

Investigating the Emergence of Chemical Complexity for Life Detection and Patent
Evolution; and Developing Earth Science Curricula for Prison Education

by

John F. Malloy

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2023 by the
Graduate Supervisor Committee:

Sara I. Walker, Chair
Darryl Reano
Elizabeth Trembath-Reichert
Hilairy Hartnett
Leroy Cronin

ARIZONA STATE UNIVERSITY

August 2023

ABSTRACT

The origin of life remains unknowable to current science. Scientists cannot see into the origin of life on Earth, and until humanity discovers life elsewhere in the universe and begin to compare this alien life to Earth, it is likely to be undiscoverable. However, alien life may be so different from life as it is currently known that it may not be recognizable when it is found. Therefore, astrobiology needs a universal theory for life to avoid detection methods being biased towards Earth-based life. This also extends to the instrumentation sent into space, which should be built to detect universal properties of life. Assembly theory, a novel measure of complexity and arguably the only testable agnostic biosignature in current science, is used here to provide precision requirements for mass spectrometry instrumentation on future spaceflight missions with the goal of finding life elsewhere.

Universal properties are not only applicable to the origins of life, but also to technologically advanced societies. Predictable patterns are found in today's industrially based society, such as energy usage as a function of population density. These patterns may serve as the basis for technosignatures that are evidence of advanced extraterrestrial civilizations. Patterns found in patent chemistry are explored, as well as predictions of chemical complexity based on assembly theory, to determine how complex chemistry is built by human society and which statistical patterns may be found in extraterrestrial civilizations.

Moving beyond astrobiology, science cannot be done in a vacuum but must be communicated and taught to others. Topics such as a universal definition of life,

biosignatures, and increasing complexity mean nothing without interest and engagement from others, particularly students. To this end, transformative pedagogical tools are used, particularly sociotransformative constructivism (sTc), to build and teach an Earth Science and Astrobiology curriculum to a classroom of high school incarcerated students. The impact of this class on their science learning and how they personally identify as scientists is studied.

ACKNOWLEDGMENTS

Every PhD is unique, but I'm sure very few had to overcome COVID and the challenges of bouncing between three different cities in two countries. I couldn't have asked for a more supportive, engaging group of friends and scientists who welcomed me to Phoenix, Glasgow and Santa Fe with open arms and supported me throughout these five years.

Thank you first and foremost to those who have truly made these a special five years.

Caitlyn Hall, for introducing me to Arizona, exploring everywhere with me and consistently being my voice of reason, comfort, and support when I needed it most. Mom and Dad, for always supporting me and making time for Saturday afternoon calls as I traipse around the world. Veronica Mierzejewski, for always being up for shenanigans, especially the spontaneous ones. Tyler McCabe, John Bello, Andrew Smith, and Scott Bohmke for running all over Arizona with me, driving me to Colorado at 2am, and showing me that physical limitations are ultimately meant to be broken. Silke Asche, Alex Telfar, and Amit Kahana for introducing me to Glasgow and giving me friendly faces to look forward to every time I came back. Mercedes Vilaso, for reminding me that home is always in Baltimore.

Thank you to all the communities I've been lucky enough to be a part of – the ELIFE research group, particularly Dylan Gagler, Pilar Vergeli, Cole Mathis, Estelle Janin, Daniel & Mara Czegel, Gage Sierbert, Hikaru Furukawa, Hannah Dromiack, Jake Hanson, Harrison Smith, Swanand Khanapurkar, Sonakshi Sharma, Tessa Fisher, and

everyone else who has been a part of the lab over the last five years; the entire La Jolla house, especially Chanel Vidal; the graduate student community at SESE, especially Jisoo Kim, Brendan Chapman, Grace Carlson, Mara Karageozian, Logan Jenson, Tyler Richey-Yowell and Mariah Heck; the 2019 & 2022 Complex Systems Summer Schools at the Santa Fe Institute, particularly Lou de la Felice; the Glasgow Climbing 2.0 group and the 3pm coffee/tea club; everyone in November Project, particularly Jackie Knoll, Rick Jessop, Christine Meyer, Kendra Flory, and Annie Dube (with special credit to Annie and Amanda Ohmer for driving me to the hospital after my very first Wednesday workout at Papago Park); Tempe Tri Club Monday swimming; Glasgow Achilles Heel Running Club, and the entire Aravaipa Running ultrarunning community.

Finally, thank you particularly to my committee members for giving me the opportunity to study at ASU and work through this highly weird PhD. Sara Walker, for the opportunities to work on fascinating topics and consistently giving me too many ideas to play with. Lee Cronin, for supporting me (twice) to come live in Glasgow and pushing me to be the best scientist I can be. Darryl Reano, for patiently walking me through an entirely new-to-me field of research in under a year and demonstrating how to be a great teacher. Elizabeth Trembath-Reichert and Hilairy Hartnett, for supporting me and providing advice along this entire journey.

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	ix
LIST OF FIGURES	x
CHAPTER	
1: INTRODUCTION	1
Origin of Life Theories	1
Assembly Theory	6
Assembly Theory and Complex Systems Science	8
Network Science	9
Chemical Scaling Patterns	10
Science Communication	12
Prison Education	14
2: EXPLORING THE MOLECULAR LIMITS OF LIFE	17
Introduction.....	17
Results.....	24
Building Chemical Space.....	24
Molecular Assembly Over Chemical Space	30
Mission Specification.....	33

CHAPTER	Page
Distinguishing Existing Biochemistry	35
Chemical Space of Life.....	38
Methods.....	42
Formula Generation	42
Fitting Formula Count.....	43
Molecular Formula Size Fitting & Distributions	44
Mean/Standard Deviation Fitting.....	45
Random Molecule Generation	46
Assembly Index Calculations	46
Number of Possible Molecules	46
Likelihood	47
m/ Δ M Calculation.....	48
KEGG Overlap.....	48
Assembly Theory Over Different Databases	48
Discussion	49
3: SOCIAL DYNAMICS SHAPE CHEMICAL INNOVATION.....	51
Introduction.....	51
Results & Methods.....	57
Network Statistics.....	57
Preferential Attachment	64

CHAPTER	Page
Tracking Preferential Attachment Across Compound Classes	72
Molecular Assembly	77
Social Factors and Molecular Assembly	85
Discussion	97
Preferential Attachment	97
Assembly	99
4: EARTH SCIENCE CURRICULA FOR INSTITUTIONALIZED YOUTH.....	102
Introduction.....	102
Theoretical Framework	106
Methods.....	109
Research Design.....	109
Classroom Setting	110
Prison Environment	111
Survey Instrument.....	112
Arizona State Standards.....	114
Place-Based Curriculum	116
Units.....	116
Arizona State University Volunteer Coordination.....	118
Final Project	119

CHAPTER	Page
Results.....	124
Survey	124
Thematic Analysis.....	128
Limitations	138
Discussion	139
5: CONCLUDING DISCUSSION	141
REFERENCES	147
APPENDIX	
A INSTITUTIONAL REVIEW BOARD APPROVAL.....	168

LIST OF TABLES

Table	Page
Table 1: KS Goodness of Fit tests of Gaussian Modelled Distributions for Upper Elemental Limits 6-15.....	27
Table 2: Average Preferential Attachment Values	70
Table 3: Compound Degree Power Law Fit Statistics.....	71
Table 4: Linear Regression Statistics For University of Arizona Patents.	92
Table 5: Survey Instrument, Based on the CLASS Survey, Given Out to Students in the Earth Science Class Developed as Part of This Project.	114
Table 6: High School Sub-Standards Within the E1 Core Idea of the Arizona State Standards. Adapted from (Arizona Science Standards, 2018, p. 80).....	115
Table 7: List of Units and Weeks When Each Unit Was Taught: Geology (Blue), Interacting Systems (Green), Climate Change (Orange), and the Final Project (Purple).	119
Table 8: Planetary Constraints Available for Students to Build Their Planet.	121
Table 9: Planning Worksheet Where Students Predict Future Outcomes of Their Planet’s Geosphere, Hydrosphere, Atmosphere, Biosphere, and Climate. Based on Their Individual Choices in Table 8.	122
Table 10: Qualitative Analysis Codes.....	129

LIST OF FIGURES

Figure	Page
Figure 1: Chapter 2 Concept Figure.....	23
Figure 2: The Number Of Chnops Chemical Formulas With Modelled Gaussian Distributions.....	26
Figure 3: Assembly Values Of 10000 Chemical Structures And (Inset) Likelihood Of High-Assembly Compounds As A Function Of Molecular Weight.	26
Figure 4: Linear Regression Predicting μ (Mean) And Σ (Standard Deviation) Values For Gaussian Distributions Of Number Of Formulae. I Performed A Linear Regression On An Upper Elemental Limit Of 6-15 Heavy Atoms, And This Regression Was Validated By Testing At An Upper Limit Of 20 (Black Dots – See Figure 5).	28
Figure 5: The Predicted μ (Mean) And Σ (Standard Deviation) (Figure 4) Accurately Predict The Computationally Generated Distribution Of Molecular Weights Where The Maximal Number Of Elements Is 20.	29
Figure 6: The Number Of Formulas At A Given Molecular Weight (Left) And The Cumulative Number Of Formulas As A Function Of Increasing Molecular Weight.	30
Figure 7: Precisions At Which High-Assembly Compounds Can Be Detected.....	32
Figure 8: The Minimum $M/\Delta M$ Resolution Necessary To Distinguish Exactly 10 High-Ma Compounds At Each Molecular Weight.....	35
Figure 9: The Number Of Overlapping Biochemical Compounds Within Kegg As A Function Of Different Mass Resolutions.	38

Figure	Page
Figure 10: Ma Distributions Over Chemical Databases Representing Differently Curated And Biased Chemical Space Distributions.	41
Figure 11: The Number Of Formulas Generated For 0-1500 Daltons, As A Function Of The Maximum Number Of Chnops Elements (6-10 Shown Here).	43
Figure 12: The Number Of Formulas Per Maximum Element Count Modelled Using An Exponential Curve.....	44
Figure 13: Assembling Chemical Networks From The Surechembl Database Over Time, Calculating Preferential Attachment, And Calculating Assembly Indices.	56
Figure 14: Network Statistics Over Surechembl Patent Chemistry.	60
Figure 15: Largest Connected Component Size (Patents + Compounds).....	61
Figure 16: Average Clustering Coefficient.....	62
Figure 17: Compound Exponential Growth Fit. The Total Number Of Compounds (In Units Of 10^7 Individual Compounds) Are Listed On The X-Axis.	63
Figure 18: Patent Exponential Growth Fit. The Total Number Of Patents (In Units Of 10^6 Individual Patents) Are Listed On The X-Axis.	63
Figure 19: Preferential Attachment Indices Across All Surechembl Data In 5-Year Increments.....	64
Figure 20: Full Preferential Attachment, 1980-2019.....	65
Figure 21: Compound Degree Distribution Power Law Fits	71
Figure 22: Psychedelic Drug Attachment Over Time.....	74
Figure 23: Sars/Hcv Drug Attachments Over Time.	75
Figure 24: Green Solvent Attachment.	76

Figure	Page
Figure 25: Changing Chemical Properties (Assembly Indices, Molecular Weights And Fragment Diversity) Of Compounds From 1980 - 2020.	81
Figure 26: Ma Over Time With Linear Fit	82
Figure 27: Molecular Weight (Daltons) Over Time With Linear Fit	82
Figure 28: Ma / Molecular Weight Correlation	83
Figure 29: Ma / Number Of Bonds Correlation	84
Figure 30: Effect Of Various Factors (Cost, Network Degree, Individual Authors, Assignees, And Patent Classifications) On Ma Increase Over Time.	85
Figure 31: Cost (Gbp Per Gram) / Ma, All Compounds	86
Figure 32: Cost (Gbp Per Gram) / Ma, Exact Ma Only	87
Figure 33: Cost Over Time (Colored By Ma Value)	88
Figure 34: Average Ma Of All 494 Patents Associated With The University Of Arizona Over Time.	91
Figure 35: Average Ma Of Patents With Linear Regression And Delta Ma Shown.	93
Figure 36: Author, Assignee, And Classification Regression Results.	95
Figure 37: Author Dropout Tests Results	96
Figure 38: Assignee Dropout Tests Results	96
Figure 39: Classification Dropout Tests Results	97
Figure 40: Mixed-Methods Study Overview.	110
Figure 41: Example Final Project Planet - This Desert Planet Includes Liquid Water And A Single Mountain Range Lining The Equator.	124

Figure	Page
Figure 42: Overall Change In Earth Science Identity Of The Five Students Who Completed Both Pre- And Post-Class Class Surveys.	126
Figure 43: Student Answers And Changes From The “Problem Solving - Confidence” Questions.....	127
Figure 44: Student Answers And Changes From The “Problem Solving - General” Questions.....	127
Figure 45: Student Answers And Changes From The “Real-World Connections” Questions.....	128

1: INTRODUCTION

Origin of Life Theories

Charles Darwin's theory of evolution, which postulated that living things evolved through selection processes over time, has not generally been challenged over the past 160 years of biology (Darwin, 1872; Minoru Kanehisa, 2019; Pagel, 1999). Technological advances in recent years have further enhanced our understanding of evolution through discoveries such as Rosalind Franklin's discovery of the structure of DNA (Sayre, 2000) and searchable sequence libraries (Chen et al., 2019; Federhen, 2012). Nevertheless, the underlying question of how evolution (and life) originated on Earth remains unanswered.

To address the origin of life, scientists often utilize at least one of four possible fields – geology & planetary science, the history of metabolism, evolutionary biology (particularly through phylogenetics), and biophysics. Here, I explain each, and ultimately propose a measure that moves past each.

Some scientists address the origin of life through geology, where they specifically study how biochemistry emerges from geological processes. The Miller-Urey experiment in 1952 is a classic example of this approach, where the mixing of water, methane, ammonia, and hydrogen with an electric current within a closed system led to the synthesis of at least four amino acids used in biochemistry – i.e., glycine, alanine, aspartic acid, and aminobutyric acid (Miller & Urey, 1959). Recent analyses of the experiment have found that over 23 amino acids and other biochemically meaningful compounds were synthesized, but not detected due to the limitations of 1950s analytical capabilities (Parker

et al., 2011). This experiment and subsequent analyses link simple, abiotic chemistry that likely existed on the primordial Earth to more complex biochemical compounds that form the basis of life as we know it today (McCollom, 2013).

The Miller-Urey experiment was done in a laboratory-based closed system, as have many recent iterations as well (Cooper et al., 2017; Parker et al., 2014). However, the environmental conditions on the prebiotic Earth are unlikely to be as pristine as those found in these controlled experiments (McCollom, 2013). Therefore, the geologic approach to the origin of life is also performed through observation and study of environments on Earth. The environments which are studied are similar to what existed roughly four billion years ago when the origin of life as we know it likely occurred (Bromberg et al., 2022). Hydrothermal systems such as those in Yellowstone (Shock, 1990) and on the seabed (Martin et al., 2008) are particularly interesting because there is a consistent supply of geothermal energy and potential for prebiotic chemical reactions which link different chemical species as electron donors (Shock et al., 2010). If these or similar environments are found elsewhere in the universe, there could be a high likelihood of finding extraterrestrial life there.

Beyond Earth-based environments, scientists also investigate the possibility of building blocks of life originating beyond our planet (panspermia). Meteorites, such as the Murchison meteorite that fell in Australia in 1969, provide valuable insights into the potential for life to arise through abiotic processes in space. The Murchison meteorite

contained eight of the 20 biochemical amino acids used in protein formation and an additional 44 amino acids not used in protein formation (Koga & Naraoka, 2017; Kvenvolden et al., 1970; Martins et al., 2008). This discovery demonstrates amino acid formation outside of Earth and suggests the widespread availability of the compounds necessary for Earth-like life.

Metabolic chemistry provides another approach to studying the origin of life, where scientists work backwards from existing chemistry and biology to predict the environmental conditions and chemical reactions present at the time when the last universal common ancestor (LUCA) lived. All life is united through central metabolic pathways, such as the tricarboxylic acid cycle (TCA), that is at least partially found in all living organisms. Using a network of existing biochemistry, Goldford et al. demonstrated that basic biological building blocks such as lipids can be generated from a variant of the TCA cycle under plausible prebiotic environmental conditions (Goldford et al., 2019). Furthermore, the central carbon metabolism, which includes the TCA cycle, has been shown to be optimally efficient at converting small metabolites into biomass (Noor et al., 2010). The prevalence of the TCA cycle, as well as the high degree of optimality, suggest environmental conditions that favor these carbon reactions were present and possibly even common 3.8 billion years ago when life first emerged on Earth. Finding a similar set of optimized chemical reactions elsewhere would suggest the presence of life.

Yet another approach to studying the origin of life is to use present-day genomic data to extrapolate possible historic environmental conditions. This top-down approach is complementary to geology or planetary science approaches, which approach the origin of life in a bottom-up manner. Gene sequencing techniques have improved scientific understanding of the tree of life that represents phylogenetic relationships between living organisms starting at LUCA (Hug et al., 2016). Weiss et al. used a consensus phylogenetic tree of life to model the genes that were likely to exist near the origin of life, which is a novel approach compared to prior studies (Weiss et al., 2018) that focused on the universality of different genes as a proxy for existence at the origin of life (Kyrpides et al., 1999). However, radiation of genes does not correspond directly with historical environments, particularly given the dramatic environmental changes over Earth's history that necessitated and favored novel chemical reactions (Anbar et al., 2007; Raup, 1986). Genetic data instead allows for extrapolation into the enzymes and reactions present within genomes (Chen et al., 2019). This suggests that LUCA had many reactions necessary to survive in a hydrothermal vent environment and had a similar biochemistry to many of today's prokaryotes which are found in the Earth's crust (Weiss et al., 2018). While this hypothesis is untestable, it shows that the origin of life can be approached through existing biology, rather than through predicting or experimenting on abiotic geologic processes.

The final approach to studying the origin of life is to predict the phenotype (expressed chemistry) of the first organisms from the physics of biology and chemistry. This approach explores the known chemical bases of present-day life, such as the set of 20 amino acids

used nearly exclusively in proteins (Lopez & Mohiuddin, 2020) or lipid membranes (Kahana et al., 2021), to infer what features were present at the origin of life. Autocatalytic sets that provide physically possible mechanisms where biochemical molecules can self-replicate and potentially begin to evolve (Hordijk et al., 2010; Xavier et al., 2020) are favored in this approach. The existence of lipid membranes that spontaneously emerge due to hydrophobic and hydrophilic interactions in a water-based system, is considered potentially essential due to the need for compartmentalization of chemical species and reactions, allowing the evolution of autocatalytic reactions (Deamer, 2017; Lancet et al., 2018). Moreover, the emergence of specific types of chemistry can provide clues to how life is structured, such as the emergence of homochiral chemical species that are predicted by physical models of chemistry (Blackmond, 2010; Gleiser et al., 2012; Gleiser & Walker, 2012). These insights suggest the properties of the first organisms and their chemistry can be predicted by their physical properties.

These four approaches - geologic and planetary science, metabolic history, phylogenetics, and biophysics - provide complementary insights into the origin of life. Two approaches focus on chemical reactions, one on biological processes and evolution, and the third on the physical basis of biology and chemistry. Together, they illuminate the complex interplay between environmental conditions, chemistry, and biology that led to the emergence of life on Earth. However, these and other physics-based hypotheses suffer from the same deep-seated problem found across the field of astrobiology: the lack of practical results and generalizability. None of these models has yet created an origin of life event,

rendering them hypotheses rather than proven theories (Walker, 2017). To definitively answer how life originated on Earth and begin to understand how evolution might arise in radically different chemical systems elsewhere in the universe, universal solutions must be generated, tested, and applied. Essentially, the fundamental question that remains unanswered by research on the origin of life is: What are the universal properties of life?

Assembly Theory

To begin answering the question of what the universal properties of life are, we must look at chemistry, as chemical species and reactions are easily identifiable elsewhere in the universe when compared to other possibly universal phenomena, such as gravity waves or dark matter. Chemical reactions and the processes they create are fundamental to the universe, regardless of the environmental or thermodynamic conditions where potential life could be found. Finding the chemical processes that result in universal properties of life is challenging. Many chemical reactions and compounds could potentially result in life (Bains, 2004), and observational limitations make it extremely difficult to detect these processes elsewhere (Seager & Bains, 2015). Assembly Theory (AT) offers a promising solution to this problem, as it abstracts chemical processes so that universal properties of life can still be based in chemistry, regardless of the elemental makeup or environmental conditions of alien chemistry. This dissertation prominently features AT, as proposed and formulated by members of my dissertation committee (Cronin & Walker, 2016; Marshall et al., 2021, 2017).

Assembly Theory is based on the idea that living systems create complex structures that are not possible through abiotic processes. This theory proposes that living systems, regardless of their chemical composition or environmental conditions, produce a level of chemical complexity that is unique to life (Cronin & Walker, 2016). While AT can be applied to any organizational system, the focus of this dissertation is on small molecules (e.g., not proteins nor macromolecules). AT works by abstracting a chemical compound into a graph, with elements as nodes and bonds between the elements representing elements. The graph is then broken down into its smallest decomposable components (usually two elements bonded together) that are used to recursively reconstruct the original graph. This reconstruction assumes unlimited energy and unlimited intermediate parts. The minimum number of steps required to build a graph is the assembly index, or the molecular assembly value (MA) when applied to chemical compounds. Previous research has shown that compounds with an MA of 15 or higher are only produced as a result of biological processes, suggesting a complexity threshold of life on Earth (Marshall et al., 2021). Although this number may be different when applied to alternative chemistries or different systems, the fact that AT can measure complexity and that complex compounds ($MA \geq 15$) are only observed through living processes suggests that AT can be used as a biosignature for life elsewhere (Marshall et al., 2021).

AT has a physical manifestation that can be useful for space exploration in addition to its clear delineation of biochemical processes. According to previous studies, there is a strong positive correlation between the MA of a compound and the output from an ion trap mass

spectrometer – specifically, the MA matches the number of second-fragmentation peaks in an Orbitrap Fusion Lumos Tribrid spectrometer (Marshall et al., 2021). This is significant because mass spectrometers have been used in space since the Apollo missions of the early 1970s (Arevalo et al., 2020). Mass spectrometry outputs can be employed to construct an accurate structural graph of a detected molecule, which can then be used to calculate its MA, even though ion trap machines have not yet been sent to space (Arevalo et al., 2018; Willhite et al., 2021), making it challenging to directly correlate existing extraterrestrial spectra with MA. Chapter 2 of this dissertation investigates the ability of mass spectrometers on spaceflight missions to differentiate molecules to enable the direct application of AT to the search for life elsewhere, considering the broad spectrum of chemistry that could exist elsewhere in the universe.

Assembly Theory and Complex Systems Science

AT offers a method for measuring complexity that can be applied to detect life elsewhere in the universe. The results of biochemical evolution and selection on a chemical level lead to a complexification process (Marshall et al., 2021; Peng et al., 2020; E. Smith & Morowitz, 2016; Williams, 1997) as well as to chemical evolution across ecosystems (Sternier & Elser, 2017). Beyond astrobiology, AT can also be applied to measuring other types of complexity, such as how societies adapt and evolve to scientific discoveries and inventions. This approach involves looking at higher-order structures, such as reactions between compounds and relationships between individuals, rather than focusing on specific elements or individuals in a system. By doing so, seemingly disparate systems can be directly compared. Complex systems science studies the interactions between individual

agents in a system and the resulting dynamics that emerge (Siegenfeld & Bar-Yam, 2020). There are various methods used to measure dynamics in complex systems science, such as large-scale language models and non-linear dynamic studies of physical systems (Bradley & Kantz, 2015; Nguyen et al., 2020). One of the most common approaches for extrapolating data in complex systems science is network science.

Network Science

The field of network science is concerned with the interactions between individuals, which are analyzed at the system level (Barabási, 2013). This methodology is applicable to a wide range of systems, including biochemical metabolism (Barkai & Leibler, 1997; Goldford et al., 2019; H. Kim et al., 2019) social networks (e.g., Twitter, (Ke et al., 2017)), and energy grids (West, 2018). Regardless of the system under study, *nodes* within a network represent individual actors (e.g., people or metabolites) and *edges* represent the relationships between them (e.g., friendship or chemical reactions). Nodes and edges form the underlying structure of networks, and this structure can reveal common features across networks. For example, a small-world network structure, where very few nodes have a high degree (i.e., more connections) as compared to low-degree nodes, is a common feature of many networks, including social networks, metabolic networks, and human-engineered networks such as airport connections (Barabási, 2013; Broido & Clauset, 2019; Kunegis et al., 2013). Specifically, a small-world structure is one where the degree distribution of all nodes can be modeled by a power-law distribution, a decreasing, heavy-tailed exponential function. The preferential attachment model, where new nodes are exponentially more likely to attach to high-degree nodes than to low-degree nodes, is often used to describe

the growth of small-world networks (Barabási, 2013; Clauset et al., 2009; Jeong et al., 2003). The presence of small-world networks in living systems suggests an evolutionary basis for this structure. In an astrobiological context, unstructured chemical reaction networks have been observed in exoplanetary atmospheres, although it is unclear if more structured chemical networks exist due to detection limitations (Fisher et al., 2022; Kiang et al., 2018).

One application of network science is to compare and measure the complexity of different systems. Fully random networks, where nodes and edges have no correlation or pattern, can be easily distinguished from networks created by living systems through the measurement of the degree distribution (Barabási, 2013). The degree distribution of living systems' networks can be categorized as a power-law distribution and a high preferential attachment index (Clauset et al., 2009; Jeong et al., 2003; Newman, 2001). The degree distribution of a network allows for comparisons of the "life-like" qualities of systems across a variety of fields, including social and biological systems.

Chemical Scaling Patterns

The higher-order levels of organization in life reveal interesting observations and trends across systems. For example, in the context of societal complexity, there is an exponential increase in the production of objects, including chemical compounds, reactions, and patents (Brooks et al., 2011; Coley et al., 2018; Guo et al., 2021; Hähnke et al., 2015; Szymkuć et al., 2021). In addition, the amount of scientific literature has also increased exponentially, leading to a decrease in the influence of individual papers (Park et al., 2023). Moreover,

the human population increased exponentially since the Industrial Revolution, but the rate of innovation has remained linear at best (Szymkuć et al., 2021; West, 2018). Taken together, there are exponential increases in various factors – production, research, and population – without a corresponding increase in innovation. Innovation can have a wide variety of meanings, such as novel chemical reaction classes created per year based on a set of author-defined reaction classes and chemical patent databases (Szymkuć et al., 2021), or inventions and novel ideas, both of which are increasing sub-exponentially (Kempes et al., 2019; West, 2018). The third chapter of my dissertation tests how exponential production leads to innovation in chemistry by using AT as an agnostic means of measuring complexity. This is an improvement on previous work, as AT does not rely upon author-defined reaction classes. The third chapter also explores how social structures, such as cost, usage rates, individuals, and companies potentially drive innovation in chemistry.

Size Limitations of AT

It is worth mentioning that both Chapters 2 and 3 exclusively focus on small molecules, rather than macromolecules or larger structures. Part of this is a computational limit of AT – the number of possible construction paths increase exponentially as the size of the final structure increases. This leads to the AT calculation likely being an NP-hard problem, a computational classification where every possible construction path must be analyzed in order for the most optimal path to be found (Johnson, 1985). Additionally, the base unit of larger structures may be different. For example, when considering proteins, the base unit of an AT calculation may be the bonds between elements, or it may be the amino acids that

form the peptide backbone. This becomes even more of an issue when considering technological structures, where the base unit may be a joining step between disparate parts, such as in the case of building a piece of IKEA furniture. Ultimately, the challenges of considering different sizes and defining base units was outside the scope of Chapters 2 and 3. I exclusively used a base unit of bonds connecting two elements, which allows comparisons between the specifications necessary to search for complex life elsewhere and the increase in patent chemistry complexity.

Science Communication

The implications of Chapters 2 and 3 in this dissertation may be profound. The possibility of unambiguously discovering life and habitable worlds elsewhere is one of the highest priority goals of NASA (Christensen et al., 2022), and the observed agnostic increase in chemical complexity can extend to predict future societal patterns. However, none of these lofty goals and implications are worthwhile without science communication and education to non-scientists. In fact, at this current stage in astrobiology where life has not been discovered elsewhere, the main output of the field of astrobiology is science communication regarding the potential rewards, impacts, and incremental progress of the search for life.

Astrobiology is a unique field of science in terms of how it influences popular culture. The idea of alien life elsewhere in the universe has been present in scientific and cultural circles for centuries. Brake and Hook link this idea to both the Copernican revolution and Darwin's Theory of Evolution (Brake & Hook, 2007). By defining the solar system as

heliocentric, Copernicus turned the cosmos into something inhuman, or “alien”. This directly contradicted religious thought, where the cosmos was centered around Earth and humanity, and led to an era of “terrestrial mediocrity”, where “the history of astronomy [and other planetary sciences] is a history of increasing humiliation” for human-centric thought (quotes from “The Information”, by Martin Amis (Amis, 1995)). Additionally, through Darwin’s Theory of Evolution (Darwin, 1872), it became well-known that humans do not occupy a special place in the hierarchy of species – we are equally as evolved as every other living species on the planet (Gould, 2002). This recognition of humanity’s non-unique place on Earth provided a fertile ground for exploration of the idea of extraterrestrial beings (Sagan & Druyan, 2011). Contemporaries of Darwin, particularly the French astronomer Camille Flammarion, hypothesized about the prevalence of other worlds where evolution could occur (Flammarion, 1980). Early science fiction writers such as H.G. Wells with “The Time Machine” (H. G. Wells, 2005) and “The War of the Worlds” (H. G. Wells, 2003) and Olaf Stapledon with “Last and First Men” (Stapledon, 2008) popularized the ideas of evolution and alien life within the emerging genre of science fiction that continues to be fertile and popular ground for exploring astrobiology in both writing and film (e.g., Chambers, 2015; Kubrick, 1968; Nekola Nováková, 2020), among many others.

The exploration of and interest in alien life in popular fiction is only one reason why astrobiology is an ideal science for science communication. As an emerging, interdisciplinary field of science (Billings, 2012; Dick, 2012), astrobiology is uniquely suited as a case study for highlighting the continual evolution and progress of science. The

ongoing search for life elsewhere is in the nascent stages of discovery, and techniques and theories are continually being proposed and discussed, in both scientific literature and popular science articles. The excitement surrounding this cutting-edge science allows for non-astrobiologists to observe and explore an emerging field of science (Fergusson et al., 2012). Additionally, astrobiology consists of a wide variety of established scientific disciplines, such as: geology (Domagal-Goldman et al., 2016); biology (Kolb, 2014; O'Malley-James & Lutz, 2013); astronomy (Shaw, 2007); physics (Walker, 2017); chemistry (Bains, 2004; Marshall et al., 2021); computer science (Gisiger, 2001; Vitas & Dobovišek, 2019); oceanography and atmospheric science (Clarke, 2020; Fisher et al., 2022); even philosophy (Chon, 2021; Dick, 2012) among many others. This incredibly wide range of possible entry points to studying the origin, emergence, and search for life gives many non-astrobiologists reasons to engage with and become interested in astrobiology (and science) in general (Impey, 2021).

Prison Education

Since astrobiology is an ideal science for introducing non-scientists to science, I use it to increase science interest and identity among juvenile prisoners in Chapter 4 of this dissertation. I developed and taught a place-based Earth Science & Astrobiology course based on transformative pedagogy at Life Learning Academy (LLA) within the Lewis-Sunrise Detention Unit in the Arizona Department of Corrections, Rehabilitation, and Re-entry (ADCRR). This course was part of the curriculum within LLA and fits within the Arizona Science Standards for Earth and Space Science. The students who were a part of this course are minors under the age of 18 who were charged as adults within the state of

Arizona. This mixed-methods study conducted both quantitative and qualitative research on the outcomes of the class in regard to the student's learning, interest, and identity with Earth Science as a whole, of which astrobiology is considered a sub-field and was mentioned extensively throughout the course.

Studying the impact of science education – in fact, education in general – on juvenile offenders is extremely rare. Most prison education opportunities and research have been conducted on adult male offenders at the expense of female and juvenile offenders (Rose & Rose, 2014). Additionally, most research performed on prison education efforts has focused on reducing recidivism, which is the rate at which former offenders are reincarcerated. The research overwhelmingly shows – again, for adult male offenders – that education is one of the highest achieving interventions for lowering recidivism (Baranger et al., 2018; Courtney, 2019; Ellison et al., 2017; Esperian, 2010; Fabelo, 2002; Gaes, 2008). However, there are relatively few efforts to explore how transformative pedagogical tools can be applied within a prison environment, particularly within science. Recent work describes how advocating for reintegration education results in improved outcomes upon release (Flynn & Higdon, 2022) and how education writ large can be a method and motivator for positive transformation for offenders (Szifris et al., 2018). This highlights how specific transformative pedagogical structure can lead to improved outcomes within a prison environment.

Specifically, we use the sociotransformative constructivism (sTc) pedagogy developed by Alberto Rodriguez (A. J. Rodriguez, 1998; Alberto J. Rodriguez, 2015; Alberto J. Rodriguez & Morrison, 2019). Sociotransformative constructivism is built using educational social justice theory (Maulucci, 2012) and transformative pedagogy, with an emphasis on the transformative results of teaching towards diversity and understanding of student's cultures and unique viewpoints. Here, building a curriculum based on sTc is specifically designed to lead to change within the established power structures in a prison environment so that students are empowered to take control of their learning, instead of having their education dictated to them by those in power (A. J. Rodriguez, 1998; Szifris et al., 2018). The use of Earth Science and astrobiology as the scientific topics are meant to assist the transformative nature of the curriculum and provide a wide, inclusive entry point to increased science interest and identity.

2: EXPLORING THE MOLECULAR LIMITS OF LIFE

Introduction

The unambiguous detection of extra-terrestrial life is a grand challenge for the scientific community (Hays et al., 2015; National Academies of Sciences & Medicine, 2019). Detecting alien life is difficult because we suspect that life beyond Earth may be radically different than life as we know it, including the possibility that it is based on completely different metabolic and genetic materials (Bains, 2004). We cannot use standard approaches from molecular biology or genomics to determine whether a sample is the product of life as those approaches are specific to our biosphere. Thus, the challenge of life detection involves one key unanswered question. How can we unambiguously determine if a chemical system is the product of life? We are at a point where this question can begin to be answered using state-of-the-art chemical detection methods (Arevalo et al., 2020), combined with assembly theory applied to molecular life detection (Y. Liu et al., 2021; Marshall et al., 2021) but it is not yet clear whether these insights can be deployed on life detection missions within the solar system. Here, we give the design constraints for the development of future life detection missions based on assembly theory that use mass spectrometry as the primary analytical technique. I use a cheminformatics approach to explore unconstrained chemical space that allows us to avoid the biases of known biochemistry to determine the mass range and resolution required for a space-flight mass spectrometer to unambiguously identify the molecular signatures of life using assembly theory. Ultimately, our starting hypothesis is that life elsewhere in the universe uses unfamiliar chemistry, and we must be prepared with methods capable of detecting life, independent of the large space of possible chemical options alien life might use.

We use assembly theory (AT), a novel measure of complexity, to determine if a chemical compound can be built from a living system (Y. Liu et al., 2021; Marshall et al., 2021). Assembly theory improves on previous complexity measures by providing an experimentally testable measure that can be interpreted as an unambiguous biosignature. Previous metrics of molecular complexity, which are mainly developed for computational drug discovery, all involve various theoretical or computational pitfalls (Méndez-Lucio & Medina-Franco, 2017). For example, structural measures – based on chiral centers, molecular weight, or compactness – are simple measures of complexity, but only consider one single measure of a compound rather than a holistic view (Sheridan & Kearsley, 2002). More complex measures, such as graph-based measures that consider subgraph counts (Bertz, 1981), quantum mechanics (Luzanov & Babich, 1995), or information theory (Böttcher, 2016) among many other possible factors (von Korff & Sander, 2013), ultimately lack correlation with experimental data, as well as with each other (Méndez-Lucio & Medina-Franco, 2017).

In contrast, AT considers the amount of information required to build a molecule from the space of possible chemistry available to create a range of complexity values from low (commonly available compounds) to high (exceedingly rare compounds that are only created by living system (Marshall et al., 2021; Sharma et al., 2022)). Specifically, the measurement of the *molecular assembly index* (MA) of a compound is done by calculating the fewest possible bonding steps necessary to build that compound from basic component

bonds. It is important to note that while these assembly steps do not correspond to actual chemical reactions used in synthesis, the MA - the integer value of the smallest number of bonding steps necessary to build the full compound – linearly correlates to fragmentation data found through mass spectroscopy (Marshall et al., 2021), providing a physical grounding and experimental verification of MA as a measure of a molecule’s complexity independent of the route of synthesis. The shortest path is calculated through a graph-based approach. The molecule is converted into a computational graph, with elements as nodes and bonds as edges. This graph is randomly fragmented into sub-graphs consisting of two elements and one bond, which are then recursively merged to create the full graph of the compound at hand. The intermediate sub-graphs created along this recursive process can be re-used, meaning symmetrical structures have a lower MA than non-symmetrical ones. Regarding the search for life, previous work has demonstrated that only living systems produce compounds with MAs of 15 or higher (Marshall et al., 2021). This is an estimate based on observed biochemistry, so for this work, we extend the definition of complex chemistry to include compounds with a MA of 20 or higher to account for potentially unknown abiotic processes elsewhere which could result in high MA values.

Additionally, AT includes *copy number* as an essential measure. This value is the number of each unique molecule in a sample, where high values represent high concentrations of molecules within an environment and high robustness of the construction processes necessary to build a given molecule. A compound with a high copy number, typically in the order of 10^4 (Marshall et al., 2021), in conjunction with a high (≥ 15) MA suggests that

a living system preferentially selects for the existence of that compound in light of all other possible compounds available within chemical space (Sharma et al., 2022).

We can measure the MA of a compound, along with a trivial measure of copy number, through mass spectrometry (Marshall et al., 2021). Mass spectrometry can be used to determine the molecular weight of a compound (or mix of compounds) from the mass-to-charge ratio (m/z) of ionized molecular fragments and has a long legacy on space missions from Apollo 15 in 1971 to proposed missions to Europa, Mercury, and elsewhere (Arevalo et al., 2020; Chou, Mahaffy, et al., 2021). Mass spectrometers require a minimum ion count in the 100s to 10000s, thereby automatically satisfying the copy number constraint of AT for any detected molecule. Developing assembly theory as a life detection method using mass spectrometry in space therefore has only two theoretical constraints: 1) determining the precision necessary for a mass spectrometry spectra signal to provide evidence of previously unknown, distinguishable, high-assembly molecules, then 2) ascertaining if the detected molecules have a sufficiently high MA to be considered a sign of life. There are many technical challenges required to building a mass spectrometer which fits these constraints, which are beyond the scope of this dissertation.

There are several challenges involved in sending high-precision instrumentation on spacecraft that make detecting complex chemical compounds difficult. These challenges include the low precision of heritage instruments (technology used on previous missions), the long development time required to add newer and more precise instruments to future

missions, and the potential for mixed samples of unknown chemistry (Arevalo et al., 2020; Merder et al., 2020; Ren et al., 2018). Given these challenges, current spaceflight mass spectrometry technology is not designed to identify the complex molecules that provide evidence of life, as both high mass range and high mass resolution ($m/\Delta M$) are necessary. The mass range corresponds to the lowest and highest values (in Daltons) where ionized fragments can be read. Different mass spectrometry designs can provide different mass ranges, with the highest ranges belonging to time-of-flight instruments that are capable of detecting fragments up to nearly 10,000 Daltons in recent spaceflight missions (Arevalo et al., 2020). Mass resolution is the closest separation between spectral peaks, where higher $m/\Delta M$ values correspond to stronger distinguishing power between peaks (G Marshall et al., 2013). Terrestrial mass spectrometers, such as Fourier transform ion cyclotron resonance machines, have incredibly high mass resolution ($> 10^8 m/\Delta M$). However, these instruments are very large and sensitive, making them functionally impossible to include as part of a spaceflight mission (Arevalo et al., 2018; Chou, Mahaffy, et al., 2021). Technological advancement of spaceflight mass spectrometry instrumentation is necessary to advance the field of biosignature detection so that future missions will be able to detect high-MA chemical structures indicative of life in unknown chemical mixtures leading to more confident assessment of positive detection.

Current mass spectroscopy approaches to identify unknown compounds involve matching spectra to known chemical formulas (Kind & Fiehn, 2007) and integration of spectra into large-scale databases such as PubChem (S. Kim et al., 2019). As mass spectrometry

technology advances, the mass accuracy is routinely at the sub-ppm level (Tamara et al., 2021), and can be as high as the sub-0.1 ppm level (Bowman et al., 2020). Previous research has determined a mass spectrometer with a mass precision of 3 ppm is required for detection of high-mass metabolic compounds (Kind & Fiehn, 2006), and that a top of the line mass spectrometers (6 ppb mass precision) can distinguish complex chemical mixtures (Merder et al., 2020). However, these ultra-high-resolution methods are limited when it comes to analyzing extra-terrestrial chemistry, as we expect alien worlds to present potentially unknown chemical environments (Bains, 2004; Méndez et al., 2021) that may present a wider range of chemicals than that found on Earth. For example, many alternative hypotheses for chemical life have been proposed, involving unique chemical conditions for their evolution (Bains et al., 2021; Cleland, 2019; Irwin & Schulze-Makuch, 2020). The enormous range of potential chemistry elsewhere is the chemical space available to life. Research by Ruediger et al found that with a limited set of elements and a size restriction of 17 atoms, over 166 billion compounds can be potentially created on Earth (Ruediger et al., 2012), and it is estimated there are over 10^{60} different chemical structures available throughout the universe (Mullard, 2017). Exploration of this chemical space is necessary to provide mass range & resolution recommendations to future spaceflight mass spectrometers to distinguish unknown extraterrestrial chemistry.

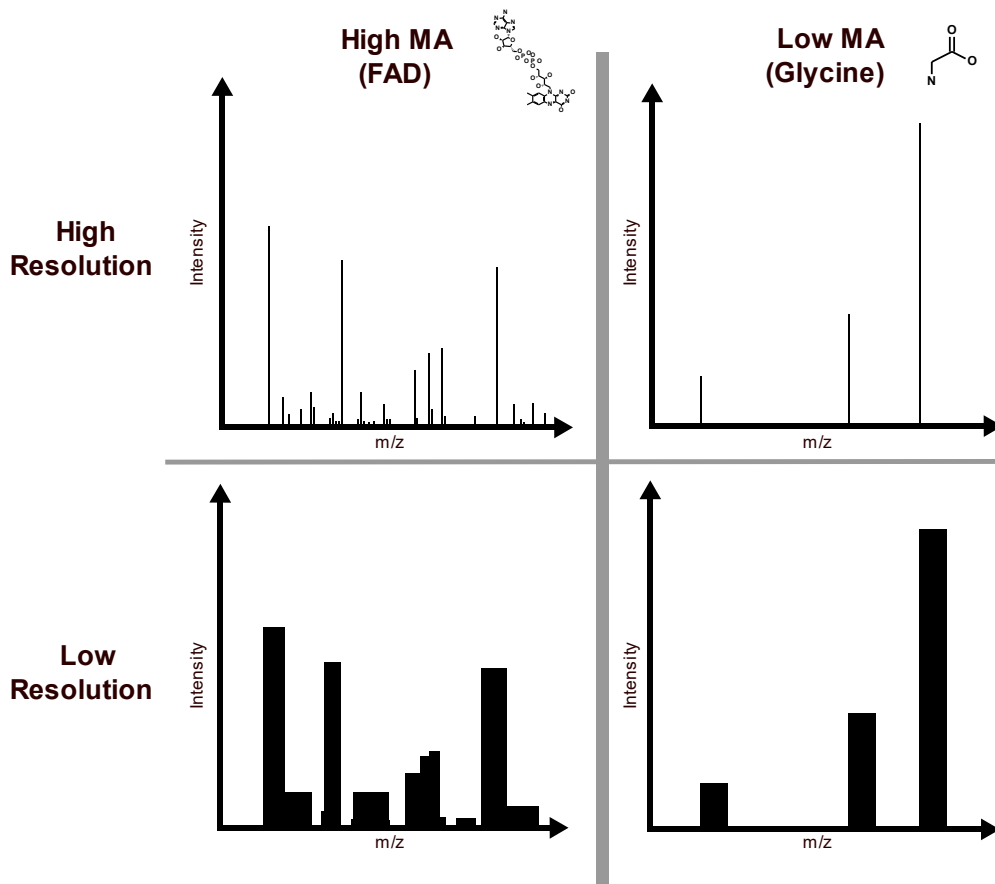


Figure 1: Chapter 2 concept figure.

Results

Building Chemical Space

The chemistry observed on Earth is an extremely small subset of all chemistry, estimated at only 10^{-50} of all possible chemical space available (Awale et al., 2017). It is very likely that living systems elsewhere are based on different chemical substrates, ranging from systems based on different energetic constraints (Bains, 2004) to unpredictable “life as we don’t know it” (Cleland, 2019; Marcheselli, 2019). To build a model of potential chemical space available to life elsewhere that could be detected by spaceflight missions, we used the cheminformatics enumeration program MOLGEN 5.0 to find structures and formulas of possible chemistry. This model included compounds that may be impossible to form on Earth due to thermodynamic and/or stability constraints, but could exist given different planetary environments (Guttenberg et al., 2021; Irwin & Schulze-Makuch, 2020; Seager & Bains, 2015). We limited our study to compounds containing carbon, hydrogen, nitrogen, oxygen, phosphorous, and sulfur (CHNOPS), as these form the bulk of Earth-based organic chemistry (Pace, 2001). Additionally, they are good targets for the bulk elemental composition of alien life because of the chemistry they can mediate and their high availability throughout the universe (Cockell et al., 2021). This element restriction also fits with current spaceflight mission directives of finding biomolecules that are similar to those on Earth (Meadows et al., 2022; National Academies of Sciences & Medicine, 2019).

MOLGEN 5.0 takes a specific number of elements, then outputs all theoretically possible chemical structures and formulas which contain that number of elements. We provided

MOLGEN 5.0 with an upper limit of atoms permitted for each of these elements in a compound. If this upper limit was set to five for all atoms, then the maximum number of each type of atom in the enumerated chemical structures allowed would be five. Structures containing fewer than five elements would be permitted. For example, with an upper limit of five CHNOPS elements, the maximum number of C atoms allowed would be five (with fewer than five permitted), the maximum number of N atoms allowed would be five, and so on. We did not apply this upper limit to Hydrogen so that as many H atoms as necessary can be added to ensure the generated structure was chemically plausible. We counted the number of possible formulas across a molecular weight distribution to categorize the size of chemical space available to each elemental limit. Here, chemical space shows an initial increase in the number of unique formulae as molecular weight increases. However, the number of unique formulae decreases as the number of atoms approaches the upper limit due to fewer combinations of atoms available (Figure 2). MOLGEN 5.0 became prohibitively computationally expensive as the upper limit of elements within a compound is increased, so we fit a Gaussian distribution to each molecular formula count distribution to predict the size of chemical space for high elemental limits (Figure 2) and validated this fit using a two-sample Kolmogorov-Smirnov (KS) test (Massey, 1951) (Table 1). This fit scales with increasing upper elemental limits (i.e., Figure 4, Figure 5, Figure 6) and is necessary to model chemical space of larger compounds.

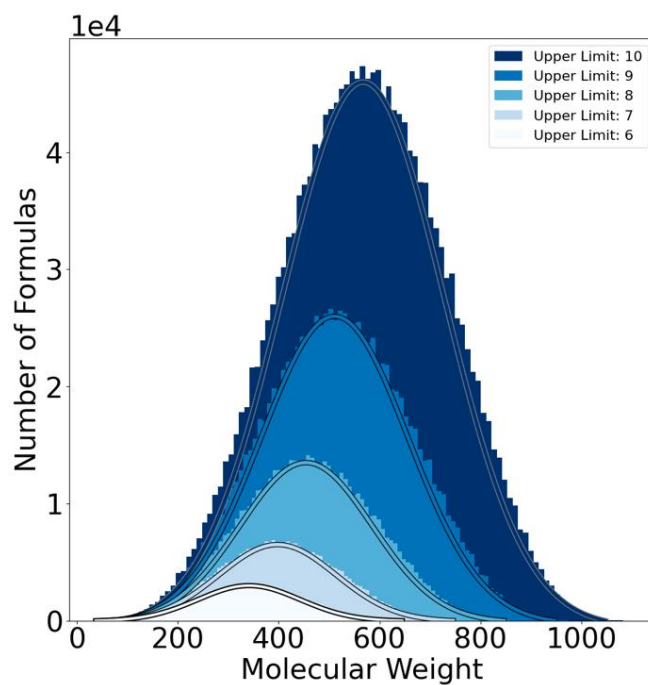


Figure 2: The number of CHNOPS chemical formulas with modelled gaussian distributions.

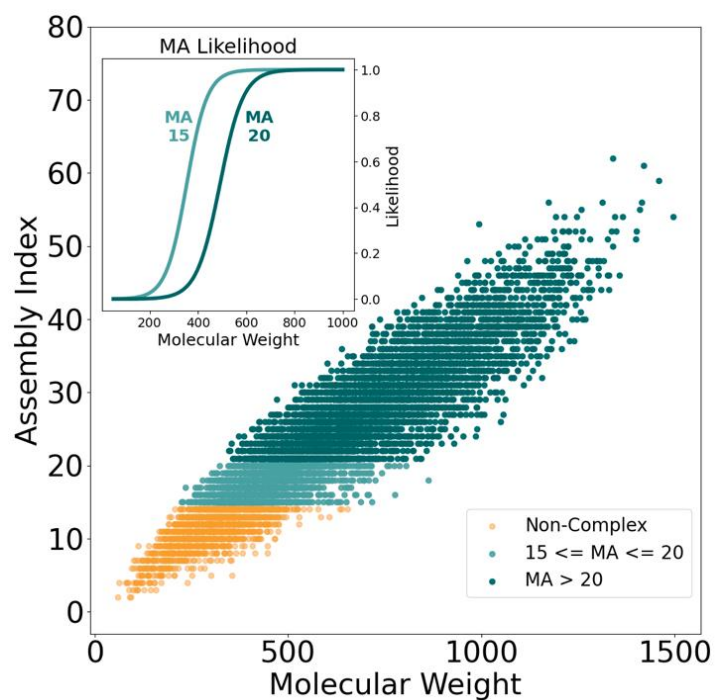


Figure 3: Assembly values of 10000 chemical structures and (inset) likelihood of high-assembly compounds as a function of molecular weight.

Table 1: KS Goodness of Fit tests of Gaussian modelled distributions for upper elemental limits 6-15.

Maximum Element Count	KS Statistic	P-value
5	0.0133	0.1037
6	0.0157	0.0200
7	0.0105	0.2361
8	0.0173	0.0056
9	0.0128	0.0777
10	0.0232	4.4011e-05
11	0.0238	2.5470e-05
12	0.0151	0.0207
13	0.0200	0.0007
14	0.0199	0.0007
15	0.0123	0.0979

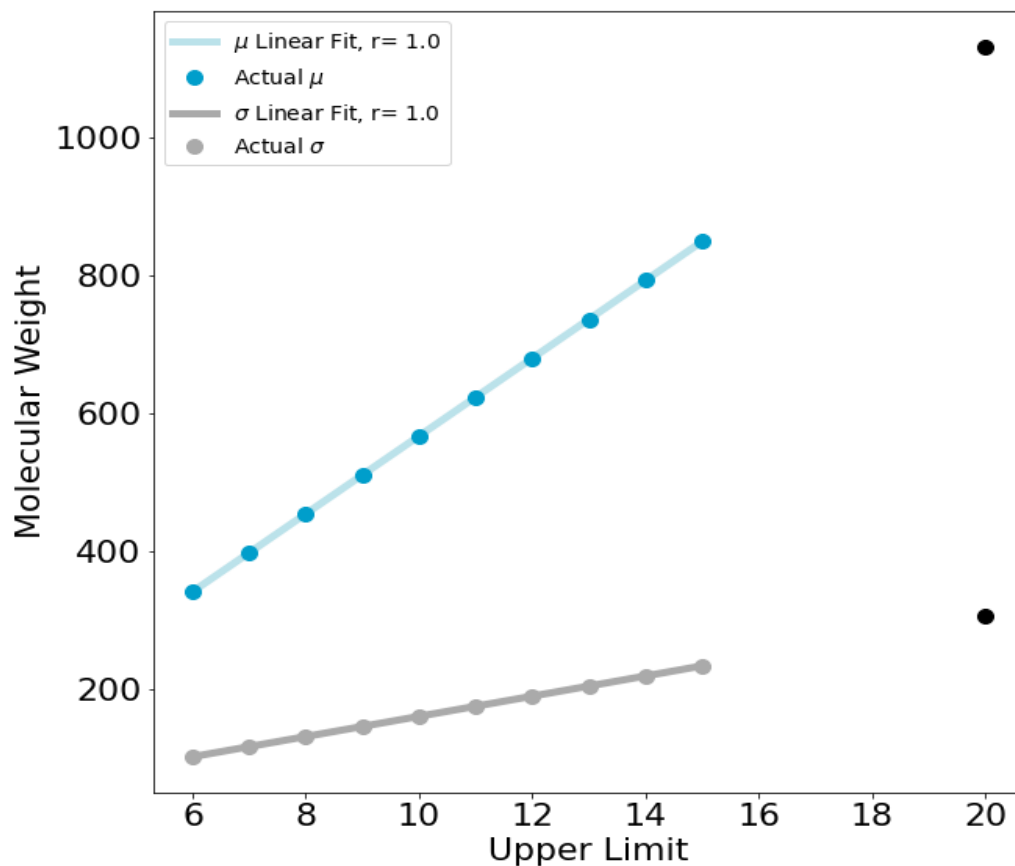


Figure 4: Linear regression predicting μ (mean) and σ (standard deviation) values for Gaussian distributions of number of formulae. I performed a linear regression on an upper elemental limit of 6-15 heavy atoms, and this regression was validated by testing at an upper limit of 20 (black dots – see Figure 5).

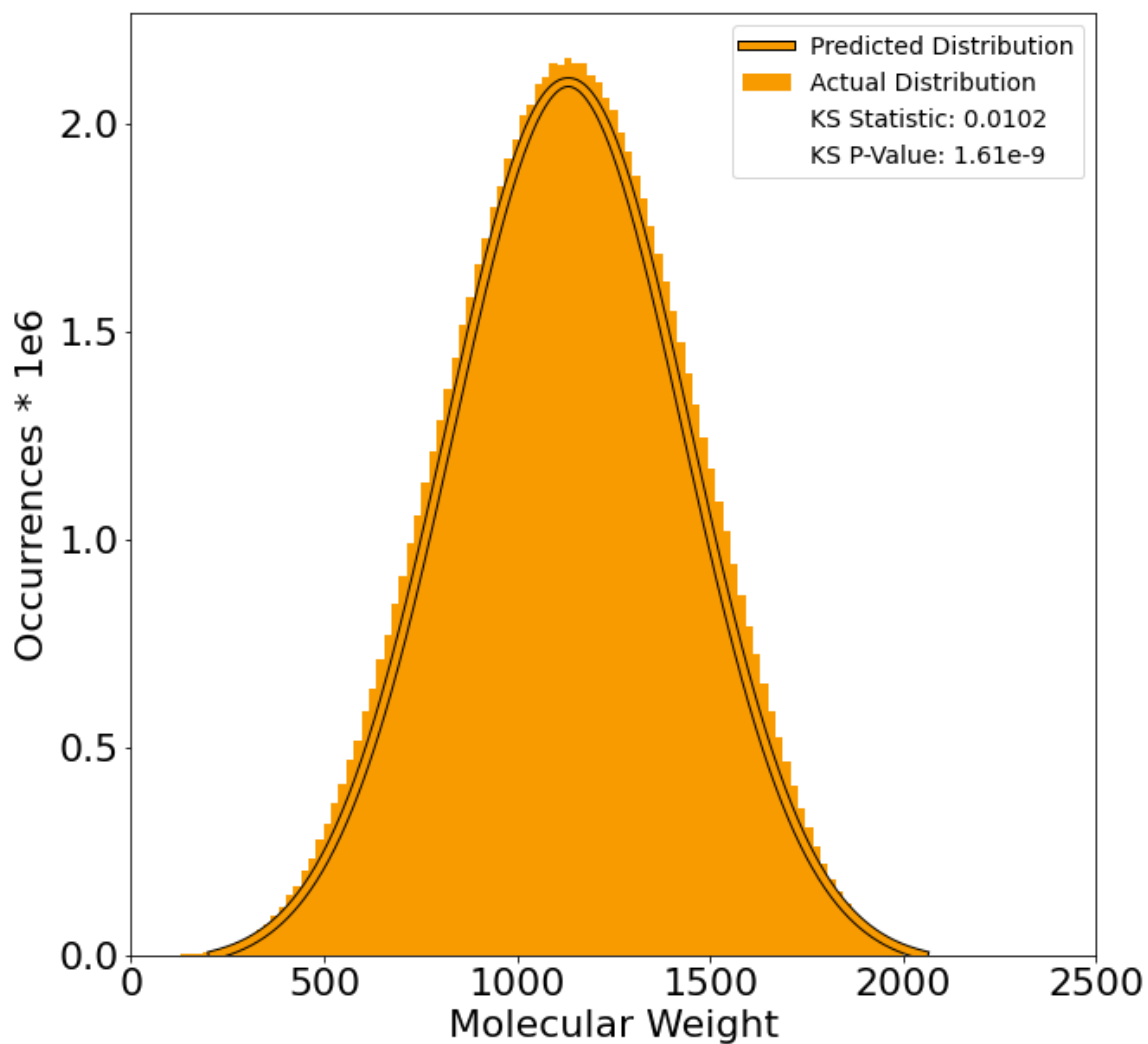


Figure 5: The predicted μ (mean) and σ (standard deviation) (Figure 4) accurately predict the computationally generated distribution of molecular weights where the maximal number of elements is 20.

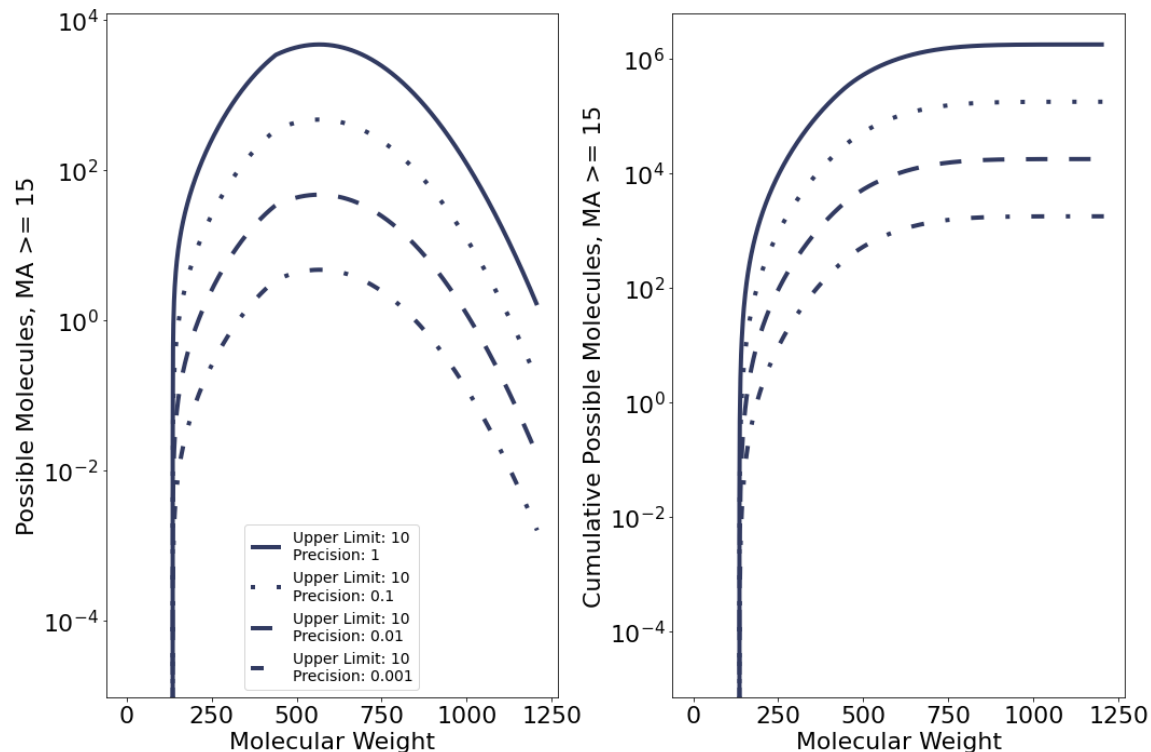


Figure 6: The number of formulas at a given molecular weight (left) and the cumulative number of formulas as a function of increasing molecular weight.

Molecular Assembly Over Chemical Space

We expect high-assembly compounds to be found more often as molecular weight increases. As the number of atoms increases, the number of bonds tends to increase, leading to an increase in the number of steps necessary to form a full chemical structure, meaning higher MA is possible (Sharma et al., 2022). We calculated the likelihood that a chemical compound has a high molecular assembly index as a function of molecular weight. An $MA \geq 15$ has been shown to be the delimitation in chemistry between compounds generated uniquely by Earth life and those that can also be generated from non-biological sources (Marshall et al., 2021). We also highlight MA values ≥ 20 to unambiguously account for unknown boundary cases where abiotic processes can potentially build a compound with a

MA between 15 and 20. We generated 10,000 random chemical structures corresponding to molecular weights between 50 and 1500 Daltons with an upper limit of 10 atoms of each element and calculated the assembly index of each compound (Figure 3). This random structure generation was done using an assembly-based algorithm. For each structure, we randomly chose a formula from the formulae output of MOLGEN 5.0, then iteratively built bonds between distinct atoms in the formula until all atoms were part of a single chemical graph. The graphs were validated using RDKit's 3D embedding tool that transforms a chemical graph into 3D space and rejects graphs where bonds cannot be mapped to physical structures due to valence rules, bonding limitations, or other physical constraints (Landrum, 2020). Only validated graphs -i.e., thermodynamically possible molecules - were used in the MA analysis. We performed a logistic regression to find the likelihood of discovering a complex compound ($MA \geq 15$ or $MA \geq 20$) at a given molecular weight sampled from this random chemical space (Figure 2b). Above 654 Daltons, the likelihood of a compound with an $MA \geq 15$ goes to 100%, and above 862 Daltons, the likelihood is 100% for compounds with an $MA \geq 20$. Any instruments utilized to search for assembly-related biosignatures should have a high precision at high molecular weights, starting with at least 654 Daltons to identify high-MA (≥ 15) compounds and 862 Daltons for unambiguously high-MA (≥ 20) compounds.

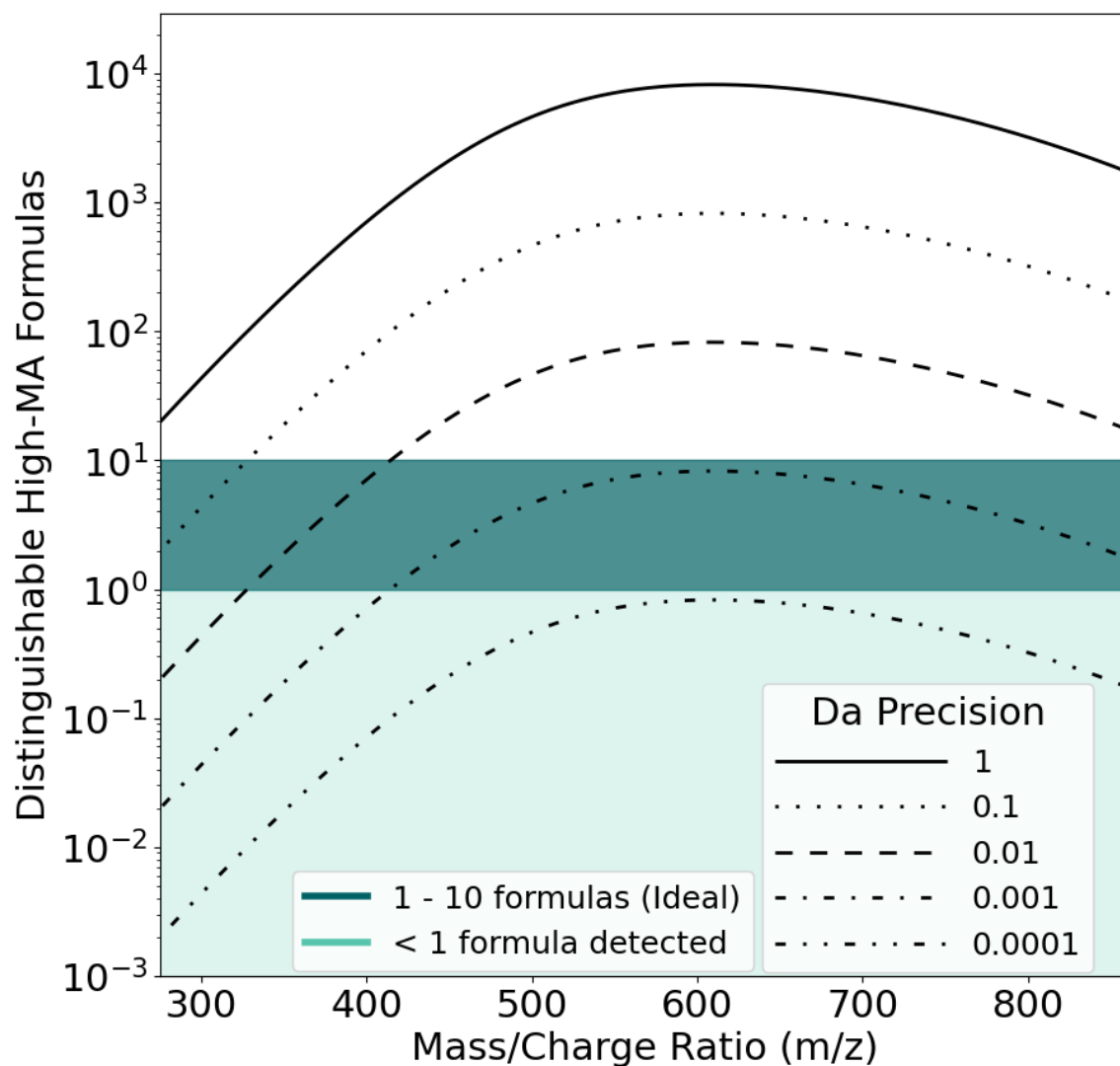


Figure 7: Precisions at which high-assembly compounds can be detected.

We calculated the number of detectable high-MA formulas by a mass spectrometer across various precisions using the theoretical number of formulas available in chemical space and the likelihood of finding a high MA compound at a given molecular weight. We first defined the mass range in which we searched for the number of unique formulae. We defined the mass range over five orders of magnitude, 10^0 to 10^{-4} Daltons. For each order of magnitude, we split our pre-defined m/z range (275 – 861 defined by the molecular

weight where the likelihood of finding of high-MA compounds is between 0 and 1 (Figure 3, inset), then subtracted by one to account for ionization) into chunks. For example, if the mass range is 1 Dalton, we calculated the number of distinct formulae with a m/z of 275-276 (exclusive), 276-277 m/z , 278-279 m/z , and so forth. We then multiplied the number of distinct formulae by the likelihood of finding a high-MA (≥ 20) compound at a particular m/z value (Figure 3). This provides the number of high-MA formulae which are theoretically detectable at a specific m/z value given a particular mass precision (Figure 7). As our goal is to find the resolution at which mass spectrometers can detect 10 or fewer high-MA formulas, we highlight the precisions where 1-10 formulae can be detected, as well as where fewer than one formula can be detected, which represents resolutions which are better than necessary for the purposes of finding high-MA compounds. Importantly, as the number of possible structures increases up through 609 m/z , it becomes less certain that a unique structure can be identified with a fixed resolution, requiring a more precise instrument. After 609 m/z , fewer high-MA structures are detected due to our constraints put on the graph enumeration, as our pre-defined upper limit of 10 CNOPS elements lead to fewer combinations of atoms and therefore fewer formulas.

Mission Specification

Mass spectrometry on spaceflight missions is steadily increasing in precision, and as a result, this leads to a steadily increasing capacity to meaningfully detect high-assembly chemical compounds (Figure 8). A quadrupole mass spectrometer (QMS) on the recent Curiosity Mars Rover has a mass detection range between 1.5 - 535.5 Daltons (Da) with a resolution of 5355 $m/\Delta M$ (Mahaffy et al., 2012). The ion trapping mass spectrometer in

the Mars Organic Molecule Analyzer (MOMA) instrument suite in the European Space Agency's ExoMars 2020 Rover allows for a mass range of 50-500 Da and a resolution over 500 m/ Δ M in gas chromatography mode, and a mass range of 50-1000 Da with the same resolution in laser desorption ionization mode (Goesmann et al., 2017; Li et al., 2017). The proposed time-of-flight mass spectrometry (TOFMS) instruments on the Jupiter Icy Moon Explorer (JUICE) mission has a mass detection range of 1-1000 Daltons with a mass resolution of 757 m/ Δ M below 642 Da (Föhn et al., 2021). The proposed Dragonfly mass spectrometer (DraMS) on the upcoming NASA New Frontiers mission to Titan has a high mass range up through 1950 Da, but a mass resolution of roughly 1375 m/ Δ M up through 550 Da, with a lower resolution beyond that (Stern et al., 2023). The upcoming Europa Clipper mission includes two mass spectrometers: a TOFMS with a mass resolution up to 23,822 m/ Δ M from 16-114 Da (Brockwell et al., 2016); and a Fourier transform mass spectrometry instrument, the Characterization of Ocean Residues and Life Signatures (CORALS) Orbitrap with mass range of 20-600 Da and 120,000 m/ Δ M mass resolution (Willhite et al., 2021). Similarly, a prototype of a space based OrbiTrap instrument, CosmOrbitrap, has been developed to have a similar mass range to CORALS, with a mass resolution as high as 140000 m/ Δ M (Arevalo et al., 2018).

The CORALS and the proposed CosmOrbitrap mass spectrometers can detect ≤ 10 formulas that correspond to high assembly, up through 448 and 461 Da, respectively. While this range is incredibly useful for the potential detection of biosignatures, we observe a mixture of low-assembly and high-assembly compounds at higher mass values through 862 Da (861 m/z) (Figure 3). Obtaining precise structural data up to 862 Da is essential, so we

recommend that future development of mass spectrometers on spacecraft missions aim for a mass precision of 552,252 $m/\Delta M$ at minimum 862 Da, or a factor of 3.9 higher precision at 1.4 times higher mass range than achieved by the proposed CosmOrbitrap instrument.

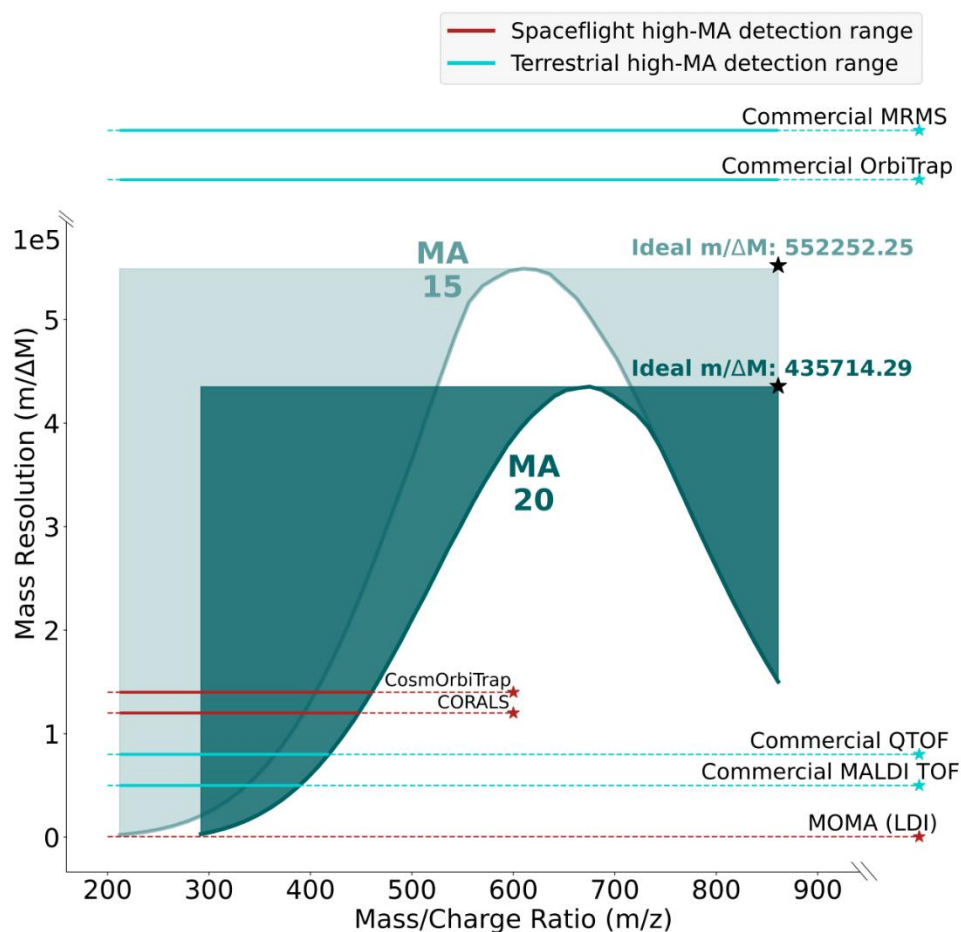


Figure 8: The minimum $m/\Delta M$ resolution necessary to distinguish exactly 10 high-MA compounds at each molecular weight.

Distinguishing Existing Biochemistry

We tested the distinguishability of terrestrial biochemistry using two existing spaceflight mass spectrometers and our recommended instrument resolution (Figure 9). As our goal is to provide a resolution which can determine 10 or fewer potential structures given unknown

chemical space, we use known biochemistry as a control to ensure that our recommendation of 552,252 $m/\Delta M$ – and the resolution of state-of-the-art mass spectrometers – can sufficiently distinguish living systems. We used the MOMA Laser Desorption/Ionization mass spectrometer on the 2020 ExoMars Rover (MOMA LDI; 500 $m/\Delta M$ mass resolution) and the Fourier transform mass spectrometer (CORALS; 120,000 $m/\Delta M$) on the soon-to-launch Europa Clipper mission to represent state-of-the-art, flight-ready instruments which are currently being used or will be soon used in life detection missions. We obtained 10208 chemical formulas and corresponding molecular weights from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (M. Kanehisa & Goto, 2000), a commonly used database for cataloguing small molecular metabolism found on Earth. Using the molecular weights of each compound, we calculated the number of formulas for whose largest mass spectrometry peaks overlap (cannot be distinguished) at three different mass resolutions. The m/z value of the largest peak is the molecular weight subtracted by one to account for ionization and does not take fragmentation into account, as the MOMA LDI and CORALS use different fragmentation methods and future mass spectrometers may also not follow the same methods. For each of the three resolutions (500 $m/\Delta M$, 120,00 $m/\Delta M$ and 552,252 $m/\Delta M$), we split the full m/z range into bins, with the bin size dependent upon the resolution. We use equation [7] to find the m/z precision of each compound, which then was used to place compounds into m/z bins. The number of KEGG formulae which appear in each bin is the overlap count. For the MOMA LDI instrument, the highest overlap occurs at a m/z of 338, where 50 unique formulae overlap. In contrast, the highest overlap value for CORALS occurs with only five formulae

overlapping. This occurs three times, at m/z values of 226, 240, and 269. Our recommendation's highest overlap occurs at 300 m/z with only three formulae, at m/z values of 227, 266, 304, and 320. Both CORALS and our recommended mass resolution have fewer than 10 overlaps throughout the m/z range of small molecule metabolism, demonstrating the viability of detecting Earth-based biochemistry using existing mass spectrometry technology. While potential chemical space is much larger than terrestrial metabolism, it is likely that living systems elsewhere do not include all potential chemistry available (Dobson, 2004). Therefore, these high-resolution instruments (both CORALS and our recommendation) are potentially sufficient for unambiguously detecting high-MA compounds elsewhere. Low-resolution mass spectrometers, such as the MOMA LDI instrument, cannot distinguish between biomolecules on Earth, meaning they are potentially unsuited for detecting biosignatures given the possibilities of chemical space elsewhere.

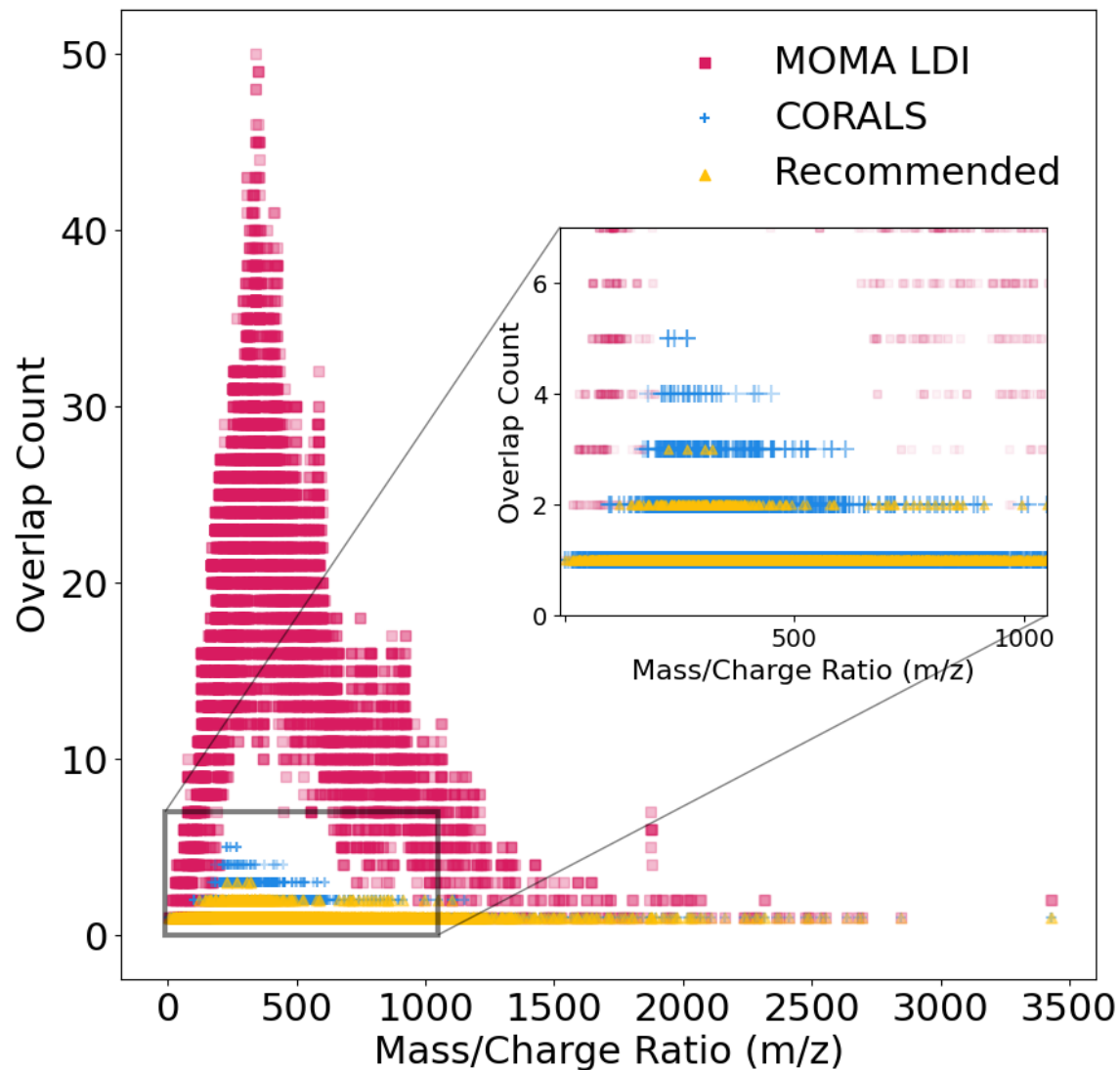


Figure 9: The number of overlapping biochemical compounds within KEGG as a function of different mass resolutions.

Chemical Space of Life

Our mass spectrometer recommendation is based on the assumption that the chemical space available on other planets is unconstrained, and therefore can be modelled by random chemical structures. Constrained systems show a similar distribution of molecular assembly values as that of random chemistry (Figure 10), justifying the use of this random chemistry as the foundation of our instrument recommendations. We compared assembly

values of chemical formulas across three structural databases: randomly generated structures; GDB17 (Reymond, 2015); and PubChem (S. Kim et al., 2019), each of which represent different sets of constraints and prior knowledge of a system. We built distributions of assembly indices for compounds sampled over structures with the same formulae. Two hundred chemical formulae were randomly selected, each of which had at least 100 isomeric structures in GDB17 and PubChem. We additionally generated 100 random structures for each formula, for a total of 300 structures for each formula (100 obtained from PubChem, 100 obtained from GDB17, and 100 randomly built). We found the MA for all structures and calculated the average and standard deviation of the MA values within each formula, separated by database. The average MA & standard deviation across each database show there is no difference between the constrained PubChem and GDB17 datasets and the less constrained randomly generated chemical structures, as all formulas occupy the same space of molecular assembly values.

PubChem is the most constrained of the three databases used, as it contains only compounds which have been experimentally verified and utilized in some form. The structures in PubChem are curated from a wide variety of sources, all of which involve real-world usage of compounds (S. Kim, Thiessen, Cheng, et al., 2016). This curation imposes a layer of physical implementation onto chemical space - if a compound is not used in real-world chemistry or if it is not entered into the database, it is not included. It is ultimately subject to the scientists who input the data, which incorporates a bias not found within the enumeration methods of the two other data sources used here. We do not consider this a negative, as PubChem still holds an extremely large amount of chemical

data (over 90 million unique chemical structures) and is useful for our purposes when recognizing the constraints of data entry. In comparison, GDB17 is built from graph-based enumeration and includes chemical and reactivity constraints, as well as size and elemental restrictions. Structures must have no more than 17 non-hydrogen atoms, must be stable, and must satisfy functional group reactivity constraints (Ruddigkeit et al., 2012). The third source of structures used here are random chemical structures, built using iteratively connecting random bonds between a given set of atoms. We used the same process as in Figure 2, but here we generated separate structures in the MA – molecular weight correlation experiment. The stability and reactivity of these random structures is not considered, so many of these structures may only be possible in other planetary environments with vastly different thermodynamic constraints.

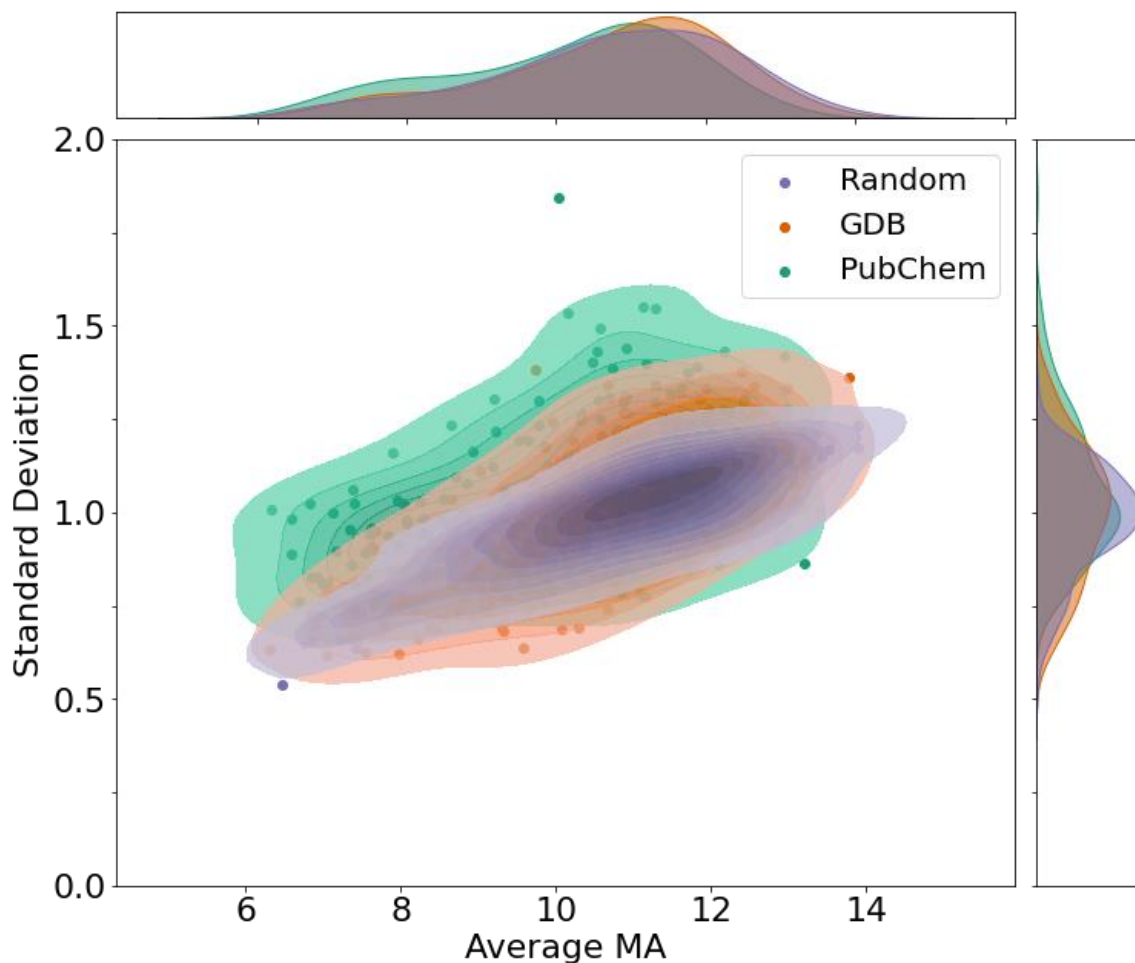


Figure 10: MA distributions over chemical databases representing differently curated and biased chemical space distributions.

Overall, the similarities in MA distributions between relatively unconstrained random chemistry and the more constrained GDB17 and PubChem databases show that the random chemical structures used to generate the mass spectrometry recommendations occupy a similar MA space to the chemistry we observe here on Earth. The overlapping distributions justify the application of randomly generated chemical structures to model and predict chemistry produced by potentially biochemical processes elsewhere, as random structures and biochemical structures share the same molecular assembly profile.

Methods

Formula Generation

A python script was used to run MOLGEN 5.0 analysis, using a MOLGEN 5.0 license belonging to the Cronin group to generate formulas and incorporates the default bad list defined by MOLGEN 5.0. The search query used was:

```
mgen C-010H0-100N0-10O0-10P0-10S0-10 -mass X -badlist badlist.sdf -o <output.txt>
```

This example query is for a formula generation where the upper element limit is 10, and where X is a range of atomic masses from 50-2000. The outputs were stored on a secure server at the University of Glasgow.

Rather than exhaustively generating all possible chemical structures for our constraints, we found all possible chemical formulas from CHNOPS elements, and limited the maximum number of atoms for each element (excluding Hydrogen) to 5 through 10, inclusive ().

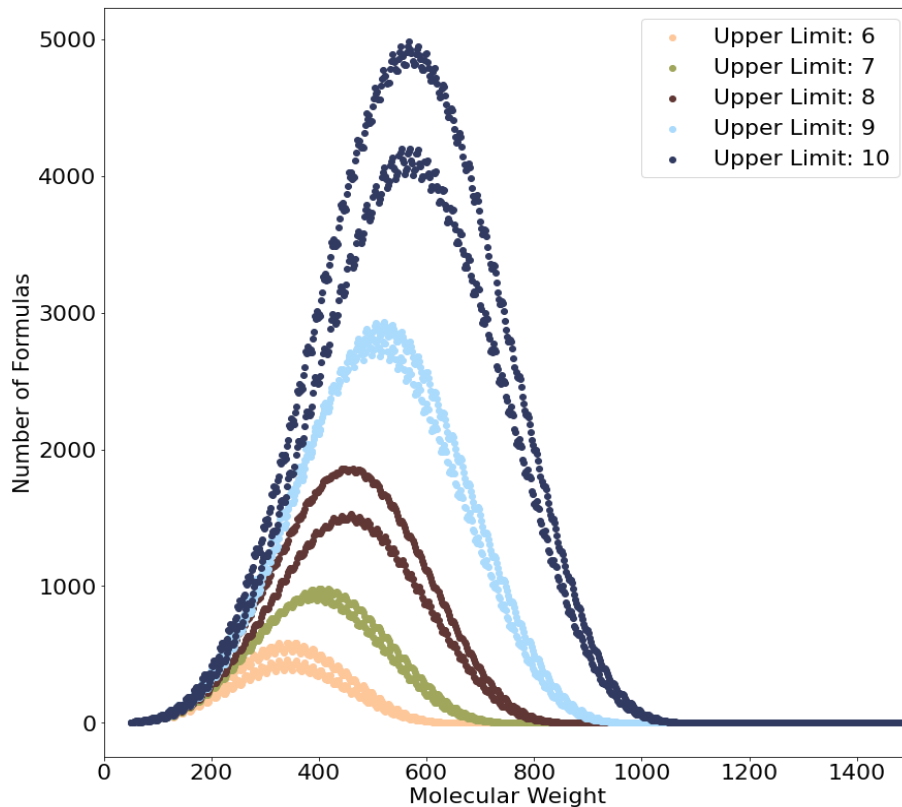


Figure 11: The number of formulas generated for 0-1500 Daltons, as a function of the maximum number of CHNOPS elements (6-10 shown here).

Fitting Formula Count

To model the number of possible molecular formulas given an arbitrary upper limit without performing the computationally expensive step of generating every molecular formula, we fit an exponential curve to the exhaustive formula counts (Figure 11) for the upper atomic limits in the range 5-15, inclusive. We used the Python library numpy (Harris et al., 2020) to calculate a weighted exponential fit for each distribution. We used linear regression to fit the exponential equation (12).

$$y = Ae^{Bx} \quad (1)$$

In practice, fitting an exponential using linear regression without weighting overfits smaller values. Therefore, we weighted the regression by y to properly fit all values. The r^2 value for this weighted fit is 0.991 (p-value = 3.667e-9). For the exponential function, $A = 0.4598$ and $B = 9.8410$. (Figure 12)

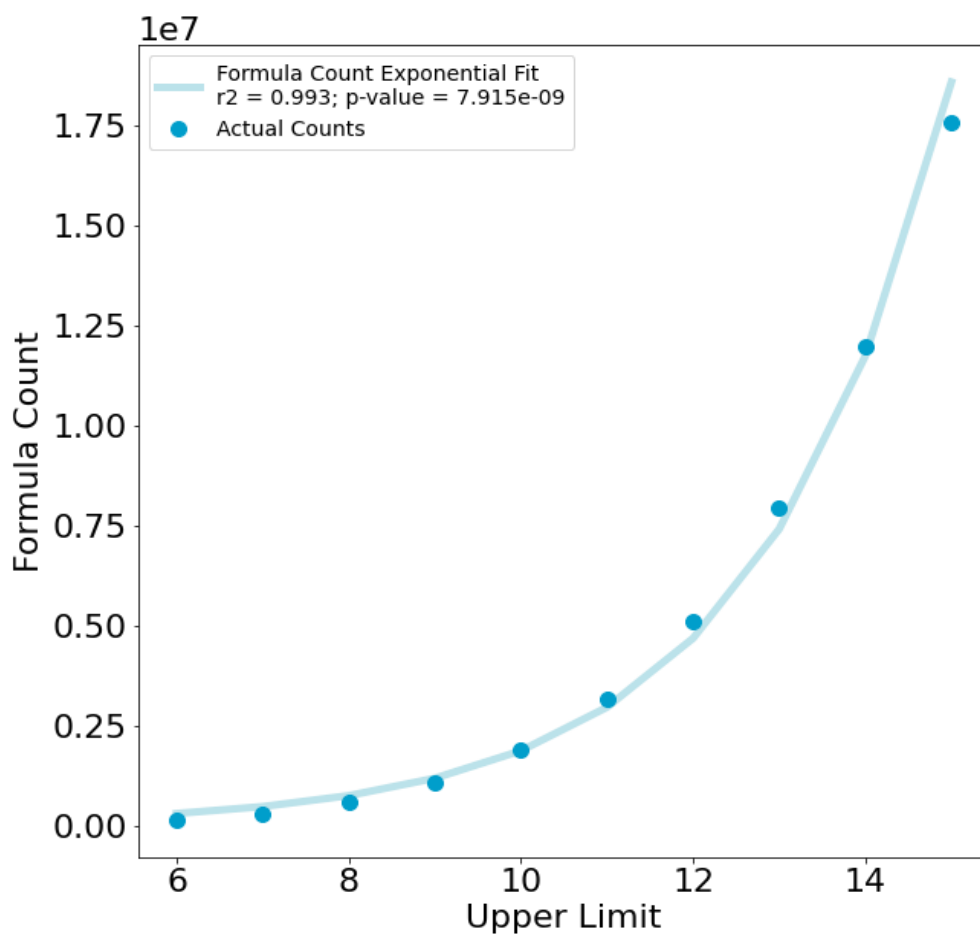


Figure 12: The number of formulas per maximum element count modelled using an exponential curve.

Molecular Formula Size Fitting & Distributions

We calculated the average molecular weight for each chemical formula for all maximum number of elements in the range 5-15, inclusive, using the molmass python package,

v10.18. For each upper limit of the number of atoms of each element, we fit a Gaussian distribution ((2) to the distribution of molecular weights, with μ as the average molecular weight and σ as the standard deviation (SI Figure 3).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (2)$$

To test the validity of this Gaussian model to the molecular weight distribution generated from MOLGEN 5.0 formulas, we used a two-sample Kolmogorov-Smirnov (KS) test (Massey, 1951) to measure similarity between the modelled Gaussian distribution and the formula-generated molecular weight distribution. This test was done using the Scipy python library (Virtanen et al., 2020). The KS test makes no assumptions about the two distributions being compared, as it compares the cumulative distribution functions of each distribution. All KS statistics are less than 0.025 (Table 1).

Mean/Standard Deviation Fitting

To model the Gaussian distribution of molecular weights given an arbitrary upper elemental atom limit without performing the computationally expensive step of generating formulas and calculating molecular weights, we fit a linear regression to the average molecular weight and standard deviation for each of the upper limits in the range 5-15, inclusive (molecular weight $m=0.999$, $b=1.003e-5$, standard deviation $m=0.999$, $b=0.0006$). The r^2 value for both the average (μ) and standard deviation (σ) are 1.000, with a pvalue of $2.200e-32$ and $4.148e-23$ for the molecular weight and standard deviation fits, respectively). Therefore, using the linear relationship calculated here, molecular weight distributions for compounds with arbitrarily large numbers of elements & atoms can be

modelled. We tested this with an upper element atom limit of 20 as a proof of concept, with the KS statistic = 0.0102 and the p-value = 1.61e-9 (Figure 4, Figure 5).

Random Molecule Generation

We built random molecules for MA calculations through a generative assembly process. Given a molecular formula, all chemically possible bonds which can be formed between the given atoms are generated. For example, given two Carbon atoms, the possible bonds would be C-C, C=C, and C≡C. Two bonds are randomly sampled with uniform probability from this distribution. If these bonds can chemically be bonded together, they are kept and are part of the growing molecule. If not, two other bonds are chosen. This process iterates through randomly sampling one bond to add to the growing molecule until all atoms in the chemical formula are accounted for. The final molecule is checked by the 3D embedding tool within RDKit (Landrum, 2020) for chemical feasibility. The random molecule generation code was written by Stuart Marshall.

Assembly Index Calculations

The MA of chemicals was found using the latest version of the Molecule Assembly code (written by Stuart Marshall in Go), with a timeout of 300 seconds.

Number of Possible Molecules

We calculated the number of theoretically possible molecules detected within increasingly smaller molecular weight ranges and specific DOFs. We calculated the likelihood of finding molecules through a two-step process. First, we found the cumulative distribution function – the probability that a random variable will have a value less than or equal to a specific value (**Error! Reference source not found.**) - at a specific molecular weight plus

a given precision, then subtracted the cumulative distribution function at the same molecular weight minus the given precision (**Error! Reference source not found.**).

$$cdf(x) = P(X) \quad (3)$$

$$\begin{aligned} &P(mw \pm precision) \quad (4) \\ &= cdf(mw + precision) - cdf(mw - precision) \end{aligned}$$

This likelihood is multiplied by the theoretically expected number of molecular formulae generated from the given atomic limit (**Error! Reference source not found.**). The exponential fit from (12) is used here, with $A = 0.4598$ and $B = 9.8410$.

$$\begin{aligned} &expected\ molecules(mw \pm precision) \quad (5) \\ &= P(mw \pm precision) * 0.4598e^{9.841 * atomic\ limit} \end{aligned}$$

These calculations are repeated for all possible ranges within four standard deviations of the average theoretical molecular weight to create distributions of the number of theoretically possible molecules which can be detected.

Likelihood

We generated the likelihood of finding a high MA compound (using $MA \geq 15$ and $MA \geq 20$ as delimiters) through logistic regression, using the GLM 4.4 method in R.

The likelihood values from the logistic regression were included in a modified version of (**Error! Reference source not found.**), resulting in the number of molecules found at a particular weight which have a molecular assembly value greater than 15 (or 20, depending on the delimiter used) (**Error! Reference source not found.**).

$$\begin{aligned} & \text{expected molecules } (MA \geq 15, mw \pm \text{precision}) & (6) \\ & = [\text{eq. 5}] * (1 - \text{likelihood}(M \leq 15, mw \pm \text{precision})) \end{aligned}$$

m/ Δ M Calculation

We calculated m/ Δ M values necessary for detecting at least 10 distinct formulae through taking the ratio of the Dalton precision required to detect 10 or fewer formulae at a given m/z value between 275 – 861, and the m/z value. This precision can be thought of as the maximum values of the “Ideal” section denoted in (Figure 7). We used (**Error! Reference source not found.**) across all m/z values in the range 275.00000 – 861.00000 m/z (inclusive), with a stepwise increase of 1e-5 m/z.

$$\frac{m}{\Delta M} = \frac{m/z}{\text{precision}} \quad (7)$$

KEGG Overlap

Formulas were taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG) in August 2022. The 10208 formulas found is an exhaustive list of formulas present in the database. The molecular weights of each formula were calculated using the molmass python library v10.18, and the overlaps at various mass resolutions were calculated using python scripts.

Assembly Theory Over Different Databases

Assembly theory results over different chemical databases (MOLGEN 5.0, GDB, and PubChem) representing differently curated and biased chemical space distributions were calculated. We found 100 isomers from each database - generated through randomly

building graphs, sampled from GDB17, and sampled from PubChem - across 200 chemical formulas. These chemical formulas were randomly selected from formulas which satisfied three constraints: 1) there were 100 or more isomers available within the publicly available GDB17 dataset (<https://gdb.unibe.ch/downloads/>); 2) the formulas contained only the elements C, H, N, O, and P; and 3) there were 100 or more structures available in PubChem. Python scripts were used to search GDB17 data and query the PubChem API in order to satisfy these constraints in September 2022. The 100 random structures were generated using the random molecular assembly code described above. Assembly values were generated in parallel using the AssemblyGo method and python scripts running on the Agave cluster at Arizona State University.

Discussion

Future life detection missions must have the capability to detect unknown biochemistry. Our work recommends that mass spectrometers used on spaceflight missions 1) have a mass resolution of at least 552,252 $m/\Delta M$; and 2) have a mass range of at least 861 Daltons. This mission specification will unambiguously distinguish high-MA (≥ 15) molecules, while a lower mass resolution of 435,714 $m/\Delta M$ is sufficient for detecting compounds with $MA \geq 20$. These recommendations are regularly achieved by terrestrial mass spectrometers but are roughly four times higher than that of proposed instruments. Recent engineering work by Arevalo et al (Arevalo et al., 2018) has been focused on developing ion trap mass spectrometers for space exploration that may dramatically increase the capabilities of future instruments and missions. We hope that this recommendation will serve as a benchmark for future instrument development focused on detecting high-MA compounds.

It is important to note that these recommendations, particularly those based on enumerated MOLGEN 5.0 formulas & randomly generated structures, make very few assumptions regarding the chemistry present in alien living systems. This is purposeful - there is a wide range of abiotic chemistry that is plausible elsewhere in the universe (Seager, 2010; Shaw, 2007), and an even larger scope of potential organic chemistry (Awale et al., 2017; Bains, 2004). Our mass spectrometry recommendations, based on random chemistry, should be treated as a useful null model for detecting complex compounds generated by life. While these recommendations have been derived within the context of assembly theory, the advantages of higher resolution across higher mass ranges would benefit many other research topics relevant to astrobiology and the broader planetary science and astronomy community (Arevalo et al., 2020; Chou, Grefenstette, et al., 2021). We additionally show that the proposed CORALS mass spectrometer can distinguish Earth-based biochemical compounds in a similar fashion to our recommended mass spectrometer, highlighting that lower resolution instruments are likely sufficient for detecting alternative chemistries that have a similar chemical space as biochemistry on Earth. However, our proposed mass resolution will ensure that complex compounds generated from living systems will be definitively detected and allow for unambiguous biosignature identification.

3: SOCIAL DYNAMICS SHAPE CHEMICAL INNOVATION

Introduction

"Life...is the greatest chemist, and evolution is her design process."(Arnold, 2019) This quote, attributed to Nobel laureate Frances Arnold, underscores the critical role that chemistry plays in the evolution of life on Earth. Over the course of 3.5 billion years, evolution has shaped the formation of novel and innovative chemical processes that underpin the functions of living organisms (E. Smith & Morowitz, 2016). However, with recent technological advancements, society has gained the ability to design and manipulate chemical reactions and compounds in ways that surpass the strict confines of biological evolution (Arnold, 2019; Derry & Williams, 1960; Judson, 2017). This revolution has placed today's chemical innovation at the intersection of science and society, presenting both opportunities and challenges for our understanding of the natural world and how society and technology contributes to its growing complexity.

Chemical innovation in modern society has been shaped by several competing forces, from scientific breakthroughs to patent protections to financial market pressures. Breakthroughs such as the synthesis of insulin or the development of nitrogen fertilizer in the early 20th century have yielded immense benefits to society (Gomollón-Bel, 2019; Smil, 2004; Vecchio et al., 2018) and importantly were not done in a scientific vacuum, but rather within a society where patents which grant sole rights to the patent-holder for usage and publication of an invention. Famously, the inventors of insulin synthesis - Fredrick Banting, James Collip, and Charles Best – sold their patent to the University of Toronto for \$1, believing it to be unethical to monetize such a discovery (Vecchio et al., 2018). This

is uncommon in chemistry, particularly for high-quality discoveries which hold potential for profits, encouraging authors to register patents (Anton & Yao, 2004; Hall & Harhoff, 2012; Moser, 2007). Today, it is estimated that as much as 77% of all novel chemical data is only found in patents instead of scientific publications and other literature (Bregonje, 2005; Senger et al., 2015). Novel compounds are more likely to be made publicly available in less time via patents than compared to scientific literature (Akhondi et al., 2019; Bregonje, 2005; Krallinger et al., 2017). Various uses of compounds are also more likely to be reported in patent literature (Bregonje, 2005), and many compounds are available only within patent-specific databases (Asche, 2017).

This amount of chemical data in patent literature makes patents the ideal data source to explore how society influences chemistry over time. There has been an increasing recognition and study of the influence of society on chemistry and science writ large - innovative science is performed as part of and in conjunction with pressures exerted by society & policy (Edler & Fagerberg, 2017; Owen et al., 2012; Ware, 2001). These pressures can take a variety of forms, such as the monetary value of patents, or author and organizational fame and recognition. Previous work has shown individual researchers are on average risk-averse, which potentially hinders their ability to develop novel inventions (Foster et al., 2015; Jia et al., 2017). Research output can also be predicted on the level of topics, where simple models have been developed which describe scientific research output within sub-categories of physics (Jia et al., 2017) and computer science, where funding strongly precedes research into specific topics (Hoonlor et al., 2013). Additionally, there

are modest correlations between the prestige of an academic institution and the scientific output by researchers there (Deville et al., 2014). At the level of patent classifications, which are given to a patent by the US Patent & Trademark Office (USPTO) and denote the type of invention, the majority of inventive effort is given to combining existing classifications rather than developing inventions which necessitate novel classifications (Strumsky et al., 2012). These results suggest a combinatorial approach to novel inventions (Youn et al., 2015), where new inventions are predominately driven by merging existing ideas.

Here, we measure the outcomes of this combinatorial, societally driven approach to patent chemistry innovation through network growth models and assembly theory (Figure 13). Network analysis over patent data allows for system-level time-series exploration into the creation and subsequent evolution of chemistry which is not possible from a snapshot of a database. For example, chemical patent databases often have high redundancy, where compounds are referenced many times across multiple databases (Yonchev et al., 2018), but high redundancy does not implicitly imply high usage. A high usage compound would be one found across many patents, such as a frequently used solvent like acetone. A network built on patent data can identify these highly used compounds through time-series connectivity measures where high-usage compounds are more connected to patents than sparingly used ones. A high connectivity over time suggests a high degree of importance for that compound. Networks are also useful for determining system-level trends, such as how the network grows over time (Broido & Clauset, 2019; Szymkuć et al., 2021). One

potential growth model would contain a system of chemical reactions which produces a single novel compound but utilizes commonly available substrates to produce that single novel invention. This example would suggest a preferential attachment model (Jeong et al., 2003; Newman, 2001) of chemistry evolution where common compounds are utilized with higher frequency than uncommon compounds to make novel chemistry, as opposed to more random explorations of chemical space (Barabasi & Albert, 1999; Reymond, 2015).

In addition to the combinatorial growth of patents, we study the evolution of patent chemistry complexity over time using assembly theory (AT, Y. Liu et al., 2021; Marshall et al., 2021). There are many metrics of molecular complexity, mainly for computational drug discovery, but all involve various theoretical or computational pitfalls (Méndez-Lucio & Medina-Franco, 2017). For example, structural measures – based on chiral centers, molecular weight, or compactness – are relatively simple to compute, but only consider a single measure of a compound rather than a holistic view (Sheridan & Kearsley, 2002). More complex measures, such as graph-based measures which consider subgraph counts (Bertz, 1981), quantum mechanics (Luzanov & Babich, 1995), or information theory (Böttcher, 2016) among many other possible factors (von Korff & Sander, 2013), ultimately lack correlation with experimental data, as well as with each other (Méndez-Lucio & Medina-Franco, 2017). As a complexity measure, AT is uniquely experimentally verifiable. Originally designed to discover biomolecules built by living systems elsewhere in the universe (Chou, Grefenstette, et al., 2021; Marshall et al., 2017), AT considers the amount of information required to build a molecule from the space of possible chemistry

available to create a range of complexity values from low to high, where high values represent compounds which are combinatorically nearly impossible to create without living systems (Marshall et al., 2021; Sharma et al., 2022).

Specifically, the measurement of the *molecular assembly index* (MA) of a compound is done through calculating the fewest possible joining steps necessary to build that compound from basic component bonds. It is important to note that while these joining steps do not correspond to actual chemical reactions used in synthesis, the MA – an integer value representing the number of steps in the theoretically shortest path to build a full compound - does correspond to fragmentation data found through mass spectroscopy (Marshall et al., 2021), providing a physical grounding and experimental verification of MA as a measure of a molecule's complexity independent of the route of synthesis. The shortest path is calculated through a graph-based approach where sub-graphs are recursively merged to create the full graph of the compound at hand. The intermediate sub-graphs created along this recursive process can be re-used, allowing symmetrical structures to have a lower MA than non-symmetrical ones. The details of the algorithm and software implementation can be found in the SI. When applied to time-series patent chemistry data, the MA of compounds over time give an agnostic measure of the changing complexity of chemistry, meaning MA does not depend on the details of how the molecules are synthesized, only how complex they are. This is an improvement over previous human-derived measures of measuring complexity over chemical patents (Szymkuć et al., 2021),

as there could be biases introduced through classifications which are not present in assembly theory.

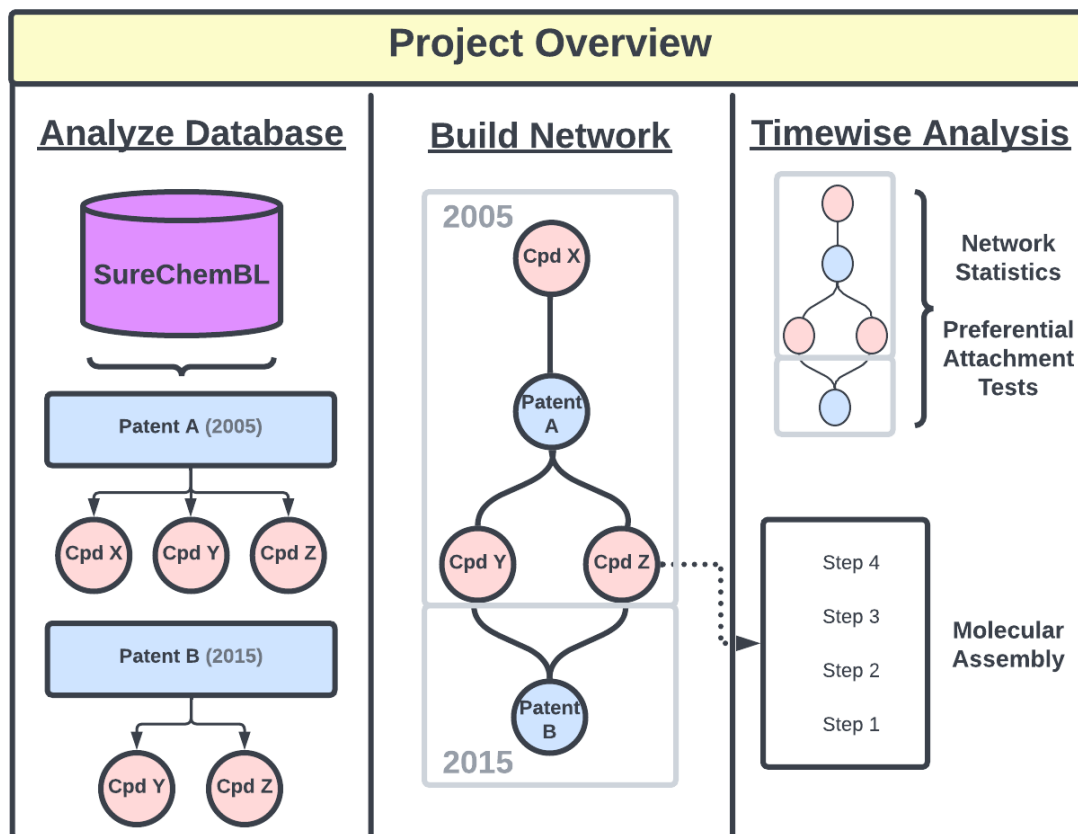


Figure 13: Assembling chemical networks from the SureChemBL database over time, calculating preferential attachment, and calculating assembly indices.

Results & Methods

Network Statistics

The SureChemBL database contains patent chemical data from 1962 through 2020, containing chemical data from the US, European, and Japanese Patent offices. The database has roughly 17 million unique compounds, with data from textual analysis since 1976 and from image recognition since 2007 (Falaguera & Mestres, 2021; Papadatos et al., 2016). We use 1976 as a starting point for the analyses in this project to consider patents with a full set of compounds. While larger publicly available chemical databases exist (Gaulton et al., 2017; S. Kim, Thiessen, Bolton, et al., 2016; Reymond, 2015), SureChemBL uniquely connects chemical structures and reactions with timestamped patent literature. This database is incomplete, as it is missing patents from China (the leading patent-developing country (Christodoulou et al., 2018)), among many others. Additionally, it is likely biased due to intricacies of patent law - in the US, for example, patents have to demonstrate a “significant, immediate, and well-defined use”, which can be difficult to prove in terms of chemistry (Seymore, 2013), while prior to 1988 Japanese patent law required multiple patents for multi-step inventions (Sakakibara & Branstetter, 2001). This wealth of metadata across such a large publicly available database makes SureChemBL a leading source of information for tracking large-scale chemical usage and innovation over time.

We built a bipartite, undirected network from SureChemBL patent data (Papadatos et al., 2016) from 1980-2020, inclusive, using the igraph python library (Csardi et al., 2006). A bipartite network is where two types of objects (nodes) are connected via some relationship

to form edges. We added both patents and chemical compounds as nodes, with edges defined as links between a patent and all compounds listed as used within that patent. These edges are undirected, as there is no causal influence between the patent and the compounds or vice versa. New patents are added to the full network and are connected to either existing compounds that have been used in prior patents or to newly added compounds. These new compounds may or may not be newly invented compounds - the task of finding novel compounds and distinguishing between these and previously known compounds is difficult (Falaguera & Mestres, 2021) and beyond the scope of this project, as we are interested in how the use of various chemical compounds changes over time within the patent record.

There are 551,235 compounds and 70,772 patents listed in SureChemBL before 1980, and these nodes and the edges between them were considered as the base network prior to adding compounds and patents from January 1980. We calculated a variety of network statistics over the evolving patent-compound network to highlight the characteristics of social chemistry growth. There is a large increase in new compounds added per month in 2008, when SureChemBL added image recognition to its database. Image recognition allowed chemical compounds which were only described visually in patents to be added to the database. Prior to 2008, only compounds which were explicitly written within patents and could be labeled using textual analysis were included. We also calculated the average compound degree (**Error! Reference source not found.**) - the average number of patents containing a randomly selected compound – which decreases exponentially across the entire time series. The average patent degree (**Error! Reference source not found.**) - the

average number of unique compounds within a randomly selected patent record - increases over time before leveling off in the early 2010s. There is an increase in average patent degree in 2008, suggesting that image recognition added roughly 10 compounds per patent. Additionally, the data suggests that the diversity of compounds within patents increased with the addition of image recognition, as there is a large increase in the number of novel compounds added per month in 2008 (Figure 14).

$$\underline{k} = \frac{1}{N} \sum_{i=1}^N k_i \quad (8)$$

The largest connected component of the network - the maximum number of nodes which are connected by edges - is also found using the igraph python library (Csardi et al., 2006). Figure 15 describes the number of nodes (compounds and patents) found within this largest connected component compared to the number of compounds and patents in the full network discovered at a given month.

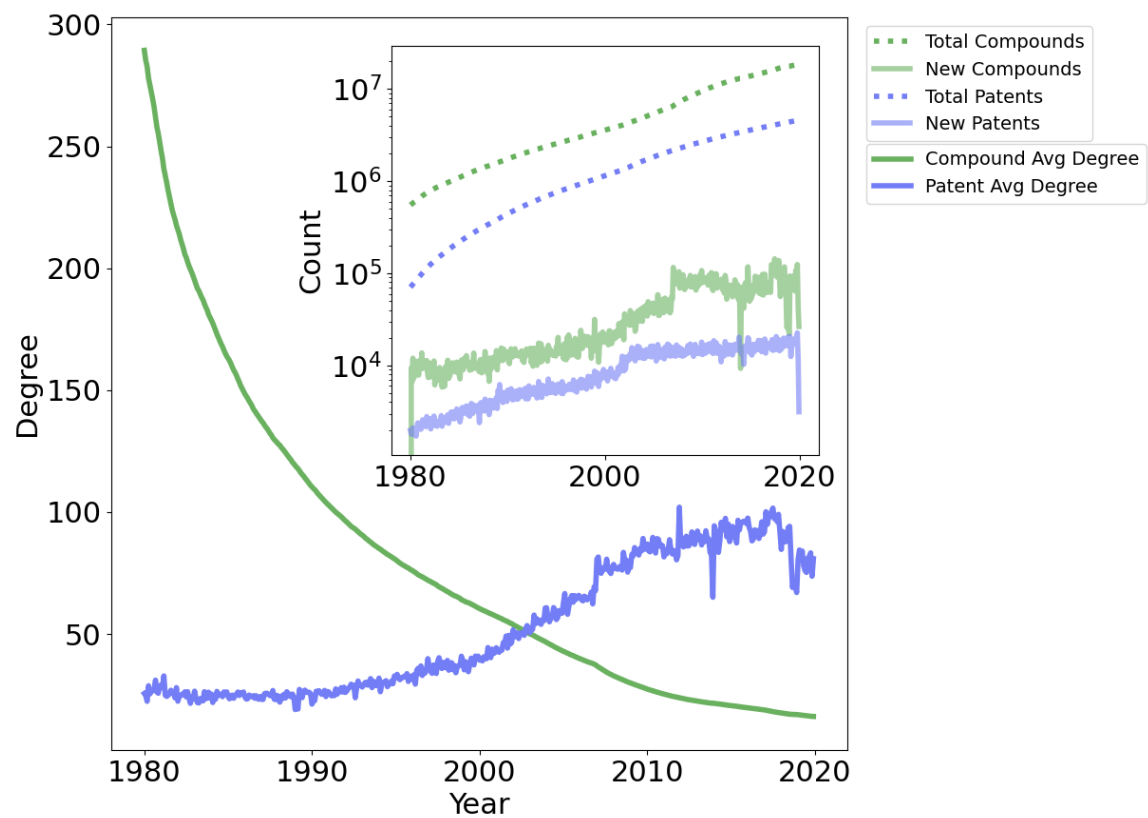


Figure 14: Network statistics over SureChemBL patent chemistry.

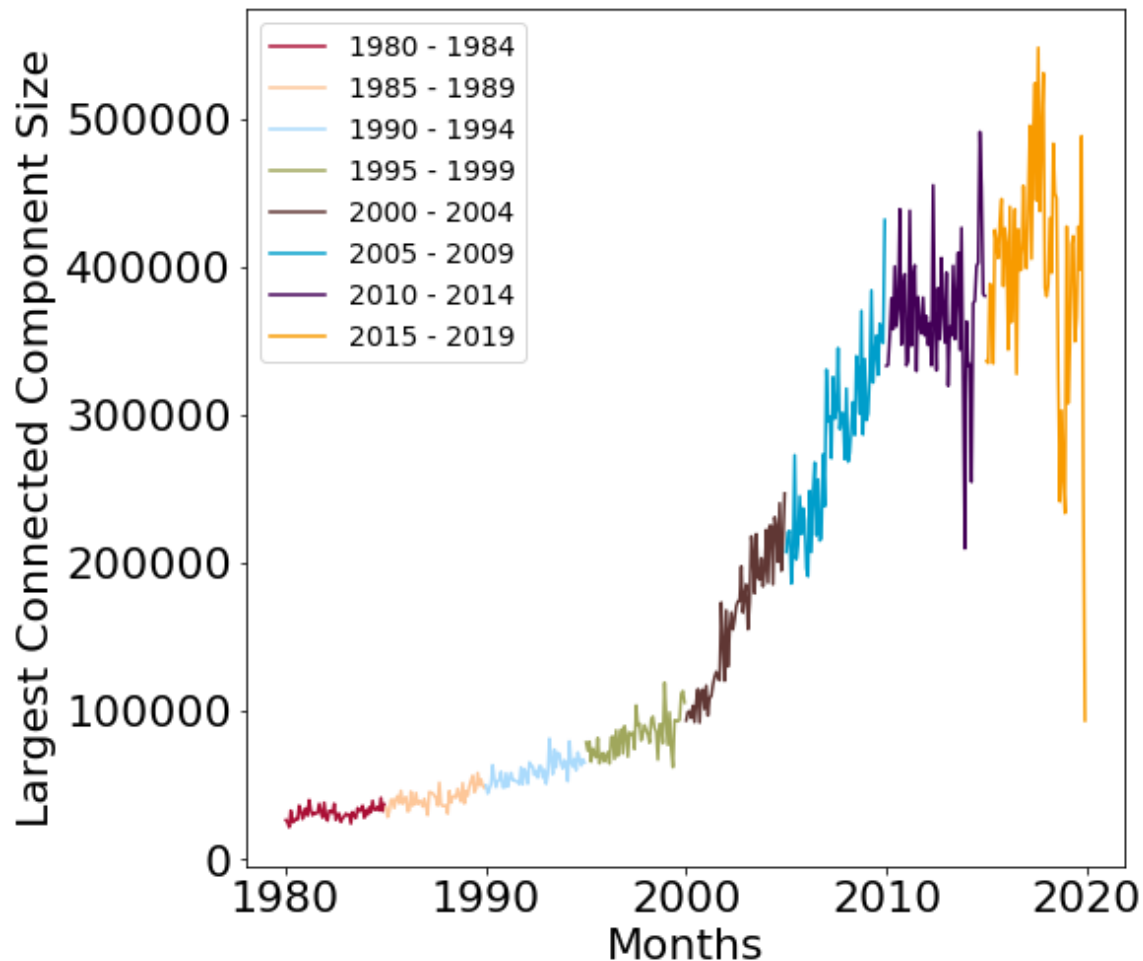


Figure 15: Largest connected component Size (patents + compounds)

The average clustering coefficient is also calculated for each node (compounds and patents) (Figure 16). The average clustering coefficient of all nodes N is calculated through **Error! Reference source not found.**, where k_i are the neighbors of node i , and L_i is the number of edges between all neighbors.

$$\underline{C} = \frac{1}{N} \sum_{i=1}^N \frac{2L_i}{k_i(k_i - 1)} \quad (9)$$

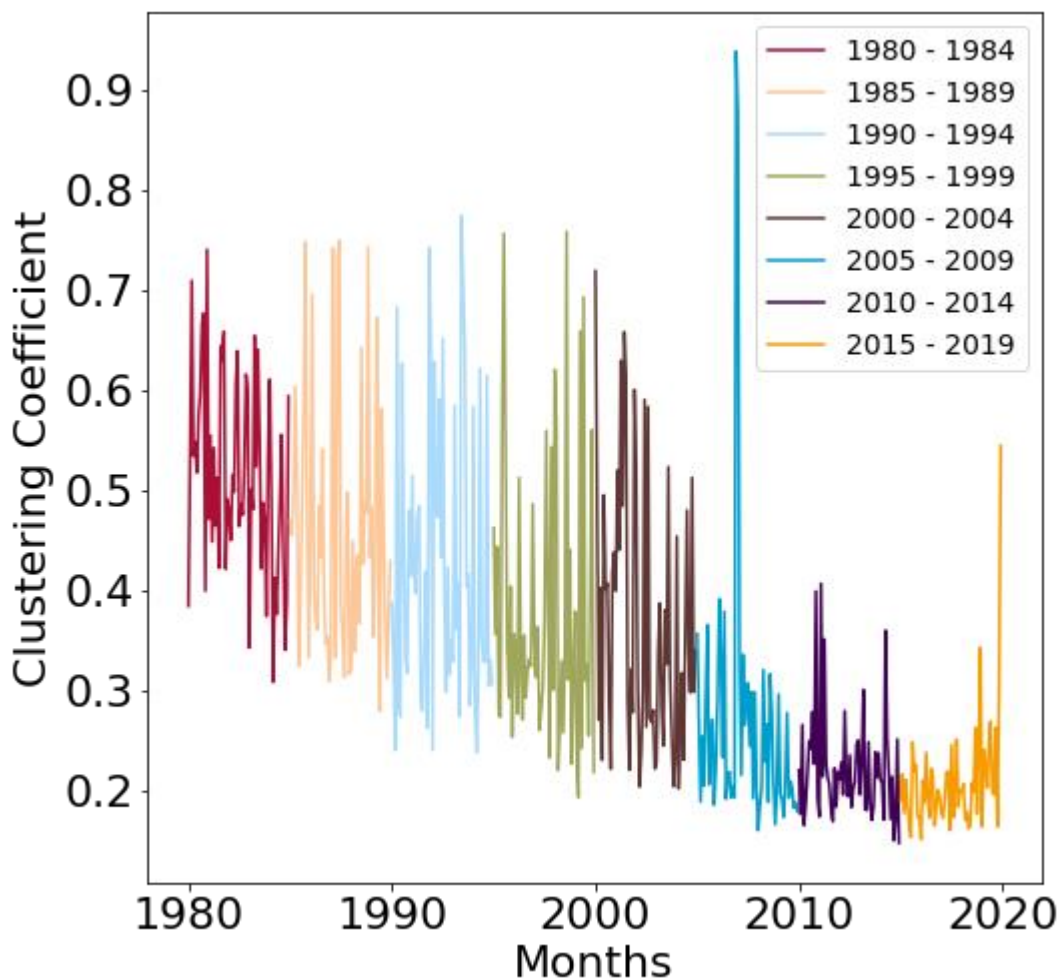


Figure 16: Average clustering coefficient

We also tested the underlying growth equation of both compounds and patents. The exponential growth function describes the number of total compounds and patents with a small r -squared value. The exponential fits were generated using the `numpy.polynomial.polynomial.polyfit` method, which generated a sum of squares of the regression (SSR) value. The total sum of squares (SST) was generated for the total number of patents and compounds, with a r^2 value (SSR / SST) of $3.30e-10$ for compounds and $3.49e-10$ for patents (Figure 17, Figure 18).

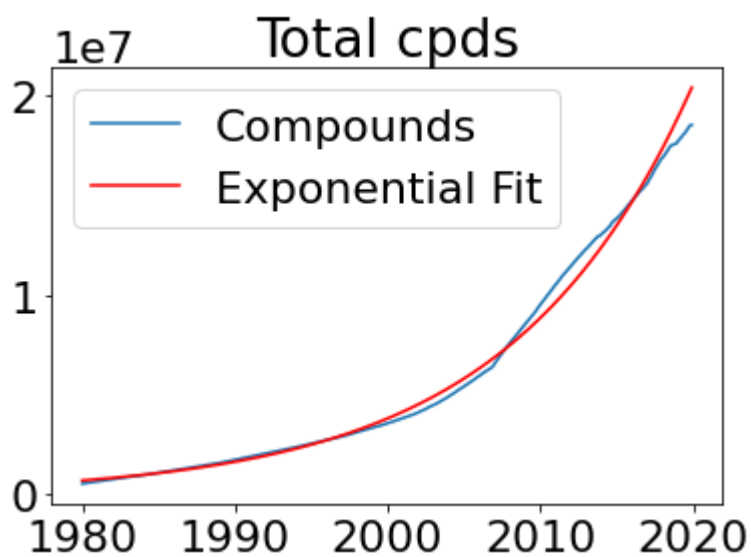


Figure 17: Compound exponential growth fit. The total number of compounds (in units of 10^7 individual compounds) are listed on the x-axis.

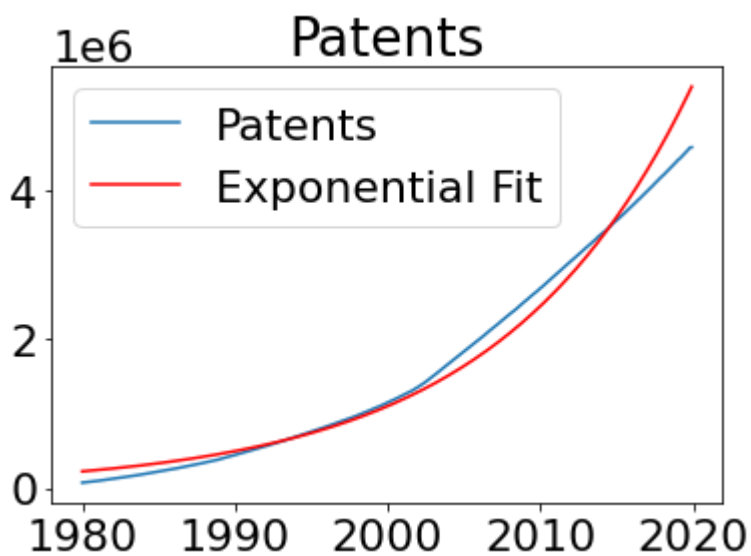


Figure 18: Patent exponential growth fit. The total number of patents (in units of 10^6 individual patents) are listed on the x-axis.

Preferential Attachment

We calculated the attachment of patents to compounds in 5-year intervals from 1980-2019 (Figure 19), as well as for the full 40-year time series (Figure 20), which measures the average number of patents using a given compound within each interval. For each compound, we found the degree of every compound at every month. The compound degree measures how many patents contain a given compound. The cumulative sum of degrees at each month across the 5-year period divided by the 60 (the number of months in one 5-year interval), returns the average preferential attachment index.

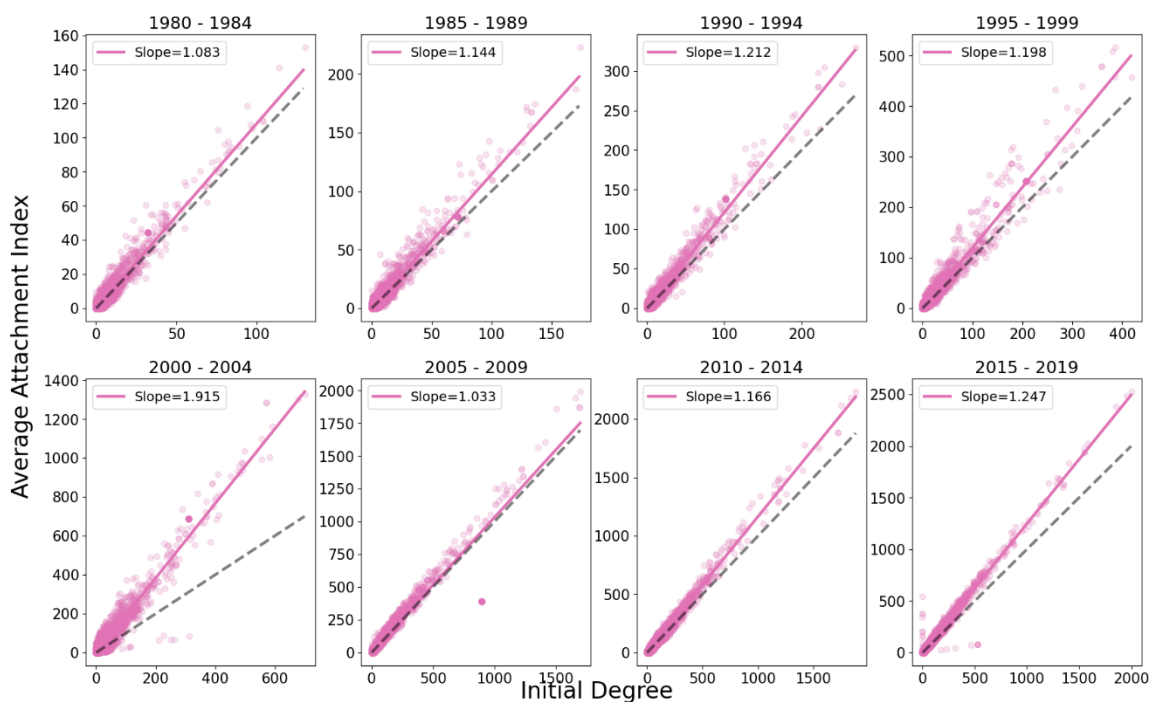


Figure 19: Preferential attachment indices across all SureChemBL data in 5-year increments.

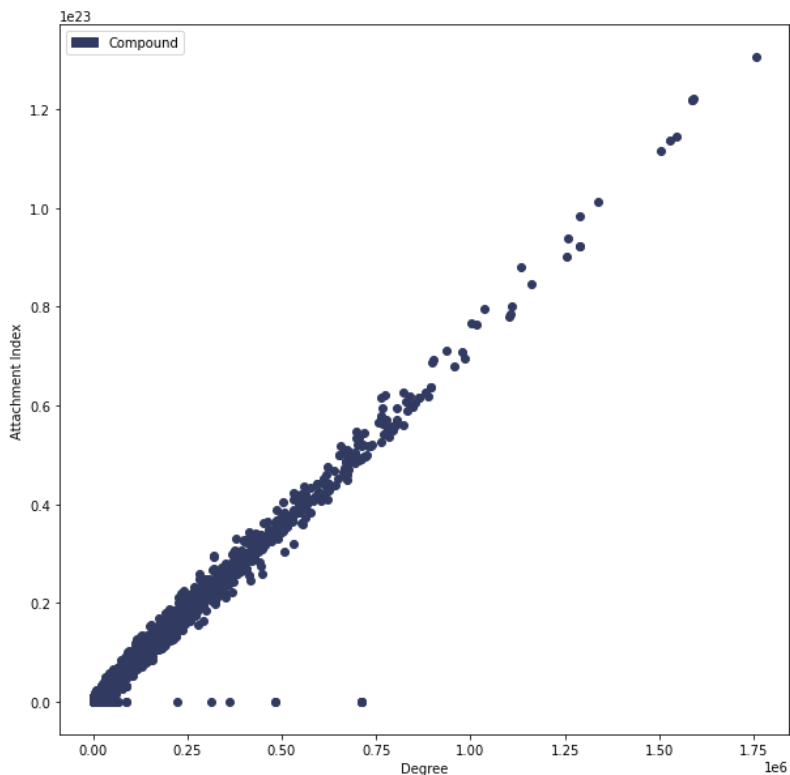


Figure 20: Full preferential attachment, 1980-2019.

The preferential attachment model is a statistical phenomenon of degree distributions within networks (Barabasi & Albert, 1999), as well as a growth model which describes how new nodes are added to existing networks over time (Jeong et al., 2003). Here, a preferential attachment model means new patents would use popular, pre-existing compounds with power-law frequency. That is, it is much more likely that new patents would use those compounds which are already present in a large number of patents than compounds which are rarely found in the patent literature. In a linear preferential attachment model, this likelihood of choosing a highly-connected compound decreases as a power-law function, as demonstrated through the Barabasi-Albert model of network growth (Barabási, 2013).

In this model, based on preferential attachment, the likelihood $\Pi(k)$ of a new node connecting to a random node i is based on the degree of node i , k_i , and the total degree of a network with N nodes (**Error! Reference source not found.**).

$$\Pi(k_i) = \frac{k_i}{\sum_j^N k_j} \quad (10)$$

In the theoretical preferential attachment model, the degree of each node increases according to (**Error! Reference source not found.**), where t is the timestep where the node k_i is added to the network, and $m(t/t_i)$ is the number of edges added to node k_i between time t and t_i .

$$k_i = m \left(\frac{t}{t_i} \right)^{\frac{1}{2}} \quad (11)$$

This equation predicts that the degree for every node in the network at hand increases at the same exponential rate (here, $1/2$) and that degree growth is sublinear, since newly added nodes have more options with which to connect to. It also predicts that earlier-added nodes will have higher degrees than later-arriving nodes ((Barabási, 2013).

An experimental test (Barabási, 2013; Redner, 2005) of preferential attachment is also performed. This test has four algorithmic steps and was calculated for all 5-year periods within the SureChemBL compound-patent network: 1980-1984, 1985-1989, 1990-1994,

1995-1999, 2000-2004, 2005-2009, 2010-2014, and 2015-2019 (all inclusive). For each step, we work through the calculation on a specific SureChemBL compound (SureChemBL229199) for the period between 1980-1984.

Step 1

The number of patents utilizing each compound present in the SureChemBL network are found at every timestep. Timesteps in this case are individual months within the five-year period, so the first timestep is January of the first year considered. There are 60 total timesteps within each period. This is done for all compounds found within the entire five-year period, so if no patents utilize a given compound during a specific timestep, that particular time step is listed as 0 in our analysis. For SureChemBL229199, these results should be interpreted as in January 1980, there were 5 patents which used this compound. In February 1980, there was only 1 patent using it, and so on through December 1984, where 0 patents used it.

SureChemBL229199 results: [5, 1, 3, 1, 2, 2, 0, 0, 1, 1, 1, 0, 1, 0, 2, 2, 4, 0, 1, 0, 1, 1, 0, 1, 1, 1, 3, 0, 0, 2, 1, 0, 2, 2, 0, 0, 1, 0, 1, 1, 0, 2, 1, 0, 0, 1, 2, 0, 1, 0, 1, 3, 1, 3, 1, 0, 2, 2, 1, 0]

Step 2

The cumulative number of patents referencing each compound is found at each timestep. This shows the cumulative degree of each compound throughout the time period, as the edges between compound and patent do not disappear once created in the network structure. The cumulative sum is built using the numpy python library (Harris et al., 2020).

For SureChemBL229199, the degree in January 1980 is 5, the same as in Step 1. The degree in February 1980 is 6 (5 + 1), and the final degree of the compound in December 1984 is 66.

SureChemBL229199 results: [5, 6, 9, 10, 12, 14, 14, 14, 15, 16, 17, 17, 18, 18, 20, 22, 26, 26, 27, 27, 28, 29, 29, 30, 31, 32, 35, 35, 35, 37, 38, 38, 40, 42, 42, 42, 43, 43, 44, 45, 45, 47, 48, 48, 48, 49, 51, 51, 52, 52, 53, 56, 57, 60, 61, 61, 63, 65, 66, 66]

Step 3

The attachment index, a , is calculated through finding the difference in degree (number of cumulative patents) at every timestep, then averaging these differences (**Error! Reference source not found.**). Here, t is the number of timesteps, and k is the degree at each timestep t .

$$a = \frac{1}{t} \sum_{i=0}^t (k_{i+1} - k_i) \quad (12)$$

SureChemBL229199 results: 1.0338983050847457

Step 4

The attachment index a is calculated for every compound present in the SureChemBL network within the time period analyzed. These results are then graphed against the initial degree - the number of patents using that particular compound within the first month of the analysis. A linear regression model is used to determine if the results experimentally

demonstrate preferential attachment. A slope of 1 or higher demonstrates preferential attachment, as it shows those compounds which have a high initial degree have a linearly (or superlinearly) related attachment index.

In theoretical preferential attachment models, there is a distinction between “internal links”, which are edges connecting two nodes which previously exist in a network, and “external links”, which link a new node to either another new node or a previously existing one (Jeong et al., 2003). In this patent network, there are no internal links, as the only possible new connections which can be made are from newly registered patents to either new or existing compounds. The preferential attachment network is calculated using compounds for this reason - once added to the network, patents cannot make new nodes. However, if a compound is used in multiple patents, then the degree of the compound increases. As we are interested in the changing usage of chemical compounds over time, we calculate preferential attachment using changes in compound degree over time.

The average attachment index is calculated through averaging the results of **Error! Reference source not found.** for every compound present within a given period. The standard deviation and standard error were also calculated for each 5-year period (Table 2).

Table 2: Average Preferential Attachment Values

Year	Average Attachment	Standard Deviation	Standard Error
1980-1984	0.0783	0.9575	0.001116
1985-1989	0.0879	1.2781	0.001305
1990-1994	0.1061	1.8071	0.001601
1995-1999	0.1312	2.6399	0.002046
2000-2004	0.1977	5.3109	0.003203
2005-2009	0.1614	5.1401	0.002149
2010-2014	0.1565	5.1103	0.001886
2015-2019	0.1660	5.5983	0.002011

In a degree/attachment index plot, such as those in Figure 19 and Figure 20, preferential attachment is distinguished by a linear or superlinear relationship between initial degree (x-axis) and average attachment index (y-axis). We demonstrate the linear-to-superlinear relationship (calculated by linear regression minimizing the chi-squared error between all data points within a 5-year increment) between initial degree and the average attachment index of a compound, denoting preferential attachment. This is additionally confirmed by power-law analysis of the compound degree distribution within each time frame (Figure 21, Table 3).

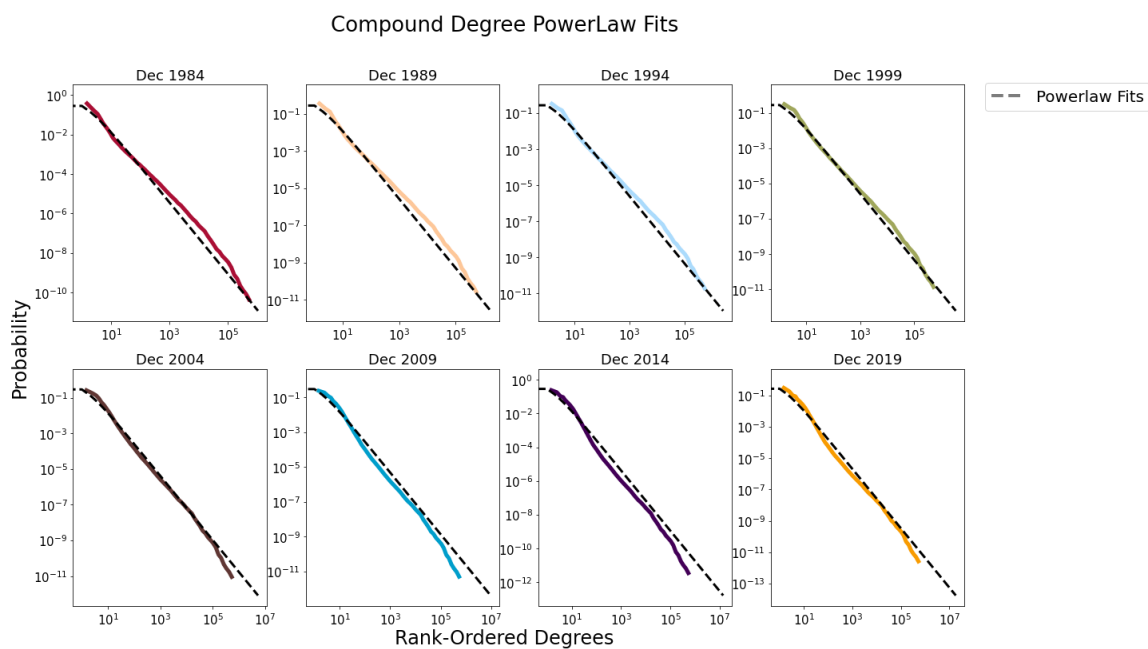


Figure 21: Compound degree distribution power law fits

Table 3: Compound Degree Power Law Fit Statistics

Month	Power Law Alpha	Power Law Sigma	Xmin	Exponential R	Exponential p-value
1984-12	1.8099	0.0007768	1.0000	3964938.268	0.0000
1989-12	1.8562	0.0006501	1.0000	5857216.654	0.0000
1994-12	1.8687	0.0005456	1.0000	7828779.998	0.0000
1999-12	1.8645	0.0004580	1.0000	9929113.35	0.0000
2004-12	1.7959	0.0003429	1.0000	12161916.29	0.0000
2009-12	1.7734	0.0002516	1.0000	16352054.54	0.0000
2014-12	1.7905	0.0002128	1.0000	20427204.44	0.0000
2019-12	1.8935	0.0002075	1.0000	27602685.79	0.0000

A preferentially built, scale-free network represents a small number of compounds being found in many patent records, while nearly all compounds are found in very few patents. This is not entirely surprising, as some compounds are extremely common across chemistry

(e.g., acetone, ethyl acetate, and hexane (Joshi & Adhikari, 2019)) and would be expected to be found within many patents. This growth model of preferential attachment also follows various other socially created networks (Kunegis et al., 2013; Pham et al., 2015), adding to the literature of networks based on society following preferential attachment growth models and resulting in scale-free networks.

Tracking Preferential Attachment Across Compound Classes

We are particularly interested in classes of compounds which are utilized more often than expected in patents and can provide context to trends in social chemistry, such as researchers and companies using certain compounds more often due to social influence. We do not focus on common solvents and other compounds which are prevalent in laboratories - these are so ubiquitous that they provide little information about social pressures within chemistry. Rather, we first explored compounds which highlight changing influences and innovation using preferential attachment. Those compounds which have increasing preferential attachment scores than expected show an outsize influence on chemistry than initially shown. The attachment index is the result from **Error! Reference source not found.**, and individual compounds of interest are found from literature. The SureChemBL dataset uses InChI representation for chemical identification, so when necessary, PubChem (S. Kim et al., 2019) is used to find the InChI representation of compounds of interest.

We tested various sets of compounds, including psychedelic drugs (Nutt, 2019) (Figure 22) and SARS/HCV drugs (Elfiky & Ibrahim, 2020) (Figure 23). We also tested green

solvents (Pacheco-Fernández & Pino, 2019), which exhibited a much higher than expected usage (Figure 24). These solvents are a result of a decades-long effort to reduce the use of toxic organic solvents, with notable successes such as sodium dodecylsulphate (SDS), an amphiphilic solvent used in tandem with magnetic nanoparticles (Qi et al., 2016). Between 2015 and 2019, over 1400 patents per month used SDS in some fashion, up from under 100 patents per month in the early 1980s. The attachment index of SDS in 2019 is 700,000 times the standard error of the average attachment index (Table 2). The use of preferential attachment to detect social trends such as this effort to increase the usage of green solvents such as SDS and others like 1-hexanol (Shen et al., 2020), 1-octanol (Chong et al., 2018), and decanoic acid (Florindo et al., 2019) in comparison with a growing environmental movement in chemistry (de Marco et al., 2019; Płotka-Wasyłka et al., 2018) shows the power of large-scale network analyses of social chemistry to observe and quantify trends.

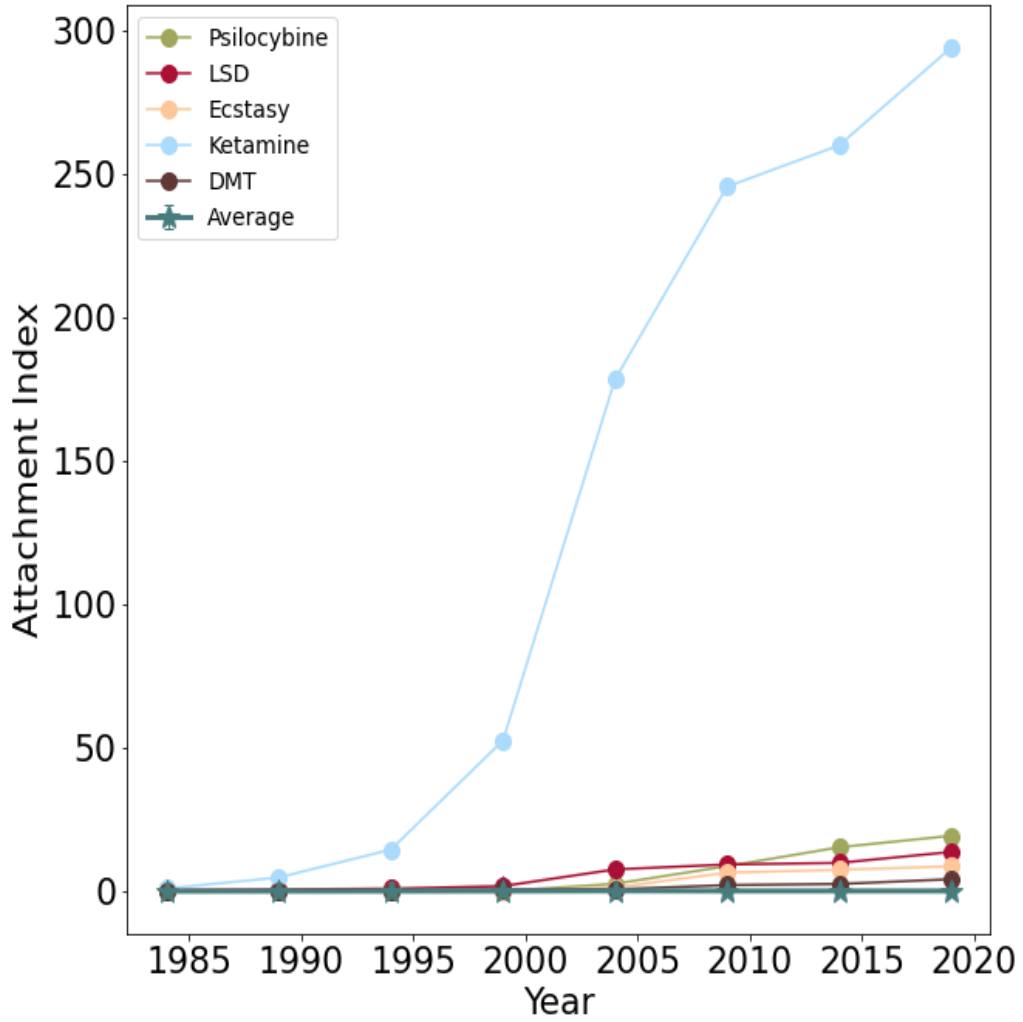


Figure 22: Psychedelic drug attachment over time.

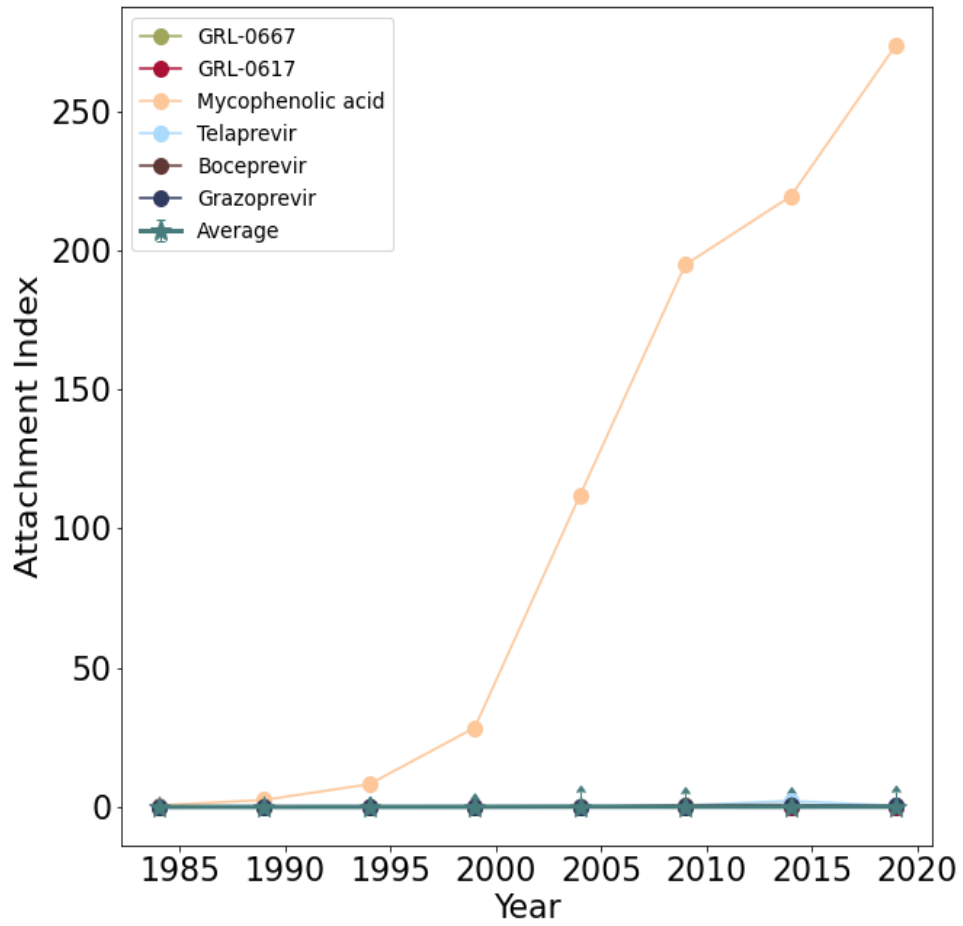


Figure 23: SARS/HCV drug attachments over time.

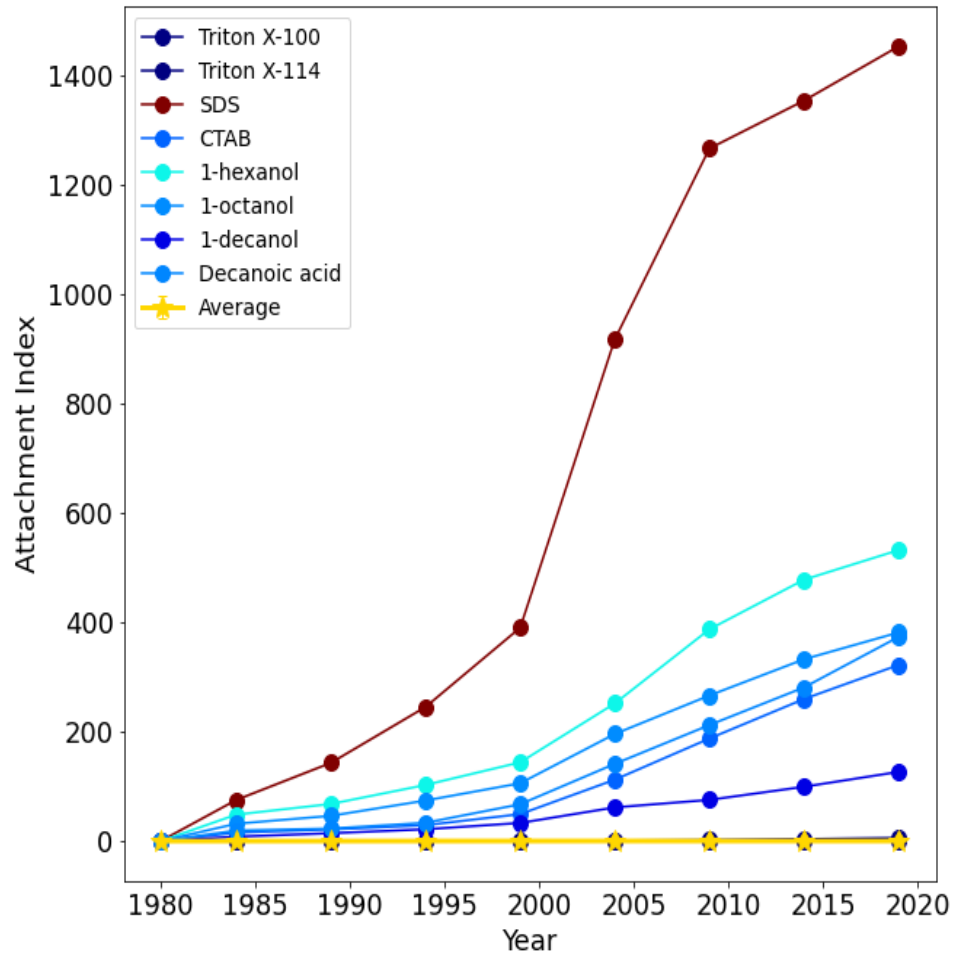


Figure 24: Green solvent attachment.

Molecular Assembly

We measured the assembly index of randomly sampled novel SureChemBL compounds within each month from 1976 to 2020. The assembly number of a compound is generated using four computational steps. First, a chemical structure is converted into a graph-based representation, using the RDKit cheminformatics python library (Landrum, 2020). This structure is then randomly fragmented into component parts, creating the fundamental building blocks of the graph structure. In chemical terms, these fundamental components are two or three atoms connected by bonds, all of which correspond to atom-atom pairings present within the compound. These fragments are then iteratively joined together, with intermediate structures and component parts available for re-use, to create the minimal assembly path of the initial graph. The number of steps in this minimal assembly path is the assembly index of the chemical compound (Marshall et al., 2021). For all analyses shown here, we used the AssemblyGo calculator (<https://gitlab.com/croningroup/cheminformatics/molecular-assembly>), which calculates the MA using a split-branch algorithm, with a timeout of 300 seconds. When this timeout is reached, the split-branch algorithm returns the best possible MA value found at that time. All MA values were calculated using the Agave supercomputing cluster at Arizona State University.

The novel fragments are built from the MA building process. A fragment is an intermediate step within an assembly pathway. These are graph-based fragments which represent partial chemical structures, which may or may not be physically plausible. Taken together, fragments give a measure of the diversity of assembly pathways and overall structures

within a dataset. Novel fragments are those which have not been utilized in prior assembly pathways, representing new substructures which have not been used before and providing a loose measure of diversity within patent chemistry. We calculated novel fragments from a randomly sampled set of 1000 compounds per month between 1980 and 2020 and built the cumulative number of fragments from the summation of these novel fragments. We find the number of novel fragments decreases and eventually stabilizes over time, suggesting the rate of finding novel structures has stagnated at best.

Molecular Assembly Compound Sampling

In total, we sampled 2,033,834 compounds from SureChemBL. These compounds came from two sampling runs – one based on random sampling within individual months, and the other based on random sampling of patent authors and assignees.

The first sampling run found roughly 960,000 compounds. We sampled 2000 compounds per month from 1980-2020, inclusive. The sampling steps are outlined below:

1. Within each set of 2000 compounds per month, 1000 compounds (full database compounds) were taken entirely by random from the set of compounds which had been added to the database by that particular month.
 - a. For example, a random set of 1000 compounds taken in July 1993 (the birth month of the author) would only include compounds listed in patents prior to July 1993.
2. The other 1000 compounds (the novel set) are compounds that are randomly sampled from those added to the database for the first time in that particular month.

- a. Using the July 1993 example again, there were roughly 10000 compounds added in July 1993 that were not previously found in SureChemBL. The novel compound set was randomly taken only from that set of 10000.

The second sampling run was performed for the patent authors and assignee analysis. In total, we obtained 1053834 compounds. These sampling steps are also outlined below.

1. We randomly sampled 10000 patents from SureChemBL, with no date or compound restrictions.
2. From those 10000 patents, we found all authors (21947 total) and assignees (6165 total) associated with those patents.
 - a. We removed all non-company assignee records, resulting in 1950 total assignees.
 - b. Assignee filtering was done by removing all assignees which did not include at least one of the following key terms: “Corp”, “Inc”, “Co”, “Ltd”, “Llc”, “Lllp”, “Rlllp”, “Corporation”, “Incorporated”, “Limited”, “Company”, “Univ”, “University”. This was done so that authors were not co-listed and therefore co-analyzed within assignees.
3. We found all patents associated with each of the 21947 authors and 1850 assignees discovered in step 2.
 - a. In total, this resulted in 7056912 patents associated with authors and 667165 patents associated with assignees.

4. In order to make MA calculations tractable, we randomly sampled 100000 patents from the 7056912 patents associated with authors and found all unique compounds associated with those patents.
 - a. In total, there were 1034394 compounds associated with authors.
5. We also randomly sampled 1000 unique assignees and found 1751784 additional unique compounds.

Molecular Assembly, Molecular Weight, and Novel Fragment Trends

For all sampled compounds, we found the earliest date of entry in SureChemBL. We use this as a proxy for invention – while we do not know if this represents the exact date of invention for each compound, it provides an estimate of when a compound was added to the growing patent record of chemistry. Figure 25 shows the increases of MA and molecular weight (in Daltons, calculated using RDKit) over time, and here we show the linear regression fits over time. Prior to 1980, the sample size of the compounds is much lower compared to future years and does not contain novel compound sampling. Additionally, prior to 1972, only 10 compounds are present, skewing the graph without adding much sampling power.

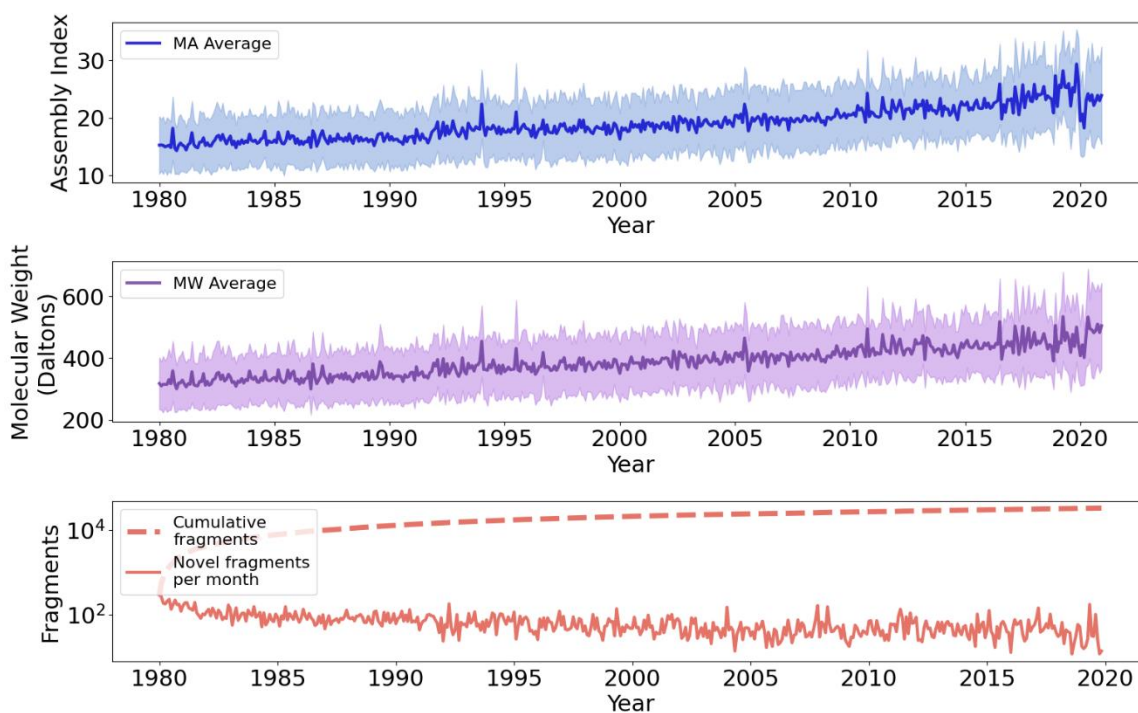


Figure 25: Changing chemical properties (assembly indices, molecular weights and fragment diversity) of compounds from 1980 - 2020.

The average MA of compounds increases linearly from 8.857 in January 1980 to 23.891 in December 2020, showing for the first time an agnostically measured increase in complexity of patent chemistry. Additionally, the average molecular weight of the compounds increases linearly over time from 231.061 to 504.431 Daltons. The linear regression r^2 values (0.933 for MA over time and 0.928 for molecular weight over time) show that both MA and molecular weight increase linearly over time (Figure 26 and Figure 27).

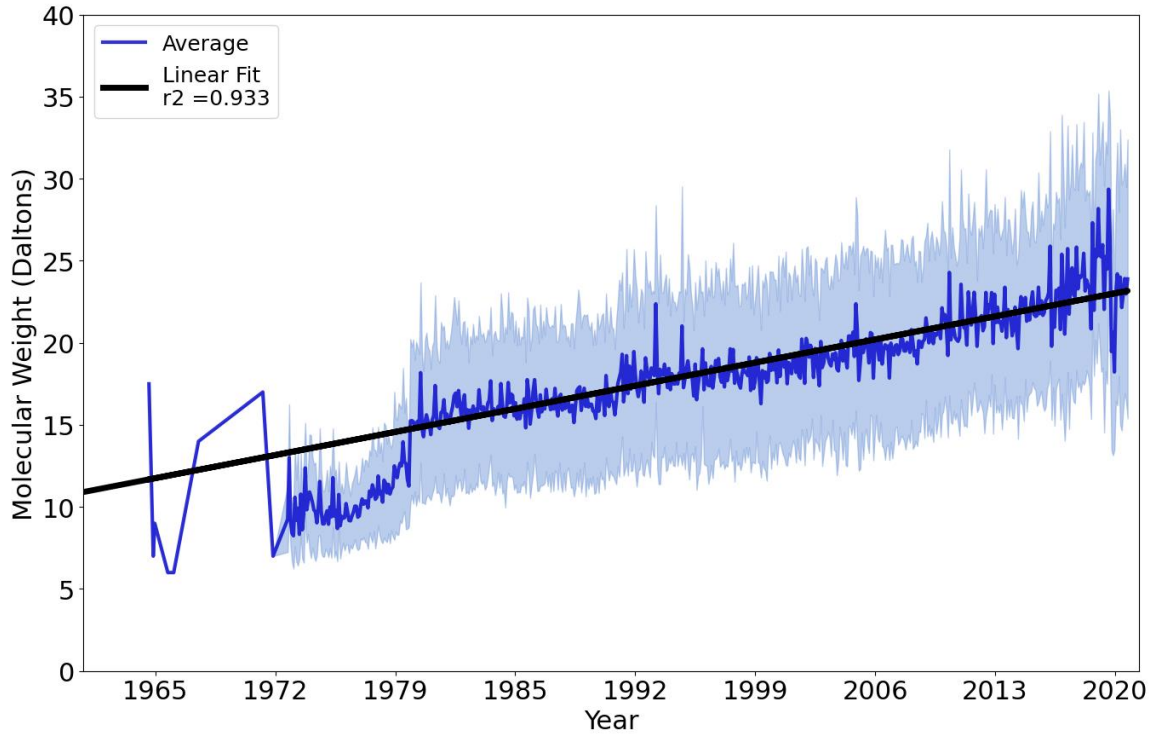


Figure 26: MA over time with linear fit

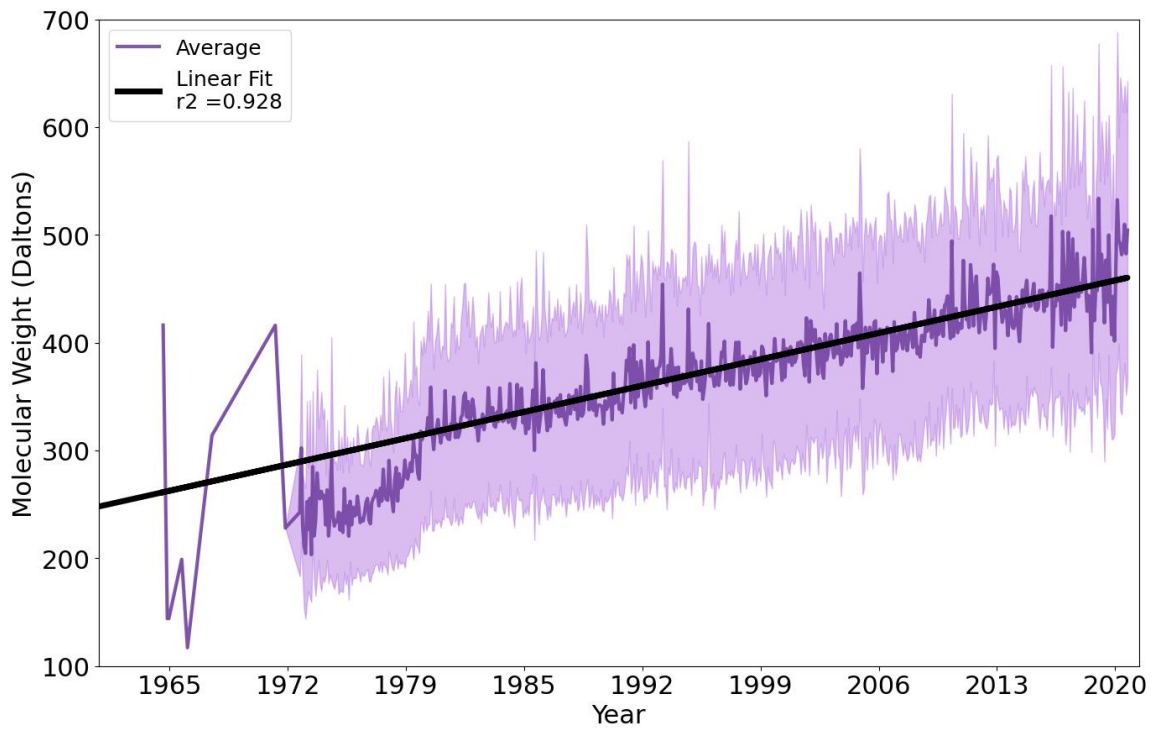


Figure 27: Molecular weight (Daltons) over time with linear fit

We are also particularly interested in how MA can be correlated with various physical molecular characteristics, as this would ground the theoretical MA measurement to the physical world. There has been observed correlations with MA and mass spectrometry spectra output, as well as with molecular weight (Marshall et al., 2021). However, the molecular weight correlation was performed with an outdated version of the MA code which estimated the MA of various compounds. Here, we use the most recent version of MA calculations, and only include compounds for which the exact MA value was computed. In total, there are roughly 600,000 compounds which have exact calculations. We computed the spearman correlation coefficient (we assumed non-gaussian distributions) between the MA of these compounds and their molecular weight (in Daltons) and number of non-Hydrogen bonds. Both molecular values were calculated using RDKit. The spearman coefficient of the MA-Molecular Weight correlation was 0.699, while the MA-Bonds correlation was 0.77 (Figure 28, Figure 29). Both show a strong positive correlation, with bonds having a tighter correlation with MA values.

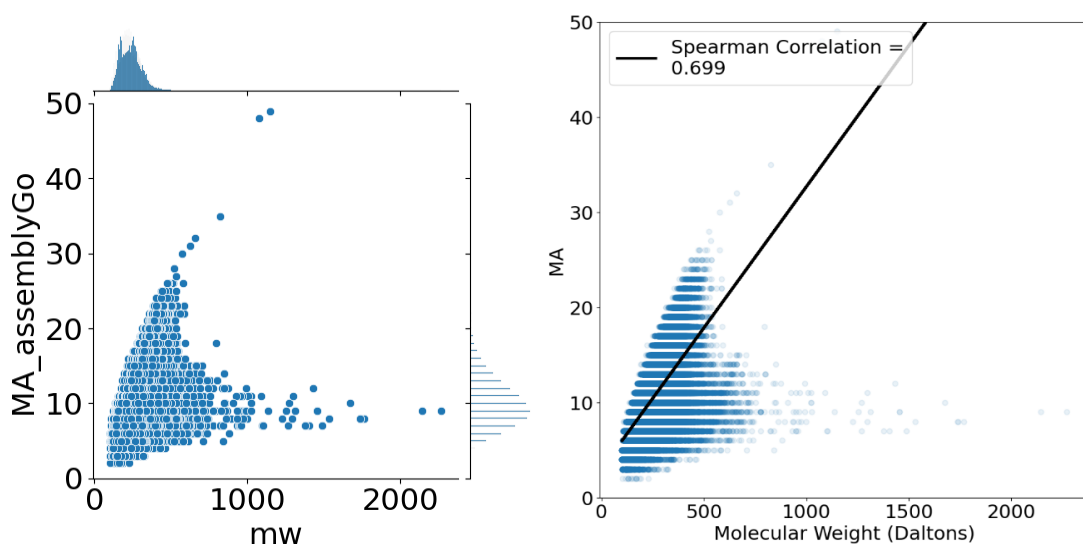


Figure 28: MA / molecular weight correlation

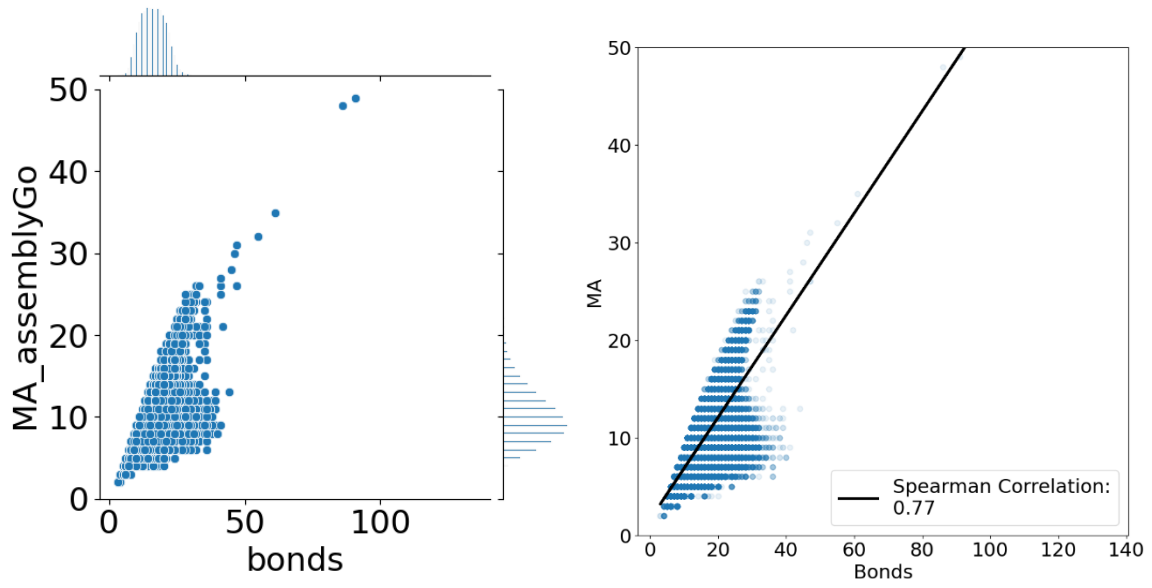


Figure 29: MA / number of bonds correlation

Social Factors and Molecular Assembly

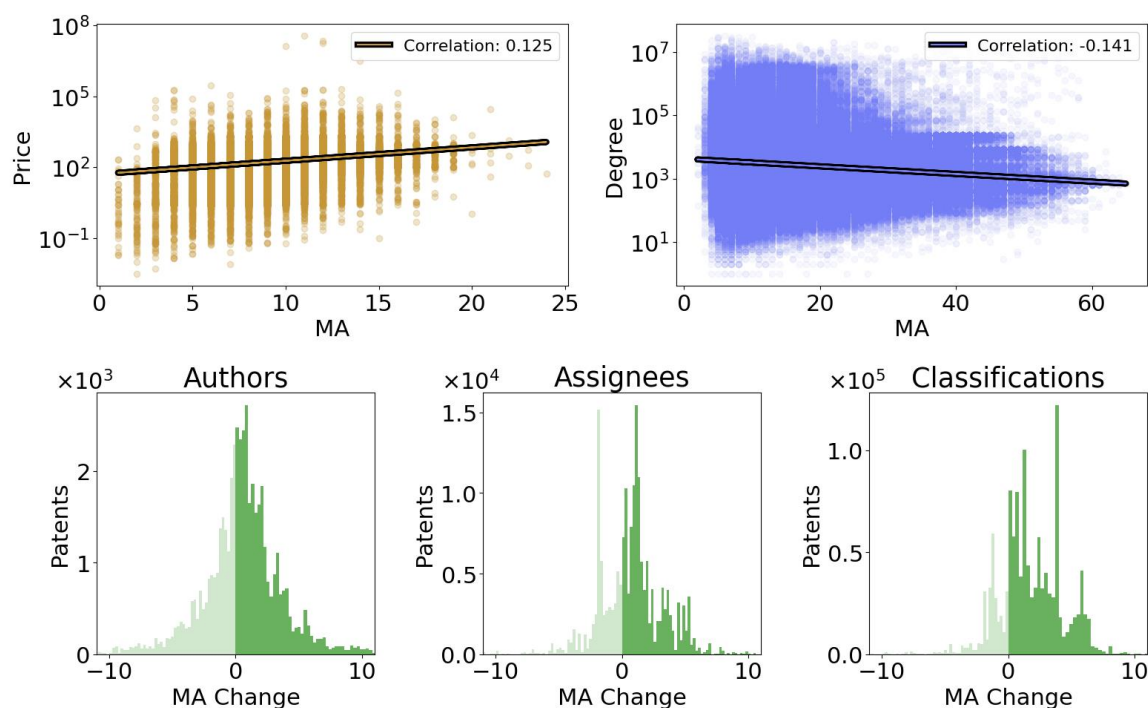


Figure 30: Effect of various factors (cost, network degree, individual authors, assignees, and patent classifications) on MA increase over time.

MA & Cost

We tested several possible factors which could potentially explain the increase in MA over time. We first explored the cost of compounds, hypothesizing that higher-MA compounds are associated with higher costs due to increased discovery and production expenses of more complex compounds. We use cost data from the Reaxys database (Lawson et al., 2014) collected from Dr. Dario Caramelli and Dr. Hessam Mehr from the University of Glasgow, where each compound has a specific price in GBP per gram. We sampled 50,000 compounds from their unpublished analysis and calculated the MA from all 50,000 using the AssemblyGo split-branch algorithm with a timeout of 300 seconds (Figure 31). In total, we calculated the exact MA for 46,187 compounds (Figure 32). From these data, we found a weak positive correlation (spearman coefficient = 0.125) between cost and MA.,

suggesting increasing cost over time plays a small factor in the observation of increasing MA.

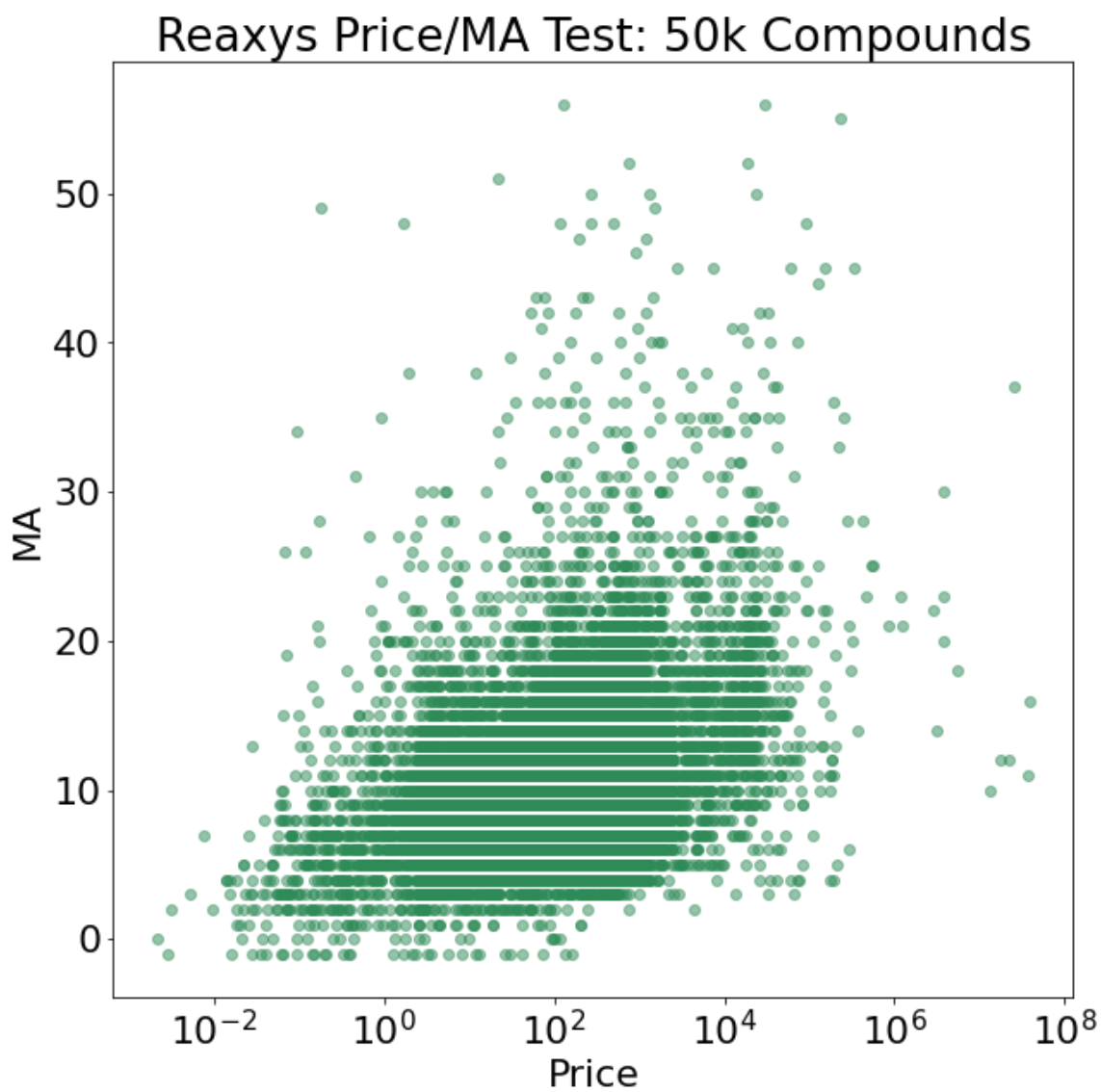


Figure 31: Cost (GBP per gram) / MA, all compounds

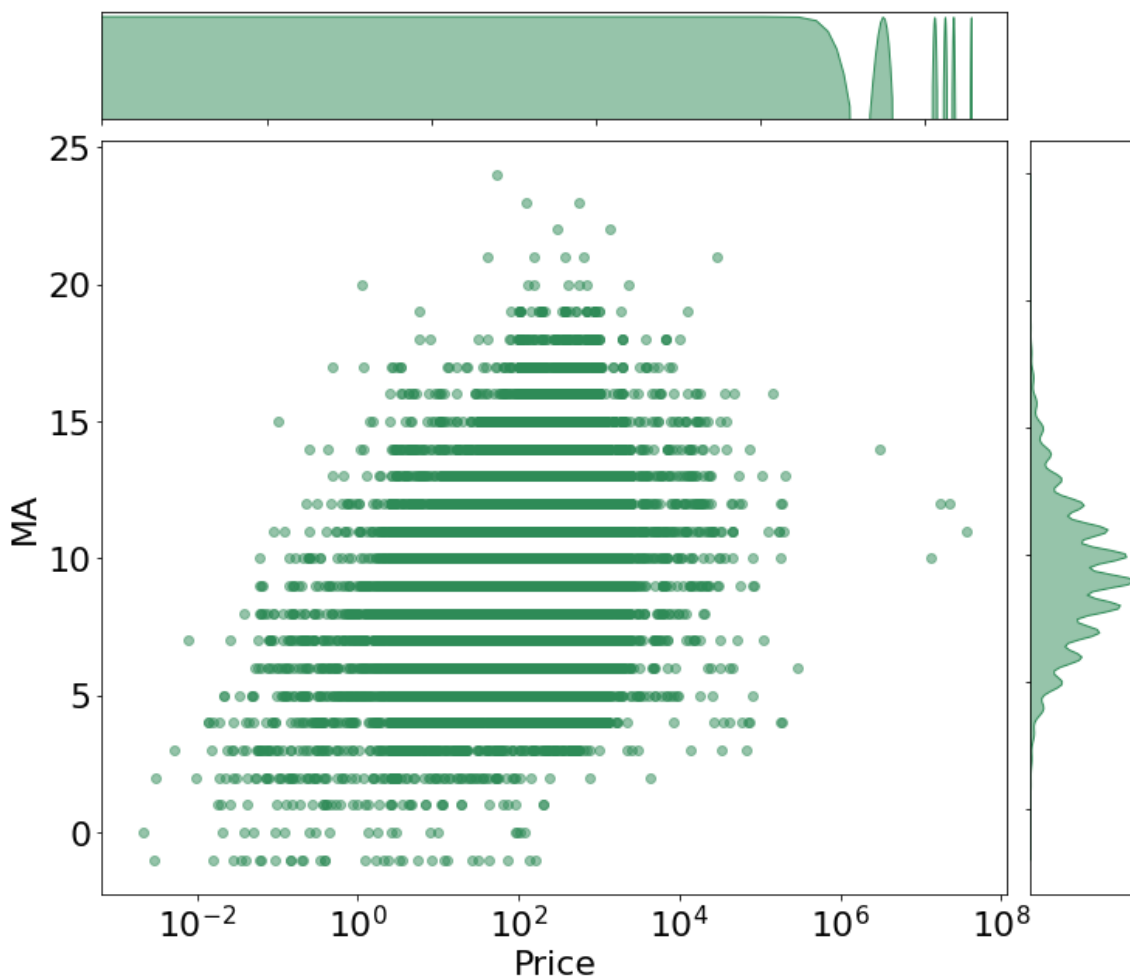


Figure 32: Cost (GBP per gram) / MA, exact MA only

Additionally, from these 50,000 original sampled compounds, we found 2185 which were also present in the SureChemBL database. This joining operation was done by comparing the InChI representations of the original 50,000 compounds to the InChI representations of the 17 million SureChemBL compound database. For the 2185 compound subset, we found the date of earliest entry in SureChemBL to test if there exists a correlation between cost and time. We find a slight positive correlation (spearman = 0.324), which is slightly

stronger than the original cost/MA correlation and further suggests cost has a small positive affect on increasing MA over time (Figure 33).

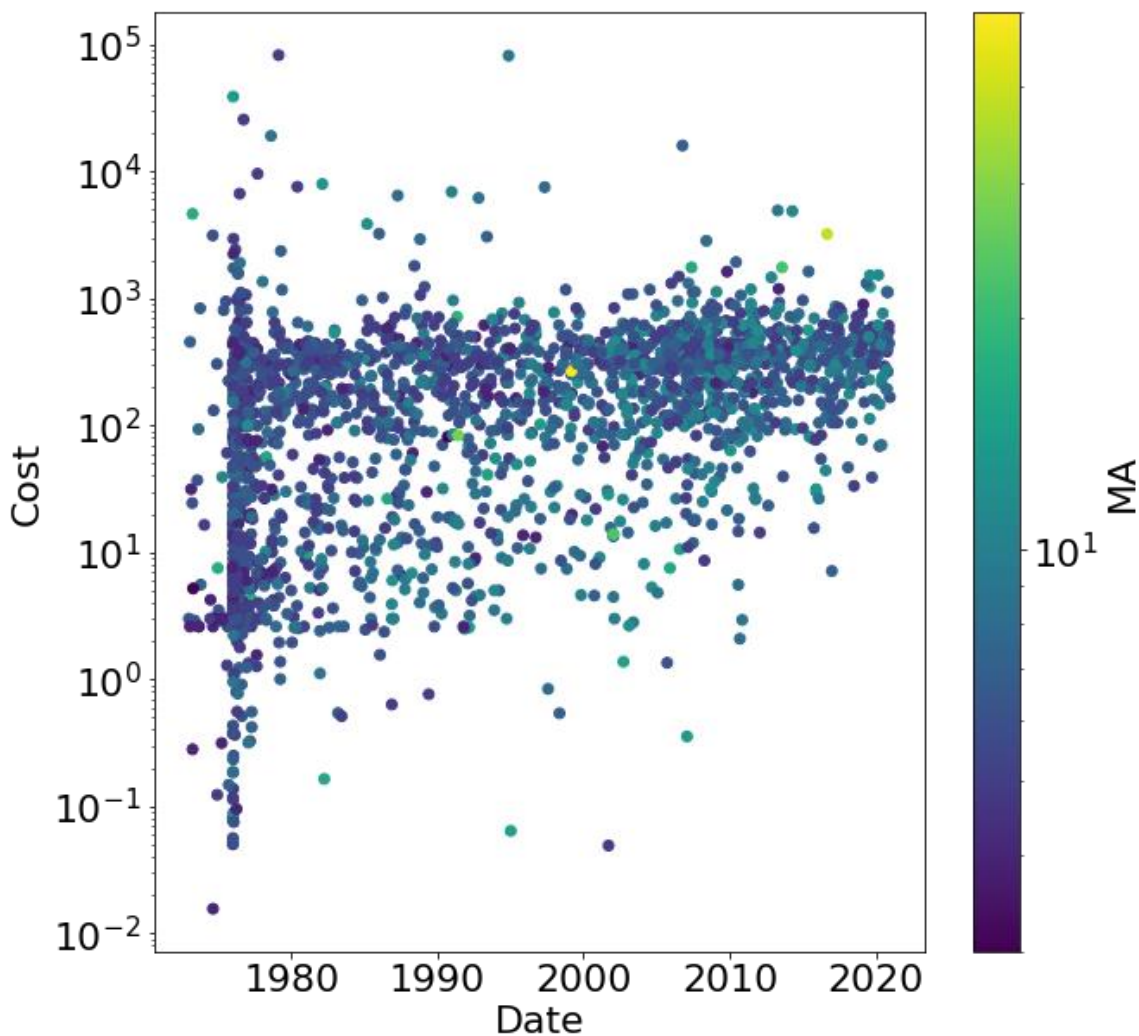


Figure 33: Cost over time (colored by MA value)

MA & Degree

We also tested the influence of how often a compound is used in patents on MA (Figure 30). We used the network from Figure 14 to obtain compound degrees, the number of patents where a compound is used as of December 2019. We used the end of the 2010s as a cutoff in order to match the preferential attachment timeframe in Figure 19. We observe

a weak negative correlation (coefficient = -0.141) between degree and MA, suggesting that higher-MA compounds are used less often in patents than lower-MA compounds. This is not unexpected for the same reasons as cost slightly correlates with MA – there appears to be a higher effort level required to create and subsequently use higher-MA compounds. Additionally, as shown in Figure 14 and Figure 21, compound degree is decreasing over time. This suggests more higher-MA compounds are being created, helping contribute to the increase in MA we observe across patent chemistry.

MA & Individuals, Companies, and Patent Classifications

We were particularly interested in applying social dynamics to MA through individual patent authors, assignees (companies, universities, or other ownership groups associated with a patent), and classifications (USPTO designation of the type of invention) associated with patents. We randomly sampled patents within the SureChemBL database and analyzed the changing MA statistics of authors, assignees, and classifications associated within. In total, we analyzed data from 21947 authors, 1850 assignees, and 2923 classifications. A minimum of 10 patents was necessary for any individual author, assignee, and classification to be included. Additionally, we limited assignees to those that were obviously companies or universities, in order to avoid double-counting authors. When multiple authors, assignees, or classifications are associated with a single patent, we included that patent within every analysis for each individual designation. For each individual analysis, we calculated the average MA of each patent associated with a single author/assignees/classification, then performed a linear regression over time to observe

how MA changes over time. The change in MA of the linear regression – either positive, negative, or zero – is recorded for each individual analysis (see SI for specific methodology and an example). This change in MA is represented on the x-axis of each histogram in Figure 30, with the total number of patents associated with that change shown on the y-axis. A detailed example of this methodology is shown below. The average MA change for authors is 0.412, showing a very small bias towards creating larger compounds over time. For assignees, this average is higher, at 0.989, and for classifications it is even higher at 1.643. Sample sizes are likely not responsible for the differences between authors, assignees and classifications, as MA change stay consistent across dropout tests within each category (shown below). These averages are supported by the percentile of the data where zero MA change appears – for authors, 0 MA change is in the 41st percentile; for assignees, 0 MA change is in the 34th percentile; and for classifications, 0 MA change is in the 22nd percentile. These results suggest different types of patents, as classified by the USPTO, are more responsible for increasing MA over time, with assignees being less responsible, but still more of a driver of increasing MA than authors, who ultimately have little influence on MA changes over time.

Social MA Methodology & Example

We walk through one specific example to explain this process – this explores patents which listed the University of Arizona as the assignee. We first filter the list of all patents to only include the University of Arizona (labeled “UNIV_ARIZONA” in the data) and calculate the average MA of the compounds associated with every patent. There are 494 patents in the dataset associated with University of Arizona in total. Figure 34 shows the average MA of patents over time.

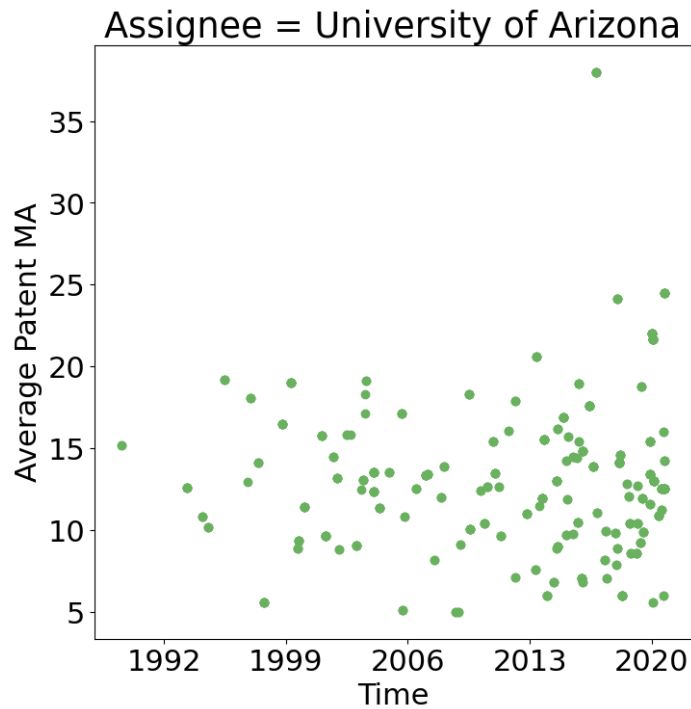


Figure 34: Average MA of all 494 patents associated with the University of Arizona over time.

We built a linear regression model to fit these data, demonstrating that patents associated with the University of Arizona have a slight positive trend over time. We calculate the delta MA (2.704 assembly units) over the time series, which is the change in average MA of the linear regression line from the time of the first patent considered to the time of the final patent (Figure 35). The linear regression statistics for the University of Arizona are shown in Table 4.

We graph the deltaMA in Figure 30, weighted by the number of patents associated for each individual. For example, since the University of Arizona had 494 patents associated with a

delta MA of 2.704, we graphed 494 occurrences of 2.704 onto the assignee histogram in Figure 30.

Table 4: Linear Regression Statistics For University of Arizona Patents.

Dep. Variable:	y	R-squared:	0.018			
Model:	OLS	Adj. R-squared:	0.016			
Method:	Least Squares	F-statistic:	9.029			
Date:	Wed, 05 Apr 2023	Prob (F-statistic):	0.00279			
Time:	11:25:10	Log-Likelihood:	-1490.6			
No. Observations:	494	AIC:	2985.			
Df Residuals:	492	BIC:	2994.			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-165.1187	59.448	-2.778	0.006	-281.921	-
48.316						
x1	0.0002	8.09e-05	3.005	0.003	8.41e-05	0.000
Omnibus:	143.136	Durbin-Watson:	0.402			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	510.705			
Skew:	1.298	Prob(JB):	1.26e-111			
Kurtosis:	7.251	Cond. No.	1.96e+08			

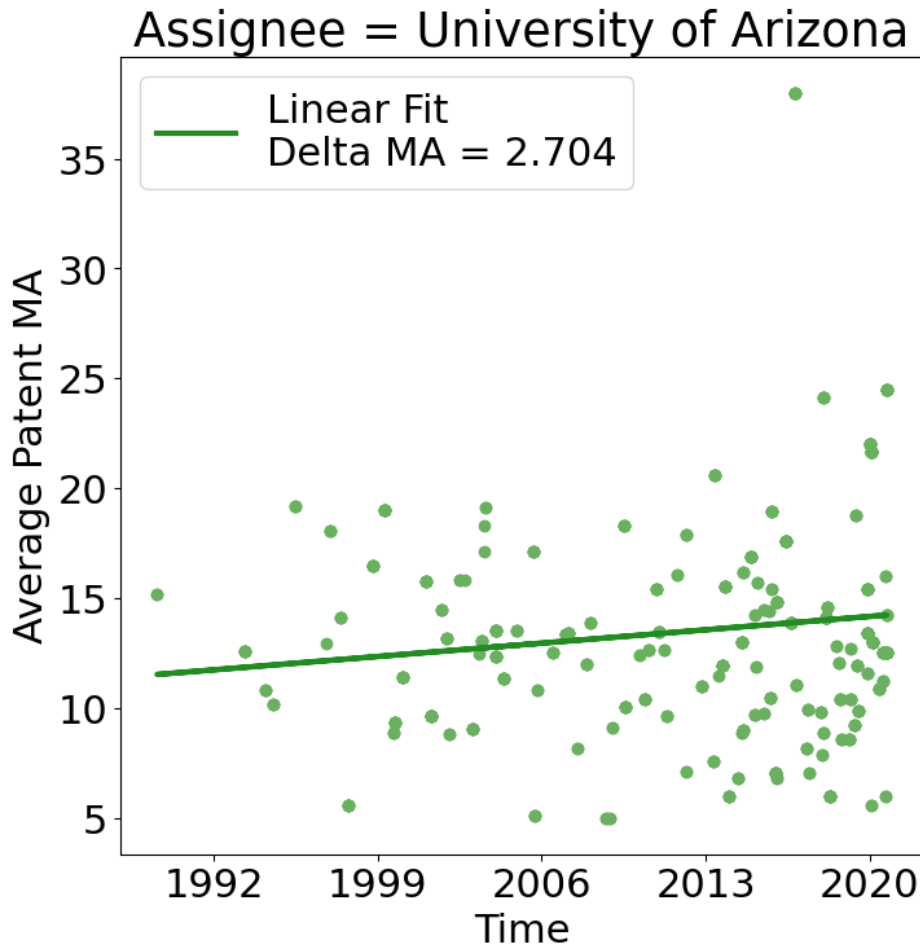
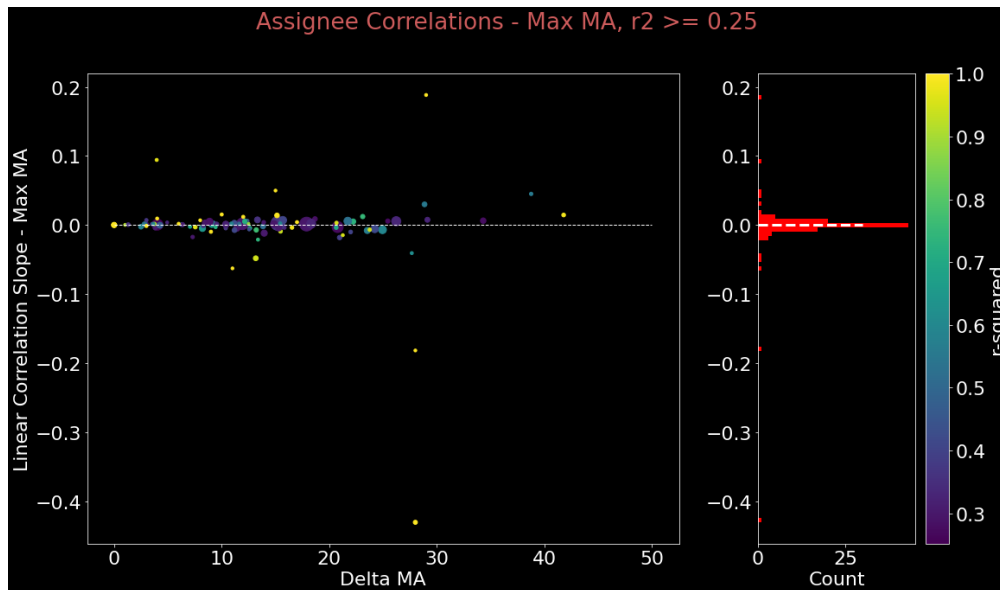
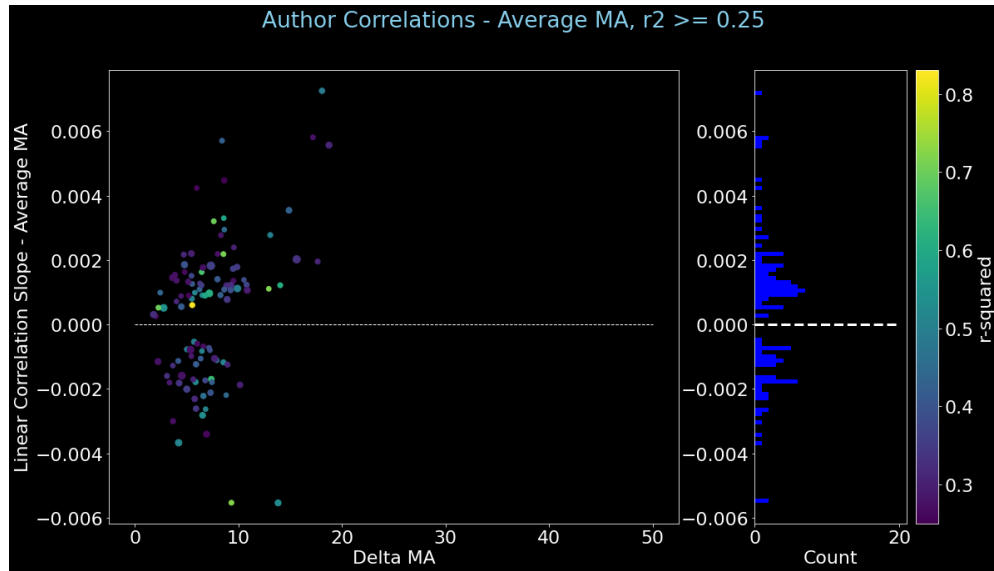


Figure 35: Average MA of patents with linear regression and delta MA shown.

For all individual authors, individual assignees and unique classifications, we performed the same tests to calculate the delta MA of average MA across patents. For each individual author, individual assignee and unique classification, we calculated the r^2 value of the linear regression in order to evaluate how close the average MA of patents over time was to a linear fit. The linear regression slopes (y-axis) by the delta MA (x-axis) and r^2 values (coloring of data points) for all authors, assignees, and classifications are shown here. Each data point is sized by the number of patents associated with each individual. Additionally,

the histogram of slopes also shown vertically here (rather than horizontally as in Figure 30), and is not weighted by the number of patents per individual (Figure 36).



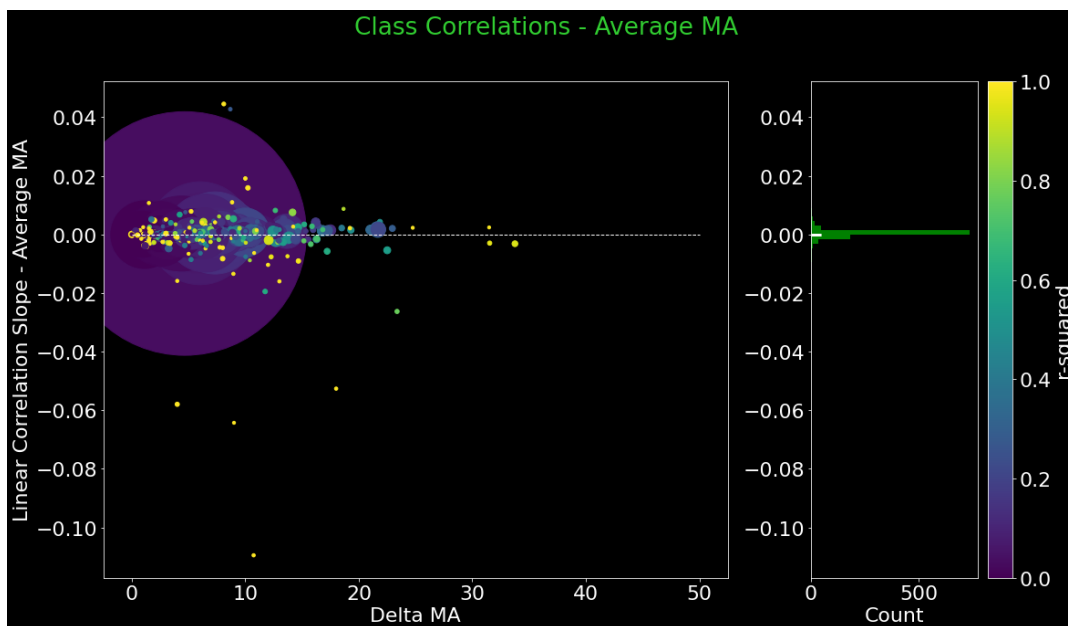


Figure 36: Author, Assignee, and Classification regression results.

We also performed dropout analyses on the weighted bulk results from each of the author, assignee and classification analyses in order to test if basic statistics – specifically, the mean, median and skew – significantly changed as a result of smaller sample size. We performed 1000 iterations of 20% dropout, where 20% of the individual authors, individual assignees or unique classifications were removed from each analysis. We removed individuals or classifications instead of patents because each individual/classification had potentially different numbers of patents associated with them, and since we are testing at the level of individuals/classifications, we wanted to ensure the data is statistically rigorous in regard to perturbations at the individual level. Across all 1000 iterations of individual authors, individual assignees and unique classifications, the mean, median, and skew are stable, showing that the results are robust in regard to individuals (Figure 37, Figure 38, Figure 39).

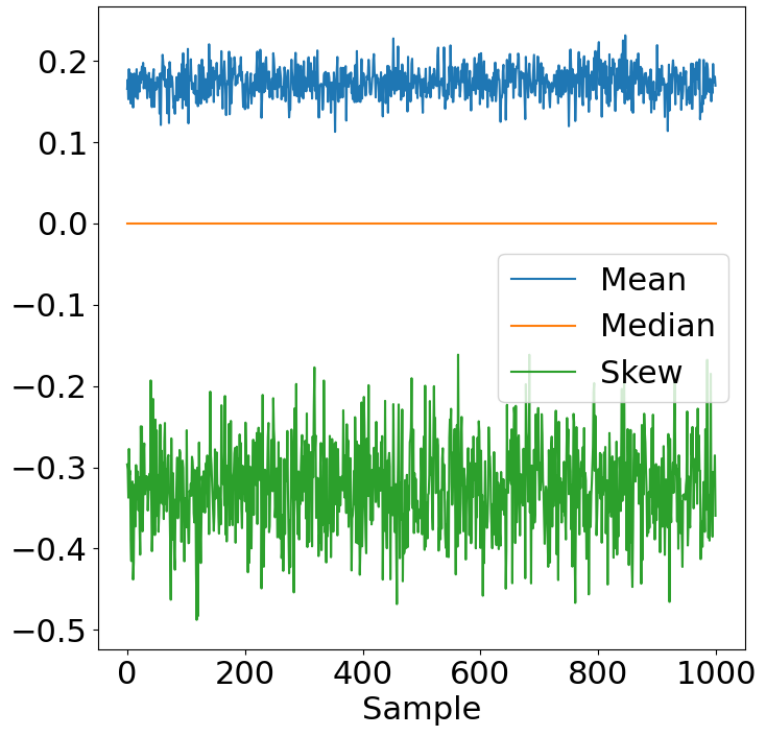


Figure 37: Author dropout tests results

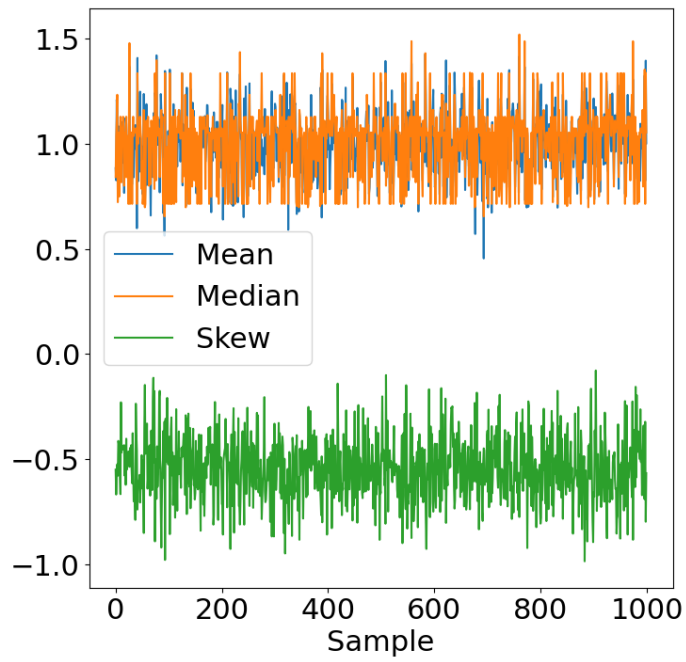


Figure 38: Assignee dropout tests results

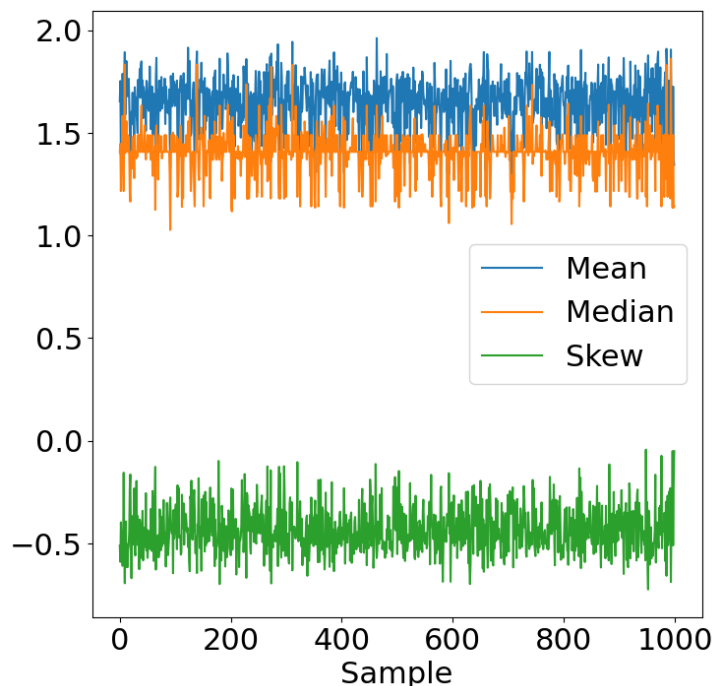


Figure 39: Classification dropout tests results

Discussion

Preferential Attachment

We found the SureChemBL patent-compound network grows according to the preferential attachment model - specifically, patents are more likely to use compounds that were already highly used in previous patents. Those compounds which were rarely utilized in patents are unlikely to be utilized in future patents. This growth model results in a scale-free network structure (SI Table 1) where relatively few compounds are ubiquitous in patent chemistry and the vast majority of compounds are used in very few patents. Scale-free networks are found in a wide variety of systems, including other chemical networks such as biological metabolism (Broido & Clauset, 2019), and this result adds patent chemical networks to the growing body of systems which have this property.

Patent chemistry networks fill an interesting niche in the landscape of network science and scale-free networks. There are constraints put on it through chemical reactions, in a similar fashion to biochemistry and other chemical reactions networks, as patents can only use chemically feasible reactions, limiting the type of compounds utilized within a patent. Additionally, there are social constraints put on patent chemistry through the nature of patents and economics - these documents protect specific compounds and reactions from use elsewhere, and as such can prevent the spread of these compounds to other patents, resulting in a low degree for protected compounds. Additionally, these patents are done to protect intellectual property of scientists and companies, possibly resulting in some form of profit for the patent authors and/or assignees. The motivation behind the patents, therefore, is driven at least in part by market forces. Most socially derived networks have some form of constraints, such as human-human interaction networks that are constrained by location, but most of these networks do not have a law- or economy-based constraint. The result that patent chemistry is limited by both physical and social constraints, yet still is built according to the preferential attachment model and results in a scale-free structure, highlights the ubiquity of these kinds of networks across systems.

The role of individual compounds within the network is also explored here, with green solvents highlighted as a particularly intriguing example of compounds which become more connected over time. The motivation for green chemistry is fairly obvious as environmental issues become more pressing and urgent (Gałuszka et al., 2013; Raccary et al., 2022), but we encourage other researchers in various chemical disciplines to research

their own compounds of interest to study the relative popularity of compounds and patents over time relating to their individual fields.

Assembly

Beyond the growth of the patent network, the compounds themselves provide intriguing insight into how chemistry grows and changes over time. Using assembly theory, we observe a growing trend in the MA of compounds over time, showing compounds have steadily become more complex over the last 40 years. The average molecular weight of compounds has also linearly increased and is positively correlated with MA (see SI Figure 25). The reasons for this linear increase are not immediately clear, however. We hypothesize that novel fragments – which also grow at a linear rate – are responsible for new combinations of atoms and bonds, resulting in a corresponding rise in MA. Despite a shallow understanding of the underlying mechanisms of increasing MA over time, we highlight the importance of this observation. It has been hypothesized that life becomes more complex over time (Bettencourt et al., 2007; Llanos et al., 2019; Szathmáry & Smith, 1995; Szymkuć et al., 2021), and by using assembly theory as an essential quantifiable definition of complexity, we demonstrate that chemistry used within human systems is, in fact, becoming more complex. It is important to note that complexity is not a synonym of disruption (Park et al., 2023). In fact, these results go further than the premise that patents are less disruptive over time proposed in (Park et al., 2023), as we observe a consistent linear increase in MA compared to an exponential growth of patents. This suggests chemistry is predominately re-using similar patents and has hewn to this model since at

least 1980, as opposed to finding more complex structures in the past and less complexity in the present.

There are many potential social reasons for increasing MA, from cost, usage rates, and various levels of usage within social hierarchies. The slight positive correlation between cost and MA highlights that as molecules have more joining steps, it presumably costs more to discover and synthesize (which, therefore, leads to a higher price). This fits the overall hypothesis of AT, which postulates that high-MA molecules are exceedingly rare to make in combinatorial chemical space, and that any high-MA molecule found in high abundance would be the result of a dedicated process (Marshall et al., 2022; Sharma et al., 2022), be that biochemistry or a group of scientists working to discover a novel drug. The inverse of this is that as MA increases, we find that the usage of compounds - measured through network degree - decreases within patent literature. This may be related to the cost, as higher-priced (higher-MA) compounds require more funding and resources to obtain, and are therefore used less frequently than established, cheaper compounds. We hypothesize that cost and usage These two factors show that social pressures, specifically market pressure, scientific research & synthesis time, and resources, have an impact on the MA of chemistry produced within society and affect the trajectory of how complex chemistry is produced at scale.

MA increase emerges at different levels of social interactions. Taken as a collective, individual patent authors have very little impact on how MA increases changes at large scales. There are individuals who have a career where the average MA of their associated

patents steady increases, but there is a roughly equal number of individuals who have the opposite and have a decrease in MA, with a plurality of authors having no trend whatsoever. However, both companies and USPTO classifications exhibit behavior which results in increased MA values. More companies and classifications have a steady increase in MA in patents over time, with classifications having a stronger increase than companies. We suggest that while individual authors have a specific skill set and possibly work on similar tasks – and therefore similar compounds – throughout their career, the long-term direction provided by companies allows for discovery of new, more complex compounds. Additionally, we different companies working within the same classification compete to discover novel, often higher-MA compounds, allowing for higher-order emergent behavior which selects for highly complex compounds.

4: EARTH SCIENCE CURRICULA FOR INSTITUTIONALIZED YOUTH

Introduction

This work highlights an Earth Science curriculum built and taught in Spring 2022 in collaboration with teachers in the Arizona Department of Corrections, Rehabilitation and Reentry (ADCRR) and the impact of this curriculum on the Science, Technology, Engineering, and Mathematics (STEM) identity of the students taking the course. We use the sociotransformative constructivism (sTc) (A. J. Rodriguez, 1998) framework and a place-based focus on the Sonoran Desert to develop the lessons and activities covered here. Topics covered include planetary formation, interacting geologic systems leading to the origin of life, potential life on other planets and possible similarities to that on Earth, and the impacts of human-driven climate change (Table 3). Graduate student and faculty volunteers were included in the creation and teaching of lessons, so that the students could learn from experts in various topic fields and be introduced to a diverse range of scientists.

The student's STEM identities towards Earth Science topics were collected using a modified version of the Colorado Learning Attitudes about Science Survey (W. Adams et al., 2006), to quantitatively study how their perception of themselves within Earth Science changed over the twelve-week course. Additionally, we developed a final project which required students to create unique planets and ecosystems based on geologic constraints. We qualitatively assessed the project reports to measure the impact of the sTc framework used to build the curriculum, as well as to measure other emergent themes - such as references to the flora and fauna found within the Sonoran Desert - that were expressed within the student projects.

While the vast majority of research on prison education is focused on the impacts of education on reducing recidivism in adult learners (Baranger et al., 2018; Courtney, 2019; Ellison et al., 2017; Esperian, 2010; Fabelo, 2002; Gaes, 2008), there are efforts to move the conversation of prison education beyond solely recidivism and instead highlight the transformative opportunities of education within the prison environment. Flynn and Higdon advocate for incarcerated persons engaging with their education in a way that ensures meaningful interactions with outside society (Flynn & Higdon, 2022), while Szifris and others describe education as a method of positive personal change for those incarcerated (Szifris et al., 2018). This study adds to this transformative effort by utilizing sociotransformative constructivism (Alberto J. Rodriguez & Morrison, 2019) within science education in a youth educational setting in ADCRR. The impact of education opportunities for incarcerated persons under 18 years old is largely unstudied, and while this study does not address large-scale issues within juvenile prison education at large, it highlights how the sociotransformative constructivism framework in conjunction with interactive, place-based teaching strategies can lead to personal identity changes towards science in prison education efforts, particularly to youth learners.

Earth science is an ideal subject for prison education, as it incorporates and informs interdisciplinary thinking from scientific disciplines such as geology, atmospheric sciences, planetary science, biology, astrobiology, and even sociology through the effects of climate change (Orion, 2019; Pennington et al., 2020). The various fields and topics

which can be covered within an earth science curriculum allow students to not only connect knowledge across multiple disciplines, but also gives intersectional perspectives and ways for students to relate to different ways of scientific thinking using real-world, place-based lessons (Gosselin et al., 2016; Núñez et al., 2020; Semken et al., 2017). For example, the study of climate change is fundamentally a study at the interface of geological processes and human impacts on them (Pennington et al., 2020). Arizona - and the Sonoran Desert in particular - is already affected by climate change through hotter summers and a changing fire regime (Aslan et al., 2021; Hantson et al., 2021). As LLA is located within the Sonoran Desert, this focus on earth science specific to Arizona is personally relatable to the students here in a way that other subjects would not be. Additionally, the Next Generation Science Standards, which were used by the Arizona Department of Education to build the Arizona Science Standards followed within this project, emphasize cross-cutting concepts so that students learn to solve problems using techniques from seemingly disparate fields (Pruitt, 2014).

This project introduces students to these various earth science topics through a sociotransformative constructivism (sTc) framework (A. J. Rodriguez, 1998). sTc is a synthesis of educational social justice (Maulucci, 2012) and constructivist learning theory, which results in educational practices that teaches towards diversity and understanding (A. J. Rodriguez, 1998). The sTc framework has been used in many educational settings (Avsar Erumit et al., 2021; Tolbert et al., 2022; Varelas et al., 2022), but rarely within a prison environment. In general, there is a lack of applied educational theory to prison education,

particularly in how education can lead to change within prison structures (Szifris et al., 2018). This work will contribute to this lack of literature through the application of sTc to a specifically designed youth prison education class.

Theoretical Framework

The sTc framework is built on four specific components for educators to shape their classrooms to teach to their student's diverse experiences and ensure that lessons are designed to facilitate student's understanding. These components are four unique - but often overlapping - elements: dialogic conversation, authentic activity, metacognition, and reflexivity (A. J. Rodriguez, 1998; Alberto J. Rodriguez, 2015). Dialogic conversation is the emphasis on understanding why a speaker (either the student or educator) chooses to speak in a specific way based on their experiences. Crucially, a dialogic conversation can only occur if there is an established trust between the speaker and listener, so that there are no power hierarchies within the conversation (Bakhtin, 2010; Howe et al., 2019). An authentic activity is a hands-on, tactile lesson which is specifically designed to teach a subject, but also is designed to incorporate student's experiences through tying it to the student's culture and everyday life. Metacognition also draws on student's experiences by encouraging them to identify where they personally fit into lessons, and additionally reflect on why and how they are learning specific topics. Reflexivity occurs when students bring their own social upbringing, location (both ideological and geographical), and beliefs to the lessons to explore their place in the subject material. All these components are based in experience-based learning and a critical approach to education (Wang et al., 2019), and force both students and educators to examine power structures embedded within the classroom (such as teacher-student relationships) and also within larger society.

The use of sTc to bring diversity to the forefront of teaching emphasizes cross-cultural education, meaning that the perspectives, knowledge and contributions of the student's

cultures are represented within the curriculum developed here to empower students within this course (Aronson & Laughter, 2016). This is the sociotransformative element of sTc, and allows students from culturally diverse backgrounds to see themselves within the material and provides a platform for these students to become empowered and have a sense of personal agency towards the course material (A. J. Rodriguez, 1998; Alberto J. Rodriguez & Morrison, 2019).

The social constructivist theory within sTc focuses on learning as an individual process which comes about through personal experience within a wider culture. The student's culture influences how they see themselves and the world around them, and when combined with their own personal views and experiences, the student is able to make better sense of the world and results in higher levels of thinking (Amineh & Asl, 2015; Vygotsky & Cole, 1978). Approaching education in a constructivist manner allows educators to teach in a way that brings their student's cultural and personal experiences to the forefront, as well as understand that these experiences are critical to learning and development.

Within a prison environment specifically, the relationship between educators as outsiders to the prison reflects the historical and cultural context of the prison system in the United States, where those in prison are often due to the result of cumulative disadvantages, such as the multiple disadvantages (economic, social, political, etc....) attached to communities such as young, male persons of color from low-income neighborhoods (Bishop & Frazier, 1988; Kurlychek & Johnson, 2019). Those who teach prison education, who are often

associated with universities, often do not have the same racial makeup of prisoners - in adult facilities, over 50% of prisoners are Black or Latinx (Gramlich, 2019), while only 11% of college prison instructors are (Krupnick, 2018). This dichotomy in racial and power relationships present in prison education work (Taylor et al., 2021) necessitate an egalitarian approach (Young, 2006) using sTc to build trust between educators and students through deconstructing traditional social power structures within the classroom (Collins & Blau, 1979; Taylor et al., 2021; Tolbert et al., 2018).

Methods

Research Design

We used a simultaneous multi-methods design for this project involving both quantitative and qualitative data analysis (Morse, 1991) (Figure 40). The two distinct methods of data analysis did not inform each other, making this a simultaneous study rather than a mixed method design. The qualitative research, which was the main focus of the study, was performed on individual project reports. These written reports were converted to electronic form and coded using the NVivo analysis software tool (Edwards-Jones, 2014). We initially used the four components of sTc - dialogic conversation, authentic activity, metacognition, and reflexivity - as well as two markers of science identity as codes. The science identity codes were based on knowing and using science. After reading the individual reports, we also included five codes based on recurrent themes across the reports and included each of these as part of one of the overarching sTc components. These codes were designed around sTc to evaluate the curriculum using the same framework in which it was built and taught.

We also conducted a survey based on the Colorado Learning Attitudes about Science Survey (CLASS) (W. K. Adams et al., 2008; W. Adams et al., 2006) at the beginning and end of the 12-week class. This quantitative survey was designed to measure the Earth Science identity of the students and how this identity changed over the course of the class.

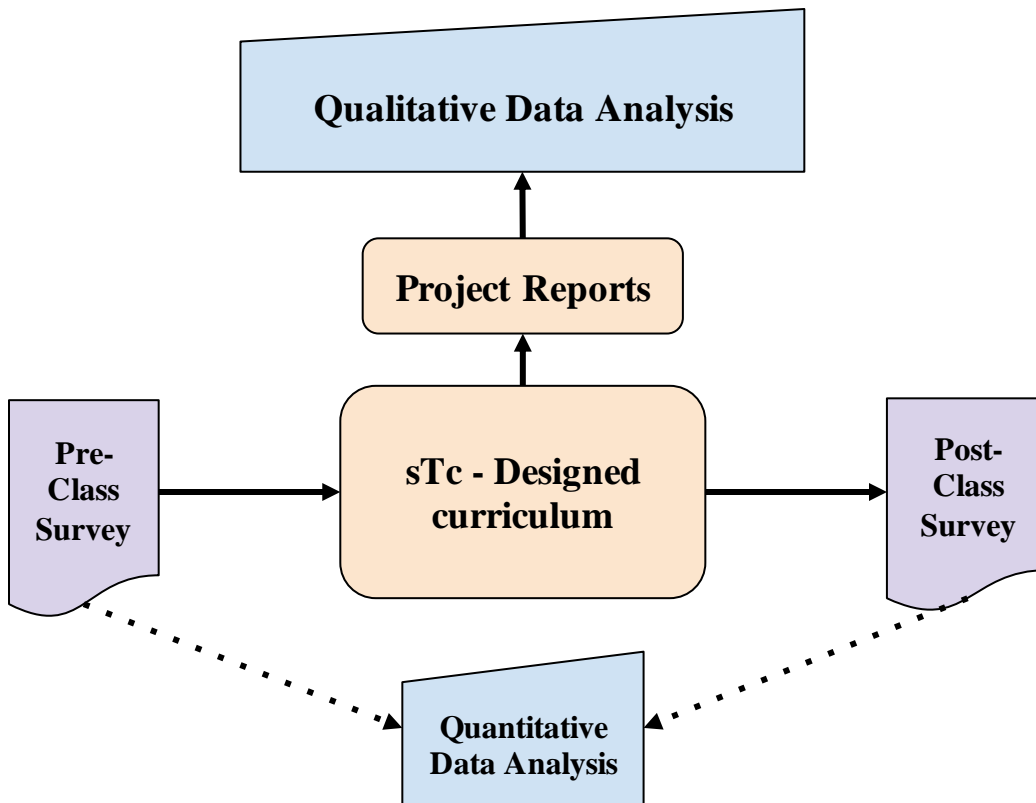


Figure 40: Mixed-methods study overview.

Classroom Setting

=The class ultimately reached over 20 students, but only five were present throughout the entire course we designed as sentence times, age limits, and new arrivals all influenced the number of students within the classroom on a day-to-day basis. The survey and project described below, administered to these five students, are designed to measure the learning outcomes and connections made by the students.

We worked in collaboration with the administrative staff and teachers within ADCRR, particularly Dr. Debra Skinner and Jordan Remold to build a curriculum for a quarter-long (12 weeks) Earth Science course. The author and Jordan Rembold co-taught this class

during spring 2022. This class started on January 3rd and ended on March 26th to two separate classes of students, teaching over 20 students in total.

Prison Environment

Throughout the course, we were challenged and sometimes limited by the structure placed by teaching within a prison. This class was designed so that lessons and activities were built on top of previous ones, but a significant number of students started in the middle of the class due to sentencing and movement within the prison. Additionally, while experimental & activity materials were nearly consistently approved to bring into the classroom, the students were not permitted to interact with technology such as computers, so technology-based lessons and activities were not permitted. The approval process to bring materials into LLA consisted of submitting lists of materials at minimum one week before the planned lesson and receiving written permission from the warden of the Lewis-Sunrise facility. The teachers were responsible for purchasing and bringing in materials. The ban on student technology allowed for a greater tactile learning experience but limited the type and diversity of activities available to the students. Also, the minors unit is a subunit of a larger prison complex, and disruptions in other units caused class delays and sometimes cancellations. One example of a disruption was a 45-minute delay due to searches in the adult wings of the prison. A guest speaker from ASU was present, and the delay turned a planned 1-hour activity into an impromptu question and answer session to give the students an opportunity to learn from the speaker and ensure their curiosity was addressed. While the students were disappointed to miss the activity, they were engaged in the short time available with the guest speaker.

Survey Instrument

The goal of the curriculum is to increase earth science knowledge and science identity of the students within the course. A student's science identity is how a student perceives themselves and navigates through learning science (Tytler, 2014). A strong science identity is related to when a student is recognized as having an affinity for science and also develops an increasing interest in a subject (Dou et al., 2019). This results in a student identifying personally and/or being recognized as a person who has an affinity for scientific topics. If a student is not recognized for their affinity or does not develop a science-specific interest, which could result from a variety of reasons, such as underrepresentation of experts who look similar to them in positions of power (Barton & Tan, 2010; Hazari et al., 2020), then they do not have a strong science identity.

We modified the CLASS survey (W. Adams et al., 2006) to test changes in STEM identity. The CLASS survey, initially designed for physics and then extended to chemistry (W. K. Adams et al., 2008) and other science subjects (Semsar et al., 2011; Wilcox & Lewandowski, 2016), tests how students view themselves compared to experts in a given scientific field. Previous work with CLASS measures how this identity compares to classroom achievement (W. Adams et al., 2006; Deslauriers et al., 2019; Hazari et al., 2010). Here, we are limited by two factors - first, the survey is not validated for earth science, and therefore cannot definitively show an increase in earth science identity. Second, there are only five students who ultimately participated in the survey, which limits the statistical power of the results. However, we can show trends over time within this limited sample size, so while we cannot give specific answers to changes in STEM identity,

we can observe the impact of the curriculum on how the surveyed students view themselves within earth science. The modified CLASS earth science survey is shown in Table 5. We gave the survey to students twice - once at the beginning of the course, and again at the very end of the course. Each question is scored on a Likert scale of 1-5, where 1 is “Strongly Disagree” and 5 is “Strongly Agree”.

Table 5: Survey Instrument, Based on the CLASS Survey, Given Out to Students in the Earth Science Class Developed as Part of This Project.

1. A significant problem in learning earth science is being able to memorize all the information I need to know.
2. After I study a topic in earth science and feel that I understand it, I have difficulty solving problems on the same topic.
3. Knowledge in earth science consists of many disconnected topics.
4. When I solve an earth science problem, I locate an equation that uses the variables given in the problem and plug in the values.
5. If I get stuck on an earth science problem on my first try, I usually try to figure out a different way that works.
6. Nearly everyone is capable of understanding earth science if they work at it.
7. If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it.
8. If I want to apply a method used for solving one earth science problem to another problem, the problems must involve very similar situations.
9. Learning earth science changes my ideas about how the world works.
10. I can usually figure out a way to solve earth science problems.
11. The subject of earth science has little relation to what I experience in the real world.
12. To understand earth science, I sometimes think about my personal experiences and relate them to the topic being analyzed.
13. If I get stuck on an earth science problem, there is no chance I'll figure it out on my own.

Arizona State Standards

In order to fit within Arizona state curriculum standards, we built the curriculum on Arizona Standards for Earth & Space Science (Arizona Science Standards, 2018, p. 80),

which are based on Next Generation Science Standards (Pruitt, 2014). These standards emphasize that inquiry and knowledge are equally important in learning scientific topics.

The two specific core standards which apply to this course are:

E1: The composition of the Earth and its atmosphere and the natural and human processes occurring with them shape the Earth’s surface and its climate.

E2: The Earth and our solar system are a very small part of one of many galaxies within the Universe.

Within the Arizona Earth & Space Science core standard E1, there are four specific sub-standards detailed within the full state standards, all of which are covered within this class.

These are detailed in Table 6 below.

Table 6: High School Sub-Standards Within the E1 Core Idea of the Arizona State Standards. Adapted from (Arizona Science Standards, 2018, p. 80).

Arizona State Sub-Standard	Description
HS.E1U1.11	The foundation for Earth’s global climate system is the electromagnetic radiation from the Sun as well as its reflection, absorption, storage, and redistribution among the atmosphere, ocean, and land systems and this energy’s reradiation into space.
HS.E1U1.12	Earth’s systems, being dynamic and interacting, cause feedback effects that can increase or decrease the original changes.
HS.E1U1.13	Continental rocks, which can be older than 4 billion years, are generally much older than rocks on the ocean floor, which are less than 200 million years old.
HS.E1U1.14	Global climate models are often used to understand the process of climate change because these changes are complex and can occur slowly over Earth’s history. Though the magnitudes of human impacts are greater than they have ever been, so too are human abilities to model, predict, and manage current and future impacts.

Place-Based Curriculum

We used the geology of Arizona, particularly the central Arizona basin and range region where the school is located, throughout the curriculum to incorporate elements of place-based learning (Semken et al., 2017; G. A. Smith, 2002), even within a prison environment. The integration of Arizona-specific volcanoes, rocks, climate change risks, and heat mitigation solutions allowed for students to meaningfully understand and relate to the material in a way that fits with their personal experiences, beliefs, and approach learning in a way that puts them at the center (Butler & Sinclair, 2020; Sheerman, 2020).

Units

The first three weeks of the curriculum consist of introductory geology topics, specifically the water cycle, carbon cycle, planetary formation, and the rock cycle. This geology unit was designed to provide a foundation for the students in terms of geological topics, so that later units can build upon this basic knowledge.

The second unit, interacting systems, focuses on the various interactions found in earth science between seemingly distinct geologic topics. We dealt specifically on geology-atmosphere and geology-biology interactions, allowing students to learn how abiotic factors interact and also how living systems are interconnected with geological processes. We used volcanoes as a specific example of geology-atmosphere interactions, as they are common to the basin-and-range geography of central Arizona. The students learned about the various types of volcanoes, from the shield volcanoes in Arizona to the stratovolcanoes in the Pacific Ring of Fire, as well as how the ash clouds lead to micro- and macro-climate perturbations.

As the final part of this unit, we introduced geology-biology interactions by discussing the origin of life on Earth, as well as predicting what extraterrestrial life would look like given different geologic restrictions. This activity, done over two class periods, allowed students to bring their own experiences into the classroom (reflexivity) and we focused on how these extraterrestrial life models can relate to various ecosystems and their own learning experiences (metacognition).

The final unit we designed was focused on climate change. This unit was the culmination of the previous two, as it utilized both geology knowledge and the interactions between different systems, with a focus on Arizona-specific climate change risks and solutions. The place-based curriculum gave students the opportunity to personally relate to the material, as well as to reflect on how climate change affects Arizona communities and why it is an important subject to study for the state's future (reflexivity). In this section, we also enabled students to create space to discuss their personal viewpoints on climate change, and how they think they have been and will be affected by climate change (dialogic conversation). One specific example from the classroom which highlights the impact of using sTc within the classroom was a discussion on drought in Arizona. This discussion occurred in Week 9, near the end of the main curriculum. As a part of the conversation, I used the example of how water distribution laws applied to the Colorado River have caused dramatic changes to water flow within Arizona, specifically how the Colorado now flows sporadically through Yuma, the southernmost US city on the river. Unbeknownst to me, two of the

students were originally from Yuma, and immediately reflected out loud on their experience of growing up with the river flowing constantly, then drying up as they got older. As the discussion leader, I gave these students space to describe their experience, and other students asked questions of them, allowing them to lead the discussion. This sharing of these personal experiences with the topic at hand showed the trust the students had in the classroom, as they were comfortable sharing their lived experience, but also demonstrated the educational usefulness of incorporating sTc components into lessons.

The materials used on a day-to-day basis, such as activity guides, assignments, and slides, can be found here:

<https://drive.google.com/drive/folders/117OPN2fpyd3XZCNTYRFdKcmoZcEhF8Kk?usp=sharing>

Arizona State University Volunteer Coordination

In addition to teaching with Jordan Rembold, I wanted to ensure the students enrolled in this course learn from a diverse range of scientists and expertise. To that end, I asked various graduate students, undergraduate students, and professors in the School of Earth and Space Exploration to co-present classes. All graduate students went through a background check performed by the ADCRR in order to be approved to teach. Additionally, I worked individually with each student to create a lesson plan which followed the sTc framework, as well as the AZ state standards and the specific earth science curriculum which the class follows.

In total, seven ASU volunteers assisted with teaching throughout the course. One faculty member, Dr. Darryl Reano, assisted with the survey implementation, while five graduate students and one undergraduate student volunteered their expertise to plan and teach a class.

Table 7: List of Units and Weeks When Each Unit Was Taught: Geology (Blue), Interacting Systems (Green), Climate Change (Orange), and the Final Project (Purple).

Week	Topic
01: Jan 3-7 (Geology Unit)	Cycles (Water & Carbon)
02: Jan 10-14 (Geology)	Planetary Formation
03: Jan 17-21 (Geology)	Rock Cycle
04: Jan 24-28 (Interacting Systems Unit)	Geologic Energy Systems & Hydrosphere
05: Jan 31 - Feb 4 (Interacting Systems)	Geologic Energy Systems & Hydrosphere
06: Feb 7-11 (Interacting Systems)	Atmosphere
07: Feb 14-18 (Interacting Systems)	Interactions Between Geologic & Living Systems
08: Feb 21-25 (Climate Change Unit)	Geologic Climate
09: Feb 28 - Mar 4 (Climate Change)	Arizona Climate Change
10: Mar 7-11 (Climate Change)	Natural Resources
11: Mar 14-18 (Climate Change)	Hazard Risks & Future Climate Change
12: Mar 21-25	Final Project

Final Project

The students were required to complete a final project which was the culmination of the class. Students had to choose different planetary characteristics (Table 8), predict the behavior of different systems, including life, on their personalized planet (Table 9), and construct their planet using various materials which were approved by ADCRR for use within the classroom. The students also had to write a final report describing the various

components of their planet, as well as the various inhabitants of their planet and how they interact with other inhabitants and the distinct geologic processes of their individual planets. The students were evaluated on their work developing their ideas, building their planets, and on the final report. This project replaced a multiple-choice exam as the summative evaluation metric for the end of the class.

This project was built using the sTc framework such that the authentic activity, reflexivity, and dialogic conversation components were incorporated. The student's experience of both imagining and building a planet based on how they envision it to be allowed their own personal experiences and beliefs to be incorporated into the project (authentic activity & reflexivity). Dialogic conversation, particularly the trust between students and teacher that was built up throughout the previous 11 weeks of the course, was included through how the students described the various components of their planets and the interactions between different lifeforms. These descriptions were often spontaneous while students were creating their project, ranging from calling across the room to the instructors to conversations between students as they moved around the classroom. Our decision to replace a multiple-choice exam with a writing-based project, as well as giving the students freedom to talk about their planets to each other, was based on the trust established between the students and teachers.

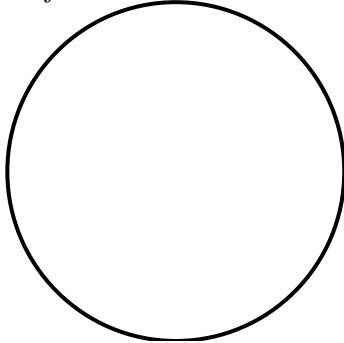
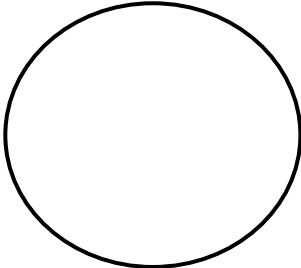
One example of a final project and report is shown in Figure 41. This student built a 0.5 Earth gravity planet with minimal water, some tectonic activity, and a thin atmosphere.

Biological life is found around small oases scattered around the planet, while there is a large mountain range around the equator, where the only tectonic activity is found.

Table 8: Planetary Constraints Available for Students to Build Their Planet.

From the following possible constraints, choose (circle) at least one (1) from each row. The circled choices will inform your model planet and the life you create on it.					
Gravity / Size <i>Circle one option</i>	¼ the gravity & size of Earth	Roughly Earth-sized	4x the gravity & size of Earth	10x the gravity & size of Earth	Other (<i>specify below</i>)
Tectonic Activity <i>Circle one option</i>	No tectonic activity	Some tectonic activity (occasional volcanoes & earthquakes)	Constant tectonic activity (constant volcanoes & earthquakes)		Other (<i>specify below</i>)
Water <i>Circle one option</i>	Minimal water (desert planet)	Some water (lakes present)	Majority water (oceans, lakes, rivers)	All water	Other (<i>specify below</i>)
Gasses <i>Circle 2 or more</i>	Carbon Dioxide	Water (<i>must align with water row</i>)	Nitrogen	Ammonia	Ozone
Gasses (continued) <i>Circle 2 or more</i>	Oxygen	Sulfur	Other (<i>specify below</i>)		

Table 9: Planning Worksheet Where Students Predict Future Outcomes of Their Planet’s Geosphere, Hydrosphere, Atmosphere, Biosphere, and Climate. Based on Their Individual Choices in Table 8.

<p>Use the spaces below to plan the construction and various processes of your planet. The answers are entirely up to you, as long as they are consistent with your planet.</p>	
<p>Geosphere: Describe in three (3) phrases the general geology of your planet (for example, “volcanic”, “one continent”, “full of craters”)</p> <hr/> <hr/> <hr/>	<p>Geosphere: What colors will the surface of your planet be? On the circle below, color what a section of your planet’s surface.</p> 
<p>Hydrosphere: How will the living organisms on your planet use & store water?</p> <hr/> <hr/>	<p>Hydrosphere: Is there a water cycle on your planet? Briefly how this is different / similar to the water cycle on Earth.</p> <hr/> <hr/>
<p>Atmosphere: Would the gasses in your atmosphere be visible from space? Draw what you think your planet’s atmosphere would look like.</p> 	<p>Atmosphere: How would the gasses in your planet affect life? (For example, more oxygen means larger organisms)</p> <hr/> <hr/> <hr/>

<p>Biosphere: Write two (2) characteristics of life on your planet. Be consistent with the <i>geosphere</i>, <i>hydrosphere</i>, and <i>atmosphere</i>!</p> <hr/> <hr/>	<p>Biosphere: Draw what a possible organism on your planet could look like</p>
<p>Climate: What is the current climate of your planet (for example, write if there are seasons, or if it stays one temperature)</p> <hr/> <hr/>	<p>Climate: What would the climate look like in 1000 years? Will it be the same or different?</p> <hr/> <hr/>

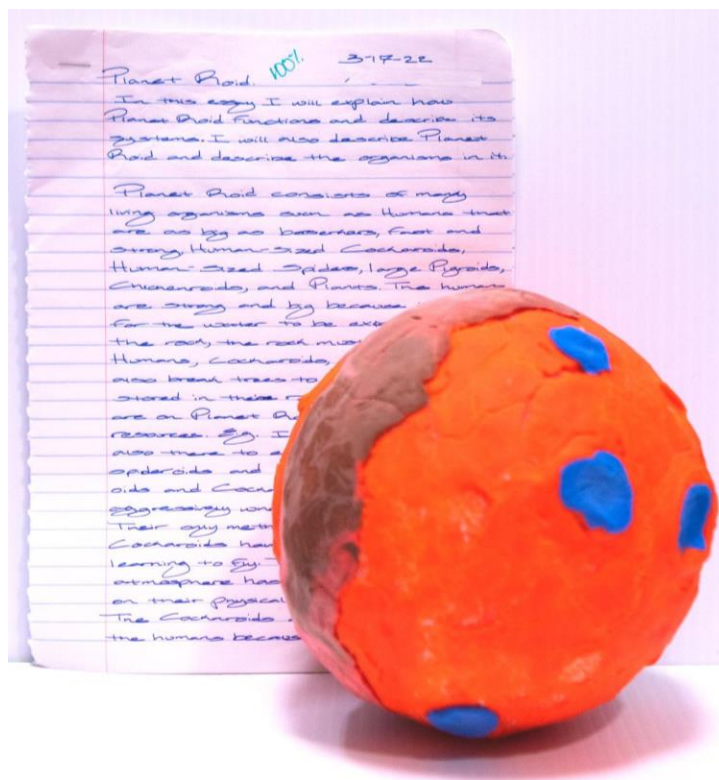


Figure 41: Example final project planet - this desert planet includes liquid water and a single mountain range lining the equator.

Results

Survey

Due to movement within the prison, only five students were able to complete both the pre- and post-class surveys. The results for these five students are shown in percentages, which reflect the answers which the students' responses agreed with expert scientist views. The expert views were taken from the validated CLASS survey results from biology and chemistry, as the questions can be generally applied to earth science. This demonstrated how much their views align with those of scientists, and therefore is a measure of science identity (W. K. Adams et al., 2008). As a whole, the agreement with scientist views increased by 7% throughout the course. This was entirely due to a large increase in Student

4, while the other students' science identity stayed the same or dropped (Figure 42). This is reflective of prior results of the CLASS survey in undergraduate courses (W. K. Adams et al., 2008) and similar other studies across STEM classes (Teichmann et al., 2022). Some proposed explanations for this trend in other studies are gender-based, where male-identifying students are more likely to drive observed decreases (Teichmann et al., 2022), or the desire for students to provide answers that the teachers "wants" in pre-tests, which fades throughout the class (W. Adams et al., 2006). When broken down by question classification, though, there were some more specific trends., Student answers from the "problem solving - confidence" showed that on average, the student's agreement with expert answers decreased by 33.3%. Students 2 and 5 had the same initial and final percent agreement (66.6% and 33.3%, respectively) (Figure 43). (C) Answers from the "problem solving - general" category demonstrated that average agreement increased by 16.7% (Figure 44), while answers from the "real world connections" category increased by 6.67% on average (Figure 45).

From a post-class interview with Jordan Rembold, the students initially had a "big misconception" about science because "it didn't apply to them because of how they grew up, or where they are at now." From this course, she observed that the students "are more open to the idea of learning and working towards a solution, not just looking at a problem and not knowing where to go", and "approach problems differently now for sure, because now they can see that there is not always one right answer, there can be many or there can be no answer at this moment." The trend towards disagreement with experts shown in the

majority of these students may not be representative of learning outcomes in this particular study.

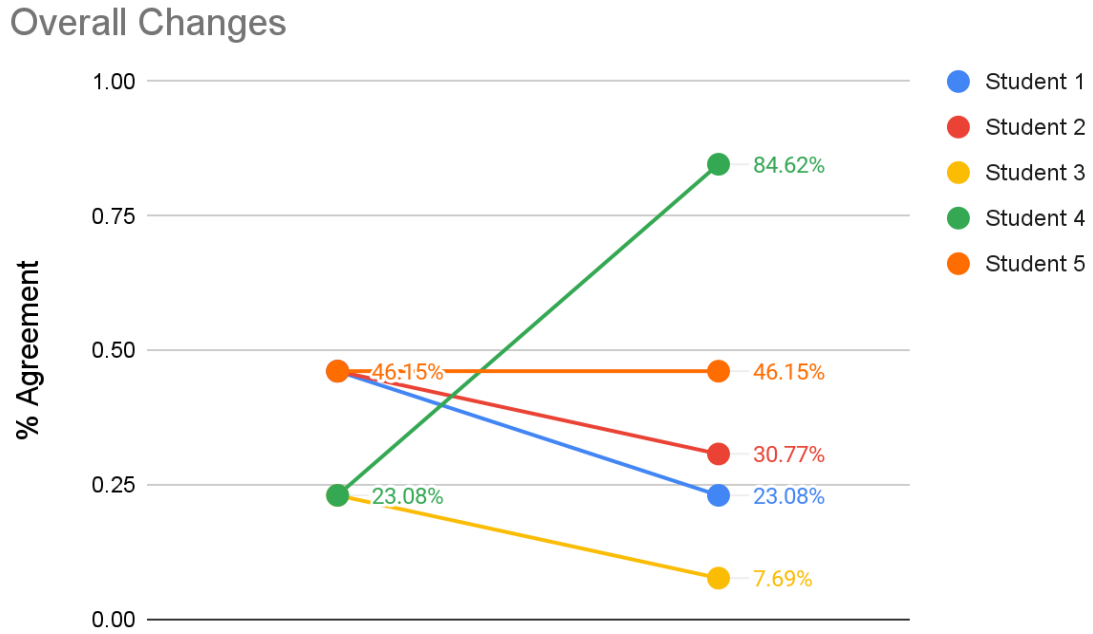


Figure 42: Overall change in Earth Science identity of the five students who completed both pre- and post-class CLASS surveys.

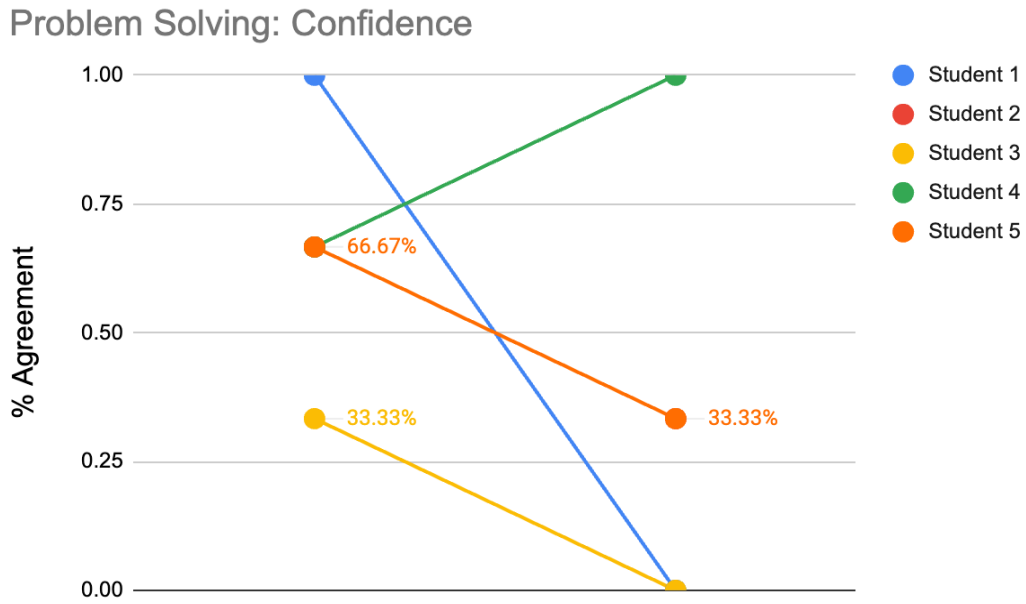


Figure 43: Student answers and changes from the “problem solving - confidence” questions.

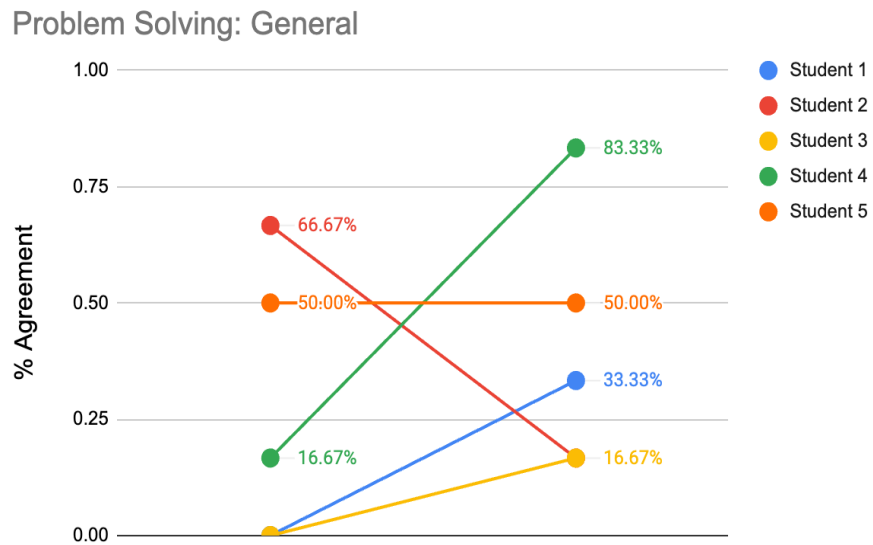


Figure 44: Student answers and changes from the “problem solving - general” questions.

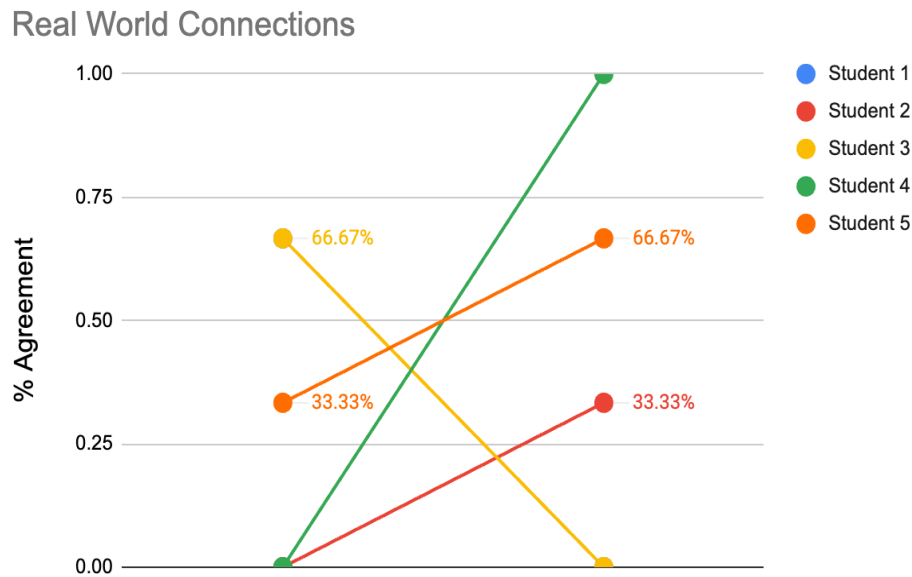


Figure 45: Student answers and changes from the “real-world connections” questions.

Thematic Analysis

The qualitative codes include geologic content knowledge that students learned throughout the course, as well as further extensions of STEM knowledge into different subjects. Three themes that we observed for different STEM subject areas included climate change, which was a focus of the last third of the course, as well as natural (Darwinian) evolution and technological adaptations. We define natural evolution as changes in an organism which solve a problem brought about by other organisms, while technological adaptation is using some form of technology to solve a problem, such as through using tools to access water. We also included the dialogic conversation component of sTc, which was demonstrated through joking and informal language within reports, which highlighted the comfort and trust students felt in the class. We also observed spiritual elements within some of the planet reports, as well as aggressive behaviors of organisms within the student’s planets. (Table 10).

Table 10: Qualitative Analysis Codes.

Codes and Subcodes	Definition
Geologic Content Knowledge	Mention of any geologic features (e.g., mountains, volcanoes, and earthquakes), earth materials (e.g., minerals, elements), erosional agents (e.g., ice, water, wind), and other abiotic agents related to earth science.
Other STEM disciplinary knowledge (e.g., biology, geoconnections to other topics, etc.)	
Climate Change <i>(included within other STEM)</i>	Identification of a changing climate - either cooling or warming - by mentioning relationships between components of the atmosphere, hydrosphere, biosphere, and geosphere.
Natural Evolution <i>(included within other STEM)</i>	Features or traits in organisms which have evolved “naturally” or through technology to adapt to specific environments and/or have influenced the behavior of organisms (both from their own species and other species).
Technology-supported adaptations <i>(included within other STEM)</i>	Technological progress to solve a problem (e.g., converting sweat to water on a dry planet).
Dialogic Conversation	
Informal language and joking / profanity <i>(included within Dialogic Conversation)</i>	Occasions of language that does not traditionally appear in academic writing (e.g., jokes) and profanity
Spirituality <i>(included within Dialogic Conversation)</i>	Identifying elements of the planet which include sacred or spiritual elements (e.g., sacred trees).
Strength / Aggression <i>(included within Dialogic Conversation)</i>	Organisms have traits or behaviors that reflect fighting and aggressive tendencies, or geologic features are potentially named after strength/aggression-related topics (e.g., “Roid” as a planet name, only first mention needs to be coded).

The **geologic content knowledge** code is where students demonstrated knowledge of the geologic content that was discussed in class. The course broadly covered geological topics, from geologic features such as mountains and volcanoes to how abiotic materials like minerals are formed to how different erosional agents (e.g., water) interact with these features and materials. The students referenced many different forms of geology in their reports, such as:

Planet Roid's plate tectonics shift occasionally, forming mountains, volcanoes, and causing roid-quakes. Consequently, it destroys water-contained rocks and trees. The water-contained rocks destroyed create a small lake. The heat then evaporates the lake, which gives the inhabitants of Planet Roid a short amount of time to gather/consume water from the lake. Planet Roid is enormous with many craters. It also has many mountains and dust.

[Essay 03]

The dust particles floating around Planet Roid mix with the H₂O and form mud, H₂O exposed to heat will evaporated a short amount of time, which is why water is usually stored in rocks or trees.

[Essay 03]

Vulcans are a breed of gigantic fire birds that live in volcanoes, they have an armored layer made of obsidian.

[Essay 04]

once every year earthquak's come and shaft the plant so sometime's different part's of the plants spreads and creates new forms in the plant

[Essay 05]

While the course broadly covered earth science, some non-earth science topics frequently emerged within their reports. The first is **climate change**, which was an extensive part of

the course and focused on applications of different earth science topics to the current climate change crisis. The students referenced how the climate of their planets can change quickly, as well as how humanity and pollution is affecting Earth [brackets added by the authors for clarity]:

About 65% of planet CZAR was frozen many years ago, and yes just like the earthlings say it was indeed an ice age. However for some reason, a reason no one quite figure out the ice melted and the water level [dropped].

[Essay 01]

The atmosphere on Planet X is very clean because there are no humans on this planet and **humans tend to harm their planet** and atmosphere just take a look at earth.

[Essay 04]

The first quote highlights how the climate warmed for an unknown reason, causing this planet to transition from an ice age to a warmer climate - even if the water levels dropped, which does not correlate with melting ice. The second quote is more direct as a criticism of anthropogenic climate change, where this student specifically blames humans for harming the Earth and uses their planet as a comparison. Their planet has no humans, and therefore is not affected by the pollution and large-scale destruction caused by them.

The topic of **biological evolution** also appeared throughout the reports. This was not a formal topic in class, but it was mentioned periodically throughout discussions in order to understand how living systems emerged and adapted to various geological events. The code definition focused on where students used evolution to solve problems or describe how their organisms interacted with other types of biological creatures:

Around wintertime is when temperatures tend to drop drastically from 15 degrees to below zero this causes Vulcans to hibernate in their volcanoes because they are very sensitive to the cold weather. This is also the time that **the Krotocanoes hunt for food because they don't have to worry about the Vulcans hunting them.**

[Essay 04]

sometime's different part's of the plants spreads and [creates] new forms in the plant but over time [trees and mountains] start to disintegrate due to the cold weather that go under 115- degrees [Celsius] so when earthquakes come there all ready-frozen so when it hit they just fall apart but they leave seed that grow and rebuild the nature.

[Essay 05]

Both quotes highlight evolution through the predator-prey relationship between the Vulcans and Krotocanoes and the life cycle of plants, but also make geologic concepts an integral part of how these organisms evolve. In both, winter plays a vital role - through removing a predator from the ecosystem and allowing a prey species to search for food in turn in the first quote, and second through providing the means for seeding new ground in the second. The seasonal aspect of both planets demonstrates that the students were able to integrate weather cycles and atmospheric concepts into evolution, a topic that was not specifically taught as part of this class.

Additionally, the organisms in these reports solved problems through **technological adaptations**. This can be thought of as an engineering theme which explores how students approached and solved a problem that their organisms were facing using some form of

technology instead of using biological evolution to adapt their organisms to fit their world.

Some examples of these problems and their technology-based solutions are:

Unexpected bacteria in the swamps and ponds could make you very sick **if you don't sterilize water.**

[Essay 01]

If we go and [visit] my planet we will have to bring [our] own oxygen.

[Essay 02]

The humans, however, have **enhanced technology that can convert the body's sweat into H₂O.** Humans of Planet Roid also extract water from mud.

[Essay 03]

All three of these quotes solved problems that the students imagined, then created solutions for. In the first example, bacteria in the water needs to be sterilized in order to be clean, showing a control of fire and technology in order to solve the problem of finding clean water. The second shows a clear link to atmospheric science, where the student recognizes that their planet does not have oxygen in the atmosphere and takes care to mention that if humans visit, we need to solve the problem of breathing. Their solution of bringing oxygen assumes some level of technology to 1) transport breathable oxygen, and 2) a mechanism for breathing oxygen. The third quote, which comes from a planet with a hot, dry environment, shows how humans have adapted to live in this harsh geology through recycling and extracting water.

We also found three themes relating to the dialogic conversation element of sTc. While the questions on the reports were primarily on scientific knowledge and the interactions

between organisms, the students were able to show their trust in us as teachers and bring in their own real-world experiences and thoughts into the project through 1) **informal language, joking, and profanity**, 2) **spirituality**, and 3) **strength & aggression**.

The **informal language, joking, and profanity** involves students writing in language that normally does not fit within a traditional science classroom. We did not police this informal language and profanity within the classroom because it allowed the students to more freely express themselves and their thoughts during discussions and written assignments. This highlights how this classroom diverges from a traditional science classroom, where this kind of language would likely be heavily censored.

The swamps are something else though. They say a world without crime is probably a boring blank world. Well if you disagree then never cross the swamp lands, the hippies are like best friends with the damn gators it's insane.

[Essay 01]

However I need to be very specific **there are no mosquitoes on planet CZAR there is no use for them.**

[Essay 01]

You name it, we got it, but **let me remind Planet CZAR has zero, natha, zilch, no mosquitoes none what so ever.** Planet CZAR in earth terms is one big fat country. You don't need a "pass port" to travel Planet CZAR. Come explore Planet CZAR you never know you may never leave.

[Essay 01]

The first quote highlights a silly anecdote which is not related to the swamp ecosystem of this planet. Rather, it deviates from the expected description of a swamp and instead gives a description of how well the "hippies" and gators get along, along with some profanity

associated with the gators. The second and third quotes are linked by the author's likely hatred of mosquitos, as this student used a running joke about ensuring their planet has no mosquitos throughout their essay (the second quote is found near the beginning, while the third quote is the final paragraph). The infomercial-style third quote also is a humorous way to end the essay.

Some reports also brought **spiritual** aspects in as well. We did not explicitly mention religion as part of the course, but it was an integral part of how some of the student's planets functioned, such as:

Across this jungle terrain in the heart of the land lies the sacred tree Balsion - (Bal-si-on). The Balsion is what gives life to Planet X from the creatures to the plants this tree is protected by a force that won't allow any harm to it. **The tree also heals plate techtonics so that this massive continent won't separate into a bunch of smaller continents** and of course this would happen because volcano eruptions tend to cause earthquakes which cause plate techtonic shifts.

[Essay 02]

So in zooland the hydrosphere in the [planet contains] water that [is] able to cutain healin and gives you energy.

[Essay 04]

The tree in the first quote heals the planet through interacting with its geology. The sacred Balsion tree not only works to heal the planet and prevent catastrophic earthquakes and volcanoes, but also demonstrates how this student has a deep understanding of geology where plate tectonics lead to continental drift. This knowledge is integrated into their planet and explained through a spiritual lens. The second quote also contains healing elements,

this time through the spiritual idea of healing water. This could show the necessity of water in biology and life on Earth but is not directly linked to biology in the same way the first quote is directly linked to geology.

Finally, many of the student's reports contained instances of **strength and aggression** shown through in various ways, such as through violence between different organisms, between humans, and within disease:

It's too dangerous to catch fish in the king shark's territory because the king shark is the peace keeper of the ocean **eating what doesn't belong.**

[Essay 01]

The spideroids and cockaroids tend to act aggressively when feeling threatened. **Their only method of survival is violence.**

[Essay 03]

The cockaroids and spideroids hate the humans **because they experiment on the cockaroids and spideroids. Also because they invaded their home.**

[Essay 03]

zefora is a type of [virus] that is air born so when creates get the [virus] they go savage but after a [certain] time it goes away.

[Essay 05]

The appearance of these quotes may be a result of the power dynamics within the prison environment, where these students are at the lowest rung of the power structure. Their teachers, guards, volunteers and other ADCRR employees all are part of a structure which removes power from the students. The aggressive tendencies found within the organisms

and humans within these reports may be a way to push against this power dynamic and use imaginary characters to act in ways that they cannot. This school project is one of the few places where these students have freedom and power to create and manipulate their world, so this is a natural place for the students to express themselves within the bounds of what they are allowed to do.

Additionally, the final quote may reflect the nature of living through COVID-19 within such a controlled environment and power structure. This class took place in Spring 2022, when the first Omicron variant wave was beginning to subside in Arizona. Social distancing was not possible in the prison, and the real health concerns of COVID-19 and the impact on the students cannot be ignored.

Limitations

The prison environment was not without its challenges. Every activity which included outside materials needed to be fully vetted and approved by the ADCRR, and all materials were required to be counted before and after an activity. Additionally, the students were permitted to only have pencil and paper - while we could show videos on TV screens to the entire class, individualized technology was not permitted and could not be used as part of this class.

Outside of classroom materials, the class was subject to facility rules and lockdowns. This was not common, but two classes over the span of the quarter were pushed back due to lockdowns. The interruptions due to the prison environment necessitated day-to-day changes to the schedule. Overall, we were able to complete the entire curriculum and complete the course, but some weeks were interrupted due to factors outside of our control. These factors also extended to the students. While over 20 students were impacted as part of this class, only five were present from the beginning of the 12-week class to the end. This is due to a variety of reasons, such as students turning 18 and moving to adult facilities, students being released, or students entering the prison in the middle of the course. This required that lessons include background material that had been covered previously or rely solely on critical thinking without much required knowledge - with students coming and going throughout the course, all with different academic backgrounds, it was nearly impossible to assume background knowledge, even of course activities done weeks previously.

Discussion

This work presents an established structure for teaching Earth Science in a unique environment and provides resources so that other educators working in prison environments or elsewhere can incorporate sTc into future lessons and curricula. The students included sTc frameworks and earth-science specific topics into their reports, highlighting how the curriculum developed here was able to influence the students' learning and thinking about earth science.

The quantitative survey data does not show the same influence, as four of the five students surveyed did not show an increase in science identity. However, my personal experience in the classroom was that throughout the 12-week class, the students became more curious, engaged with myself and guest speakers more, and were able to more closely related earth science to their personal experiences. From an interview with Jordan Rembold, the student's education has "opened their eyes to a new world that they would have never been exposed to in the outside world at a public school." One specific activity that seemed to stick with the students was touching and interacting with different types of rocks - some students "say that looking at the different rocks was a new experience that they had never had and never realized how interesting rocks and the Earth can be." This activity, and the lessons throughout the curricula, were designed to build that curiosity and personal connection, and from our observations in the classroom, these lessons were able to achieve that goal.

Current and future work on this project involves expanding it to different subjects and continuing the partnership between ASU and ADCRR. While it was initially set up and launched as a self-contained 12-week science course, this class has many themes and topics that can be expanded in other subjects, such as including culturally specific stories of historical geologic events in English class or creating topological maps as part of History. Also, the students consistently requested additional graduate student volunteer instructors throughout the course. Future planning will involve more graduate students and train them to build their own lessons and modules within the sTc framework, so that they can incorporate sTc over multiple lessons and allow the juvenile students to have a more in-depth understanding of the research that they do.

5: CONCLUDING DISCUSSION

To definitively find life beyond our Earth, astrobiology must undergo a fundamental paradigm shift. It is imperative that researchers reach a consensus on what life can look like elsewhere. In my dissertation, I propose that assembly theory represents this crucial paradigm shift, capable of enabling the discovery of life within the professional lifetimes of today's scientists, provided that life indeed exists elsewhere in our Solar System. Moreover, it is important to note that assembly theory holds value beyond its potential in discovering life. This theory sheds light on the factors driving the complexification of societies, which could include the intriguing possibility of understanding the societal structures of extraterrestrial civilizations. By applying assembly theory, researchers could unravel the forces that shape the evolution of societies and foster a deeper understanding of our place in the universe.

Additionally, the utility of assembly theory extends beyond the realm of scientific discovery. It can be employed as an invaluable tool for educators to teach about the field of astrobiology and its connection to society. By incorporating assembly theory into sTc-driven curricula, instructors can effectively engage students in an inclusive, culturally sensitive manner and cultivate a broader appreciation for the implications of astrobiology beyond the boundaries of traditional scientific disciplines.

The second chapter in my dissertation focuses on the application of assembly theory to space missions. Most space missions do not focus on directly discovering – for example,

the upcoming Europa Clipper mission is not classified as a life detection mission. Rather, the mission objectives are focused on assessing the potential habitability of Europa. This is consistent across nearly all recent space missions, with the exception of the Viking rover missions in the 1970s, which were categorized as life detection missions, but ultimately did not discover life. Assembly theory, by contrast, is explicitly designed to feature as the core component of future life detection missions, providing that mass spectrometry technology is sufficiently precise to distinguish individual molecules. The engineering goals specified within my dissertation provides NASA with specific, tangible measurements that will definitively detect molecules created as a part of a living system.

The second chapter of my dissertation delves into the practical application of assembly theory in the context of space missions. It is noteworthy that the primary objectives of most space missions do not directly revolve around the search for extraterrestrial life (Neveu et al., 2018). A case in point is the forthcoming Europa Clipper mission, which is not classified as a life detection mission, but instead aims to only assess the potential habitability of Jupiter's moon, Europa (Pappalardo et al., 2015). This mission's focus on habitability assessment aligns with the overarching trend observed in recent space missions. Aside from the Viking rover missions in the 1970s, which were explicitly categorized as life detection missions, most contemporary missions have not pursued direct life discovery (Dick, 2006). It is worth noting that despite the efforts of the Viking missions, they did not yield conclusive evidence of life on Mars. Considering the de-emphasis on life detection missions, assembly theory is designed explicitly to serve as the

core framework for future missions which emphatically focus on life detection, provided that the advances in mass spectrometry technology recommended by my dissertation are made.

Beyond providing NASA with specific engineering goals, the mass spectrometer recommendations in my dissertation are useful for life detection because it will give a binary yes/no answer to if there is life elsewhere on the timeline of current scientific careers. Current means of searching for life elsewhere cannot give a binary answer, as highlighted by the continued detection and arguments surround phosphine on Venus highlight. Additionally, some theories regarding intelligent life cannot be resolved within the lifetime of current scientists. The Search for Extraterrestrial Intelligence (SETI) and similar organizations look for signs of intelligent life that humanity could potentially communicate with. However, given the vastness of space and the limits of today's communication technology, as well as the lack of evidence of actively communicating civilizations, it is unlikely that any such effort will result in the discover of life elsewhere. Simon Conway Morris backs up this idea that intelligent life is difficult to find in his book "Life's Solution: Inevitable Humans in a Lonely Universe", where he argues that life is likely common, but intelligent life that humanity can interact with is rare, if present at all (Morris, 2003). An alternate hypothesis is the Dark Forest Hypothesis, which is featured prominently in Cixin Liu's The Dark Forest trilogy (C. Liu, 2015). This hypothesis uses game theory to argue that communicating on a galactic level is ultimately an antagonistic and fatal process. Regardless of the various search methods and hypotheses, assembly

theory using mass spectrometers is nearly ready to be used in life detection missions, giving scientists a testable means of answering the question of if there is life elsewhere in the solar system.

The third chapter in my dissertation is more theoretical in nature, as it discovers patterns of complexity within the extremely wide space of chemistry within patents. The study of patent chemistry draws from many different fields of study. These include experimental chemistry and medical research, where public and private funding is available to develop new compounds, to sociology, as patents are discovered and registered in a competitive, social environment. Finding overarching, meaningful patterns within such a broad field of study is difficult, but through multiple avenues of research – network science & assembly theory – I found that the organization of individual agents within the invention of patents has a strong effect on the complexity of the compounds found within them. As a whole, the molecular assembly index of chemical compounds used within patents increases over time. However, different agents operate at different levels of inventions. Individual patent authors work usually work in small teams, either as part of research groups in academia or within dedicated companies. These authors have very little effect on the increase of patents – that is, on average individual authors do not invent or work with increasingly more complex compounds over the course of their careers. This intuitively makes sense, as individual scientists may have a defined skill set when it comes to chemical reactions and molecules and not branch out from that skill set to discover different, more complex, molecules.

Assignees and social classifications, in contrast, have more of an effect on complexity. On average, the organization responsible for patents (the assignee) sees an increase in MA of compounds over time within its patents. Potentially, assignees can pivot and respond to market forces, funding sources, and other societal pressures. These societal pressures may be responsible for driving increases in complexity. This increase is seen on an even higher level, where compounds within a particular patent classification provided by the USPTO increase over time at an even greater rate than assignees. Since classifications demonstrate a higher rate of increase, it would be interesting to explore how interest or funding within different classifications – measured through grants, investment, or stock market growth - affects complexification of compounds. This hypothesis is not covered within this dissertation, but I believe this area of research could be fruitful in predicting future increases in complexity in chemistry.

There is also an astrobiological bent to increasing complexity in patent chemistry. In order to communicate elsewhere in the universe, a civilization most likely needs a high level of technology. This could involve metalworking, radio waves, and so on – essentially, this civilization will need to be industrialized in order to communicate outside its home solar system. Given the potential universality of assembly theory as a biosignature (Marshall et al., 2021), as well as the commonalities shared among cities and industrial processes (Bettencourt et al., 2007; West, 2018), it is possible that extraterrestrial societies would exhibit the same complexification patterns demonstrated here.

The educational research presented in this dissertation provides a wrapper for the astrobiological research I completed. Education is essential for science, as without it, there is no shared understanding and appreciation for scientific discoveries. This is particularly true in the field of astrobiology, as the impacts of discovering alien life are ripe for misinterpretation. It is easy to imagine a scenario where life is definitively discovered elsewhere, but the hard science behind its discovery is incomprehensible to the average non-scientist. In this scenario, the news of such a discovery is filtered through various science communicators, reporters, and likely other non-scientists on social media channels. The critical thinking and understanding required to parse this wide variety of information (and potential misinformation) comes from a holistic and engaging education in science, which is where the application of sociotransformative constructivism to astrobiology is useful. My dissertation gives a resource that demonstrates that sTc can be effectively applied to astrobiology through providing educators with an example of how to use sTc within a science classroom, including lessons on astrobiology.

Taken together, my research shown here demonstrates the use of assembly theory to both future life detection missions and the complexification of society, but also the importance of education in scientific understanding.

REFERENCES

- Adams, W. K., Wieman, C. E., Perkins, K. K., & Barbera, J. (2008). Modifying and Validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry. *Journal of Chemical Education*, 85(10), 1435–1439.
- Adams, W., Perkins, K., Podolefsky, N., Dubson, M., Finkelstein, N., & Wieman, C. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics - Physics Education Research*, 2(1), 01–14.
- Akhondi, S. A., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M., Gregory, M., & Kors, J. A. (2019). Automatic identification of relevant chemical compounds from patents. *Database: The Journal of Biological Databases and Curation*, 1–14.
- Amineh, R. J., & Asl, H. D. (2015). Review of constructivism and social constructivism. *Journal of Social Sciences, Literature and Languages*, 1(1), 9–16.
- Amis, M. (1995). *The Information*. Harmony Books.
- Anbar, A. D., Duan, Y., Lyons, T. W., Arnold, G. L., Kendall, B., Creaser, R. A., Kaufman, A. J., Gordon, G. W., Scott, C., Garvin, J., & Buick, R. (2007). A whiff of oxygen before the great oxidation event? *Science*, 317(5846), 1903–1906.
- Anton, J. J., & Yao, D. A. (2004). Little Patents and Big Secrets: Managing Intellectual Property. *The Rand Journal of Economics*, 35(1), 1–22.
- Arevalo, R., Jr, Ni, Z., & Danell, R. M. (2020). Mass spectrometry and planetary exploration: A brief review and future projection. *Journal of Mass Spectrometry: JMS*, 55(1), 4454–4471.
- Arevalo, R., Jr, Selliez, L., Briois, C., Carrasco, N., Thirkell, L., Cherville, B., Colin, F., Gaubicher, B., Farcy, B., Li, X., & Makarov, A. (2018). An Orbitrap-based laser desorption/ablation mass spectrometer designed for spaceflight. *Rapid Communications in Mass Spectrometry: RCM*, 32(21), 1875–1886.
- Arnold, F. H. (2019). Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angewandte Chemie*, 58(41), 14420–14426.
- Aronson, B., & Laughter, J. (2016). The Theory and Practice of Culturally Relevant Education: A Synthesis of Research Across Content Areas. *Review of Educational Research*, 86(1), 163–206.
- Asche, G. (2017). “80% of technical information found only in patents” – Is there proof

- of this [1]? *World Patent Information*, 48, 16–28.
- Aslan, C. E., Souther, S., Stortz, S., Sample, M., Sandor, M., Levine, C., Samberg, L., Gray, M., & Dickson, B. (2021). Land management objectives and activities in the face of projected fire regime change in the Sonoran desert. *Journal of Environmental Management*, 280, 111644–111654.
- Avsar Erumit, B., Akerson, V. L., & Buck, G. A. (2021). Multiculturalism in higher education: experiences of international teaching assistants and their students in science and math classrooms. *Cultural Studies of Science Education*, 16(1), 251–278.
- Awale, M., Visini, R., Probst, D., Arús-Pous, J., & Reymond, J.-L. (2017). Chemical Space: Big Data Challenge for Molecular Diversity. *Chimia*, 71(10), 661–666.
- Bains, W. (2004). Many chemistries could be used to build living systems. *Astrobiology*, 4(2), 137–167.
- Bains, W., Petkowski, J. J., Zhan, Z., & Seager, S. (2021). Evaluating Alternatives to Water as Solvents for Life: The Example of Sulfuric Acid. *Life*, 11(5), 400–434.
- Bakhtin, M. M. (2010). *The Dialogic Imagination: Four Essays*. University of Texas Press.
- Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- Barabási, A.-L. (2013). *Network Science* (S. Capelin & R. M. Munnell, Eds.). Cambridge University Press.
- Baranger, J., Rousseau, D., Mastrorilli, M. E., & Matesanz, J. (2018). Doing Time Wisely: The Social and Personal Benefits of Higher Education in Prison. *The Prison Journal*, 98(4), 490–513.
- Barkai, N., & Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, 387(6636), 913–917.
- Barton, A. C., & Tan, E. (2010). We Be Burnin’! Agency, Identity, and Science Learning. *Journal of the Learning Sciences*, 19(2), 187–229.
- Bertz, S. H. (1981). The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12), 3599–3601.
- Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17), 7301–7306.

- Billings, L. (2012). Astrobiology in culture: the search for extraterrestrial life as “science.” *Astrobiology*, *12*(10), 966–975.
- Bishop, D. M., & Frazier, C. E. (1988). The Influence of Race in Juvenile Justice Processing. *The Journal of Research in Crime and Delinquency*, *25*(3), 242–263.
- Blackmond, D. G. (2010). The origin of biological homochirality. *Cold Spring Harbor Perspectives in Biology*, *2*(5), 002147–002165.
- Böttcher, T. (2016). An Additive Definition of Molecular Complexity. *Journal of Chemical Information and Modeling*, *56*(3), 462–470.
- Bowman, A. P., Blakney, G. T., Hendrickson, C. L., Ellis, S. R., Heeren, R. M. A., & Smith, D. F. (2020). Ultra-High Mass Resolving Power, Mass Accuracy, and Dynamic Range MALDI Mass Spectrometry Imaging by 21-T FT-ICR MS. *Analytical Chemistry*, *92*(4), 3133–3142.
- Bradley, E., & Kantz, H. (2015). Nonlinear time-series analysis revisited. *Chaos*, *25*(9), 097610–097635.
- Brake, M., & Hook, N. (2007). Darwin to the double helix: astrobiology in fiction. *International Journal of Astrobiology*, *6*(4), 273–280.
- Bregonje, M. (2005). Patents: A unique source for scientific technical information in chemistry related industry? *World Patent Information*, *27*(4), 309–315.
- Brockwell, T. G., Meech, K. J., Pickens, K., Waite, J. H., Miller, G., Roberts, J., Lunine, J. I., & Wilson, P. (2016). The mass spectrometer for planetary exploration (MASPEX). *2016 IEEE Aerospace Conference*, 1–17.
- Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, *10*(1), 1017–1027.
- Bromberg, Y., Aptekmann, A. A., Mahlich, Y., Cook, L., Senn, S., Miller, M., Nanda, V., Ferreiro, D. U., & Falkowski, P. G. (2022). Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer. *Science Advances*, *8*(2), 3984–3997.
- Brooks, W. H., Guida, W. C., & Daniel, K. G. (2011). The significance of chirality in drug design and development. *Current Topics in Medicinal Chemistry*, *11*(7), 760–770.
- Butler, A., & Sinclair, K. A. (2020). Place Matters: A Critical Review of Place Inquiry and Spatial Methods in Education Research. *Review of Research in Education*, *44*(1), 64–96.

- Chambers, B. (2015). *The Long Way to a Small, Angry Planet: Wayfarers 1*. Hodder & Stoughton.
- Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Eloë-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research*, *47*(D1), D666–D677.
- Chon, O. (2021). Astrobioethics: Epistemological, Astrotheological, and Interplanetary Issues. *Astrobiology: Science, Ethics, and Public Policy*, 1–16.
- Chong, Y. T., Mohd Ariffin, M., Mohd Tahir, N., & Loh, S. H. (2018). A green solvent holder in electro-mediated microextraction for the extraction of phenols in water. *Talanta*, *176*, 558–564.
- Chou, L., Grefenstette, N., Johnson, S. S., Graham, H., Mahaffy, P., Kempes, C., Elsila, J. E., Libby, E., Ellington, A., Anslyn, E., Hoehler, T., Girguis, P., Cronin, L., Brinkerhoff, W., & Lollar, B. S. (2021). Towards a more universal life detection strategy. *Bulletin of the American Astronomical Society*, *53*(4), 1–9.
- Chou, L., Mahaffy, P., Trainer, M., Eigenbrode, J., Arevalo, R., Brinkerhoff, W., Getty, S., Grefenstette, N., Da Poian, V., Fricke, G. M., Kempes, C. P., Marlow, J., Sherwood Lollar, B., Graham, H., & Johnson, S. S. (2021). Planetary Mass Spectrometry for Agnostic Life Detection in the Solar System. *Frontiers in Astronomy and Space Sciences*, *8*, 755100–755120.
- Christensen, P., Canup, R., & Smith, D. H. (2022). *Report on the National Academies' Decadal Survey on Planetary Science and Astrobiology*. 44.
- Christodoulou, D., Lev, B., & Ma, L. (2018). The productivity of Chinese patents: The role of business area and ownership type. *International Journal of Production Economics*, *199*, 107–124.
- Clarke, J. (2020). Speculative Oceanography: An Historical Survey of the Literature. *Advances in Oceanography & Marine Biology*, *2*(1), 1–3.
- Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *Society for Industrial and Applied Mathematics*, *51*(4), 661–683.
- Cleland, C. E. (2019). *The Quest for a Universal Theory of Life: Searching for Life As We Don't Know It*. Cambridge University Press.
- Cockell, C. S., Wordsworth, R., Whiteford, N., & Higgins, P. M. (2021). Minimum Units of Habitability and Their Abundance in the Universe. *Astrobiology*, *21*(4), 481–

489.

- Coley, C. W., Green, W. H., & Jensen, K. F. (2018). Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research*, 51(5), 1281–1289.
- Collins, R., & Blau, P. M. (1979). Inequality and heterogeneity: A primitive theory of social structure. *Social Forces; a Scientific Medium of Social Study and Interpretation*, 58(2), 677–683.
- Cooper, G. J. T., Surman, A. J., McIver, J., Colón-Santos, S. M., Gromski, P. S., Buchwald, S., Suárez Marina, I., & Cronin, L. (2017). Miller-Urey spark-discharge experiments in the deuterium world. *Angewandte Chemie*, 56(28), 8079–8082.
- Courtney, J. A. (2019). The Relationship Between Prison Education Programs and Misconduct. *Journal of Correctional Education (1974-)*, 70(3), 43–59.
- Cronin, L., & Walker, S. I. (2016). Beyond prebiotic chemistry. *Science*, 352(6290), 1174–1175.
- Csardi, G., Nepusz, T., & Others. (2006). The igraph software package for complex network research. *Complex Systems*, 1695(5), 1–9.
- Darwin, C. R. (1872). *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. Gale and the British Library.
- de Marco, B. A., Rechelo, B. S., Tófoli, E. G., Kogawa, A. C., & Salgado, H. R. N. (2019). Evolution of green chemistry and its multidimensional impacts: A review. *Saudi Pharmaceutical Journal : SPJ : The Official Publication of the Saudi Pharmaceutical Society*, 27(1), 1–8.
- Deamer, D. (2017). The Role of Lipid Membranes in Life's Origin. *Life*, 7(1), 5–12.
- Derry, T. K., & Williams, T. I. (1960). *A Short History of Technology from the Earliest Times to A.D. 1900*. Courier Corporation.
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences of the United States of America*, 116(39), 19251–19257.
- Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V. D., & Barabási, A.-L. (2014). Career on the move: geography, stratification, and scientific impact. *Scientific Reports*, 4, 4770–4777.

- Dick, S. J. (2006). NASA and the search for life in the universe. *Endeavour*, 30(2), 71–75.
- Dick, S. J. (2012). Critical issues in the history, philosophy, and sociology of astrobiology. *Astrobiology*, 12(10), 906–927.
- Dobson, C. M. (2004). Chemical space and biology. *Nature*, 432(7019), 824–828.
- Domagal-Goldman, S. D., Wright, K. E., Adamala, K., Arina de la Rubia, L., Bond, J., Dartnell, L. R., Goldman, A. D., Lynch, K., Naud, M.-E., Paulino-Lima, I. G., Singer, K., Walther-Antonio, M., Abrevaya, X. C., Anderson, R., Arney, G., Atri, D., Azúa-Bustos, A., Bowman, J. S., Brazelton, W. J., ... Wong, T. (2016). The Astrobiology Primer v2.0. *Astrobiology*, 16(8), 561–653.
- Dou, R., Hazari, Z., Dabney, K., Sonnert, G., & Sadler, P. (2019). Early informal STEM experiences and STEM identity: The importance of talking science. *Science Education*, 103(3), 623–637.
- Edler, J., & Fagerberg, J. (2017). Innovation policy: what, why, and how. *Oxford Review of Economic Policy*, 33(1), 2–23.
- Edwards-Jones, A. (2014). Qualitative data analysis with NVIVO. *Journal of Education for Teaching*, 40(2), 193–195.
- Elfiky, A., & Ibrahim, N. S. (2020). Anti-SARS and anti-HCV drugs repurposing against the Papain-like protease of the newly emerged coronavirus (2019-nCoV). In *Research Square*. <https://doi.org/10.21203/rs.2.23280/v1>
- Ellison, M., Szifris, K., Horan, R., & Fox, C. (2017). A Rapid Evidence Assessment of the effectiveness of prison education in reducing recidivism and increasing employment. *Probation Journal*, 64(2), 108–128.
- Esperian, J. H. (2010). The Effect of Prison Education Programs on Recidivism. *Journal of Correctional Education*, 61(4), 316–334.
- Fabelo, T. (2002). The Impact of Prison Education on Community Reintegration of Inmates: The Texas Case. *Journal of Correctional Education*, 53(3), 106–110.
- Falaguera, M. J., & Mestres, J. (2021). Identification of the Core Chemical Structure in SureChEMBL Patents. *Journal of Chemical Information and Modeling*, 61(5), 2241–2247.
- Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, 40(DI), D136-143.
- Fergusson, J., Oliver, C., & Walter, M. R. (2012). Astrobiology outreach and the nature

- of science: the role of creativity. *Astrobiology*, 12(12), 1143–1153.
- Fisher, T., Kim, H., Millsaps, C., & Line, M. (2022). Inferring exoplanet disequilibria with multivariate information in atmospheric reaction networks. *The Astronomical Journal*, 164(53), 1–40.
- Flammarion, C. (1980). *The Plurality of Inhabited Worlds* (R. L. Jones, Trans.). Robert L. Jones, Jr.
- Florindo, C., Branco, L. C., & Marrucho, I. M. (2019). Quest for green-solvent design: From hydrophilic to hydrophobic (deep) eutectic solvents. *ChemSusChem*, 12(8), 1549–1559.
- Flynn, N., & Higdon, R. (2022). Prison Education: Beyond Review and Evaluation. *The Prison Journal*, 102(2), 196–216.
- Föhn, M., Galli, A., Vorburger, A., Tulej, M., Lasi, D., Riedo, A., Fausch, R. G., Althaus, M., Brüngger, S., Fahrer, P., Gerber, M., Lüthi, M., Munz, H. P., Oeschger, S., Piazza, D., & Wurz, P. (2021). Description of the Mass Spectrometer for the Jupiter Icy Moons Explorer Mission. *2021 IEEE Aerospace Conference (50100)*, 1–14.
- Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. *American Sociological Review*, 80(5), 875–908.
- G Marshall, A., T Blakney, G., Chen, T., K Kaiser, N., M McKenna, A., P Rodgers, R., M Ruddy, B., & Xian, F. (2013). Mass resolution and mass accuracy: how much is enough? *Mass Spectrometry*, 2(Spec Iss), S0009-S0014.
- Gaes, G. G. (2008). The impact of prison education programs on post-release outcomes. *Reentry Roundtable on Education*, 31, 1–31.
- Gałaszka, A., Migaszewski, Z., & Namieśnik, J. (2013). The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices. *Trends in Analytical Chemistry*, 50, 78–84.
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1), D945–D954.
- Gisiger, T. (2001). Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biological Reviews of the Cambridge Philosophical Society*, 76(2), 161–209.

- Gleiser, M., Nelson, B. J., & Walker, S. I. (2012). Chiral polymerization in open systems from chiral-selective reaction rates. *Origins of Life and Evolution of the Biosphere*, 42(4), 333–346.
- Gleiser, M., & Walker, S. I. (2012). Life's chirality from prebiotic environments. *International Journal of Astrobiology*, 11(4), 287–296.
- Goesmann, F., Brinckerhoff, W. B., Raulin, F., Goetz, W., Danell, R. M., Getty, S. A., Siljeström, S., Mißbach, H., Steininger, H., Arevalo, R. D., Jr, Buch, A., Freissinet, C., Grubisic, A., Meierhenrich, U. J., Pinnick, V. T., Stalport, F., Szopa, C., Vago, J. L., Lindner, R., ... van Amerom, F. H. W. (2017). The Mars Organic Molecule Analyzer (MOMA) Instrument: Characterization of Organic Material in Martian Sediments. *Astrobiology*, 17(6–7), 655–685.
- Goldford, J. E., Hartman, H., Marsland, R., 3rd, & Segrè, D. (2019). Environmental boundary conditions for the origin of life converge to an organo-sulfur metabolism. *Nature Ecology & Evolution*, 3(12), 1715–1724.
- Gomollón-Bel, F. (2019). Ten Chemical Innovations That Will Change Our World: IUPAC identifies emerging technologies in Chemistry with potential to make our planet more sustainable. *Chemistry International*, 41(2), 12–17.
- Gosselin, D., Burian, S., Lutz, T., & Maxson, J. (2016). Integrating geoscience into undergraduate education about environment, society, and sustainability using place-based learning: three examples. *Journal of Environmental Studies and Sciences*, 6(3), 531–540.
- Gould, S. J. (2002). *The Structure of Evolutionary Theory*. Harvard University Press.
- Gramlich, J. (2019, April 30). *The gap between the number of blacks and whites in prison is shrinking*. Pew Research Center. <https://www.pewresearch.org/fact-tank/2019/04/30/shrinking-gap-between-number-of-blacks-and-whites-in-prison/>
- Guo, J., Ibanez-Lopez, A. S., Gao, H., Quach, V., Coley, C. W., Jensen, K. F., & Barzilay, R. (2021). Automated Chemical Reaction Extraction from Scientific Literature. *Journal of Chemical Information and Modeling*, 62(9), 2034–2045.
- Guttenberg, N., Chen, H., Mochizuki, T., & Cleaves, H. J., 2nd. (2021). Classification of the Biogenicity of Complex Organic Mixtures for the Detection of Extraterrestrial Life. *Life*, 11(3), 234–255.
- Hähnke, V. D., Bolton, E. E., & Bryant, S. H. (2015). PubChem atom environments. *Journal of Cheminformatics*, 7, 41–78.
- Hall, B. H., & Harhoff, D. (2012). Recent Research on the Economics of Patents. *Annual Review of Economics*, 4(1), 541–565.

- Hantson, S., Huxman, T. E., Kimball, S., Randerson, J. T., & Goulden, M. L. (2021). Warming as a driver of vegetation loss in the sonoran desert of California. *Journal of Geophysical Research. Biogeosciences*, 126(6), 1–15.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362.
- Hays, L., Achenbach, L., Bailey, J., Barnes, R. K., Baross, J. A., Bertka, C., & Wordsworth, R. D. (2015). NASA Astrobiology Strategy 2015. Retrieved from the NAI Website https://nai.nasa.gov/media/medialibrary/2015/10/NASA_Astrobiology_Strategy_2015_15100, 8, 1–256.
- Hazari, Z., Chari, D., Potvin, G., & Brewe, E. (2020). The context dependence of physics identity: Examining the role of performance/competence, recognition, interest, and sense of belonging for lower and upper female physics undergraduates. *Journal of Research in Science Teaching*, 57(10), 1583–1607.
- Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M.-C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *Journal of Research in Science Teaching*, 47(8), 978–1003.
- Hoonlor, A., Szymanski, B. K., & Zaki, M. J. (2013). Trends in computer science research. *Communications of the Association for Computing Machinery*, 56(10), 74–83.
- Hordijk, W., Hein, J., & Steel, M. (2010). Autocatalytic Sets and the Origin of Life. *Entropy*, 12(7), 1733–1742.
- Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–Student Dialogue During Classroom Teaching: Does It Really Impact on Student Outcomes? *Journal of the Learning Sciences*, 28(4–5), 462–512.
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hemsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1, 16048.
- Impey, C. (2021). Astrobiology education: Inspiring diverse audiences with the search for life in the universe. *Astrobiology: Science, Ethics, and Public Policy*, 135–156.
- Irwin, L. N., & Schulze-Makuch, D. (2020). The Astrobiology of Alien Worlds: Known

- and Unknown Forms of Life. *Universe*, 6(9), 130–162.
- Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(4), 567–572.
- Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, 1(4), 1–7.
- Johnson, D. S. (1985). The NP-completeness column: an ongoing guide. *Journal of Algorithms & Computational Technology*, 6(3), 434–451.
- Joshi, D. R., & Adhikari, N. (2019). An Overview on Common Organic Solvents and Their Toxicity. *Journal of Pharmaceutical Research International*, 28(3), 1–18.
- Judson, O. P. (2017). The energy expansions of evolution. *Nature Ecology & Evolution*, 1(6), 138–147.
- Kahana, A., Maslov, S., & Lancet, D. (2021). Dynamic lipid aptamers: non-polymeric chemical path to early life. *Chemical Society Reviews*, 50(21), 11741–11746.
- Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30.
- Kanehisa, Minoru. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science: A Publication of the Protein Society*, 28(11), 1947–1951.
- Ke, Q., Ahn, Y.-Y., & Sugimoto, C. R. (2017). A systematic identification and analysis of scientists on Twitter. *PloS One*, 12(4), e0175368–e0175385.
- Kempes, C. P., Koehl, M. A. R., & West, G. B. (2019). The Scales That Limit: The Physical Boundaries of Evolution. *Frontiers in Ecology and Evolution*, 7, 242–262.
- Kiang, N. Y., Domagal-Goldman, S., Parenteau, M. N., Catling, D. C., Fujii, Y., Meadows, V. S., Schwieterman, E. W., & Walker, S. I. (2018). Exoplanet Biosignatures: At the Dawn of a New Era of Planetary Observations. *Astrobiology*, 18(6), 619–629.
- Kim, H., Smith, H. B., Mathis, C., Raymond, J., & Walker, S. I. (2019). Universal scaling across biochemical networks on Earth. *Science Advances*, 5(1), 149–161.
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1), D1102–D1109.

- Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., & Bryant, S. H. (2016). PubChem Substance and Compound databases. *Nucleic Acids Research*, *44*(D1), D1202-13.
- Kim, S., Thiessen, P. A., Cheng, T., Yu, B., Shoemaker, B. A., Wang, J., Bolton, E. E., Wang, Y., & Bryant, S. H. (2016). Literature information in PubChem: associations between PubChem records and scientific articles. *Journal of Cheminformatics*, *8*, 32–47.
- Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, *7*, 234–244.
- Kind, T., & Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, *8*, 105–125.
- Koga, T., & Naraoka, H. (2017). A new family of extraterrestrial amino acids in the Murchison meteorite. *Scientific Reports*, *7*(1), 636–644.
- Kolb, V. M. (2014). *Astrobiology: An Evolutionary Approach*. CRC Press.
- Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., & Valencia, A. (2017). Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews*, *117*(12), 7673–7761.
- Krupnick, M. (2018, October 2). *After colleges promised to increase it, hiring of black faculty declined*. The Hechinger Report: Covering Innovation & Inequality in Higher Education. <https://hechingerreport.org/after-colleges-promised-to-increase-it-hiring-of-black-faculty-declined/>
- Kubrick, S. (1968). *2001: A Space Odyssey*. Metro-Goldwyn-Mayer.
- Kunegis, J., Blattner, M., & Moser, C. (2013). Preferential attachment in online networks: measurement and explanations. *Proceedings of the 5th Annual ACM Web Science Conference*, 205–214.
- Kurlychek, M. C., & Johnson, B. D. (2019). Cumulative Disadvantage in the American Criminal Justice System. *Annual Review of Criminology*, *2*(1), 291–319.
- Kvenvolden, K., Lawless, J., Pering, K., Peterson, E., Flores, J., Ponnampereuma, C., Kaplan, I. R., & Moore, C. (1970). Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. *Nature*, *228*(5275), 923–926.
- Kyrpides, N., Overbeek, R., & Ouzounis, C. (1999). Universal protein families and the

- functional content of the last universal common ancestor. *Journal of Molecular Evolution*, 49(4), 413–423.
- Lancet, D., Zidovetzki, R., & Markovitch, O. (2018). Systems protobiology: origin of life in lipid catalytic networks. *Journal of the Royal Society, Interface / the Royal Society*, 15(144), 20180159–20180195.
- Landrum, G. (2020). *RDKit*. <https://www.rdkit.org/>
- Lawson, A. J., Swienty-Busch, J., Géoui, T., & Evans, D. (2014). The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information* (Vol. 1164, pp. 127–148). American Chemical Society.
- Li, X., Danell, R. M., Pinnick, V. T., Grubisic, A., van Amerom, F., Arevalo, R. D., Jr, Getty, S. A., Brinckerhoff, W. B., Southard, A. E., Gonnissen, Z. D., & Adachi, T. (2017). Mars Organic Molecule Analyzer (MOMA) laser desorption/ionization source design and performance characterization. *International Journal of Mass Spectrometry*, 422, 177–187.
- Liu, C. (2015). *The Dark Forest*. Macmillan.
- Liu, Y., Mathis, C., Bajczyk, M. D., Marshall, S. M., Wilbraham, L., & Cronin, L. (2021). Exploring and mapping chemical space with molecular assembly trees. *Science Advances*, 7(39), 2465–2475.
- Llanos, E. J., Leal, W., Luu, D. H., Jost, J., Stadler, P. F., & Restrepo, G. (2019). Exploration of the chemical space and its three historical regimes. *Proceedings of the National Academy of Sciences of the United States of America*, 116(26), 12660–12665.
- Lopez, M. J., & Mohiuddin, S. S. (2020). *Biochemistry, essential amino acids*. StatPearls Publishing.
- Luzanov, A. V., & Babich, E. N. (1995). Quantum-chemical quantification of molecular complexity and chirality. *Journal of Molecular Structure: Theory and Computation in Chemistry*, 333(3), 279–290.
- Mahaffy, P. R., Webster, C. R., Cabane, M., Conrad, P. G., Coll, P., Atreya, S. K., Arvey, R., Barciniak, M., Benna, M., Bleacher, L., Brinckerhoff, W. B., Eigenbrode, J. L., Carignan, D., Cascia, M., Chalmers, R. A., Dworkin, J. P., Errigo, T., Everson, P., Franz, H., ... Mumm, E. (2012). The Sample Analysis at Mars Investigation and Instrument Suite. *Space Science Reviews*, 170(1), 401–478.
- Marcheselli, V. (2019). *Life as-we-don't-know-it: research repertoires and the emergence of astrobiology* (J. Calvert & A. Street, Eds.) [Ph.D.]. University of

Edinburgh.

- Marshall, S. M., Mathis, C., Carrick, E., Keenan, G., Cooper, G. J. T., Graham, H., Craven, M., Gromski, P. S., Moore, D. G., Walker, S. I., & Cronin, L. (2021). Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nature Communications*, *12*(1), 3033–3041.
- Marshall, S. M., Moore, D. G., Murray, A. R. G., Walker, S. I., & Cronin, L. (2022). Formalising the Pathways to Life Using Assembly Spaces. *Entropy*, *24*(7), 884–904.
- Marshall, S. M., Murray, A. R. G., & Cronin, L. (2017). A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *375*(2109), 20160342–20160355.
- Martin, W., Baross, J., Kelley, D., & Russell, M. J. (2008). Hydrothermal vents and the origin of life. *Nature Reviews. Microbiology*, *6*(11), 805–814.
- Martins, Z., Botta, O., Fogel, M. L., Sephton, M. A., Glavin, D. P., Watson, J. S., Dworkin, J. P., Schwartz, A. W., & Ehrenfreund, P. (2008). Extraterrestrial nucleobases in the Murchison meteorite. *Earth and Planetary Science Letters*, *270*(1), 130–136.
- Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), 68–78.
- Maulucci, M. S. R. (2012). Social Justice Research in Science Education: Methodologies, Positioning, and Implications for Future Research. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second International Handbook of Science Education* (pp. 583–594). Springer Netherlands.
- McCullom, T. M. (2013). Miller-Urey and Beyond: What Have We Learned About Prebiotic Organic Synthesis Reactions in the Past 60 Years? *Annual Review of Earth and Planetary Sciences*, *41*(1), 207–229.
- Meadows, V., Graham, H., Abrahamsson, V., Adam, Z., Amador-French, E., Arney, G., Barge, L., Barlow, E., Berea, A., Bose, M., Bower, D., Chan, M., Cleaves, J., Corpolongo, A., Currie, M., Domagal-Goldman, S., Dong, C., Eigenbrode, J., Enright, A., ... Young, L. (2022). Community Report from the Biosignatures Standards of Evidence Workshop. In *arXiv [astro-ph.IM]*. arXiv. <http://arxiv.org/abs/2210.14293>
- Méndez, A., Rivera-Valentín, E. G., Schulze-Makuch, D., Filiberto, J., Ramírez, R. M., Wood, T. E., Dávila, A., McKay, C., Ceballos, K. N. O., Jusino-Maldonado, M., Torres-Santiago, N. J., Nery, G., Heller, R., Byrne, P. K., Malaska, M. J., Nathan,

- E., Simões, M. F., Antunes, A., Martínez-Frías, J., ... Haqq-Misra, J. (2021). Habitability Models for Astrobiology. *Astrobiology*, 21(8), 1017–1027.
- Méndez-Lucio, O., & Medina-Franco, J. L. (2017). The many roles of molecular complexity in drug discovery. *Drug Discovery Today*, 22(1), 120–126.
- Merder, J., Freund, J. A., Feudel, U., Niggemann, J., Singer, G., & Dittmar, T. (2020). Improved Mass Accuracy and Isotope Confirmation through Alignment of Ultrahigh-Resolution Mass Spectra of Complex Natural Mixtures. *Analytical Chemistry*, 92(3), 2558–2565.
- Miller, S. L., & Urey, H. C. (1959). Organic compound synthesis on the primitive earth. *Science*, 130(3370), 245–251.
- Morris, S. C. (2003). *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge University Press.
- Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing Research*, 40(2), 120–123.
- Moser, P. (2007). *Why Don't Inventors Patent?* (No. 13294). National Bureau of Economic Research . <http://www.nber.org/papers/w13294>
- Mullard, A. (2017). The drug-maker's guide to the galaxy. *Nature*, 549, 445–447.
- National Academies of Sciences, Engineering, & Medicine. (2019). *An Astrobiology Strategy for the Search for Life in the Universe*. The National Academies Press.
- Nekola Nováková, J. (2020). *Strangest of All: An Anthology of Astrobiological Science Fiction* (pp. EPSC2020-1010). European Astrobiology Institute.
- Neveu, M., Hays, L. E., Voytek, M. A., New, M. H., & Schulte, M. D. (2018). The Ladder of Life Detection. *Astrobiology*, 18(11), 1375–1402.
- Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 64(2 Pt 2), 025102.
- Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How We Do Things With Words: Analyzing Text as Social and Cultural Data. *Frontiers in Artificial Intelligence*, 3, 62.
- Noor, E., Eden, E., Milo, R., & Alon, U. (2010). Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Molecular Cell*, 39(5), 809–820.

- Núñez, A.-M., Rivera, J., & Hallmark, T. (2020). Applying an intersectionality lens to expand equity in the geosciences. *Journal of Geoscience Education*, 68(2), 97–114.
- Nutt, D. (2019). Psychedelic drugs—a new era in psychiatry? *Dialogues in Clinical Neuroscience*, 21(2), 139–147.
- O'Malley-James, J. T., & Lutz, S. (2013). From Life to Exolife: The Interdependence of Astrobiology and Evolutionary Biology. In P. Pontarotti (Ed.), *Evolutionary Biology: Exobiology and Evolutionary Mechanisms* (pp. 95–108). Springer Berlin Heidelberg.
- Orion, N. (2019). The future challenge of Earth science education research. *Disciplinary and Interdisciplinary Science Education Research*, 1(1), 3.
- Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science & Public Policy*, 39(6), 751–760.
- Pace, N. R. (2001). The universal nature of biochemistry. *Proceedings of the National Academy of Sciences of the United States of America*, 98(3), 805–808.
- Pacheco-Fernández, I., & Pino, V. (2019). Green solvents in analytical chemistry. *Current Opinion in Green and Sustainable Chemistry*, 18, 42–50.
- Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756), 877–884.
- Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S. A., Pettersson, J., Goncharoff, N., Hersey, A., & Overington, J. P. (2016). SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research*, 44(D1), D1220–D1228.
- Pappalardo, R., Senske, D., Prockter, L., Paczkowski, B., Vance, S., Goldstein, B., Magner, T., & Cooke, B. (2015). Science and Reconnaissance from the Europa Clipper Mission Concept: Exploring Europa's Habitability. *European Planetary Science Congress*, 8155–8156.
- Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942), 138–144.
- Parker, E. T., Cleaves, H. J., Dworkin, J. P., Glavin, D. P., Callahan, M., Aubrey, A., Lazcano, A., & Bada, J. L. (2011). Primordial synthesis of amines and amino acids in a 1958 Miller H₂S-rich spark discharge experiment. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), 5526–5531.

- Parker, E. T., Cleaves, J. H., Burton, A. S., Glavin, D. P., Dworkin, J. P., Zhou, M., Bada, J. L., & Fernández, F. M. (2014). Conducting miller-urey experiments. *Journal of Visualized Experiments: JoVE*, 83, e51039–e51052.
- Peng, Z., Plum, A. M., Gagrani, P., & Baum, D. A. (2020). An ecological framework for the analysis of prebiotic chemical reaction networks. *Journal of Theoretical Biology*, 507, 110451–110466.
- Pennington, D., Ebert-Uphoff, I., Freed, N., Martin, J., & Pierce, S. A. (2020). Bridging sustainability science, earth science, and data science through interdisciplinary education. *Sustainability Science*, 15(2), 647–661.
- Pham, T., Sheridan, P., & Shimodaira, H. (2015). PAFit: A Statistical Method for Measuring Preferential Attachment in Temporal Complex Networks. *PloS One*, 10(9), e0137796–e0137809.
- Plotka-Wasyłka, J., Kurowska-Susdorf, A., Sajid, M., de la Guardia, M., Namieśnik, J., & Tobiszewski, M. (2018). Green Chemistry in Higher Education: State of the Art, Challenges, and Future Trends. *ChemSusChem*, 11(17), 2845–2858.
- Pruitt, S. L. (2014). The Next Generation Science Standards: The Features and Challenges. *Journal of Science Teacher Education*, 25(2), 145–156.
- Qi, P., Liang, Z.-A., Wang, Y., Xiao, J., Liu, J., Zhou, Q.-Q., Zheng, C.-H., Luo, L.-N., Lin, Z.-H., Zhu, F., & Zhang, X.-W. (2016). Mixed hemimicelles solid-phase extraction based on sodium dodecyl sulfate-coated nano-magnets for selective adsorption and enrichment of illegal cationic dyes in food matrices prior to high-performance liquid chromatography-diode array detection detection. *Journal of Chromatography A*, 1437, 25–36.
- Raccary, B., Loubet, P., Peres, C., & Sonnemann, G. (2022). Evaluating the environmental impacts of analytical chemistry methods: From a critical review towards a proposal using a life cycle approach. *Trends in Analytical Chemistry: TRAC*, 147, 116525–116237.
- Raup, D. M. (1986). Biological extinction in earth history. *Science*, 231, 1528–1533.
- Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, 58(6), 49–54.
- Ren, Z., Guo, M., Cheng, Y., Wang, Y., Sun, W., Zhang, H., Dong, M., & Li, G. (2018). A review of the development and application of space miniature mass spectrometers. *Vacuum*, 155, 108–117.
- Reymond, J.-L. (2015). The chemical space project. *Accounts of Chemical Research*, 48(3), 722–730.

- Rodriguez, A. J. (1998). Strategies for counterresistance: Toward sociotransformative constructivism and learning to teach science for diversity and for understanding. *Journal of Research in Science Teaching*, 35(6), 589–622.
- Rodriguez, Alberto J. (2015). What about a dimension of engagement, equity, and diversity practices? A critique of the next generation science standards. *Journal of Research in Science Teaching*, 52(7), 1031–1051.
- Rodriguez, Alberto J., & Morrison, D. (2019). Expanding and enacting transformative meanings of equity, diversity and social justice in science education. *Cultural Studies of Science Education*, 14(2), 265–281.
- Rose, K., & Rose, C. (2014). Enrolling in college while in prison: Factors that promote male and female prisoners to participate. *Journal of Correctional Education (1974-)*, 65(2), 20–39.
- Ruddigkeit, L., van Deursen, R., Blum, L. C., & Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11), 2864–2875.
- Sagan, C., & Druyan, A. (2011). *Pale Blue Dot: A Vision of the Human Future in Space*. Random House Publishing Group.
- Sakakibara, M., & Branstetter, L. (2001). Do Stronger Patents Induce More Innovation? Evidence from the 1988 Japanese Patent Law Reforms. *The Rand Journal of Economics*, 32(1), 77–100.
- Sayre, A. (2000). *Rosalind Franklin and DNA*. W. W. Norton & Company.
- Seager, S. (2010). *Exoplanet Atmospheres*. Princeton University Press.
- Seager, S., & Bains, W. (2015). The search for signs of life on exoplanets at the interface of chemistry and planetary science. *Science Advances*, 1(2), e1500047.
- Semken, S., Ward, E. G., Moosavi, S., & Chinn, P. W. U. (2017). Place-Based Education in Geoscience: Theory, Research, Practice, and Assessment. *Journal of Geoscience Education*, 65(4), 542–562.
- Semsar, K., Knight, J. K., Birol, G., & Smith, M. K. (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in Biology. *CBE Life Sciences Education*, 10(3), 268–278.
- Senger, S., Bartek, L., Papadatos, G., & Gaulton, A. (2015). Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. *Journal of Cheminformatics*, 7(1), 49.

- Seymore, S. B. (2013). Making Patents Useful. *Minnesota Law Review*, 98, 1046–1109.
- Sharma, A., Czégel, D., Lachmann, M., Kempes, C. P., Walker, S. I., & Cronin, L. (2022). Assembly Theory Explains and Quantifies the Emergence of Selection and Evolution. In *arXiv [physics.bio-ph]*. arXiv. <http://arxiv.org/abs/2206.02279>
- Shaw, A. M. (2007). *Astrochemistry: From Astronomy to Astrobiology*. John Wiley & Sons.
- Sheerman, J. (2020). *The Impact of Place-Based Pedagogy on Teachers' Professional Identities, Instructional Decisions, and Collaborations* (A. A. Lannin, Ed.) [Ph.D.]. University of Missouri - Columbia.
- Shen, Y., Fu, Y., Yao, J., Lao, D., Nune, S., Zhu, Z., Heldebrant, D., Yao, X., & Yu, X.-Y. (2020). Revealing the structural evolution of green rust synthesized in ionic liquids by in situ molecular imaging. *Advanced Materials Interfaces*, 7(15), 2000452.
- Sheridan, R. P., & Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today*, 7(17), 903–911.
- Shock, E. L. (1990). Geochemical constraints on the origin of organic compounds in hydrothermal systems. *Origins of Life and Evolution of the Biosphere: The Journal of the International Society for the Study of the Origin of Life*, 20(3–4), 331–367.
- Shock, E. L., Holland, M., Meyer-Dombard, D., Amend, J. P., Osburn, G. R., & Fischer, T. P. (2010). Quantifying inorganic sources of geochemical energy in hydrothermal ecosystems, Yellowstone National Park, USA. *Geochimica et Cosmochimica Acta*, 74(14), 4005–4043.
- Siegenfeld, A. F., & Bar-Yam, Y. (2020). An Introduction to Complex Systems Science and Its Applications. *Complexity*, 2020, 1–16.
- Smil, V. (2004). *Enriching the Earth: Fritz Haber, Carl Bosch, and the Transformation of World Food Production*. MIT Press.
- Smith, E., & Morowitz, H. J. (2016). *The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere*. Cambridge University Press.
- Smith, G. A. (2002). Place-Based Education: Learning to Be Where We are. *Phi Delta Kappan*, 83(8), 584–594.
- Stapledon, O. (2008). *Last and First Men*. Courier Corporation.
- Stern, J. C., Trainer, M. G., Brinckerhoff, W. B., Grubisic, A., Danell, R. M., Kaplan, D.,

- van Amerom, F. H. W., Li, X., Suero Amparo, W., Freissinet, C., Szopa, C., Buch, A., Teinturier, S., Moulay, V., Boulesteix, D., Malespin, C. A., Barfknecht, P. W., Stysley, P. R., Coyle, D. B., ... Turtle, E. P. (2023). Development of the Dragonfly Mass Spectrometer (DRAMS) for Titan. *54th Lunar and Planetary Science Conference*, 1532–1533.
- Sterner, R. W., & Elser, J. J. (2017). *Ecological Stoichiometry*. Princeton University Press.
- Strumsky, D., Lobo, J., & van der Leeuw, S. (2012). Using patent technology codes to study technological change. *Economics of Innovation and New Technology*, 21(3), 267–286.
- Szathmáry, E., & Smith, J. M. (1995). The major evolutionary transitions. *Nature*, 374(6519), 227–232.
- Szifris, K., Fox, C., & Bradbury, A. (2018). A Realist Model of Prison Education, Growth, and Desistance: A New Theory. *Journal of Prison Education and Reentry*, 5(1), 41–62.
- Szymkuć, S., Badowski, T., & Grzybowski, B. A. (2021). Is organic chemistry really growing exponentially? *Angewandte Chemie*, 133(50), 26430–26436.
- Tamara, S., den Boer, M. A., & Heck, A. J. R. (2021). High-Resolution Native Mass Spectrometry. *Chemical Reviews*, 122(8), 7269–7326.
- Taylor, S., Holder, J., Muhammad, B., Jones, T., & Haynes, L. (2021). Why Race Matters for Higher Education in Prison. *Peabody Journal of Education*, 96(5), 588–597.
- Teichmann, E., Lewandowski, H. J., & Alemani, M. (2022). Investigating students' views of experimental physics in German laboratory classes. *Physical Review Physics Education Research*, 18(1), 010135.
- Tolbert, S., Gray, S., Rivera, M., & Schindel, A. (2022). Teaching science to transgress: Portraits of feminist praxis. *Journal of Research in Science Teaching*, 59(1), 127–165.
- Tolbert, S., Schindel, A., & Rodriguez, A. J. (2018). Relevance and relational responsibility in justice-oriented science education research. *Science Education*, 102(4), 796–819.
- Tytler, R. (2014). Attitudes, identity, and aspirations toward science. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of Research on Science Education, Volume II* (pp. 96–117). TaylorFrancis.

- Varelas, M., Segura, D., Bernal-Munera, M., & Mitchener, C. (2022). Embracing equity and excellence while constructing science teacher identities in urban schools: Voices of new Teachers of Color. *Journal of Research in Science Teaching*, 2023(60), 196–223.
- Vecchio, I., Tornali, C., Bragazzi, N. L., & Martini, M. (2018). The Discovery of Insulin: An Important Milestone in the History of Medicine. *Frontiers in Endocrinology*, 9, 613–621.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272.
- Vitas, M., & Dobovišek, A. (2019). Towards a General Definition of Life. *Origins of Life and Evolution of the Biosphere: The Journal of the International Society for the Study of the Origin of Life*, 49(1–2), 77–88.
- von Korff, M., & Sander, T. (2013). About Complexity and Self-Similarity of Chemical Structures in Drug Discovery. *Chaos and Complex Systems*, 301–306.
- Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.
- Walker, S. I. (2017). Origins of life: a problem for physics, a key issues review. *Reports on Progress in Physics*, 80(9), 092601–092622.
- Wang, V. X., Torrisi-Steele, G., & Hansman, C. A. (2019). Critical theory and transformative learning: Some insights. *Journal of Adult and Continuing Education*, 25(2), 234–251.
- Ware, S. A. (2001). Teaching chemistry from a societal perspective. *Journal of Macromolecular Science, Part A: Pure and Applied Chemistry*, 73(7), 1209–1214.
- Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V., & Martin, W. F. (2018). The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genetics*, 14(8), 1007518–1007537.
- Wells, H. G. (2003). *The War of the Worlds*. Broadview Press.
- Wells, H. G. (2005). *The Time Machine*. New York, Berkley.
- West, G. (2018). *Scale: The Universal Laws of Life, Growth, and Death in Organisms, Cities, and Companies*. Penguin.

- Wilcox, B. R., & Lewandowski, H. J. (2016). Students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for experimental physics. *Physical Review Physics Education Research*, 12(1), 010123.
- Willhite, L., Ni, Z., Arevalo, R., Bardyn, A., Gundersen, C., Minasola, N., Southard, A., Briois, C., Thirkell, L., Colin, F., Grubisic, A., Fahey, M., Yu, A., Hernandez, E., Ersahin, A., Danell, R., & Makarov, A. (2021). CORALS: A Laser Desorption/Ablation Orbitrap Mass Spectrometer for In Situ Exploration of Europa. *2021 IEEE Aerospace Conference (50100)*, 1–13.
- Williams, R. J. (1997). The natural selection of the chemical elements. *Cellular and Molecular Life Sciences: CMLS*, 53(10), 816–829.
- Xavier, J. C., Hordijk, W., Kauffman, S., Steel, M., & Martin, W. F. (2020). Autocatalytic chemical networks at the origin of metabolism. *Proceedings. Biological Sciences / The Royal Society*, 287(1922), 20192377–20192387.
- Yonchev, D., Dimova, D., Stumpfe, D., Vogt, M., & Bajorath, J. (2018). Redundancy in two major compound databases. *Drug Discovery Today*, 23(6), 1183–1186.
- Youn, H., Strumsky, D., Bettencourt, L. M. A., & Lobo, J. (2015). Invention as a combinatorial process: evidence from US patents. *Journal of the Royal Society, Interface / the Royal Society*, 12(106), 20150272–20150280.
- Young, I. M. (2006). Responsibility and Global Justice: A Social Connection Model. *Social Philosophy & Policy*, 23(1), 102–130.

APPENDIX A
INSTITUTIONAL REVIEW BOARD APPROVAL



APPROVAL:CONTINUATION

[Darryl Reano](#)

CLAS-NS: Earth and Space Exploration, School of (SESE)

-

Darryl.Reano@asu.edu

Dear [Darryl Reano](#):

On 12/1/2022 the ASU IRB reviewed the following protocol:

Type of Review:	Continuing Review
Title:	Developing Earth Science Curricula to Increase STEM Identity in Juvenile Institutionalized Youth
Investigator:	Darryl Reano
IRB ID:	STUDY00014872
Category of review:	Expedited (7)(a) Behavioral research (7)(b) Social science methods
Funding:	Name: NASA: Arizona Space Grant Consortium (AZSGC)
Grant Title:	None
Grant ID:	None
Documents Reviewed:	<ul style="list-style-type: none"> • Arizona Department of Correction Research Proposal, Category: Other; • CITI certificate REANO 102022.pdf, Category: Non-ASU human subjects training (if taken within last 3 years to grandfather in); • Darryl Reano CITI certification, Category: Non-ASU human subjects training (if taken within last 3 years to grandfather in); • IRB Social Behavioral Application (1).docx, Category: IRB Protocol; • Letter of Support, Category: Off-site authorizations (school permission, other IRB approvals, Tribal permission etc); • Malloy Offer Letter, Category: Sponsor Attachment; • Student Assent, Category: Consent Form;

	<ul style="list-style-type: none">• supporting documents 03-11-2021.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);• Survey Instrument.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);
--	---

The IRB approved the protocol from 12/1/2022 to 11/30/2023 inclusive. Three weeks before 11/30/2023 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 11/30/2023 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: John Malloy
John Malloy
Sara Walker
Darryl Reano