Nanoscale

Autonomous Generation of Single Photon Emitting Materials

Journal:	*Nanoscale*
Manuscript ID	NR-ART-10-2023-004944.R2
Article Type:	Paper
Date Submitted by the Author:	02-Apr-2024
Complete List of Authors:	Tempke, Robert; West Virginia University, Mechanical, Materials & Aerospace Engineering Musho, Terence; West Virginia University, Mechanical, Materials & Aerospace Engineering

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Autonomous Generation of Single Photon Emitting Materials[†]

Robert Tempke,[*a] and Terence Musho,[b‡]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

The utilization of machine learning in Materials Science has highlighted that trained models' effectiveness is dependent on the quality and quantity of data utilized for training. Unlike fields such as image processing and natural language processing, there is limited availability of atomistic datasets, leading to biases in training datasets. Particularly in the domain of materials discovery, there exists an issue of continuity in atomistic datasets. Experimental data sourced from literature and patents is usually only available for a select number of atomistic data, resulting in bias in the training dataset. This study focuses on developing a language-based model for generating a synthetic dataset of quantum materials using a variational autoencoder approach. The study centers on generating a synthetic dataset of quantum materials specifically for quantum sensing applications, with a focus on two-level quantum molecules demonstrating dipole blockade. The proposed technique offers an improved sampling algorithm by incorporating newly generated materials into the sampling algorithm to create a more normally distributed dataset. Through this technique, the study was able to generate over 1,000,000 candidate quantum materials from a small dataset of only 3,000 materials. The generated dataset identified several iodine-containing molecules as promising single photon emitting materials for potential quantum sensing applications.

1 Introduction

In recent years, the use of machine learning techniques in chemistry has become increasingly prevalent. Despite this, it has become apparent that many available chemical and materials science datasets suffer from bias[1–4]. This is primarily because these datasets are often composed of a collection of patents and research articles that exist on the internet, rather than representing a continuous material space[4–6]. In response to this issue, the present study seeks to leverage the predictive abilities of deep learning to generate a chemically diverse dataset that is less biased and more robust than those currently available.

To achieve this goal, this study employs a deep learning technique known as a variational autoencoder (VAE), which is capable of synthesizing chemical species with specific chemical properties[7–10]. The VAE forms a custom chemical compression intelligence that provides efficient generation of new specific chemical species by sampling the latent space of the VAE, which can be thought of as a representation of compressed chemical information[11–13]. By training the neural network to learn the chemical and structural similarities of species with specific physical properties, we enable it to identify patterns in a higher dimensionality.

This research is focused on the generation of candidate quantum materials, specifically single-photon source (SPS) materials, also referred to as UV/vis materials[14–18]. These SPS materials rely on a two-level system of electronic states with the added complexity a secondary interaction to form a resonant behavior. While it is beyond the scope of this paper to discuss all of the potential quantum material frameworks the focus of this study is on the discovery of a material that exhibit SPS behavior with a strong dipole interaction. The creation of a chemically diverse dataset is critical to the development of accurate machine learning algorithms. In the context of machine learning, the phrase "garbage in, garbage out" highlights the importance of high-quality data[19,20]. The machine learning algorithm must be taught on a range of inputs and outputs, as the space is continuous, and it must learn everything to know everything.

Several studies have emphasized the issue of bias in experimental design and data collection, which ultimately leads to skewed and unreliable data. Griffiths et al. investigated biases in the natural sciences, focusing on the impact of data splitting, noisy datasets, and contextual variables on the outcome of experiments[21]. Similarly, Kovacs et al. highlighted the direct effects that biased and unbiased datasets can have on the quality of machine learning outputs[22]. Glavatskikh et al. demonstrated how the lack of diversity in data limits machine learning's potential to

[a] Department of Mechanical, Materials, & Aerospace Engineering, West Virginia University, P.O. Box 6106, Morgantown, WV, USA.
[b] Department of Mechanical, Materials, & Aerospace Engineering, West Virginia University, P.O. Box 6106, Morgantown, WV, USA. E-mail: tdmusho@mail.wvu.edu

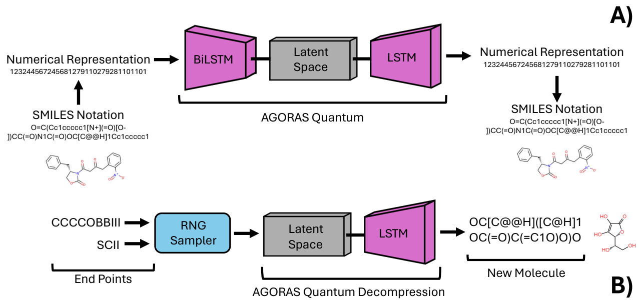


Fig. 1 Subfigure A is an illustration of the AGoRaS network illustrating how a chemical species is encoded and used as training data for the network. Chemical database information is compressed and decompressed to form a high-dimensional latent space. Subfigure B is an illustration of how the trained latent space can be sampled to generate new single-photon emitting materials.

predict[23].

The creation of an experimentally unbiased, continuous dataset is a costly and challenging task. However, our use of deep learning techniques, specifically the VAE, shows promise in generating a chemically diverse dataset with specific chemical properties[8,24,25]. Through our research, we aim to contribute to the development of more accurate and reliable machine learning algorithms in chemistry and materials science.

In the field of SPS materials, it is common to encounter incomplete or inaccurate data in the literature, making it unsuitable for machine learning algorithms. Zakutayev et al. have highlighted the significance of having sufficiently large and diverse datasets for the training of advanced machine learning algorithms in materials science[26] emphasizes the importance of having a robust and extensive dataset to develop machine learning algorithms that can predict the structure, stability, and properties of various materials. Furthermore, it illustrates that existing machine learning algorithms can be quickly and easily adopted to address material science problems, provided there is a suitable dataset for training purposes[5,27,28].

In the context of UV/vis research, the work of Beard et al. stands out for their comprehensive collection of available materials and corresponding relevant calculations[29]. The authors conducted an extensive search of over 400,000 scientific documents to extract a database of just over 8,000 unique compounds. Despite the use of state-of-the-art tools such as ChemDataExtractor, the process of creating a database for quantum materials is challenging. This is due to the wide variety of formatting among different scientific journals, discrepancies within the tools being used, and the lack of a standard set of ground truth rules for representing materials using the SMILES notation. Nevertheless, the database created by Beard et al. is currently the most complete UV/vis material dataset available.

Single photon materials are crucial for a variety of applications, such as quantum communication, quantum computing, quantum information, and quantum precious metrology[16,17,30,31]. SPS materials have found use in various applications, including quantum computing materials, remote sensing, and dipole gates[2,32–34]. These single photon materials exhibit a two-level system behavior, where a resonance is formed between a ground state and a excited state or two excited states. In the application of quantum sensing or computing the defining metric is the coherence time. Often this involves several aspects of the material that are deeply rooted in the atomic coordination on the atoms. One of the targeted metric in organic based quantum sensing materials, is discovering a material that exhibits a strong photo absorption strength and a strong dipole interaction. This type of molecule when interacting with neighboring molecules will exhibit a quantum resonance in which a single photon can exist on only one of the molecules at a time. The dipole interaction will shift or change the neighboring molecule's excited state energy level. This single photon-dipole interaction will give rise to a resonance that can be exploited for quantum sensing and other quantum applications. One potential application is remote sensing, which has gained significant attention in recent years as quantum materials technology has improved. These remote sensing methods have been used to monitor contaminants in water, air quality trends, dissolved nutrients in surface water, and many other advanced techniques[35–37]. For example, Spangenberg et al. demonstrated how quantum materials could be combined to detect relative concentrations of mixtures within water in real-time[36]. Fei et al. demonstrated that the right combination of machine learning algorithms and SPS material, the monitoring of groundwater contamination could be achieved[37]. However, the limited dataset of 1,665 materials used in Fei et al.'s work highlights the need for much larger datasets. Moreover, Mamede et al. further demonstrated the potential of machine learning to be applied with quantum materials by focusing on finding the UV/vis absorption spectrum of organic molecules using fingerprints generated from 2D chemical structures. Their work yielded a sample size of approximately 75,000 molecules using only information about the chemical structures[38].

Recent studies by De Leonardis et al. and Richter et al. have demonstrated the potential of quantum materials in overcoming phase-matching challenges in remote sensing applications[32,39]. These recent research studies demonstrate the growing demand for new quantum materials that can advance different fields . Highlighting the need for more extensive datasets to facilitate machine learning algorithms in SPS materials science studies.

The hypothesized approach in this study is to create a material compression intelligence via the latent space representation of an existing SPS materials database using a VAE with a similar structure as AGoRaS[7]. The latent representation of the materials can be sampled at various points to generate new SPS materials with desirable characteristics. The latent space can be viewed as a representation of the compressed structural and chemical information inherent in the species used to train the network. By populating the latent space with structurally and chemically similar SPS materials, the network can learn the underlying similarities between the materials, resulting in the generation of new materials that share the desired characteristics[40].

The VAE's ability to represent data in n-dimensions and fit input nodes to probabilistic distributions, typically multivariate Gaussian distributions, offers a significant advantage over traditional methods of data representation[8,12,13,41]. The methodology proposed in this study demonstrates the practicality and flexibility of the AGoRaS network in creating organic materials for quantum applications, with the generation of single complex materials replacing the balanced chemical reactions generated in the original AGoRaS study[7]. The generated materials can serve as a vast and robust dataset for the training of other machine learning algorithms.

The generation of SPS materials has be rigorously validated, following a similar methodology to the one outlined by Beard et al., using a workflow that facilitates data collection, network training, network testing, material generation, and material testing[29]. The modular nature of this workflow enhances code quality and robustness while also enabling non-data scientists to employ the methodology with ease, thereby expanding the network's utility to researchers in different fields who lack data science expertise. This approach has been successfully utilized in other generative networks such as ChatGPT, enabling non-data scientists to generate text with background knowledge beyond their expertise.

2 Methodology

2.1 Processing the Database

In this study, the chemical species used to train the AGoRaS-Quantum network were obtained from the dataset created by Beard et al. The data manipulation and network were written in Python. An illustration of the workflow can be found in the original AGoRaS publication for chemical reaction generation[7]. The dataset, which contains approximately 8,000 different species, was downloaded in JSON format[29]. This is relatively small dataset compared to the generated data, which will be in the millions. To provide a scale, the input data is approximately 0.08% the size of the output generated. To ensure the quality and consistency of the dataset, each species was validated using RDKit to verify that its provided SMILES string matched its IUPAC name[42–45]. RDKit is an open-source cheminformatic software that has a series of tools for checking the validity of SMILES strings. It is necessary to check the SMILES notation to avoid training on invalid SMILES strings. There are a set of standards that can be checked by RDKit and an error code is returned if the SMILES is invalid. All species were then read into a Pandas dataframe with relevant information, including SMILES string, excitation wavelength, intensities, and dipole moments.

To simplify the network's predictions and improve reproducibility, the AGORAS network focused solely on predicting SMILES species and not on their associated properties, such as dipole moments and excitation wavelength[46]. This decision allowed the network to focus on learning the underlying physical and chemical structural patterns rather than extending the prediction to properties, which could be calculated using density functional theory (DFT)[46–48]. To generate new species, the network continued to use character-level embedding due to its advantages over word-level embedding in natural language processing. This allowed the generative network to use the information learned during training to generate new species based on the universal alphabet created from all the species in the dataset[49,50].

The use of molecule embedding can improve the predictive power of machine learning models by formulating inputs into sequence embeddings[8,51–53]. In this study, TensorFlow's built-in embedding techniques were used to create embeddings based on the universal alphabet created from the chemical species[54]. This approach was inspired by Gaspar et al.'s work, which demonstrated that molecule embedding can be similar to NLP embeddings[52]. By using sequence embeddings, the network can capture more of the structural and chemical similarities between the species, allowing it to generate new species with similar properties. The use of embedding techniques and a universal alphabet enables the AGORAS network to accurately represent chemical species and generate new species, making it a useful tool for materials science research.

2.2 AGORAS Structure for Quantum Materials

The AGoRaS algorithm is designed to generate new chemical species based on a vector representation of the longest SMILES string in the training dataset. The vector is passed through an Embedding layer in TensorFlow, which projects the input into a higher dimensionality space. This is a critical step as the intrinsic values of the numeric values are removed in the higher-dimensional space. The projected vectors are then passed through a bidirectional LSTM layer, with a recurrent dropout of 0.2[55–57]. The mean and log variance of the output are used to sample the solution space using a sampling function.

The sampled solution space is then decoded using a RepeatVector layer wrapped around the output of the latent space, which turns the data into a tensor vector that an LSTM layer can read. The LSTM layer's output is projected into a vector of length n, and this projection is used to calculate the loss of the network. AGoRaS uses a sequence-to-sequence style loss function typical of variation autoencoders, and the kl loss is used as the monitoring metric during training. 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 SoftMax. The optimizer function was Adam, and the learning rate was set at $1x10^{-5}$. This structure closely mimics that of the original AGORAS network for chemical reaction prediction, except for the input vector's length.

The model takes in a vector representation of the longest SMILES string in the training dataset, which is then projected into a higher dimensionality space. The projected vectors are passed through a bidirectional LSTM layer with a recurrent dropout to extract the mean and log variance, which are then used to sample the solution space. A sequence-to-sequence style loss function is used to calculate the loss of the network, with the kl loss serving as the monitoring metric during training. The model's performance is governed by several hyperparameters, including batch size, embedding and latent dimensionality, kl weight, activation function, optimizer function, and learning rate. The AGoRaS algorithm's combination of deep learning techniques and chemical domain knowledge allows it to generate new chemical species accurately and efficiently.

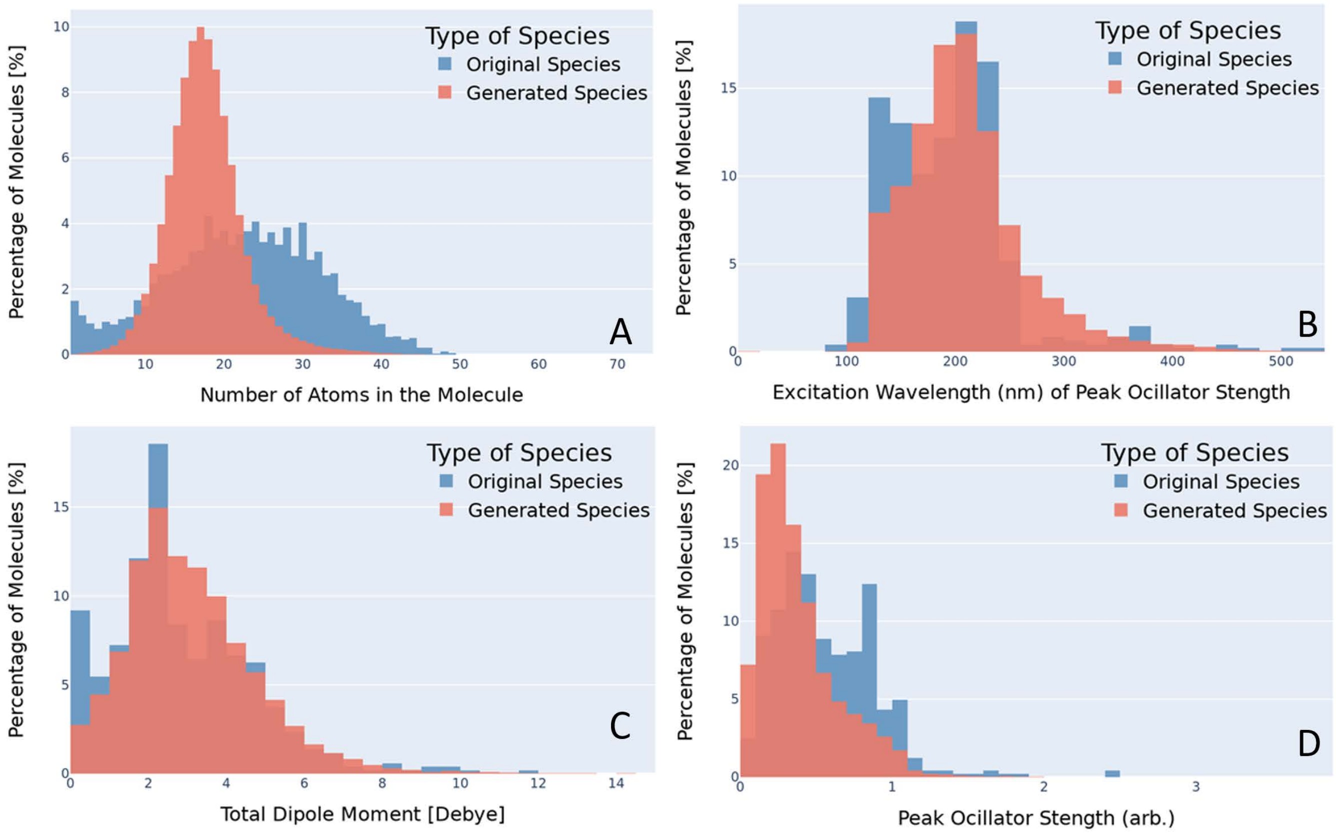


Fig. 2 Histogram comparing the number atoms (A), excitation wavelength (B), the total dipole moment (C), and the peak oscillator strength (D). The blue bins are the training (original) species and the red bins are the generated species from the VAE generated in this study. The goal is to identify small molecules with an excitation wavelength near 500 nm, a strong dipole moment, and strong oscillator strength.

2.3 Training AGORAS for Quantum Materials

Once the chemical data had been pre-processed and converted into a numerical format suitable for neural network architecture, it was divided into three separate datasets: the training set, the validation set, and the test set. The training set comprised 70 percent of the available data, the validation set comprised 20 percent, and the remaining 10 percent was used for the test set. A k-folding approach was used to cross-validate the data over the dataset. Although AGoRaS-Quantum is a generative model, it can be validated using traditional methods. The VAE used in this study was evaluated by its ability to encode the validation set and decode it back to the original string construction with no loss of information.

During the training process, a sequence-to-sequence loss function was employed to score the reconstructed string versus the original string. This approach enabled the validation of the VAE's ability to reconstruct the chemical equations with zero loss of information, which is indicative of a stable latent space. Given the small size of the data used in this study, it was essential to validate the stability of the latent space as much as possible. After it had been demonstrated that the network could reconstruct the test data, the remaining 90 percent of the data were also tested to further validate the stability of the latent space. Although the network should be able to reconstruct all the data used in training, this additional test served as a further validation of the latent space's stability.

Overall, validating the stability of the latent space is critical for this study. By demonstrating that the VAE can encode and decode the original chemical equations with zero loss of information, it is possible to confirm that the latent space is stable. This validation is especially important given the small size of the data used in this study.

2.4 Autonomously Sample the Latent Representation for Quantum Materials

After a neural network has been trained, it is possible to create a sampler that can interface with the latent representation using real species. This can be achieved by selecting two materials and using them to access the latent space. The sampler then returns a new species located at some equidistant point between the two selected materials. Another way to sample the latent representation is to randomly select a species and have the decoder part of the network construct new species based on the equidistant points between the two materials. This directed sampling approach has a significant advantage over continuous sampling of the latent space. It enables researchers to focus on areas of the latent space where the decoded species possess characteristics of interest.

Due to the probabilistic nature of the latent representation, an almost unlimited number of sample points can be taken to generate new species. However, this approach has diminishing returns as there are only a limited number of chemically feasible species that can be generated. Nevertheless, the directed sampling approach can still be a powerful tool for researchers to generate new species with specific characteristics of interest.

2.5 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[44]. This is a common practice when using a neural network with generated SMILES species. 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.

The performance of the AGoRaS-Quantum networks is determine by comparing the number of generated species to the number of unique species. It was determined that 10% of the generated species are unique on the first iteration. That means nearly 10 million species need to generate in order to discover 1 million unique species. Success was around 10% at first because the latent space was limited, since we started with only 3,000 species. As the latent space was sampled more and more stable species were added to the list of stable species, the predictions became better. This increased the stability of the sampling. In the end it was around 40% of species generated were stable. Another 5% were repeats of previously generated species. This theoretically would continue to improve as we sampled more of the latent space.

2.6 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[58–60].

The semi-empirical model that was implemented in the automated script was based on the Materials Studio provided VAMP software package[61]. 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)[62,63]. 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 data preparation steps prior to the semi-empirical calculation. After data was read using SMILES format the SMILES was checked for consistency, followed by making and cleaning of the molecule. The cleaning steps included centering the molecule, adding hydrogen, and conducting 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 and relaxing the spin state in the case of failure and retrying the calculation. In the event that the semi-empirical calculation fails for each spin state, the molecule was assumed unstable and removed from the dataset.

The semi-empirical calculation was chosen because it throughput and robustness of the calculation. Compared to all-electron density functional theory calculations, the semi-empirical calculations take between one to two orders of magnitude less time to provide a prediction. This is critical to this approach where we are aiming to predict the properties of hundreds of thousands of molecules. The reasoning was to use the semi-empirical to provide a quick estimate of the molecule properties, which could be investigated using higher fidelity models after the initial screening using this process.

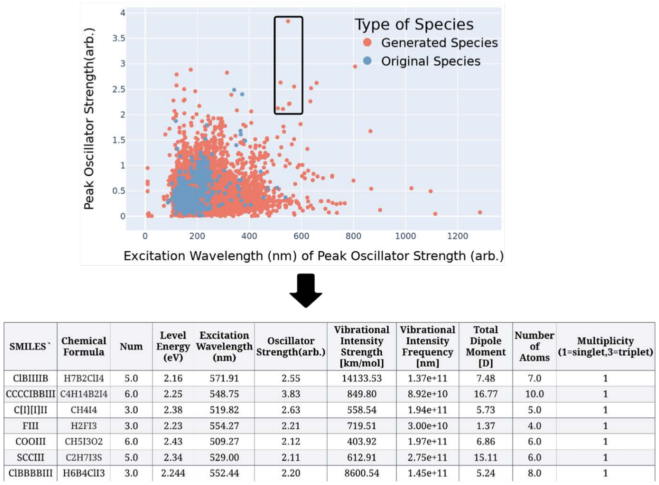
The semi-empirical model provides an estimate of several properties of the molecules, not limited to the formation energy, dipoles, and UV/VIS properties. VAMP can determine the molecular wavefunction of a species, which can then be used to derive the dipole moment and associated properties. This is done using the LCAO method of molecular orbitals rather than the standard MNDO Hamiltonian calculation[61,64]. VAMP is also able to calculate accurate dipole moments using the Natural Atomic Orbital-Point Charge model for molecular electrostatic properties. Furthermore, the UV/VIS properties can be predicted using an empirical methodology that is approximate, providing an estimate of the optical properties.

2.7 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 approximately 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. A histogram for the dipole moments with a comparison between the original species (blue) and the generated species (red) is provided in Figure 2.

2.8 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 discussed in the previous sections. For this study we have identified two materials with strong dipoles and a single frequency characteristics in the 500-600 nm range with a strong peak compared to other peaks in the UV/VIS spectrum. These criteria are selected due to their direct interest to researchers in the field of quantum sensing, where single photon materials with strong dipole moments provide an opportunity for quantum sensors that operate on the principle of dipole blockade.



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

Fig. 3 Example of the ability to search the generated solution space for molecules of potential interest. The inset table within the figure outline seven molecules in the 500-600 *nm* wavelengths with an oscillator strength (pronounced optical peak) above 2.0. These seven correspond to the black box in the histogram. Note, most materials exhibit low excitation wavelength and oscillator strength.

2.9 Validate Material with TDDFT

Based on the seven materials outlined in Figure 3 for the remainder of this study, the focus will be on the two most promising, which are FIII and C[I][I]II. We believe these two materials warrant experimental consideration, however, before this, it is our approach to put these molecules through more rigorous computational calculations. This involves conducting a time-dependent density functional theory (TDDFT). The higher accuracy of the TDDFT method than that of the semi-empirical methods provides greater confidence in the predicted value. This higher level of validation also allows for experimentation with these molecules to help identify what is the molecular origin or mechanism of these strong peaks and associated oscillator strength. The molecule's emission spectra can be investigated by adding or removing elements to see the response in the strength and the peak wavelength. The TDDFT approach allows the researcher to visualize the wavefunction and local density of states as shown in Figure 4. In this figure for FIII (A and B) and C[I][I]II (C and D) the ground state density of states (A and C) and the excited states (B and D) can be investigated. The yellow is associated with the spin-up states and the blue isosurfaces are associated with the spin-down states. An important aspect is that the overlap of the wavefunction is similar to provide good coherence of the excited electron. Table 1 is an outline of the most promising excited state transitions. As can be associated between this Table and Figure 4 a strong oscillator strength (f-osc) is desirable with a large overlap of the wavefunctions. While these may not be the best optical absorbers within the 500-600 *nm* range this proves that there is potential to find these materials. As seen here for these two molecules the C[I][I]II molecule provides slightly better oscillator strength when compared to FIII, at the expense of a larger molecule.

From	To	TD-ex [eV]	TD-ex [nm]	f-osc	Overlap
85	86+	1.31	950	0.000026	0.50
83	86-	1.46	851	0.000110	0.87
81	86+	2.17	571	0.001437	0.39
80	86+	2.76	450	0.000796	0.44

From	To	TD-ex [eV]	TD-ex [nm]	f-osc	Overlap
111	112+	1.83	676	0.001886	0.61
109	112-	1.99	622	0.000295	0.59
107	112+	2.25	550	0.082783	0.56
106	112+	2.32	535	0.003263	0.46

Table 1 List of most probable transition states and their associated oscillator strength (f-osc). The top table is for the molecule FIII and the bottom table is for molecule C[I][I]II. The transitions with optical transitions between 500-600 *nm* with large f-osc and overlap are desired.

Results

After the SMILES strings had been embedded and the VAE was trained, the latent space was sampled to generate new chemical species. The approach of sampling the latent is a unique approach that saw success in a previous study[7] where bias in the training dataset was corrected. Sampling was stopped once approximately a 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 consistency checker 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 Figures 2A. The number of atoms within the molecule 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.

The distribution of these atom counts can be seen in Figures 2, which allows for a more in-depth analysis. Again, 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 2 that the generated species and original species share an approximately normal distribution with two main differences. The first is

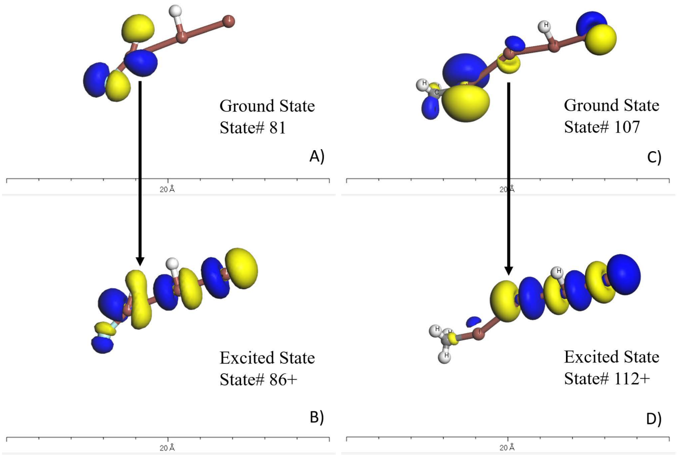


Fig. 4 TDDFT results for the FIII (H_2FI_3) molecule (A and B) and C[I][I]II (C and D). Subfigure A and C illustrate the ground states local density of states surfaces. The yellow is associated with spin-up and blue is associated with spin-down electrons. Subfigure B and D are the excited states local density of states. Comparing the ground state to the exited states provide spatial information of the electron transition outlined in Table 1. It is desirable to have a high overlap of these states to avoid decoherence of the states.

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 cannot 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.

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. 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 (high oscillator strength) at some wavelength in the visible spectrum.

Using semi-empirical calculations it is possible to calculate the wavelength and emission strength of excited electrons. It is important to note that the intensity of the light being emitted is difficult to compare between molecules but can be compared between other peaks in the spectrum for a given material. 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. These calculations at each energy level led to the calculated emission spectra that will be used as validation of these materials. 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 total dipole moment, vibrational spectrum, and vibrational strength.

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 2D illustrates 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.

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.

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, only the best material i.e., the strongest emitters 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 generates 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.

Another important aspect of these types of materials is at which frequency these peaks occur. Figure 2B 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 100 and 1400 nm, which helps to enforce an equal distribution of values within that range. An interesting find from Figure 2B is that 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.

The total dipole moments of the real and generated materials were also calculated, which was important in selecting a molecule that could potentially be operated in the dipole blockade quantum sensing application. The dipole is based on the partial charge and positions of the atoms. The overlaid histogram for the total dipole moments of the original and generated species can be seen in Figure 2C. 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 zero dipole moment. It is interesting to note, in Figure 2B, where the original data has a bit of an uneven distribution, however, the generated data is a single peak distribution. The network cannot generate any more single element materials with new elements from the periodic table so the zero dipole materials are limited.

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. 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 5. 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 5A has all of the original data 8,000 species while only showing a randomly selected 8,000 of the generated species. Meanwhile, Figure 5B also illustrates 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 Figure 5 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. Figure 5B clearly illustrates 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.

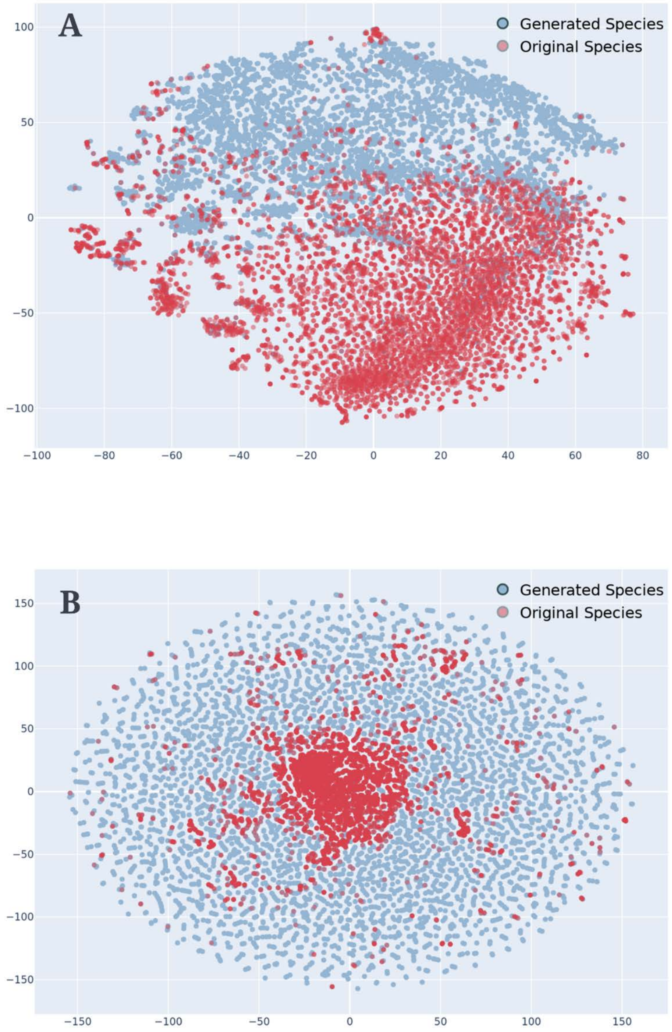


Fig. 5 t-SNE plot of the training dataset and the generated dataset with the oscillator strength representative of the size of the points. Subfigure A has all of the 8,000 original species and a randomly sampled 8,000 generated species. Subfigure B has all of the 8,000 original species and 80,000 randomly sampled generated species.

2.10 Candidate Quantum Materials

While this study focused more on the framework for the generation and discovery of new materials it was interesting to point out that there were a few species that show promise for the targeted application of quantum sensing. The most impressive of the species is the iodine containing structure shown in Figure 4. Two iodine containing molecules were identify, FIII and C[I][I]II. Both of these molecule had two of the strongest oscillator strengths that was confirmed with TDDFT calculations. These molecules also exhibited a strong dipole. The peak wavelength of FIII was 851 nm, which is in the infrared. But C[I][I]II was 550 nm, which is yellow color in the visible and is close to the wavelengths used in fiber optic communication. A brief literature reveals that iodine has extensive use in the field of photochemistry, which is encouraging and supports the predictive capability of AGoRaS-Quantum.

Conclusions

In this study, the AGoRaS network[7] was extended from the generation of gas-phase chemical reactions to the generation of quantum materials. This study was designed to demonstrate how an AGoRaS-style VAE could be used in other applications of nanoscale materials science. The primary purpose was to demonstrate how with a small dataset of materials with specific characteristics it would be possible to synthetically generate a large number of new materials. The focus materials system for this study was single photon emitting materials. AGoRaS-Quantum was used 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. The utility of the generated data was demonstrated in this study by indentifying several iodine containing structures that exhibited promising quantum material attributes.

The novel aspect 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. This was an improvement of the previous AGoRaS sampling method in the ability to use the SMILES representations of the molecular species as starting points in sampling the latent space. This allowed 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 and to generate new molecular species beyond the size and descriptions contained in the training data.

This generational approach opens the possibilities for more in-depth analysis of these quantum materials. For example, there is potential for a traditional machine learning analysis to be performed 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 the overfitting of the latent space via the network's variance. Another interesting approach that could be taken to improve the network speed and efficiency, is the autonomous design of the network parameters. While this study was based on hand-tuned parameters until a stable network could be created. This leaves a great opportunity for the design of a more memory-efficient network.

Conflicts of Interest

There are no conflicts to declare.

Acknowledgments

R.T. would like to acknowledge the Department of Energy ORISE Fellowship for support in completing this research. T.M. would like to acknowledge his attendance to the inaugural 2023 US Quantum Information Science School that was hosted by the SQMS Center and U.S. Department of Energy's (DOE's) Fermi National Accelerator Laboratory (Fermilab). Computational resources were provided by the WVU Research Computing Dolly Sods HPC cluster, which is funded in part by NSF OAC-2117575. This manuscript is being submitted to celebrate the 150th anniversary of Vanderbilt University.

Data Availability

Input and Output Data, which includes generated structures will be available on NOMAD online materials repository at the following address: `https://nomad-lab.eu/`

The source code corresponding to the machine learning model and the Pipeline Pilot script can be downloaded on GitHub at the following address: `https://github.com/Dr-Musho-Research-Group/AGORAS_QUANTUM`

References

1 Y. Zhang and C. Ling, *npj Computational Materials*, 2018, **4**, 28–33.

2 S. Go, J. Kim, S. S. Park, M. Kim, H. Lim, J. Y. Kim, D. W. Lee and J. Im, *Remote Sensing*, 2020, **12**, 1–34.

3 M. Shepperd and M. Cartwright, *IEEE Transactions on Software Engineering*, 2001, **27**, 987–998.

4 A. Amini, W. Schwarting, G. Rosman, B. Araki, S. Karaman and D. Rus, *IEEE International Conference on Intelligent Robots and Systems*, 2018, 568–575.

5 K. K. Yalamanchi, M. Monge-Palacios, V. C. Van Oudenhoven, X. Gao and S. M. Sarathy, *Journal of Physical Chemistry A*, 2020, **124**, 6270–6276.

6 H. A. Carroll, Z. Toumpakari, L. Johnson and J. A. Betts, *PLoS ONE*, 2017, **12**, 1–19.

7 R. Tempke and T. Musho, *Nature Communications Chemistry*, 2022, 1–10.

8 A. B. L. Larsen, S. K. Sønderby, H. Larochelle and O. Winther, *Icml*, 2016, **4**, 2341–2349.

9 R. Burks, K. A. Islam, Y. Lu and J. Li, *2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2019*, 2019, 0660–0665.

10 S. Semeniuta, A. Severyn and E. Barth, *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2017, 627–637.

11 W. Jin, R. Barzilay and T. Jaakkola, *arXiv*, 2018.

12 L. Yu, W. Zhang, J. Wang and Y. Yu, *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017, 2852–2858.

13 S. Rajeswar, S. Subramanian, F. Dutil, C. Pal and A. Courville, *arXiv*, 2017.

14 C. L. Degen, F. Reinhard and P. Cappellaro, *Reviews of Modern Physics*, 2017, **89**, 1–39.

15 S. L. Tomta, 2021.

16 M. D. Eisaman, J. Fan, A. Migdall and S. V. Polyakov, *Review of Scientific Instruments*, 2011, **82**, year.

17 A. Slachter, *PhD thesis*.

18 A. B. D. al jalali wal ikram Shaik and P. Palla, *Scientific Reports*, 2021, **11**, 1–27.

19 L. T. Rose and K. W. Fischer, *Measurement*, 2011, **9**, 222–226.

20 H. Sanders and J. Saxe, *Proceedings of Blackhat 2017*, 2017,

6.

21 R. R. Griffiths, P. Schwaller and A. A. Lee, *ChemRxiv*, 2018.

22 D. P. Kovács, W. McCorkindale and A. A. Lee, *Nature Communications*, 2021, **12**, 1–9.

23 M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy and B. Da Mota, *Journal of Cheminformatics*, 2019, **11**, 1–15.

24 M. A. Kayala, P. Baldi, S. Rajeswar, S. Subramanian, F. Dutil, C. Pal, A. Courville, J. Zhao, Y. Kim, K. Zhang, A. M. Rush, Y. LeCun, H. Catherine, M. L. Cook, E. Mckone, R. R. Griffiths, P. Schwaller, A. A. Lee, M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy, B. Da Mota, T. F. Cova, A. A. Pais, A. C. Mater, M. L. Coote, M. J. Kusner, J. M. Hernández-Lobato, R. D. Camino, C. A. Hammerschmidt, R. State, E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, J. Sun, P. Potash, A. Rumshisky, R. D. Camino, C. A. Hammerschmidt, R. State, M. A. Kayala, C. A. Azencott, J. H. Chen, P. Baldi, C. McCutchen, J. Schmidt, M. A. M. R. Marques, S. Botti, M. A. M. R. Marques, D. P. Kovács, W. McCorkindale, A. A. Lee, P. L. Kang, Z. P. Liu, M. A. Kayala, P. Baldi, L. Yu, W. Zhang, J. Wang and Y. Yu, *arXiv*, 2019, **68**, 1–9.

25 M. Ryo and M. C. Rillig, *Ecosphere*, 2017, **8**, year.

26 A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas and C. Phillips, *Scientific Data*, 2018, **5**, 1–12.

27 D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge and P. W. Chung, *Scientific Reports*, 2018, **8**, 1–12.

28 M. Hirohara, Y. Saito, Y. Koda, K. Sato and Y. Sakakibara, *BMC Bioinformatics*, 2018, **19**, year.

29 E. J. Beard, G. Sivaraman, Á. Vázquez-Mayagoitia, V. Vishwanath and J. M. Cole, *Scientific Data*, 2019, **6**, 1–11.

30 P. Rath, 2017.

31 J. Q. Jiandong Qiao, F. M. Fuhong Mei and Y. Y. Yu Ye, *Chinese Optics Letters*, 2019, **17**, 020011.

32 A. Richter and T. Wagner, *Solar Backscattered Radiation: UV, Visible and Near IR - Trace Gases*, 2011, pp. 67–122.

33 Agilent Technologies, *Models in the 85071E Materials Measurement Software*, `na.support.keysight.com/materials/docs/85071Emodels.pdf`.

34 K. K. Kärkkäinen, A. H. Sihvola and K. I. Nikoskinen, *IEEE Transactions on Geoscience and Remote Sensing*, 2000, **38**, 1303–1308.

35 S. I. Ahmad, P. Koteshwar Rao and I. A. Syed, *Journal of Taibah University for Science*, 2016, **10**, 381–385.

36 M. Spangenberg, J. I. Bryant, S. J. Gibson, P. J. Mousley, Y. Ramachers and G. R. Bell, *Scientific Reports*, 2021, **11**, 1–8.

37 C. Fei, X. Cao, D. Zang, C. Hu, C. Wu, E. Morris, J. Tao, T. Liu and G. Lampropoulos, 2021, 46.

38 R. Mamede, F. Pereira and J. Aires-de Sousa, *Scientific Reports*, 2021, **11**, 1–11.

39 F. De Leonardis, R. A. Soref, M. Soltani and V. M. Passaro, *Scientific Reports*, 2017, **7**, 1–9.

40 T. Mikolov, K. Chen, G. Corrado and J. Dean, *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013, 1–12.

41 Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu and J. Hu, *npj Computational Materials*, 2020, **6**, 1–16.

42 W. Grenier and M. Vinokurskaya.

43 J. He, H. You, E. Sandström, E. Nittinger, E. J. Bjerrum, C. Tyrchan, W. Czechtizky and O. Engkvist, *Journal of Cheminformatics*, 2021, **13**, 1–17.

44 Rdkit.org, *RDKit:Open-source cheminformatics*, `http://www.rdkit.org/`.

45 A. I. Sarker, A. Aroonwilas and A. Veawab, *Energy Procedia*, 2017, **114**, 2450–2459.

46 J. Jornet-Somoza and I. Lebedeva, *Journal of Chemical Theory and Computation*, 2019, **15**, 3743–3754.

47 F. Sottile, F. Bruneval, A. G. Marinopoulos, L. K. Dash, S. Botti, V. Olevano, N. Vast, A. Rubio and L. Reining, *TDDFT from molecules to solids: The role of long-range interactions*, 2005.

48 J. S. Smith, O. Isayev and A. E. Roitberg, *Chemical Science*, 2017, **8**, 3192–3203.

49 M. J. Kusner, B. Paige and J. Miguel Hernández-Lobato, *arXiv*, 2017.

50 A. Nigrin, *Neural networks for pattern recognition*, MIT Press, Massachusetts, 1st edn, 1993, p. 413.

51 S. Jaeger, S. Fulle and S. Turk, *Journal of Chemical Information and Modeling*, 2018, **58**, 27–35.

52 H. A. Gaspar, M. Ahmed, T. Edlich, B. Fabian, Z. Varszegi, M. Segler, J. Meyers and M. Fiscato, *ChemRxiv*, 2021, 1–10.

53 J. D. Prusa and T. M. Khoshgoftaar, *Journal of Big Data*, 2017, **4**, year.

54 Google, *ImageDataGenerator*, `https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator`.

55 S. Metlek, K. Kayaalp, I. B. Basyigit, A. Genc and H. Dogan, *International Journal of RF and Microwave Computer-Aided Engineering*, 2021, **31**, 1–10.

56 M. Dwarampudi and N. V. S. Reddy, 2019.

57 S. Gajendran, D. Manjula and V. Sugumaran, *Journal of Biomedical Informatics*, 2020, **112**, year.

58 A. Jinich, B. Sanchez-Lengeling, H. Ren, R. Harman and A. Aspuru-Guzik, *ACS Central Science*, 2019, **5**, 1199–1210.

59 J. Choi, L. Quaglito, S. Lee, J. Shin and N. Kim, *International Journal of Fatigue*, 2021, **145**, year.

60 H. Catherine, M. L. Cook and E. Mckone, *Iclr*, 2017, **15**, 401–437.

61 M. M. Datasheet, *VAMP*, `http://www.addlink.es/images/pdf/agdweb937.pdf`.

62 J. J. Goings, F. Ding, M. J. Frisch and X. Li, *Journal of Chemical Physics*, 2015, **142**, year.

63 K. K. Ni, T. Rosenband and D. D. Grimes, *Chemical Science*, 2018, **9**, 6830–6838.

64 J. Furthmüller, *VAMP - Vienna Ab initio Molecular dynamics Package*, 1994.