Identifying Nearby Molecules through Fingerprint Bit Flipping

Final Presentation

Nils Dunlop, Francisco Erazo, Qi Chen
Supervisor: Shirin Tavara, Christian Tyrchan

Agenda

1. Terminology
2. MolForge
3. Research problems 1 & 2
4. Literature review
5. Methodology research problem 1
6. Results for research problem 1
    - Valid, invalid and unique SMILES
    - 1 Bit flipping
    - 2 Bit flipping
    - 4 Bit flipping
    - 8 Bit flipping
    - Relation to Atoms and Molecular weight
7. Methodology for research problem 2
8. Results for research problem 2
9. Conclusions

Terminology

- SMILES = Cn1c(=O)c2c(ncn2C)n(C)c1=O
- Fingerprints (ECFP4) = [0 0 1 ... 0 1 0]
- Tanimoto Similarity

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

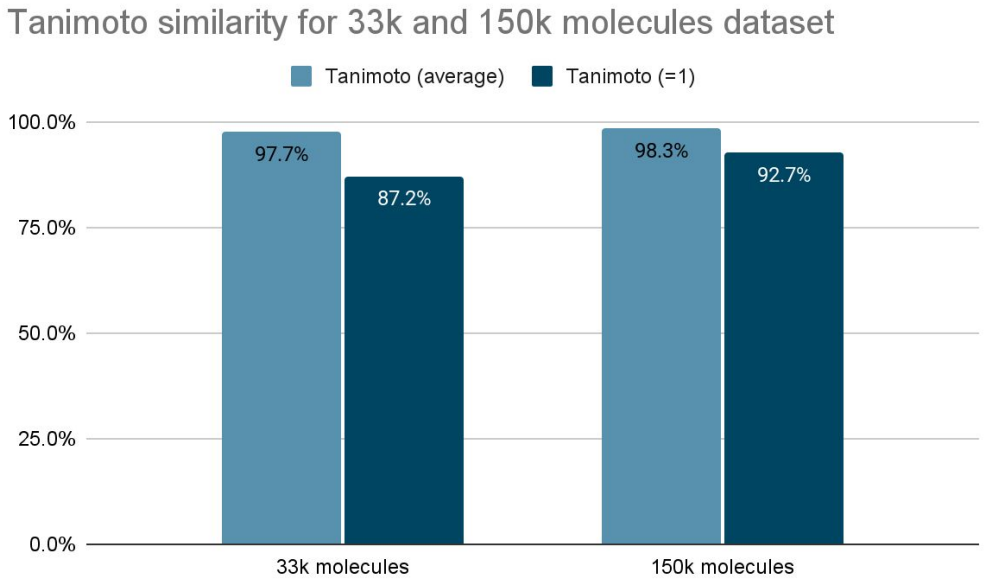
- MutiDimensional Scaling (MDS) = Dimensionality reduction technique that preserves distances between points.

Terminology

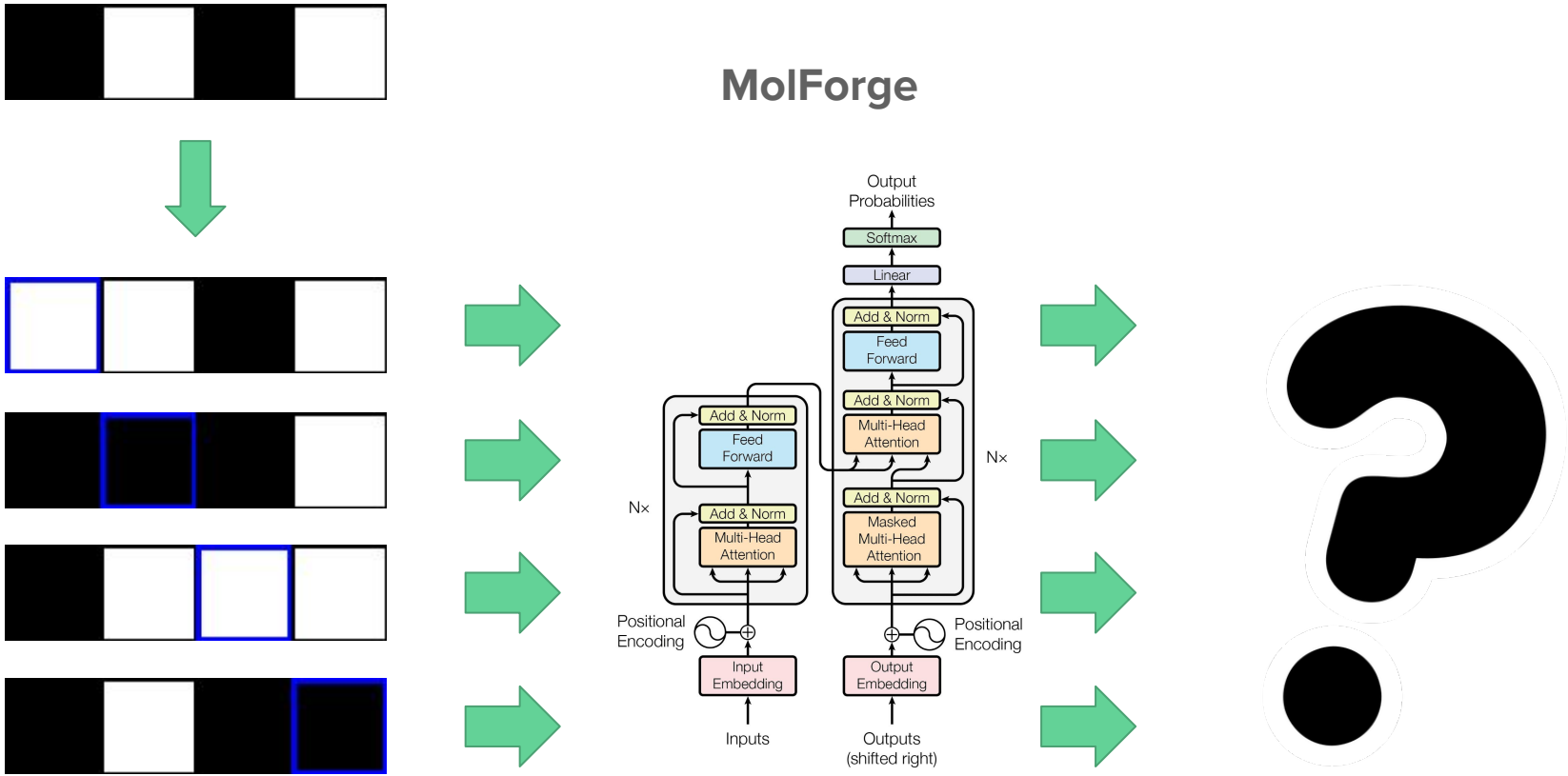
- COX-2 inhibitors = class of drugs used for treating pain and inflammation.
- Janus Kinase inhibitors = class of drugs that block JAK enzymes to treat autoimmune and inflammatory diseases.
- Molecular Weight = the sum of atomic weights of all atoms in a molecule.

MolForge

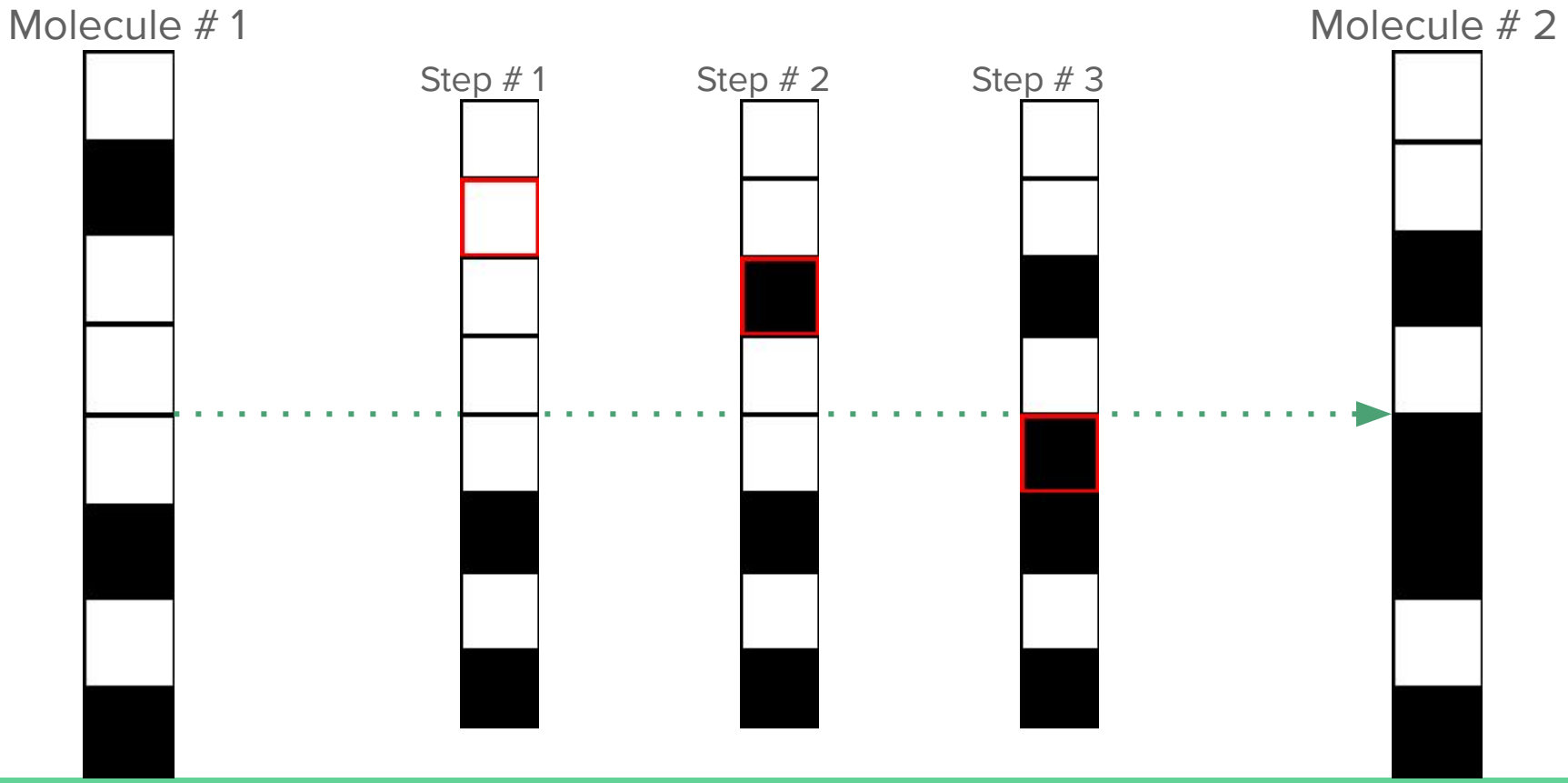
We increased the test set from 33k molecules (mid-term presentation) to 150k molecules in order to evaluate the reconstruction capability of MolForge.



Research problem 1 - Nearby molecules



Research problem 2 - Molecular pathway

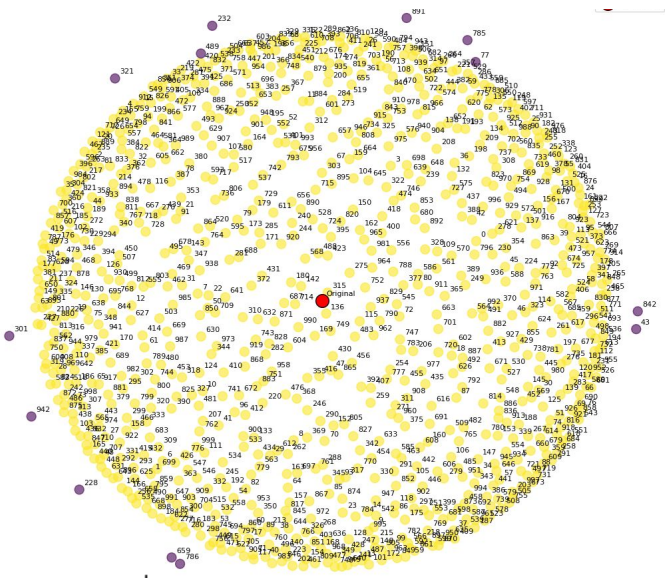


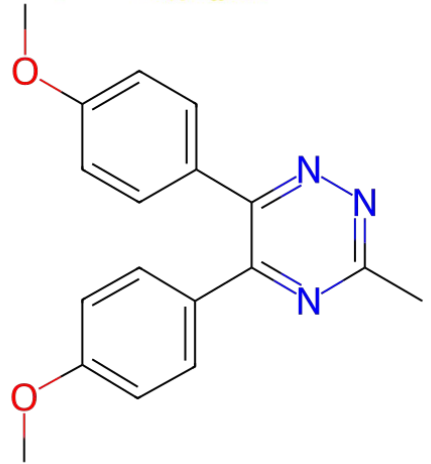
Literature review findings

Paper	Method	Method	Dataset	Key Results
Bit Silencing in Fingerprints (Wang and Bajorath, 2008)	"Bit silencing" to assess the contribution of each bit in molecular fingerprints to similarity search performance.	MACCS (166 structural fragments) Hit rate profile Weight vector $w_i = (1 + (hr_o - hr_i) \cdot sf) \cdot 100\%$ Weighted Tanimoto Similarity	21 activity classes from MDDR (Molecular Drug Data Report) 5000 compounds randomly selected from ZINC as background set	Average hit rate increased from 5% (conventional Tc) to 12% (bw_Tc) Average recovery rate increased from 8% to 20% Not all bits are equally useful For 8 classes where Tc failed to identify active compounds, bw_Tc achieved hit rates up to 20% and recovery rates up to 40%

Methodology - Nearby molecules

- Flipped ECFP4 fingerprint bits (1, 2, 4, 8, 128, 1024) times to examine nearby molecules
- Reduced dimensionality for singular flips by flipping only 10, 100, 1000 and 2048 bits
- Used MolForge to predict SMILES from modified fingerprints
- Calculated Tanimoto similarity between original and generated fingerprints
- Visualized similarity using MDS (Multi-Dimensional Scaling)





Results - valid and invalid SMILES

- As the size of the bit group increases, so the number of invalid SMILES

# of valid SMILES

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 # FPs	**2048**	**1024**	**512**	**256**	**16**	**2**

% of invalid SMILES

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%

Results - unique SMILES

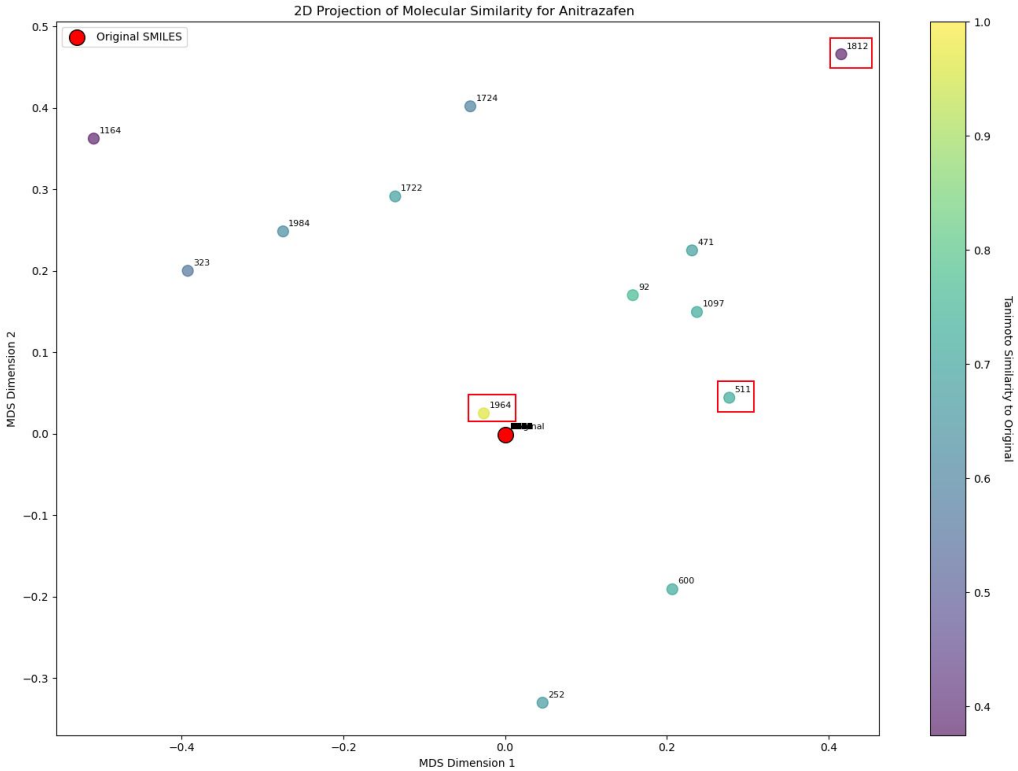
- As the size of the bit group increases, the number of unique SMILES increase until a certain point.

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

Anitrazafen 1 Bit Change 2048 SMILES

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

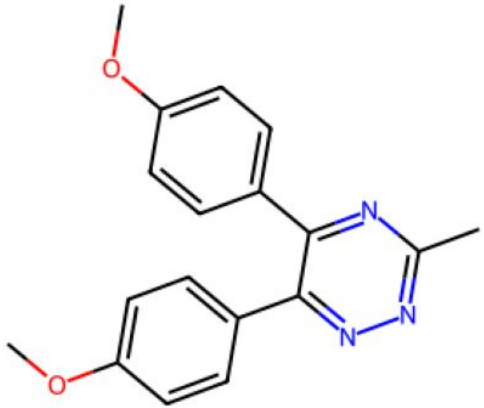
Anitrazafen - MDS - 1 bit flipped



Overlapping SMILES: 2034, Unique SMILES: 14 (PubChem: 1)

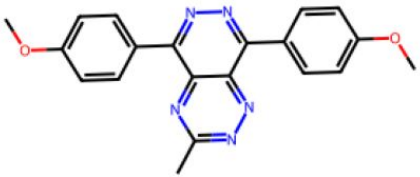
Original Molecule

Original Molecule: Anitrazafen
SMILES: COc1ccc(cc1)c2nnc(C)nc2c3ccc(OC)cc3

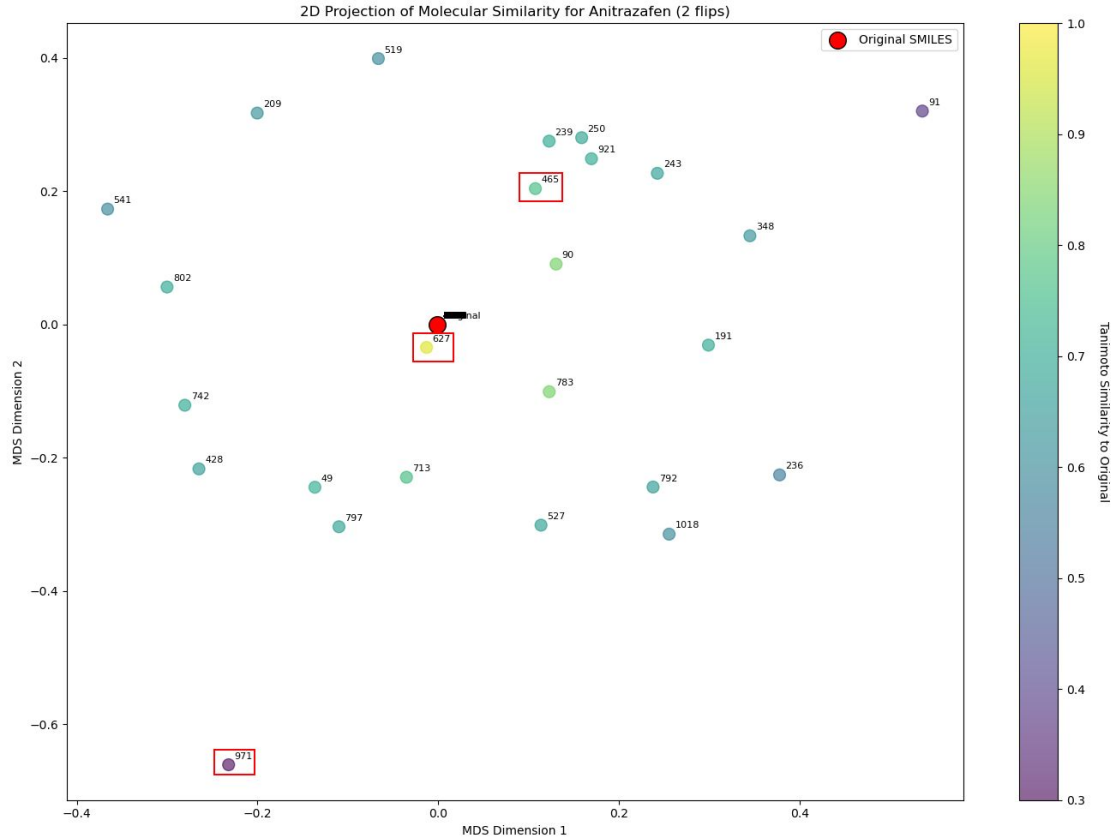


Generated Molecules

Generated Molecule 511 (Flipped Bits: 511)
SMILES: COc1ccc(-c2nnc(-c3ccc(OC)cc3)c3nc(C)nnc23)cc1
Tanimoto = 0.72



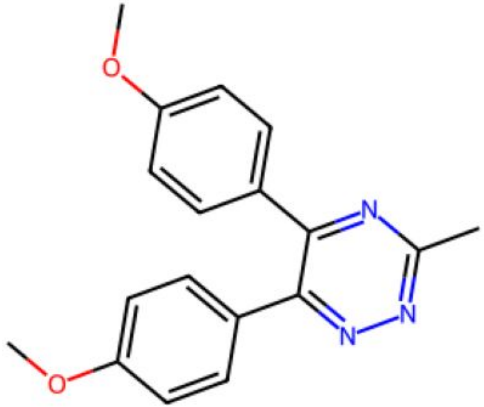
Anitrazafen - MDS - 2 bits flipped



Overlapping SMILES: 1000, Unique SMILES: 24 (PubChem: 3)

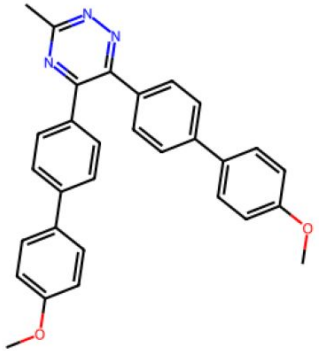
Original Molecule

Original Molecule: Anitrazafen
SMILES: COc1ccc(cc1)c2nnc(C)nc2c3ccc(OC)cc3

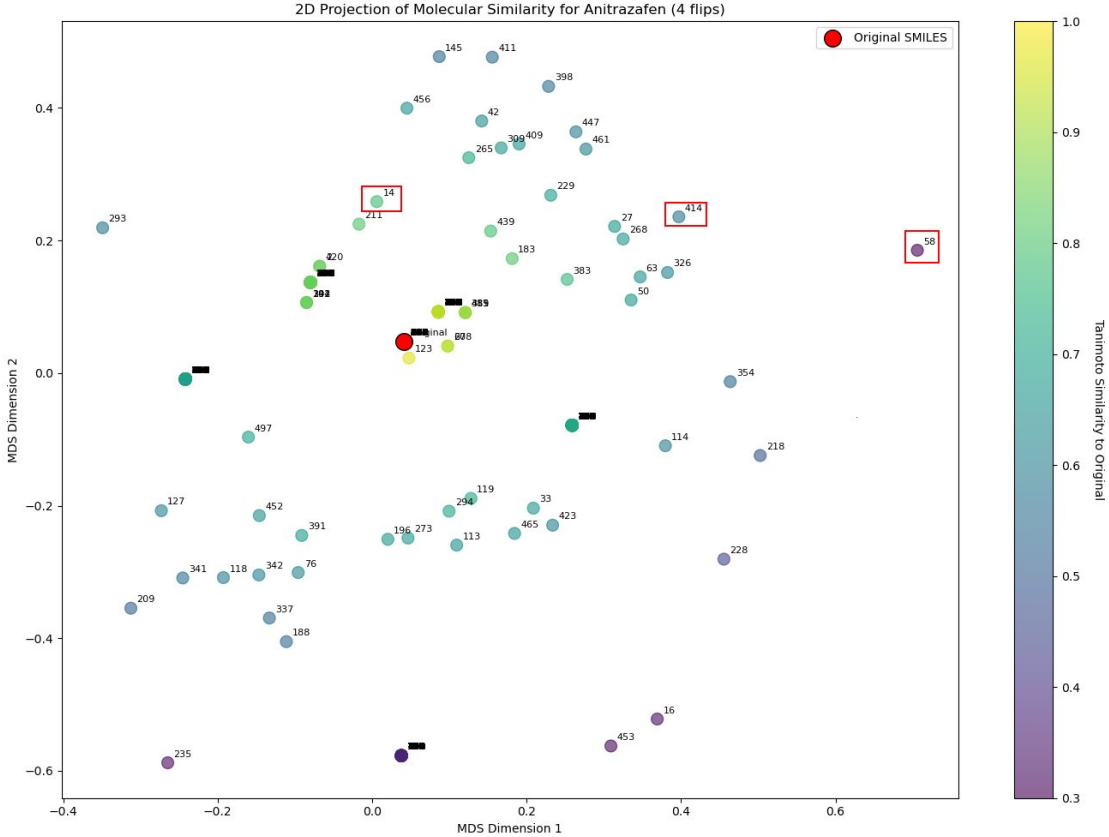


Generated Molecules

Generated Molecule 627 (Flipped Bits: [1964 1189])
SMILES: COc1ccc(-c2ccc(-c3nnc(C)nc3-c3ccc(-c4ccc(OC)cc4)cc3)cc2)cc1
Tanimoto = 0.96

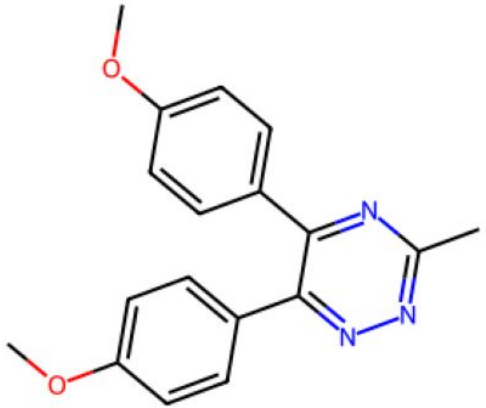


Anitrazafen - MDS - 4 bits flipped



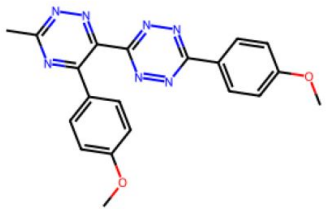
Original Molecule

Original Molecule: Anitrazafen
SMILES: COc1ccc(cc1)c2nnc(C)nc2c3ccc(OC)cc3



Generated Molecules

Generated Molecule 14 (Flipped Bits: [1855 1781 462 604])
SMILES: COc1ccc(-c2nnc(-c3nnc(C)nc3-c3ccc(OC)cc3)nn2)cc1
Tanimoto = 0.78

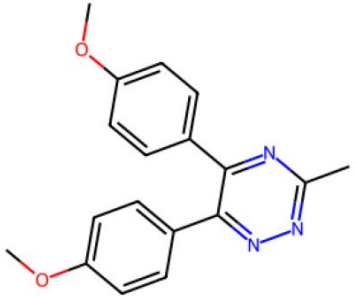


Overlapping SMILES: 451, Unique SMILES: 61 (PubChem: 2)

Anitrazafen- MDS - 8 bits flipped

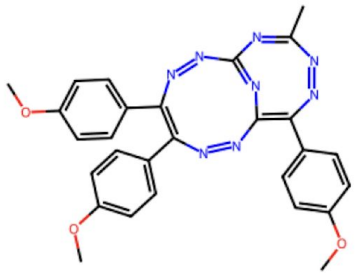
Original Molecule

Original Molecule: Anitrazafen
SMILES: COc1ccc(cc1)c2nnc(C)nc2c3ccc(OC)cc3



Generated Molecules

Generated Molecule 108 (Flipped Bits: [1332 1521 289 27 1664 945 1295 299])
IILES: COc1ccc(C2=C3N=NC(c4ccc(OC)cc4)=C(c4ccc(OC)cc4)N=NC(=N3)N=C(C)N=N2)cc1
Tanimoto = 0.35

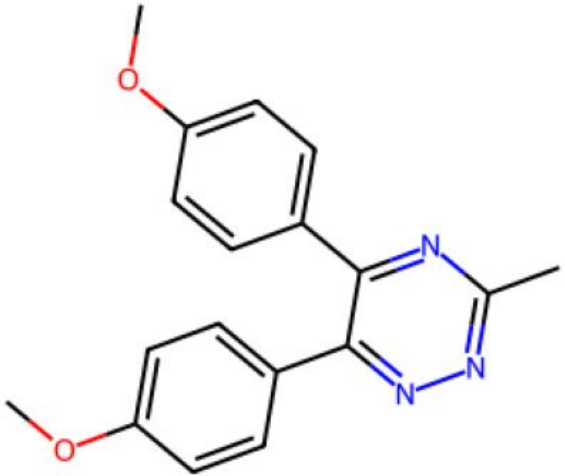


Overlapping SMILES: 141, Unique SMILES: 115 (PubChem: 1)

Anitrazafen- MDS - 128 bits flipped

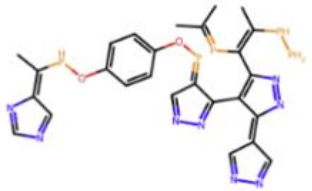
Original Molecule

Original Molecule: Anitrazafen
SMILES: COc1ccc(cc1)c2nnc(C)nc2c3ccc(OC)cc3



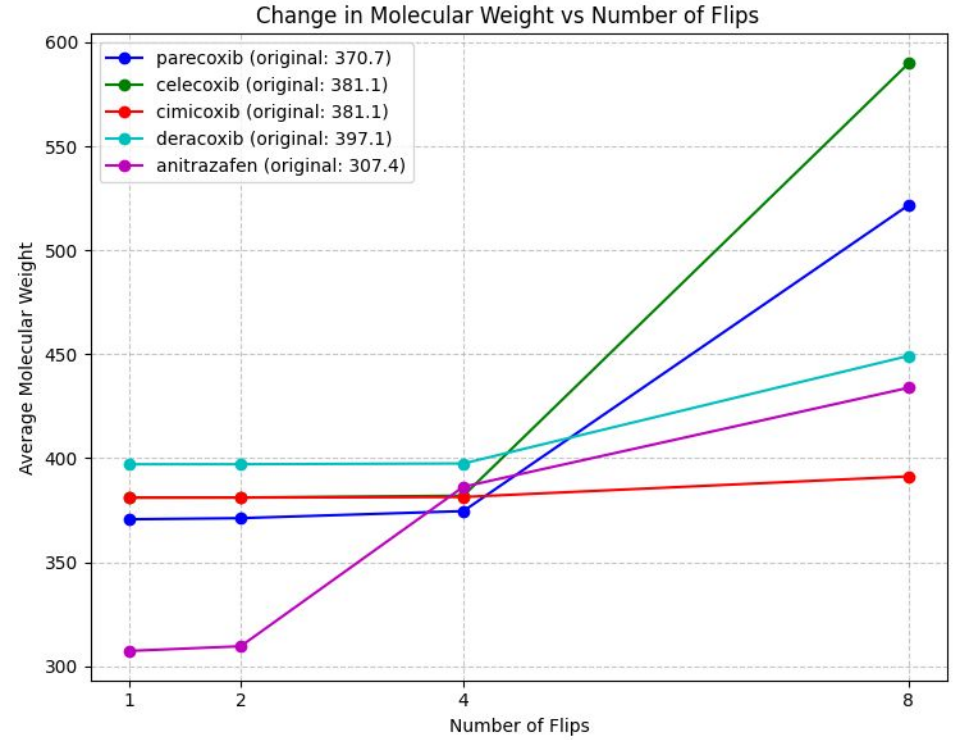
Generated Molecules

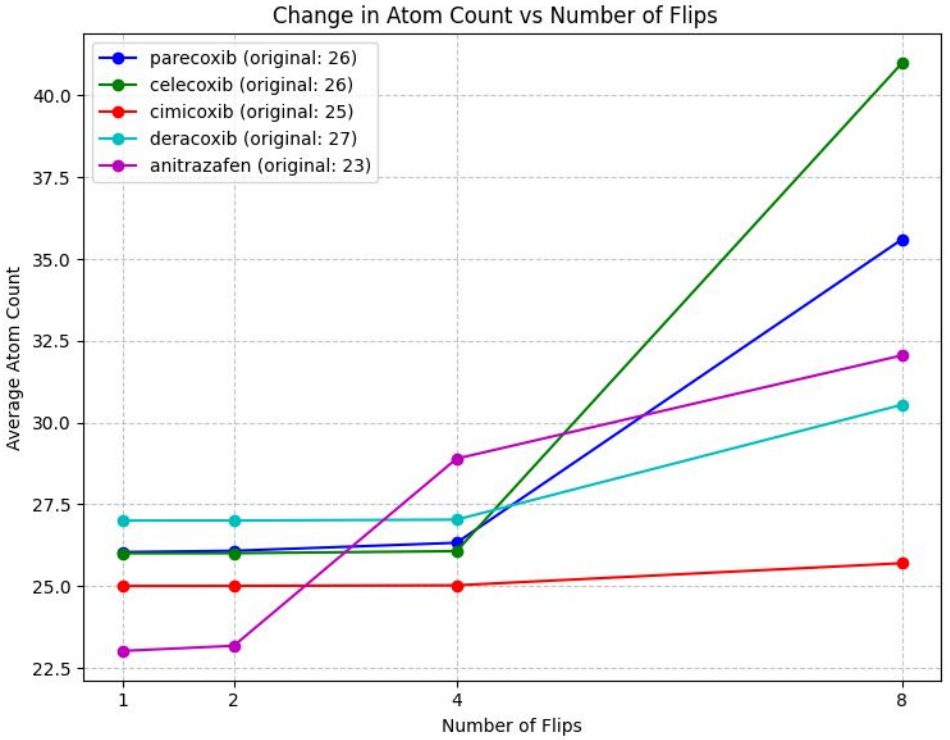
Generated Molecule 0 (Flipped Bits: [464 177 1603 1094 829 1300 1101 1146 851 191 1971 183 1415 1923
744 417 619 1550 1878   17 1608 1762 1963 1369 678 1973 528 1674
1816 1921 598 1389 188 1104 517 1521 1234 286 1741 1598 1703 202
496 866   77 1995 1780 111 440 1116 991   58 1475 1175 356 487
1505 1677 1799 1815 820 460 1988 350 1853 284 529 795 2029 909
1160 857 366 974 961 524 1745 1348 1319 1644 1546 1836 468 1489
327 1201 1960 1273 1620 763 1303 596 430 536 1839   63 577 1597
1099 1763 1159 1012 1738 1169 1367 1941 427 238 781 557 133 1038
955 527 1615 1523 785 1789 1017 134 834 722 751 1586 1454 583
1406 1080])
SMILES: CC(C)=PC(=C(C)PP)C1=C(C2=NN=CC2=POc2ccc(OPC(C)=C3C=NC=N3)cc2)C(=C2C=NN=C2)N=N1
Tanimoto = 0.11



Atoms and Molecular Weight

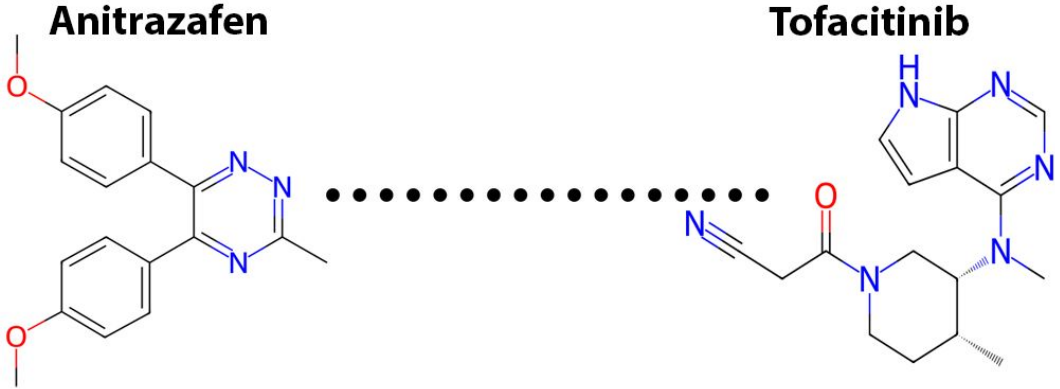
- Increasing fingerprint bit flips raised average atom and molecular weight



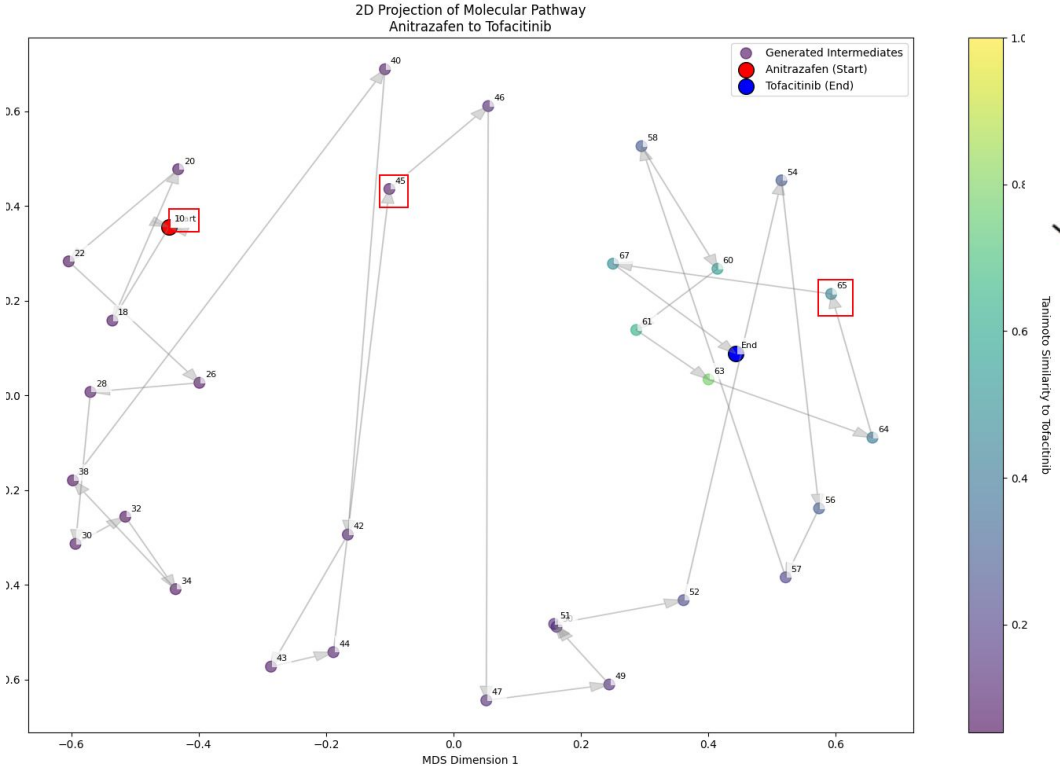


Methodology - Molecular pathway

- Studied pathway from COX-2 (Anitrazafen) ➜ Janus Kinase inhibitor (Tofacitinib)
- Modified Anitrazafen’s fingerprint bits by adding and removing step-by-step
- Predicted SMILES at each modification stage using MolForge
- Reached Tofacitinib Fingerprint by adjusting Anitrazafens bits
- Calculated Tanimoto similarity to both starting molecules

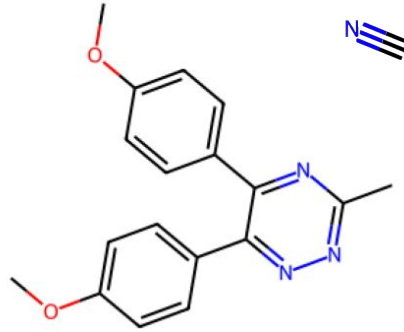


Anitrazafen ➜ Tofacitinib Pathway



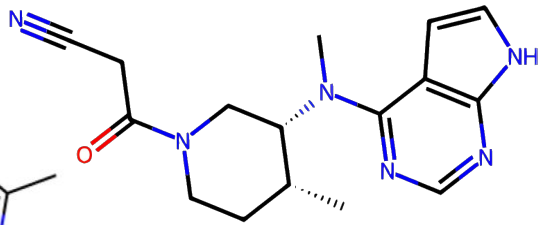
Start Molecule

Original Molecule: Anitrazafen
SMILES: COc1ccc(cc1)c2nnc(C)nc2c3ccc(OC)cc3

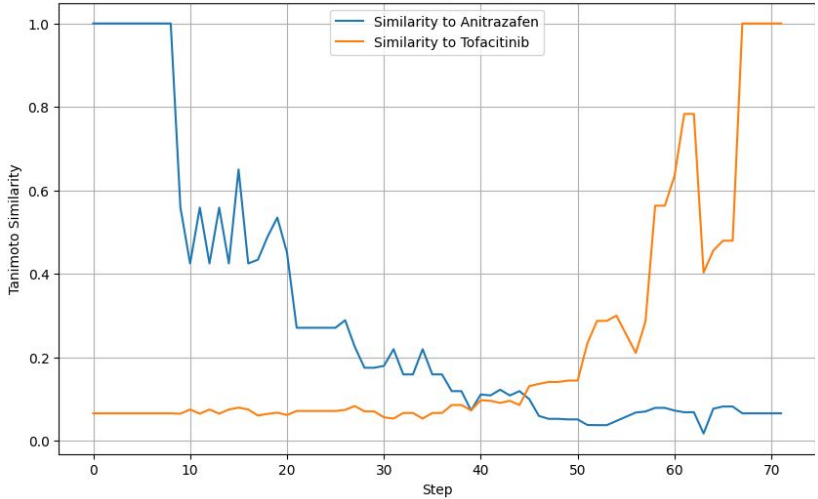


End Molecule

Final Molecule: Tofacitinib
SMILES: CC1CCN(C(=O)CC#N)CC1N(C)c1ncnc2[nH]ccc12



Generated Molecules



Valid SMILES: 71/72, Unique SMILES: 40/71, ChEMBL: 14/71

Conclusions

- MolForge is good at reconstructing smiles from ECFP4 vectors. However, it does not a good job at recognizing singular bit flips.
  - Most of generated SMILES overlap the original SMILE.
  - ECPF4 from the generated SMILES match in most of the cases with the ECPF4 of the original molecule rather than the flipped ECP4.
  - This is related in how MolForge was trained.
- MolForge is able to generate unique and valid molecules. However,
  - Most of them are not available in PubChem
  - Singular bit flipping often results in the original molecule
  - Generated molecules are repeating the SMILES structure
- More bit changes (more 1s in ECFP4) correlate with higher molecular weight and atom count on average

Thanks for listening! :)

Final Presentation

Nils Dunlop, Francisco Erazo, Qi Chen