Evaluating the COATI Model: Generative Capabilities and Performance Assessments

Stefan Hangler
Report - Practical Work in AI
Institute for Machine Learning
Johannes Kepler University Linz
stefan.hangler@outlook.com

Abstract

This report examines the COATI model, a novel generative model designed for molecular design, leveraging multimodal data and contrastive learning to overcome limitations commonly associated with learning molecular representation. The model's architecture incorporates a Graph Neural Network (GNN) and a transformer, enabling effective representation and generation of molecules from SMILES strings and 3D point cloud data. Extensive experiments were conducted to validate the model's generative capabilities and assess its performance through linear probing tasks using the ADMET dataset. Despite challenges posed by the unavailability of training data, which complicates the validation process, the newly developed COATI 2 model demonstrated promising enhancements in performance across several metrics. Additionally, the model was trained from scratch on the GuacaMol dataset, providing insights into its potential when adequately trained. The code and data supporting this work can be found at https://github.com/StefanHangler/COATI-Model-Evaluation.

Story Summary

1. What is the central question?
Can the COATI model effectively generate and represent molecular structures, and how do its capabilities compare with other molecular related models on linear probing tasks?

2. Why is this question important?
If the COATI model excels in linear probing regression and classification tasks, it indicates that the model has effectively learned a robust molecular representation. Evaluating its generative capabilities is crucial to determine its applicability and utility in practical scenarios.

3. What evidence/data (variables) are needed to answer this question?
Data on molecular validity, uniqueness, novelty, and computational metrics like RMSE and Spearman correlation from models trained on the GuacaMol dataset are required.

4. What methods are used to get this evidence/data?
In this report, embeddings are computed using COATI models for ADMET data, which are subsequently employed in linear probing tasks. The GuacaMol dataset, once downloaded, serves as the basis for training the COATI model. The trained model is then utilized to generate molecular samples that are evaluated against benchmark metrics.

5. What analyses must be applied for the data to answer the central question?
Analyses involve linear probing for property prediction, evaluation of generative capabilities through metrics like FCD, and comparative analysis against established benchmarks.

6. What evidence/data (values for the variables) were obtained?
The data included molecular generation metrics such as validity, uniqueness, novelty, FCD and performance scores on linear probing tasks from multiple models.

7. What were the results of the analyses?
Results indicated that COATI, especially the newer COATI 2 model, demonstrated superior generative capabilities and effective representation learning compared to other models with keeping in mind that there is a very likely dataset-leakage and therefore an overestimation of model performances. The trained model from scratch with the GuacaMol Dataset cannot achieve baseline results with this less training

8. How did the analyses answer the central question?
The analyses demonstrated that COATI models consistently outperform CLAMP in most regression tasks, illustrating their superior molecular representation capabilities. The latest iteration, COATI 2, excels or closely competes for the top spot in all classification tasks. Additionally, preliminary results from the model trained from scratch suggest potential to match or exceed baseline performance with extended training.

9. What does this answer tell us about the broader field?
The findings affirm the potential of advanced machine learning models like COATI to transform molecular design, highlighting the importance of multimodal learning and the need for open-source data for validation.

1 Introduction

Molecular design is a challenging field that involves navigating a vast and complex chemical space to identify molecules with specific properties and functions. This field integrates approaches from drug discovery, materials science and synthetic chemistry. Generative models have revolutionised the approach to molecular design by enabling the automated generation of novel molecular structures. Despite advances, these models often face limitations related to molecular representation, adequacy of training data, and latent space optimisation. To address these challenges, the COATI model (Kaufman et al., 2023) presents a novel architecture that exploits multimodal data representations and contrastive learning for enhanced generative capabilities, which are analysed in this report.

A review of the relevant literature reveals that the representation of molecular structures has undergone a significant evolution, with string-based formats such as SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2020) becoming standard tools for encoding molecular topologies. These representations facilitate the application of deep learning models to predict molecular properties and generate new molecules, as evidenced by the work of Honda et al. (2019) and Olivecrona et al. (2017). In addition, molecular fingerprints, which provide a method for capturing essential chemical information, have been widely used in quantitative structure–activity relationship (QSAR) modelling and database searching (Morgan, 1965; Muratov et al., 2020).

In recent years, contrastive learning has emerged as a promising approach to enhance the quality of molecular representations by aligning data across different modalities. This approach has been exemplified by the use of augmentation techniques, as demonstrated in the work of Radford et al. (2021) and Stärk et al. (2021). The COATI model builds upon this foundation by integrating a multimodal encoder-decoder architecture that efficiently traverses and optimises latent space.

Other notable developments include CLAMP (Seidl et al., 2023), which aligns SMILES string embeddings with scientific assay descriptions, and MolReactGen (Holzgruber, 2024), which demonstrates robust molecular generation capabilities with an GPT2 architecure. These models, together with COATI, form the basis of a comparative analysis in later sections.

This work is dedicated to a comprehensive evaluation of the COATI model's generative capabilities. Key activities include:

- **Training the model from scratch** on the GuacaMol dataset to benchmark its performance against other models.
- **Analyzing the generative capabilities** of different COATI models.
- **Conducting linear probing** on both regression and classification tasks using the ADMET dataset from the authors to assess representational learning.

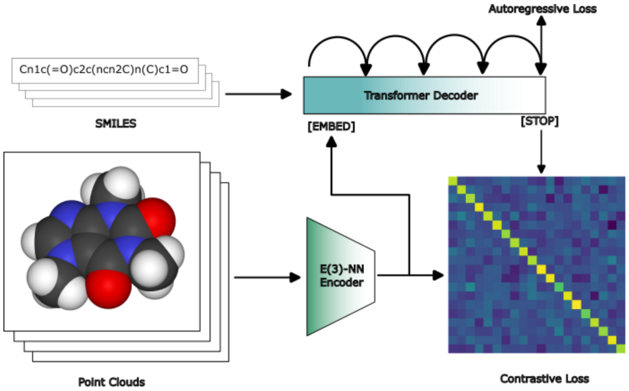

Figure 1: The model architecture features a Transformer decoder that processes SMILE strings and a GNN encoder for corresponding point cloud data. These components are jointly trained using a combination of contrastive and autoregressive losses (Kaufman et al., 2023).

It is noteworthy that the authors of the COATI paper have facilitated reproducibility by providing extensive resources in their repository. This includes a variety of Jupyter notebooks that reproduce most of the results presented in their study, covering topics from generation and autoencoding to chemical space exploration. They have also shared comprehensive codebases for replicating their experiments on conditional generation of therapeutics.

2 Model Architecture

The model integrates a multi-modal encoder-decoder architecture comprising a Graph Neural Network (GNN) and a transformer. The GNN processes a molecule's point cloud, and the transformer processes its SMILES string, with both models learning the 3D-textual relationship via contrastive loss, enabling molecule generation (shown in figure 1).

2.1 Text Encoder

The COATI model employs a rotary transformer to encode SMILES strings and graphs with a trie encoder. The encoder uses Byte Pair Encoding (BPE) to develop a vocabulary from basic SMILES characters to more complex combinations, forming open and closed sets of tokens for structural balance and validity. Special tokens enhance dynamic information integration, improving molecular structure understanding. In addition, RoFormer incorporates Rotary Position Embedding (RoPE) (Su et al., 2021), which manages the positions of tokens through a rotation matrix, thereby maintaining vector norms and improving computational efficiency.

2.2 Point Encoder

Employing the E(3)-equivariant Graph Neural Network (GNN), the point encoder adeptly manages 3D molecular structures by respecting spatial symmetries. It uses 'periodic one-hot' vectors for atom nodes, allowing the model to handle chiral information variably. This GNN is noted for its computational efficiency and its ability to maintain high-quality molecular representations without requiring intensive computational resources.

2.3 Learning Objectives and Dataset

The model's training combines contrastive loss and autoregressive cross-entropy, optimizing it for encoding and decoding molecular structures. Using the InfoNCE and Barlow cross-correlation losses, it aims to maximize matched embeddings' cosine similarities while minimizing unmatched pairs:

$$\mathcal{L}_{\text{Barlow}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum j \neq i C_{ij}^2 \tag{1}$$

Table 1: **Overview of Model Specifications:** This table summarizes key characteristics of each COATI model used in this report including the number of tokens processed, parameter counts, loss functions used, and dimensionality of model components.

Model	**Tokens**	**Params**	**Loss**	**Dimensions**
Grande Closed	1,152,504,376	xformer: 17.92M Total: 20.36M	InfoNCE + AR	E(3)-GNN: 5×256 Transformer: $16 \times 16 \times 256$ Latent: 256
Barlow Closed	4,649,953,856	xformer: 17.92M Total: 20.36M	Barlow + AR	E(3)-GNN: 5×256 Transformer: $16 \times 16 \times 256$ Latent: 256
Autoreg Only	6,796,944,213	xformer: 17.92M Total: 20.36M	AR	Transformer: $16 \times 16 \times 256$ Latent Dim: 256
COATI 2	$\sim$ 2x more data	xformer: 54.81M Total: 54.81M	InfoNCE + AR	Chiral-aware 3D GNN: ? Transformer: $16 \times 16 \times 256$ Latent Dim: 512 (new!)

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2b} \sum_{i} \left(\ln \frac{\exp\left(z_{s,i}^{\top} zp, i\right)}{\sum_{j=0}^{K} \exp\left(z_{s,i}^{\top} z_{p,j}\right)} + \ln \frac{\exp\left(z_{s,i}^{\top} z_{p,i}\right)}{\sum_{j=0}^{K} \exp\left(z_{s,j}^{\top} z_{p,j}\right)} \right) \tag{2}$$

The training set comprises over 140 million (SMILES, geometry) tuples from diverse sources, though it's not open source, posing significant challenges in model evaluation.

2.4 Models

The models detailed in Table 1 are used extensively in the experiments to assess their performance in linear probing and molecule generation tasks.

It is stated from the autors that the Grande Closed model is highlighted as the top performer across general applications, excelling due to its balanced approach to both text and graphical molecular data representation. Conversely, the Barlow Closed model demonstrates superior results specifically in linear probing tasks, which can be attributed to its effective utilization of the Barlow Twins loss method that emphasizes learning distinct and informative features from the data.

In molecule generation tasks, the Autoreg Only model, which operates without the benefit of contrastive loss, shows exceptional performance. This model relies solely on autoregressive loss to enhance its capability in generating coherent and chemically valid molecular structures.

A notable addition to the model lineup is the COATI model, released in March 2024 (COATI 2). This model has been trained on approximately twice the data volume compared to its predecessors, though the dataset remains proprietary, posing significant challenges in evaluation due to potential data leakage issues. The COATI model notably boasts over 54.81 million parameters, a substantial increase that suggests enhanced learning capabilities. Moreover, this model introduces a chiral-aware 3D GNN and doubles the latent dimension to 512, underscoring its advanced capacity to handle complex molecular data. However, the unavailability of its training code limits the ability for replication.

3 Experiments

3.1 Train from Scratch - GuacaMol Dataset

This section focuses on the training of the COATI model architecture from scratch, utilizing the GuacaMol dataset, which is a benchmark designed to evaluate the generative capabilities of molecular models. An overview of the GuacaMol benchmark is provided, followed by detailed preprocessing steps tailored for this training. The section concludes with a comprehensive presentation of the training parameters and specific computational aspects involved.

Table 2: **Distribution of GuacaMol Dataset Splits:** Details of molecule distribution across training, validation, and test sets, showing the number of molecules and their corresponding percentages.

Data Split	Number of Molecules	Percentage
Training	1,273,103	80%
Validation	79,567	15%
Test	238,705	5%
Total	1,591,375	100

3.1.1 GuacaMol Benchmark

The GuacaMol dataset by Brown et al. (2019) is a standardized benchmark used to assess the ability of generative models to create molecules with desired properties. It is based on the ChEMBL database and provides molecules in the SMILES format. This benchmark is crucial in the field of computational drug discovery, providing a variety of tests that simulate real-world challenges in generating viable molecular candidates.

In this work the GuacaMol benchmark is employed to train the models using the predefined train, valid and test splits of the dataset, ensuring that all models are exposed to the same training conditions, which are listed in table 2. This setup allows for a direct, fair comparison in the evaluation phase, where the COATI model's performance is assessed against other models trained under similar conditions. The benchmark effectively measures how well each model captures and replicates the distribution characteristics of the training data, emphasizing the creation of chemically valid structures. This approach helps in determining the effectiveness of the models in producing novel and practical molecular structures suitable for further drug development processes.

3.1.2 Data Preprocessing

To effectively train the combined Graph Neural Network (GNN) and Transformer model, preprocessing steps are tailored to prepare both SMILES strings and their associated atomic coordinates:

- **Molecular Fingerprints:** Molecules are transformed into Morgan fingerprints via the `mol_to_morgan` function. This function encodes the molecular structure into a fixed-size vector of 2048 bits with a radius of 3, capturing essential substructural features.
- **Atomic Coordinates and Features:** The `mol_to_atoms_coords` function converts SMILES strings into RDKit molecule objects, adding hydrogens if necessary and optimizing the molecular conformation using the Merck Molecular Force Field (MMFF) when requested. It outputs atomic numbers and 3D coordinates essential for the GNN's input. Additional outputs, such as adjacency matrices and the conformer's lowest energy, are also available if required.

All named functions are from the library `rdkit.Chem.AllChem` (rdk). These preprocessing steps are used for all data splits to ensure that the model is provided with accurate structural and textual data representations, facilitating robust learning and effective molecular generation.

3.1.3 Training

The training process was executed using the script in the github repository provided by the authors of the referenced paper, with hyperparameters explicitly detailed in Table 3. These parameters were carefully selected to align with the recommendations from the paper, ensuring that the training regimen was both standard and reproducible.

A slight modification was made to the dataset class to accommodate fixed training, validation, and testing splits. The original dataset class did not support such partitions, which are essential for correct model evaluation with other models trained on the GuacaMol benchmark.

The training was conducted over 7 epochs, processing a total of 128 million tokens. This was executed on an NVIDIA L4 GPU, equipped with a computational capacity of 121 TFLOPs and 24GB of memory, supplemented by 16 CPU cores. Despite the substantial hardware resources, the training

Table 3: **Hyperparameters Utilized in Model Training:** List of hyperparameters, their values, and descriptions, highlighting the configuration settings for the GuacaMol training of the model.

Hyperparameter	**Values**	**Description**
n_epochs	7	Number of complete passes through the entire training dataset.
batch_size	160	Number of samples processed before the model parameters are updated.
lr	5.0e-04	Learning rate for the optimizer.
weight_decay	0.1	Weight decay to apply (used in regularization).
clip_grad	10	Maximum value for gradient clipping.
n_hidden_xformer	256	Number of hidden units in the transformer layers.
n_layer_xformer	16	Number of layers in the transformer model.
n_head	16	Number of attention heads in the transformer model.
n_layer_e3gnn	5	Number of layers in the E(3)-equivariant Graph Neural Network.
n_hidden_e3nn	256	Dimensionality of hidden layers in the E(3)-equivariant neural network.
msg_cutoff_e3nn	12.0	Cutoff distance for message passing in the E(3)-NN.
max_n_seq	250	Maximum sequence length for inputs to the model.
dtype	float32	Data type for tensors, typically float or double.
norm_clips	True	Whether to apply normalization and clipping to inputs.
token_mlp	True	Whether to use a Multi-Layer Perceptron for token processing.
p_randsmiles	0.3	Probability of using randomized SMILES transformations.
p_clip	0.9	Probability of using Clip Token Augmentation.
p_clip_emb_smi	0.5	Probability of clipping embeddings specifically for SMILES.
tokenizer_vocab	"mar"	Specifies the vocabulary set for the tokenizer.
p_dataset	0.2	Probability to augment data with dataset information.
test_interval	2	Frequency (in epochs) of testing/validation.
biases	True	Whether to use biases in the neural network layers.

was halted after approximately 12 hours due to limitations in available compute resources. The progression of the training, including the loss reduction over time, is depicted in Figure 2, illustrating the model's learning trajectory until the point of interruption.

3.2 Linear Probing

3.2.1 Datasets

For the linear probing tasks I used the dataset which is provided by the authors, who combined publicly available data with proprietary data from the Terray platform to create a robust and comprehensive dataset suitable for regression analysis of molecular properties. This hybrid dataset is specifically curated to overcome the common limitations associated with the small size of typical chemical datasets, which often contain fewer than 1000 data points. By augmenting this with millions of data points from diverse sources, the dataset provides a rich basis for evaluating the predictive power of molecular representations.

The primary focus of the linear probing was on regression tasks related to critical molecular properties such as potency against drug targets and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) characteristics. These properties are crucial for determining the viability of molecules as they progress through the drug development pipeline. The dataset includes detailed molecule-target binding data, which offers millions of scalar values that indicate the binding affinity of molecules to specific protein targets.

The rich variety and volume of data in this dataset ensure that the findings from the linear probing tasks are robust and generalizable across various molecular properties.

Detailed specifications, including size and descriptions of the datasets, are available in the appendix of the original paper (Kaufman et al., 2023) under Table 5. The used datasets for the linear probing tasks are listed in the Tables 6, 7, and 8.

3.2.2 Training

For the new COATI 2 model, embeddings were specifically computed for each dataset. Unfortunately, due to computational resource constraints, it was not feasible to recompute embeddings for all other

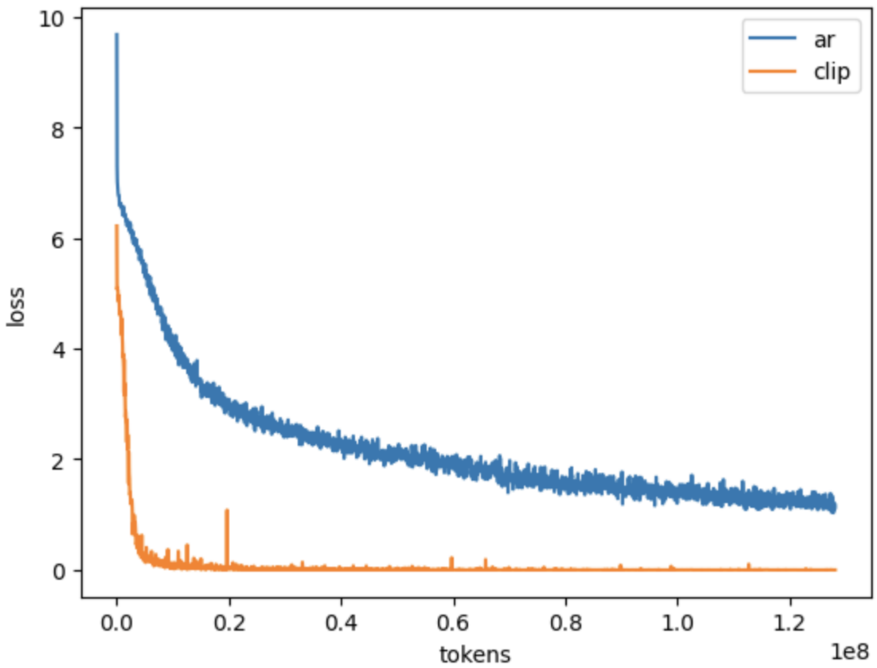

Figure 2: Loss curves over 7 epochs which means the model was trained on 128M tokens from the GuacaMol train set. The blue curve represents the autoregressiv loss and the orange curve the InfoNCE Loss.

models. Instead, precomputed embeddings provided by the authors were used under the assumption that they were correctly generated. However, the provided codebase does allow for the recomputation of all embeddings, enabling future work to update or validate these embeddings as necessary.

The datasets were split using the scaffold splitting method. Scaffold split is widely recognized as both the best and most challenging method for splitting molecular datasets because it groups molecules based on their chemical scaffold. This method ensures that structurally similar molecules are grouped together, providing a rigorous test of a model's ability to generalize to new, structurally diverse molecules. Scaffold splitting thus represents a realistic challenge that models might face in real-world drug discovery scenarios.

Simple models were employed to evaluate the linear probing tasks. This approach ensures that the differences in performance across various embeddings are attributed to the quality of the embeddings themselves, rather than the capacity of a complex regression or classification model. This strategy prevents any model from gaining an undue advantage due to the expressive power of the regressor.

- **Classification Tasks:** For tasks involving classification, a logistic regression model was utilized, configured with parameters for maximum 1500 iteration, L2 regularization, lbfgs solver and balanced class weights.

- **Regression Tasks:** Regression tasks were conducted using a random forest regressor, set up with 100 estimators to provide robust and stable predictions across different molecular datasets.

To enhance the robustness and generalizability of the results, each probing task was run across 10 different seeds. This procedure ensures that the findings are not overly dependent on any particular random initialization or split of the data, thereby providing more reliable insights into the efficacy of the embeddings.

4 Results

4.1 GuacaMol Benchmark

4.1.1 Metrics

In evaluating the performance of molecular generative models on the GuacaMol benchmark, several key metrics are employed to assess different aspects of the generated molecules:

- **Validity:** This metric measures the percentage of generated SMILES strings that represent chemically valid molecules. A higher validity score indicates that a model generates fewer erroneous or uninterpretable molecular structures.
- **Uniqueness:** This metric assesses the diversity within the set of generated molecules by calculating the proportion of unique valid molecules. High uniqueness suggests that the model is capable of generating a varied array of molecular structures, rather than repeating a limited set of motifs.
- **Novelty:** Novelty quantifies the proportion of generated molecules that are not found in the training data, indicating the model's ability to extrapolate beyond observed examples and generate new molecular entities.
- **FCD (Fréchet ChemNet Distance) (Preuer et al., 2018):** The FCD measures the distance between the distribution of generated molecules and the distribution of molecules in a reference dataset (typically the training set) in a learned feature space. The metric is calculated using a neural network trained to predict properties of molecules. A lower FCD value is preferable, indicating that the distribution of generated molecules closely resembles the reference dataset.
- **FCD_GuacaMol:** The GuacaMol benchmark provides a modified interpretation of the FCD, expressed as $\exp(-0.2 \times \text{FCD})$. This transformation results in a metric where higher values (up to 1) are more desirable, reflecting better performance. For consistency and ease of interpretation across different contexts, FCD_GuacaMol values can be reverse-engineered using the formula $\text{FCD} = -5 \times \log(\text{FCD\_GuacaMol})$. This transformation allows the integration of GuacaMol's FCD interpretation with traditional FCD calculations.

4.1.2 Compared Methods

This subsection provides an quick overview of the models compared for the GuacaMol benchmark, highlighting their architectural details and contributions to the field of molecular generation.

- **GuacaMol:** GuacaMol (Brown et al., 2019) employs a Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997), which is implemented in Python and available as open-source software. The LSTM architecture enables effective learning of dependencies in sequence prediction tasks, making it suitable for generating valid molecular structures based on the SMILES notation.
- **MolGPT:** MolGPT (Bagal et al., 2022) utilizes a Transformer-Decoder architecture, specifically designed for the task of molecular generation The Transformer architecture leverages self-attention mechanisms to capture complex relationships within molecular data, offering significant improvements over traditional sequence-based models.
- **MolReactGen:** MolReactGen (Holzgruber, 2024) is built upon a GPT2 architecture and incorporates a specialized tokenizer known as "Char Wordpiece 176." This tokenizer is tailored to efficiently process molecular data, optimizing the model's ability to learn and generate molecular structures. The GPT2 architecture, known for its powerful generative capabilities, is adapted to the specific requirements of chemical compound generation, ensuring high fidelity and novelty in the generated outputs.

4.1.3 Performance Analysis

The evaluation of COATI within the GuacaMol benchmark framework reveals mixed results across several critical metrics used to assess the quality of generated molecules (shown in Table 4). Notably,

Table 4: **Performance Metrics on GuacaMol Benchmark:** Compares validity, uniqueness, novelty, and two versions of the FCD scores across four different molecular generation models. Every model generated 100,000 sample molecules. The green highlighted values denote as the best for each metric.

Metrics	**GuacaMol**	**MolGPT**	**MolReactGen**	**COATI**
Validity ↑	0.959	0.981	0.976	0.768
Uniqueness ↑	1.000	0.998	0.999	0.910
Novelty ↑	0.994	1.000	0.935	1.000
FCD ↓	0.455	0.488	0.219	3.873
$\text{FCD}_{\text{GuacaMol}}$ ↑	0.913	0.907	0.957	0.461

Table 5: **Detailed Evaluation of COATI Variants:** Performance of different COATI model configurations on validity and uniqueness metrics. The inverse temperature parameter is used for the GPT-2's top-k generation scheme.

Model	**Inv. Param**	Validity ↑	Uniqueness ↑
Autoreg Only	2	0.999	0.001
Autoreg Only	1.7	0.921	0.160
Grande Closed	2	0.994	0.828
COATI 2	2	0.883	0.999
COATI Guacamol	2	0.768	0.910

COATI exhibits a Validity score of 0.768, which is lower compared to other models such as MolGPT and MolReactGen.

Despite this, COATI excels in generating novel molecular structures, achieving a perfect Novelty score of 1.000. This indicates that all molecules generated by COATI during the benchmark were not present in the training data, highlighting its capacity to explore new areas of the chemical space, which is vital for discovering potential drug candidates with unique properties.

However, in terms of Uniqueness, where COATI scored 0.910, there is room for improvement. Although the score is high, it lags slightly behind other models, suggesting that while COATI is capable of generating novel molecules, there may be some redundancy in its outputs.

The FCD scores present a more complex picture. COATI's FCD score of 3.873 is significantly higher than those of other models, indicating a greater deviation from the distribution of molecules in the training dataset. In contrast, its $\text{FCD}_{\text{GuacaMol}}$ score is 0.461, the lowest among the compared models, reflecting a less favorable performance when the metric is adjusted to emphasize similarity to the training set's chemical space. This could suggest that while COATI is innovative in exploring new molecular configurations, its outputs may not always align with existing molecular data, which could impact its practical utility in drug discovery where relevance to known compounds is often crucial.

In summary, while COATI demonstrates exceptional capability in generating entirely novel structures, its practical application might be hindered by lower validity and a tendency to generate molecules that deviate from the typical chemical profiles encountered in pharmaceutical contexts.

4.1.4 Exploring Uniqueness in Molecule Generation

While the original paper presents validity scores across various COATI models, it does not address their uniqueness (shown in Table 5). This lack prompted an exploration into the uniqueness metrics. Unfortunately, due to the absence of training data, other metrics could not be assessed. Notably, the uniqueness scores for the Grande Closed model were unexpectedly low compared to other models like MolGPT or MolReactGen. In contrast, the COATI 2 model exhibited nearly perfect uniqueness scores but performed poorly on validity, suggesting potential limitations in its practical applicability for real-world molecular generation tasks. Moreover, the Autoregressive Only model showed remarkably low uniqueness, indicating significant challenges in generating diverse molecular structures.

4.2 Linear Probing

4.2.1 Metrics

Linear probing tasks in drug discovery rely on a set of metrics to evaluate model performance, including RMSE, Spearman correlation, AUROC, and Average Precision (AP).

- **RMSE (Root Mean Squared Error)** quantifies the average magnitude of prediction error, providing a clear measure of performance accuracy where lower values indicate better predictions.
- **Spearman Correlation** assesses how well the relationship between the predicted and true values can be described using a monotonic function, which is crucial for ranking predictions in drug efficacy.
- **AUROC (Area Under the Receiver Operating Characteristic Curve)** measures the ability of a model to distinguish between classes and is widely used in classification tasks. It measures the area under the false-positive rate against the true-positive rate.
- **AP (Average Precision)** approximates the area under the precision-recall curve, reflecting the precision of retrieving relevant instances throughout the range of possible recall levels.

The Relevance of ΔAP over AUROC: In the realm of drug discovery, the primary goal often involves identifying active compounds, which tend to be rare. This scenario is better captured by the Precision-Recall Curve (PRC) rather than the ROC curve, as PRC focuses more directly on the success of retrieval among rare positive cases. Average Precision (AP) is highly sensitive to the proportion of positive samples in the test set (base rate), which can distort performance assessments on tasks with severe class imbalance.

ΔAP addresses this by adjusting the AP value by subtracting the base rate of the task, effectively normalizing the metric across different datasets and ensuring that a random classifier scores zero. This adjustment makes ΔAP a more robust and fair measure, especially in scenarios where positive samples are a minority, as often seen in molecular screening for bioactive compounds.

4.2.2 Compared Methods

In the linear probing tasks, frozen embeddings from several models were compared to assess their efficacy in predicting molecular properties. The models which are evaluated include Autoregressive Only (Autoreg. Only), Barlow Closed, Grande Closed, COATI 2, (these four models are already described in section 2.4 and CLAMP.

CLAMP, introduced by (Seidl et al., 2023), is particularly notable for its innovative approach to aligning embeddings of SMILES strings with descriptions of scientific assays. This model is designed to predict molecule activity by embedding chemical data and natural language descriptions into a joint embedding space. The dual-encoder architecture of CLAMP utilizes descriptor-based fully-connected networks for the SMILES strings and a Latent Semantic Analysis (LSA) encoder for the assay descriptions. The model is trained using the InfoNCE loss, which enhances its ability to generalize to new bioassays that are described in human language. This capability is especially valuable for predicting activities in newly developed wet-lab procedures.

The embeddings from these models were rigorously compared to understand their performance across different datasets and to highlight which embeddings are most effective for specific types of molecular property predictions.

4.2.3 Performance Analysis

The validity of this performance evaluation is constrained due to the unavailability of the training datasets for all COATI models, leading to potential dataset leakage and likely overestimation of model performance. Despite these limitations, the evaluation was conducted to replicate the experimental setup described in the paper.

Regression Tasks: The COATI models, particularly Grande Closed and COATI 2, exhibit robust representations for regression tasks, suggesting effective learning of the underlying molecular features. However, in two specific scenarios, the CLAMP model surpasses the performance of both, which is shown in Table 6.

Table 6: **Regression Tasks:** Linear probing results of 4 COATI models and CLAMP with respect to the root-mean-squared-error (RMSE) or Spearman. Green highlighted values indicate the highest value for a dataset and yellow highlighted values indicates the second best result. Rank-avg shows the average rank over all tasks.

Dataset	**Metric**	**Autoreg Only**	**Barlow Closed**	**Grande Closed**	**CLAMP**	**COATI 2**
BACE Regression	RMSE ↓	0.78	0.72	0.69	0.82	0.70
Caco-2	RMSE ↓	0.71	0.62	0.58	0.66	0.54
Clearance Hepatocyte	Spearman ↑	0.14	0.32	0.34	0.27	0.37
Clearance Microsome	Spearman ↑	0.18	0.40	0.46	0.34	0.045
Delaney	RMSE ↓	0.83	0.74	0.64	0.64	0.60
Half Life	Spearman ↑	0.10	0.15	0.24	0.15	0.017
LD50	RMSE ↓	0.84	0.79	0.79	0.82	0.78
PPBR	RMSE ↓	18.12	17.17	16.55	15.40	15.74
Volume of Distribution	Spearman ↑	0.13	0.34	0.40	0.50	0.45
	rank-avg	4.9	3.45	1.98	3.1	1.58

Table 7: **Classification Tasks with AUROC metric:** Linear probing results of 4 COATI models and CLAMP with respect to AUROC. Green highlighted values indicate the best value for a dataset and yellow highlighted values are within the standard-deviation to the highest value. Because of a low number of re-runs the bootstrapped standard deviation are represented as superscript. Rank-avg shows the average rank over all tasks.

Dataset	**Autoreg Only**	**Barlow Closed**	**Grande Closed**	**CLAMP**	**COATI 2**
Ames Mutagenicity	$62.27^{\pm 0.8}$	$70.58^{\pm 0.8}$	$66.46^{\pm 0.8}$	$72.95^{\pm 0.7}$	$72.94^{\pm 0.7}$
BACE Classification	$66.61^{\pm 1.7}$	$72.91^{\pm 1.6}$	$71.67^{\pm 1.6}$	$68.61^{\pm 1.7}$	$74.83^{\pm 1.6}$
Bioavailability	$56.49^{\pm 3.2}$	$58.28^{\pm 3.2}$	$58.86^{\pm 3.3}$	$56.10^{\pm 3.2}$	$61.31^{\pm 3.1}$
ClinTox	$64.46^{\pm 3.4}$	$64.09^{\pm 3.4}$	$64.36^{\pm 3.5}$	$63.06^{\pm 3.3}$	$65.88^{\pm 3.3}$
CYP P450 2C9 Inhib.	$64.30^{\pm 0.7}$	$77.09^{\pm 0.6}$	$76.11^{\pm 0.6}$	$82.89^{\pm 0.5}$	$78.70^{\pm 0.6}$
CYP P450 3A4 Inhib.	$66.84^{\pm 0.6}$	$76.05^{\pm 0.5}$	$74.83^{\pm 0.6}$	$80.84^{\pm 0.5}$	$78.48^{\pm 0.5}$
DILI	$64.74^{\pm 3.1}$	$73.52^{\pm 2.8}$	$74.01^{\pm 2.8}$	$74.56^{\pm 2.8}$	$76.92^{\pm 2.7}$
hERG Blockers	$60.20^{\pm 3.0}$	$74.15^{\pm 2.7}$	$72.96^{\pm 2.7}$	$75.46^{\pm 2.6}$	$76.14^{\pm 2.6}$
hERG, Karim et al.	$57.80^{\pm 0.6}$	$68.61^{\pm 0.6}$	$67.37^{\pm 0.6}$	$69.41^{\pm 0.5}$	$72.04^{\pm 0.5}$
HIA	$71.89^{\pm 4.1}$	$82.33^{\pm 3.6}$	$80.40^{\pm 3.8}$	$84.40^{\pm 3.5}$	$85.21^{\pm 3.5}$
HIV	$60.57^{\pm 1.0}$	$68.75^{\pm 0.9}$	$66.99^{\pm 0.9}$	$82.40^{\pm 0.8}$	$70.32^{\pm 0.9}$
Pgp Inhibition	$69.76^{\pm 1.8}$	$81.15^{\pm 1.6}$	$82.32^{\pm 1.5}$	$82.17^{\pm 1.5}$	$80.09^{\pm 1.6}$
rank-avg	4.7	3.1	3.3	2.3	1.6

Classification Tasks: Across the results presented in Tables 7 and 8, the CLAMP model generally outperforms all other COATI models, except when including the newly developed COATI 2. This model, which benefits from double the training data and enhancements in model parameters and embedding dimensions, shows superior performance. Conversely, the Grande Closed and Barlow Closed models achieve comparable outcomes, while the Autoregressive Only model consistently underperforms across nearly all tasks.

Table 8: **Classification Tasks with ΔAP metric:** Linear probing results of 4 COATI models and CLAMP with respect to ΔAP. Green highlighted values indicate the best value for a dataset and yellow highlighted values are within the standard-deviation to the best value. Because of a low number of re-runs the bootstrapped standard deviation are represented as superscript. Rank-avg shows the average rank over all tasks.

Dataset	**Autoreg Only**	**Barlow Closed**	**Grande Closed**	**CLAMP**	**COATI 2**
Ames Mutagenicity	$07.74^{\pm0.6}$	$14.22^{\pm0.7}$	$10.92^{\pm0.6}$	$16.19^{\pm0.7}$	$15.95^{\pm0.7}$
BACE Classification	$11.49^{\pm1.5}$	$17.78^{\pm1.7}$	$16.27^{\pm1.5}$	$13.30^{\pm1.7}$	$19.78^{\pm1.8}$
Bioavailability	$02.46^{\pm1.3}$	$03.11^{\pm1.3}$	$03.39^{\pm1.3}$	$02.27^{\pm1.2}$	$04.32^{\pm1.3}$
ClinTox	$03.81^{\pm1.3}$	$05.81^{\pm2.2}$	$04.50^{\pm1.6}$	$05.07^{\pm2.1}$	$09.23^{\pm3.2}$
CYP P450 2C9 Inhib.	$09.02^{\pm0.5}$	$22.34^{\pm0.7}$	$21.07^{\pm0.7}$	$30.85^{\pm0.8}$	$24.79^{\pm0.7}$
CYP P450 3A4 Inhib.	$11.54^{\pm0.5}$	$20.74^{\pm0.6}$	$19.33^{\pm0.6}$	$26.80^{\pm0.7}$	$23.89^{\pm0.7}$
DILI	$09.70^{\pm2.5}$	$17.38^{\pm2.9}$	$17.75^{\pm2.9}$	$18.00^{\pm2.8}$	$20.54^{\pm3.0}$
hERG Blockers	$04.90^{\pm1.6}$	$12.71^{\pm1.8}$	$12.11^{\pm1.8}$	$13.66^{\pm1.9}$	$13.90^{\pm1.9}$
hERG, Karim et al.	$04.50^{\pm0.4}$	$12.81^{\pm0.5}$	$11.71^{\pm0.5}$	$13.52^{\pm0.5}$	$15.94^{\pm0.5}$
HIA	$05.03^{\pm1.2}$	$07.62^{\pm1.5}$	$07.11^{\pm1.4}$	$08.21^{\pm1.5}$	$08.38^{\pm1.6}$
HIV	$01.14^{\pm0.2}$	$02.94^{\pm0.3}$	$02.44^{\pm0.2}$	$13.36^{\pm0.7}$	$03.94^{\pm0.3}$
Pgp Inhibition	$13.64^{\pm1.6}$	$24.70^{\pm1.8}$	$25.80^{\pm1.8}$	$25.70^{\pm1.8}$	$23.40^{\pm1.9}$
rank-avg	4.9	2.9	3.4	2.1	1.6

5 Discussion and Critical Assessment

This section serves as the final address of a recurrent issue. Due to the absence of the training data, it is extremely challenging to thoroughly evaluate the model, resulting in an unavoidable data leakage problem. This limitation raises significant doubts about whether the experimental results can be validly replicated or if the training data was entirely excluded from testing in the tasks like linear probing, which is fundamental for authentic evaluation.

With regard to the learning efficacy, it appears that relying solely on the autoregressive component without the complementary contrastive learning involving the GNN leads to suboptimal representation learning. The paper initially posits that the Barlow Closed model exhibits superior performance across the models tested. However, the results demonstrate that only the newer COATI 2 model outperforms CLAMP, contradicting the claim that Barlow Closed is universally superior.

Additional challenges were faced with the ADMET datasets. Of the 29 intended for download from their AWS bucket, only 21 were successfully retrieved. Furthermore, the pretraining with the GuacaMol dataset highlighted significant vocabulary deficiencies. The vocabulary failed to recognise the punctuation mark '.' in canonical SMILES, which denotes disconnected molecular structures – a common occurrence in salts and other complex compounds. Furthermore, the absence of entries such as '*' (used to indicate points where different molecular fragments attach to each other) and 'Ac' (used as a shorthand for an acetyl group), as well as the inability to process SMILES longer than 349 characters, indicate deficiencies in the handling of diverse molecular representations.

Additionally, the structure of the training dataset was not readily apparent from the provided code. Following an email exchange with the authors, a bug in the dataset downloading script was identified and rectified by me. This enabled the loading of data with a similar data structure to the training data, which in turn facilitated the resolution of some column and type issues.

The results presented in Table 1 indicates that the GuacaMol model was trained on a significantly smaller number of tokens compared to other iterations of COATI models. This discrepancy suggests that extended training periods, potentially 10 to 20 times longer, could align its performance more closely with established GuacaMol baselines, particularly for the FCD score.

Finally, the potential for improved training results with the inclusion of an embedded graph token for data augmentation remains an area of interest. Unfortunately, due to constraints in computational resources, further experiments in this direction were not feasible.

6 Conclusion and Further Work

This work has critically evaluated the ability of the COATI model to generate and represent molecular structures, using extensive linear probing to measure performance on a variety of tasks. The lack of training data poses a significant challenge to the validation of the model, undermining confidence in the reproducibility of published results and highlighting the need for dataset accessibility for rigorous scientific validation. Despite these hurdles, the new COATI 2 model showed promising results, notably outperforming other models on several metrics. The training from scratch using the GuacaMol dataset also shows that if the model is trained 20 times longer, it could probably reach the benchmark results.

For the authors of the original paper, the development of a bigger model named COATI 2, featuring a chiral-aware GNN and an expanded vocabulary for conditional generation, represents a significant advancement. Additionally, experimenting with different tokenization schemes may further refine the model's performance.

For future work on my side and for my master's thesis, I am inspired by the challenges of conditional molecule generation also trough the work of COATI and the emerging role of diffusion models in this area. There are promising new avenues to explore, particularly in applying embeddings from models like COATI or CLAMP for conditional molecule generation. This approach could leverage the nuanced representation capabilities of these models to innovate in the design of novel compounds.

References

Benjamin Kaufman, Edward Williams, Carl Underkoffler, Ryan Pederson, Narbe Mardirossian, Ian Watson, and John Parkhill. COATI: multi-modal contrastive pre-training for representing and traversing chemical space, August 2023. URL https://chemrxiv.org/engage/chemrxiv/article-details/64e8137fdd1a73847f73f7aa.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988. URL https://api.semanticscholar.org/CorpusID:5445756.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100*Machine Learning: Science and Technology*, 1(4):045024, oct 2020. doi: 10.1088/2632-2153/aba947. URL https://dx.doi.org/10.1088/2632-2153/aba947.

Shion Honda, Shoi Shi, and Hiroki R. Ueda. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery. *CoRR*, abs/1911.04738, 2019. URL http://arxiv.org/abs/1911.04738.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9, 09 2017. doi: 10.1186/s13321-017-0235-x.

H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi: 10.1021/c160017a018. URL https://doi.org/10.1021/c160017a018.

Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alan Aspuru-Guzik, David A. Winkler, Dimitris Agrafiotis, Artem Cherkasov, and Alexander Tropsha. Qsar without borders. *Chem. Soc. Rev.*, 49:3525–3564, 2020. doi: 10.1039/D0CS00098A. URL http://dx.doi.org/10.1039/D0CS00098A.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. *CoRR*, abs/2110.04126, 2021. URL https://arxiv.org/abs/2110.04126.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing Activity Prediction Models in Drug Discovery with the Ability to Understand Human Language, June 2023. URL http://arxiv.org/abs/2303.03363. arXiv:2303.03363 [cs, q-bio, stat].

Stephan Holzgruber. *MolReactGen: Generating Molecules and Reaction Templates with a Transformer Decoder Model / Author Mag. Stephan Holzgruber*. 2024. URL http://epub.jku.at/obvulihs/9527348.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864, 2021. URL https://arxiv.org/abs/2104.09864.

Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, March 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00839. URL https://doi.org/10.1021/acs.jcim.8b00839. Publisher: American Chemical Society.

rdkit.Chem.AllChem module — The RDKit 2024.03.4 documentation. URL https://www.rdkit.org/new_docs/source/rdkit.Chem.AllChem.html.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, September 2018. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00234. URL https://doi.org/10.1021/acs.jcim.8b00234. Publisher: American Chemical Society.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, May 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00600. URL https://doi.org/10.1021/acs.jcim.1c00600. Publisher: American Chemical Society.