D&AI - AI Innovation Lab

ToDD: Topological Compound Fingerprinting in Computer-Aided Drug Discovery (NeurIPS ’22)

Andac Demir, Baris Coskunuzer, Bülent Kızıltan

NOVARTIS | Reimagining Medicine

ROADMAP

1. Virtual Screening and Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Enrichment Factor
8. Results
9. Ablation Study
10. Potential Future Work

NOVARTIS | Reimagining Medicine

VIRTUAL SCREENING

1. Early phases of biomedical research involve the **identification of targets** for a disease, followed by **high-throughput screening (HTS)** experiments to determine hits within the compound library.

2. Then, **these compounds are optimized** to increase potency and other desired target properties.

3. In the final phases of the R&D pipeline, drug candidates have to pass a series of rigorous controlled tests in **clinical trials** to be considered for regulatory approval.

4. On average, this process takes 10-15 years end-to-end and costs in excess of ~2 billion US dollars.

5. HTS is **highly time and cost-intensive**. Therefore, it is critical to find good potential compounds effectively for the HTS step in short period of time for novel compound discovery.

NOVARTIS | Reimagining Medicine

KEY TAKEAWAYS

1. **Novelty**: We developed a new compound fingerprinting method based on topological features.

2. **Performance**: We develop and benchmark ML approaches; and outperform the state-of-the-art by a wide and statistically significant margin: 93% gain for Cleves-Jain and 54% gain for DUD-E Diverse dataset.

3. **Small data sets**: effective few-shot classification (only 2-3 active ligands per drug target for training)

4. **ML integration**: features suited for SoTA Neural Networks, as well as traditional ML methods

5. **Computational efficiency**: Full training + analysis on a laptop ~7 minutes (for a library of 1100 compounds, distributed across the 8 cores of an Intel Core i7 CPU (100GB RAM))

6. **Native integration of new information**: in contrast with all other fingerprinting approaches.

VIRTUAL SCREENING METHODS FOR HIT IDENTIFICATION

Structure-based methods

1. Molecular docking is the most common method.
2. Molecular docking approach models the interaction between a small molecule and a drug target at the atomic level. → This allows us to characterize the behavior of small molecules in the binding site of drug targets.
3. It requires:
    1. Knowledge of the binding site before docking process.
    2. Prediction of the ligand conformation as well as its position and orientation in binding site.

Ligand-based methods

1. We know a set of active ligands that can inhibit a drug target.
2. There is little or no structural information available for those drug targets.
3. Drug candidates are compared against a library of dozens/hundreds of active ligands and thousands of decoys (inactive ligands).

NOVARTIS | Reimagining Medicine

KEY CONTRIBUTIONS

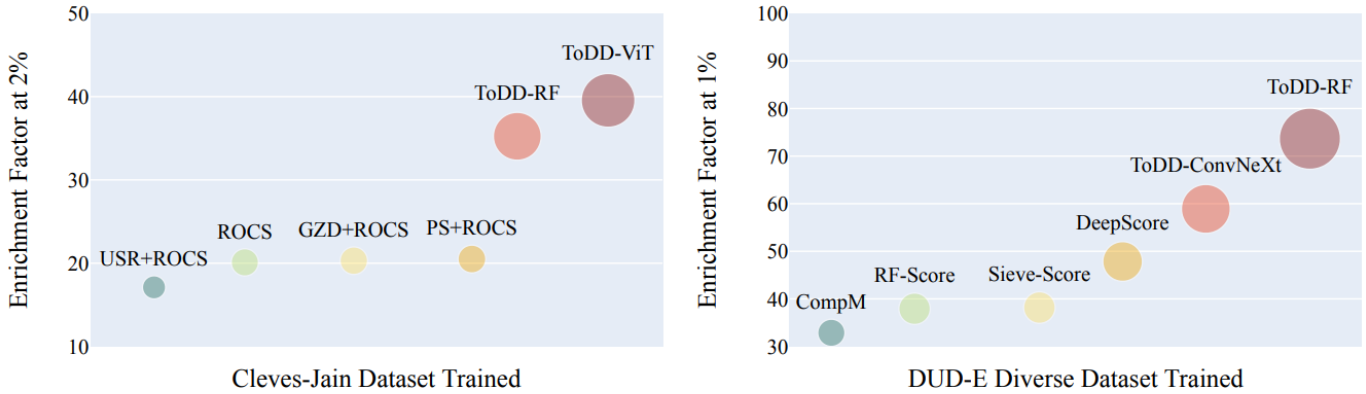


Figure 1: **Comparison of virtual screening performance.** Each bubble's diameter is proportional to its EF score. ToDD offers significant gain regardless of the choice of classification model such as random forests (RF), vision transformer (ViT) or a modernized ResNet architecture ConvNeXt. The standard performance metric $EF_{\alpha\%}$ is defined as $\frac{100}{\alpha}$, and therefore the maximum attainable value is 50 for $EF_{2\%}$, and 100 for $EF_{1\%}$.

NOVARTIS | Reimagining Medicine

KEY METRIC: ENRICHMENT FACTOR

- **Enrichment Factor** (EF) is the most common performance evaluation metric for Virtual Screening methods.

- VS method φ ranks compounds in the database by the **similarity score**. We measure the similarity score using the **inverse of Euclidian** distance between the embeddings of an anchor and drug candidate.

- Let N be the total number of ligands in the dataset, A_φ be the number of true positives (i.e., correctly predicted active ligands) in the first α% of all ligands and $N_{actives}$ be the number of active ligands in the whole dataset. Then,

$$EF_{\alpha\%} = \frac{A_\varphi / N_{actives}}{\alpha / 100}$$

→ a.k.a. ***prediction at k***

- Notice that with this definition, the **maximum score** for $EF_{\alpha\%}$ is $\frac{100}{\alpha}$, i.e., 100 for $EF_{1\%}$ and 20 for $EF_{5\%}$.

NOVARTIS | Reimagining Medicine

RESULTS

Table 1: Comparison of EF 2%, 5%, 10% and AUC values between ToDD and other virtual screening methods on the Cleves-Jain dataset.

Model	EF 2% (max. 50)	EF 5% (max. 20)	EF 10% (max. 10)	AUC
USR [7]	10.0	6.2	4.1	0.76
GZD [83]	13.4	8.0	5.3	0.81
PS [42]	10.7	6.6	4.9	0.78
ROCS [36]	20.1	10.7	6.2	0.83
USR + GZD [75]	13.7	7.7	4.7	0.81
USR + PS [75]	13.1	7.9	5.0	0.80
USR + ROCS [75]	17.1	9.1	5.4	0.83
GZD + PS [75]	16.0	9.1	5.9	0.82
PH_VS [48]	18.6	NA	NA	NA
GZD + ROCS [75]	20.3	10.8	5.3	0.83
PS + ROCS [75]	20.5	10.7	6.4	0.83
ToDD-RF	35.2±2.3	15.6±1.0	8.1±0.4	**0.94**±0.02
ToDD-ViT	**39.6**±1.4	**18.6**±0.4	**9.9**±0.1	0.90±0.01
Relative gains	92.9%	83.7%	54.1%	13.3%

Relative gains are relative to the next best performing model. Mean and standard deviation of EF scores evaluated by 5-fold cross-validation.

NOVARTIS | Reimagining Medicine

RESULTS

Table 2: Comparison of EF 1% (max. 100) between ToDD and other virtual screening methods on 8 targets of the DUD-E Diverse subset.

Model	AMPC	CXCR4	KIF11	CP3A4	GCR	AKT1	HIVRT	HIVPR	Avg.
Findsite [90]	0.0	0.0	0.9	21.7	34.2	39.0	1.2	34.7	16.5
Fragsite [91]	4.2	42.5	0.0	32.9	29.1	47.1	2.4	48.7	25.9
Gnina [78]	2.1	15.0	38.0	1.2	39.0	4.1	11.0	28.0	17.3
GOLD-EATL [87]	25.8	20.0	33.5	17.9	34.6	29.2	28.7	23.4	26.6
Glide-EATL [87]	35.5	20.8	30.5	15.1	24.0	31.6	29.0	22.0	26.1
CompM [87]	32.3	25.0	35.5	33.6	37.1	44.2	30.2	25.0	32.9
CompScore [66]	39.6	51.6	51.3	14.0	27.1	37.6	21.8	18.2	32.7
CNN [68]	2.1	5.0	11.2	28.7	12.8	84.6	12.2	9.9	20.8
DenseFS [44]	14.6	5.0	4.3	44.3	20.9	89.4	12.8	8.4	25.0
SIEVE-Score [88]	30.7	61.1	53.4	6.7	33.3	42.1	39.8	38.3	38.2
DeepScore [85]	28.1	56.8	54.3	37.1	40.9	59.0	43.8	62.8	47.9
RF-Score-VSv3 [88]	32.3	60.9	4.5	25.9	32.5	41.9	39.8	65.7	37.9
ToDD-RF	42.9±4.5	**92.3**±3.2	**75.0**±5.0	**67.6**±3.4	**78.9**±4.0	**90.7**±1.3	**64.1**±2.3	**92.1**±1.5	**73.7**
ToDD-ConvNeXt	**46.2**±3.6	84.6±2.8	72.5±3.6	28.8±2.8	46.0±2.0	81.2±2.5	37.5±3.6	74.6±1.0	58.9
Relative gains	16.7%	51.1%	38.1%	52.6%	92.9%	1.5%	46.3%	40.2%	53.9%

Relative gains are relative to the next best performing model. Mean and standard deviation of EF scores evaluated by 5-fold cross-validation.

NOVARTIS | Reimagining Medicine

KEY CONTRIBUTIONS

1. We use **multiparameter persistent homology (PH)** of the **Vietoris-Rips complexes** to produce **topological fingerprints** for compounds.

2. Our multiparameter PH approach can successfully **incorporate more than one domain function to the PH process** such as atomic mass, partial charge, bond type, ionization energy, electron affinity.

3. We perform extensive numerical experiments in **Virtual Screening (VS)**, showing that our ToDD models **outperform all state-of-the-art methods by a wide margin**.

4. The strong hierarchical topological representations enable ToDD to become a **model agnostic** method that is extensible to state-of-the-art neural networks (**ConvNeXt, Vision Transformer**) as well as ensemble methods like random forests (RF).

5. Transfer learning by finetuning triplet networks where pretrained ConvNeXt and Vision Transformer models serve as the backbone produce **successful results on few-shot classification tasks**.

ROADMAP

1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

NOVARTIS | Reimagining Medicine

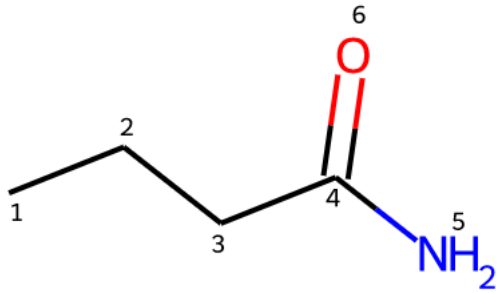
RELATED WORK

- **ECFP** (Extended Connectivity Fingerprints) -> Most popular compound fingerprinting technique based on Morgan algorithm.
- **Step 1**: A set of numbers is assigned to each atom in the compound using the Daylight atomic variants (7 properties of an atom):
    1. # of non-hydrogen immediate neighbors
    2. Valency minus the # of connected hydrogens
    3. Atomic number
    4. Atomic mass
    5. Formal charge
    6. # of attached hydrogens
    7. Whether the atom is part of a ring or not (1, if yes, 0 no)

NOVARTIS | Reimagining Medicine

RELATED WORK

- **Step 1**: We convert the features of an atom to an integer using a hashing function. Consider the atom 4 in our compound Butyramide:
  1. # of non-hydrogen immediate neighbors = 3
  2. Valency minus the # of connected hydrogens = 4
  3. Atomic number = 6
  4. Atomic mass = 12
  5. Formal charge = 0
  6. # of attached hydrogens = 0
  7. Whether the atom is part of a ring or not (1, if yes, 0 no) = 0




```
identifier = hash((3, 4, 6, 12, 0, 0, 0))
print(identifier)
# -2155244659601281804
```

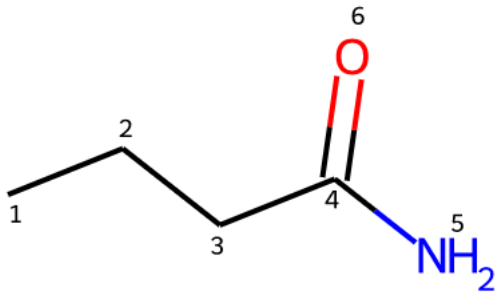
Hash values are unique like fingerprints,
But the original values cannot be recovered from hash values

NOVARTIS | Reimagining Medicine

RELATED WORK

- **Step 1**: We repeat this process for all the 6 vertices in the compound and get a set of hash values:

```
1: -4080868480043360372
2:  8311098529014133067
3:  8311098529014133067
4: -2155244659601281804
5: -3602994677767288312
6:  8573586092015465947
```

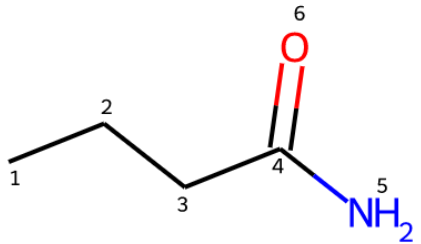


Initial representation of Butyramide with a set of hash values (identifiers).
Notice that atom 2 and 3 have the same integer identifier after the first step.

NOVARTIS | Reimagining Medicine

RELATED WORK

- **Step 2**: To include information about the neighborhood of the atoms, we update these identifiers:
  1. First, an array is initialized containing the iteration number and the initial identifier of the atom in question.
  2. Then, we add to this array the identifier of each non-hydrogen nearest neighbor along with the bond order with that particular atom:



[(1, -2155244659601281804), (1, -3602994677767288312), (1, 8311098529014133067), (2, 8573586092015465947)]

Identifier (hash value)

Iteration number | Single bond | Single bond | Double bond

  3. Same as before, this list is hashed to get an updated integer identifier.
  4. This process is repeated for all atoms for a prespecified number of iterations.

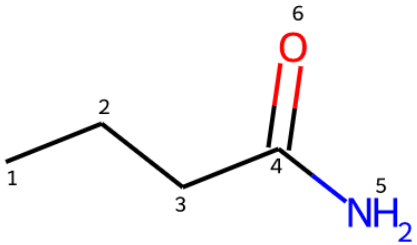
NOVARTIS | Reimagining Medicine

RELATED WORK

- **Step 2**: To include information about the neighborhood of the atoms, we update these identifiers:

  4. This process is repeated for all atoms for a prespecified number of iterations. The updated identifiers after first iteration will be:

```
1: -3879702859024654160
2:  2648074263463118673
3:  9209025387859845960
4:  3790237506519639747
5: -8399737669368778010
6:  3271801898087186516
```



Notice that atom 2 and 3 do not have the same integer identifier after the second step. This is because the atoms are identical when we look at them individually, but become different when using their neighborhood information.

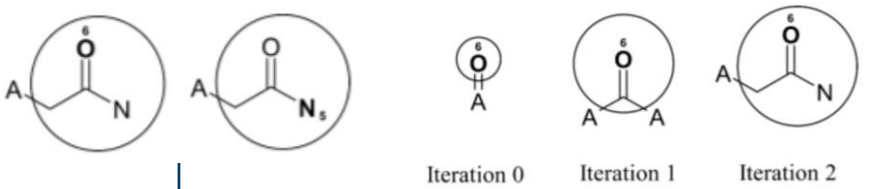
Note: After each iteration, the identifiers are appended to a feature list.

So if the compound has 6 atoms and if we have done 2 iterations of update, then our feature set will become a vector of 12 identifiers (hash values).

NOVARTIS | Reimagining Medicine

RELATED WORK

- **Step 3**: As we increase the radius to include information from K nearest neighbors, we end up generating different identifiers for the same substructures.



- Before the iteration begins: identifier is simply double-bonded oxygen.
- After one iteration, the identifier represents a carbonyl group.
- After two iterations, the identifier represents an aliphatic carboxylic acid amide.

- For instance, after 2 iterations the identifiers generated for Oxygen and Nitrogen are: -5964710996914813053 and 8916398073441202914 respectively.
- This difference is expected since the regions started at different atoms.
- For each compound, we deduplicate the identifiers that represent the same substructure.

NOVARTIS | Reimagining Medicine

RELATED WORK

- **Step 4**: The last step is to convert these identifiers into a computer-usable bit vector:

1. First, the user has to choose the length of the fingerprint vector. Traditionally, a length of 1024 is used.
2. Once the length is decided, initialize a zero-vector of the decided length.

```
import numpy as np
fp = np.zeros(1024)
print(fp)
# array([0., 0., 0., ..., 0., 0., 0.])
```

3. Divide each identifier with the vector length (1024) and calculate the remainder.

   remainders = [908, 331, 244, 520, 475, 176, 849, 840, 707, 742, 84, 553, 632, 358]

4. Lastly, set the values in the bit vector to one at the indices equal to the remainders. In other words, set the values to 1 in the positions 908, 331, …, 358.

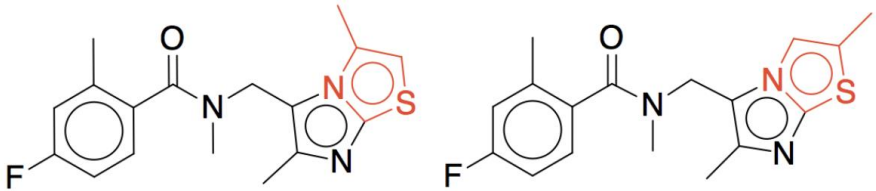
```
for x in remainders:
 fp[x] = 1
```

- Terrible way, because different hash values can give the same remainder when divided by the vector length. → **Bit Collision** → Increasing the vector length to avoid bit collision causes **curse of dimensionality**.

- Also: **Hash function is irreversible. No way back from hash identifiers back to the compound topology!!**

NOVARTIS | Reimagining Medicine

RELATED WORK



Cc1cn2c(CN(C)C(=O)c3ccc(F)cc3C)c(C)nc2s1
Cc1cc(F)ccc1C(=O)N(C)Cc1c(C)nc2scc(C)n12

Figure 1. Two almost identical molecules with markedly different canonical SMILES in RDKit. The edit distance between two strings is 22 (50.5% of the whole sequence).

- Another popular technique **SMILES** formulated the compound fingerprinting task as string generation problem.
- SMILES strings are reversible i.e., they can be translated into graphs.

- However. SMILES has 2 limitations:

1. Two molecules with similar chemical structures may be encoded into markedly different SMILES strings.

2. Essential chemical properties such as molecule validity are easier to express on graphs rather than linear SMILES representations.

NOVARTIS | Reimagining Medicine

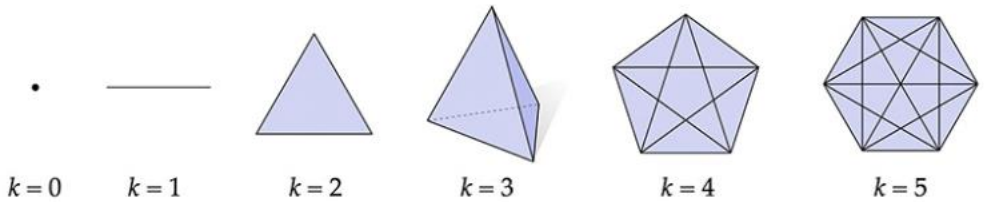
ROADMAP

1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Extracting Multiparameter Persistence Signatures of Compounds
5. Datasets
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

NOVARTIS | Reimagining Medicine

HOMOLOGY

- **Homology**: Mathematical way of counting connected components and loops in topological spaces.
  - Homology is a **topological invariant**, you can't change the number of connected components or holes of an object by bending/stretching it.
  - **Simplex**: Represents the simplest polytope in any given space. (Generalization of the notion of a triangle or tetrahedron to arbitrary dimensions.) --> **convex hull of n+1 independent points**

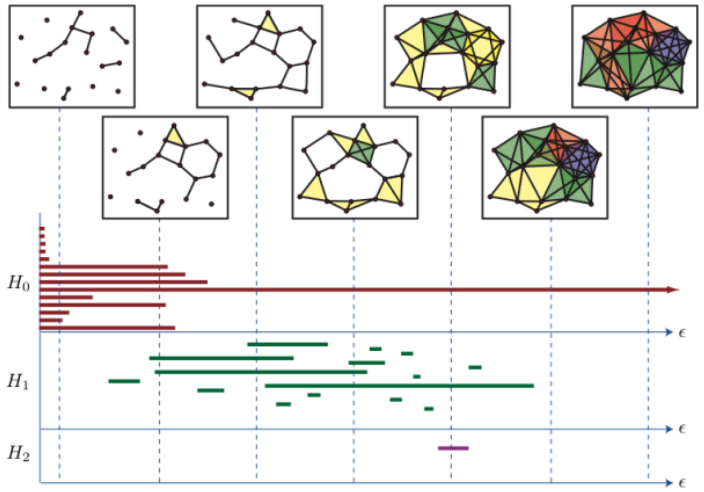


k-simplex is a point, a line segment, a triangle, a tetrahedron ,etc..

NOVARTIS | Reimagining Medicine

PERSISTENT HOMOLOGY

- **Persistent Homology**: A topological data analysis tool to study qualitative features of the data across the increasing sequence of simplex complexes.
  - **Input**: Increasing sequence of simplex complexes.
  - **Output**: Persistence homology barcodes, which show the homology at each stage.
  - **Methods**:
    1. Vietoris-Rips Complexes
    2. Alpha Complexes
    3. Cech Complexes
    4. Delaunay Complexes



NOVARTIS | Reimagining Medicine

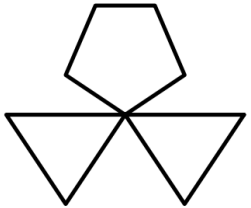
PERSISTENT HOMOLOGY

- **Persistent Homology** is our main tool to distinguish shapes of substructures.
- **i-dim homology** counts the # of i-dim holes.
- **0-dim hole**s are just connected components.
- First example has 6 disks, so 0-dim homology has rank 6 reflecting those 6 connected components.

i-dimensional homology H_i "counts the number of *i*-dimensional holes"

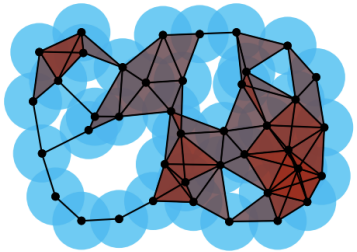
0-dimensional homology H_0: rank 6

1-dimensional homology H_1: rank 0



0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 3



0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 6

NOVARTIS | Reimagining Medicine

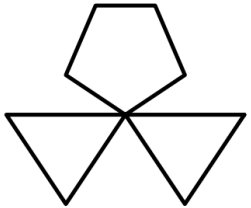
PERSISTENT HOMOLOGY

- **Persistent Homology** is our main tool to distinguish shapes of substructures.
- **i-dim homology** counts the # of i-dim holes.
- **1-dim homology** counts the # of holes that are like a circle/loop/cylinder.
- Second example has 3 loops, hence 1-dim homology has rank 3.
- Third example has 6 loops, hence 1-dim homology has rank 6.

i-dimensional homology H_i "counts the number of *i*-dimensional holes"

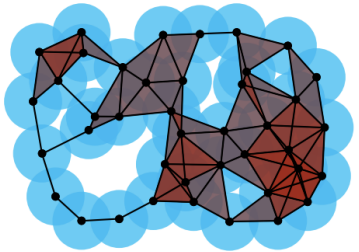
0-dimensional homology H_0: rank 6

1-dimensional homology H_1: rank 0



0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 3



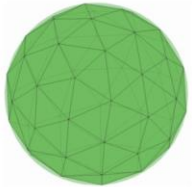
0-dimensional homology H_0: rank 1

1-dimensional homology H_1: rank 6

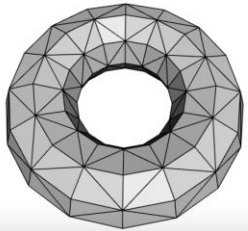
NOVARTIS | Reimagining Medicine

PERSISTENT HOMOLOGY

- **Persistent Homology** is our main tool to distinguish shapes of substructures.
- **i-dim homology** counts the # of i-dim holes.
- We also can count the 2-dim holes (in case we use the 3D conformations of compounds instead of their 2D geometry).
- A hollow sphere or taurus has each 1 2-dim homology.



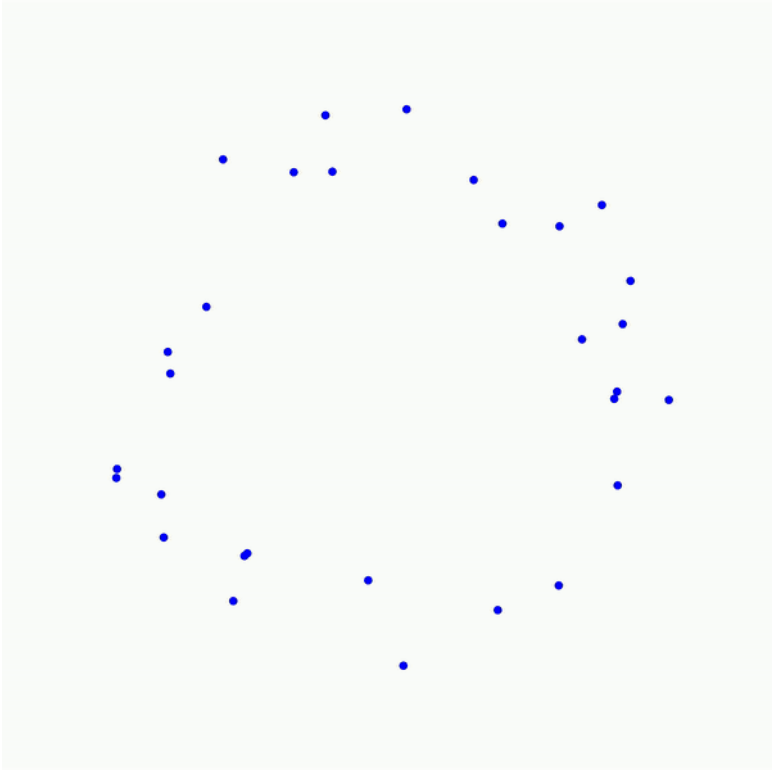
0-dimensional homology H_0: rank 1
1-dimensional homology H_1: rank 0
2-dimensional homology H_2: rank 1



0-dimensional homology H_0: rank 1
1-dimensional homology H_1: rank 2
2-dimensional homology H_2: rank 1

NOVARTIS | Reimagining Medicine

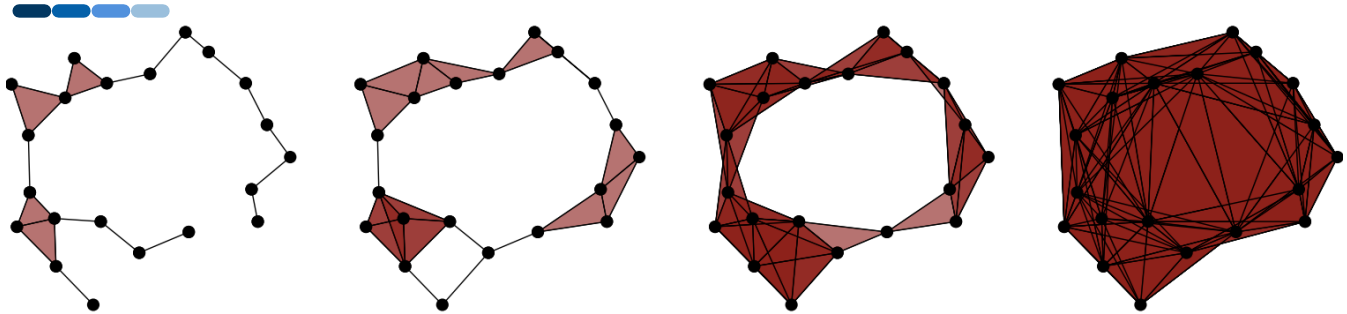
PERSISTENT HOMOLOGY



- **Key Idea**: Construct a graph piece-by-piece and track the topological changes.
- We capture 2 topological summaries: connected components (H_0) and loops (H_1).

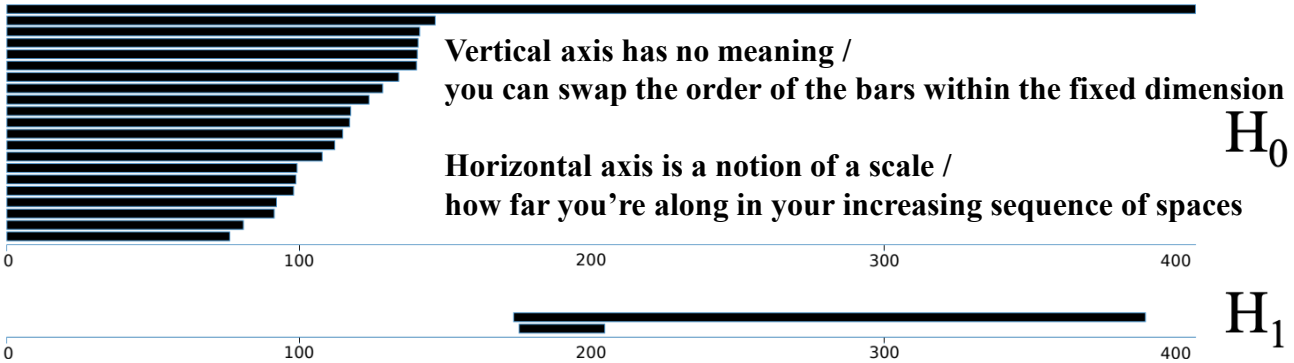
NOVARTIS | Reimagining Medicine

PERSISTENCE BARCODES



Input: Increasing spaces (as more edges and triangles get added)
Output: Barcode

Cubic computation time in the number of spaces.



Vertical axis has no meaning /
you can swap the order of the bars within the fixed dimension

H_0

Horizontal axis is a notion of a scale /
how far you're along in your increasing sequence of spaces

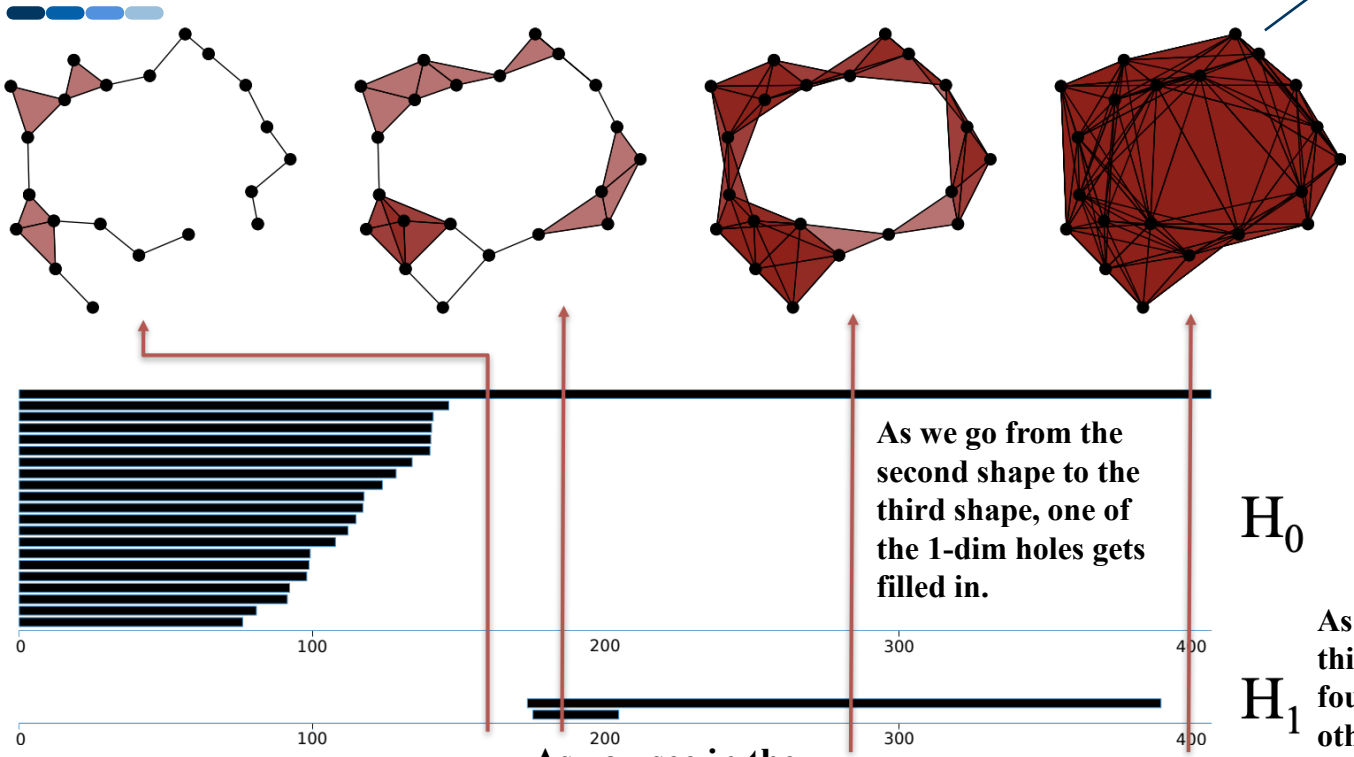
H_1

Barcodes track the number of 0 and 1-dim holes as the spaces increase.

Longer bars are interpreted as more features &
Shorter parts are more representative of noise

NOVARTIS | Reimagining Medicine

PERSISTENCE BARCODES



As the space increases, the disparate points connect up together finally into one connected component and remain as 1 connected component forever.

As we go from the second shape to the third shape, one of the 1-dim holes gets filled in.

H_0

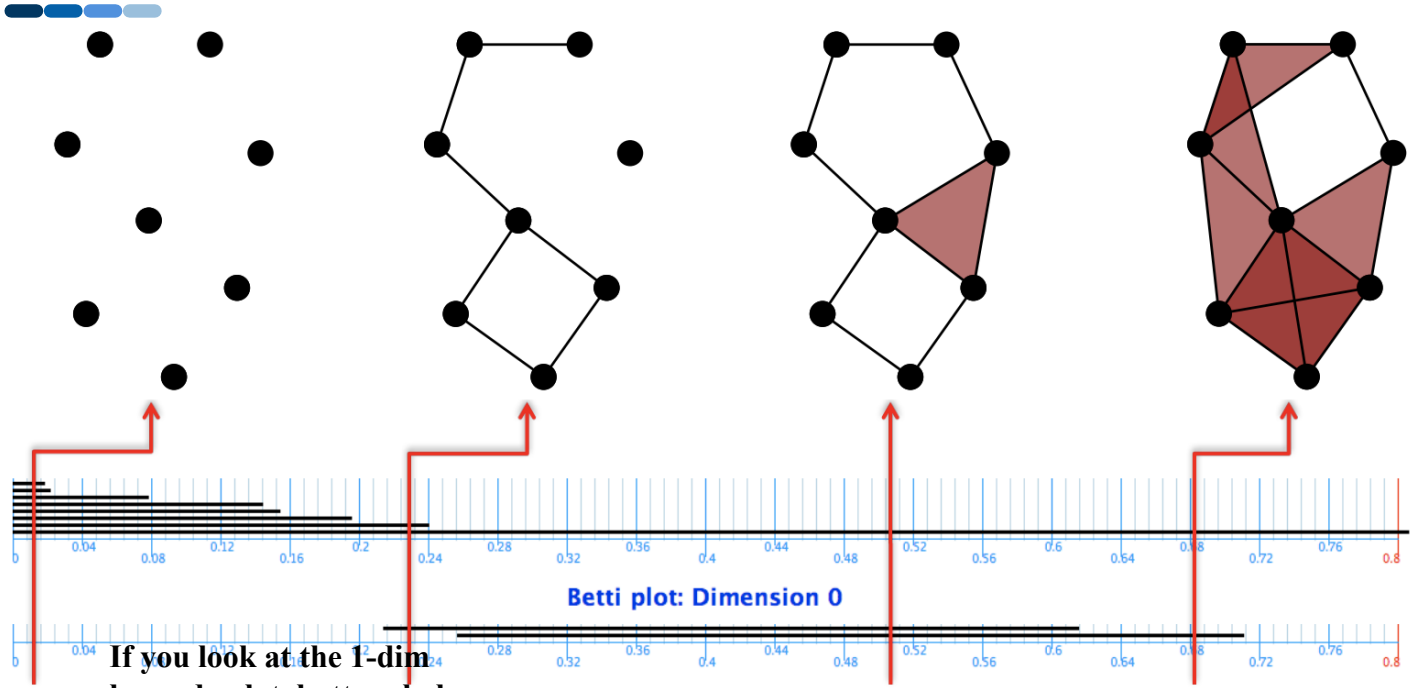
H_1

As we go from the third shape to the fourth shape, the other 1-dim hole also gets filled in.

As you see in the second shape, we have 2 1-dim holes.

NOVARTIS | Reimagining Medicine

PERSISTENCE BARCODES



If you look at the 1-dim barcode plot, bottom hole which is born first corresponds to this 1-dim bar which is born first.

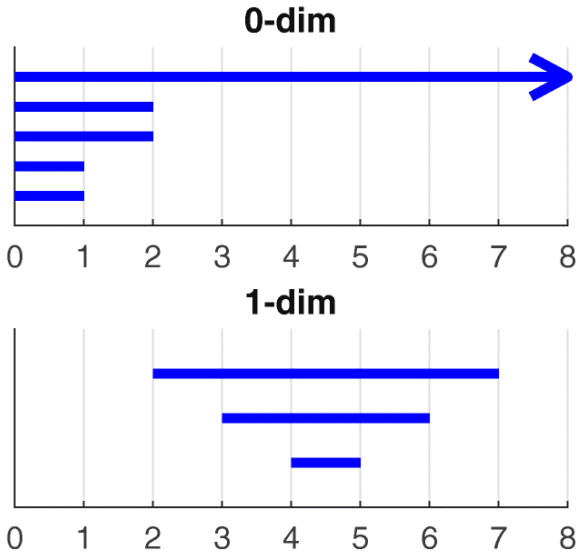
The other 1-dim hole which is born second corresponds to the other 1-dim bar which is born second.

The 1-dim that is born first also dies first in this example.

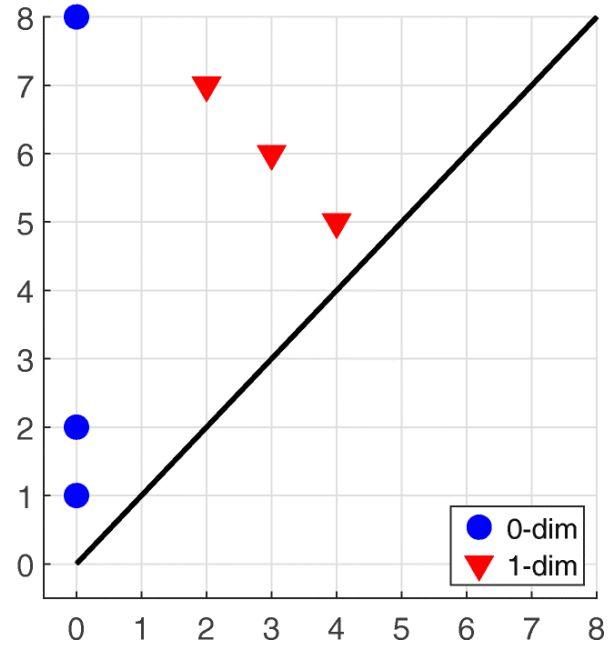
NOVARTIS | Reimagining Medicine

FROM PERSISTENCE BARCODES TO PERSISTENCE DIAGRAMS

a

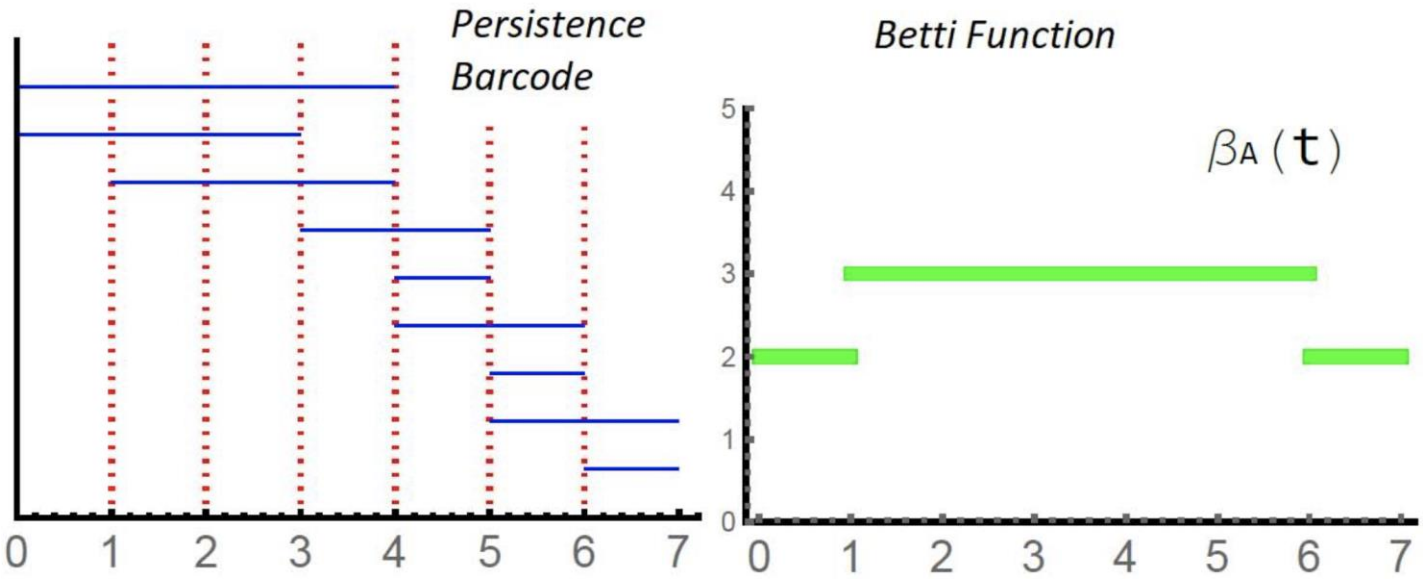


b



NOVARTIS | Reimagining Medicine

BETTI VECTORIZATION OF PERSISTENCE BARCODES



NOVARTIS | Reimagining Medicine

ROADMAP

1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

NOVARTIS | Reimagining Medicine

DATASETS

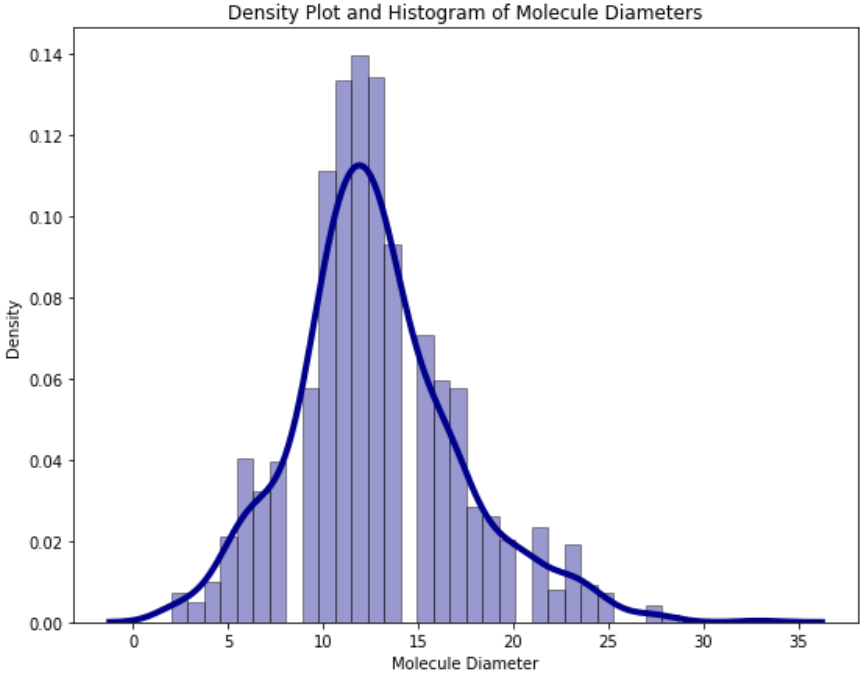
1. **Cleves-Jain Dataset**:
   - Relatively small dataset that contains 1149 compounds.
   - There are 22 different drug targets, and for each one of them the dataset provides only 2-3 active compounds dedicated for model training, which presents a **few-shot learning task**.
   - All targets are associated with 4 to 30 active compounds dedicated for model testing.
   - Additionally, the dataset contains 850 decoy compounds (used for all 22 targets).
   - The aim is to find the active compounds among the pool of active compounds and decoys.

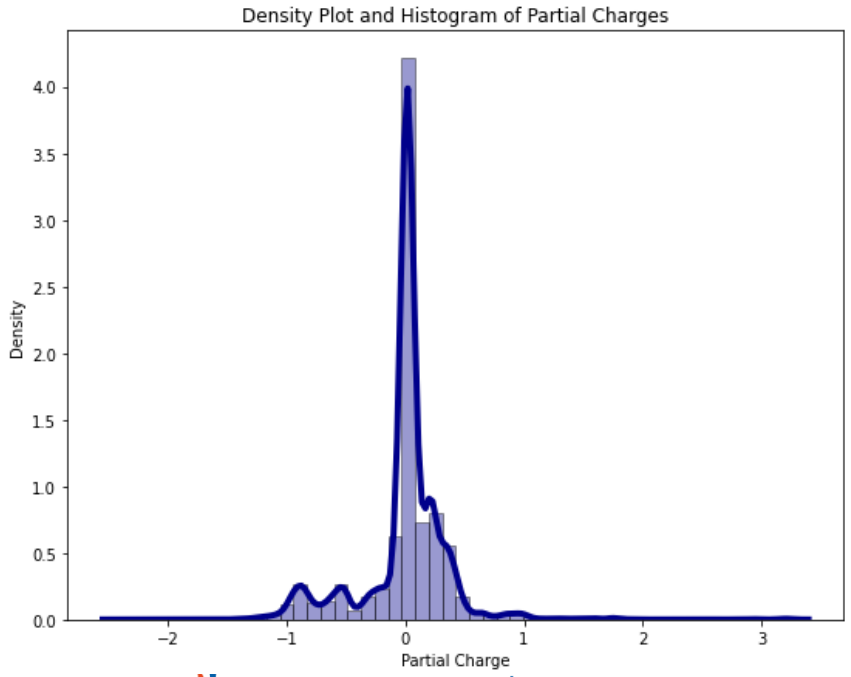
NOVARTIS | Reimagining Medicine

DATASETS

1. **Cleves-Jain Dataset**:

List of unique atoms (10): C, O, N, H, S, Cl, F, P, Br, I





NOVARTIS | Reimagining Medicine

DATASETS

1. Cleves-Jain Dataset:

```
# superlevel
charge_levels = [x[0] for x in list_charges]
charge_levels
```

```
[-2.41836,
 -0.5122,
 -0.05004,
 -0.01444,
 0.01239,
 0.02301,
 0.03973,
 0.07499,
 0.19865,
 0.32739]
```

```
# sublevel
charge_levels = [x[-1] for x in list_charges]
charge_levels
```

```
[-0.51226,
 -0.05015,
 -0.01444,
 0.01239,
 0.02301,
 0.03973,
 0.07499,
 0.19863,
 0.32736,
 3.26758]
```


NOVARTIS | Reimagining Medicine

DATASETS

2. DUD-E Dataset (Diverse Subset):

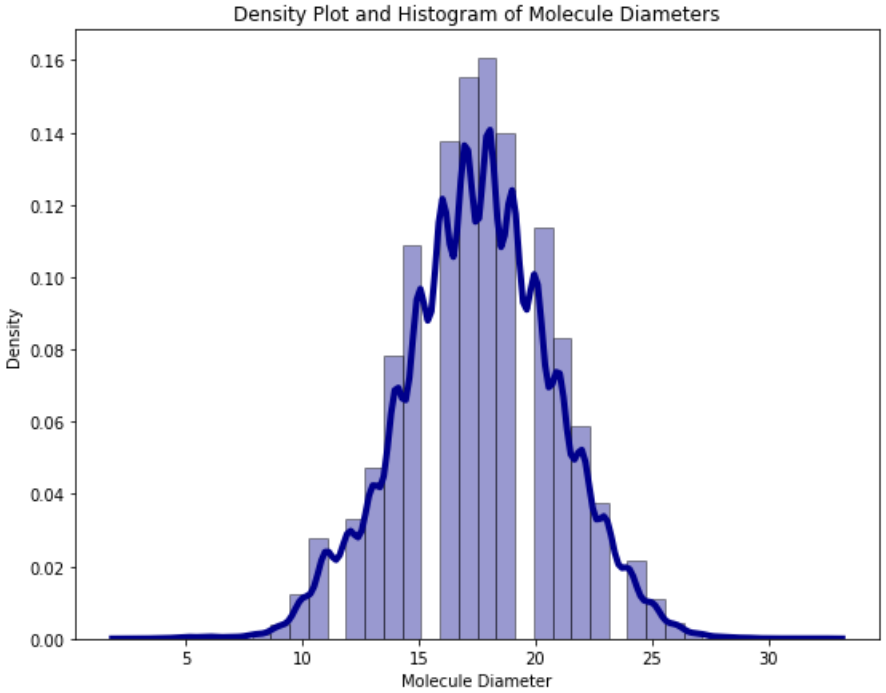
- DUD-E (Directory of Useful Decoys, Enhanced) is a comprehensive ligand dataset with 102 targets and approximately 1.5 million compounds.
- The targets are categorized into 7 classes with respect to their protein type.
- The "Diverse subset" of DUD-E contains targets from each category to give a balanced benchmark dataset for Virtual Screening methods.
- Diverse subset contains 116,105 compounds from 8 target and 8 decoy sets. One decoy set is used per target.

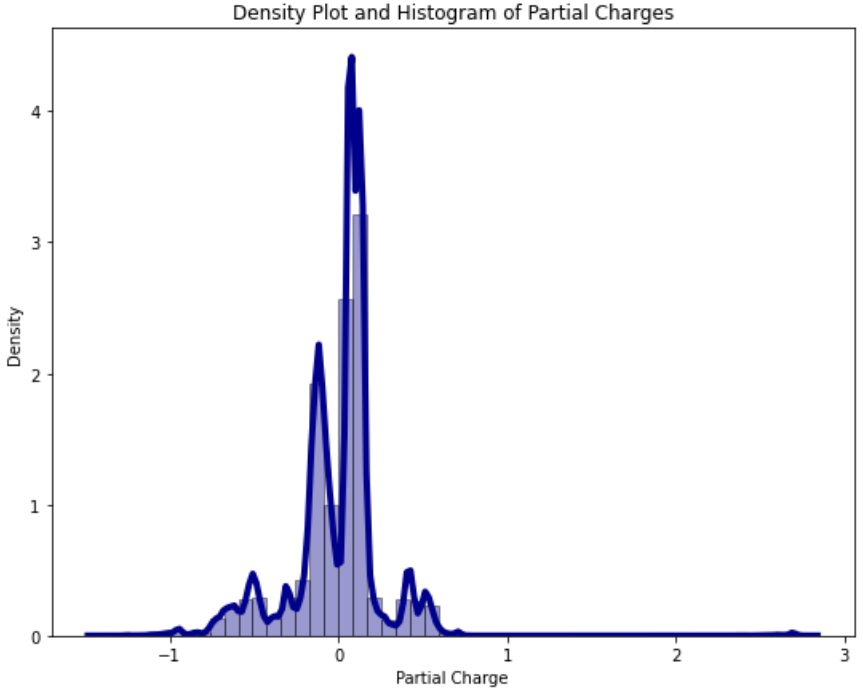
NOVARTIS | Reimagining Medicine

DATASETS

List of unique atoms (11): C, O, N, H, S, Cl, F, P, Br, I, Si

2. DUD-E Dataset (Diverse Subset):





NOVARTIS | Reimagining Medicine

DATASETS

2. DUD-E Dataset (Diverse Subset):

```
# superlevel
charge_levels = [x[0] for x in global_partial_charges]
charge_levels
```

```
[-1.4402,
 -0.3208,
 -0.1436,
 -0.0965,
 -0.0108,
 0.0621,
 0.0792,
 0.1095,
 0.1339,
 0.1773]
```

```
# sublevel
charge_levels = [x[-1] for x in global_partial_charges]
charge_levels
```

```
[-0.3208,
 -0.1436,
 -0.0965,
 -0.0108,
 0.0621,
 0.0792,
 0.1095,
 0.1339,
 0.1773,
 2.7958]
```


NOVARTIS | Reimagining Medicine

DATASETS

Table 3: Summary statistics of the Cleves-Jain dataset.

Target	# Training Samples	# Test Samples
a	3	6
b	3	22
c	2	13
d	3	6
e	2	5
f	2	4
g	2	5
h	2	5
i	2	5
j	3	14
k	3	14
l	3	10
m	3	9
n	2	10
o	3	30
p	3	23
q	3	11
r	2	14
s	3	15
t	2	5
u	3	9
v	3	7
Decoy	0	850

Table 4: Summary statistics of the DUD-E Diverse dataset.

Target	Description	# Active	# Decoy
AMPC	beta-lactamase	62	2902
CXCR4	C-X-C chemokine receptor type 4	122	3414
KIF11	kinesin-like protein 1	197	6912
CP3A4	cytochrome P450 3A4	363	11940
GCR	glucocorticoid receptor	563	15185
AKT1	serine/threonine-protein kinase Akt-1	423	16576
HIVRT	HIV type 1 reverse transcriptase	639	19134
HIVPR	HIV type 1 protease	1395	36278

NOVARTIS | Reimagining Medicine

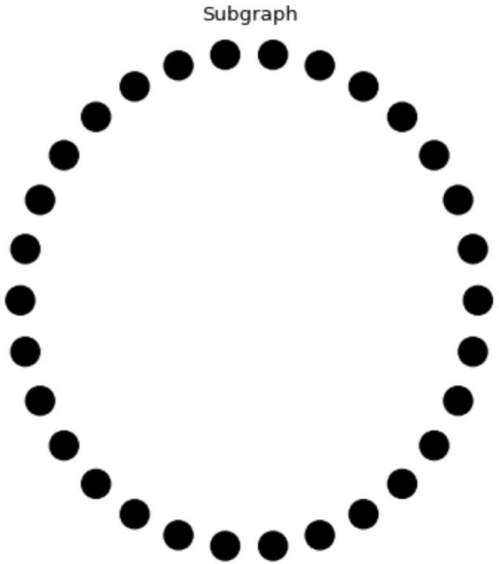
ROADMAP

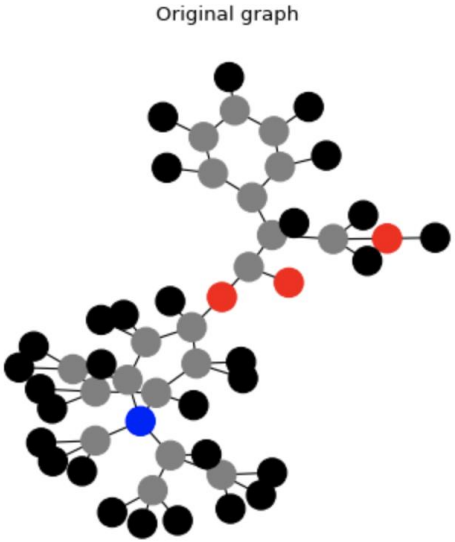
1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

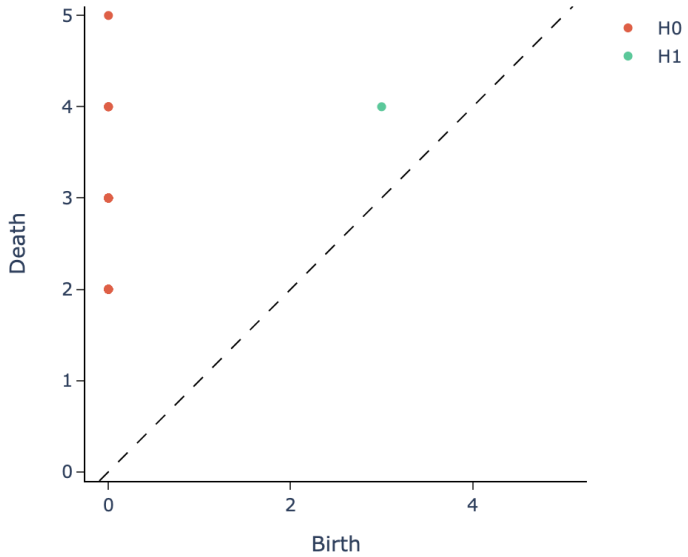
NOVARTIS | Reimagining Medicine

ATOMIC MASS FILTRATION (SUBLEVEL ORDER: 0)

```
List of atoms in subgraph 0:
['H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H',
'H', 'H', 'H', 'H', 'H', 'H', 'H']
```







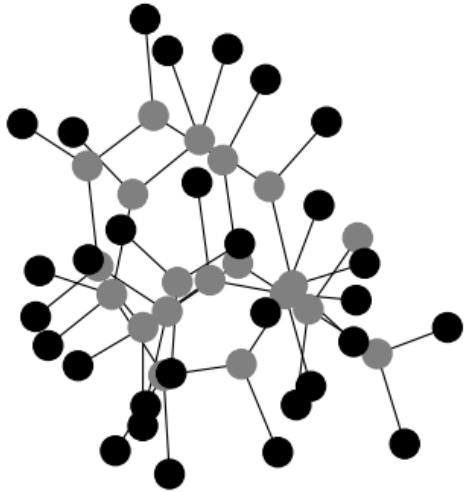

```
Feature matrix:  [[[29 29 29 29 29 18 18 18  4  4  1  1  0]
  [ 1  1  1  1  1  1  1  1  1  1  1  1  0]]]
Shape of the features matrix:  (1, 2, 13)
```


NOVARTIS | Reimagining Medicine

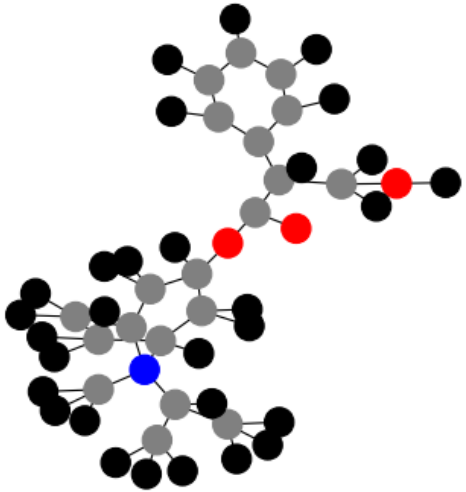
ATOMIC MASS FILTRATION (SUBLEVEL ORDER: 1)

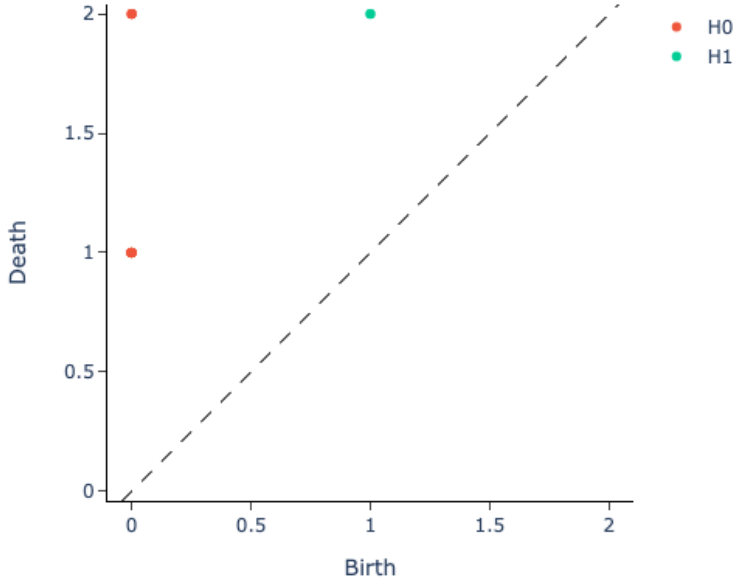
```
List of atoms in subgraph 1:
['C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'H', 'H', 'H',
'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H',
'H', 'H', 'H', 'H']
```

Subgraph



Original graph



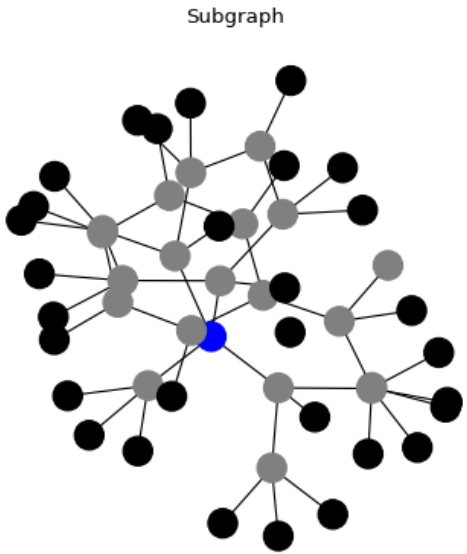


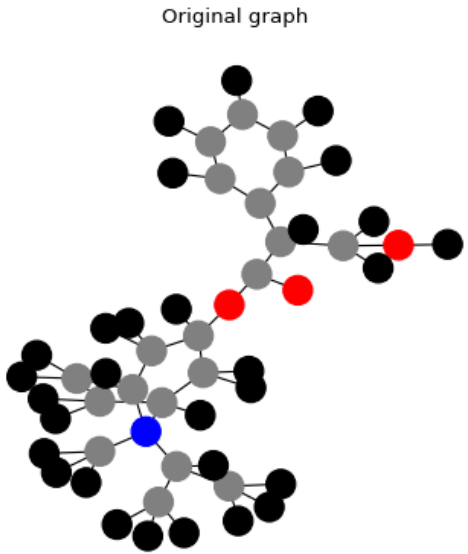

```
Feature matrix:  [[[49 49 49 49 49 49  4  4  4  4  4  4  0]
  [ 2  2  2  2  2  2  2  2  2  2  2  2  0]]]
Shape of the features matrix:  (1, 2, 13)
```

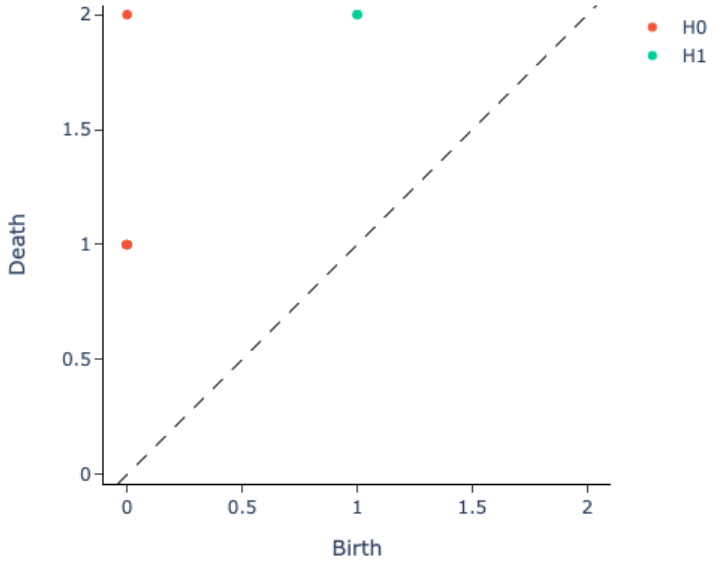

NOVARTIS | Reimagining Medicine

ATOMIC MASS FILTRATION (SUBLEVEL ORDER: 2)

```
List of atoms in subgraph 2:
['C', 'N', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'H', 'H',
'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H',
'H', 'H', 'H', 'H', 'H']
```







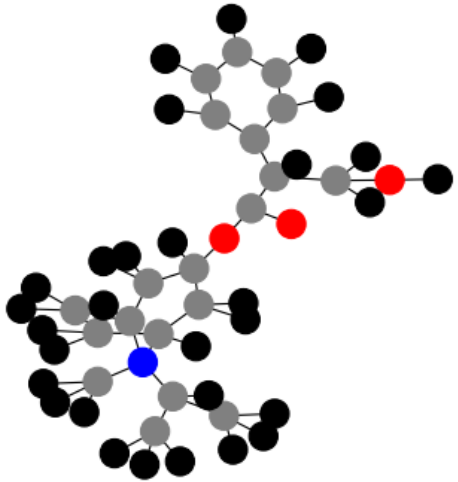

```
Feature matrix:  [[[50 50 50 50 50 50  2  2  2  2  2  2  0]
  [ 3  3  3  3  3  3  3  3  3  3  3  3  0]]]
Shape of the features matrix:  (1, 2, 13)
```


NOVARTIS | Reimagining Medicine

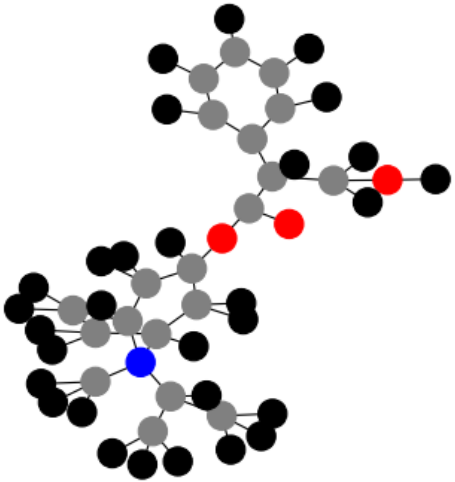
ATOMIC MASS FILTRATION (SUBLEVEL ORDER: 3-9)

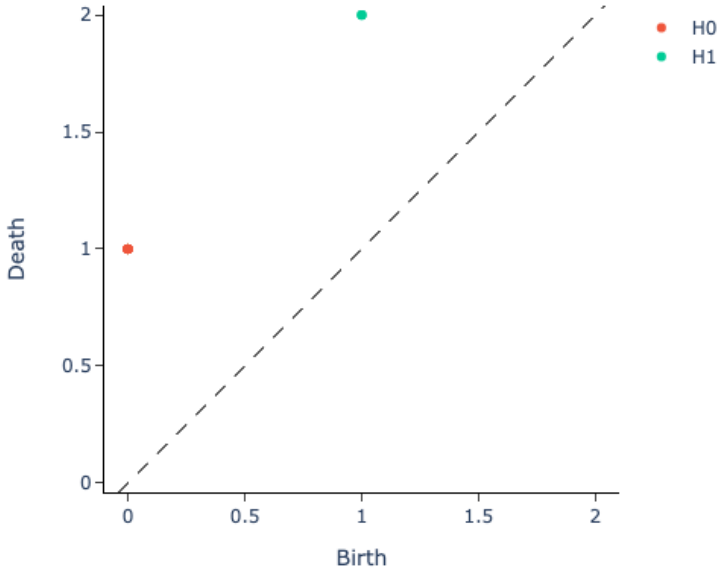
```
List of atoms in subgraph 3:
['C', 'N', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'O', 'C', 'C', 'O', 'C', 'C', 'C', 'C', 'O', 'C', 'C',
'C', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H',
'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H']
```

Subgraph



Original graph



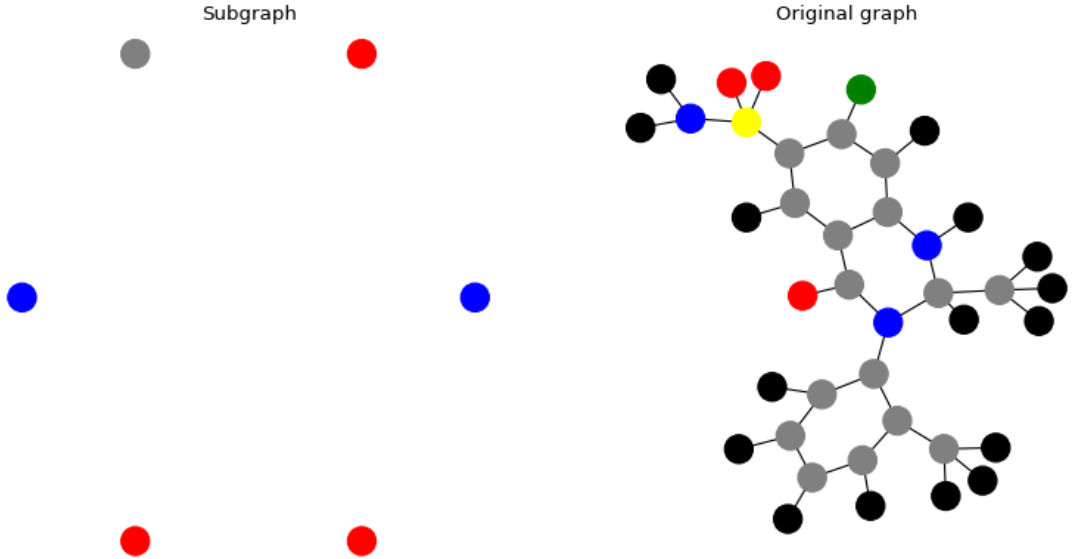


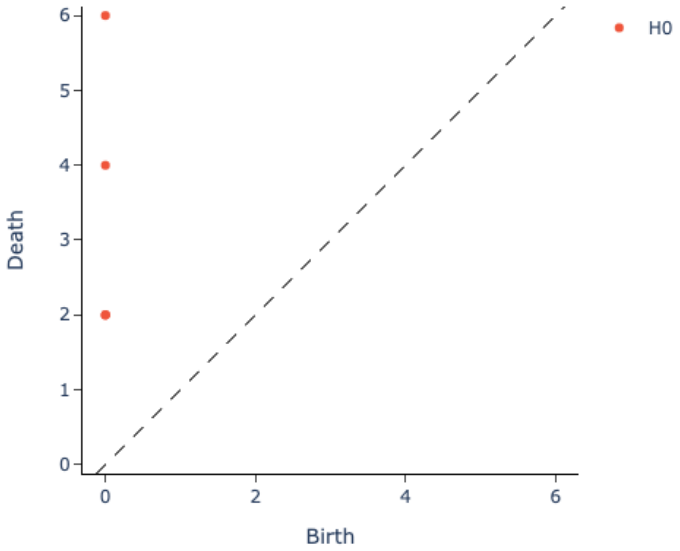

```
Feature matrix:  [[[53 53 53 53 53 53 53 53 53 53 53 53  0]
  [ 3  3  3  3  3  3  3  3  3  3  3  3  0]]]
Shape of the features matrix:  (1, 2, 13)
```


NOVARTIS | Reimagining Medicine

PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 0)

```
List of atoms in subgraph 0:
['N', 'O', 'C', 'N', 'O', 'O']
```



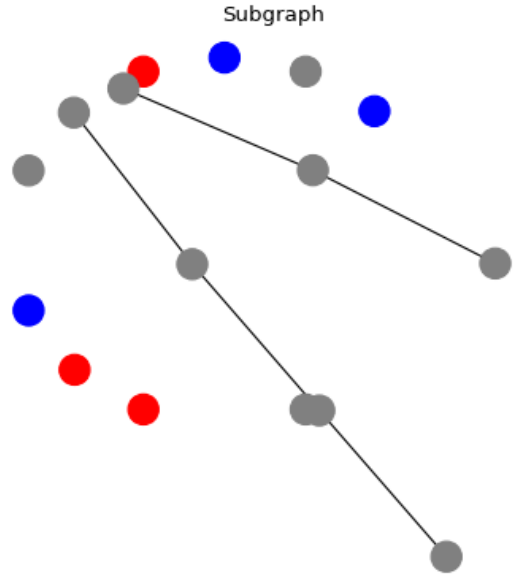


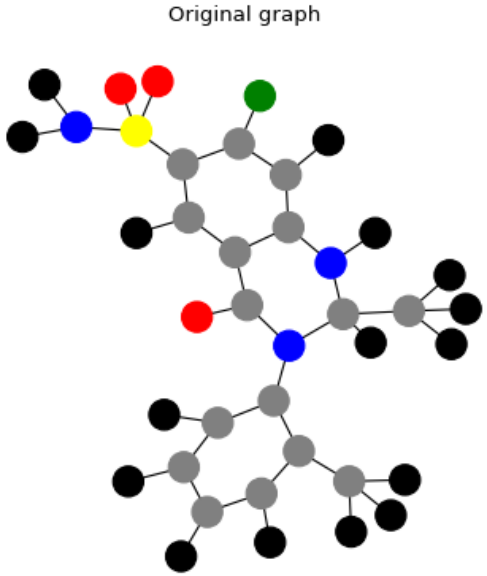

```
Feature matrix:  [[[5 5 5 5 2 2 2 2 1 1 1 1 0]
  [0 0 0 0 0 0 0 0 0 0 0 0 0]]]
Shape of the features matrix:  (1, 2, 13)
```

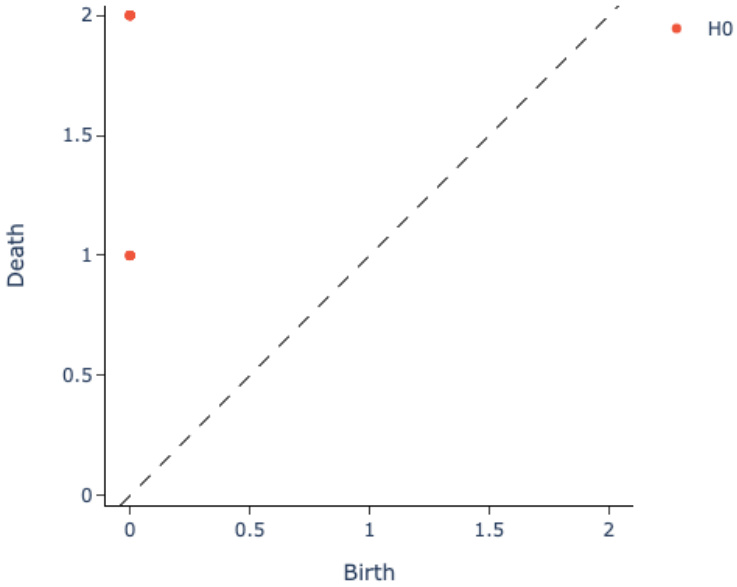

NOVARTIS | Reimagining Medicine

PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 1)

```
List of atoms in subgraph 1:
['C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C']
```





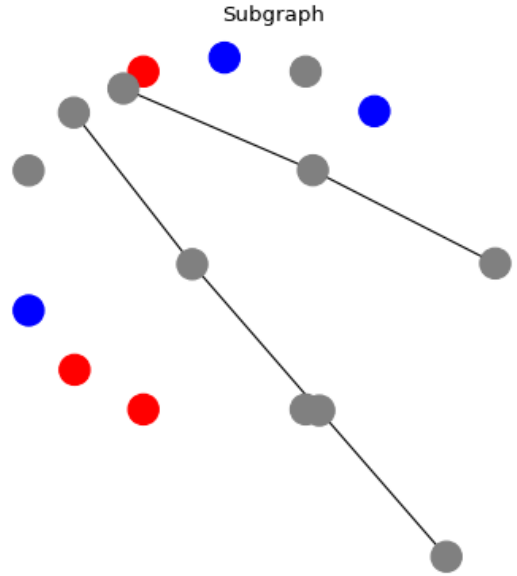


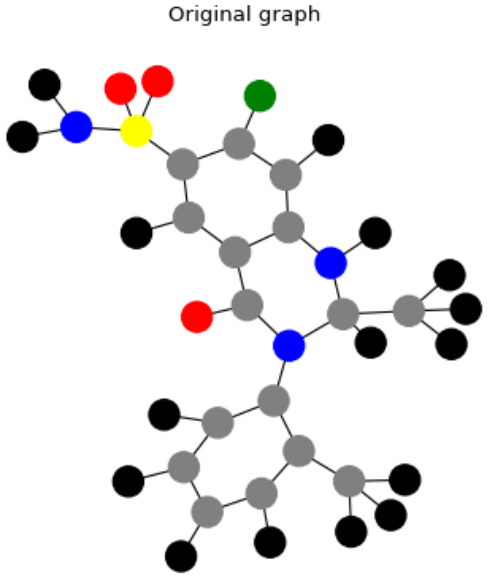

```
Feature matrix:  [[[15 15 15 15 15 15 10 10 10 10 10 10  0]
  [ 0  0  0  0  0  0  0  0  0  0  0  0  0]]]
Shape of the features matrix:  (1, 2, 13)
```

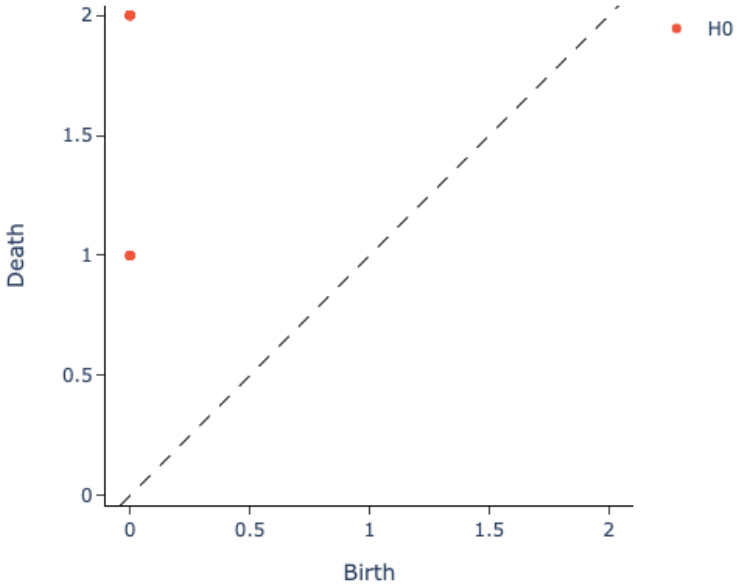

NOVARTIS | Reimagining Medicine

PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 2)

```
List of atoms in subgraph 1:
['C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C']
```







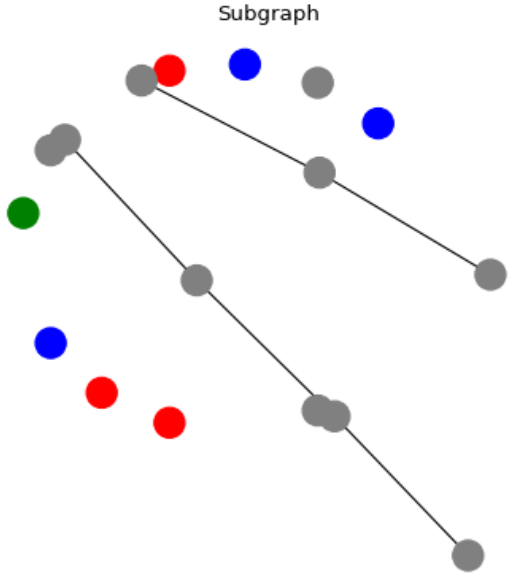

```
Feature matrix:  [[[15 15 15 15 15 15 10 10 10 10 10 10  0]
  [ 0  0  0  0  0  0  0  0  0  0  0  0  0]]]
Shape of the features matrix:  (1, 2, 13)
```

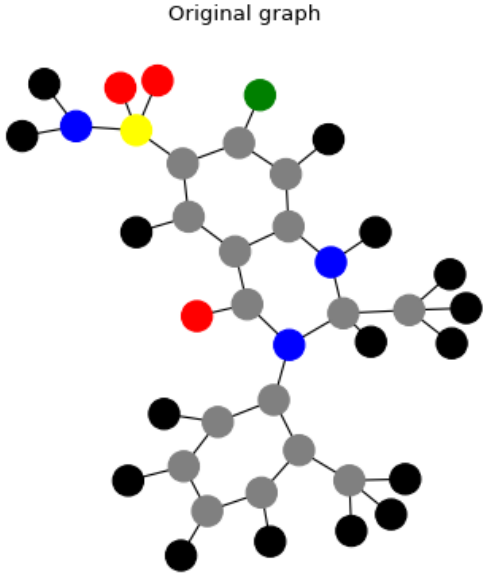
Same as Level 1

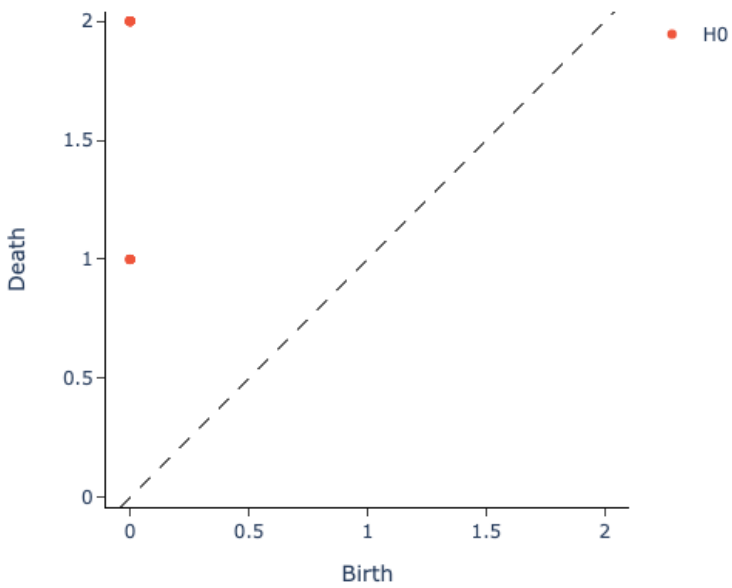
NOVARTIS | Reimagining Medicine

PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 3)

```
List of atoms in subgraph 3:
['C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'Cl', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C']
```





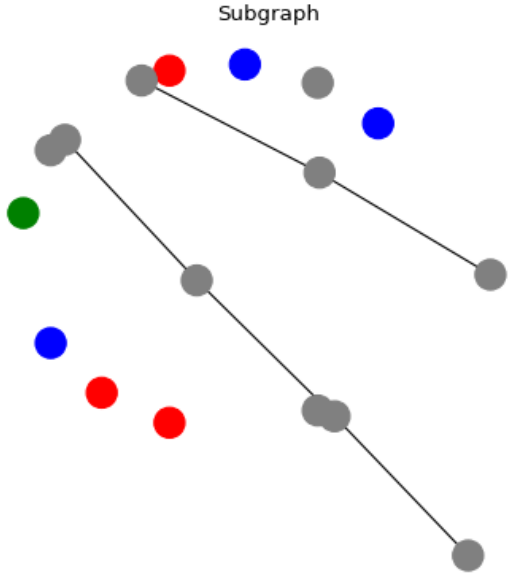


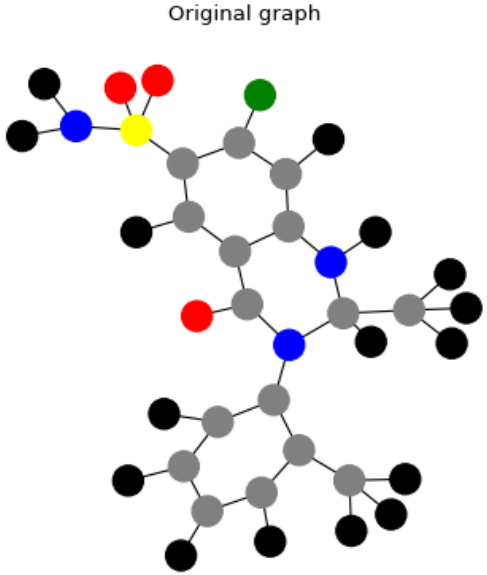

```
Feature matrix:  [[[16 16 16 16 16 16 11 11 11 11 11 11  0]
  [ 0  0  0  0  0  0  0  0  0  0  0  0  0]]]
Shape of the features matrix:  (1, 2, 13)
```

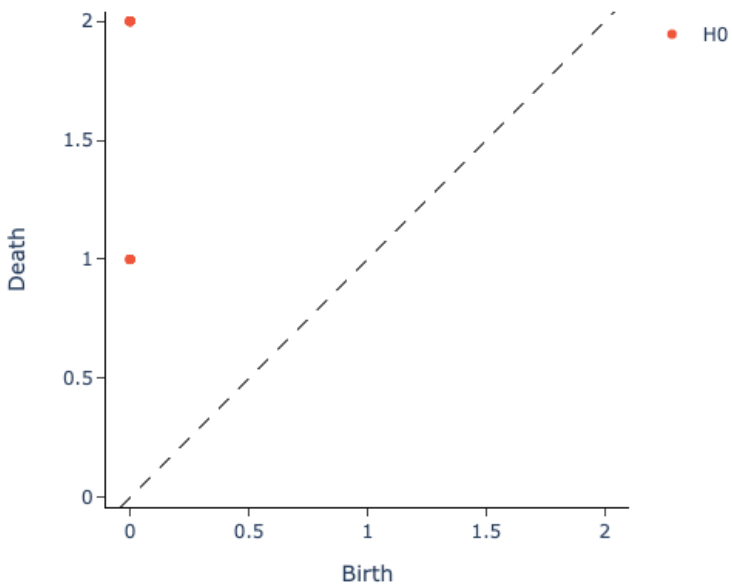

NOVARTIS | Reimagining Medicine

PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 4)

```
List of atoms in subgraph 3:
['C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'Cl', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C']
```







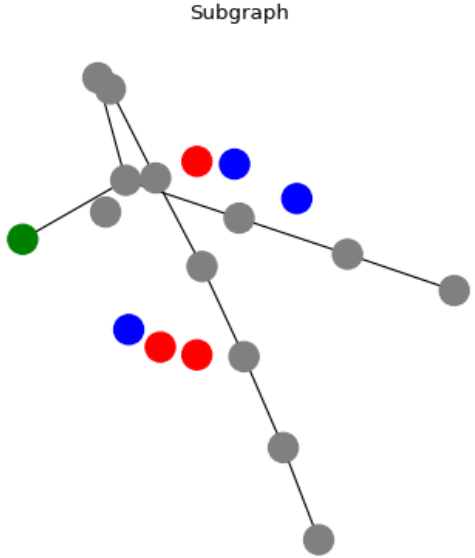

```
Feature matrix:  [[[16 16 16 16 16 16 11 11 11 11 11 11  0]
  [ 0  0  0  0  0  0  0  0  0  0  0  0  0]]]
Shape of the features matrix:  (1, 2, 13)
```

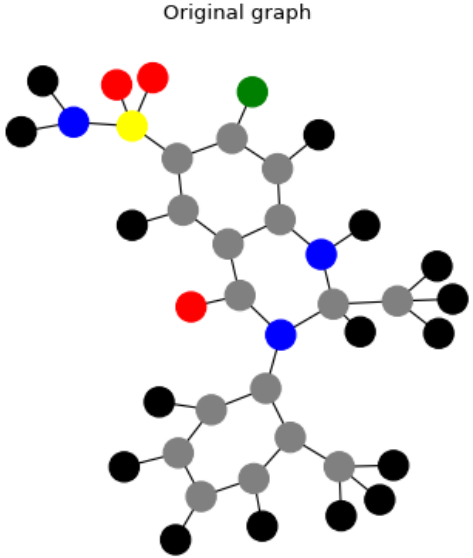
Same as Level 3

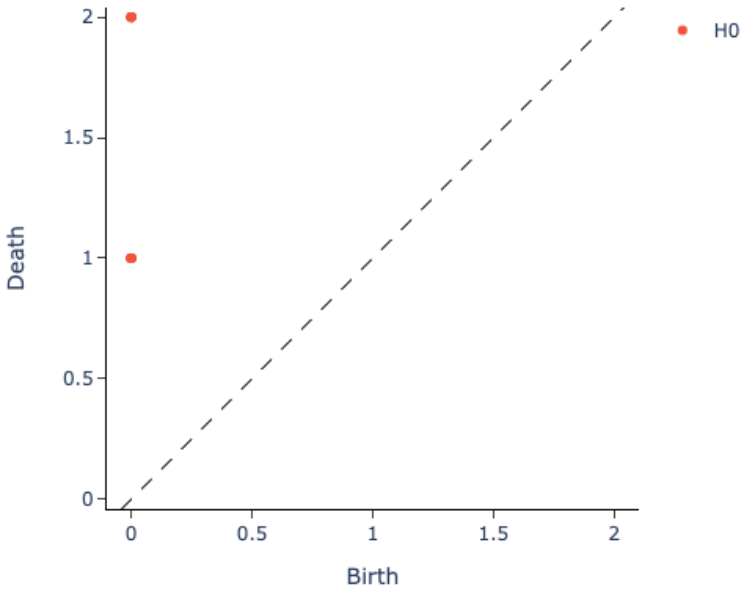
NOVARTIS | Reimagining Medicine

PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 5)

```
List of atoms in subgraph 5:
['C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'C', 'Cl', 'C', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C']
```





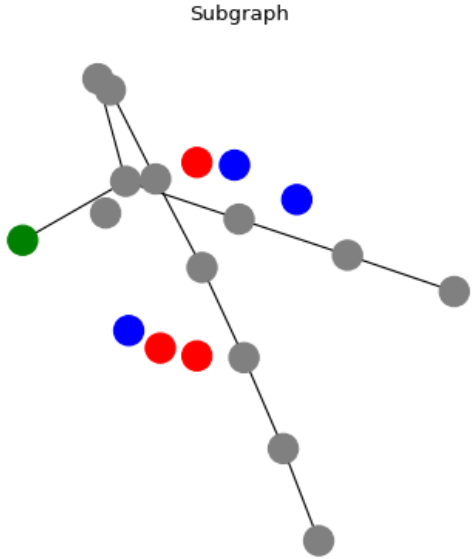


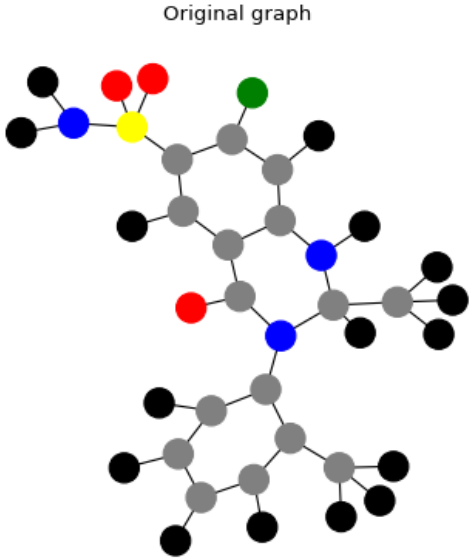

```
Feature matrix:  [[[18 18 18 18 18 18  8  8  8  8  8  8  0]
  [ 0  0  0  0  0  0  0  0  0  0  0  0  0]]]
Shape of the features matrix:  (1, 2, 13)
```

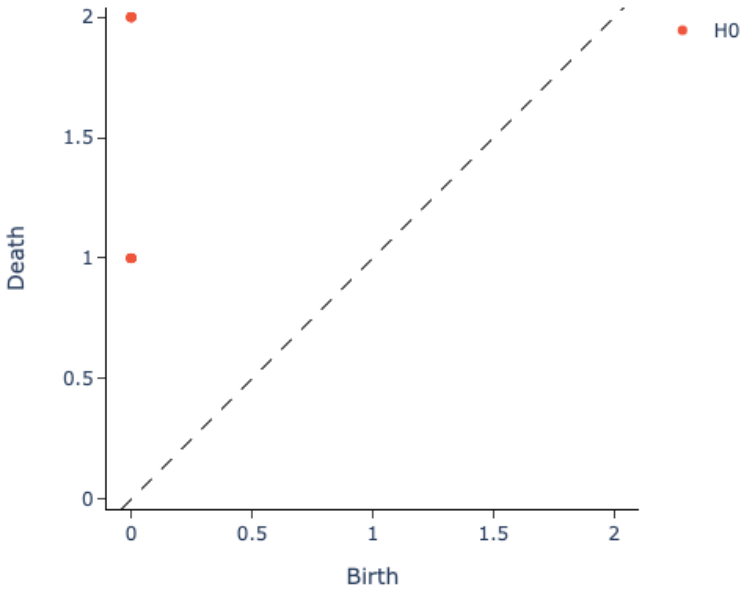

NOVARTIS | Reimagining Medicine

PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 6)

```
List of atoms in subgraph 5:
['C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'C', 'Cl', 'C', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C']
```








```
Feature matrix:  [[[18 18 18 18 18 18  8  8  8  8  8  8  0]
  [ 0  0  0  0  0  0  0  0  0  0  0  0  0]]]
Shape of the features matrix:  (1, 2, 13)
```

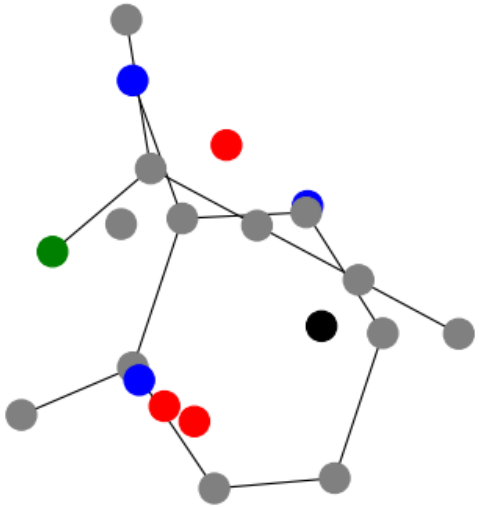
Same as Level 5

NOVARTIS | Reimagining Medicine

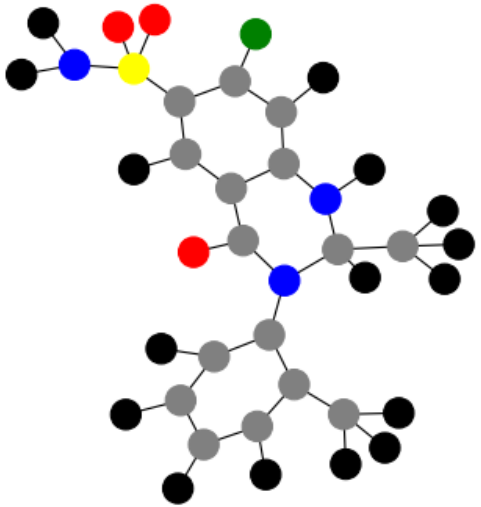
PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 7)

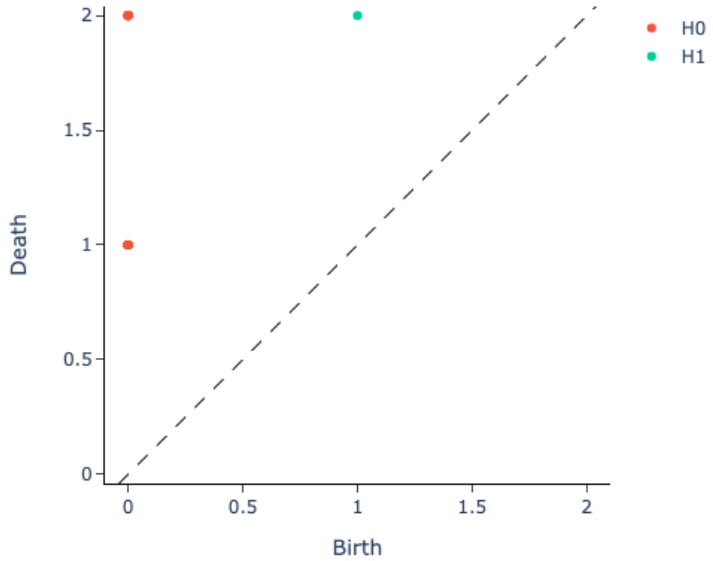
```
List of atoms in subgraph 7:
['C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'C', 'C', 'Cl', 'C', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C', 'H']
```

Subgraph



Original graph





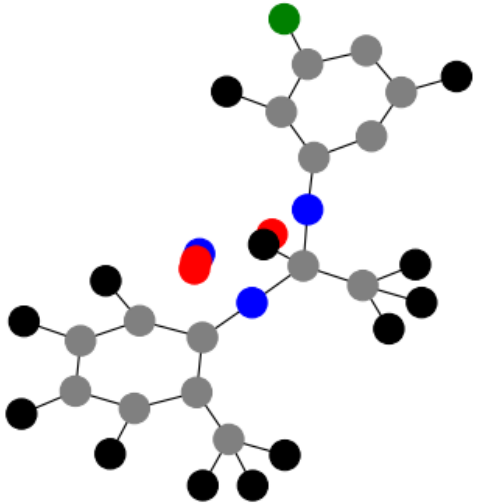

```
Feature matrix:  [[[20 20 20 20 20 20  8  8  8  8  8  8  0]
  [ 1  1  1  1  1  1  1  1  1  1  1  1  0]]]
Shape of the features matrix:  (1, 2, 13)
```


NOVARTIS | Reimagining Medicine

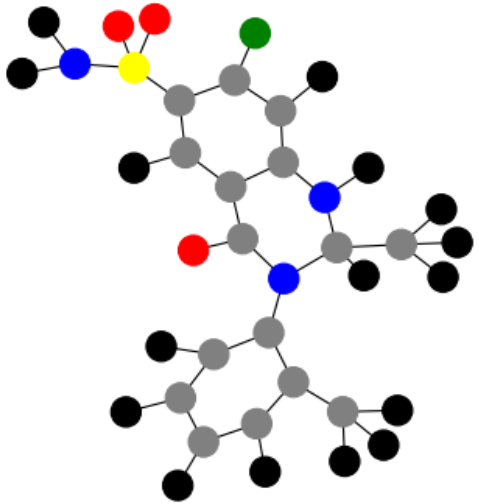
PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 8)

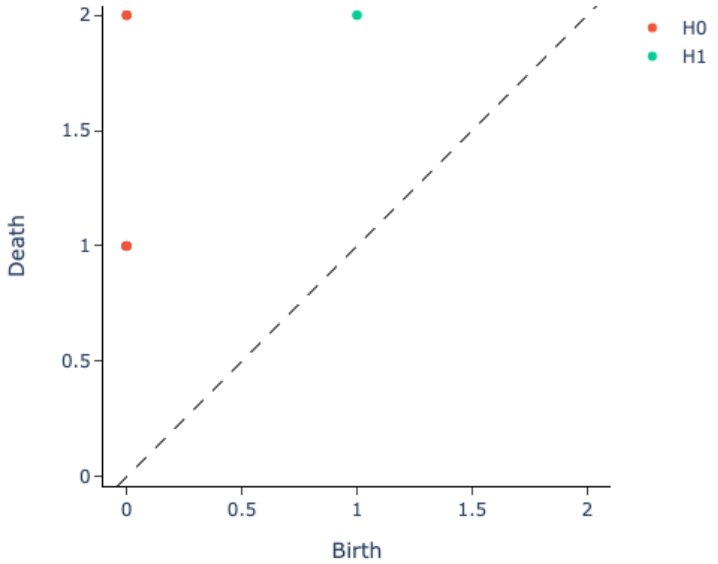
```
List of atoms in subgraph 8:
['C', 'C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'C', 'C', 'C', 'Cl', 'C', 'C', 'N', 'O', 'O', 'C', 'C', 'C', 'C', 'H',
'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H']
```

Subgraph



Original graph





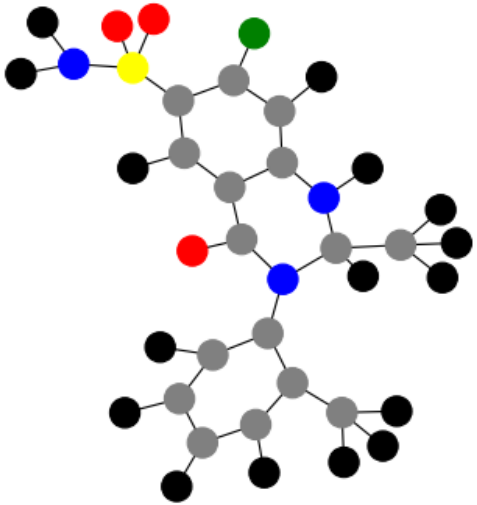

```
Feature matrix:  [[[34 34 34 34 34 34  4  4  4  4  4  4  0]
  [ 2  2  2  2  2  2  2  2  2  2  2  2  0]]]
Shape of the features matrix:  (1, 2, 13)
```


NOVARTIS | Reimagining Medicine

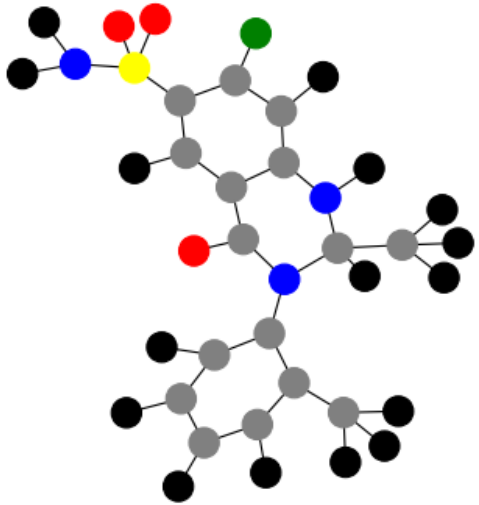
PARTIAL CHARGE FILTRATION (SUBLEVEL ORDER: 9)

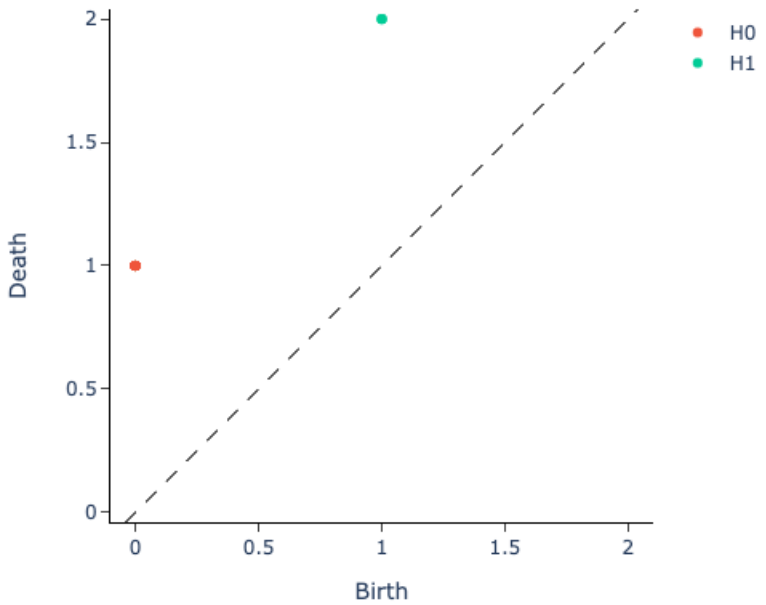
```
List of atoms in subgraph 9:
['C', 'C', 'C', 'C', 'N', 'C', 'N', 'O', 'C', 'C', 'C', 'C', 'S', 'C', 'Cl', 'C', 'C', 'N', 'O', 'O', 'C', 'C', 'C',
'C', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H', 'H']
```

Subgraph



Original graph






```
Feature matrix:  [[[39 39 39 39 39 39 39 39 39 39 39 39  0]
  [ 3  3  3  3  3  3  3  3  3  3  3  3  0]]]
Shape of the features matrix:  (1, 2, 13)
```


NOVARTIS | Reimagining Medicine

ROADMAP

1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

NOVARTIS | Reimagining Medicine

ML PIPELINE

- We construct different ToDD models, namely: ToDD-ViT, ToDD-ConvNeXt and ToDD-RF to test the generalizability and scalability of topological features.
- ToDD-ViT and ToDD-ConvNeXt are Triplet network architectures with Vision Transformer (ViT_b_16) and ConvNeXt_tiny models pretrained on ILSVRC-2012 ImageNet, serving as the backbone of the Triplet network.
- MP signatures of compounds are applied **nearest neighbour interpolation** to increase their resolutions to 224^2, followed by normalization.
- We only use **GaussianBlur** with kernel size 5^2 and standard deviation 0.05 as a data augmentation technique.
- **Transfer learning via fine-tuning** 260 ViT_b_16 and ConvNeXt_tiny models using **Adam optimizer** with a **learning rate of 5e-4**, **no warmup or layerwise learning rate decay**, **weight decay of 1e-4, and a batch size of 64 for 10 epochs** led to significantly better performance in Enrichment Factor and AUC scores compared to training from scratch.
- The performance of all models was assessed by 5-fold cross-validation (CV).

NOVARTIS | Reimagining Medicine

ML PIPELINE

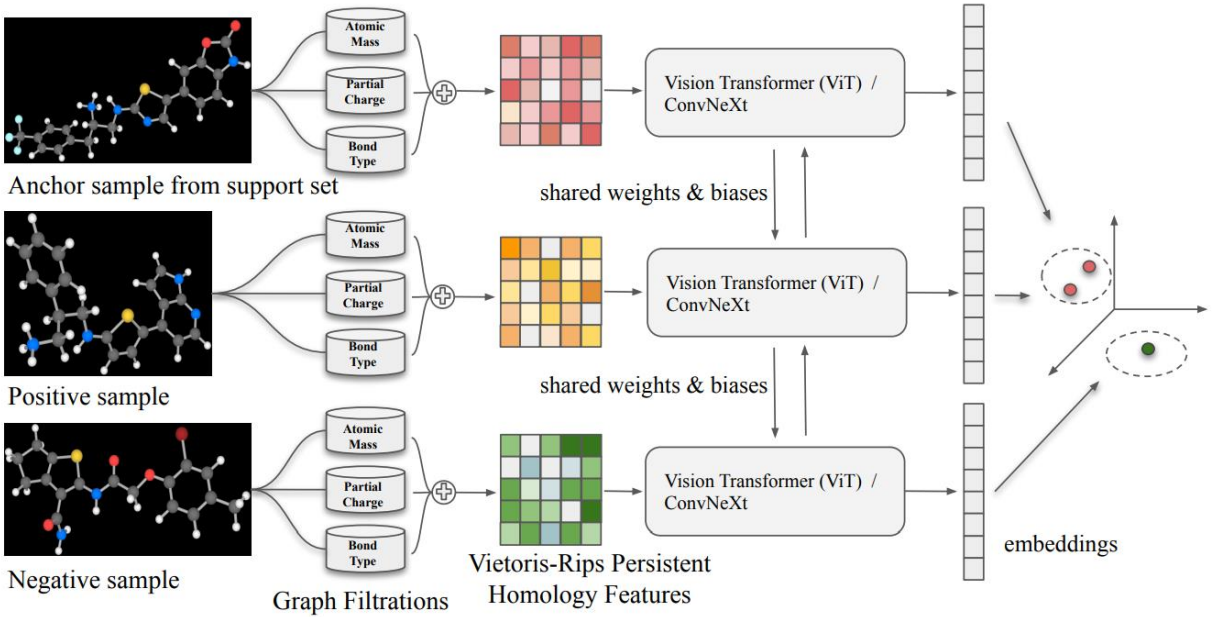
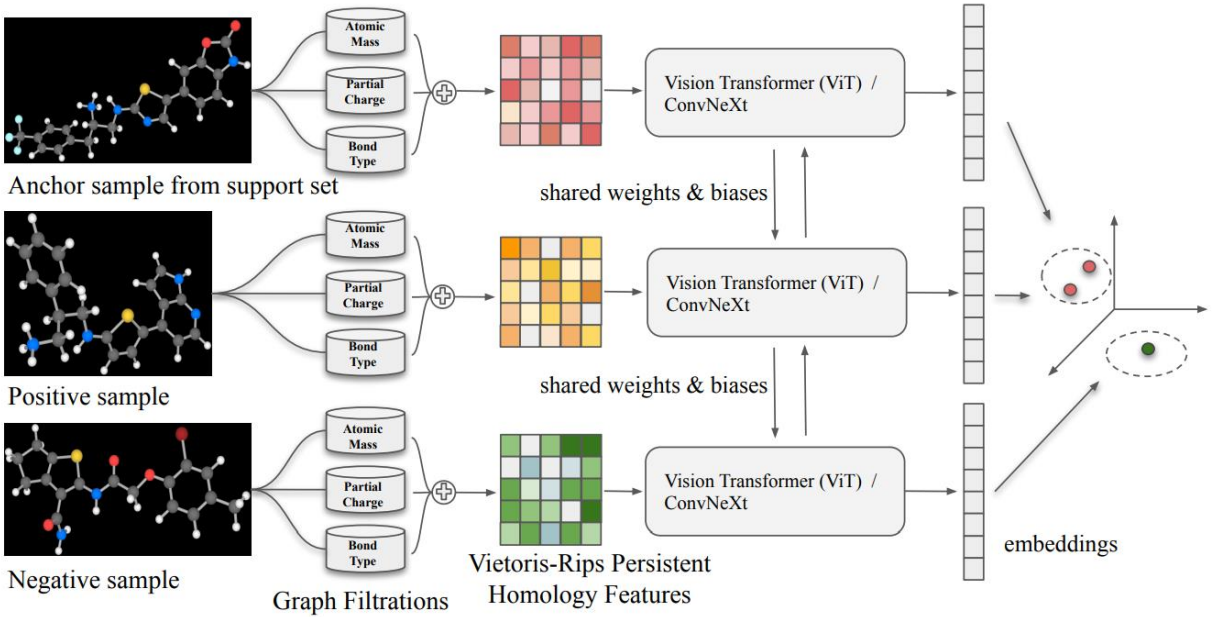


Figure 2: **End-to-end model pipeline.** Anchor sample, x, and positive sample, x^+, are compounds that can bind to the same drug target, whereas negative sample, x^-, is a decoy. $2D$ graph representation of each compound is decomposed into subgraphs induced by the periodic properties: atomic mass, partial charge and bond type. Potentially these domain functions can be augmented using other periodic properties such as ionization energy and electron affinity at the cost of computational complexity. Subgraphs may have isolated nodes and edges. Our MP framework establishes Vietoris-Rips complexes for each subgraph and provides MP signatures (topological fingerprints) of the compounds. Both ToDD-ViT and ToDD-ConvNeXt can encode the pair of distances between a positive query and a negative query against an anchor sample from the support set.

ML PIPELINE



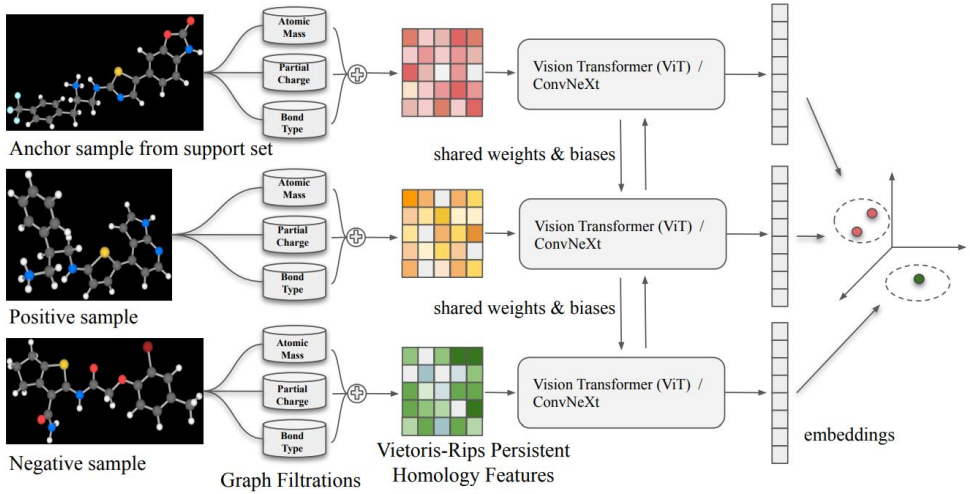
$$L(x, x^+, x^-) = \max(0, \alpha + \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^+)\|_p - \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^-)\|_p)$$

NOVARTIS | Reimagining Medicine

ML PIPELINE

- **Sampling Strategy**

- Learning metric embeddings via triplet margin loss on large-scale datasets poses a special challenge in sampling all distinct triplets (x, x^+, x^-), and collecting them into a single database causes excessive overhead in computation time and memory.



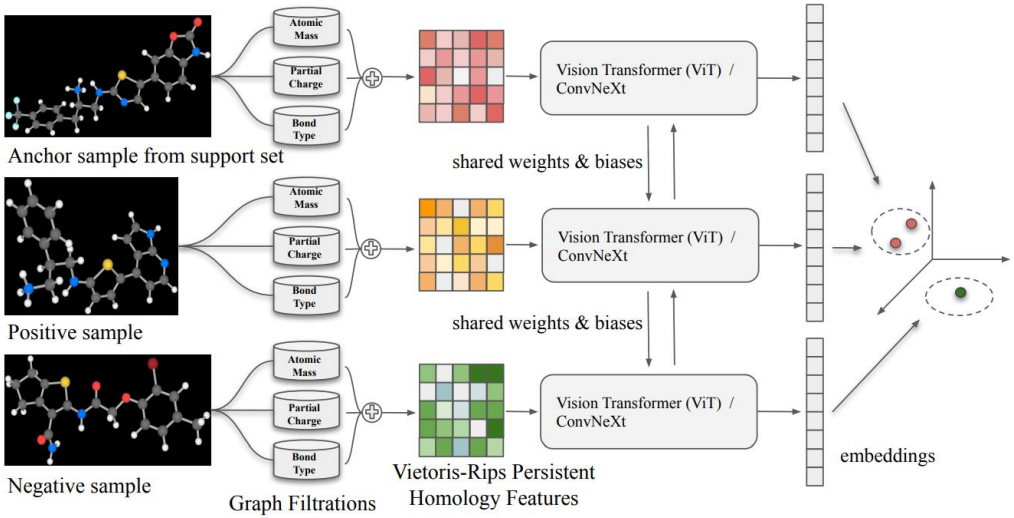
NOVARTIS | Reimagining Medicine

ML PIPELINE

- **Sampling Strategy**

- Since triplets (x, x^+, x^-) with $\mathbf{d(x, x^-) > d(x, x^+) + \alpha}$ have already negative queries sufficiently distant to the anchor compounds from the support set in the embedding space, they are not sampled to create the training dataset.

- We only sample triplets that satisfy $\mathbf{d(x, x^-) < d(x, x^+)}$ (where negative query is closer to the anchor than the positive) and $\mathbf{d(x, x^+) < d(x, x^-) < d(x, x^+) + \alpha}$ (where negative query is more distant to the anchor than the positive, but the distance is less than the margin).



NOVARTIS | Reimagining Medicine

ROADMAP

1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

NOVARTIS | Reimagining Medicine

RESULTS

- ToDD models consistently achieve the best performance on both Cleves-Jain and DUD-E Diverse datasets across all targets and $EF_{\alpha\%}$ levels.

- ToDD learns hierarchical topological representations of compounds using their atoms' periodic properties, and captures the complex chemical properties essential for high-throughput VS. These strong hierarchical topological representations enable ToDD to become a model agnostic method that is extensible to state-of-the-art neural networks as well as ensemble methods like random forests (RF).

- For small-scale datasets such as Cleves-Jain, RF is less accurate than ViT despite regularization by bootstrapping and using pruned, shallow trees, because small variations in the data may generate significantly different decision trees. For large-scale datasets such as DUD-E Diverse, ToDD-RF and ToDD-ConvNeXt exhibit comparable performances except for: CP3A4, GCR and HIVRT. We conclude that transformer-based models are more robust than convolutional models despite increased computation time.

NOVARTIS | Reimagining Medicine

RESULTS

Table 1: Comparison of EF 2%, 5%, 10% and AUC values between ToDD and other virtual screening methods on the Cleves-Jain dataset.

Model	EF 2% (max. 50)	EF 5% (max. 20)	EF 10% (max. 10)	AUC
USR [7]	10.0	6.2	4.1	0.76
GZD [83]	13.4	8.0	5.3	0.81
PS [42]	10.7	6.6	4.9	0.78
ROCS [36]	20.1	10.7	6.2	0.83
USR + GZD [75]	13.7	7.7	4.7	0.81
USR + PS [75]	13.1	7.9	5.0	0.80
USR + ROCS [75]	17.1	9.1	5.4	0.83
GZD + PS [75]	16.0	9.1	5.9	0.82
PH_VS [48]	18.6	NA	NA	NA
GZD + ROCS [75]	20.3	10.8	5.3	0.83
PS + ROCS [75]	20.5	10.7	6.4	0.83
ToDD-RF	35.2±2.3	15.6±1.0	8.1±0.4	**0.94**±0.02
ToDD-ViT	**39.6**±1.4	**18.6**±0.4	**9.9**±0.1	0.90±0.01
Relative gains	92.9%	83.7%	54.1%	13.3%

Relative gains are relative to the next best performing model. Mean and standard deviation of EF scores evaluated by 5-fold cross-validation.

NOVARTIS | Reimagining Medicine

RESULTS

Table 2: Comparison of EF 1% (max. 100) between ToDD and other virtual screening methods on 8 targets of the DUD-E Diverse subset.

Model	AMPC	CXCR4	KIF11	CP3A4	GCR	AKT1	HIVRT	HIVPR	Avg.
Findsite [90]	0.0	0.0	0.9	21.7	34.2	39.0	1.2	34.7	16.5
Fragsite [91]	4.2	42.5	0.0	32.9	29.1	47.1	2.4	48.7	25.9
Gnina [78]	2.1	15.0	38.0	1.2	39.0	4.1	11.0	28.0	17.3
GOLD-EATL [87]	25.8	20.0	33.5	17.9	34.6	29.2	28.7	23.4	26.6
Glide-EATL [87]	35.5	20.8	30.5	15.1	24.0	31.6	29.0	22.0	26.1
CompM [87]	32.3	25.0	35.5	33.6	37.1	44.2	30.2	25.0	32.9
CompScore [66]	39.6	51.6	51.3	14.0	27.1	37.6	21.8	18.2	32.7
CNN [68]	2.1	5.0	11.2	28.7	12.8	84.6	12.2	9.9	20.8
DenseFS [44]	14.6	5.0	4.3	44.3	20.9	89.4	12.8	8.4	25.0
SIEVE-Score [88]	30.7	61.1	53.4	6.7	33.3	42.1	39.8	38.3	38.2
DeepScore [85]	28.1	56.8	54.3	37.1	40.9	59.0	43.8	62.8	47.9
RF-Score-VSv3 [88]	32.3	60.9	4.5	25.9	32.5	41.9	39.8	65.7	37.9
ToDD-RF	42.9±4.5	**92.3**±3.2	**75.0**±5.0	**67.6**±3.4	**78.9**±4.0	**90.7**±1.3	**64.1**±2.3	**92.1**±1.5	**73.7**
ToDD-ConvNeXt	**46.2**±3.6	84.6±2.8	72.5±3.6	28.8±2.8	46.0±2.0	81.2±2.5	37.5±3.6	74.6±1.0	58.9
Relative gains	16.7%	51.1%	38.1%	52.6%	92.9%	1.5%	46.3%	40.2%	53.9%

Relative gains are relative to the next best performing model. Mean and standard deviation of EF scores evaluated by 5-fold cross-validation.

NOVARTIS | Reimagining Medicine

ROADMAP

1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

NOVARTIS | Reimagining Medicine

ABLATION STUDY

1. **Multimodal Learning**

- We first address the question of how adding different domain information improves the model performance.

- We demonstrate one-by-one the importance of each modality (atomic mass, partial charge and bond type) used for graph filtration to the classification of each target.

- We find that their importance varies across targets in a unimodal setting, but the orthogonality of these information sources offers significant gain in EF scores when the topological fingerprints learned from each modality are integrated into a joined multimodal representation.

NOVARTIS | Reimagining Medicine

ABLATION STUDY

2. Model Choice

- For small-scale datasets such as Cleves-Jain, RF is less accurate than ViT despite regularization by bootstrapping and using pruned, shallow trees, because small variations in the data may generate significantly different decision trees.

- For large-scale datasets such as DUD-E Diverse, ToDD-RF and ToDD-ConvNeXt exhibit comparable performances except for: CP3A4, GCR and HIVRT.

- We conclude that transformer-based models are more robust than convolutional models despite increased computation time.

NOVARTIS | Reimagining Medicine

ABLATION STUDY

2. Model Choice

- In order to effectively handle the large-scale datasets that have long-tailed distributions, we undersample from the majority class (decoys).
- Specifically, while training RF for the binary classification task on the drug targets of DUD-E Diverse, we use 80% of the active compounds and the same number of randomly chosen decoys for training.
- Undersampling decoys to avoid heavy class imbalance achieves better trade-offs between the accuracies of active compounds and decoys.

NOVARTIS | Reimagining Medicine

ABLATION STUDY

3. Network Architecture

- We investigated ways to leverage deep metric learning by architecting:
  i) a Siamese network trained with contrastive loss,
  ii) a Triplet network trained with triplet margin loss, and
  iii) a Triplet network trained with circle loss.

- Based on our preliminary experiments, the embeddings learned by i and iii provide sub-par results for compound classification, hence we use ii.

NOVARTIS | Reimagining Medicine

ABLATION STUDY

Table 5: EF 2% values and AUC scores across different modalities on Cleves-Jain dataset using **ToDD-RF**.

Target	Atomic Mass	Partial Charge	Bond Type	Atomic Mass & Partial Charge	All Modalities
a	33.3	33.3	33.3	33.3	41.7
b	25.0	29.5	31.8	27.3	25.0
c	19.2	7.7	15.4	26.9	34.6
d	33.3	33.3	41.7	50.0	50.0
e	30.0	30.0	30.0	40.0	40.0
f	25.0	50.0	37.5	50.0	37.5
g	30.0	30.0	30.0	30.0	40.0
h	40.0	50.0	30.0	50.0	50.0
i	40.0	40.0	30.0	40.0	40.0
j	17.9	39.3	35.7	28.6	28.6
k	21.4	21.4	17.9	35.7	32.1
l	15.0	15.0	15.0	30.0	25.0
m	44.4	50.0	33.3	50.0	38.9
n	15.0	25.0	10.0	25.0	10.0
o	21.7	20.0	25.0	23.3	23.3
p	10.9	8.7	13.0	17.4	26.1
q	45.5	27.3	22.7	40.9	40.9
r	42.9	42.9	42.9	39.3	32.1
s	26.7	16.7	20.0	20.0	30.0
t	30.0	50.0	50.0	50.0	50.0
u	33.3	38.9	27.8	38.9	50.0
v	21.4	28.6	28.6	28.6	28.6
Mean	28.3	31.3	28.3	35.2	35.2
AUC	0.92	0.90	0.88	0.94	0.93

ABLATION STUDY

Table 6: EF 2% values and AUC scores across different modalities on Cleves-Jain dataset using **ToDD-ViT**.

Target	Atomic Mass	Partial Charge	Bond Type	Atomic Mass & Partial Charge	All Modalities
a	25.0	33.3	33.3	33.3	50.0
b	20.5	6.8	34.1	6.8	34.1
c	11.5	15.4	23.1	7.7	46.2
d	25.0	41.7	50.0	33.3	50.0
e	40.0	20.0	30.0	20.0	30.0
f	25.0	25.0	37.5	25.0	50.0
g	20.0	30.0	40.0	30.0	50.0
h	40.0	50.0	50.0	50.0	50.0
i	30.0	20.0	20.0	20.0	50.0
j	10.7	14.3	21.4	17.9	21.4
k	21.4	21.4	25.0	21.4	39.3
l	15.0	30.0	30.0	40.0	35.0
m	22.2	44.4	22.2	44.4	50.0
n	10.0	25.0	15.0	20.0	35.0
o	20.0	16.7	16.7	18.3	20.0
p	26.1	17.4	26.1	15.2	32.6
q	36.4	36.4	50.0	36.4	18.2
r	14.3	17.9	42.9	21.4	32.1
s	30.0	13.3	23.3	13.3	33.3
t	20.0	30.0	50.0	30.0	50.0
u	33.3	38.9	38.9	33.3	50.0
v	21.4	28.6	28.6	21.4	42.9
Mean	23.5	26.2	32.2	25.4	39.5
AUC	0.87	0.85	0.86	0.84	0.90

ABLATION STUDY

Table 7: EF 1% values and AUC scores across different modalities on DUD-E Diverse using **ToDD-RF**.

	Atomic Mass		Partial Charge		Bond Type		Atomic Mass & Partial Charge		All Modalities	
Model	**EF %1**	**AUC**	**EF %1**	**AUC**	**EF %1**	**AUC**	**EF %1**	**AUC**	**EF %1**	**AUC**
AMPC	42.9	0.90	42.9	0.92	28.6	0.84	42.9	0.84	28.6	0.86
CXCR4	84.6	0.98	76.9	0.99	84.6	0.97	92.3	0.99	92.3	0.99
KIF11	55.0	0.96	55.0	0.98	70.0	0.97	70.0	0.98	75.0	0.98
CP3A4	54.1	0.96	48.6	0.87	40.5	0.93	62.2	0.95	67.6	0.96
GCR	54.4	0.96	43.9	0.95	57.9	0.97	63.2	0.97	78.9	0.97
AKT1	62.8	0.97	86.0	0.99	79.1	0.98	81.4	0.98	90.7	0.99
HIVRT	53.1	0.90	46.9	0.93	64.1	0.97	54.7	0.91	64.1	0.95
HIVPR	85.0	0.99	86.4	0.99	78.6	0.99	87.9	0.99	92.1	0.99
Mean	61.5	0.95	60.8	0.95	62.9	0.95	69.3	0.95	73.7	0.96

Table 8: EF 1% values and AUC scores across different modalities on DUD-E Diverse using **ToDD-ConvNeXt**.

	Atomic Mass		Partial Charge		Bond Type		Atomic Mass & Partial Charge		All Modalities	
Model	**EF %1**	**AUC**	**EF %1**	**AUC**	**EF %1**	**AUC**	**EF %1**	**AUC**	**EF %1**	**AUC**
AMPC	30.8	0.90	15.4	0.83	30.8	0.73	38.5	0.92	46.2	0.81
CXCR4	52.0	0.98	44.0	0.92	32.0	0.94	60.0	0.97	84.0	0.99
KIF11	47.5	0.96	45.0	0.88	37.5	0.92	60.0	0.96	72.5	0.97
CP3A4	19.2	0.86	17.8	0.86	15.1	0.86	28.8	0.90	28.8	0.91
GCR	25.7	0.95	30.1	0.95	19.5	0.94	43.4	0.96	46.0	0.97
AKT1	60.0	0.91	51.8	0.91	41.2	0.96	77.6	0.99	81.2	0.98
HIVRT	26.6	0.93	21.9	0.89	17.2	0.94	35.9	0.94	37.5	0.95
HIVPR	65.6	0.98	50.9	0.97	45.5	0.94	70.3	0.99	74.6	0.99
Mean	40.9	0.93	34.6	0.90	29.9	0.90	51.8	0.95	58.8	0.95

RTIS | Reimagining Medicine

ROADMAP

1. Key Contributions
2. Related Work (Mainly ECFP / Morgan Fingerprints)
3. Introduction to Persistent Homology
4. Datasets
5. Extracting Multiparameter Persistence Signatures of Compounds
6. ML Pipeline
7. Results
8. Ablation Study
9. Potential Future Work

NOVARTIS | Reimagining Medicine

POTENTIAL FUTURE WORK

1. Computing multiparameter persistent homology using **computationally cheap Clique Complex filtrations** instead of Vietoris-Rips complexes.

2. Testing the performance of ToDD on **ultra-large Virtual Screening datasets** with millions of compounds such as MUV, DUD-E and custom datasets of Novartis.

3. Using transfer learning to adapt state-of-the-art convolutional and transformer based computer vision models to extract complex chemical properties of compounds, specifically for few-shot learning problems.

4. There are other subdomains in chemistry that ToDD can be benchmarked and tested such as: **property and activity prediction** in addition to affinity of binding.

NOVARTIS | Reimagining Medicine