2022

Artificial Intelligence based Approach for Rapid Material Discovery: From Chemical Synthesis to Quantum Materials

Robert Tempke
rstempke@mix.wvu.edu

Artificial Intelligence based Approach for Rapid Material Discovery: From Chemical Synthesis to Quantum Materials

Robert Sean Tempke

Dissertation submitted to the Benjamin Statler College of Engineering at West Virginia University

in partial fulfillment of the requirements of the degree of

**Doctor of Philosophy
in
Mechanical Engineering**

Terence Musho, Ph.D., Chair
Ali Baheri, Ph.D.
Debangsu Bhattachayya, Ph.D.
Edward Sabolsky, Ph.D.
Christina Wildfire, Ph.D.

Department of Mechanical and Aerospace Engineering

Morgantown, West Virginia
November 2022

Keywords: Artificial Intelligence, Machine Learning, Dielectrics, Prediction, Gas-phase, Catalyst, Variational Autoencoder, Single Photon

ABSTRACT

Artificial Intelligence based Approach for Rapid Material Discovery: From Chemical Synthesis to Quantum Materials

Robert Sean Tempke

With the advent of machine learning (ML) in the field of Materials Science, it has become obvious that trained models are limited by the amount and quality of the data used for training. Where researchers do not have access to the breadth and depth of labeled data that fields like image processing and natural language processing enjoy. In the specific application of materials discovery, there is the issue of continuity in atomistic datasets. Often if one relies on experimental data mined from literature and patents this data is only available for the most favorable of atomistic data. This ultimately leads to bias in the training dataset. In providing a solution, this research focuses on investigating the deployment of ML models trained on synthetic data and the development of a language-based approach for synthetically generating training datasets. It has been applied to three material science-related problems to prove these approaches work. The first problem was the prediction of dielectric properties, the second problem was the synthetic generation of chemical reaction datasets, and the third problem was the synthetic generation of quantum material datasets. All three applications proved successful and demonstrated the ability to generate continuous datasets that resolve the issue of dataset bias.

This first study investigated the synthetic generation of complex dielectric properties of granular powders and their ability to train a ML network. The neural network was trained using a supervised learning approach and a common backpropagation. The network was double-validated using experimental data collected from a coaxial airline experiment.

The second study demonstrated the synthetic generation of a chemical reaction database. An artificial intelligence model based on a Variational Autoencoder (VAE) has been developed and investigated to synthetically generate continuous datasets. The approach involves sampling the latent space to generate new chemical reactions that were assembled into the synthetic dataset. This developed technique is demonstrated by generating over 7,000,000 new reactions from a training dataset containing only 7,000 reactions. The generated reactions include molecular species that are larger and more diverse than the training set.

The third study investigated a similar variational autoencoder approach to the second study but with the application of generating a synthetic dataset for quantum materials focusing on quantum sensing applications. The specific quantum sensors of interest are two-level quantum molecules that exhibit dipole blockade. This study offers an improved sampling algorithm by continuously feeding newly generated materials into a sampling algorithm to help generate a more normally distributed dataset. This technique was able to generate over 1,000,000 new quantum materials from a small dataset of only 8,000 materials. From the generated dataset it was identified that several iodine-containing molecules are candidate quantum sensor materials for future studies.

Thank you to my best friend, supporter, and travel companion, my amazing wife Rhiannon Tempke. She has been by my side for nearly every major moment in my life and supported all of my hopes and my dreams. I am looking forward to this next chapter with you.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor and mentor Dr. Terence D. Musho for all his support and guidance in leading me through my research. His guidance and advice were paramount to my success and growth over these last few years. I have known and worked with Dr. Musho since I was an undergraduate, he has guided me through both my Masters' degree and now my Ph.D. His vision and thoughtfulness have helped me to achieve things I did not think would be possible when I first met him. He has always encouraged me to pursue whatever path I was interested in and has carved out a space for me within his research group.

I would also like to give special thanks to Dr. Christina Wildfire and Dr. Dushyant Shekhawat who have also been paramount in my success. Through both research support and mentorship. They have also seen me through every degree I have every achieved and believed in me even when I did not believe in myself. I made frequent use of their open door to ask them a constant barrage of questions which they tirelessly answered. They have always encouraged me to be bold and to go after problems that no one has even thought of yet. For this, I will be forever grateful to them.

Special thanks are required for all my colleagues at NETL, who despite some less-than-conventional motivational methods were of great help in all my research. One thing I would like to remember is donuts are not free.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

Introduction

Every day, at almost any given moment most humans are interacting with something that is powered by artificial intelligence (AI) in some capacity. It is one of the most powerful tools of our age and has helped to facilitate some of the greatest advancements of this century[1–3]. As it has become more widely accessible it has worked its way into almost every field of research. Where researchers in fields such as material science are hoping it can help them make evermore advanced breakthroughs in material synthesis. AI.

The study of material synthesis has been around for decades, yet despite talented researchers and strong theory, most material discovery is due to experience-guided trial and error methodologies[4]. Even with tools such as density functional theory material discovery is typically a high manpower and timely cost of resources. Researchers have been attempting to automate the material discovery processes for many years. They have used tools like rapid structural phase identification, combinatorial materials synthesis, and high throughput experimentation[5,6]. While these methods have helped immensely, they all have the same flaw of requiring human oversight for the design and execution of experiments.

One of the areas that AI can have a large impact on the field of material science is in the prediction of material properties. Historically these properties have been simulated using things like Monte Carlo methods and Molecular Dynamics[7–9]. Unfortunately, these simulations are often heavy on assumptions and require a lot of computational time and complexity. AI however while training can be costly once a model is trained it is typically very fast to predict properties at a low computational cost. AI has already had major success predicting electrical properties for materials such as band gaps, superconductivity, and topological states[10,11].

The major problem holding back AI in its pursuit of material predictions is the lack of available datasets. Unlike other fields that have seen wide adoption of AI, material science has a distinct lack of available data. The available data is typically taken from journal publications which leads to a strongly biased dataset that is curated only with the “best” materials. Despite great work being done by the materials genome project and others like it, these small and biased datasets are still holding back AI in material research[12,13]. To make the same breakthroughs in materials research that other fields have enjoyed, the paramount problem to solve is the creation of large and diverse datasets.

Study 1: Dielectric Properties Prediction

Study one investigates and implements a machine learning (ml) algorithm for the use of calculating the dielectric properties of solid materials in a coaxial airline using the observable parameters S11 and S21. This approach like Tuck and Coad's will allow for the calculation of the dielectric material to take place directly from the measured S-parameters without the need to de-embed the air in the coaxial airline. Thus, simplifying the mathematics and intrinsic error. The approach is to use supervised learning to teach and validate an algorithm using simulated S-parameters and dielectric data. The advances in modern computational power allow for high-fidelity simulated data for almost all feasible combinations of this observable-property relationship.

Once trained the CNN model is validated on experimentally collected S-parameters for known dielectric materials. This approach was selected because of the precise control of the input and outputs being used in training the system and its reproducibility as well as its ability to capture a large portion of the solution space. The system attempts to achieve increased accuracy over previous models by utilizing the additional input parameters available when testing in the coaxial airline. The previous studies were only able to utilize the reflected coefficient of the S-parameters (S11) because of the limitations of the coaxial probe method. The coaxial airline method provides both the forward and reflected coefficients of the S-parameters (S11 and S21).

The methodology for this study is to generate simulation solutions for a wide variety of dielectrics. In addition to varying the dielectric properties in the simulation, the specimen length within the coaxial line was also varied. It is postulated that for any sample length and dielectric constant there exists a unique set of inputs (S-parameters) that generate that solution for a given testing geometry. This would allow the ANN method to be length invariant, a considerable advantage over the precise length measurements needed by some of the classical measuring techniques. By teaching a machine learning algorithm these relationships based on multiple conversion methods a more robust and accurate solution can be obtained than previously existed. Utilizing the simulation results with machine learning can potentially result in a much faster and less computationally intensive solution methodology. Together these techniques can provide a new solution method for converting S-parameters to dielectric properties.

Study one differs from the next two studies in that it serves as proof that ANN can be trained on synthetic data in order to make real-world predictions. It utilizes widely accepted and tested software to generate the synthetic data so that the focus of the paper can remain on the ANN predictions and not on the synthetic data production. In the following studies, the focus is instead shifted to dataset creation using machine learning. While utilizing the results of this study to prove that synthetic data can be used to extend what is currently possible in machine learning.

1.1 Study 1 Significance

With the increased rise of machine learning throughout the field of material science it has become increasingly important to make advancements with this new tool to increase the fidelity of models being created. This will allow for a more accurate predictive machine learning model especially for high dimensional problems, over traditional mathematical models. Allowing the field of material science to keep pace with other fields such as the medical and image recognition fields. This can be exemplified in the area of dielectrics and their synthesis techniques. Where there has been a growing need to increase inverse models to predict dielectric properties of materials based on measurable observables. For the case of dielectrics, this observable is scattering parameters recorded using a coaxial transmission line testing method for dielectrics. Traditional mathematical solutions for solving scattering parameters to dielectric solutions are often iterative or unstable at one-half wavelengths[14]. Also, two measurements taken minutes apart can return varying results. This leads to uncertainty in the calculated values of these dielectric values. The fundamental design of devices in the field of microwave engineering needs to have confidence in the measurements of a material's dielectric properties. It was hypothesized that a machine learning approach could be designed that transformed scattering parameters into dielectric property predictions.

1.2 Study 1 Background

Artificial Intelligence has rapidly become one of the most important tools in the scientific community used for everything from fundamental characterization to real-world applications. This is especially true in the field of chemistry and chemical engineering where machine learning has been used in conjunction with density functional theory (DFT) as well as on its own to produce unique insight and provide rapid results[15,16]. Machine Learning has also been used in the chemical engineering field to help search the solution space of possible reactions to help save both time and resources[17–22]. It has also shown great promise in the field of catalyst development, where it has demonstrated help in the creation of state-of-the-art catalyst[23–25]. This study focuses on three main areas where artificial intelligence (AI) can help to advance the field of materials research and engineering. The first is using machine learning to overcome traditional mathematical obstacles in the characterization of dielectric materials which is key when developing materials for us in EM fields.

Study one focuses on the prediction of dielectric properties, where it is shown that the use of AI can produce highly accurate and stable predictions over that of traditional mathematically derived techniques. This research will rely on the use of electromagnetic computational solutions to generate synthetic data to train the ML model.

At GHz frequencies, the electromagnetic (EM) interactions are quantized by a material's dielectric properties or the dynamics of dipole interactions. The dielectric constant of a material is the ability of the material to store electrical energy. While the loss tangent of a material is a quantification of the energy loss. The complex dielectric is defined as ε_r= ε'-iε", where ε' is the real portion and ε" is the imaginary portion. For microwave material engineering, the characterization of material dielectrics as a function of frequency and temperature are critical to understanding the response. Dielectric properties can often be difficult to ascertain for granular materials as the shape and distribution influence the polarizability and subsequently the dielectric response[26–28].

To date, the characterization of these properties has been done using a multitude of inverse mathematical techniques. Many of the techniques require initial guesses to avoid discontinuities arising from the resonance of the system[29–31]. Schwab et al. have recently shown what a powerful tool machine learning can be for solving inverse problems using latent information in high dimensions[32]. Recent computational advancements in the ability to conduct large numbers of permutations of solutions with high accuracy have ushered in the potential to revisit many of these inverse methods using a machine learning approach. By utilizing the inverse space of a problem Sahoo et al. showed that machine learning could be used to understand functional relationships between data to extract underlying equations showing great agreement with the paper published by Schwab et al[33]. Moreover, Zhao et al. showed that machine learning methods could be applied to a wide variety of microwave device modeling techniques, both active and passive devices achieving efficient and fast characterization[34].

The complexity of many dielectric materials such as in the case of microwave chemistry and the use of heterogeneous (multi-component, macroscopic, granular) catalysts leads to inaccuracies in dielectric constant calculations, steaming from non-standard synthesis procedural approach and therefore high dimensionality of inverse space. In Mueller et al.'s review of current machine learning approaches in materials, they establish how supervised learning techniques like the one used in this paper have achieved excellent structure-property predictions[35]. The most direct method of determining an observable-property relationship, in the case of this study for the complex dielectric properties, is from calculations based on scattering parameters (S-parameters). Multiple measurement techniques utilize S-parameters, such as a rectangular free-space waveguide, open-ended probe, free space, resonant cavity, parallel plate, and coaxial precision airline[26,36–38]; Zangwill 2013). These different methods utilize different inverse techniques such as Nicholson-Ross-Weir (NRW), NRW polynomial, The National Institute of Standards and Technology (NIST) Iterative, NIST non-iterative, and the short circuit line (SCL) methods[29,31,39]. All these techniques suffer from intrinsic errors such as the need to de-embed the void space within each testing apparatus as well as mathematical discontinuities that must be eliminated or ignored.

Artificial neural networks (ANNs) and convolutional neural networks (CNN) are a subset of machine learning that lend themselves well to material science problems. The usage of these networks has been steadily on the rise over the past decade, with more and more studies investigating the possibilities of ANNs to map non-linear relationships[40–46]. ANNs are part of the biologically inspired computational techniques used in different artificial intelligence applications[47–49]. ANNs have been used in many applications of chemistry, material science, and microwave engineering[19,21,35,50]. In Li et al.'s paper, they proved that machine learning models are more accurate than traditional linear and non-linear statistical regression methods when dealing with high dimensional inputs[51]. This advantage of machine learning algorithms only increases as the dimensionality and non-linearity of the relationships increases[52–54].

Machine learning models and ANNs, in particular, have recently been used in the literature to try and relate complex geometric parameters or certain material characteristics to dielectric constants[48,55–57]. However, Tuck and Coad[58] showed that ANNs can be used to calculate the dielectric properties of liquids directly from the S-parameters using a coaxial probe method without the need for de-embedding the data first. By calculating the dielectric properties directly from the recorded S-parameters without needing to de-embed the data in the time domain Tuck and Coad were able to achieve a significant reduction in the intrinsic error. This was because the ANN was able to capture the realities of the non-ideal system by training the ANN on vectors of reflected coefficient data and correlating that to the permittivities of the substance being studied. This method avoided the need for any parametric models of the cable such as those used by Stuchly et al. in their attempt to solve the inverse problem[59,60].

Chen et al. demonstrated that this same de-embedding approach was suited for different geometries[61]. More importantly, Chen et al. were able to show that finite different time-domain (FDTD) simulation data can be used to accurately train an ANN for prediction on experimental data. They even postulate that this method would work for granular materials and at high temperatures. These ANNs however, were limited in scope to only liquids due to both the computational restrictions of the time and the method chosen to solve the inverse problem. The probe method that Tuck et al and Chen et al utilized only looks at the non-linear and complex relationships between a single observable, reflected coefficient of the scattering parameters (S11) and the dielectric properties. In more advanced measurement techniques utilizing different methods such as the coaxial airline, in which two observables, S11 and S22 are utilized it is believed more accurate results are possible. Regardless the results of both studies were fast and extremely accurate calculations of the dielectric properties of a combination of different liquids. These studies have been followed up and expanded upon in different directions all to use machine learning to replace traditional mathematical models for the perdition of dielectric properties[6,14,62–64].

1.2.1 Microwaves

Microwaves are a form of electromagnetic radiation with a varying range of wavelengths from 3 m to 3 mm possessing frequencies of 1 GHz to 100 GHz. Microwaves unlike radio frequencies travel by line of sight rather than as ground waves or as reflections from the ionosphere[65]. Microwaves fall within the inferred and radio waves in the electromagnetic spectrum which is shown in subfigure 3A[66].

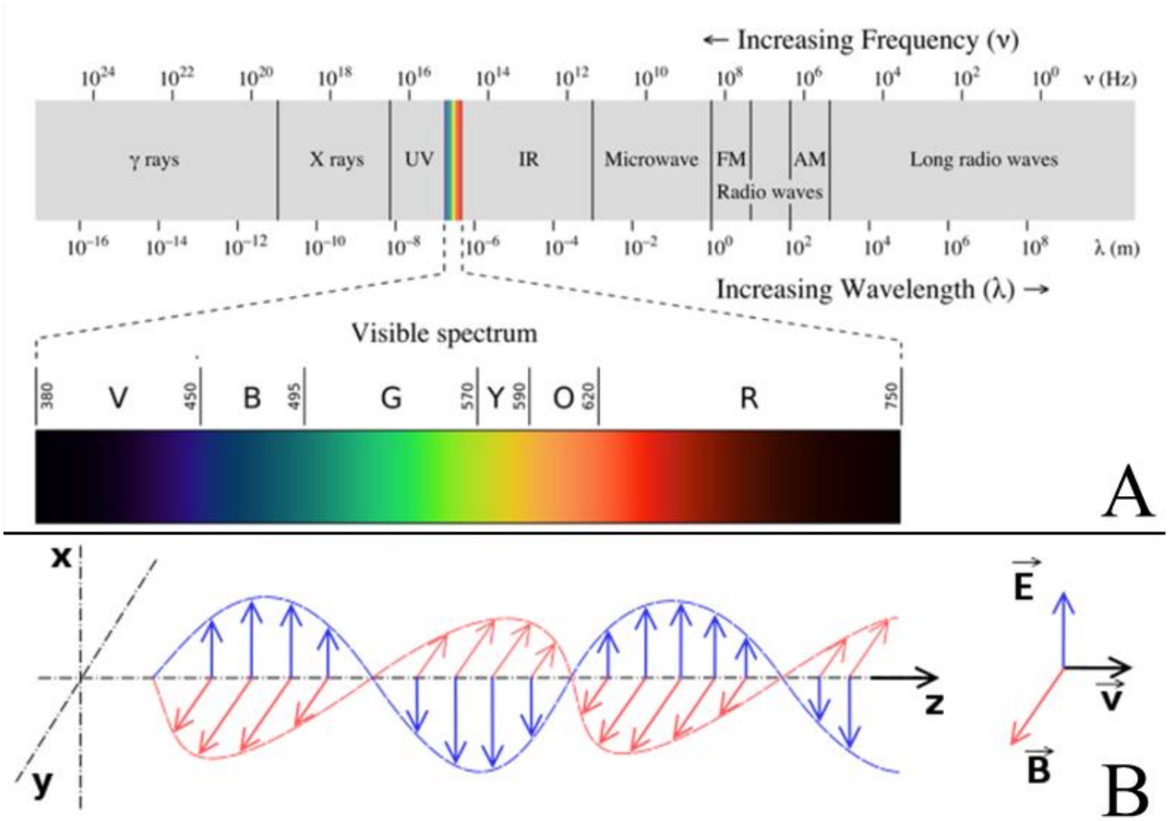

← Increasing Frequency (ν)
10^{24} 10^{22} 10^{20} 10^{18} 10^{16} 10^{14} 10^{12} 10^{10} 10^{8} 10^{6} 10^{4} 10^{2} 10^{0} ν (Hz)
γ rays
X rays
UV
IR
Microwave
FM
AM
Radio waves
Long radio waves
10^{-16} 10^{-14} 10^{-12} 10^{-10} 10^{-8} 10^{-6} 10^{-4} 10^{-2} 10^{0} 10^{2} 10^{4} 10^{6} 10^{8} λ (m)
Increasing Wavelength (λ) →
Visible spectrum
380
V
450
B
495
G
570
Y
590
O
620
R
750
A
x
y
z
$\vec{E}$
$\vec{v}$
$\vec{B}$
B

Figure 1: Subfigure A: Electromagnetic spectrum with a visualization of the visible spectrum shown as a subset of electromagnetic radiation[66]. Subfigure B**:** An electromagnetic wave propagating in the +z direction through a homogenous, isotropic, dissipationless medium. The wave is linearly polarized, where the electric field is shown in blue and the magnetic field is shown in red. The electric field oscillates in the ±x direction while the magnetic field oscillates in the ±y direction[66].

Microwaves are used in a wide variety of applications such as point-to-point communication links, wireless networks, microwave radio relay networks, radar, medical diathermy, cancer treatments, remote sensing, satellite communication, spacecraft communication, radio astronomy, spectroscopy, industrial heating, collision avoidance systems, particle accelerators, garage door openers and keyless entry systems, and for cooking food in microwave ovens[67].

Microwaves consist of electromagnetic waves made up of two components, an electric field, and a magnetic field. Microwaves are synchronized oscillations of the electric and magnetic fields both of which propagate at the speed of light. These two waves are commonly perpendicular to one another and the direction of the energy, with this perpendicular wave propagation forming a transverse wave, this is shown in subfigure 1B[66].

Microwaves possess several unique characteristics and advantages over what can be seen at other wavelengths. These electromagnetic waves drastically reduce the time of heat conduction in a sample by directly heating the material[68]. These quick heating rates have been used in a variety of applications used in everyday life. Especially important for this study is the use of microwaves in the fields of organic chemistry, catalytic chemistry, inorganic material chemistry, and analytical chemistry. Microwaves are especially useful in these fields over conventional heating as they drive regioselectivity, regular radical reaction, molecule orientation, high crystallization, anisotropic crystal, specific solid-phase diffusion, and strong reducing reaction [68]. The formation of temperature gradients at the microscale when using microwaves for catalyzed reactions if controlled would allow for huge advances in the field of catalytic chemistry. Microwaves are a type of non-ionizing radiation meaning that they do not contain enough energy to ionize or change substances.

1.2.2 Microwave Material Interactions

High-frequency electromagnetic waves affect solid materials in a multitude of ways depending on their different material classifications. Most materials used in microwave applications are designed to either pass a conduction current or prevent its flow as completely as possible. Conductors reflect microwaves from their surface without being effectively heated by the microwave. The electric field generated moves electrons freely from the surface of the material thereby heating the material via the resistivity of the heating material [69,70]. The conductive material can be regarded as a nonconducting dielectric with resistance in parallel.

Dielectric materials are characterized as materials that have changeable dipole interactions which result in heat generation[70]. The passage of microwave radiation through the medium generates absorption and heat generation throughout.[16,69]. Electromagnetic waves can be applied to heat dielectric materials by applying the electric field to induce polarization of the charges within the material being heated. The polarization cannot match the rapid reversals of the electric field and thus induces the heating of the irradiated media. It can also result in dipolar moments, which are merely localized reorganization of polar molecules. The magnetic component of electromagnetic waves introduces magnetic moments into the material. This local reorganization of bonded and free charges is what is commonly known as the polarization phenomenon. The polarization phenomena have two main points, the storage of electromagnetic energy within the irradiated medium and the conversation of thermal energy in relation to the frequency of the electromagnetic stimulation[26,71,69].

The reorganization of linked and free charges is the physical origin of polarization phenomena which is clearly explained using quantum theory. The interaction between an electric and or magnetic field and a dipole can be explained using quantum theory. Weak coupling between dipole and electric field leads to no quantified orientations existing. Dipoles are typically associated with chemical bonds, and the movement of the dipole induces a correlative motion in the molecular bonds. The motion of the magnetic moment is independent of this molecular motion[2669].

The physical orientation of polarization can be expressed by the quantity $\vec{\mathrm{P}}$ which gives the contribution of matter with regard to that of a vacuum. The electric field and the polarization are linked with Maxwell's equations. The displacement and the electric field can be expressed as $\vec{\mathrm{D}}$

and $\vec{\mathrm{E}}$ respectively, their relationship can be seen in Equation 1. The dielectric permittivity is the ratio of the electric displacement to the electric field[26,69].

$$\vec{D} = \overrightarrow{\bar{\varepsilon}\,E} \qquad \text{Eq. 1}$$

The contribution of matter to polarization can be given as $\vec{\mathrm{P}}$ and the dielectric medium can be characterized by $\bar{\varepsilon}$. $\vec{\mathrm{P}}$ describes a polarization process relating to the response of dipoles and charges applied to the field. The relationships between these different fields are expressed in Equation 2[72].

$$\vec{D} = \overrightarrow{\bar{\varepsilon}\,E} = \varepsilon_0\vec{E} + \vec{P} \qquad \text{Eq. 2}$$

Insulating materials allow microwaves to penetrate the material without absorption, losses, or heat generation. The electronic reorientation or distortions of the induced and/or permanent dipole can result in heat generation within the material[69,70,73]. The material can polarize within the electric field generated by the microwave.

1.2.3 Permittivity

The capacitance encountered in the formation of an electric field of a medium is denoted as the absolute permittivity. This can be expressed as the amount of charge needed to generate one unit of electric flux in the medium being studied. Permittivity in essence is a material's ability to store an electric field in the polarization of the medium[26]. A material's dielectric medium usually is expressed as the relative permittivity of the material, this term is commonly called the dielectric constant in literature[39,74]. It can be expressed as kappa κ which is the ratio of the absolute permittivity to the electric constant. The dielectric constant is not typically constant, it varies with position in the medium, the frequency of the field applied, humidity, temperature, and other parameters. In a nonlinear medium, the dielectric constant can vary with the strength of the applied electric field as shown in Equation 3[26].

$$\kappa = \varepsilon_r = \frac{\varepsilon}{\varepsilon_0} \qquad \text{Eq. 3}$$

The dielectric constant is directly proportional to the electric susceptibility χ, which is a measurement of how easily a dielectric polarizes in response to an electric field. The relation of these terms is given in Equations 4 and 5.

$$\chi = \kappa - 1 \qquad \text{Eq. 4}$$

$$\varepsilon = \varepsilon_r\varepsilon_0 = (1+\chi)\,\varepsilon_0 \qquad \text{Eq. 5}$$

The two main points of wave-matter interactions can be expressed by the two components of the dielectric constant as seen in Equation 6.

$$\varepsilon = \varepsilon' - j\varepsilon'' = \varepsilon_0\varepsilon_r' - j\varepsilon_0\varepsilon_r'' \qquad \text{Eq. 6}$$

Where ε', ε'', ε_r' , and ε_r'' are the real and imaginary parts of the complex dielectric permittivity and the real and imaginary parts of the relative complex dielectric permittivity. The ability of a material to store electromagnetic energy is expressed as the real part and the thermal conversion potential is proportional to the imaginary part[75].

The electric loss of a material can be expressed in terms of the real and imaginary parts of the relative complex dielectric permittivity. This term is commonly referred to in the literature as the loss tangent of a material. The most important aspect of the loss tangent is that it is proportional to a material's ability to absorb heat[26,75,76]. The loss tangent is illustrated below in Equation 7.

$$\tan \delta_e = \frac{\varepsilon''}{\varepsilon'} \qquad \text{Eq. 7}$$

1.2.4 Measurement Models for Material Properties

The characterization of microwave material interactions can be defined as a mathematical model, where the waves that are reflected and transmitted through the material at a certain frequency are measured. In a transmission/reflection (TR) measurement, a material is placed into a waveguide or coaxial line and subjected to microwave with a known frequency[38]. The material reflects part of that wave while allowing for some of it to pass through. The study of this effect reveals the specimen's dielectric properties. The reflection and transmission data are known as scattering data. The scattering data must be solved using the electromagnetic boundary-value problem to determine the material's properties[38].

1.2.5 Scattering Parameters

Scattering parameters (S-parameters) are a type of small-signal AC commonly used to characterize RF components. S-parameters establish small-signal characteristics of a device at a specific bias and temperature[77]. They are measured by making the measuring device impeded between a 50-ohm load and a source, drastically reducing the chance of oscillations to occur. S-parameters have the distinct advantage of not varying in magnitude at points along a lossless transmission line because they are traveling waves, not terminal voltages[77]. A signal wave for a two-port electrical element is represented in figure 2, where a_1 is the wave into port 1, a_2 is the wave into port 2, b_1 is the wave out of port 1, and b_2 is the wave out of port 2.

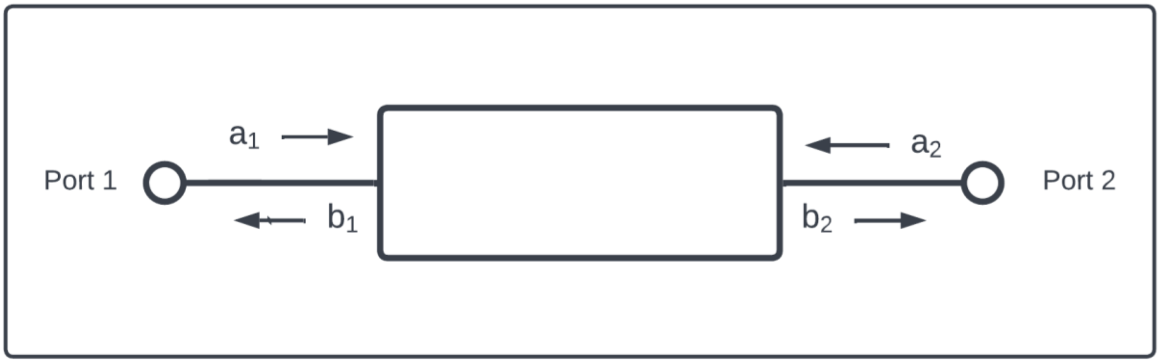


Figure 2: Single wave in a two-port electrical-element. Simple representation of a standard 2 port measurement for S-parameters.

A conventional element S-parameters for microwaves can be defined as in Equations 8 and 9[71]. Where S_{11} is the port-1 reflection coefficient, S_{22} is the port-2 reflection coefficient, S_{21} is the forward transmission coefficient, and S_{12} is the reverse transmission coefficient. For a 50-ohm system with the two-port setup, each port is terminated at 50 ohms and the S_{21} parameter represents the voltage gain of the element from port 1 to port 2[78]. S-parameters are commonly displayed as the magnitude plus phase of the wave being measured or as a real plus imaginary number converted from the magnitude and phase. Where Equations 8 and 9 show the general scattering parameter equations.

$$b_1 = a_1 s_{11} + a_2 a_{12} \qquad \text{Eq. 8}$$

$$b_2 = a_1 s_{21} + a_2 a_{22} \qquad \text{Eq. 9}$$

1.2.6 Instrumentation

There are several methods used in literature to measure the above-mentioned material properties each with different strengths and weaknesses. No one technique can characterize every material for every frequency leading to a need for a plethora of techniques based on several key factors. The selection of measuring techniques depends on several significant factors such as frequency, accuracy, temperature, material nature, sample size and or thickness, containing or non-contacting, destructive or non-destructive, and cost. Most dielectric property measurement techniques can be broken into two categories; resonant and non-resonant[79,80]. Resonant methods characterize materials at discreet frequency points where the dielectric material is used as a resonant element[79]. The drawbacks to this method are that the sample must be a low-loss material. The resonant method technique also known as the perturbation method requires samples to be placed into a resonant cavity[79]. The resonant cavity causes perturbations that result in resonant frequency shifts. This form of the resonant method is used for low to moderate-loss samples[79,80].

The non-resonant method is used to measure frequencies over a broad range. This technique is the more prevalent one in literature as it utilized the transmission and reflection coefficients discussed in Section 2.3.2[79]. Some of the most common non-resonant techniques in literature are the coaxial airline method, waveguide method, free space method, and coaxial probe method[79,80].

1.2.7 Coaxial Line

The coaxial airline is part of the transmission line method, utilizing the measurement of a reflected signal and the transmitted signal. A coaxial transmission line is a cylindrical test cell with a center conductor running concentric, cut to the exact length as the test cell, a schematic is shown in subfigure 3A. The coaxial line is characterized by the material filling the entire cross-section with no air gaps existing at the walls[79]. The coaxial line technique can measure magnetic materials, it is however limited at its lower frequencies based on the sample lengths. The coaxial airline technique can cover a broad frequency range and is best for lossy machinable solids[77].

The coaxial airline method is one of the transmission line methods commonly used in literature and has its associated advantages and disadvantages. These advantages include the ability to cover a wide frequency range from 50 MHz to more than 100 GHz as well as to measure anisotropic materials. Coaxial airlines are used to measure solid and powdery materials with low to medium loss[79,31]. It is one of the most accurate measuring techniques available for the testing of material properties. This is because the bandwidth of the coaxial airline is smaller than that of some of the other techniques discussed in this study. The presence of a center conductor in the coaxial airline method avoids the creation of any higher-order modes that would cause increased error[79,31].

The disadvantages of the transmission/reflection line methods include the air-gap effects and the difficulties of discontinuities associated with samples that are multiples of one-half wavelengths[79,35]. The presence of the center conductor makes the creation of testing samples slightly more difficult than other methods as well as reduces the amount of power that the testing cell can handle[79,31].

1.2.8 Network Analyzers

Network analyzers are the preferred method for the collection of data on the electromagnetic wave and material interactions. Network analyzers work by measuring the scattering parameters to characterize a material. Vector network analyzers (VNA) measure both amplitude and phase, allowing for more detailed information to be gathered about the material being measured, an example is shown in subfigure 3B. Network analyzers are subject to various sources of error such as nonlinearity of mixers, gain and phase drifts in amplifiers, noise introduced by the analog to digital converter, imperfect tracking in dual channel systems, imperfect matching at connectors, and imperfect calibration standards[38,81].

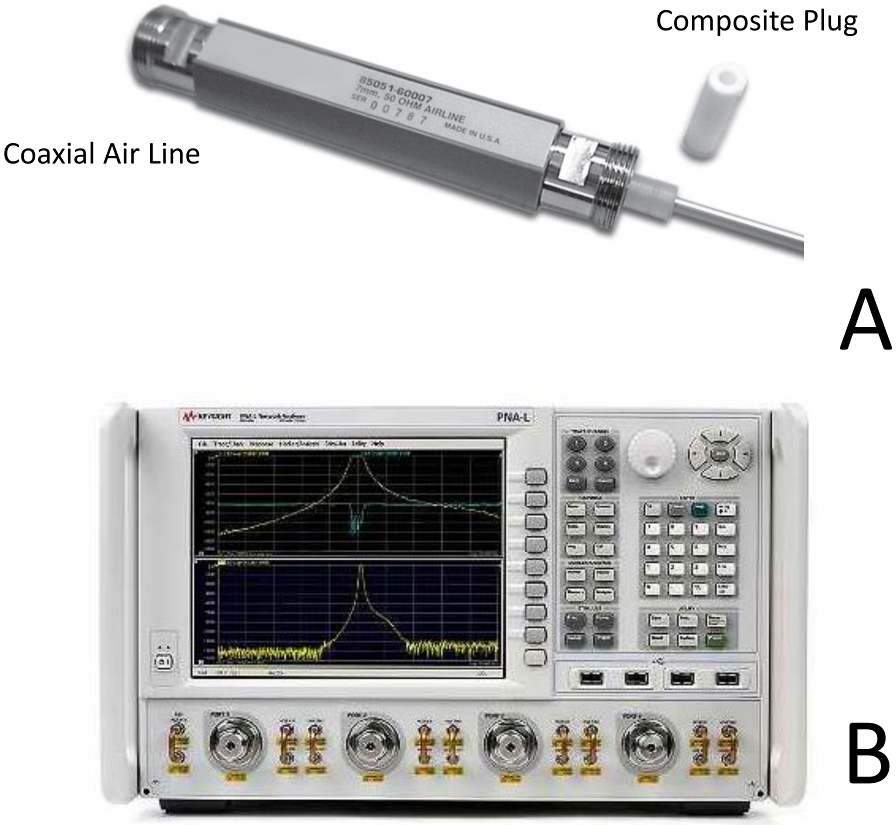


Figure 3: Subfigure A Keysight Network high-precision coaxial airline[81]. Used to measure the scattering parameters and calculate the associated material properties. An example of a testable material is shown as composite[69]. Subfigure B: Keysight Network Analyzer is part of the vector network analyzer family of machinery. Used primarily to measure the scattering parameters and calculate the associated material properties.

1.2.9 Random Uncertainties and Error

In transmission/reflection measurement techniques there are several different types of error, one of them being random uncertainties of the calibration and from the specimen itself. The three main types of random uncertainties and error sources typically related to transmission/reflection measurements are errors in measuring the magnitude and phase of the scattering parameters, errors in specimen length, and errors in reference plane positions[38]. To counteract this problem a differential uncertainty analysis can be applied to both S_{11} and S_{21} separately. For both S_{11} and $S_{21,}$ the dominant uncertainty is the phase, with longer specimens having less uncertainty. It has been found in the literature that at higher frequencies S-parameters have larger uncertainties in phase[38].

1.2.10 Systematic Uncertainties

The other type of error associated with transmission/reflection measurements is systematic uncertainties. These uncertainties can be broken down into several main types, gaps between the specimen and specimen holder and specimen holder dimensional variations and line losses, and connector mismatch[38]. There are standard equations in the literature that are made to handle the first type of uncertainties for gaps around the specimen[27,28]. Along with airgaps, other systematic uncertainties include short-circuit and waveguide wall imperfections and losses[38]. Waveguide losses can be corrected by taking a measurement of an empty waveguide and calculating the appropriate correction factor or attenuation coefficient. For air gaps, additional measurements using a resonator of the same material in the frequency band being measured will determine the required gap in the correction formula.

1.2.11 Corrections to Data

With the many possible errors associated with dielectric measurement testing and the difficulty of data collection, corrections must be made once a measurement has been obtained. The corrections must account for the systematic uncertainties and if possible the random uncertainties and errors. Known uncertainties associated with transmission/reflection measurements are air gaps around the sample, wall imperfections, and losses. Airgap corrections are most important when considering the coaxial method of testing with particular emphasis on the center conductor[38]. For both coaxial and waveguide methods, airgap correction is particularly important in whichever region has the strongest electromagnetic field [27,28]. Both waveguides and coaxial lines at ambient temperature will experience some power loss because they are not perfectly conducting. The different propagation modes will be attenuated to some degree because of this power loss [38].

1.2.12 Nicolson-Ross-Weir

The most widely used method for calculating permittivity and permeability from S-parameters is the Nicolson, Ross, Weir method (NRW)[29,30]. This method applies over the range of 100 MHz to 18 GHz when using a computer-controlled network analyzer. However, it is subject to singularities when the specimen length is a multiple of one-half wavelength in the material[38]. This is especially prevalent for low-loss materials as it is impossible to measure the phase of S_{11} accurately. NRW takes measurements in the frequency domain rather than the time domain[75,76] which avoids the need of using a Fourier transform to calculate permittivity and permeability. The NRW method works for both waveguides and coaxial lines using discrete frequencies in less than 20 kHz steps[30].

1.2.13 Finite Difference Frequency Domain

There are several advantages of the finite element frequency domain (FEFD) method over a finite difference time domain (FDTD) method, the advantage is two-fold. First, the FE approach is an implicit technique that relies on an energy minimization method while the FD involves an explicit stability criterion dependent on the mesh characteristics[82,83]. Second, by solving in the frequency domain the computational time is significantly reduced by eliminating time-stepping criteria. While these advantages do not apply to all problems, especially large (time and spatial) non-linear problems, the frequency domain was appropriate for the following 2D axis-symmetric linear (steady-state, non-temperature dependent properties) study[74].

Finite difference frequency domain is typically describing all frequency-domain finite methods. Where the methodology applied is the numerical solution of a finite-difference approximation of the derivative operations of the targeted differential being solved. This methodology works by applying a partial differential equation such as Maxwell's equations. There are at least two categories of recognized frequency-domain problems in electromagnetism[26,68,84].

One category is to find the response to a current density J with a constant frequency ω, or a similar time-harmonic source which can be described in Equation 10. This frequency-domain response problem lends itself to a simple matrix system of linear equations described by Equation 11**.** The second category of problems is to find the normal modes of a structure in the absence of sources. This can be applied to waveguides, coaxial lines, and many other common electromagnetic structures[83,85,86]. To solve this problem the frequency is treated as a variable and the eigenvalues and vectors are calculated. This can be described by the mathematical equation seen in Equation 12. Where lambda is the eigenvalues of the matrices.

$$J(x) = e^{i\omega t} \qquad \text{Eq. 10}$$

$$Ax = b \qquad \text{Eq. 11}$$

$$Ax = \lambda b \qquad \text{Eq. 12}$$

1.2.14 Multilayer Perceptron

Multilayer perceptron (MLP) was one of the first conceived neural networks in the field of artificial intelligence. MLPs are a class of feedforward artificial neural networks (ANN) that comprise layers of perceptions. In the field of machine learning, perceptrons are any neuron that can employ an activation function that allows it to perform a regression or classification task. Neurons take in a weighted input and produce an output signal based on these activation functions. MLP consists of at least 3 layers, an input layer, a hidden layer, and an output layer. Hidden layers are where neural networks get their strength. The hidden layers are not directly exposed to the system input, and they can be chained together to create deep networks. The only limit on the number of neurons and layers possible in a network is the computational power of the machine being used for the network.

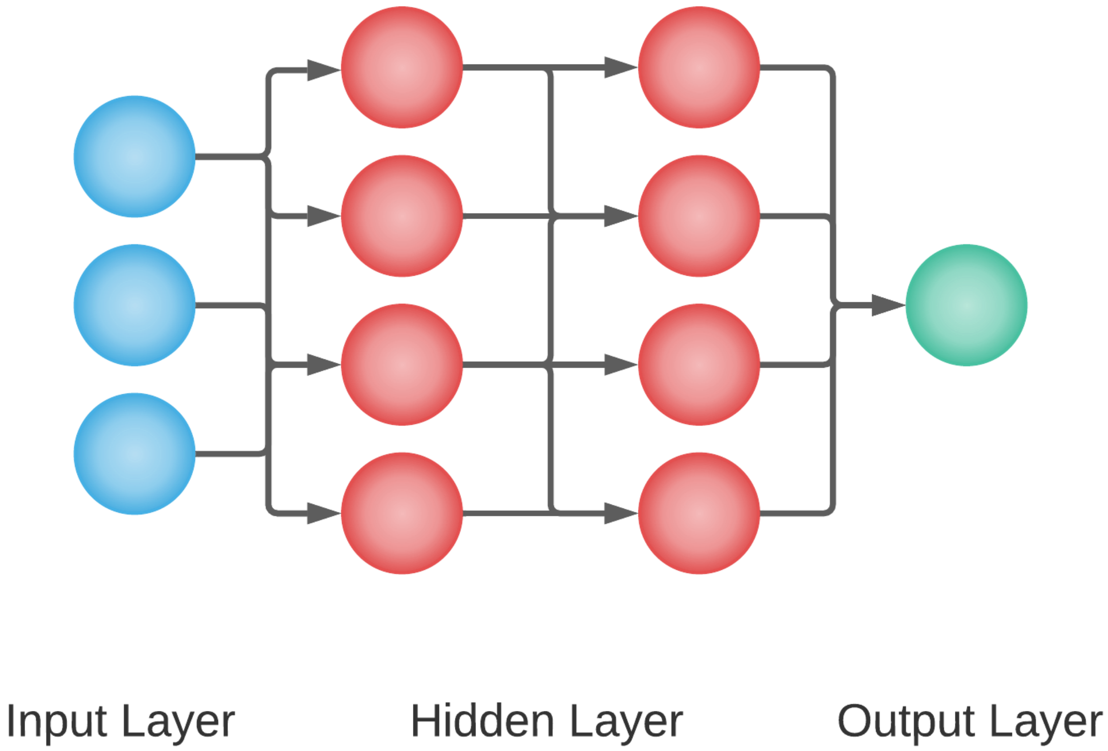


Figure 4: General example of a multilayer perceptron neural network. Where the blue circles represent input nodes, the red circles represent the hidden layer nodes, and the green circle is the output node.

Once a network has been created the weights of the neurons need to be adjusted and fitted to allow for the network to produce usable results. Typically, data is split into three categories, training data, validation data, and test data. These 3 subsets should be randomly distributed from the same source of data. Training data is usually the largest of the 3 subsets followed by validation data and then testing data. ANN work by taking training data in from the input layer and adjusting the weights of the neurons to try and predict the correct output. After each neuron's weight correction, the ANN is then scored using the validation data. Where the input data is fed through the network without any correction of weights and the result is scored against the ground truth. After some threshold has been reached, either a metric achieved or a number of iterations the network is considered trained. Typically, the test set is then utilized to score how the network would do in a real-life example. The test inputs are fed into the network without the correct outputs. The ANN predictions are then compared against the actual results, this is the last step to validating an ANN’s performance and is done by a human.

1.2.15 Deep Leaning

Deep learning is a machine learning method that stems from ANNs and utilizes representation learning. Deep learning networks can be deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks. The term “deep” comes from the usage of multiple hidden layers within a network. The strength of these networks is that they can model very complex non-linear relationships. This is accomplished by each subsequent layer of the network attempting to exact features from lower-level representations of the input data. Theoretically, this allows for complex data to be represented using fewer input parameters than for a shallower network. Like ANNs, Deep networks are typically feedforward networks with data only going in one direction.

1.2.16 Convolutional Neural Network

Convolutional neural networks (CNN) are a class of deep learning networks that is a regularized version of an MLP. CNN uses a shared weight architecture of convolution filters to create a feature map of the input data. They achieve this by using a sliding window technique that looks at a smaller subsection of the input data at a given time to provide a translationally equivariant response. CNN uses this feature map to create a hierarchical pattern of the data. By using the smaller and simpler subsections of the input data and assembling patterns of increasing complexity CNNs can achieve create generalizations. Without the computation cost typically associated with MLPs.

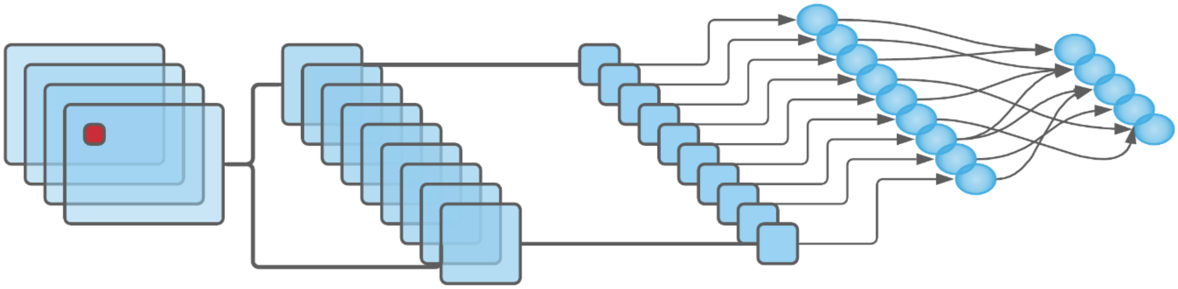


Figure 5: General example of a convolutional neural network. Where the first several layers depicted by the squares are the convolutional layers. The last two layers depicted by the circles are fully connected layers. Similar in nature to the MLP described in Section 1.2.14

1.3 Study 1 Methodology

1.3.1 Material Dataset Generation

The calculations from S-parameters for varying dielectric properties were determined using a finite element (FE) EM wave modeling software COMSOL Multiphysics® [87]. All solutions were solved in the frequency domain, using the FEFD approach. Using the FEFD model a series of parametric sweeps of both the real and imaginary portions of the dielectric properties was performed to encompass the most naturally occurring dielectric materials. Properties were swept from a real dielectric constant of 1 to 100 in increments of 0.5 while the imaginary portion of the dielectric properties was varied from 0 to 2 in increments of 0.05. These parametric sweeps were done in conjunction with a gradually increasing sample length. The length of the sample was increased from an initial 10 mm to 50 mm in 1 mm increments. These dielectric properties were evaluated in correspondence with a frequency range of 0.1 to 13.5 GHz at 51 equally spaced points.

The computational model was set up to represent a real-world two-port vector network analyzer utilizing a high-precision coaxial airline of length 150 mm. The airline is modeled based on the experimental airline used for validation. The coaxial airline is an HP model no. 85051-60010 with a 0.70 cm diameter[81].

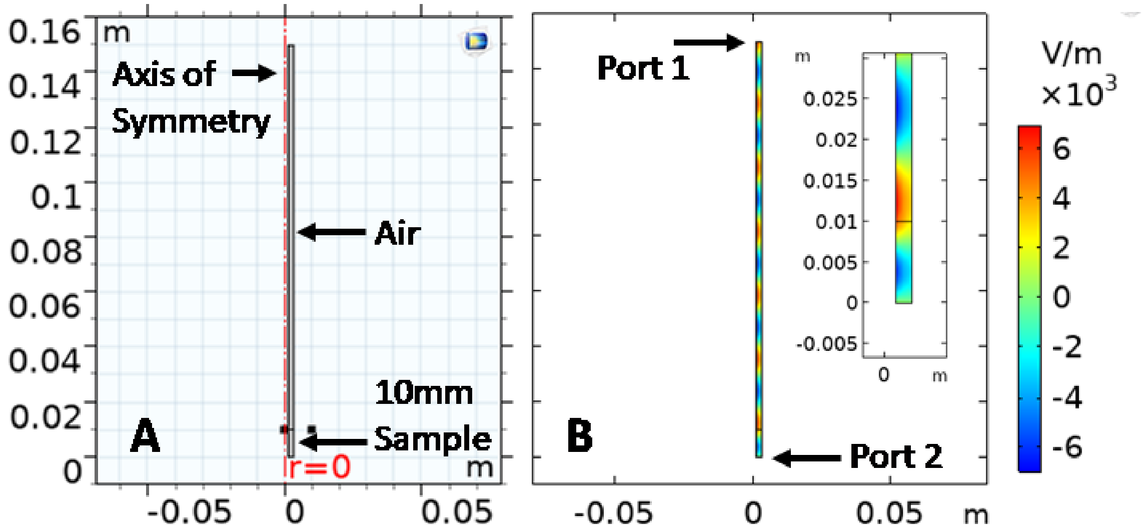


Figure 6: Illustration of the axis-symmetric coaxial airline model with a 10mm dielectric sample embedded within the airline. Subfigure A is an illustration of the geometry of the 150mm coaxial airline. Subfigure B is a contour plot of the radial electric field (Er) at 13.0 GHz@1W with a 3:2 CeO2:Parafin 10mm plug. Contours visually confirm a TEM standing wave within a coaxial transmission line.

A 2D axis-symmetric FE model was constructed, which represented the 150mm coaxial airline that is used in the experimental measurement setup. Figure 6 is an illustration of the coaxial airline modeled with the FE solver software. The walls of the airline, as well as the center electrode, were assumed to be perfect electrical conductors. Due to the axis-symmetric assumption, it is assumed only a 2D representation of a slice in the +r and +z directions needed to be constructed. The plane was partitioned at an initial z=10 mm to form material regions increasing after each full parametric sweep. The region between z=0 mm and z=10-50 mm will be defined as the sample region. The remaining will be assigned air (vacuum, $\varepsilon_r=\mu_r=1$). Two ports were defined at extremes in the z-dir. Port 1 was defined at z=150 mm and Port 2 was defined at z=0 mm. A coaxial boundary condition (TEM mode) was specified for both ports. The scattering parameter was measured at both port planes without the de-embedding of the void air space, as would be representative of experimental measurement where de-embedding has taken place during the calculation of the dielectric properties from the S-parameters observed at the ports. Subfigure 6B is a contour plot of the radial electric field (Er) at 13 GHz and 1W of input power at Port 1. For the remainder of the study, 0.1W will be used as the input power with the understanding that 1) material properties are linear and are not changed by field strength or temperature, 2) scattering parameters are a function of normalized power, and 3) the experimental network analyzer will utilize much lower port power. The color contours of subfigure 6B and the inset image confirm the radial electric field is synonymous with a TEM mode.

1.3.2 Artificial Neural Network Implementation

The ANN was developed using open-source TensorFlow developed by Google in Python for ease of implementation with all data scaled to be within the same power factor. The ANN used two different approaches one in which all 5 input features of, frequency, the magnitude of S11 and S21, and the phase-in radians of S11 and S21 were studied independently of one another. This first approach would allow researchers to get discrete answers at any frequency point independent of the solutions to previous frequencies. The other approach was to look at all the inputs for a given dielectric at once, in this case, all 51 data points from 0.1 to 13.5 GHz. This second approach attempts to pull the latest information that exists in the transition between wavelengths to achieve a better characterization across the whole frequency range. The data was broken down into 3 different sets 60% was allocated to training data, 20% to validation data, and 20% to test data. The experimental data was kept separate until a suitable algorithm had been created. This breakdown allows the algorithm to be tested on unseen data ensuring that it was not overfitted to the training and validation data before it was tested on experimental data. To achieve the ideal performance of this network multiple different loss functions were looked at as well as different combinations of the number of neurons and number of hidden layers. Different regularizes were investigated to help encourage convergence. These different combinations were evaluated using the mean squared error (MSE) and the mean absolute error (MAE).

To ensure the optimization of these parameters the ANN utilized the ReLU activation function and the Adadelta optimizer. The network also introduced Gaussian noise into the training data to represent real-world errors in experimental setups. The ReLU activation function was chosen because of its proven ability to represent sparsity[51,88]. Sparsity is useful in ANN because of its ability to imitate a biological neural network. Sparsity in an ANN allows for models to have better predictive power with less noise and overfitting by encouraging neurons to only process meaningful aspects of the problem[53,88]. In the work by Maas et al. and Narang et al., they demonstrated that increased sparsity helped to improve an ANNs training performance and reduce computational time for several problems[89,90].

The Adadelta optimizer is a gradient descent method that utilizes a dynamic updating system. The system adapts using first-order information and stochastic gradient descent which reduces its computational cost over many of the other optimizers available[91]. One of the key advantages of this optimizer for the ANN system of interest is that it requires no human training in the learning rate and can handle training data that may have lower signal-to-noise ratios. These two hyperparameters were chosen after some initial data testing and held constant for the remainder of the study. The first scheme when considering frequency points independently employed, two convolutional layers, and two fully connected layers. The second scheme in which an array of frequency points is used employed, two convolutional layers, one max-pooling layer, and two fully connected layers.

1.3.3 Experimental Data Collection Method

Data was collected on a Teflon (PTFE) plug, replicating the standard NRW test that validated their measurement model. To take a measurement the sample was loaded into the airline with the center electrode in place, as shown in subfigure 3A. All interfaces between the airline and cables were thoroughly cleaned using isopropyl alcohol and dried using dry compressed air. Each test was conducted with a frequency range from 0.1 to 13.5 GHz. The scattering parameters were recorded at 51 equally spaced points within this range. The relative dielectric constant for each point was calculated using the NRW method. All measurements reported in the study were conducted using a 0.70 cm diameter coaxial airline (HP model no. 85051-60010), as shown in subfigure 3A, and connected to a Keysight N5231A PNA-L microwave network analyzer shown in subfigure 3B.

1.4 Study 1 Results

1.4.1 ANN Result for Dielectric Predictions

The simulation-derived dielectric datasets consisted of 330,813 values with real portions of the dielectric ranging from 1 to 100 and the imaginary portion ranging from 0 to 0.2. The corresponding inputs of S-parameters include the magnitude and phase of S11 and S21. Because the system is symmetric S11=S22 and S21=S12 and therefore only S11 and S21 are necessary. To create an efficient machine learning model a statistical analysis of the input features needed to be performed to determine their significance on output targets. Strong correlations can be both good and bad for ANNs, strong correlations can help to reduce the number of input features needed for the network. They can also skew the network towards harmful bias creating multicollinearity with a single input and the target feature. Which can result in small changes to the input data leading to large changes in the model. To check on these traits a Pearson correlation was performed between all the input features and the output targets[92,93]. Table 1 is a summary of the correlation between the different inputs and outputs. With a 1.0 meaning a very strong positive correlation and a -1.0 corresponding to an inversely related parameter. The complex dielectric is defined as ε= ε'+i ε", where ε' is the real portion and ε" is the imaginary portion. The scattering parameter magnitude is denoted as |S11| and |S21|. The associated phase angle of the scattering parameter is denoted as ∠S11 and ∠S21

	|S11|	**|S21|**	**∠S11**	**∠S21**	**ε'**	**ε"**
|S11| (Input)	1.0	---	---	---	---	---
|S21| (Input)	-0.9	1.0	---	---	---	---
∠S11 (Input)	0.0	0.0	1.0	---	---	---
∠S21 (Input)	0.0	0.0	0.0	1.0	---	---
ε' (Output)	0.8	-0.9	0.0	0.0	1.0	---
ε" (Output)	0.0	-0.2	0.0	0.0	0.0	1.0

Table 1: Correlation matrix for input and output parameters. Values range from -1 to 1. Negative values are associated with inverse correlation.

As seen in Table 1 the magnitudes of S11 and S21 are strongly correlated to the real part of the dielectric constant. The features that correlate linearly to a dielectric constant are the magnitude of the wave that is reflected from a material and the magnitude of the same wave that passes through the material. This correlation of magnitudes is expected since the real part of the dielectric constant is the ability of a material to store energy. It is noted from Table 1 that there is no linear correlation between the phase angle and the magnitude of scattering parameters. There is also a lack of correlation between the phase angle and the dielectric constant. While this is correct for the assumed linear materials and constant sample geometry (plug length) some caution must be taken with this correlation. If the phase angles of the scattering parameters were eliminated it would make this system highly linearly correlated to the magnitude of the scattering parameters. This would result in more system performance as small changes in the magnitude of the scattering parameters would have large effects on the output of the system (dielectric properties). CNNs are uniquely suited to this type of problem because not only do they perform their calculations in high dimensionality, but they use convolutional math applied over the input data. Therefore, the inclusion of the phase angle allows the network to eliminate its dependence on the magnitude of the scattering parameters as seen in other multi-layer perceptron networks. To achieve a more extensive understanding of the different relationships between input features and output targets joint plots were created for each input. These are shown in subfigure 7A-D where the entire spectrum of dielectric properties as a function of the inputs. The darker regions of the contour plot represent a stronger correlation. From these plots, a better understanding of the correlation coefficient from Table 1 can be gained. The strong positive and negative correlation for the magnitudes of S11 and S21 can be seen in subfigures 7A and 7B. However, the figure illustrates that there is a direct effect on correlation based on the magnitude of the dielectric properties (|S11| and |S21|). Lower dielectric constants (ε' < 40) show an increasingly strong correlation between the inputs and the output as the dielectric constant approaches 100. This growing correlation will provide a unique challenge to the design of the ANN architecture as traditional approaches to strong and weak correlation architecture will be insufficient to capture the unique relationship.

The trained neural network was used to predict randomly generated test data that the ANN was not explicitly trained or validated on. Multiple models with a varying number of convolutional layers, hidden layers, neurons, and loss metrics were evaluated for their applicability in the calculation of complex dielectric properties. Each test was run for 500 epochs to allow for convergence to an optimized set of weights and used the Relu activation function along with the adadelta optimizer. The training set used in the network had a mean dielectric constant of 50.4 and a standard deviation of 28.6 while the test set had a mean of 50.8 and a standard deviation of 28.7. This similarity confirms that the test datasets contain a good representation of the whole dataset. Demonstrating that the data was well randomized and ANN performance was not due to the selection of a certain sub-dataset.

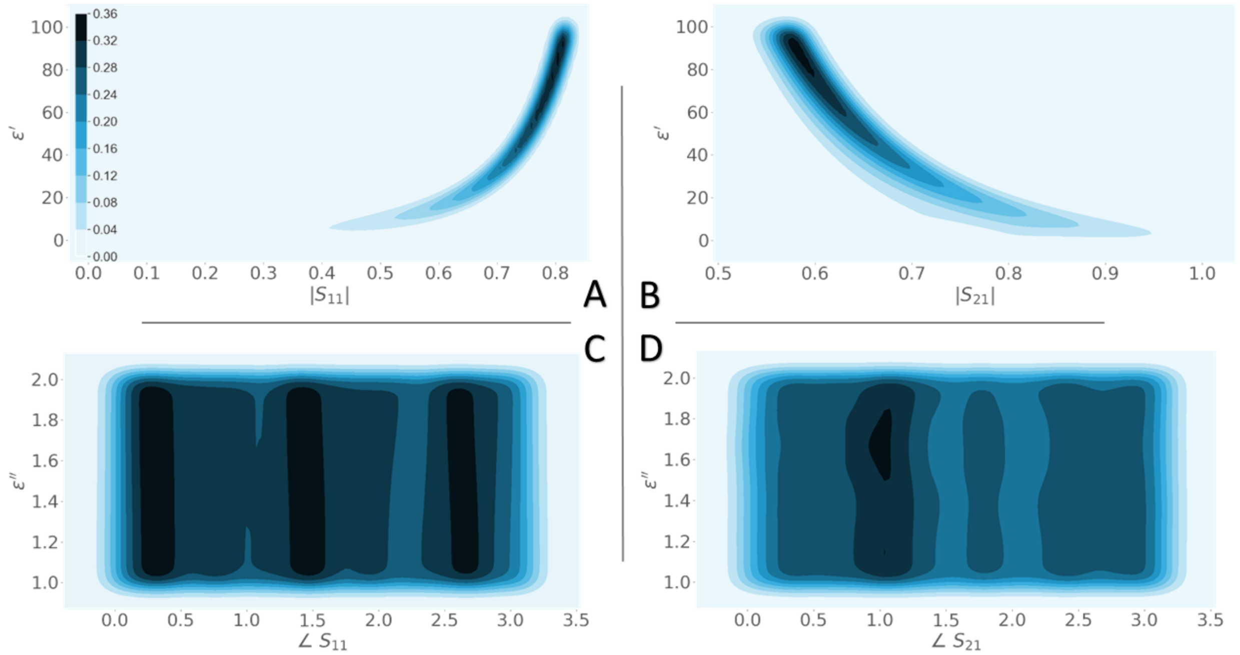


Figure 7: Collection of correlation density maps, these plots provide a qualitative understanding of the correlation between input features and output targets. Subfigure A illustrates a strong positive sloping correlation. Subfigure B illustrates an inverse correlation. Subfigures C and D illustrate a weak correlation between phase angle and imaginary dielectric properties.

A comparison between subfigure 10A and subfigures 9A and B illustrates this, due to the adverse relationship between how strongly input and output features are correlated, the predictive accuracy of the ANN starts to degrade because of an increase in multicollinearity. As the dielectric constant approaches, 100, and the correlation between the magnitudes of S11 and S21 shows a much stronger relationship the predictive accuracy of the ANN goes down due to the strong dependence on a single input parameter. This multicollinearity is unfavorable at high dielectric constants where variables can be linearly predicted from the others with a high degree of accuracy resulting in erratic responses to small changes. The strong correlation skews the values of predictions with small changes in the weight resulting in larger responses in the output neurons. At smaller dielectric constants the predictive accuracy of the ANN is much greater showing little scattering from the regression line.

The second approach results where the entire frequency spectrum is used as an input to the CNN are shown in subfigure 8B. As with the results shown in subfigure 8A, the network can accurately predict the dielectric constant for all values looked at in this study. However, a comparison of subfigures 8A and 8B confirms that this second approach has a much smaller spread of predictions, especially at high dielectrics. Statistically, the results between these two approaches are very similar to this approach having an MSE of 0.430 and an MAE of 0.511 for the real portion of the dielectric. While the imaginary portion had an MSE of 0.002 and an MAE of 0.035. It can be seen that the second approach while limited to frequency-independent dielectrics results in a higher accuracy across a wider range of dielectrics.

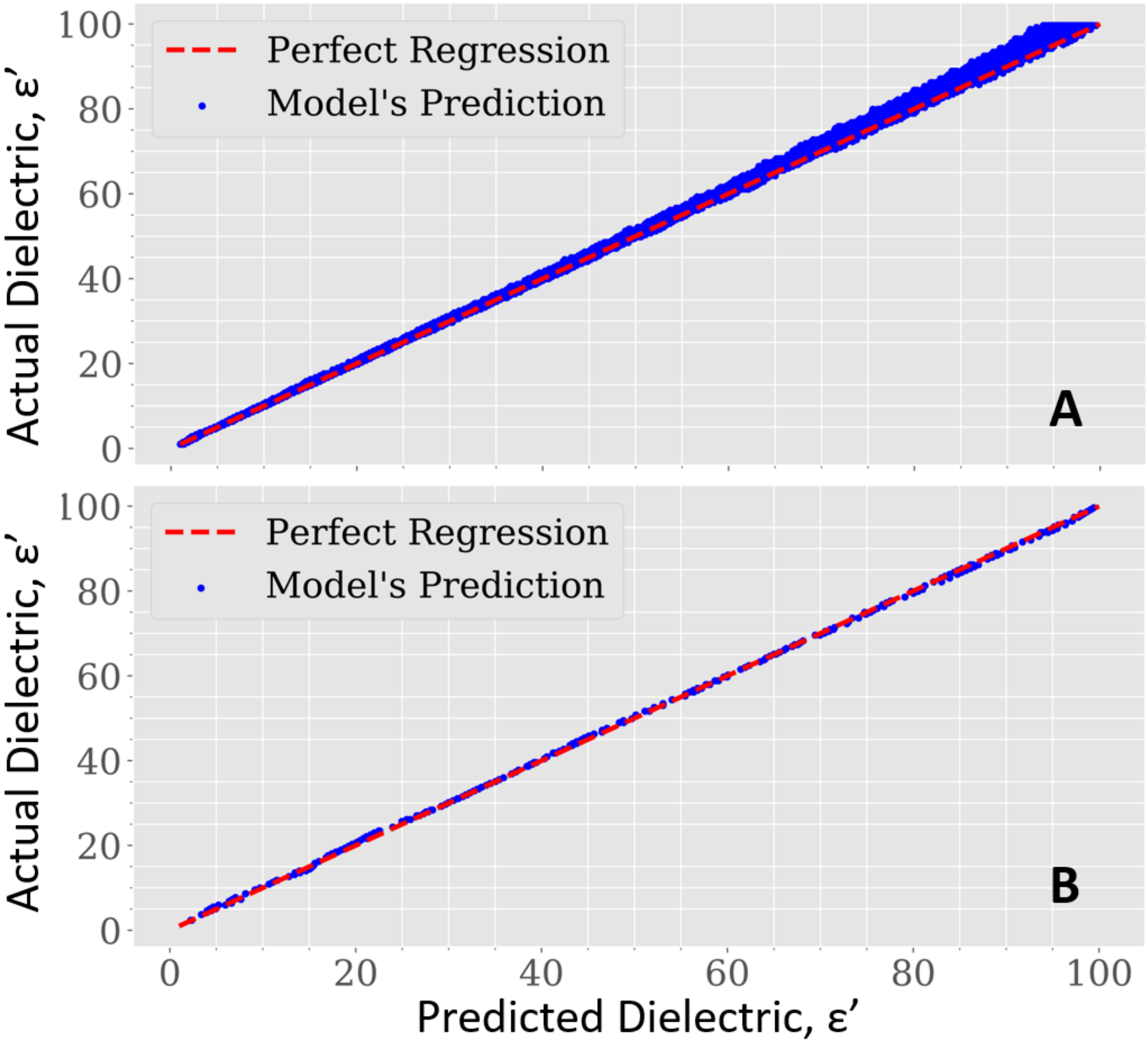


Figure 8: Plots of the predicted dielectric versus the actual dielectric. Subfigure A is the ANN results based on training each dielectric and frequency independently. Subfigure B is the ANN results based on training with dielectric associated with an array of frequencies ranging from 0.1 to 13.5 GHz.

1.4.2 Experimental Data Results

To validate that the ANN architecture that was selected could be used in future applications experimentally collected data needed to be tested on it. This was accomplished using a Teflon piece of 44.45 mm in length and machined to fit the high-precision coaxial airline. The validation metrics used previously in this study were performed on the dielectric constant of the Teflon piece as well as other dielectric properties such as the loss tangent. The scattering parameters from the Teflon piece were evaluated using the different CNN approaches. The pre-trained ANN was loaded into Python as a json file with the weights saved as an h5 file. The Teflon’s scattering parameters were evaluated over the frequency range and compared to the NRW results for evaluation. The performance of ANN at predicting the dielectric constant and the loss tangent of the Teflon is shown in Table 2, once again the system was evaluated using the MSE and MAE. The equation for loss tangent is shown in Equation 7, epsilons are the associated components of the complex dielectric.

	MAE Model 1	**MSE Model 1**	**MAE Model 2**	**MSE Model 2**
ε'	0.24	0.22	0.56	0.66
tan(δ)	0.19	0.16	0.59	0.64

Table 2: Comparison of the predicted dielectric properties with experimentally determined dielectric properties of Teflon. The experimental values are based on the NRW method with ε' = 2.16 and loss tangent = 0.0007. MSE=mean squared error and MAE=mean absolute error. Model 1 is associated with subfigure 8A and Model 2 is associated with subfigure 8B.

Study 2: Synthetic Generation of Chemical Reaction

The first study (A) introduces a method referred to as Autonomous Generation of Reactions and Species Variational Autoencoder simplified to AGoRaS, for the synthetic production of large quantities of balanced chemical reactions that can be utilized for the unbiased training of artificial intelligence techniques. As well as for help in generating new targeted reactions, which can help guide experimental studies. Previous research by Amini et al. demonstrated how VAE can be used to generate unbiased synthetic machine learning training data for the training of more robust and accurate algorithms[94].

The work by Iovanac et al. is a great example of how an encoder and decoder-style neural network can be used to represent a continuous chemical latent space. Their work uses both real and density functional theory-predicted models to predict properties of various pKa predictions of moderately sized molecular species[95]. However, this is just an illustrative example what they clearly show is how scarce data can be overcome in the chemical space by using a latent representation of chemical molecules. They even show how that latent space can be used as training data for a neural network. This study takes this work a step farther and expands this methodology not only to single molecules but to whole equations containing a variety of molecules of different sizes. Unlike previous studies that focus on reactants predicting products or vice versa, this study instead allows for the latent representation of the data to predict both parts of an equation.

Other methods used in literature have shown the ability of neural networks to autonomously predict chemical attributes of equations. Take the work from Zhang et al. where they use a combination of unsupervised K-means clustering and support vector machines to help in the prediction of activation energy of catalytic chemical reactions. They use the outputs of these unsupervised methods as inputs to a trained neural network. This is another way of generating a latent space [95]. In comparison to the study by Iovanac et al. the K-means clustering, and the classification can be thought of as the “encoder” while the predictive neural networks would be the “decoder”. Unlike this paper’s study, the focus of Zhang et al.'s work was on using whole chemical reactions and their physical quantities to predict the activation energy. While this study attempts to capture as much of that latent information as possible in the SMILES descriptions of gas-phase reactions.

Study one proves that VAEs can be applied to chemistry to generate a large number of both new reactions and new species. This will allow not only for the improvement of machine learning algorithms but also offer significant time and cost savings to experimental studies. The ability of these networks to generate a near-continuous and infinite space of new reactions will allow researchers to sort the data for specific products, and reactants, for a given thermodynamics condition.

The AGoRaS method was developed to synthetically generate chemical equations without the need for human interaction. A pipeline will be created to accomplish the following steps outlined in figure 13. The motivation behind having a pipeline set up was so that new data could easily be ingested into the pipeline to start quickly and effectively start training AGoRaS on a variety of data. The use of data pipelines and their effect on code quality and robustness is a well-studied topic[96–98]. This ability to take a lot of the data cleaning out of the hands of the user will allow AGoRaS to be more accessible to a non-data scientist. It will instead allow chemists and engineers to utilize the network using their data. Which has historically been a problem, where the lack of data science skills can hold back a field from utilizing the power of artificial intelligence[99].

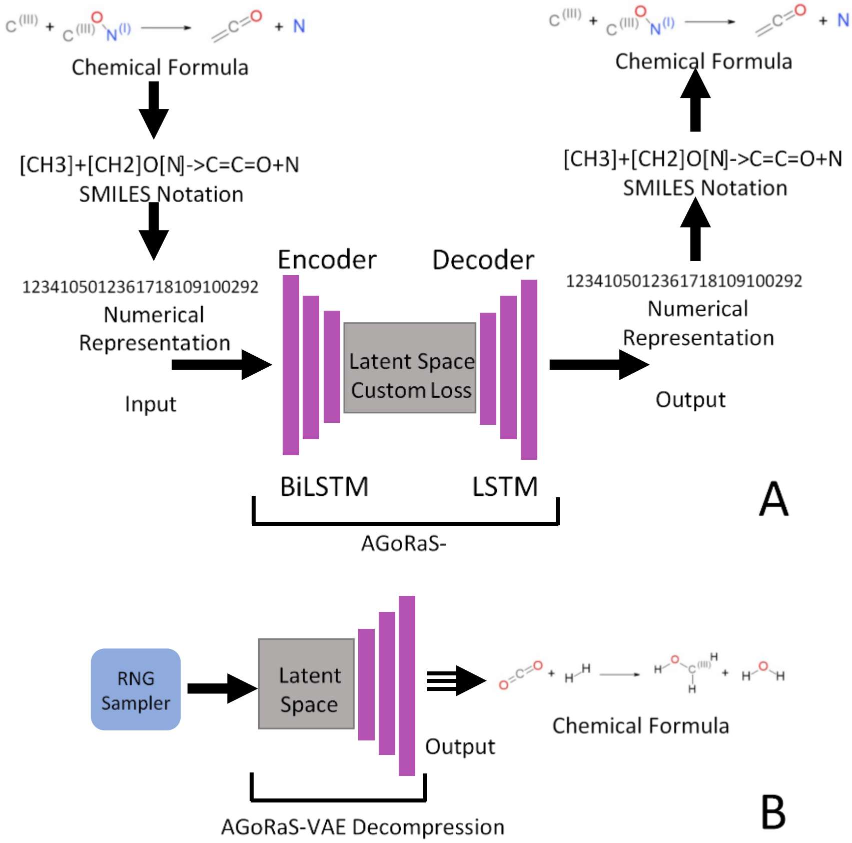


Figure 9: Subfigure A is an illustration of the AGoRaS based VAE during training. Chemical database information is compressed and decompressed to form a high-dimensional latent space. Subfigure B is an illustration of how the trained latent space can be sampled to generate new chemical formulas

2.1 Study 2 Significance

Along with the increased use of machine learning in chemistry, researchers have begun to see that a biased problem exists in every available dataset of chemical reactions[15,21,50,100]. Most datasets for chemistry come from a collection of patents and research articles that exist on the internet, however, this does not complete the continuous chemical reaction solution space [100,101]. This study aims to utilize the predictive abilities of deep learning to synthetically generate a chemical reaction dataset that is less biased and more robust than the ones currently available or can be data mined. This is accomplished using a generation deep learning technique known as a variational autoencoder (VAE)[102]. It is hypothesized that a VAE will form a custom chemical compression intelligence that will provide efficient generation of new reactions by sampling the latent space of the VAE. Where the latent space can be thought of as a representation of compressed chemical information. Where a neural network will learn the chemical and structural similarities of various equations. This is important because it allows the neural network to look at data in a way that it can find patterns in a higher dimensionality.

The power of using a VAE over other traditional methods mentioned above such as CNN is that a VAE represents its latent space as a probabilistic distribution. This can be thought of as a multivariate Gaussian distribution. This allows the VAE to be able to map a range of inputs for each possible input vector. This is a unique approach that relies on artificial intelligence to generate new reactions rather than retrosynthesizing reactions and tracking molecular species[8,19,20,103]

This study introduces a method referred to as Autonomous Generation of Reactions and Species Variational Autoencoder (AGoRaS), for the synthetic production of large quantities of balanced chemical reactions that can be utilized for the unbiased training of artificial intelligence techniques. As well as for help in generating new targeted reactions, which can help guide experimental studies.

Study one study proves that VAEs can be applied to chemistry to generate a large number of both new reactions and new species. This will allow not only for the improvement of machine learning algorithms but also offer significant time and cost savings to experimental studies. The ability of these networks to generate a near-continuous and infinite space of new reactions will allow researchers to sort the data for specific products, and reactants, for a given thermodynamics condition

2.2 Study 2 Background

The issue with biased datasets in the teaching of machine learning algorithms comes down to a common principle of inherited bias evident in most artificial intelligence techniques, sometimes colloquially known as, “garbage in garbage out”[104]. Having good data is the most important aspect of any artificial intelligence technique. However, it is not just having optimized data. Machine learning techniques do much better when taught on a range of inputs and outputs and the space is continuous[104,105]. In essence, a machine learning technique needs to learn everything, to know everything. In the paper by Jia et al., a detailed explanation is given of how the researcher’s bias goes into the creation and design of almost all experiments. They go on to show how this can be combined with other types of biases to skew reported data[100]. Their work is further backed up by Griffiths et al.'s study of biases in the natural sciences. Where a detailed explanation is provided on the effects of data splitting noisy datasets as well as the influence contextual variables can have on the outcome of experiments[103]. Kovacs et al. do an excellent job of illustrating the direct effects that a biased and unbiased dataset can have on the quality of machine learning's outputs [21]. All of these different studies combine to show that the current methodology for creating datasets from publications or other open-source resources is flawed and will limit the ability of future machine learning algorithms. Glavatskikh et al. demonstrated how the lack of diversity in data limits machine learning's potential to predict[106]. Meanwhile, the cost of building an experimentally unbiased, continuous dataset is potentially an infinite feat.

Despite the limited dataset availability and biases, research is currently demonstrating the employment of artificial intelligence techniques in the field of reaction chemistry. Inorganic chemistry, Kayala, and Baldi show how machine learning can be utilized to predict multistep reactions that take place over a range of reactants and conditions[107]. Moreover, they demonstrate how machine learning can be employed despite the limitations of the training dataset, still being able to predict mechanistic reactions. In their study, they discuss the necessity of manually sorting the dataset for the reactions they could utilize[18]. A follow-on study by Kovacs et al. dissects their paper and concluded that the Clever Hans effect is apparent. This is essentially a failure of the double-blind condition, or in this study, the correct prediction of the reactions was reached but only because of a reaction bias[21]. Similarly, Mater and Coote emphasize the common problem of bias in chemical reaction datasets[108]. These examples show that while it is understood that bias plays an important role in the predictions of machine learning and other artificial intelligence techniques no consensus has been reached on how to deal with this problem.

However, in many different fields that faced similar dataset issues the use of generational techniques has shown excellent results. The most famous examples of these are typically found in the medical field, where it is notoriously difficult to share patient data (due to Health Information Privacy) among scientists and researchers. Choe et al. came up with a way to generate realistic synthetic data that contained high-dimensional discrete variables[109]. Their network known as MedGan has shown excellent comparative performance to real-world datasets. Rigorous analysis has been done on the generated datasets including distribution statistics, predictive modeling tasks as well as expert review[109]. These findings are backup up by the research of Camino et al. and Gulrajani et al. who demonstrated that generative networks can be used to generate multi-categorical outputs completely synthetically[110,111]. This is an important aspect of generational techniques, especially VAE. It allows for the autonomous synthetic generation of data for both discrete and continuous variables[112,113].

The other popular generative technique is known as Generative Adversarial Networks (GAN). These have notably outperformed VAE on various use cases such as image generation and discrete variable representations of problems[114]. However, they are notoriously hard to work with resulting in very unstable networks. Most importantly for this study is that GAN needs a lot of data and a lot of tuning. This makes it undesirable for smaller datasets typically used in chemical reaction engineering. In addition, there is no latent space generation when using GAN. This limits the usability in further use cases such as using latent space as an input into a generational network[114].

This is an active area of research with Yu et al. contributing by demonstrating that generative networks can be used to model long sequences that rely on earlier segments to be semantically correct[115]. More importantly to this work, generative networks can generate missing data within a dataset[94]. Meaning that these deep learning techniques can interpret between existing data points to generate new synthetic points that share the same characteristics as the original dataset.

One approach that has been employed recently for the generation of chemical reactions is the approach of retrosynthesizing known reactions[19,20,116]. The idea of retrosynthizing involves subsequently breaking the reaction products down by putting them back into the reaction process as reactants. This provides a complete continuous space of intermediate molecular species, but it is limited to the generation of new reaction chemistries that are not bound by the initial reaction chemistries in the database. Chemists have been taking retrosynthesis further by combining artificial intelligence into the process. The work by Segler et al. shows a promising route of combining Monte Carlo tree search with symbolic artificial intelligence. Which leads to a speed-up of 30 times over traditional methods[8]. The approach taken in the following study is a forward approach that relies upon the forward prediction from the VAE model.

Other work has been done by several research groups that focus on improving the outcome of chemical reaction predictions. Work by Shields et al. shows how Bayesian optimization can be used to fine-tune neural networks in synthetic chemistry [117]. In their paper, they show how Bayesian optimization can essentially be thought of as an autonomous tool for limiting human biases. They then go on to show how removing expert chemists and engineers in real-life experiments instead of using autonomous optimization of experiments can get both a higher optimization efficiency and a better consistency. Importantly though this Bayesian optimization only optimizes parameters that a human has set and can't create new parameters as it "learns".

A review of reaction prediction methods by Gale and Durand shows how almost all areas of machine learning in chemistry require improvement and are active areas of research[118]. Some of the important topics they touch on are the need for datasets to have not only error-free reactions but also negative results. They also discuss the difficulty in encoding chemical information into a machine-readable format. Another key point is their discussion on how you can use generative, discriminative deep neural networks to assess chemical reaction synthesis. Unfortunately, there is no discussion on how latent representations of chemical data can be used to generate feasible reaction synthesis.

2.2.1 Laws of Thermodynamics

The laws of thermodynamics are a set of empirical facts that give the basis for the interaction of temperature, energy, and entropy in a thermodynamic system that is in thermodynamic equilibrium. Traditionally the laws of thermodynamics are broken into 3 laws, however, there is a fourth law termed the "zeroth law" that came after the initial definition of the 3 laws of thermodynamics. This law helps to explain some of the physical interactions in the proceeding laws. The zeroth law of thermodynamics defines a definition of temperature that importantly, does not rely on entropy[119]. By defining temperature this way, the definition can be thought of as empirical.

The first law of thermodynamics is the law of thermal conservation, which states that the total energy of an isolated system must be constant. Energy can change forms but cannot be created or destroyed[119]. In an idealized closed system where there is no transfer of matter in or out of the said system this first law of thermodynamics can be generalized to Equation 13. Where ΔU is the internal energy of the system, Q is the heat supplied to the system and W is the work done on the system by its surrounding[119,120].

$$\Delta U = Q + W \qquad \text{Eq. 13}$$

The second law of thermodynamics is that heat does not spontaneously pass from a colder body to a warmer body. In any natural thermodynamic process, the entropy sum of n number of thermodynamic systems never decreases[119]. This helps to describe the tendency of a natural system to favor spatial homogeneity when it comes to temperature. Again, in an idealized closed system two initially separate systems put together will reach thermodynamic equilibrium. This new system will have a sum of the entropies equal to or greater than the total entropy of the isolated system.

The third law of thermodynamics is that as the temperature of a system goes to absolute zero its entropy approaches a single constant value[119,120]. Entropy is increased during irreversible and reversable processes, due to entropy being a state function. Regardless of if the process is reversible or not the increase in entropy would be the same. Since energy always flows downhill, entropy will increase during these processes. As molecules become larger, possessing more energy levels the entropy of molecules also increases. This is due to the ability off these higher energy levels to be significantly occupied[119,121]. For nearly all systems the entropy of a system at absolute zero would typically be close to zero.

2.2.2 Entropy

Entropy is defined in thermodynamics by a measurable physical property such as temperature, volume, pressure, and bulk mass. Entropy is defined by statistical mechanics as the statistics of motions of microscopic constituents of a system[119]. These two approaches form the basis for the second law of thermodynamics, allowing it to have a unified view of the same phenomena. In classic thermodynamics, the state function entropy can be defined as a reversible system, with Equation 14[120]. Where S is the entropy and Q is the heat supplied to the system and T is the temperature.

$$dS = \frac{\delta Q_{rev}}{T} \qquad \text{Eq. 14}$$

This property of entropy allows for entropy to be thought of as a state function of thermodynamics[120]. Meaning that entropy is only dependent on the current state of the system and is not affected by how the system entered its current state.

2.2.3 Enthalpy

Enthalpy is a property in thermodynamics that describes a system's internal energy. In thermodynamics, this is defined by Equation 15. Where H is the enthalpy, U is the internal energy, p is the pressure and V is the volume of the system[119,120].

$$H = U + pV \qquad \text{Eq. 15}$$

$$h = \frac{H}{m} \qquad \text{Eq. 16}$$

$$H = \int (\rho h) dV \qquad \text{Eq. 17}$$

The pressure-volume term (pV) is a representation of the work required to achieve the system's physical dimensions. This can be very small for solids, liquids, and gasses under ambient conditions[119,122]. Therefore, typically makes Enthalpy mostly reliant on the chemical energy of a system. For a closed system, a term-specific enthalpy can be described in Equation 16 where it is referenced against a unit of mass m. This definition allows for enthalpy to be described as an integral representing the sum of the enthalpies of all the elements in a given volume[120]. This is shown in Equation 17 where ρ is density, dV is a small element of volume within a thin horizontal layer.

2.2.4 Gibb's Free Energy

In thermodynamics, the Gibbs free energy is the maximum reversible work that can be done by a thermodynamic system at a specific temperature and pressure. If the process being described is reversible then the Gibbs free energy describes the maximum amount of non-expansion work that can be extracted from a system. Gibb's free energy is described in Equation 18[119,123]. Where G is the Gibbs free energy, H is the enthalpy of the system, S is the entropy of the system and T is the temperature.

$$\Delta G = \Delta H - T\Delta S \qquad \text{Eq. 18}$$

In chemistry, the Gibbs energy is the thermodynamic potential that is minimized when a system enters a state of chemical equilibrium while at a constant temperature and pressure[120]. The derivative of Equation 18 approaches zero as the system enters a state of chemical equilibrium (at constant temperature and pressure). This means that the lower the ΔG of a reaction the more likely that reaction is to happen spontaneously.

This is a quantitative way to measure the favorability of a given reaction at constant temperature and pressure. For a reaction to happen the ΔG must be smaller than the non-pressure and volume work (nonPV). Often this is equal to zero meaning that ΔG must be negative for the reaction to happen spontaneously. If ΔG is greater than the nonPV then additional work would need to be added to the system to allow for the reaction to happen[123].

2.2.5 Chemical Reactions

A chemical reaction is traditionally defined as a set of chemicals called reactants/reagents being chemically transformed into another set of chemicals called products. These product(s) typically have different chemical and/or physical differences from the reactants. These changes happen in classical chemistry through the forming and breaking of chemical bonds between atoms via the positions of the electrons[124]. There is no change in the elements present from the products to the reactants[7]. This allows for reactions to happen in forward or reverse directions until a state of equilibrium is reached.

The rate at which chemical reactions take place typically has a direct correlation to the temperate and pressure of the reaction environment. At higher temperatures, there is more energy in the form of thermal energy available[124]. This allows for a reaction to reach the activation energy required to initiate the change. This change allows for the breaking of bonds between atoms and the formation of new species[124].

These chemical reactions are described using symbols to represent the starting and end materials. However, they sometimes include things like elementary reactions that are needed to reach the final product. Elementary reactions are sequential sub-reactions that can provide information on the course of the reaction mechanism[125].

2.2.6 Chemical Equations

Chemical equations are a set of symbols representing elements, that are used to graphically illustrate reactions. The products and reactants of a chemical equation are separated by either an arrow (→) or a double arrow (⇌)[126]. Where the single arrow represents a reaction that proceeds in only one direction, while the double arrow represents equilibrium reactions. Equations typically have the form of Equation 19, where uppercase letters are representing elements and lowercase letters are representing the scaling units. A key point of chemical reactions is that they need to be balanced, meaning the same number of elements need to appear on both sides of the equation[127]. This is termed stoichiometry, and it plays a key role in all forms of chemistry. This balancing is achieved by scaling the number of elements on either side of the equation until the sums are equal[127].

$$aA + bB \rightarrow \mathrm{cC} + \mathrm{dD} \qquad \text{Eq. 19}$$

2.2.7 Thermodynamics of Reactions

The laws of thermodynamics discussed earlier play an essential role in chemical reactions. In particular, the enthalpy and entropy quantities describe the free energy of the reactions. Known as the Gibbs free energy is given in Equation 18. If an equation is exergonic and releases energy a reaction can happen spontaneously. These reactions are exothermic releasing heat, meaning that the enthalpy is negative[119,127].

On the opposite end of the spectrum if a reaction is endothermic and absorbs energy from the system the entropy of the system would most likely need to increase. A typical way to achieve this is by forming high entropy products such as gaseous products. As discussed earlier entropy increases with temperature therefore, endothermic reactions typically take place at high temperatures[119,127].

2.2.8 Kinetics

Reaction kinetics is the rate at which a reaction occurs and is dependent on several factors. These include the temperature, pressure, activation energy, surface area, reactant, and the presence of a catalyst. As discussed in the thermodynamics section, higher temperatures increase the energy of the system. This in turn increases the number of atomic collisions happening within a given time frame. An increase in pressure also increases the reaction kinetics by reducing the volume of space that a given number of atoms can occupy. This in turn increases the collision rate of the molecules leading to higher reaction kinetics[127].

Similarly, if the reactant concentration is raised the reaction kinetics will increase due to an increase in the collision in a given time frame[127]. This is achieved by putting more molecules into a given volume. The activation energy is the amount of energy needed to allow for a reaction to start and to continue on its own[127]. The higher the activation energy the less favorable the reaction and the more energy it will need to continue. Another way to control reaction kinetics is via the available surface area[127]. This plays a much more important role in catalyst reactions and other gas, solid reactions. The higher the surface area the more chance the surface has to interact with the molecules. The existence of a catalyst can play a huge role in reaction kinetics. Catalysts can change the reaction mechanisms and greatly increase the rate of reactions by subsequently lowering the activation energy needed[127,128].

2.2.9 Reaction types

Four main types of reactions are synthesis reactions, decomposition reactions, single replacement reactions, and double replacement reactions[7,129].

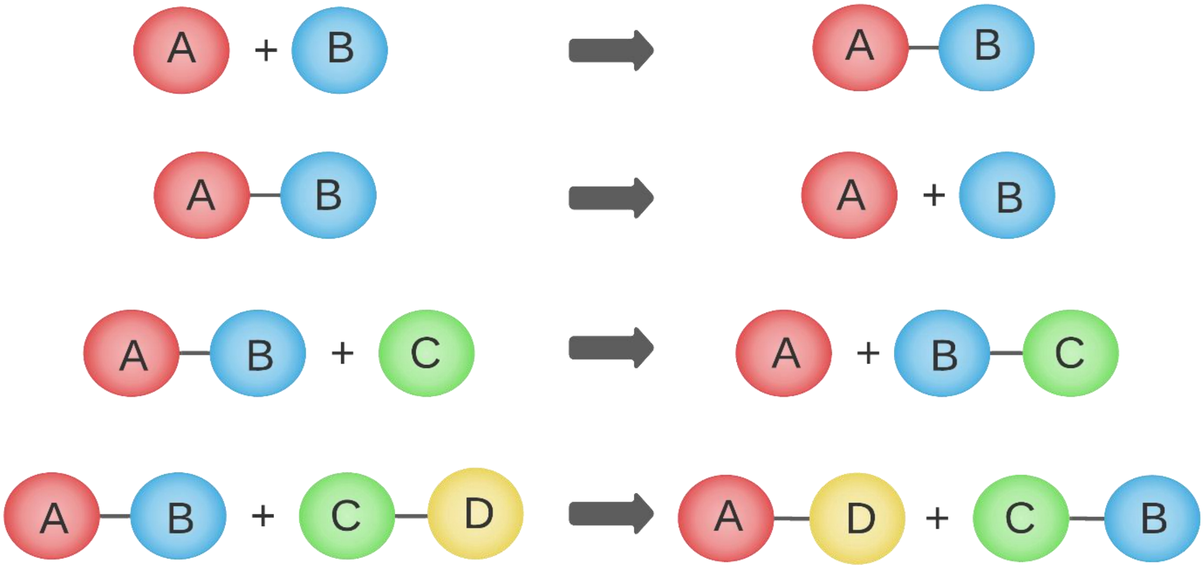


Figure 10: Simple representation of the four main types of chemical reactions. The first is a synthesis reaction representing the simple combination of two substances. The second is a decomposition reaction representing the breakdown of a more complex substance into its parts. The third is a single displacement reaction representing when a single substance is replaced in a complex substance. The fourth is a double displacement reaction where two compounds change place.

Synthesis reactions are when two or more substances combine and form a more complex substance, as seen in Equation 20. Decomposition reactions are the opposite when a complex substance is broken down into its constituent parts, as seen in Equation 21. Single replacement reactions are when a free-floating element replaces another element in an already-existent compound, as seen in Equation 22. Double replacement reactions are when two compounds swap elements to form two brand-new compounds, as seen in Equation 23[129].

$$A + B \rightarrow \ AB \qquad \text{Eq. 20}$$

$$AB \ \rightarrow \ A \ + \ B \qquad \text{Eq. 21}$$

$$A + BC \rightarrow AC + B \qquad \text{Eq. 22}$$

$$AB + CD \rightarrow AD + CB \qquad \text{Eq. 23}$$

2.2.10 Gas-Phase Reactions

Gas-phase reactions are chemical reactions in which the reactants and the products are gaseous under the conditions in which the reaction is occurring. For a homogenous gas-phase reaction this means that all reactants and products that occur will be in a gaseous form[127,130]. While in a heterogenous gas-phase reaction, at least one reactant or one product will be in a solid or liquid phase when the reaction is in equilibrium. Gas-phase reactions are classified as being either intramolecular or intermolecular reactions[130]. Intramolecular reactions involve a reaction between two or more atoms in the same reactant molecule. Intermolecular reactions involve a reaction between two or more reactant molecules[130].

2.2.11 Collision Theory

Collision theory plays a key role in chemical reactions and especially in gas-phase reactions. Collision theory aims to explain how reactions occur via the collision of particles. It states that particles must hit each other at a specific orientation at an energy level greater than or equal to that of the activation energy[131]. If this occurs then the collision can break the existing chemical bonds and form new ones[126,127]. This is why increasing temperature or pressure allows for a higher kinetic rate. The collision theory is derived from the kinetic theory of gases with the assumption of ideal gasses[131]. Some of the other assumptions that collision theory uses are that, molecules travel through space in a straight line, all the molecules are rigid spheres, and the reactions only occur between two molecules.

From the collision theory of gasses, it is possible to derive the rate constant for gas-phase reactions, which is equal to the rate of successful collisions[127,131]. Successful collisions are defined as a proportion of successful collisions multiplied by the overall collision frequency. This helps to explain why reactions occur at different rates from one another[131]. It also provides a template for ways to increase or decrease the rates of the chemical reaction.

2.2.12 Simplified Molecular Input Line Entry System

As discussed in section 3.2.5 chemical equations are a set of symbols representing elements, that are used to graphically illustrate reactions. However, this is not the only way to represent chemical reactions. A newer method called simplified molecular-input line-entry system (SMILES) was introduced in the late 1980s[132]. SMILES is a better specification of chemical species in the form of a line notation for describing the structure of chemical species using short ASCII strings. This notation works by encoding the molecular structure and specific instance of a molecule into a string[132,133]. This is an extremely powerful way to represent molecules as it allows molecule editors to convert a string into a 2- or 3-dimensional model of a molecule. A common application of SMILES is indexing and ensuring the uniqueness of molecules in a database.

A drawback of smiles is that several equally valid SMILES strings can be written for a molecule. However, another useful benefit of SMILES notation is that it allows the specification of configuration at tetrahedral centers, as well as double bond geometry. These are structural features that cannot be specified by the bonding sequences alone[132,134]. Any SMILES string which encodes this information is termed isomeric SMILES. A notable feature of these rules is that they allow rigorous partial specification of chirality.

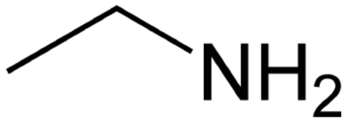
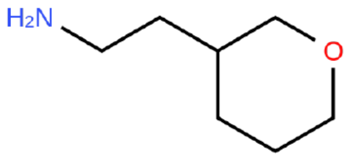
Name	SMILES	Structure
Ethylamine	NCC	
2-(Tetrahydrofuran-3-yl) ethanamine	C1COCC1CCN	

Figure 11: SMILES Representation of Chemical Compounds with name, SMILES structure, and 2D structure is shown side by side for comparative purposes.

2.2.13 Context-free language

When looked at from a formal language theory standpoint a SMILES string can be thought of as a word that contains symbols from an alphabet with inherent value. Formal language theory states that an alphabet must exist that contains symbols, letters, or tokens that can be merged into strings of a language[135]. In the case of SMILES context-free grammar is used to define the rules of this language[136]. This type of definition lends itself well to allowing for SMILES words to be computer ingestible. Where the rules of SMILES represent the grammar of the chemical language in which the words represent concepts that are associated with a particular meaning[136].

By looking at SMILES in this way many tools and theories from natural language theory can be applied to chemistry. This led to the rise of cheminformatics, where the basic theory is that similar molecules have similar properties[137]. Like words in human languages have synonyms that are closely related. These representations can be grouped in high dimensions to create a unique representation of the language. This has been proved in the work by Mikolov et al. which was followed up in the field of cheminformatics by mol2vec the seminal work by Jaeger et al.[138,139] These works allow for predictive models to implement a syntactic pattern recognition approach as well as a more robust scheme based on statistical pattern recognition[133,134].

2.2.14 Chemical descriptors

In a SMILES string atoms are represented by their chemical symbols which are unique for every element. Atoms typically have brackets surrounding them but as per the grammatical rules of SMILES, these can be excluded if an element meets five different criteria[132,140]. The first is that the element needs to not have a formal charge, second, they must be the normal isotopes of the element, third they need to not have chiral centers, fourth they need to be an organic compound in the subset of B, C, N, O, P, S, F, Cl, Br, or I, and fifth they must have the hydrogens implied by the SMILES valence model[132,136]. If the element does not meet these criteria they have to have charges, and hydrogens explicitly stated along with being put inside of brackets. When the brackets need to be used the hydrogen (H) symbol is also added in the case where the bracketed atom is bonded to at least one H. Additionally a + or – sign is added to the end of the element to denote a positive or negative charge[136].

The bonds between these atoms are always assumed to be single unless explicitly specified otherwise. Single bonds are implied whenever two atoms are adjacent to one another in a SMILES string. They can however be specified by the – symbol. For other types of bonds, the symbols in Table 3 are used to specify the bond type. Three special symbols are the ".", ":", and "/" symbols. Where the "." symbol represents non-bonded adjacent elements, ":" represents aromatic one-and-a-half bonds, and "/" represents single bonds that are adjacent to double bonds. This last symbol helps to indicate stereochemical configurations[132,136].

Symbol	**Bond**
.	Non-Bond
-	Single Bond
=	Double Bond
#	Triple Bond
$	Quadruple Bond
:	Aromatic Bond
/	Single Bonds adjacent to Double Bonds

Table 3: List of chemical bond representations available in the SMILES language.

Chemical ring structures can be represented using the SMILES notation by breaking each ring at some arbitrary point. These broken rings are used to make acyclic structures while an addition of a numerical ring closure allows for information on connectivity between non-adjacent atoms[132,136].

2.2.15 Variational Autoencoders

Variational autoencoders (VAE) are a type of ANN architecture that utilizes probabilistic graphical models and variation Bayesian methods. Where graphical models refer to the methodology of a graph structure expressing the conditional dependence structure between random variables. While the Bayesian methods refer to a family of techniques for approximating intractable integrals that arise in machine learning. The strength of VAEs comes from the ability of variational Bayesian methods to model the relationship between observed variables and unknown parameters and latent variables.

VAEs are designed to compress input information into a contained multivariate latent distribution, referred to as the encoding. While also being able to reconstruct that compressed data accurately, referred to as decoding. This is accomplished by encoding variable inputs as a distribution over the latent space rather than a single point in the latent space. Training of VAE is slightly different than what has been previously described in section 2.13.1. For a VAE the first step is to encode the input as a distribution over the latent space. The second step is to sample a point from that latent space distribution. The third step is to decode that distribution and calculate the reconstruction error. Then the error is backpropagated through the network so that the neuron weights can be adjusted. A depiction of this encoding and decoding training is shown in figure 14.

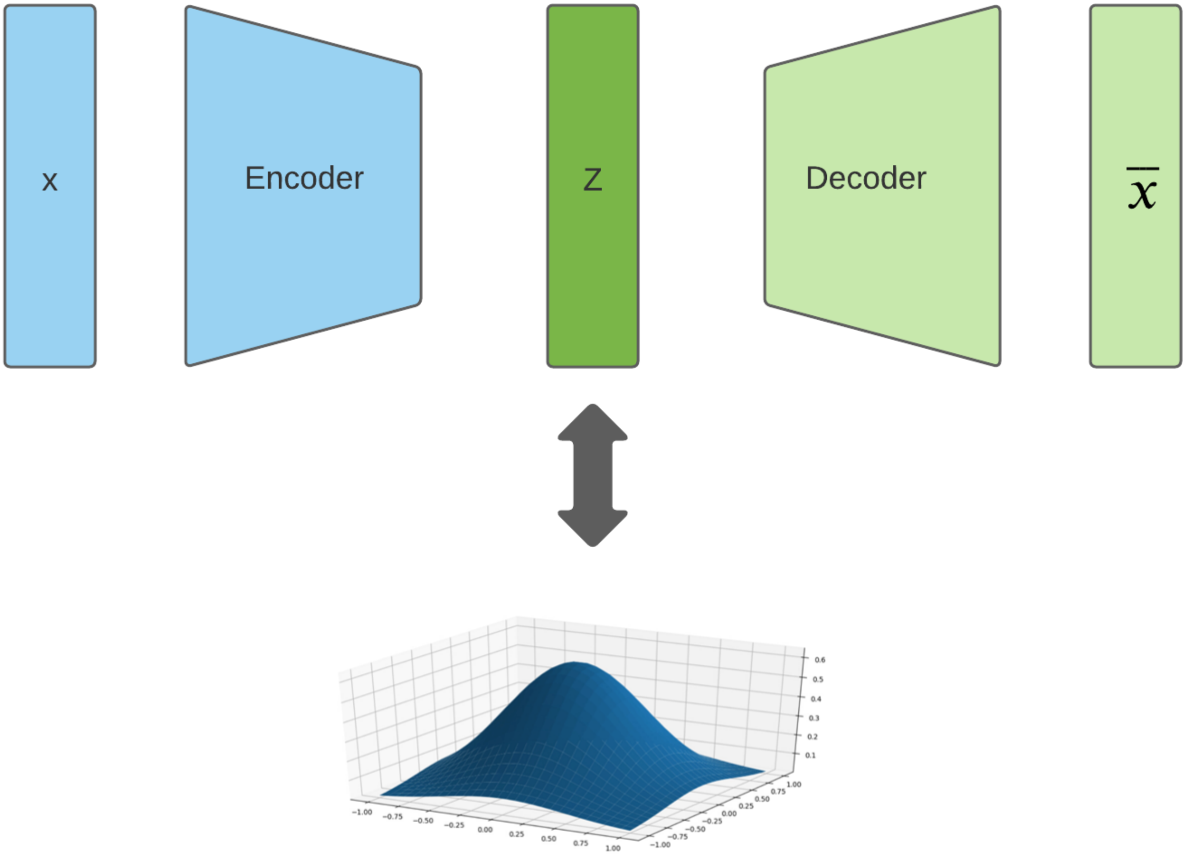


Figure 12: General representation of a variation autoencoder style of the network. The left side of the figure represents the input (X) and the encoder block of the network, while the right side is the decoder block of the network and the reconstructed input ($\bar{X}$). While Z represents the latent space of the network which is developed through training and is illustrated as 3D gaussian here.

2.2.16 Generative Adversarial Network

A popular machine learning network-style commonly used in generative tasks is a generative adversarial network (GAN). A GAN network is two separate neural networks that compete to try and outperform each other. They are separated into a generator network and a discrimination network. Given some training data, the goal is for the generative network to learn the data distribution to create new outputs that mimic that of the real data. While the goal of the discriminator network is to differentiate between real and synthetic data. This competition takes place in the form of a zero-sum contest, where the losing network is forced to update its weights and attempt to get better at the designated task.

2.2.17 T-Distributed Stochastic Neighbor Embedding

A common tool for visualizing high-dimensional data and specifically outputs from machine learning is t-distributed stochastic neighbor embedding (t-SNE) plots. This methodology is considered an unsupervised machine learning network based on Stochastic Neighbor Embedding. Where each data point is given a location in a two/three-dimensional unitless map. T-SNE is a nonlinear dimensionality reduction technique, which specializes in low-dimensional embeddings of high-dimensional data. T-SNE plots have two steps, the first is to construct a probability distribution over every two inputs. The second step is to define a similar probability distribution within the 2/3-dimensional map. There are several benefits to approaching the problem in these ways. The first is that objects with high similarity are assigned a higher probability while in high dimensional space. Next, when this probability is mapped down to a lower dimensionality the distance between points can be minimized.

2.2.18 Dataset Bias

In machine learning, database bias is a type of error in which the dataset is not equally representative of real-world data. Typically, that means the data is skewed towards certain types of inputs, giving them greater weight in the model. This typically results in skewed outcomes, analytical errors, and low accuracy levels. This is because nearly all machine learning models are data-driven models, which means they learn and make decisions purely based on the training data provided to them. So, if the human building the dataset, and or the data itself is biased it will directly affect the results of the models. Due to the high cost and long time spans typically associated with building datasets, it is extremely important to be able to identify and eliminate biases in machine learning. Some common types of data biases are sample bias, exclusion bias, measurement bias, confirmation bias, association bias, recall bias, and outliers.

Sample bias is when a dataset fails to reflect the actual real-world data that it will be required to process. Exclusion bias is when variables of a dataset are deleted because the creator of the dataset deemed them unimportant. Measurement biases are when the data collected specifically for the training of machine learning models differs from that of actual real-world data. It can also refer to when the tools used to measure the data are faulty or differ from the intended use case. Confirmation bias is when an observer biases the inputs and outputs of a model based on what they are expecting to see rather than what the data truly represents. Association bias refers to when a dataset reinforces an observation that has no causation in the real world. Recall bias is typical for poorly labeled data, where similar types of data are labeled inconsistently and/or input data is structured differently. Outliers in machine learning are data points that are extremely far from the normal data flow and therefore eliminated. However many fields of machine learning such as anomaly detection rely on outliers to accurately predict an outcome.

2.2.19 Chemistry Bias

The problem of biased datasets is an extremely important factor in chemistry machine learning. Several factors make chemistry unique to bias problems, the first is that nearly all chemical experiments are planned by a human scientist. This makes the studies and results subject to human cognitive biases, heuristic biases, and social influences. The second problem is that in the history of chemistry scientists have only been encouraged and rewarded for sharing their most optimal results. This means that for every paper published with a new chemical reaction there are likely dozens or hundreds more reactions that worked as well but produced a lower yield, slower reaction time, or happened under less favorable conditions. But since this data isn't reported and is part of the public domain knowledge it is therefore inaccessible to machine learning algorithms. As stated in section 2.14.3, machine learning algorithms need the training data to be representative of real-world conditions. These historical and human factors lead chemistry AI and chemistry datasets, in particular, to be extremely vulnerable to producing biased results.

2.2.20 Data Collection

Currently, the most widely used tool for developing chemistry databases for machine learning is by using data mining[141,142]. Where data mining refers to the methodology for extracting specific information from various sources and transforming them into a coherent structure for the use of machine learning. In chemistry, the main sources of information in data mining are research articles and patents[25,100]. As discussed earlier these sources are already inherently biased toward favorable reactions and therefore not representative of the real world.

2.3 Study 2 Methodology

2.3.1 Data Collection AGORAS chemical equations

The chemical equations used to train the AGoRaS network for chemical equations were provided by the NIST chemical kinetics database[143]. The data was received from them as two separate CSV files. One file contained a column of 15000 id numbers and several other columns that all had different naming conventions. These included chemical formula, hill sorted formula, and IUPAC name. The second CSV file contained 68702 rows corresponding to different chemical reactions in the dataset. Each of the reactions was represented by the id numbers from the first CSV. The first CSV was read into Python using a pandas dataframe[144] where each possible naming convention was a different column.

2.3.2 Conversion to SMILES notation AGORAS chemical equations

SMILES was chosen to be used as the ground truth representation of chemical species due to its ability to represent species with the same chemical formula but different structures uniquely[132]. Smiles are widely used in chemical artificial intelligence and have shown excellent results from other techniques[133,134,145–148]. Despite its popularity, there is no easy and free way to convert different chemical names into SMILES. The pipeline utilizes several different techniques to convert the different names into SMILES notation. Since the available names of the 1500 species were uploaded into a pandas data frame it is it was possible to ping each one of a website for conversion or to utilize a Python package. In the end, two main servers were used for conversions due to their large repositories and their trustworthy data sources, the CADD Group

Chemoinformatics Tools and User Services[149] and Pubchem[150]. Each possible name of the 15,000 different species was put through both Pubchem and the CADD groups databases.

2.3.3 Preprocessing the data for AGORAS chemical equations

Once all species were converted a comparison was done between both database's results and any species that could not be converted or that was not in agreement between the two was eliminated from our pool of valid species. This was part of our attempts to be both rigorous and autonomous in our species data pipeline. It would have been possible to manually go through the different disagreements and correct them, but this fell outside of the aims of this study. Since one of the main goals was reproducibility it was deemed this strategy was the best one. A tertiary check of the species was performed, and each of the species was checked for stability and feasibility using the RDkit package in Python[151]. This is another widely used and well-verified package for computational chemistry that is often paired with SMILES notation machine learning algorithms[17,100,142,145,147,148,152].

Once the number of available species had been reduced, the remaining species could be mapped to the CSV file containing the actual equations. Where any equation that did not have a SMILES, conversion for each species was eliminated from the dataset. An additional check was performed where each of the reaming equations was checked to be sure they were balanced. This was done to catch any equations that may have had incorrect ids for reactants or products. A further criterion was placed on the equations, where all equations with more than three species on the product or reactant sides of the equations were eliminated. This was done to facilitate the convergence of the VAE during training by having similar-length character vectors. Work by Dwarampudi et al. and Prusa et al. show that padding LSTM and Neural networks can cause instability and result in poor network performance[142,153]. To this end, it was decided that using character-level embedding offered an advantage over word-level embedding[142,154]. This would allow for the generative network to not just use the same species it had been trained with but instead allow it to generate new species based on information gained in the training process. The embedding was done using TensorFlow's built-in embedding techniques and was based on a universal alphabet created from all the equations[155]. Additional work by Gaspar et al. shows how molecule embedding can mirror that of NLP embeddings. They show in their work how improvements can be made to the predictive power of machine learning models by formulating the inputs into sequence embeddings[156]. This is an extremely interesting avenue to pursue and could lead to an increased number of attractive and viable reactions for AGoRaS.

2.3.4 AGoRaS structure for chemical equations

AGoRaS takes in a vector representation of length n, where n is the maximum length character-level representation of any equation in the training dataset. This vector is then fed forward into a TensorFlow Embedding layer that projects the input into a higher dimensionality. This is an important step because the projection fits numeric values to a high dimensionality space removing the intrinsic values of the values themselves. This is done because in our 2D space the numeric values have no intrinsic value i.e. 5 is not greater than 6 they just represent two different characters. The projected vectors are then fed into a bidirectional LSTM (BiLSTM) layer with a recurrent dropout of 0.2. The mean and log variance is created from the output of the BiLSTM layer. A sampling function is used to randomly sample this solution space based on the mean and log variance. The network then decodes the sampled solution space using a RepeatVector layer that is wrapped around the output of the latent space, thus turning it into a tensor vector that an LSTM layer could read. The RepeatVector is fed as an input into an LSTM layer, whose output is projected into a vector of length n. The output of this projection is what is used to calculate the loss of the network. AGoRaS uses a sequence-to-sequence style loss function common to variation autoencoders. The metric used for monitoring the network during training is also the standard kl loss. The network was trained for 500 epochs using a batch size of 25, an embedding dimensionality of 500, and a latent dimensionality of 350. The kl weight used was 0.1 and the activation function was a softmax function. The optimizer function was Adam and the learning rate was 1x10−5.

2.3.5 Training AGoRaS for chemical equations

Once the training data had been cleaned and embedded in a numerical format readable by a neural network architecture it is split into three sections. The training set, validation set, and test set where the training set was 70% of available data, the validation set was 20% of available data, and the test data was 10% of the data. Even though AGoRaS utilizes a generative technique it can still be validated using traditional methods. The way VAEs are validated is by the ability of the network to encode the validation set and decode it back to the original string construction with no loss of information. The sequence-to-sequence loss function was used during the training process. This allows for scoring of the reconstructed string versus the original string. This ability to encode and decode the original chemical equations with zero loss of information is typically indicative of a stable latent space. Do to the small data nature of this problem it is extremely important to validate as much as possible the stability of your latent space. To that end, after it had been proven that the network could reconstruct the test data, the remaining 90% were also tested again. Of course, since this data was used in training the network should be able to reconstruct them all. But it serves as a further validation of the stability of the latent space.

2.3.6 Autonomously sample the latent representation for chemical equations

Once a trained network has been created it is possible to construct a sampler that directly interfaces with the latent representation. This is shown in subfigure 9B where the neural net layers responsible for the encoding are removed from the network. It is possible to then start sampling the different latent representations randomly using a randomized point for each element and having the decoder part of the network construct equations. The decoder takes the randomized points and applies the learned weights to construct new equations. Due to the probabilistic nature of the latent representations, it is possible to take an almost unlimited number of sample points and continue to generate new equations. Of course, there would be diminishing returns on this as there are only so many chemically feasible equations possible.

2.3.7 Validating Generated Equations

The methodology for determining the chemical validity of equations was extremely like that of the data cleaning process. The first step was to check duplicate equations were eliminated. The second step was to check if the generated equations were balanced. This offered a computationally inexpensive way to cut down on the number of equations present in the generated dataset. The third step was to check each species for chemical validity using RDKit, where any physically or chemically unstable species should be rejected by the software[151]. This is a common practice when using a neural network with generated SMILES species[134,140,145,146,148,157–159]. The ability of the network to generate valid chemical species helps to further prove that the latent space is stable and representative of the original dataset.

2.3.8 Preform semi-empirical methods

Using the SMILES notion provided in the generated dataset output a custom Pipeline Pilot protocol was written that would take the SMILES entry and convert it to an atomistic description. Once the data was converted to an atomistic description a semi-empirical density functional theory calculation was conducted. Pipeline Pilot is a powerful tool capable of manipulating and analyzing large quantities of scientific data and is provided by Dassault Systems. Over 30,000 molecular species were processed through the automated protocol to generate the thermodynamics data for this study.

The semi-empirical model that was implemented in the automated script was based on the Materials Studio provided VAMP software package[9,160]. Geometry optimization was conducted with a diatomic differential overlap (NDDO) and PM6 Hamiltonian, Auto multiplicity, and a spin state unrestricted Hartree-Fock (UHF), restricted Hartree-Fock (RHF), or annihilated unrestricted Hartree-Fock (A-UHF). Several spin states were tested based on convergence. A Paulay/IIS convergence scheme with a convergence energy tolerance of $2x10{-}4$. The thermodynamics information and total dipole moment were output.

The pipeline pilot script conducted a series of preparation steps prior to the DFT calculation. After data was read using SMILES format the SMILES was checked for consistency, followed by making and cleaning of the molecule. The molecule was generated from the SMILES, centered, hydrogens were added, and the molecule was cleaned. The cleaning step conducted a quick empirical elastic relaxation of the structure to refine the initial geometry. The structure was provided to a programmed series of VAMP calculations starting with the most rigorous spin state.

2.3.9 Calculate Thermodynamic Properties

Once the quantities of interest for each molecular species had been calculated using the semi-empirical methods model, the thermodynamics for each reaction were determined. For each molecular species, the entropy, enthalpy, and dipole moments were provided at the standard state at room temperature. The Gibbs free energy for each reaction was calculated using Equation 18. Any equation that exceeded +/−5 eV was eliminated due to stability concerns. Once this part of the pipeline was completed the equations were deemed safe to publish and for engineers to interpret.

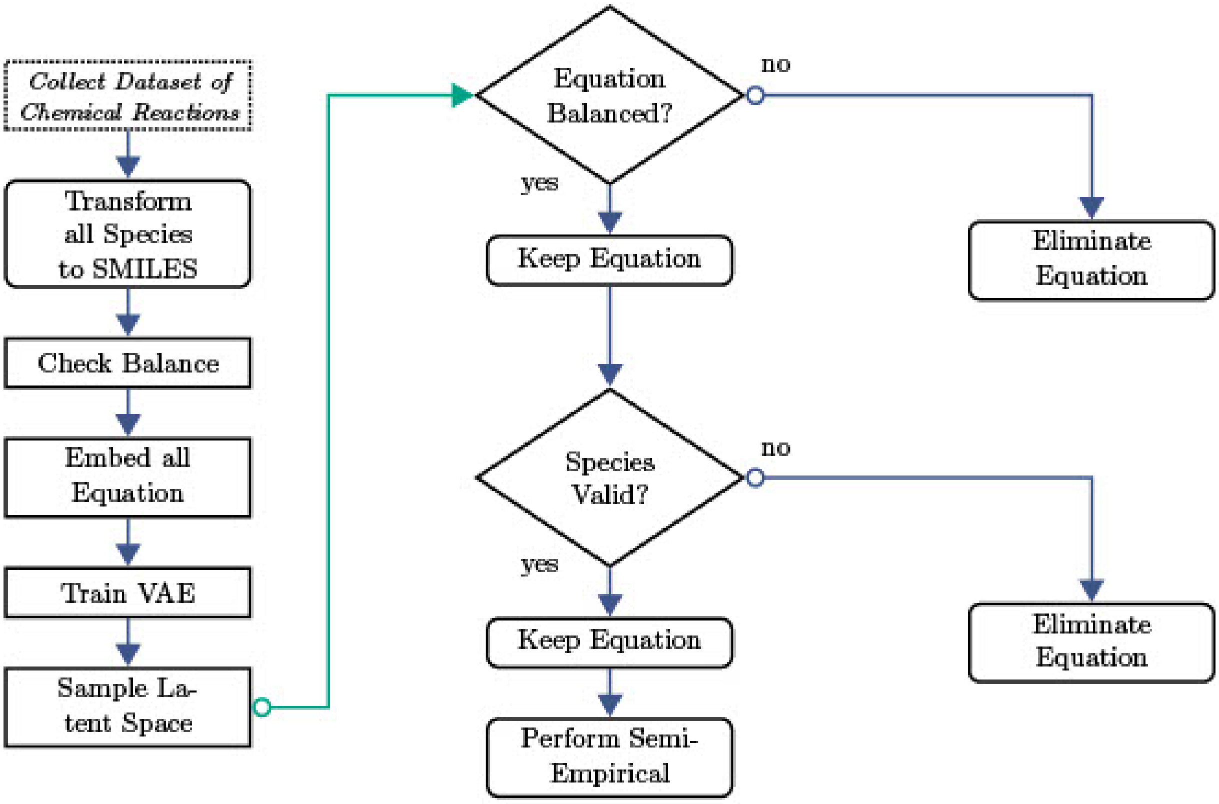


Figure 13: Flow diagram of the data collection, training, and validation steps taken by AGoRaS to generate synthetic data. The generated equation is checked for both balance and existence. The semi-empirical calculation receives a SMILES descriptor and conducts an independent series of processes to generate a clean atomistic description before the calculation of thermodynamics data.

2.4 Study 2 Results

Study one utilized the probabilistic sampling of the latent space by VAE to explore the solution space for a dataset of gas-phase reactions. The model VAE model that was created was named AGoRaS. The VAE procedure employed is illustrated in subfigure 9A. Once the VAE is trained the new reactions can be generated by sampling the latent space, as illustrated in subfigure 9B. The only information given to AGoRaS is the encoded SMILES string and the only output is an encoded SMILES string, in which hydrogen atoms are implicit. The VAE approach can be thought of as a custom compression technique for these chemical reactions. Moreover, the latent space can be thought of as the memory of artificial intelligence. While it is not apparent by looking at the weights of the system, artificial intelligence is empirically learning about the underlying physics and chemistry of the process.

At the very essence, this is why it is necessary to remove bias. With bias present, there is a skew in the memory of the network. A unique approach applied in this research is that instead of each latent variable being encoded with a discrete value, as with a traditional autoencoder, instead, the latent variable has an associated probability distribution. Once trained, the probability distributions of each variable can be sampled to generate new output features or chemical equations. By using continuous values instead of a discrete value it forces a smooth latent space representation of the data, Gaussian in nature. In the case of AGoRaS, each SMILES character is represented uniquely as a digit and the entire equation string is encoded. By sampling different points along the latent variable Gaussian curves, a new reaction can be generated. While this new reaction is not guaranteed to be balanced, it is necessary to check for compliance. Once trained, AGoRaS can generate new reactions quickly with the gained knowledge of the input dataset.

The steps outlined in the methodology section were used to validate that the equations were balanced and that the SMILES species represented physically possible chemical species. Figure 13 illustrates a simplified flow chart of the steps needed to accomplish the training and validation of AGoRaS. The semi-empirical modeling technique used the generated SMILES notation that is converted to an atomistic description to predict the thermodynamic properties. To allow for the independent validation of the network, it was decided to use an easily accessible dataset for training. To that end, the NIST chemical kinetics database was selected based on being open access and well-documented[143]. Since the dataset is comprised of solely published reactions, it was determined that it would make an excellent case study of the ability of AGoRaS to generate a wide variety of equations from a sparse dataset. After data collection and cleaning a core number of 7000 chemical reactions, this resulted in ~2000 unique gas-phase molecular species.

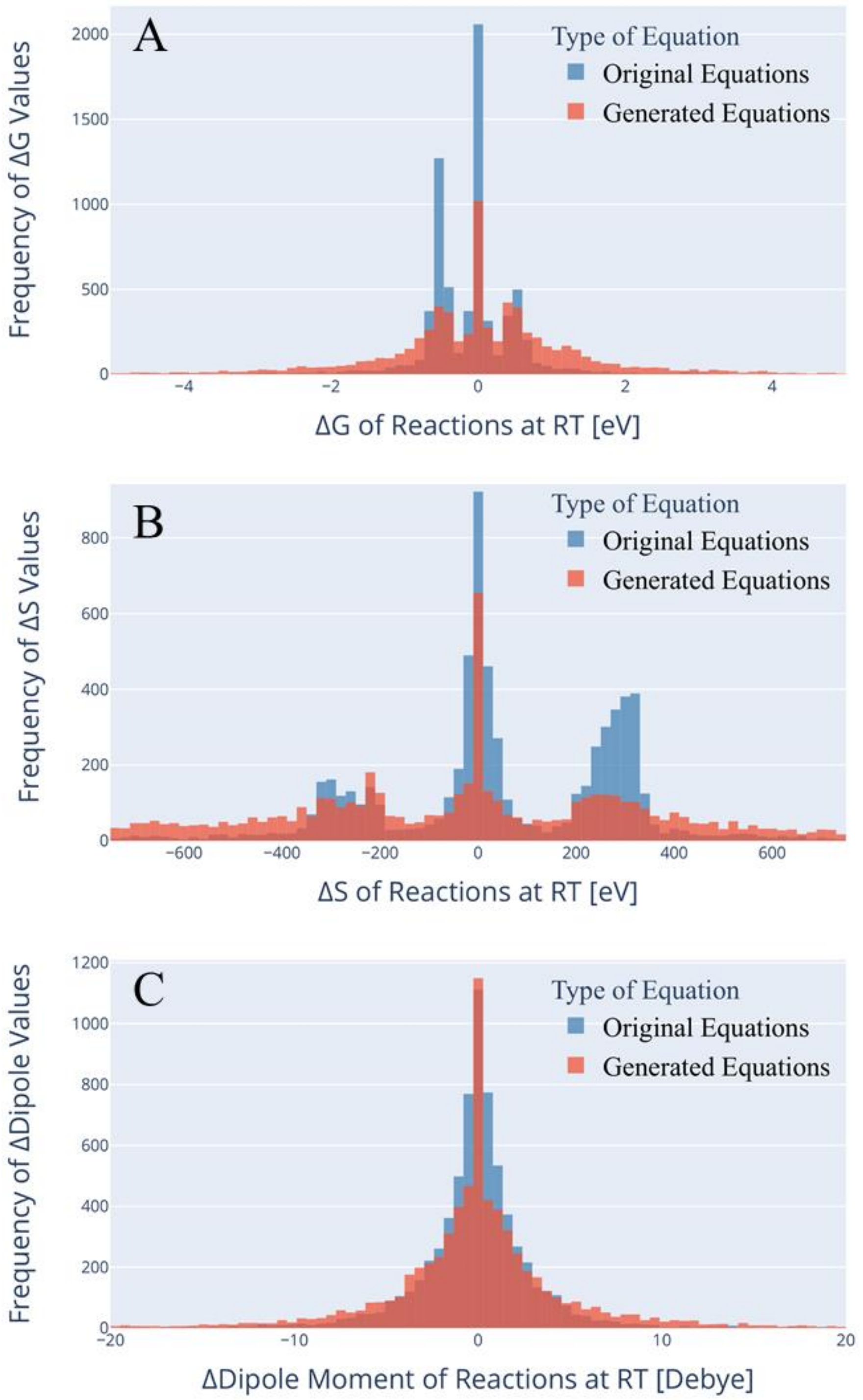


Figure 14: Histograms of the thermodynamics values comparing the training data (original equations) and the generated data (generated equations). Subfigure A is the Gibbs energy associated with Eq. 25. Subfigure B is the difference in entropy between the product and reactant molecular species. Subfigure C is the difference between the product and reactant dipole moment

Once the original dataset had been created and cleaned the VAE was trained, and the latent space was sampled. Sampling was stopped once 7,000,000 valid equations had been disseminated. This was selected as an arbitrary stopping criterion of 1000 times the size of the original dataset. Among these new equations, AGoRaS was also able to generate ~20,000 new species, with both subspecies and completely new molecular species. The ability to predict new and larger molecular species owes to the success of this approach over other approaches, such as retrosynthesizing. All species were checked using the validation pipeline to ensure their uniqueness and stability. To apply utility to the generated dataset and check the stability of individual molecular species, the Gibbs free energy of each molecular species, along with the overall reaction energy, were determined at the standard state using a semi-empirical computational technique[9,160] Furthermore, the difference in the dipole moments of the overall reactions was also predicted. The Gibbs free energy was selected because it is a good quantitative measure of the thermodynamics of the reaction. The dipole moment was also selected because of its thermodynamic relationship and application to electrocatalysis. To calculate the Gibbs free energy of each reaction, the individual molecular species' entropy and enthalpies were calculated, where the Gibbs energy of the reaction is defined in Equation 25, where the Gibbs energy of the reactant and products are based on the following relationship:

$$\Delta G^0 = \Delta H^0 + T\Delta S^0 \qquad \text{Eq. 24}$$

Here the change in enthalpy (ΔH) is based on the total energy of the semi-empirical calculation that is calibrated to the standard state in the semi-empirical calculation. The zero-point energy in this case is encapsulated within the enthalpy. The change in entropy (ΔS) is based on the vibrational frequencies, moments of inertia of the molecule, and its symmetry number. Similarly, the difference in entropy is calculated between the product and reactant species as follows in equation 26:

$$\Delta S^0_{reaction} = \Delta S^0_{product} - \Delta S^0_{reactant} \qquad \text{Eq. 25}$$

where the entropy terms ($\Delta S^0_{p,r}$) are corrected for the standard state. While this metric is encompassed in the overall Gibbs energy, it provides quantitative values for the amount of energy that is contributing to vibrational energy. Moreover, it provides a quantitative measure of the size of the molecules. While more advanced, first principle DFT calculations are potentially desired in this application, the computational expense of 7,000,000 simulations outweighs the accuracy required for this study.

The assumed approach will be to conduct a higher fidelity modeling approach for promising reactions when applied to practice. Due to the reporting bias in the original dataset, the selection of the best reaction, or thermodynamically favorable reactions (±1 eV) was known to be skewed towards reactions with low reaction energies. This was assumed to generate a bias in the network. However, it was found that generated equations ~97% fell within a ΔG_0 of ±5 eV. This is important for the practical use cases of AGoRAS, as it shows that even without providing any context on thermodynamic properties during training the network can produce stable results over a much larger range than the training data. The dipole moment was also derived from the semi-empirical calculation. The dipole moment is calculated based on the partial charge of the atoms and the position of the atom. The following expression is used in the calculation shown as equation 27:

$$\mu = \sum_{a=1}^{N} q_a r_a \qquad \text{Eq. 26}$$

where N is the total number of atoms, $q_{a,i}$ is the partial charge of the atom a, and $r_{a,i}$ is the position vector of the atom a. In this study, the difference between the product and reactant dipole moments was calculated using the following equation 28:

$$\Delta\mu = \mu_{products} - \mu_{reactants} \qquad \text{Eq. 27}$$

To visually and statistically compare trained and generated datasets of vastly different sizes, a random sample of 7000 equations was subselcted out of the generated dataset. An overlapping comparison between datasets was performed to confirm a correlation between the two datasets. This comparison is shown in figure 14, where A is a plot of the Gibbs free energy, B is a plot of the entropy, and C is a plot of the dipole moment. In all of these plots, the training dataset is pictured in blue in the background with the generated dataset in the foreground in red. The y-axis in all the plots is the frequency of occurrence and the x-axis is the associated thermodynamic property. It should be noted that 7,000,000 reaction equations were generated and this is a flat random selection of those generated.

In subfigure plots 14A and 14B, there is a noted tri-modal distribution, which is a result of the tri-modal distribution of the training data. This is evidence of the inherited bias in the training data that is being used. The AGoRaS is picking up on this knowledge as it generates a similar tri-modal distribution. The important region is the regions between and outer peaks of the input data. The AGoRaS can fill in these regions, which are either new or intermediate reactions. This generation ability is not based on retrosynthesizing previous reactions, but rather on the knowledge it has gained during training of the latent space variables. It can be shown if all seven million generated reactions are plotted, that the distribution becomes more continuous. Another interesting finding is apparent in the entropy plot of subfigure 14B. It is noted in this figure that AGoRaS has a flat distribution below −400 eV and above 400 eV. As seen in figure 14, a negative difference means the entropy is greater on the reactant side or smaller on the product side.

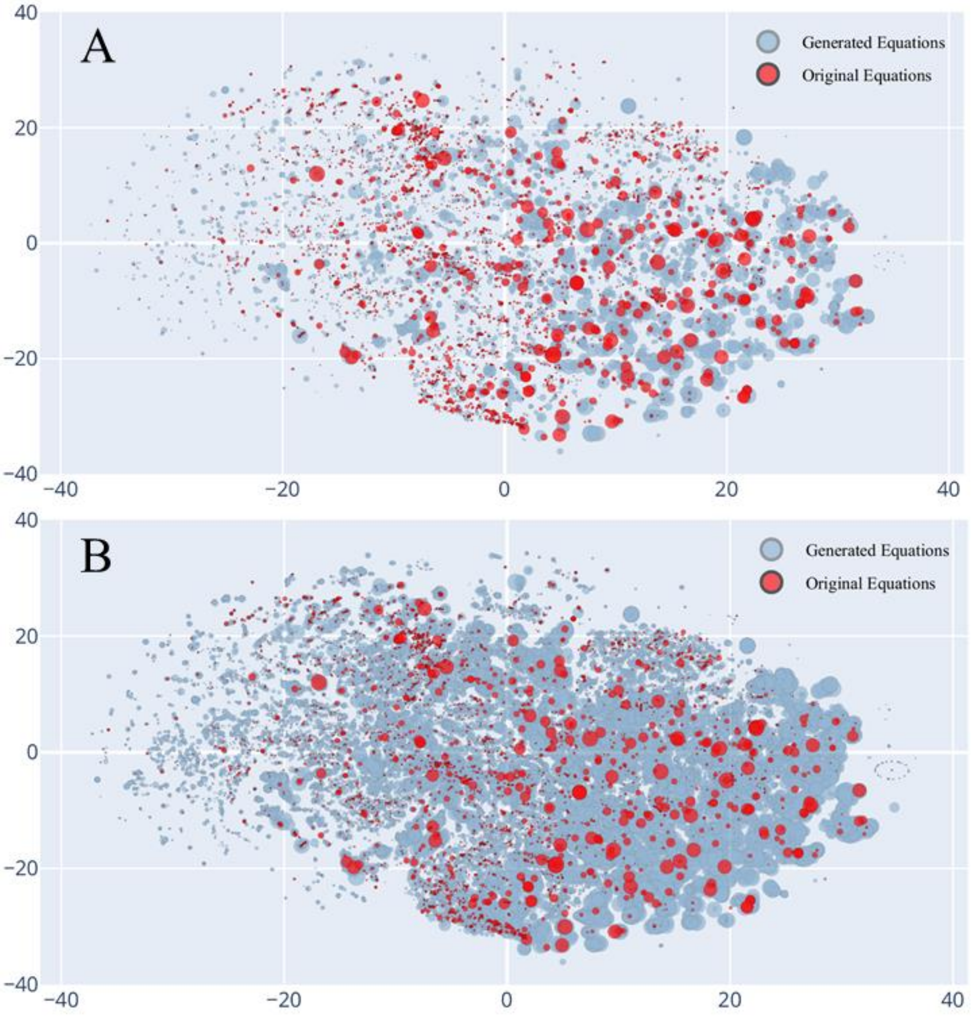


Figure 15: T-SNE plot of the training dataset (original equations) and the generated dataset (generated equations). The size of the sphere is proportional to ΔG at standard conditions. Subfigure A all 7000 equations from the training dataset and randomly sampled 7000 equations from the generated dataset. Subfigure B 70,000 equations from the generated dataset.

A thorough inspection of the generated dataset confirms that AGoRaS is generating larger molecules species on the reactant side. The size of these new species is beyond the size (number of atoms) of the input molecules species in the training dataset. This is a significant advantage of this approach, the ability to generate larger molecular species. Likewise, in subfigure 14A, there is bias arising from the training data, but the AGoRaS can fill in between the trimodal peaks and extend beyond. It is also able to generate larger species on the product size as demonstrated with large positive values in subfigure 14B. Subfigure 14C is a plot of the difference in dipole defined by Equation 28. From a molecular point of view, this has the opposite trend of the entropy, as the molecule gets larger (increased number of atoms) it is more likely the charge will be neutral for the molecule and have a corresponding negligible dipole moment. Therefore, as seen in subfigure

14C, both the original and generated datasets have a single peak near zero. This means most molecular species have near-zero dipole moments, as expected. Moreover, the generated dataset contains some cases which have large negative differences in dipole moment. This means the reactants have a larger dipole moment than the products. This is significant because, based on Equation 27, the position and charge of the atoms are critical to the dipole. This alludes to the fact that the AGoRaS has the ability to place atoms with ionic bonding tendencies further out from the center of charge, resulting in increased dipole moment. More succinctly said, AGoRaS is placing atoms beyond the radii provided in the training data.

An important aspect of VAEs and AGoRaS is the ability to map the latent probabilistic solution space of the problem using unsupervised learning and then sampling that solution space to generate new data. The latent space is hard to visualize and conceptualized, especially as the size of the data being represented in it grows exponentially. As discussed in the previous section, it is the aim to demonstrate that the newly generated equations are filling in the solution space compared to the input and ultimately removing bias. Using a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot, as seen in figure 15, it was possible to get a two-dimensional representation of what was happening in the high-dimensional latent space. This is possible because t-sne plots use an unsupervised learning method of stochastic neighbor embedding to give high-dimensional data to a single point on a two-dimensional grid.

Figure 15 illustrates the t-SNE plot for AGoRaS. The blue circles represent the generated data set and the red circles represent the training data. The size of the sphere is proportional to the Gibbs free energy equivalent to Equation 25. The t-SNE algorithm groups data together based solely on their SMILES representations. After their placement on the t-SNE plot, their thermodynamic properties were applied to the radius of each point. Subfigure 15A is a plot of the original 7000 equation in the training dataset with a sub-selection of an equivalent number of points for the generated dataset. Subfigure 15B is a similar plot with an increased sampling of 70,000 generated equations. It is noted from these plots that generated data sets are not only filling between the training data but also extending out beyond the training data.

Another important application of AGoRaS is the ability to conduct targeted reaction searches. Researchers can utilize AGoRaS' ability to generate massive datasets comprised of new, unique reactions to search for targeted molecular species and reactions. To demonstrate this application, the original and generate datasets were searched for reactions containing CO_2 and reactions containing CH_4. The original dataset had approximately 150 unique reactions that contained CO_2, while the generated dataset had approximately 6000 reactions. For CH4, the training dataset had approximately 700 reactions and the generated dataset had approximately 91,000 reactions.

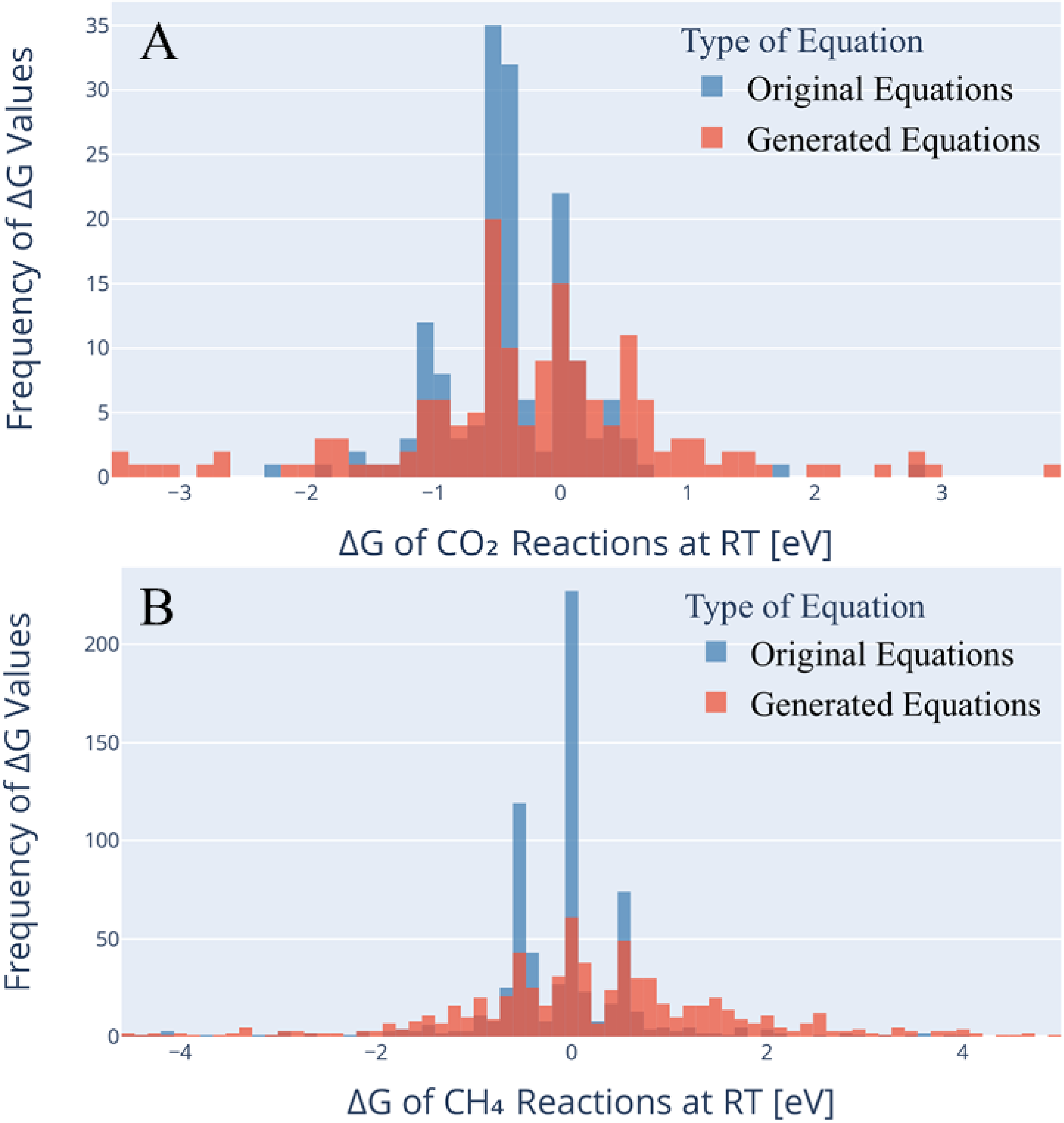


Figure 16: Histogram of the Gibbs energy at room temperature for specific molecular species. Subfigure A has 168 unique equations containing CO_2 selected from both the training dataset (original equations) and generated dataset (generated equations). Subfigure B has 688 unique equations containing CH_4 selected from both the training dataset and generated dataset.

The generated data provided 40 times as many reactions to examine further from a thermodynamic perspective. As discussed above, AGoRaS can generate intermediate molecular species and new molecular species beyond the descriptions of the training dataset. Figure 16 is a histogram of the selection of both CO_2 and CH_4 for the training dataset and the generated dataset. The selection for both cases was down-selected to Gibbs reaction energies, ranging between ΔG ±

5 eV. Again, the network can avoid the bias of the training dataset demonstrating the utility of this approach compared to other approaches. Table 4 is a summary of a select number of these reactions that demonstrate the complexity of the reactions that were discovered. Note the equations displayed in Table 4 were not part of the training data.

SMILES EQUATION	CHEMICAL EQUATION	ΔG^o [eV]
[O]=[C]=[O] + 2[H][H] -> [H][C]([H])=[O] + [H][O][H]	$CO_2 + 2\ H_2$ -> $CH_2O + H_2O$	0.204
1[H][C] + 2H][O][H] -> 4[H][H] + 1[O]=[C]=[O]	$1\ CH_4 + 2\ H_2O$ -> $4\ H_2 + 1\ CO_2$	0.047
[H][O][C]([H])([H])[C](=[O])[O][C]([H])([H])[H] -> [H][O][C]([H])([H])[C]([H])([H])[H] + [O]=[C]=[O]	$C_3H_6O_3$ -> $C_2H_6O + CO_2$	-0.486
The rightmost column is the associated predicted Gibbs energy from the semi-empirical calculation.		

Table 4: Selected results from the AGoRaS network that illustrates the SMILES notation and equivalent chemical equation representation.

Study 3: Synthetic Generation of Quantum Materials

The third study of this dissertation shows the practicality and flexibility of the AGoRaS network's methodology on other real-world problems such as the creation of organic materials for quantum applications. The objective is to apply almost the same steps and logic seen in study one, but instead of the outcome being balanced chemical reactions, it will instead be compositions of a single complex material. This will allow not only for a direct pool of materials to choose from for researchers but also can serve as a large and robust dataset for the training of other machine learning algorithms.

Study three proves that VAE can be used to overcome the problem of limited material availability in the field of quantum materials, specifically in single-photon source materials. By using the carefully curated dataset created by Beard et al. An AGoRaS network can be trained on materials that only have the appropriate physical and chemical properties of the single-photon source material. This will allow for the rapid generation of materials fitting a specific criterion needed for fields utilizing quantum materials. As proven by Zakutayey et al, this increase in available materials will greatly improve available machine learning algorithms. More importantly, this material generation will allow for significant time and cost savings for experimental studies or other data collection methods.

This aim will be accomplished by using the ability of VAEs to generate a near-continuous and infinite space of materials that share the desired characteristics. As in the case of the AGoRaS network's generation of chemical equations, the generation of SPS materials will also need to be rigorously checked using a similar methodology to the one laid out by Beard et al. As in the previous study a pipeline will be created to accomplish the data collection, network training, network testing, new material generation, and new material testing. The purpose of setting up a modular data pipeline like this is to allow non-data scientists to quickly apply the same methodology to any type of problem. The use of data pipelines with modular easily replaceable pieces and their effect on code quality and robustness is a well-studied topic[96–98]. This modular approach to taking a lot of the data cleaning out of the hands of the user while also allowing for greater flexibility will allow the AGoRaS methodology to be more accessible to a non-data scientist. It will instead allow chemists and engineers to utilize the network using their data. Which has historically been a problem, where the lack of data science skills can hold back a field from utilizing the power of artificial intelligence[99].

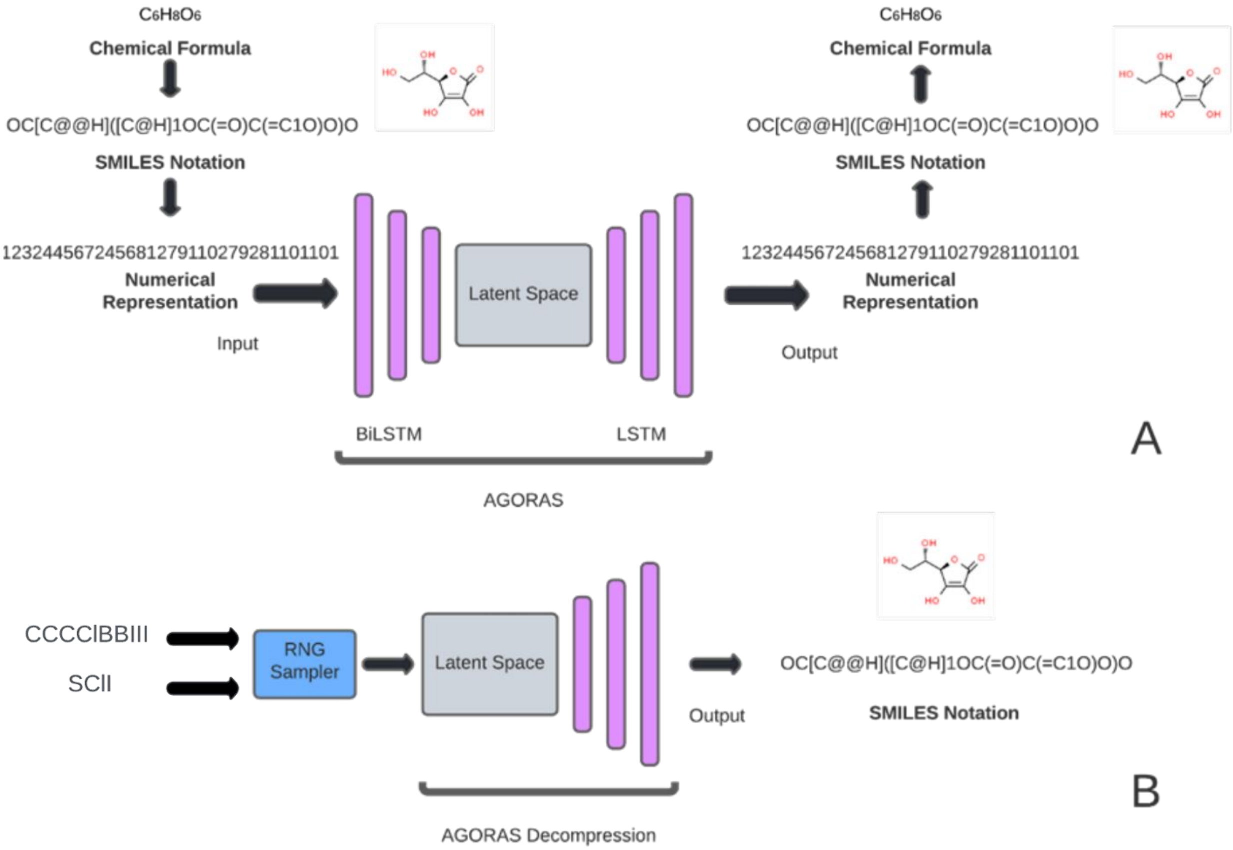


Figure 17: Subfigure A is an illustration of the AGoRaS network illustrating how a chemical species is encoded and used as training data for the network. Chemical database information is compressed and decompressed to form a high-dimensional latent space as with the original AGoRaS network. Subfigure B is an illustration of how the trained latent space can be sampled to generate new single-photon emitting materials.

3.1 Study 3 Significance

As with study two, study three also looks to help solve the biased problem that exists in chemical datasets comprised of research articles and patents[101,161]. With study three, however, the goal is changed from general chemical equations to specific species with targeted uses. In this case, quantum materials, particularly single-photon source (SPS) materials also sometimes referred to as UV/vis materials. These SPS materials have very few datasets that are already compiled and of those most are particularly small[162,163]. The two-part problem of small amounts of data being available, as well as limited open source datasets, leads SPS materials to be an excellent choice to prove the applicability of AGoRaS to generate large datasets that can be used for the training of machine learning algorithms. This study utilizes the same AGoRaS designed for study two to synthetically generate SPS materials that are less biased and more robust than the ones currently available in literature and patents.

SPS materials have a varying array of uses, and therefore require a huge amount of different materials with different emission spectra to meet those needs[164–167]. They are used in frequency-domain optical storage[167], quantum communications, optical quantum information processing, and metorology[166], as well as quantum sensing[168] and dipole gates[169]. They are also used for checking water and ground water content for contaminants[42,170] to name just a few. This huge range of different use cases requires a wide variety of different materials that can satisfy the different single spectra emissions requirements of each use case. As of today, there exists no comprehensive database that covers the range of single-photon emitting materials needed for all these use cases.

It is hypothesized that a VAE sharing the same structure as AGoRaS can create a material compression intelligence via the latent space representation of a currently available SPS material database. This latent representation can be sampled at different points to generate new SPS materials with strong peaks. The latent space can be thought of as a representation of the compressed structural and chemical information inherent in the species being used to train the network. The network will be able to learn the structural and chemical similarities between the SPS materials and populate the latent space with them. This is the power of a VAE type of network, it does not look at and represent data in 2 or 3 dimensions but instead in n number of dimensions.

While representing the data in a higher dimension offers a VAE network a unique advantage, the real advantage of a VAE is that instead of each input node having a unique value in these high dimensions they are fitted to a probabilistic distribution. Most often this distribution is in a multivariate gaussian distribution. This means that each of the input nodes has a range of possible inputs.

Study three proves that AGoRaS can be reutilized for a targeted approach to a specific type of problem in the materials field. This will allow not only for large amounts of new materials to be discovered but also lead to improved machine learning algorithms. It will also offer cost and time savings over DFT and experimental studies. As well as allowing the field of quantum materials to catch up to other fields in the adaptation of machine learning algorithms. The ability of this network to generate a near-continuous and infinite space of new reactions will allow researchers to speed up the research and deployment of cutting-edge machine learning techniques.

3.2 Study 3 Background

Just like in study two, the field of SPS material is filled with “garbage”. In so much as the data currently available in the literature is not often suitable for machine learning algorithms. Zakutayev et al. showed that the lack of sufficiently large and diverse datasets is a huge hindrance to the development of advanced machine learning algorithms in the field of material science[171]. Their study drives home the point of the importance of large datasets in the training of machine learning algorithms, as well as the ability of these machine learning algorithms to predict the structure, stability, and properties of various types of materials. Importantly, Zakutayey et al. show how advanced machine learning algorithms that exist already can be adopted quickly and easily to material science problems, as long as a robust and extensive dataset exist for their training.

In UV/vis research and its application in SPS materials, the work by Beard et al. stands out for their thorough and exhaustive collection of available materials and corresponding relevant calculations[162]. They looked at over 400,000 scientific documents to extract a database of just over 8,000 unique compounds. This highlights the difficulties that exist in creating a database for quantum materials, even when using state-of-the-art tools such as ChemDataExtractor as was used in the work by Beard et al[172]. The small return of materials from such an exhaustive search has several root causes. The first is the wide variety of formatting among different scientific journals. The second is the discrepancies that exist within the tools being used, such as RDKIT's various APIs. A third problem is that there is no one set of ground truth rules to use when representing a material using the smile's notation, which can lead to some errors when phrasing materials. Despite these obstacles, the database created by Beard et al. represents the most complete UV/vis material dataset to date.

Single photon generational materials are a requirement for applications of quantum communication, quantum computing, quantum information, and quantum precious metrology. SPS materials are used in a wide range of applications, from quantum computing materials to remote sensing, and dipole gates. Remote sensing has gained a lot of traction in the last few years as technology has improved around quantum materials. These remote sensing methods have been used for monitoring contaminants in water, monitoring air quality trends, dissolved nutrients in surface water, and many other advanced techniques[168,170,173–176]. Spangenberg et al. show how using quantum materials could be combined to detect relative concentrations of mixtures within the water in real time[170]. Meanwhile, Fei et al. showed how with the right combination of machine learning algorithm and SPS material this same contamination monitoring of ground water. However, even Fei et al. work could only find 1665 materials to be used, highlighting the problem with small datasets[177]. In the work, by Mamede et al. it is continued to be shown how machine learning can be utilized with quantum materials. Their work focuses on finding the UV/vis adsorption spectrum of organic molecules using the fingerprints generated from 2D chemical structures. They were able to achieve a sample size of ~75 k molecules by using only chemical structures[173].

Finding different quantum materials that work for a varying array of conditions is a very active area of research. With De Leonardis et al. showed that by using quantum materials they can overcome the challenges of phase-matching conditions between an optical pump, signal, and idler photon pairs when operating short-wavelength macroing resonators[174]. Meanwhile, Richter et al. showed that quantum materials can help with remote sensing of tropospheric compositions from space[175]. These finds lead to an increasing need for more quantum materials that can help advance many different fields.

3.2.1 Quantum Materials

Quantum material is a term from condensed matter physics that encompasses all materials whose essential properties cannot be described in terms of semiclassical particles and low-level quantum mechanics[178]. These are generally materials where non-generic quantum effects are linked to electronic properties. They can also be materials that have some type of electronic order or electronic correlations. Quantum materials often have properties that have no equivalent in the macroscopic material world. These properties can often be extremely difficult to understand. Some of these properties include quantum entanglement and quantum fluctuations[178,179].

3.2.2 Single Optical Photon Source

A material that emits light at a single particle or photon is known as a single-photon source (SPS). These materials are a subset of quantum materials and are interested in numerous scientific fields. A material is considered an ideal SPS if it meets two conditions, the first is if it produces single-photon states with 100% probability. The second condition is if it produces multi-photon states with 0% probability[180]. In general, if a system has an electric field amplitude that is distributed over a narrow bandwidth, this will allow Fock states to be studied[11]. This is how single-photon sources can get around the Heisenberg uncertainty principle. Photons from these sources exhibit quantum mechanical characteristics, including photon antibunching[181]. This means that the time between two consecutive photons can never be less than some specific value. Typically, an excited electron stays in a higher state for a few nanoseconds and then returns to the ground state. It is during this return to the ground state that the photon releases energy in the form of radiation[11].

A simplified SPS material could be explained by considering a single atom that contains exactly two energy levels and a single electron. Therefore, the electron could only be in either the ground state (lower energy) or the excited state (higher energy). When the electron transitions from the excited state down to the ground state the atom would emit a photon with an energy level equal to the total energy difference between the atomic state of the excited level and the ground level. A schematic of this can be seen in figure 19. Once the electron has entered the ground state, it would require something to re-excite it to a higher energy level in order to be able to produce another photon. In this way, it is clear that the atom could not produce two photons at once[182,183]. Some examples of SPS materials are generally broken down into two types of materials. The first is single emitters include single atoms, ions, and molecules. The other is solid-state emitters such as quantum dogs, color centers, and carbon nanotubes. The second category of material is considered on-demand materials, as they can be directed to emit a photon on light as needed[184].

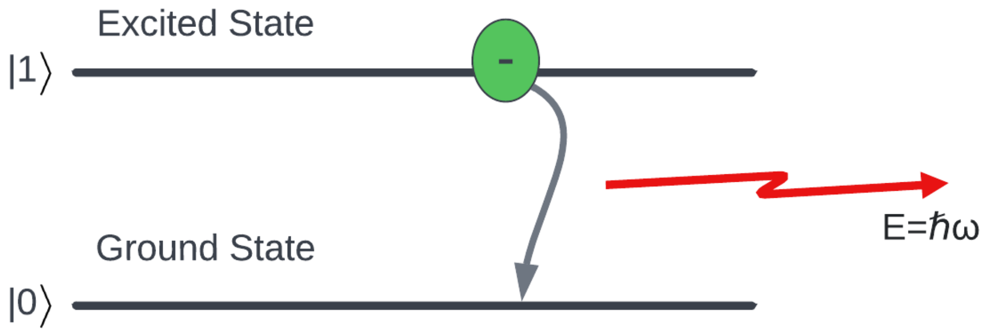


Figure 18: Electronic levels of a simplified atom with one ground state and one excited state and a single electron. As the electron falls to the ground state a photon is emitted. Where |1> indicates the excited state and |0 > indicates the ground state.

3.2.3 Ultraviolet-visible Spectroscopy

Adsorption spectroscopy or reflectance spectroscopy known as ultraviolet-visible spectroscopy(UV/Vis) studies light in the pat of ultraviolet and adjacent visible regions of the electromagnetic spectrum. The perceived colors seen in the visible light spectrum due to absorption or reflectance effects the color of the chemicals involved. Within this region of the visible light spectrum, molecules and atoms undergo electronic transitions between states. With fluorescence studying electrons moving from an excited state to a ground state. Meanwhile, absorption is the study of electrons moving from a base state to an excited state.

3.2.4 Semi-Empirical Quantum Chemistry Method

Semi-empirical quantum chemistry methods are all based on approximations for the determination of the wave function and the energy of a quantum many-body system in a stationary state[179]. This approximation used most often is the Hartree-Fock formalism[185]. A popular semi-empirical method is the VAMP package. VAMP is a fast reliable system that utilizes the MINDO/3, MNDO, AMI, and PM3 semi-empirical methods[186]. It can be used for the evaluation of many chemical and physical properties of organic and inorganic systems. The VAMP software is commonly used before other methods such as density functional theory to find good starting geometry and properties. By using a semi-empirical approach, the throughput of the number of predictions that can be executed is significantly increased. Evaluating millions of small molecules using a semi-empirical approach is possible using facilities at WVU, while a DFT approach would be time prohibitive. The throughput of the number of species that can be evaluated outweighs the accuracy of the solution. The approach was to use the semi-empirical method to evaluate all species in the design spaces and then sub select good performing material and conduct high fidelity simulation, such as TDDFT simulations.

3.2.5 Quantum Sensing

Quantum sensing is the act of using a quantum object to measure a physical quantity through the use of quantum systems, quantum properties, or quantum phenomena. When describing quantum sensing it is typical to describe three main use cases. The first is when a Quantum object is being used to measure a physical object. The second is the use of quantum coherence to measure a physical quantity. The third is to improve the precision or sensitivity of a measurement through the use of quantum entanglement[187]. Typically, when using SPS materials this third use case is the most common. This third use case is considered the truly quantum definition out of the three. It allows for operations to take place at nanoscales that are not possible though the use of classical sensors.

3.2.6 Time-Dependent Density Functional Theory

Time-dependent density functional theory (TDDFT) is a type of quantum mechanical theory with uses in chemistry and physics. Its usages are mainly in investigating the properties and dynamics of many-body systems with time-dependent potentials[188]. Time-dependent potentials are commonly things like electric or magnetic fields. By looking at these many-bodied systems in time it is possible to see the effects of such fields on various molecules and solids[189]. This allows for the extraction of features such as excitation energies, frequency-dependent response properties, and photo absorption spectra. TDDFT is an extension of the widely used density functional theory (DFT). The computational and conceptual fundamentals of these techniques are the same. However, TDDFT systems tend to be much more complex due to the time-dependent effective potential of any given moment being dependent on the values of all previous times[190]. The TDDFT software used in this study was DMOL3.

As with the semi-empirical methods predicting the visible and near-visible UV spectra with TDDFT involves calculations of excitations between discrete electronic levels in a system. These evaluations are done by calculating transition dipole moments between the sates involved in the various excitations. This allows the calculation of the adsorption peaks in the spectrum due to excitation energies[188,189,191]. A formal definition of TDDFT is outside the scope of this study but the work by Barone et al.[192] does and excellent job of formally defining the problem. Some of the relevant equations for this study can be seen in equations 29-31. The eigenvalue problem shown in equation 29 can be used to calculate the poles of the frequency-dependent dynamic polarization tensor and its response[188,191].

$$Q\boldsymbol{F}_I = \mathit{\Omega}_1^2\,\boldsymbol{F}_I \qquad \text{Eq. 28}$$

Where $\mathit{\Omega}_1$ are the excitation energies and $\boldsymbol{F}_I$ the eigenvectors that describe the multi-determinantal excited sates. The subscript I is the index of the electronic excitations. To calculate the response density for the converted excitation vector ($\boldsymbol{F}(I)$) when taken as being discretized on the real space grid can be used to calculate the dipole matrix elements and oscillator strengths using equation 29[191,193].

$$M_{I,x} = \sqrt{2} \int \rho \ \boldsymbol{F}(I)(\boldsymbol{r})\text{xd}\mathbf{r} \qquad \text{Eq. 29}$$

Which allows for the oscillator strength to be expressed as equation 30[191].

$$f_I = \frac{2}{3}\left(M_{1,x}^2 + M_{1,y}^2 + M_{1,z}^2\right) \qquad \text{Eq. 30}$$

In order to have a complete orbital basis set the oscillator strength has to satisfy the sum rule shown in equation 31. Which states that the number of electrons (N_e) is the number of electrons in the molecule[191].

$$\sum_I f_I = N_e \qquad \text{Eq. 31}$$

3.3 Study 3 Methodology

3.3.1 Processing the Database

The chemical species used in this study to train the network were taken from the Dataset created by Beard et al[162]. The dataset was downloaded in the json format with ~8,000 different species. The species were then further divided up to separate species with a single lambda and a dipole moment and species with multiple lambda peaks, and a dipole moment. Each species already had SMILES strings attached along with an IUPAC name. For validation, each IUPAC name was converted using RDKIT to verify that it matched the provided SMILES string. All species were read into a pandas dataframe with the important information as columns, such as SMILES string, lambda values, intensities, and dipole moments.

It was decided that for simplicity and reproducibility, the AGORAS network would only focus on the predictions of SMILES species and not on the associated properties. This was done because it allows the network to focus on learning the underlying physical and chemical structural patterns rather than also trying to extend the prediction to properties. The physical properties such as dipole moments, lambda values and intensity could be calculated using DFT.

Since the idea of this iteration of AGORAS was to generate new species it was decided to continue to use character-level embedding, due to its advantages over word-level embedding in the field of natural language processing[142,154]. This would allow for the generative network to not just use the same species it had been trained with but instead allow it to generate new species based on information gained in the training process. The embedding was done using TensorFlow's built-in embedding techniques and was based on a universal alphabet created from all the species[155]. This was done in part because of the work by Gaspar et al. who show how molecule embedding can mirror that of NLP embeddings. They show in their work how improvements can be made to the predictive power of machine learning models by formulating the inputs into sequence embeddings[156].

3.3.2 AGORAS Structure for Quantum Material

AGoRaS takes in a vector representation of length n, where n is the maximum length character-level representation of the longest SMILES string in the training dataset. This vector is then fed forward into a TensorFlow Embedding layer that projects the input into a higher dimensionality. This is an important step because the projection fits numeric values to a high dimensionality space removing the intrinsic values of the values themselves. This is done because in our 2D space the numeric values have no intrinsic value i.e., 5 is not greater than 6 they just represent two different characters. The projected vectors are then fed into a bidirectional LSTM (BiLSTM) layer with a recurrent dropout of 0.2. The mean and log variance is created from the output of the BiLSTM layer.

A sampling function is used to randomly sample this solution space based on the mean and log variance. The network then decodes the sampled solution space using a RepeatVector layer that is wrapped around the output of the latent space, thus turning it into a tensor vector that an LSTM layer could read. The RepeatVector is fed as an input into an LSTM layer, whose output is projected into a vector of length n. The output of this projection is what is used to calculate the loss of the network. AGoRaS uses a sequence-to-sequence style loss function common to variation autoencoders. The metric used for monitoring the network during training is also the standard kl loss. The network was trained for 500 epochs using a batch size of 25, an embedding dimensionality of 500, and a latent dimensionality of 350. The kl weight used was 0.1 and the activation function was a SoftMax function. The optimizer function was Adam and the learning rate was 1x10−5. This structure exactly mimics that of the AGORAS network for chemical reaction prediction other than the length of the input vector.

3.3.3 Training AGORAS for Quantum Material

Once the training data had been cleaned and embedded in a numerical format readable by a neural network architecture it is split into three sections. The training set, validation set, and test set where the training set was 70% of available data, the validation set was 20% of available data, and the test data was 10% of the data. Even though AGoRaS utilizes a generative technique it can still be validated using traditional methods. The way VAEs are validated is by the ability of the network to encode the validation set and decode it back to the original string construction with no loss of information. The sequence-to-sequence loss function was used during the training process. This allows for scoring of the reconstructed string versus the original string. This ability to encode and decode the original chemical equations with zero loss of information is typically indicative of a stable latent space. Do to the small data nature of this problem it is extremely important to validate as much as possible the stability of your latent space. To that end, after it had been proven that the network could reconstruct the test data, the remaining 90% were also tested again. Of course, since this data was used in training the network should be able to reconstruct them all. But it serves as a further validation of the stability of the latent space.

3.3.4 Autonomously Sample the Latent Representation for Quantum Material

Once a trained network has been created it is possible to construct a sampler that utilized real species to interface with the latent representation. This is shown in subfigure 17B where it is shown how two materials can be used to access the latent space and return a new species at some equidistant point between them. It is still possible to start sampling the different latent representations randomly using a randomized species and having the decoder part of the network construct species. The decoder takes the equidistant points between the two species and applies the learned weights to construct new equations. This directed sampling has a powerful advantage over the previous version. Instead of attempting to just continuously sample the latent space, it is now possible to sample the latent spaces in areas for which the decoded species have characteristics researchers might be interested in. Due to the probabilistic nature of the latent representations, it is possible to take an almost unlimited number of sample points and continue to generate new species. Of course, there would be diminishing returns on this as there are only so many chemically feasible species possible.

3.3.5 Validating Generated Species

The methodology for determining the chemical validity of species was extremely like that of the data cleaning process. The first step was to check duplicate species were eliminated. The second step was to check each species for chemical validity using RDKit, where any physically or chemically unstable species should be rejected by the software[151]. This is a common practice when using a neural network with generated SMILES species[134,140,145,146,148,157–159]. The ability of the network to generate valid chemical species helps to further prove that the latent space is stable and representative of the original dataset.

3.3.6 Preform Semi-Empirical Methods on Generated Species

Using the SMILES notion provided in the generated dataset output a custom Pipeline Pilot protocol was written that would take the SMILES entry and convert it to an atomistic description. Once the data was converted to an atomistic description a semi-empirical density functional theory calculation was conducted. Pipeline Pilot is a powerful tool capable of manipulating and analyzing large quantities of scientific data and is provided by Dassault Systems

The semi-empirical model that was implemented in the automated script was based on the Materials Studio provided VAMP software package[9,160]. Geometry optimization was conducted with a diatomic differential overlap (NDDO) and PM6 Hamiltonian, Auto multiplicity, and a spin state unrestricted Hartree-Fock (UHF), restricted Hartree-Fock (RHF), or annihilated unrestricted Hartree-Fock (A-UHF). Several spin states were tested based on convergence. A Paulay/IIS convergence scheme with a convergence energy tolerance of $2\text{x}10{-}4$. The thermodynamics information and total dipole moment were output.

The pipeline pilot script conducted a series of preparation steps prior to the DFT calculation. After data was read using SMILES format the SMILES was checked for consistency, followed by making and cleaning of the molecule. The molecule was generated from the SMILES, centered, hydrogens were added, and the molecule was cleaned. The cleaning step conducted a quick empirical elastic relaxation of the structure to refine the initial geometry. The structure was provided to a programmed series of VAMP calculations starting with the most rigorous spin state.

Using the VAMP software, the total formation energy of a structure can be determined. To do so it is necessary to determine the total energy of the system minus the total energy of each element. This can be seen in Equation 32 shown below. Where the formation energy represents the energy required to dissociate a structure into its individual components.

$$E_{form} = E_{tot} - \sum_{x} E_{tot}(x) \qquad \text{Eq. 32}$$

Using the VAMP software it is possible to also calculate the dipole moments at the same time as the total energy. VAMPs can determine the molecular wavefunction of a species, which can then be used to derive the dipole moment. This is done using the LCAO method of molecular orbitals rather than the standard MNDO Hamiltonian calculation. VAMP is also able to calculate accurate dipole moments using the Natural Atomic Orbital-Point Charge model for molecular electrostatic properties. Another function of the VAMP software is the ability to calculate molecular properties such as optical spectra. A summarized version of these steps can be seen in the flow chart in figure 16.

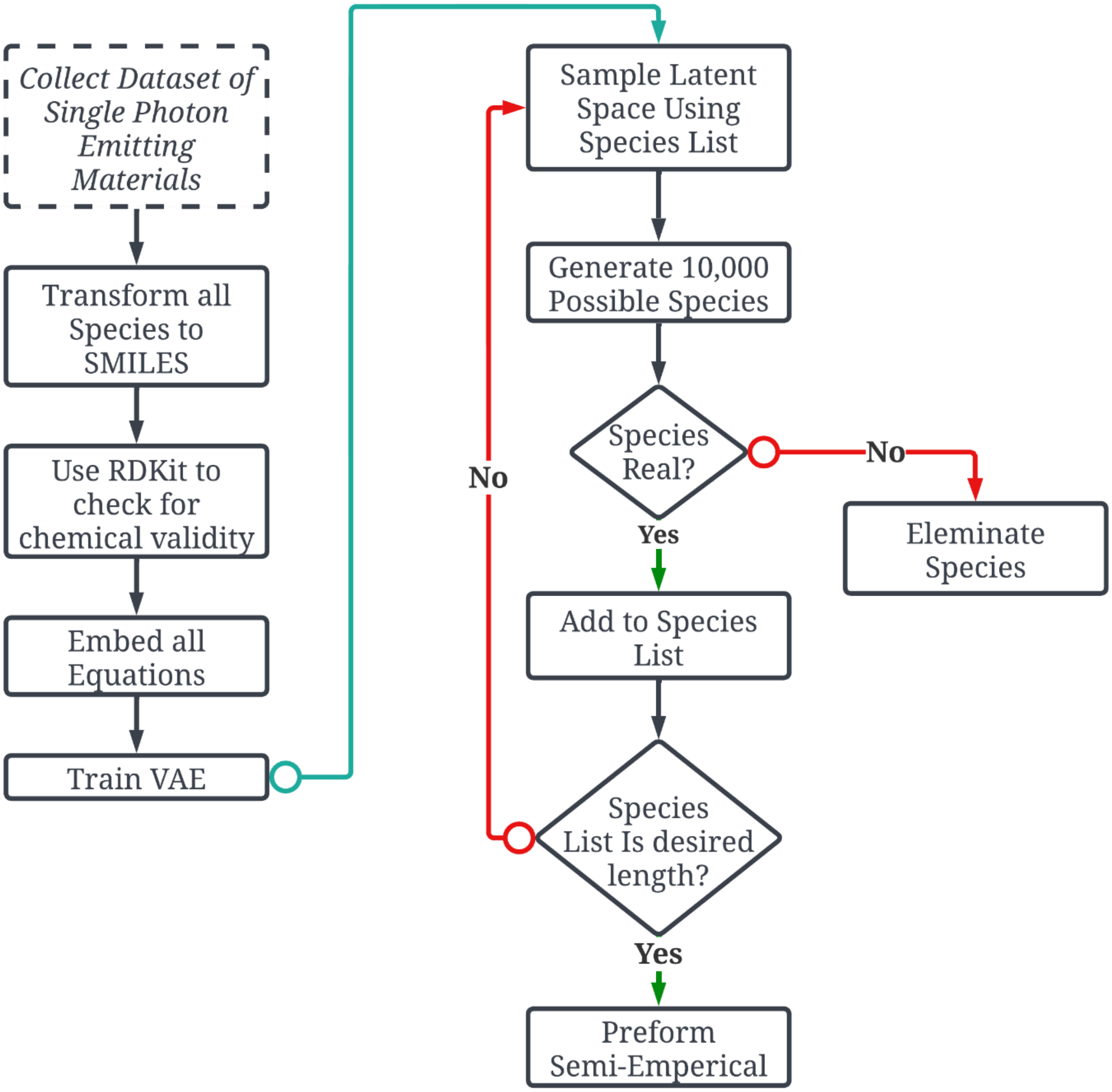

Collect Dataset of Single Photon Emitting Materials
Transform all Species to SMILES
Use RDKit to check for chemical validity
Embed all Equations
Train VAE
Sample Latent Space Using Species List
Generate 10,000 Possible Species
Species Real?
No
Eleminate Species
Yes
Add to Species List
Species List Is desired length?
No
Yes
Preform Semi-Emperical

Figure 19: Flow diagram of the data collection, training, and validation steps taken by AGoRaS to generate synthetic data for Single Photon Emitting Materials. After 10,000 new species have been created, they are checked for chemical feasibility. These new species are then added to the

original species list to help sample the latent space more effectively. The semi-empirical calculation receives a SMILES descriptor and conducts an independent series of calculations to validate the species as stable and, physically possible.

3.3.7 Compare Training Data with Generated Data

The above calculation for optical spectra, dipole moments, and total energy is done for both the original dataset and the generated data. To determine if our generated species offer a good representation of real-world conditions, these three properties needed to be compared and contrasted. In addition, this study also looks at the number of atoms for both the generated materials and the original materials. Due to the disparity of the dataset sizes a percentage normalized comparison between the larger generated dataset and the original dataset containing ~8,000 materials, was taken. It is then possible to compare the total number of atoms, dipole moments, optical spectra, and total energy directly between percent normalized datasets. This is done using histograms, to get a view of the distribution of values for each dataset.

3.3.8 Identify Promising Material

Once the dataset has been proven to contain realistic values and material distributions it is possible to sort the data for promising materials. The data is sorted by the three criteria already looked at previously. The aim is to identify 2 materials with strong dipoles and a single lambda frequency in the 500-600 nm range with a high intensity. These criteria are selected due to their direct interest to researchers in the field of quantum sensing. It was decided that using materials with direct practical applications would help make this work more inviting to a wider range of researchers.

3.3.9 Validate Material with TDDFT

To determine if these 2 materials warrant experimental consideration it is important to put them through a more rigorous form of calculation, TDDFT. The higher accuracy of the TDDFT method to that of the semi-empirical methods used earlier, allow for researchers to have greater confidence in the predicted value. This higher level of validation also allows for experimentation with these molecules to help identify what is causing these strong peaks and particular wavelengths. The molecule's emission spectra can be investigated by adding or removing elements to see the response in the strength and the peak wavelength.

3.4 Study 3 Results

Since study three was meant to be an extension of study two to show AGoRaS's versatility both networks utilized the same training methodology. However, study three utilized a slightly different approach to that of study two when sampling the latent space of the VAE. As well as for generating the synthetic dataset. Therefore, it has been named AGoRaS-Quantum to differentiate it from the original AGoRaS network. The reason for this difference is twofold, the first being the different nature of the desired outputs, and the second is the natural development of software where better tools are developed to help reach desired goals faster. In the original AGoRaS, a probabilistic sampling of the latent space was used to generate materials and continued to use that until we reached some arbitrary number of new equations.

In AGoRaS-Quantum to sample the latent space, two different species are chosen and embedded as a starting and end point in the latent space. A set number of equidistant points between them are then returned and decoded into possible species. This directed sampling has a powerful advantage over the previous version. Instead of attempting to just continuously sample the latent space, it is now possible to sample the latent spaces in areas for which the decoded species have characteristics researchers might be interested in. This improvement led to a slightly different methodology for generating the synthetic dataset. In study three after the network had produced 10,000 new species that were verified using RDKit, those species were appended to the list of species being used as starting and endpoints. This helped to stabilize the sampling of the latent space and increased the number of stable species being returned. It should be noted that for the generation of the synthetic dataset in this study that random species were selected as the starting and end species. This allows for the synthetic dataset to remain unbiased in its generation while still maintaining the chemical characteristics of the original dataset.

As in the previous study, the only information that the AGoRaS-Quantum network receives is an encoded SMILES sting and the only output is an encoded SMILES string, in which hydrogen atoms are implicit. This can be seen in subfigure 17B, where the AGoRaS-Quantum network is illustrated. In the previous study, the claim was made that the AGoRaS network could and indeed was learning the underlying physics and chemistry of the reaction processes. By targeting the AGoRaS-Quantum network to generate single molecules that had very distinct chemical properties it was possible to validate that this claim was true. Unless the AGoRaS-Quantum network was learning the chemical and physical combinations of atoms and structures that generated single photon emitting materials then the species being created would fail to have those distinct characteristics. Instead, it would have just learned what characters to put together to make some valid SMILES string.

Just as in the previous study each character of the smiles string is represented in the latent space as a probability distribution rather than as a discrete value. This allows for the trained network to sample this distribution for the desired number of points to generate the corresponding number of new species. This is possible due to the unique nature of the VAE network over that of a traditional AE network. As previously discussed, a VAE’s latent space is a smooth representation of data due to the continuous nature of the input. However, what this does is creates a latent space in which points that are in proximity to each other are similar in nature. This is not the case in an AE network's latent space, where inputs are represented as discrete values.

The steps outlined in the methodology section were used to gather the training dataset, transform it into the correct format, train the network, sample the latent space, and validate the chemical and physical validity of the generated species. A simplified flow chart is shown in figure 19 that outlines the straightforward and easily repeatable steps needed to accomplish the training and validation of the AGoRaS-Quantum network. It can be seen that figure 19 is remarkably similar to figure 13 which outlines the training and validation of the original AGoRaS network. This is, of course, intentional, it provides an easy-to-understand and recreates flow for researchers looking to recreate this experiment. As well as showing how the same steps and validation that were used for a network tuned to chemical reactions can easily be applied to a network tuned to a different purpose. It also serves to easily highlight any additions or improvements made to the flow.

Just as in the previous study the semi-empirical modeling technique used the generated SMILES notation that is converted to an atomistic description to predict the chemical properties. In an effort to allow others to easily recreate this study, it was decided to use the database created by Beard et al. who have already collected suitable materials from various sources and made them easily accessible in several different formats[162]. They used a combination of a literature search of experimental data, RDKit, and TDDFT to generate the original corpus of data used in this experiment. As with the previous study, since all of the species were extracted from the literature it was determined that this would be a good comparison dataset to show the ability of AGoRaS-Quantum to generate a wider-ranging and more balanced dataset. After the collection of data and phrasing using RDKit, we were left with 8,000 molecular species.

After the SMILES strings had been embedded and the VAE was trained, the latent space was sampled to generate new chemical species. Sampling was stopped once ~1,000,000 million valid species had been disseminated. This arbitrary stopping criterion was selected because it was greater than 10 times the size of the original dataset. All species were run through the data pipeline to check their uniqueness and stability. The new species contained species with atoms ranging from 1 to 74 with an average of 18 atoms while the original dataset had molecules with atoms between 1 and 49 with an average of 22. A comparison of these distributions can be seen in figure 20. The number of atoms was calculated using RDKit. It should be noted that this ability to predict larger molecules shows again the benefits of this approach over approaches such as retrosynthesizing. We can see statistically most of the larger molecules are outliers when compared to the rest of the molecules. It is hypothesized that if the latent space were to continue to be sampled, especially if the large molecule species were targeted sampling the area could produce many large molecules.

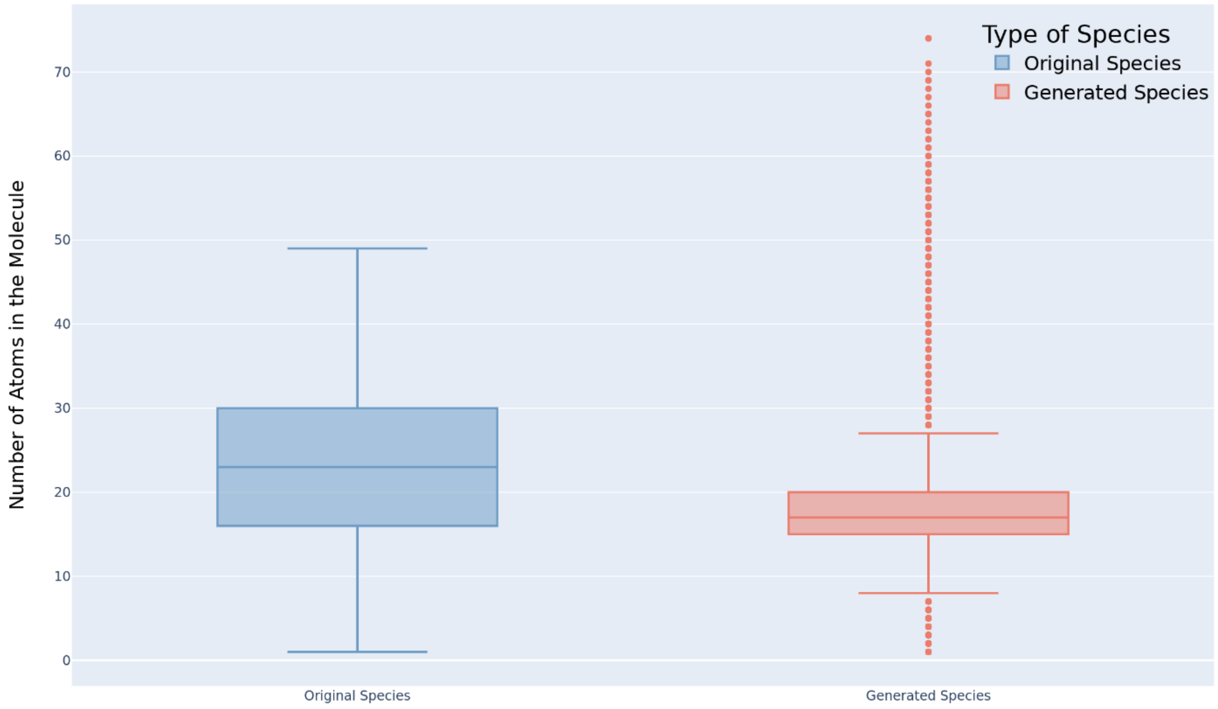


Figure 20: Box plot statistics showing the Original and Generated Species atom count.

The distribution of these atom counts can be seen in figure 21, which allows for a more in-depth analysis. In all plots for this study, the training dataset is pictured in blue in the background with the generated dataset in the foreground in red. The y-axis in all the plots is the percentage of all molecules within that bin and the x-axis is the associated chemical property. It can be seen from figure 21 that the generated species and original species share an approximately normal distribution with two main differences. The first is the high percentage of species in the first two bins and the second is that the distributions are not centered around the same number of atoms. Both of these differences are due to the original dataset containing single-element molecules. Since they are already included in the original species and AGoRaS-Quantum can't come up with new elements it is impossible for it to share that feature. However, due to the existence of these small molecules, it biased the network into creating species that were on average smaller than the median number of atoms for the original dataset.

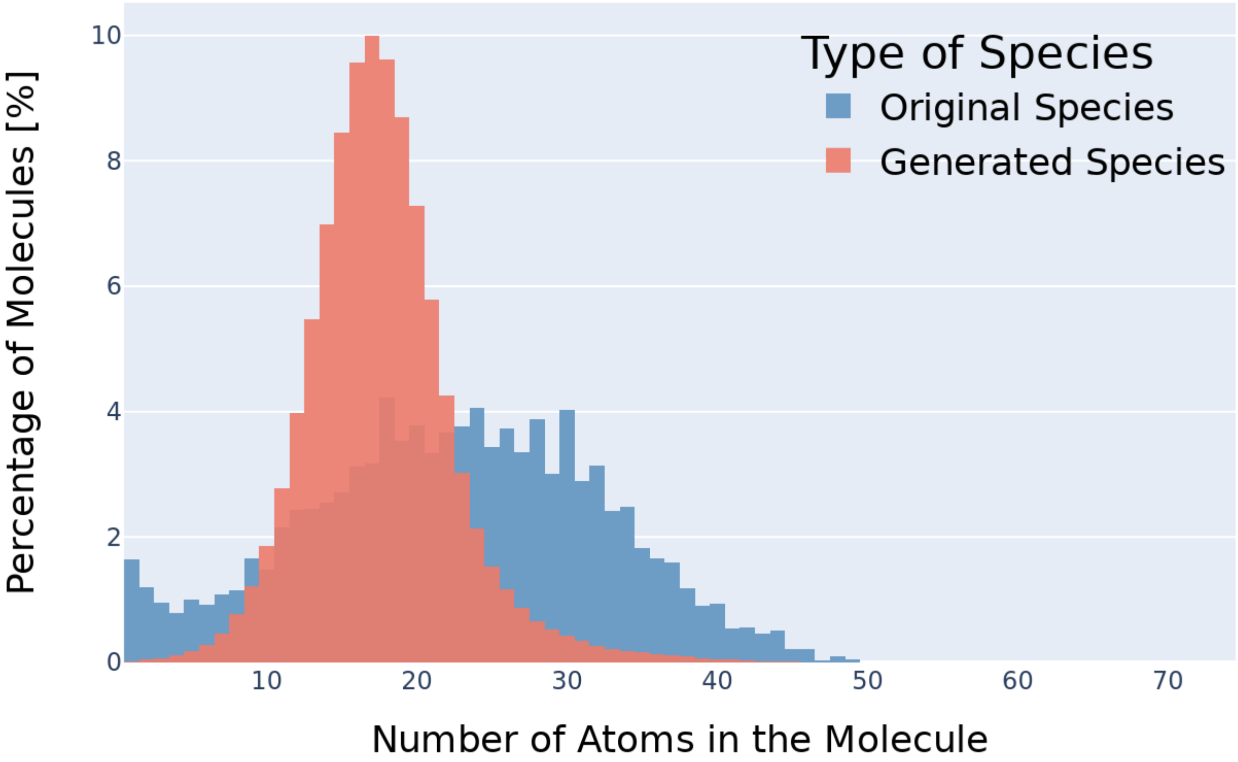


Figure 21: Histogram comparing the number of atoms per molecule between the original and the generated equations.

Of course, for these molecules to be useful as either a dataset for machine learning or as a database for potential experimentalists it had to be proved that these generated species shared the same properties as the original species. It was decided to look at molecules containing no more than 10 atoms. This was due to the computational complexity and cost associated with the semi-empirical calculations[160,179]. The criteria that were deemed the most important to compare between the datasets were those that can identify if a material is a single photon emitter. That is whether or not the calculated emission spectra of the molecule exhibit a single strong peak at some wavelength in the visible spectrum.

Using semi-empirical calculations it is possible to calculate the wavelength and emission strength of photons at each electronic level. It is important to note that the intensity of the light being emitted is "arbitrary". It is considered arbitrary due to the Franck-Condon Principle which explains the relative intensities of vibronic transitions. These intensities are the relation between the probability of a vibrational transition to the overlap of the vibrational wave functions[194]. This is shown in equations 33 and 34. In equation 33, μ is the molecular dipole operator, e is the charge of the electrons, r_i is the location of the electrons $Z_j e$ is the charge of the nuclei and, R_j is the position of the nuclei. In equation 34, P is the probability amplitude for the transition between the two states, ψ is the wavefunction of the original state, and ψ′is the wavefunction of the final state[195].

$$\mu = \mu_e + \mu_N = -e\sum_i r_i + e\sum_j Z_j R_j \qquad \text{Eq. 33}$$

$$P = \langle \psi' | \mu | \psi \rangle = \int \psi^{*\prime} \mu \psi d\tau \qquad \text{Eq. 34}$$

These calculations at each energy level led to the calculated emission spectra that will be used as validation of these materials. An example of this can be seen in table 5. All electronic levels of each molecular species were determined at the standard state using the semi-empirical computational technique described earlier. Other properties of interest that were calculated to help identify promising materials were the molecules, total dipole moment, vibrational frequency, and strength.

Num	Level Energy	Excitation Wavelength (nm)	Oscillator Strength (arb.)	Multiplicity (1 = singlet, 3= triplet)
0	0.00	0.00	0.000000	1
2	2.36	525.08	0.000000	3
3	3.05	404.50	0.000049	1
4	5.73	261.64	0.000000	3

Table 5: Example of the first few energy levels of the molecule NCCCCC=0 (C5H11NO). These are actual examples, but this table serves as a reference to the reader.

Since it would be unrealistically time-consuming to manually check all of the generated materials' electronic levels to see if they possessed a single strong peak a simple but effective methodology was implemented. For each material, the quantile ranges of the oscillator strength were calculated and the number of values greater than the 75th percentile was counted. It was found that ~99% of all generated materials' oscillator strength had no more than 3 values above the 75th percentile. This gives a clear indication that almost all the generated materials exhibited the desired chemical property of being a single photon-emitting material. This is extremely vital for the practicality of AGoRaS-Quantum going forward, as it showed it can generate materials with a very specific use case and structure. This will allow for its use in other areas of chemistry and material science where materials with specific properties are desired.

Once it was confirmed that the generated data exhibited single peak behavior it was necessary to perform further analysis to confirm that the generated data shared similar value ranges as the original data. For this overlapping histograms were determined to be an ideal way to show that the generated data had properties similar to that of the real data. Figure 22 shows a histogram of the original and generated species' peak oscillator strength. It can be seen that both the generated and original species have a semi-normal distribution with a slightly left skew to the values. However, it appears that on average the generated materials have a slightly weaker peak oscillator strength.

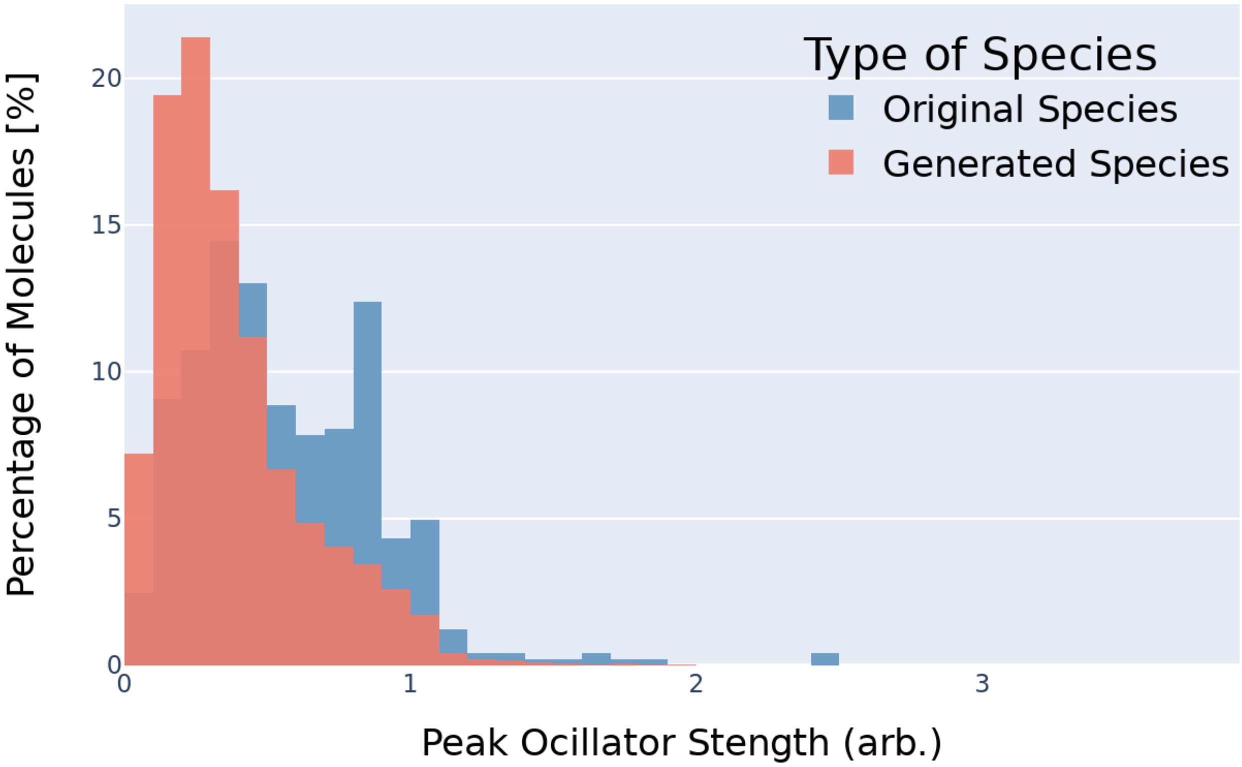


Figure 22: Histogram comparing the Peak Oscillator Strength between the original and the generated equations.

In this type of behavior, both the lower average peak strength and the identical distribution of values are expected due to the bias in the training data. Since the network uses all of the original data as a starting point for sampling the latent space it will always return data of a similar distribution. This problem could easily be overcome by sampling only data from the underrepresented regions until a uniform distribution was created. Which of course was not possible when originally generating the dataset as the semi-empirical calculations had not been run yet.

The high percentage of molecules being generated that produce weaker strengths is also a byproduct of the inherited bias. Due to the training data being sourced from experimental results, where only the best material i.e., the strongest emitter are reported. This leaves a lot of materials for AGoRaS-Quantum to be able to generate that still meet the chemical and physical requirements but do not produce as strong of a peak. Simply put the area of the latent space that generate strong peak materials is crowded, while the rest of the latent space is sparsely populated. As shown in study two, however, if the latent space were to continue to be sampled until we reached 100X the number of generated materials to the original material. Then the generated distributions would be exactly that of the original materials.

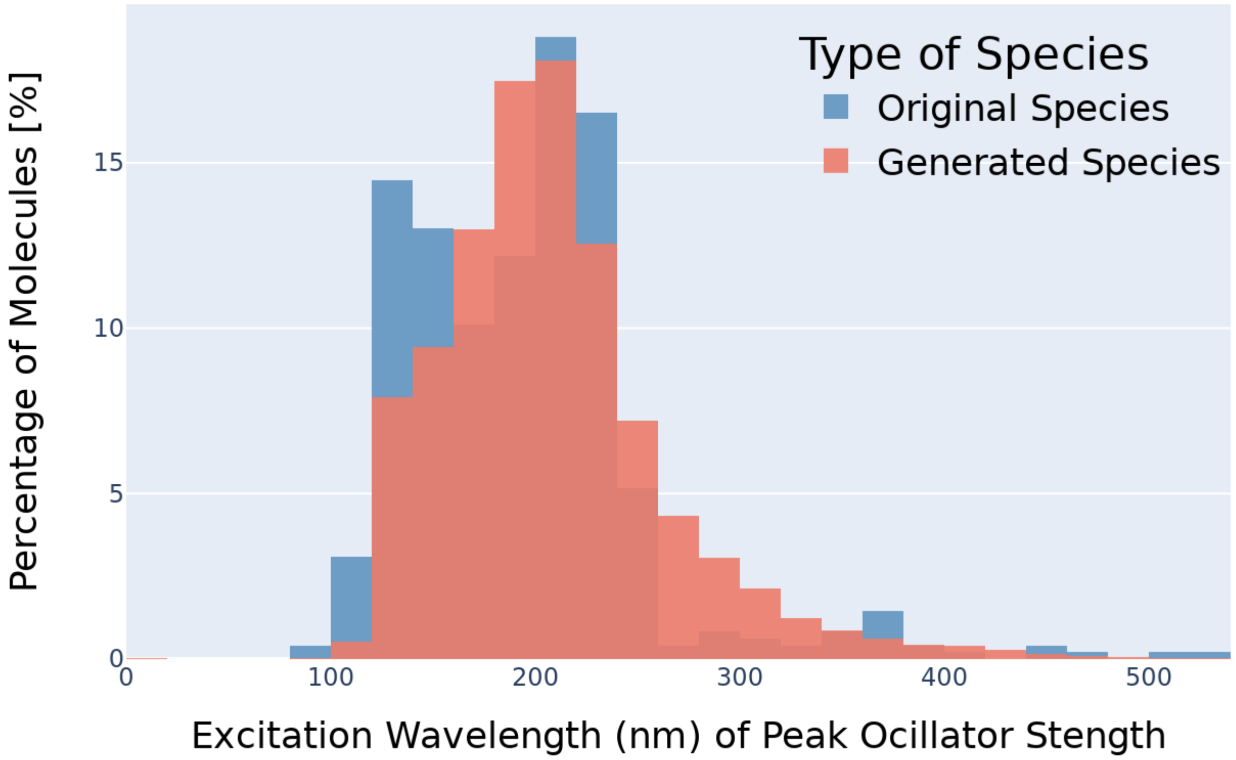


Figure 23: Histogram comparing the Excitation Wavelength at which the Peak Oscillator Strength occurs between the original and the generated equations.

Another important aspect of these types of materials is at which frequency these peaks occur. Figure 23 depicts the excitation wavelength at which the molecule's peak oscillator strength occurs for the original and generated species. Once again it can immediately be seen that the original and generated materials follow a similar distribution of semi-normal with a left skew. Like with the previous figure this could be corrected with a more directed sampling methodology. Another factor in the similarity of these distributions is that unlike the other histograms that have been shown in this study their values could theoretically be anything. The excitation wavelengths are calculated between 0 and 1400 nm, which helps to enforce an equal distribution of values within that range. An interesting find from figure 23 is that while the original data has a disjointed distribution of values when the excitation wavelength is greater than 250 nm. The generated data shows a much more normal distribution as the values tail out to 1200 nm. This helps to suggest that even if the training data has a disjointed distribution that a VAE will be able to generate a smooth distribution of the data.

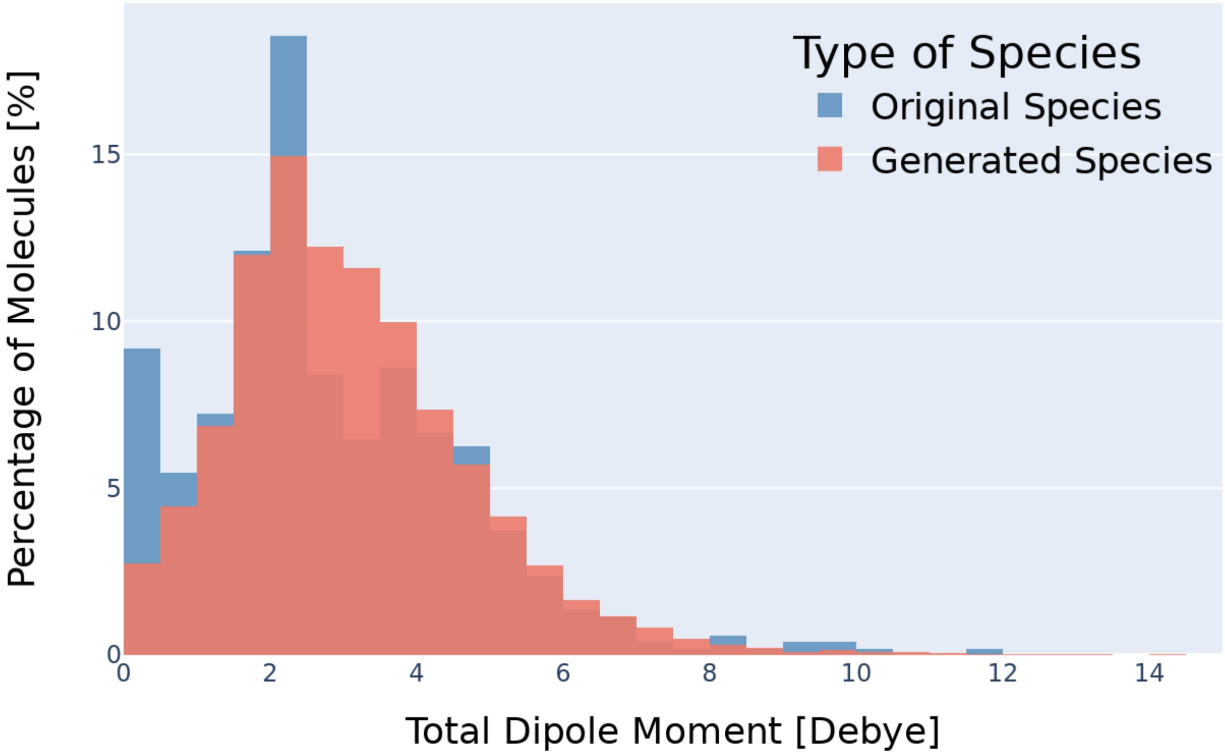


Figure 24: Histogram comparing the total dipole moment between the original and the generated equations.

It can be seen from figures 22 and 23 as well as the statistical analysis done, that the generated materials have the desired characteristics that make materials a single photon emitting material. However, it is important to check that the network was able to generate materials that mimicked the real dataset in various other chemical and physical aspects. To that end, it was decided to look at the total dipole moments of the real and generated materials. The dipole was also calculated using semi-empirical calculations. Which is based on the partial charge and positions of the atoms. Equation 27 was used to calculate the total dipole moment. The overlaid histogram for the total dipole moments of the original and generated species can be seen in figure 24. As we have seen previously it is a semi-normal distribution with a left skew. Something to note here is the high percentage of dipole values around 0 Debye for the original species. This is due to the original dataset containing single atom species which would have 0 Dipole. An interesting thing to note about figure 24 is as in figure 23 where the original data has a bit of an uneven distribution, the generated data is much smoother. This again helps to validate that for all chemical and physical features the VAE will be able to produce a more homogenous distribution.

As discussed earlier a key aspect of SPS materials is at what wavelength a peak appears and how strong that peak is. It can be seen in figures 22 and 22 that AGoRaS-Quantum can produce materials that have similar peak strength and wavelength values. While histograms are very good at showing distributions of data they lack the ability to show information about single points or about how two variables relate to one another. In figure 25 a scatter, plot is shown which shows the original and generated species peak oscillator strength mapped against the wavelength it is produced at. This helps to tie together figures 21 and 22 and indeed provides further insights that can greatly benefit users of this dataset. As in the previous figures, the red color dots represents the generated species, while the blue color dots represent the original species.

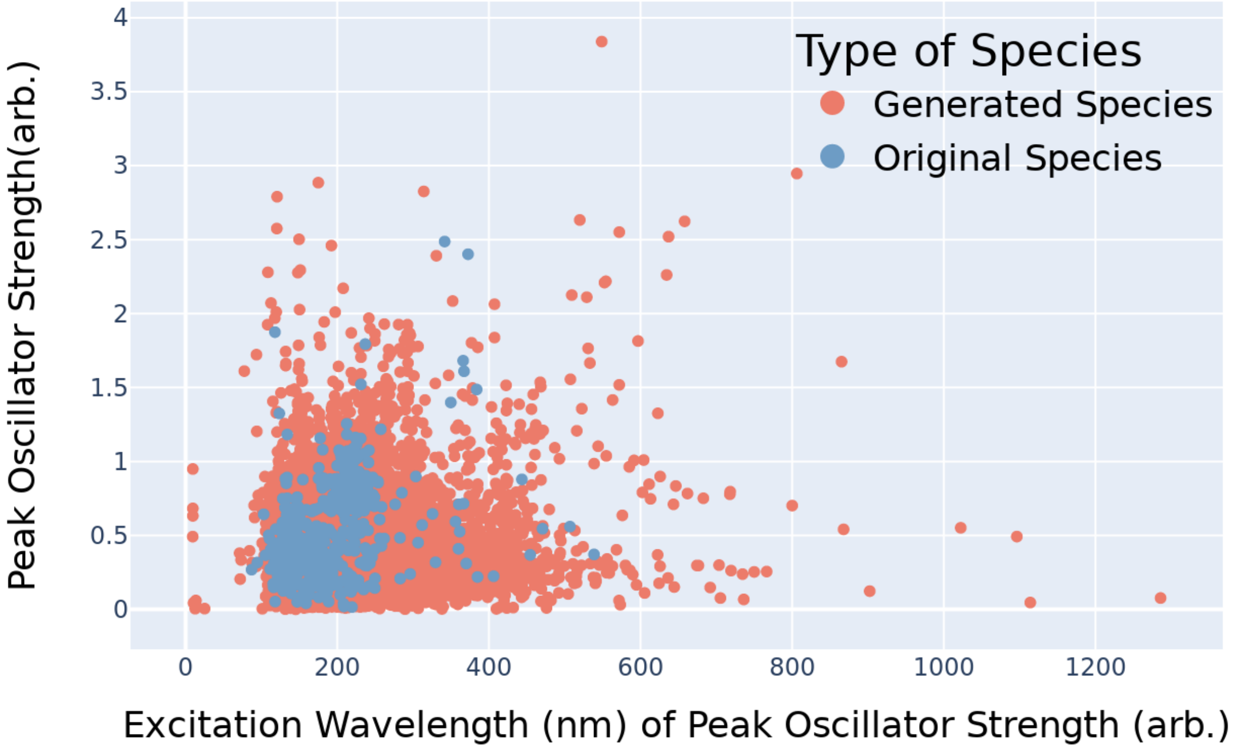


Figure 25: Scatter plot of the generated and original species peak oscillator strength against the excitation wavelength where it occurs

It can be immediately seen from figure 25 that the generated species have a much wider range of wavelengths that produce their peak oscillator strength. While the original species all have peak strengths somewhere between ~100 nm and 550 nm. The generated species produce peak values all the way from ~0 to 1,300 nm. It is very interesting to see that despite being trained on materials only in the visible light spectrum that the network was able to produce materials that had peaks in the infrared light spectrum. An encouraging sign of the usefulness of this network to scientists as a tool is the ability of the network to produce materials with very strong peak strengths. The original dataset only contained 2 materials that had a peak oscillator strength greater than 2. While the generated materials already have 29 materials with a strength above 2. This would allow a scientist to search the materials and find candidates for further investigation. Another thing that this graph helps to show us is how the generated materials are "filling up" the graph. We can see that between many of the blue dots the space is completely filled with red dots. This is an extremely important aspect for any future machine learning application that will be built off of this data. This helps to show that AGoRaS-Quantum is closing the gaps in the data, which will lead to more robust and powerful machine learning algorithms.

This filling in of the data represents an extremely important aspect of VAEs and especially of the AGoRaS-Quantum network. Which is the ability of the network to map the latent probabilistic solution space of these materials. By sampling all of the latent space the network would be able to fill in all of the gaps between points seen in figure 25. It is the aim to show that the new materials are filling in the solution space and therefore allowing for the removal of bias. To do this a t-Distributed Stochastic Neighbor Embedding (t-SNE) was undertaken and shown in figure 26. The t-SNE algorithm is used primarily to be able to explore and visualize high-dimensional data such as text. At its most simple level, it allows a user to get an understanding of how data is arranged in high-dimensional space. The algorithm accomplishes this through an unsupervised learning method of stochastic neighbor embedding to give high-dimensional data a single point on a two-dimensional grid.

For this t-SNE algorithm, the only input was the SMILES representations of the molecules embedded as numbers just as in the original training for the AGoRaS-Quantum network. The blue circles represent the generated data set and the red circles represent the training data. Figure 26 subfigure A has all of the original data ~8,000 species while only showing a randomly selected 8,000 of the generated species. Meanwhile, subfigure B is also showing the ~8,000 generated species but has 80,000 randomly selected generated species. This is done to illustrate how as we sample more species we can fill in the latent space. It can be seen from subfigure A that AGoRaS-Quantum is starting to fill in the blank spaces in the latent space. It is very interesting to note that most of the species selected belong to the larger emptier area within the latent space. Subfigure B it is clearly illustrated how the network is beginning to fill in all of the available space with generated materials. It appears the areas around the original species are the most densely populated with generated materials. This would make sense as species were used as entry points into the latent space to begin sampling. So, a high proportion of the early generated species would be located near the original species. Due to the memory cost, it was not possible to show how using 800,000 species would show an even more densely packed latent space.

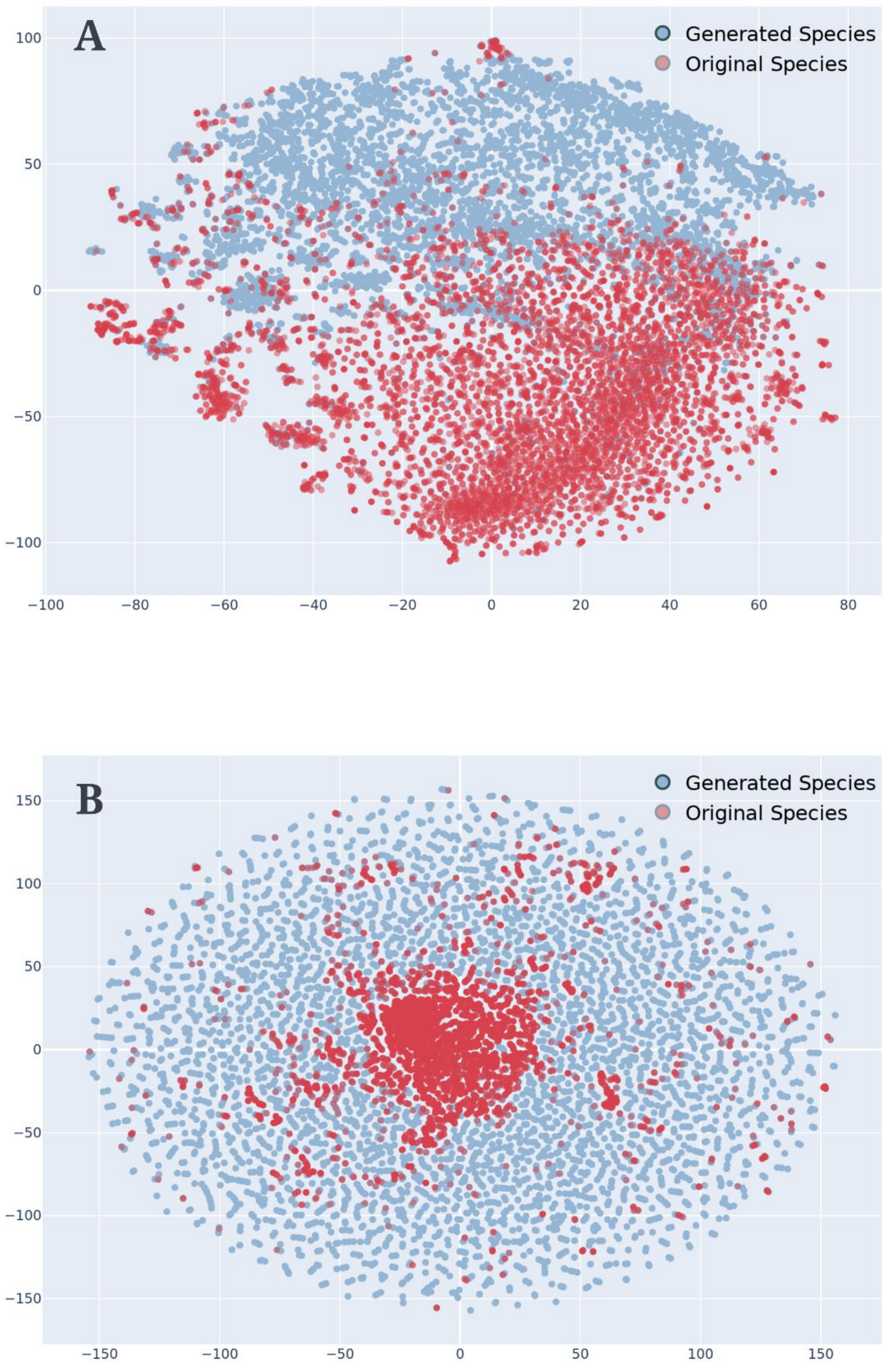
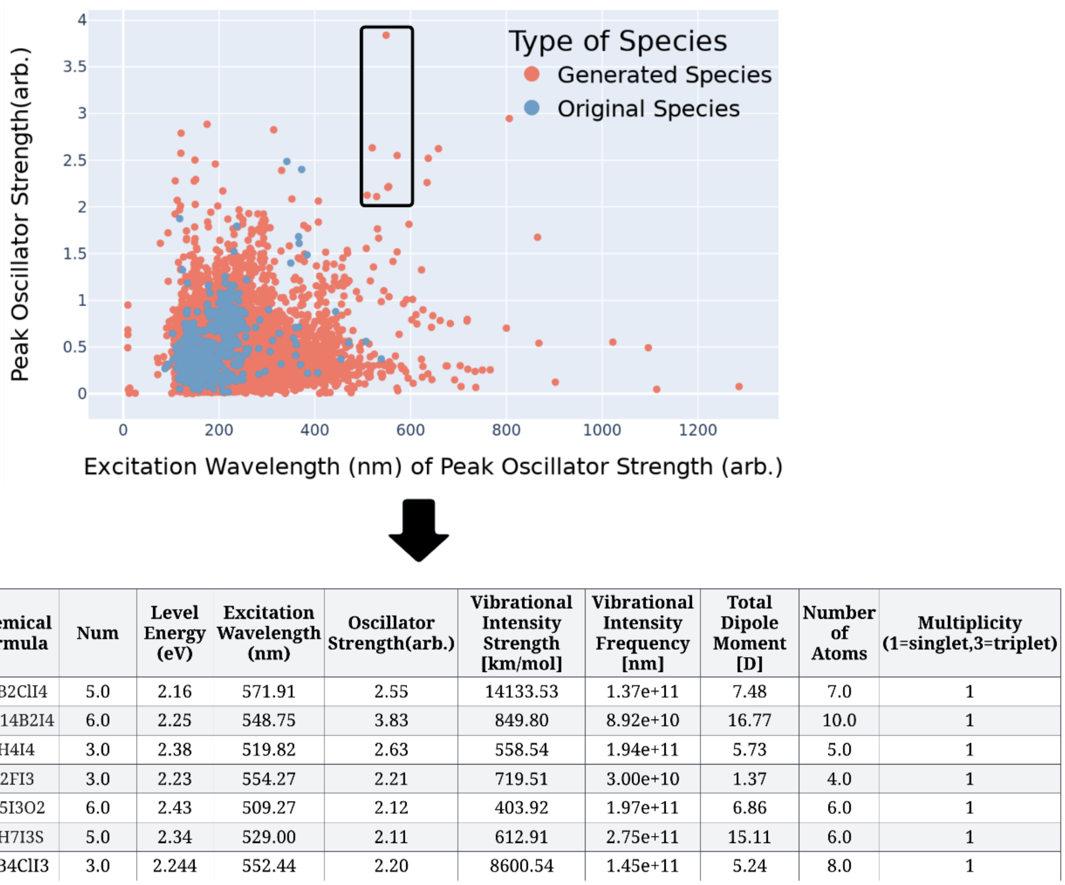


Figure 26: t-SNE plot of the training dataset and the generated dataset. Subfigure A All of the ~8,000 original species and a randomly sampled 8,000 generated species. Subfigure B All of the ~8,000 original species and 80,000 randomly sampled generated species.

Besides the ability to generate new datasets for machine learning training, the most important and immediately impactful application of AGoRaS-Quantum is the ability to conduct a targeted search of the generated materials. In this way, researchers can utilize the massive amount of new materials being created that exhibit behavior of interest to them. To demonstrate this application the generated dataset was searched for specific criteria. In this case, it was materials with a peak oscillator strength greater than 2 at a frequency between 500 – 600 nm, which corresponds to green and yellow light. These values were selected due to the high amount of use cases for materials in that frequency range, along with the strong peak strength. The results of this search can be seen in figure 27.



SMILES`	Chemical Formula	Num	Level Energy (eV)	Excitation Wavelength (nm)	Oscillator Strength(arb.)	Vibrational Intensity Strength [km/mol]	Vibrational Intensity Frequency [nm]	Total Dipole Moment [D]	Number of Atoms	Multiplicity (1=singlet,3=triplet)
ClBIIIIB	H7B2ClI4	5.0	2.16	571.91	2.55	14133.53	1.37e+11	7.48	7.0	1
CCCCIBBIII	C4H14B2I4	6.0	2.25	548.75	3.83	849.80	8.92e+10	16.77	10.0	1
C[I][I]II	CH4I4	3.0	2.38	519.82	2.63	558.54	1.94e+11	5.73	5.0	1
FIII	H2FI3	3.0	2.23	554.27	2.21	719.51	3.00e+10	1.37	4.0	1
COOIII	CH5I3O2	6.0	2.43	509.27	2.12	403.92	1.97e+11	6.86	6.0	1
SCCIII	C2H7I3S	5.0	2.34	529.00	2.11	612.91	2.75e+11	15.11	6.0	1
ClBBBBIII	H6B4ClI3	3.0	2.244	552.44	2.20	8600.54	1.45e+11	5.24	8.0	1

Figure 27: Example of the ability to search the generated solution space for molecules of potential interest. In this example molecules in the 500-600 nm wavelengths were chosen if they had a strength above 2.0.

The most immediate thing that stands out is that all returned species are from the generated dataset and none from the original dataset. In fact, to get data from the original dataset the peak oscillator strength threshold would need to be lowered to 0.5. Of course, just the peak oscillator strength and frequency are not enough to determine if a material is suitable for further investigation. To help identify more promising materials the peak vibrational intensity strength and corresponding frequency were also calculated along with the total dipole moment and the size of the molecule. For instance, the dipole moment can help a researcher identify if a material would be a promising candidate for a quantum sensor. As stated earlier each material's emission spectrum is not studied as it would be too manually intensive. So once some promising materials have been identified it is important to perform further investigation and plot the excitation wavelength and oscillator strength as well as the vibrational intensity strength and frequency. An example of these plots for the CCCClBBIII molecule can be seen in figure 28.

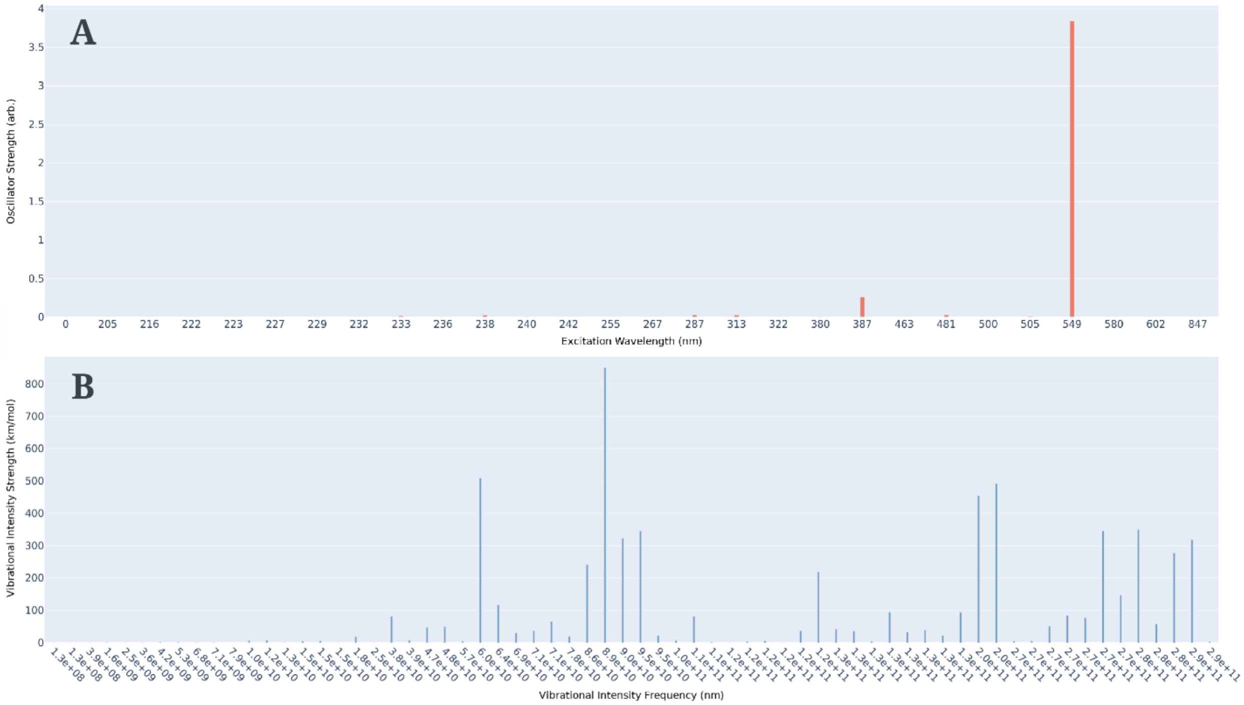


Figure 28: Subfigure A Oscillator strength versus excitation wavelength for the CCCClBBIII molecule. Subfigure B Vibrational intensity frequency versus vibrational intensity strength for the CCCClBBIII molecule.

The first molecule investigated was the CCCClBBIII molecule. It was chosen due to its high oscillator strength. An analysis of figure 28, subfigure A shows that the CCCClBBIII molecule has a peak excitation wavelength at 549 nm. It also shows several other smaller peaks all of them relatively low in comparison. The main peak is approximately 20 times stronger than the next highest peak. A review of subfigure B shows a lot of photon peaks throughout the spectra. None of these directly correlate to an integer of 549 nm. However, due to the abundance of peaks, it is could be suggested that a closer examination take place of the CCCClBBIII molecule. This higher abundance of peaks could create a scattering of the light phonon making it a less desirable molecule.

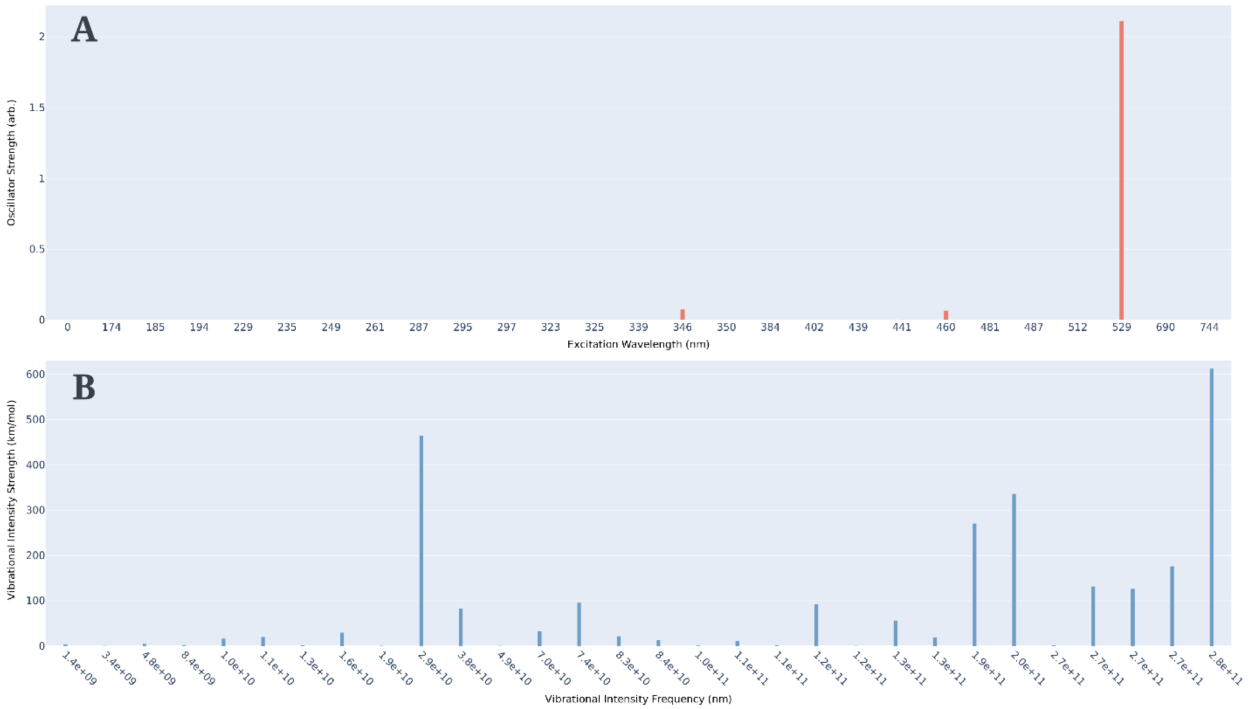


Figure 29: Subfigure A Oscillator strength versus excitation wavelength for the SClI molecule. Subfigure B Vibrational intensity frequency versus vibrational intensity strength for the SClI molecule.

The next molecule chosen to investigate was SCII as it had the strongest dipole moment. In figure 29 it is possible to see the results of the analysis. Compared to the CCCClBBIII molecule it is clear that the SCII molecule has approximately the same number of peaks and those that do exist are at a lower intensity level. The peak intensity of the SCII molecule is 20 times greater than that of the next highest peak. Again, a quick analysis can be done to see if any of the peaks in subfigure B are an integer of the SCII molecules peak at 529 nm.

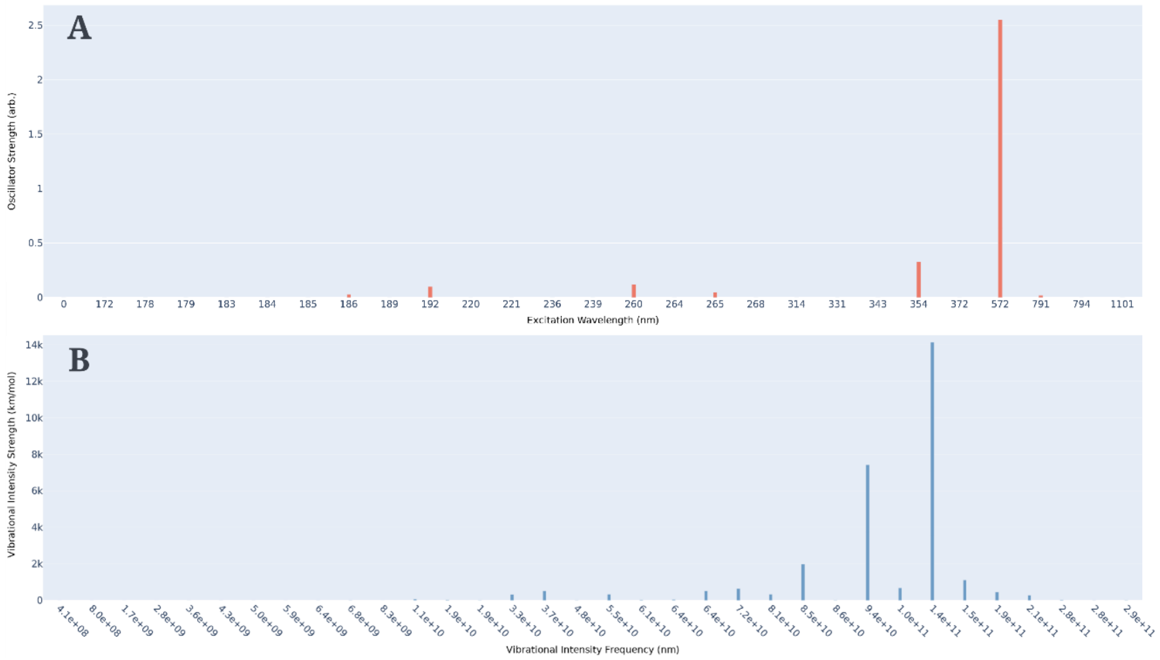


Figure 30: Subfigure A Oscillator strength versus excitation wavelength for the ClBIIIIB molecule. Subfigure B Vibrational intensity frequency versus vibrational intensity strength for the ClBIIIIB molecule.

It can be seen from figure 30 subfigure A that the ClBIIIIB molecule has several small peaks but one very dominating peak at 572 nm. The peak located at 572 nm is approximately 6 times as strong as the other peaks found in the excitation spectra. This ratio is noticeable smaller than that of the other two molecules. While it still appears to be an excellent molecule special note should be taken to make sure no other light spectra are emitted to a degree that would interfere with its purpose. Subfigure B can be investigated to see if any of the photon frequencies is an integer of the optical absorption peak excitation wavelength. If it is this can indicate that the material might have photon scattering.

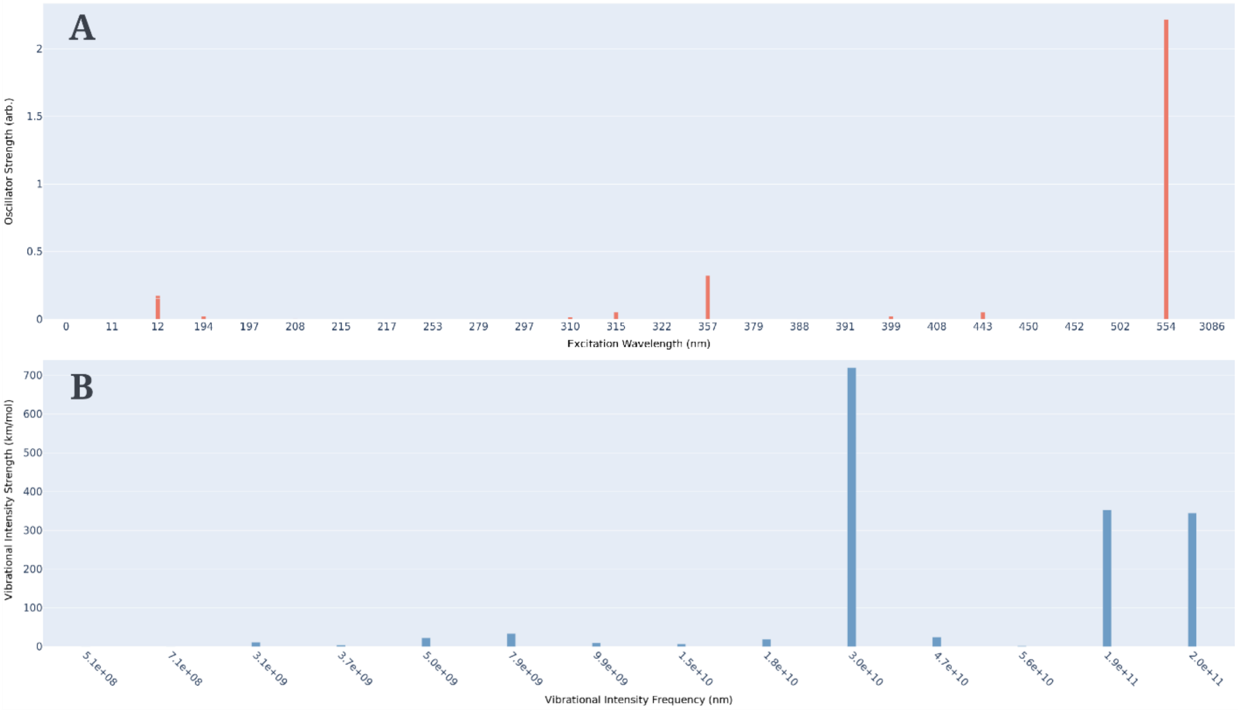


Figure 31: Subfigure A Oscillator strength versus excitation wavelength for the FIII molecule. Subfigure B Vibrational intensity frequency versus vibrational intensity strength for the FIII molecule

Another molecule of interest was the FIII molecule due to its simplicity and relatively small size. As with the other molecules, it exhibits the desired behavior of a single dominate peak, this one at 554 nm. This peak like the previous molecule studied is around 6 times as strong as any other peak in the spectrum. It is believed that this would still make an excellent material for single photon emitting materials due to the magnitude increase in the peaks strength as well as the small size of the molecule. This combined with the relatively few photon peaks seen in subfigure B make it likely that this material would produce a strong clear light.

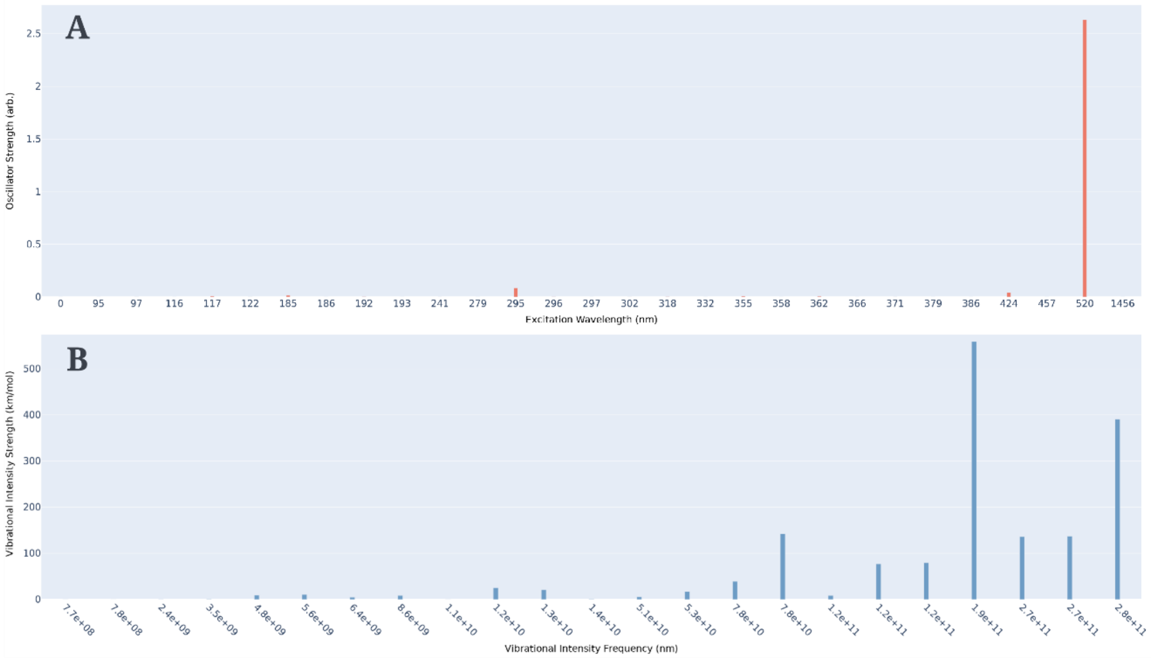


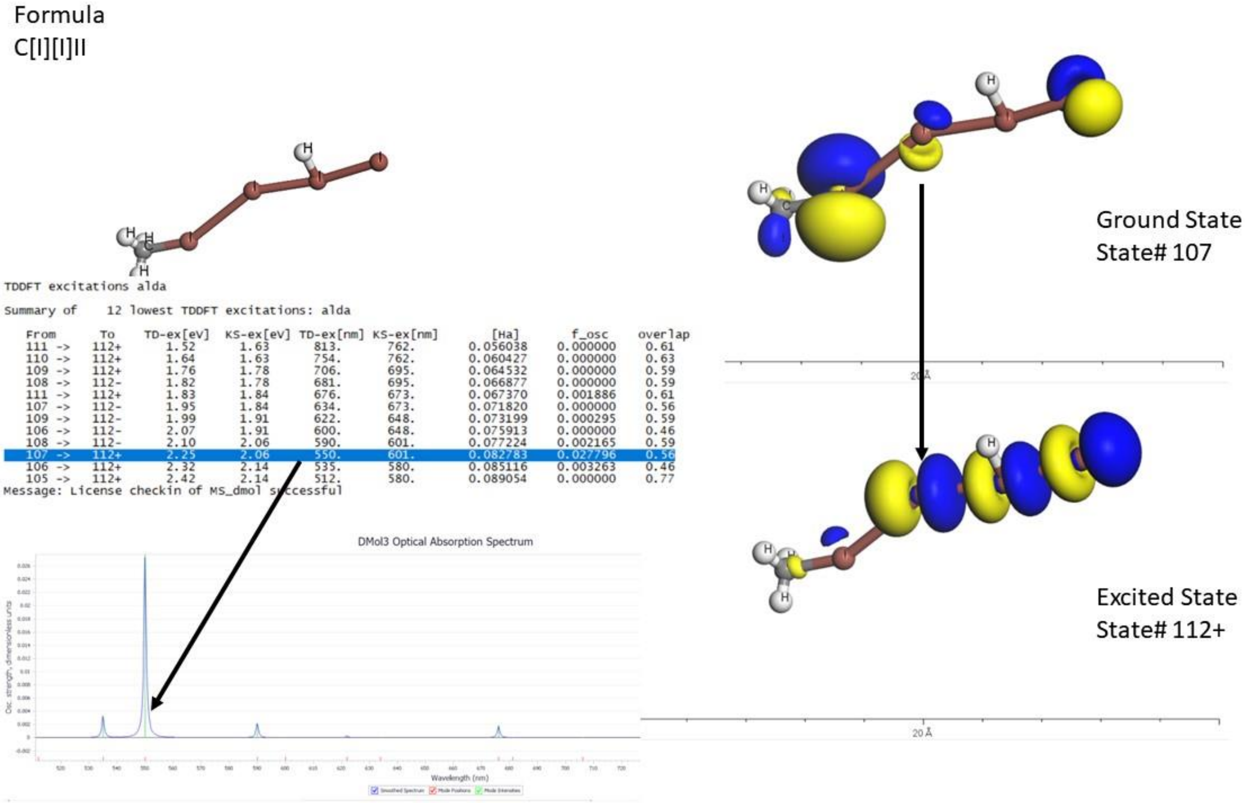
Figure 32: Subfigure A Oscillator strength versus excitation wavelength for the C[I][I]II molecule. Subfigure B Vibrational intensity frequency versus vibrational intensity strength for the C[I][I]II molecule

The final molecule chosen from the molecules of interest for study here was the C[I][I]II molecule. It was chosen because it had a good combination of peak strength, molecule size, and dipole moment. These 3 things made the molecule a good overall chose to study in more detail. In figure 32 subfigure A it is clear that this molecule has a single photon peak at 520 nm. It is also interesting that this molecule exhibits almost no other peaks of significance. It only has very small peaks at 295 nm and 424 but is otherwise free of other light-emitting peaks. The dominate peak is approximately 26 times larger than the other peaks in the spectra. This indicates that C[I][I]II might be an excellent SPS material. It is also clear from subfigure B that none of the photon peaks are integers of the peak wavelength.

This analysis of the five molecules helps to validate our early assertion that all materials generated by the AGoRaS-Quantum network have the characteristic single strong peak that was intended. If a researcher is further intrigued by these types of materials, they will be able to go into the network and use these species as points for directed sampling. Which should allow the network to create new materials from the latent space that have these same properties. Allowing for an exponential increase in the possible materials for study.

The final step in this study was to choose two promising materials and perform further analysis using TDDFT. To that end, the FIII and C[I][I]II molecules were chosen for the TDDFT analysis. This additional analysis is done for several reasons, the first is to provide greater confidence to researchers in AGoRaS-Quantum's ability to generate real and interesting materials.

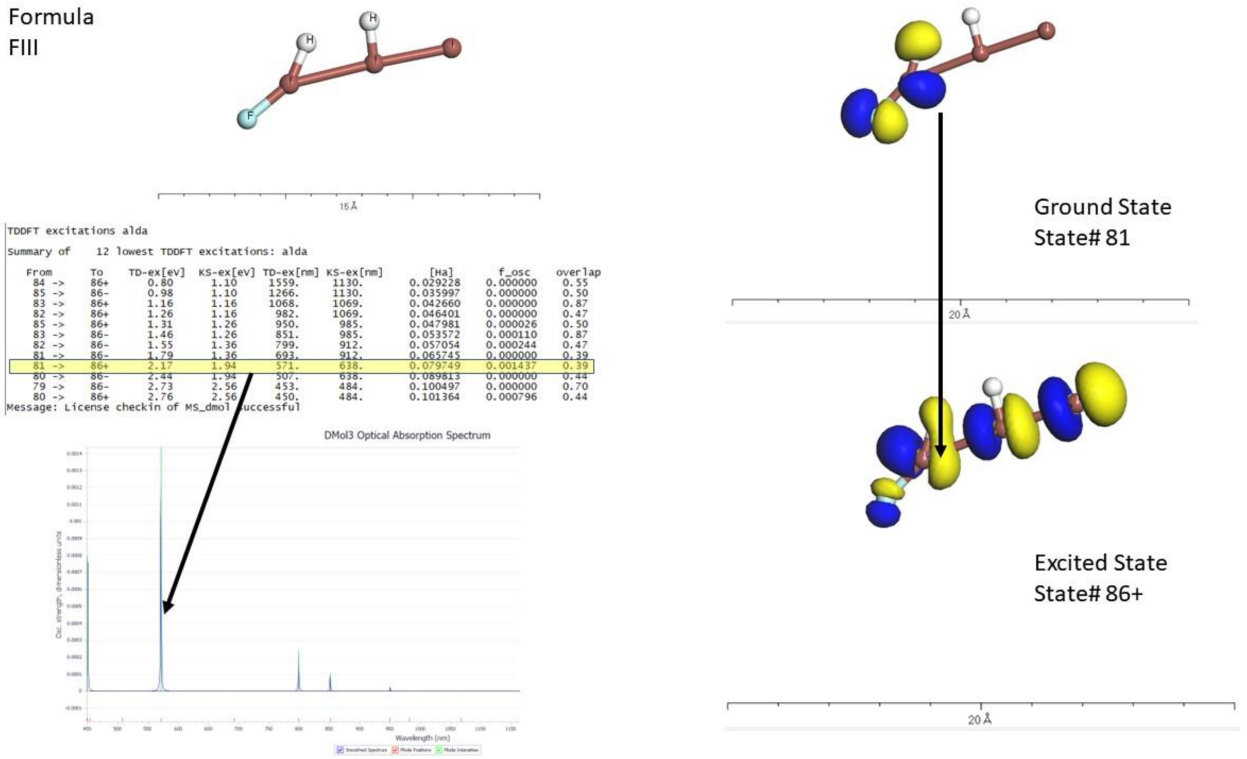
The other is to showcase how once the network has generated some interesting material researchers can possibly manipulate the molecule in order to have some desired wavelength. It also allows researchers to better study what is happening within the molecule when it generates the photon. Which helps to facilitate the design of systems such as quantum sensors. The excitation energies indicate the location of absorption peaks in the spectrum. While calculating the transition dipole moments between the different states involved in the different excitations gives the peak intensities[188,191]. An important thing to note for TDDFT methods is that there is a strong correlation between the excitation energy errors and the degree of spatial overlap. This spatial overlap is between the occupied and virtual orbitals involved in an excitation. As the overlap gets smaller the errors increase[196,197].



```
TDDFT excitations alda

Summary of    12 lowest TDDFT excitations: alda

   From      To    TD-ex[eV]  KS-ex[eV] TD-ex[nm] KS-ex[nm]       [Ha]       f_osc     overlap
   111 ->   112+      1.52       1.63      813.      762.      0.056038   0.000000    0.61
   110 ->   112+      1.64       1.63      754.      762.      0.060427   0.000000    0.63
   109 ->   112+      1.76       1.78      706.      695.      0.064532   0.000000    0.59
   108 ->   112-      1.82       1.78      681.      695.      0.066877   0.000000    0.59
   111 ->   112+      1.83       1.84      676.      673.      0.067370   0.001886    0.61
   107 ->   112-      1.95       1.84      634.      673.      0.071820   0.000000    0.56
   109 ->   112-      1.99       1.91      622.      648.      0.073199   0.000295    0.59
   106 ->   112-      2.07       1.91      600.      648.      0.075913   0.000000    0.46
   108 ->   112-      2.10       2.06      590.      601.      0.077224   0.002165    0.59
   107 ->   112+      2.25       2.06      550.      601.      0.082783   0.027796    0.56
   106 ->   112+      2.32       2.14      535.      580.      0.085116   0.003263    0.46
   105 ->   112+      2.42       2.14      512.      580.      0.089054   0.000000    0.77
Message: License checkin of MS_dmol successful
```

Figure 33: TDDFT results for the C[I][I]II (CH4I4) molecule. The excitation spectra are shown in the bottom left-hand corner, with the corresponding table shown above as well as the original molecule. The right-hand side of the image depicts the ground state of the molecule and the excited state at which the molecule emits the strongest light.

The first molecule studied using TDDFT was the C[I][I]II molecule due to the results seen in the semi-empirical analysis. The TDDFT spectra does not cover the wide frequency range as in the semi-empirical method. Instead, it is more fine-tuned into the range of frequencies that it is suspected that this material emits light at. The first thing to note is that the TDDFT calculations show the peak wavelength at 550 nm while the semi-empirical method had the peak at 520 nm. This is considered a negligible difference. In the table shown in figure 33, the “f_osc” column indicates the relative strength of the peaks. It is also important to note the overlap column, as mentioned earlier the smaller the overlap the higher the chance of error in the excitation energy. Once again in figure 33, it is possible to see a single dominate peak and several other smaller peaks. This is an excellent validation that the generated material exhibits the desired material properties. The right-hand side of figure 33 shows the molecule in its ground state with filled electron orbitals (top) and the molecule in the excited state (bottom). This excited state corresponds to the state at which the molecule will produce light at 550 nm. The overlap between them is one of the lower values of any depicted in figure 33. This gives increased confidence in the predictions.



```
TDDFT excitations alda

Summary of    12 lowest TDDFT excitations: alda

  From      To    TD-ex[eV]  KS-ex[eV] TD-ex[nm] KS-ex[nm]      [Ha]        f_osc     overlap
   84 ->    86+     0.80       1.10     1559.     1130.      0.029228    0.000000    0.55
   85 ->    86-     0.98       1.10     1266.     1130.      0.035997    0.000000    0.50
   83 ->    86+     1.16       1.16     1068.     1069.      0.042660    0.000000    0.87
   82 ->    86+     1.26       1.16      982.     1069.      0.046401    0.000000    0.47
   85 ->    86+     1.31       1.26      950.      985.      0.047981    0.000026    0.50
   83 ->    86-     1.46       1.26      851.      985.      0.053572    0.000110    0.87
   82 ->    86-     1.55       1.36      799.      912.      0.057054    0.000244    0.47
   81 ->    86-     1.79       1.36      693.      912.      0.065745    0.000000    0.39
   81 ->    86+     2.17       1.94      571.      638.      0.079749    0.001437    0.39
   80 ->    86-     2.44       1.94      507.      638.      0.089813    0.000000    0.44
   79 ->    86-     2.73       2.56      453.      484.      0.100497    0.000000    0.70
   80 ->    86+     2.76       2.56      450.      484.      0.101364    0.000796    0.44
Message: License checkin of MS_dmol successful
```

Figure 34: TDDFT results for the FIII (H2FI3) molecule. The excitation spectra are shown in the bottom left-hand corner, with the corresponding table shown above as well as the original molecule. The right-hand side of the image depicts the ground state of the molecule and the excited state at which the molecule emits the strongest light.

The second molecule studied using TDDFT was the FIII molecule, chosen due to the results calculated in the semi-empirical method. The initial analysis shows that the TDDFT calculations determined that the peak oscillator strength for this molecule occurs at 571 nm. While the semi-empirical methods calculated the peak to be at 554 nm. Since TDDFT is the more accurate method for calculating the emission spectra it is excepted that this molecule would produce light at 571 nm. This difference is roughly the same as the previous molecule and is no cause for concern, as these two wavelengths would produce almost the same light color. An analysis of the results from the TDDFT shows a strong peak at 450 nm for FIII as seen in figure 34. This peak is still around half as strong as the dominate peak seen at 571 nm, but further study is warranted. Overall, however, this material still exhibits the desired behavior of having distinct peaks at discrete wavelengths. Once again the molecule is depicted on the right-hand side in the ground state (top) and the excited state (bottom) at which it will emit the strongest light.

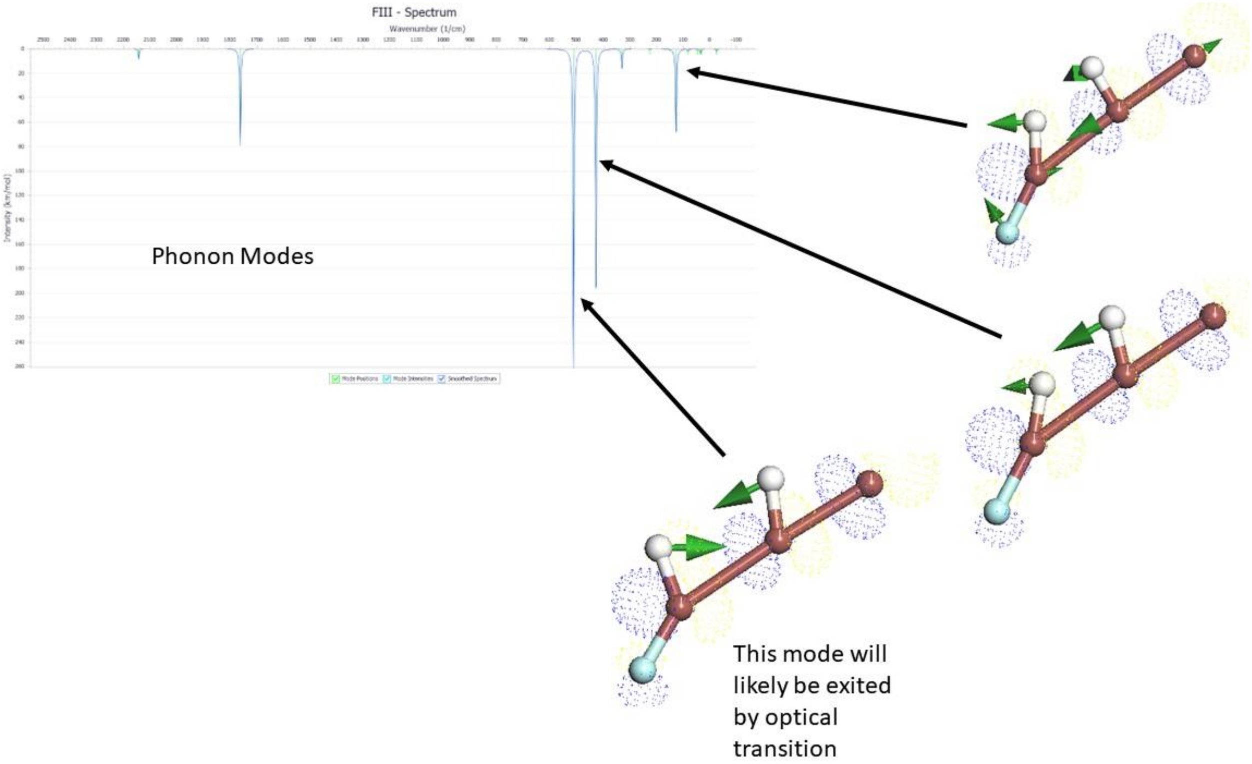
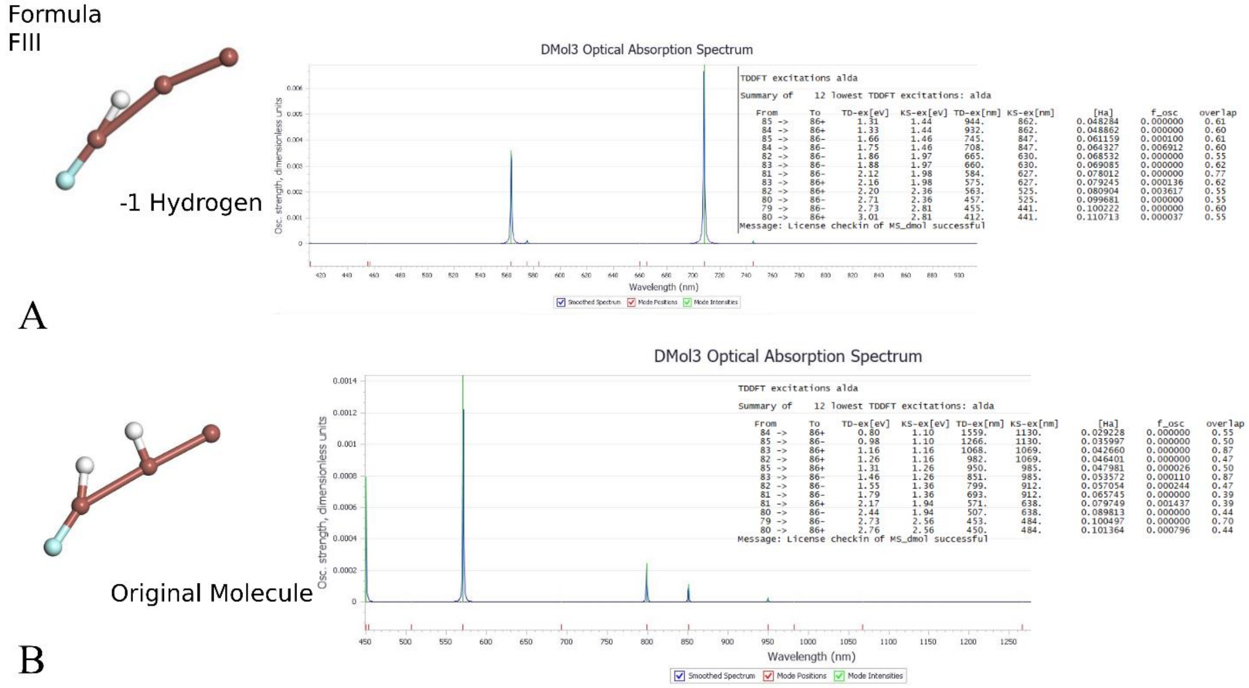


Figure 35: TDDTF results of the photon modes for the FIII molecule. The molecule is depicted showing what happens to it if the photon is excited at one of the three highest modes.

As with the semi-empirical methods, it is once again important to look at the photon modes of the molecules. However, since it is not the main focus of this study it was chosen to only examine the FII molecules' photon modes to show agreement with the semi-empirical calculations. The first thing to note from figure 35 is that the x-axis is wavenumber rather than wavelength as in the FII photon emission graphs from figure 31 subfigure B. The main takeaway from a comparison between the two methods is that they both result in 3 distinct peaks for the photon spectrum. This indicates good agreement between the methods and still helps to indicate that this could make an excellent SPS material.

So far this study has looked at only the materials being generated directly by AGoRaS-Quantum. With the TDDFT study, it is possible to look beyond just these molecules and investigate how they can be changed to produce peaks at different wavelengths. This portion of the study is designed to show the robustness of the generated molecules while also showing researchers a way to directly use the results of this network. However, since it goes beyond the main focus of validating the generated materials as single photon emitting materials it was decided to only look at the effects that removing single hydrogen atoms had on the molecules. This will provide further justification of the networks generated molecules' stability. While also helping to make this work more widely usable.



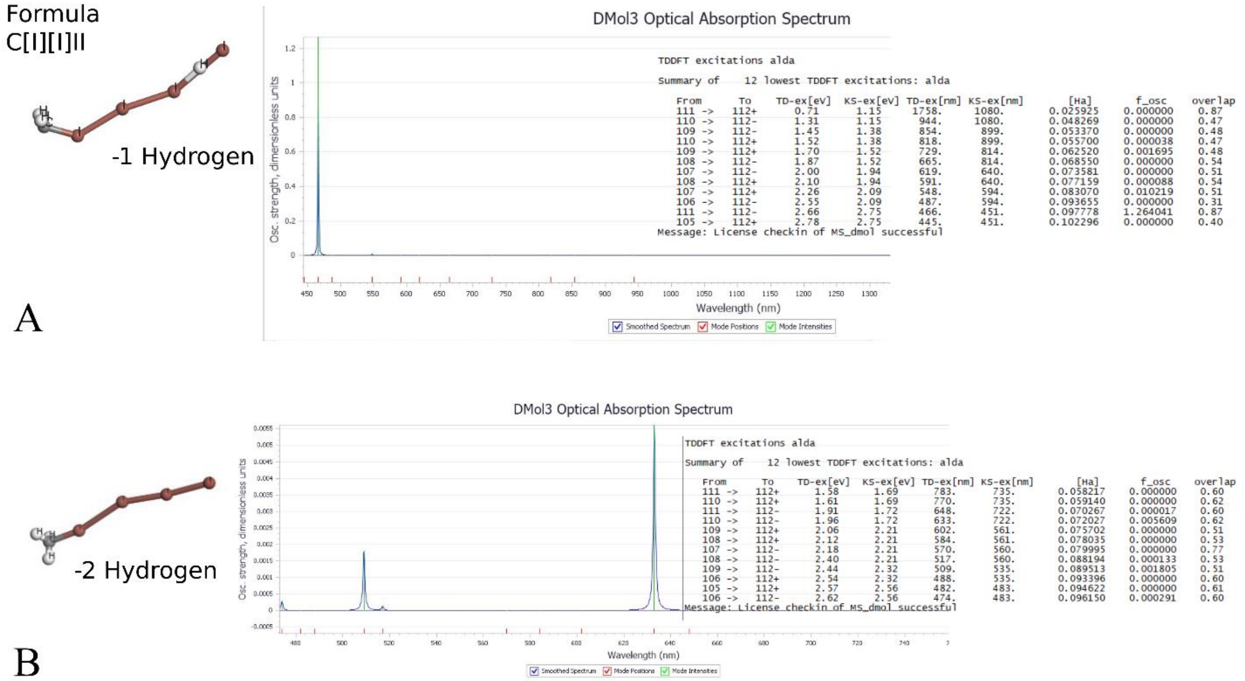
```
TDDFT excitations alda

Summary of    12 lowest TDDFT excitations: alda

  From      To    TD-ex[eV]  KS-ex[eV] TD-ex[nm] KS-ex[nm]       [Ha]      f_osc    overlap
   85 ->    86+      1.31       1.44      944.      862.      0.048284   0.000000    0.61
   84 ->    86+      1.33       1.44      932.      862.      0.048862   0.000000    0.60
   85 ->    86-      1.66       1.46      745.      847.      0.061159   0.000100    0.61
   84 ->    86-      1.75       1.46      708.      847.      0.064327   0.006912    0.60
   82 ->    86-      1.86       1.97      665.      630.      0.068532   0.000000    0.55
   83 ->    86-      1.88       1.97      660.      630.      0.069085   0.000000    0.62
   81 ->    86-      2.12       1.98      584.      627.      0.078012   0.000000    0.77
   83 ->    86+      2.16       1.98      575.      627.      0.079245   0.000136    0.62
   82 ->    86+      2.20       2.36      563.      525.      0.080904   0.003617    0.55
   80 ->    86-      2.71       2.36      457.      525.      0.099681   0.000000    0.55
   79 ->    86-      2.73       2.81      455.      441.      0.100222   0.000000    0.60
   80 ->    86+      3.01       2.81      412.      441.      0.110713   0.000037    0.55
Message: License checkin of MS_dmol successful
```

```
TDDFT excitations alda

Summary of    12 lowest TDDFT excitations: alda

  From      To    TD-ex[eV]  KS-ex[eV] TD-ex[nm] KS-ex[nm]       [Ha]      f_osc    overlap
   84 ->    86+      0.80       1.10     1559.     1130.      0.029228   0.000000    0.55
   85 ->    86-      0.98       1.10     1266.     1130.      0.035997   0.000000    0.50
   83 ->    86+      1.16       1.16     1068.     1069.      0.042660   0.000000    0.87
   82 ->    86+      1.26       1.16      982.     1069.      0.046401   0.000000    0.47
   85 ->    86+      1.31       1.26      950.      985.      0.047981   0.000026    0.50
   83 ->    86-      1.46       1.26      851.      985.      0.053572   0.000110    0.87
   82 ->    86-      1.55       1.36      799.      912.      0.057054   0.000244    0.47
   81 ->    86-      1.79       1.36      693.      912.      0.065745   0.000000    0.39
   81 ->    86+      2.17       1.94      571.      638.      0.079749   0.001437    0.39
   80 ->    86-      2.44       1.94      507.      638.      0.089813   0.000000    0.44
   79 ->    86-      2.73       2.56      453.      484.      0.100497   0.000000    0.70
   80 ->    86+      2.76       2.56      450.      484.      0.101364   0.000796    0.44
Message: License checkin of MS_dmol successful
```

Figure 36: Subfigure A: FIII molecule with one hydrogen atom removed emission spectra. On the left-hand side is a generated image of the molecule and on the right-hand side is the absorption spectra. Subfigure B: FIII molecule emission spectra shown for the originally generated molecule. On the left-hand side is a generated image of the molecule and on the right hand side is the absorption spectrum.

As would be expected when an atom is removed from one of the molecules the peak wavelength shifts. For the FII atom depicted in figure 36, the removal of a Hydrogen atom shifts the peak from 571 nm to 709 nm. This is a large shift in the color of the light being emitted, potentially opening up this material for different use cases. It can be seen that the secondary peak that was depicted at the edge of the spectrum in figure 34 has also shifted from 450 nm to 575 nm. It is interesting to note, as mentioned earlier the H2FI3 molecule does not exist in the PubChem database but the molecules for $HFI3^{-}$ and FI3 both do.

A similar analysis was carried out on the C[I][I]II molecule, with the removal of hydrogen atoms. The results are shown in figure 37, where the top molecule and spectra are C[I][I]II with a single hydrogen removed and the bottom is the molecule with two hydrogen atoms removed. Once again it is possible to see a very large shift in the peak wavelength due to the removal of the hydrogen atom. In the original molecule, the peak wavelength is shown at 550 nm but with one hydrogen atom removed that peak shifts to 466 nm. It also appears to be a much stronger peak than in the original molecule. An examination of figure 37 also shows that if two hydrogen atoms are removed the peak shifts strongly in the other direction. Resulting in a peak wavelength of 633 nm. Another check revels that both the molecules with hydrogen(s) removed $CH3I4^{-}$ and CH2I4 exists in PubChem along with the original CH4I4 molecule. This helps to validate the network can generate interesting and real species.



```
TDDFT excitations alda

Summary of    12 lowest TDDFT excitations: alda

  From      To    TD-ex[eV]  KS-ex[eV] TD-ex[nm] KS-ex[nm]       [Ha]       f_osc     overlap
  111 ->   112+     0.71       1.15     1758.     1080.     0.025925   0.000000    0.87
  110 ->   112-     1.31       1.15      944.     1080.     0.048269   0.000000    0.47
  109 ->   112-     1.45       1.38      854.      899.     0.053370   0.000000    0.48
  110 ->   112+     1.52       1.38      818.      899.     0.055700   0.000038    0.47
  109 ->   112+     1.70       1.52      729.      814.     0.062520   0.001695    0.48
  108 ->   112-     1.87       1.52      665.      814.     0.068550   0.000000    0.54
  107 ->   112-     2.00       1.94      619.      640.     0.073581   0.000000    0.51
  108 ->   112+     2.10       1.94      591.      640.     0.077159   0.000088    0.54
  107 ->   112+     2.26       2.09      548.      594.     0.083070   0.010219    0.51
  106 ->   112-     2.55       2.09      487.      594.     0.093655   0.000000    0.31
  111 ->   112-     2.66       2.75      466.      451.     0.097778   1.264041    0.87
  105 ->   112+     2.78       2.75      445.      451.     0.102296   0.000000    0.40
Message: License checkin of MS_dmol successful
```

```
TDDFT excitations alda

Summary of    12 lowest TDDFT excitations: alda

  From      To    TD-ex[eV]  KS-ex[eV] TD-ex[nm] KS-ex[nm]       [Ha]       f_osc     overlap
  111 ->   112+     1.58       1.69      783.      735.     0.058217   0.000000    0.60
  110 ->   112+     1.61       1.69      770.      735.     0.059140   0.000000    0.62
  111 ->   112-     1.91       1.72      648.      722.     0.070267   0.000017    0.60
  110 ->   112-     1.96       1.72      633.      722.     0.072027   0.005609    0.62
  109 ->   112+     2.06       2.21      602.      561.     0.075702   0.000000    0.51
  108 ->   112+     2.12       2.21      584.      561.     0.078035   0.000000    0.53
  107 ->   112-     2.18       2.21      570.      560.     0.079995   0.000000    0.77
  108 ->   112-     2.40       2.21      517.      560.     0.088194   0.000133    0.53
  109 ->   112-     2.44       2.32      509.      535.     0.089513   0.001805    0.51
  106 ->   112+     2.54       2.32      488.      535.     0.093396   0.000000    0.60
  105 ->   112+     2.57       2.56      482.      483.     0.094622   0.000000    0.61
  106 ->   112-     2.62       2.56      474.      483.     0.096150   0.000291    0.60
Message: License checkin of MS_dmol successful
```

Figure 37: Subfigure A: C[I][I]II molecule with one hydrogen atom removed emission spectra. On the left-hand side is a generated image of the molecule and on the right-hand side is the absorption spectra. Subfigure B: C[I][I]II molecule with two hydrogen atoms removed emission spectra. On the left-hand side is a generated image of the molecule and on the right hand side is the absorption spectrum.

Conclusion

Study one proves the viability of training a ML algorithm on synthetically generated data and using that algorithm to predict real work material properties. This approach is fundamental to the credibility of the following studies as it serves as proof that synthetic data is not a limiting factor for machine learning models' ability to predict on real-world data. In study one, A machine learning ANN has been designed that predicts the dielectric properties of any material inside of coaxial airline geometry using only the standard inputs of S11 and S21. The training data utilized was created synthetically using COMSOL, a well-known and tested electromagnetic simulation software. This software was chosen because it could generate input and output pairs covering the entire solution space for this problem. Synthetic data were generated for all realistic possible output scenarios. This covered all real and imaginary combinations of the dielectric constant as well as all realistic lengths of the sample.

This extensive data generation was done for two main reasons. The first was the lack of available experimental data that covered that same range of combinations. The second was to show that if the synthetic solution space is well representative of the real solution space than a ML model will be just as effective with either data source. Both styles of neural networks discussed in study one showed excellent results. The system showed exceptional performance on training datasets and experimentally collected datasets. Input data required very little prepossessing, with scaling being the only numeric manipulation done to the datasets.

It should be noted that the ANN in this study has only been tested on a few real experimental values. While the results were very promising, further testing is needed. However, it can immediately be seen from the results the benefits that this method has over the other traditional mathematical models. The most important being that solutions are not unstable at one-half wavelength, no initial guess is required, and there is no need to de-embedded air to get a more accurate measurement. Another distinct advantage that ANNs have over the mathematical models is that other variables can be included in the predictions going forward. These can include things like a chemical formula, atomic information, or any physical or chemical information.

Study one shows that with a high-fidelity model of a given geometry an ANN can be created on computational data that will allow the prediction of dielectric properties without the need to de-embed air. It should be noted that as the dielectric constant increased the ANN had a harder time predicting. This problem could be eliminated using a filtering system with multiple downstream neural networks that train on smaller ranges of data to increase accuracy within ranges of interest.

In study two, a new method for synthetic data generation of chemical equations known as AGoRaS was introduced. This study focused on AGoRaS' application for only gas-phase reactions but has applicability well beyond. The aim of AGoRaS was the avoidance of bias inherent to all training data in the generation of continuous synthetic datasets. In this study, AGoRaS was trained on a core dataset comprised of only ~7000 reactions and 2000 molecular species and was able to generate 7,000,000 reactions with 20,000 molecular species. This is an extremely high return for such a small training set. The most exciting aspect of the results of AGoRaS was its ability to create a large quantity of new molecular species and overall reactions that were stable. AGoRaS implemented a unique approach of selecting from the latent space to generate new reactions. This is a different approach as the typical usually involves retrosynthesis of existing reactions. An advantage of this approach is that the VAE gathers the knowledge of the physics and chemistry that it is trained on. This allows AGoRaS to generate new molecular species and reactions beyond the size and description of the training data.

It should not be overlooked that the generated results of AGoRaS have not been experimentally synthesized. Other studies in the past have noted that many molecules derived from generative networks cannot be synthesized[198]. This is even in light of the generated molecules scoring well on various quantitative benchmarks. While this may be the case for some of the molecules generated in the study. By using a semi-empirical technique, which has been calibrated with experiments, there is added confidence many of these reactions and molecular species can be synthesized Moreover, this study attempted to mitigate this risk by checking all molecules against existing databases such as RDKit and PubChem, in addition to conducting the semiempirical calculations. Reactions that did not pass this check with both RDKit and the semi-empirical technique were eliminated from the final dataset.

To demonstrate the utility of AGoRaS the generated reaction equations were filtered for reactions containing CO_2. Table 4 is a summary of a select number of these reactions that demonstrate the complexity of the reactions that were discovered. Note the equations displayed in Table 4 were not part of the training data. One of the limitations of AGoRaS is that the synthetic generated data still may share similar distributions and biases with the training data. However, this approach is an improvement on the current techniques as it can expand beyond the molecular description of the training dataset. As the AGoRaS is trained on larger datasets and potentially fed back generated data, the bias will diminish. It would be trivial to analyze the training dataset for biases and introduce synthetic equations that balance out the biases. The network could be retrained and sampled again to check for biases. This process could be repeated until the network is no longer producing biased outputs. While the uniqueness of the dataset can be satisfied, the completeness of the dataset will never be achieved due to the continuous nature of the data. However, this provides a tool for achieving completeness over a bound design space and satisfying the local continuity. But this may be acceptable for engineering applications where absolute knowledge is not necessary

In study three, the AGoRaS network was extended from the generation of gas-phase chemical reactions to the generation of quantum materials. This study was designed to show how an AGoRaS-style VAE could be used in other applications of chemical engineering and material science. The primary purpose was to show how with a small dataset of material with specific characteristics it would be possible to then synthetically generate a large number of new materials. Single photon emitting materials were used in this study because it is relatively quick to determine if a material possesses those characteristics. Speed and computational complexity become increasingly important as the number of generated materials begin to rise into the millions.

As in the previous study, the aim of AGoRaS-Quantum was to generate a continuous dataset that would allow for future training datasets to be unbiased. AGoRaS-Quantum was trained on a core dataset containing ~8,000 molecular species. A sampling of the latent space was stopped after ~1,000,000 new molecular species were created. This was an arbitrary stopping point and sampling could have continued until the latent space was saturated. Once again this greater than 10x return on species is an extremely high return for such a small and biased dataset. As with the previous study, AGoRaS-Quantum can create new molecules that are larger than in the original dataset.

The truly exciting aspect of the results of the AGoRaS-Quantum network was its ability to generate a large quantity of new molecular species that were both stable and shared the same defining feature as the training dataset. A nice improvement of the previous study's sampling method is the ability to use the SMILES representations of the molecular species as starting points in sampling the latent space. This will allow for targeted sampling of the latent space to generate materials with specific types of properties. This is possible due to the ability of the VAE to gather knowledge of physics and chemistry from the dataset it is trained on. This allows AGoRaS to generate new molecular species beyond the size and descriptions contained in the training data.

A further development in the third study was the use of TDDFT to provide a higher level of accuracy for the generated molecules. Additionally, the use of TDDFT allowed us to verify that the molecules would converge and therefore be more likely to be synthesizable. It also allowed for further investigation into specific molecules. In this study, we chose to show how the molecules could be manipulated to shift the peak wavelengths around the visible light spectra. Once AGoRaS-Quantum was able to generate materials that were of interest, it becomes trivial for a scientist to tune them to specific use cases.

Another exciting aspect of the results of the AGoRaS-Quantum network is its ability to generate molecules that have very strong peak values. As well as generating molecules that have peaks outside of the same wavelengths from the original data. This can be seen in figure 25 where it is clear that the AGoRaS-Quantum network is generating molecules with stronger peaks and at a wider range of frequencies. This will allow researchers to easily identify potentially excellent replacement materials for various use cases where they are using weaker materials. It will also allow them to identify materials that can strongly produce visible spectrum lights at a wider range of frequencies. Once materials with strong peaks have been "discovered" using the AGoRaS-Quantum network it is possible to manipulate them so that they generate light at the desired frequencies. This is clearly shown in the TD-DFT results for the same materials the subtractions of hydrogen atoms move the peak emission spectra into different ranges.

Again, it is important to note that these generated species have not been experimentally synthesized and therefore need to be put through continued testing. As stated earlier other methods have noted that numerous molecules created using generative networks have been unable to be physically synthesized. But as with the last study the use of a semi-empirical technique, calibrated using experimental data lends added confidence that many of these species will be chemically synthesizable. In addition the use of RDKit, PubChem, and ChemSypder[199] help to mitigate the risk that these molecules will be synthesizable.

Unlike study two, study three utilized a feedback loop of generated materials when sampling the latent space to synthesize the generated materials. This helped to reduce the bias that was present in the initial dataset. This can be seen in the histograms where the generated data has a much more uniform distribution. Of course, since the sampling was random some bias still existed in the generated dataset, but it was noticeably improved from study two. It would be possible to analyze the original and generated species and come up with a new dataset that was unbiased with normal distributions. Sampling that dataset at random would generate an unbiased dataset as an output. As before it can be assured that the dataset contains unique species. However, due to the continuous nature of the data, the completeness of the sampled latent space can never be assured.

These studies have all been designed to prove two specific points about artificial intelligence in material science. The first point is that machine learning algorithms can be trained on synthetic data to help researchers to make important and impactful breakthroughs in the field of material science. The second main focus is that through the use of a langue-based VAE model and careful data curation it is possible to synthetically produce new and interesting datasets. This can ultimately lead to more advanced machine learning models and even faster development and deployment of machine learning algorithms on a large scale in material science.

Moreover, this research has focused on how machine learning, and especially VAE can be used to overcome the data shortcomings that are currently holding back the advancement of machine learning in the field of materials research. It has shown that a VAE can synthetically reproduce chemical equations and species. Even being able to be tuned to develop specific types of species for practical use cases. It has accomplished this by showing how a small well-curated material dataset can be used to generate a much larger and more diverse dataset. This dissertation has shown how a VAE can take SMILES formatted textual data and learn the underlying chemistry and physics in order to generate new materials. This will help material research to start making the same types of groundbreaking advancements that have been seen in the past few years in fields like image processing and NLP where labeled data is abundant and high quality.

Now that the applicability and feasibility of these language based networks for material science have been proven it opens the possibilities for more in depth analysis. Some traditional machine learning analysis could be preformed in order to gain a better understanding of the underlying processes. Mainly things such as the covariant estimates of the different parameters within the network. This would also help with determining overfitting of the latent space via the network's variance. Some other interesting approaches that could be taken to improve the networks speed and efficiency, is the autonomous design of the network parameters. This study was based on hand tuned parameters until a stable network could be created. This leaves great opportunity for the design of a more memory efficient network.

Future Work

Going forward several important areas should be addressed to make the studies here more applicable and more accessible. The first problem to address would be the time and computational intensity of the semi-empirical calculations. Processing either dataset took months of almost continuous calculations to generate the necessary information to compare the datasets. Going forward the first step would be to build a machine learning model that can predict the chemical properties of interest. Since we now have two very large diverse datasets with multiple properties calculated for various species. It is an ideal scenario to continue testing if the synthetically generated data can be used to train machine learning models. The main properties of interest would be the Gibbs free energy at room temperature and the dipole moment for gas phase species. For quantum materials, the properties would be the wavelength that produces the strongest peak and its strength.

Many different avenues can be taken going forward to increase the applicability of this study to experimentalists. One of the simplest would be to build in more checks for the post-processing of the generated materials. In such a way that materials outside of the realm of experimentally creatable materials would not be part of the generated dataset. This could be done with a combination of cheminformatics tools and subject matter experts to craft the more extensive rules needed to filter the dataset. Another easy-to-implement step could for the reactions dataset is to curate the dataset to contain only “realistic” reactions based on the Gibbs free energy.

Looking particularly at the network used for predicting the quantum material's properties, it would be possible to create an extremely powerful reinforcement learning algorithm. Since reinforcement learning is outside of the scope of this dissertation I will not go into great detail on the inner workings. In its most basic form, it would be possible to create a reinforcement learning algorithm that optimized for both the stable creation of molecular species but also for the wavelength and strength at which it produced light. It would be possible to almost instantly check if a species was real and stable using RDKit. While then predicting its peak strength and wavelength using the previously developed machine learning model. This would allow for the development of a network that could perform targeted generation of new species.

An additional avenue for future work in regard to the AGoRaS-Quantum is the promising materials identified in study three. All of the materials were Iodine based materials with excellent peak strength and similar emittance wavelengths. Due to the strong dipoles associated with some of them, they make an excellent candidate for quantum sensing applications. The high dipoles allow for a potential materials system of these molecules to trade an electron back and forth at a resonance known as the Rabi frequency[71,197]. This allows for a system of SPS to continuously emit light until there is a change in the environment. The TDDFT analysis will allow for further verification and testing of these molecules. Which can help experimentalists quickly work at creating these materials. The AGoRaS-Quantum network can of course continue to be sampled using these promising materials to generate additional Iodine based materials with similar properties. That can then be further analyzed using TDDFT to analyze their feasibility and to experiment with different configurations.

Both of the variational autoencoders used in studies two and three were trained on what is considered small data. An interesting avenue of research going forward would be to monitor the failed generation of molecules or reactions. To be able to use them to help restructure the latent space, so that the probabilities of steady materials would be increased. As of now those reactions and species are just eliminated from the list of stable outputs. But they could be used as guidance for where the latent space is unstable. This would be an interesting research approach that has not been attempted to stabilize the latent space post-training.

Another important step to take to increase the impact of these studies is to make the data as widely accessible and available as possible. To that end due to the author's connections on a professional level, several software engineers have become interested in this project and would like to help make it an open-source application. This would allow an easy to use interface for accessing the data as well as searching through the various structures. It will also allow for all data to be downloadable either as a bulk download or a curated download. The application will also allow other research institutions to upload their real or experimental reactions. Putting an emphasis on uploading any and all experimental reactions, not just the best ones. Ideally, it will also host a version of AGoRaS where people can train it using their own data. This application will help to push the field of materials research forward by making data widely available to anyone who wants to use it.

Code Availability

The machine learning models are available in GitHub repositories with accompanying filtered experimental datasets and semi-empirical datasets. Unfiltered datasets and Pipeline Pilot protocols used to generate filtered datasets can be requested on GitHub. All code can be found on GitHub under the following link, https://github.com/Dr-Musho-Research-Group/AGoRaS. Future releases and versions of AGoRaS will also be released here.

References

(1) Marr, B. *The Most Amazing Artificial Intelligence Milestones So Far*; **2018**.

(2) Allen, D. M. W. J. R. *How Artificial Intelligence Is Transforming the World*; **2018**.

(3) Jones, L. D.; Golan, D.; Hanna, S. A.; Ramachandran, M. Artificial Intelligence, Machine Learning and the Evolution of Healthcare: A Bright Future or Cause for Concern? *Bone Jt. Res.* **2018**, *7* (3), 223–225.

(4) Zhang, A. How Does AI Reshape Material Science? *Medium*. **2021**.

(5) Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. Accelerating Materials Discovery Using Artificial Intelligence, High Performance Computing and Robotics. *npj Comput. Mater.* **2022**, *8* (1).

(6) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3*, 1–6.

(7) Higham, D. J. Modeling and Simulating Chemical Reactions. *SIAM Rev.* **2008**, *50* (2), 347–368.

(8) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610.

(9) Furthmüller, J. VAMP - Vienna Ab Initio Molecular Dynamics Package. **1994**.

(10) Zhang, Y.; Ling, C. A Strategy to Apply Machine Learning to Small Datasets in Materials Science. *npj Comput. Mater.* **2018**, *4* (1), 28–33.

(11) Shaik, A. B. D. al jalali wal ikram; Palla, P. Optical Quantum Technologies with Hexagonal Boron Nitride Single Photon Sources. *Sci. Rep.* **2021**, *11* (1), 1–27.

(12) NIST. Materials Genome Initiative https://www.nist.gov/mgi.

(13) Berkeley Lab. The Materials Project https://materialsproject.org/.

(14) Tempke, R.; Thomas, L.; Wildfire, C.; Shekhawat, D.; Tempke, R.; Thomas, L.; Wildfire, C.; Shekhawat, D. Machine Learning Approach to Transform Scattering Parameters to Complex Permittivities. **2021**.

(15) Kayala, M. A.; Baldi, P.; Rajeswar, S.; Subramanian, S.; Dutil, F.; Pal, C.; Courville, A.; Zhao, J.; Kim, Y.; Zhang, K.; Rush, A. M.; LeCun, Y.; Catherine, H.; Cook, M. L.; Mckone, E.; Griffiths, R. R.; Schwaller, P.; Lee, A. A.; Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B.; Cova, T. F. G. G.; Pais, A. A. C. C.; Mater, A. C.; Coote, M. L.; Kusner, M. J.; Hernández-Lobato, J. M.; Camino, R. D.; Hammerschmidt, C. A.; State, R.; Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W. F.; Sun, J.; Potash, P.; Rumshisky, A.; Camino, R. D.; Hammerschmidt, C. A.; State, R.; Kayala, M. A.; Azencott, C. A.; Chen, J. H.; Baldi, P.; McCutchen, C.; Schmidt, J.; Marques, M. A. L. M. R. G.; Botti, S.; Marques, M. A. L. M. R. G.; Kovács, D. P.; McCorkindale, W.; Lee, A. A.; Kang, P. L.; Liu, Z. P.; Kayala, M. A.; Baldi, P.; Yu, L.; Zhang, W.; Wang, J.; Yu, Y. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns.

arXiv **2019**, *68* (1), 1–9.

(16) Li, Z.; Ma, X.; Xin, H. Feature Engineering of Machine-Learning Chemisorption Models for Catalyst Design. *Catal. Today* **2017**, *280*, 232–238.

(17) Kang, P. L.; Liu, Z. P. Reaction Prediction via Atomistic Simulation: From Quantum Mechanics to Machine Learning. *iScience* **2021**, *24* (1), 102013.

(18) Kayala, M. A.; Baldi, P. A Machine Learning Approach to Predict Chemical Reactions. *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011* **2011**, 1–9.

(19) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using Transformer-Based Models and a Hyper-Graph Exploration Strategy. *Chem. Sci.* **2020**, *11* (12), 3316–3325.

(20) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583.

(21) Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative Interpretation Explains Machine Learning Models for Chemical Reaction Prediction and Uncovers Bias. *Nat. Commun.* **2021**, *12* (1), 1–9.

(22) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23* (6), 1241–1250.

(23) Qiao, Z.; Wang, Z.; Zhang, C.; Yuan, S.; Zhu, Y.; Wang, J.; Yang, W.; Fidelis, T. T.; Sun, W. H.; Williams, T.; McCullough, K.; Lauterbach, J. A. Machine Learning in Catalysis, from Proposal to Practicing. *ACS Omega* **2020**, *32* (1), 157–165.

(24) Yang, W.; Fidelis, T. T.; Sun, W. H. Machine Learning in Catalysis, from Proposal to Practicing. *ACS Omega* **2020**, *5* (1), 83–88.

(25) Goldsmith, B. R.; Esterhuizen, J.; Liu, J. X.; Bartel, C. J.; Sutton, C. Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE J.* **2018**, *64* (7), 2311–2323.

(26) Zangwill, A. *Modern Electrodynamics*; Cambridge University Press: Cambridge, **2013**.

(27) Busseyt, H. E. Measurement and Standardization of DieIectric Sam&; **1962**, No. 4, 4–7.

(28) Bussey, H. E. Measurement of RF Properties of Materials A Survey. *Proc. IEEE* **1967**, *55* (6), 1046–1053.

(29) Nicolson, A. M.; Ross, G. F. Measurement of the Intrinsic Properties of Materials by Time Domain Techniques. *IEEE Trans. Instrum. Meas.* **1970**, *19* (4), 377–382.

(30) Blakney, T. L.; Weir, W. B. Automatic Measurement of Complex Dielectric Constant and Permeability at Microwave Frequencies. *Proc. IEEE* **1975**, *63* (1), 203–205.

(31) Yaw, K. C. Measurement of Dielectric Material Properties Application Note, Rohde & Schwarz. *Meas. Tech.* **2006**, 1–35.

(32) Schwab, J.; Antholzer, S.; Haltmeier, M. Deep Null Space Learning for Inverse Problems:

Convergence Analysis and Rates. *Inverse Probl.* **2019**, *35* (2).

(33) Sahoo, S.; Lampert, C.; Martius, G. Learning Equations for Extrapolation and Control. *Proc. 35th Int. Conf. Mach. Learn.* **2018**, *80*, 4442–4450.

(34) Zhao, Z.; Qu, Y.; Zhang, Y.; Shen, X. Overview of Microwave Device Modeling Techniques Based on Machine Learning. *Int. J. Adv. Comput. Technol.* **2013**, *5* (9), 299–306.

(35) Mueller, T.; Kusne, A.; Ramprasad, R. Science : Recent Progress. *Rev. Comput. Chem.* **2016**, *29* (i).

(36) Baker-jarvis, J.; Janezic, M. D.; Domich, P. D.; Geyer, R. G. Analysis of an Open-Ended Coaxial Probe with Lift-Off for Nondestructive Testing. **1994**, *43* (5), 1–8.

(37) Bois, K. J. Analysis of an Open-Ended Coaxial Probe with Lift-off for Nondestructive Testing. *IEEE Trans. Instrum. Meas.* **1999**, *48* (6), 1141–1148.

(38) Geyer, R. G.; Grosvenor, C. A.; Holloway, C. L.; Janezic, M. D.; Johk, R. T.; Kabos, P.; Baker-Jarvis, J. *Measuring the Permittivity and Permeability of Lossy Materials :*; Boulder, 2005.

(39) Bartley, P. G.; Begley, S. B. A New Technique for the Determination of the Complex Permittivity and Permeability of Materials. *2010 IEEE Int. Instrum. Meas. Technol. Conf. I2MTC 2010 - Proc.* **2010**, 54–57.

(40) Nigrin, A. *Neural Networks for Pattern Recognition*, 1st ed.; MIT Press: Massachusetts, **1993**.

(41) E, W.; Han, J.; Jentzen, A. Algorithms for Solving High Dimensional PDEs: From Nonlinear Monte Carlo to Machine Learning. **2020**, 1–40.

(42) Li, L.; Rong, S.; Wang, R.; Yu, S. Recent Advances in Artificial Intelligence and Machine Learning for Nonlinear Relationship Analysis and Process Control in Drinking Water Treatment: A Review. *Chem. Eng. J.* **2021**, *405* (June 2020).

(43) Wilson, Z. T.; Sahinidis, N. V. The ALAMO Approach to Machine Learning. *Comput. Chem. Eng.* **2017**, *106*, 785–795.

(44) Tang, Y.; Kurths, J.; Lin, W.; Ott, E.; Kocarev, L. Introduction to Focus Issue: When Machine Learning Meets Complex Systems: Networks, Chaos, and Nonlinear Dynamics. *Chaos* **2020**, *30* (6).

(45) Ryo, M.; Rillig, M. C. Statistically Reinforced Machine Learning for Nonlinear Patterns and Variable Interactions. *Ecosphere* **2017**, *8* (11).

(46) Li, H.; Cao, J.-N.; Love, P. E. . USING MACHINE LEARNING AND GA TO SOLVE TIME-COST TRADE-OFF PROBLEMS. **1999**, *125* (October), 347–353.

(47) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2015**, *07-12-June*, 1–9.

(48) Scott, D. J.; Coveney, P. V.; Kilner, J. A.; Rossiny, J. C. H.; Alford, N. M. N. Prediction of the Functional Properties of Ceramic Materials from Composition Using Artificial Neural Networks. *Journal of the European Ceramic Society*. **2007**, pp 4425–4435.

(49) Raff, L.; Komanduri, R.; Hagan, M.; Bukkapatnam, S. Applications of Neural Network Fitting of Potential -Energy Surfaces. In *Neural Networks in Chemical Reaction Dynamics*; Oxford University Press: Oxford, **2012**; p 283.

(50) Schwaller, P.; Hoover, B.; Reymond, J. L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Sci. Adv.* **2021**, *7* (15).

(51) Li, Y.; Yuan, Y. Convergence Analysis of Two-Layer Neural Networks with ReLU Activation. In *Advances in Neural Information Processing Systems 30*; NeurIPS Proceedings: Long Beach, **2017**; pp 1–11.

(52) Guo, D.; Wang, Y.; Xia, J.; Nan, C.; Li, L. Investigation of BaTiO3 Formulation: An Artificial Neural Network (ANN) Method. *Journal of the European Ceramic Society*. 2002, pp 1867–1872.

(53) Agostinelli, F.; Hoffman, M.; Sadowski, P.; Baldi, P. Learning Activation Functions to Improve Deep Neural Networks. In *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*; ICLR 2015: San Diego, **2015**; pp 1–9.

(54) Curry, B.; Morgan, P. H. Model Selection in Neural Networks: Some Difficulties. *European Journal of Operational Research*. **2006**, pp 567–577.

(55) Dong, Y.; Liu, J.; Liu, Y.; Li, H.; Zhang, X.; Hu, X. Creep–Fatigue Experiment and Life Prediction Study of Piston 2a80 Aluminum Alloy. *Materials (Basel).* **2021**, *14* (6).

(56) Xue, D.; Xue, D.; Yuan, R.; Zhou, Y.; Balachandran, P. V.; Ding, X.; Sun, J.; Lookman, T. An Informatics Approach to Transformation Temperatures of NiTi-Based Shape Memory Alloys. *Acta Mater.* **2017**, *125*, 532–541.

(57) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5* (1).

(58) Tuck, D.; Coad, S. Neurocomputed Model of Open-Circuited Coaxial Probes. *Ind. Res.* **1995**, *5* (4), 5–7.

(59) Stuchly, M. A.; Stuchly, S. S. Coaxial Line Reflection Methods for Measuring Dielectric Properties of Biological Substances at Radio and Microwave Frequencies—A Review. *J. Microw. Power* **1980**, *43* (3), 165–172.

(60) Gajda, G.; Stuchly, S. S. An Equivalent Circuit Ofan Open-Ended Coaxial Line. *IEEE Trans. Instrum. MEASIJREMENT*, **1983**, *IM* (4), 367–368.

(61) Chen, Q.; Huang, K.-M.; Yang, X.; Luo, M.; Zhu, H. AN ARTIFICIAL NERVE NETWORK REALIZATION IN THE MEASUREMENT OF MATERIAL PERMITTIVITY. *Prog. Electromagn. Res.* **2011**, *116* (April), 347–361.

(62) Metlek, S.; Kayaalp, K.; Basyigit, I. B.; Genc, A.; Dogan, H. The Dielectric Properties Prediction of the Vegetation Depending on the Moisture Content Using the Deep Neural Network Model. *Int. J. RF Microw. Comput. Eng.* **2021**, *31* (1), 1–10.

(63) Takahashi, A.; Kumagai, Y.; Miyamoto, J.; Mochizuki, Y.; Oba, F. Machine Learning Models for Predicting the Dielectric Constants of Oxides Based on High-Throughput First-Principles Calculations. *Phys. Rev. Mater.* **2020**, *4* (10).

(64) Qin, J.; Liu, Z.; Ma, M.; Li, Y. Machine Learning Approaches for Permittivity Prediction and Rational Design of Microwave Dielectric Ceramics. *J. Mater.* **2021**, *7* (6), 1284–1293.

(65) Vainshtein, L. A. Electromagnetic Waves. *Izd. Radio i Sviaz* **1988**, *00*, 440.

(66) Wikipedia. Electromagnetic Radiation https://en.wikipedia.org/wiki/Electromagnetic_radiation.

(67) Purcell, E. M. *Electricity and Magnetism*, Second.; McGraw-Hill Book Company: New York, **1985**.

(68) Horikoshi, S.; Schiffmann, R. F.; Fukushima, J.; Serpone, N. *Microwave Chemical and Materials Processing: A Tutorial*; **2017**.

(69) Stuerga, D. *Microwave-Material Interactions and Dielectric Properties, Key Ingredients for Mastery of Chemical Microwave Processes*; **2008**; Vol. 1.

(70) Surati, M. A.; Jauhari, S.; Desai, K. R. A Brief Review : Microwave Assisted Organic Reaction. *Arch. Appl. Sci. Res.* **2012**, *4* (1), 645–661.

(71) Fung, A. K.; Chen, K.-S.; Schutzer, and D. *Microwave Scattering and Emission Models for Users*, 1st ed.; Artech House: Norwood, **2010**.

(72) Sugawara, H.; Kashimura, K.; Hayashi, M.; Matsumuro, T.; Watanabe, T.; Mitani, T. Temperature Dependence and Shape Effect in High-Temperature Mi- Crowave Heating of Nickel Oxide Powders. *Phys. B Phys. Condens. Matter* **2015**, *458*, 35–39.

(73) Mishra, R. R.; Sharma, A. K. Microwave-Material Interaction Phenomena: Heating Mechanisms, Challenges and Opportunities in Material Processing. *Composites Part A: Applied Science and Manufacturing*. **2016**, pp 78–97.

(74) Musho, T. D.; Wildfire, C.; Houlihan, N. M.; Sabolsky, E. M.; Shekhawat, D. Study of Cu2O Particle Morphology on Microwave Field Enhancement. *Materials Chemistry and Physics*. **2018**, pp 278–284.

(75) Arthur R. Von Hipple. *Dielectric Materals and Applications*; Arthur R. Von Hipple, Ed.; The MIT Press, **1952**.

(76) Arthur R. Von Hipple, S. Ramo, J. R. W. Dielectrics and Waves. *Fields and Waves in Mordern Radio* **1955**, *8* (2).

(77) Anderson, R. W. *S-Parameter Techniques For Faster, More Accurate Network Design*; **1967**.

(78) Agilent Technologies. *Advanced Design System 1.5 Circuit Simulatione*; Palo Alto, **2000**.

(79) Jilani, M. T.; Rehman, M. Z. ur; Khan, A. M.; Khan, M. T.; Ali, S. M. A Brief Review of Measuring Techniques for Characterization of Dielectric Materials. *Int. J. Inf. Technol. Electr. Eng.* **2012**, *1* (1), 1–5.

(80) Venkatesh, M. S.; Raghavan, G. S. V. An Overview of Dielectric Properties Measuring Techniques. *J. Can. Soc. Bioeng.* **2005**, *47* (40), 7.15-7.30.

(81) Technologies, K. N5231A PNA-L Microwave Network Analyzer, 13.5 GHz https://www.keysight.com/main/techSupport.jspx?cc=US&lc=eng&nid=-32497.1150488&pid=x201914&pageMode=PL.

(82) Karl, H.; Oszkár, B.; Gernot, M.; Kurt, P.; Christian, S.; Bernhard, W.; Peter, C.; Gerhard, P. Comparison of FETD and FDTD to Simulate Micro-Strip Structures on PCBs. *6th Int. Conf. Comput. Electromagn. CEM 2006 - Proc.* **2006**, 209–210.

(83) Soares, D.; Mansur, W. J.; Tsai, H. S.; York, R. A.; Meng, H. T.; Nie, B. L.; Wong, S.; Macon, C.; Jin, J. M.; Feng, J.; Santamouris, M. FDTD Analysis of CPW-Fed Folded-Slot and Multiple-Slot Antennas on Thin Substrates. *IEEE Trans. Antennas Propag.* **1996**, *44* (2), 217–226.

(84) Joannopoulos, J.; Johnson, S.; Winn, J.; Meade, R. *Photonic Crystals: Molding the Flow of Light Second Edition*, Second Edi.; Princeton University Press, **2008**.

(85) Meng, H. T.; Nie, B. L.; Wong, S.; Macon, C.; Jin, J. M. GPU Accelerated Finite-Element Computation for Electromagnetic Analysis. *IEEE Antennas Propag. Mag.* **2014**, *56* (2), 39–62.

(86) Feng, J.; Santamouris, M. Numerical Techniques for Electromagnetic Simulation of Daytime Radiative Cooling: A Review. *AIMS Mater. Sci.* **2019**, *6* (6), 1049–1064.

(87) COMSOL AB. EM Module. COMSOL: Stockholm, Sweden.

(88) Ramachandran, P.; Zoph, B.; Le, Q. V. Searching for Activation Functions. *arXiv* **2017**, *1* (1), 1–13.

(89) Maas, A. L.; Hannun, A. Y.; Ng, A. Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. *Proc. 30 th Int. Conf. Mach. Learn.* **2013**, *30* (1), 3.

(90) Narang, S.; Elsen, E.; Diamos, G.; Sengupta, S. Exploring Sparsity in Recurrent Neural Networks. *Int. Conf. Learn. Represent.* **2017**, *1* (1), 1–10.

(91) Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. *arXiv* **2012**, *1* (1), 1–6.

(92) Mittlböck, M.; Schemper, M. Explained Variation for Logistic Regression. *Stat. Med.* **1996**, *15* (19), 1987–1997.

(93) Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. Applying Machine Learning Techniques to Predict the Properties of Energetic Materials. *Sci. Rep.* **2018**, *8* (1), 1–12.

(94) Amini, A.; Schwarting, W.; Rosman, G.; Araki, B.; Karaman, S.; Rus, D. Variational Autoencoder for End-to-End Control of Autonomous Driving with Novelty Detection and Training De-Biasing. *IEEE Int. Conf. Intell. Robot. Syst.* **2018**, 568–575.

(95) Iovanac, N. C.; Savoie, B. M. Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *J. Phys. Chem. A* **2019**, *123* (19), 4295–4302.

(96) Chen, L. Continuous Delivery. **2016**, 84–84.

(97) Akoglu, A.; Vargas-Solar, G. Putting Data Science Pipelines on the Edge. **2021**, 1–13.

(98) Rovinelli, A.; Sangid, M. D.; Proudhon, H.; Ludwig, W. Using Machine Learning and a Data-Driven Approach to Identify the Small Fatigue Crack Driving Force in Polycrystalline Materials. *npj Comput. Mater.* **2018**, *4* (1), 1–10.

(99) Anyoha, R. *The History of Artificial Intelligence*; **2017**.

(100) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573* (7773), 251–255.

(101) Carroll, H. A.; Toumpakari, Z.; Johnson, L.; Betts, J. A. The Perceived Feasibility of Methods to Reduce Publication Bias. *PLoS One* **2017**, *12* (10), 1–19.

(102) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc.* **2014**, No. Ml, 1–14.

(103) Griffiths, R. R.; Schwaller, P.; Lee, A. A. Dataset Bias in the Natural Sciences: A Case Study in Chemical Reaction Prediction and Synthesis Design. *ChemRxiv* **2018**.

(104) Rose, L. T.; Fischer, K. W. Garbage In, Garbage Out: Having Useful Data Is Everything. *Measurement* **2011**, *9* (4), 222–226.

(105) Sanders, H.; Saxe, J. Garbage in, Garbage out (How Purportedly Great ML Models Can Be Screwed up by Bad Data). *Proc. Blackhat 2017* **2017**, 6.

(106) Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's Chemical Diversity Limits the Generalizability of Machine Learning Predictions. *J. Cheminform.* **2019**, *11* (1), 1–15.

(107) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52* (10), 2526–2540.

(108) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**.

(109) Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W. F.; Sun, J. Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks. *arXiv* **2017**, *68*, 1–20.

(110) Camino, R. D.; Hammerschmidt, C. A.; State, R. Generating Multi-Categorical Samples with Generative Adversarial Networks. *arXiv* **2018**.

(111) Zhao, J.; Kim, Y.; Zhang, K.; Rush, A. M.; LeCun, Y. Adversarially Regularized Autoencoders. *35th Int. Conf. Mach. Learn. ICML 2018* **2018**, *13*, 9405–9420.

(112) Kusner, M. J.; Paige, B.; Miguel Hernández-Lobato, J. Grammar Variational Autoencoder. *arXiv* **2017**.

(113) Kusner, M. J.; Hernández-Lobato, J. M. GANS for Sequences of Discrete Elements with the Gumbel-Softmax Distribution. **2016**, 1–6.

(114) Burks, R.; Islam, K. A.; Lu, Y.; Li, J. Data Augmentation with Generative Models for Improved Malware Detection: A Comparative Study. *2019 IEEE 10th Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf. UEMCON 2019* **2019**, No. October, 0660–0665.

(115) Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *31st AAAI Conf. Artif. Intell. AAAI 2017* **2017**, 2852–2858.

(116) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted Noise-Reduction of Chemical Reactions Data Sets. *Nat. Mach. Intell.* **2021**, *3*, 485–494.

(117) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96.

(118) Gale, E. M.; Durand, D. J. Improving Reaction Prediction. *Nat. Chem.* **2020**, *12* (6), 509–510.

(119) NAG, P. *Engineering Thermodynamics*, 6th ed.; MC Graw Hill India, **2017**.

(120) Seddon, J.; Gale, J. *Thermodynamics and Statistical Mechanics*, 1st ed.; Wiley-RSC, 2002.

(121) Armandi, M.; Garrone, E.; Areán, C. O.; Bonelli, B. Thermodynamics of Carbon Dioxide Adsorption on the Protonic Zeolite H-ZSM-5. *ChemPhysChem* **2009**, *10* (18), 3316–3319.

(122) Frank Ogletree, D.; Bluhm, H.; Hebenstreit, E. D.; Salmeron, M. Photoelectron Spectroscopy under Ambient Pressure and Temperature Conditions. *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.* **2009**, *601* (1–2), 151–160.

(123) Ghassemi, M.; Shahidian, A. *Nano and Bio Heat Transfer and Fluid Flow*; Academic Press: London, 2017.

(124) Scott, S. K.; Johnson, B. R.; Taylor, A. F.; Tinsley, M. R. Complex Chemical Reactions - a Review. *Chem. Eng. Sci.* **2000**, *55* (2), 209–215.

(125) Kayala, M. A.; Azencott, C. A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51* (9), 2209–2222.

(126) Talanquer, V. Macro, Submicro, and Symbolic: The Many Faces of the Chemistry "Triplet." *Int. J. Sci. Educ.* **2011**, *33* (2), 179–195.

(127) Zumdahl, S.; Zundahl, S. *Chemistry*, 5th ed.; Boston : Houghton Mifflin: Boston, 2000.

(128) Walsh, R.; Morgan, D. H.; Bollmann, A.; Dixon, J. T. Reaction Kinetics of an Ethylene Tetramerisation Catalyst. *Appl. Catal. A Gen.* **2006**, *306*, 184–191.

(129) Stains, M.; Talanquer, V. Classification of Chemical Reactions: Stages of Expertise. *J. Res. Sci. Teach.* **2008**, *45* (7), 771–793.

(130) Nikitin, E. E. *Theory of Thermally Induced Gas Phase Reactions*, 17th ed.; Indiana

University Press.: Bloomington, **1966**.

(131) Child, M. S. *Molecular Collision Theory*, 4th ed.; Academic Press: London, 1974.

(132) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.

(133) He, J.; You, H.; Sandström, E.; Nittinger, E.; Bjerrum, E. J.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Molecular Optimization by Capturing Chemist's Intuition Using Deep Neural Networks. *J. Cheminform.* **2021**, *13* (1), 1–17.

(134) Karwath, A.; De Raedt, L. SMIREP: Predicting Chemical Activity from SMILES. *J. Chem. Inf. Model.* **2006**, *46* (6), 2432–2444.

(135) Jäger, G.; Rogers, J. Formal Language Theory: Refining the Chomsky Hierarchy. *Philos. Trans. R. Soc. B Biol. Sci.* **2012**, *367* (1598), 1956–1970.

(136) Grenier, W.; Vinokurskaya, M. Canonical SMILES for Context Free Grammars.

(137) Leach, A.; Valerie, G. *An Introduction to Chemoinformatics*, Rev.; Springer International Publishing: London, **2007**.

(138) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58* (1), 27–35.

(139) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.* **2013**, 1–12.

(140) Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, Products, and Transition States of Elementary Chemical Reactions Based on Quantum Chemistry. *Sci. Data* **2020**, *7* (1), 1–8.

(141) Saxén, H.; Pettersson, F. Method for the Selection of Inputs and Structure of Feedforward Neural Networks. *Comput. Chem. Eng.* **2006**, *30* (6–7), 1038–1045.

(142) Prusa, J. D.; Khoshgoftaar, T. M. Improving Deep Neural Network Design with New Text Data Representations. *J. Big Data* **2017**, *4* (1).

(143) Manion, J. A.; Huie, R. E.; Levin, R. D.; Burgess Jr, D. R.; Orkin, V. L.; Tsang, W.; McGivern, W. S.; Hudgens, J. W.; Knyazev, V. D.; Atkinson, D. B.; Chai, E.; Tereza, A. M.; Lin, C.-Y.; Allison, T. C.; Mallard, W. G.; Westley, F.; Herron, J. T.; Hampson, R. F.; Frizzell, D. H. NIST Chemical Kinetics Database https://kinetics.nist.gov/.

(144) McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*; Van Der Walt, S., Millma, J., Eds.; **2010**; pp 56–61.

(145) Pinheiro, G. A.; Mucelini, J.; Soares, M. D.; Prati, R. C.; Da Silva, J. L. F.; Quiles, M. G. Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset. *J. Phys. Chem. A* **2020**, *124* (47), 9854–9866.

(146) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-

Learning of Chemical Structures. *ChemRxiv* **2018**, 1–9.

(147) Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional Neural Network Based on SMILES Representation of Compounds for Detecting Chemical Motif. *BMC Bioinformatics* **2018**, *19* (Suppl 19).

(148) Honda, S.; Shi, S.; Ueda, H. R. SMILES Transformer: Pre-Trained Molecular Fingerprint for Low Data Drug Discovery. **2019**.

(149) NCI; CADD; Chemical Biology Laboratory. CADD Group Chemoinformatics Tools and User Services.

(150) Kim, S.; Chen, J.; Chen, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.; Thiessen, P.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. PubChem in 2021: new data content and improved web interfaces.

(151) rdkit.org. RDKit:Open-source cheminformatics http://www.rdkit.org/.

(152) Cova, T. F. G. G.; Pais, A. A. C. C. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Front. Chem.* **2019**, *7* (November), 1–22.

(153) Dwarampudi, M.; Reddy, N. V. S. Effects of Padding on LSTMs and CNNs. **2019**.

(154) Gajendran, S.; Manjula, D.; Sugumaran, V. Character Level and Word Level Embedding with Bidirectional LSTM – Dynamic Recurrent Neural Network for Biomedical Named Entity Recognition from Literature. *J. Biomed. Inform.* **2020**, *112*.

(155) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Greg, C.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jozefowicz, R.; Jia, Y.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Schuster, M.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Viégas, V.; Fernanda, V.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. Large-Scale Machine Learning on Heterogeneous Systems. **2021**, 1-8

(156) Gaspar, H. A.; Ahmed, M.; Edlich, T.; Fabian, B.; Varszegi, Z.; Segler, M.; Meyers, J.; Fiscato, M. Proteochemometric Models Using Multiple Sequence Alignments and a Subword Segmented Masked Language Model. *ChemRxiv* **2021**, 1–10.

(157) Jinich, A.; Sanchez-Lengeling, B.; Ren, H.; Harman, R.; Aspuru-Guzik, A. A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315 »000 Redox Reactions. *ACS Cent. Sci.* **2019**, *5* (7), 1199–1210.

(158) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv* **2018**.

(159) Toropov, A. A.; Toropova, A. P.; Benfenati, E. QSPR Modeling for Enthalpies of Formation of Organometallic Compounds by Means of SMILES-Based Optimal Descriptors. *Chem. Phys. Lett.* **2008**, *461* (4–6), 343–347.

(160) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification

of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13* (12), 1173–1213.

(161) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60* (12).

(162) Beard, E. J.; Sivaraman, G.; Vázquez-Mayagoitia, Á.; Vishwanath, V.; Cole, J. M. Comparative Dataset of Experimental and Computational Attributes of UV/Vis Absorption Spectra. *Sci. Data* **2019**, *6* (1), 1–11.

(163) Commerce, U. S. D. of. NIST Chemistry WebBook https://webbook.nist.gov/chemistry/.

(164) Sun, X.; Wang, P.; Wang, T.; Chen, L.; Chen, Z.; Gao, K.; Aoki, T.; Li, M.; Zhang, J.; Schulz, T.; Albrecht, M.; Ge, W.; Arakawa, Y.; Shen, B.; Holmes, M.; Wang, X. Single-Photon Emission from Isolated Monolayer Islands of InGaN. *Light Sci. Appl.* **2020**, *9* (1).

(165) Slachter, A. Single Photon Emitters. *Sci. Rep*. **2003**, 1-11

(166) Jiandong Qiao, J. Q.; Fuhong Mei, F. M.; Yu Ye, Y. Y. Single-Photon Emitters in van Der Waals Materials. *Chinese Opt. Lett.* **2019**, *17* (2), 020011.

(167) Moerner, W. E.; Levenson, M. D. Can Single-Photon Processes Provide Useful Materials for Frequency-Domain Optical Storage? *J. Opt. Soc. Am. B* **1985**, *2* (6), 915.

(168) Go, S.; Kim, J.; Park, S. S.; Kim, M.; Lim, H.; Kim, J. Y.; Lee, D. W.; Im, J. Synergistic Use of Hyperspectral Uv-Visible Omi and Broadband Meteorological Imager Modis Data for a Merged Aerosol Product. *Remote Sens.* **2020**, *12* (23), 1–34.

(169) Ni, K. K.; Rosenband, T.; Grimes, D. D. Dipolar Exchange Quantum Logic Gate with Polar Molecules. *Chem. Sci.* **2018**, *9* (33), 6830–6838.

(170) Spangenberg, M.; Bryant, J. I.; Gibson, S. J.; Mousley, P. J.; Ramachers, Y.; Bell, G. R. Ultraviolet Absorption of Contaminants in Water. *Sci. Rep.* **2021**, *11* (1), 1–8.

(171) Zakutayev, A.; Wunder, N.; Schwarting, M.; Perkins, J. D.; White, R.; Munch, K.; Tumas, W.; Phillips, C. An Open Experimental Database for Exploring Inorganic Materials. *Sci. Data* **2018**, *5*, 1–12.

(172) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894–1904.

(173) Mamede, R.; Pereira, F.; Aires-de-Sousa, J. Machine Learning Prediction of UV–Vis Spectra Features of Organic Compounds Related to Photoreactive Potential. *Sci. Rep.* **2021**, *11* (1), 1–11.

(174) De Leonardis, F.; Soref, R. A.; Soltani, M.; Passaro, V. M. N. Broadband Biphoton Generation and Statistics of Quantum Light in the UV-Visible Range in an AlGaN Microring Resonator. *Sci. Rep.* **2017**, *7* (1), 1–9.

(175) Richter, A.; Wagner, T. *Solar Backscattered Radiation: UV, Visible and Near IR - Trace Gases*; **2011**.

(176) Kärkkäinen, K. K.; Sihvola, A. H.; Nikoskinen, K. I. Effective Permittivity of Mixtures: Numerical Validation by the FDTD Method. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38* (3), 1303–1308.

(177) Fei, C.; Cao, X.; Zang, D.; Hu, C.; Wu, C.; Morris, E.; Tao, J.; Liu, T.; Lampropoulos, G. Machine Learning Techniques for Real-Time UV-Vis Spectral Analysis to Monitor Dissolved Nutrients in Surface Water. **2021**, No. March 2021, 46.

(178) Cava, R.; Leon, N.; Xie, W. Introduction: Quantum Materials. *Chem. Rev.* **2021**, No. 121, 2777–2779.

(179) Dral, P. O.; Wu, X.; Thiel, W. Semiempirical Quantum-Chemical Methods with Orthogonalization and Dispersion Corrections. *J. Chem. Theory Comput.* **2019**, *15* (3), 1743–1760.

(180) Khazali, M.; Heshami, K.; Simon, C. Single-Photon Source Based on Rydberg Exciton Blockade. *J. Phys. B At. Mol. Opt. Phys.* **2017**, *50* (21).

(181) Hedley, G. J.; Schröder, T.; Steiner, F.; Eder, T.; Hofmann, F. J.; Bange, S.; Laux, D.; Höger, S.; Tinnefeld, P.; Lupton, J. M.; Vogelsang, J. Number of Chromophores. *Nat. Commun.* No. **2021**, 1–10.

(182) Rath, P. Integrated Optomechanics and Single-Photon Detection in Diamond Photonic Integrated Circuits. **2017**, No. November 2016.

(183) Tomta, S. L. Single-Photon Emitters in Silicon for Quantum Technology Towards Revealing the Nature of Recently Observed Quantum Emitters. **2021**.

(184) Eisaman, M. D.; Fan, J.; Migdall, A.; Polyakov, S. V. Invited Review Article: Single-Photon Sources and Detectors. *Rev. Sci. Instrum.* **2011**, *82* (7).

(185) Goings, J. J.; Ding, F.; Frisch, M. J.; Li, X. Stability of the Complex Generalized Hartree-Fock Equations. *J. Chem. Phys.* **2015**, *142* (15).

(186) Datasheet, M. M. VAMP http://www.addlink.es/images/pdf/agdweb937.pdf.

(187) Degen, C. L.; Reinhard, F.; Cappellaro, P. Quantum Sensing. *Rev. Mod. Phys.* **2017**, *89* (3), 1–39.

(188) Jornet-Somoza, J.; Lebedeva, I. Real-Time Propagation TDDFT and Density Analysis for Exciton Coupling Calculations in Large Systems. *J. Chem. Theory Comput.* **2019**, *15* (6), 3743–3754.

(189) Sottile, F.; Bruneval, F.; Marinopoulos, A. G.; Dash, L. K.; Botti, S.; Olevano, V.; Vast, N.; Rubio, A.; Reining, L. TDDFT from Molecules to Solids: The Role of Long-Range Interactions. *International Journal of Quantum Chemistry*. 2005, pp 684–701.

(190) Wang, Y. G. Examination of DFT and TDDFT Methods II. *J. Phys. Chem. A* **2009**, *113* (41), 10873–10879.

(191) Delley, B. Time Dependent Density Functional Theory with DMol3. *J. Phys. Condens. Matter* **2010**, *22* (38).

(192) Barone, V.; Bencini, A.; Fantucci, P. C. *Recent Advances In Density Functional Methods*; River Edge, NJ : World Scientific. **2002**.

(193) Abozeed, A. A.; Younis, O.; Al-Hossainy, A. F.; El-Mawla, N. A.; Sayed, M.; M. Kamal El-dean, A.; Tolba, M. S. Combined Experimental and TD-DFT/DMOl3 Investigations, Optical Properties, and Photoluminescence Behavior of a Thiazolopyrimidine Derivative. *Sci. Rep.* **2022**, *12* (1), 1–16.

(194) Bernath, P. *Spectra of Atoms and Molecules*; Oxford University Press: New York, 1995.

(195) Franck, J.; Dymond, E. G. Elementary Processes of Photochemical Reactions. *Trans. Faraday Soc.* **1926**, *21*, 536–542.

(196) Moreva, E.; Bernardi, E.; Traina, P.; Sosso, A.; Tchernij, S. D.; Forneris, J.; Picollo, F.; Brida, G.; Pastuović, Ž.; Degiovanni, I. P.; Olivero, P.; Genovese, M. Practical Applications of Quantum Sensing: A Simple Method to Enhance the Sensitivity of Nitrogen-Vacancy-Based Temperature Sensors. *Phys. Rev. Appl.* **2022**, *30*.

(197) Huang, J.; Chen, S.; Zhuang, M.; Lee, C. Robust Interaction-Enhanced Sensing via Antisymmetric Rabi Spectroscopy. *arXiv* **2022**.

(198) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; Desjarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63* (16), 8667–8682.

(199) Royal Society of Chemisty. ChemSpider http://www.chemspider.com/.