MPEGO: A toolkit for multi-level performance evaluation of generative models for material discovery domains

Girmaw Abebe Tadesse*
IBM Research - Africa
girmaw.abebe.tadesse@ibm.com

Jannis Born
IBM Research - Zurich
jab@zurich.ibm.com

Celia Cintas
IBM Research - Africa
celia.cintas@ibm.com

Matteo Manica
IBM Research - Zurich
tte@zurich.ibm.com

Komminist Weldemariam
IBM Research - Yorktown Heights
kommy@ibm.com

ABSTRACT

The ability to generate candidate molecules with certain chemical properties is key in novel material discovery. The generation capability has been improved using larger training data, sophisticated generative models, and sampling techniques. However, the evaluation of generative models for material discovery and the characterization of the generated candidate molecules is premature otherwise largely relies on costly and error-prone expert involvement. Such evaluations help to improve understanding of the generative process, differentiate across models, and facilitate interaction between machine learning researchers and materials scientists. To this end, we propose a toolkit for Multi-level Performance Evaluation of Generative mOdels (MPEGO) for material discovery applications. MPEGO aims to hierarchically characterize and quantify the capability of generative models across the chemical and biological properties of molecules. The toolkit is validated with two generative models: Graph Convolutional Policy Network (GCPN) and a Flow-based Autoregressive (GraphAF) trained on ZINC-250K molecules. Preliminary results show that the GCPN generated molecules achieve higher independence from the training molecules compared to GraphAF's, across multi-level evaluation metrics, whereas GraphAF molecules are found to achieve higher independence in scaffolding and molecular weight features. Finally, as MPEGO is model-agnostic, it can be integrated with any generative models for material discovery and beyond.

KEYWORDS

generative models, material discovery, performance evaluation

ACM Reference Format:
Girmaw Abebe Tadesse, Jannis Born, Celia Cintas, Matteo Manica, and Komminist Weldemariam . 2022. MPEGO: A toolkit for multi-level performance evaluation of generative models for material discovery domains. In *Proceedings of 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD Workshop).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding Author

KDD Workshop, 2022, Washington DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Machine learning methods, particularly generative models, have shown to provide a promising potential for generating and optimizing novel molecules across material domains (e.g., drug, polymer, etc.); by combining data-driven techniques and domain knowledge to efficiently search the space of all plausible molecules and generate new and valid ones [1, 10, 20, 24]. Traditional high-throughput wet-lab experiments, physics-based simulations, and bio-informatics tools for the molecular design process heavily depend on human expertise. These processes require significant resource expenditure to propose, synthesize and test new molecules, thereby limiting the exploration space [6, 12, 15].

Generative models have been applied to facilitate the material discovery process by employing inverse molecular design problem. This approach transforms the conventional and slow discovery process by mapping the desired set of properties to a set of structures. The generative process is then optimized to encourage the generation of molecules with those selected properties. Multiple approaches have suggested the use of latent representation learning coupled with different sampling techniques to achieve efficient discovery. These examples range from Variational Autoencoders (VAE) with different sampling techniques [2, 4, 7, 9] to Generative Adversarial Networks (GANs) [18].

The generation capability has been improved recently using larger training data and sophisticated generative methodologies, the evaluation of generative models in the natural sciences remains a grand challenge [5]. Some of the reasons include the multi-objective nature of real discovery problems, the intricacy of evaluating relevant properties *in-silico* and the lack of widely accepted, model- and property-agnostic success metrics for generative models. Quantifying a specific subset of the evaluation metrics, e.g., using distance measures between distributions of generated samples have been discussed in prior work such as [17, 22]. These approaches are limited to understand the generation performance at sub-feature level, e.g., *in a specific range of a molecular weight.*

In this paper, we introduce a Multi-level Performance Evaluation of Generative mOdels (MPEGO) toolkit (see Fig. 1), which aims to hierarchically characterize and quantify the generation capability of models across the chemical and biological properties of molecules. To that end, MPEGO is a model- and property-agnostic toolkit and

its core design is derived from two main requirements: representative *examples* (of training and generated molecules) and one or multiple *properties* (extracted from these molecules). Metrics derived from MPEGO are also interpretable and provides multi-level abstractions of the generation process. Specifically, the contributions of this paper are as follows:

(1) We provide low-level evaluation of generative models using a novel Sub-feature-level Independence Score (SIS) between the generated and training molecules.
(2) We derive high-level evaluation of generative models using a principled and hierarchical aggregation of SIS values.
(3) We employ multi-dimensional subset scanning (MDSS) [13] to automatically identify and characterize generation bias, i.e., the types of molecules generated by each model with extreme frequency.
(4) We validate MPEGO toolkit on two state-of-the-art generative models: GCPN [25] and GraphAF [19] trained with molecules from ZINC-250K [11].

The organization of the paper is as follows. Section 2 summarizes the related work on evaluation of generative models for molecular discover. Section 3 formulates the problem concisely and presents the core elements of the proposed MPEGO toolkit. Section 4 details the experimental setup, including the datasets, feature extracted and generative models used for validation. Section 5 presents and discusses the results obtained from comparing the generation performance of generative models using the MPEGO toolkit. Section 6 concludes the paper and outlines future work.

2 RELATED WORK

The state-of-the-art evaluation approaches for generative models (and their generated molecules) aim to quantify pre-determined requirements, such as diversity and validity, using a variety of metrics. Frechet ChemNet Distance (FCD) [17] is one of such metrics, and it measures the distance between hidden representations drawn from sets of generated and training molecules using the ChemNet architecture, which is limited in disclosing low-level (sub-feature-level) performance.

Benchmark platforms were also proposed to evaluate the generation process across a variety of metrics and generation tasks. Examples of such benchmarks include GuacaMol [3] and Molecular Sets (MOSES) [16]. GuacaMol [3] is one of the early benchmark platforms for new molecule discovery, which aims to evaluate a generative models across different tasks. These tasks include the fidelity of the models to reproduce the property distribution of the training sets, the ability to generate novel molecules, the exploration and exploitation of chemical space, and a variety of single and multi-objective optimization tasks. Molecular Sets (MOSES) [16] is another benchmarking framework to evaluate the distribution learning of generative models. To this end, MOSES provides training and testing datasets, and a set of metrics to evaluate the quality and diversity of generated structures to standardize training and model comparisons.

However, automated characterization of generated molecules, and model-agnostic, quantitative and multi-level evaluation of the generative models still remains challenging. These challenges could be summarized as follows. First, multiple evaluation metrics are model-dependent. For example, Frechet ChemNet Distance (FCD) [17] depends on latent representation in a neural network, and Maximum-mean discrepancy (MMD) [8] is more specifically used to evaluate graph-based generative models. State-of-the-art metrics also suffer from limited generalizability (across different level of abstractions) and interpretability, e.g., by domain experts. In addition, existing evaluation metrics are susceptible to potential flaws in predictive models used in goal-oriented or constrained generation. Moreover, existing evaluation strategies lack a generic and standalone evaluation metric that combines both distributional metrics (e.g., uniqueness and diversity) and property-based metrics that score single property (e.g., aromatic). The dependency on single-constraint-objective lacks a principled approach to incorporate multiple chemical target properties (e.g., molecular weight), structural details (e.g., scaffold) and, synthetic metrics, such as Qualitative Estimate of Drug-likeness (QED). This becomes a significant challenge when a single and inaccurate evaluation metric is used, which oversimplifies real discovery problems and hence less practical.

The proposed MPEGO toolkit aims to provide an effective and multi-level characterization of generated molecules (compared to the training) and evaluation of their generative models. The multi-level evaluation of MPEGO toolkit starts from sub-feature-based performance evaluation (at the bottom), which provides low-level evaluation of generative models. The sub-feature-based performance scores are later aggregated hierarchically and across multiple-properties to provide high-level evaluations and characterizations.

3 THE MPEGO TOOLKIT

In this section, we present the details of the proposed MPEGO toolkit (see Fig. 1) that aims to evaluate the performance of generative models for the material discovery and to provide insights that are interpretable to improving interactions between machine learning researchers and experts in material science. We, first, formulate the critical research questions MPEGO toolkit is designed to address, followed by the details on MPEGO's core components: *Feature extraction and preprocessing* and *Multi-level performance evaluation*.

3.1 Problem statement

Let $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_k, \cdots, \mathcal{G}_K$ are datasets of molecules that are generated from K black-box generative models $(\Theta_1, \Theta_2, \cdots, \Theta_k, \cdots, \Theta_K)$ trained on a dataset of existing molecules, $\mathcal{T}$. Can we evaluate the generation capability of each black box model in a scalable, easily intepretable and multi-objective manner? Specifically, we aim to address two specific questions.

Q1: Given a set of different chemical and biological features, $\mathcal{F} = f_1, f_2, \cdots, f_m, \cdots, f_M$, extracted from each molecule, how do we quantify the generation capability of each model conditioned on one or more of these features, i.e., at different levels of abstractions?
Q2: What are the characteristics of molecules being generated with extreme frequencies (least or most) by each of the generative models, i.e., generation bias?

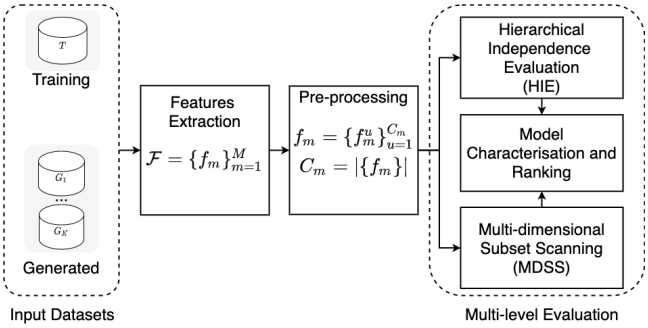


Figure 1: Overview of the MPEGO toolkit for material discovery.

To address Q1, we propose Hierarchical Independence Evaluation (HIE) that aims to quantify the independence of generative models compared to the training molecules. The hierarchy begins with a sub-feature level evaluation at the bottom to the aggregation of performance across all the features at the top. To address Q2, we employ multi-dimensional subset scanning (MDSS) [13] that aims to automatically identify and characterize generation bias, i.e., identifying molecules being generated by each model with extreme frequency.

3.2 Feature extraction and pre-processing

Given representative examples of generated and training molecules, MPEGO starts with the extraction of multiple features that encode different chemical and biological properties of these molecules (see Fig. 1). Examples of features include domain-expert driven characteristics (e.g., Lipinski criteria), Molecular Weight, structural attributes (Scaffold, Ring, Morgan bits), and synthetic metrics (e.g., QED, ESOL, and LogP). The detailed list of these features, along with their description, is shown in Table 1 in Section 4.

The type of feature values could be binary (e.g., Aromatic where *Aromatic:1* represents the presence of an aromatic ring, and *Aromatic:0* represent the lack of aromatic ring) or continuous (e.g., Molecular Weight). Thus, a preprocessing step is employed where the continuous values are discretized into a number of bins that provide low- or sub-feature-level evaluation of models. MDSS also requires the discretization of feature values to allow the exploration across different levels of combinations of their values.

3.3 Multi-level evaluation of generated molecules and their models

Figure 2 provides the details of MPEGO's multi-level evaluation pipeline to quantify the performance of generative models, and its core components are *Hierarchical Independence Evaluation (HIE)* and *Multi-dimensional Subset Scanning (MDSS)*, which are tailored to evaluate the generative models from different perspectives (formulated as Q1 and Q2 above).

3.3.1 Hierarchical Independence Evaluation (HIE). HIE follows a bottom-up approach (from a sub-feature to a global aggregation levels) in order to address Q1, i.e., multi-level performance evaluation of generative models for material discovery across different properties encoded as features. HIE is based on a sub-feature-level normalized objective measure (SNOM) between the sets of training and generated molecules. While odds ratio is a commonly used objective measure, its values are unbounded $(0, \infty)$ and hence it does not directly satisfy the first key properties of a good association measure outlined by [14] as *if two variables are independent, their association needs to be zero*. To this end, our SIS computation employs Yule's Y coefficients [26] that bounds the association measure values in $[-1, 1]$. Yule's Y measure is selected as it satisfies all the key properties of objective measures in addition to anti-symmetry under row or column permutation[21]. Below are the steps to utilize SNOM and compute the hierarchical performance metrics as a form of independence scores between the generated and training sets of molecules.

Let $f_m \in \mathcal{F}$ is a feature with C_m uniqe values or ranges, i.e., $f_m = \{f_m^u\}, u \in [1, 2, \cdots, C_m]$. Note that $C_m = 2$ for binary features, and C_m is the number of unique values for a categorical features or the number of bins after discretization of continuous features. SNOM computation requires the stratification of both the generated $(\mathcal{G}_k)$ and training $(\mathcal{T})$ datasets per each unique value/range f_m^u resulting $\mathcal{G}_{km}^u$ and $\mathcal{T}_m^u$, respectively. The remaining subsets are $\widetilde{\mathcal{G}_{km}^u}$ and $\widetilde{\mathcal{T}_m^u}$, respectively, which are the subsets of molecules that are not characterized by f_m^u, i.e., $\widetilde{\mathcal{G}_{km}^u} = \mathcal{G}_{km}^u | (f_m \neq f_m^u) = \mathcal{G}_k - \mathcal{G}_{km}^u$ and $\widetilde{\mathcal{T}_m^u} = \mathcal{T}_m^u | (f_m \neq f_m^u) = \mathcal{T} - \mathcal{T}_m^u$. Accordingly, a 2×2 pivot table is generated for each f_m^u as:

	$(f_m = f_m^u)$	$(f_m \neq f_m^u)$
$\mathcal{G}_{km}$	α	β
$\mathcal{T}_m$	δ	γ

where α is the number of generated molecules in $\mathcal{G}_{km}^u$ that are characterized by the feature value $f_m = f_m^u$, β is the number of generated molecules in $\widetilde{\mathcal{G}_{km}^u}$ with $f_m \neq f_m^u$. Similarly, δ and γ are the numbers of training molecules that satisfy $f_m = f_m^u$ and not in $\mathcal{T}_m^u$ and $\widetilde{\mathcal{T}_m^u}$, respectively. Note that $\alpha + \beta$ is total number of generated molecules, i.e., $|\mathcal{G}_k|$. Similarly, $\delta + \gamma$ is the number of training molecules, i.e., $|\mathcal{T}|$. Then SNOM per f_{km}^u is computed as a form of Yule's Y coefficient from the pivot table as $o_{km}^u \in [-1, 1]$:

$$o_{km}^u = \frac{\sqrt{P(\mathcal{G}_{km}^u)P(\widetilde{\mathcal{T}_m^u})} - \sqrt{P(\widetilde{\mathcal{G}_{km}^u})P(\mathcal{T}_m^u)}}{\sqrt{P(\mathcal{G}_{km}^u)P(\widetilde{\mathcal{T}_m^u})} + \sqrt{P(\widetilde{\mathcal{G}_{km}^u})P(\mathcal{T}_m^u)}} \tag{1}$$

$$o_{km}^u = \frac{\sqrt{\alpha\gamma} - \sqrt{\beta\delta}}{\sqrt{\alpha\gamma} + \sqrt{\beta\delta}} \tag{2}$$

Sub-feature-based Independence Score (SIS) is then computed from SNOM value as $I_{km}^u = 1 - |o_{km}^u|$, where $I_{km}^u \in [0, 1]$ and higher I_{km}^u reflects higher independence between the sets of generated and training molecules, i.e., given a sub-feature f_m^u, it will be difficult to infer whether a molecule is from a generated or training set. While low-level evaluation of generative models is provided via SIS values, higher level independence evaluation scores are then obtained from a principled aggregation of SIS values. Feature-level Independence Score (FIS) is computed as $I_{km} = \sum_{u=1}^{C_m} \lambda_{km}^u I_{km}^u$ via a weighted aggregation of SIS values, where $\sum_{u=1}^{C_m} \lambda_{km}^u = 1$ and each λ_{km}^u weights the SIS of its corresponding unique value f_m^u. In a

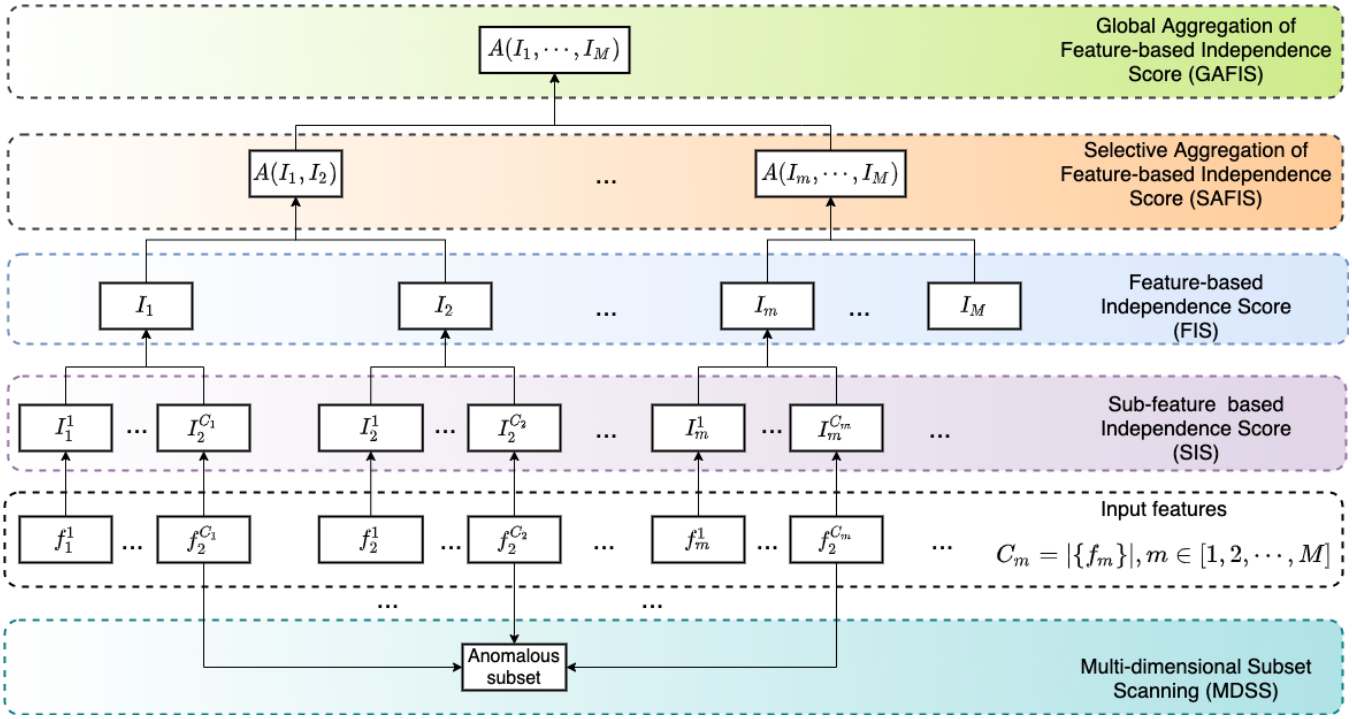


Figure 2: Details of Multi-level evaluation component of the MPEGO toolkit that comprises Hierarchical Independence Evaluation (i.e., SIS, FIS, SAFIS and GAFIS) and Multi-dimensional Subset Scanning (MDSS). $A(\cdots)$ represents aggregation operation, and anomalous subset refers to the logical combinations of features that characterize molecules generated with extreme frequencies.

simple average aggregation, $\lambda_{km}^u = 1/C_m$. Similarly, FIS values can be aggregated for even higher-level evaluation of the models. Selective Aggregation of Feature-level Independence Score (SAFIS) and Global Aggregation of Feature-level Independence Score (GAFIS) are computed as $\hat{I} = \sum_{r=1}^{R} \eta_r I_{kr}$, where $\sum_{r=1}^{R} \eta_r = 1$, and $R < M$ for SAFIS and $R = M$ for GAFIS computation, respectively.

3.3.2 Multi-dimensional Subset Scanning (MDSS). Generative models trained on similar set of molecules will hardly generate molecules with exact characteristics, partly due to the variations in model architecture, hyper-parameter setting and training methodologies. Thus, there is potential generation bias in each model. To this end, we employ MDSS to identify the characteristics of molecules being generated with extreme frequencies (least or most) by each of the generative models. Such approach requires unconstrained interactions of different sub-feature values, e.g., Aromatic:1 and Molecular Weight > 500 Daltons.

MDSS addresses the question by transforming the exploration of divergent subset of generated molecules across the interaction of multiple features into a search problem. MDSS utilizes additive linear time subset scanning (ALTSS) property [13] to efficiently search across potentially exponential combination of different features values, i.e., given M features, the search space could be as large as $2^M - 1$. Thus, the goal of MDSS is to automatically identify the subset (also known as the anomalous subset) of molecules $\mathcal{S}$ that are divergent compared to our expectation.

Specifically, to identify a subset of molecules being generated by a generative model M_k, we first merge the corresponding $\mathcal{G}_k$ and $\mathcal{T}$ datasets as $\mathcal{D} = \mathcal{G}_k \cup \mathcal{T}$, and an outcome label ($y$) is generated for each molecule, i.e., $y = 1$ for molecules in $\mathcal{G}_k$ and $y = 0$ for molecules in $\mathcal{T}$. if there are $N_g = |\mathcal{G}_k|$ generated and $N_t = |\mathcal{T}|$ training molecules in $\mathcal{D}$, the expectation of generated molecules in $\mathcal{D}$ is $e_g = \frac{N_g}{N_g + N_t}$. Thus, scanning for a subset of more frequently generated molecules aims to identify molecules with extreme deviation in their observation rates compared to e_g.

The deviation between the expectation and observation is evaluated by maximizing a Bernoulli likelihood ratio scoring statistic $\Gamma(\cdot)$ for a binary outcome, i.e., $y \in [0, 1]$. The null hypothesis assumes that the odds of the generated molecule in any subgroup $\mathcal{S}$ is similar to the expected, i.e., $H_0 : odds(\mathcal{S}) = \frac{e_g}{1-e_g}$; while the alternative hypothesis assumes a constant multiplicative increase in the odds of the generated molecules in the anomalous or extremely divergent subgroup $\mathcal{S}^a$, i.e., $H_1 : odds(\mathcal{S}^a) = q\frac{e_g}{1-e_g}$ where $q \neq 1$ ($q > 1$ for a subset of molecules generated with extremely high frequency (over-generated), and $0 < q < 1$ for a subset of molecules generated with extremely low frequency (under-generated), compared with the training set of molecules. The anomalous scoring function for a subgroup ($\mathcal{S}$) with reference $\mathcal{D}$ is formulated as, $\Gamma(\mathcal{S}, \mathcal{D})$ and computed as:

$$\Gamma(\mathcal{S}, \mathcal{D}) = \max_q log(q) \sum_{i \in S} y_i - N_s * log(1 - e_g + qe_g), \quad (3)$$

Table 1: Features extracted per each molecule (existing and generated) that encode physical, biological and chemical properties of these molecules. The features are then used to analyze the association measures with the generated and existing molecules via normalized odds ratio. In addition, the divergent characteristics of these molecules is identified and characterized using these interpretable features.

Feature	Description
QED	Qualitative Estimate of Drug-likeness
ESOL	Estimated SOLubility
SCScore	Synthetic Complexity Score
SAS	Synthetic Accessibility Score.
Scaffold	A molecule is identical to its scaffold
Ring	A molecule contains a ring structure
LargeRing	A molecule contains a ring structure with more than 6 atoms
Aromatic	A molecule contains an aromatic ring
SteroCharacter	SMILES string contains a stereochemistry string.
Stereocenter	Whether a molecule contains a stereocenter,
Heterocycle	A molecule has at least one ring with at least two different atoms
Lipinski	A molecule adheres to the Lipinski's rule of five.
HBondDonor	A molecule has not more than 5 hydrogen bond donors
HBondAcceptor	A molecule has not more than 10 hydrogen bond acceptors
MolecularWeight	The molecular mass in Daltons
LogP	Logarithmic partition coefficient

Table 2: Summary of hyper-parameters set-up to train the two graph-based models: GCPN and GraphAF using ZINC-250K dataset.

Setting	Models	
	GCPN	GraphAF
Input Dimension	18	9
Number of relation	3	3
Batch normalization	False	True
Atom types	[6-9, 15-17, 35, 53]	[6-9, 15-17, 35, 53]
Hidden dimensions	[256, 256, 256, 256]	[256, 256, 256]

where N_s is the number of molecules in $\mathcal{S}$. The anomalous subset, $\mathcal{S}^a$, identification is iterated until convergence to a local maximum is found, and the global maximum is subsequently optimized using multiple random restarts.Thus, $\mathcal{S}^a$ is a subset of molecules with the largest $\Gamma(\mathcal{S}, \mathcal{D})$, which encodes both the divergence and size of the subset compared to the expectation.

4 EXPERIMENTAL SETUP

We utilize the publicly available ZINC-250K[1] dataset to train the two graph-based generative models selected for MPEGO's validation: GCPN [25] and GraphAF [19]. ZINC-250K contains 249, 455 small molecules in Simplified Molecular-Input Line-Entry System (SMILES) representation. Details on ZINC tool is available in [11]. We also generated 10, 000 small molecules (in SMILES) from each of trained GCPN and GraphAF models. We utilize PyTorch implementation of the two graph generated models with experimental set-ups shown in Table 2.

[1]https://www.kaggle.com/datasets/basu369victor/zinc250k

In addition to the SMILES representation, a few molecular properties are also provided for each molecule in ZINC-250K. These properties include logP (water–octanal partition coefficient), SAS (synthetic accessibility score) and QED (Qualitative Estimate of Drug-likeness). We later extracted more features from both training and generated molecules using RDKit Chemoinformatics Software[2] that helps to automatically extract domain-specific chemical and biological properties as features. A total of a total of 16 features are extracted (as shown in Table 1). We excluded the *SteroCharacter* feature from our analysis since both GCPN and GraphAF were unable to encode the sterochemistry due to the inductive bias nature of the graph-models. For continuous features, we employ quintile based discretization that segments the values into five bins.

We employ the Hierarchical Independence Evaluation (consisting of SIS, FIS, SAFIS and GAFIS) to quantify the performance of generative models with respect to the training molecules. We also utilize the histogram of features to provide qualitative comparison of the distributions from generated and training molecules in reference to the quantitative sub-feature-based normalized objective measure (SNOM). SNOM directly reveals over- or under-generation of molecules at different sub-feature levels (ranges). The characterization of the identified anomalous subgroup ($\mathcal{S}^a$) of MDSS includes providing the logical combination of anomalous feature values, the size of the subgroup N_s, the odds ratio between $\mathcal{S}^a$ and $\widetilde{\mathcal{S}^a} = \mathcal{D} - \mathcal{S}^a$ and its 95% Confidence Interval (CI) and empirical p value.

5 RESULTS AND DISCUSSION

In this Section, we present and discuss results obtained from hierarchical performance of generative models in the form of independence scores (SIS, FIS, SAFIS and GAFIS) and automated detection and characterization of generation bias (obtained via MDSS).

5.1 Hierarchical Performance Evaluation of Generative Models

Table 3 provides detailed analysis of hierarchical quantification of the generation performance of GCPN and GraphAF models using the training (ZINC-250K) dataset as a baseline. To demonstrate the step-by-step evaluation of these models , two continuous features were selected as examples: QED and Molecular Weight. These features were discretized into five bins in the pre-processing step. Sub-feature-level normalized objective measure (SNOM) is derived for each bin via Yule's Y Coefficient as $o^u_{km} \in [-1, 1]$, and the $-ve$ sign suggests the direction of dependency (i.e., negative correlation) and its magnitude reflects the level of dependency. Note that ideal independence translates to zero valued SNOM values[21]. We derive SIS that directly quantifies the sub-feature-based independence score ($I^u_{km} \in [0, 1]$), and the higher SIS value reflects higher Independence. The results show that GCPN generated molecules achieve higher independence than GraphAF's molecules compared to the training ZINC-250K, across multiple QED bins (I, II and V). On the other hand, GraphAF generated molecules demonstrates higher independence across the majority of Molecular Weight bins (I, VI and V). These findings could be qualitatively understood from the histogram of the distributions of the two features shown in Fig 3,

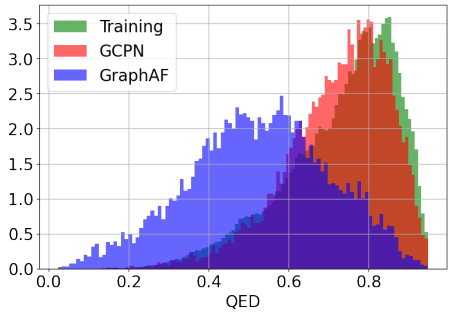
[2]https://www.rdkit.org/

Table 3: Multi-level evaluation metrics to compare the performance of the two graph-based generative models: GCPN and GraphAF that are trained on ZINC-250K dataset. QED and Molecular Weight features are selected as examples. First, each feature is discretized into five bins (I-V) and range of values for each bin is shown. The multi-level metrics are SNOM, SIS and FIS. SNOM provides the sub-feature based normalized odds ratio computing using Yule's Y coefficients; SIS: sub-feature-level independence score derived for each bin; and FIS: feature-level independence score that aggregates the SIS values of all the bins in a given feature.

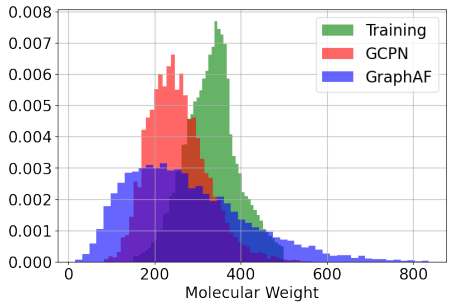
			Multi-level Evaluation Metrics					
			SNOM		SIS		FIS	
Feature	Bins	Range	GCPN	GraphAF	GCPN	GraphAF	GCPN	GraphAF
	I	(0.023, 0.607]	**-0.017**	0.502	**0.983**	0.498		
	II	(0.607, 0.717]	0.075	**-0.029**	0.925	**0.971**		
QED	III	(0.717, 0.789]	**0.034**	-0.259	**0.966**	0.741	**0.953**	0.653
	IV	(0.789, 0.850]	**-0.017**	-0.403	**0.983**	0.597		
	V	(0.850, 0.948]	**-0.090**	-0.540	**0.910**	0.460		
	I	(16.042, 272.372]	0.505	**0.430**	0.495	**0.570**		
	II	(272.372, 313.466]	**-0.051**	-0.209	**0.949**	0.791		
Weight	III	(313.466, 345.402]	**-0.273**	-0.317	**0.727**	0.683	0.677	**0.739**
	IV	(345.402, 377.259]	-0.394	**-0.339**	0.606	**0.661**		
	V	(377.259, 836.218]	-0.393	**0.012**	0.607	**0.988**		

where closer resemblance of the QED distributions from GCPN generated and training molecules is shown in Fig. 3 (a). The higher-than-training QED histogram of GraphAF molecules for bin-I (i.e., QED < 0.6) is detected by SNOM= 0.502 (i.e., over-generation) in Table 3, whereas lower-than-training QED histogram for bin-V (i.e.,QED > 0.8 is detected with SNOM $= -0.540$ (i.e., under-generation). When Molecular Weight is considered (see Fig. 3 (b)), GCPN's distribution resembles that of the training, particularly in the bins II and III, i.e., in the weight range between 270 $<$ and $<$ 354 Daltons). However, GraphAF's histogram shows closer resemblance in the tails of the training, i.e., Weight $>$ 345 Daltons. The feature-based independence score (FIS) is aggregated from SIS values of all the bins in a feature as I_{km}. The QED-based GCPN's FIS = 0.953 and GraphAF's FIS = 0.653 demonstrate that GCPN model has superior capability of generating molecules with similar QED values as the training molecules, i.e., higher independence. On the other hand, GraphAF has achieved higher Molecular Weight-based (FIS =0.739) than that of GCPN (FIS = 0.677) due to its higher SIS values for extreme bins (tails of the distribution).

Table 4 provides extended FIS values across each of the features considered for the analysis in this work. Based on FIS values, the results show that GCPN generated molecules demonstrate higher independence compared to GraphAF's molecules. These features include synthetic metrics, such as QED, ESOL, SCScore, SAS and LogP. In addition, GCPN is shown to outperform GraphAF in features such as Ring, Aromatic, Lipinski, Heterocenter and Heterocycle. GraphAF molecules are also shown to demonstrate higher independence from the training set of molecules across features such as Scaffold, HBondAccepter and MolecularWeight. The last two rows in Table 4 represent high-level aggregation of FIS values. The penultimate row represents an examples of SAFIS that aggregated the FIS of selected structural features (i.e., Scaffold, Ring, LargeRing and Aromatic). GCPN's SAFIS = 0.694 compared to GraphAF's



(a)



(b)

Figure 3: Visualization of the comparison of the GCPN and GraphAF models as per their QED values. The QED distribution of the training molecules is also provided to visualize the similarity/divergence of the generated molecules.

Table 4: Feature-based Independence Score (FIS) for all the 15 features considered in the comparative analysis of the two generative models (GCPN and GraphAF). The penultimate row represents an example of Selective Aggregation of Feature-level Independence Score (GAFIS), where Scaffold, Ring, Large Ring and Aromatic features are purposely selected to compare the models in their structural details. The last row represent Global Aggregation of Feature-level Independence Score (GAFIS) that is aggregated across all the features.

Independence Score		Models GCPN	GraphAF
FIS	QED	**0.953**	0.653
	ESOL	**0.910**	0.858
	SCScore	**0.83**5	0.758
	SAS	**0.955**	0.738
	Scaffold	0.644	**0.778**
	LogP	**0.951**	0.938
	Ring	**0.806**	0.669
	LargeRing	**0.580**	0.472
	Aromatic	**0.747**	0.431
	HBondAcceptor	0.054	**0.938**
	HBondDonors	**0.273**	0.248
	MolecularWeight	0.677	**0.739**
	Lipinski	**0.77**1	0.543
	Heterocenter	**0.964**	0.905
	Heterocycle	**0.834**	0.789
SAFIS	Structural features	**0.694**	0.588
GAFIS	All features	**0.730**	0.697

SAFIS = 0.588 reflects GCPN's capability to generate molecules with higher structural independence from the training molecules. The last row represent GAFIS that aggregated the FIS values from all the features considered, and GCPN (GAFIS = 0.730) has shown to slightly outperform GraphAF (GAFIS = 0.697). Note that the results above are generated using equal weighting among different bins. However, the proposed approach is flexible to utilize different weighting strategies that aim to encourage generation molecules with a specific range(s) of feature values.

5.2 Automated Detection and Characterization of Generation Bias

MDSS is applied to detect and characterize generation bias in these models, i.e., the types of molecules generated by each of the generative models under extreme frequency, and the results are shown in Table 5. The first row in Table 5 shows the details of molecules that are most frequently generated by GCPN models (Training vs. GCPN). The results show that the anomalous subset of molecules is described by the combinations of anomalous features as: absence of Scaffold with smaller Molecular Weight ($\leq$ 272.372 Daltons) and LogP $\geq$ 1.303. Compared to the complimentary subset of the molecules (i.e., those not characterized by the identified anomalous subset), these molecules are generated by GCPN with an odds ratio of 11.46.4, 95% CI: (10.99, 11.94, $p < 0.01$). We repeated the same experiment but to identify molecules more frequently generated by GraphAF, i.e., Training vs. GraphAF (second row). The anomalous subset of molecules by GraphAF are characterized by higher SAS ($>$ 3.135) and lower QED ($<$ 0.607) and Ring ($\leq$ 3) values. Compared to the training molecules, GraphAF model generates these types of molecules of with an increased odds ratio of 29.4, 95% CI: (28.11, 30.75), $p < 0.01$.

5.3 Limitations

While our MPEGO toolkit provides a multi-level evaluation of generative models by characterizing the generated molecules at different levels of abstractions, the following limitations need to be considered. First, MPEGO heavily depends of the discretization of continuous features, and the quintile based binning adopted in this work is generic, and it could greatly benefit from domain expert insight. The paper is also lacks direct comparison with existing comparison benchmarks, such as GuacaMol [3] and MOSES [16], which is our immediate next step. Importantly, findings obtained using the proposed MPEGO approach and characteristics driven from the generative models still requires further validations from domain experts.

6 CONCLUSION AND FUTURE WORK

We proposed MPEGO - a simple, generalizable, and model-agnostic evaluation toolkit of generative models for material discovery. Given examples of training and generated molecules from a model, MPEGO employs extraction of physical, chemical, and biological properties of these molecules as features for the analysis. MPEGO consists of two main performance evaluation blocks: Hierarchical Independence Evaluation (HIE) and Multi-dimensional Subset Scanning (MDSS). HIE follows a bottom-up approach to quantify the generation performance of a model, starting from per sub-feature level (at the bottom) to the global aggregation of features (at the top). Thus, HIE provides a flexible performance evaluation of generative models. MDSS is applied to detect and characterize generation bias, i.e., to identify the types of molecules a particular model is more likely to generate. The proposed MPEGO toolkit was validated with ZINC-250K training dataset and two graph-based deep generative models (GCPN and GraphAF). The results show that GCPN generated molecules exhibit higher independence across multiple features, compared to GraphAF's molecules using the training baseline. MDSS results show that GCPN generated molecules, compared to GraphAF's molecules, are found to have lower Molecular Weight, higher QED, and lower SAS values.

Generally, the proposed toolkit provides encouraging evaluation metrics and further insights to understand better the under- or over-generation of molecules and their characterizations at different levels of evaluation. Such insights could help to introspect the generation process and to further improve the generation quality via improved interactions between machine learning researchers and domain experts in material science. Future work aims to extend MPEGO with more functionalities, to evaluate more generative models and molecule representations, to directly compare with

Table 5: Automated characterization of the identified subsets of molecules that were generated by GCPN and GraphAF models with extreme frequency, compared to the training set (ZINC-25OK).

MDSS Input				MDSS Anom. Output			Characterization
Molecules	Size	Outcome	Expected	Subset	Size	Observed	OR (95% CI), p
Training vs. GCPN	259, 455	GCPN	0.039	Scaffold=0 and Weight $\leq$ 272.4 and LogP $\geq$ 1.3	28, 555	0.189	11.46 (10.99, 11.94), $p < 0.01$
Training vs. GraphAF	259, 455	GraphAF	0.039	Ring $\leq$ 3 and SAS $>$ 3.1 and QED $\leq$ 0.6	13, 420	0.376	29.4 (28.11, 30.75), $p < 0.01$

existing benchmarks (e.g., MOSES and GuacaMol), and then to integrate with existing open-source resources to accelerate hypothesis generation in the scientific discovery, such as Generative Toolkit for Scientific Discovery (GT4SD) [23]. Furthermore, the MPEGO-driven insights will be utilized to improve the generation capability, particularly in goal-oriented generation process.

REFERENCES

[1] Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. 2017. Machine learning unifies the modeling of materials and molecules. *Science advances* 3, 12 (2017), e1701816.

[2] Jannis Born, Matteo Manica, Joris Cadow, Greta Markert, Nil Adell Mill, Modestas Filipavicius, Nikita Janakarajan, Antonio Cardinale, Teodoro Laino, and María Rodríguez Martínez. 2021. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on SARS-CoV-2. *Machine Learning: Science and Technology* 2, 2 (2021), 025024. https://doi.org/10.1088/2632-2153/abe808

[3] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. 2019. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* 59, 3 (2019), 1096–1108.

[4] Vijil Chenthamarakshan, Payel Das, Samuel Hoffman, Hendrik Strobelt, Inkit Padhi, Kar Wai Lim, Benjamin Hoover, Matteo Manica, Jannis Born, Teodoro Laino, et al. 2020. Cogmol: target-specific and selective drug design for covid-19 using deep generative models. *Advances in Neural Information Processing Systems* 33 (2020), 4320–4332.

[5] Connor W Coley, Natalie S Eyke, and Klavs F Jensen. 2020. Autonomous discovery in the chemical sciences part II: outlook. *Angewandte Chemie International Edition* 59, 52 (2020), 23414–23436.

[6] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics* 47 (2016), 20–33.

[7] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. 2018. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* 4, 2 (2018), 268–276.

[8] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[9] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. 2020. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical science* 11, 2 (2020), 577–586.

[10] Samuel C Hoffman, Vijil Chenthamarakshan, Kahini Wadhawan, Pin-Yu Chen, and Payel Das. 2022. Optimizing molecules using efficient queries from property evaluations. *Nature Machine Intelligence* 4, 1 (2022), 21–31.

[11] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. 2012. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling* 52, 7 (2012), 1757–1768.

[12] Matthew D Lloyd. 2020. High-throughput screening for the discovery of enzyme inhibitors. *Journal of Medicinal Chemistry* 63, 19 (2020), 10742–10772.

[13] Daniel B Neill, Edward McFowland III, and Huanian Zheng. 2013. Fast subset scan for multivariate event detection. *Statistics in Medicine* 32, 13 (2013), 2185–2208.

[14] Gregory Piatetsky-Shapiro. 1991. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases* (1991), 229–238.

[15] Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. 2013. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of computer-aided molecular design* 27, 8 (2013), 675–679.

[16] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. 2020. Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in pharmacology* 11 (2020), 1931.

[17] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. 2018. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling* 58, 9 (2018), 1736–1741.

[18] Benjamin Sanchez-Lengeling, Carlos Outeiral, Gabriel L Guimaraes, and Alan Aspuru-Guzik. 2017. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). (2017).

[19] Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. 2020. Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382* (2020).

[20] Tiago Sousa, João Correia, Vítor Pereira, and Miguel Rocha. 2021. Generative deep learning for targeted compound design. *Journal of Chemical Information and Modeling* 61, 11 (2021), 5343–5361.

[21] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. 2004. Selecting the right objective measure for association analysis. *Information Systems* 29, 4 (2004), 293–313.

[22] Taffee T Tanimoto. 1958. Elementary mathematical theory of classification and prediction. *International Business Machines Corp.* (1958).

[23] GT4SD Team. 2022. *GT4SD (Generative Toolkit for Scientific Discovery).* https://github.com/GT4SD/gt4sd-core

[24] O Anatole von Lilienfeld and Kieron Burke. 2020. Retrospective on a decade of machine learning for chemical discovery. *Nature communications* 11, 1 (2020), 1–4.

[25] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. 2018. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems* 31 (2018).

[26] G Udny Yule. 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* 75, 6 (1912), 579–652.