Title: Explainable-AI driven feature engineering of CRISPR-Cas9 sgRNA efficiency leads to quantum biological insights into genome editing

Authors: Jaclyn Noshay[1]* (noshayjm@ornl.gov), Ashley Cliff[1,2], Jonathan Romero[1,2], Stephan Irle[3], David Kainer[1], Daniel Jacobson[1], and **Paul Abraham**[1]

Institutions: [1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [2]The Bredesen Center, University of Tennessee, Knoxville, TN; [3]Computational Chemistry and Nanomaterial, Oak Ridge National Laboratory, Oak Ridge, TN

Website URL: https://seed-sfa.ornl.gov/

Project Goals: The Secure Ecosystem Engineering and Design (SEED) Science Focus Area (SFA), led by Oak Ridge National Laboratory, combines unique resources and expertise in the biochemistry, genetics, and ecology of plant-microbe interactions with new approaches for analysis and manipulation of complex biological systems. The long-term objective is to develop a foundational understanding of how non-native microorganisms establish, spread, and impact ecosystems critical to U.S. Department of Energy missions. This knowledge will guide biosystems design for ecosystem engineering while providing the baseline understanding needed for risk assessment and decision-making.

Abstract Text:

High-throughput genome-wide studies using CRISPR-Cas tools have transformed capabilities for genetic manipulation in the laboratory. However, the performance of these tools is prone to error and increased uncertainty beyond model organisms and laboratory-controlled conditions. This is because current models for sgRNA prediction have primarily been trained on a narrow range of model species including human, mouse, drosophila, and zebrafish. While general "rules" have been defined in these organisms, the mechanistic underpinnings are not well understood. Additionally, with extreme variation in the genetic architecture between species and kingdoms, mammalian data alone lacks the information to generate an accurate predictive model in another organism. Herein, we utilized a feature engineering machine learning model, iterative random forest (iRF), to better understand the features of importance when considering an effective sgRNA. Using recently published data assessing genome-wide activity of gRNA for *E. coli*, we identified traits important for sgRNA design in a bacterial species and observed immense variation when tested with the same feature set in human and fungal species. We show the influence of expanding positional encoding to larger k-mers to capture intricate interactions in local and neighboring nucleotides. Most interestingly, we have utilized a novel feature set, quantum chemical tensors, that can be applied across species to improve sgRNA efficiency prediction as well as greatly enhance our understanding of the intricate quantum biological processes involved in CRISPR-Cas9 machinery. These advancements will provide a means to improve the safety and reliability of biosystem design and ecosystem engineering using non-model organisms.

Funding Statement: The Secure Ecosystem & Engineering Design Science Focus Area is sponsored by the Genomic Science Program, U.S Department of Energy, Office of Science, Biological and Environmental Research, under FWP ERKPA17. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract no. DE-AC05-00OR45678. This program is supported by the U. S. Department of Energy, Office of Science, through the Genomic Science Program, Office of Biological and Environmental Research, under FWP ERKP123.