Identifying Nearby Molecules through ECFP4 Fingerprint Bit Flipping

Nils Dunlop, 20010127-2359
gusdunlni@student.gu.se

Francisco Erazo Piza, 19930613-921
guserafr@student.gu.se

Qi Chen, 20001016-8342
gusqichr@student.gu.se

December 3, 2024

Abstract

This study investigates the generation of structurally similar molecules through systematic manipulation of Extended-Connectivity Fingerprint (ECFP4) representations using the MolForge transformer model. The research focuses on five COX-2 inhibitors, examining how bit flipping in their ECFP4 fingerprints affects molecular structure and similarity. The study analyzes the relationship between bit modifications and molecular similarity using Tanimoto coefficients by incrementally modifying 1, 2, 4, 8, 128, and 1024 bits in the 2048-bit fingerprint vectors. Results indicate that small bit changes (1-4 bits) maintain high structural similarity, while larger modifications (8+ bits) generate more diverse molecules but with decreased validity. The optimal balance occurs at 8-bit flips, producing structurally diverse yet valid SMILES strings. However, extensive modifications (128+ bits) predominantly yield invalid or highly divergent structures. Analysis reveals that increased bit flipping correlates with higher molecular weights and atom counts, suggesting greater structural complexity. While the approach demonstrates the potential for exploring chemical space, limitations include MolForge's training data bias and challenges in generating valid SMILES from substantial bit modifications. Future work should focus on retraining MolForge with datasets specifically designed for small-scale molecular variations.

1 Introduction

The exploration of chemical space through computational methods has become increasingly important in drug discovery and molecular design. Among these methods, molecular representations play a crucial role in capturing and manipulating chemical structures digitally. This study focuses on a novel approach to generating structurally similar molecules through systematic manipulation of Extended-Connectivity Fingerprint (ECFP4) representations using the MolForge transformer model. Molecular fingerprints, particularly ECFP4, have traditionally been considered irreversible due to information loss during the conversion of complex molecular structures into fixed-size bit vectors. However, recent advances in deep learning, specifically through models like MolForge, have demonstrated that these fingerprints retain sufficient information for molecular reconstruction. This breakthrough opens new possibilities for exploring chemical space through direct manipulation of fingerprint bits.

The manipulation of fingerprint bits to explore chemical space remains relatively unexplored, with limited previous research focusing primarily on similarity searching. This study addresses this gap by systematically investigating how bit flipping in ECFP4 fingerprints affects molecular structure and similarity. By focusing on a set of COX-2 inhibitors, the research examines the relationship between varying degrees of bit modification and the resulting molecular structures. The approach involves systematically flipping bits in increments (1, 2, 4, 8, 128, and 1024 bits) in the 2048-bit ECFP4 fingerprint vectors of selected molecules. These modified fingerprints are then processed through MolForge to generate new SMILES strings, which are analyzed for structural validity and similarity to the original molecules. This methodology allows for a comprehensive examination of how fingerprint modifications translate to structural changes in the resulting molecules.

Through this investigation, the study aims to advance our understanding of molecular fingerprint manipulation and its potential applications in drug discovery and molecular design. The findings provide insights into the relationship between bit flipping and molecular structure, offering a foundation for future research in targeted molecular generation and optimization. The remainder of this paper details the methodology employed, presents comprehensive results of the bit flipping experiments, discusses the implications and limitations of the findings, and suggests directions for future research in this promising area of computational chemistry.

2 Literature Review

2.1 Background

Molecular representations involve the digital encoding of molecules, which are used as input for training deep learning models. While not mandatory, two desirable features of representations are uniqueness and invertibility. Uniqueness ensures that each molecular structure corresponds to a single representation, whereas invertibility means each representation maps back to a specific molecule, establishing a one-to-one relationship. Although most representations for molecular generation are invertible, many lack uniqueness [2].

Over the years, various molecular representations have been developed to meet the diverse needs of drug discovery and chemical informatics. Two of the most widely used representations are Simplified Molecular Input Line Entry System (SMILES) and Extended-Connectivity Fingerprints (ECFP).

2.2 Simplified Molecular Input Line Entry System (SMILES)

Introduced by Weininger in 1988 [13], SMILES provides a linear string format that captures molecular structures, including atom connections, stereochemistry, and bond types, using ASCII characters. This notation significantly advanced cheminformatics by enabling efficient storage, retrieval, and manipulation of chemical information. While SMILES representations are invertible—they can be converted back to molecular structures—they are inherently non-unique because a single molecule can be represented by multiple valid SMILES strings due to different traversal orders of the molecular graph [2].

2.3 Extended-Connectivity Fingerprints (ECFP)

The foundational Morgan algorithm, introduced by Morgan in 1965 [10], established a pioneering method for generating unique molecular descriptions through an iterative process of computing “connectivity values” for each atom. This technique, developed at Chemical Abstracts Service (CAS), was revolutionary in its ability to create machine-readable chemical structure representations by assigning numerical identifiers to atoms based on their local environment and connectivity patterns.

Building upon this core concept, Rogers et al. [3] introduced the Extended-Connectivity Fingerprints (ECFP), specifically designed for molecular activity modeling. ECFP generates circular fingerprints by iterating over atoms and their neighboring atoms to capture molecular substructures. These substructures are then hashed into a fixed-size bit vector, traditionally 2048 bits. ECFP4, a variant with a maximum diameter of four bonds, offers a refined approach to molecular representation compared to SMILES. While the Morgan algorithm was originally developed to detect molecular isomorphism, ECFP extends its utility to tasks like similarity search and property prediction, becoming a general-purpose tool in cheminformatics.

Historically, ECFP was considered irreversible due to the information loss inherent in converting complex molecular structures into fixed-size bit vectors. However, recent advances in deep learning have challenged this assumption, demonstrating that these fingerprints retain sufficient information for molecular reconstruction. This breakthrough has opened new possibilities for molecular design and optimization, leading to various approaches for reconstructing molecules from their fingerprint representations.

2.4 Molecular Reconstruction from Fingerprints

Recent years have seen significant progress in reconstructing molecular structures from fingerprints, with varying degrees of success and importance for molecular design applications. The ability to accurately reconstruct molecules from fingerprints is crucial for bit flipping experiments, as it determines how reliably structural changes can be generated from modified fingerprints.

Le et al. [6] proposed an early method using neural networks to predict molecular structures from ECFP. Their approach involved comparing the ECFP and SMILES representations of molecules. If no exact match was found, pairwise Tanimoto similarities between ECFPs were calculated and molecular structures with over 90% similarity were returned. The model was trained on the ChEMBL25 dataset, which initially contained nearly 2 million molecules, though preprocessing reduced the sample size to around 1.5 million. Their architecture consisted of a recurrent autoencoder to convert SMILES strings into Continuous Data-Driven Descriptors (CDDD), which were then fed into a fully connected feedforward neural network.

When evaluated on the unseen ChEMBL26 test dataset, the model achieved a 70% success rate in deducing molecular structures from ECFP vectors of length 4096. This achievement, while significant as a proof of concept, left substantial room for improvement in reconstruction accuracy. Subsequent approaches would significantly improve upon this baseline, with reconstruction accuracies reaching over 90%.

A significant breakthrough came from Ucak et al. [5] with their MolForge model. They investigated the reconstruction of molecular structures from various fingerprints including ECFP, atom pairs (AP), topological torsion (TT), and atomic environments (AEs). To address limitations with SMILES, they utilized SELFIES, which provide a more robust alternative by ensuring chemically valid molecular graphs. MolForge's transformer-based architecture proved particularly effective, achieving over 93% reconstruction accuracy for SMILES strings from standard ECFP4 fingerprints. This high reconstruction rate is crucial for bit flipping experiments, as it suggests that modifications to fingerprint bits are more likely to result in valid and meaningful structural changes. This capability, combined with its robust handling of chemical validity through SELFIES, made MolForge an ideal choice for our investigation of bit flipping effects.

Bilsland et al. [1] took a different approach with their dual SMILES autoencoder model for fragment-based hit identification (FBHI), achieving 98% SMILES reconstruction accuracy. Their model incorporated transfer learning in the fingerprint decoder layers. However, its computational intensity and focus on fragment generation made it less suitable for our bit flipping experiments.

Kotsias et al. [4] developed a conditional recurrent neural network (cRNN) approach that achieved up to 72% reconstruction accuracy using four decoding layers. Despite being a more recent approach, its lower accuracy compared to MolForge and Bilsland's model made it less suitable for investigating the subtle structural changes that might arise from bit flipping.

These advances in fingerprint reconstruction, particularly MolForge's high accuracy, provide the foundation for our investigation into bit flipping. The ability to reliably reconstruct molecules from modified fingerprints is essential for exploring how systematic bit changes translate to structural variations in the resulting molecules.

2.5 Bit Manipulation in Fingerprints

Despite the potential applications in drug discovery and molecular design, manipulating fingerprint bits to explore chemical space has been relatively unexplored. This limited exploration may be due to the historical assumption that ECFP fingerprints were irreversible, making bit manipulation seem impractical for generating valid molecular structures.

While bit manipulation has not been explored with ECFP4 fingerprints, Wang and Bajorath's [7] ”bit silencing” study demonstrated the potential of this approach using MACCS keys. Their method systematically evaluated bit positions' contributions to similarity searching by setting individual bits to zero, improving hit rates from 5% to 12%. While their focus was on enhancing similarity searches using MACCS keys, their work demonstrated the potential impact of bit manipulation on molecular representation and similarity assessment.

Recent advances in molecular reconstruction from fingerprints, particularly with models like MolForge achieving high accuracy with ECFP4, present an opportunity to explore bit manipulation in more complex fingerprint systems. Combining efficient bit manipulation strategies with accurate reconstruction capabilities makes it possible to systematically investigate the relationship between fingerprint modifications and molecular structure in ways not previously attempted with ECFP4 fingerprints.

2.6 Objectives

Building upon the limited exploration of fingerprint bit manipulation in previous research and leveraging the advanced reconstruction capabilities of modern models, our project aimed to investigate the relationship between fingerprint modifications and molecular structure systematically. Using MolForge, a transformer model capable of reconstructing SMILES strings from ECFP4 fingerprints with high accuracy, we sought to expand beyond simple bit silencing to explore how various degrees of bit flipping influence molecular structure and could identify potential nearby molecules.

COX-2 inhibitors were selected as a focus for this study due to their critical role in managing inflammatory diseases such as arthritis and cancer and their ability to minimize gastrointestinal side effects compared to older anti-inflammatory drugs. Recent advancements have highlighted their promise for improved effectiveness and safety, making them an essential area for further exploration [11, 12]. The study specifically aims to:

1. **Investigate Systematic Bit Flipping**: Examine the effects of incrementally increasing bit modifications (1, 2, 4, 8, 128, and 1024 bits) on ECFP4 fingerprint vectors and analyze the validity of resulting SMILES.

2. **Discover and Evaluate Similar Molecules**: Identify structurally similar molecules through bit flipping and quantify their relationship to original molecules using Tanimoto similarity scores.

3. **Analyze Structural Changes**: Assess how bit flipping affects molecular properties, including molecular weight, atom count, and overall structural complexity.

4. **Visualize Chemical Space**: Represent the spatial distribution of generated molecules using multidimensional scaling (MDS) to understand the relationship between bit modifications and molecular similarity.

These objectives are designed to bridge the gap between the computational manipulation of fingerprints and practical applications in molecular design and drug discovery.

3 MolForge

MolForge 1 uses a transformer-based architecture optimized explicitly for molecular structure reconstruction. The model comprises an encoder-decoder framework with six identical layers in each stack. The input is processed through molecular fingerprints, such as ECFP4 or MACCS keys, which first undergo an embedding transformation and positional encoding. These embeddings are then scaled by $\sqrt{512}$ (dim_model) to maintain an appropriate magnitude. Each encoder layer implements an 8-headed self-attention mechanism operating in 512 dimensions. This is followed by a feed-forward neural network that expands the representation from 512 to 2048 neurons before projecting back to 512 dimensions. Layer normalization and residual connections are applied at each step to ensure stable gradient flow.

The decoder mirrors the encoder in complexity but introduces an additional cross-attention mechanism between the encoder and decoder stacks. It begins by applying masked self-attention, which restricts position `i` from attending to positions greater than `i` during training. Next, it performs cross-attention, allowing the decoder to focus on relevant parts of the encoded fingerprint representation. The model concludes with a linear transformation that projects the 512-dimensional decoder output to the target vocabulary size, which varies between SMILES and SELFIES representations. LogSoftmax activation is then applied to generate the final molecular string representation. A dropout rate of 0.1 is used consistently throughout the architecture to prevent overfitting. Additionally, all attention mechanisms are scaled by $1/\sqrt{dim_k}$, following the standard transformer architecture.

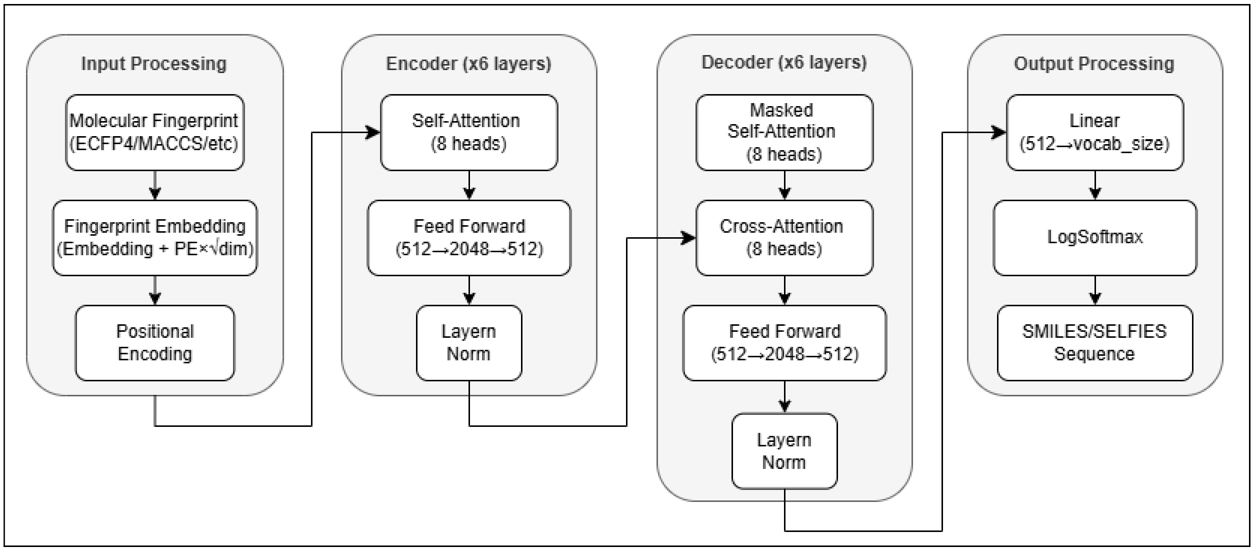

Input Processing
Molecular Fingerprint (ECFP4/MACCS/etc)
Fingerprint Embedding (Embedding + PE×√dim)
Positional Encoding
Encoder (x6 layers)
Self-Attention (8 heads)
Feed Forward (512→2048→512)
Layern Norm
Decoder (x6 layers)
Masked Self-Attention (8 heads)
Cross-Attention (8 heads)
Feed Forward (512→2048→512)
Layern Norm
Output Processing
Linear (512→vocab_size)
LogSoftmax
SMILES/SELFIES Sequence

Figure 1: MolForge Architecture Summary.

4 Methodology

This study analyzed a selection of COX-2 inhibitors: anitrazafen, celecoxib, celecoxib, deracoxib, and parecoxib to investigate how ECFP4 fingerprint bit flipping affects molecular similarity. We aimed to identify structurally similar molecules generated from small and large changes in ECFP4 fingerprint bit vectors. This methodology comprised five essential steps.

First, we generated Extended-Connectivity Fingerprint (ECFP4) representations for each molecule. We converted their SMILES strings into 2048-bit fingerprint vectors with a radius of 2, capturing a simplified version of each molecule's structural features. These fingerprints served as the basis for our bit-flipping analysis.

Next, we systematically altered the fingerprints by flipping bits in increments. We iterated through all 2048 bits for each fingerprint, flipping one bit at a time. We repeated this process for flips of 2, 4, 8, 128, and finally 1024 bits per ECFP4 bit vector. It is important to highlight that the flipping is not cumulative. In other words, each new vector only contains 1, 2, 4, 8, 128, or 1024 flipped bits while keeping the rest as the original molecule. This approach allowed us to produce a range of modified fingerprints for each molecule, each with a different degree of structural perturbation.

After generating these modified fingerprints, we input each into MolForge to predict a new SMILES string. Each newly generated SMILES represented a potential novel molecule with altered structural features. After confirming the validity of the derived SMILES (using RDKit), we recalculated the ECFP4 fingerprints for these to determine their similarity to the original molecules.

We calculated the Tanimoto similarity score for each pair of original and generated fingerprints to evaluate the similarity between the original and modified molecules. This score indicated how altered fingerprints resembled their originals, providing a molecular similarity measure for each variant generated through bit flipping.

Additionally, we visualized the unique molecules generated by MolForge. We created a multidimensional scaling (MDS) plot to represent the Tanimoto similarity between the generated and original SMILES in two dimensions. This visualization illustrated the spatial distribution of generated molecules relative to the originals, offering insights into molecular proximity based on fingerprint modifications.

Finally, we analyzed the molecular properties of the generated compounds, including molecular weight and atom count, to understand how structural complexity changes with increasing bit flips. We calculated these properties using RDKit for each valid SMILES and tracked their distribution across different bit flip levels. We also systematically documented the number of valid and unique SMILES generated at each bit flip level and cross-referenced selected structures against the PubChem database to identify known compounds.

Each step was conducted independently for bit flips of 1, 2, 4, 8, 128, and 1024 to ensure a comprehensive analysis across a broad range of fingerprint modifications.

5 Results and Discussion

5.1 Valid and Invalid SMILES

Table 1 shows the number of chemically valid SMILES (validated using RDKit) generated for five COX-2 inhibitors when varying numbers of bits are flipped in their molecular fingerprints. The last row ("Expected # SMILES") represents the theoretical maximum number of new SMILES that could be generated based on the bit flip combinations. For example, when flipping 1 bit in a 2048-bit fingerprint, we expect up to 2048 new unique fingerprints, and in this case, all generated fingerprints successfully translate to valid SMILES structures. As more bits are flipped simultaneously (e.g., 2, 4, 8 bits), the theoretical expectations and the number of valid SMILES decrease proportionally.

The results show a clear pattern: as simultaneously flipped bits increase, the number of valid SMILES decreases more significantly than expected. This trend becomes particularly noticeable at higher bit-flip counts. For instance, when flipping 8 bits simultaneously, Anitrazafen generates only 206 valid SMILES compared to the expected 256, representing 19.5% of invalid SMILES (Table 2). The effect becomes even more significant when flipping 128 bits. In this case, Celecoxib generates only eight valid SMILES, while Anitrazafen produces 7. This is compared to the expected 16 valid SMILES, leading to 50.0% invalid SMILES for Celecoxib and 56.3% for Anitrazafen. When flipping 1024 bits, half of the fingerprint, valid SMILES are nearly nonexistent across all compounds. Only Cimicoxib generates a single valid structure, resulting in 50% invalid SMILES. All other compounds display 100% invalid SMILES. This suggests that extensive modifications to the fingerprint typically lead to chemically invalid structures, with the percentage of invalid SMILES increasing dramatically as more bits are flipped simultaneously.

5.2 Tanimoto Similarity

Table 3 shows the average Tanimoto similarity between the generated SMILES and the original SMILES for five COX-2 inhibitors at different bit flip levels of their ECFP4 fingerprint vectors. The similarity per-

Compound	1 bit	2 bits	4 bits	8 bits	128 bits	1024 bits
Parecoxib	2048	1024	512	255	12	0
Celecoxib	2048	1024	512	256	8	0
Cimicoxib	2048	1024	512	256	11	1
Deracoxib	2047	1023	511	256	13	0
Anitrazafen	2048	1024	508	206	7	0
Expected # SMILES	2048	1024	512	256	16	2

Table 1: Number of valid SMILES (validated using RDKit) generated through bit flipping of molecular fingerprints for COX-2 inhibitors. The bottom row shows the expected number of valid SMILES for each bit flip scenario.

Compound	1 bit	2 bits	4 bits	8 bits	128 bits	1024 bits
Parecoxib	0.00%	0.00%	0.00%	0.39%	25.00%	100.00%
Celecoxib	0.00%	0.00%	0.00%	0.00%	50.00%	100.00%
Cimicoxib	0.00%	0.00%	0.00%	0.00%	31.25%	50.00%
Deracoxib	0.00%	0.10%	0.20%	0.00%	18.75%	100.00%
Anitrazafen	0.00%	0.00%	0.78%	19.53%	56.25%	100.00%

Table 2: Percentage of invalid SMILES generated through bit flipping operations for COX-2 inhibitors. Values represent the proportion of chemically invalid structures obtained when compared to the expected number of SMILES for each bit flip scenario.

centages reflect how incremental changes to the fingerprint affect molecular similarity.

The results indicate that minor changes, such as one or 2-bit flips, yield very high Tanimoto similarity values above 99%, suggesting a minimal structural impact. As the number of bit flips increases, Tanimoto similarity values decrease, with significant drops at 128 and 1024-bit flips, indicating greater structural differences. This trend suggests that small bit flips maintain the molecular identity, while increased flips generate more structurally diverse molecules. The molecules with four or 8-bit changes could be interesting to explore for novel chemical spaces.

Compound	1 bit	2 bits	4 bits	8 bits	128 bits	1024 bits
Parecoxib	99.96%	99.92%	99.72%	88.73%	30.07%	N/A
Celecoxib	99.99%	99.98%	99.88%	86.85%	15.37%	N/A
Cimicoxib	99.98%	99.97%	99.92%	98.37%	23.75%	5.22%
Deracoxib	99.99%	99.99%	99.86%	95.86%	17.20%	N/A
Anitrazafen	99.78%	99.19%	78.94%	64.72%	27.91%	N/A

Table 3: Average Tanimoto Similarity between original molecules and their bit-flipped variants. For each compound, values represent the mean Tanimoto similarity calculated between the original molecule and all valid SMILES generated through bit flipping. N/A indicates scenarios where no valid SMILES were generated.

5.3 Unique SMILES

Table 4 outlines the number of unique SMILES MolForge could generate for each compound at every bit flip level. Demonstrating the diversity in the molecular structures, one can derive from modifications in bits in ECFP4 fingerprint vectors.

Parecoxib and Anitrazafen produce the highest number of unique SMILES, with the greatest variety from Anitrazafen. Potentially relating to the molecular structure of Anitrazafen being more flexible to modification, hence having more structurally distinct nearby molecules.

When only one or two bits are flipped, many SMILES generated are identical to the original SMILES. For example, Table 5 lists the unique SMILES for Anitrazafen with a 1-bit flip given that out of 2048 SMILES generated, 2034 were identical to the original string. That is to say, small changes to the fingerprint vectors frequently lead to little structural change where the resulting molecule is similar to the original one.

The more bits flipped, the more unique are the SMILES. This peaks for most compounds at 8-bit flips. 8-bit flips seem to be an inflection point for most compounds beyond which these modifications create valid yet distinct SMILES from the starting molecule. Going much higher than 8-bit (128 and 1024-bit) flips, the number of valid SMILES decreases substantially. The higher levels of bit flips are generated, but the SMILES tend to be more unique since they portray major structural divergence.

Compound	1 bit	2 bits	4 bits	8 bits	128 bits	1024 bits
Parecoxib	8	8	11	63	12	0
Celecoxib	3	3	5	52	8	0
Cimicoxib	4	3	4	19	11	1
Deracoxib	3	2	5	39	13	0
Anitrazafen	14	24	61	115	7	0
Total SMILES	2048	1024	512	256	16	2

Table 4: Number of unique SMILES generated for each compound at different bit flip levels in the ECFP4 fingerprint vectors. Each cell shows the count of unique SMILES generated by MolForge after flipping the specified number of bits. The ”Total SMILES” row indicates the maximum possible SMILES for each bit flip level, illustrating the proportion of unique structures generated as bit flips increase.

Count	SMILES
2034	COc1ccc(-c2nnc(C)nc2-c2ccc(OC)cc2)cc1
2	COc1ccc(-c2nnc(-c3ccc(OC)cc3)c3nc(C)nnc23)cc1
1	COC1=NC(c2ccc(OC)cc2)=C(c2ccc(OC)cc2)N=C(C)N1
1	COc1ccc(-c2ccc(-c3nnc(C)nc3-c3ccc(OC)cc3)cc2)cc1
1	COc1ccc(-c2nc(C)nc(-c3ccc(OC)cc3)c2C)cc1
1	COc1ccc(-c2nc(C)nc3nc(C)nnc23)cc1
1	COc1ccc(-c2nc(C)nnc2C)cc1
1	COc1ccc(-c2nc3nnc(C)nc3nc2-c2ccc(OC)cc2)cc1
1	COc1ccc(-c2nnc(-c3ccc(OC)cc3)c3c(-c4ccc(OC)cc4)nc(C)nc23)cc1
1	COc1ccc(-c2nnc(C)nc2-c2ccc(-c3nnc(C)nc3-c3ccc(OC)cc3)cc2)cc1
1	COc1ccc(-c2nnc(OC)nc2-c2ccc(OC)cc2)cc1
1	COc1ccc(-c2nnc3c(-c4ccc(OC)cc4)nc(C)nc3c2-c2ccc(OC)cc2)cc1
1	COc1ccc(-c2nnnc(C)n2)cc1
1	COc1ccc(C2=C3N=C(C)N=C3N=N2)cc1

Table 5: Anitrazafen 1-Bit Change: Unique SMILES and Their Occurrences

Table 6 presents a detailed assessment of SMILES strings generated by MolForge across varying bit-flip levels. Each generated SMILES is compared to existing entries in PubChem. The results highlight MolForge's ability to reproduce recognizable molecular structures, especially for well-known compounds like Celecoxib and Anitrazafen. Several matches with PubChem entries confirm that MolForge can replicate known structures at low bit-flip levels.

However, as the bit-flip level increases, fewer SMILES align with known PubChem compounds, suggesting that MolForge explores more novel or proprietary chemical spaces at higher bit-flip settings. In some cases, this may also reflect inaccurate or "hallucinated" outputs by the model. The table shows that with minimal bit changes (1–4 bits), generated SMILES closely match the original molecules, yielding valid matches in PubChem. There are no PubChem matches at higher levels, such as 128 or 1024 bits. This result implies that these structures might represent new compounds or that the model struggles to generalize accurately with high bit changes. Overall, the findings suggest MolForge's potential for novel chemical exploration. Retraining the model with a focus on minimal bit changes may help improve precision and reduce unintended outputs.

Compound	1 bit	2 bits	4 bits	8 bits	128 bits	1024 bits
Anitrazafen	1	4	3	1	0	0
Celecoxib	3	3	3	3	0	0
Cimicoxib	2	1	1	1	0	0
Deracoxib	1	2	2	1	0	0
Parecoxib	2	2	2	1	0	0

Table 6: Number of PubChem entries found for each compound at different bit flip levels in the ECFP4 fingerprint vectors. Each cell shows the count of PubChem entries for unique SMILES generated by MolForge after flipping the specified number of bits.

5.4 Multidimensional Scaling (MDS) Plots

The MDS plots in Figures 4, 5, 6, 7, 8 show the Tanimoto similarity of bit-flipped molecules, represented in two dimensions to highlight their structural differences from the original molecule. Each plot point represents a unique SMILES string generated through bit flipping. The color scale reflects similarity to the original molecule, with yellow points indicating high similarity and purple points indicating lower similarity. The number above each point corresponds to the bit that was flipped.

We observe that even a single bit flip can introduce some structural variation in the Anitrazafen Figure 4 plots. Within the plot, we can see 13 points scattered outside the central cluster. These scattered points suggest that most of the 2048 generated SMILES remain almost identical to the original molecule, while only a few show distinct changes. As the number of flipped bits increases, the spread of points expands, illustrating an exponential rise in structural diversity compared to the original molecule.

With 4 and 8-bit flips, the plots show groupings of molecules with similar structures. This clustering indicates moderate changes produce structurally distinct groups while retaining some molecular similarity. At 128-bit flips, the number of valid molecules drops sharply. This reduction suggests that more extensive modifications often result in invalid SMILES.

Figures 5, 6, 7 show limited structural variation at lower bit flip levels for Celecoxib, Cimicoxib and Deracoxib. For 1, 2, and 4-bit flips, nearly all generated SMILES are highly similar to the original molecule with only 1-4 generated SMILES being different. Compared to Anitrazafen, Celecoxib, Cimicoxib and Deracoxib, small bit changes seem to create more SMILES identical to the original SMILE.

At 8-bit flips, there is a noticeable increase in diversity, with points spread more sporadically and only a few small clusters. Suggesting that moderate bit flips introduce distinct structural changes without forming large groups of similar molecules.

With 128-bit flips, the number of valid SMILES drops significantly. The remaining points are highly unique and structurally distant from the original SMILE, indicating substantial divergence.

In Figure 8, we observe that Parecoxib produces a higher number of unique SMILES even at low bit flip levels. For 1, 2, and 4-bit changes, there are 8, 8, and 11 visually distinct SMILES, respectively. Parecoxib's structure responds more sensitively to minor modifications, resulting in a greater diversity of SMILES than other compounds at these levels. At 8 bits, we see many distinct groups, as indicated by overlapping numbers in certain areas. It further suggests that 8 bits effectively produce structurally varied yet valid SMILES. At 128-bit flips, as with other compounds, the generated SMILES show greater structural differences and a lower count of valid configurations.

In summary, different COX-2 inhibitors exhibit varying levels of structural flexibility in response to bit flips. Anitrazafen and Parecoxib generate the highest number of unique SMILES, with Anitrazafen producing the most overall. Generally, increasing the number of bit flips results in more unique SMILES, although flips of 128 bits or more often lead to invalid or highly divergent structures. Certain bit positions appear frequently in the unique SMILES, suggesting that these bits may be vital in driving molecular changes.

5.5 Molecular Weight and Atom Count Correlation

Figure 2 shows a clear trend: as we increase the number of bits flips in the molecular fingerprints, the resulting SMILES structures tend to have higher molecular weights and more significant atom counts. This increase is likely due to a higher proportion of ”1s” in the 2048-bit ECFP4 fingerprint as more bits are flipped, introducing additional structural features into the generated molecules. Figure 3 further supports this observation, showing a steady rise in the proportion of ”1s” with more flips.

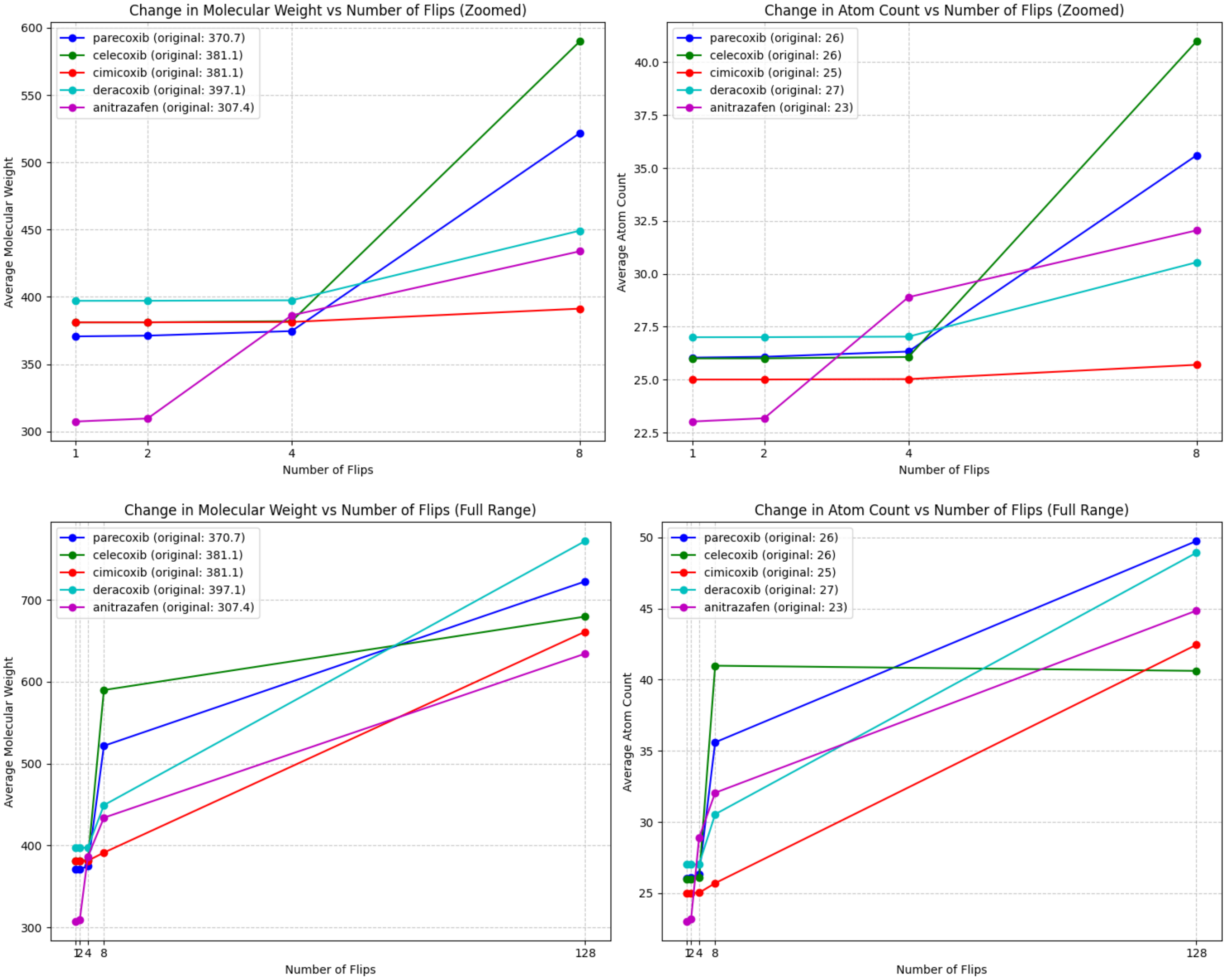


Figure 2: Effect of flip count on molecular weight and atom count for COX2 Inhibitors. Top: Zoomed-in range (1, 2, 4, 8) for detailed examination of early changes. Bottom: Full range of flips (1, 2, 4, 8, 128) showing overall trends.

The left panel of Figure 2 highlights this trend over a smaller range of flips (1, 2, 4, and 8), capturing the initial molecular weight and atom count changes. The right panel, which covers a broader range (1, 2, 4, 8, and 128 flips), illustrates the overall upward trend more clearly as more bits are altered. The increase in molecular weight and atom count also aligns with more extended SMILES strings, suggesting greater structural complexity. The data indicates that more bit flips enhance the complexity of the generated SMILES structures, resulting in molecules with higher molecular weights and atom counts.

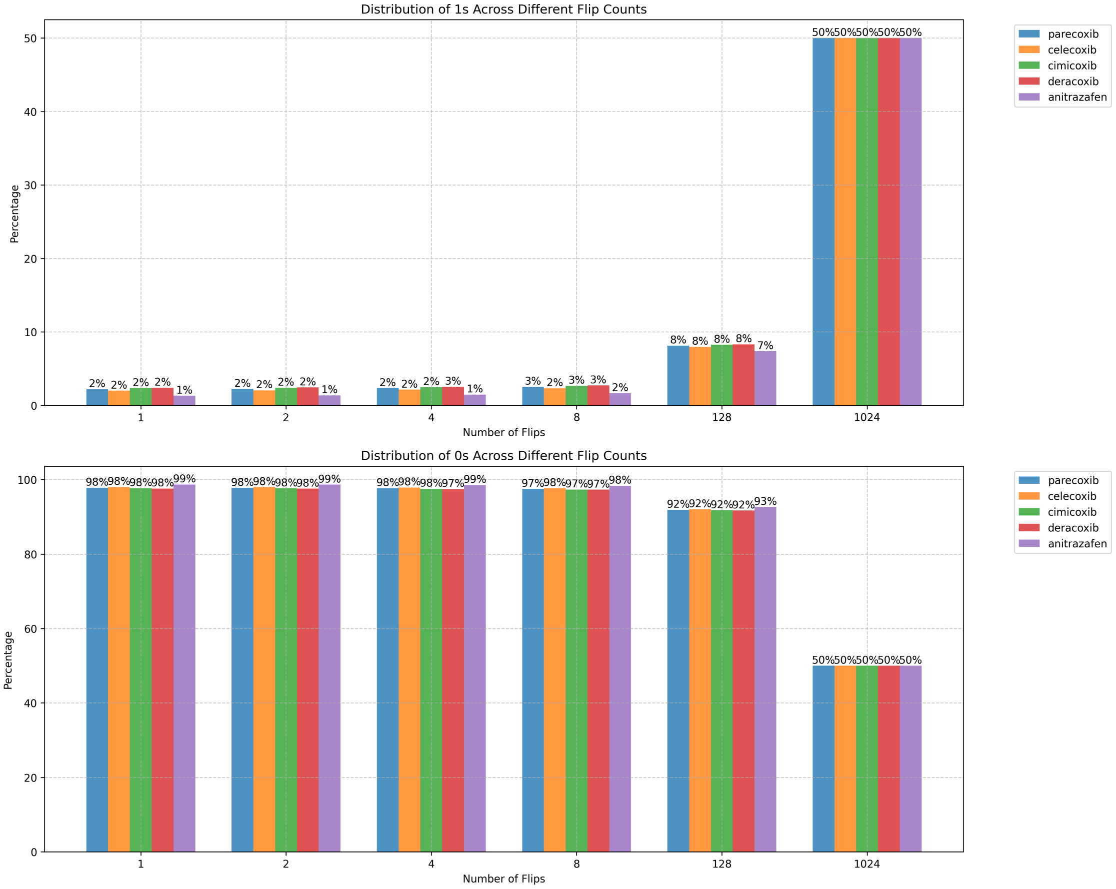


Figure 3: Distribution of 1s and 0s across different flip counts for COX2 inhibitors. The top subplot shows the percentage of 1s, while the bottom subplot shows the percentage of 0s for each flip count.

6 Limitations and Future Work

This study presents several limitations related to the methodology and tools used. First, MolForge was not trained on a dataset designed for generating molecules with minor, one-bit differences between their molecular fingerprints. Instead, it was trained on the standard ChEMBL25 and ChEMBL26 datasets, which do not correlate to our use case. MolForge would need to be retrained on a dataset containing molecules with incremental bit changes to generate SMILES with more subtle variations. Additionally, when comparing the flipped fingerprint with the fingerprint of the generated SMILES, we often found that the expected flipped bit was absent. Further explaining explains why many data points in the MDS plots are located at the same position as the original SMILES, with a Tanimoto similarity of 1.

Another limitation is the small proportion of generated SMILES in the PubChem database. This observation suggests that the generated molecules are often novel or invalid rather than recognized structures. However, this does not necessarily indicate MolForge's poor performance. In fact, when evaluated on the ChEMBL26 dataset, MolForge demonstrated an average Tanimoto similarity of 98.3%, showing high predictive accuracy. The issue likely lies in MolForge's capacity to generate valid SMILES from small-bit changes rather than an inherent flaw in the model itself.

We also encountered challenges with the SMILES representation. SMILES strings are not always unique; different molecular structures can sometimes share the same SMILES string. Identifying molecules with very close structural similarities to the original molecule using single-bit changes was difficult. In many cases, as shown in Figure 2, bit flips resulted in SMILES that were nearly identical to the original structure, suggesting that a single-bit modification may not generate a distinct molecule.

As the number of bit flips increased to 8, 128 and 1024 we observed that the generated SMILES became more distinct from the original structures. However, with more bit changes, especially at 128 and 1024 flips, we encountered SMILES strings that appeared incorrect or implausible when visualized as molecular structures. It suggested that changing too many bits may lead MolForge to generate ”hallucinated” molecules that lack structural validity. Retraining MolForge on a dataset that accommodates more minor bit variations could mitigate this issue.

For future work, retraining MolForge on a dataset curated for bit-flip modifications in molecular fingerprints could improve its accuracy in generating structurally related molecules. Additionally, further research could explore the generative capabilities of ECFP4 bit flipping using a diverse set of molecules beyond COX-2 inhibitors to test the generalizability of our findings. Finally, an alternative approach could involve making minor modifications directly to the original SMILES strings instead of altering the ECFP4 fingerprints. This approach may preserve specific molecular details lost during the fingerprint generation process and could lead to more diverse and valid molecular structures.

7 Conclusion

This study explored generating new SMILES strings by flipping bits in ECFP4 molecular fingerprints and using MolForge's ECFP4-to-SMILES transformer to interpret these modified fingerprints. Our results demonstrate that this approach can produce new SMILES but with varying effectiveness depending on the number of bits flipped. With small numbers of flips, MolForge frequently generates SMILES that are identical or nearly identical to the original molecule, suggesting that it struggles to interpret subtle fingerprint changes. This limitation is likely due to MolForge's training on the general ChEMBL25 and ChEMBL26 datasets, which may not have prepared it to detect and generate minor structural variations.

As the number of bit flips increases, the diversity of generated SMILES grows. However, we observe a corresponding increase in invalid SMILES, especially with more bit changes (e.g., 128 or 1024 bits). This high invalidity rate suggests that extensive fingerprint modifications lead MolForge to generate structurally implausible or ”hallucinated” molecules. Such results indicate a need to modify MolForge's training to handle substantial fingerprint changes better, as the current model struggles to maintain chemical validity when too many bits are altered.

Despite these limitations, our approach demonstrates the potential for exploring structurally similar molecules within a chemical space. This is particularly evident at lower bit-flip levels, such as 1, 2, 4, and 8 bits. The current MolForge setup seems to reach an optimal point with 8-bit flips. At this level, it balances generating structurally diverse and valid SMILES.

Small bit flips for the tested COX-2 inhibitors resulted in valid SMILES that could serve as close structural analogs. This offers potential value for early-stage molecular research. However, retraining MolForge on a dataset specifically designed to capture systematic, small-scale molecular changes will likely be necessary to achieve more reliable results. This would enhance its ability to generate accurate structural neighbors without reverting to the original molecule.

While established methods such as Recurrent Neural Networks (RNNs) [14], Bidirectional Encoder Representations from Transformers (BERT) [15], Variational Autoencoders (VAEs) [16,17,19], Transformers [18], Generative Adversarial Networks (GANs) [19], Graph Neural Networks (GNNs) [22], and Evolutionary Algorithms [20,21] have been widely used to explore chemical space and generate new molecules. Their focus has often been on modifying SMILES strings. This study, in contrast, introduces molecule generation through bit manipulations of ECFP4 fingerprints as an alternative approach.

For future work, retraining MolForge with a dataset tailored for small-bit modifications could enhance its sensitivity to subtle molecular differences. Expanding the bit-flipping approach to include other molecular groups could test its broader applicability beyond COX-2 inhibitors. Investigating alternative architectures beyond transformers may further refine molecule generation. Finally, exploring direct modifications to SMILES strings instead of ECFP4 fingerprints could retain more molecular information, potentially enabling the generation of more diverse and valid molecular structures. These approaches may enhance the utility of bit-flipping as a tool for exploring chemical space.

References

[1] A. E. Bilsland, K. McAulay, R. West, A. Pugliese, and J. Bower, "Automated generation of novel fragments using screening data, a dual SMILES autoencoder, transfer learning and syntax correction," *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 2547–2559, May 2021, doi: 10.1021/acs.jcim.0c01226.

[2] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, "Deep learning for molecular design—a review of the state of the art," *Molecular Systems Design & Engineering*, vol. 4, no. 4, pp. 828–849, Jan. 2019, doi: 10.1039/c9me00039a.

[3] D. Rogers and M. Hahn, "Extended-Connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, Apr. 2010, doi: 10.1021/ci100050t.

[4] P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan, and E. J. Bjerrum, "Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks," *Nature Machine Intelligence*, vol. 2, no. 5, pp. 254–265, May 2020, doi: 10.1038/s42256-020-0174-5.

[5] U. V. Ucak, I. Ashyrmamatov, and J. Lee, "Reconstruction of lossless molecular representations from fingerprints," *Journal of Cheminformatics*, vol. 15, no. 1, Feb. 2023, doi: 10.1186/s13321-023-00693-0.

[6] T. Le, R. Winter, F. Noé, and D.-A. Clevert, “Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures,” *Chemical Science*, vol. 11, no. 38, pp. 10378–10389, Jan. 2020, doi: 10.1039/d0sc03115a.

[7] Y. Wang and J. Bajorath, “Bit silencing in fingerprints enables the derivation of compound Class-Directed similarity metrics,” *Journal of Chemical Information and Modeling*, vol. 48, no. 9, pp. 1754–1759, Aug. 2008, doi: 10.1021/ci8002045.

[8] C. Hawkey, “COX-2 inhibitors,” *The Lancet*, vol. 353, no. 9149, pp. 307–314, Jan. 1999, doi: 10.1016/s0140-6736(98)12154-2.

[9] Z. Ju, M. Li, J. Xu, D. C. Howell, Z. Li, and F.-E. Chen, “Recent development on COX-2 inhibitors as promising anti-inflammatory agents: The past 10 years,” *Acta Pharmaceutica Sinica B*, vol. 12, no. 6, pp. 2790–2807, Jan. 2022, doi: 10.1016/j.apsb.2022.01.002.

[10] H. L. Morgan, “The generation of a Unique Machine Description for Chemical Structures-A technique developed at Chemical Abstracts Service.,” *Journal of Chemical Documentation*, vol. 5, no. 2, pp. 107–113, May 1965, doi: 10.1021/c160017a018.

[11] Z. Ju, M. Li, J. Xu, D. C. Howell, Z. Li, and F. E. Chen, “Recent development on COX-2 inhibitors as promising anti-inflammatory agents: The past 10 years,” *Acta Pharmaceutica Sinica B*, vol. 12, no. 6, pp. 2790–2807, Jun. 2022, doi: 10.1016/j.apsb.2022.01.002.

[12] C. J. Hawkey, “COX-2 inhibitors,” *The Lancet*, vol. 353, no. 9149, pp. 307–314, Jan. 1999, doi: 10.1016/S0140-6736(98)12154-2.

[13] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: https://doi.org/10.1021/ci00057a005.

[14] F. Grisoni, M. Moret, R. Lingwood, and G. Schneider, “Bidirectional Molecule Generation with Recurrent Neural Networks,” *Journal of Chemical Information and Modeling*, vol. 60, no. 3, pp. 1175–1183, Jan. 2020, doi: 10.1021/acs.jcim.9b00943.

[15] Z. Wu et al., “Knowledge-based BERT: a method to extract molecular features like computational chemists,” *Briefings in Bioinformatics*, vol. 23, no. 3, Mar. 2022, doi: 10.1093/bib/bbac131.

[16] R. N. Tazhigulov, J. Schiller, J. Oppenheim, and M. Winston, “Molecular fingerprints for robust and efficient ML-Driven molecular generation,” *arXiv (Cornell University)*, Nov. 16, 2022. doi: 0.48550/arXiv.2211.09086

[17] X.-C. Zhang et al., “Pushing the Boundaries of Molecular Property Prediction for Drug Discovery with Multitask Learning BERT Enhanced by SMILES Enumeration,” *Research*, vol. 2022, Jan. 2022, doi: 10.34133/research.0004.

[18] S. Honda, S. Shi, and H. R. Ueda, “SMILES Transformer: Pre-trained molecular fingerprint for low data drug discovery,” *arXiv (Cornell University)*, Jan. 2019, doi: 10.48550/arxiv.1911.04738.

[19] L. Schoenmaker, O. J. M. Béquignon, W. Jespers, and G. J. P. Van Westen, “UnCorrupt SMILES: a novel approach to de novo design,” *Journal of Cheminformatics*, vol. 15, no. 1, Feb. 2023, doi: 10.1186/s13321-023-00696-x.

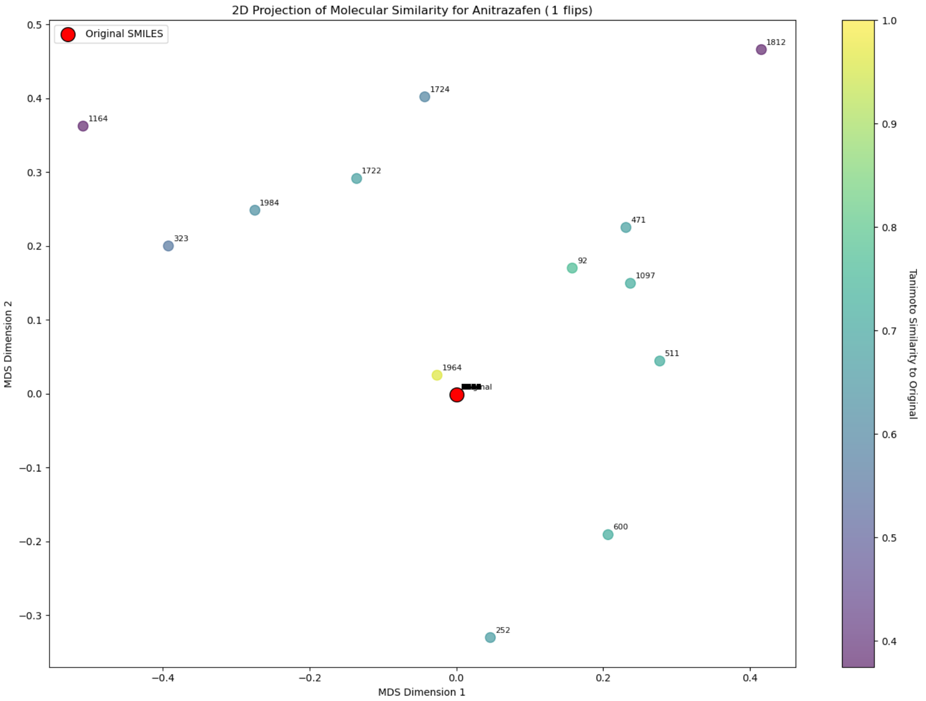
[20] Y. Kwon and J. Lee, “MolFinder: an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using SMILES,” *Journal of Cheminformatics*, vol. 13, no. 1, Mar. 2021, doi: 10.1186/s13321-021-00501-7.

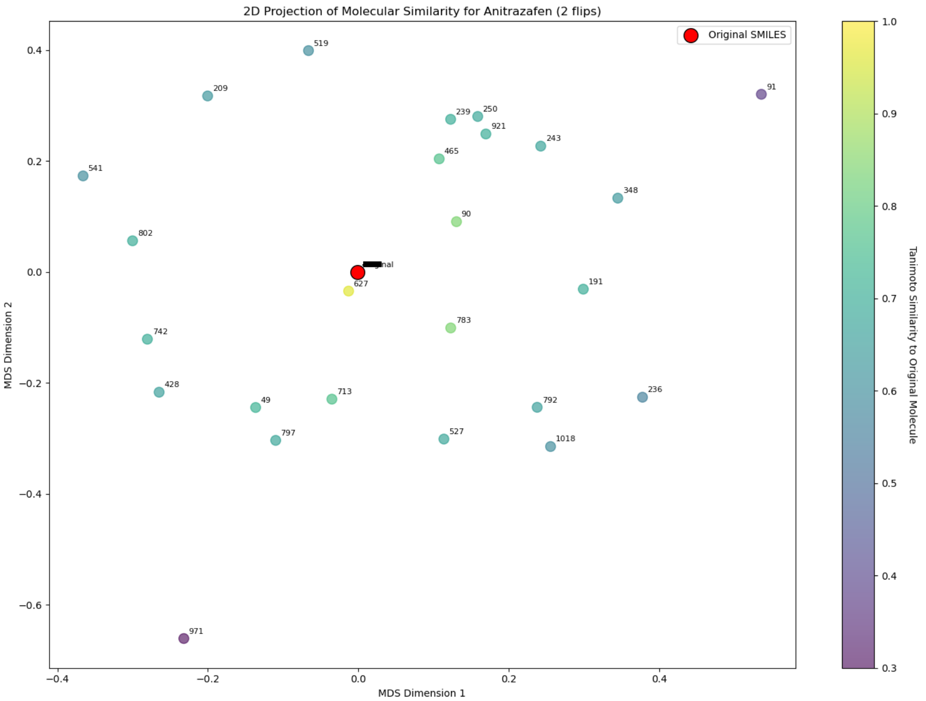
[21] H. Wang et al., “Efficient Evolutionary Search Over Chemical Space with Large Language Models,” *arXiv (Cornell University)*, Jun. 23, 2024. https://arxiv.org/abs/2406.16976

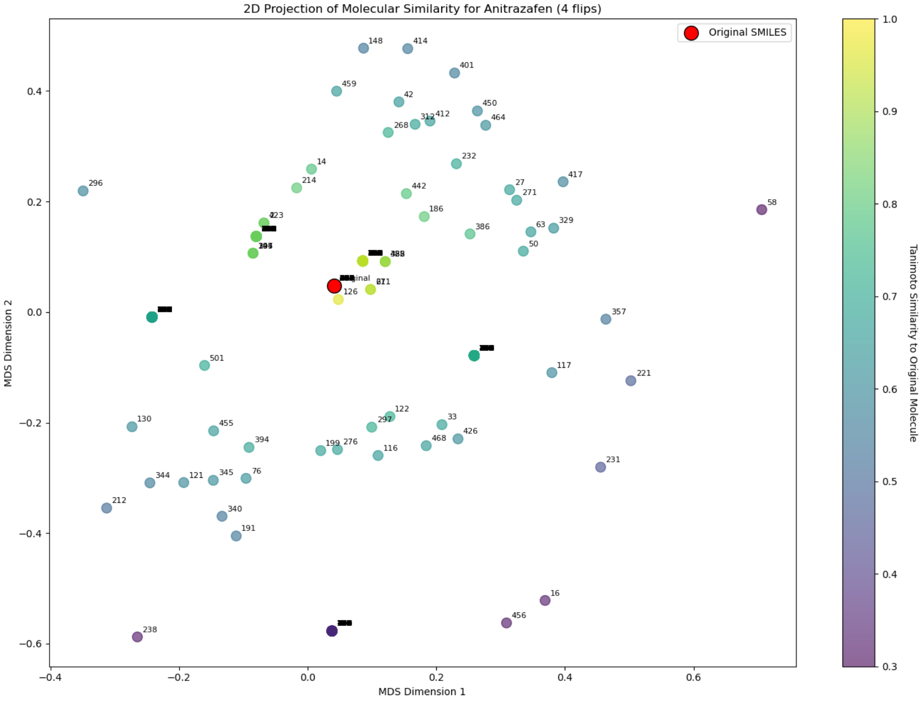
[22] T. Xie and J. C. Grossman, “Hierarchical visualization of materials space with graph convolutional neural networks,” *The Journal of Chemical Physics*, vol. 149, no. 17, Nov. 2018, doi: 10.1063/1.5047803.

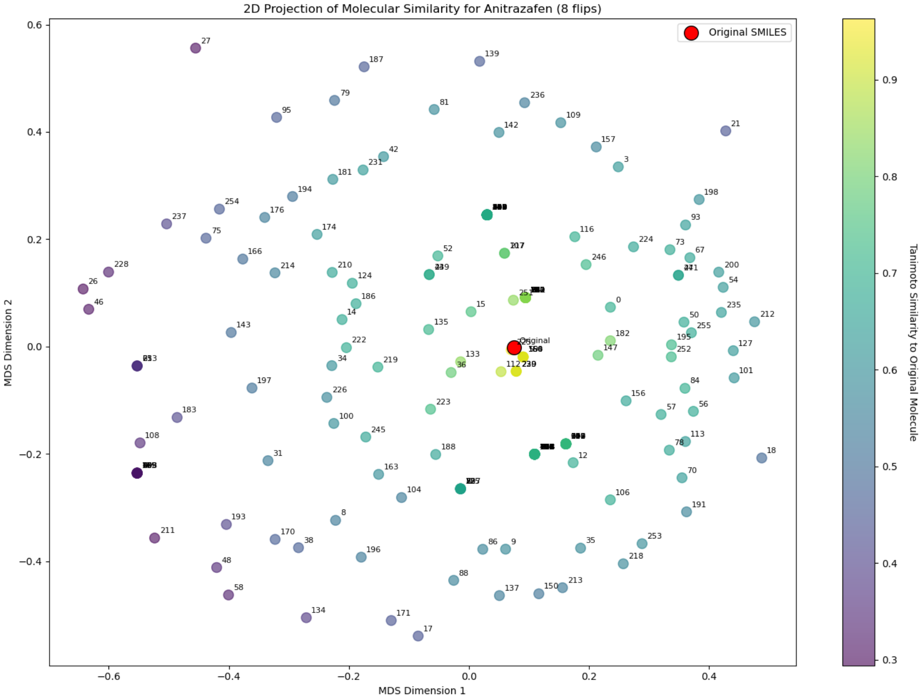
A Appendix

A.1 Anitrazafen






2D Projection of Molecular Similarity for Anitrazafen (4 flips)
Original SMILES
MDS Dimension 1
MDS Dimension 2
Tanimoto Similarity to Original Molecule
−0.4
−0.2
0.0
0.2
0.4
0.6
−0.6
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
148
414
401
459
42
450
312
412
464
268
14
232
296
214
442
27
417
58
271
186
223
386
329
63
50
126
357
501
117
221
122
33
130
455
297
426
394
199
276
468
116
231
344
121
345
76
212
340
191
16
456
238


2D Projection of Molecular Similarity for Anitrazafen (8 flips)
Original SMILES
MDS Dimension 1
MDS Dimension 2
Tanimoto Similarity to Original Molecule
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.3
0.4
0.5
0.6
0.7
0.8
0.9
27
187
139
79
81
236
95
109
142
21
157
42
3
231
181
198
194
254
176
237
174
116
93
75
52
224
73
67
166
246
228
214
210
200
124
54
26
186
15
0
235
46
14
212
143
135
50
255
222
182
195
127
147
252
133
36
112
34
219
197
101
226
84
156
57
56
183
100
223
245
113
108
188
18
31
12
163
78
70
104
106
191
8
193
211
170
38
196
86
9
35
253
218
48
88
213
150
137
58
134
171
17

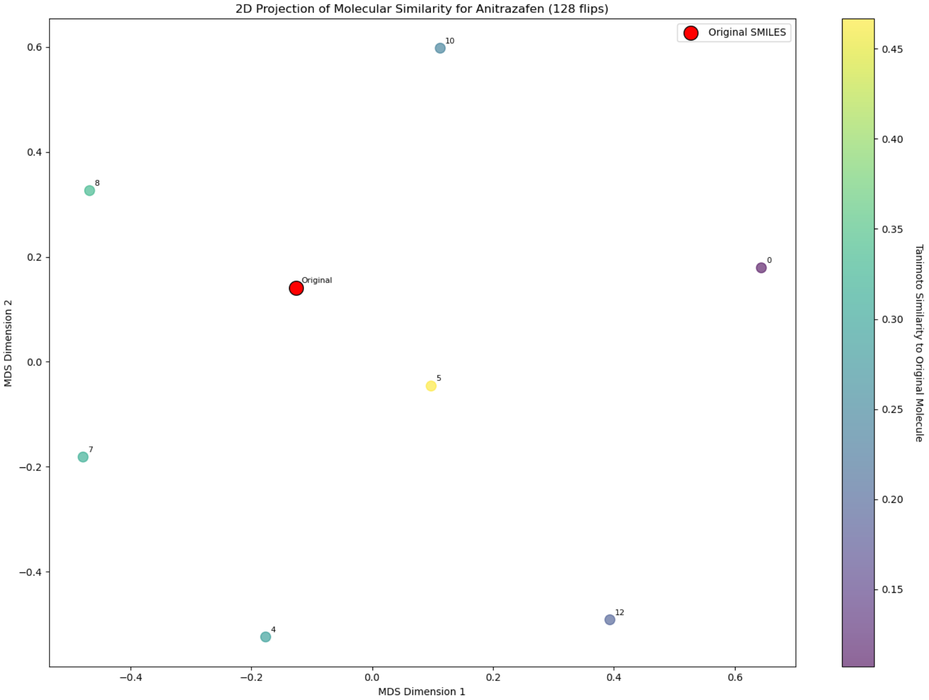
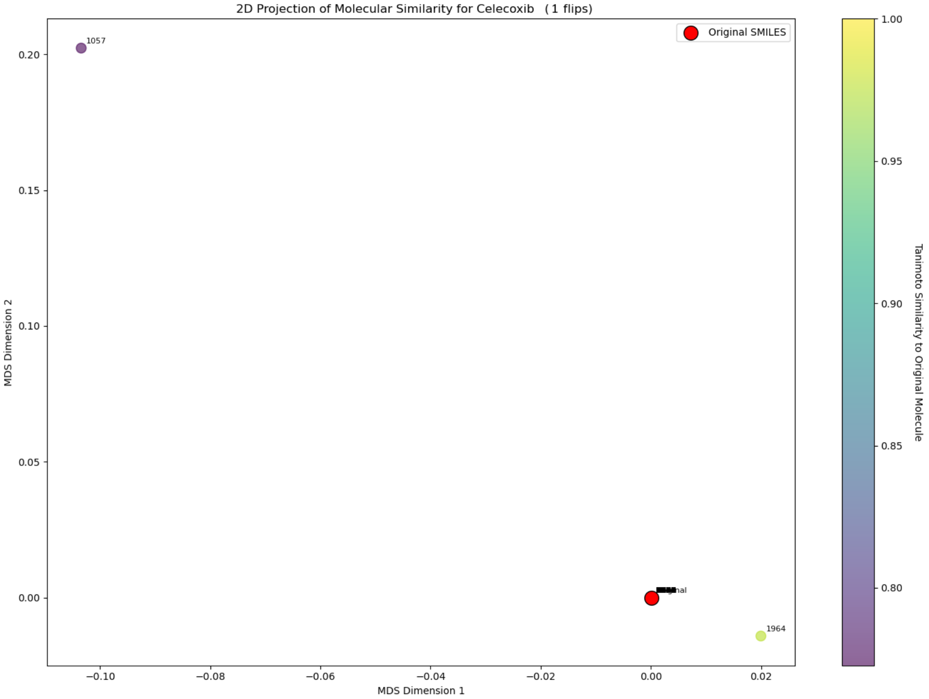
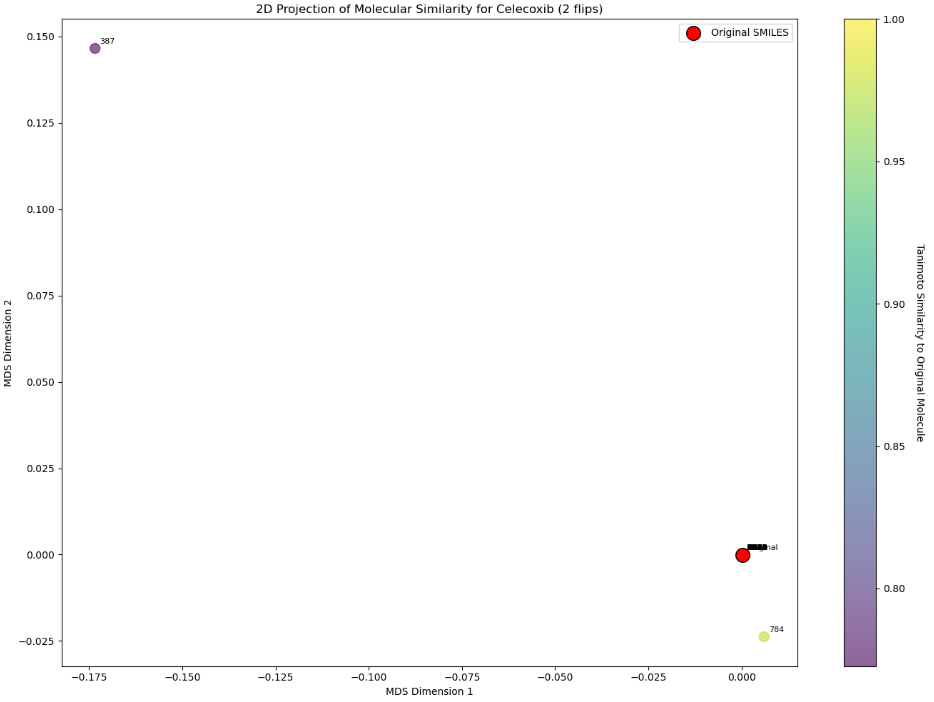
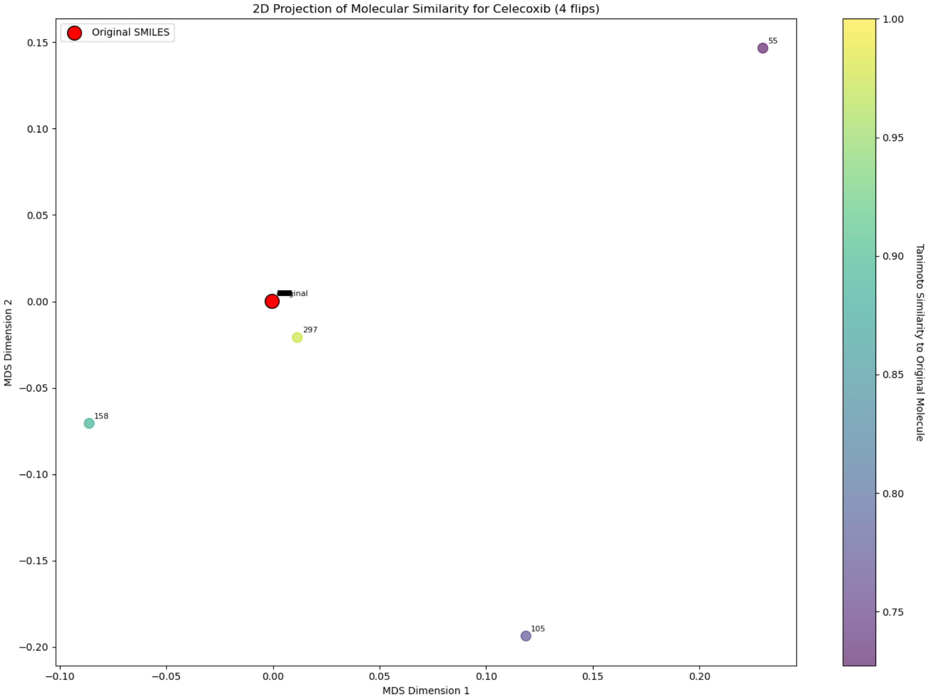


Figure 4: MDS plots for Anitrazafen with varying numbers of bit flips in ECFP4 fingerprints: 1 bit, 2 bits, 4 bits, 8 bits, and 128 bits.

A.2 Celecoxib




2D Projection of Molecular Similarity for Celecoxib (2 flips)
Original SMILES
387
784
MDS Dimension 1
MDS Dimension 2
−0.175
−0.150
−0.125
−0.100
−0.075
−0.050
−0.025
0.000
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
Tanimoto Similarity to Original Molecule
1.00
0.95
0.90
0.85
0.80


2D Projection of Molecular Similarity for Celecoxib (4 flips)
Original SMILES
55
297
158
105
MDS Dimension 1
MDS Dimension 2
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Tanimoto Similarity to Original Molecule
1.00
0.95
0.90
0.85
0.80
0.75

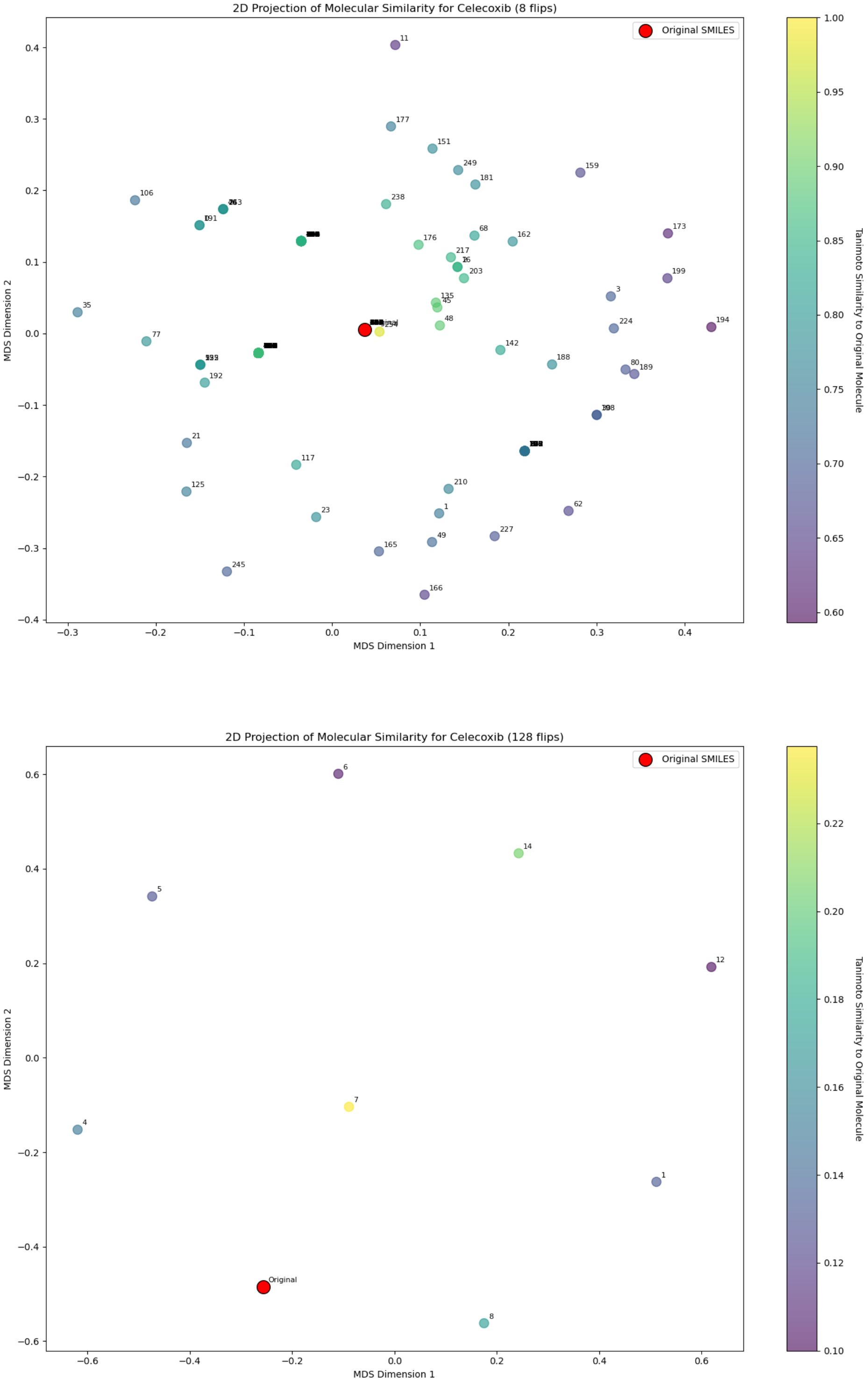
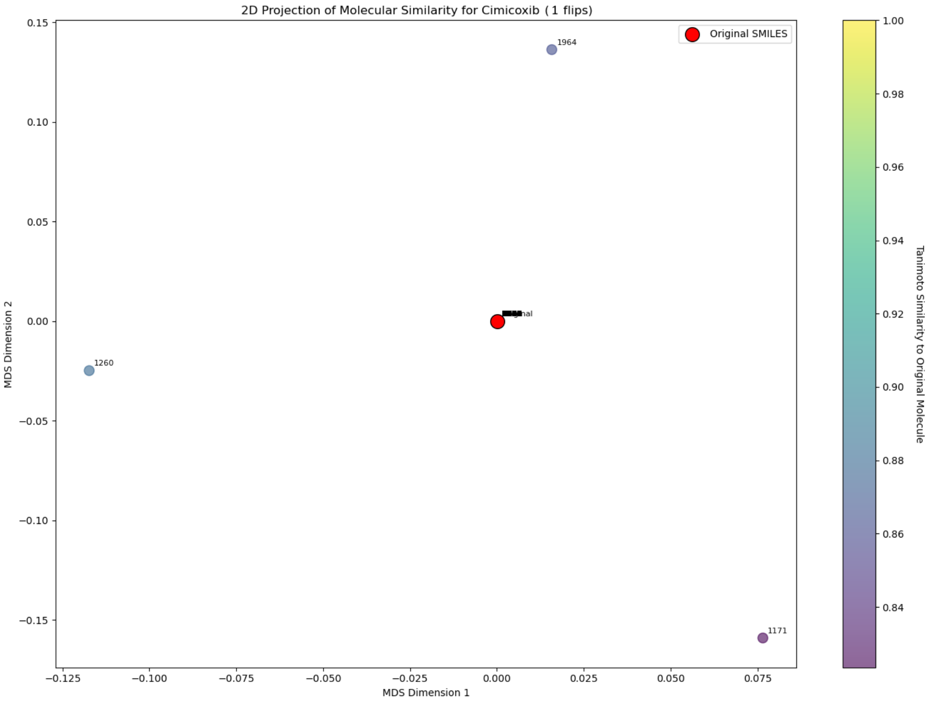
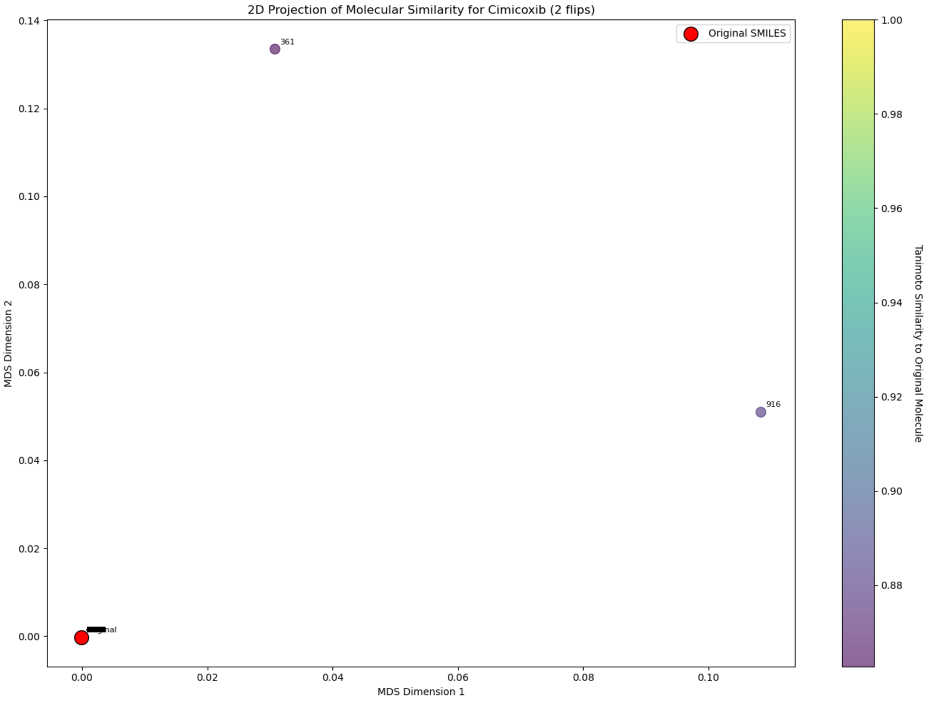
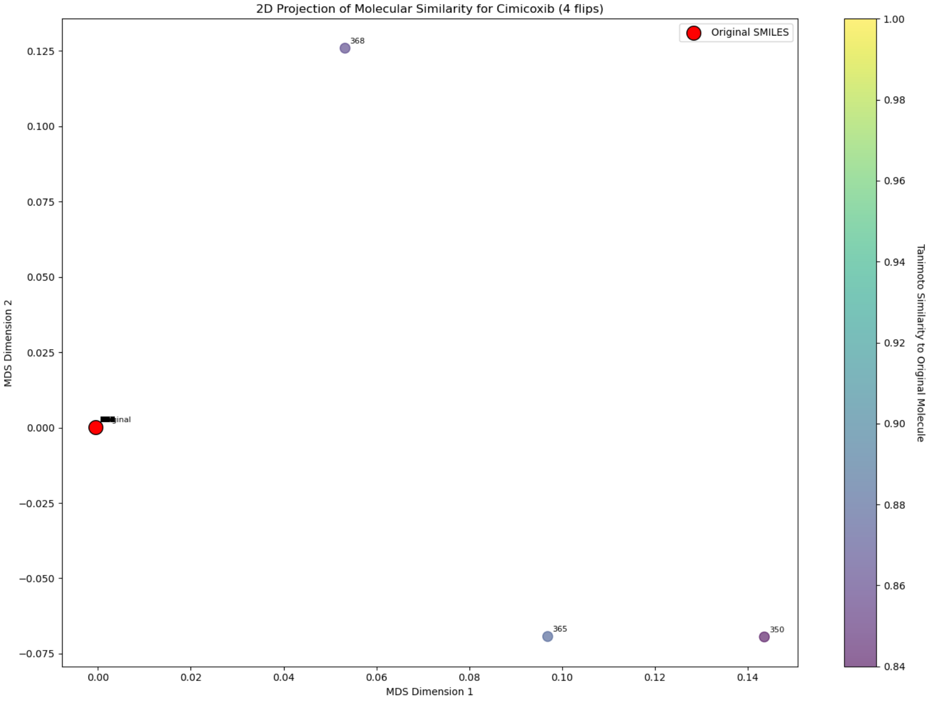


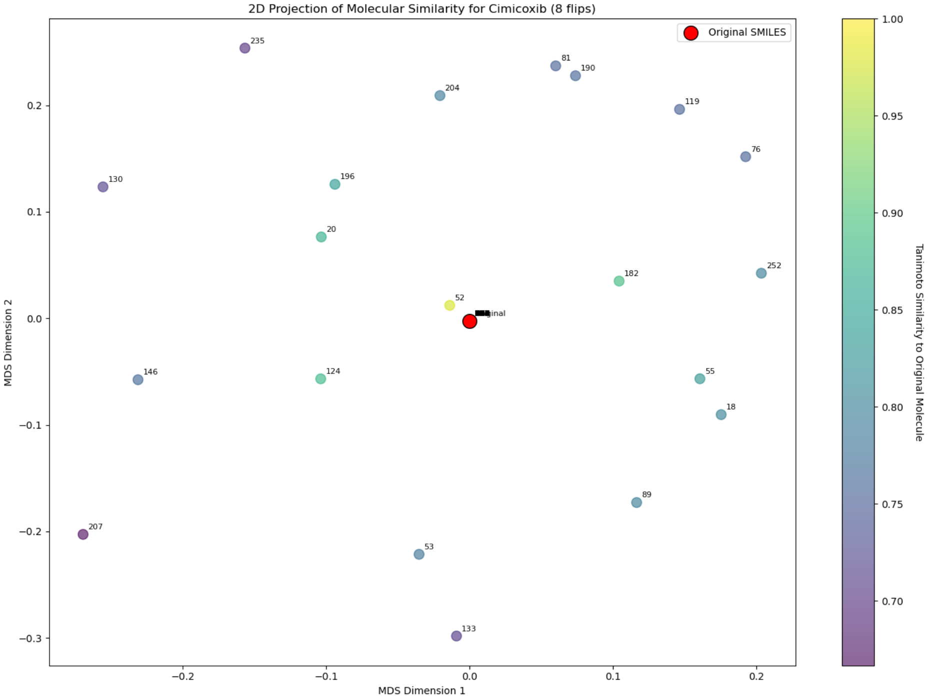
Figure 5: MDS plots for Celecoxib with varying numbers of bit flips in ECFP4 fingerprints: 1 bit, 2 bits, 4 bits, 8 bits, and 128 bits.

A.3 Cimicoxib






2D Projection of Molecular Similarity for Cimicoxib (4 flips)
Original SMILES
368
[redacted]ginal
365
350
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
MDS Dimension 2
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
MDS Dimension 1
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
Tanimoto Similarity to Original Molecule


2D Projection of Molecular Similarity for Cimicoxib (8 flips)
Original SMILES
235
81
190
204
119
76
130
196
20
182
252
52
[redacted]ginal
146
124
55
18
89
207
53
133
0.2
0.1
0.0
−0.1
−0.2
−0.3
MDS Dimension 2
−0.2
−0.1
0.0
0.1
0.2
MDS Dimension 1
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Tanimoto Similarity to Original Molecule

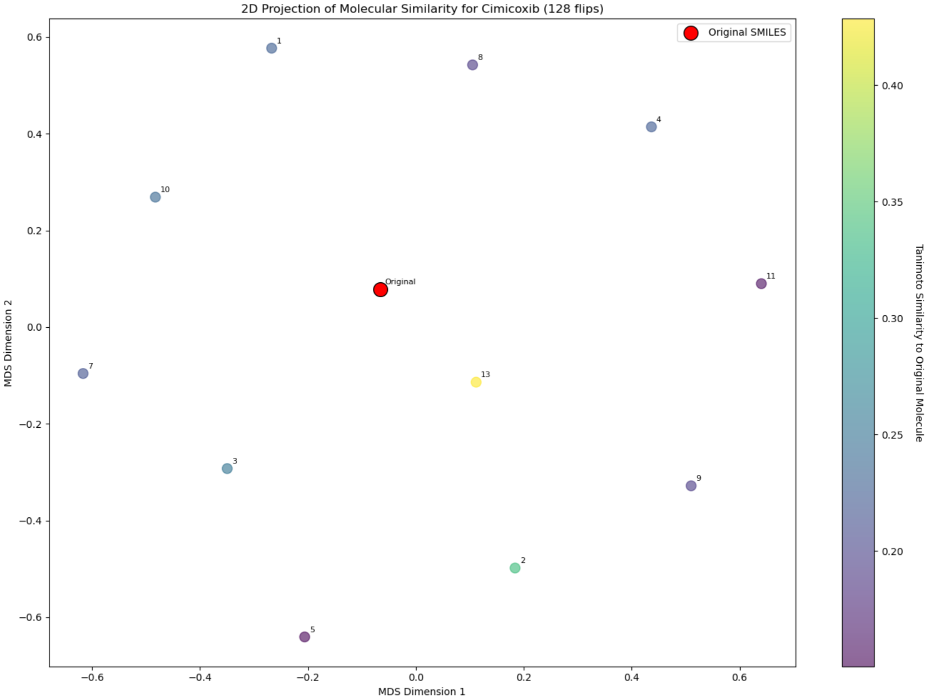
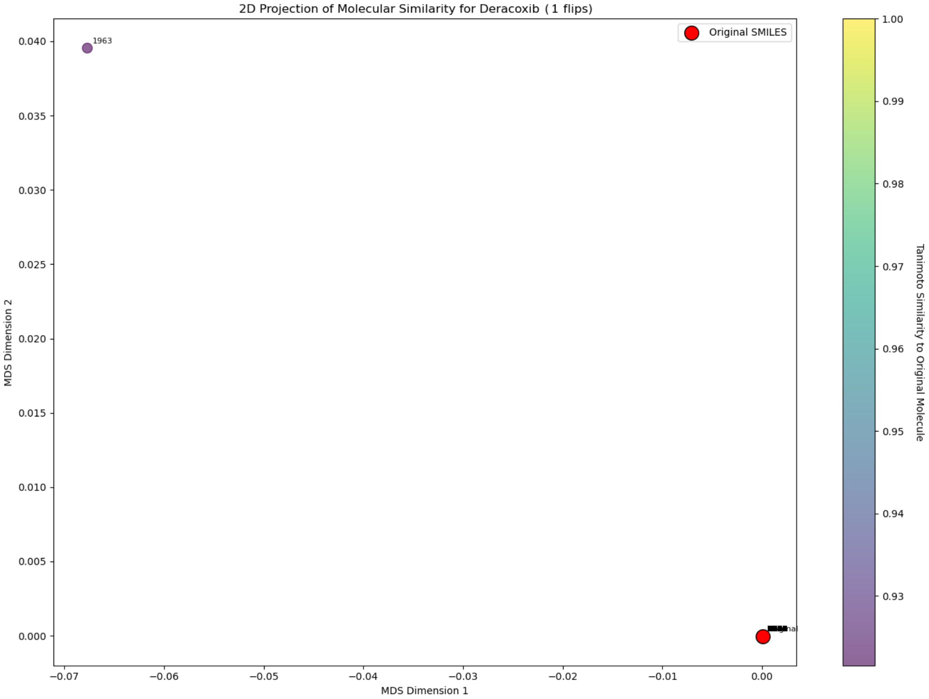
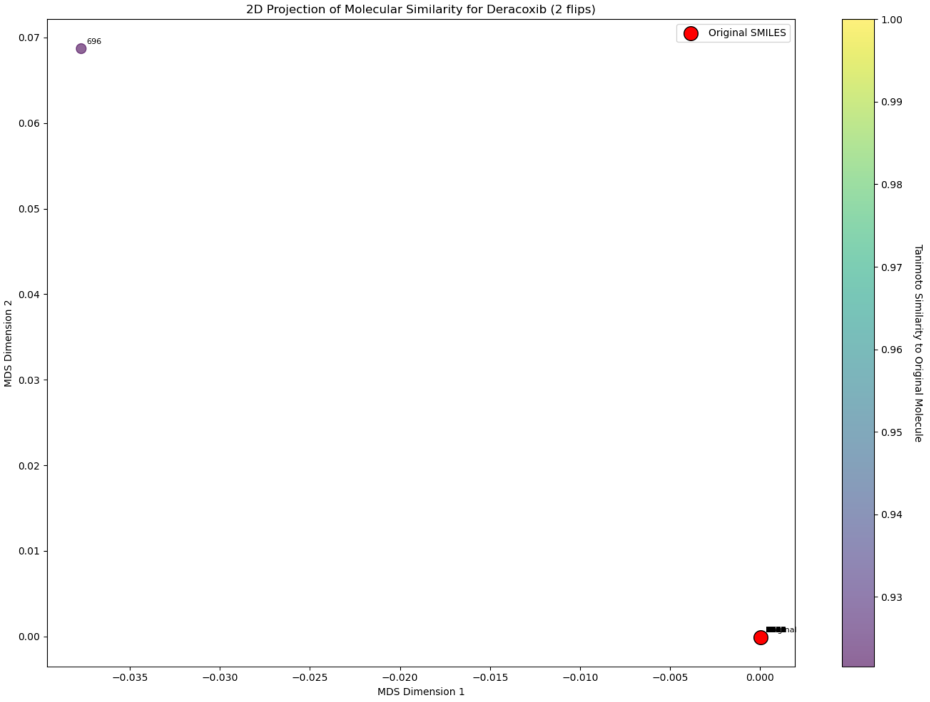
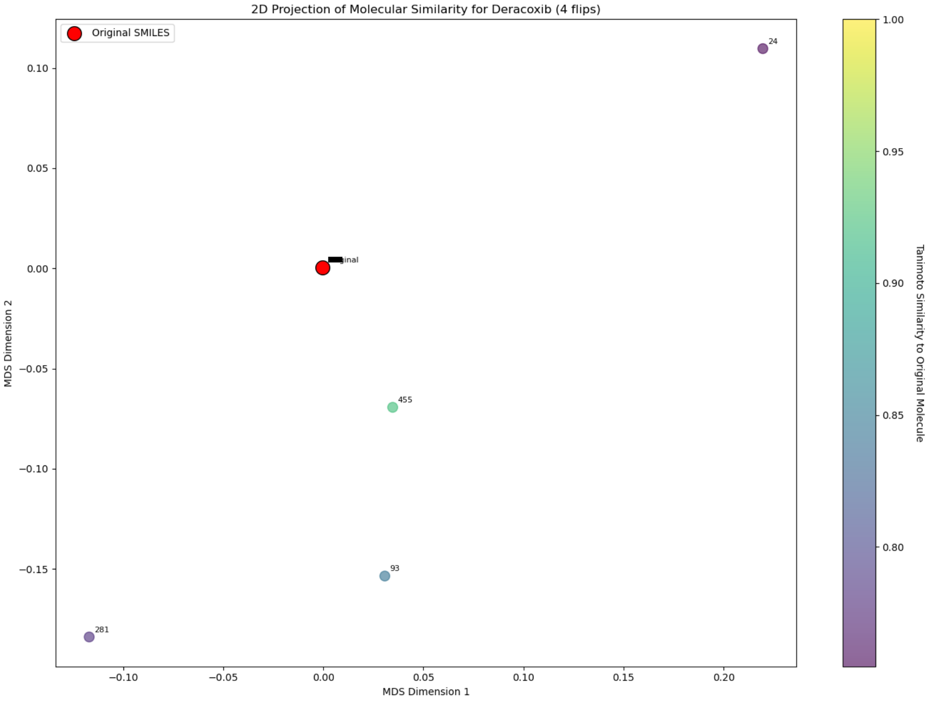


Figure 6: MDS plots for Cimicoxib with varying numbers of bit flips in ECFP4 fingerprints: 1 bit, 2 bits, 4 bits, 8 bits, and 128 bits.

A.4 Deracoxib




2D Projection of Molecular Similarity for Deracoxib (2 flips)
Original SMILES
696
Original
MDS Dimension 1
MDS Dimension 2
−0.035
−0.030
−0.025
−0.020
−0.015
−0.010
−0.005
0.000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Tanimoto Similarity to Original Molecule
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1.00


2D Projection of Molecular Similarity for Deracoxib (4 flips)
Original SMILES
24
Original
455
93
281
MDS Dimension 1
MDS Dimension 2
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.10
0.05
0.00
−0.05
−0.10
−0.15
Tanimoto Similarity to Original Molecule
0.80
0.85
0.90
0.95
1.00

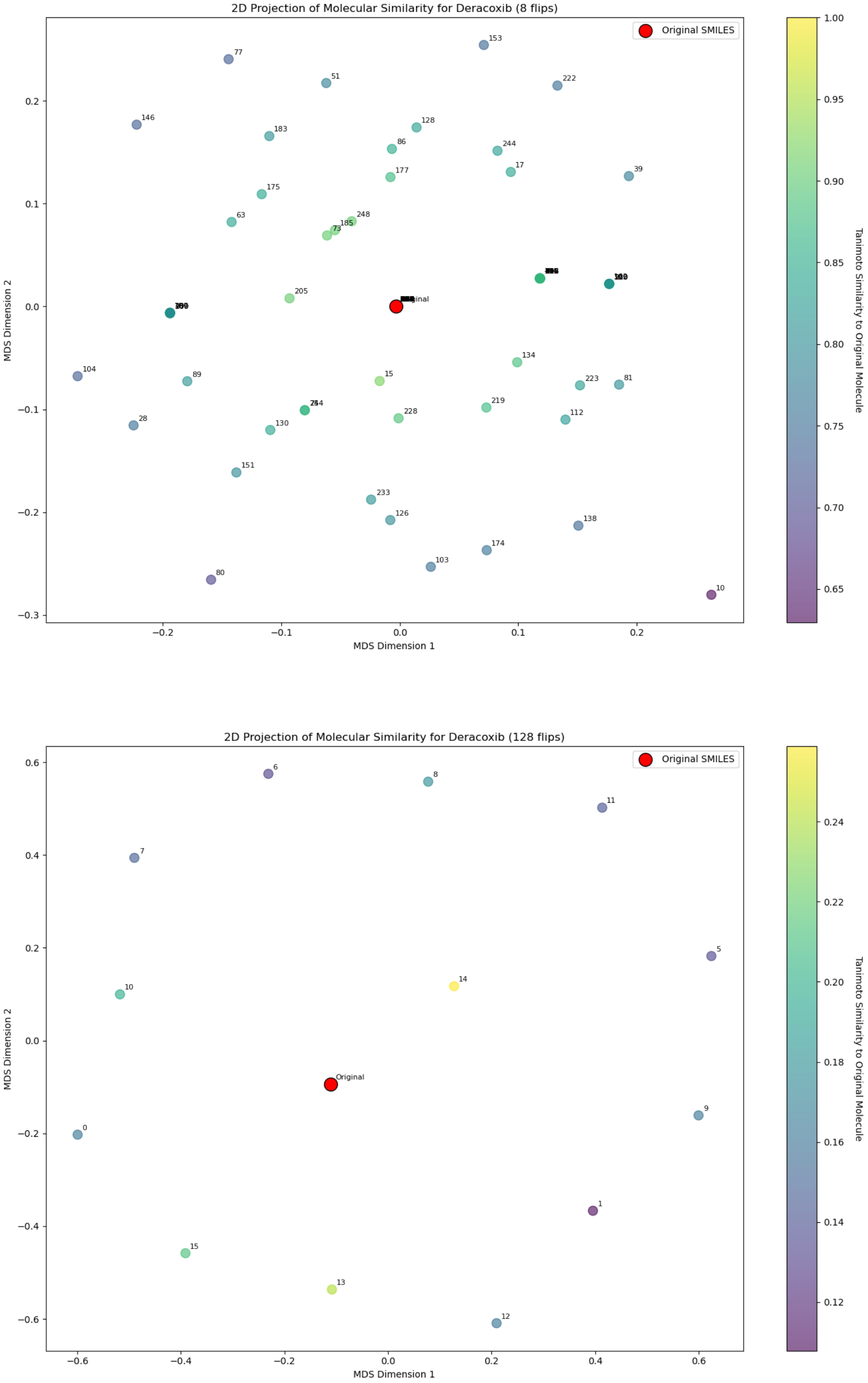
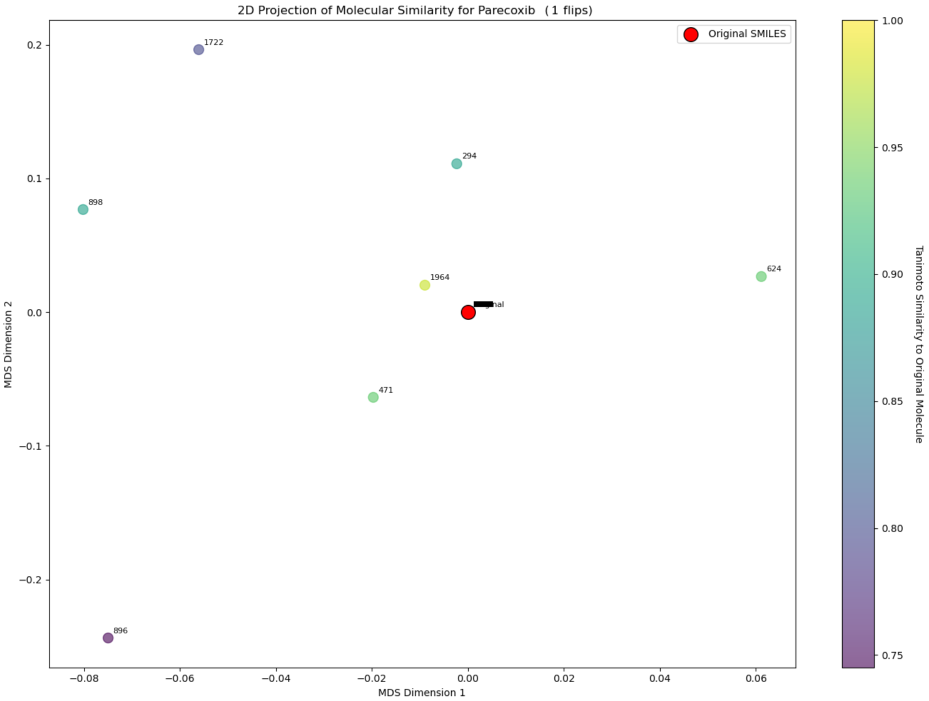
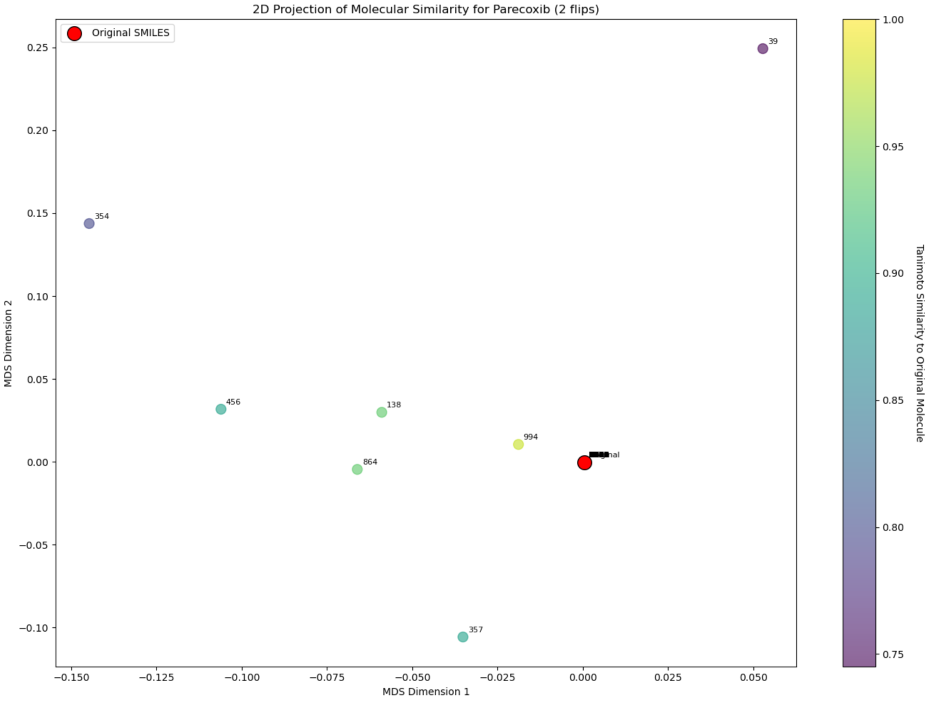
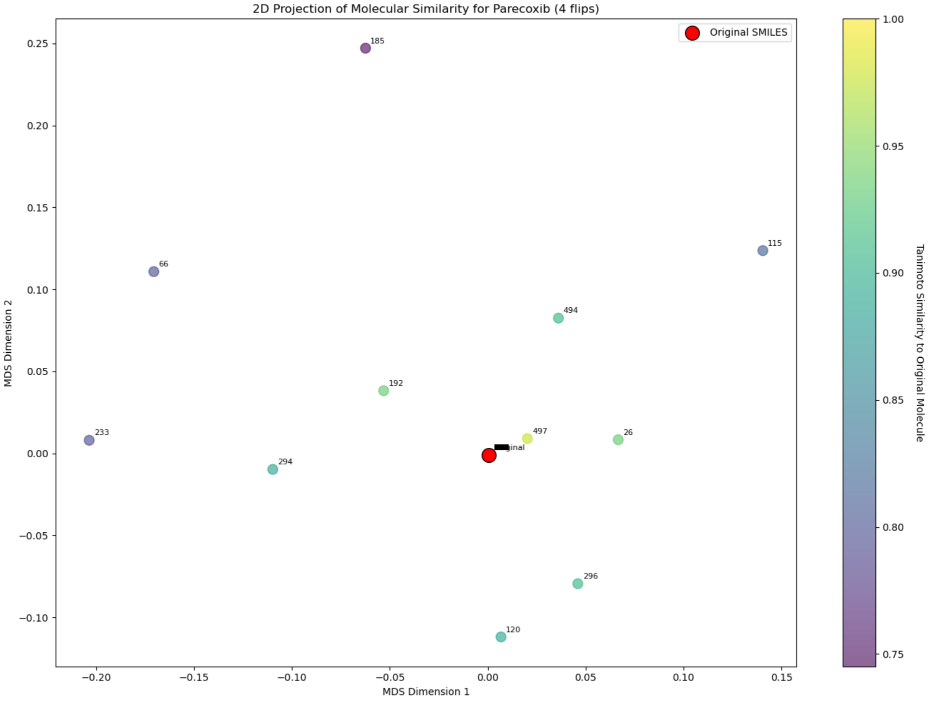


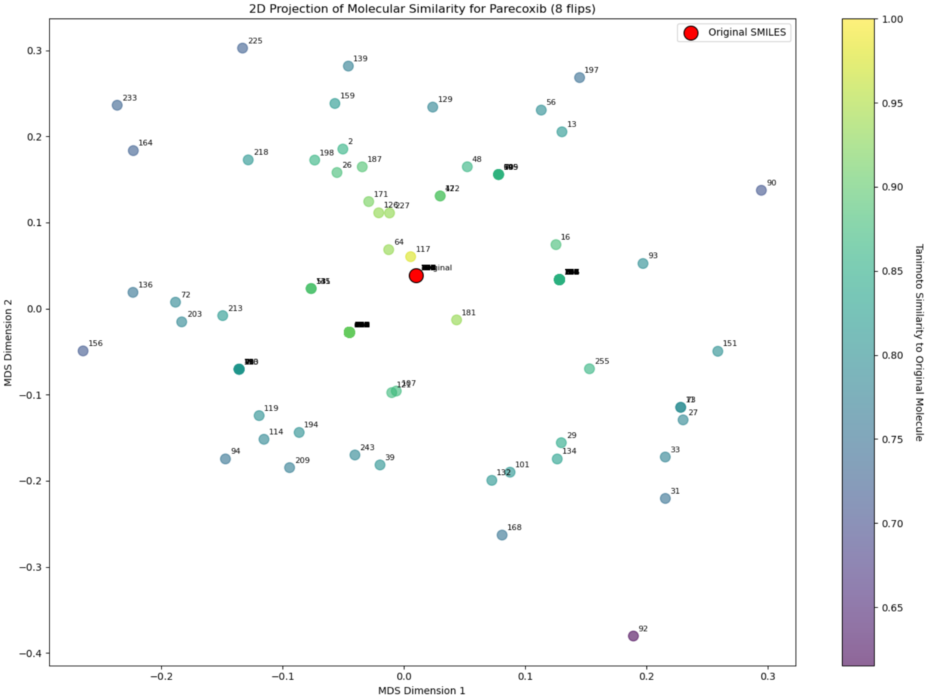
Figure 7: MDS plots for Deracoxib with varying numbers of bit flips in ECFP4 fingerprints: 1 bit, 2 bits, 4 bits, 8 bits, and 128 bits.

A.5 Parecoxib






2D Projection of Molecular Similarity for Parecoxib (4 flips)
Original SMILES
185
66
115
494
192
233
294
497
26
Original
296
120
MDS Dimension 1
MDS Dimension 2
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Tanimoto Similarity to Original Molecule
1.00
0.95
0.90
0.85
0.80
0.75


2D Projection of Molecular Similarity for Parecoxib (8 flips)
Original SMILES
225
139
197
233
159
129
56
13
164
218
198
2
187
48
26
90
171
122
126
227
16
64
117
93
136
72
203
213
Original
181
156
151
255
119
114
194
94
209
243
39
29
134
33
27
101
132
31
168
92
MDS Dimension 1
MDS Dimension 2
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.2
−0.1
0.0
0.1
0.2
0.3
Tanimoto Similarity to Original Molecule
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65

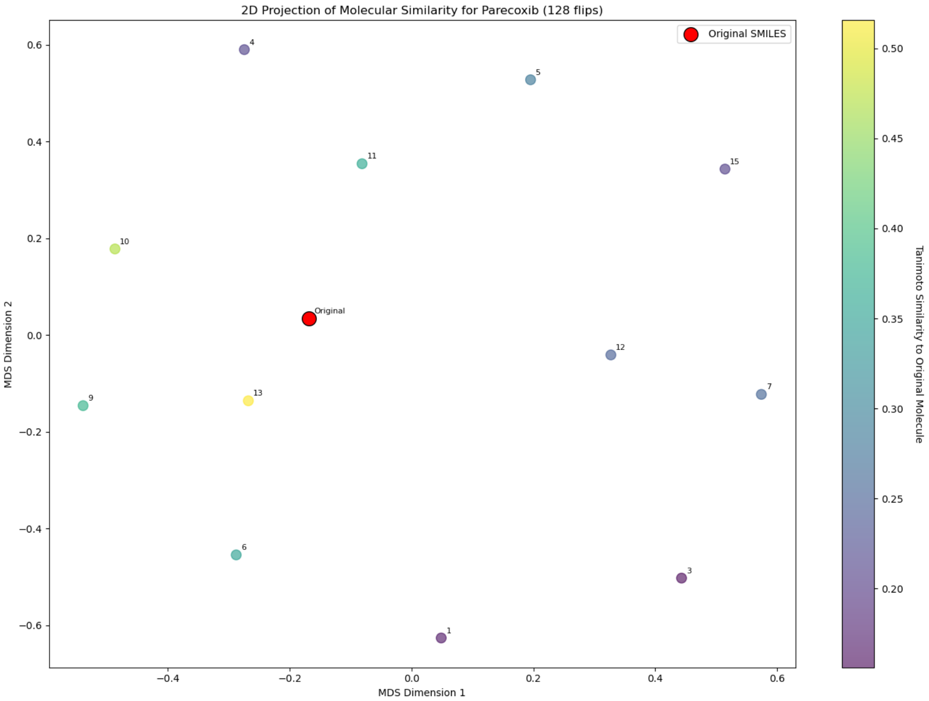


Figure 8: MDS plots for Parecoxib with varying numbers of bit flips in ECFP4 fingerprints: 1 bit, 2 bits, 4 bits, 8 bits, and 128 bits.