2025 11th International Conference on Computing and Artificial Intelligence (ICCAI) | 979-8-3315-2491-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCAI66501.2025.00116

Generative and Evolutionary Approaches for Safe and Sustainable by Design Material Development: A Case Study on Acid Alternatives for Advanced Material Coatings

Andrea Gregores Coto
Artificial Intelligence and Data Analytics. Smart Systems and Smart Manufacturing
AIMEN Technology Centre
O Porriño, Spain
andrea.gregores@aimen.es

Christian Eike Precker
Artificial Intelligence and Data Analytics. Smart Systems and Smart Manufacturing
AIMEN Technology Centre
O Porriño, Spain
christian.precker@aimen.es

Leticia Hernando Rodrıgúez
Intelligent Systems. Maths Department
University of the Basque Country
Leioa, Spain
leticia.hernando@ehu.eus

Santiago Muiños Landín *
Artificial Intelligence and Data Analytics. Smart Systems and Smart Manufacturing
AIMEN Technology Centre
O Porriño, Spain
santiago.muinos@aimen.es
*Corresponding author

***Abstract*—The intersection of technological advancement and environmental sustainability presents a challenge in contemporary society. This study explores this intersection within the framework of the MOZART project, a research initiative funded by the European Union's Horizon 2021 program. At its core, the MOZART project seeks to advance Safe and Sustainable by Design (SSbD) strategies, with a specific focus on replacing environmentally harmful Hard Chromium coatings. These coatings, while offering exceptional performance, pose significant environmental risks due to the carcinogenic nature of Chromium Trioxide, a byproduct of their production process. The proposed solution centers on nickel matrix nano-composite coatings. However, the traditional use of boric acid in the process to obtain them, contradicts sustainability principles, necessitating the exploration of sustainable alternative acids with similar performance. Using a Conditional Variational Autoencoder (CVAE), a specialized generative machine learning model, this study investigates potential boric acid substitutes. A comparative analysis of molecular representations is conducted to assess their suitability as inputs for the CVAE. Subsequently, an optimization process employing Genetic Algorithms is undertaken to identify regions within the CVAE's latent space where sustainable molecules similar to boric acid are located.**

***Index Terms*—Safe and Sustainable by Design, Inverse Molecular Design, Machine Learning, Generative models, Variational Autoencoder, Genetic Algorithms, Latent Space, Molecular Representation**

I. Introduction

Technological advancements and environmental sustainability are increasingly intertwined as society seeks innovation with minimal ecological impact [1]. The European Union's Horizon 2021 MOZART project [1] addresses this need, targeting the replacement of toxic hard chromium coatings with nickel-matrix nanocomposite alternatives. However, the electroplating process to obtain these coatings relies on boric acid as a pH regulator, driving the need for sustainable alternatives compliant with EU regulatory standards.

This study employs an Inverse Molecular Design approach, using Conditional Variational Autoencoders (CVAE) to explore potential boric acid substitutes based on their molecular properties [2]. By evaluating two molecular representations—SMILES and SELFIES—this research assesses their reliability in a CVAE framework for molecular generation tasks [3], [4]. The most robust representation will be applied to explore viable electroplating alternatives.

A novel aspect of this research is the combination of CVAEs with Genetic Algorithms (GA), which optimizes molecules

[1](https://www.mozart-project.eu/)

979-8-3315-2491-3/25/$31.00 ©2025 IEEE
DOI 10.1109/ICCAI66501.2025.00116

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

within the latent space, targeting sustainable molecules with properties aligned to boric acid [5].

The study's contributions are threefold: (1) an innovative generative-optimization methodology for sustainable molecular design, (2) an empirical comparison of molecular representations for generative tasks, and (3) insights into sustainable replacements for industrial chemicals in alignment with EU goals.

II. Background and Related Work

The pursuit of sustainable industrial practices has gained momentum since the 1987 "Our Common Future" report defined sustainable development as *"meeting the needs of the present without compromising future generations"* [1], [6]. Current EU initiatives, including the European Green Deal and Chemical Strategy for Sustainability, promote Safe and Sustainable by Design (SSbD) approaches [7], [8]. These initiatives highlight a need for sustainability-focused innovations, such as replacing carcinogenic hard chromium (HC) coatings with nickel-matrix nanocomposite coatings—a key goal of the MOZART project. However, replacing boric acid, which presents regulatory restrictions under EU REACH guidelines [9], as a pH regulator in electroplating processes is essential to the development of a fully sustainable alternative. Inverse molecular design has proven effective in identifying sustainable alternatives for industrial chemicals [2], [10]–[12]. This approach utilizes computational methods to predict molecules with target properties, aligning well with SSbD goals [2], [5]. Among various techniques, generative models have gained popularity for their ability to learn complex data distributions and generate valid, property-aligned structures [2]. Variational encoders (VAE) are particularly suitable for this study due to their controlled generation capabilities within a defined latent space [2]. Compared to other generative models like Generative Adversarial Networks (GAN), VAEs allow for more direct optimization of molecular structures without additional reinforcement learning requirements [13].

Choosing an appropriate molecular representation is critical for training generative models. The Simplified Molecular Input Line Entry System (SMILES) is widely used due to its accessibility in datasets, but it has limitations, such as generating invalid structures and lacking robustness [3], [14], [15]. To overcome these issues, Self-Referencing Embedded Strings (SELFIES) was introduced, offering a more reliable alternative by ensuring valid molecular outputs through its formal grammar [4], [16]. In this study, SMILES and SELFIES are compared to evaluate their effectiveness in training models, aiming to identify the more robust representation for generating valid molecular structures.

After training the VAE, latent space exploration is critical to identifying regions of interest where potential alternative molecules can be found. Previous works have applied Gaussian Process (GP) models to smooth the latent space for molecular property optimization [14]. However, in this study, the focus is on using genetic algorithms (GAs), building upon a parallel population-based approach [5]. This method allows for a comprehensive exploration of the chemical space while exploiting the best-performing candidates, ultimately leading to a more refined understanding of the latent space and the identification of promising regions for alternative molecules.

III. Methodology

A. Conditional Variational Autoencoder

Variational Autoencoders (VAE) consist of two neural networks: an encoder and a decoder. In the context of inverse molecular design, the encoder maps the molecular representation to a fixed-length continuous vector, called a latent vector, which resides in a lower-dimensional space known as the latent space [2], [14], [17]. The core idea behind VAEs is to introduce a probabilistic element into the encoding process by sampling from a probability distribution over the latent space. This regularization encourages the model to explore broader regions of the latent space, allowing the decoder to reconstruct valid molecular structures from various points within this space. Consequently, a molecule is not mapped to a single point but to a probability distribution, enhancing the model's ability to discover diverse molecular representations [2], [14].

For this study, a Conditional Variational Autoencoder (CVAE) is used. Unlike a standard VAE, the latent vector in a CVAE incorporates additional conditions or properties, enabling more controlled generation. In the molecular inverse design context, this means the latent vector includes both the molecular structure and specific properties, such as the dissociation constant (pK_a) and the temperature of measurement, as shown in Fig. 1. This setup allows for the independent control of molecular structures and their properties, making it easier to generate diverse molecular structures with specific desired properties without requiring further optimization [15].

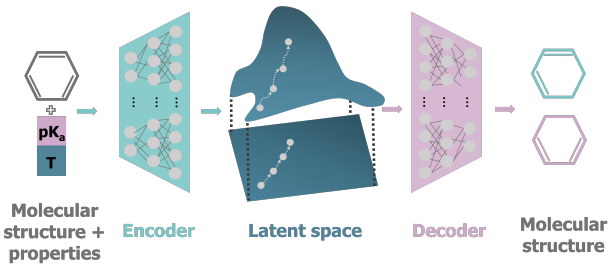

Fig. 1. Representation of a Conditional Variational Autoencoder. The corresponding vector representing each molecular structure in training set, concatenated with their properties conforms the inputs of the model.

The CVAE's loss function is similar to that of a standard VAE, with an added term for conditions. The loss is composed of two parts:

$$E_{z\sim Q}[logP(X|z,c)] - \mathcal{D}[Q(z|X,c)||P(z|c)], \qquad (1)$$

where the first term represents the reconstruction error, encouraging the decoder to accurately reconstruct the molecule

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

from the latent space, and the second term is the Kullback-Leibler (KL) divergence, which regularizes the latent space by aligning the encoder's distribution with a prior distribution [13], [17], [18]. This regularization ensures smooth transitions within the latent space, promoting better generalization.

Both the encoder and decoder in the CVAE are implemented using a recurrent neural network (RNN) with Long Short-Term Memory (LSTM) cells [14], [15], [19], [20]. LSTMs are well-suited for sequential data, such as text-based molecular representations like SMILES, as they capture long-range dependencies between characters in the sequence [19]. This architecture allows the decoder to generate molecular structures character by character, conditioned on the previous part of the sequence. The decoding process is sequential, starting with an initial character, and each subsequent character is predicted based on the previously decoded portion of the sequence [14], [15].

The dataset used for training this model is available in [21]. It contains the SMILES representation of each acid molecule, its dissociation constant (pK_a), and the temperature of measurement. For comparison purposes, one model is trained with the SMILES representation directly, while another model is trained using SELFIES, a more robust molecular representation, after translating the SMILES strings using the corresponding libraries [4].

B. Genetic Algorithm for Latent Space Optimization

Genetic algorithms are based in the principles of biological evolution dictated by Darwin, so that populations evolve through natural selection across generations. In this study, this is employed to find the region of interest in latent space where optimal molecules are located. The goal of the optimization process is to identify registered molecules that exhibit similar behavior to boric acid while adhering to Safe and Sustainable by Design (SSbD) principles, as outlined in this study.

The GA used in this work is inspired by the JANUS algorithm [5], which employs two parallel populations: an explorative population to cover a wide range of potential solutions, and an exploitative population that focuses on improving the best-performing solutions. This dual approach ensures both breadth and depth in exploring the latent space.

The fitness function in this study is a multi-objective function that evaluates molecules based on two key similarity metrics:

- Tanimoto coefficient: Quantifies the overlap between the features given by two molecules [22]–[25]. Formally, given a fingerprint A of a molecule and anoher fingerprint B, the Tanimoto similarity is computed as

$$T(A,B) = \frac{A \cap B}{|A| + |B| - A \cap B} \tag{2}$$

- Distance in latent space: Measured as the Euclidean distance between two latent vectors given the formula

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3}$$

  Due to the probabilistic nature of encoding and decoding in the Conditional Variational Autoencoder (CVAE), the Euclidean distance between a molecule and boric acid is calculated multiple times to capture the variability across iterations, and both the mean and standard deviation are considered.

To guide the optimization process, the fitness function combines these metrics to identify molecules similar to boric acid while considering sustainability. Table I contains 14 acids initially proposed as alternatives which were chosen based on literature [26], [27], previous experience and because they are all REACH compliant. Importantly, boric acid is a monoprotic acid [28] with a pK_a of 9.14 [29]. Therefore, these 14 selected acids ensure that they are viable replacements for boric acid, and thus they serve as reference points for the optimization process.

TABLE I
14 ACIDS INITIALLY PROPOSED IN THE PROJECT AS POSSIBLE CANDIDATES TO REPLACE BORIC ACID. THEIR pK_a AND TEMPERATURE OF MEASUREMENT ARE ALSO INCLUDED.

	pK_a	**Temperature of measurement (°C)**
Citric acid	2.79	25
Succinic acid	4.21	25
Adipic acid	3.90	25
Lactic acid	3.86	20
Heptanoic acid	4.40	25
Lauric acid	5.30	20
Terephthalic acid	3.54	25
Glutaric acid	4.34	25
Ascorbic acid	4.70	10
Maleic acid	1.83	25
Fumaric acid	3.03	18
Oxalic acid	1.46	25
Acetic acid	4.76	25
Tartaric acid	2.89	25

The mathematical expression of the objective is articulated as follows

$$\max F(x,p) = w_D \cdot F_D(x) + w_T \cdot F_T(p), \tag{4}$$

where $F_D(x)$ represents the fitness component based on Euclidean distance x in the latent space and $F_T(p)$ the component based on the Tanimoto coefficient p. In this study, the Euclidean distance is given a weight of 0.6, while the Tanimoto coefficient is weighted at 0.4. Adjusting these weights could shift the emphasis between broad molecular similarity and latent space distance, affecting the overall optimization outcome.

For each acid, its corresponding metric—either the Tanimoto coefficient or the Euclidean distance—with respect to boric acid is computed. The mean of all metrics provides a reference point. Additionally, the standard deviation, maximum, and minimum values are calculated for both metrics. These thresholds define penalizations for values that fall below the minimum or exceed the maximum, ensuring that deviations from the reference point are accounted for in the fitness function.

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

The fitness function component, whether related to the **Tanimoto coefficient** or the **Euclidean distance**, incorporates a Gaussian distribution that yields lower values for solutions with higher deviations from the mean as

$$f(m) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(m-\mu)^2}{2\sigma^2}\right) \tag{5}$$

Here:

- m represents the metric, which can be the **Tanimoto coefficient** p or the **Euclidean distance** x.
- μ is the mean of the metric (either μ_T for Tanimoto or μ_D for Euclidean distance).
- σ is the standard deviation (either σ_T for Tanimoto or σ_D for Euclidean distance).

Additionally, penalization terms are applied for values exceeding the maximum or falling below the minimum thresholds for each metric:

$$g(m) = \max(0, m - m_{\text{max}}), \tag{6}$$

$$k(m) = \max(0, m_{\text{min}} - m), \tag{7}$$

where m_{max} and m_{min} represent the upper and lower thresholds, respectively, for each metric (Tanimoto or Euclidean distance).

The final fitness function component for each metric is computed as:

$$F(m) = f(m) - \lambda \cdot (g(m) + k(m)) \tag{8}$$

where λ is a penalization weight that can be adjusted based on the desired importance of the penalization terms.

Fig. 2 shows the key steps in the Genetic Algorithm that can be summarized as follows:

- Initialization: The GA starts by initializing the population with individuals (molecules) from the dataset used to train the CVAE [21]. Each molecule's fitness score is calculated, the population is sorted in descending order of fitness, and the top-performing molecules are selected, constrained by a predefined generation size. This limitation ensures computational efficiency and prevents excessively long execution times.
- Selection and Mutation: The GA then divides the population into two subsets: *keep smiles*, which are individuals that automatically advance to the next generation, and *replace smiles*, which undergo mutation or crossover for diversity. The division is determined by a *keep ratio*, which is based on the top percentage of the population's fitness scores.

  The probability of an individual being assigned to *replace smiles* is computed using the following formula:

$$P = \left(3^{\frac{(\text{F50} - \text{fitness})}{(\text{F50} - \text{F}_{25})}} + 1\right)^{-1} \tag{9}$$

  Here, F_{50} and F_{25} represent thresholds based on fitness values. Individuals with fitness scores above the F_{50} threshold advance, while the remaining are subject to mutation and crossover.
- Crossover and Mutation: The *replace smiles* subset undergoes genetic diversity enhancement through mutation and crossover operations. Half of this subset is mutated, and the other half undergoes crossover, where the first parent is chosen from *replace smiles* and the second randomly from *keep smiles*. These operations are constrained by the generation size to maintain efficiency.
- Evaluation and Filtering: After mutation and crossover, the resulting offspring are evaluated to check if they are already present in the initial dataset or registered in external repositories such as PubChem [30]. Additionally, only molecules with known pK_a and temperature of measurement are retained, ensuring they meet the sustainability criteria. The fitness function is recalculated for these valid molecules, contributing to the next explorative population.
- Exploitation Population: The best-performing molecules from the explorative population are transferred to the exploitative population, where they undergo further mutation. Only those molecules with higher similarity to the top-performing molecule (based on the Tanimoto coefficient) are retained. Similar to the values and temperatures of measurement. Their fitness scores are then recomputed.
- Replacement and Iteration: Finally, the worst-performing molecules in the explorative population are replaced by the top-performing molecules from the exploitative population. The updated explorative population then becomes the starting point for the next iteration. This process repeats until the algorithm converges or reaches a predefined number of generations.

IV. Results

A. CVAE Performance

In certain models, particularly those designed for prediction tasks, performance metrics such as accuracy play an important role in evaluating the model once trained [31]. However, in the present scenario, the objective is to assess the latent space's ability to capture the intrinsic characteristics and properties of molecules. Consequently, there is not a specific metric tailored for this purpose, necessitating the use of diverse evaluation methods.

To commence, it is essential to verify the convergence of the loss function for both models. Once achieved, a comparative analysis between the SMILES and SELFIES models will be conducted to discern any disparities in terms of robustness. Following this, considering the superior model identified from this evaluation, further comparisons will be made, including an examination of various sampling methods. Additionally, to assess their performances, Tanimoto coefficient and Euclidean distance will be computed between given molecules and boric acid.

1) Training convergence: The convergence of the loss function is a critical metric for evaluating the effectiveness of the training process in generative models. The two key

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

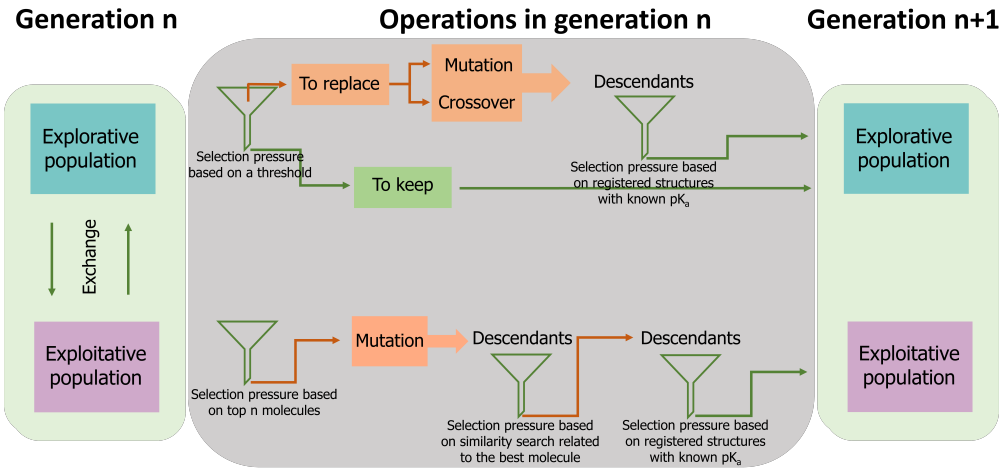

Fig. 2. Representation of the Genetic Algorithm based on two populations. Mutation and crossover are illustrated in orange. Green funnels represent the selection pressures imposed on each population to pass to the next generation. Lines in green indicate the definitive selection of descendants.

factors influencing convergence are batch size and epochs. Batch size refers to the number of samples processed in each iteration of the model, while epochs denote the number of times the model iterates over the entire training dataset during the training process [32].

In this study, both the SMILES and SELFIES models were trained under different conditions to assess their respective convergence behaviors. The SMILES model was trained for 100 epochs with a batch size of 128, exhibiting convergence in its loss function, as shown in Fig. 3(a). Similarly, the SELFIES model was trained for 150 epochs with the same batch size, achieving convergence as illustrated in Fig. 3(b).

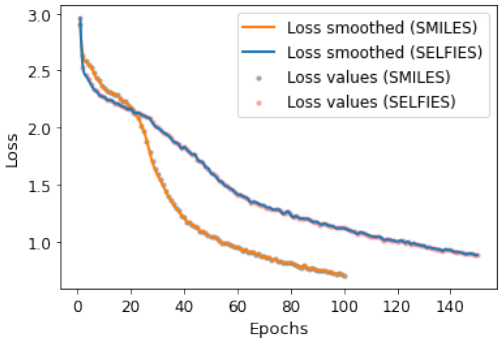

Fig. 3. Convergence of the loss function for SMILES and SELFIES representation models.

Following the completion of training, both models demonstrated the ability to generate molecular structures from appropriate input data. Given the probabilistic nature of the encoding and decoding processes, multiple iterations were performed for each input, allowing for a thorough analysis of the generated molecules and their proximity to the target structures.

2) Comparison between SMILES and SELFIES: A comparison between SMILES and SELFIES representations was conducted to evaluate the validity and diversity of decoded molecules generated by each model. Random sampling, was used to generate inputs for both models, so that the molecular properties of boric acid were set as the target, and latent vectors were randomly sampled from a Gaussian distribution. Three different cases were studied based on the standard deviation of the Gaussian distribution:

- Case 1: Mean of 0 and standard deviation of 1
- Case 2: Mean of 0 and standard deviation of 3
- Case 3: Mean of 0 and standard deviation of 7

For each case, the percentage of non-duplicated molecular representations was calculated. As shown in Table II, across all cases, the SELFIES representation consistently exhibited a higher percentage of non-duplicated molecules compared to SMILES. This discrepancy is likely due to the fact that a single molecule can have more SMILES representations than SELFIES, which leads to a less diverse set of decoded molecules for the SMILES model.

The validity of molecules was also evaluated using RDKit [33], translating each decoded molecular representation into a molecular object. The results indicate that the SELFIES model achieved a 100% validity rate across all cases, attributed to its robust internal syntax, which ensures valid molecular outputs [25]. In contrast, the SMILES model produced a significant number of invalid molecules, especially as the standard deviation of the Gaussian distribution increased.

These results, summarized in Table II, demonstrate that the SELFIES representation not only provides a more diverse set of molecules but also ensures a higher percentage of valid molecules compared to SMILES. Therefore, the SELFIES model was selected for further evaluations and validations in subsequent stages of this study.

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

TABLE II
COMPARISON OF THE VALIDITY OF RESULTS BETWEEN THE SELFIES AND SMILES MODELS. THE EVALUATION PROCESS INVOLVED COMPUTING THE PERCENTAGE OF NON-DUPLICATED MOLECULAR REPRESENTATIONS, FOLLOWED BY THE PERCENTAGE OF VALID MOLECULES.

	Case 1		Case 2		Case 3	
	SMILES	SELFIES	SMILES	SELFIES	SMILES	SELFIES
% non duplicated	25.47	95.31	77.50	99.97	80.94	96.88
% valid molecules	100	32.13	100	10.85	100	5.97

3) Comparison of Sampling Methods for Latent Vector Generation: Various sampling methods were applied to generate molecular structures from the latent space of the Conditional Variational Autoencoder (CVAE) model:

- **Random Sampling**: Molecular structures are generated by sampling from a Gaussian distribution (mean 0, standard deviations 1, 3, and 7). These vectors are concatenated with target properties.
- **Random Sampling from Dataset**: Structures are selected from the training set, perturbed with Gaussian noise, and concatenated with target properties.
- **Target Sampling**: The boric acid structure is perturbed with Gaussian noise and concatenated with target properties.

To validate these methods, the generated molecules were evaluated for registered pK_a values using a local database, PubChem [30], and the training dataset. It was also analyzed the most frequently decoded molecules and checked for the presence of any initially proposed acids (see Table I).

4) Random Sampling Results: In Random Sampling, three Gaussian distributions (standard deviations 1, 3, 7) were used to sample latent vectors. Results show that increasing the standard deviation yields more unique molecules (Table III). Case 2 is particularly interesting, with butyric acid as the most frequently decoded molecule, which is a registered one.

TABLE III
GENERAL RESULTS FOR RANDOM SAMPLING METHOD OF THE MOLECULAR STRUCTURE FOR THE THREE DIFFERENT CASES.

	Case 1	Case 2	Case 3
% of non duplicated SELFIES decoded	16.17	74.53	78.44
% of registered structures	3.36	14.61	6.02
Most decoded structure	`Cc1ccc(O)c(O)c1C(=O)O` (19 %)	Butyric acid (4.1 %)	`CCC` (4.1 %)

Focusing on registered molecules, the pK_a analysis (Fig. 4) indicates that case 1 does not meet the expected pK_a value, likely due to lower variability. However, cases 2 and 3 fall within the expected range, albeit with high standard deviations. Among registered acids, decanoic acid is the most frequent in case 1, while butyric acid and acetic acid appear in cases 2 and 3 (Table IV).

Random Sampling from Dataset was also assessed. However, results varied depending on the random structure chosen from the dataset. While this approach remains a valid methodology for generating novel molecular structures, it is less relevant in the current context, which emphasizes known molecules for boric acid substitution.

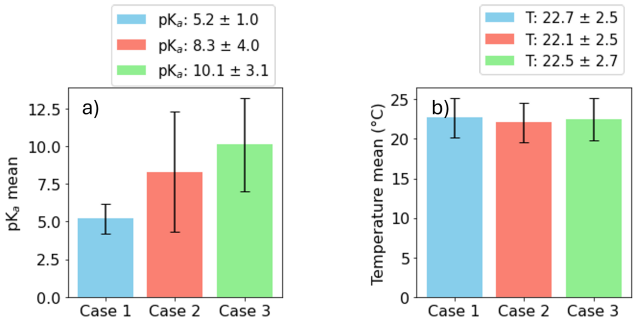

Fig. 4. pK_a and temperature mean values along with their standard deviations for the three different cases of generated molecules registered with known pK_a using random sampling.

TABLE IV
RESULTS FOR THE THREE DIFFERENT CASES FOR REGISTERED STRUCTURES GENERATED BY RANDOM SAMPLING.

	Case 1	Case 2	Case 3
% of unic structures	18.60	13.37	25.97
Most decoded structure	Decanoic acid (28 %)	Butyric acid (28 %)	`CCCCS` (17%)
Proposed acids	Heptanoic acid Lauric acid (7 %, 2.3 %)	Acetic acid Heptanoic acid Lauric acid (2.14 %, 1.6 %, 0.53 %)	Acetic acid (1.3 %)

5) Target Sampling Results: Target Sampling was initially restrictive, producing poor diversity. To address this, Gaussian noise was added to boric acid properties. For this, perturbation on the boric acid structure was set for all cases with a standard deviation of 7. Three cases were considered: case 1 with no noise in properties, case 2 with a standard deviation of 3, and case 3 with a standard deviation of 5 in target properties.

As shown in Table V, increasing the noise in target properties results in enhanced diversity of generated SELFIES. In terms of registered structures, a slight decrease is observed, possibly due to increased flexibility in altering structures.

TABLE V
GENERAL RESULTS FOR TARGET SAMPLING OF THE MOLECULAR STRUCTURE AND PERTURBATION OF TARGET PROPERTIES WITH GAUSSIAN NOISE.

	Case 1	Case 2	Case 3
% of non duplicated SELFIES decoded	1.02	1.25	1.4
% of registered structures	72.5	74.53	64.84
Most decoded structure	`CCCCCCS` (68.7 %)	`CCCCCCS` (71.3 %)	`CCCCCCS` (60.2 %)

Turning to the registered structures (Table VI), the diversity of unique structures increases with more noise in target properties. The consistent appearance of propionic and butyric acids across all cases is noteworthy. In the highest-noise case, acetic acid, one of the proposed replacements, emerges as well, which is particularly noteworthy.

6) Comparison between Tanimoto coefficient and distance in latent space: To better understand how the latent space captures molecular features, it was compared the Euclidean distance in latent space to boric acid (see 3) with the Tanimoto

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

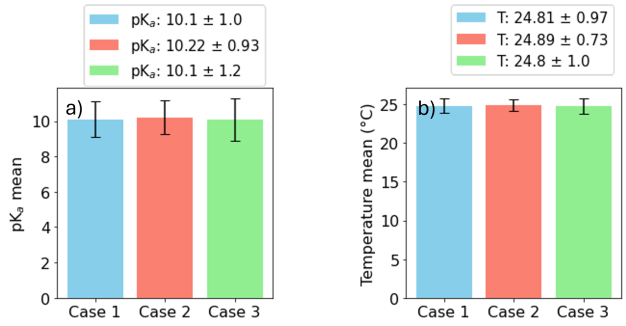

Fig. 5. pK_a values along with their standard deviations for the three different cases for target sampling with and without noise in target properties.

TABLE VI
RESULTS OF REGISTERED STRUCTURES GENERATED BY TARGET SAMPLING WITH GAUSSIAN NOISE IN TARGET PROPERTIES.

	Case 1	Case 2	Case 3
% of unique structures	0.54	0.52	0.72
Unique structures	Propionic acid Butyric acid CCCCCCS CCCCCCN CCCCO	Propionic acid Butyric acid CCCCCCS CCCCCCN CCCCO	Propionic acid Butyric acid Acetic acid CCCCCCN CCCCCCS,CCCCO

coefficient (see 2) for a set of molecules. These molecules are specified in Table VII, encompassing the initially proposed acids and the most frequently decoded registered acids.

Due to the probabilistic nature of encoding and decoding in the CVAE, multiple iterations were performed for each molecule to compute the Euclidean distance from boric acid, and the mean and standard deviation of these values were considered.

The Tanimoto coefficient ranges from 0 to 1, representing the proportion of molecular features shared between molecules [22], [25]. As the number of features used in fingerprints can vary, it was fixed a 2048-bit restriction for all Tanimoto coefficient computations to ensure consistency in comparison [24].

As shown in Table VII, lactic and oxalic acids have the same Tanimoto coefficient (0.10526), and their mean Euclidean distances (19.93 and 19.92, respectively) are also closely aligned. Similarly, maleic and fumaric acids share the same Tanimoto coefficient (0.07692), though their Euclidean distances are slightly less aligned (19.98 and 20.03). These results suggest that the latent space captures structural similarity effectively in these cases.

In terms of Tanimoto coefficients, lactic and oxalic acids have the highest values. Notably, these acids do not appear in any of the sampling cases, potentially suggesting they are not as close to boric acid as other structures in terms of Euclidean distance. This is confirmed by checking that lactic and oxalic acids, despite having shorter Euclidean distances compared to other structures, do not have the shortest distances overall. Acetic, butyric, and propionic acids, exhibit the shortest distances.

Conversely, lauric and decanoic acids have the lowest Tanimoto coefficient. Although their distances are not the greatest, they cannot be considered short when compared to the average. Ascorbic, citric, and tartaric acids exhibit the largest Euclidean distances.

Overall, while the Tanimoto coefficient and Euclidean distance in latent space are not fully correlated, patterns emerge when identical Tanimoto coefficients are observed. These results highlight that the two metrics are complementary in assessing molecular similarity, justifying their combined use in the Genetic Algorithm's fitness function.

TABLE VII
ANALYSIS OF THE TANIMOTO COEFFICIENT FOR THE PROVIDED MOLECULE SET COMPARED TO BORIC ACID. ADDITIONALLY, MEAN AND STANDARD DEVIATION OF THE EUCLIDEAN DISTANCE IN LATENT SPACE BETWEEN THE GIVEN MOLECULES AND BORIC ACID.

	Tanimoto coefficient	Mean of Euclidean distance
Citric acid	0.08108	20.39 ± 0.96
Succinic acid	0.07692	19.98 ± 0.98
Adiptic acid	0.0625	20.07 ± 0.97
Lactic acid	0.10526	19.93 ± 0.99
Heptanoic acid	0.03226	20.0 ± 1.0
Lauric acid	0.02174	20.10 ± 0.97
Terephthalic acid	0.05263	20.18 ± 0.99
Glutaric acid	0.06897	20.03 ± 0.98
Ascorbic acid	0.08333	20.51 ± 0.96
Maleic acid	0.07692	19.98 ± 0.99
Fumaric acid	0.07692	20.03 ± 0.99
Oxalic acid	0.10526	19.92 ± 0.99
Acetic acid	0.06667	19.89 ± 0.99
Tartaric acid	0.10345	20.21 ± 0.98
Butyric acid	0.04545	19.9 ± 1.0
Propionic acid	0.0556	19.87 ± 0.99
Decanoic acid	0.025	19.98 ± 0.98

B. Genetic Algorithm Performance

During the initial implementation of the Genetic Algorithm, computational challenges emerged, particularly due to the extensive search required in PubChem for registered compounds. To mitigate this, optimizations were made to the algorithm's source code, eliminating unnecessary operations such as redundant identifier checks. Furthermore, an extended dataset was employed, incorporating both local databases and previously generated molecules. These adjustments significantly reduced the runtime, enabling multiple iterations within a 24-hour period.

However, even with these improvements, execution time remained considerable, particularly when incorporating the Euclidean distance component in the fitness function. Consequently, an alternative approach was adopted, focusing solely on the Tanimoto coefficient which improved the runtime in a 2900 % compared to the original multi-objective approach that included Euclidean distance.

Both the training of the CVAE and the execution of the GA were conducted locally on a system with a 10th generation Intel Core i7 processor. This demonstrates that the approach is feasible on non-specialized, moderately powerful hardware, making it accessible to a wide range of users. Nevertheless,

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

larger datasets or higher-dimensional latent spaces could benefit from more powerful hardware, such as GPUs. Running both algorithms on GPUs would reduce training time and enable handling more complex datasets

1) Comparison of Multi-objective and Tanimoto-Only Algorithms: A comparison between the two Genetic Algorithms was conducted, focusing on the first two generations of the explorative population, where both the Euclidean and Tanimoto algorithms overlapped. To visualize the results, Principal Component Analysis (PCA) was applied, reducing the latent space into two dimensions.

- Multi-objective Algorithm (Euclidean + Tanimoto): The first generation shows a cluster located between (0.4, 0.6) in the x-axis and (0.2, 0.4) in the y-axis. In the second generation, the cluster splits into two, with reduced dispersion (Fig. 6a, b).
- Tanimoto-Only Algorithm: The first generation presents a more dispersed set of points along the diagonal (x=0.6, y=0.6) with a cluster between (0.2, 0.4) for the x-axis and (0.2, 0.4) for the y-axis. The second generation shows increased dispersion between 0.6 and 1 on the x-axis, with clustering similar to the multi-objective approach (Fig. 6c, d).

Both algorithms tend to form clusters within the interval 0.3 to 0.6 on the x-axis and 0.2 to 0.4 on the y-axis. However, the Tanimoto-only algorithm evolves more slowly, displaying greater dispersion across generations.

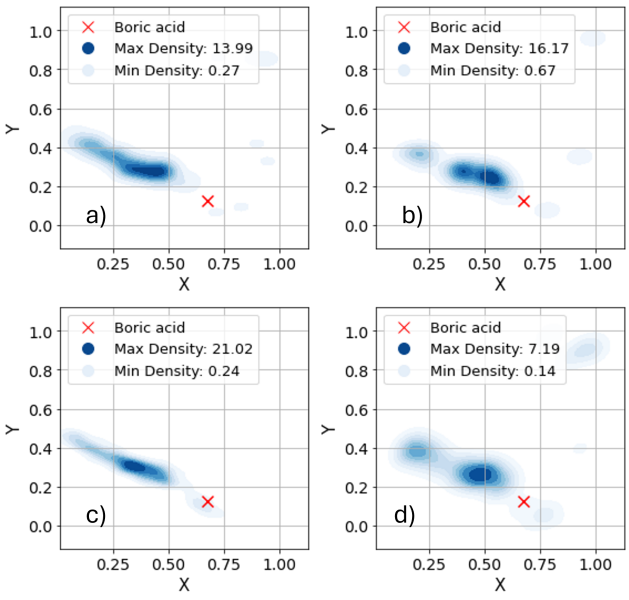

Fig. 6. a) Generation 1 of the first algorithm. b) Generation 2 of the first algorithm. c) Generation 1 of the second algorithm. d) Generation 2 of the second algorithm.

The Tanimoto-only algorithm was evaluated over 90 generations. After generation 30, fewer than three registered molecules were observed in each iteration. By generation 90, only a single registered molecule was found in the expected region (Fig. 7 (j)).

While both algorithms capture essential behaviors in the latent space, the Tanimoto-only algorithm shows a slower convergence, evidenced by the slower evolution of clusters. Despite this, both approaches indicate that the latent space effectively captures molecular similarity to boric acid. The use of both metrics in the original fitness function ensures more balanced optimization.

Finally, locations in latent space of butyric, propionic and decanoic acids are studied. Butyric and propionic acid appeared several times in the decoding process . On the other hand, decanoic appeared only in one of the cases. This difference of frequency in the decoding process can be explained by having a look in Fig. 8. There, it can be seen that butyric and propionic are really close to boric acid in latent space, however, they are out of the interval of interest, suggesting that is possible that they would not meet the sustainability criteria. Here arises again the need of a metric to measure sustainability in a way which can also be included as a condition in the input vector. Contrary, although decanoic acid is not as close in the latent space as the previous two, it is located in the defined region of interest.

V. Conclusions

This methodology serves as a decision-support tool for identifying sustainable alternatives to boric acid in industrial processes. It can be integrated into existing chemical screening pipelines by generating molecular candidates that meet predefined performance and sustainability criteria, which can then be tested through high-throughput or virtual screening.

Moreover, the approach is highly generalizable. With the appropriate training data, it can be extended to explore alternatives to a variety of industrial chemicals, facilitating the design of safer and more sustainable materials across different domains.

While the tool provides valuable candidates, regulatory validation is required to ensure compliance with Safe and Sustainable by Design (SSbD) principles, including safety and environmental standards. The approach offers a promising foundation for developing safer, more sustainable materials and processes across various industries.

This study underscores the critical importance of selecting an appropriate molecular representation as input for Machine Learning models. A significant divergence in the percentage of valid generated molecules arises when comparing the use of SMILES versus SELFIES representations. Employing SMILES requires a greater number of decodings due to the substantial proportion of invalid generated molecules, resulting in increased computational time. This highlights the overarching significance of data quality for any Machine Learning model. Furthermore, it is demonstrated the suitability of Variational Autoencoders (VAEs) as a model for identifying potential replacements for boric acid. The latent space of VAEs offers insights into the structural and property-related relationships between molecules. Different sampling

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

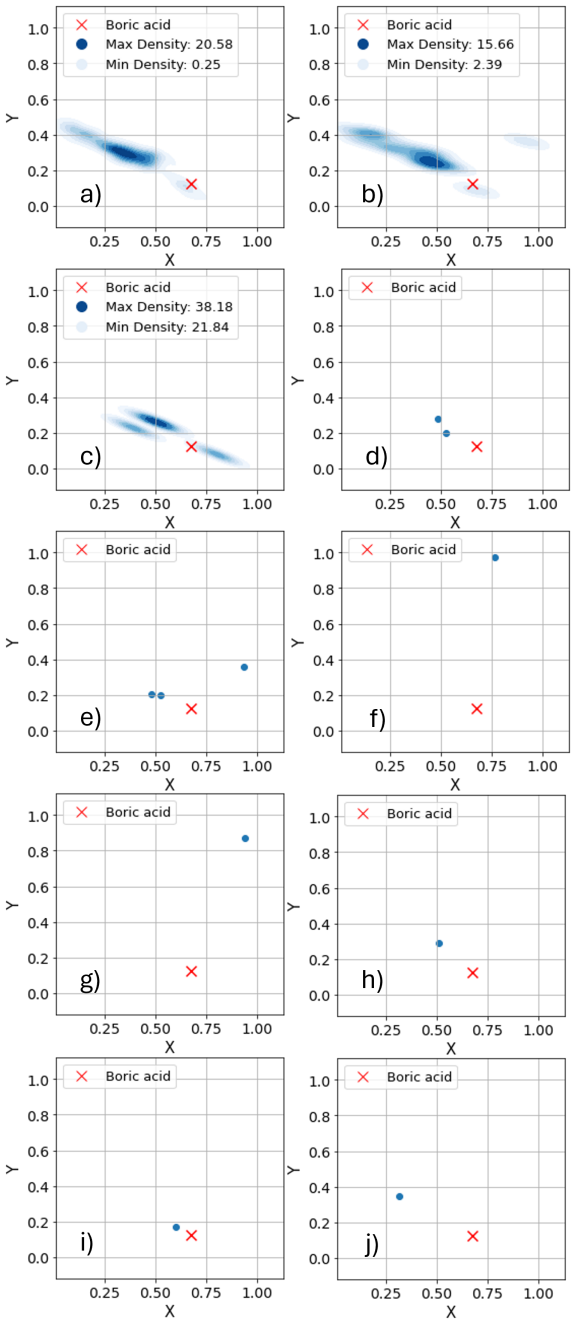

Fig. 7. Representation of the evolution of the density of points in latent space trough 90 generations of the second algorithm.

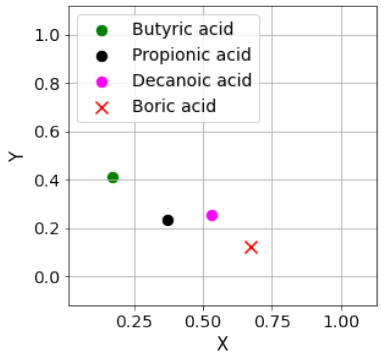

Fig. 8. Location of Butyric, Propionic and Decanoic acids in the latent space.

methods reveal how the restrictiveness of inputs influences the breadth of output variety. The generation of molecules from Table I under various sampling methods suggests that latent variables adeptly capture complex relationships among molecules, as evidenced by the ability to generate acids with similar performance to boric acid. This is corroborated by evaluating Euclidean distances in the latent space among these molecules. The definition of molecular similarity metrics is pivotal. While the Euclidean distance effectively captures structural and property-related molecular relationships, it does not entirely mirror the information encapsulated by the Tanimoto coefficient. This discrepancy may stem from utilizing different molecular representations and accounting for distinct molecular features. Hence, both metrics are deemed complementary. Genetic Algorithms (GAs) emerge as a viable optimization method for exploring molecular evolution in latent space across generations. Utilizing two parallel populations—one for exploration and one for exploitation—enhances evolutionary processes, yielding superior candidates in the explorative population. Notably, incorporating both Tanimoto and Euclidean distance metrics accelerates evolution towards regions of interest within populations, further emphasizing their complementary nature. However, there remain areas for improvement in future stages. A primary consideration is the development of a metric to gauge sustainability. Ideally, integrating sustainability scores into the dataset used for VAE training would enhance model performance. Achieving this without necessitating the inclusion of acids from Table I as targets presents a significant challenge.

As an alternative to methodologies combining GANs and Reinforcement Learning (RL) which optimize latent spaces indirectly, the combination of CVAEs and GAs allows for a direct search in the molecular space, reducing computational complexity. In addition, GAs offer the advantage of customizable fitness functions, enabling the focus on specific properties, such as those aligned with Safe and Sustainable by Design (SSbD) principles. This direct approach also enhances interpretability, providing better insights into the latent space. Overall, the approach based on CVAEs + GAs presented in this work, offers a straightforward method, efficient, and well-

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.

suited for SSbD, making it a robust alternative to GANs+RL approaches for sustainable molecular design.

Nonetheless, the objective is to incorporate sustainability into the Genetic Algorithm's fitness function to guide evolution towards regions of latent space containing sustainable acids. Finally, a recurring challenge throughout this work is the absence of a comprehensive global database containing all molecules alongside their properties. Instances where molecules were registered but lacked essential properties like pK_a, or were absent from certain databases like PubChem, underscore the need for a unified database. While time-consuming, efforts to consolidate information across databases are essential for facilitating research endeavors across various fields.

ACKNOWLEDGMENT

This work has been supported by the project "METAL MATRIX NANO-COMPOSITE COATINGS UTILIZATION AS ALTERNATIVE TO HARD CHROMIUM" (MOZART), which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No.101058450.

REFERENCES

[1] A. Azapagic and S. Perdan, "Indicators of sustainable development for industry: a general framework," *Process Safety and Environmental Protection*, vol. 78, no. 4, pp. 243–261, 2000.

[2] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360–365, 2018.

[3] D. S. Wigh, J. M. Goodman, and A. A. Lapkin, "A review of molecular representation in the age of machine learning," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 12, no. 5, p. e1603, 2022.

[4] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, 2020.

[5] A. Nigam, R. Pollice, and A. Aspuru-Guzik, "Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design," *Digital Discovery*, vol. 1, no. 4, pp. 390–404, 2022.

[6] B. R. Keeble, "The brundtland report: 'Our common future'," *Medicine and War*, vol. 4, no. 1, pp. 17–25, 1988. [Online]. Available: https://doi.org/10.1080/07488008808408783

[7] C. Fetting, "The european green deal," *ESDN report*, vol. 53, 2020.

[8] M. Herzler, P. Marx-Stoelting, R. Pirow, C. Riebeling, A. Luch, T. Tralau, T. Schwerdtle, and A. Hensel, "The "EU chemicals strategy for sustainability" questions regulatory toxicology as we know it: is it all rooted in sound scientific evidence?" *Archives of Toxicology*, vol. 95, no. 7, pp. 2589–2601, 2021.

[9] P. Fisk, *Chemical risk assessment: a manual for REACH*. John Wiley & Sons, 2013.

[10] J. R. Broach, J. Thorner *et al.*, "High-throughput screening for drug discovery," *Nature*, vol. 384, no. 6604, pp. 14–16, 1996.

[11] A. Supady, V. Blum, and C. Baldauf, "First-principles molecular structure search with a genetic algorithm," *Journal of Chemical Information and Modeling*, vol. 55, no. 11, pp. 2338–2348, 2015.

[12] H. Ikebata, K. Hongo, T. Isomura, R. Maezono, and R. Yoshida, "Bayesian molecular design with a chemical language model," *Journal of computer-aided molecular design*, vol. 31, pp. 379–391, 2017.

[13] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[14] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS central science*, vol. 4, no. 2, pp. 268–276, 2018.

[15] J. Lim, S. Ryu, J. W. Kim, and W. Y. Kim, "Molecular generative model based on conditional variational autoencoder for de novo molecular design," *Journal of cheminformatics*, vol. 10, no. 1, pp. 1–9, 2018.

[16] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka *et al.*, "SELFIES and the future of molecular string representations," *Patterns*, vol. 3, no. 10, 2022.

[17] The theory behind latent variable models: formulating a variational autoencoder. Website: https://theaisummer.com/latent-variable-models/.

[18] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[19] E. J. Bjerrum and R. Threlfall, "Molecular generation with recurrent neural networks (RNNs)," *arXiv preprint arXiv:1705.04612*, 2017.

[20] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[21] A. Gregores-Coto, C. Precker, and S. Muiños Landin, "Hybrid pK_a Dataset for Acid Molecules," *Zenodo*, Jan. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10551522

[22] D. Bajusz, A. Rácz, and K. Héberger, "Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?" *Journal of cheminformatics*, vol. 7, no. 1, pp. 1–13, 2015.

[23] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[24] A. Nigam, R. Pollice, M. Krenn, G. dos Passos Gomes, and A. Aspuru-Guzik, "Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES," *Chemical science*, vol. 12, no. 20, pp. 7079–7090, 2021.

[25] M. Vogt and J. Bajorath, "ccbmlib–a python package for modeling tanimoto similarity value distributions," *F1000Research*, vol. 9, 2020.

[26] H. Riechmann, M. Müller, and A. Meyerovich, "Boric acid free nickel electrolyte for plating of electrical contacts," *Transactions of the IMF*, vol. 95, no. 2, pp. 73–77, 2017.

[27] F. Saito, K. Kishimoto, Y. Nobira, K. Kobayakawa, and Y. Sato, "Nickel electroplating bath using malic acid as a substitute agent for boric acid," *Metal finishing*, vol. 105, no. 10, pp. 192–204, 2007.

[28] J. Jordan and W. Dumbaugh, "Thermometric titration of weak monoprotic acids," *Analytical Chemistry*, vol. 31, no. 2, pp. 210–213, 1959.

[29] A. Asuero, "On the buffer index of a single weak monoprotic acid system," *International journal of pharmaceutics*, vol. 134, no. 1-2, pp. 1–13, 1996.

[30] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu *et al.*, "Pubchem 2019 update: improved access to chemical data," *Nucleic acids research*, vol. 47, no. D1, pp. D1102–D1109, 2019.

[31] B. J. Erickson and F. Kitamura, *Magician's corner: 9. Performance metrics for machine learning models*. Radiological Society of North America, 2021, vol. 3, no. 3.

[32] Difference between a batch and a epoch in a neural network. Website: https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/.

[33] G. Landrum, "Rdkit documentation," *Release*, vol. 1, no. 1-79, p. 4, 2013.

Authorized licensed use limited to: UNIVERSIDAD DE VIGO. Downloaded on May 06,2026 at 10:33:28 UTC from IEEE Xplore. Restrictions apply.