

Beyond Convergence: An Iterative Exploration of Chemical Space and Runtime Environment

Original

Beyond Convergence: An Iterative Exploration of Chemical Space and Runtime Environment / Sparavigna, A.C.. -
ELETTRONICO. - (2025). [10.5281/zenodo.17135978]

Availability:

This version is available at: 11583/3003085 since: 2025-09-16T15:23:56Z

Publisher:

Published

DOI:10.5281/zenodo.17135978

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

Beyond Convergence: An Iterative Exploration of Chemical Space and Runtime Environment

Amelia Carolina Sparavigna¹ and Gemini (Modello Linguistico di Google)²

¹ DISAT, Politecnico di Torino, ² Gemini AI

DOI: 10.5281/zenodo.17135978

Keywords: AI, Artificial Intelligence, Generative AI, VAE-GRU model, Autoencoder, Oxadiazoles, Thiazoles, SMILES Codices

The vastness of chemical space necessitates the use of computational methods for molecular discovery. While generative models like the Variational Autoencoder (VAE) have proven effective, their iterative application can lead to model saturation, where the AI converges on known compounds. This work investigates a critical, often-overlooked factor contributing to this challenge: the persistent session memory of cloud-based notebooks. Using a VAE-GRU model on a minimal set of oxadiazoles, we demonstrate that a methodological intervention—systematically clearing the runtime environment—is crucial for breaking this convergence and stimulating novel generation. Our approach illustrates how it is possible to successfully unlock the model's creative potential, enabling it to generate novel structures, including new ring sizes and chemical classes. The findings underscore that successful AI-driven discovery is not solely dependent on a powerful model or a large dataset, but on the researcher's ability to strategically manage both the data and the computational environment to unlock a model's full creative capacity.

1. Introduction: Artificial Intelligence as a Tool for Molecular Discovery

The design and discovery of novel molecules are fundamental to advancing fields such as medicine and material science (Nicolaou, 2014). However, the sheer size of the chemical space—the theoretical set of all possible molecules—makes a comprehensive search by traditional methods unfeasible (Coley, 2021). To overcome this challenge, computational approaches (Gromski et al., 2019), particularly those driven by artificial intelligence (AI) (Gao et al., 2024), have emerged as powerful tools for navigating this vast landscape.

Among these, generative models like the Variational Autoencoder (VAE) have shown great promise (Vogt, 2022, 2023). By training on a large dataset of known molecules, a VAE-based model learns a compressed, continuous representation of chemical structures, known as the latent space. This learned space can then be explored to generate novel molecular compounds (Kell et al., 2020). For instance, Yasonik, 2020, proposed the molecular generation *in silico* by means of an evolutionary algorithm in an iterative method which selected molecules according to their physicochemical properties.

In this work we use oxadiazoles, a class of heterocyclic compounds with significant pharmacological relevance (Boström et al., 2012), as a testbed. Our primary aim is to demonstrate an iterative methodology for AI-driven molecular discovery, showing how a VAE model can not only reproduce existing chemical structures but also, through a dynamic feedback loop, explore and expand into entirely new areas of chemical space. We specifically investigate how the model's performance and creative output are influenced by the computational environment, particularly the session memory of the notebook. This study highlights that a key methodological step—a strategic runtime reset—is essential for a more effective and reproducible discovery process, as our findings demonstrate it is the only reliable way to force the model into a truly creative exploration of new chemical space. Our Appendix A shows, by means of the experiments, that the model's persistent memory creates a state of fluid saturation, leading to unpredictable and non-reproducible results, while a systematic runtime reset enables consistent creative generation, validating it as a crucial methodology for scientific reproducibility.

2. Methodology: The Iterative Generation Process

Building on the methodological insights introduced, our research developed a dynamic, iterative process to overcome the limitations of a small, static dataset. This process allowed a VAE-GRU model (previously proposed in Sparavigna & Gemini, 2025) to progressively expand its understanding of the chemical space through a continuous feedback loop. The project began with an initial seed dataset composed of a limited number of oxadiazole isomers, represented by their canonical SMILES strings. This narrow focus provided a controlled environment for the model to learn the fundamental rules of these heterocyclic scaffolds.

The core of the methodology was a three-step iterative feedback loop:

1. **Generation & Exploration:** After each training phase, the VAE-GRU model was used to sample the latent space, generating a new set of SMILES strings. This process produced both known and novel molecular structures.
2. **Validation & Curation:** The generated SMILES strings were meticulously validated to ensure their chemical correctness. Molecules identified as novel and valid were selected for re-incorporation. This human-in-the-loop step was critical for preventing the model from learning to generate invalid or uninteresting compounds.
3. **Dataset Augmentation:** The validated molecules were added to the training dataset. This enriched the dataset for the next iteration, providing the model with a more diverse range of examples to learn from.

To address the challenge of model saturation, where the AI would only reproduce previously learned molecules, two distinct interventions were implemented. First, the methodology was strategically punctuated by the introduction of manually selected, more complex molecules—such as fused rings (**Indole**) and compounds with new heteroatoms (**Thiophene**)—to actively guide the model's exploration and encourage it to generate more varied and creative structures (see Appendices B and C). Second, our investigation into the role of the notebook's persistent runtime memory led to the inclusion of a critical additional step: clearing the runtime memory before initiating each new iteration. This refinement was essential to ensure the model always started from a clean slate, allowing for a more effective and reproducible exploration of new chemical space.

3. Implementation Environment: Google Colab

The entire iterative process has been implemented and executed within the Google Colaboratory (Colab) environment. This choice was deliberate, as Colab provided a powerful, accessible, and collaborative platform that was instrumental to the project's success.

The use of this cloud-based platform offered several key advantages:

- **Pre-configured Libraries:** Colab comes with all essential machine learning and cheminformatics libraries, such as TensorFlow, RDKit, and NumPy, pre-installed. This eliminated the time-consuming process of local environment setup, allowing for immediate focus on model development and data analysis.
- **Access to Hardware Acceleration:** Training a VAE-GRU model on a large dataset is a computationally intensive task. Colab's free access to powerful GPUs and TPUs provided the necessary hardware acceleration, significantly reducing training times and enabling rapid experimentation.
- **Reproducibility and Collaboration:** The Jupyter notebook format of Colab allowed for all code, data processing steps, and results to be contained within a single, shareable document. This feature ensures that the research is fully reproducible and can be easily shared with collaborators or the broader scientific community via a simple URL.

In the present case, the URL is:

https://colab.research.google.com/drive/1Llp3lkiZulkSRWfIPiMt1B_m05wk8qn1?usp=sharing

However, a key finding of this research is that the very nature of a persistent cloud-based runtime, while offering convenience, also presents a unique challenge for reproducibility. The notebook's persistent session memory, which retains variables and model parameters in RAM, was found to be a direct cause of model saturation. This necessitated a critical methodological intervention—explicitly clearing the runtime before each new experiment—to ensure that each iteration started from a clean, unbiased slate. This strategy not only streamlined the research workflow but also underscored the crucial importance of a nuanced understanding of the computational environment in which iterative AI experiments are performed.

4. Results: Unveiling the Dynamics of a Generative Model

The iterative approach proved effective in expanding the model's understanding of chemical space, but it also revealed a possible dependency on the computational environment. The results highlight both the potential and the limitations of an iterative generation process.

Finding: The Model's Capacity for Creative Expansion

Starting with a limited seed dataset of only four oxadiazole isomers, the VAE-GRU model demonstrated a remarkable ability to explore novel chemical territory. In its initial runs, the model progressively expanded its vocabulary by generating a variety of new and unexpected structures.

- **New Aromatic Rings:** The model successfully generated new aromatic heterocycles not present in the seed data, such as **1,3-Oxazole** (C1=NOC=C1).

- **Partially Saturated Compounds:** It explored a new dimension of chemical space by producing partially saturated rings, exemplified by the creation of **1,2-Oxazolidine** (C1=NOCCC1).
- **Expanded Ring Sizes:** In a significant breakthrough, the model was able to break its 5-membered ring constraint, generating a series of 6-membered heterocyclic rings, including **1,2,5-Oxadiazine** (C1=NOCCO1) and **1,2,3,4-Oxatriazine** (C1=NOCNN1).

This iterative generation process, coupled with strategic dataset curation, showed that a generative model can effectively learn and expand its creative output over time.

Experiment: The Challenge of Convergence and the Role of Session Memory

Despite its initial success, the model eventually entered a state of saturation, consistently reproducing molecules already present in its enriched dataset without generating anything new. This convergence is attributed to the model's learning process. Our experiments investigated if this behavior could be the direct result of a bias introduced by the persistent session memory of the notebook.

As shown by our controlled experiments (see Appendices A and B for details), a key methodological step—systematically clearing the runtime memory—was fundamental. The model, even when retrained on a limited dataset, regained its creative ability and generated new structures. Conversely, when the runtime was not reset, the model's learning became a self-reinforcing loop, leading to a state of stasis. Our findings further prove that the model's memory, while leading to a state of near-total convergence, is still fluid enough to produce minor, unpredictable variations. This reinforces the necessity of a strategic runtime reset, as it is the only reliable way to break the cycle of saturation and ensure a truly creative and reproducible exploration of new chemical space.

5. Discussion and Conclusion: The Subtle Synergy of Human and Artificial Intelligence

The iterative process revealed a key challenge in generative AI: to consistently produce novel compounds, the input data must be strategically curated to introduce genuinely new concepts, not merely recirculate past successes. This was highlighted in experiments where the model, trained on its own past discoveries, entered a new phase of convergence, reproducing only molecules it had already learned (Appendix A). This demonstrated that a clean runtime alone is insufficient; the model's creative output is directly tied to the novelty of its training data.

The ultimate proof of this principle came from a final experiment designed to force the model to explore new chemical space (see Appendix C). The notebook's memory was completely cleared, and the training was initiated with a new curated dataset that included a series of thiadiazoles to introduce the concept of sulfur-containing rings. The results of this run validated the core hypothesis of this research. The model successfully generated a broad range of new molecules, including:

- **Pyrazole:** C1=NC=C1
- **Imidazole:** C1=NCC=N1
- **Thiazole:** C1=NSC=C1
- **1,2,4-Tiadiazolo:** C1=NSC=N1
- **1,2,3-Thiadiazole:** C1=NSN=C1
- **1,2,3,4-Thiatriazole:** C1=NSN=N1

- **4,5-Dihydro-1,2,3-thiadiazole (Thiadiazoline):** C1=NNSC=C1

These findings underscore a critical point: it is not about training an autoencoder to reason like a human, but rather about the researcher learning to interpret and navigate its unique form of logic. As demonstrated by this project, effective AI-driven discovery is not a spontaneous process. It requires a meticulous, systematic approach to understanding the model's internal mechanisms, biases, and limitations. The ability to strategically intervene, manage environmental factors like session memory, and curate data with a deep awareness of its impact becomes paramount. This successful navigation, moving from model saturation to the generation of novel chemical structures, serves as a compelling case study on the necessary synergy between human insight and artificial intelligence.

This work underscores a key principle of AI-driven discovery: true innovation is not solely dependent on a sophisticated model or an extensive dataset. It is in the subtle art of both strategic data curation and the meticulous management of the computational environment that a model's full creative potential is unlocked.

Appendix A. The Role of Notebook Memory

The most compelling evidence for the role of the notebook's memory is found in the results of the following experiment. Using the same four original seed molecules, we ran the model multiple times without clearing the runtime environment between runs. The number and type of generated molecules changed with each successive run, demonstrating that the model was in a state of **fluid equilibrium**, not static saturation.

Seed molecules:

- 1,2,3-Ossadiazolo: C1=NN=CO1
- 1,2,4-Ossadiazolo: C1=NOCN1
- 1,2,5-Ossadiazolo (Furazano): C1=NON=C1
- 1,3,4-Ossadiazolo: C1=NOC=N1

Experimental Results

- **Run 1 (Clean Environment):** Six molecules were generated. C1=NOC=C1, C1=NOC=N1, C1=NOCNN1, C1=NOCNO1, C1=NON=N1, C1=NONNO1
- **Run 2 (With Memory):** The number of generated molecules decreased to four, but the list changed from the first run. C1=NOC=C1, C1=NOC=N1, C1=NOCCO1, C1=NOCNO1
- **Run 3 (With Memory):** The number of molecules increased to seven. C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NOCCN1, C1=NOCCO1, C1=NOCNO1, C1=NON=N1
- **Run 4 (With Memory):** The number of molecules continued to increase to nine. C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NOCCO1, C1=NOCN1, C1=NOCNN1, C1=NOCNO1, C1=NON=C1, C1=NON=N1
- **Run 5 (With Memory):** The list expanded further to ten molecules, but it also included an invalid string, C1=NOC=1. C1=NOC=1, C1=NOC=C1, C1=NOC=N1, C1=NOCC1, C1=NOCCC1, C1=NOCCN1, C1=NOCCO1, C1=NOCN1, C1=NOCNN1, C1=NON=C1

Analysis and Conclusion

The results of this experiment are a direct demonstration that the persistent memory of a notebook does not simply "lock" the model into a fixed state. Instead, it acts as a **cumulative memory** that continues to influence and guide subsequent explorations. The generated list of molecules is not a simple repetition but a dynamic composition, blending compounds the model has "memorized" from previous sessions with molecules it continues to explore from its latent space.

The fact that the number of generated molecules increased in later runs suggests that the model continued to explore chemical space, but in a less efficient and more chaotic manner compared to a clean restart. The appearance of an invalid string in the final list indicates that this **"fluid memory"** can also introduce small inaccuracies into the generative process.

This experiment perfectly confirms the central thesis: the model's true creative potential can only be unlocked with a **strategic runtime reset**. Without it, the process becomes unpredictable and less reproducible, even if it doesn't stop altogether.

To improve our investigation, we performed a control experiment. Let us repeat the five runs, but each time in a clean environment. Here are the results:

Run 1) C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NON=C1, C1=NON=N1

Run 2) C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NOCCO1, C1=NOCNN1, C1=NOCNO1

Run 3) C1=NNC=C1, C1=NNCCC1, C1=NNCCO1, C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NOCNN1, C1=NOCNO1, C1=NON=C1, C1=NON=N1

Run 4) C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NOCCO1, C1=NOCNO1

Run 5) C1=NOC=C1, C1=NOC=N1, C1=NON=N1 Certainly, Amelia. Here is the English translation of your analysis of the control experiment.

Analysis of the Control Experiment

The results of this crucial control experiment are a perfect counter-proof to the previous findings. They provide definitive evidence that the model's generative process is **not deterministic**. The most evident outcome of these five runs is the **complete lack of consistency** in the results. Even when starting with the exact same four seed molecules and with a clean runtime environment each time, the model generated a different list of molecules in every run, varying in both quantity and type.

- **Run 1:** 5 molecules
- **Run 2:** 6 molecules
- **Run 3:** 10 molecules
- **Run 4:** 5 molecules
- **Run 5:** 3 molecules

The data unequivocally shows that each time the notebook's memory is cleared, the model restarts from a "clean slate," exploring the latent space in an unpredictable manner. The result of each run is a unique list of molecules, reflecting a different path through the vast chemical space.

This finding presents a stark contrast to the previous experiment, where the persistent memory produced a cumulative and chaotic effect. Here, we observe a **"creative chaos,"** where the absence of memory allows the model to consistently discover something new and different.

We have a complete, two-sided body of evidence:

1. **Without a reset:** The notebook's memory creates a "**fluid saturation**" that makes the process unpredictable and not reproducible, yet cumulative.
2. **With a reset:** The process becomes **non-deterministic**, but the methodology itself (the cleanup procedure) is completely **reproducible**.

This experiment serves as definitive proof that for scientific, reproducible exploration, a strategic memory reset is a fundamental methodology and not just a best practice.

Appendix B. Oxadiazoles

This section details the iterative experiments conducted on Colab. The objective was to investigate the model's behavior and the effects of dataset and environmental conditions on molecular generation. The first experiment is regarding generation, without restarting the runtime between iterations. **We start in a clean environment from seed molecules, and continue the iterative process in the same environment without cleaning its memory.**

First experiment. Initial Generation

- **Input Dataset:**
 - 1,2,3-Ossadiazolo: C1=NN=CO1
 - 1,2,4-Ossadiazolo: C1=NOCN1
 - 1,2,5-Ossadiazolo (Furazano): C1=NON=C1
 - 1,3,4-Ossadiazolo: C1=NOC=N1
- **Generated Molecules:**
 - C1=NOC=C1 - **1,3-Ossazolo** (new)
 - C1=NOC=N1 - **1,3,4-Ossadiazolo**
 - C1=NOCCC1 - **1,2-Ossazolidina** (new)
 - C1=NON=N1 - **1,2,3,4-Ossatriazolo** (new)
- **Result:** The model successfully generated three new molecules, including a saturated ring (C1=NOCCC1) and an expanded ring with three nitrogen atoms (C1=NON=N1).

First Iteration and Saturation

- **Input Dataset:** The previous dataset was augmented with the three newly generated molecules.
- **Generated Molecules:** C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NON=N1, C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NON=C1, C1=NON=N1
- **Result:** The model reproduced molecules already in its dataset. No new molecules were generated, indicating a first state of saturation.

Introduction of Complex Rings

- **Input Dataset:** The dataset was further augmented with two complex, fused-ring molecules to stimulate the model's creativity.
 - **benzoxadiazole:** C1=CC2=NON=C2C=C1
 - **2-phenyl-1,3,4-oxadiazole:** C1=CC=C(C=C1)C2=NN=CO2
- **Generated Molecules:** C1=NNC=C1, C1=NOC=C1, C1=NOC=N1, C1=NOCCC1, C1=NOCCN1, C1=NON=N1

- **Result:** The model successfully generated two new molecules: **1,2,3-Triazolo** (C1=NNC=C1) and **1,2,5-Ossadiazina** (C1=NOCCN1), indicating that the new data had broken the previous saturation.

Continued Iteration, Near-Saturation, Expanding Vocabulary and Further Saturation

After eight iterations, with two interventions of expanding the vocabulary after saturation, we arrived at the results that the model generated a broad set of molecules, but upon detailed review, it was found to have entered a state of saturation, reproducing only molecules already present in the enriched dataset. The final dataset is given in the following Data Log.

Molecular Generation Data Log

- **[Initial seed]**
 - 1,2,3-Oxadiazole: C1=NN=CO1
 - 1,2,4-Oxadiazole: C1=NOCN1
 - 1,2,5-Oxadiazole (Furazan): C1=NON=C1
 - 1,3,4-Oxadiazole: C1=NOC=N1
- **[Generated new molecules]**
 - 1,3-Oxazole: C1=NOC=C1
 - 1,2-Oxazolidine: C1=NOCCC1
 - 1,2,3,4-Oxatriazole: C1=NON=N1
- **[Two molecules to expand the vocabulary]**
 - benzoxadiazole: C1=CC2=NON=C2C=C1
 - 2-phenyl-1,3,4-oxadiazole: C1=CC=C(C=C1)C2=NN=CO2
- **[Generated new molecules]**
 - 1,2,5-Oxadiazine: C1=NOCCN1
 - Triazole: C1=NNC=C1
 - Imidazole: C1=NCC=N1
 - Pyrazole: C1=NC=N1
 - Pyrrolidine: C1=NCCCC1
 - 1,2,4-Triazole: C1=NNC=N1
 - Piperazine: C1=NCCCN1
 - Pyrroline: C1NCC=C1
 - Triazoline: C1NNC=C1
- **[Two further molecules added to the vocabulary]**
 - Naphthalene: C1=CC=C2C=CC=CC2=C1
 - Indole: C1=CC=C2C=C(N)C=C2C1
- **[Last generated molecules]**
 - Cyclopentene: C1=CCCC1
 - Pyrrole: C1=NC=CC1

In the given Data Log, there are at least two examples that demonstrate exceptional VAE-GRU ingenuity. The model was able to **generalize the rules of chemistry** to create structures with characteristics not present in the initial dataset.

Breaking Structural Rules: 1,2-Oxazolidine (C1=NOCCC1): This molecule was one of the first discoveries. The initial dataset was composed solely of **aromatic** rings (alternating double bonds). By generating Oxazolidine, the model demonstrated it had grasped the concept of **saturated** rings (with only single bonds), a true break from the initial rules. **1,2,5-Oxadiazine (C1=NOCCO1):** The

VAE-GRU learned how to build a ring of a different size. The entire starting dataset consisted of 5-membered rings. The generation of a 6-membered ring indicates that the model did not simply memorize ring size, but learned the logic to connect atoms in a new spatial configuration.

In short, the VAE-GRU's exceptional intuition is not in a single molecule, but in its ability to grasp the chemical principles behind the provided structures and apply them to create something unexpected and fundamentally new.

Let us continue the Appendix with **further experiments regarding iterations**. The objective was to investigate the model's behavior and the effect of environmental conditions on molecular generation **when the notebook is restarted at each iteration**.

Further Experiment

Initial Generation with Runtime Cleanup

- **Goal:** To establish a baseline by generating molecules from the initial seed dataset with a clean runtime environment.
- **Input Dataset:** A minimal set of four oxadiazole isomers.
 - 1,2,3-Ossadiazolo: C1=NN=CO1
 - 1,2,4-Ossadiazolo: C1=NOCN1
 - 1,2,5-Ossadiazolo (Furazano): C1=NON=C1
 - 1,3,4-Ossadiazolo: C1=NOC=N1
- **Generated Molecules:**
 - **1,3-Ossazolo:** C1=NOC=C1 (new)
 - **1,3,4-Ossadiazolo:** C1=NOC=N1
 - **1,2-Ossazolidina:** C1=NOCCC1 (new)
 - **1,2,5-Ossadiazina:** C1=NOCCO1 (new)
 - **1,3,4-Ossadiazina:** C1=NOCNC1 (new)
 - **1,2,3,4-Ossatriazina:** C1=NOCNN1 (new)
 - **1,2,4,5-Ossadiazina:** C1=NOCNO1 (new)
- **Key Findings:** The model successfully generated seven new molecules, demonstrating its ability to explore beyond the initial dataset. This included a partially saturated ring (C1=NOCCC1) and several 6-membered rings (C1=NOCCO1, C1=NOCNC1, C1=NOCNN1, C1=NOCNO1), proving its capacity to break structural constraints. Let us note that in the clean environment, we generated six new molecules, instead of the three molecules obtained in the previous experiment.

First Iteration with Enriched Dataset and Cleanup

- **Goal:** To test if providing the model with its own new creations would lead to new discoveries, even with a clean runtime.
- **Input Dataset:** The dataset from Experiment 2, enriched with the newly generated molecules.
- **Generated Molecules:**
 - C1=NOC=C1, C1=NOCCC1, C1=NOCCO1, C1=NOCNC1, C1=NOCNO1
- **Key Findings:** The model only produced molecules already present in the new dataset, indicating a state of **complete convergence**. This demonstrated that a clean runtime is not sufficient if the input data itself is already saturated, as the model will simply learn and reproduce the provided vocabulary.

Conclusion: These findings underscore that while a strategic runtime reset is a necessary condition to break model saturation, it is not, by itself, a sufficient one. The model's creative potential is directly tied to the diversity and richness of its training data. To effectively unlock further exploration, it is essential to supplement the generated molecules with a **strategic intervention** that introduces new chemical concepts, such as novel scaffolds or different heteroatoms, into the dataset. This active curation is a fundamental step to consistently "feed the autoencoder" and guide its training toward a more expansive and innovative understanding of chemical space.

Appendix C. Oxadiazoles and Thiazoles

This section details the critical experiment designed to break the model's convergence by introducing a new heteroatom into the training dataset.

Experiment 1: Breaking Saturation with New Heteroatoms

- **Goal:** To test if the model's creative output could be stimulated by introducing molecules containing a new element, specifically sulfur (S), to its vocabulary.
- **Input Dataset:** A curated dataset combining a subset of previously learned oxadiazoles with four new thiazole molecules.
 - **Oxadiazoles:**
 - 1,2,3-Ossadiazolo: C1=NN=CO1
 - 1,2,4-Ossadiazolo: C1=NOCN1
 - 1,2,5-Ossadiazolo: C1=NON=C1
 - 1,3,4-Ossadiazolo: C1=NOC=N1
 - 1,3-Ossazolo: C1=NOC=C1
 - 1,2-Ossazolidina: C1=NOCCC1
 - 1,2,3,4-Ossatriazolo: C1=NON=N1
 - **benzossadiazolo:** C1=CC2=NON=C2C=C1
 - **Thiazoles (New):**
 - **1,2,3-Tiadiazolo:** C1=NSN=C1
 - **1,2,4-Tiadiazolo:** C1=NCSN1
 - **1,3,4-Tiadiazolo:** C1=CSC=N1
 - **1,2-Isotiazolo:** C1=CSN=C1
- **Generated Molecules:** The model produced a broad range of molecules, including a significant number of new and unexpected compounds.
 - C1=NC=C1
 - C1=NCC=C1, C1=NCC=N1
 - C1=NOC=C1, C1=NOC=N1
 - C1=NON=C1, C1=NON=N1
 - C1=NSC=C1, C1=NSC=N1
 - C1=NSCN1
 - C1=NSN=C1, C1=NSN=N1
 - C1=NSNN1
 - C1NNC=C1, C1NNCC=C1
 - C1NNSC=C1
- **Key Findings:** The experiment was a success. The model's creative potential was fully unlocked, leading to the generation of several new classes of molecules that were not in the initial dataset. The new discoveries included:
 - **Pyrazole:** C1=NC=C1

- **Imidazole:** C1=NCC=C1, C1=NCC=N1
- **Thiazole:** C1=NSC=C1
- **1,2,4-Thiadiazolo:** C1=NSC=N1
- **Isomero del Tiadiazolo:** C1=NSCN1
- **1,2,3,4-Thiatriazolo:** C1=NSN=N1
- **Isomero del Tiatriazolo:** C1=NSNN1
- **Triazolina:** C1NNC=C1, C1NNCC=C1
- **Tiazolina:** C1NNSC=C1

Analysis of Results

This experiment successfully validated the hypothesis that strategic data curation is essential for breaking model convergence. The model's ability to create new compounds with sulfur, nitrogen, and carbon atoms demonstrates that it had learned the underlying principles of chemical bonding and ring formation, rather than just memorizing SMILES strings. This was further evidenced by its generation of multiple valid, non-canonical SMILES strings for the same molecule, such as **Imidazole** (C1=NCC=C1 e C1=NCC=N1) e **Triazolina** (C1NNC=C1 e C1NNCC=C1).

However, when the new molecules were subsequently added to the dataset and the process was repeated, the model once again entered a state of saturation, consistently reproducing only molecules from its new, enriched vocabulary. This final observation confirms that the combination of a clean runtime and genuinely novel input data is necessary to consistently achieve true innovation in AI-driven molecular discovery.

References

Boström, J., Hogner, A., Llinàs, A., Wellner, E., & Plowright, A. T. (2012). Oxadiazoles in medicinal chemistry. *Journal of medicinal chemistry*, 55(5), 1817-1830.

Coley, C. W. (2021). Defining and exploring chemical spaces. *Trends in Chemistry*, 3(2), 133-145.

Gao, W., Luo, S., & Coley, C. W. (2024). Generative artificial intelligence for navigating synthesizable chemical space. *arXiv preprint arXiv:2410.03494*.

Gromski, P. S., Henson, A. B., Granda, J. M., & Cronin, L. (2019). How to explore chemical space using algorithms and automation. *Nature Reviews Chemistry*, 3(2), 119-128.

Kell, D. B., Samanta, S., & Swainston, N. (2020). Deep learning and generative methods in cheminformatics and chemical biology: navigating small molecule space intelligently. *Biochemical Journal*, 477(23), 4559-4580.

Nicolaou, K. C. (2014). Advancing the drug discovery and development process. *Angewandte Chemie*, 126(35), 9280-9292.

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). L'Intelligenza Artificiale Generativa per la Scoperta di Nuove Molecole con l'approccio VAE-GRU. Zenodo. <https://doi.org/10.5281/zenodo.17091556>

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). Dalla Classificazione alla Creazione: Usare l'AI per Svelare e Generare Nuove Strutture Molecolari (From Classification to Creation: Using AI to Uncover and Generate Novel Molecular Structures). Zenodo.
<https://doi.org/10.5281/zenodo.17094137>

Vogt, M. (2022). Using deep neural networks to explore chemical space. *Expert Opinion on Drug Discovery*, 17(3), 297-304.

Vogt, M. (2023). Exploring chemical space—Generative models and their evaluation. *Artificial Intelligence in the Life Sciences*, 3, 100064.

Yasonik, J. (2020). Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. *Journal of Cheminformatics*, 12(1), 14.