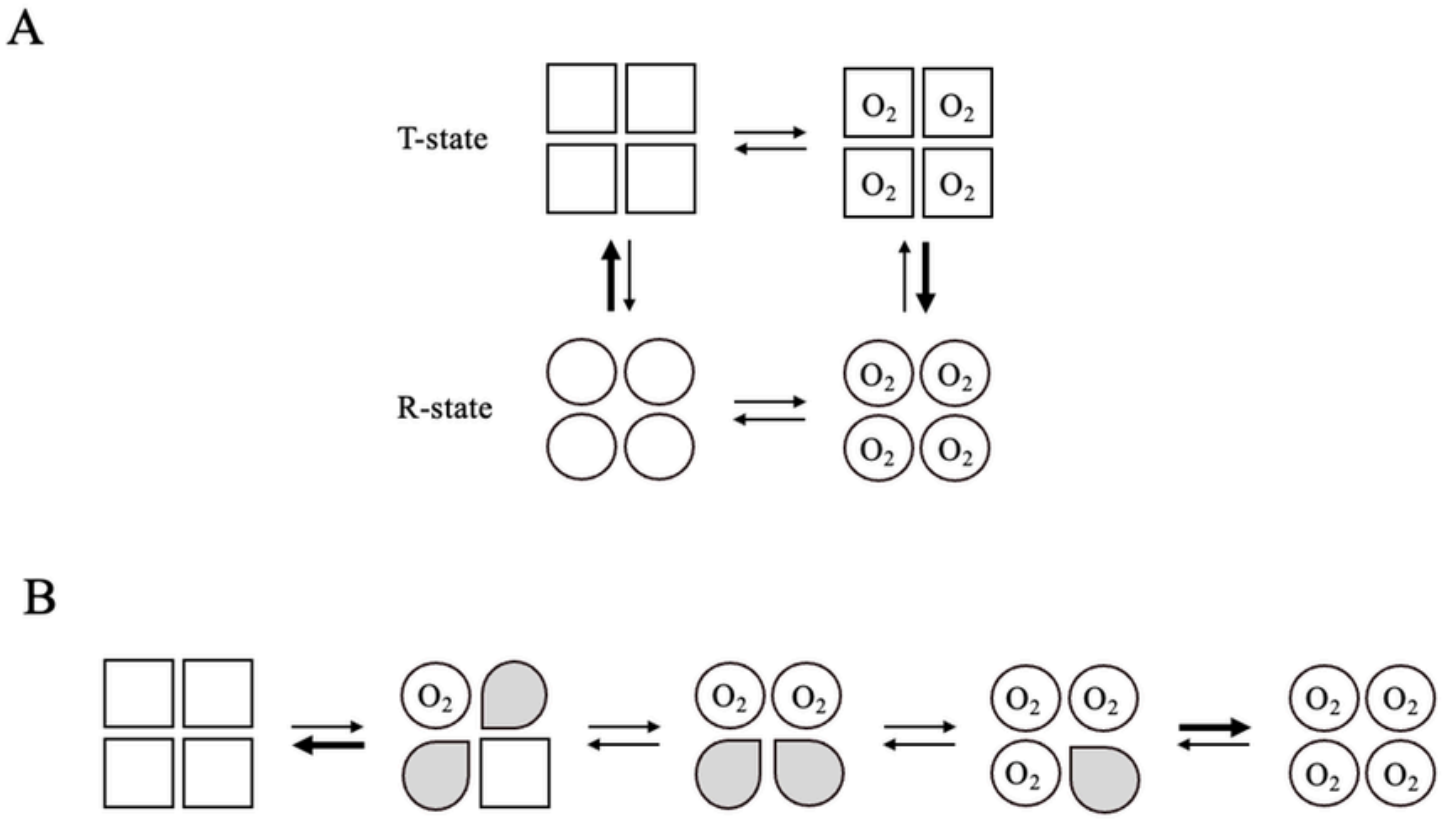


Figure 2. Models of allostery. Both panels use hemoglobin's four subunit structure as an example, with squares representing the tense (T) state, circles the relaxed (R) state, and pointed circles as an intermediate state with higher affinity for oxygen (O_2). **A. Monod-Wyman-Changeux (MWC) model.** In this concerted model, all subunits transition between the T and R states. Without oxygen, the R-T equilibrium favours the T-state. On the other hand, oxygen preferentially binds the R state due to higher affinity. Once oxygen binds to all subunits, the equilibrium shifts to favour the R-state. All subunits are thus either in the T- or R-state conformations, depending on substrate binding. **B. Koshland-Nemethy-Filmer (KNF) model.** In this sequential model, binding of oxygen to one subunit causes a conformational shift and a change in affinity in the neighbouring subunits. This conformational shift is to an intermediate state other than the T- or R-states considered within the MWC model. Panel adapted from Figure 1 in Monstserrat-Canals, Cordara, and Krengel (2025) (12).

others held constant; changing the concentration of one solute/ligand will directly affect the activity and distribution of water too. In a physiological medium, the binding of these 60 water molecules is "osmotic work" and requires 0.2 kcal/mol of energy. Water therefore acts thermodynamically as an allosteric ligand and contributes energetically to a protein's conformational selection and functional regulation (16). Next, Arnone et al.'s discovery redefined the entire "T- versus R-state" depiction of hemoglobin. Intermediate conformations do exist and are energetically accessible structures (17), demonstrating how protein conformations are continuous and how alternative conformations have functional relevance. As such, both findings support an ensemble view for protein conformations.

Because an ensemble view for protein conformations describes proteins as dynamic, heterogenous populations of different conformational states that constantly interconvert (as depicted in Equation 32), ligands are not seen as binding to a single static state. Instead, they interact with a population of states and stabilize specific conformations (this depicts conformational selection) and induce population shifts towards certain states within the ensemble (18, 19). Both conformational changes and binding/unbinding events in proteins require the crossing of free-energy barriers since they are thermally activated processes. However, the transition time needed to cross a free-energy barrier is shorter than the "dwell times" in conformational states before/after energy barrier crossing. This means that transition times are typically poorly resolved in experiments, with conformational changes instead appearing as sudden "jumps" between established conformational states (19).

With this in mind, conformational selection and the KNF model become more similar. In the KNF model, conformational changes occur, or are induced, after a ligand binds to the unbound ground-state conformation of a protein. On the other hand, in conformational selection, the conformational change occurs before ligand binding, and then the ligand will select a given conformation for binding and stabilize it. Nevertheless, a conformational selection-based mechanism (i.e., conformational excitation from a ground-state, low-energy confirmation to a

higher-energy conformation) essentially becomes an induced change-based mechanism (i.e., conformational relaxation from an excited-stated to a lower-energy, ground-state conformation) when binding/unbinding direction is reversed (Figure 3) (19). The MWC and KNF models are thus not invalidated, but redefined and broadened within a higher-dimensional energy landscape view of protein allostery and function, as described by conformational selection.

Functional consequences of conformational ensembles: Intrinsically disordered proteins

Interestingly, function may emerge entirely from ensembles. This is the case for intrinsically disordered proteins (IDPs). IDPs are functional proteins that lack a fixed three-dimensional structure, often characterized instead by their amino acid sequence and simplicity. Although IDPs are unable to fold spontaneously into unique, stable, globular structures, their undefined structures continuously interconvert between a continuum of conformations, and they are still thermodynamically stable (20, 21). This means that their highly entropic disorder (i.e., the equilibrium of their distinct conformations) is lower in free energy than a single conformation or a select set of conformations (21). Unlike many proteins, IDPs satisfy the "disorder-function paradigm," whereby proteins may still perform cellular functions without ever achieving a stable, physiological, three-dimensional structure. It is this conformational flexibility that allows IDPs to be involved in cell signalling, regulation, and binding (22).

IDPs have significant conformational entropy due to their ability to sample many conformations. A large amount of intra- and intermolecular interactions would conversely result in structure and limit sampling, thus restricting conformational entropy. Resultantly, the same biophysical forces that apply to normal proteins also apply to IDPs, but with varying relative importances.

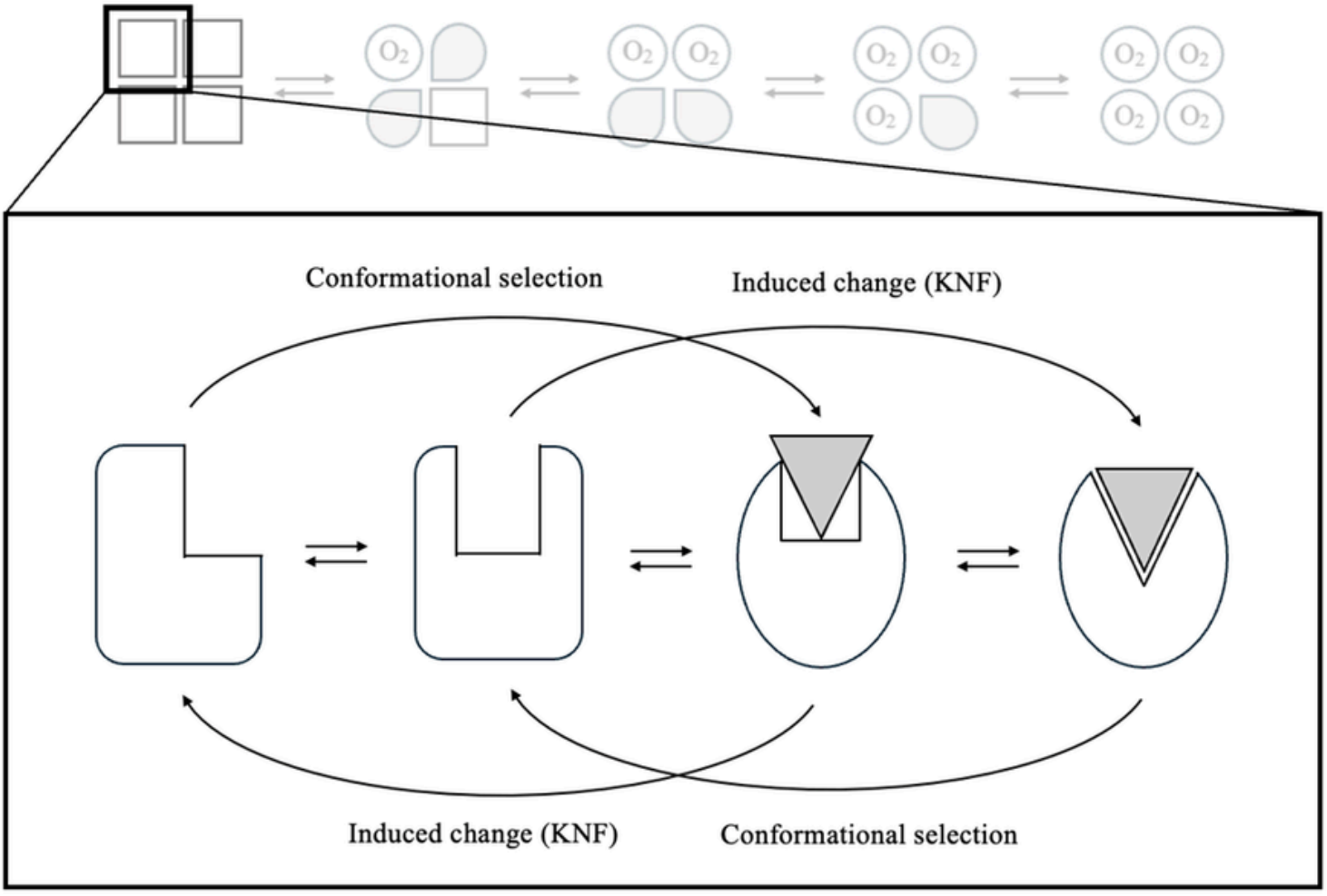


Figure 3. The relationship between the KNF and conformational selection dynamic ensemble models. This figure looks at one T-state subunit (shown in Figure 2's KNF model) in the context of dynamic ensembles as an example. In the dynamic ensembles model, this subunit will continuously interconvert between conformations, which in this figure, are the ground (T) or excited (R) states. Both the KNF and conformational selection models represent mechanisms where ligand binding is coupled to changes in conformation. These models appear equivalent when the reaction is considered in forward versus reverse directions. In the forward direction, induced fit (KNF) occurs when ligand binding precedes and drives the conformational change, while conformational selection occurs when the ligand binds a pre-existing conformation. In the reverse direction, a conformation formed after binding becomes a pre-existing state prior to unbinding, and a ligand-stabilized conformation undergoes conformational change after ligand release.

For example, IDP sequences are rich in charged and polar residues and glycine and proline, but they lack hydrophobic residues. The uncommon presence of hydrophobic residues is instead usually found within motifs that allow IDPs to recognize binding partners. As such, the hydrophobic effect is not a driving force for IDP dynamics, function, or structure. Functionally, this is advantageous since IDPs contain motifs for recognition by enzymes that carry out post-translational modifications, and the disorder and accessibility of these motifs allows IDPs to be rich in post-translational modifications and this, combined with their ability to bind many targets, facilitates IDPs' ability to act as protein hubs (i.e., central nodes within protein-protein interaction networks that accommodate many binding partners) (20, 21).

Furthermore, dynamic conformational sampling also facilitates the averaging of electrostatic fields within IDPs, so their binding or structural properties are more dependent on net charge or charge distributions. Notably, their high proportion of charged and polar residues enhances IDP solubility and even makes it so that some IDPs can function as proteinaceous detergents, depending on their sequence distribution (21, 23).

Moreover, conformational selection within IDPs does not only impact their function, but also their implication in health and disease. IDPs are involved in many signalling pathways, including the regulation of transcription, translation, and the cell cycle (20), and the involvement of disorder within these processes makes it so that IDPs are associated with many diseases, particularly those characterized by a loss of biological regulation (e.g., cancer), and those characterized by the formation of protein aggregates (e.g., Alzheimer's and Parkinson's diseases) (21). These pathological effects may arise from different imbalances within IDPs, with one being missense mutations that result in disorder-to-order transitions (21, 24). In this case, it becomes evident that disorder and a lack of native structure is beneficial for IDPs.

IDPs thereby serve as an example that principles governing protein folding and function are not dependent on a well-defined three-dimensional structure. In the case where no native fold exists, the biophysical principles still exist but are altered in their relative importance and driving force capabilities. Even with following the same dynamic laws but having no set three-dimensional structure, IDPs favour continuous interconversion between a spectrum of conformations within an ensemble, which provides them with their many functions, roles, and pathological risks. IDPs thus represent how biological function is not only encoded in structure, but also in conformational selection.

Non-spontaneous protein folding

Of course, the laws that govern spontaneous protein folding cannot be considered independently without also considering the cellular environment. Protein folding is a balance between thermodynamic stability and conformational flexibility. In the cell, where proteins tend to be marginally stable, this means that protein folding is in constant competition with protein aggregation. As aforementioned, aggregation (and other proteostatic processes) contributes to the development of disease, so molecular chaperones are essential in ensuring proper protein folding and the undertaking of certain native states (25, 26). In vivo, protein folding is complicated by its coupling to translation, the need for many newly translated proteins to be transported into subcellular compartments, and a crowded cellular environment (25).

In vivo folding after synthesis by the ribosome is energetically restricted. The ribosome's exit tunnel prevents large-scale folding and only allows for the formation of smaller tertiary structural elements. This makes it so that C-terminal amino acids cannot participate in more distal interactions essential for cooperative domain folding. Productive protein folding is thus delayed until entire protein domains are translated. This sequential exit and then folding of domains within a protein prevents non-native interactions between simultaneously folding domains, thus preventing the formation of unproductive structural intermediates. However, under stress conditions (e.g., heat shock, oxidative stress), proteins become destabilized (25). As aforementioned, under normal conditions, the free-energy surface that must be navigated by a folding protein is also rugged, creating kinetic traps that can become populated by partially folded states. These trapped intermediates tend to undergo hydrophobic collapse into disorganized globules or become "misfolded states" (if stabilized by non-native interactions), which tend to aggregate in a concentration-dependent manner (26).

Molecular chaperones, proteins that interact with and help in the folding/refolding or assembly of other proteins without being present in the final structure, are thus crucial interactors. Chaperones can be ATP-dependent, like the larger heat shock proteins Hsp70s or Hsp90s, or ATP-independent, like smaller heat shock proteins, depicting how chaperone-assisted folding is not purely equilibrium-driven like conformational ensembles, but active energetic processes (25, 26).

In brief, chaperones recognize non-native protein states after binding to hydrophobic segments. Binding to chaperones prevents aggregation and reduces the amount of freely folding intermediates, although transient release of the hydrophobic regions is necessary for folding to continue. Successful folding is achieved if the rate of folding is greater than the rate of chaperone rebinding (and thus the rate of aggregation). If folding is slower than either process, then the protein may be transferred to a different chaperone system or to degradation machinery (25). Overall, while the dynamics and biophysics of folding are still applicable to proteins in vivo, it can be more accurately described as a kinetically partitioned process, where environmental surroundings and interactors like chaperones bias the folding outcome towards a native state fold.

Conclusion

Both biophysical and biochemical frameworks discussed within this review highlight that protein folding and function cannot be accurately described as transitions between a small set of structures. Instead, they are a product of conformational selection from a continuum of states, which arise from dynamic conformational ensembles, governed by thermodynamics, entropy, and kinetics. When extended to cellular contexts, this further elucidates how biological function is not simple encoded in a single three-dimensional structure, but in population shifts and averages amongst a heterogenous pool of conformations. Overall, when taken together, these biophysical and biochemical insights connect folding and function, making them both consequences of the same continuous exploration process of an energy landscape within conformational ensembles.

Editorial Conflict of Interest Statement

Ishaan S. Goswami and Isra F. Omar are members of the OSURJ editorial team. Both authors were fully recused from all aspects of the editorial process for this manuscript, including reviewer selection, peer review, and final decision-making. The manuscript was handled independently by other members of the editorial board.

References

1. M. L. Boas, Mathematical Methods in the Physical Sciences (Wiley, Hoboken, NJ, 2006).
2. R. Phillips, J. Kondev, J. Theriot, H. G. Garcia, N. Orme, Physical Biology of the Cell (Garland Science, London; New York, NY, 2013), pp. 187-236, 311-354.
3. D. Schroeder, Introduction to Thermal Physics. (Oxford Univ Press, 2020), pp. 220-256.
4. J. Kuriyan, Boyana Konforti, D. Wemmer, The Molecules of Life: Physical and Chemical Principles (Garland Science, Taylor & Francis Group, New York, 2013), pp. 191-238, 220-256.
5. P. C. Nelson, D. S. Goodsell, K. Chen, S. Bromberg, Biological Physics: Energy, Information, Life (Chiliagon Science, Philadelphia, PA, 2020) pp. 184-376.
6. B. Alberts, R. Heald, A. Johnson, D. Morgan, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell (W. W. Norton & Company, New York, ed. 7, 2022), pp. 115-182.
7. N. Kresge, R. D. Simoni, R. L. Hill, The Thermodynamic Hypothesis of Protein Folding: the Work of Christian Anfinsen. J. Biol. Chem. 281, e11–e13 (2006). 10.1016/S0021-9258(19)56522-X
8. R. Zwanzig, A. Szabo, B. Bagchi, Levinthal's paradox. Proc. Natl. Acad. Sci. U.S.A. 89, 20–22 (1992).
9. S. S. Plotkin, J. Wang, P. G. Wolynes, Statistical mechanics of a correlated energy landscape model for protein folding funnels. J. Chem. Phys. 106, 2932–2948 (1997).
10. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins. 21, 167–195 (1995).
11. N. Onuchic, H. Nymeyer, A. E. García, J. Chahine, N. D. Socci, "The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios" in Protein folding mechanisms (Academic Press, 2000; vol. 53, pp. 87–152.
12. M. Montserrat-Canals, G. Cordara, U. Krengel, Allostery. Q. Rev. Biophys. 58, e5 (2025). 10.1017/S0033583524000209
13. L. Zhang, M. Li, Z. Liu, A comprehensive ensemble model for comparing the allosteric effect of ordered and disordered proteins. PLoS Comput. Biol. 14, e1006393 (2018). 10.1371/journal.pcbi.1006393
14. E. R. Henry, C. M. Jones, J. Hofrichter, W. A. Eaton, Can a two-state MWC allosteric model explain hemoglobin kinetics? Biochemistry 36, 6511–6528 (1997).
15. T. Yonetani, M. Laberge, Protein dynamics explain the allosteric behaviors of hemoglobin. Biochim. Biophys. Acta Proteins Proteom. 1784, 1146–1158 (2008).
16. M. F. Colombo, D. C. Rau, V. A. Parsegian, Protein solvation in allosteric regulation: A water effect on hemoglobin. Science256, 655–659 (1992).
17. M. M. Silva, P. H. Rogers, A. Arnone, A third quaternary structure of human hemoglobin A at 1.7-Å resolution. Journal of Biological Chemistry267, 17248–17256 (1992).
18. H. Frauenfelder, S. G. Sligar, P. G. Wolynes, The energy landscapes and motions of proteins. Science (1979). 254, 1598–1603 (1991).
19. T. R. Weikl, F. Paul, Conformational selection in protein binding and function. Protein Science 23, 1508–1518 (2014).
20. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. 16, 18–29 (2015).
21. J. D. Forman-Kay, T. Mittag, From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure 21, 1492–1499 (2013).
22. R. Trivedi, H. A. Nagarajaram, Intrinsically Disordered Proteins: An Overview. Int. J. Mol. Sci. 23, e14050 (2022). 10.3390/ijms232214050
23. R. W. Bailey, A. K. Dunker, C. J. Brown, E. C. Garner, M. D. Griswold, Clusterin, a binding protein with a molten globule-like region. Biochemistry40, 11828–11840 (2001).
24. V. Vacic, P. R. L. Markwick, C. J. Oldfield, X. Zhao, C. Haynes, V. N. Uversky, L. M. Iakoucheva, Disease-Associated Mutations Disrupt Functionally Important Regions of Intrinsic Protein Disorder. PLoS Comput. Biol. 8, e1002709 (2012). 10.1371/journal.pcbi.1002709
25. Y. E. Kim, M. S. Hipp, A. Bracher, M. Hayer-Hartl, F. Ulrich Hartl, Molecular chaperone functions in protein folding and proteostasis. Annu. Rev. Biochem.82, 323–355 (2013).
26. F. U. Hartl, A. Bracher, M. Hayer-Hartl, Molecular chaperones in protein folding and proteostasis. Nature 475, 324–332 (2011).

Review

Neurofilaments and Beyond: Multi-Modal Biomarkers and the Hidden Biology of ALS

Neurofilaments et au-delà : biomarqueurs multimodaux et biologie cachée de la SLA

Maïka Harvey[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: mharv082@uottawa.ca

Abstract | Résumé

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterized by progressive upper and lower motor neuron loss and death from respiratory failure within a few years of symptom onset. The marked clinical and biological heterogeneity of ALS has hindered the development of effective therapies, prompting intense interest in biomarkers that can improve diagnosis, prognostication, and trial efficiency. This review summarizes recent advances in fluid, imaging, genetic, and digital biomarkers of ALS, with a particular focus on neurofilament light chain as a leading candidate for prognostic and pharmacodynamic use. It highlights how current biomarkers, while promising, remain imperfect surrogates for underlying disease biology, and often expose fundamental gaps in understanding ALS pathogenesis. It argues that to fully realize the potential of biomarkers, the field must invest more heavily in mechanistic and longitudinal research that links biomarker dynamics to cellular pathways, genetics, and patient outcomes, ultimately enabling more rational, personalized therapeutic strategies.

La sclérose latérale amyotrophique (SLA) est une maladie neurodégénérative fatale caractérisée par la perte progressive des motoneurones supérieurs et inférieurs, entraînant le décès par insuffisance respiratoire quelques années après l'apparition des symptômes. L'hétérogénéité clinique et biologique marquée de la SLA a entravé le développement de thérapies efficaces, suscitant un intérêt considérable pour les biomarqueurs susceptibles d'améliorer le diagnostic, le pronostic et l'efficacité des essais cliniques. Cette revue résume les avancées récentes concernant les biomarqueurs liquides, d'imagerie, génétiques et numériques de la SLA, en mettant particulièrement l'accent sur la chaîne légère du neurofilament (NF-L), considérée comme un candidat majeur pour des applications pronostiques et pharmacodynamiques. Elle souligne que les biomarqueurs actuels, bien que prometteurs, demeurent des substituts imparfaits de la biologie sous-jacente de la maladie et révèlent souvent des lacunes fondamentales dans la compréhension de la pathogenèse de la SLA. Elle soutient que, pour exploiter pleinement le potentiel des biomarqueurs, le domaine doit investir davantage dans des recherches mécanistiques et longitudinales reliant la dynamique des biomarqueurs aux voies cellulaires, à la génétique et aux issues cliniques, afin de permettre l'élaboration de stratégies thérapeutiques plus rationnelles et personnalisées.

Keywords: Amyotrophic lateral sclerosis, biomarkers, neurofilament light chain, imaging, clinical trials, neurodegeneration, research priorities

Introduction

Amyotrophic lateral sclerosis (ALS) is the most common adult-onset motor neuron disease and presents with progressive weakness, muscle wasting, spasticity, and eventual respiratory failure (1). Median survival is approximately two to five years from symptom onset, although individual trajectories vary widely, from fulminant disease to slow progression over a decade (1). Currently approved disease-modifying therapies, including riluzole and edaravone, provide at best modest extensions in survival or slowing of functional decline, leaving an urgent unmet need for more effective treatments (2).

Heterogeneity in clinical presentation, genetics, and rate of progression complicates both diagnosis and therapeutic development (1). This variability reflects the multifactorial nature of ALS pathology, encompassing protein misfolding and aggregation, glutamate-mediated excitotoxicity, mitochondrial dysfunction, impaired RNA metabolism, and neuroinflammation, with the relative contribution of these mechanisms differing between patients. Diagnostic delay often exceeds 12 months, during which time patients may already have substantial motor neuron loss, reducing the window in which neuroprotective interventions could be effective (1). Delay is further compounded by the initially focal and frequently subclinical spread of neurodegeneration, the overlap of early symptoms with more common musculoskeletal or neuromuscular conditions, and the absence of specific biomarkers in routine clinical practice.

Traditional trial endpoints such as survival and functional scales are slow to change and are strongly influenced by baseline disease stage and progression rate, making large, lengthy, and expensive trials necessary to detect treatment effects (3). These challenges explain why robust biomarkers are central to the next phase of ALS research and why their development must go hand in hand with deeper study of disease mechanisms (3).

What counts as a biomarker in ALS?

Biomarkers are measured indicators of normal biological, pathogenic processes, or responses to therapeutic interventions (4). In ALS, candidate biomarkers span multiple categories, including diagnostic biomarkers, which help distinguish ALS from mimicking conditions, prognostic biomarkers, which predict disease courses independent of treatment, and predictive and pharmacodynamic biomarkers, which indicate likely response to therapy or capture biological effects of an intervention (4). For example, cerebrospinal fluid or blood neurofilament light chain (NfL) supports both diagnosis and prognosis, higher baseline NfL levels predict faster functional decline and shorter survival, and reductions in NfL or mutant protein levels in response to a targeted therapy can serve as pharmacodynamic readouts in clinical trials (1, 5-7).

Biomarkers can be derived from blood and cerebrospinal fluid (CSF), neuroimaging, electrophysiology, genetics, and digital or computational measures such as speech and movement patterns (5). The recent literature emphasizes that no single biomarker is sufficient; instead, multi-modal panels integrating fluid, imaging, electrophysiological, and digital measures are likely needed to capture the complexity of ALS and to link underlying biology with clinical outcomes (3, 5-7).

Fluid biomarkers: neurofilaments and beyond

Neurofilament light chain (NfL) has emerged as the most extensively validated fluid biomarker in ALS (8). NfL is a structural protein of myelinated axons, released into CSF and blood when axonal damage occurs. In ALS, CSF and blood NfL concentrations are several-fold higher than in healthy controls and ALS mimics, with levels often approximately 4–10 times those seen in controls, and higher NfL values are associated with faster functional decline and shorter survival (8). Ultrasensitive assays now allow reliable measurement of NfL in serum, which strongly correlates with CSF concentrations, making it practical for repeated sampling in large cohorts and clinical trials (9). Large cohort studies show that baseline serum NfL levels predict subsequent decline in ALS Functional Rating Scale–Revised (ALSFRS-R) scores and overall survival, indicating robust prognostic utility (9).

In addition, reductions in NfL in response to experimental therapies, most notably SOD1-lowering antisense oligonucleotides, have been used as pharmacodynamic readouts, contributing to regulatory decisions and supporting its use as a surrogate of target engagement and possibly neuroprotection (8, 10). Consensus papers now argue that there is compelling evidence for NfL as a prognostic and response biomarker in ALS therapy development, even though formal regulatory qualification is still in progress (10). However, NfL also illustrates the limits of our current understanding. Elevated NfL reliably signals axonal injury but does not specify which molecular pathways are active or why some patients with high NfL progress more slowly than others (6, 9). This gap underscores the need for mechanistic studies linking NfL kinetics to cell-type-specific pathology, genetics, and other biomarkers, rather than relying on NfL as a black-box marker of neurodegeneration.

Other protein and metabolic markers

Beyond neurofilaments, multiple protein and metabolic biomarkers are under investigation. CSF phosphorylated neurofilament heavy chain (pNfH) may have slightly better specificity for differentiating ALS from mimics in early disease, though serum pNfH has shown more variable performance (7). This likely reflects that pNfH in CSF more directly reflects neuroaxonal damage within the central nervous system, whereas blood concentrations are influenced by additional biological and analytical factors leading to greater variability in serum pNfH performance. The TAR DNA-binding protein 43 (TDP-43), the major protein component of pathological inclusions in most ALS cases involved in RNA processing, gene expressing and neuron function, can be detected in plasma and CSF, where higher levels appear to be associated with faster disease progression and may decline as disease advances, suggesting possible roles as both monitoring and prognostic markers (11).

Markers of muscle and systemic metabolism also show promise (11, 12). Plasma creatinine, a proxy for muscle mass and metabolism, declines over time in ALS and has shown lower variability than ALSFRS-R in some studies, supporting its potential as a surrogate endpoint for disease progression (11). Inflammatory proteins and complement components, as well as cytokine profiles reflecting peripheral immune activation, correlate with disease severity and rate of decline, but reproducibility across cohorts and platforms remains a major challenge (12). Recent work has also explored retroviral elements such as HERV-K (the Human Endogenous RetroVirus-family K, that can become reactivated in neurons in ALS), oxidative stress markers, and metabolic signatures, yet these remain at earlier stages of validation (3, 12). Together, these fluid biomarkers highlight a critical point: they reveal that ALS is not purely a motor neuron problem but a systemic disorder involving immune, metabolic, and muscle changes, and thus motivate further research into how these systems interact over the course of the disease.

Imaging biomarkers: seeing ALS in the brain and spinal cord

Neuroimaging offers a complementary window into ALS pathology (13). Structural MRI consistently demonstrates atrophy of the primary motor cortex and corticospinal tract, along with involvement of extra-motor regions in many patients, particularly those with cognitive or behavioral impairment (1, 13). Diffusion tensor imaging (DTI) metrics, such as reduced fractional anisotropy along the corticospinal tract and corpus callosum, correlate with disease severity and can differentiate ALS from

controls at the group level (13).

More advanced techniques, including functional MRI and PET, are uncovering alterations in functional connectivity and metabolic activity across motor and frontotemporal networks (13). PET tracers targeting neuroinflammation (e.g. TSPO ligands) or synaptic markers can detect microglial activation and synaptic loss in vivo, potentially serving as mechanistic and pharmacodynamic biomarkers (13). However, variability across scanners, protocols, and analysis pipelines, along with limited accessibility and high cost, has slowed translation into routine clinical practice and large multicenter trials (13).

The current generation of imaging biomarkers thus underscores both progress and limitation: they visualize ALS-related changes in living patients, but often lack disease specificity, standardization, and individual-level predictive power (13). Addressing these issues will require coordinated, longitudinal imaging studies linked tightly to fluid biomarkers, genetics, and detailed clinical phenotyping (3, 5, 13).

Genetic and digital biomarkers

Genetic testing is now an integral component of ALS evaluation in many centers, particularly for patients with a family history or early onset (1). Pathogenic variants in ALS-associated genes, such as the superoxide dismutase 1 gene (SOD1), which encodes the SOD1 enzyme that protects cells from oxidative stress, and C9orf7, a hexanucleotide repeat expansion, can inform counselling and, in selected contexts, guide therapeutic decisions as gene-targeted approaches emerge (1,3,8). Although only a minority of ALS cases have a clear family history and known pathogenic variants explain only a fraction of apparently sporadic disease, such genetic findings still represent an important entry point for precision therapies and mechanistic studies (1,3,5,6). From a biomarker perspective, these variants act as risk and stratification markers, defining biologically distinct subgroups for trial enrichment and mechanistic work (3,5-6)

Digital biomarkers, derived from speech analysis, kinematic assessments, wearable sensors, and smartphone-based tasks, represent a rapidly expanding frontier (5-14). Early studies suggest that quantitative measures of speech rate, articulatory precision, limb acceleration, and fine motor control can capture subtle changes before they are apparent on traditional scales, offering high-frequency, low-burden longitudinal monitoring (14). Nevertheless, most digital biomarkers remain in proof-of-concept stages, and their validation across devices, languages, and real-world environments is incomplete (5,14). Importantly, the emergence of genetic and digital biomarkers reinforces the need for intensive basic and translational work: genetic markers raise mechanistic questions about how specific variants drive degeneration, while digital outputs must be linked back to underlying neurobiology if they are to do more than describe surface-level function (14, 15).

Biomarkers in clinical trials: success and limitations

The integration of biomarkers into ALS clinical trials has accelerated over the past decade (3). NfL is now frequently included as a secondary or exploratory endpoint, enabling early readouts of biological effect that may precede changes in ALSFRS-R or survival (3,8,10). Systematic reviews emphasize that biomarkers can improve trial design by enabling the following three aspects (3). Enrichment of cohorts for faster progressors, increasing statistical power (3). Stratification based on biology (e.g. genetic status, baseline NfL) to reduce heterogeneity (3,8,10). Finally, adaptive designs where interim biomarker changes inform continuation, modification, or termination of trial arms (3).

Recent trials have used biomarker-based inclusion criteria, such as minimum NfL levels or specific genetic variants, and have interpreted NfL reductions as supportive evidence of target engagement and potential efficacy (8, 10). These developments represent a shift towards more efficient, biology-driven trials and demonstrate how biomarker research can directly impact therapy development (3).

At the same time, biomarker use in trials highlight several gaps (3-12). Many candidate biomarkers show promising associations but lack standardized assays, reference ranges, or clear thresholds for clinical decision-making (3, 12). Moreover, a biomarker's responsiveness to treatment does not guarantee that modifying the measured process will yield meaningful clinical benefit, especially in a multifactorial disease like ALS (15). Future work must therefore embed mechanistic studies and multi-marker panels within trials, so that biomarker dynamics are interpreted in the context of pathways and networks rather than isolated analytes (3,15).

How biomarkers reveal that ALS is under-studied

A central theme emerging from the biomarker literature is that our current tools, while increasingly sensitive, often raise more questions than they answer about ALS biology (6, 12). For instance, NfL clearly reflects axonal damage, yet current research lacks a detailed understanding of why some genetic backgrounds or environmental exposures yield higher or lower NfL at similar clinical stages, or how NfL trajectories differ between phenotypes such as bulbar-onset and limb-onset disease (6, 8).

Similarly, inflammation-related biomarkers reveal robust immune activation, but the relative contributions of peripheral versus central immune responses and the balance between neuroprotective, and neurotoxic glial states, remain incompletely defined (12). Imaging shows widespread network involvement beyond the motor system; however, a full understanding of how these changes relate to cognitive and behavioral symptoms, as well as non-motor manifestations such as pain, weight loss, and autonomic dysfunction, remains lacking (1, 13).

These gaps reflect a broader issue: ALS has historically received less research attention and funding than more prevalent neurodegenerative diseases, despite its devastating prognosis (1).

To move beyond descriptive biomarkers towards truly mechanistic, predictive, and actionable tools, the field must invest more in longitudinal, multi-modal cohort studies that track patients from pre-symptomatic or very early stages through the disease course, integrating fluid, imaging, genetic, and digital data (3, 5). Complementary experimental models are equally important: patient-derived cell-based systems and induced pluripotent stem cell–derived motor neurons and glia, as well as genetically engineered animal models, can be used to test how ALS-associated variants or environmental stresses influence candidate biomarkers and to dissect cell-type specific pathways that are difficult to resolve in vivo (5, 6, 15). More investments should as well be considered for basic research linking biomarker alterations to cellular pathways in neurons and glia, using models that incorporate patient-specific genetics (15). Diverse, global cohorts to understand how ancestry, environment, and healthcare access shape biomarker profiles and ALS risk are likewise important to consider (1,3). In this sense, biomarkers are not just tools for trials; they are signals pointing to where ALS remains under-studied and where deeper investigation is most urgently needed (3,12,15).

Conclusion

Biomarker research has transformed the conceptual and practical landscape of ALS by providing objective measures of neurodegeneration, prognosis, and therapeutic response, with neurofilament light chain at the forefront (8, 9). Fluid, imaging, genetic, and digital biomarkers collectively demonstrate that ALS is a multisystem, heterogeneous disease and are beginning to enable more efficient, biology-driven clinical trials (1, 3, 5, 13, 14). At the same time, the limitations and unanswered questions surrounding these biomarkers expose how much remains unknown about ALS pathogenesis, progression, and response to treatment (6, 8, 12, 13, 15). To fully realize the promise of biomarkers, future ALS research must emphasize longitudinal, multi-modal studies, mechanistic work that ties biomarker dynamics to cellular pathways, and inclusive cohorts that capture global diversity (3, 5, 15). Such efforts will not only refine biomarker panels but also deepen our understanding of ALS itself, ultimately supporting the development of rational, personalized combination therapies and, crucially, offering patients earlier diagnosis, better prognostication, and more effective treatment options.

References

1. O. Hardiman, A. Al-Chalabi, A. Chio, E. M. Corr, G. Logroscino, W. Robberecht, P. J. Shaw, D. Simmons, L. H. van den Berg, Amyotrophic lateral sclerosis. Nat. Rev. Dis. Primers 3, 17071 (2017).
2. M. K. Jaiswal, Riluzole and edaravone: A tale of two amyotrophic lateral sclerosis drugs. Med. Res. Rev. 39, 733–748 (2019).
3. T. Roy, A. Al-Chalabi, A. Iacoangeli, A. Al Khleifat, Biomarkers in ALS trials: From discovery to clinical utility. Front. Neurosci. 19, 1636303 (2025).
4. M. Benatar, K. Boylan, A. Jeromin, S. B. Rutkove, J. Berry, N. Atassi, L. Bruijn, ALS biomarkers for therapy development: State of the field and future directions. Muscle Nerve 53, 169–182 (2016).
5. R. S. Alshehri, A. M. Alhammad, A. Almaghrabi, F. Alabdali, R. Algarni, A. Alkhaldi, A review of biomarkers of amyotrophic lateral sclerosis: A pathophysiologic approach. Int. J. Mol. Sci. 25, 10900 (2024).
6. J. M. Al-Khayri, M. Ravindran, A. Banadka, C. D. Vandana, K. Priya, P. Nagella, K. Kukkemane, Amyotrophic lateral sclerosis: Insights and new prospects in disease pathophysiology, biomarkers and therapies. Pharmaceuticals 17, 1391 (2024).
7. K. E. Irwin, U. Sheth, P. C. Wong, T. F. Gendron, Fluid biomarkers for amyotrophic lateral sclerosis: A review. Mol. Neurodegener. 19, 9 (2024).
8. M. Benatar, J. Wuu, M. R. Turner, Neurofilament light chain in drug development for amyotrophic lateral sclerosis: A critical appraisal. Brain 146, 2711–2720 (2023).
9. M. Benatar, E. A. Macklin, A. Malaspina, M.-L. Rogers, E. Hornstein, V. Lombardi, J. Wuu, Prognostic clinical and biological markers for amyotrophic lateral sclerosis disease progression: Validation and implications for clinical trial design and analysis. eBioMedicine 108, 105323 (2024).
10. M. Benatar, L. W. Ostrow, J. W. Lewcock, F. Bennett, J. Shefner, R. Bowser, P. Larkin, L. Bruijn, J. Wuu, Biomarker qualification for neurofilament light chain in amyotrophic lateral sclerosis: Theory and practice. Ann. Neurol. 95, 211–216 (2024).
11. J. Lv, Y. Li, S. Shi, X. Xu, H. Wu, B. Zhang, Q. Song, Blood diagnostic and prognostic biomarkers in amyotrophic lateral sclerosis. Front. Mol. Neurosci. 17, 1423895 (2024).
12. E. Sturmey, A. Malaspina, Blood biomarkers in ALS: Challenges, applications and novel frontiers. Acta Neurol. Scand. 146, 375–388 (2022).
13. P. Bede, O. Hardiman, Lessons of ALS imaging: Pitfalls and future directions—a critical review. NeuroImage Clin. 4, 436–443 (2014).
14. J. D. Berry, S. Paganoni, K. Carlson, K. A. Burke, H. Weber, J. Staple, R. Wheeler, J. Chandler, K. Hurley, M. Kam, D. Bhatt, Design and results of a smartphone-based digital phenotyping study to quantify ALS progression. Ann. Clin. Transl. Neurol. 6, 873–881 (2019).
15. M. Benatar, C. McDermott, M. R. Turner, R. P. A. van Eijk, Rethinking phase 2 trials in amyotrophic lateral sclerosis. Brain 148, 1106–1111 (2025).

Review

Statistical Significance Reconsidered: The Role of Bayesian Methods in the Life Sciences

Reconsidération de la signification statistique : le rôle des méthodes bayésiennes dans les sciences de la vie

Ishaan S. Goswami[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: igosw085@uottawa.ca

Abstract | Résumé

Statistical reasoning underpins the interpretation of experimental results in the life sciences. For decades, frequentist hypothesis testing has dominated research practice, with the p-value serving as a benchmark of significance. Yet, p-values are often misunderstood and limited in what they reveal, and issues of statistical power further complicate inference. Bayesian methods offer an alternative framework that addresses several of these limitations. By incorporating prior knowledge and comparing the probabilities of competing hypotheses, Bayesian inference yields richer measures such as Bayes Factors, posterior probabilities, and credible intervals. However, Bayesian approaches also present challenges, including the subjectivity of priors and the computational intensity of complex models. This review synthesizes the conceptual and practical differences between frequentist and Bayesian approaches, with particular attention to their implications for experimental design, ethical considerations, and interpretation in modern biological research. It further examines how disciplinary norms, data complexity, and regulatory constraints shape methodological preferences across fields. Rather than positioning the two paradigms as competing frameworks, this work argues for statistical bilingualism as a necessary foundation for transparent, context-aware, and reproducible scientific inference.

———

Le raisonnement statistique sous-tend l'interprétation des résultats expérimentaux dans les sciences de la vie. Pendant des décennies, le test d'hypothèses fréquentistes a dominé la pratique de la recherche, la valeur p servant de référence significative. Pourtant, les valeurs p sont souvent mal comprises et limitées dans ce qu'elles révèlent, et les questions de puissance statistique compliquent encore davantage l'inférence. Les méthodes bayésiennes offrent un cadre alternatif qui répond à plusieurs de ces limitations. En incorporant des connaissances préalables et en comparant les probabilités d'hypothèses concurrentes, l'inférence bayésienne permet de produire des mesures plus riches telles que les facteurs bayésiens, les probabilités postérieures et les intervalles crédibles. Cependant, les approches bayésiennes présentent également des défis, notamment la subjectivité des a priori et l'intensité computationnelle des modèles complexes. Cette revue synthétise les différences conceptuelles et pratiques entre les approches fréquentistes et bayésiennes, en portant une attention particulière à leurs implications pour la conception expérimentale, les considérations éthiques et l'interprétation dans la recherche biologique moderne. Il examine également comment les normes disciplinaires, la complexité des données et les contraintes normatives influencent les préférences méthodologiques selon les domaines. Plutôt que de présenter les deux paradigmes comme des cadres concurrents, ce travail plaide en faveur du bilinguisme statistique comme fondation nécessaire pour une inférence scientifique transparente, consciente du contexte et reproductible.

Keywords: Bayesian inference; frequentist statistics; p-values; Bayes Factors; statistical power; experimental design; reproducibility; model comparison; hypothesis testing

Introduction

The importance of statistical significance in the life sciences

Interpreting whether experimental results are meaningful, rather than merely interesting, is a question of statistics. R.A. Fisher demonstrated the vitality of statistics through his "lady tasting tea" experiment. He outlined that "[e]very experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis" (1). The significance of data can be assessed through statistical tests, treating experimental outcomes as classical statistical events.

Frequentist Hypothesis Testing

In frequentist statistics, the p-value quantifies how compatible the observed data are with the null hypothesis, which is the assumption that there is no true effect. Specifically, it represents the probability of obtaining results as extreme as those observed, if the null hypothesis were true. A small p-value indicates that the observed results would be unlikely under the null. Conventionally, $p < 0.05$ is taken to indicate statistical significance, meaning that assuming the null hypothesis were true, results at least as extreme

as those observed would occur with probability less than 5%.

If one were to simulate experiments repeatedly under the null hypothesis, p-values would follow a uniform distribution between 0 and 1. In practice, however, this uniformity depends on correct model specification and data independence. When the null hypothesis is false, p-values cluster toward smaller values, and their distribution shape depends on factors such as effect size, sample size, and statistical power.

Power and sample size

In classical statistics, there are two types of errors: Type I (α) and Type II (β) errors. Type I error, or the false positive, occurs when the null hypothesis is rejected, even though it is actually true. It is typically set at 0.05, which means that if the null is true, it will be wrongly rejected 5% of the time. Type II error, or the false negative, occurs when one fails to reject the null hypothesis, even though the alternative is true. Type 2 error relies on the sample size, effect size, and variability of the data. A smaller sample size inherently increases the standard error, which makes the test less sensitive. Statistical power (1- β) is the probability that a test correctly detects an effect when one exists. It is often set at 80% (2). This means that high power means there is a lower chance of a Type 2 error.

Power analysis helps determine the sample size required (n) to detect a specified effect size with a desired significance level (α) and power (1-β). Power analysis should be conducted a priori to determine an adequate sample size, rather than after the fact. This ensures the study is designed to detect the effect of interest with sufficient sensitivity. An underpowered study is ethically questionable, as it may entail wasted reagents, time, and potentially unnecessary harm to animal and human participants without yielding interpretable results. If an underpowered study is not presented transparently, it could drive further research and policy towards a direction that is fundamentally inconclusive or misleading. As Gaskill and Garner (3) emphasize, underpowered studies not only waste resources but risk ethical violations in animal research. However, this does not mean that an excessively large sample should be collected. This can lead to the detection of statistically significant effects which may have no real biological or clinical significance. Both extremes undermine the ethics and efficiency of scientific and medical research.

Limitations of frequentist hypothesis testing

A small p-value doesn't prove that the null hypothesis is incorrect, and a large p-value does not prove the null hypothesis is correct. Rather, they indicate how comparable the observed data are with the assumption that there is no true effect between the variables. Gigerenzer (4) argues that statistical significance often substitutes for substantive reasoning. A large p-value may simply reflect insufficient data, insufficient statistical power, a poorly specified model, such as variance misspecification, or high variability in the data.

It is important to note that one can never accept the null hypothesis. Rather, you can only fail to reject it. As summarized by the American Statistical Association's official statement, a p-value near 0.05 alone provides only weak evidence against the null hypothesis and should not be treated as a bright-line criterion for discovery (5). That may not seem like a major semantic difference, but it has massive implications in how one can analyze data. The absence of evidence for the null hypothesis at high p-values is not evidence of absence. Thus, while p-values can hint at inconsistencies between the data and the null, it cannot confirm a competing hypothesis. This is a limitation that Bayesian methods are designed to address.

Bayesian Hypothesis Testing

In Bayesian statistics, the likelihood of data being obtained under both the null and alternative can be compared. Bayesian methods allow for the quantification of the degree of support for one hypothesis over another, rather than rejecting based on thresholds. This requires an assignment of prior probabilities and likelihoods under both hypotheses. These can be informed by how a scientist thinks the data would work or can be uniform priors if there are no strong assumptions. Probability is interpreted as a degree of belief, not just frequency. This allows one to update their beliefs about a hypothesis using the observed data. This is governed by Bayes' theorem:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \quad \text{(Eq. 1)}$$

$P(H|D) = Posterior: probability\ of\ hypothesis\ H\ given\ data\ D$

$P(D|H) = Likelihood: probability\ of\ data\ given\ H$

$P(H) = Prior: belief\ about\ H\ prior\ to\ data$

$P(D) = Marginal\ Likelihood: hard\ to\ compute, normalizing\ factor$

After defining two competing hypotheses, probability models must also be defined under both hypotheses. For example, under the null hypothesis, (H_0), it is assumed there is no true effect. In this case, the observed data X are modeled as arising from a normal distribution centered at zero:

$$X \sim \mathcal{N}(0, \sigma^2) \quad \text{(Eq. 2)}$$

Where σ^2 represents the variance of the observation. This formulation reflects the assumption that any observed deviation from zero is due solely to random noise, and that there is no mean difference.

The alternative hypothesis (H_1), on the contrary, allows for a non-zero effect size wherein the data are modeled as:

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{(Eq. 3)}$$

Where μ represents the true, unknown, effect. In Bayesian statistics, μ is treated as a random variable and assigned a prior distribution:

$$\mu \sim \mathcal{N}(0, \tau^2) \quad \text{(Eq. 4)}$$

This prior reflects the uncertainty about the effect size before observing the data, with τ^2 controlling the expected magnitude of plausible effects. The choice of prior thus influences the resulting inference, particularly in cases with limited data.

Now that the hypotheses' probability models have been defined, likelihoods for the data appearing under both hypotheses can be generated. P(Data | Null) finds the likelihood of the data under the null. This is a simple density evaluation. For P(Data | Alternative), one calculates the marginal likelihood, which is the evidence for the alternative. This is defined as:

$$P(D|H_1) = \int P(D|\theta) \cdot P(\theta|H_1) d\theta \quad \text{(Eq. 5)}$$

Under the alternative hypothesis, the marginal likelihood of the observed data is given by Equation (5). This quantity represents the probability of the data after integrating over all possible parameter values. In this expression, denotes the marginal likelihood of the data under the alternative hypothesis. The term is the likelihood, representing the probability of observing the data given a specific parameter value . The term is the prior distribution over the parameter , reflecting our uncertainty about its value under . The integral sums (or averages) the likelihood over all possible values of , weighted by the prior distribution.

This integrates the likelihood over the prior distribution of parameters under the alternative. The more parameters or flexibility a model has, the more "room" it has to explain data, even when the effect is not real. Hence, this integration favours simpler models such as those with fewer parameters or more constrained prior structures, in line with Ockham's Razor, which advises against unnecessary complexity unless the data demands it. However, the life sciences have ample systems that are complex, necessitating balance, not oversimplification.

The Bayes Factor (BF_{10}) compares how well two competing hypotheses predict the observed data. It is defined as the likelihood of the data under the alternative hypothesis to the data under the null. The Bayes Factor is not the posterior probability of a hypothesis; rather, it is a relative measure of the strength of the evidence for one hypothesis over the other. The Bayes Factor is defined as:

$$BF_{10} = \frac{P(Data|Alternative)}{P(Data|Null)} \quad \text{(Eq. 6)}$$

The Bayes Factor can be combined with prior odds to yield the posterior odds, which quantifies how much more likely one hypothesis is than another after also accounting the prior beliefs regarding data distribution.

There are certain heuristic guidelines to interpret the Bayes Factor. These are not hard statistical laws and are meant to act as tools to facilitate analysis. Jeffreys (6) introduced his scale to facilitate comparative analyses. Kass and Raftery (7) introduced stricter cutoffs than Jeffreys and their scale is widely cited in applied works. Lee and Wagenmakers (8) introduced their scale for applications in psychology. These are summarized by Taboga (9). According to these conventions, a Bayes Factor between 1 and 3 is considered anecdotal or barely worth mentioning, 3 to 10 indicates moderate or substantial evidence, 10 to 30 reflects strong evidence, 30 to 100 suggests very strong evidence, and values above 100 represent extreme or decisive support for the alternative hypothesis. It is necessary to emphasize that conclusions should always integrate experimental design, data quality, and prior plausibility

Once the Bayes Factor has been computed, it can be combined with prior odds to yield posterior probabilities, which quantifies the updated degree of belief in each hypothesis given the data. This requires specification of prior probabilities for both the null and alternative hypotheses. When interpreting results, thus, the Bayes Factor and the assumptions encoded in the prior must be explicitly stated. Unlike p-values, which measure compatibility with the null hypothesis, Bayes factors directly quantify the relative evidence that the data provides for one hypothesis compared with another.

For simple models that have known distributions and linear relationships, this process can be done in an analytical manner using closed-form solutions. This, however, is only the case when the prior and the likelihood distributions are conjugate or belong to the same distributional family as each other. One classic conjugate prior-likelihood pair is the normal likelihood with a normal prior which results in a normal posterior. This is used to model continuous variables with unknown mean but known variance. Another commonly used pair is the binomial likelihood with a beta prior which yields a beta posterior. However, the life sciences rarely follow these straightforward models. Oftentimes, Bayesian inference must rely on numerical approximation methods. The most widely used method is the Markov Chain Monte Carlo algorithm, which allows for sampling from the posterior distribution, even if it is computationally difficult to compute exactly. Markov Chain Monte Carlo algorithms, such as Metropolis-Hastings and Hamiltonian, generate a sequence of samples that approximate the posterior distribution via different means. By generating enough samples, it will converge to the true posterior.

In Bayesian statistics, the corollary to confidence intervals is the credible interval. A 95% credible interval means that given the observed data and the specified prior, there is a 95% probability that the parameter lies within that interval.

Posterior Predictive Distributions are a standard tool in the modern Bayesian workflow used to estimate the probability distribution of future or replicated data conditional on what has already been observed, as outlined by Gelman et al. (10). They are obtained by averaging predictions across all possible parameter

values, weighted by their posterior probability. Posterior predictive checks allow scientists not just to test a hypothesis, but to simulate new experiments. This is valuable in the life sciences, where replication is costly. This is also a part of the model criticism workflow, where experiments can be simulated a priori to evaluate expected performance.

The pitfalls of Bayesian priors

Bayesian inference is powerful as it can quantify updated belief in a hypothesis given the model and prior assumptions. The double-edged sword of Bayesian statistics is the prior model. The Bayesian inference about the alternative hypothesis can be biased and misleading based on the selection of the prior. This is one of the central critiques in Bayesian statistics. The Bayesian updating equation, which explains the reliance of the posterior upon the prior is defined as:

$$Posterior \propto Likelihood \times Prior \quad \text{(Eq. 7)}$$

The prior encodes what one already believes about a parameter or a hypothesis before seeing the data. There are three major ways a prior can be biased: I) it is too narrow, II) it is too wide, and III) it is too biased. If a prior is too narrow, the probability distribution is tightly centred around 0, which may fail to detect real effects that occur farther away in the tails. The data must show very strong deviation from zero to show that the alternative is more probable than the null. If the prior is too wide, like a uniform distribution, evidence for an effect may be inflated, even if the data generated is weak. This dilutes any evidence, as probability mass is spread across many implausible values, contrary to Ockham's Razor. If a prior is too biased, the posterior can be biased towards an expected effect due to the proportional relation between them. This may make weak data appear more supportive of a hypothesis than it really is. These three scenarios pose an important ethical dilemma to Bayesian statisticians. A model may be too conservative to accurately model complicated data, or too general to ascertain significant results. If one thinks the data will fit a certain distribution, they could fit the prior to make that hypothesis valid.

This can be overcome, though. A sensitivity analysis can be run, where the Bayesian model is run under multiple different priors. If the conclusion changes dramatically, then it shows the result is fragile. But it is important that these are run at the same time to avoid selective reporting of results that depend on a particular prior specification. Priors should exclude implausible values while still remaining sufficiently flexible to allow the data to inform the posterior.

Comparing the Two Schools & Conclusions

Having reviewed both the frequentist and Bayesian frameworks, it is valuable to directly compare the two. While both approaches rely on probability theory, they interpret probability very differently. This leads to different manners of designing experiments, analyzing the data, and drawing conclusions about the underlying biological systems.

Fundamentally, the two schools diverge on what probability means. Frequentists view probability as the long-run frequency of events. Here, randomness lies only in the data that is observed. Inference focuses on error rates across hypothetical repeated experiments. Bayesians, on the other hand, view probability as a degree of belief about uncertainty. Parameters are treated as random variables with their own probability distributions, and inference comes from updating prior beliefs about the distribution of results with new data to obtain a posterior distribution. With respect to reporting information, the frequentist would say with their confidence intervals, that if an experiment was conducted 1000 times, 95% of confidence intervals would contain the true success rate. A Bayesian would instead say that given their selected prior and the observed data, there is a 95% probability that the true success rate lies within a certain range.

The differences between the two schools are more evident in hypothesis testing. Classical frequentist statistics focuses on null hypothesis significance testing. As outlined above, this framework cannot directly provide information on the validity of the alternative hypothesis. The Bayesian school instead uses the Bayes Factor and posterior odds to directly compare two hypotheses given collected data with the incorporation of prior biological knowledge. This makes Bayesian statistics more robust, but also more open to being abused.

Uncertainty is treated in different ways in the two schools. Frequentists use the theoretical confidence interval, wherein if the experiment was repeated an infinite number of times, 95% of the intervals would contain the true parameter. This is often misunderstood to be a probability statement about the true value of the parameter itself. Bayesians instead use the credible interval, which directly represents the probability of the parameter; given the observed data and the prior, there is an X% chance that the parameter lies within the interval.

Experimental design also looks quite different under the two schools. In the frequentist tradition, researchers rely on a power analysis to ensure that an experiment is sensitive enough to detect a true effect. The emphasis is on controlling error probabilities over the long run, so that conclusions drawn across many repetitions of the same study are statistically reliable.

The Bayesian perspective allows experimental design to incorporate prior distributions and simulate posterior predictive outcomes before data collection begins. However, simulation-based planning can also be conducted within frequentist frameworks. This approach helps researchers anticipate how informative a study will likely be, given both prior knowledge and the planned sampling. Bayesian adaptive designs thus have the potential to stop trials early when strong evidence emerges, saving resources and reducing unnecessary exposure of animals or patients to ineffective treatments. However, regulatory adoption is

still ongoing. Bayesian adaptive trials in clinical research exemplify how probabilistic updating can improve both ethical and statistical efficiency (11).

Frequentist statistics is often praised for its simplicity and objectivity, since it relies on minimal prior assumptions. The use of standardized thresholds, such as the familiar $p < 0.05$, makes results easy to compare across studies and facilitates regulatory acceptance. In contrast, the strength of Bayesian methods lies in their flexibility, as they allow researchers to incorporate prior biological knowledge, update inferences as new data arrives, and generate richer outputs such as probabilities of hypotheses, model comparisons, and predictive distributions.

Yet these strengths come with trade-offs. Frequentist methods, while simple, cannot assign probabilities directly to hypotheses and can foster rigid reliance on significance thresholds, hence obscuring biological nuance. Bayesian approaches, on the other hand, are criticized for being sensitive to the choice of prior, which can introduce bias if not carefully justified, and for the computational intensity of techniques like Markov Chain Monte Carlo when applied to large or high-dimensional datasets. In practice, the perceived advantages and drawbacks of each framework often depend on the goals of the study, the complexity of the data, and the broader scientific or regulatory context.

Neither framework is universally superior. Frequentist methods remain dominant in the life sciences for historical, institutional, and regulatory reasons. At the same time, Bayesian methods are steadily gaining traction in areas such as genomics, neuroscience, epidemiology, and adaptive clinical trials, where richer probabilistic statements and the incorporation of prior knowledge offer clear advantages.

A pragmatic stance is to view the two schools as complementary tools rather than competitors. Frequentist methods provide standardized inference and well-defined error control, which support comparability and regulatory acceptance. Bayesian methods, on the other hand, allow deeper integration of prior information, adaptive study designs, and more intuitive probability statements. For modern life scientists, fluency in both schools offers a richer and more flexible toolkit for interpreting complex, noisy, and high-dimensional data.

One consideration is that the preference for frequentist or Bayesian methods across scientific disciplines is not purely statistical, but reflects the structure of the questions being asked, as well as institutional and practical constraints. Frequentist approaches remain dominant in fields such as clinical research, where regulatory frameworks demand standardized thresholds, controlled error rates, and comparability across studies (12). In contrast, Bayesian methods have gained traction in domains such as genomics, neuroscience, and systems biology, where data are high-dimensional, prior knowledge is informative, and adaptive or iterative experimental designs are advantageous (13-15). These differences are further reinforced by disciplinary training, computational accessibility, and historical precedent. As a result, methodological preference often emerges not from theoretical superiority, but from alignment with the epistemological and logistical demands of a field.

Statistical reasoning sits at the heart of modern life sciences. Regardless of the question being investigated, researchers must parse out what the data shows and determine how confident they are in those conclusions. Frequentist and Bayesian statistics offer different answers and methodologies to these challenges. However, they both provide unique insights and pose limitations that must be acknowledged, not disregarded. For life scientists, the key lesson is that statistics is not just a set of mechanical tests but a framework for reasoning under uncertainty. As biological research increasingly grapples with big data, complex models, and ethical constraints on experimentation, fluency in both frequentist and Bayesian methods will be essential to navigate the increasing complexity, uncertainty, and scale of modern life sciences research.

Editorial Conflict of Interest Statement

Ishaan S. Goswami is Co-Editor-in-Chief of the *University of Ottawa Science Undergraduate Research Journal*. He was fully recused from all aspects of the editorial process for this manuscript, including reviewer selection, peer review, and final decision-making. The manuscript was handled independently by other members of the editorial board.

References

1. R. A. Fisher, The Design of Experiments (Oliver & Boyd, 1935).
2. C. C. Serdar, M. Cihan, D. Yücel, M. A. Serdar, Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia Medica 31, 010502 (2021).
3. B. N. Gaskill, J. P. Garner, Power to the people: Power, negative results, and sample size. J. Am. Assoc. Lab. Anim. Sci. 59, 9–16 (2020).
4. G. Gigerenzer, Mindless statistics. J. Socio-Econ. 33, 587–606 (2004).
5. R. L. Wasserstein, N. A. Lazar, The ASA's statement on p-values: Context, process, and purpose. Am. Stat. 70, 129–133 (2016).
6. H. Jeffreys, Theory of Probability (Oxford Univ. Press, ed. 3, 1939).
7. R. E. Kass, A. E. Raftery, Bayes factors. J. Am. Stat. Assoc. 90, 773–795 (1995).
8. M. D. Lee, E.-J. Wagenmakers, Bayesian Cognitive Modeling: A Practical Course (Cambridge Univ. Press, 2014).
9. M. Taboga, Lectures on Probability Theory and Mathematical Statistics (Kindle Direct Publishing, 2021).
10. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, Bayesian Data Analysis (CRC Press, ed. 3, 2013).
11. D. A. Berry, Bayesian clinical trials. Nat. Rev. Drug Discov. 5, 27–36 (2006).

12. Center for Drug Evaluation and Research, “Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry” (U.S. Food and Drug Administration, 2020); https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.
13. E. C. Goligher, A. Heath, M. O. Harhay, Bayesian statistics for clinical research. The Lancet 404, 1067–1076 (2024).
14. H. P. Kärkkäinen, M. J. Sillanpää, Back to basics for Bayesian model building in genomic selection. Genetics 191, 969–987 (2012).
15. K. P. Kording, Bayesian statistics: Relevant for the brain? Curr. Opin. Neurobiol. 25, 130–133 (2014).

Review

Werner Syndrome: Symptoms, Hallmarks of Aging, Molecular Mechanisms and Therapeutic Pathway Inhibitors

Syndrome de Werner : symptômes, marques du vieillissement, mécanismes moléculaires et inhibiteurs des voies thérapeutiques

Tabassum Howlader[1*], Annie Xiang[1], Kamron Yunusov[1]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: thowl012@uottawa.ca

Abstract | Résumé

Werner Syndrome (WS) is a rare autosomal recessive progeroid disorder characterized by accelerated aging and the premature onset of age-related conditions, such as stunted growth, cataracts, cardiovascular disease, malignancies, sarcopenia, osteoporosis, and diabetes. Clinical disease manifestations typically begin in adolescence to early adulthood, and result in a reduced lifespan compared to healthy individuals. WS arises from loss-of-function mutations in the WRN gene, which encodes a RecQ family helicase that has implications in DNA repair, replication, and telomere maintenance. Deficiency in functional RecQ helicase activity results in dysfunction that links WS to the hallmarks of aging, including genomic instability, telomere attrition, and premature cellular senescence. To date, there is no cure for WS, with current therapies primarily focusing on disease management through inhibition of important proteins in aging- and stress-related signaling pathways, namely mTOR and p38 MAPK. These emerging approaches have shown promising results in cellular models, but have yet to be tested in human clinical studies. This review therefore examines WS as a potential model for understanding the mechanisms of aging, and the implications of existing findings for informing new therapeutic strategies.

———

Le syndrome de Werner (SW) est une maladie progéroïde autosomique récessive rare, caractérisée par un vieillissement accéléré et par l'apparition précoce de problèmes normalement associés à l'âge, comme un retard de croissance, des cataractes, des maladies cardiovasculaires, des cancers, la sarcopénie, l'ostéoporose et le diabète. Les manifestations cliniques commencent généralement entre l'adolescence et le début de l'âge adulte, puis entraînent une espérance de vie plus courte que chez les personnes en bonne santé. Le SW est causé par des mutations avec perte de fonction dans le gène WRN, qui code une hélicase de la famille RecQ impliquée dans la réparation de l'ADN, la réplication de l'ADN et le maintien des télomères. Lorsque l'activité fonctionnelle de cette hélicase RecQ est déficiente, plusieurs dysfonctionnements apparaissent et relient le SW aux grandes marques du vieillissement, notamment l'instabilité génomique, le raccourcissement des télomères et la sénescence cellulaire prématurée. À ce jour, il n'existe aucun traitement curatif pour le SW. Les thérapies actuelles visent surtout la prise en charge de la maladie, entre autres par l'inhibition de protéines importantes dans les voies de signalisation liées au vieillissement et au stress, comme mTOR et p38 MAPK. Ces approches émergentes ont donné des résultats prometteurs dans des modèles cellulaires, mais elles n'ont pas encore été testées dans des études cliniques chez l'humain. Cette revue examine donc le SW comme modèle potentiel pour mieux comprendre les mécanismes du vieillissement, ainsi que les implications des résultats actuels pour orienter de nouvelles stratégies thérapeutiques.

Keywords: Werner syndrome, WRN helicase, genomic instability, telomere attrition, cellular senescence, hallmarks of aging, mTOR signaling, p38 MAPK.

Introduction

Aging is accompanied by the gradual accumulation of DNA damage and genomic instability, which has emerged as a hallmark of the aging process (1). Failure of genomic maintenance systems can accelerate damage accumulation, contributing to age-related pathologies such as cancer (1). The importance of maintaining genome integrity in aging is underscored by rare hereditary disorders in which mutations in DNA repair pathways lead to premature aging syndromes (1). In these conditions, an excess of unrepaired DNA damage manifests as early-onset degenerative changes, directly linking genomic instability to accelerated aging. Premature aging (progeroid) syndromes serve as powerful biological models for studying aging mechanisms. These rare diseases recapitulate many features of normal aging and provide natural opportunities to probe how specific molecular defects drive aging phenotypes.

Among the DNA instability-driven progeroid disorders, Werner syndrome (WS) is a quintessential example. WS is a rare autosomal recessive disorder caused by loss-of-function (LOF) mutations in the WRN gene, first identified in 1996, which encodes a 1,432

amino acid RecQ-family DNA helicase (2). WS prevalence is estimated to be 1:1,000,000–1:10,000,000, with high prevalence in certain populations, such as Japan, due to founder effects that increase the frequency of disease-causing variants within a population (3). WRN is the only human gene whose LOF mutations give rise to WS, and loss of WRN function leads to genomic instability and a characteristic premature-aging phenotype (4). Affected individuals typically develop normally until adolescence, then begin to show early signs of aging such as loss and greying of hair, skin atrophy, and bilateral cataracts, followed by the onset of age-related diseases in early adulthood (5).

Despite its rarity, WS remains highly relevant to aging research because it directly links loss of genome maintenance to accelerated tissue decline. Recent studies have renewed interest in WS by highlighting both its mechanistic importance and its therapeutic potential. These findings not only provide a foundation for potential interventions in WS, but also offer broader insights into targeting aging-related pathways.

Discussion

Characteristics of the Disease

Age of onset and disease progression

WS is an autosomal recessive progeroid disorder with onset in adolescence or early adulthood. Typical development usually occurs until puberty, then WS patients fail to undergo the pubertal growth spurt, resulting in short stature as the first indicator of disease (5). Characteristic aging-like features begin to appear in their twenties, and by their thirties, most WS patients display greying and loss of hair, as well as an aged facial appearance with thin, atrophic skin (6). Bilateral cataracts also develop in most cases, often requiring surgery by their late twenties and into their early thirties (7). As the disease progresses through the third and fourth decades of life, patients start to accumulate multiple age-related pathologies including diabetes mellitus, osteoporosis, atherosclerosis, myocardial infarctions, and malignant tumours (6,7). As a result of this accelerated multi-system deterioration, WS patients display markedly shorter lifespans, with the median age at death of approximately 54 years (7).

Clinical features

WS patients display a broad spectrum of clinical features affecting dermatologic, endocrine, musculoskeletal, and metabolic systems. Patients develop scleroderma-like skin changes with tight, thin skin and a loss of subcutaneous fat, giving an aged and pinched facial appearance. Skin pigment changes and skin ulcers are also common, with approximately 40% of patients displaying skin ulcers in the distal one-third of the lower legs (8). These persistent ulcers are often linked to extensive calcification of soft-tissues and can lead to serious complications such as infections, with approximately 15% of WS patients eventually requiring a foot or lower leg amputation (6). In terms of musculoskeletal features, disproportionately short stature and premature osteoporosis are also observed in WS patients (7). Patients tend to have slender extremities with decreased muscle mass as a result of sarcopenia or can exhibit truncal obesity due to visceral fat accumulation (9). Mobility can be impaired due to joint contractures and tendon/soft-tissue calcifications.

Endocrine and metabolic features are also hallmarks of WS. Type 2 diabetes mellitus and dyslipidemia occur in a majority of patients by their thirties (6). In one study cohort, over 60% of participating WS patients had impaired glucose tolerance or diabetes along with hypertriglyceridemia (6). This diabetic tendency occurs despite a relatively low body mass index (BMI), due to severe insulin resistance and altered fat distribution as a result of the relative loss of peripheral subcutaneous fat. Hypogonadism is also observed in WS patients. These individuals experience gonadal atrophy and infertility, leading to premature menopause in women and testicular failure in men (5). Thyroid function is also impaired, with patients at risk of developing thyroid neoplasms (10). Other systemic features can include a high-pitched hoarse voice, cataracts, senile dementia, and brain atrophy; although not of Alzheimer's type (5).

Common complications and causes of mortality: Due to the accelerated nature of aging in WS, patients are predisposed to many complications that typically occur much later in normal aging. Cancer and cardiovascular disease are the most common causes of mortality in WS patients (10). Overall cancer risk is dramatically increased, but unlike ordinary age-related cancer, WS shows a distinct tumour spectrum. Uncommon tumour types are predominant, specifically those of mesenchymal or endocrine origins. Soft-tissue sarcomas, osteosarcomas, melanomas, and thyroid carcinomas account for about 57% of all reported WS cancers, compared to roughly 2% of cancers in an age-matched general population study conducted by Goto et al. (11). Meningiomas, leukemias, and bone malignancies are also overrepresented in WS (10). Since multiple primary tumours can occur in the same patient, this high incidence plays a role in significantly reducing the lifespan of WS patients. Tumour predisposition in WS patients can primarily be attributed to genome instability and telomere dysfunction, which together promote the accumulation of oncogenic mutations and chromosomal aberrations that drive malignant transformation. In addition to cancer, atherosclerotic cardiovascular disease is another primary fatal complication (10). Due to the aggressive and premature damage caused to artery walls, WS patients are at a significantly higher risk of myocardial infarction and death (10). Together, cancer and cardiovascular disease account for the majority of mortality cases in WS patients, leading to the relatively low median life expectancy (7). Less common causes of mortality include stroke, infection complications, and organ failure secondary to diabetes. Overall, the multisystem involvement in WS leads to an elevated mortality risk well before the seventh decade of life, in stark contrast to the normal aging population.

Molecular mechanism of Werner Syndrome

WRN deficiency in WS

The gene linked to the onset of WS is known as WRN, which encodes a 1432 amino acid protein product (2). The disease phenotype is caused by loss-of-function mutations in the WRN gene that are most commonly associated with small indels, premature stop codons or splice-site mutations, leading to truncated transcripts that are undetectable in patient-derived cells independent of mutation type (7,12). The resulting null alleles, which do not produce functional WRN proteins or have measurable enzymatic activity, are the direct cause of WS.

Helicase and exonuclease activity in WRN

At the molecular level, the WRN gene belongs to the RecQ family of helicases (13). Biochemical characterization confirmed that its unwinding activity occurs in the 3' → 5' direction and requires the presence of ATP. In a study by Moser et al (12), no WRN protein or immune-precipitable helicase activity was detected in patient cell lines, while WRN heterozygous cells showed reduced amounts of WRN protein and helicase activity (12). This suggests that WRN deficiency could directly impact DNA unwinding, impairing replication fork progression and stability. Consequently, stalled replication forks are more susceptible to collapsing, leading to increased DNA damage and genomic instability, which contributes to the onset of WS (14). Intrinsic 3' → 5' exonuclease activity was also found in the N-terminal region of WRN (15). Its activity is physically and functionally separate from the helicase domain of WRN. This domain preferentially catalyzes degradation of specific DNA secondary structures, such as bubbles, loops or stem-loops (16). Its exonuclease activity is further stimulated during non-homologous end joining (NHEJ), where WRN physically interacts with factors such as Ku70/80, DNA-PKcs and DNA ligase IV/XRCC4 to process DNA ends during repair (17). Loss of functional WRN protein extends the amount of time that cells need to complete the cell cycle, establishing the necessity of WRN for effective fork restart following DNA damage and replication arrest (18). WS patient and WRN-knockout cells also exhibited elevated levels of DNA breaks, suggesting genomic instability and/or impairment of DNA repair systems (19). The helicase and exonuclease functions of the WRN protein contribute to its functions during this DNA repair process, as well as during replication and recombination, however, how their combined dysfunction contributes to the disease phenotype of WS remains unclear (20).

Significance of WRN in telomere maintenance: The WRN protein also facilitates maintenance of telomeres, which contributes to the normal progression of aging. WRN's helicase activity is stimulated through its interaction with TRF2, a critical telomere maintenance protein that stabilizes T- and D-loop secondary structures in telomeres (21). In cooperation with Bloom syndrome helicase (BLM) and replication protein A (RPA) (21), WRN plays a role in unwinding this telomeric DNA such that it can be efficiently replicated (22). Telomeres replicated by lagging-strand synthesis were found to be affected by loss of WRN helicase activity, exhibiting loss of telomeres from individual sister chromatids. The shortening of telomeres can contribute to cellular senescence and trigger apoptosis by activating DNA damage responses, leading to tissue dysfunction (23). Loss of these sequences therefore directly contributes to the premature aging phenotype of WS, suggesting the importance of the WRN helicase activity.

Hallmarks of aging

The clinical and molecular features of WS align with several recognized hallmarks of aging, reinforcing the value of WS as a translational model for studying mechanisms that may also contribute to normal physiological aging. Central to WS pathology is genomic instability, one of the key hallmarks of aging. The WRN gene encodes a RecQ helicase involved in DNA repair and replication. Loss-of-function (LOF) mutations in WRN lead to DNA replication stress and accumulation of DNA damage. WS cells exhibit a mutator phenotype supported by increased chromosomal aberrations and large DNA deletions (10). WS also highlights the hallmark of telomere attrition. Studies have shown that telomere dysfunction is a major driver of premature aging in WS, with WRN-deficient fibroblasts displaying defective telomere lagging-strand synthesis, resulting in telomere shortening and instability (10). Since telomere attrition synergizes with WRN loss to induce premature aging, the consequence of this is an early onset of cellular senescence in proliferative tissues. WS patient cells have a reduced replicative lifespan and express senescence markers (e.g. DEC1 and p16) at a younger age than typical aging (10). This combination of accumulated DNA damage, telomere dysfunction, and cellular senescence likely drives many WS clinical features, mirroring the processes of normal aging but at an accelerated scale. In terms of other hallmarks, WS patients' metabolic disorders reflect deregulated nutrient sensing. Mitochondrial dysfunction has also been postulated, although the studies show no significant changes in mtDNA mutations from control groups (10). WS demonstrates how disruption of key aging mechanisms can generate an aging-like phenotype. Studies of WS have therefore provided important insight into human aging, supporting the idea that the hallmarks of aging play a causal role in disease development.

Pathway-based inhibitory therapeutic approaches

Therapeutic challenges

To date, there is no cure for WS, and effective treatment remains elusive. Despite extensive research, the mechanisms and genetic programs underlying human aging, including those disrupted in WS, remain unclear (24). As explored, WRN deficiency affects multiple cellular pathways across many tissues, making it challenging to design therapies that can safely and effectively correct the defect throughout the body. In addition, the genomic instability characteristic of WS cells raises concerns regarding the long-term safety and durability of gene-editing approaches, thus studies investigating genetic correction strategies in WS are relatively scarce, with most studies being limited to cellular models and pathway-based interventions (25). However, recent biochemical studies have shown that the WS protein functions in

several DNA metabolic pathways, highlighting the complexity of its role in cellular processes (24). As a result, current clinical management of WS centers on treating disease manifestations, preventing secondary complications, and screening for acquired diseases common to WS. Specifically, novel pathway-specific therapies have been proposed, most of which involve the inhibition of aging and stress-related pathway proteins.

mTOR Inhibitors

One notable therapy is the inhibition of mechanistic Target of Rapamycin (mTOR) (26). The mTOR signaling pathway is critical to aging and senescence as it contributes to the formation of protein aggregates, oxidative damage, and defective mitochondria and vacuoles, which are all hallmark features of cellular aging. Importantly, when it is inhibited in eukaryotic model organisms, prolonged lifespan is observed, reinforcing its role as a driver of aging (26).

Moreover, the mTOR signaling pathway has been directly linked to the cellular pathology of WS, particularly when the mechanistic Target of Rapamycin Complex 1 (mTORC1) complex is hyperactivated. In a study by Talaei et al. (27), WS-phenotype fibroblasts derived from WS patients showed increased phosphorylation of mTOR at Ser448, which reflects activation of mTORC1, and subsequently the cascade that leads to the phosphorylation of the downstream effector ribosomal protein kinase S6. These results were determined through Western blot analysis of enhanced expression of phosphorylated mTOR and S6 proteins in WS fibroblasts relative to normal fibroblasts, demonstrating elevated mTORC1 signaling in WS. This increased mTORC1 signaling was accompanied by the observation of typical hallmarks of aging in WS fibroblasts including excessive intracellular protein aggregation, elevated oxidative damage, and abnormal cellular morphology.

Talaei et al (27) treated these WS fibroblasts with hydrogen sulfide (H_2S) in the form of sodium hydrosulfide (NaHS), and compared results to WS fibroblasts treated with rapamycin, a known pharmaceutical inhibitor of mTORC1. Following treatment, mTOR pathway activity was assessed by Western blot analysis measuring levels of phosphorylated mTOR at Ser2448 and protein S6. In both NaHS and rapamycin-treated WS fibroblasts, it was observed that there were reduced levels of phosphorylated mTOR and S6, indicating lower mTORC1 signaling. In addition to this reduced marker expression, they observed the restoration of WS fibroblasts towards a more normal morphology as compared to the abnormal morphology accompanied by untreated WS fibroblasts (27).

p38 MAPK Inhibitors

Another emerging therapy is the inhibition of the p38 mitogen-activated protein kinase (p38 MAPK) pathway. In WS fibroblasts, senescence is accelerated and cells have a reduced cellular replicative life span (28-30), and therefore act similarly to fibroblasts of elderly individuals. It has been observed that WS is not only indicative of accelerated aging, but is also stress-associated, through the observation of slow growth rates, an extended cell cycle, and a normal aged morphology (31-34). The morphology of young WS cells has also been compared to fibroblasts subjected to premature senescence from oncogenic ras or arsenic, (35) both of which activate the p38 MAPK pathway through map kinase kinase 6 (MKK6) (35, 36). This leads to stabilization and upregulation of the kinase inhibitor p21, resulting in cell-cycle arrest. This is hypothesized to play a role in the premature senescence observed in WS (36-38).

In examining if the p38 MAPK pathway contributes to the cellular WS phenotype, Davis et al. (36) observed markers of stress-related senescence in patient-derived WS fibroblasts, including increased expression of phosphorylated p38 that is correlated with enhanced activation of p38 MAPK, as well as consequent enhanced expression of p21. The authors also treated WS fibroblasts with SB20358, a cytokine-suppressive, anti-inflammatory that inhibits p38 activity. Following treatment, they observed a reduction of p38 and p21 activity, and a reversion of the WS phenotype cell morphology back to the morphology of young normal fibroblasts, which was not observed in young normal cells. These results suggested that p38-mediated growth arrest is likely a contributing factor to premature aging in WS cells (39).

Implications of exploring WS therapeutics for aging research

The therapeutic potential of mTOR and p38 MAPK inhibitors can be applied beyond WS itself and have broader implications for aging research. Since these pathways are involved in processes such as cellular senescence and stress responses, the improvements seen in WS cells suggest that some aging-related changes may be slowed or modified through targeted therapies. Because WS displays many features of normal aging over a shorter period of time, it can also serve as a useful model for studying potential anti-aging treatments. As a result, research on pathway inhibitors in WS may help guide the future development of therapies for age-related diseases in the general population. However, although mTOR and p38 MAPK inhibitors have shown promising results in WS cellular models, they have not been tested in human clinical studies (25). The aforementioned findings should thus be applied to the progression of WS therapeutics with caution, as they are limited to fibroblast models and may not accurately predict therapeutic success in humans.

Conclusion

Werner Syndrome is a disease characterized by accelerated aging in humans, stemming from mutations in the WRN gene that produce null alleles. The resulting WRN deficiency negates its helicase and exonuclease activities, which impact the efficacy of DNA replication, repair, recombination and telomere maintenance in WS patients, though the impact of the combined loss of these functions remains relatively unexplored. Although no preventative cures currently exist, pathway inhibitors, such as mTOR and p38 MAPK inhibitors, have demonstrated effective reduction in senescence within experimental cellular systems. Therapies for Werner Syndrome directly targeting telomeric dysfunction and

compensatory models for loss of WRN function remain largely unexplored in humans, highlighting crucial gaps that require further investigation.

References

1. M. D'Amico, T. T. Li, D. Wylie, G. Wang, K. M. Vasquez, Aging alters genomic instability at endogenous mutation hotspots in mice. Sci. Rep. 15, (2025).
2. E. Yu, J. Oshima, Y. H. Fu, E. M. Wijsman, F. Hisama, R. Alisch, S. Matthews, J. Nakura, T. Miki, S. Ouais, G. M. Martin, J. Mulligan, G. D. Schellenberg, Positional cloning of the Werner's syndrome gene. Science 272, 258–262 (1996).
3. F. Coppedè, The epidemiology of premature aging and associated comorbidities. Clin. Interv. Aging 8, 1023–1032 (2013).
4. P. L. Opresko, G. Sowd, H. Wang, The Werner syndrome helicase/exonuclease processes mobile D-loops through branch migration and degradation. PLoS One 4, e4825 (2009).
5. M. Goto, Hierarchical deterioration of body systems in Werner's syndrome: Implications for normal ageing. Mech. Ageing Dev. 98, 239–254 (1997).
6. M. Koshizaka et al., Time gap between the onset and diagnosis in Werner syndrome: a nationwide survey and the 2020 registry in Japan. Aging (Albany NY) 12, 24940–24956 (2020).
7. S. Huang et al., The spectrum of WRN mutations in Werner syndrome patients. Hum. Mutat. 27, 558–567 (2006).
8. H. Peng et al., Case report: A novel WRN mutation in Werner syndrome patient with diabetic foot disease and myelodysplastic syndrome. Front. Endocrinol. 13, 918979 (2022).
9. M. Yamaga et al., Werner syndrome: a model for sarcopenia due to accelerated aging. Aging (Albany NY) 9, 1738–1744 (2017).
10. M. Tokita et al., Werner syndrome through the lens of tissue and tumour genomics. Sci. Rep. 6, 32038 (2016).
11. M. Goto, R. W. Miller, Y. Ishikawa, H. Sugano, Excess of rare cancers in Werner syndrome (adult progeria). Cancer Epidemiol. Biomarkers Prev. 5, 239–246 (1996).
12. M. J. Moser et al., WRN helicase expression in Werner syndrome cell lines. Nucleic Acids Res. 28, 648–654 (2000).
13. M. D. Gray et al., The Werner syndrome protein is a DNA helicase. Nat. Genet. 17, 100–103 (1997).
14. S. Mukherjee et al., Werner Syndrome Protein and DNA Replication. Int. J. Mol. Sci. 19, 3442 (2018).
15. S. Huang et al., The premature ageing syndrome protein, WRN, is a 3'–5' exonuclease. Nat. Genet. 20, 114–116 (1998).
16. J. C. Shen, L. A. Loeb, Werner syndrome exonuclease catalyzes structure-dependent degradation of DNA. Nucleic Acids Res. 28, 3260–3268 (2000).
17. R. Kusumoto et al., Werner protein cooperates with the XRCC4-DNA ligase IV complex in end-processing. Biochemistry 47, 7548–7556 (2008).
18. J. M. Sidorova et al., The RecQ helicase WRN is required for normal replication fork progression after DNA damage or replication fork arrest. Cell Cycle 7, 796–807 (2008).
19. P. Pichierri et al., Werner's syndrome protein is required for correct recovery after replication arrest and DNA damage induced in S-phase of cell cycle. Mol. Biol. Cell 12, 2412–2421 (2001).
20. W.-H. Cheng, V. A. Bohr, Diverse dealings of the Werner helicase/nuclease. Sci. Aging Knowl. Environ. 2003, PE22 (2003).
21. P. L. Opresko et al., Telomere-binding protein TRF2 binds to and stimulates the Werner and Bloom syndrome helicases. J. Biol. Chem. 277, 41110–41119 (2002).
22. F. d'Adda di Fagagna, Living on a break: cellular senescence as a DNA-damage response. Nat. Rev. Cancer. 8, 512–522 (2008).
23. L. Crabbe et al., Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. Science 306, 1951–1953 (2004).
24. K. Kyng, D. L. Croteau, V. A. Bohr, Werner syndrome resembles normal aging. Cell Cycle 8, 2323 (2009).
25. J. Oshima, J. M. Sidorova, R. J. Monnat Jr., Werner syndrome: Clinical features, pathogenesis and potential therapeutic interventions. Ageing Res. Rev. 33, 105–114 (2017).
26. Fischbach et al., mTOR signaling controls protein aggregation during heat stress and cellular aging in a translation- and Hsf1-independent manner. J. Biol. Chem. 301, 108172 (2025).
27. F. Talaei et al., Hydrogen sulfide restores a normal morphological phenotype in Werner syndrome fibroblasts, attenuates oxidative damage and modulates mTOR pathway. Pharmacol. Res. 74, 34–44 (2013).
28. Salk et al., Systematic growth studies, cocultivation, and cell hybridization studies of Werner syndrome cultured skin fibroblasts. Hum. Genet. 58, 310–316 (1981).
29. T. O. Tollefsbol, H. J. Cohen, Werner's syndrome: An underdiagnosed disorder resembling premature aging. AGE 7, 75–88 (1984).
30. R. G. Faragher et al., The gene responsible for Werner syndrome may be a cell division "counting" gene. Proc. Natl. Acad. Sci. U.S.A. 90, 12030–12034 (1993).
31. Y. Fujiwara et al., A retarded rate of DNA replication and normal level of DNA repair in Werner's syndrome fibroblasts in culture. J. Cell Physiol. 92, 365–374 (1977).
32. F. Takeuchi et al., Prolongation of S phase and whole cell cycle in Werner's syndrome fibroblasts. Exp. Gerontol. 17, 473–480 (1982).
33. M. Poot et al., Impaired S-phase transit of Werner syndrome cells expressed in lymphoblastoid cell lines. Exp. Cell Res. 202, 267–273 (1992).
34. M. Rodríguez-López et al., Asymmetry of DNA replication fork progression in Werner's syndrome. Aging Cell 1, 30–39 (2002).
35. W. Wang et al., Sequential activation of the MEK–extracellular signal-regulated kinase and MKK3/6–p38 pathways mediates oncogenic ras-induced premature senescence. Mol. Cell. Biol. 22, 3389–3403 (2002).
36. Q. Deng et al., High intensity ras signaling induces premature senescence by activating p38 pathway in primary human fibroblasts. J. Biol. Chem. 279, 1050–1059 (2004).
37. G. Y. Kim et al., The stress-activated protein kinases p38α and JNK1 stabilize p21(Cip1) by phosphorylation. J. Biol. Chem. 277, 29792–29802 (2002).

38. R. Haq et al., Constitutive p38HOG mitogen-activated protein kinase activation induces permanent cell cycle arrest and senescence. Cancer Res. 62, 5076–5082 (2002).
39. T. Davis et al., Prevention of accelerated cell aging in Werner syndrome using a p38 mitogen-activated protein kinase inhibitor. J. Gerontol. A Biol. Sci. Med. Sci. 60, 1386–1393 (2005).

Commentary

Exercise Under the Microscope: How Physical Activity Reshapes Aging Muscle Biology

Exercice sous la loupe: comment l'activité physique remodele la biologie musculaire du vieillissement

Zoha Fatima[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: zfati011@uottawa.ca

Abstract | Résumé

Sarcopenia is an age-related musculoskeletal disease characterized by the progressive loss of skeletal muscle mass, strength, and regenerative capacity, contributing to frailty, impaired mobility, and reduced independence in older adults. While pharmacological therapies targeting muscle degeneration continue to emerge, exercise remains one of the most effective and accessible interventions for preserving musculoskeletal health during aging. This commentary discusses the study by Liu et al., which used single-cell transcriptomic analyses to demonstrate that exercise modulates aging tissue biology through both direct and indirect mechanisms. Exercise directly influenced inflammatory signaling, stem cell communication, and regenerative pathways while indirectly improving broader physiological processes associated with healthy aging and functional recovery. By reducing inflammatory signatures and restoring intercellular communication within aged skeletal muscle, exercise may improve tissue regeneration beyond its effects on muscle mass alone. These findings emphasize the growing importance of rehabilitation-based strategies for preserving mobility, improving recovery, and maintaining functional independence in aging populations.

———

La sarcopénie est une maladie musculosquelettique liée à l'âge caractérisée par une perte progressive de masse musculaire squelettique, de force et de capacité régénératrice, contribuant à la fragilité, à la mobilité réduite et à une diminution de l'indépendance chez les personnes âgées. Bien que les thérapies pharmacologiques ciblant la dégénérescence musculaire continuent d'émerger, l'exercice reste l'une des interventions les plus efficaces et accessibles pour préserver la santé musculosquelettique au fil du vieillissement. Ce commentaire aborde l'étude de Liu et al., qui a utilisé des analyses transcriptomiques unicellulaires pour démontrer que l'exercice module la biologie tissulaire vieillissante par des mécanismes directs et indirects. L'exercice a directement influencé la signalisation inflammatoire, la communication avec les cellules souches et les voies régénératrices tout en améliorant indirectement les processus physiologiques plus larges associés au vieillissement sain et à la récupération fonctionnelle. En réduisant les signatures inflammatoires et en restaurant la communication intercellulaire au sein des muscles squelettiques vieillissants, l'exercice peut améliorer la régénération tissulaire au-delà de ses seuls effets sur la masse musculaire. Ces résultats soulignent l'importance croissante des stratégies basées sur la rééducation pour préserver la mobilité, améliorer la récupération et maintenir l'indépendance fonctionnelle des populations vieillissantes.

Keywords: Aging, sarcopenia, exercise, skeletal muscle regeneration, rehabilitation, inflammaging, inflammation, muscle stem cells, tissue homeostasis.

Introduction

Aging is associated with a progressive decline in skeletal muscle mass, strength, and regenerative capacity, contributing to impaired mobility, increased risk of falls, frailty, and reduced quality of life(1, 2). Sarcopenia is a musculoskeletal disease characterized by the age-associated decline in skeletal muscle tissue and affects approximately 10–16% of the aging population and represents an increasing burden on healthcare and rehabilitation systems (3–5). While muscle degeneration has traditionally been associated with the loss of muscle mass alone, aging also impairs the ability of skeletal muscle to effectively regenerate following injury or physiological stress (4). Reduced regenerative capacity contributes not only to muscle weakness, but also to prolonged recovery following injury, illness, or hospitalization, increasing the risk of long-term dependence in older adults (6, 7). As populations continue to age, there is growing interest in identifying therapeutic strategies capable of preserving muscle and improving regeneration to maintain functional independence in older adults.

Current therapeutic approaches aimed at combating age-related muscle degeneration include physical, nutritional and pharmacological interventions (8). Drug-based methods target specific molecular pathways such as PI3K/Akt/mTOR and myostatin signalling, involved in inflammation, protein turnover,

and muscle growth (9–11). Myostatin inhibitors, for example, aim to increase muscle mass by blocking negative regulators of muscle hypertrophy (10, 12). However, despite significant research efforts, there are currently no United States Food and Drug Administration (FDA)-approved drugs specifically for the treatment of sarcopenia (8). Furthermore, many of these pharmacological therapies focus on isolated mechanisms and may fail to address the broader physiological and regenerative changes that occur during aging. On the other hand, exercise remains the most effective strategy and is capable of simultaneously influencing multiple biological processes associated with muscle health, including inflammatory signalling, stem cell activity, metabolism, vascularization, and tissue remodelling (13). Beyond its effects on muscle mass and strength, exercise may act as a biologically active regulator of tissue regeneration and healthy aging (11). Furthermore, understanding the molecular pathways through which exercise promotes regeneration may also contribute to the development of exercise mimetics for individuals unable to participate in regular physical activity. Addressing functional decline through rehabilitation, and preventative strategies is therefore crucial for preventing the effects of aging, preserving independence, and reducing long-term strain on healthcare and rehabilitation systems (14).

The recent study by Liu et al. provides important mechanistic insight into how exercise reshapes aging stem cell environments and improves tissue homeostasis across multiple organ systems (15). Unlike previous studies that primarily examined exercise-induced changes in muscle mass or bulk tissue signalling, this study provides cell-type specific insight into how exercise remodels aged stem cell niches across multiple tissues. Using integrative single-cell transcriptomic analyses, the authors examined over 435,000 single cells collected from skeletal muscle, neural stem cell compartments, hematopoietic stem and progenitor cells, and peripheral immune cells in both young and aged mice subjected to voluntary exercise. Their findings demonstrate that exercise significantly reduces inflammatory signalling across multiple stem cell compartments while restoring more youthful intercellular signalling within aged skeletal muscle. This work suggests that exercise has both direct and indirect effects, as it not only preserves muscle tissue superficially but also alters the regenerative landscape of aging tissue at the molecular and cellular levels. The use of single-cell transcriptomics allows for a more comprehensive understanding of how exercise influences intercellular communication and inflammatory signalling during aging. In this context, studies such as Liu et al. are valuable because they not only reinforce the importance of exercise in promoting regeneration but also provide mechanistic insight into the signalling pathways and intercellular communication networks involved, potentially helping identify targets that could mimic the regenerative effects of exercise.

Results

One of the most significant findings of this study is the observation that exercise reverses several age-associated inflammatory changes. Research shows aging is commonly characterized by chronic low-grade inflammation, often referred to as “inflammaging,” which contributes to impaired stem cell function and reduced regenerative capacity (16). In this paper, increased inflammatory signalling pathways involving interferon gamma (IFNγ), interferon alpha (IFNα), interleukin-6 (IL-6), and tumour necrosis factor alpha (TNFα) were observed across multiple aged cell populations. Exercise reduced the expression of many of these pathways, although these effects varied between cell types, suggesting that physical activity may partially restore a more regenerative environment within aged tissue by reducing inflammation.

Since chronic inflammation can disrupt communication between regenerative cell populations, these findings are particularly relevant in the context of skeletal muscle regeneration. Effective muscle repair depends not only on muscle stem cells (MuSCs), also known as satellite cells, but also on the surrounding stem cell niche and communication between neighbouring cell populations (17). Aging disrupts this cellular communication network, impairing the coordinated signalling required for efficient tissue repair (18). Liu et al. showed that exercise restored several of these disrupted communication pathways, particularly interactions involving monocytes, macrophages, fibro-adipogenic progenitors (FAPs), and myofibers. The restoration of these networks suggests that exercise may improve regeneration by reshaping the broader regenerative landscape rather than acting only on muscle fibers themselves.

This concept shows one of the major advantages of exercise-based interventions compared to many pharmacological therapies. Rather than targeting a single pathway, exercise influences multiple interconnected biological systems at the same time. In this sense, exercise may be viewed not simply as a method of maintaining physical fitness, but as a form of regenerative rehabilitation capable of modifying the biological processes underlying aging itself.

Moving Forward

The broader implications of these findings extend beyond skeletal muscle biology alone. Exercise-induced reductions in inflammatory signalling were also observed within neural stem cell compartments and hematopoietic tissues, supporting the idea that exercise promotes systemic tissue homeostasis during aging. Liu et al. also showed changes within hematopoietic stem cell niches, suggesting that exercise may influence regeneration through broader immune-related mechanisms. This may be particularly relevant in the context of sarcopenia, as immune aging contributes to chronic low-grade inflammation and altered immune signalling that can impair muscle stem cell function and regenerative capacity (19). These findings support the idea that the benefits of exercise extend beyond muscle growth and cardiovascular fitness, influencing regenerative processes throughout the body. Understanding how exercise modulates these interconnected systems may help identify future therapeutic

targets that mimic or enhance the regenerative benefits of physical activity.

Importantly, the findings presented by Liu et al. also emphasize the importance of rehabilitation-based approaches in aging populations. Preservation of mobility and functional independence remains one of the primary goals of geriatric healthcare and rehabilitation. Exercise-based rehabilitation strategies not only improve strength and balance but may also directly influence the molecular mechanisms responsible for tissue maintenance and repair. By improving regenerative capacity and reducing chronic inflammation, exercise may help delay frailty, reduce fall risk, and improve recovery following injury or illness in older adults.

Despite these promising findings, limitations and challenges remain. Exercise responsiveness varies considerably between individuals and may be a less suitable option in older populations with advanced frailty or chronic disease (20). Adherence to long-term exercise programs may also present difficulties due to mobility limitations, pain, or comorbidities (21). Furthermore, while exercise improves regenerative signaling, it may not fully reverse all age-associated declines in muscle function.

Additional limitations arise from the translational nature of the study itself. Many of the reported findings are based on transcriptomic changes observed in murine models, which may not fully reflect functional outcomes in humans. Although altered gene expression profiles suggest improved regenerative potential, transcriptional changes do not always directly translate into enhanced tissue repair, mobility, or sustained functional recovery. Furthermore, because the mice engaged in voluntary exercise, exercise levels varied and may not accurately represent exercise behavior in aging human populations. Future research should therefore build on this paper's findings to focus on identifying clinically effective exercise regimens and exploring how exercise-based rehabilitation strategies may be integrated with pharmacological or regenerative therapies to optimize recovery outcomes in older adults.

Overall, the study by Liu et al. provides important molecular insight into how exercise modulates aging tissue biology. Using single-cell transcriptomic analyses, the authors demonstrated that exercise reduces inflammatory signaling, restores intercellular communication, and improves stem cell niche environments across multiple aged tissues. These findings suggest that exercise promotes regeneration both directly and indirectly: directly by influencing molecular pathways involved in inflammation, stem cell function, and tissue repair, and indirectly by improving broader physiological processes associated with healthy aging and functional recovery. Understanding how exercise influences these signaling networks may help guide the development of more effective rehabilitation and pharmacological strategies for preserving mobility, improving recovery, and maintaining functional independence in aging populations.

References

1. E. Volpi, R. Nazemi, S. Fujita, Muscle tissue changes with aging. Curr Opin Clin Nutr Metab Care 7, 405–410 (2004).
2. L. J. Melton, S. Khosla, C. S. Crowson, M. K. O'Connor, W. M. O'Fallon, B. L. Riggs, Epidemiology of sarcopenia. J Am Geriatr Soc 48, 625–630 (2000).
3. W. J. Evans, Skeletal muscle loss: cachexia, sarcopenia, and inactivity. The American Journal of Clinical Nutrition 91, 1123S-1127S (2010).
4. Y. Zhang, A. Desai, S. Y. Yang, K. B. Bae, M. I. Antczak, S. P. Fink, S. Tiwari, J. E. Willis, N. S. Williams, D. M. Dawson, D. Wald, W.-D. Chen, Z. Wang, L. Kasturi, G. A. Larusch, L. He, F. Cominelli, L. Di Martino, Z. Djuric, G. L. Milne, M. Chance, J. Sanabria, C. Dealwis, D. Mikkola, J. Naidoo, S. Wei, H.-H. Tai, S. L. Gerson, J. M. Ready, B. Posner, J. K. V. Willson, S. D. Markowitz, TISSUE REGENERATION. Inhibition of the prostaglandin-degrading enzyme 15-PGDH potentiates tissue regeneration. Science 348, aaa2340 (2015).
5. S. Yuan, S. C. Larsson, Epidemiology of sarcopenia: Prevalence, risk factors, and consequences. Metabolism 144, 155533 (2023).
6. J. G. Tidball, I. Flores, S. S. Welc, M. Wehling-Henricks, E. Ochi, Aging of the immune system and impaired muscle regeneration: A failure of immunomodulation of adult myogenesis. Experimental Gerontology 145, 111200 (2021).
7. V. E. Arango-Lopera, P. Arroyo, L. M. Gutiérrez-Robledo, M. U. Perez-Zepeda, M. Cesari, Mortality as an adverse outcome of sarcopenia. The Journal of nutrition, health and aging 17, 259–262 (2013).
8. J. Y. Kwak, K.-S. Kwon, Pharmacological Interventions for Treatment of Sarcopenia: Current Status of Drug Development for Sarcopenia. Ann Geriatr Med Res 23, 98–104 (2019).
9. D. J. Glass, "PI3 Kinase Regulation of Skeletal Muscle Hypertrophy and Atrophy" in Phosphoinositide 3-Kinase in Health and Disease, C. Rommel, B. Vanhaesebroeck, P. K. Vogt, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010; http://link.springer.com/10.1007/82_2010_78)vol. 346 of Current Topics in Microbiology and Immunology, pp. 267–278.
10. J. Rodriguez, B. Vernus, I. Chelh, I. Cassar-Malek, J. C. Gabillard, A. Hadj Sassi, I. Seiliez, B. Picard, A. Bonnieu, Myostatin and the skeletal muscle atrophy and hypertrophy signaling pathways. Cell. Mol. Life Sci. 71, 4361–4371 (2014).
11. C. Liu, X. Wu, G. Vulugundam, P. Gokulnath, G. Li, J. Xiao, Exercise Promotes Tissue Regeneration: Mechanisms Involved and Therapeutic Scope. Sports Med - Open9, 27 (2023).
12. X. Zhu, S. Topouzis, L. Liang, R. L. Stotish, Myostatin signaling through Smad2, Smad3 and Smad4 is regulated by the inhibitory Smad7 by a negative feedback mechanism. Cytokine 26, 262–272 (2004).
13. L. Huang, M. Li, C. Deng, J. Qiu, K. Wang, M. Chang, S. Zhou, Y. Gu, Y. Shen, W. Wang, Z. Huang, H. Sun, Potential Therapeutic Strategies for Skeletal Muscle Atrophy. Antioxidants 12, 44 (2022).
14. V. Gianfredi, D. Nucci, F. Pennisi, S. Maggi, N. Veronese, P. Soysal, Aging, longevity, and healthy aging: the public health approach. Aging Clin Exp Res37, 125 (2025).

15. L. Liu, S. Kim, M. T. Buckley, J. M. Reyes, J. Kang, L. Tian, M. Wang, A. Lieu, M. Mao, C. Rodriguez-Mateo, H. D. Ishak, M. Jeong, J. C. Wu, M. A. Goodell, A. Brunet, T. A. Rando, Exercise reprograms the inflammatory landscape of multiple stem cell compartments during mammalian aging. Cell Stem Cell 30, 689-705.e4 (2023).
16. C. Franceschi, P. Garagnani, P. Parini, C. Giuliani, A. Santoro, Inflammaging: a new immune–metabolic viewpoint for age-related diseases. Nat Rev Endocrinol14, 576–590 (2018).
17. H. Yin, F. Price, M. A. Rudnicki, Satellite Cells and the Muscle Stem Cell Niche. Physiological Reviews 93, 23–67 (2013).
18. M. Thorley, A. Malatras, W. Duddy, L. Le Gall, V. Mouly, G. Butler Browne, S. Duguez, Changes in Communication between Muscle Stem Cells and their Environment with Aging. JND 2, 205–217 (2015).
19. J. G. Tidball, I. Flores, S. S. Welc, M. Wehling-Henricks, E. Ochi, Aging of the immune system and impaired muscle regeneration: A failure of immunomodulation of adult myogenesis. Experimental Gerontology 145, 111200 (2021).
20. M. O. Whipple, E. N. Schorr, K. M. C. Talley, R. Lindquist, U. G. Bronas, D. Treat-Jacobson, Variability in Individual Response to Aerobic Exercise Interventions Among Older Adults. Journal of Aging and Physical Activity26, 655–670 (2018).
21. S. Nayab, M. Bilal Elahi, The Impact of Exercise Interventions on Pain, Function, and Quality of Life in Patients With Osteoarthritis: A Systematic Review and Meta-Analysis. Cureus, doi: 10.7759/cureus.74464 (2024).

Galileo as Physicist and Polemicist: A Commentary on an Unpublished Mid-Twentieth-Century Pedagogical Essay

Galilée en tant que physicien et polémiste : commentaire sur un essai pédagogique inédit du milieu du XXe siècle

Ishaan S. Goswami[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: igosw085@uottawa.ca

Abstract | Résumé

This commentary examines an unpublished mid-20th-century pedagogical essay on Galileo Galilei authored by physicist and educator Mam Chand Jain. Written within a classical physics teaching tradition that integrated history, philosophy, and mathematical derivation, the essay presents Galileo not merely as a source of foundational laws but as a persuasive figure whose scientific arguments effected a broader worldview shift. The original work combines biographical reflection, mechanical analysis of Galileo's contributions to classical mechanics and astronomy, and a mathematical–philosophical treatment of the so-called Galileo–Plato problem concerning the "abode of God."

In this commentary, the essay is situated within its historical and educational context, highlighting both its conceptual strengths and its limitations in light of subsequent historiography and physical theory. Particular attention is given to its pedagogical method, which foregrounds the development of physical intuition through historical argumentation rather than rote formalism. This commentary argues that such integrative approaches remain valuable for contemporary scientific pedagogy, offering an integrated framework for understanding how scientific knowledge is discovered, communicated, and taught.

———

Ce commentaire examine un essai pédagogique inédit du milieu du XXe siècle portant sur Galileo Galilei, rédigé par le physicien et éducateur Mam Chand Jain. Rédigé dans une tradition d'enseignement de la physique classique intégrant l'histoire, la philosophie et la dérivation mathématique, cet essai présente Galilée non seulement comme une source de lois fondamentales, mais aussi comme une figure persuasive dont les arguments scientifiques ont provoqué un changement plus large de vision du monde. Le texte original combine une réflexion biographique, une analyse mécanique des contributions de Galilée à la mécanique classique et à l'astronomie, ainsi qu'un traitement mathématique et philosophique du problème dit de Galilée–Platon concernant la « demeure de Dieu ».

Dans ce commentaire, l'essai est replacé dans son contexte historique et éducatif, en mettant en évidence à la fois ses forces conceptuelles et ses limites à la lumière de l'historiographie et des théories physiques ultérieures. Une attention particulière est accordée à sa méthode pédagogique, qui met en avant le développement de l'intuition physique à travers l'argumentation historique plutôt que par un formalisme répétitif. Ce commentaire soutient que de telles approches intégratives demeurent pertinentes pour la pédagogie scientifique contemporaine, en offrant un cadre unifié pour comprendre comment le savoir scientifique est découvert, communiqué et enseigné.

Keywords: Galileo Galilei; History and Philosophy of Science; Physics Pedagogy; Classical Mechanics; Scientific Revolution; Scientific Communication

Introduction

This article presents a commentary on and transcription of an unpublished pedagogical essay (See Appendix A) on Galileo Galilei written by Mam Chand Jain, a physicist and educator trained in the mid-20th century in both the Indian and British academic traditions. The manuscript was begun in the 1950s and formally typed in 1972. It was intended as a pedagogical synthesis combining a first-principles exposition of Galileo's discoveries in classical mechanics with historical analysis of the Scientific Revolution. Jain's paper serves as a biographical portrait of Galileo

Figure 1. Mam Chand Jain (1923–2014), physicist and educator. Photograph taken during his teaching career in Leader, Saskatchewan (c. 1966). Author of the 1972 paper "Galileo."

as a personality; a technical exposition of Galileo's physics and astronomy; and a mathematical and philosophical analysis of the Galileo-Plato problem - the "place of God" problem. This paper reflects a period when physics education emphasized conceptual foundations, history, and philosophy, rather than highly formalized problem-set-driven instruction. Galileo is not portrayed as the founder of several axioms of classical physics, but rather as a pedagogical example in the history of science. The essay reproduced here is a record of how science was once taught as a unified intellectual enterprise.

Beyond its pedagogical value, the essay contains a detailed mathematical analysis of Galileo's speculative claim regarding a single "abode of God" from which planets were set into motion. Using Keplerian relations and later Newtonian reasoning, Jain demonstrates that Galileo's proposal is internally inconsistent and systematically examines successive reformulations of the problem by later thinkers. This analysis does not refute Galileo's reasoning within his historical context, nor does it challenge the legitimacy of his speculative premise. Rather, it examines the assumptions underlying his proposal using physical laws and relations developed after his time.

Mam Chand Jain (1923–2014), B.Sc., M.Sc., P.G.C.E., was trained in physics and mathematics, earning a bachelor's degree in chemistry, physics, and mathematics and a master's degree in physics with a focus on astronomy and spectroscopy, followed by postgraduate teaching qualifications from the Institute of Education, University of London. He spent over three decades teaching physics and mathematics at the secondary and post-secondary levels in India, Ethiopia, Ghana, the United Kingdom, and Canada. His academic formation emphasized classical mechanics, mathematical derivation, and the historical development of physical theory, reflecting a pedagogical tradition in which history and philosophy of science were integral to physics instruction. This pedagogical orientation is reflected in one of the texts Jain used both as a student and later as a source for his essay: The Pre-University and Intermediate Physics by Basu and Chatterji (2), in which dense mathematical derivations are interwoven with historical context and epistemic reflection

Conceptual Strengths of the Essay

Galileo was a persuader, not just an experimenter
Before the notion that science communication was a noteworthy pursuit, Galileo was a polemicist who fought for his interpretation of the physical universe, as opposed to previous Aristotelian notions of cosmology. Galileo was not afraid of confrontation, and rather, was willing to engage in fierce debates with his opponents, ultimately leading to his trial and condemnation by the Catholic Church. Galileo understood that science had to advance socially, not just logically, as his arguments were opposed to the neo-Platonic and Aristotelian cosmology accepted by the church. Long before 'science communication' became a formal discipline, this essay recognizes that Galileo's success depended as much on persuasion as on proof.

Physics as worldview shift
The essay understands Galileo as dismantling Aristotelian cosmology, not merely adding new observational data. While heliocentrism is now accepted as a physical fact, in Galileo's era it represented a profound epistemic and existential rupture. The displacement of the Earth from the center of the cosmos challenged deeply held assumptions about order, purpose, and humanity's place in creation. The emotional and intellectual shock of heliocentrism is thus foregrounded as a necessary component of scientific transformation, not a peripheral consequence.

Serious engagement with mechanics
Jain explicitly outlined the mechanical derivation of Galileo's observations of projectile motion, inertia, inclined plane reasoning, and refinements under Newtonian dynamics. This paper was structured as such to situate the mathematics and physics in the world in which it was discovered. The mechanics are presented as arguments constructed in response to physical phenomena – not laws divorced from the reality they were meant to describe.

Historical and Scientific Limitations

The essay necessarily reflects the historiographical and scientific context of its time. Subsequent scholarship has questioned the historicity of the Tower of Pisa experiment (3). As well, this paper does not address Galileo's findings in the context of later developments in thermodynamics and relativity. However, for a physicist trained in the classical tradition, these limitations can be understood as products of historical circumstance. Read as a pedagogical document rather than a contemporary research contribution, the essay remains internally coherent and intellectually rigorous within its intended framework.

Pedagogical Significance Today

Modern science education is increasingly fragmented, with students learning techniques before understanding meaning, and how the axioms and theorems they assume to be true were discovered. This essay models an integrated approach, wherein history, philosophy, mathematics, and experimentation are all acknowledged in order to provide a deeper intuition behind these discoveries. Revisiting such pedagogical texts is a reminder that science is not merely a body of results, but a way of thinking about nature, evidence, and truth.

The essay exemplifies an educational philosophy in which the history and philosophy of science functioned as tools for developing physical intuition. By tracing Galileo's arguments, errors, and rhetorical strategies, students were encouraged to understand why modern mechanics emerged, not merely how to apply its equations.

Editorial Conflict of Interest Statement

The author of the original Galileo essay, Mam Chand Jain, was a physicist and educator and the maternal grandfather of the a author Ishaan S. Goswami. This commentary was handled independently by the OSURJ editorial team, and the author was not involved in the review or acceptance decision. Publication of the original essay and related archival materials was undertaken with the permission of the estate of Mam Chand Jain.

References

1. M. C. Jain, Galileo (unpublished manuscript, 1972).
2. N. Basu, J. Chatterjee, The Pre-University and Intermediate Physics (H. Chatterji & Co., Calcutta, 1954). Available from: https://archive.org/details/dli.ernet.240843
3. M. Segre, Galileo, Viviani and the tower of Pisa. Stud. Hist. Philos. Sci. A 20, 435–451 (1989).

Appendix A: Transcribed Text of “Galileo” (1972)

This appendix presents a faithful transcription of Mam Chand Jain’s unpublished 1972 pedagogical essay “Galileo.” The text has been re-typed for clarity and accessibility. Spelling, punctuation, and mathematical notation have been preserved wherever possible. No substantive content has been altered. The manuscript reflects the conventions of physics pedagogy and historical interpretation of its time. Mathematical notation, terminology, and historiographical claims are preserved as originally presented. The text is provided as a historical document and should be read within its pedagogical and temporal context.

Galileo
August 4, 1972
M. C. Jain

Today, I will honor a great man who was born just four centuries ago, a man whose achievements have touched on almost all fields of human effort. Galileo was a powerful, passionate figure, a man who dominated every room and every discussion he entered. His excitement over the new world he saw opening up, and his blistering intolerance of those who would not see it as he did, break through in every page of his writings. As we read his letters we can hardly help falling into step behind his banners, we laugh with him, lock swords with his enemies, rejoice at his triumphs. His is no clean cut world of concepts and theorems, but a brawling world of clashes and schemes, no place for a scientist, I would say. But a sort of place in which someone who has set himself to tearing down an ageold system of thought and replacing it with a new is likely to feel at home.

Galileo loved the fray. Not for him the laborious hours of observation of a Tycho; not for him the endless calculations and curve fittings of a Kepler. He was a man with a vision of the way, the universe had to be, and a talent for communicating that vision to others. None knew better than Galileo that, whereas theorems have to be proved, people had to be persuaded. To someone as strongly convinced of his own righteousness as Galileo was proof is at best secondary. But when someone has a message as novel and as far reaching as Galileo had, the ability to persuade others of its worth is altogether vital. Most of his professional life was spent, not in observing, not in calculating, not in proving, but simply in persuading. He had to convince reluctant hearers that what he had to say about Nature made far more sense than anything that had ever been said before. His historic role was to change a world view, and this demanded talents of a far more diverse order than would be required by a simple establishment of a new theory. His works are thus characterized by a vigor and an immediacy that set them apart in the annals of scientific writing. They are exuberant, brash, speculative and wheedling by turn. His way of deployment of scientific method turned out almost immediately to have a power that set off the new science sharply from the old.

Galileo was the man who was a legend almost in his own life time and who has never since ceased to light up man's imagination. His life, his work, his death were all of a piece; there was the sort of symbolic unity about his career that sets up an immediate resonance in anyone who shares even a part of the vision that animated it. In Galileo, I see a vast energy directed single-mindedly to the changing of history.

Galileo's work marked the watershed between old and new. Galileo is portrayed as the pioneer of a new spirit, the father of the new sciences. Galileo's science had to be conceived as a violent creative break with all that had gone before.

Galileo published the great Discourse that contains the germ of Newtonian Mechanics both in its results and its methods, this mode of approach suggests that the transition from medieval to modern science was accomplished almost single-handedly by Galileo.

Galileo was born in Pisa in 1564, within a year of Shakespeare's birth. When Galileo was 11 years old he was sent to school in a monastery. At the age of 15, his father sent him to study medicine under a famous doctor Cesalpino but medicine was not Galileo's best love. At 19 years of age, he left university without taking examination and managed to study algebra from Professor Ostilio Ricci at Florence. He discovered the beauty of Mathematics and thereafter devoted his life to it, Physics, and Astronomy. Not only was he a great mathematician but also an excellent musician in both performance and composition. Galileo was an adept experimenter. As a boy he was fond of making mechanical toys. The training in Manual dexterity, which he thus acquired, was invaluable asset in all his scientific work.

He was also brilliant student, his intellectual curiosity was avid, his perception quick and his retention tenacious.

During his studies as a medical student he made his first invention, the pulsilogium for reading the pulse of a patient. The pulsilogium was essentially a pendulum of adjustable length. The doctor had to synchronize the pulse beat with the oscillations of the pulsilogium and read the post directly from a scale. He was a genius in the selection of a problem and method of attack.

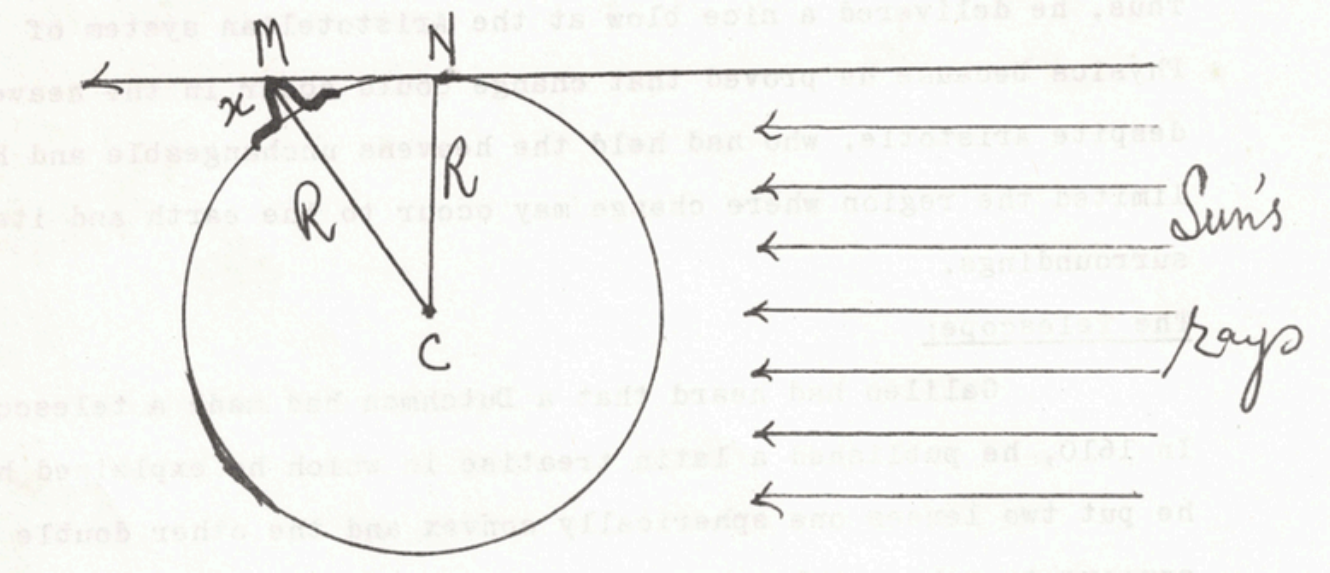
Galileo made his first contribution to Astronomy before he ever used a telescope. In 1604, Galileo showed a nova to be a true star located out in the celestial spaces and not inside the sphere of the moon. Galileo showed that this new star had no measurable parallax and so was very far from the earth. Thus, he delivered a nice blow at the Aristotelian system of Physics because he proved that change could occur in the heavens despite Aristotle, who had held the heavens unchangeable and had limited the region where change may occur to the earth and its surroundings.

<u>The Telescope:</u>

Galileo had heard that a Dutchman had made a telescope. In 1610, he published a Latin treatise in which he explained how he put two lenses one spherically convex and the other double concave to make a telescope of his own. Objects when looked through his

third telescope appeared magnified almost a thousand-fold in area and thirty times nearer. With the help of this telescope he made astonishing discoveries of sun spots, the mountains and craters on the surface of the moon. He found that the surface of the moon is not smooth, uniform and precisely spherical as a great number of philosophers believed it to be, but it is uneven, rough and full of cavities and prominences, being not unlike the face of the earth, relieved by chains of mountains and deep valleys. Not only did Galileo describe the appearance of mountains on the moon but also measured them. Galileo's determination of the height of the mountains on the moon has withstood the test of time and even today we agree with his estimate of their maximum height

Galileo's Measurements of the Height of Mountains on moon:



N = the terminator (boundary) between the illuminated and non-illuminated positions of the moon.
M = bright spot observed in the shadowed region.

$(R + x)^2 = R^2 + MN^2$ (Pythagorean Theorem)

$R^2 + 2Rx + x^2 = R^2 + MN^2$

$x^2 + 2Rx - MN^2 = 0$

Solved for x, the altitude of the mountain.

Radius of the moon was calculated from the distance of the moon from the earth which was known to Galileo.

Galileo also discovered the phases of Venus like those of our moon and four planets of Jupiter. Jupiter is now known to have at least 12 moons. He found that the milky way consisted of a great number of individual stars. He also saw the morning star.

In 1610, he left Padua and took a position as court philosopher to the Duke of Tuscany. In 1611, he discovered the handles of Saturn. His telescope was not powerful enough to define the phenomena as a ring. The same year, he saw spots on the Sun and realized that planets rotated. Copernicus had described planets rotated but for the first time a man had seen rotations of the planets.

Galileo also deduced, from study of the moon in its various phases that the earth turned about its axis and revolved around the sun as Copernicus had said and that it reflected light as the other planets. He saw the phases of Venus and struck great blow against the old astronomy. If Venus did travel in epicycles as held before that she would have no quarter, half and full phases as seen by Galileo.

Prior to 1609, the Copernican system of the universe had seemed to men a mere mathematical speculation. The basic supposition that the earth was merely another planet had been so contrary to all the dictates of experience and common sense that very few men had faced up to the awesome consequences of the heliostatic system. But after 1609, when men discovered through Galileo's eyes what the universe was like, they had to accept the fact that the telescope showed the world to be non-Ptolemaic, non-Aristotelian, in that the uniqueness attributed to the earth could not fit the facts. Thus, gone forever was the concept that the earth had a fixed spot in the centre of the universe, but it was now conceived to be in motion. Gone also was the comforting thought that the earth is unique, that it is an individual object without any likeness anywhere in the universe, that the uniqueness of man had given a uniqueness to his habitation.

Galileo's Other Works:

In 1597, he invented a compass for the direction of long distance. He elaborated and wrote about his theories of motion. He experimented in ballistics and the velocity of projectiles. He studied and described the strengths of materials, vital to construction of buildings and machinery. Galileo has been called father of dynamics. He defined various types of motion. Amongst others the definitions of uniform motion and uniform acceleration are of particular interest. The following paragraph is a description in his own words:

> "I consider a motion steady or uniform if the distances traversed by the moving body during any equal time intervals are equal..... I say that motion is steadily or uniformly accelerated which acquires, in any equal time equal increments of velocity."

Aristotle believed that the bodies fall freely with speeds that are proportional to their weights, but Galileo proved it to be wrong by allowing a card board placed over a coin with no edges projecting outwards that both gained speed at the same rate. Thus, he introduced the idea of air friction or friction of the medium.

In this connection, a historical passage from his book, "The Dialogue Concerning the Two New Sciences" is given... the book is presented in the form of a dialogue between three persons, Salviati, the teacher and a man on knowledge is the mouthpiece of Galileo. Salgredo is the neutral who asks intelligent questions. The character Simplico, Galileo claims that he derived from the name Simplicius, a student of Aristotle in the sixth century.

He performed the greatest historical experiment and showed the world that two pieces of iron of different weights allowed to fall together from the top of the leaning tower of Pisa touch the ground together.

Galileo's biographer, Viviani, who knew Galileo during his last years told a fascinating story which had taken root in the Galileo legend. According to Viviani, Galileo, desiring to confute Aristotle, ascended the Leaning Tower of Pisa, "in the presence of all other teachers, philosophers and all students and by repeated experiments proved that the velocity of moving bodies of the same composition, unequal in eight, moving through the same medium, do no attain the proportion of their weight as Aristotle assigned it to them, but rather they move with equal velocity.

He, with the help of a pail of water, determined time up to $1/10^{th}$ of a second.

In order to find how did the velocity change with time, he had to modify his experiment and let a ball roll down a sloping plank as he said to dilute the gravity i.e. to slow the motion or in other words decrease the acceleration.

With the aid of these experiments, Galileo demonstrated that:

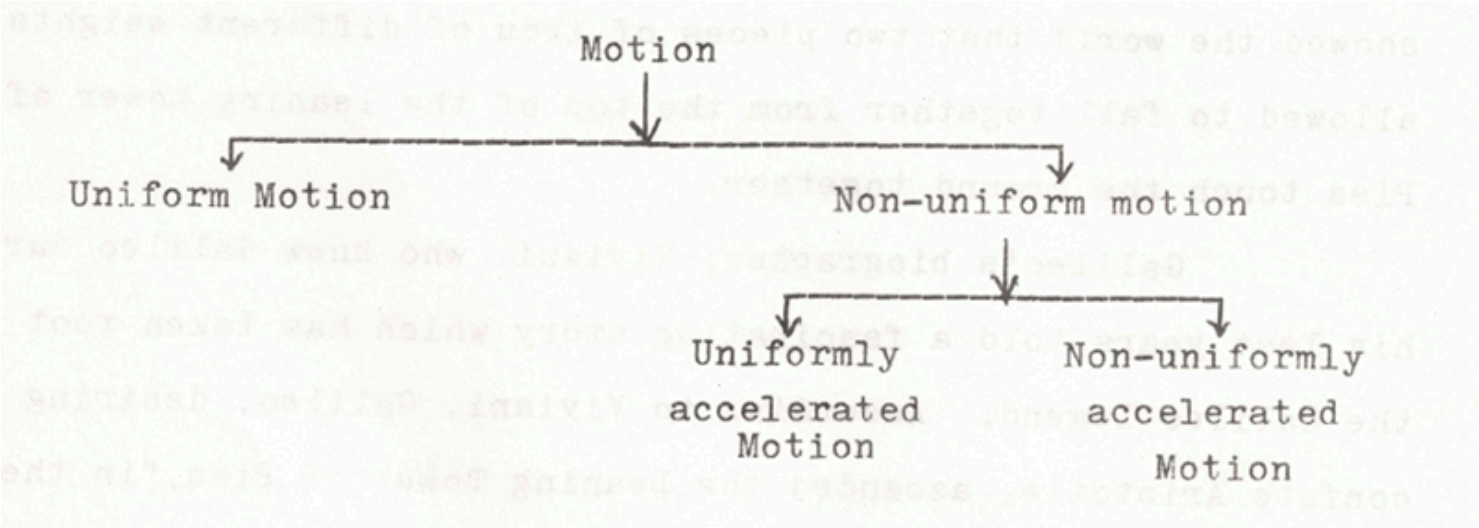
$$v \propto t$$

$$s \propto t^2$$

Where
v = speed,
t = time,
s = distance

For any inclination of the plane, however steep.

Galileo presented the following scheme to correct the Aristotelian Law of Motion



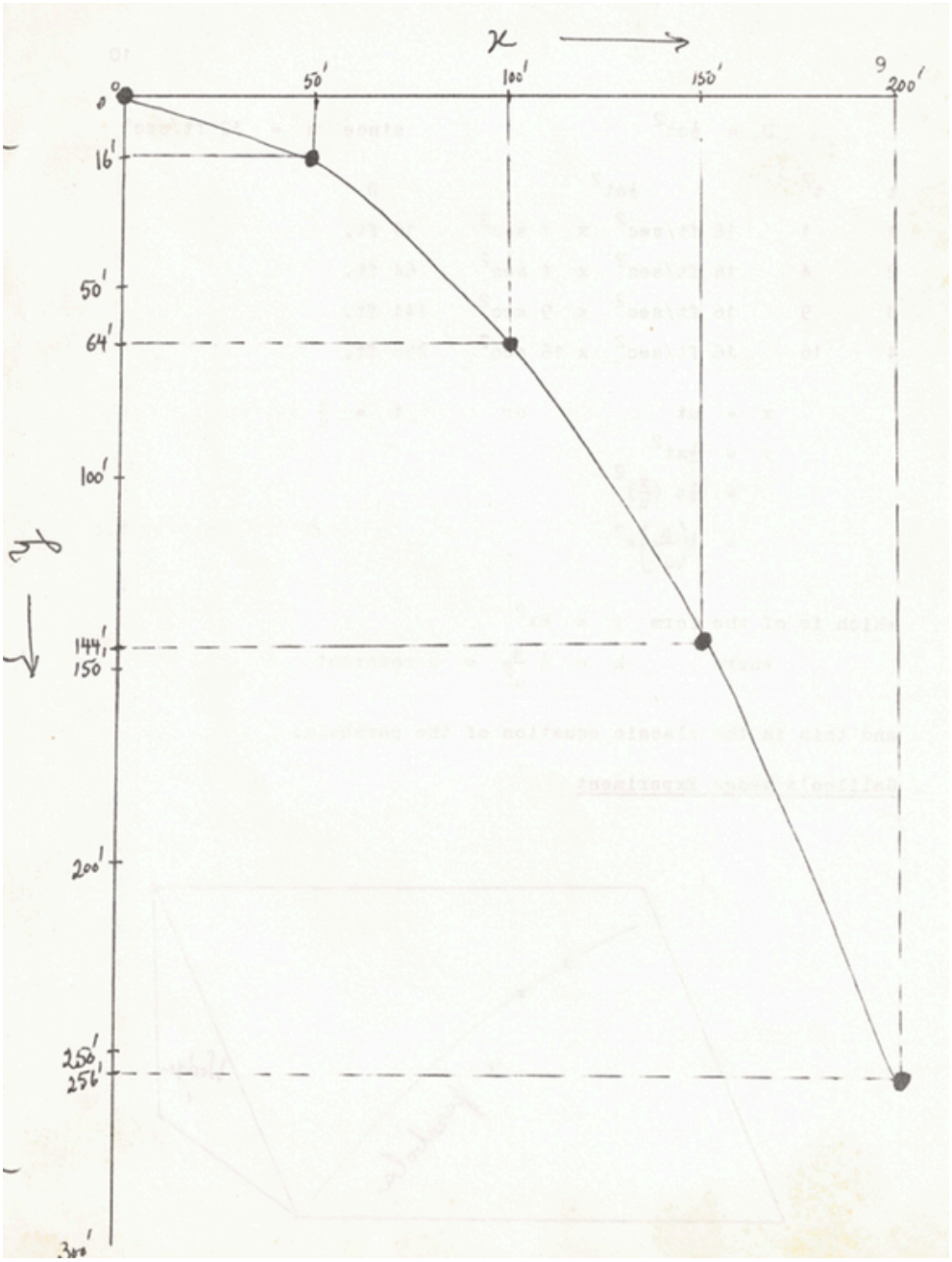
Galileo applied the fourteenth century “mean speed rule” to his Experiment of a “free fall of a body” and showed that”

$$s = \left(\frac{v_1 + v_2}{2}\right) t$$

Galileo also demonstrated that a projectile follows the path of a parabola because the projectile has simultaneously a combination of two independent motion:

(i) A uniform motion in a forward direction
(ii) A uniformly accelerated motion downward

Galileo analyzed the projectile motion considering a shell fired horizontally from a cannon at the edge of a cliff at a speed of 50 ft. per second.



$$D = \frac{1}{2}at^2, since\ a = 32\ ft/sec^2$$

t	t^2	$1/2\ at^2$	D
1	1	16 ft/sec^2 x 1 sec^2	16 ft.
2	4	16 ft/sec^2 x 4 sec^2	64 ft.
3	9	16 ft/sec^2 x 9 sec^2	144 ft.
4	16	16 ft/sec^2 x 16 sec^2	256 ft.

$$x = ut\ or\ t = \frac{x}{u}$$

$$y = \frac{1}{2}at^2$$

$$y = \frac{1}{2}a\left(\frac{x}{u}\right)^2$$

$$y = \frac{1}{2}\left(\frac{a}{u^2}\right)x^2$$

which is of the form

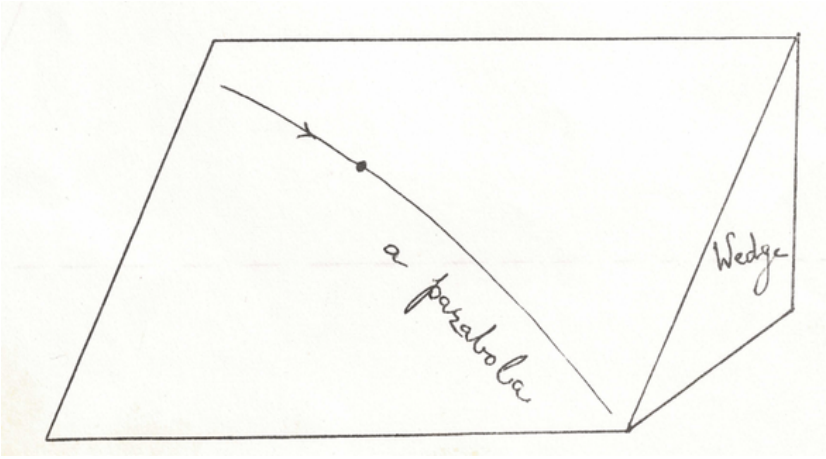
$$y = kx^2$$

where

$$k = \frac{1}{2}\left(\frac{a}{u^2}\right) = a\ constant$$

and this is the classic equation of the parabola.

Galileo's edge experiment



Inertia:

In one of the experiments, Galileo used two balls of the same size, one of lead and the other of oak, and both were let fall from a height of 200 cubits. Galileo found that balls reached the earth with slight difference in speed. He proved by this experiment that the resistance of the air increases in some proportion to the speed until the resistance of the air equals and offsets the weight, pulling the body down to the earth. Thus, Galileo concluded that when the resistance becomes so great that it equals the weight of the falling body, the resistance of the air will prevent any increase in speed and will render the motion uniform. This is anti-Aristotelian, because Aristotle held that when the motive force equals the resistance, the speed is zero. Galileo's principle is in limited form, a statement of Newton's first law of motion or the principle of inertia. This may be considered one of the major foundations of Modern Newtonian Physics.

In Dialogue, Galileo has hit upon the principle of inertia:

(i) A ball on a sloping plane would accelerate spontaneously
(ii) A ball on an upward slope would need a force to go up or to remain still
(iii) A ball placed upon a surface with no slope upward or downward would remain at rest. But if it is pushed in a direction it will move in that direction. If the surface is unbounded the motion would be boundless or the motion will be perpetual.

Galileo used the term quantity of motion (M) for momentum and 'w' for mass:

$$\therefore M = w \times u$$

He had keen sense of observation and interpretation. He studied vibrations, particularly in churches, of hanging lamps, etc. that the frequency of a simple free pendulum is constant and independent of its length provided the angle of swing is not greater than about 20^{o}. As there were no watches he used his pulse as a watch.

Aristotle had told that water is ten times heavier than air. Galileo made a comparative study of the weights of equal volumes of water and air with a crude but elegant experiment. He did the experiment with utmost patience and care using a single grain of sand at a time as a unit of mass. He verified the results with a modified apparatus.

He obtained that water should be 400 times heavier than air volume to volume. The modern ratio of density of water to density of air = 776. This experiment shows how good he was in his ability to predict and then to prove.

He had such an insight that even when he was blind he could sense a good problem from afar. He saw a lift pump which would work perfectly if the level of water stood above a certain level but below that level it failed to work. Aristotelian explanation for the working of pump was that "nature abhors vacuum." Discussing the pump with his students, Torricelli and Viviani, Galileo remarked in his usual gay tone that Damn Nature's horror of the void, by some mysterious whim seemed to peter-out suddenly at about 18 cubits. He suggested if they investigated the matter they would likely learn something important and useful. The work of Torricelli in the field is well known.

During his last years, he worked on several problems, among them was the construction of a pendulum clock which had a dial graduated in minutes and a hand that was operated by the pendulum but it had to be given a nudge every now and then to keep it swinging. Working on the problem on his deathbed in 1642, to Viviani he said, "quick while there is yet breath" and gasped out the specifications of the new instrument. Many years later, the Dutch scientist, Christian Huygen, completed the invention and took out a patent.

As I already stated, he had keen sense of observation and whatever he observed he recorded it. He once saw Chaldni figures and his statement of the phenomena is as follows:

> "As I was scraping a brass plate with a sharp chisel to remove some spots, I heard the plate emit a clear whistling tone. I noticed a long row of streaks (of small particles) parallel to each other. When the tone was higher, the streaks were closer together."

As we know now that these streaks will be formed at nodal lines.

In 1592, Galileo gave the importance to the study of heat, for he constructed a thermometer which expresses temperature numerically and thereby brought the subject to a quantitative stage. As explained by Viviani, he took a glass bulb about the size of a hen's egg with a tube about 2 spans in length and width of a straw-stem, warmed the glass with his hands, and turned it so that the end of the tube is dipped into a tumbler placed underneath. When the air in the bulb cooled, the water rose more than a span above the surface of the liquid in the tumbler. He also showed if the bulb was cooled, with wine or alcohol, the level of water rose still higher. This tube is called Galileo's tube.

Even when sealed thermometer was constructed for the first time, Galileo' device was considered more accurate. Later on in 1810, the Galileo's tube was modified to make a differential thermometer.

Galileo Traces the place of abode of God, The Creator

The Galileo-Plato Problem:

In the Dialogo under the guise of a suggestion by Plato, Galileo states that

God made each planet at His place of abode and gave it a straight and accelerated motion so that starting from rest, it gradually attained the particular velocity, He intended to confer upon it. when that velocity had been attained, God converted that straight line motion into a circular motion at the very same speed: a circular motion to be kept perpetually uniform forever after.

Thus, knowing the dimensions of the planetary orbits and the speeds of the planets, the Divine Order may be discovered.

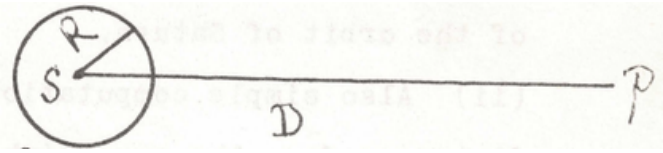
Let R = average distance of a planet from the sun

And T = periodic time of revolution of the planet about the sun

P = place of God

D = distance of P from the sun

a = uniform acceleration of the planets towards the sun



$$V = \frac{2\pi R}{T}$$

$$V^2 = u^2 + 2as$$

$$V^2 = 0 + 2a(D - R)$$

$$\left(\frac{2\pi R}{T}\right)^2 = 2a(D - R)\ (eq\ 1)$$

or

$$\left(\frac{4\pi^2 R^2}{T^2}\right) \times \frac{R}{R} = 2a(D - R)$$

$$4\pi^2 \frac{R^3}{T^2} = 2aR(D - R)$$

$$4\pi^2 K = 2aR(D - R)$$

As K = $\frac{R^3}{T^2}$ (Kepler's Third Law)

$$\frac{4\pi^2 K}{2a} = DR - R^2$$

$$\therefore DR - R^2 = constant$$

$$\therefore DR_1 - DR_1^2 = DR_2 - DR_2^2$$

$$D(R_1 - R_2) = R_1^2 - R_2^2$$

$$D = R_1 + R_2$$

Now R1 for Mercury = 0.4 A.U.

And R2 for Venus = 0.7 A.U

$$\therefore D = 0.4 + 0.7$$

$$= 1.1\ A.U.$$

Analysis of Galileo's Conception of Heavens:

So God must live a little more than the orbital distance of the earth from the sun.

(i) Which is absurd because by definition, P must lie outside of the orbit of Saturn.

(ii) Also simple computations show that there are no two planetary distances from the sun which can be added together to give the same distance D.

Conclusion:

Thus, there is no single place from which all the planets could be let fall from rest toward the sun with the same uniform acceleration so as to arrive at their respective orbits with speeds of the magnitude of their observed orbital speeds.

Paul Mansion's Analysis (1894) of Galileo's Problem:

Galileo's assumptions

According to equation (1)

$$\left(\frac{2\pi R}{T}\right)^2 = 2a(D-R)$$

$$\frac{4\pi^2 R^2}{T^2 2a} = D - R$$

$$R + \frac{2\pi^2}{a}\frac{R^2}{T^2} = D$$

$$R + m\frac{R^2}{T^2} = D$$

where $m = \frac{2\pi^2}{a}$ = a constant

$$\therefore\ R_1 + m\frac{K}{R_1} = R_2 + m\frac{K}{R_2}$$

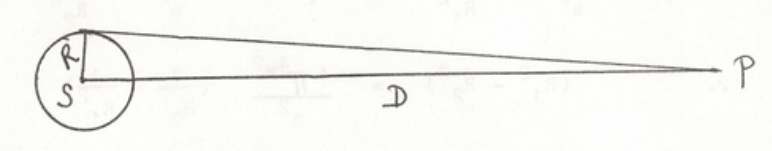
$$R_1 - R_2 = mK\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = mK\left(\frac{R_1 - R_2}{R_1 R_2}\right)$$

$$R_1 R_2 = mK$$

i.e. the product of any pair of sun-planet distances = a constant is plainly absurd.

Mansion considered two further possibilities.

1) All planets do not fall toward the sun but rather that each one drops along a tangent until it meets its proper orbit.



$$D^2 = R^2 + d^2$$

$$d_2 = \frac{1}{2}at^2$$

Where t = time for a planet to fall from P to its orbit

$$V = at$$

$$V = \frac{2\pi R}{T}$$

$$V^2 = a^2t^2$$

$$t^2 = \frac{V^2}{a^2}$$

$$d = \frac{1}{2}at^2$$

$$d = \frac{1}{2}a\frac{V^2}{a^2}$$

$$d = \frac{1}{2}\frac{V^2}{a} = \frac{1}{2}\left(\frac{2\pi R}{T}\right)^2\frac{1}{a}$$

$$d = \frac{2\pi^2R^2}{aT^2}$$

as $D^2 = R^2 + d^2$

$$D^2 = R^2 + \left(\frac{2\pi^2R^2}{aT^2}\right)^2$$

$$D^2 = R^2 + \frac{4\pi^4R^4}{a^2T^4}$$

$$D^2 = R^2 + \frac{4\pi^4}{a^2}K^2\left(\frac{1}{R^2}\right) \ (eq\ 2)$$

For two planets:

$$R_1^2 + \frac{4\pi^4}{a^2}K^2\left(\frac{1}{R_1^2}\right) = R_2^2 + \frac{4\pi^4}{a^2}K^2\left(\frac{1}{R_2^2}\right)$$

$$\therefore (R_1^2 - R_2^2) = \frac{4\pi^4}{a^2}K^2\left(\frac{1}{R_1^2} - \frac{1}{R_2^2}\right)$$

$$(R_1^2 - R_2^2) = \frac{4\pi^4K^2}{a^2}\left(\frac{R_1^2 - R_2^2}{R_1^2R_2^2}\right)$$

$$R_1^2R_2^2 = \frac{4\pi^4K^2}{a^2}$$

$$R_1R_2 = \frac{2\pi^2K}{a} \ (eq\ 3)$$

$$= a\ constant$$

There is no constant value obtained by multiplying the pairs of R_1 and R_2.

To continue further
From equation (3)

$$R_2 = \frac{2\pi^2 K}{a}\left(\frac{1}{R_1}\right)$$

But, from equation (2)

$$D^2 = R_1^2 + \frac{4\pi^4}{a^2}K^2\left(\frac{1}{R_1^2}\right)$$

$$\therefore D^2 = R_1^2 + R_2^2 \ (eq\ 4)$$

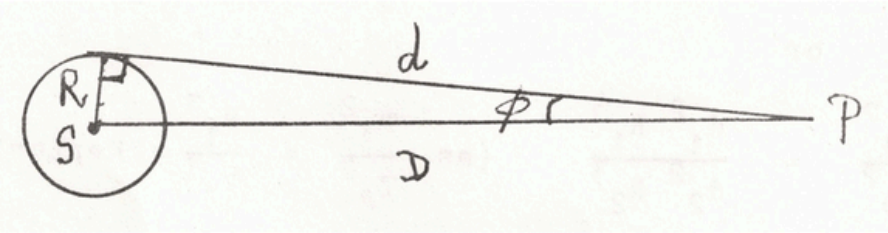
By substituting $R_2 = \frac{2\pi^2 K}{a}\left(\frac{1}{R_1}\right)$

There can be no single value of D for which this equation is valid and hence, no single point from which all the planets may be dropped according to the conditions of the problem.

Finally, suppose each planet falls along a tangent from P to its orbit but its acceleration is not a but component a directed from P to the sun.
Acceleration of the planet:

$$= a\cos\varphi$$

$$= a\frac{d}{D}$$



$$\therefore d = \frac{1}{2}a\cos\varphi\, t^2$$

$$d = \frac{1}{2}a\frac{d}{D}t^2$$

$$\frac{2D}{a} = t^2$$

$$t = \sqrt{\frac{2D}{a}}$$

But

$$V = g\cos\varphi\, t$$

$$V = g\frac{d}{D}t$$

As g, D and t are constant

$$V \propto d$$

Or

$$\frac{V_1}{V_2} = \frac{d_1}{d_2}$$

$$\frac{V_1}{V_2}\frac{T_1}{T_2} = \frac{d_1}{d_2}\frac{T_1}{T_2}$$

$$2\pi R_1 = V_1 T_1$$

$$2\pi R_2 = V_2 T_2$$

$$\therefore \frac{R_1}{R_2} = \frac{V_1}{V_2}\frac{T_1}{T_2} = \frac{d_1}{d_2}\frac{T_1}{T_2}$$

$$\therefore \frac{R_1^2}{R_2^2} = \frac{d_1^2 T_1^2}{d_2^2 T_2^2}$$

$$\frac{R_1^2}{R_2^2} = \frac{d_1^2 R_1^3}{d_2^2 R_2^3}$$

As $\frac{T_1^2}{T_2^2} = \frac{R_1^2}{R_2^2}$ (Kepler's Third Law)

$$\therefore\ d_1^2 R_1 = d_2^2 R_2\ (eq\ 5)$$

From the diagram

$$d_1^2 = D^2 - R_1^2$$

$$d_2^2 = D^2 - R_2^2$$

Substituting in equation (5) we get:

$$(D^2 - R_1^2)R_1 = (D^2 - R_2^2)R_2$$

$$\therefore\ D^2 R_1 - R_1^3 = D^2 R_2 - R_2^3$$

$$D^2(R_1 - R_2) = R_1^3 - R_2^3$$

$$D^2 = \frac{R_1^3 - R_2^3}{R_1 - R_2}$$

$$D^2 = R_1^2 + R_1 R_2 + R_2^2$$

Which is incompatible for the data for any pair of planets.

Newton's Calculations to search for the location of Heavens:

Newton calculated the distance to the Divine Order supposing if a planet is torn off the gravitational force of the sun, how far will it travel before it comes to rest where it was created by God.

Sun's gravitational force on the planet

$$F = ma$$

Where m = mass of the planet, a = centripetal acceleration of the planet

$$F = m\frac{4\pi^2 R}{T^2}$$

$$F = m\frac{4\pi^2 R}{T^2}\frac{R^2}{R^2}$$

$$F = m\frac{4\pi^2}{R^2}\frac{R^3}{T^2}$$

$$F = 4\pi^2 K\frac{m}{T^2}$$

If the planet is made to stop its circular motion and move away from the sun in a straight line to a distance d before coming to rest will gain in potential energy ΔU.

$$4\pi^2 K\frac{m}{R} - 4\pi^2 K\frac{m}{d} = 4\pi^2 Km\left(\frac{1}{R} - \frac{1}{d}\right)$$

But

$$Gain\ in\ U = Loss\ in\ E_k$$

$$\Delta U = \Delta E_k\ (eq\ 6)$$

But

$$\Delta E_k = \frac{1}{2}mV^2$$

$$\Delta E_k = \frac{1}{2}\left(\frac{2\pi R}{T}\right)^2$$

$$\Delta E_k = \frac{2\pi^2 R^2}{T^2}\frac{R}{R}$$

$$\Delta E_k = 2\pi^2 Km\frac{1}{R}\ (eq\ 7)$$

Substitute the values of ΔU and ΔE_k in equation 6 we get:

$$4\pi^2 Km\left(\frac{1}{R} - \frac{1}{d}\right) = 2\pi^2 Km\frac{1}{R}$$

$$2\left(\frac{1}{R} - \frac{1}{d}\right) = \frac{1}{R}$$

$$\frac{2(d-R)}{Rd} = \frac{1}{R}$$

$$2(d-R) = d$$

$$d = 2R$$

If the planet were created at a distance 2R from the sun, the planet would have moved a distance d to be in an orbit round the sun with its present speed such that

$$d = 2R$$

Thus, according to Newton's condition, each planet is dropped from a different point just twice its normal distance from the sun while Galileo observed all planets were let fall from one and the same point.

Next, Newton suggests that if the gravitational power of the sun be diminished by one half.

In this case:

Gain in U as the planet moves away from the sun from the original distance R_1 to a new distance R_2

$$\Delta U = \frac{1}{2}\Delta U_1$$

$$= \frac{1}{2}4\pi^2 Km\left(\frac{1}{R} - \frac{1}{d}\right)$$

$$= 2\pi^2 Km\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

$$Gain\ in\ U = Loss\ in\ E_k$$

$$\Delta U = \Delta E_k$$

$$2\pi^2 Km\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = 2\pi^2 Km\frac{1}{R_1}$$

This equation will hold only when R_2 approaches infinity. So the planet will now ascend perpetually. Now, if planet 2 moves outward to the orbit of planet 3 and if during this outward motion the gravitational attracting power of the sun once again becomes one half of its actual value.

$$Gain\ in\ U = Loss\ in\ E_k$$

$$\Delta U = \Delta E_k$$

$$\frac{1}{2}4\pi^2 Km_2\left(\frac{1}{R_2} - \frac{1}{R_3}\right) = \frac{1}{2}m_2 V_2^2 - \frac{1}{2}m_2 V^2$$

Where V is the speed of planet 2 when it has reached the orbit normally occupied by planet 3.

$$\therefore 2\pi^2 K\left(\frac{1}{R_2} - \frac{1}{R_3}\right) = \frac{1}{2}V_2^2 - \frac{1}{2}V^2$$

$$= \frac{1}{2}\left(\frac{2\pi R_2}{T_2}\right)^2 - \frac{1}{2}V^2$$

$$= \frac{2\pi^2 R_2^2}{T_2^2}\frac{R_2}{R_2} - \frac{1}{2}V^2$$

$$= 2\pi^2\frac{K}{R_2} - \frac{1}{2}V^2$$

$$\therefore 2\pi^2 K\frac{1}{R_3} = \frac{1}{2}V^2$$

$$V^2 = 4\pi^2 K\frac{1}{R_3} = V_3^2$$

So, we conclude that if a planet moves away from the sun in a straight line with its normal orbital speed but if the sun's gravitational force is ½ of what it actually is then each planet reaches the orbit of any outer planet and it will have there a linear speed exactly equal to the orbital speed of the planet that occupies that orbit.

Thus, Newton concluded if all planets ascend at once and ascend in the same line, they will constantly in ascending become nearer and nearer together and their motion will constantly approach to an equality and become at length zero.

The converse of this statement will apply to Galileo-Plato concept:

When the planets reach zero motion position, let the motion of all these planets be reversed and let fall. Each planet would then arrive at its own orbit with its proper normal orbital speed. As the planets reach their respective orbits then, their motion turned sideways and at the same time the gravitational power of the sun doubled.

Newton said at this point that the Divine power is here, required in a double respect.

1. To turn the straight line motion of the falling planets into a side motion.
2. To double the attractive power of the sun at the same time.

If the attractive power of the sun is not doubled, then the sun will not be able to hold the planets and they will go into the highest heavens in parabolic lines.

The unity of Galileo's scientific life, combining observational astronomy and mathematical physics, comes from his dedication to a sun-centred universe, a dedication reinforced in some way by every major discovery he made in either physics or astronomy. Having been the instrument by which the glorious aspects of the creation in the heavens first had been fully revealed to a mortal, Galileo must have has a special sense of urgency to convert all his fellowmen to the true i.e. the Copernican system of the universe. His conflict with the Roman Catholic Church arose because deep in his heart, Galileo was a true believer. There was for him no path of compromise. No way to have separate secular and theological cosmologies. If the Copernican system was true, as he believed, then what else could Galileo do but fight with every weapon he had, logic, rhetoric, scientific observation, mathematical theory and cunning insight, to make his church accept a new system of the universe. We may catch a glimpse of the spirit of this great man, as I think of him, after his trial and condemnation, living under a kind of house arrest, completing his greatest scientific work, "Discourse and Demonstrations Concerning Two New Sciences." This book was the base from which Newton began his great exploration of the dynamical principles of a sun-centred universe.

Bibliography

1. Basu & Chatterji; The Pre-University and Intermediate Physics.
2. Borowitz & Bornstein; A Contemporary View of Elementary Physics.
3. Cohen, Bernard J.; The Birth of a New Physics.
4. Cooper, Lane; Aristotle, Galileo and the Tower of Pisa.
5. Drake, Stillman; Discoveries and Opinions of Galileo.
6. Drake, Stillman; Galileo Galilei, A Biography and Inquiry into his Philosophy of Science.
7. McMullin, Ernan; Galileo, Man of Science.
8. Seeger, Raymond J.; Galileo Galilei, His Life and his Works.
9. Shortley and Williams; Elements of Physics.

Appendix B: Facsimile of the Original Manuscript (1972)

This appendix reproduces scanned images of the original manuscript pages of "Galileo" (1972), provided to preserve the historical form and provenance of the document. The fascimile can be found online at https://github.com/ishgosw/Galileo-as-Physicist-and-Polemicist

L'opéron lac : d'un modèle bactérien à un langage pour la régulation génique

The lac Operon: From a Bacterial Model to a Language for Gene Regulation

Maïka Harvey[1*]

1. Université d'Ottawa, Ottawa, ON, Canada
*Auteur correspondant. Courriel : mharv082@uottawa.ca

Résumé | Abstract

Cet article propose un commentaire historique et conceptuel sur le modèle d'opéron formulé par Jacob et Monod à partir de l'étude du système lac chez Escherichia coli en 1961. Il rappelle comment l'introduction de gènes régulateurs, de séquences opératrices et d'un messager instable, l'ARNm, a permis de concevoir la régulation de l'expression génique comme un problème de logique conditionnelle fondé sur des circuits de décision ON/OFF qui intègrent plusieurs signaux, plutôt que comme une simple lecture linéaire de l'ADN. Le texte montre que l'opéron lac reste un exemple canonique dans l'enseignement de la biologie moléculaire, au cœur des manuels et d'outils d'évaluation et qu'il structure encore les premières représentations étudiantes de la régulation génique. Enfin, il met en évidence l'influence durable du modèle d'opéron sur la biologie de synthèse et la biologie des systèmes, où la logique répresseur-opérateur, les architectures modulaires et la notion de circuits régulateurs continuent d'inspirer la conception de réseaux génétiques artificiels, parmi d'autres sources conceptuelles.

This article offers a historical and conceptual commentary on the operon model formulated by Jacob and Monod from the study of the lac system in Escherichia coli in 1961. It recalls how the introduction of regulatory genes, operator sequences, and an unstable messenger, mRNA, made it possible to conceive gene expression regulation as a problem of conditional logic based on ON/OFF decision circuits that integrate multiple signals, rather than as a simple linear reading of DNA. The text shows that the lac operon remains a canonical example in molecular biology teaching, at the heart of textbooks and assessment tools, and that it still structures students' initial representations of gene regulation. Finally, it highlights the lasting influence of the operon model on synthetic biology and systems biology, where repressor-operator logic, modular architectures, and the notion of regulatory circuits continue to inspire the design of artificial genetic networks, among other conceptual sources.

Mots-clés: Opéron lac, régulation génique, Jacob et Monod, régulation négative, répresseur-opérateur, biologie de synthèse

Introduction

Au début des années 1960, les biologistes savaient que l'ADN porte l'information héréditaire, mais ils ne disposaient pas encore d'un cadre clair pour expliquer comment les cellules activent ou répriment sélectivement leurs gènes selon l'environnement. Des expériences de croissance bactérienne, par exemple, des courbes de croissance d'Escherichiacoli montrant une phase de latence lors du passage d'un sucre préféré à un sucre secondaire, indiquaient que la production d'enzymes suit la disponibilité des nutriments, sans mécanisme formalisé pour expliquer de tels choix d'expression génique. En 1961, François Jacob et Jacques Monod comblent cette lacune en proposant le modèle d'opéron pour expliquer l'expression inductible de groupes de gènes chez les bactéries. En introduisant des gènes régulateurs, des opérateurs, des promoteurs et un messager instable, plus tard identifié comme l'ARNm, ils transforment la régulation de l'expression génique en problème logique de décision plutôt qu'en lecture passive de l'ADN (1-3).

Le modèle de l'opéron et le système lac

Le modèle d'opéron de Jacob et Monod part d'une énigme précise : comment E.coli coordonne-t-il la synthèse des enzymes nécessaires au métabolisme du lactose (2, 4) ? Des observations montrent que des enzymes apparentées peuvent être co-induites ou co-réprimées par le même métabolite, suggérant un contrôle commun de groupes de gènes (2, 4, 5). Par des analyses génétiques, notamment de mutants constitutifs et de diploïdes partiels, ils distinguent des éléments qui codent la structure des enzymes et d'autres qui en gouvernent la synthèse (2, 3). Ils proposent ainsi un modèle de régulation négative dans lequel un répresseur, produit par un gène régulateur, se lie à une séquence opératrice et bloque la transcription jusqu'à ce qu'un signal approprié l'inactive (2, 3). L'architecture de base comprend des gènes structuraux, une région de contrôle, y compris un promoteur (site de liaison de

l'ARN polymérase) et un opérateur, ainsi qu'un gène régulateur produisant un répresseur diffusible (2, 4). Dans l'opéron lac, le répresseur se fixe à l'opérateur en absence de lactose, empêchant la transcription des gènes nécessaires à l'utilisation de ce sucre (2, 4). La présence de lactose ou d'un analogue agit comme inducteur ; il se lie au répresseur, réduit son affinité pour l'opérateur et lève la répression, ce qui autorise la transcription (2, 4). Cependant, cette régulation est également modulée par un mécanisme de régulation positive impliquant le complexe CAP-AMPc. Le complexe CAP-AMPc est un facteur de transcription bactérien formé de la protéine activatrice du catabolisme (CAP) liée à l'adénosine monophosphate cyclique (AMPc). En conditions de faibles concentrations en glucose, l'augmentation du AMPc permet la formation du complexe CAP-AMPc, qui se fixe en amont du promoteur et favorise le recrutement de l'ARN polymérase, amplifiant ainsi la transcription (2, 4, 6). À l'inverse, en présence de glucose, ce mécanisme est inhibé, ce qui limite l'expression de l'opéron, même si le lactose est disponible. La régulation du système lac repose donc sur une intégration de signaux négatifs et positifs, permettant une réponse fine aux conditions environnementales. L'information circule de l'ADN vers les protéines via un messager transitoire, l'ARNm, qui transmet les décisions régulatrices de l'opéron à la machinerie de traduction (1, 2, 4).

L'impact conceptuel et les circuits géniques

L'élégance du modèle d'opéron tient autant à son explication du métabolisme du lactose qu'à sa portée générale pour comprendre la régulation de l'expression génique (1, 3). Jacob et Monod proposent que des combinaisons variées de gènes régulateurs, d'opérateurs et de gènes structuraux puissent générer des profils d'expression très différents (2, 3). Des études ultérieures sur d'autres systèmes bactériens montrent que la régulation coordonnée de groupe de gènes est une stratégie courante pour contrôler la synthèse d'enzymes (4, 6). Dans cette perspective, les gènes peuvent être envisagés comme des éléments de circuits régulateurs intégrant différentes conditions biologiques. Leur expression résulte alors de l'interaction de signaux multiples, permettant une réponse adaptée à l'environnement cellulaire. Le système lac illustre ce principe : son expression est activée en présence de lactose et réprimée lorsque des sources de carbone préférées, comme le glucose, sont disponibles (4, 6, 7). Des approches de modélisation mathématique ont également revisité les principes de conception du circuit lac, confirmant la robustesse de son architecture (5). Cette manière de représenter la cellule comme un ensemble de circuits de décision précède de plusieurs décennies le vocabulaire de la biologie des systèmes et des réseaux (1, 3, 7).

Le rôle dans l'enseignement

Plus de 60 ans plus tard, l'opéron lac reste un exemple canonique dans l'enseignement de la biologie moléculaire et cellulaire, omniprésent dans les manuels et les cours d'introduction (4, 8). Il sert à illustrer des promoteurs, opérateurs, répresseurs et inducteurs, ainsi que la distinction entre les gènes structuraux et les gènes régulateurs (4, 8). Un instrument d'évaluation conceptuelle consacré à l'opéron lac a même été développé pour évaluer systématiquement la compréhension des étudiants, signe de sa place centrale dans les programmes de génétique. L'opéron inspire aussi de nombreux travaux pratiques, des dosages de β-galactosidase aux systèmes d'expression inductibles par l'isopropyl β-D-1-thiogalactopyranoside en clonage et production de protéines (4, 9). Pour plusieurs étudiants, leur premier contact avec la régulation génique passe par ce modèle, qui reste un schéma mental même lorsqu'ils abordent ensuite la chromatine eucaryote ou des réseaux de facteurs de transcription plus complexes (4, 8).

De l'opéron à la biologie de synthèse et des systèmes

Au-delà de son rôle pédagogique, le modèle d'opéron a profondément marqué la biologie de synthèse et la biologie des systèmes (1, 7). Les circuits de biologie de synthèse réutilisent explicitement la logique répresseur-opérateur et des architectures proches de celles proposées par Jacob et Monod (6, 7). Des revues de biologie de synthèse soulignent que l'idée de parties modulaires et recombinables, incluant opérateurs, régulateurs et promoteurs, s'inscrit dans la continuité du modèle d'opéron, tout en s'appuyant sur d'autres avancées conceptuelles et technologiques (6, 7). Cette vision modulaire sous-tend aujourd'hui la conception de réseaux régulateurs aux comportements ciblés, de simples interrupteurs et dispositifs de mémoire jusqu'à des portes logiques plus complexes (6, 7). Même lorsque l'on intègre des niveaux supplémentaires, comme la régulation épigénétique ou des protéines et ARN programmables, l'idée d'assembler des composants régulateurs en circuits reste fortement inspirée par cette première formulation, parmi d'autres influences majeures telles que le développement des technologies d'ADN recombinant et des outils de régulation génétique programmable (6, 7).

Réflexion et l'actualité du modèle

L'influence durable de l'article de 1961 se voit de la manière dont les étudiants décrivent la régulation génique : souvent sous forme de cartoon, avec un répresseur posé sur un opérateur comme un barrage et le lactose comme clé qui libère l'expression des gènes (1, 8). Ce dessin simplifié ouvre sur des idées plus profondes de flux d'information, de rétroaction et de prise de décision au sein des cellules (4, 8). Dans les laboratoires d'enseignement, des systèmes d'expression inductibles dérivés du contrôle lac sont utilisés de façon routinière, souvent sans rappeler qu'ils prennent racine dans un modèle de métabolisme bactérien proposé en 1961 (4, 9). Relire l'article original met en évidence à la fois l'audace de ses hypothèses et la précision de ses prédictions et montre comment un modèle conceptuel peut orienter des décennies d'expériences (1, 2). Le modèle d'opéron de Jacob et Monod reste ainsi une référence parce qu'il offre un cadre logique et visuel pour penser la régulation génique, accessible aux étudiants, mais suffisamment riche pour nourrir la biologie de synthèse et des systèmes contemporains (1, 3). Dans le contexte actuel des données omiques et des paysages régulateurs très complexes, revenir à l'opéron lac rappelle combien une grande partie de notre langage actuel (régulateurs, opérateurs, circuits, logique) était déjà là en 1961 (1-

3, 7). Pour les apprenants comme pour les chercheurs, il demeure plus qu'un exemple de manuel. Il montre comment un modèle bien construit peut révéler des principes qui traversent les générations en biologie moléculaire.

Références

1. M. Yaniv. The 50th anniversary of the publication of the operon theory in the Journal of Molecular Biology: past, present and future. J. Mol. Biol. 409, 1-6 (2011).
2. F. Jacob, J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318-356 (1961).
3. M. Lewis. A tale of two repressors. J. Mol. Biol. 409, 14-27 (2011).
4. F. C. Neidhardt, R. Curtiss III, Eds. Escherichia coli and Salmonella: Cellular and Molecular Biology (ASM Press, Washington, DC, ed. 2, 1996).
5. M. A. Savageau. Design of the lac gene circuit revisited. Math. Biosci. 231, 19-38 (2011).
6. I. Bervoets, D. Charlier. Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. FEMS Microbiol. Rev. 43, 304-339 (2019).
7. M. A. English, R. V. Gayet, J. J. Collins. Synthetic biology within the operon model and beyond. Annu. Rev. Biochem. 90, 221-244 (2021).
8. K. M. Stefanski, G. E. Gardner, R. L. Seipelt-Thiemann. Development of a Lac Operon Concept Inventory (LOCI). CBE Life Sci. Educ. 15, ar24 (2016).
9. F. W. Studier. Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif. 41, 207-234 (2005).

Rethinking Physician Wellness: The Role of Artificial Intelligence in Addressing Burnout in Canada

Repenser le bien-être des médecins : le rôle de l'intelligence artificielle dans la lutte contre l'épuisement professionnel au Canada

Parastoo Golzarian[1*]

1. Toronto Metropolitan University, Toronto, ON, Canada
*Corresponding author. Email: parastoo.golzarian@gmail.com

Abstract | Résumé

Burnout among Canadian physicians, and medical trainees is a pressing issue although the drivers, and manifestations of burnout differ across these groups. Worsening workforce shortages, and deteriorating patient care quality underscore the urgency. Traditional wellness programs have shown inconsistent success, highlighting the need for innovation. This review explores how artificial intelligence (AI) could support mental health across the medical education, and practice continuum. The current evidence base spans multiple stages of development: initial pilot studies, and feasibility trials demonstrate promising engagement, and symptom reduction, while a smaller number of randomized controlled trials provide early support for effectiveness. However, large-scale implementation studies remain limited, and many tools have yet to be validated in real-world clinical environments. Current findings should therefore be interpreted as preliminary rather than definitive. Some machine learning models have shown early promise in identifying healthcare workers at risk of burnout, though the evidence remains preliminary, and warrants cautious interpretation. Combining wearables with workplace data offers continuous monitoring. AI may strengthen wellness strategies with earlier detection, and tailored support; however, ethical considerations remain critical, including data privacy, algorithmic bias, informed consent for monitoring technologies, and the risk of a surveillance-oriented culture, all of which require robust institutional governance.

———

L'épuisement professionnel parmi les médecins canadiens et les stagiaires en médecine est un problème est urgent, même si les facteurs et les manifestations de l'épuisement professionnel diffèrent selon les groupes. L'aggravation de la pénurie de personnel et la détérioration de la qualité des soins aux patients soulignent l'urgence de la situation. Les programmes de bien-être traditionnels ont connu un succès inconsistant, ce qui met en évidence la nécessité d'innover. Cet article de synthèse explore comment l'intelligence artificielle (IA) pourrait soutenir la santé mentale tout au long du parcours de formation et de pratique médicale. Les données probantes actuelles couvrent plusieurs stades de développement : des études pilotes initiales et des essais de faisabilité démontrent un engagement prometteur et une réduction des symptômes, tandis qu'un nombre plus restreint d'essais contrôlés randomisés apportent des preuves préliminaires d'efficacité. Cependant, les études d'implantation à grande échelle restent peu nombreuses et de nombreux outils doivent encore être validés en situation clinique réelle. Les résultats actuels doivent donc être interprétés comme préliminaires et non comme définitifs. Certains modèles d'apprentissage automatique se sont révélés prometteurs pour identifier les professionnels de santé à risque d'épuisement professionnel, mais ces résultats demeurent préliminaires et nécessitent une interprétation prudente. L'association des objets connectés aux données du lieu de travail permet une surveillance continue. L'IA peut renforcer les stratégies de bien-être grâce à une détection plus précoce et un accompagnement personnalisé ; toutefois, les considérations éthiques demeurent essentielles, notamment la protection des données, les biais algorithmiques, le consentement éclairé pour les technologies de surveillance et le risque d'une culture de la surveillance, autant d'éléments qui nécessitent une gouvernance institutionnelle rigoureuse.

Keywords: Physician burnout, medical trainees, Artificial intelligence, Wellness programs, Predictive analytics

Introduction

The Canadian medical community is becoming increasingly concerned about the well-being of doctors, and medical students. The Canadian Medical Association (CMA) has formally identified physician wellness as an essential issue, citing its growing impact on patient care, healthcare costs, and workforce sustainability (1). Recent studies underscore the urgency of this issue. In 2019, a

survey of Canadian physicians highlighted concerning levels of burnout, with the majority meeting criteria for overall burnout, and reporting high rates of emotional exhaustion, depersonalization, and diminished personal accomplishment (2). Similarly, the 2021 National Physician Health Survey (NPHS) reported comparable results, further linking high burnout rates to the ongoing physician shortage in Canada (3). This has led to more than half of family doctors saying they are likely to cut back on or change their clinical hours (4), and some are retiring early or quitting altogether (5, 6).

Aside from the physician departures, fewer medical students are pursuing family medicine (7). This is taking place as the population in Canada ages, and patient dem, ands get more complex (8), hospitals all around the nation are fighting to keep their emergency departments open, an estimated 6.5 million people lack a regular family doctor (9).

These trends paint a concerning picture of a workforce under significant strain. Physician burnout carries profound consequences, not only for the physicians themselves, but for the patients they serve, and the sustainability of Canada's health system. As traditional wellness programs continue to fall short, there is a growing need for innovative, evidence-based approaches that can identify at-risk physicians earlier, and provide more targeted support (2, 10). This article argues that while systemic reforms remain essential, artificial intelligence (AI)-enabled interventions represent an underexplored yet promising adjunct to improve wellness among medical learners, and physicians in Canada.

A. Limitation of current approaches

Despite increased awareness, many wellness programs now in place, like peer support groups, resilience training, and counseling services face major obstacles to providing effective support to doctors (11-14).

1.Stigma

Despite the existence of wellness programs, poor participation rates are frequently caused by deeply ingrained stigma, and confidentiality issues. Systematic research found that medical professionals, and students are severely discouraged from getting help due to both self-stigma, and public stigma, which hinders the effectiveness of wellness programs (11, 12).

2.Limited evidence of impact

Another significant limitation of current wellness programs is the quality, and consistency of evidence supporting their impact. Systematic reviews, and meta-analyses have shown that interventions such as resilience training, mindfulness workshops, and peer support often produce only small or short-term improvements in stress, and burnout, and their long-term effectiveness remains uncertain (13, 14). These limitations point to the need for complementary approaches.

B. Role of AI

Integration of AI into wellness programs has the potential to address the limitations described above across multiple dimensions: reducing barriers to access, improving measurement, and enabling earlier intervention.

The following figure summarizes how each AI intervention type addresses a corresponding limitation in current wellness programs (Figure 1).

1. Stigma Reduction

AI-based solutions, such as chatbots for mental health, and conversational agents, provide private assistance 24/7. When compared to control groups, randomized trials have shown that AI

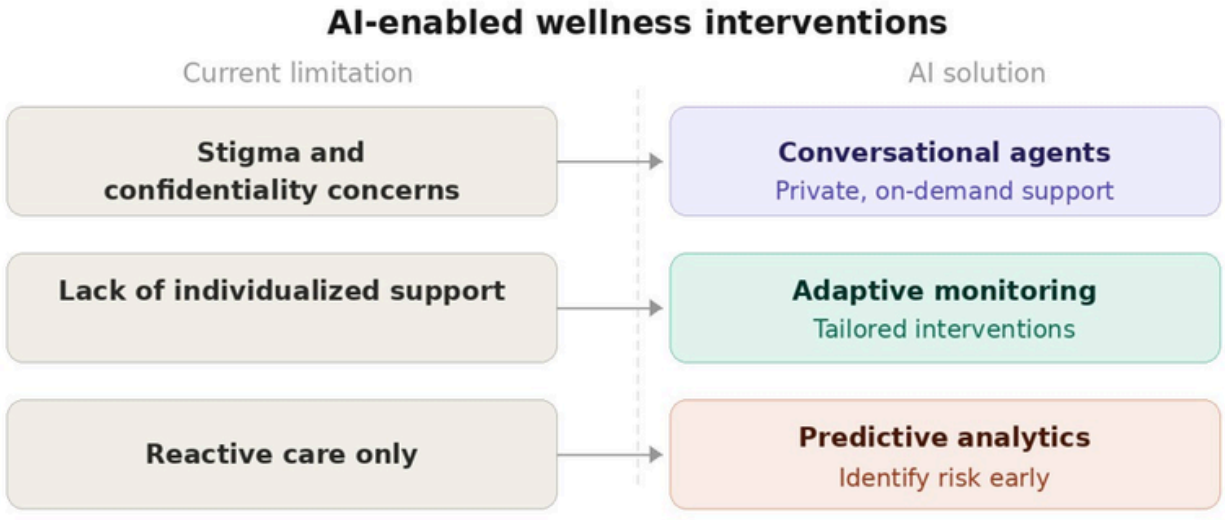


Figure 1. AI-enabled interventions addressing key limitations of current wellness programs.

chatbots can provide cognitive behavioral therapy (CBT)-based interventions that dramatically reduce symptoms of anxiety, and depression (15, 16). A randomized controlled trial in Hong Kong compared an AI chatbot with a nurse hotline. Although overall satisfaction levels of both services were similar, AI chatbot groups showed significant reduction in depression, and anxiety when pre-post scores were compared, while the nurse hotline demonstrated no significant change in the scores (17). Moreover, according to a recent meta-analysis of AI conversational agents, these technologies can offer significant help in ways that are more approachable, and less stigmatizing than in-person counseling (18). AI-enabled platforms that provide private, on-dem, and access could potentially inspire doctors, and students who might otherwise shy away from institutional wellness programs out of concern for confidentiality violations or social stigma to seek care.

2. Personalization

Beyond accessibility, AI can provide unbiased, customized, and trustworthy proof of wellness benefits. In contrast to conventional workshops that rely on self-reports at the group level, AI systems can continuously monitor, and assess user involvement, mood swings, and stress indicators in real time (18). A significant shortcoming of current wellness programs is addressed by this data-driven capability, which makes it possible to standardize outcome measurements across various programs, and circumstances.

AI can also customize interventions by adjusting to each person's unique requirements. Learning resources or coping mechanisms, for instance, can be modified based on usage trends, and monitored stress levels. Research indicates that this kind of flexible personalization not only improves user involvement but also reinforces the long-term viability of wellness advantages (19). These discoveries can be used in wellness programs of many universities. AI integration into the academic calendar or clinical rotation schedule could be greatly beneficial for tracking the student's wellbeing.

3. Early Detection & Preventive Monitoring

AI systems allow for early detection of burnout by monitoring the activity of physicians, and students. A study demonstrates that HiPAL, which is a burnout prediction model based on activity logs, can use AI-driven predictive analytics to identify doctors who are at risk of burnout.

These techniques enable early diagnosis, and initiative-taking support, in contrast to traditional approaches that respond only after symptoms appear. However, predictive models carry important limitations that must be acknowledged: false positives may unnecessarily alarm or stigmatize individuals, false negatives may create a false sense of security, and models trained on specific populations may not generalize across institutions with different demographics, workflows, or cultural norms. External validation across diverse settings remains limited (20). In 2025, researchers conducted a study to see if burnout among medical professionals could be predicted using AI. They collected information through a standardized burnout questionnaire along with basic demographic, and workplace data. This dataset was then used to train different machine learning models to recognize patterns that separate staff with high burnout risk from those with minimal risk. Among these, some models achieved the strongest performance, with prediction accuracies close to 80% in classifying healthcare workers at high versus minimal risk for burnout. Although, accuracy alone is an incomplete metric, and must be interpreted alongside sensitivity, specificity, and external validation, this study demonstrates the feasibility of applying artificial intelligence methods to clinician wellness by moving beyond descriptive statistics, and toward predictive modeling (21).
The *Burnout Prediction Using Wearables, and Artificial Intelligence* (BROWNIE) study protocol introduces an innovative approach that combines occupational, psychological, and physiological data in an AI framework to present a novel method of predicting burnout among registered nurses. This is decentralized, divided into three cohorts for training, testing, and validation. A smartwatch will be worn by each participant to continuously record information about their heart rate, sleep habits, and level of physical activity. BROWNIE has not yet reported outcome data, as the trial remains ongoing; nonetheless, it illustrates the potential of AI-enhanced wearable monitoring to move beyond retrospective surveys toward continuous, objective burnout prediction.

These findings have important implications for physician wellness at the institutional level.

Similar tools could be embedded into hospital systems or occupational health platforms, combining wearable metrics, workload indicators, and periodic wellness check-ins to proactively identify physicians at risk. This predictive capacity represents a meaningful shift from reactive crisis management toward preventive support, giving healthcare organizations the ability to intervene before burnout escalates (22). Furthermore, by connecting early burnout detection to patient outcomes, and institutional costs, these tools present a compelling case for health systems to invest in AI-enabled physician wellness monitoring as a matter of both workforce sustainability, and patient safety.

Conclusion

Physician's burnout is not simply an individual issue but a systemic challenge with wide-reaching implications for patient care, workforce stability, and the sustainability of Canada's health system. The persistence of high burnout rates demonstrates that current approaches are insufficient. Traditional initiatives such as peer support groups, resilience workshops, counseling remain hindered by stigma, inconsistent engagement, and limited evidence of durable outcomes (11-14).

Multiple studies demonstrated how AI can intervene at multiple points in the wellness continuum. Conversational agents offer a private, stigma-free entry point to mental health support, while predictive models demonstrate that burnout risk can be identified with increasing accuracy before symptoms escalate. Together,

these developments point to an emerging paradigm where AI shifts wellness from reactive treatment toward proactive prevention (15-22).

Nevertheless, enthusiasm for AI must be tempered with caution (23). The use of sensitive mental health data introduces concerns regarding privacy, data security, and the appropriate handling of personal health information. Algorithmic bias is also a key concern, as models trained on non-representative datasets may inaccurately assess burnout risk across different demographic groups, potentially reinforcing existing disparities. Furthermore, the use of continuous monitoring tools, such as wearable devices or digital activity logs, raises questions about informed consent, and the extent to which individuals are aware of, and agree to how their data are collected, and used (23,24).

There is also a risk that such systems may contribute to a culture of surveillance, particularly if data are used by institutions to monitor performance rather than to support well-being. This highlights the importance of clearly defining institutional responsibility, ensuring that data are used ethically, transparently, and solely for supportive purposes. As such, robust governance frameworks, transparency in algorithm design, and strict safeguards around data use are essential to ensure that AI-enabled wellness interventions are implemented in a manner that is ethical, equitable, and aligned with the interests of healthcare workers, and trainees (25).

Beyond individual-level tools, it is important to acknowledge that burnout is fundamentally driven by structural conditions, excessive workloads, staffing shortages, EMR burden, and organizational culture, that AI alone cannot resolve. While AI can monitor physician well-being, and flag early signs of distress, these signals must be received by institutions willing to act on them: redistributing workloads, addressing staffing gaps, reducing administrative burden, and fostering cultures of psychological safety. Without this institutional responsiveness, AI risks becoming a sophisticated monitoring system that identifies suffering without addressing its root causes. The value of AI in physician wellness therefore lies not in replacing structural reform, but in making the need for it visible, and urgent (25, 26).

Looking forward, investment in rigorous, large-scale trials will be critical to validate effectiveness, identify best practices, and build trust among physicians, learners, and institutions. If guided by evidence, and implemented responsibly, AI-enabled interventions have the potential to transform wellness initiatives from fragmented, short-term fixes into sustainable, system-level solutions. In doing so, they can help secure not only the resilience of Canada's medical workforce but also the quality, and accessibility of patient care for the generations ahead.

References

1. Canadian Medical Association, "CMA National Physician Health Survey: A national snapshot" (2022); https://digitallibrary.cma.ca/link/digitallibrary18.
2. L. S. Rotenstein, M. Torre, M. A. Ramos, R. C. Rosales, C. Guille, S. Sen, et al., Prevalence of burnout among physicians. JAMA 320, 1131–1132 (2018).
3. Canadian Medical Association, "Physician wellness: New 2021 National Physician Health Survey findings—burnout, short-staffing, and an overburdened system take their toll" (2022); https://www.cma.ca/physician-wellness-hub/content/physician-wellness-new-2021-national-physician-health-survey-findings-burnout-short-staffing-and-overburdened-system-take-their-toll.
4. Canadian Medical Association, "CMA 2021 National Physician Health Survey" (2022); https://digitallibrary.cma.ca/link/digitallibrary17.
5. Government of Canada, "Experiences of health care workers during the COVID-19 pandemic, September to November 2021" (2022); https://www150.statcan.gc.ca/n1/daily-quotidien/220603/dq220603a-eng.htm.
6. R. Walsh, D. Telner, D. A. Butt, P. Krueger, K. Fleming, S. MacDonald, et al., Factors associated with plans for early retirement among Ontario family physicians during the COVID-19 pandemic: A cross-sectional study. BMC Prim. Care 25, 67 (2024). https://doi.org/10.1186/s12875-024-02374-9.
7. K. Li, A. Frumkin, W. G. Bi, J. Magrill, C. Newton, Biopsy of Canada's family physician shortage. Fam. Med. Community Health11, e002236 (2023). https://doi.org/10.1136/fmch-2023-002236.
8. Public Health Agency of Canada, "Aging and chronic diseases: Profile of Canadian seniors report" (2022); https://www.canada.ca/en/public-health/services/publications/diseases-conditions/aging-chronic-diseases-profile-canadian-seniors-report.html.
9. D. Duong, L. Vogel, National survey highlights worsening primary care access. CMAJ 195, E547 (2023). https://doi.org/10.1503/cmaj.1096049.
10. J. E. Wallace, J. B. Lemaire, W. A. Ghali, Physician wellness: A missing quality indicator. Lancet 374, 1714–1721 (2009). https://doi.org/10.1016/S0140-6736(09)61424-0.
11. A. J. Bannatyne, C. Jones, B. M. Craig, D. Jones, K. Forrest, A systematic review of mental health interventions to reduce self-stigma in medical students and doctors. Front. Med. 10, 1204274 (2023). https://doi.org/10.3389/fmed.2023.1204274.
12. M. Berliant, N. Rahman, C. Mattice, C. Bhatt, K.-A. Haykal, Barriers faced by medical students in seeking mental healthcare: A scoping review. medRxiv [Preprint] (2022). https://doi.org/10.21203/rs.3.rs-1344013/v1.
13. C. Regehr, D. Glancy, A. Pitts, V. R. LeBlanc, Interventions to reduce the consequences of stress in physicians. J. Nerv. Ment. Dis. 202, 353–359 (2014). https://doi.org/10.1097/NMD.0000000000000130.

14. M. Panagioti, E. Panagopoulou, P. Bower, G. Lewith, E. Kontopantelis, C. Chew-Graham, et al., Controlled interventions to reduce burnout in physicians. JAMA Intern. Med. 177, 195–205 (2017). https://doi.org/10.1001/jamainternmed.2016.7674.
15. K. K. Fitzpatrick, A. Darcy, M. Vierhile, Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Ment. Health 4, e19 (2017). https://doi.org/10.2196/mental.7785.
16. B. Inkster, S. Sarda, V. Subramanian, An empathy-driven conversational artificial intelligence agent (WYSA) for digital mental well-being: Real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 6, e12106 (2018). https://doi.org/10.2196/12106.
17. C. Chen, K. T. Lam, K. M. Yip, H. K. So, T. Y. Lum, I. C. Wong, et al., Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: Pilot randomized controlled trial. JMIR Hum. Factors 12, e65785 (2025). https://doi.org/10.2196/65785.
18. A. S. Miner, A. Milstein, S. Schueller, R. Hegde, C. Mangurian, E. Linos, Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Intern. Med. 176, 619–625 (2016). https://doi.org/10.1001/jamainternmed.2016.0400.
19. R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, M. Rauws, Using psychological artificial intelligence (TESS) to relieve symptoms of depression and anxiety: Randomized controlled trial. JMIR Ment. Health 5, e9782 (2018). https://doi.org/10.2196/mental.9782.
20. H. Liu, S. S. Lou, B. C. Warner, D. R. Harford, T. Kannampallil, C. Lu, HiPAL: A deep framework for physician burnout prediction using activity logs in electronic health records. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022), pp. 3377–3387. https://doi.org/10.1145/3534678.3539056.
21. C. Liu, Y.-C. Chuang, L. Qin, L. Ren, C.-W. Chien, T.-H. Tung, Machine-learning-based model for analysing and accurately predicting factors related to burnout in healthcare workers. BMJ Public Health 3, e000777 (2025). https://doi.org/10.1136/bmjph-2023-000777.
22. A. R. Wilton, K. Sheffield, Q. Wilkes, S. Chesak, J. Pacyna, R. Sharp, et al., The burnout prediction using wearable and Artificial Intelligence (BROWNIE) study: A decentralized digital health protocol to predict burnout in registered nurses. BMC Nurs. 23, 56 (2024). https://doi.org/10.1186/s12912-024-01711-8.
23. T. Pham, Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. R. Soc. Open Sci. 12, 241873 (2025). https://doi.org/10.1098/rsos.241873.
24. K. C. Kellogg, M. A. Valentine, A. Christin, Algorithms at work: The new contested terrain of control. AI Ethics 3, 703–720 (2023). https://doi.org/10.1007/s43681-023-00275-8.
25. I. Dankwa-Mullan, Health equity and ethical considerations in using artificial intelligence in public health and medicine. Prev. Chronic Dis. 21, 240245 (2024). https://doi.org/10.5888/pcd21.240245.
26. T. D. Shanafelt, C. P. West, L. N. Dyrbye, et al., Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2023. Mayo Clin. Proc. 100 (2025). https://doi.org/10.1016/j.mayocp.2024.11.031.

Rethinking the Central Dogma: Protein Amyloids acting as Transgenerational Epigenetic Memory Carriers

Commentary on "Noncanonical Inheritance of Phenotypic Information by Protein Amyloids" by Matthew Eroglu et al. (September 2, 2024)

Repenser le dogme central : les amyloïdes protéiques agissant comme vecteurs de mémoire épigénétique transgénérationnelle

Commentaire sur « Hérédité non canonique de l'information phénotypique par les amyloïdes protéiques » par Matthew Eroglu et al. (2 septembre 2024)

Shreya Pal[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: spal009@uottawa.ca

Abstract | Résumé

Nucleic acids remain the primary mechanism for transmitting hereditary information across generations. Despite advances in genome-wide association studies and epigenetic reprogramming mechanisms, many familial traits and disease susceptibilities remain unexplained, a gap known as "missing heritability" (1). Conventionally, epigenetic inheritance is attributed to small RNA or chromatin/histone modifications. However, a revolutionary finding by Matthew Eroglu and colleagues identified amyloid-like protein aggregates in *Caenorhabditis elegans* (*C. elegans*) that persist across generations to influence developmental phenotypes. This suggests proteins can act as independent carriers of transgenerational epigenetic memory. This commentary examines how these findings challenge the Central Dogma, expand inheritance models, and redefine amyloids beyond disease pathology.

———

Les acides nucléiques restent le principal mécanisme de transmission d'informations héréditaires à travers les générations. Malgré les avancées dans les études d'association à l'échelle du génome et les mécanismes de reprogrammation épigénétique, de nombreux traits familiaux et susceptibilités aux maladies restent inexpliqués, un écart connu sous le nom de « manque d'héritabilité » (1). Conventionnellement, l'hérédité épigénétique est attribuée à de petits ARN ou à des modifications de chromatine/histones. Cependant, une découverte révolutionnaire de Matthew Eroglu et de ses collègues a identifié des agrégats protéiques amyloïdes chez *Caenorhabditis elegans* (*C. elegans*) qui persistent à travers les générations pour influencer les phénotypes développementaux. Cela suggère que les protéines peuvent agir comme porteuses indépendantes de la mémoire épigénétique transgénérationnelle. Ce commentaire examine comment ces résultats remettent en question le dogme central, élargissent les modèles d'hérédité et redéfinissent les amyloïdes au-delà de la pathologie des maladies.

Keywords: Protein amyloids, epigenetic inheritance, central dogma, prions, transgenerational memory, noncanonical inheritance

Introduction

The phenomenon of "missing heritability" arises from puzzling relationships in heritable conditions that cannot be justified by known transmissible molecular mechanisms. While transgenerational epigenetic inheritance can transmit adaptive traits independently of DNA sequence, most epigenetic marks are extensively erased during gametogenesis and embryonic development. This eradication limits their long-term stability across generations, suggesting that additional inheritance mechanisms exist beyond conventional nucleic acid-based models.

Additionally, sexually reproducing organisms transmit parental factors such as gametic proteins to progeny during early embryogenesis. The dynamic regulation of a balanced, functional proteome, encompassing the cellular mechanisms that control protein synthesis, folding, trafficking, and degradation, is referred to as proteostasis. These protein quality control systems are highly upregulated in stem and germline cells, suggesting that proteostasis is crucial for self-renewal.

Amyloids are solid-phase fibrillar protein oligomers characterized by cross β-sheets that allow them to self-propagate by transferring their structure onto native proteins. A notorious sub-class of amyloids called prions can convert native proteins into distinct prion conformations, disrupting normal cellular functions and infecting nearby cells (2). However, amyloids are not solely pathogenic; they play essential roles in hormone regulation, such

as facilitating storage and secretion of peptide hormones (3). Furthermore, controlled transitions between amyloid aggregation and disaggregation are strictly required for proper embryonic development in yeast, worms, and flies (4). Despite these developmental functions, this specific mechanism of heritability was not likely to be investigated further, as higher-order organisms clear such aggregates during embryogenesis, allowing for no method of transmission (5).

During early development, most parental epigenetic information, including DNA methylation and chromatin modifications, is erased and replaced by zygotic material, thereby limiting the long-term stability of inherited epigenetic states (6). Consequently, amyloid-like aggregates observed in metazoan germ cells and embryos were generally assumed to be transient structures that are cleared during development. However, this paper challenged this assumption by demonstrating that amyloid-like protein aggregates in *C. elegans* persist beyond the maternal-to-zygotic transition and are stably inherited across generations (7).

Structural findings

The authors identified cytoplasmic amyloid-like puncta within developing oocytes that subsequently travel into embryos and spread throughout somatic and germline tissues during development (7). Unlike transient protein aggregates associated with cellular stress or degeneration, these structures persist across generations, producing distinct developmental phenotypes. This transgenerational persistence challenges the long-standing assumption that protein aggregates are fully cleared during embryogenesis, suggesting instead that proteins can carry heritable biological memory independent of genomic DNA sequences.

To investigate their functional significance, the authors examined two conserved AN1-domain proteins, mstr-1 and mstr-2, which regulate 26S proteasomal selectivity to maintain proper amyloid homeostasis. Absence of MSTR proteins disrupted normal proteasomal regulation, allowing heritable amyloid aggregates to progressively accumulate and gradually transform sperm-producing germ cells into functional oocytes over multiple generations. Remarkably, this phenotype intensified gradually across generations and could be reversed under permissive environmental conditions — hallmarks that are highly consistent with epigenetic inheritance rather than irreversible genomic mutation (7). The reversibility of the phenotype strongly supports a dynamic, transmissible, protein-mediated inheritance system rather than a purely genetic mechanism.

Among the most compelling findings in the study were the aggregate injection experiments, which provided strong evidence for protein-mediated inheritance. Injection of amyloid-like aggregates isolated from mstr-deficient feminized worms into naïve hermaphrodites caused reduced fertility and progressive feminization that persisted for at least five generations. In contrast, aggregates isolated from wild-type worms maintained under normal conditions produced minimal developmental effects, suggesting that the heritable influence of the aggregates depends not only on their presence, but also on their conformational or compositional state (7). The self-propagating behavior of these aggregates resembles fungal prions, in which protein conformations template similar structural states onto native proteins. However, unlike pathogenic prions associated with neurodegenerative disease, the amyloid-like structures identified in this study appear to function physiologically in developmental regulation and reproductive fitness.

Importantly, the feminization phenotype intensified progressively across generations yet reverses when worms return to permissive growth conditions following maintenance at elevated temperature. Worms propagated for multiple generations at 25°C gradually regained normal spermatogenesis and fertility after being shifted back to 20°C, demonstrating that the phenotype was not permanently fixed (7). This reversibility represents one of the strongest pieces of evidence supporting an epigenetic rather than genetic mechanism of inheritance. If the phenotype were caused solely by irreversible genomic mutation, restoration of normal spermatogenesis across subsequent generations would not be expected. Instead, the gradual accumulation and subsequent reversal of the phenotype suggest that the inherited factor is dynamic, environmentally responsive, and capable of being remodeled over time. These observations support the authors' proposal that amyloid-like protein aggregates function as transmissible epigenetic memory carriers whose abundance or conformational state can be altered by environmental conditions such as temperature. Collectively, the findings demonstrate not only the heritability of protein-mediated phenotypes, but also the remarkable plasticity of this inheritance system, a defining characteristic of epigenetic regulation.

Under normal conditions, proteasomes selectively degrade misfolded or damaged proteins to preserve cellular proteostasis. However, mutations affecting MSTR proteins disrupted proteasomal selectivity, leading to altered amyloid accumulation and compensatory activation of proteasomal pathways (7). Genetic suppressors targeting 26S proteasomal regulatory subunits restored normal phenotypes, further supporting the conclusion that regulated proteostasis is central to maintaining this protein-based epigenetic memory system.

Mechanism of Action for MSTR Proteins in Proteasomal Selectivity

Further investigation of how MSTR proteins regulate sex-determining pathways through proteasomal activity demonstrated the regulatory function of MSTR proteins. The MSTR proteins maintained a critical balance between the sex-regulatory proteins GLD-1 and TRA-1 by controlling selective protein degradation through the 26S proteasome. In wild-type worms, MSTR-1 expression was highest in spermatogenic germ cells, while GLD-1 expression remained restricted to regions associated with oogenesis (7). However, in mstr mutant worms, GLD-1

accumulated abnormally in spermatogenic regions and progressively increased over generations, coinciding with increasing germline feminization (7). Simultaneously, expression of TRA-1 progressively declined across generations, further shifting germ cells toward oocyte differentiation.

Pharmacological inhibition of proteasomal activity partially restored TRA-1 expression and spermatogenesis in early generations, supporting the conclusion that altered proteasomal regulation drives the inherited phenotype (7). The progressive nature of these molecular changes provides additional evidence that the observed feminization is mediated by an epigenetically inherited factor rather than irreversible genetic mutation. By linking protein homeostasis to transgenerational developmental regulation, the study strengthens the argument that amyloid-associated protein states may function as stable carriers of heritable biological information.

Amyloid-Like Aggregates as Epigenetic Memory Carriers

A major strength of the study is that the authors systematically eliminated several established mechanisms of transgenerational epigenetic inheritance before proposing amyloid-like protein aggregates as the heritable factor responsible for germline feminization. In *C. elegans*, inheritance is commonly mediated through chromatin modifications or small RNA (sRNA)-dependent pathways (8). However, expression levels of the sex-determining genes gld-1 and tra-1 remained unchanged across generations and environmental conditions, suggesting that transcriptional regulation through DNA methylation, histone modifications, or direct mRNA inheritance was unlikely to explain the observed phenotype (7). Furthermore, disruption of multiple endogenous sRNA pathways, including miRNA, piRNA, 22G RNA, and 26G RNA systems, failed to suppress germline feminization in mstr mutant worms (7). These findings support the conclusion that the inherited developmental effects observed in the study occur independently of currently established epigenetic inheritance pathways.

While imaging feminized worms, the authors identified prominent green autofluorescent puncta within the germline that progressively accumulated over generations. These were colocalized with multiple amyloid-specific dyes, including Proteostat, Thioflavin T, and Amytracker, indicating that the structures possessed amyloid-like properties (7).

To determine whether amyloid accumulation directly contributed to germline feminization, the researchers treated worms with structurally distinct anti-amyloid compounds, including curcumin, baicalein, and epigallocatechin-3-gallate (EGCG) (7). These treatments reduced amyloid accumulation and restored self-fertility in MSTR worms, providing functional evidence that amyloid formation contributes to the inherited phenotype. Further investigation into the composition and inheritance of these structures revealed proteins previously associated with aggregation-prone assemblies, including RHO-1 and vitellogenins, alongside proteasomal subunits implicated in proteostasis regulation (7). Although wild-type and MSTR worms displayed similar amyloid compositions, aggregates isolated from MSTR mutants exhibited altered structural properties and greater resistance to proteolytic digestion, suggesting that differences in aggregate conformation, rather than protein identity alone, may underlie the inherited phenotypic effects observed across generations.

Using fluorescently tagged VIT-2 reporters and photoconversion experiments, the authors demonstrated that maternally derived amyloid-associated proteins persisted throughout embryogenesis and larval development in offspring (7). Similar persistence was observed following injection of fluorescently labelled amyloids into wild-type germlines, providing direct evidence that physiological amyloid-like proteins are physically inherited and remain stable across developmental stages.

To directly trace inheritance of these structures, the authors generated fluorescently tagged VIT-2 reporters that colocalized with amyloid-like bodies in the germline. Through irreversible photoconversion experiments, maternally derived VIT-2 aggregates were visualized persisting throughout embryogenesis and into larval development in offspring. Importantly, these inherited protein puncta remained detectable in the germline during later stages of gametogenesis, where newly synthesized amyloid-associated proteins accumulated around pre-existing parental aggregates (7). Similar persistence was observed following injection of fluorescently labelled isolated amyloids into wild-type germlines, with labelled aggregates subsequently detected in both somatic and germline tissues of progeny. Collectively, these experiments provide direct evidence that physiological amyloid-like proteins are physically transmitted between generations and remain stable throughout development.

Limitations

Despite the study's strengths, the precise molecular composition of the amyloid-like structures remains incompletely defined, and it is unclear whether a single core aggregate species is responsible for inheritance or whether multiple heterogeneous protein assemblies contribute collectively. Similarly, the injection experiments utilized mixed amyloid populations, preventing definitive identification of the minimal factor required for phenotypic transmission. The anti-amyloid compounds used in the study also possess antioxidant activity, introducing the possibility that some observed effects may partially arise from altered oxidative stress rather than amyloid disruption alone. Nevertheless, the convergence of genetic, biochemical, imaging, pharmacological, and inheritance-based evidence supports the authors' central conclusion that amyloid-like protein aggregates participate in transgenerational epigenetic regulation.

Implications for Current Scientific Perspective

Perhaps most importantly, the study demonstrates that inherited protein aggregates can coexist with newly synthesized proteins in progeny and influence developmental outcomes long after fertilization. This persistence challenges the long-standing assumption that protein-based cellular states are transient and erased during embryogenesis. Instead, the work supports a model in which proteins may store conformational information capable of acting as a stable epigenetic memory system. While additional research is necessary to determine whether similar mechanisms operate in vertebrates or humans, the findings presented by Eroglu and colleagues substantially broaden current understanding of inheritance beyond nucleic acids and suggest that regulated amyloid formation may represent an evolutionarily conserved mechanism for transmitting adaptive phenotypic information across generations.

References

1. L. J. Matthews, E. Turkheimer, Three legs of the missing heritability problem. Studies in history and philosophy of science 93, 183-191 (2022).
2. A. I. P. Taylor, R. A. Staniforth, General Principles Underpinning Amyloid Structure. Frontiers in Neuroscience 16, e878869 (2022).
3. S. K. Maji, M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. R. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. River, P. Sawchenko, W. Vale, R. Riek, Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules. Science 325, 328–332 (2009).
4. J. S Fassler, S. Skuodas, D. L. Weeks, and B. T. Phillips. Protein Aggregation and Disaggregation in Cells and Development. Journal of Molecular Biology 433, e167215 (2021).
5. M. H. Hayes, Daniel E Weeks. Amyloids Assemble as Part of Recognizable Structures during Oogenesis in Xenopus. Biology Open 5, 801–806 (2016).
6. Q. Liu, X. Ma, X. Li, X. Zhang, S. Zhou, L. Xiong, Y. Zhao, D. Zhou, Paternal DNA methylation is remodeled to maternal levels in rice zygote. Nature Communications 14, e6571 (2023).
7. M. Eroglu, A. Zocher, J. McAuley, R. Webster, M. Z. X. Xiao, B. Yu, C. Mok, W. B. Derry, Noncanonical inheritance of phenotypic information by protein amyloids. Nat Cell Biol 26, 1712-1724 (2024).
8. R. M. Woodhouse, A. Ashe, How do histone modifications contribute to transgenerational epigenetic inheritance in C. elegans?. Biochemical Society Transactions 48, 1019-1034 (2020).

Strengthening Social Determinants of Health Education in Canadian Medical Schools

Renforcement de l'éducation des déterminants sociaux à la santé dans les facultés de médecine canadiennes

Parastoo Golzarian[1*]

1. Toronto Metropolitan University, Toronto, ON, Canada
*Corresponding author. Email: parastoo.golzarian@gmail.com

Abstract | Résumé

The social determinants of health (SDOH), including income, housing, and education, are major drivers of health outcomes in Canada. Ignoring these factors can lead to misinterpretation of patient behaviours and reinforcing inequities. Housing insecurity and poverty increase risks of mortality, poor perinatal outcomes, and cancer disparities, with disadvantaged, immigrant, Indigenous, and racialized groups disproportionately affected. Canadian medical schools teach SDOH through lectures, case discussions, and service-learning, yet current approaches are fragmented and often optional. This paper argues for structured, longitudinal advocacy training to better prepare future physicians to address inequities and advance social accountability in healthcare.

Les déterminants sociaux de la santé (DSS), notamment le revenu, le logement et l'éducation, sont des moteurs majeurs des résultats de santé au Canada. Ignorer ces facteurs peut conduire à une mauvaise interprétation des comportements des patients et renforcer les inégalités. L'insécurité du logement et la pauvreté augmentent les risques de mortalité, de mauvais résultats périnataux et les disparités liées au cancer, affectant de manière disproportionnée les populations défavorisées, immigrées, autochtones et racialisées. Les écoles de médecine canadiennes enseignent le DSS à travers des conférences, des discussions de cas et l'apprentissage par le service, mais les approches actuelles sont fragmentées et souvent optionnelles. Cet article plaide pour une formation structurée et longitudinale au plaidoyer afin de mieux préparer les futurs médecins à lutter contre les inégalités et à promouvoir la responsabilité sociale dans le secteur de la santé.

Keywords: Medical education; Social accountability; Social determinants of health; Curricula; Clinical reasoning

Introduction

In Canada, medical schools teach students about the social determinants of health (SDOH). These are factors such as income, housing, and education that strongly shape patient outcomes (1). If physicians overlook these factors, they may misinterpret a patient's situation. Oftentimes, physicians label someone "non-compliant" with medications when the real issue is that they cannot afford the prescription or that they lack transportation to the pharmacy (2).

Housing instability is strongly linked to higher mortality, yet these risks are often overlooked in clinical encounters if physicians are not trained to ask about them (3). Living in rental housing versus ownership is associated with higher rates of preterm birth, stillbirth, and infant mortality, showing how housing insecurity directly shapes clinical outcomes (4). Research shows moving patients into stable social housing significantly improves mental health and overall patient outcome, showing how attention to SDOH can transform outcomes beyond what can be achieved by medical treatment alone (5, 6).

Housing is one example of a broader pattern, across multiple health domains, of social conditions that predict who gets sick, who gets diagnosed, and who survives. Cancer is one of the most common diseases in Canada. It remains the leading cause of death, responsible for an estimated 85,100 deaths in 2022 and accounting for nearly one in four deaths nationwide, with recent reports projecting a continued substantial burden on patients and the healthcare system (7, 8). Advances in screening, diagnosis, and treatment have reduced mortality rates for many cancer patients (8), yet significant disparities in morbidity and mortality remain. These disparities are strongly linked to SDOH. From screening to diagnosis to treatment, patients from socioeconomically disadvantaged communities in Canada experience worse cancer outcomes (9). For instance, lower-income and rural Canadians face a higher risk of both developing and dying from cancer (10). Persistent inequities in cancer screening are evident, as recent immigrants and Indigenous communities (often facing socioeconomic disadvantages) consistently show lower

participation rates (11–13). These disparities are not limited to socioeconomic status or immigration experience; racialized communities face compounding disadvantages even when controlling for income. Black, African-Caribbean, and South Asian Canadians are screened less often and experience higher cancer and mortality rates (14, 15), reflecting how race and ethnicity independently shape access to care. Beyond these groups, Two-Spirit, lesbian, gay, bisexual, transgender, queer, intersex, and other sexual/gender minority patients with cancer often report worse healthcare experiences and greater unmet support needs compared to other patients (16), illustrating how inequity in cancer care cuts across multiple dimensions of identity and social positioning. These disparities are perpetuated in part through clinical encounters shaped by inadequate training. Addressing these inequities requires structured training. Without proper training in SDOH, physicians are not equipped to take a meaningful patient social history, recognize structural barriers, or connect patients to appropriate community resources (2). A physician who does not know to ask about housing, income, or immigration status cannot identify the conditions driving a patient's poor health outcomes, and without that identification, no referral and no systemic intervention can follow. Patients from lower-income, Indigenous, immigrant, and racialized communities bear the greatest burden of these missed opportunities as their health outcomes are most sensitive to whether a physician can look beyond the clinical presentation and address the conditions shaping it (1). The training gaps described in this paper are therefore not abstract educational concerns, but instead translate directly into the inequities that disadvantaged Canadians continue to experience.

This gap exists not for lack of formal recognition. Frameworks such as the Future of Medical Education in Canada and the CanMEDS Health Advocate role have long required students to understand these issues, and many schools highlight social accountability in their curricula (1, 17). Students are exposed to SDOH through lectures, case discussions, community projects, and some clinical placements. This paper argues that while Canadian medical schools have made progress in teaching social determinants of health, more structured and longitudinal clinical reasoning training and social accountability are required to prepare future physicians to address these inequities in patient care and improve their clinical reasoning (18).

Current Efforts and Gaps in Education

Most SDOH curricula include a mix of classroom lectures and community exposure, with many implemented longitudinally over time (1). Some Canadian medical students also participate in extracurricular programs, such as the Longitudinal Advocacy Training Series, which offers virtual workshops on skills developed by the Canadian Federation of Medical Students (19). Schools also use service-learning models, where students engage with communities outside of clinical settings to connect theory and practice (20).

Across Canada, the depth and structure of SDOH training varies considerably between institutions. The University of British Columbia's Flexible Enhanced Learning program offers students self-directed scholarly activities across first to fourth years that can include community health and advocacy projects, with social accountability explicitly embedded as a curricular goal. However, students self-direct their topic, meaning SDOH content is not guaranteed (20). The University of Toronto runs a mandatory two-year longitudinal program called Health in Community, connecting students with community partners and co-educators to develop SDOH competencies in real-world settings (21). McMaster University integrates service-learning into its curriculum as a structured community placement experience focused on SDOH and social accountability; however, it is registered as a horizontal elective and requires a minimum of only four hours, making it effectively optional and leaving the depth of engagement entirely up to individual students (22).

The University of Ottawa delivers SDOH content primarily through didactic lectures in the first and second years, complemented by 10 mandatory experiential learning logs completed during third-year clerkship that require students to reflect on social accountability in clinical settings. Additional community exposure is available through the student-led Health Initiative Partnership Clinic; however, participation remains voluntary and dependent on student initiative rather than being embedded in the formal curriculum (23, 24). While each of these models reflects institutional commitment to SDOH, none consistently progress students through all three competency levels, and most focus on knowledge and exposure rather than applied or structural change.

A challenge with many courses on SDOH are that they are short and uneven in depth, often treating SDOH as "facts to be known" rather than conditions to be changed, which leaves students with little practice in applying this knowledge at the policy or system level (2, 25). Moreover, knowledge about SDOH is also usually embedded only in public health or social medicine courses and community clerkships, and less often built across all years of training in a cohesive progression. This causes many students to never get the opportunity to apply what they learn (26). Service-learning projects also help students understand community health, but they vary between schools, and in many cases, they are optional, so not every student benefits from them (27).

This raises an important equity concern within medical education itself: optional experiences disproportionately benefit students who have more available time, prior experience, and fewer financial constraints. Students who work part-time jobs, have caregiving responsibilities, or who come from backgrounds with less exposure to community advocacy are less likely to participate. As a result, clinical reasoning skills become unevenly distributed across the graduating physician workforce. This has implications for which communities ultimately receive care from physicians equipped to address structural barriers. Making service-learning and experiences mandatory rather than optional is therefore not only a curricular decision but an equity imperative (27).

Addressing Gaps in Education

To strengthen current efforts, medical schools could adopt structured, longitudinal pathways so students progressively build skills rather than receive isolated exposure. For example, integrating clinical reasoning competencies into core clinical rotations, adding mandatory workshops or modules on structural determinants and policy, and embedding SDOH assessment tasks into patient encounters would help. Schools could also expand partnerships with community organizations and legal, housing, and social services to give students real opportunities to intervene. Further, aligning assessment and evaluation so that SDOH competencies are graded and included in big exams or high-stakes assessments (instead of being optional) would signal their importance institutionally. Currently, SDOH-related clinical reasoning skills are rarely included in formal evaluations, signalling to students that they are supplementary rather than core (2). Concrete strategies to address this include I) Objective Structured Clinical Examination stations that assess students' ability to identify social risk factors and respond appropriately during a patient encounter (28, 29), II) Entrustable Professional Activities tied specifically to SDOH-informed care, such as conducting a social history or connecting a patient to community resources (30), III) structured reflective assignments following community placements, requiring students to analyze the structural conditions affecting the populations with which they worked, and IV) community-based assessments, evaluated jointly by faculty and community partner organizations, giving weight to reasoning skills that cannot be measured in a clinical setting alone. Embedding these tools into formal grading would institutionalize understanding SDOH factors as a core expectation rather than an optional enhancement and it would provide the kind of accountability that encourages faculty to consistently model and teach these skills.

Conclusion

Canadian medical schools made meaningful progress in teaching SDOH, but the evidence presented in this paper makes it clear that awareness alone is insufficient. Housing instability, cancer disparities, and inequities in screening and survival among Indigenous, immigrant, racialized, and Two-Spirit, lesbian, gay, bisexual, transgender, queer, intersex, and other sexual/gender minority Canadians are not inevitable. Instead, they are shaped by structural conditions that a well-trained physician can help identify, address, and challenge (3–16).

Closing this gap requires a deliberate, three-part curricular model. First, all medical students should develop a foundational understanding of SDOH, not as background knowledge, but as clinical information essential to accurately diagnosis and provide patient-centered care (1, 20). Second, this knowledge must be translated into clinical skills: taking social histories, screening for housing instability and food insecurity, recognizing when "non-compliance" reflects structural barriers, and connecting patients to appropriate community resources (2, 21). Third, students must be trained to advocate for structural change and health system reform, and engage with policy and community organizations to address the root causes of the inequities they encounter in practice (21, 25).

Achieving this requires concrete institutional commitments: mandatory rather than optional service-learning; Objective Structured Clinical Examination stations and Entrustable Professional Activities that formally assess SDOH competencies; structured reflective assignments following community placements; faculty development programs that equip teachers to model these skills; sustained partnerships with community organizations and legal, housing, and social services (19). Without these structural supports, curricular reform will remain aspirational.

Physicians who are trained to see beyond the clinical presentation, to ask about housing, income, immigration status, and identity, are better equipped to interrupt the pathways through which SDOH produces preventable harm. By making structural competency core expectations rather than supplementary experiences, Canadian medical education can fulfill its broader mandate of social accountability and begin to reduce the health disparities that disadvantaged Canadians continue to experience (1, 17).

References

1. K. A. Hunter, B. Thomson, A scoping review of social determinants of health curricula in post-graduate medical education. Can. Med. Educ. J.10, e61709 (2019). 10.36834/cmej.61709
2. N. Nour, D. Stuckler, O. Ajayi, M. E. Abdalla, Effectiveness of alternative approaches to integrating SDOH into medical education: A scoping review. BMC Med. Educ. 23, e18 (2023). 10.1186/s12909-022-03899-2
3. E. C. Draper, H. J. Burgess, C. Chisholm, E. L. Mazerolle, C. Barker, Front-line insights into the social determinants of health in housing instability: A multi-province study. J. Prim. Care Community Health 15, e21501319241292131 (2024). 10.1177/21501319241292131
4. A. Mehrabadi, G. D. Shapiro, J. S. Kaufman, S. Yang, Housing and preterm birth, stillbirth and neonatal death in Canada: A population-based study using 2006 and 2016 National Census Data. Epidemiology 36, 647–655 (2025).
5. J. R. Dunn, K. L. W. Smith, P. Smith, R. Moineddin, F. I. Matheson, S. W. Hwang, C. Muntaner, M. Janus, P. O'Campo, Does receipt of social housing impact mental health? Results of a quasi-experimental study in the Greater Toronto Area. Soc. Sci. Med. 362, e117363 (2024). 10.1016/j.socscimed.2024.117363
6. M. E. Marziali, S. Hansen, K. W. Kooij, M. Budu, M. Ye, C. Tam, T. McLinden, S. D. Emerson, J. S. G. Montaner, S. Parashar, R. S. Hogg, Housing matters: The long-term impact of stable housing on mortality among people with HIV in Brisitsh Colobia, Canada. Soc. Sci. Med. 367, e117713 (2025). 10.1016/j.socscimed.2025.117713

7. D. R. Brenner, J. Gillis, A. A. Demers, L. F. Ellison, J.-M. Billette, S. X. Zhang, J. Q. L. Liu, R. R. Woods, C. Finley, N. Fitzgerald, N. Saint-Jacques, L. Shack, D. Turner, Projected estimates of cancer in Canada in 2024. CMAJ 196, e625–e634 (2024). 10.1503/cmaj.240095
8. M. T. Warkentin, Y. Ruan, L. F. Ellison, J.-M. Billette, A. Demers, F.-F. Liu, D. R. Brenner, Progress in site-specific cancer mortality in Canada over the last 70 years. Sci. Rep. 14, e5688 (2024). 10.1038/s41598-024-56150-x
9. "Annual Report 2015/16" (Canadian Partnership Against Cancer, 2016). _www.partnershipagainstcancer.ca/wp-content/uploads/2016/07/cpac-annual-report-2015-16-en-final2.pdf
10. P. Tope, S. Morais, M. El-Zein, E. L. Franco, T. Malagón, Differences in site-specific cancer incidence by individual- and area-level income in Canada from 2006 to 2015. Int. J. Cancer 153, 1766–1783 (2023).
11. K. A. Benjamin, N. Lamberti, M. Cooke, Predictors of non-adherence to cervical cancer screening among immigrant women in Ontario, Canada. Prev. Med. Rep. 36, e102524 (2023). 10.1016/j.pmedr.2023.102524
12. H. Yang, A. Letendre, M. Shea-Budgell, L. Bill, B. A. Healy, B. Shewchuk, G. Nelson, J. Newsome, B. Chiang, C. R. Rahul, K. A. Kopciuk, Cervical cancer screening outcomes among First Nations and non–First Nations women in Alberta, Canada. Cancer Epidemiol. 93, e102672 (2024). 10.1016/j.canep.2024.102672
13. Statistics Canada, "Cancer screening tests" (Government of Canada, 2025). www150.statcan.gc.ca/n1/daily-quotidien/250806/dq250806a-eng.htm
14. J. Hwee, E. Bougie, Do cancer incidence and mortality rates differ among ethnicities in Canada? Health Reports 32, e8 (2021). 10.25318/82-003-x202100800001-eng
15. D. A. Ezeife, G. Padmore, M. Vaska, T. H. Truong, Ensuring equitable access to cancer care for Black patients in Canada. CMAJ 194, e1412–e1417 (2022). 10.1503/cmaj.212076
16. D. Comeau, C. Johnson, N. Bouhamdani, Review of current 2SLGBTQIA+ inequities in the Canadian health care system. Front. Public Health 11, e1183284 (2023). 10.3389/fpubh.2023.1183284
17. "The Future of Medical Education in Canada (FMEC): A Collective Vision for MD Education" (The Association of Faculties of Medicine of Canada, 2015). www.afmc.ca/wp-content/uploads/2022/10/2015-FMEC-MD_EN.pdf
18. J. M. Metzl, H. Hansen, Structural competency: Theorizing a new medical engagement with stigma and inequality. Soc. Sci. Med. 103, 126–133 (2014).
19. C. Hardy, M. E. Boulos, S. Bhargava, L. A. Cooper-Brown, M. Hackett, J. Hearn, E. Rowe, J. Shapiro, J. Speidel, A. Srajer, S. Suleman, Longitudinal advocacy training for medical students: A virtual workshop series. Can. Med. Educ. J. 13, 67-69 (2022).
20. Faculty of Health Sciences, "Service Learning in Undergraduate Medical Education" (McMaster University, 2025); ugme.healthsci.mcmaster.ca/education/service-learning/
21. Faculty of Medicine, "Flexible and Enhanced Learning (FLEX)" (University of British Colombia, 2025). mednet.med.ubc.ca/teaching/flex/course-overview/
22. L. Cohen, F.-H. Leung, C. Oriuwa, R. Wright R, Service-learning curriculum design and implementation at the University of Toronto Faculty of Medicine. MedEdPORTAL21, e141 (2019). 10.15694/mep.2019.000141.1.
23. O. W. Fung, A. Mulholland, M. Bondy, M. Driedger, C. E. Kendall, Implementing experiential learning logs addressing social accountability into undergraduate medical clerkship education. Can. Med. Educ. J. 14, 146-149 (2023).
24. L. Scherer. Student-led initiative promotes social medicine in Canada. Medscape, 2025. www.medscape.com/viewarticle/student-led-initiative-promotes-social-medicine-canada-2025a1000kgz
25. M. Sharma, A. D. Pinto, A. K. Kumagai,Teaching the social determinants of health: A path to equity or a road to nowhere? Acad. Med. 93, e25–e30 (2018). 10.1097/ACM.0000000000001689
26. F. E. de Bok, J. Hermans, R. J. Duvivier, D. Wolff, S. A. Reijneveld, Conceptualization and teaching health advocacy in undergraduate medical education: A document analysis. BMC Med. Educ. 24, e6039 (2024). 10.1186/s12909-024-06039-0
27. B. Tuohy, L. Olsen, H. Calvelli. How medical students learn about the social: Opportunities and limitations in service learning and volunteering. Soc. Sci. Med. 374, e118018 (2025). 10.1016/j.socscimed.2025.118018
28. K. A. Mangold, T. R. Bartell, A. A. Doobay-Persaud, M. D. Adler, K. M. Sheehan, Assessment and evaluation in social determinants of health education: A national survey of US medical schools and physician assistant programs. J. Gen. Intern. Med. 37, 2700–2708 (2022).
29. E. Krupat, J. L. Dienstag, Teaching the social determinants of health in undergraduate medical education: A scoping review. J. Gen. Intern. Med. 34, 720–730 (2019).
30. C. Gummesson, S. Alm, A. Cederborg, M. Ekstedt, J. Hellman, H. Hjelmqvist, M. Hultin, K. Jood, C. Leanderson, B. Lindahl, R. Möller, B. Rosengren, A. Själander, P. J. Svensson, S. Särnblad, A. Tejera, Entrustable professional activities (EPAs) for undergraduate medical education: Development and exploration of social validity. BMC Med. Educ. 23, e635 (2023). 10.1186/s12909-023-04621-6

We've Been Putting People to Sleep for 175 Years. And We Still Don't Fully Know How

A Commentary on the Mechanistic Gap in General Anesthesia

Cela fait 180 ans qu'on endort des patients. Et on ne comprend toujours pas complètement comment cela fonctionne.
Commentaire sur le manque d'explication mécanistique en anesthésie générale

Ayman Assaaoudi[1*]

1. University of Ottawa, Ottawa, ON, Canada
*Corresponding author. Email: aassa087@uottawa.ca

Abstract | Résumé

General anesthesia is one of modern medicine's most impressive successes. Every day, patients are made unconscious, kept still and pain-free during surgery, and then brought back to awareness with no memory of the operation itself. Clinically, this works remarkably well. The harder question is why it works. We can describe many of the drugs, receptors, and brain patterns involved, but the explanation is still incomplete when we try to connect a molecular drug effect to the lived disappearance of awareness.
This commentary follows that gap in three steps. It first looks at the early shift from the Meyer-Overton lipid theory to receptor-based explanations involving targets such as GABA-A and NMDA receptors. It then considers why receptor pharmacology alone does not fully explain the anesthetic state, especially because unconsciousness, immobility, amnesia, and analgesia are partly separable effects. Finally, it turns to the deeper problem: anesthesia removes consciousness, but neuroscience still does not have a complete theory of what consciousness is. The result is not a claim that anesthesia is unsafe or mysterious in every respect, but rather that an important mechanistic gap remains even after major clinical and scientific progress.

———

L'anesthésie générale est l'une des plus grandes réussites de la médecine moderne. Chaque jour, des patients sont rendus inconscients, maintenus immobiles et sans douleur pendant une chirurgie, puis ramenés à l'état d'éveil sans aucun souvenir de l'opération elle-même. Sur le plan clinique, cette pratique fonctionne remarquablement bien. La question plus difficile est de comprendre pourquoi elle fonctionne. Nous pouvons décrire plusieurs médicaments, récepteurs et modèles d'activité cérébrale impliqués, mais l'explication reste incomplète lorsqu'on essaie de relier l'effet moléculaire d'un anesthésique à la disparition vécue de la conscience.
Ce commentaire examine cet écart en trois étapes. Il présente d'abord le passage de l'ancienne théorie lipidique de Meyer-Overton vers des explications centrées sur les récepteurs, comme les récepteurs GABA-A et NMDA. Il explique ensuite pourquoi la pharmacologie des récepteurs, à elle seule, ne suffit pas à expliquer complètement l'état anesthésique, surtout parce que l'inconscience, l'immobilité, l'amnésie et l'analgésie sont des effets qui peuvent être en partie séparés. Enfin, il aborde un problème plus profond : l'anesthésie retire la conscience, mais les neurosciences ne possèdent toujours pas de théorie complète de ce qu'est la conscience. L'idée n'est donc pas de dire que l'anesthésie est dangereuse ou entièrement mystérieuse, mais plutôt de montrer qu'un écart mécanistique important demeure, malgré les grands progrès cliniques et scientifiques.

Keywords: General anesthesia, Meyer-Overton correlation,GABA-A receptors, Consciousness, Thalamocortical networks · Xenon anesthesia, Intraoperative awareness

Introduction

General anesthesia was first used publicly in surgery by the dentist William Morton in 1846, when he demonstrated ether anesthesia at Massachusetts General Hospital in front of a group of skeptical surgeons. He anesthetized a patient long enough for a neck tumor to be removed without pain. Within months, the use of general anesthesia spread internationally. Nearly 180 years later, anesthesia is one of the safest areas of hospital medicine. Every day, anesthesiologists provide anesthesia and pain control for hundreds of millions of procedures worldwide each year. The drugs are better, the monitoring is better, and the safety record is extraordinary. That clinical success is exactly what makes the remaining scientific uncertainty so striking.

Anesthetics work, but explaining how they work is harder than it first appears. It is fairly easy to describe the first step: an anesthetic molecule binds to, blocks, or modulates a target in the nervous system. The difficulty is the next part. Somehow, those molecular changes alter circuits, reshape large-scale brain

communication, and produce a patient who is unconscious, immobile, amnesic, and unable to experience pain. The explanation becomes less secure as we move upward from receptors to circuits to the whole brain. Some of the missing pieces are experimental. Others are conceptual, especially because we still do not have a settled account of what consciousness is. This paper follows that chain level by level, pointing out where the evidence is strong and where the explanation still thins out.

The Lipid Theory: A Great Idea That Wasn't Quite Right

The first systematic attempt to explain anesthesia came from a striking observation. In 1899 and 1901, two pharmacologists working independently, Hans Meyer in Germany and Charles Overton in England, noticed something remarkably consistent across a wide range of chemically different anesthetic compounds (1, 2). If you measured how readily each one dissolved into fat versus water, the ranking almost perfectly predicted their potency. More fat-soluble meant more powerful, nearly every time.

The obvious interpretation was that anesthetics worked by dissolving into the fatty membranes of neurons and disrupting them in a nonspecific physical way, much like a drop of oil changes the surface of water. It was an appealing idea because it gave one simple explanation for many different compounds, from ether to chloroform to alcohol. There were researchers who suspected early on that proteins might matter more than lipids, but the lipid theory had something those alternatives did not yet have: a clean, predictive relationship between fat solubility and anesthetic potency. For that reason, it dominated the field until later experiments made the protein-target explanation harder to ignore.

The cracks began to appear slowly. The most convincing evidence against a purely lipid-based explanation came from studying what are known as stereoisomers. These are molecules with identical chemical formulas and identical fat solubility, but with mirror-image three-dimensional shapes. If anesthesia were simply about dissolving into fatty membranes, mirror-image versions of the same molecule should work equally well. But they do not. Stereoisomers of the same compound can differ meaningfully in their anesthetic potency despite being chemically indistinguishable on paper (3).

This finding carries a clear implication. Lipid membranes are indifferent to molecular shape, but proteins are exquisitely sensitive to it. The fact that shape affected potency pointed strongly toward specific protein structures as the real targets, not the membranes themselves.

What Receptors Can and Can't Explain

When researchers started narrowing their focus to specific proteins, two kept surfacing in the literature. GABA-A receptors are a type of ion channel found in the neuronal membrane. Their role is inhibitory, meaning that when they open up, the neuron has a harder time firing. NMDA receptors work in the other direction, promoting excitation and contributing to memory encoding. Pull one up, the other down, and most anesthetics are essentially playing that game in one form or another.

Propofol, which is probably what most people get before surgery these days, works by pushing heavily on the GABA-A side. It does this by boosting inhibitory signals throughout the brain, gradually suppressing neural activity until the patient is under. Clinicians like it because it works fast and wears off cleanly, which is not something every anesthetic can claim.

Ketamine is a stranger case. It does not touch GABA-A much at all. Instead it blocks NMDA receptors, and what follows looks less like sleep and more like a dissociative state, where normal perception just kind of comes apart. That made it useful in settings where you needed something fast and did not have a full anesthesia setup handy, emergency medicine, field surgery, pediatric procedures. What nobody expected was that psychiatrists would eventually come knocking. At sub-anesthetic doses, ketamine turns out to act on depression surprisingly quickly, sometimes within hours, in patients who had failed multiple other treatments. The NMDA receptor, it seems, is involved in more than memory.

Volatile agents such as isoflurane and sevoflurane cast a wider net, affecting GABA-A, NMDA, and several other targets (4, 5). This receptor-level work has been genuinely important, but it does not solve the whole problem. Part of the gap is biological: we still do not fully know which mechanisms are responsible for which parts of the anesthetic state. Part of it is conceptual: if the endpoint is loss of consciousness, then we also need to know what consciousness is before we can fully explain what has been removed.

What we call "general anesthesia" is not one thing. It is actually four distinct phenomena happening together: unconsciousness, immobility, amnesia, and pain relief. And these four components appear to involve different mechanisms in different parts of the nervous system. The immobility component is primarily a spinal cord effect. The unconsciousness component is cortical. Amnesia involves hippocampal circuits specifically. These are anatomically and mechanistically separate processes.

Why does this matter? Because a drug that only targets spinal mechanisms could produce complete immobility in a fully conscious patient. That patient would be awake, aware, and experiencing everything but not able to move or speak. It is what intraoperative awareness is, and it happens to approximately one or two patients per thousand procedures (11). The fact that we cannot reliably prevent it, despite decades of effort, reflects how incompletely we understand which mechanisms are actually responsible for which components of the anesthetic state.

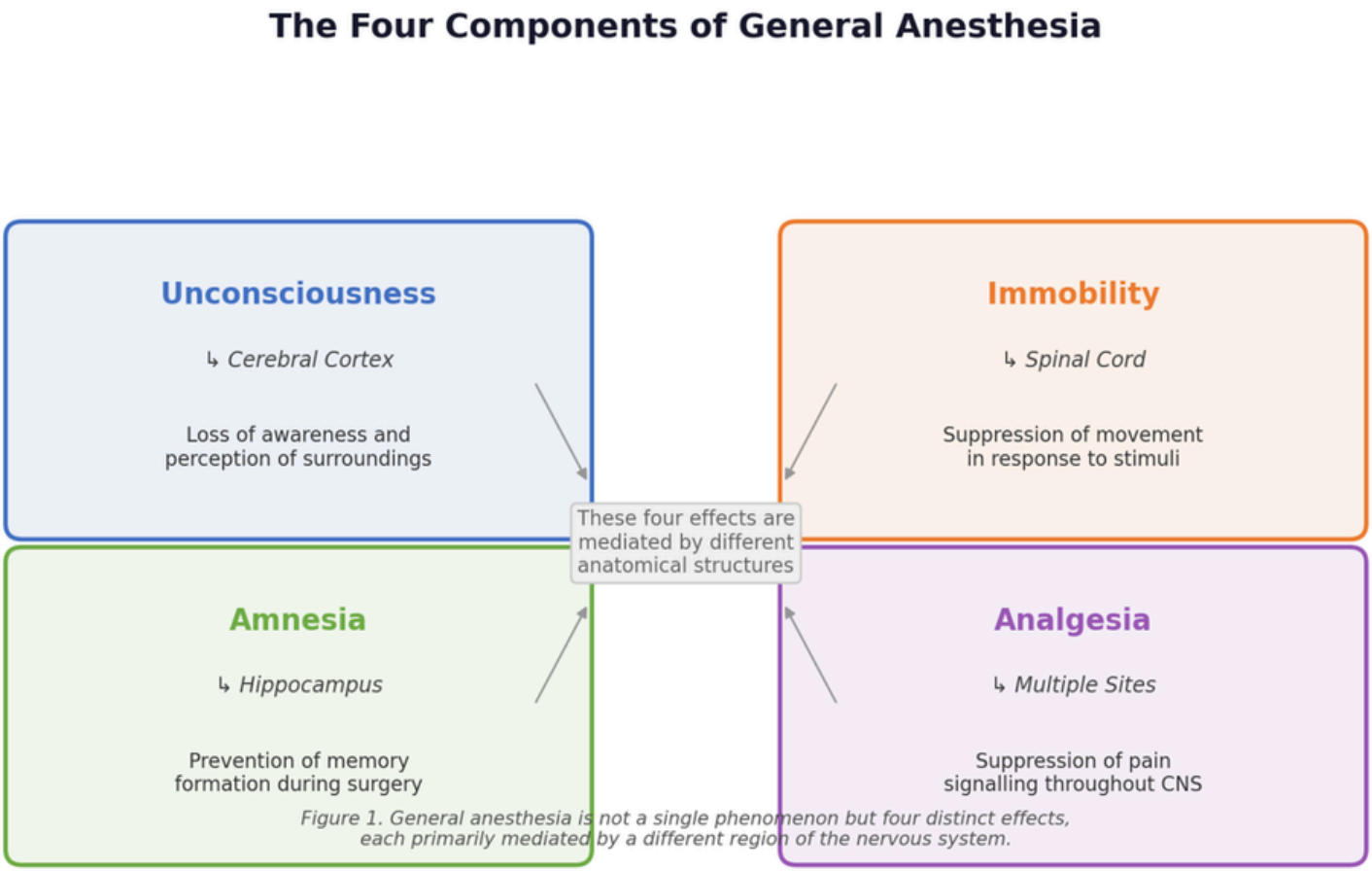


Figure 1. The four components of general anesthesia.

Xenon: The Part That Should Bother Us More

Xenon makes this problem especially clear, although it should not be treated as magic. It is a noble gas, in the same family as helium and neon, and it does not react with biological molecules in the usual chemical sense. It has no functional groups, no charge, and no obvious chemical handle that would make it look like a conventional drug. It is also not metabolised; when it is removed, the gas leaves the body essentially unchanged. Yet at a sufficient concentration, xenon can produce general anesthesia (6).

The best current explanation is still physical rather than chemical in the usual sense. Xenon can occupy hydrophobic pockets in certain receptor proteins and, by sitting there, change how those proteins move or function (7). That explanation is plausible and supported by structural and pharmacological evidence. Still, xenon is useful because it forces the question into sharper focus.

If a chemically inert atom can help produce unconsciousness, then anesthesia cannot be explained only by the familiar picture of drugs forming specific chemical interactions with targets. The mechanism has to be broad enough to include both complex pharmaceutical molecules and a simple noble gas. That does not mean receptor biology is wrong. It means the deeper explanation has to connect physical binding, protein dynamics, neural networks, and consciousness in a way we still have not fully achieved.

The Consciousness Problem Underneath All of This

This is where anesthesia becomes more than a standard pharmacology problem. We have many drugs and gases that reliably produce unconsciousness, but explaining that effect requires knowing what has actually been lost. That is still difficult. We can measure brain activity as consciousness fades, and this has been extremely useful. For example, Emery Brown and colleagues have shown that different anesthetics produce different EEG patterns. Propofol is often associated with slow oscillations around 1 Hz and frontal alpha activity, whereas ketamine is associated with higher-frequency gamma activity that is more posterior (8). These patterns help clinicians and researchers track anesthetic depth, but they are still correlates of unconsciousness, not a complete explanation of consciousness itself.

At a broader anatomical level, one influential account focuses on the thalamus, a deep brain structure that helps relay information between sensory systems and the cortex. The thalamocortical hypothesis proposes that consciousness depends on ongoing communication between the thalamus and the cortex, and that anesthetics disrupt this loop (9). That idea is important, but it is probably not the whole story. Cortical connectivity, the default mode network, and brainstem arousal systems also appear to matter, and different anesthetics may reach unconsciousness through somewhat different network routes.

The same disruptions of thalamocortical interactions can occur during some stages of normal sleep or during seizures, and are not typically followed by a complete loss of consciousness. Thus, thalamocortical disruption may be a necessary component of the underlying causes of coma, but it is not sufficient. And competing theories of consciousness — Integrated Information Theory, Global Workspace Theory, others — make somewhat different predictions about what should happen under anesthesia that the current evidence has not cleanly resolved (10).

Why Does This Matter?

This is not just a theoretical issue. It matters for patient safety, informed consent, and the way anesthetic depth is monitored. Intraoperative awareness is the clearest example. The main monitoring tool used to reduce this risk is the Bispectral Index, or BIS, which turns EEG activity into a score meant to estimate anesthetic depth. BIS can be useful, but it is not perfect. It was developed from observed EEG patterns rather than from a complete biological theory of what anesthetic depth really is. As a result, two patients can have similar BIS values while being in different underlying brain states, especially under different anesthetic drugs.

There is also a practical dosing problem. People vary in how much anesthetic they require, and this variability can be influenced by age, genetics, prior drug exposure, medical history, and baseline neurological function. Some patients need more than the average dose, while others need less. Because this variability cannot yet be predicted from first principles, dosing still relies heavily on population averages, clinical signs, and monitoring tools. That creates room for error in both directions: too little anesthesia risks awareness, while too much can increase physiological stress and delay recovery.

Future work should therefore treat anesthesia as a network-level brain state, not only as a list of receptor effects. High-density EEG and fMRI are now much better suited to studying how anesthetics change communication among brain regions. This kind of work could also help with the long-term goal of separating the four components of anesthesia more precisely: unconsciousness, immobility, amnesia, and analgesia. A drug or monitoring strategy that targets those components separately would require a much clearer map of how each one is produced (12).

Conclusion

Anesthesia is safe and effective, and that achievement is remarkable. Still, nearly 180 years after Morton's ether demonstration, a complete mechanistic account remains out of reach. The field has learned a great deal about receptor targets,

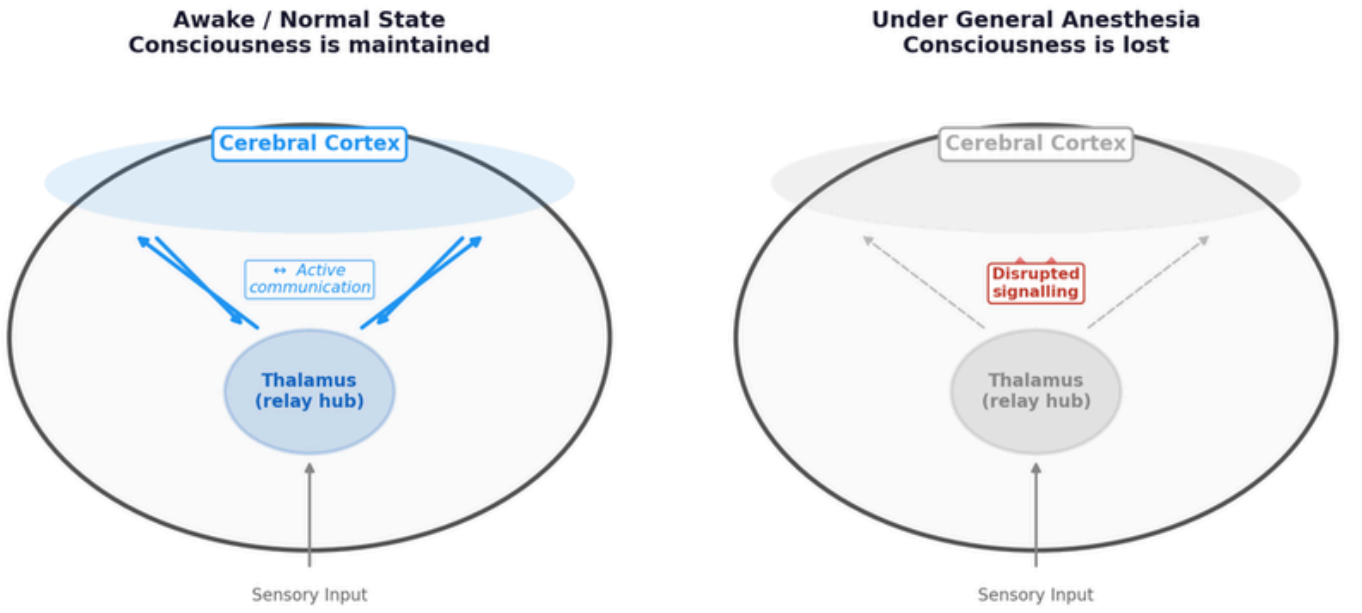


Figure 2. The thalamocortical hypothesis of anesthesia. consciousness depends on constant two-way communication between the thalamus and the cortex (left). Anesthetic drugs disrupt this communication (right), interrupting the neural activity thought to underlie awareness.

EEG signatures, and large-scale brain networks. The difficulty is that these findings do not yet join into a full explanation of how an anesthetic removes consciousness. That question sits between neuroscience, pharmacology, and philosophy, which is partly why it has been so hard to close.

There is a real difference between using anesthesia safely through empirical dosing and monitoring, and understanding it well enough to predict each patient's response from first principles. That gap has practical consequences, including intraoperative awareness, dosing variability, and the future design of anesthetics that could separate the different components of the anesthetic state. The science of anesthesia has come an extraordinary distance since 1846. It simply has not come all the way.

References

1. Meyer, H. H. (1899). Zur Theorie der Alkoholnarkose. Archiv für experimentelle Pathologie und Pharmakologie, 42, 109–118. (No DOI — predates digital era)
2. Overton, C. E. (1901). Studien über die Narkose. Gustav Fischer. (No DOI — book)
3. Franks, N. P., & Lieb, W. R. (1991). Stereospecific effects of inhalational general anesthetic optical isomers on nerve ion channels. Science, 254(5030), 427–430. https://doi.org/10.1126/science.1925602
4. Franks, N. P., & Lieb, W. R. (1994). Molecular and cellular mechanisms of general anaesthesia. Nature, 367(6464), 607–614. https://doi.org/10.1038/367607a0
5. Franks, N. P. (2008). General anaesthesia: From molecular targets to neuronal pathways of sleep and arousal. Nature Reviews Neuroscience, 9(5), 370–386. https://doi.org/10.1038/nrn2372
6. Sanders, R. D., & Maze, M. (2005). Xenon: From stranger to guardian. Current Opinion in Anaesthesiology, 18(4), 405–411. https://doi.org/10.1097/01.aco.0000174957.97759.f6
7. Dickinson, R., Peterson, B. K., Banks, P., Simillis, C., Martin, J. C. A., Valenzuela, C. A., Maze, M., & Franks, N. P. (2007). Competitive inhibition at the glycine site of the N-methyl-D-aspartate receptor by the anesthetics xenon and isoflurane. Anesthesiology, 107(5), 756–767. https://doi.org/10.1097/01.anes.0000287061.77674.71
8. Purdon, P. L., Pierce, E. T., Mukamel, E. A., Prerau, M. J., Walsh, J. L., Wong, K. F. K., Salazar-Gomez, A. F., Harrell, P. G., Sampson, A. L., Cimenser, A., Ching, S., Kopell, N. J., Tavares-Stoeckel, C., Habeeb, K., Merhar, R., & Brown, E. N. (2013). Electroencephalogram signatures of loss and recovery of consciousness from propofol. Proceedings of the National Academy of Sciences, 110(12), E1142–E1151. https://doi.org/10.1073/pnas.1221180110
9. Alkire, M. T., Hudetz, A. G., & Tononi, G. (2008). Consciousness and anesthesia. Science, 322(5903), 876–880. https://doi.org/10.1126/science.1149213
10. Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. Biological Bulletin, 215(3), 216–242. https://doi.org/10.2307/25470707
11. Mashour, G. A., Orser, B. A., & Avidan, M. S. (2011). Intraoperative awareness: From neurobiology to clinical practice. Anesthesiology, 114(5), 1218–1233. https://doi.org/10.1097/ALN.0b013e31820fc9b6
12. Hemmings, H. C., Jr., Riegelhaupt, P. M., Kelz, M. B., Solt, K., Eckenhoff, R. G., Orser, B. A., & Goldstein, P. A. (2019). Towards a comprehensive understanding of anesthetic mechanisms of action: A decade of discovery. Trends in Pharmacological Sciences, 40(7), 464–481. https://doi.org/10.1016/j.tips.2019.05.001