Logbook Total Retype

July - August was done on Notion:
https://www.notion.so/Science-Fair-23c41b505b058091a132d0b78faefa15
Unfortunately as I lost access to that notion I had to switch to another format. There is a lot of good stuff on this notion though, so please give it a look.

Sunday September 21st 2025

As unfortunately the online service I've been using online refuses to work for me anymore, I must return to my analog logbook roots. This is a change for me to reimagine my project as I really want to find an impactful area that I am passionate about this year.

The current areas of consideration are astrophysics, sustainability, statistical analysis, biology, and computer science. I *really* want to do something that I can do a statistical analysis on. I honestly might do something related to biology but I need to find a specific problem that I can design an experiment for. There are various parts of biology so I might as well do a quick analysis and description of all of them and what I'm interested in.

Branch	**Description**	**Interest**
Eccology	Relation of organisms to one another and their physical surroundings	
Astrobiology	Origins of life in the universe, deterministic conditions, contingent events	
Anatomy	Internal organism structure	
Bioinformatics	Develops software tools for understanding biological data	
Parasitology	The study of parasites, their hosts, the relationship between them	
Genetics	Genes, genetic variation, heredity	

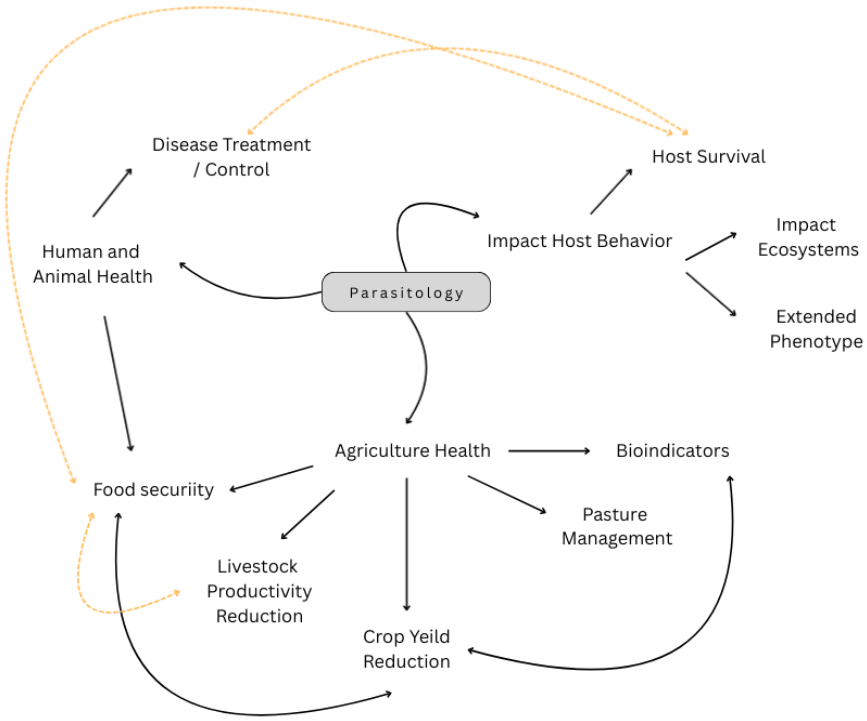
Biophysics	Something about proteins and their interactions	?
Geobiology	Interaction between physical earth and biosphere	😐
Developmental	Process in which animals and plants grow and develop	😁
Zoology	Animals, structure, embryology, habits	🙂
Cryobiology	How low temperatures impact organisms	?
Histology	Microscopic anatomy of biological tissues	😐
Biochemistry	Chemical process within living organisms	😐
Evolutionary	Evolutionary process	🙂

Time spent on project: 30 minutes

Monday September 22nd 2025

I'm aware that I shouldn't but I can't help but feel behind. I want to quickly find a specific question, but I also want it to be quality. Today's exercise will be finding problems / important topics in the fields of study that I was passionate about.

1. **P A R A S I T O L O G Y**
   - ➔ Studies parasites that cause millions of deaths
   - ➔ Studies can reduce agricultural loss
   - ➔ Understanding of ecosystems
   - ➔ Impact on host behavior



- ➔ "How can parasites impact entire ecosystems?"
- ➔ "What factors contribute to a parasite's level of control?"

2. B I O I N F O R M A T I C S

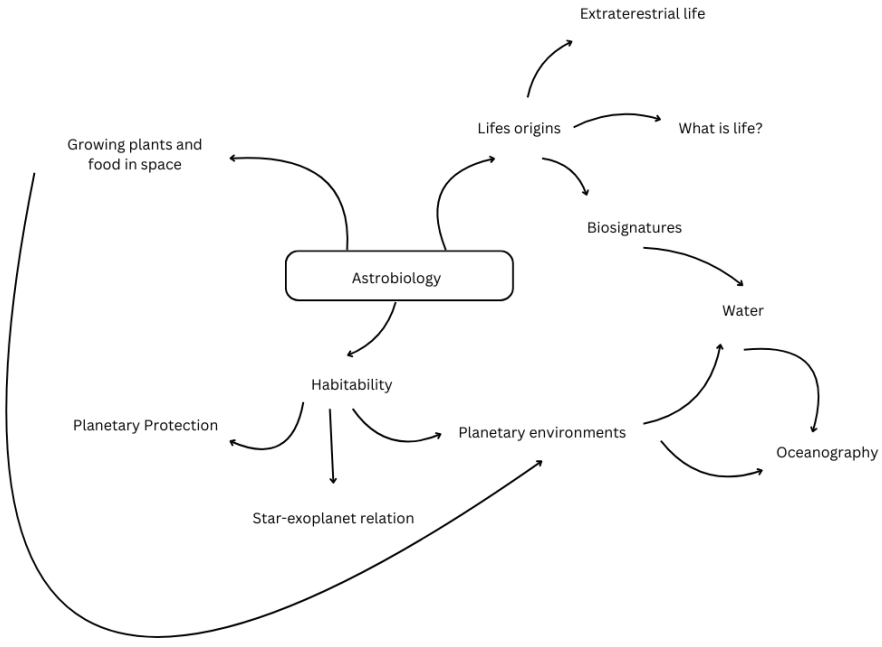
- ➔ Personalized Medicine
- ➔ Drug development
- ➔ Pest resistant Crops
- ➔ Genomics

More of a tool than a field of study. The tools are cool but I find using bioinformatics more interesting than studying it.

3. A S T R O B I O L O G Y

- ➔ Life's origins
- ➔ Extraterrestrial life
- ➔ Biosignatures
- ➔ Planetary surfaces

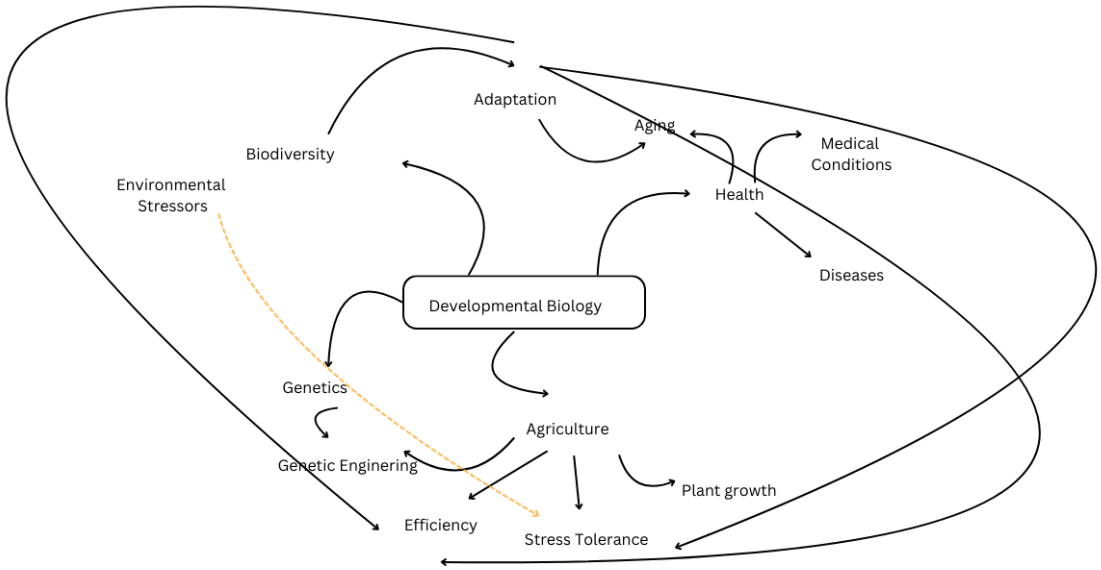
Ended Here: Monday 22nd
Time Spent Working: 30min
Started Here: Tue 23rd



- ➔ “How can oceanography of our planets lead to discoveries across the galaxy?”
- ➔ “How can planetary environments be manipulated to the benefit or detriment of mankind?”
- ➔ “How can plant growth be better modified to suit an extraterrestrial environment?”

4. DEVELOPMENTAL BIOLOGY

- ➔ Health
- ➔ Agriculture
- ➔ Biodiversity
- ➔ Genetics



- ➔ “What effect can environmental stressors have on the health of an entire ecosystem??”
- ➔ “How can humans benefit from other organisms' adaptations to solve current problems?”
- ➔ “What are the dangers of current developmental biology techniques on organisms’ health?”

Overall , all of the topics are still incredibly interesting so I’ll pick a few questions to write a one page report on, with problems in the field and potential experiments or solutions. This should of course keep track of sources as ideally I’ll eventually create an annotated bibliography for my final research question. I want to find a few quality sources this year rather than many unreliable sources.

Q U E S T I O N S

- ➔ “How can parasites impact entire ecosystems?”
  - ◆ Host population dynamics change
  - ◆ Food webs
  - ◆ Biodiversity
  - ◆ Competition
- ➔ “How can oceanography of our planets lead to discoveries across the galaxy?”
  - ◆ Habitable Environments
  - ◆ “Is water the key to life?”
  - ◆ Using tools in extreme conditions
- ➔ “How can planetary environments be modified to the benefit or detriment of mankind?”
  - ◆ Bioengineering habitats (potentially planets)
  - ◆ Teraforming
  - ◆ Ethically
  - ◆ Geoengineering

Ended here: 23rd Tue
Time Spent: ~35mn
Total Time: 1.58hrs
Started Here: 24th Wed

- ➔ “What effect can environmental stressors have on the health of an entire ecosystem??”
  - ◆ Biodiversity
  - ◆ Food Web
  - ◆ Types of Stressors (anthropogenic, chronic)
  - ◆ Types of analysis (morphological)
- ➔ “How can humans benefit from other organisms' adaptations?”
  - ◆ Biomimicry
  - ◆ Behavioral Knowledge
  - ◆ Drug Development

How Can Parasites Affect Entire Ecosystems?

PARASITES IN FOOD WEBS

- ➔ Lafferty, D.D, AllesinaS, et al. (2008). Parasites in food webs: the ultimate missing links. *Ecology letters*, 11(6), 533-546. https://doi.org/10.111/j.1461-0248.2008.01174.x
- ➔ Do parasites affect food web stability?
- ➔ Food webs + infectious disease dynamics
- ➔ Food web typology
- ➔ Food web dynamics

Ended Here: 24th Wed
Time Spent: 30min
Total Time: 125min or 2.08hrs

Start: 4:20 15 Oct

Wednesday October 15th

I look for a break. A *very* long break. About a month. This is okay, honestly it may have been necessary for my personal health. However it also had an impact on this project. I once again felt behind and for a bit I just wanted to give up. I also remembered what I thought to myself in April of last year. "What is I just worked harder and spent more days?". I have 179 days until April 12th. I bet that I can get 300hrs on this project.

That's twelve days just working on this project. That's a bit more than 1.5hrs per day. That will 100% be a lot of work.

For today I'm just going to choose an interesting question. Anything that I truly find myself passionate about because if I will be spending 300hrs or 12.5 days on this then I better enjoy it.

I'm going to research parasitic diseases. At least for this week. I find them interesting and while finding the initial ideas this is what interested me the most.

Annotated Bibliography Pg. 1

1. Citation:

Short, E. E., Caminade, C. & Thomas, B. N. (2017). Climate Change Contribution to the Re-Emergence of Parasitic Diseases. *Infectious diseases,* 10, 1178633617732296.
https://doi.org/10.1177/1178633617732296

In text citation (short et al. 2017)

Perhaps I could learn to create simulations or other technology to test & demonstrate this

this reminds me of the mind-map I did

Summary:

The article aims to show how climate change will impact public health issues and economic stagnation due to parasitic diseases. Evidence shows that climate change is altering the spread of diseases. Some areas that may have once not been hospitable to parasites may now be hospitable due to the change in climate. Development of communities may slow due to an infection of crops and livestock. This would / will impact economic growth. Parasitic diseases can be very debilitating, preventing economic growth and keeping areas in poverty. Areas in the world that are still developing, have inadequate waste systems, medical supplies, and education are most likely to be affected. It is crucial to prepare for the exacerbation that climate change may have on the world's diseases.

Application: For this project I currently have the idea of using this knowledge to create a simulation or some sort of technology to demonstrate this concept. It could also be cool to just do research on highly impacted areas and figure out how to best help with my skills.

2. Citation:

Parasites *Cleveland Clinic,* 26 Sept. 2023, my.clevelandclinic.org/health/diseases/24911-parasites.
In text: (Cleveland Clinic)

Creating tech to limit those conversations in at-risk areas

Understanding the effectiveness of various methods of treatment

Summary:
Parasites are organisms that live on, in, with, another organism referred to as a 'host'. Parasites harm their host while they benefit. The host is necessary for the parasite's survival. There are three main types: ectoparasites, helminths, and protozoans. Ectoparasites live on the exterior of its host. They are vectors, meaning that they carry diseases between animals and humans. Typically they carry infections through blood. Helminths live in your gastrointestinal tract (GI) and are once adults. Protozoans are uni-cellular organisms. They may live in one's intestines or blood and tissues. Contaminated food or water, person to person contact, or a vector bite may spread them. They cannot be seen without a microscope. People can get parasites through contaminated water, soil, blood, feces, food, not washing hands before eating, contaminated bug bites, sexual contact, or having a weak immune system amongst other causes. Parasites are treated in varying manners depending on the type such as by using antiparasitic drugs, antibiotics, shampoos. And ointments.

Applications:
Background knowledge to guide future research.
Ended here: 5:50pm
Time Spent: 90min
Total Time: 215min ~3.58hrs

3. Citation:

Godkin, A. & Smith, K. A. (2017) Chronic Infections with viruses or parasites breaking bad to make good. *Immunology,* 150(4), 389-396. https://doi.org/10.1111/imm.12703.
In text: (Godkin & Smith, 2017)

Summary:
Chronic viral and parasite infections can positively and negatively impact a host's immune system.

Sunday October 18th 2025
Today's goal is to start designing experiments or at least have several concrete ideas.

1. Ontario Vector-Borne Disease Tool
    - Information on vector-borne diseases in Ontario
    - Shows established risk areas
    - Will need to request more data if used

Sectors

Surveillance	Behavior	Treatment
☐ Could create an Ontario Vector-Borne Disease Tool for Albera or All of Canada but improved		

2. Canadian Database for Disease Vectors
   - ➔ Will need a lot of data-cleaning, includes contact emails

3. WHO Global Vectors Control Response 2017 - 2030
   - ➔ The plan is to look at this and find a part that I am willing to tackle

EFFECTIVE VECTOR CONTROL

---

1. No Poverty
3. Good Health and Well-being
6. Clean Water & Sanitation
11. Sustainable Cities & Communities
13. Climate Action
17. Partnerships For The Goals

CHALLENGES

- ➔ Systemic issue of limited public health entomology capacity & poor infrastructure.
- ➔ Structural issue of how resources are managed and issues are responded to.
- ➔ Information issue of lacking resources, support and surveillance in third-world countries. Entomology, epidemiological data are managed separately causing issue

Ended Here: Saturday Oct 8 2025
Time worked : 45min

November 11th 2025 Tuesday

You can guess what happened. That's right, I got distracted by school work & other shiny things & neglected this project. Instead of feeling shameful however, I need to get something done & fix this reoccuring problem. #1, one problem with this is that I don't work. I get mad at myself, but also feel ashamed and hide from this project. That won't do, and I need to confront my crippling fear of failure and simply work on this project. #2. I need a set time to work on this, maybe every Tuesday & every Sunday. #3. I need to be able to work on this even without my book, meaning creating a digital logbook and finding a way to add that work here. Finally, I need to set goals & deadlines. As Parkinston's law states: Work expands to fill the time you give it.

November

- ➔ Solidify project goal and scope
- ➔ Find dataset
- ➔ All background information completed

December

- ➔ Create everything

January

- ➔ Ask for Feedback
- ➔ Edit Everything

February

- ➔ School Science Fair
- ➔ Polish

March

- ➔ Final Push
- ➔ Polish

Now let's set specific deadlines until the end of the year.

November 11th - 16th

- ➔ Finish project ideation. Know exactly what I want to create / research. This means having a design doc started, a mission statement, being able to clearly define the goal. This should be about three pages. I need to be passionate about it, and it needs to be possible to do in essentially three months. This is due on the 16th. If I get it done I can buy a fun meal

November 17th - 30th

- ➔ Complete design document. I need to know exactly how to complete the project. Have the resources prepared, have some experts contacted for advice if relevant, and be organized. This is due on the 30th. If done I get one rest day from schoolwork.

I can't really plan December until the design document is done. So without further ado I shall start making the document.

Basically, looking through my work from the past TWO MONTHS all I chose was 'vectors' and that's IT. So gotta lock in for the next week.

So here's the TLDR. Vectors are really important for medicine and agriculture. So I want to study them and figure out how to deal with them, or how they affect us.

My 'expertise' (I use that word loosely) is in computer science. I'm not a biology major. Therefore it may be hard to do anything 'lab' related. So I'll try to avoid that if possible. Essentially I want to figure out how to reduce the spread of diseases by vectors for both humans and agriculture. Is this novel? No? Can I achieve it? Maybe not. But like, I'm trying to get over my fear of failure so this is perfect. Let's do it.

I think it could be cool to do something on how vectors will impact us over the next 10 years.

December 10th 2025

- ➔ Restarting for what is like the billionth (3rd) time
- ➔ Creating a logbook I can look back and be proud of
- ➔ The topic is dark clouds in the solar system
- ➔ I will use machine learning, statistical analysis, and talk to experts for this project
- ➔ The goal is to stop shame from not working from preventing me from working
- ➔ Imported libraries
- ➔ Downloaded (Nasa) dataset
- ➔ Read paper by Marc- Antoine Miville - Deschênes "Physical Properties of Molecular Clouds for the Entire Milky Way Disk,"
- ➔ Create goals / deadlines

Deadlines

December 30th
- All Machine learning done

January 10th
- Clear plan for stats

January 20th
- All stats done

Rest of January
- (School) Science Fair Prep

Feb 10th
- Talk with expert

Feb 20th
- All materials done

March 1st
- All components done only polish

December 11th 2025

- Cloud # column not necessary should be dropped
- Going to do more research
- Birds eye view of molecular clouds in the Milky Way (aanda.org)
- Density contrast of clouds linked to star formation more so than dense gas mass
- Do some histograms
- Remove rows with NaN
- None of the rows have words therefore dictionaries are not required
- Can confirm no NaNs by counting with code
- Started making histograms
- Make scatter plots instead
- Made both

December 15th 2025

- Worked on some understanding
- Reviewed code, no changes

December 30th

- Scatter plots
- Used logarithms for the output
- Dnear vs Snear
- Snear vs WCO
- Snear is physical size
- Dnear is distance
- WCO is (integrated) carbon monoxide emission
- Mnear is mass
- Utilises the matplotlib import as well as pandas

January 4th

- Attempted to add GMM clustering (Gaussian Mixture Model)
- Attempted to add Davies Bouldin score
- Attempted to add Silhouette Score
- Attempted to add Kmeans Clustering
- The code 'worked' but the graph was visibly incorrect

January 13th

- Fixed the error by using logarithms
- Created a new feature for each line

```
df["log_WCO"] = np.log10(df["WCO"])
df["log_Sigma"] = np.log10(df["Sigma"])
```

- Log10 is required as some coding languages automatically use lan
- Requires pandas & numpy libraries (plus matplotlib)
- Did the same for Snear, Mnear
- Will start the regression soon

January 19th

- Explanation on code
    1) Imports

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# import seaborn as sns
from sklearn.model_selection import train_test_split
```

- Pandas = data manipulation & cleaning
- Numpy = computing
- Matplotlib = data visualization
- Seaborn = data visualization
- Sklearn = training

2) Df (datafame)

```
df = pd.read_csv("table1_full - table1_full.csv")
df
```

	Cloud	Ncomp	Npix	Area	GLON	Sigb	GLAT	Sigl	theta	WCO	...
0	1	6	6	0.093750	-0.560716	0.153933	0.063824	0.125000	28.802170	257.122300	...
1	2	20	20	0.312500	-0.277576	0.184638	-0.241459	0.138789	10.758150	1115.180000	...
2	3	25	25	0.390625	-0.857003	0.194934	-0.259487	0.147484	-17.895130	1439.139000	...
3	4	107	87	1.359375	-19.656110	0.321626	-0.096034	0.271538	9.468965	856.856600	...
4	5	46	45	0.703125	-0.437401	0.272261	-0.316237	0.202676	9.319308	2062.619000	...
...	...	...	...	...	...	...	...	...	...	...	...
8102	8103	5	5	0.078125	5.725857	0.125000	-2.165318	0.125000	-25.910900	2.438211	...
8103	8104	8	8	0.125000	-95.802930	0.135788	-1.387736	0.125607	10.228730	4.422103	...
8104	8105	7	7	0.109375	-123.493200	0.125000	-1.356942	0.125000	19.763220	4.571887	...
8105	8106	12	11	0.171875	-129.332200	0.174926	-0.471745	0.174427	-12.111170	8.538548	...
8106	8107	5	5	0.078125	17.413410	0.211755	-3.376859	0.125000	-25.601160	4.728827	...

8107 rows × 29 columns

-
- Making df attributed to the code, Prints out dataframe

```
df.columns
```

expand

```
Index(['Cloud', 'Ncomp', 'Npix', 'Area', 'GLON', 'Sigb', 'GLAT', 'Sigl',
       'theta', 'WCO', 'NH2', 'Sigma', 'Vcent', 'SigV', 'Rmax', 'Rmin', 'Rang',
       'Rgal', 'INF', 'Dnear', 'Dfar', 'znear', 'zfar', 'Snear', 'Sfar',
       'Rnear', 'Rfar', 'Mnear', 'Mfar'],
      dtype='object')
```

```
#Remove Cloud Column
df = df.drop(columns = ['Cloud'])
```

-
- Prints out columns, gets rid of cloud column

3) How to make a histogram

- plt.hist(df[“A”], df[“B”])
- Where A = any column desired for the histogram

- Where B = any other column desires
- plt.ylabel(“B”)
- plt.xlabel(“A”)
- plt,title(“A vs B”)
- plt.xscale(“log”)*
- plt.yscale(“log”)*
- *depends on if it needs to be in logarithms
- plt.show()
- Essentially creates the graph

4) The code for the ML algorithms is messy so I might do that later

January 27th, 2026
Worked on filling out various graphs and finishing some of the unsupervised learning

January 29th, 2026
I almost freaked out because I was getting Mean absolute (error) scores of 18.21 and R-squared scores negative values which are horrendous but it worked out after I found the problem. I just forget to log the values which properly represent the large range by stretching the graph out to properly represent the data. I also added the criteria ‘absolute-error’ to try and ‘balance’ it out. It was a bit too slow so I might try it again later for reference it was: criterion = ‘absolute-error’. Something else that I added than ran faster was max-features = 5, or really any integer. It had a minimal effect but I tried it out. I also tried out Adaboost Regressor which is supposed to be better. Oddly enough I keep on getting numbers around 2.3 for mean absolute and 0.62 for R-squared. The regressors are doing FAR better in comparison to the clustering which looks very artificial. Overall, the good R^2 values essentially mean that there any completely different parts, everything is in the same continuum or group, so it is smoother anyways. There is no obvious clustering. Once all the supervising is done we can better understand monoxide emissions and size and other parameters. You can also see / quantitatively see the negative vs positive relationships. The R^2 value greater than 0.5 tells us that the larger the mass the more emissions are typically produced. We can make that generalization which isn’t always true due to the increase in heat gained from being less massive.

January 30th, 2026
I did some work before my hair appointment and did WCO vs Snear. I remembered to use log this time but the values weren’t ideal so some modifications are required. I might as well add what each variable means so that will be on the next page. Even if I don’t make it to cities this year I’ll do something with the project.

Table 1
Entries of the Molecular Cloud Catalog

Entry	Units	Description
C	⋯	Cloud number
$N_{\rm comp}$	⋯	Number of Gaussian components
$N_{\rm pix}$	⋯	Number of pixels on the sky
A	$\deg^2$	Angular area
l	deg	Baricentric Galactic longitude
σ_l	deg	Galactic longitude standard deviation
b	deg	Baricentric Galactic latitude
σ_b	deg	Galactic latitude standard deviation
θ	deg	Angle with respect to $b = 0°$
$W_{\rm CO}$	K km s^{-1}	Integrated CO emission
$N_{\rm H_2}$	cm^{-2}	Average column density
Σ	$M_\odot$ pc^{-2}	Surface density
$v_{\rm cent}$	km s^{-1}	Centroid velocity
σ_v	km s^{-1}	Velocity standard deviation
$R_{\rm max}$	deg	Largest eigenvalue of the inertia matrix
$R_{\rm min}$	deg	Smallest eigenvalue of the inertia matrix
$R_{\rm ang}$	deg	Angular size
$R_{\rm gal}$	kpc	Galactocentric radius
$I_{\rm NF}$	⋯	Near or far distance flag
D	kpc	Kinematic distance
z	kpc	Distance to Galactic midplane
S	pc^2	Physical area
R	pc	Physical size
M	$M_\odot$	Mass

Feb 1st 2025

I basically just did the annoying tasks today (and a bit of yesterday). I cleaned my code and made everything understandable. I might add more comets for those reading later. Everything is now in digestible sections and most of the ML (machine learning) is done. I will probably take time tomorrow to add all the figures into the logbook (I didn't), as well as some key bits & what is necessary to understand the project. For the code itself I need to better detail what M, S, R, WCO, and Sigma mean.

Feb 2nd 2025

Slideshow, edit 5 page analysis, data, research, conclusions, edit problem, mentor meeting, citations

- Key figures
- WCO vs M plot vs WCO vs S plot
- This helps understand how many different life forming environments exist within the milky way
- WCO vs Sigma, S, R, M, Z, Rgal
- This answers how complex the relationship between organic molecules and other molecular cloud properties is. How complex can the building blocks of life be?

Feb 3rd 2025

Problem - "What is the true diversity of the milky way galaxy star forming regions?"
Hypothesis - "By looking at a large # of these star forming regions"
Take home - "Since there aren't any distinct star forming regions the way Earth formed hopefully isn't any different.
The very straight line essentially is a pre-programmed relation.
The results mainly showed a large correlation between WCO and surface density.
Non-linear models do worse. More earth-like planets should not be hard to find.

Print slides # 1,2,3,4,5,6,7,8,16,17,19,21,22,23

Feb 6th TO - DO

- FInish online paper
- Connection current paper + previous paper
- Use a Neural Network
- Use another dataset
- Do statistical analysis
- Reprint trifold using 3 pages per panel excluding title
- Contact U of C prof

Feb 9th 2026

HOW TO DO STATS

1) **Descriptive, kind of already done, tables + graphs, do more?**
2) **Associational stats which requires software. What software? How can a high school student do it with only 20 level mathematics?**
3) **Common methods are standard deviation (SD), regression (Already done) (indeed)**
4) **Read paper Vandever C. (2020). Introduction to Research Statistical Analysis: An Overview of the Basics. HCA healthcare journal of medicine, 1(2), 71–75. https://doi.org/10.36518/2689-0216.1062 (did not do that)**

Tue 10 February

- Connection between CO and other organic compounds
- How does CO tell us about other elements

- ___CO and ___ mass = ___ amount of organic compound
- Be careful because we have not seen life in other places but probability we should be able to find it / it should exist
- Extreme UV can break apart organic compounds