

Dataset Report

June 27, 2024

Dataset Overview

Datasets - Moses, Guacamol, Zinc 250k, Zinc 1M, Zinc 10M, Zinc 270M, ChemBL

- **Moses:** The Moses dataset is a curated collection from the ZINC database, focusing specifically on the ZINC Clean Leads collection. It comprises 4,591,276 molecules, each selected based on specific criteria: a molecular weight between 250 and 350 Daltons, no more than 7 rotatable bonds, and an XlogP value of 3.5 or less. The dataset excludes molecules with charged atoms or atoms other than C, N, S, O, F, Cl, Br, H. It also omits molecules with cycles longer than 8 atoms. Additionally, the selection process involved the application of medicinal chemistry filters (MCFs) and PAINS filters, ensuring the dataset's relevance for benchmarking in medicinal chemistry and drug discovery.
- **Guacamol:** The GuacaMol dataset is derived from the ChEMBL 24 database, known for its synthesized and biologically tested molecules. This dataset offers a more realistic representation of drug-like molecules compared to others like ZINC or QM9. The refining process includes removing salts, neutralizing charges, excluding molecules with overly long SMILES strings or less frequently occurring elements, and filtering based on similarity to a set of known drugs. The result is a dataset tailored for benchmarking in drug discovery, available for download with reproducible creation through a provided docker container.
- **Zinc Datasets(250k, 1M):** The ZINC database is a comprehensive collection of commercially available chemical compounds for virtual screening and drug discovery. It includes over 35 million compounds, with information on their structures, properties, and commercial availability.
- **ChemBL:** The ChEMBL database is a comprehensive resource for drug discovery, offering detailed bioactivity data, chemical structures, and target information for a wide range of drug-like compounds. It includes quantitative measurements such as IC50 and EC50, data on approved drugs, and is regularly updated. Widely accessible to researchers, ChEMBL is invaluable for medicinal chemistry and pharmacological research.

Dataset Statistics after Processing

The datasets were first cleaned to remove any duplicate entries within each dataset. Next, we standardized the SMILES (Simplified Molecular Input Line Entry System) strings and added a new column that shows the SELFIES (Self-referencing Embedded Strings) for each molecule. Lastly, we got rid of all the columns in the dataset that were not needed for our analysis.

Dataset	Number of Rows	File Size
ZINC_250k	249,455	18.18 MB
ZINC_1M	999,998	72.13 MB
MOSES	1,936,962	93.42 MB
GuacaMol	1,591,011	140.94 MB
ChEMBL	2,066,232	189.25 MB
ZINC_10M	9,999,971	722.37 MB
PubChem	110,993,226	4.20 GB
ZINC_270M	269,536,671	12.5 GB

Table 1: Datasets Information Sorted by File Size

Dataset Overlap with Moses-test set

Dataset	Overlap (Absolute)	Overlap (%)
ZINC_250k	1287	0.52%
ZINC_1M	184	0.02%
MOSES	0	0.00%
GuacaMol	6462	0.41%
ChEMBL	6591	0.32%
ZINC_10M	1673	0.02%
PubChem	116245	0.10%
ZINC_270M	46977	0.02%

Table 2: Dataset Overlap with Moses-test set

Inter-Dataset Overlap

Dataset Pair	Overlap (Absolute)	Overlap (%)
moses & guacamol	71,675	2.07%
moses & zinc_250k	13,907	0.64%
moses & zinc_1m	1,868	0.06%
moses & zinc_10m	18,579	0.16%
moses & chembl	73,316	1.87%
guacamol & zinc_250k	2,528	0.14%
guacamol & zinc_1m	204	0.01%
guacamol & zinc_10m	2,135	0.02%
guacamol & chembl	1,093,236	42.64%
zinc_250k & zinc_1m	82	0.01%
zinc_250k & zinc_10m	908	0.01%
zinc_250k & chembl	3,162	0.14%
zinc_1m & zinc_10m	966,612	9.63%
zinc_1m & chembl	285	0.01%
zinc_10m & chembl	2944	0.02%

Table 3: Overlap of SMILES Data between Various Datasets

Visual Representation of Inter-Dataset Overlap

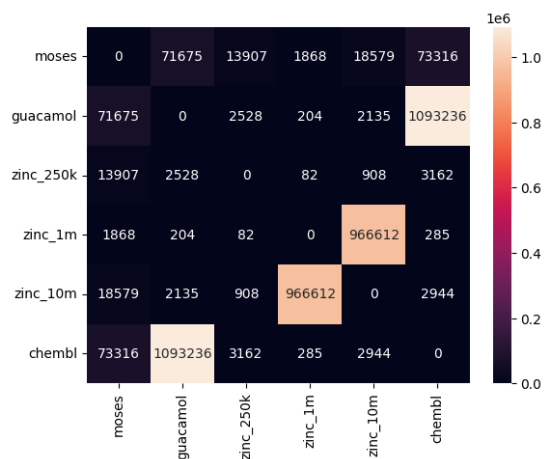


Figure 1: Heatmap showing the total overlap of SMILES strings between all datasets.

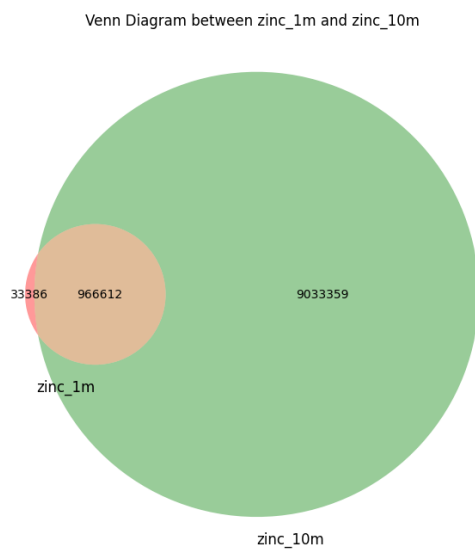


Figure 2: Venn diagram showing the absolute overlap of SMILES strings between Zinc 1M and Zinc 10M datasets.

Venn Diagram between chembl, guacamol, and mores

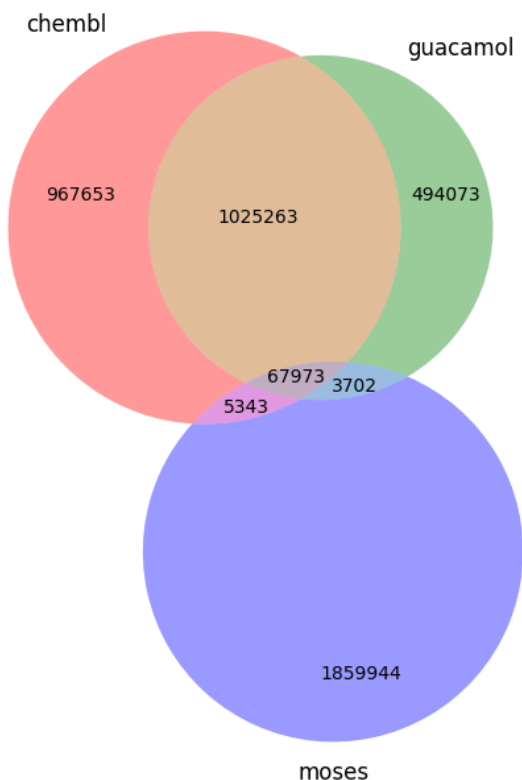


Figure 3: Venn diagram showing the absolute overlap of SMILES strings between Guacamol and ChemBL datasets.

Evaluation

In evaluating synthesis-aware generative models for molecular generation, it is imperative to employ comprehensive benchmarking methodologies that adequately reflect the complexities inherent in molecular discovery. While computational benchmarks, such as enrichment factors in virtual screening, standardized tests like Guacamol, and the MOSES (Molecular Sets) benchmark, have significantly propelled the field forward, they often fall short of capturing the entire scope of the discovery process, as highlighted by the implications of "Goodhart's law." These benchmarks serve as proxies and might not entirely encompass the nuances of molecular discovery. To address these limitations and enhance the validity of these models, the following multi-dimensional benchmarking strategies are recommended:

- **Objective Maximization and Ligand Rediscovery:** Assess the capacity of virtual screening or de novo design algorithms to identify molecules that optimize given objectives, along with their ability to rediscover known ligands. Notable benchmarks in this category include Guacamol, various virtual screening benchmarks, and the MOSES benchmark, which is specifically designed for assessing the quality of generative models.
- **Synthesis Prediction and Feasibility:** Employ CASP tools to predict synthetic routes or utilize synthesizability scores for generated molecules to ensure practical feasibility in a laboratory setting.
- **Molecule Quality Assurance:** Implement quality filters akin to those used in Guacamol and MOSES to ascertain the reasonableness of the generated molecules. It is crucial to include visualizations of random, non-cherry-picked molecular samples in machine learning publications to provide a transparent and accurate representation of the model’s output.
- **Evaluation of Synthesis Planning Algorithms:** Conduct both quantitative and qualitative assessments of synthesis planning algorithms to ensure they are efficient, practical, and innovative.

The discussion further delves into the translational impact of these improvements, questioning the real-world applicability of marginal gains observed in computational benchmarks. Given the often sparse and diverse nature of data in drug discovery, along with the occurrence of distribution shifts, the need for robustness in models is paramount. The authors advocate for a balanced approach towards benchmarking, one that encourages ongoing refinement and innovation in benchmarking practices without necessarily mandating experimental validation due to the varying resource capabilities of computational groups.

Recent prospective validations of virtual screening and de novo design offer promising examples of the field’s progress. These include the application of large enumerated on-demand libraries in virtual screening and the integration of synthesis planning with computational algorithms, showcasing their utility particularly in the early stages of discovery. However, the visibility and publication of such innovations are often delayed in the industrial context due to proprietary concerns or lack of incentives, highlighting an additional layer of complexity in benchmarking and validating these models.

In conclusion, while current benchmarks such as Guacamol, virtual screening benchmarks, and the MOSES benchmark are instrumental in driving advancements in molecular generation models, there is a clear and ongoing need for developing more nuanced, robust, and comprehensive methodologies. These should not only reflect the theoretical and computational excellence but also align closely with practical, real-world utility in the ever-evolving landscape of drug discovery.

Evaluation Metrics

MOSES Evaluation Metrics

- Validity (\uparrow): Measures the percentage of generated molecules that are chemically valid.
- Uniqueness (\uparrow): Assesses whether the model generates diverse molecules by calculating the proportion of unique molecules in the generated set.
- Novelty (\uparrow): Evaluates the model’s ability to generate molecules that are not present in the training set, indicating the model’s creativity.
- Internal Diversity (\uparrow): Quantifies the chemical diversity within the generated set of molecules.
- External Diversity (\uparrow): Compares the diversity of the generated set to the diversity of an external set, often the test set.
- Fréchet ChemNet Distance (FCD) (\downarrow): Uses a deep neural network to capture chemical and biological properties of compounds and measures the distance between the generated and real molecules in this learned feature space.
- Fragment and Scaffold Similarity (\uparrow): Measures how closely the distribution of molecular fragments and scaffolds in the generated set matches that of the reference set.
- Similarity to Nearest Neighbor (SNN) (\uparrow): Calculates the average Tanimoto similarity between the generated molecules and their closest counterparts in the reference dataset.
- Properties Distribution (\downarrow): Compares the distribution of certain molecular properties (like molecular weight, logP, etc.) between the generated and reference sets using Wasserstein-1 distance.

PyTDC Evaluation Metrics

- Diversity (\uparrow): The diversity of a set of molecules is defined as the average pairwise Tanimoto distance between the Morgan fingerprints.
- KL divergence (\downarrow): KL divergence between the probability distributions of a variety of physicochemical descriptors for the training set and a set of generated molecules. Models able to capture the distributions of molecules in the training set will lead to small KL divergence values. To increase diversity, we want high KL.
- Frechet ChemNet Distance (FCD) (\downarrow): FCD first takes the means and covariances of the activations of the penultimate layer of ChemNet are calculated for the reference set and for the set of generated molecules. The

FCD is then calculated as the Frechet distance for both pairs of values. Similar molecule distributions are characterized by low FCD values.

- Novelty (\uparrow): Novelty is the fraction of the generated molecules that are not present in the training set.
- Validity (\uparrow): Validity is calculated using RDKit’s molecular structure parser that checks atoms’ valency and consistency of bonds in aromatic rings.

Molecule Generation Oracles

- The goal of molecule generation is to create novel molecules with desired properties, evaluated by oracles based on a machine learning task that learns from a large dataset.
- Oracles serve as user-defined scoring functions to measure chemical properties.
- These tools allow for controlled generation, steering the exploration of chemical space based on user-specified oracle guidance.
- This approach facilitates the discovery of novel candidates by leveraging the internal representation learned from data.

Physicochemical Properties

- QED (Quantitative Estimate of Drug-likeness): Measures how ”drug-like” a compound is based on its physicochemical properties.
- LogP: Measures the octanol-water partition coefficient, related to compound solubility and permeability.

Synthetic Accessibility

- SAScore: Evaluates the ease of synthesizing drug-like molecules in virtual screening. Ranges from 1 (easy to synthesize) to 10 (hard to synthesize). This score reflects the presence of common fragments in a molecule and structural complexities. .
- SCScore: The SCScore model rates the synthetic complexity of molecules on a scale from 1 to 5. Based on the premise that on average, the products of published chemical reactions should be more synthetically complex than their corresponding reactants

- SYBA: It is a fragment-based method for the rapid classification of organic compounds as easy- (ES) or hard-to-synthesize (HS). Based on a Bernoulli naïve Bayes classifier that is used to assign SYBA score contributions to individual fragments based on their frequencies in the database of ES and HS molecules. Trained on ES molecules available in the ZINC15 database and on HS molecules generated by the Nonpher methodology
- RAscore: RAscore is derived using machine learning models that are specifically trained on data generated by an AI-driven retrosynthetic planning tool called AiZynthFinder. This training involves predicting whether a synthetic route can be found for a molecule.
- MPO (Multi-Property Objective): Evaluates compounds against multiple criteria relevant to drug discovery - contains seven drugs (Osimertinib, Fexofenadine, Ranolazine, Perindopril, Amlodipine, Sitagliptin, Zaleplon) where each has various objectives.

Step by Step Guide

1. Install the Hugging Face `evaluate` library using `pip`:

```
pip install evaluate
```

2. Generate lists `ls_gen` and `ls_train`, which represent the generated list of SMILES and the SMILES training dataset, respectively.
3. Load the `MolGen/MolEvalMetrics` from Hugging Face using:

```
MolEvalMetrics = evaluate.load("MolGen/MolEvalMetrics")
```

4. Get the evaluation metrics for the generated set of SMILES by calling the `compute` method:

```
print(MolEvalMetrics.compute(gensmi = ls_gen, trainsmi = ls_train))
```

References

- <https://arxiv.org/pdf/1811.12823v5.pdf>
- <https://pubs.acs.org/doi/epdf/10.1021/ci3001277>
- <https://pubs.acs.org/doi/epdf/10.1021/acs.jcim.8b00839>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245175/>