2025 ASEE Northeast Section Conference, March 22, 2025, University of Bridgeport, Bridgpeort, CT, USA.

Shannon Entropy Research Tool for Analyzing DNA/Amino Acid Sequences and Solar Eruptions Data with Application to Generative AI Diffusion Model of Text To Image Technology

Julianne Torreno, Nealesh Guha, Mashtura Rahman,
Michael Ventouratos, David Lee, Shivansh Sharma,
Sunil Dehipawala, Guozhen An, Tak Cheung

Physics Department
CUNY Queensborough Community College New York City USA

***Abstract*— The Shannon entropy formula has been accepted as a calculation to measure the amount of information in a random trial. The recent success of Generative AI diffusion model of text to image technology offers much inspiration for STEM students to learn about random processes in research projects beyond the regular classes in a community college setting. The 1-dimensional DNA/amino acid sequences on PubMed serve as signal examples of information content analysis using Shannon entropy calculations. The extension to 2-dimensional image entropy calculations could use the Solar Dynamics Observatory data as practice examples. The computation experience learned on the PubMed and SDO data analysis components was found to be sufficient to start an analysis of using entropy calculation to classify the output of the diffusion model used in Generative AI. The research projects using PubMed data (such as the ZENK gene involved in songbird singing) and SDO data (such as the NOAA 13664 driven 2024 Mother's Day Solar Storm) yielded satisfactory results and fulfilled the criterion of authentic experience in undergraduate student research. For the Generative AI diffusion model component, the entropy classification of the forward diffusion was found to be acceptable, while the backward diffusion remains as work-in-progress. Future research topics using entropy calculation are discussed, together with sustainability.**

Keywords*— *Shannon entropy; DNA/amino acid sequence entropy; solar image entropy; diffusion model Generative AI

I. INTRODUCTION

The success of AI-assisted tools has generated vast interest among students in different majors in our community college with the mandated policy of open admissions. A research project to substantiate students' interest on Generative AI should contain a "safe" research component with guaranteed new results and exploratory research without guarantee of any new results. We have identified the use of entropy as a research tool in student projects. The Shannon entropy formula has been accepted as a calculation to measure the amount of information in a random trial. The recent success of Generative AI diffusion model of text to image technology offers much inspiration for STEM students to learn about random processes in research projects beyond the regular classes in a community college setting.

II. IMPLEMENTATION I

The 1-dimensional DNA/amino acid sequences on PubMed serve as signal examples of information content analysis using Shannon entropy calculations. A student research project using PubMed data is presented here. The relationship between the entropy of the DNA sequence and the corresponding amino acid sequence methodology was used [1, 2]. The ZENK gene involved in songbird singing was studied. The interest on the ZENK gene was initially raised by an advanced high school student taking our algebra physics course because the student was enrolled in the songbird hormone project course at Hofstra University for the Regeneron competition [3]. The Shannon entropy calculation of the sum of -p*log(p), with p representing probability, was done in Excel for the Regeneron competition and adapted to Python and Matlab for regular college students. The data was downloaded from NCBI [4]. An example of data download is shown in Fig. 1.

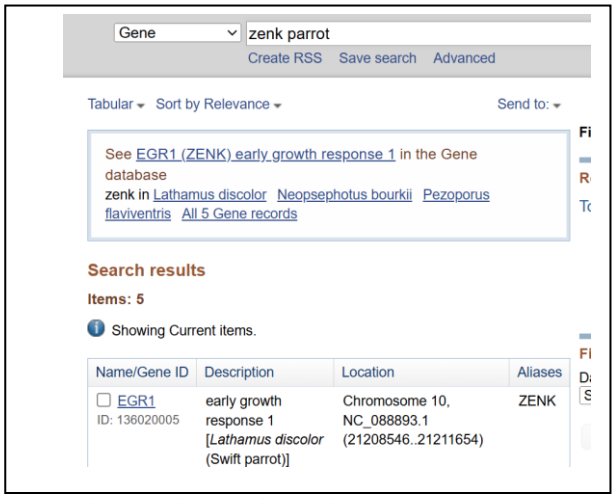

Fig. 1. The NCBI data page for "Zenk parrot"

The DNA coding sequences were used, together with the corresponding amino sequences. The entropy calculation results are shown in Fig. 2.

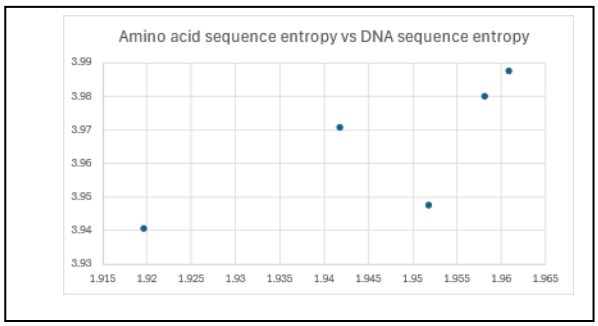

Fig. 2. Results of ZENK study. Amino acid sequence entropy (y-axis in unit of bits) versus DNA sequence entropy (x-axis in unit of bits). Mouse the outlier (1.952, 3.948), Zebra Fish (1.942, 3.971), Zebra Finch Songbird (1.961, 3.987), Night Parrot (1.958, 3.980) and Human (1.919, 3.941) were shown.

The objective was to show the students the reality of the Shannon entropy calculation in a simple procedure with minimum math requirement. An enthusiastic student can use Microsoft Word to count the frequency of each symbol and use a simple calculator to calculate -p*log(p), after downloading the sequences. The proof of concept of Shannon entropy calculation was presented with limited data. Speculation on this exploratory research result could generate a ZENK hypothesis that the fish was the earliest animal, and the songbird, night parrot, and human evolved with a pattern, while the mouse is an outlier, perhaps the mouse is lacking complex speech behavior when compared to songbird, parrot, and human.

The development of entrepreneurship in engineering education can be enhanced using a budget perspective. The collection of data, analysis of data, and working with an external professional in the interpretation of data, etc. would require budget planning in the business world. The fact that a college budget would cover the expenses of student skill learning projects and student research projects with tuitions and fees is a proof that activities usually have some monetary associations. For instance, a faculty mentor could ask a college administration for more student stipends to extend an analysis when an ongoing exploratory research is generating promising results. A faculty mentor can also explain data science activity to the students as an activity in a start-up company. In the case of bioinformatics, a data science start-up company could persuade a client company to pay for more data analysis to include more mammals in a more comprehensive study. This is an acceptable pedagogy for career development in terms of entrepreneurship.

III. Implementation II

The extension to 2-dimensional image entropy calculations was performed using the data of Solar Dynamics Observatory [5]. The entropy analysis of solar images is a good skill learning project. When there are new results, the interested students are encouraged to read on the solar physics background and transform the skill learning project to a research project with an interpretation pertinent to solar physics. Furthermore, in our community college with open admissions, there are students interested to develop careers in the weather forecast industry.

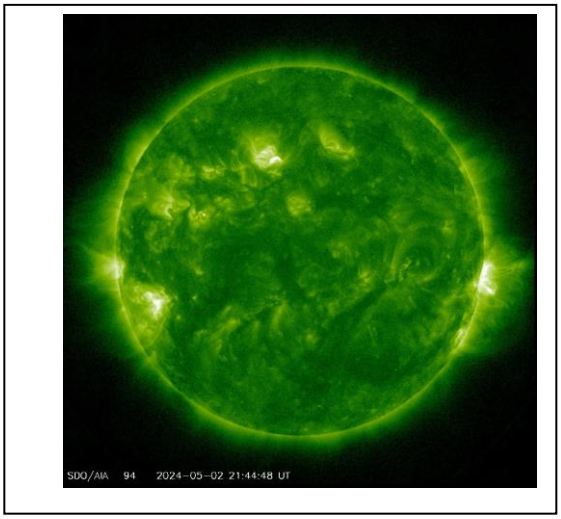

Fig. 3. An example of a solar image from Solar dynamics Observatory, 9.4 nm, the grren color was used as a pseudo-color for display.

A solar eruption can carry coronal mass ejection in terms of high energy particles such as protons and electrons, in addition to solar flares as electromagnetic waves taking only 8 minutes to reach Earth. The high energy solar particles would take a day or two to reach the Earh's magnetosphere. The trapped particles in the magnetosphere generate a ring current which would suppress the geomagnetic field. The disturbance storm time (DST) index was developed for space weather application.

The DST geomagnetic index of -420 nT was recorded by Kyoto University on May 11 [6], with the causation active region of 13664 starting around May 2 [7, 8, 9]. The analysis of the 13664 driven 2024 Mother's Day Solar Storm based on the Solar Dynamics Observatory SDO images showed a threshold process on May 6, shown in Fig. 4.

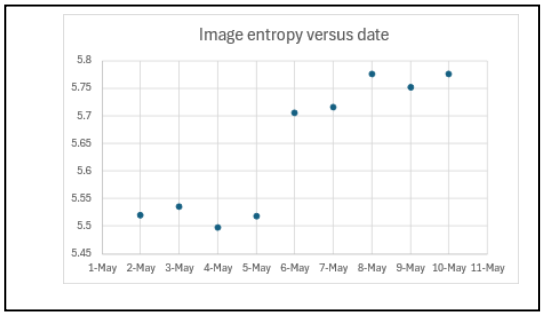

Fig. 4. Solar Dynamics Observatory 9.4 nm image entropy (y-axis in unit of bits) versus calendar date (x-axis). The studied period spanned from May 2, 2024 (about 11 pm) to May 10, 2024 (about 11 pm), consistent with the work of Kondrashova, et al. [7].

The image entropy was calculated in Colab using the following 5 commands, namely, "import cv2", "img = cv2.Imread('filename')", "import skimage.measure" and "entropy = skimage.measure.shannon_entropy(img)", and "print (entropy)".

The threshold feature of Fig. 4 suggested that the entire Sun was involved in the evolution of the active region 13664. The new result fulfilled the criterion of authentic experience in undergraduate student research project. If the analysis did not generate new results, the skill learning project is still beneficial for the students to gain experience in image entropy analysis.

IV. Implementation III

The computation experience learned on the PubMed and SDO data analysis components was found to be sufficient to start an analysis of using entropy calculation to classify the output of the diffusion model used in Generative AI. For the Generative AI diffusion model component, the entropy classification of the forward diffusion was found to be acceptable, while the backward diffusion remains as work-in-progress. The classic 2015 publication had an illustration of adding noise to a ground-truth image, shown in Fig. 5 as B/W images of different noise levels [10].

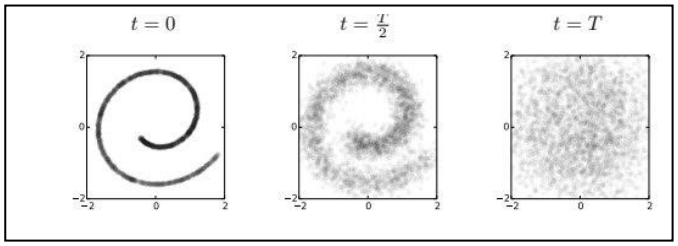

Fig. 5. A 2-dimensional spatial illustration of the forward diffusion process in B/W, adapted from Sohl-Dickstein et al. [10].

The image entropy values were computed with the Swiss-roll data example of Sohl-Dickstein et al. [10], and shown in Fig. 6. The entropy trend in forward trend (shown in blue) is reversed in the reverse diffusion (shown in orange). A research question whether continued diffusion would generate asymptotic convergence is open in this specific student research project stopping at 500 timesteps. From a data science start-up company perspective, more budget is needed.

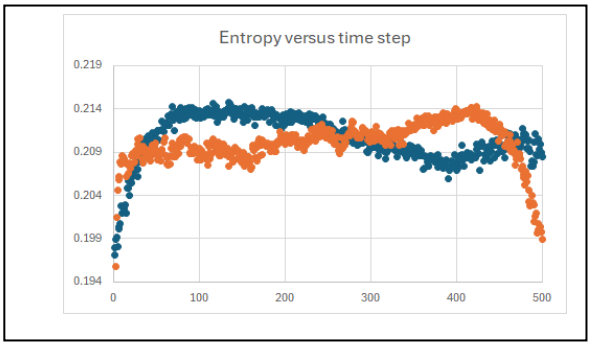

Fig. 6. Python Skimage.measure.shannon entropy values (y-axis in unit of bits) versus time steps (x-axis), using the Swiss-roll data example of Sohl-Dickstein et al. [10].

V. Sustainability

The students' interests in Generative AI text to image technology and AI technology in lab automation will support the continuation of using AI tools in student research projects. With faculty effort to include more coverage of diffusive heat transfer in Calculus Physics One, together with lab automation in engineering technology courses, the student research group using AI assisted tools will continue with new student members.

The AI Large Language Models (AI LLM) to translate C++ codes to Python codes is an efficient feature in some student projects, for instance, some drone control programs written in C++ codes could be transformed to become Python codes with the advanced AI LLM versions that require subscription fees of 10 to 20 dollars per user per month. Although our Community College Administration does not allow an instructor to use subscription-required AI tools in classrooms, faculty can translate the C++ codes to Python codes and deliver student skill learning projects, not necessarily student research projects which require new results. In any event, the skill learning project students and research project students can be a single learning community with faculty doing the organization.

VI. Discussion

The faculty members aim to create a learning community of three attributes, namely, vibrant, sharing and welcoming with a common theme of AI data science using entropy calculation as a versatile tool with a guarantee of some new results. The above student research projects can be extended when introducing fractal dimension calculations as complexity analysis, Langevin dynamics and Fokker Planck equation as diffusion analysis, etc.

For instance, the fractal dimension analysis of the images studied in Figure 4 yielded the following data trend, shown in Fig. 7. The entire Sun seemed to relax/decay to a minimal complexity state after the eruption that caused the 2024 Mother's Day Geomagnetic Storm. Continued research using entropy and fractal dimension calculations are warranted with additional budget request in the perspective of a data science company.

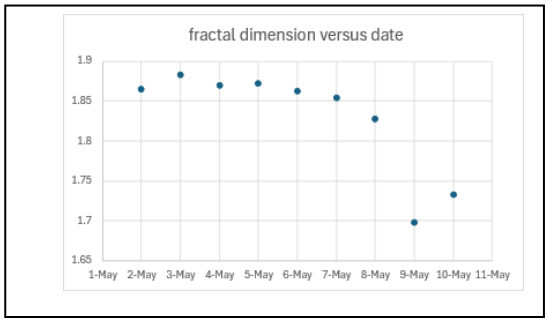

Fig. 7. Solar Dynamics Observatory 9.4 nm fractal dimension (y-axis) versus calendar date (x-axis), using the same dataset of Fig. 4.

The faculty members do not discourage students to continue the research in the perspective of a subject matter. It is important to encourage the choices of students in their initial majors. The above learning community strategy is consistent with the student population. There are about 5 physics majors, 50 engineering majors, and 500 ET majors in which 20 of them are considering transition to engineering. None of the participants in this report is majoring in physics. In fact, all the participants started as skill learning project participants.

AI instrumentation, AI data science, AI LLM are all included as project topics in the learning community. AI data science using entropy gives new results, AI LLM is relatively riskier given low GPU computing in our budget, and AI instrumentation needs a sizable budget and could end up as skill learning projects when there is no evidence of any new results.

VII. Conclusions

The Shannon entropy as a research tool is effective in holding up a learning community. The paper described three types of research projects, namely, bioinformatics research, solar physics research, and diffusion model Generative AI research using entropy calculations. The bioinformatics research component showed that entropy classification could support new hypotheses. The solar image entropy analysis results support a threshold hypothesis model in strong geomagnetic events. The diffusion model Generative AI analysis results encourage students to respect AI technology with a deeper understanding of the mechanisms. Although our GPU capacity is well below those R-1 University Computing Centers, the recent release of DeepSeek AI, etc. requiring less GPU capacity could be a viable practice to continue student skill learning projects and student research projects in diffusion model Generative AI in a community college setting.

Acknowledgment

We thank those sharing their works using open access platforms. The professor co-authors SD, GA, and TC contributed equally.

References

[1] J. Enright, Z. Dickson, and G. Golding, “Low Complexity Regions in Proteins and DNA are Poorly Correlated”, Mol Biol Evol. 2023 Apr 4;40 (4).

[2] L. Teekas, S. Sharma, and N. Vijay, “Terminal regions of a protein are a hotspot for low complexity regions and selection”, Open Biol. 2024 Jun;14(6):230439.

[3] B. Leung, unpublished data, Year of 2025 Great Neck South High School, New York State (QCC physics students 2023 Summer and 2024 Summer)

[4] National Center for Biotechnology Information NBCI. https://www.ncbi.nlm.nih.gov/gene/?term=zenk

[5] Solar Dynamics Observatory. https://sdo.gsfc.nasa.gov/

[6] Kyoto University Geomagentic DST index. https://wdc.kugi.kyoto-u.ac.jp/dst_provisional/202405/index.html

[7] N. Kondrashova, M. Pasechnik, and S. Osipov. “Evolution and flare activity of Carrington-class Solar Active Region NOAA 13664 and its impact on the Earth”, November 2024 Odessa Astronomical Publications 37:112-117

[8] R. Jarolim, A. Veronig, S. Purkhart, P. Zhang, and M. Rempel, “Magnetic Field Evolution of the Solar Active Region 13664”, The Astrophysical Journal Letters, 2024 April 976 L12.

[9] O. Kruparova, V. Krupar, A. Szabo, D. Lario, T. Nieves-Chinchilla, and J. Oliveros, “Unveiling the Interplanetary Solar Radio Bursts of the 2024 Mother's Day Solar Storm”, The Astrophysical Journal Letters, 2024 April 970 L13.

[10] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep Unsupervised Learning using Nonequilibrium Thermodynamics”, 2016 https://arxiv.org/abs/1503.03585