Predicting Biodegradability and Generating Biodegradable Molecules

Amish Saini
asaini68@gatech.edu

Nathan Wang
nwang334@gatech.edu

Varun Hegde
varun.hegde@gatech.edu

Abstract

Many environmental challenges can be attributed to types of materials, such as plastics and their polymer components, and their corresponding toxicity and biodegradability. New materials need to be innovated, but the process of doing so manually in labs becomes extremely expensive and time-intensive. This paper explores advanced machine learning models, including specialized WGANs and VGAEs, to create new molecular structures, trained on a dataset of various existing molecules and their corresponding biodegradability information. These models will be able to generate new molecular structures that can be used for materials that are stable and possess qualities that are more environmentally safe. Code for our project can be found at https://github.com/Nathan-Wang23/MLC_Final_Project.

1 Introduction and Background

Many materials pose significant risks to humans and ecosystems due to their inherent toxicity or persistence in an environment. Biodegradability refers to the extent and rate that microorganisms break down substances. Ideal biodegradable molecules break down into harmless substances quickly. However, conventional plastics are not readily biodegradable, so microplastics build up in environments and increase long-term exposure to harmful materials. Thus, the conversation has moved towards the discovery of less toxic materials that are more readily biodegradable for a multitude of purposes, like cosmetics, packaging, and more. By generating environmentally sustainable molecules, we provide insight for future work on innovating better plastics, polymers, and substances.

Generative models are a new and rapidly improving space, where the relationship between material structures and their environmental sustainability can be learned, and eventually predicted. While current machine learning discoveries are promising for new material design, there has been little exploration for employing generative models to build novel molecules that have low toxicity and high biodegradability. We developed a framework that generates these structures while maintaining similarity to our inputs and conducted comparative analysis on baseline models with models in which we added a conditioning factor to the latent space.

2 Existing Works and Limitations

2.1 Generative Models

Literature regarding generative models, especially for novel structure tasks, have become increasingly relevant. There have been 3 major types of generative models that have shown success in molecular generation.

GANs

One type of model that has shown a lot of progress are Generative Adversarial Networks. Just earlier this year, Macedo et al. (2024) developed MedGAN to improve on the previously successful and traditional MolGAN (Cao et al., 2022) architecture and even its advancement L-MolGAN (Tsujimoto et al., 2021) through the use of RCGNs to represent bond relationships well (Macedo et al., 2024). These papers all use a Wasserstein GAN and a gradient penalty and generate molecules for medicine, focusing on synthesizability and solubility by training on the Zinc 15 dataset. MedGAN was able to improve the connectivity, uniqueness, solubility, and synthesizability of its molecules.

GAN-based molecular generation has also recently been expanded through new methods and new domains. Anoshin et al. (2023) developed the Quantum Cycle GAN which incorporated quantum computing, successfully improving the efficiency of generation and the diversity as a result (Anoshin et al., 2023). Changing the domain from medicine focused molecules, Tan et al. (2023) developed a GAN that was targeted to generate fluorescent molecules, succeeding with high validity and using a custom model to predict fluorescence (Tan et al., 2023).

VGAEs

Another popular model for these kinds of tasks are Variational Auto Encoders, especially after they have been applied to generative tasks for graphs. In their proposal of the GraphVAE, the authors Simonovsky and Komodakis discuss using a variational auto-encoder architecture to generate graphs directly, and even apply it to molecular generation as a testing application (Simonovsky and Komodakis, 2018). The work done by Iwata et. al. introduces a novel approach for molecular generation, which combines deep learning and reinforcement learning techniques to learn and encode molecular structures into a latent space, and then uses Monte Carlo Tree Search (MCTS) to explore and optimize molecular structures based on this learned representation (Iwata et al., 2023). They were also able to generate and optimize novel molecules with targeted properties, outperforming several benchmark studies. Additionally, Gao et. al. proposes a graph-based variational autoencoder model (MRGVAE) for molecular generation, but this approach uniquely leverages a hierarchical chemical graph representation to construct molecular structures (Gao et. al., 2023). The model represents molecules at three different levels: atoms, fragments, and fragment clusters. The generation process divides the molecule into smaller chemical fragments, which are then grouped into clusters. These clusters help in generating diverse and complex molecular structures.

Diffusion Models

Complex Diffusion Models have shown good progress in the area of molecular generation as well. Wu et al. (2022) developed a Diffusion Model using informative prior bridges which essentially helps focus on desirable properties during molecular generation (Wu et al., 2022). Their model achieved novelty and chemical validity rates that were 25-30 percentage higher than many current methods. A more recent model by Hua et al. (2023) further improved these results by generating both atoms and bonds in one combined diffusion process and increased the structural complexity of molecules by 35 percent in comparison to other diffusion models (Hua et al., 2023). Both models focused on generating molecules for drug discovery.

The 3-dimensional generation space was explored by Xu et al. (2023) who developed a diffusion model that integrated many geometric features to generate 3D-molecules. Their molecule understood geometric attributes very well with a 40 percent increase in the precision of geometric features with a 95 percent validity (Xu et al., 2023). Recently Huang et al. (2024) made improvements to 3-dimensional generation through a dual diffusion model. It had a larger focus on application by incorporating information about fit and optimizing for fitting with target protein pockets (Huang et al., 2024).

2.2 Biodegradability Prediction

Research on biodegradability prediction and classification is relatively new and understudied. Papers like MIT's "High-throughput experimentation for discovery of biodegradable polyesters" emphasizes the significance of creating polymers with a more sustainable life cycle but notes that research into polymer and molecule biodegradation has been limited to a small number of polymers because current biodegradation testing methods are time- and resource-intensive (2023).

The promise of graphical model approaches to biodegradability prediction from static structures has been demonstrated by Lee and Min (2022). Results from this paper demonstrate the effectiveness of Graphical Convolutional Networks in classification tasks of biodegradability for given molecules. The GCN demonstrated stable and robust performance with 84% balance accuracy for binary classification of biodegradability. Therefore, our proposed method of classifying biodegradability metrics of our generated molecule structures with such a graphical model is viable. Nevertheless, we aim to improve upon Lee and Min's GCN approach by incorporating state of the art encoders and visualize the relationship between molecular features and biodegradability, as this classification is especially important since it will act as our threshold for determining success of our generations.

One such example is the Relational Graph Convolutional Network (R-GCN), as introduced by Schlichtkrull et. al (2017). R-GCNs are an improvement on traditional Graph Convolutional Networks, as they are designed to process multi-relational graphs. In the paper, these neural networks are applied to two tasks – link prediction and entity classification – and boast competitive results, outperforming standard benchmarks. We aim to apply this model to a third task: biodegradability classification, with the idea that this will perform better because molecular structures have different bond types that impact the relationship between atoms.

3 Novelty and Significance of the Study

Our proposed project aims to generate novel molecular structures, trained with low toxicity and high biodegradability in mind. Rather than generating novel crystal structures with GANs/VGAEs or predicting a molecule's toxicity and biodegradability, our approach is novel in that we will construct new molecules conditioned to be similar to those that are within the low toxicity and/or high biodegradability spaces, and then predict their toxicity and biodegradability in a multi-stage process.

4 Data Compilation and Preprocessing

4.1 Data Sources

Our data for molecule structures with biodegradability and toxicity features are derived from multiple datasets, compiled and published in research papers for public use. We note the lack of data for polymer specific biodegradability, as green plastics and polymers is a new and upcoming field of study. That is why our goal is to develop a new set of potential molecules that can be used for material-specific predictions. However, we found datasets for biodegradability and toxicity of molecules that we propose to employ when generating new molecule structures.

Lunghini et al. has compiled public and industrial data to formulate the largest existing dataset of 3,192 compounds, consisting of structure data in SMILES and Molfile formats and a Ready Biodegradability (RB) binary classification score that corresponds to slow or fast biodegradation. According to the paper, new models trained on their data results in a high Balance Accuracy in the range of 0.74 to 0.79, which speaks to the promising nature of this dataset for our task. After initial data analysis, we determined 1,133 entries within the dataset that biodegradade quickly and 2,059 compounds that do not biodegrade.

We gathered our toxicity data from Tox21 (2016), which offers 12,060 training samples, with 5,330 unique molecules and their chemical structures in Molfile format. Data from the Tox21 dataset and the All-Public dataset consists of chemical structures stored in Molfile format, which offers information of atom coordinates, structure, and bonds.

4.2 Data Preprocessing

Our data preprocessing workflow for our project followed three main steps: First, we converted Molfile formats to ASE Atoms. Second, we converted ASE Atoms into our graph representation data points. And finally, we converted the graph data representation into atomic number and bond adjacency lists of shapes [BATCHSIZE, 63, 11] and [BATCHSIZE, 63, 63, 5] respectively. We selected 63 as our maximum atom count in a given molecule because our preliminary data analysis, as shown in Figure 1, revealed the 99th percentile of atom counts in our training dataset was 62.5 atoms.

Following insight from MedGAN (Macedo et al., 2024), appending a ”no bond” classification to the traditional Single, Double, Triple, and Aromatic bond types shows some promise in generative models. Rather than using a 4-class one hot encoding for each bond type, we added the fifth dimension so our model can more accurately predict the absence of bonds between atoms.

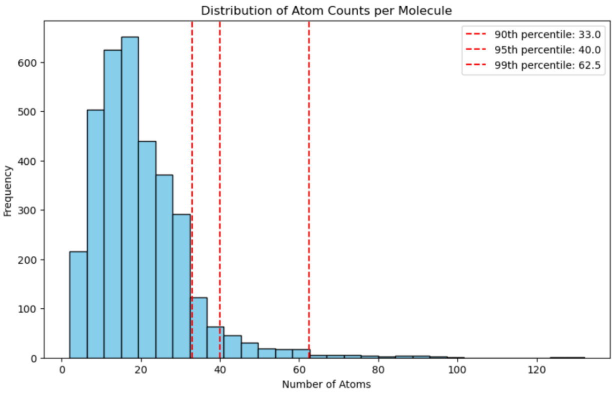


Figure 1: Graph of Number of Atoms by Frequency in the AllPublicNew Biodegradability Dataset.

We applied a similar technique for our categorization of atomic numbers. From the AllPublic biodegradability dataset, we identified ten distinct atomic types: Carbon (C), Nitrogen (N), Oxygen (O), Fluorine (F), Silicon (Si), Phosphorus (P), Sulfur (S), Chlorine (Cl), Bromine (Br), and Iodine (I). We extended the one hot encoding size from 10 to 11 with the addition of a ”no-atom” class. By including this extra dimension for atomic numbers, we enhanced our model’s flexibility in generating diverse molecules while still maintaining a structured data format for easy usage.

After data format conversions and preprocessing, our final phase in our data workflow was creating two different datasets for training models aimed at predicting biodegradability classification, and for use in generative modeling. The datasets for the biodegradability prediction models incorporate the graph representation data in a one-hot-encoded format with the original four bond dimensions in addition to their respective classification labels. The input format for our generative models uses the extended graph representations, processed just before training into adjacency matrices, which promoted straightforward conversion to graphical or chemical representations.

5 Methodological Approach and Evaluation Metrics

Our methodology is be based on a 3-step process similar to current research in structure generation. Starting from adjacency matrices converted from Molfile notations, we first generate molecules,

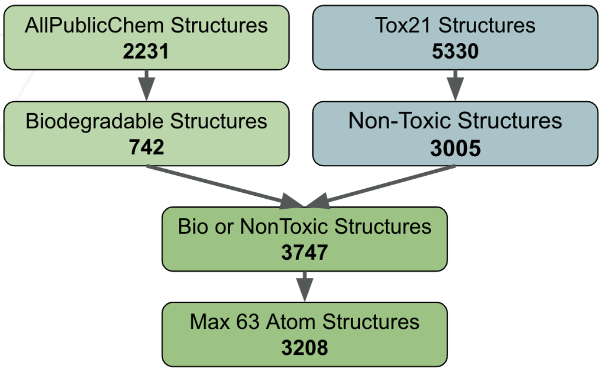


Figure 2: Components and details of our generative model dataset. We include a total of 3,208 unique molecules filtered by Biodegradable and Non-toxic.

then validate their stability and chemical formula, and finally determine the generated molecule's biodegradable properties.

5.1 Featurization

Molfile Encoding: We will utilize Molfile format to represent molecular structures. Molfiles provide comprehensive details including atom types, bond types, and 3D spatial coordinates, offering a rich dataset for our generative models. This format allows for a detailed representation of the molecular structure, enabling our models to leverage spatial information in both the generation and biodegradability prediction tasks.

5.2 Relational Graph Convolutional Network

To improve on the existing graph-based neural network solutions that were discussed, we propose using a Relational Graph Convolutional Network (R-GCN) to account for the multiple kinds of relationships between nodes in the graphs we are constructing. R-GCNs are an extension of traditional Graph Convolutional Networks that account for edge features and attributes by relationship type (Sclichtkrull et al., 2017).

This is done by learning a distinct weight matrix for each type of edge relation in the graph, as opposed to the single edge weight matrix present in traditional GCNs. The neighboring node feature matrices are obtained and transformed by the corresponding edge weight matrix based on the relationship type between a node and all its neighbors. This is repeated for every node in the graph. Finally, these transformations are aggregated and traditional techniques, like nonlinearity, are applied to output a representation of a neighborhood within a graph. This method can capture richer edge information in graphs that is frequent in the molecular graphs being constructed, which is especially useful for the four bond types we are using, as previously discussed.

This approach is relatively novel, as the main use for R-GCNs are for link predictions in cases like knowledge graphs, or social networks. As far as we have seen, there have been no results published of the usage of R-GCNs for the classification of molecular property prediction.

5.3 Modeling

The modeling process is organized into sequential steps, each building upon the insights and outputs of the previous one:

5.3.1 Step 1 - Generative Modeling Overview (WGANs and VGAEs)

The first step is using generative models trained on a dataset taken from the Molfiles for molecules that are non-toxic and/or biodegradable. The objective is to generate adjacency lists for molecules that can be both biodegradable and non-toxic by generating from a distribution of these structures. Many models we mentioned in the past literature have been successful in constructing architectures for molecule or structure generation, so we utilized transfer learning on our WGAN model to maximize our performance.

Variational Graphical Autoencoder (VGAE): We designed a specialized architecture that handles conversions between our graphical data format and an atomic number and bond adjacency matrices format. Our encoder employs two RGCN layers with a latent dimension of 256 to compress graphs into a lower dimension latent space. After global add pooling, we incorporate linear layers for mean and log standard deviation of our prior distribution space as per the VAE paper by Kingma and Welling [2013]. We sample and reparameterize Z using a standard Gaussian epsilon, which we then pass through two different linear decoders that output the input atomic number matrix and bond adjacency matrix. [Figure 3]

Wasserstein Generative Adversarial Network (WGAN): GPT We utilized the architecture of the current state-of-the-art Wasserstein Generative Adversarial Network outlined in Macedo et al.'s 2024 study, "Optimized generative adversarial network with graph convolutional networks for novel molecule design," published in Nature Scientific Reports. The generator attempts to create bond adjacency matrices and atomic number matrices from a latent space of random gaussian noise. This generator consists of six linear and dropout layers to decode the latent space into a molecule representation that can fool the dis-

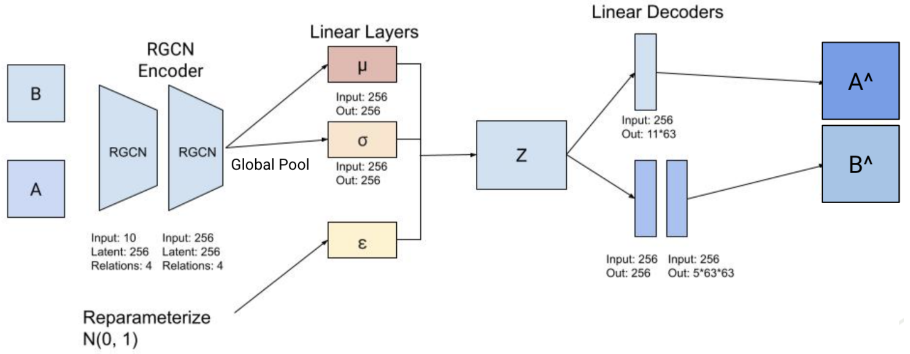


Figure 3: Our VGAE Architecture includes a two RGCN layer encoder and two decoders for atomic number and bond reconstructions.

criminator. The adjacency matrix and atom matrix get converted to graph objects before being fed into the discriminator. The discriminator consists of four RCGNs followed by four pairs of linear and dropout layers that result in a single value [Figure 4]. The discriminator evaluates the generated molecules and actual molecules and tries to tell them apart.

A key aspect of the model is that it is a WGAN (Wasserstein GAN) and uses gradient penalty. The loss is modified to use Wasserstein distance, which allows for a smoother gradient throughout training and helps ensure that the entire training distribution is being generated. This avoids the common problem of mode collapse and the smoother gradients prevent exploding or vanishing gradients. Additionally, a gradient penalty is added to the loss function which ensures that gradients of the discriminator are limited to 1, which is known as the 1-Lipschitz constraint. This penalization causes training to be more smooth and better converging since there are not drastic changes in the gradients. This enhances the impact of a WGAN with even smoother gradients and preventing explosion or vanishing (Gulrajani et al., 2017). Because of this loss format, the discriminator's output is a score rather than a 1 or 0 for real or fake. Therefore, the discriminator acts as a critic.

5.3.2 Step 2 - Stability Assessment

After we generate molecules we have to assess their stability to ensure they can even exist in real life. The validity of each candidate molecule is evaluated using the following method:

We revert generated adjacency matrices back into lists of atomic numbers by index and bonds with start index, end index, and bond type information. First, we add each atom seen into a new Mol data type. Then, we loop over each bond within the existing bonds list and only append the bond if its corresponding start and end point indices exist in the Atoms structure. If there exists an atom without any bonds connecting it to the molecular structure, we remove the atom. Finally, we run RDKit's SanitizeMol method to clean atomic features, fix electron information, and add Hydrogen atoms (as Python chemistry libraries exclude Hydrogens to avoid overcomplicated data structures with too many Hydrogen atoms). The RDKit SanitizeMol method also checks the validity of the structure in terms of energy and other chemical properties and throws errors if the given molecule is not viable in nature. With these steps, we perform a double check for the stability of the structure and ensure validity through the RDKit's sanitize checks.

5.3.3 Step 3 - Biodegradability Evaluation

Following the stability assessment, we measured the biodegradability of each of the new generated molecules. In order to do so, we constructed classification models that surpassed the current state-of-the-art for biodegradability property prediction, and can accurately predict whether a generated molecular structure will be . In (Lee and Min, 2022), the authors evaluate traditional QSAR methods, along with a novel Graph Convolutional Network. As mentioned previously, their GCN proved to perform better than the QSAR methods for biodegradability prediction. To improve on top of this, we propose using the aforementioned Relation Graph Convolutional Layer to better account for edge relationship types, as individual edge weight relationships are learned. The hypothesis for this was based on the idea that there are different bond types within molecules, and understanding the relationships between atoms of different bond types will strengthen the model's understanding of molecular structure, and increase accuracy of property prediction.

The architecture for this model is shown in Figure 5. After converting our input data into graph format,

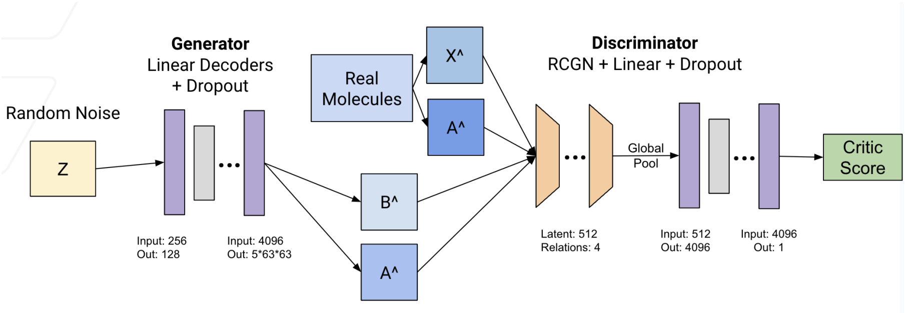


Figure 4: WGAN Architecture adapted from MedGAN paper (Macedo et al., 2024).

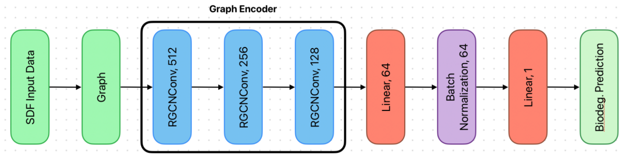


Figure 5: Model architecture for our biodegradability prediction model. We implemented a 3-layer RGCN encoder followed by two fully connected linear layers to output a probability score of our binary biodegradability class.

the model comprises of 3 R-GCN layers that help encode the graph information, while accounting for different bond types for edges. Then, the output is passed through 2 fully connected layers that provide an output that we use as a prediction for classification of a molecule as biodegradable or not. The molecules that were inputted into these models were the candidates that passed the stability threshold, and are then classified as biodegradable or non-biodegradable.

The accuracy of these models are extremely important. The outputted results that are biodegradable are the goal molecules that we are trying to generate to promote green chemistry. The priority is maintaining a high precision in order to reduce the amount of false positive biodegradability classifications. For these evaluation models, our goal is to achieve test accuracies at or above state of the art model accuracies by incorporating new model architecture that hasn't been utilized before.

5.4 Evaluation Metrics

5.4.1 Quantitative Metrics

Generative Models:

There are several metrics that generative models focus on (such as the ones in our existing works that we will compare to): Validity, Novelty, and Uniqueness/Diversity . For validity, we calculate the number of stable molecules out of the number of molecules outputted (assessed by the stability procedure described previously). To reiterate, our validity is calculated by double checking for existence of both bond and atoms while converting to a RDKit Mol object and using the RDKit validity function that (Macedo et al., 2024) employed in their evaluation. Therefore, our validity evaluations account for both our atomic number and bond generations. We also included chemical formulas to provide clearer, more concise visualizations of molecules. For novelty, we will calculate the number of new molecules in comparison to the training set over the total number outputted. For uniqueness, we will measure how different the generated molecules are from each other. This is simply the unique count of molecules among the set of generated molecules. We compare our metrics to the baseline of MedGAN due to its similarity in metrics calculation and our use of it.

Following these metrics, the average biodegradability is calculated based on our classifier's outputs. There is no baseline for this because no papers exist that generate biodegradable molecules.

Biodegradability:

For our evaluation of generated molecules, we require robust, highly accurate classification models that take a molecular structure as input and outputs predictions of whether it's toxic and/or biodegradable. Evaluation metrics for these models include balanced accuracy, precision, sensitivity, specificity, and error rate. Balanced accuracy accounts for accuracies in both the minority and majority class, which is especially important when considering our class imbalance. Specificity and sensitivity refer to reducing false positives and increasing true positives respectively. Although we need to acknowledge uncertainties and potential model inaccuracies, our goal is to generate a comprehensive list of readily biodegradable molecules that will need to by synthesized and tested for future materials.

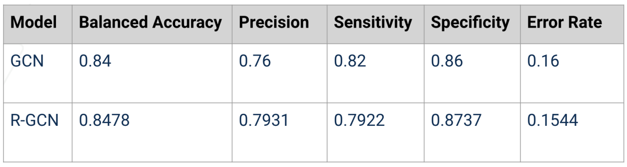
6 Results

6.0.1 Biodegradability Prediction

We used the existing GCN model (Lee and Min, 2022) as our benchmark to compare results for our Readily Biodegradable prediction model. This model had already surpassed traditional QSAR methods on these tasks. Our model surpassed state of the art models for the same task across standard evaluation metrics. As illustrated in Figure 6, our RGCN model achieves a balanced accuracy of 0.8478 compared to Lee and Min's 0.84. Additionally, we scored higher in precision and specificity while maintaining a lower sensitivity score.

These metrics are important because high precision means we are accurately predicting biodegradability for molecules that are not actually biodegradable. Similarly, our specificity score was high, which meant the model could accurately predict true negatives, and sensitivity was lower than the benchmark, meaning the model was more invariant to various changes.

Replacing the GCN with an RGCN allowed for higher accuracy for biodegradability prediction, and also demonstrates its effectiveness at understanding atom-bond interactions because of the individual weights it tracks for various edge types.



Model	Balanced Accuracy	Precision	Sensitivity	Specificity	Error Rate
GCN	0.84	0.76	0.82	0.86	0.16
R-GCN	0.8478	0.7931	0.7922	0.8737	0.1544

Figure 6: Chart depicting the our results labeled ”RGCN” compared with the state of the art GCN approach [Lee and Min, 2022]

7 Results & Experiments

7.1 General Results

The majority of our results can be found in Table 1, in which we compare our models to the MedGAN paper (Macedo et al., 2024) due to the similarity in metrics and design. The first row is the original MedGAN trained and generated on Zinc 15 database for drug discovery. Our MedGAN implementation was our version based on that paper and trained/generated on our combined data. The experiments for VGAE are explained below.

7.2 Loss Experiments

Experiments with the loss function for our VGAE model included the addition and ablation of loss regularizers and modifications to weights of the loss parts to determine the best architecture and hyperparameters for our generative task. Metrics we used to judge a given experiment's success were measured in the following order from most important to least considered: Validity, Average atoms in molecule, Uniqueness, Novelty. As seen with the results from our WGAN, validity of a molecule is important to be able to evaluate on our biodegradability model. A large problem with our VGAE generations was that the number of atoms generated in a given sample was often too many (over 57) or too few (less than 3). Initially, we experimented by adding a loss regularizer for atom count, which penalizes generated atomic number matrices that have too few atoms. More precisely, in this first experimental iteration we created a step-wise function where $atomic\_regularizer = 150 - (atom\_count * 10)$ if $atom\_count <= 7$. From this initial experiment, we saw a Uniqueness score of 30%, Molecular Validity of 100%, Novelty of 97%, Average Biodegradability score of 81.08%, and an average number of atoms per molecule of 5.66.

Prior to incorporating the atomic number and atomic bond regularizers, we generated molecules with up to 63 atoms, so our implementation of the regularizers was successful. Ideally, we want to generate larger molecules and hypothesized we could do so by increasing the atom count penalization. We performed three experiments overall: the first was the 7 atom count limit, the second was 15 atoms (which only resulted in worse uniqueness), and the last was a 50 atom count limit. Further experiments on increasing only the atom count regularizer function resulted in worse novelty, uniqueness, and biodegradability scores because it would overlearn a few types of the longer molecules generated from our normal distribution sample space. For our second loss experiment, we increased the KL Divergence weighting, decreased loss weights for the most common element types, and increased the number of bonds in the bond regularizer. We saw a Uniqueness score of 61%, Molecular Validity of 70.0%, Novelty of 78%, Average Biodegradability score of 74.05%, and an average number of atoms per molecule of 17.29.

7.3 Conditioning Experiments

We also conducted a series of experiments with conditioning with latent space concatenation. In this method, we employed a Coulomb matrix featurizer during or data processing step, ran it through a Linear layer to reduce dimensionality, and then concatenated the output to the Z sample in the latent space. We explored multiple variations on conditioning with Coulomb matrices, including various latent dimen-

Table 1: Performance Comparison of Molecular Generation Models

Model	Novelty	Validity	Uniqueness	AVG Biodegradability	AVG Atom Count
MedGAN (Zinc 15)	0.93	0.35	0.95	N/A	N/A
Our MedGAN	0	0	0	0	27
VGAE (Loss 1)	0.97	1	0.3	0.81	5.66
VGAE (Loss 2)	0.78	0.7	0.61	0.74	17.29
VGAE (Cond 1)	0.99	0.04	1	0.18	21.26
VGAE (Cond 2)	0.99	0.03	1	0.19	18.7

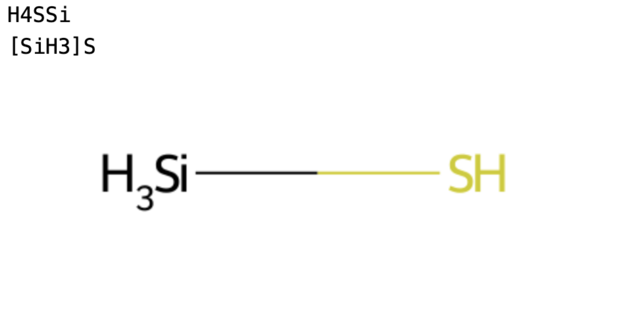


Figure 7: Example of a randomly generated molecule with chemical formula and SMILES format from our Loss Experiment 1.

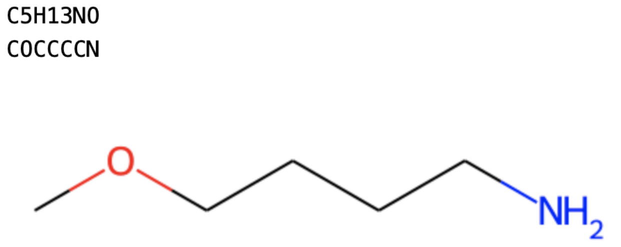


Figure 8: Example of a randomly generated molecule with chemical formula and SMILES format from our Loss Experiment 2.

sions of the original model and the conditioning features. For example, we experimented with latent dimensions of 256 and 512 with a conditioning addition of size 252. Our best model without conditioning has latent dimension of 512, so we examined the impact that different ratios of original latent space to featurizer concatenations had on our final generations.

The findings from these experiments are listed below: With a latent dimension of 256 and conditioning dimensions of 252 (overall 508 latent dimension size), we achieved Uniqueness of 100%, Molecular Validity of 4.0%, Novelty of 99%, Average Biodegradability score of 18.15%, and average atom count of 21.26.

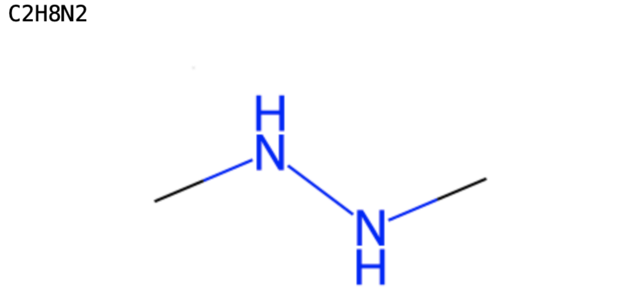


Figure 9: Example of a randomly generated molecule with chemical formula and SMILES format from our Conditioning Experiment 1.

For our experiment with latent dimension of 512 and conditioning dimension of 252, we achieved Uniqueness of 100%, Molecular Validity of 3.0%, Novelty of 99%, Average Biodegradability score of 19.06%, and average atom count of 18.7.

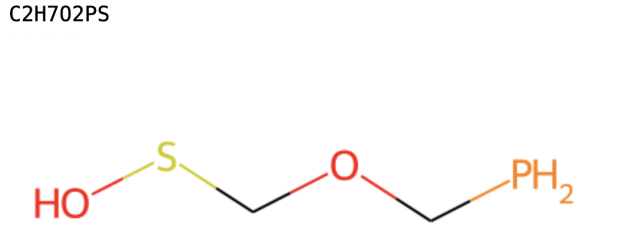


Figure 10: Example of a randomly generated molecule with chemical formula and SMILES format from our Conditioning Experiment 2.

8 Discussion

Our WGAN Implementation In comparison to the baseline and our VGAEs as shown in Table 1, our WGAN implementation failed miserably due its inability to generate valid molecules. This is likely because of representation chosen to represent bonds and atoms. Since no-atom and no-bond were encoded in matrices, the model was not able to learn that bonds should be somewhat rare given the maximum possibility of of every bond occurring. Our representation forces the model to learn sparse matrices which can be difficult to learn without additional loss experimentation as done for VGAE. Additionally the lack of data/limited training distribution could have caused some sort of mode collapse or issues where the generator comes up with loopholes to fool the discriminator due to the lack of learning by the discriminator.

Loss Experiment Results In comparison to the baseline, our loss experiments for the VGAE showed that our Novelty remained high and even beat the baseline when smaller were generated. Regarding validity, our loss experiments succeeded far past the paper's MedGAN achieving high validity for both loss experiments. Our uniqueness suffered likely due to the limited size of our molecules and potential overfitting.

From the Loss experiments, we saw that moderately increasing the atom count regularizer resulted in worse Novelty, Uniqueness, and Biodegradability scores. Our experiment with a 50 atom count regularizer improved our Uniqueness from 30% to 61% and increased the average number of atoms generated from 5.66 to 17.29. The increase in Uniqueness can be explained by the increase of molecular complexity. However, our results from our second loss experiment saw a decrease in Validity, Novelty, and Average Biodegradability. Upon exploration of our results and data, the decrease in Validity can be attributed to the average atom count change from 5.66 to 17.29. Due to the increased complexity of the molecules generated, we see worse generation of valid bonds between more atoms, supported by the bond adjacency loss. For future experiments, we would focus on improving bond reconstruction in our VGAE by increasing bond loss weights, increasing the number of learnable parameters in our bond decoder, and employing distance based featurizers or molecular fingerprints as our conditioning concatenations. The decrease in novelty signifies our model overlearned the training distribution space. As for the decrease in Biodegradability, it could be attributed to multiple possible reasons. Our generated molecules may be out of the distribution of our biodegradability training set, are overly complex, or have a larger invalid molecule subset.

Conditioning Experiment Results In comparison to the baseline, our conditional experiments for the VGAE were not nearly as successful. Our high novelty and uniqueness was due to the limited validity of our molecules.

Our findings from the Conditioning Experiments were interesting in that concatenating Coulomb information improved our Uniqueness scores to 100% as well as Novelty scores to 99%. Based on the decrease in Validity to 3.0 and 4.0% compared to our initial models Validity scores of 100% and 70%, it is reasonable to claim that the Uniqueness and Novelty scores improved significantly while Biodegradability decreased due to the lack of valid molecules generated with conditioning. As stated earlier, our future work lies in improving bond reconstruction via different featurizers for conditioning. More specifically, we want to explore the possibility of using distance based or extended connectivity fingerprints featurizers, as well as measuring electron density with a Gaussian Multipole featurizer.

9 Conclusion

As a review, in this paper, we discussed: the construction of our molecular graphs from datasets of non-toxic and readily biodegradable molecules, the creation of multiple generative models that synthesized new molecular structure candidates, and the classification and evaluation of these candidates as biodegradable and valid molecules. We achieved promising generation results that are biodegradable, novel, and diverse, which was the overall goal for this paper. Additionally, we reached state of the art results for classification on the compiled biodegradability dataset (Lunghini et al., 2020). We believe that this methodology can be extended and further used to reduce lab testing and costs, and improve the overall development process for new sustainable materials.

References

[Arjovsky et al. 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. arXiv [Stat.ML]. Retrieved from http://arxiv.org/abs/1701.07875

[Cao et al. 2022] Cao, N. D., Kipf, T. (2022). MolGAN: An implicit generative model for small molecular graphs. arXiv [Stat.ML]. Retrieved from http://arxiv.org/abs/1805.11973

[Fransen et al. 2023] Fransen, K. A., Av-Ron, S. H. M., Buchanan, T. R., Walsh, D. J., Rota, D. T., Van Note, L., Olsen, B. D. (2023). High-throughput experimentation for discovery of biodegradable polyesters. Proceedings of the National Academy of Sciences, 120(23), e2220021120. doi:10.1073/pnas.2220021120

[Iwata et al. 2023] Iwata, H., Nakai, T., Koyama, T., Matsumoto, S., Kojima, R., and Okuno, Y. (2023). VGAE-MCTS: A New Molecular Generative Model Combining the Variational Graph Auto-Encoder and Monte Carlo Tree Search. Journal of Chemical Information and Modeling, 63(23), 7392–7400. doi:10.1021/acs.jcim.3c01220

[Ketkar et al.2023] Ketkar R, Liu Y, Wang H, Tian H. A Benchmark Study of Graph Models for Molecular Acute Toxicity Prediction. International Journal of Molecular Sciences. 2023; 24(15):11966. https://doi.org/10.3390/ijms241511966

[Kingma and Welling 2013] Diederik P Kingma, and Max Welling. (2013). Auto-Encoding Variational Bayes.

[Kipf et al. 2016] Thomas N. Kipf, and Max Welling. (2016). Variational Graph Auto-Encoders.

[Lee and Min 2022] Lee, M., Min, K. (2022). A Comparative Study of the Performance for Predicting Biodegradability Classification: The Quantitative Structure Activity Relationship Model vs the Graph Convolutional Network. ACS Omega, 7(4), 3649–3655. doi:10.1021/acsomega.1c06274

[Lunghini et al.2020] Lunghini F., Marcou, G., Gantzer, P., Azam, P., Horvath, D., Van Miert, E., Varnek, A. (2020). Modelling of ready biodegradability based on combined public and industrial data sources. SAR and QSAR in Environmental Research, 31(3), 171–186. doi:10.1080/1062936X.2019.1697360

[Macedo et al. 2024] Macedo, B., Ribeiro Vaz, I. Taveira Gomes, T. MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design. Sci Rep 14, 1212 (2024). https://doi.org/10.1038/s41598-023-50834-6

[Mercant et al.2023] Merchant, A., Batzner, S., Schoenholz, S.S. et al. Scaling deep learning for materials discovery. Nature 624, 80–85 (2023). https://doi.org/10.1038/s41586-023-06735-9

[Mayr et al.2016] Mayr, A., Klambauer, G., Unterthiner, T., Hochreiter, S. (2016). DeepTox: Toxicity Prediction using Deep Learning. Frontiers in Environmental Science, 3:80.

[Goodfellow et al. 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Networks. arXiv [Stat.ML]. Retrieved from http://arxiv.org/abs/1406.2661

[Gulrajani et al. 2017] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved Training of Wasserstein GANs. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/1704.00028

[Huang et al. 2016] Huang, R., Xia, M., Nguyen, D. T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S., Rossoshek, A., Simeonov, A. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Frontiers in Environmental Science, 3:85.

[Sclichtkrull et al. 2017] Schlichtkrull, M., Kipf, T., Bloem, P., van den Berg, R., Titov, I., Welling, M. (2017) Modeling Relational Data with Graph Convolutional Networks

[Simonovsky and Komodakis 2018] Martin Simonovsky, and Nikos Komodakis. (2018). GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders.

[Tsujimoto et al. 2021] Tsujimoto, Y., Hiwa, S., Nakamura, Y., Oe, Y., Hiroyasu, T. (2021). L-MolGAN: An improved implicit generative model for large molecular graphs. ChemRxiv.

[Xiong et al.2020] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., … Zheng, M. (2020). Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. Journal of Medicinal Chemistry, 63(16), 8749–8760. doi:10.1021/acs.jmedchem.9b00959

[Blanchard et al. 2021] Blanchard, A. E., Stanley, C., & Bhowmik, D. (2021, February 23). Using GANs with adaptive training data to search for new molecules - journal of Cheminformatics. BioMed Central

[Huang et al. 2024] Huang, L., Xu, T., Yu, Y. et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. Nat Commun 15, 2657 (2024). https://doi.org/10.1038/s41467-024-46569-1

[Anoshin et al. 2023] Matvei Anoshin and Asel Sagingalieva and Christopher Mansell and Vishal Shete and Markus Pflitsch and Alexey Melnikov. Hybrid quantum cycle generative adversarial network for small molecule generation. https://doi.org/10.48550/arXiv.2402.00014

[Tan et al. 2023] Tan Z, Li Y, Wu X, Zhang Z, Shi W, Yang S, Zhang W. De novo creation of fluorescent molecules via adversarial generative modeling. RSC Adv. 2023 Jan 4;13(2):1031-1040. doi: 10.1039/d2ra07008a. PMID: 36686951; PMCID: PMC9811934.

[Wu et al. 2022] Lemeng Wu and Chengyue Gong and Xingchao Liu and Mao Ye and Qiang Liu. Diffusion-based Molecule Generation with Informative Prior Bridges. https://doi.org/10.48550/arXiv.2209.00865

[Xu et al. 2023] Minkai Xu and Alexander Powers and Ron Dror and Stefano Ermon and Jure Leskovec. Geometric Latent Diffusion Models for 3D Molecule Generation. https://doi.org/10.48550/arXiv.2305.01140

[Hua et al. 2023] Chenqing Hua and Sitao Luan and Minkai Xu and Rex Ying and Jie Fu and Stefano Ermon and Doina Precup. MUDiff: Unified Diffusion for Complete Molecule Generation. https://doi.org/10.48550/arXiv.2304.14621

[Gao et. al. 2023] Fragment-based deep molecular generation using hierarchical chemical graph representation and multi-resolution graph variational autoencoder. https://onlinelibrary.wiley.com/doi/10.1002/minf.202200215