Pacific Northwest
NATIONAL LABORATORY

PNNL-29478

Automated novel molecule structure determination for discovering pathways critical to soil microbiomes at the terrestrial-aquatic interface – 6 month project report

September 2019

Ryan S. Renslow
Sean M. Colby
Tino J. Wells
Madison R. Blumer

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-29478

Automated novel molecule structure determination for discovering pathways critical to soil microbiomes at the terrestrial-aquatic interface – 6 month project report

September 2019

Ryan S. Renslow
Sean M. Colby
Tino J. Wells
Madison R. Blumer

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Acknowledgments

This research was supported by the Earth & Biological Sciences Investment, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

1.0 Specific Aims

This 6-month $75k project focused on delivering the Foundational Aim and initiating work for Aim 1, detailed below. The Foundational Aim lays the groundwork to complete Aims 1-3, which are anticipated to be funded fully under a future project (either externally or internally).

Foundational Aim. Support tools and data required for automated structure elucidation
The deep learning methods detailed below, which are at the core of delivering all Aims of this future project, require capabilities to 1) pre-process data (2D NMR spectra) so it is amenable to deep learning input and training, 2) appropriately structure a machine learning network to convert 2D NMR spectra into chemical structures, and 3) validate the chemical soundness of the molecular structure output from the network. For this aim to succeed, we need to demonstrate successful data consumption (**input**), conversion to structures (**network architecture**), and sensible predicted chemicals (**output**). This foundational aim will deliver the tools and data required to ramp up for a full project (pending funding) that will deliver all Aims.

Quick 6-month project deliverables:

Deliverable 1. Input Training Data Format: automated conversion of ISiCLE density functional theory data into associated 2D NMR spectra (e.g. COSY, NOESY, HMBC) using spin dynamics simulation (*spinach*). Input structures will initially come from the Universal Natural Product Database, focusing on simple flavonoids.
*****Success is defined as demonstrating a new automated software tool that converts raw ISiCLE output into human- and machine- readable 2D NMR spectra.

Deliverable 2. Operational Deep Learning Network Architecture: initial deep learning network (Keras-based) that consumes multiple 2D NMR spectrum and predicts molecular structure (SMILES).
*****Success is defined as demonstrating a deep learning network that can create chemically-sound SMILES output from an input of 2D NMR spectra. Note that there is no expectation at this point that the predicted molecular structure will be experimentally validated, because the network will not have been fully trained on the thousands of real datasets required to achieve accurate predictions, which is a full Aim 1 outcome. This Foundational Aim is only laying the groundwork for subsequent transfer-learning-based training on thousands of datasets.

Deliverable 3. Output Molecular Structural Soundness: development of a deep learning discriminator, validated against a gold-standard SMILES structure validators (ChemAxon, rdkit), to ensure predicted molecules have realistic structures (i.e., not breaking known chemical bonding rules).
*Success is defined as demonstrating a discriminator trained on over 1 million SMILES and can operate under 1 second per molecule (single node run).

The Foundational Aim lays the foundation for success in these Aims under a longer project:

Aim 1. Build and validate deep learning tool for automated NMR structure elucidation

Deep learning methods require abundant, diverse training data in order to realize a robust, generalized model and to avoid overfitting to an underrepresented subset of chemical space. To create sufficient 2D training spectra (COSY, NOESY, HMBC, HSQC, and TOCSY), known molecular structures from the Universal Natural Product Database and MetaCyc will be processed by our recently developed quantum chemistry-based property prediction pipeline – ISiCLE, the *in silico* chemical library engine – to calculate ^{13}C and ^{1}H chemical shifts with high fidelity, followed by processing with *spinach*, an open-source spin dynamics simulation library. We pioneered ISiCLE to identify known compounds and calculate chemical properties, which led to a large, now-funded NIH grant, and recently demonstrated sub-0.1 and sub-1.0 ppm error for ^{13}C and ^{1}H chemical shifts, respectively. This is sufficiently accurate for deep learning applications, with direct relevance to actual experimental data. Additionally, after initial training, transfer learning with experimental data (BMRB and in-house) will be employed to further refine the model. All tools in this project will be developed in Python, leveraging appropriate packages, for use across desktop, HPC (*Constance*/*Cascade*), and cloud compute resources (Azure, AWS, and as a stretch goal, Google Compute Engine). Our team, which includes computer science masters, has experience with modern software development practices (version control, continuous integration, testing), with exception of cloud computing, for which we have begun collaboration with M. Macduff and T. Martin (PNNL Cloud Resource team).

Details for the Deep Learning Neural Net: The deep learning core of this software will be built using Keras (a high-level neural network Python package), leveraging the expertise gained from the recent and continued success of DarkChem, a product of the DeepScience agile investment. This will include a (suite of) semi-supervised variational autoencoder(s), powered by 2D convolutional layers coupled to a dense prediction layer, capable of encoding the targeted 2D NMR spectra for each experiment (COSY, NOESY, HMBC, HSQC, and TOCSY) into a continuous numerical — or latent — representation of structure and property information. Furthermore, the neural net will include a valid structure discriminator to ensure points in latent space map to syntactically correct and chemically feasible structures. The beauty of our approach is that our custom training methods will result in a latent representation of chemical space, shaped according to chemical shifts, and subsequently used to generate structures from raw experimental data. The neural net will be designed to remove spurious noise, initially added synthetically, and then added from actual experimental samples, prior to decoding the correlated peaks directly into a molecular structure.

As the software progresses, we will begin increasingly stringent assessments, in which a series of blinded controlled tests will be conducted to validate our approach. EMSL NMR facilities will be used to collect the COSY, NOESY, HMBC, HSQC, and TOCSY spectra (additional experimental details below, Aim 2). An associate will create of a set of test samples, without revealing the comprised molecule to either the computational or experimental teams. Initially, validation samples will consist of a single analyte out of a pre-determined set of 10 molecules purposefully removed from the training sets and associated libraries. As success is demonstrated with these out-of-sample, but known, compounds, the out-of-sample set size will be increased to 100. To add complexity and simulate real samples, after success with the pure blinded samples, the approach will be evaluated on a complex, but muted, background (e.g., diluted MPLEx extract of soil) in addition to the blinded analyte. This will directly test the denoising capability of the software.

Aim 2. Apply approach to find novel molecules and pathways in SPRUCE soil samples

The boreal peatland soil samples are in-hand, and will be provided by Kirsten Hofmockel. All samples will be extracted with the soil MPLEx extraction method developed in EMSL, followed by vacuum liquid chromatography using increasing increments (25%) of CH_2Cl_2 in hexane followed

by MeOH in CH_2Cl_2. Fractions will be monitored and further purified using TLC, focusing on high intensity peak regions determined from the in-hand LC-MS data. Samples will be analyzed using a Varian 600 to generate COSY, NOESY, HMBC, HSQC, and TOCSY spectra. The Varian 600 (EMSL) is equipped with a HCN z-gradient cold probe, including a cold ^{13}C preamp. The sample will be dissolved in 260 µL $CDCl_3$ and transferred to a susceptibility-matched Shigemi tube. Experiments will be collected using standard parameters. Finally, the raw data will be fed into our software for novel structure determination. When possible, new identified molecules will be placed within the context of (new) microbial metabolic pathways (via Pathway Tools, using the already obtained metagenomics data).

Aim 3. Refine approach to find the minimum set and minimum quality of NMR experiments required

Following initial tests, validation, and application through Aims 1 and 2, the neural network will be improved to decrease the number of experiments required to consistently generate correct structures. There is the potential that, e.g., only COSY/TOCSY, HSQC, and HMBC data will be required, meaning that less data would need to be captured in the lab. Using a Monte Carlo approach, individual data sets will be removed until we determine the minimum set of data required. Furthermore, noise will be added at increasingly high levels to understand the tradeoff between noise level and required number of experiments, as will tradeoffs between peak picking methods and peak picking delinquencies.

2.0 Objective

The exact structures of most molecules in the environment are unknown and have never been identified or synthesized. Unambiguous molecular structure determination is currently restricted by the time and effort necessary to isolate compounds and perform *de novo* structure elucidation, either using NMR or x-ray crystallography. For example, (i) approaches for data analysis remain predominantly manual and (ii) elucidation success is sensitive to data quality and data processing. Both issues impede throughput and increase the likelihood of errors, especially for extracts of complex samples, which can have rich background signals. Deep learning has been demonstrated to excel in these cases by intelligently handling noise and enabling automation of previously intractable manual tasks.

As the agents of the underlying chemical processes, small molecules mediate complex systems such as ecosystems, human physiology, and Earth's atmospheric- and geo-chemical cycles. The full project that is primed by the foundation aim, will primarily couple molecular models and deep artificial neural networks to elucidate the novel structures comprising these complex systems. Furthermore, for our future specific application in boreal peatland soil microbiomes, the new molecules will be assigned to their associated metabolic pathway and in context of their ability to govern mesoscale nutrient-cycling rates in a relevant terrestrial-aquatic interface system.

3.0 Background and Significance

Inadequate understanding of small molecules' role in soil microbial functions limits accurate prediction of carbon (C) and nutrient cycling in terrestrial-aquatic ecosystems. Natural products (secondary or specialized metabolites), lignin and humic fragments, and degradation products are organic small molecules produced by plants and microbes, some of which are not directly involved in growth and reproduction, but are produced as a consequence of interactions with other organisms and/or the environment. These molecules enter the soil via leaf litter leachate, root exudate, and aboveground volatile emissions, or may be created directly in soil through microbes and biogeochemical reactions. Experimental evidence has shown that small molecules can drastically influence C and nutrient cycling in soils, for example, by reducing rates of litter decomposition or by directly suppressing bacterial and fungal growth (e.g., as with phenolics and terpenoids). Small molecules can affect C turnover rates to such an extent that their use has been proposed as a potential means for reinforcing C sequestration techniques.

We currently do not know how the molecules interact or which compounds significantly impact biological activity in these systems. Our insufficient knowledge of small molecule composition and associated impact on microbial activity hinders accurate C cycling models. Our limited understanding is primarily due to the inability to identify the vast majority of molecules, track their fate, and assess their impact on metabolism through soil microbial communities. For this proposed work, target soils will come from the Spruce and Peatland Responses Under Changing Environments (SPRUCE) site. This boreal peatland climate change experiment represents an important terrestrial-aquatic interface, with minimal nutrient inputs, making decomposition and microbial metabolism the primary source of nutrients for plant growth. These ecosystems store 1/3 of the global soil C pool and are particularly vulnerable to C release with increasing temperatures. Characterizing the small molecules and associated microbial metabolic networks in this system will enable us to generate hypotheses about how peatlands will retain or release C and cycle nutrients in future climates.

4.0 Technical Milestones for the 6-month LDRD project

- **Deliverable 1. Input Training Data Format**
  Success is defined as demonstrating a new automated software tool that converts raw ISiCLE output into human and machine readable 2D NMR spectra.
- **Deliverable 2. Operational Deep Learning Network Architecture**
  Success is defined as demonstrating a deep learning network that can create chemically-sound SMILES output from an input of 2D NMR spectra.
- **Deliverable 3. Output Molecular Structural Soundness**
  Success is defined as demonstrating a discriminator trained on over 1 million SMILES and can operate under 1 second per molecule (single node run)

5.0 Mission Relevance

A large portion of DOE research programs seeks to understand the biological, biogeochemical, and physical processes at the molecular scale, requiring knowledge of the molecules present in complex samples. Specifically, DOE is interested in metabolic pathways, biological systems, active phenotypes/functions, industrial reactions, and these topics as they relate to biofuels, bioproducts, earth systems, and climate. This project wholly aligns with the specific goal of BER BSSD: gain a predictive understanding of complex biological systems. Furthermore, this project will demonstrate characterization of small molecules that may mediate pathogenesis in microbial pathogens (NSD), control molecular communication in microbiomes (for both NIH and DOE), and facilitate synthetic biology efforts across EBSD and NSD and biofuels work in EED. Initial proposals for follow-on funding after completion of all Aims will target the DOE (Early Career Research Program), NIH, and DHS. Within PNNL, this project clearly addresses key elements of the Decoding the Molecular Universe and Harnessing the Microbiome directorate objectives, and our initial application with soil from the SPRUCE aligns with terrestrial-aquatic interface goals.

This proposed project pushes our team in a very different direction than any currently funded and anticipated projects, and is out of scope of recent DeepScience agile investment. Whereas our growing standards-free approaches rely on predicting chemical properties from molecular structures of known chemicals, the proposed work will focus entirely on the true unknowns, that is, the molecules in samples whose structures have never been determined. There are no known existing groups pursuing the approach described here, and success in this project would represent a major leap forward for NMR-based *de novo* structure elucidation.

6.0 Results and Discussion form Project

Neural networks are increasingly being used in computational chemistry for property prediction and new molecule generation. Previous work by this group has created a variational autoencoder called DarkChem that takes in SMILES strings, learns how to represent in a compressed vector form, and then decode that representation back to a SMILES string. However, new molecules generated by creating slight variations in the compressed vector tend to be invalid SMILES strings and thus invalid molecules. Through this project, our team has been working on an additional network to determine whether a SMILE string is syntactically valid to improve DarkChem's output.

To create this network, data representing both 'valid' and 'invalid' SMILES is needed. Valid SMILES were collected from various chemical databases representing the chemical space of interest. Invalid SMILES are created by three methods: simple string perturbations, a Markov chain model, and DarkChem's invalid output. The perturbing method attempts to make invalid SMILES as close to valid SMILES as possibly by making changes to valid SMILES strings: switching characters, deleting characters, and combinations thereof. The Markov chain model is trained on SMILES from the database and then constructs SMILES character-by-character based on the learned probability of the next character given the previous one. It creates mostly invalid SMILES.

The neural network uses the encoder structure from the DarkChem network, then adds a dropout layer to prevent overfitting on the data and a fully connected dense layer. A sigmoid activation is used to make the final prediction probability of 'valid' or 'not valid'.

As numerous invalids may be generated based on the fixed set of valids, we tested different methods for training on imbalanced data, where one class (invalids) is overrepresented in training data compared to the other class. Repeating the valids to equal the number of invalids (upsampling) produces the fewest false positives, or invalid SMILES identified as valid. The lowest number of false predictions occurs when networks are combined in an ensemble, where predictions are averaged from five networks train on upsampled data and five train on downsampled data, where invalid SMILES are randomly selected to equal the number of valids.

Initial results have been promising, with most training methods achieving above 98% accuracy on a balanced-class reserved set of data . Further investigation is needed to make sure the network is learning the difference between valid and invalid SMILES syntax rather than differences in the methods of construction of the dataset.

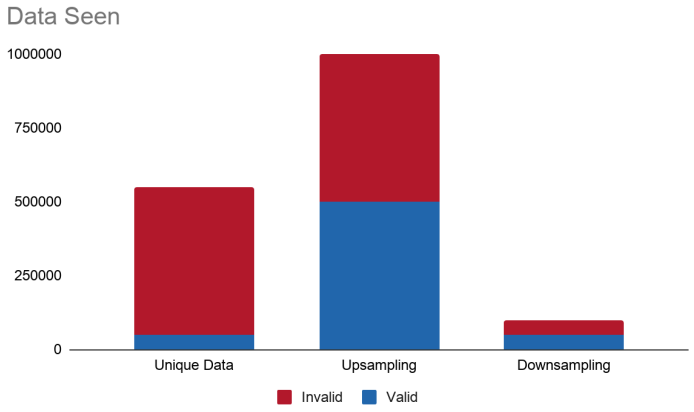

Figure 1. The first column represents the amount of unique valid (red) and invalid (blue) SMILES available, followed by the number of samples seen by the network in the two best-performing individual networks.

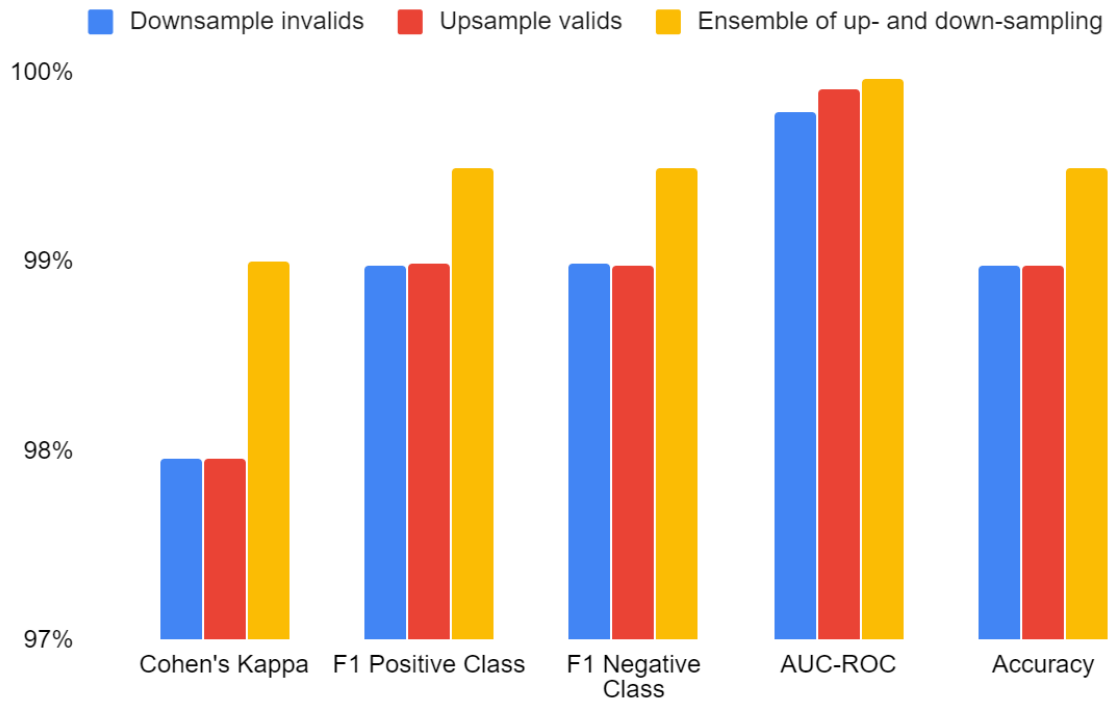

Figure 2. Five different measures comparing the two most successful individual networks with their ensembled counterpart, evaluated on the same reserved set of 50,000 valids and 50,000 invalids. F1-score is the harmonic mean of precision and recall, or the percent of SMILES predicted valid that are actually valid and the percent of actually valid SMILES predicted to be valid. In the negative case, this is the harmonic mean of the percent of SMILES predicted invalid that are actually invalid and the percent of actually-invalid data that is predicted invalid. The AUC-ROC is the area under the AUC-ROC curve, which shows the true positive and false positive rates at all possible classification thresholds. Accuracy is the number of correctly identified observations divided by the total number of observations.

Another deliverable of this project was to develop a new automated software tool that converts raw ISiCLE output into human and machine readable 2D NMR spectra. The primarily focus was developing a software stack that converts quantum chemical (i.e., density functional theory) data into multiple types of two-dimensional spectra. Currently, this stack is based in MATLAB, but will soon be fully-integrated into python in the near future due to being more compatible and flexible with other tools our team has/are developed/developing. A major piece of this stack being *spinach*; a fast, polynomial-complexity scaling spin-dynamic simulation, to wit, we run NWChem output data through this stack—and thus *spinach*; a tool were calling N2S. The resulting spectra from this process will eventually be concatenated as an individual input (i.e., batch) to a multi-channeled convolutional neural network in subsequent projects. Each channel will consist of a single two-dimensional spectra, followed by a series of convolution and pooling layers. It is important to note that this sequence of layers is architecturally-based on Google's Inception (v2) framework, and not the conventional VGG16 (i.e., repeated convolution-pooling layers). This architecture was chosen due to objectively outperforming VGG16 models, but also other major contributors' systems, such as Facebook's DenseNet, in terms of accuracy and computational

efficiency. Currently unnamed, this neural network will allow us to represent chemical space from another perspective, and explore said space in ways that have not been performed before, thus potentially providing tremendous amounts of insight.

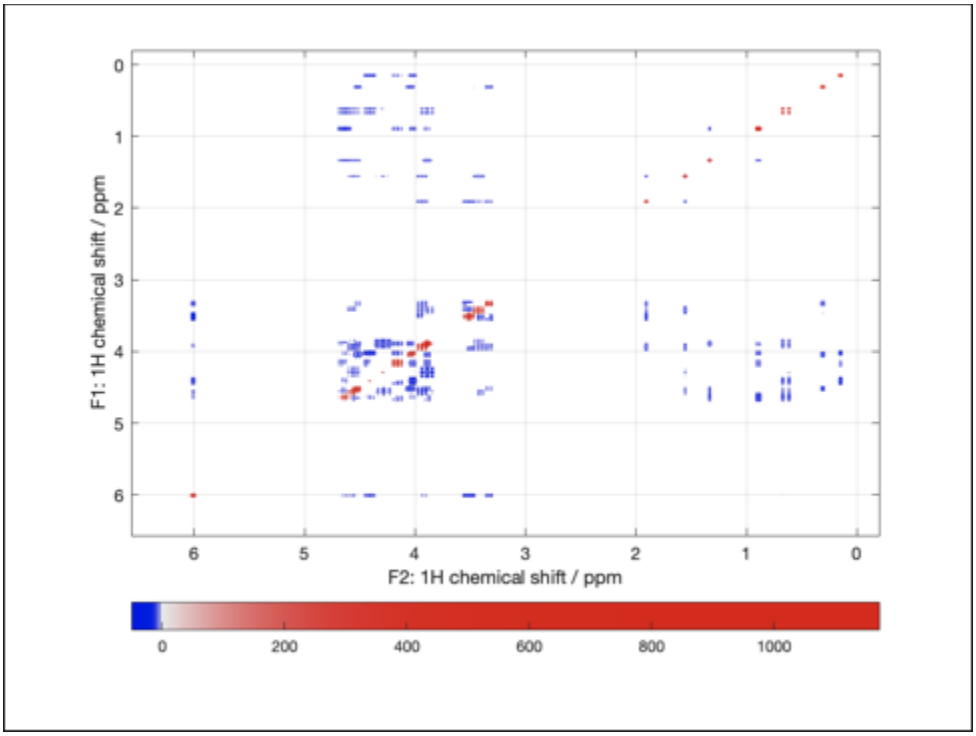

Figure 3: Individual spectra, known as nuclear magnetic resonance spectra (NMR), of molecules provide insight to said molecule. Above, a NOESY spectra generated using our tool developed under this project (NWChem output to *spinach* software) of sucrose shows spin polarization from one nuclei to another through cross-relaxation, shining light on its resonance; something the convolutional neural network will be able to learn and understand in order to automatically determine the underlying molecular structure.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov