A Topological Molecular Representation for Molecular Machine Learning Based on the Euler Characteristic Transform

Victor Toscano-Duran[1][0009−0006−1316−9026] (✉) and
Bastian Rieck[2][0000−0003−4335−0302]

[1] Department of Applied Mathematics I, University of Seville, Seville, Spain
vtoscano@us.es
[2] AIDOS Lab, University of Fribourg, Fribourg, Switzerland
bastian.grossenbacher@unifr.ch

Abstract. The shape of a molecule determines its physicochemical and biological properties. However, it is often underrepresented in standard molecular representation approaches. Here, we propose using the *Euler Characteristic Transform* (ECT) as a geometrical-topological representation. Computed directly from molecular graphs constructed using handcrafted atomic features, the ECT enables the extraction of *multiscale* structural features, offering a novel way to encode molecular shape in the feature space. As a proof of concept, we evaluate the predictive performance of our topological descriptor on a regression task using a single benchmark dataset: the ADRA1A dataset, which focuses on predicting the inhibition constant K_i. Preliminary results show that our ECT-based descriptor achieves *competitive performance* compared with traditional molecular representations and methods, such as molecular fingerprints, descriptors and graph neural networks, outperforming them. More importantly, combining our topological descriptor with established representations, particularly with the AVALON fingerprint, further enhances predictive performance. These findings highlight the complementary value of *multiscale* topological information and its potential for being combined with established techniques.

Keywords: Euler Characteristic Transform · Topological Data Analysis · Molecular Machine Learning · Molecular Representation

1 Introduction

Understanding and predicting the properties of molecules is at the core of modern medicine and healthcare. Almost all pharmaceutical agents, like small-molecule drugs or diagnostic agents, are molecular in nature, and their therapeutic efficacy is critically dependent on their physicochemical and biological properties. From drug discovery to personalized medicine, the ability to predict how a molecule will behave in the body (its solubility, toxicity, bioavailability, and binding affinity), as well as predicting molecule properties, is essential for developing safe and effective treatments. Given the high costs and long timelines associated with traditional drug discovery, predictive modeling has emerged as a vital tool for improving the efficiency of the early stages of drug development, enabling faster screening of potential candidates and reducing the need for extensive experimental testing [4,9,32]. However, despite significant advancements, *predicting* molecular properties remains a challenge due to the complexity of molecular behavior. Traditional molecular representations, such as fingerprints or descriptors, often fail to fully capture the rich structure and shape information that influence a molecule's function in a biological context [8,33]. This gap is particularly pronounced when trying to predict properties like the inhibition constant (K_i), a key parameter in drug–target interactions, where even small changes in molecular shape can lead to large differences in binding affinity [10,30]. Recent developments in molecular machine learning (MML) have demonstrated the potential of deep learning techniques in general and graph-based models in particular to improve predictive performance [2,16]. However, many of these models still fail to fully capture the topological and geometric aspects of molecular shape, proving detrimental to the learning outcome. For example, methods like Graph Neural Networks (GNNs) [14,35] may capture *some* of the structural information, but they often overlook the importance of representing the shape in a comprehensive and multiscale manner [18].

Algebraic topology [15,24] and topological data analysis [12,21] can provide powerful tools to analyze the shape of molecular data, as shown by recent studies [28]. The *Euler Characteristic Transform* (ECT) [23,26], for example, could be a topological representation of molecules that captures the molecular shape by tracking changes in the Euler characteristic [20] across different scales and directions of the feature space. By incorporating topological shape information, the ECT allows us to represent and encode the molecular structure in a multiscale and novel way, capturing essential shape details that traditional representations may miss. Recent studies have shown the promise of the ECT in other fields, such as biology and material science, where shape plays a crucial role in determining functional properties [1,27].

This paper presents a preliminary version of our ongoing research. Our primary goal is to develop a topology-based representation of molecular graphs using the ECT. While this is a work in progress that is going to be submitted to a conference, more specifically to CIABiomed 2025[1], for thorough review and a more comprehensive presentation, our preliminary results on the prediction of a specific molecular property exhibits that our ECT-based descriptor approach already achieves competitive performance, even outperforming traditional molecular representations and methods. Moreover, our preliminary results shows that combining our ECT-based approach with established techniques, particularly with the AVALON fingerprint, can further boost predictive performance. This highlights the complementary nature of multiscale topological and geometric information in molecular machine learning.

[1] `https://2025.iabiomed.org/`

The remainder of this paper is organized as follows: Firstly, Section 2 introduces the foundational concepts relevant to our study, including molecular structures and their representations, traditional molecular descriptors, and the Euler Characteristic Transform. Then, in Section 3, the dataset used for this preliminary exploration is described, as well as the traditional representations included for comparison in the preliminary experiments, our proposed ECT-based approach, and the overall experimental setup. The results of our preliminary experiments, with a comprehensive discussion, are presented in Section 4. Finally, conclusions and future work are discussed in Section 5.

2 Background

Molecules [6] are the fundamental building blocks of chemical and biological systems. In the context of drug discovery [11], small molecules are designed or screened for their ability to modulate the activity of specific biological targets, typically proteins. The interaction between a drug and its target is governed by a complex interplay of properties, including molecular shape, electronic distribution, and physicochemical characteristics such as hydrophobicity, polarity, and charge. Accurately modeling and predicting these properties is a central challenge in cheminformatics and molecular machine learning [33].

In practice, molecules are often represented in formats that facilitate both human readability and algorithmic processing. One of the most widely used textual encodings is the SMILES (Simplified Molecular Input Line Entry System) notation, which encodes a molecule as a string describing the atoms and their connectivity through a series of characters and symbols. For example, the SMILES string `CC(O)=O` represents acetic acid. This format is compact, easily parsed, and widely supported in cheminformatics toolkits. However, SMILES do not directly convey geometric or spatial information, and small changes in the string can correspond to large structural differences. From molecular structures (e.g. SMILES, although alternative encodings exist), it is common to derive graph-based representations [14], where atoms are modeled as nodes and covalent bonds as edges, possibly enriched with additional features such as atomic types, bond orders, or aromaticity indicators. These molecular graphs serve as the foundation for numerous machine learning models, facilitating the use of graph-based algorithms and neural networks [9]. They also provide the basis for computing molecular representations such as fingerprints, descriptors, and our ECT-based representation. Fig. 1a. shows an example of the molecular graph of acetic acid derived from its SMILES string.

Traditionally, computational models for molecular machine learning tasks have relied on handcrafted molecular representations [33]. Among these are *molecular descriptors* (e.g., FGCount or 2DAP), which are numerical features derived from molecular graphs, such as atom counts, topological indices, or electronic properties, and *molecular fingerprints* (e.g., AVALON or MACCS), which represent the presence of specific substructures or chemical motifs as binary or count vectors. Both types of representations are computed directly from the molecular graph structure. These representations have been successfully applied in a variety of tasks such as quantitative structure–activity relationship (QSAR) modeling, molecule classification, and property prediction. However, they often suffer from a lack of expressiveness and poor generalization to out-of-distribution chemical spaces, especially in the presence of subtle variations in molecular geometry [3,8,34].

In recent years, graph-based models, particularly Graph Neural Networks[2] (GNNs) [14], have become a dominant paradigm for molecular property prediction. These models are a class of neural networks designed to operate on graphs-based data. The core mechanism of most GNNs is message passing, which iteratively propagates information across the graph elements. While GNNs can capture relational and structural information more flexibly than fixed descriptors, they often lack explicit access to multiscale shape information. Moreover, they may struggle to distinguish between molecules that are topologically or geometrically distinct but share similar local connectivity [16].

In this context, methods from *Topological Data Analysis* (TDA) [12] provide a complementary perspective. Instead than relying on raw atomic positions, handcrafted atomic fetures or purely local graph structures, TDA extracts global shape features from data, capturing relevant geometric information at multiple scales [21], making it thus particularly suitable for applications in the life sciences [31]. A particularly relevant tool within this framework is the *Euler Characteristic Transform* (ECT) [23,26,27], a method that combines ideas from algebraic topology with geometric data analysis. The ECT operates by *filtering* a shape, such as a molecular graph, across multiple scales and directions in the feature space, and computing the *Euler characteristic* [20], denoted by χ. This geometrical-topological quantity encodes information about the number of connected components, holes, and voids at each step (scale) of the filtration. Moreover, it is an *invariant*, i.e., it will remain unchanged under any smooth transformation applied to a shape. Put briefly, the Euler characteristic can be seen as a summary statistic of the shape of a graph or simplicial complex. To fully understand this tool and how molecules are represented as graphs, we next provide a more detailed introduction about graphs and simplicial complexes, serving as the starting point for computing the Euler characteristic and the ECT.

Graphs. Graphs are powerful mathematical tools for modeling real-world systems by focusing on *dyadic relationships* between elements. Formally, a *graph* $G = (V, E)$ consists of a finite set of vertices $V = \{v_1, v_2, \ldots, v_n\}$, which represent entities (e.g., atoms in a molecule), and a set of edges $E \subseteq \{\{u, v\} \mid u, v \in V \text{ and } u \neq v\}$, which represent pairwise relationships (e.g., chemical bonds). In cheminformatics, molecules are naturally represented by graphs, where nodes corresponds to atoms, typically encoded as a feature vector, and edges to chemical bonds. However, to extract richer geometric and topological information, we

[2] General introducion about graph theory, graph representation learning, and graph neuralnetworks can be found, for example, in [14].

can generalize graphs into structures called *simplicial complexes*. The graph of the acetid acid molecule shown in Fig. 1a (visualized in $2D$) contains 8 vertices, which corresponds to 2 carbon atoms, 2 oxygen atoms, and 4 hydrogen atoms and 7 edges, with $\{H, O\}$ being an example of an edge between two vertices (representing a chemical bond).

Simplicial complexes. An (abstract) simplicial complex K is a data structure for representing topological spaces, which generalizes a graph by permitting more than mere dyadic relations. It is defined as a family of sets (*simplices*) that is closed under taking subsets, meaning that if a set (like a triangle) is part of the complex, then so are all its faces (edges and vertices). More formally, a simplicial complex is obtained by a nested family of simplices, which are the elementary building blocks, for example: a 0-simplex can be thought of as a point (vertex), a 1-simplex as an edge, a 2-simplex as a filled triangle, and a 3-simplex as a filled tetrahedron. Each k-simplex has $k+1$ faces obtained by removing one of the vertices. For example, the acetic acid graph of Fig. 1a can also be seen as a simplicial complex with 8 0-simplices and 7 1-simplices.

The *Euler characteristic* is a key topological invariant of a simplicial complex K, being defined as:

$$\chi(K) = \sum_{k=0}^{n} (-1)^k |K^{(k)}|, \tag{1}$$

where $|K^{(k)}|$ denotes the number (cardinality) of k-simplices in the simplicial complex K. Hence, the Euler characteristic is an alternating sum of the number of simplices (elements) in each dimension. For simple graphs, such as molecular ones, which only consist of 0-simplices (vertices) and 1-simplices (edges), this is reduced to:

$$\chi = |V| - |E|. \tag{2}$$

In particular, for graphs without cycles (i.e., trees), χ corresponds exactly to the number of connected components. More generally, for arbitrary graphs, including molecular graphs with ring structures, the Euler characteristic corresponds to the number of connected components minus the number of independent cycles. For example, the Euler characteristic of the acetic acid graph presented in Fig. 1a is 1 (8 vertices - 7 edges), which matches its number of connected components since the graph has one connected component and no cycles.

Taken on its own, the Euler characteristic lacks sufficient complexity to fully describe a shape, but if we think of computing it at different *scales* (thresholds), considering we have a dynamic object, which grows in the number of their components (vertices, edges, etc.) across time, we may observe significant changes on it. This leads to the concept of the *Euler Characteristic Curve* (ECC), which tracks how the topological complexity of a shape evolves over different scales. To formalize this dinamyc view and to compute the Euler characteristic at different scales, we use the notion of a filtration, leading to a *filtered simplicial complex.* More formally, a *filtered simplicial complex* is a collection of subcomplexes $\{K(t) \mid t \in \mathbf{R}\}$ of a simplicial complex K such that $K(t) \subseteq K(s)$ for

$t < s$ and there exists $t_{\max} \in \mathbf{R}$ such that $K(t_{\max}) = K$. The *filtration time* (or filtration value) of a simplex $\sigma \in K$ is the smallest t such that $\sigma \in K(t)$. To illustrate the concept of a filtration, consider the graph representation of the acetic acid molecule shown in the Fig 1b. Each node (atom) is assigned a scalar value, which in this case corresponds to its projection onto a fixed direction, which corresponds to the x-axis direction for that example (the red one shown in Fig. 1a). These scalar values determine the filtration times of the nodes: a node enters the filtration at the time equal to its value. Formally, this means we construct a sequence of subgraphs (subcomplexes), $K(t) \mid t \in \mathbf{R}$ such that each $K(t)$ contains all nodes and edges whose filtration time is less than or equal to t. Since an edge can only appear after both its incident nodes have appeared, the filtration is nested, we have $K(t) \subseteq K(s)$ for $t < s$. An illustrative example is shown in Fig. 1c, where the molecular graph of Fig. 1a evolves as the filtration parameter increases along the x-axis direction. At each step, new atoms (nodes) and bonds (edges) are incorporated according to their associated filtration values, as shown in Fig. 1b, progressively revealing the full molecular topology. From this point on, the focus will be on graphs, as molecules are naturally represented in this way. However, the simplicial complex framework helps us understand why tools like Euler characteristic, ECC, and ECT are meaningful and how graphs can be treated as multi-scale dynamic objects for topological data analysis.

Euler Characteristic Transforms. While the Euler Characteristic Curve (ECC) already provides a summary of how the topological complexity of a shape evolves with respect to a *single* filtration parameter, it is often insufficient to fully characterize a high-dimensional or intricate structure like a molecule. This is where the *Euler Characteristic Transform* (ECT) comes into play, making it possible to characterize shapes based on multiple filtrations, parameterized using a *direction vector*. Specifically, each direction provides a different view of the data, capturing how the topological features of the structure unfold when the complex is filtered using a different criterion. To better understand this, imagine a 2D or 3D object (like a molecule embedded in space). If we project the molecule along a certain direction, for example along the x-axis, we can define a filtration by sweeping a hyperplane orthogonally to that axis and including simplices as their associated values fall below a certain threshold. This process gives us an ECC in the x-direction. Now, if we repeat this process in another direction, for example along the y-axis, we will generally obtain a different ECC, since the structure of the molecule may appear differently from that angle. The key idea is that *each* direction typically provides complementary topological information. Thus, the ECT is formally defined as the collection of Euler characteristic curves obtained by filtering a shape along a family of directions, which for molecule graphs are sampled randomly in the high-dimensional node feature space. In other words, the ECT is built by stacking multiple ECCs, each corresponding to a distinct direction. For example, for a given molecule and its graph, the ECT produces a collection of curves (ECCs), each corresponding to a different direction in the node feature space. By aggregating all these curves into a single descriptor, the ECT captures how the topology of the molecule evolves as a function of spatial thresholds, providing a rich and compact descriptor of molecular shape. Refer to Fig. 1d for the ECCs of the molecule graph and along the directions shown in Fig. 1a, and to Fig. 1e for the resulting ECT of the acetic acid graph along these direction, built by stacking the ECCs shown in Fig. 1d, visualized as a matrix, where each columns corresponds to a direction, each row to a threshold level in the filtration process, and the color intensity encodes the Euler characteristic value. The resulting feature set derived from the ECT encodes both multiscale and directional information, resulting in a highly expressive representation that can be used in machine learning models as input features for both classification and regression tasks, and that can complement traditional descriptors and graph-based methods.

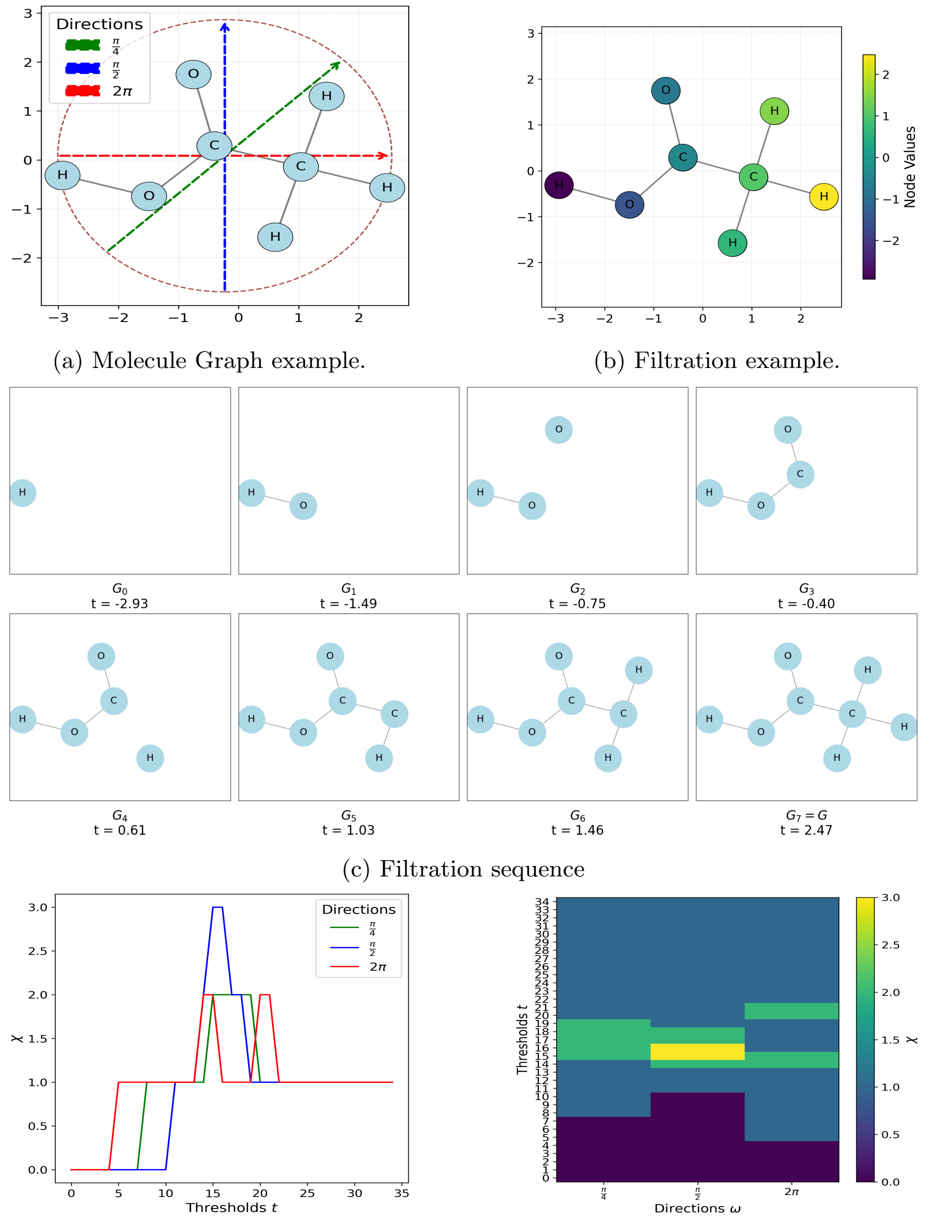
Hence, the ECT has been successfully applied for various machine-learning tasks [1,27].While the Euler characteristic itself has previously been used in molecular dynamics [29,19], to our knowledge, our work is the *first* to explore the use of the ECT in the context of molecular machine learning. In this study, we explore the potential of the ECT-based molecular representation in molecular machine learning, specifically focusing on predicting the inhibition constant (K_i), a continuous measure of binding affinity between a molecule and its biological target. Predicting K_i accurately is critical in drug design, as it can guides compound selection and optimization.

While there are compelling reasons to favor the ECT over other TDA methods, such as its invertibility [13] and computational efficiency, as it relies on a simple alternating sum of element counts, it is important to recognize that topological information can also be extracted using complementary tools from TDA, such as Betti numbers or persistence-based descriptors derived from persistent homology. These methods may be employed independently or in conjunction with the ECT to enrich the resulting topological representations. In our study, we have opted to use the ECT due to its favorable computational properties and its ability to capture multiscale and directional topological information. Nonetheless, a particularly promising direction for future work lies in the development of a more comprehensive topological descriptor that integrates multiple TDA tools. Such an approach could lead to more expressive, informative, and robust feature representations for downstream tasks.

3 Materials and Methods

Having described a novel and multiscale topology-based approach to represent molecules, we are interested in understanding how well our ECT-based representation performs in the context of molecule machine learning tasks. We conducted a preliminary experiment to explore and assess its effectiveness against established methods such as molecular fingerprints, descriptors and graph neural networks.

Dataset. On this paper, we concentrate on a single molecular dataset, focusing on predicting the constant inhibition K_i. Specifically, we use the alpha1A



(a) Molecule Graph example.

(b) Filtration example.

(c) Filtration sequence

(d) Euler Characteristic Curves.

(e) Euler Characteristic Transform.

Fig. 1: **Computing the Euler Characteristic Transform from molecular graphs**. (a): Exemplary 2D graph representation of the acetic acid molecule, derived from its SMILES string `CC(O)=O`. The graph contains 8 vertices (2 carbon, 2 oxygen, and 4 hydrogen atoms) and 7 edges, resulting in an Euler characteristic of 1 ($\chi = 1$). The three colored lines indicate the directions used to compute the ECCs and the corresponding ECT, as is shown in (d) and (e). (b): Filtration values of the nodes along the x-axis (red one in (a)) direction. (c) Filtration sequence along the x-axis direction. (d): ECCs computed for the different directions shown in (a). Each line color matches the directions shown in (a). In the figure, the x-axis represent the filtration thresholds (scales) values (35 in total), and the y-axis shows the corresponding Euler characteristic values. (e): ECT example, built by stacking the ECCs of (d) along the 3 directions shown in (a). Columns corresponds to directions, rows to thresholds, and color intensity encodes the Euler characteristic value.

adrenergic receptor (ADRA1A) dataset, which is sourced from BindingDB[3] and cover a variety of biologically relevant proteins, containing a total of 1959 molecules after preprocessing. In this step, molecules were retained only if they had a valid and unique SMILES representation, and if the associated K_i binding affinity value could be successfully parsed and converted to a numerical format. Molecules with missing or non-numeric K_i values were excluded. Additionally, duplicates based on identical SMILES strings were removed, keeping only the first occurrence.

Methods. We distinguish our description of the methods based on the type of molecular representation they use, since this directly influences the type of machine learning methods we can apply. Table 1 lists the different representations, as well as their category and dimensionality (i.e., the dimension of the resulting feature vector). Note that for graphs the concept of dimensionality as a feature vector is meaningless. In addition, note that there is an existing descriptor called *TOPO*[22], which is based in topological information but not in a multiscale sense.

In total, we have 5 different categories of representations, the molecular graph, fingerprints, descriptors, the ECT, and the ECT + fingerprint. The latter two denote our two novel ECT-based approaches. We denote by "ECT + fingerprint" a representation consisting of the ECT combined with the AVALON molecular fingerprints (via concatenation). This has been selected because the AVALON fingerprint is one the most commonly used fingerprints and generally performs well. In future work, we aim to study whether there are specific combinations of representations that can perform even better; please refer to our discussion in (Sec. 5).

When using the molecular graph representation, we use graph neural networks as the underlying machine learning models, since they are specialized for this type of data input. Concretely, we used two standard GNN models, a graph attention network (GAT) and a graph convolutional network (GCN), as well as an specialized graph neural network for molecular learning, named "AttentiveFP" [34]. For all other representations, we use an XGBoost [5] model, which has consistently shown strong performance in variety of tasks.

Experimental setup. Molecule graphs are extracted from the SMILES strings provided in the ADRA1A dataset using the `PyTorch Geometric`[3] Python package, which encoded nodes (atoms) as a 9-dimensional handcrafted feature vector. The ECT is then computed over this multidimensional feature space. Based on a prior sensitivity analysis, we fixed the number of directions and filtration thresholds to 158 and 16, respectively. This analysis revealed that the number of directions has a substantially *greater* impact on predictive performance than the number of filtration thresholds. For instance, increasing the number of directions from 20 to 30 led to noticeable improvements in model accuracy. However, performance gains plateaued at higher values, with little to no improvement observed between 180 and 200 directions. In contrast, varying

[3] `https://www.bindingdb.org/`

[3] `https://pytorch-geometric.readthedocs.io/`

the number of filtration thresholds had minimal effect on performance across a wide range of values. Therefore, we selected 158 directions to ensure sufficient expressiveness while limiting the number of thresholds to 16 to reduce computational cost without sacrificing predictive accuracy. The resulting ECT-based feature vector has a dimensionality of 2528, as summarized in Table 1. All ECTs have been computed using the `DECT` Python package [27].

To fairly evaluate the predictive performance of both traditional vector-based and ECT-based methods, we designed two experimental setups: one using XGBoost for vector representations (fingerprints, descriptors, and ECT), and another using graph neural networks (GNNs) operating directly on the molecular graphs. For vector-based representations, we employed the XGBoost regressor [5], trained with 1000 estimators, a learning rate of 0.01, and a maximum tree depth of 5. For the graph-based models, we explored three architectures: Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and the "AttentiveFP" model. The GCN was configured with two convolutional layers and 64 hidden channels. The GAT model used eight attention heads, each with eight hidden units. Both GCN and GAT architectures followed the design principles of Platonov et al. [25], including a two-layer perceptron after each neighborhood aggregation step, skip connections, and layer normalization. The "AttentiveFP" model was configured with 64 hidden units, a single output channel, four message-passing layers, and two attention-based update steps. Dropout was set to 0.2. All GNN models were trained for 100 epochs using the ADAM optimizer with a learning rate of $10^{-2.5}$ and a weight decay of 10^{-5}. A 10-fold cross-validation strategy was applied for all the methods with shuffling and a fixed random seed to ensure reproducibility. Model evaluation was conducted using two metrics: root mean squared error (RMSE), and coefficient of determination (R2).

4 Results

In this section, results are depicted. Table 1 shows the results for the different representations for the commented ADRA1A dataset. The results shows that the ECT on its own provides meaningful and effective features for predicting the constant inhibition K_i, outperforming all alternative methods. Moreover, its combination with traditional techniques, specifically with the AVALON molecular fingerprint, yields a substantial performance improvement over using ECT alone. These findings suggest that molecular data contains relevant shape information that can be effectively integrated with traditional approaches, highlighting its potential for generating generalizable features and enhancing predictive performance in molecular machine learning tasks.

Something we found particularly interesting is that, the most complex models, as GNNs are, that directly take the molecular graph as its input, have the worst performance, even considering the specific GNN for molecular machine learning, which is far from the performance of the best methods. This observation suggests that the use of complex deep learning models is not inherently

Table 1: Results (RMSE, reported as mean ± standard deviation across 10-fold cross-validation) for the ADRA1A dataset. Rows corresponds to different representation methods grouped by category (third column) (GNNs, fingerprints, descriptors, and ECT-based). The fourth column indicates the dimensionality of the resulting feature vector for each methods. Note that for graphs the concept of dimensionality as a feature vector is meaningless.

Method	RMSE	Representation Category	Dimension
AttentiveFP	2.01 ± 0.05	Graph	-
GAT	2.65 ± 0.17	Graph	-
GCN	2.75 ± 0.15	Graph	-
AVALON	1.81 ± 0.14	Fingerprint	1024
CATS2D	1.88 ± 0.09	Fingerprint	189
ECFP4	1.81 ± 0.11	Fingerprint	1024
EState	2.2 ± 0.16	Fingerprint	79
KR	1.82 ± 0.14	Fingerprint	4860
MACCS	1.88 ± 0.15	Fingerprint	166
MAP4	1.88 ± 0.15	Fingerprint	1024
Pharm2D	1.8 ± 0.11	Fingerprint	1024
PubChem	1.87 ± 0.13	Fingerprint	881
RDKit	1.83 ± 0.13	Fingerprint	1024
2DAP	1.84 ± 0.14	Descriptor	1596
ConstIdx	1.99 ± 0.17	Descriptor	50
FGCount	1.93 ± 0.16	Descriptor	153
MolProp	2.05 ± 0.05	Descriptor	14
RingDesc	2.12 ± 0.12	Descriptor	35
TOPO	1.89 ± 0.13	Descriptor	74
WalkPath	1.97 ± 0.13	Descriptor	46
ECT (ours)	**1.78 ± 0.12**	Topological	2528
ECT+ FP (ours)	**1.74 ± 0.12**	Topological	3552

justified by their performance in this context. Rather, it appears that the choice of molecular representation plays a much more critical role in determining model effectiveness, highlighting that more complex architectures do not necessarily lead to better results. This observation is in line with other studies and might be attributed to the limited availability of data in the biomedical domain.

5 Conclusion

In this paper, we propose and implement an effective new molecular representation method based on the Euler Characteristic Transform (ECT), which captures multiscale topological information relevant to molecular machine learning. Our preliminary results on the ADRA1A dataset show that our ECT-based representation provides meaningful and effective features, improving baseline methods performance for molecular property prediction tasks, specifically for predicting the constant inhibition K_i. Moreover, its combination with traditional approaches, specifically with AVALON molecular fingerprint, can improve prediction performance even further. This suggest that our multiscale topological approach based on the ECT, which can capture shape characteristics across different resolutions, encodes structural information, which were hitherto not being considered or fully exploited by existing methods, presenting a promising avenue for molecular machine learning research.

Future work. An exploration of *how* the number of directions and thresholds used in the computation of the ECT affects its representational power is needed. We believe that identifying optimal or data-adaptive strategies for selecting these parameters will lead to more expressive and discriminative topological signatures, thus potentially improving performance in molecular machine learning tasks. This question is not only relevant for cheminformatics and biomedical applications, but could also benefit a wide range of domains where data can be modeled geometrically. In addition, we plan on extend and test the proposed approach to other datasets and molecular machine learning tasks. Another promising avenue is to investigate whether specific combinations of the ECT with traditional molecular representations yield more accurate or robust models. Future studies could benchmark multiple hybrid strategies to identify synergies between topological and conventional representations. Moreover, a particularly exciting direction for future work involves the development of a more comprehensive topological descriptor that integrates information from multiple tools of TDA. Rather than relying solely on the ECT, this extended descriptor could incorporate additional topological summaries, such as Betti numbers or persistence-based descriptors from persistent homology. This may lead to more powerful, expressive, and robust representations. Finally, it will be interesting to extend this approach with different variants of the ECT, as the Weighted Euler Characteristic Transform (WECT) [17] or the Smooth Euler Characteristic Transform (SECT) [7].

Code Availability All data and code used to generate the results presented in this manuscript are publicly available in a GitHub repository[4].

Acknowledgments. This work was partially supported by REXASI-PRO H-EU project, call HORIZON-CL4-2021-HUMAN-01-01, Grant agreement ID: 101070028, by the Research Budget (VII PPIT-US) of the University of Seville, and by the Swiss State Secretariat for Education, Research, and Innovation (SERI).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amézquita, E.J., Quigley, M.Y., Ophelders, T., Landis, J.B., Koenig, D., Munch, E., Chitwood, D.H.: Measuring hidden phenotype: quantifying the shape of barley

[4] `https://github.com/victosdur/ECTforMoleculeMachineLearning-ADRA1A-MLGWorkshop.git`

seeds using the euler characteristic transform. in silico Plants **4**(1), diab033 (12 2021). https://doi.org/10.1093/insilicoplants/diab033

2. Atz, K., Grisoni, F., Schneider, G.: Geometric deep learning on molecular representations. Nature Machine Intelligence **3**(12), 1023–1032 (2021). https://doi.org/10.1038/s42256-021-00418-8
3. Baptista, D., Correia, J., Pereira, B., Rocha, M.: Evaluating molecular representations in machine learning models for drug response prediction and interpretability. Journal of Integrative Bioinformatics **19**(3), 20220006 (2022). https://doi.org/10.1515/jib-2022-0006
4. Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F.J., Carballal, A., Maojo, V., Pazos, A., Fernandez-Lozano, C.: A review on machine learning approaches and trends in drug discovery. Computational and structural biotechnology journal **19**, 4538–4558 (2021). https://doi.org/10.1016/j.csbj.2021.08.011
5. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785
6. Corey, E., Czakó, B., Kürti, L.: Molecules and Medicine. Wiley (2012), https://books.google.es/books?id=jz2GN6DYoOoC
7. Crawford, L., Monod, A., Chen, A.X., Mukherjee, S., Rabadán, R.: Functional data analysis using a topological summary statistic: the smooth euler characteristic transform. arXiv preprint arXiv:1611.06818 (2016), https://arxiv.org/abs/1611.06818v4
8. David, L., Thakkar, A., Mercado, R., Engkvist, O.: Molecular representations in ai-driven drug discovery: a review and practical guide. Journal of cheminformatics **12**(1), 56 (2020). https://doi.org/10.1186/s13321-020-00460-5
9. Deng, J., Yang, Z., Ojima, I., Samaras, D., Wang, F.: Artificial intelligence in drug discovery: applications and techniques. Briefings in Bioinformatics **23**(1), bbab430 (2022). https://doi.org/10.1093/bib/bbab430
10. Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., Wang, F.: A systematic study of key elements underlying molecular property prediction. Nature Communications **14**(1), 6395 (2023). https://doi.org/10.1038/s41467-023-41948-6
11. Drews, J.: Drug discovery: a historical perspective. science **287**(5460), 1960–1964 (2000). https://doi.org/10.1126/science.287.5460.1960
12. Edelsbrunner, H., Harer, J.L.: Computational topology: an introduction. American Mathematical Society (2010), https://webhomes.maths.ed.ac.uk/~v1ranick/papers/edelcomp.pdf
13. Ghrist, R., Levanger, R., Mai, H.: Persistent homology and euler integral transforms. Journal of Applied and Computational Topology **2**, 55–60 (2018). https://doi.org/10.1007/s41468-018-0017-1
14. Hamilton, W.L.: Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning **14**(3) (2020). https://doi.org/10.1007/978-3-031-01588-5
15. Hatcher, A.: Algebraic topology. Cambridge University Press (2002), https://pi.math.cornell.edu/~hatcher/AT/AT+.pdf
16. Jiang, D., Wu, Z., Hsieh, C.Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., Hou, T.: Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. Journal of cheminformatics **13**, 1–23 (2021). https://doi.org/10.1186/s13321-020-00479-8

17. Jiang, Q., Kurtek, S., Needham, T.: The weighted euler curve transform for shape and image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 844–845 (2020). https://doi.org/10.1109/CVPRW50498.2020.00430.
18. Koke, C., Shen, Y., Saroha, A., Eisenberger, M., Rieck, B., Bronstein, M.M., Cremers, D.: Graph networks struggle with variable scale. In: ICLR Workshop 'I Can't Believe It's Not Better: Challenges in Applied Deep Learning' (2025), https://openreview.net/forum?id=N5n6SAfnU0
19. Laky, D.J., Zavala, V.M.: A fast and scalable computational topology framework for the Euler characteristic. Digital Discovery **3**(2), 392–409 (2024). https://doi.org/10.1039/d3dd00226h
20. Leinster, T.: The euler characteristic of a category. Documenta Mathematica **13**, 21–49 (2008). https://doi.org/10.4171/DM/240
21. Lum, P.Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., Carlsson, G.: Extracting insights from the shape of complex data using topology. Scientific reports **3**(1), 1236 (2013). https://doi.org/10.1038/srep01236
22. Mauri, A.: alvadesc: A tool to calculate and analyze molecular descriptors and fingerprints. In: Ecotoxicological QSARs, pp. 801–820. Springer (2020). https://doi.org/10.1007/978-1-0716-0150-1_32
23. Munch, E.: An invitation to the euler characteristic transform. The American Mathematical Monthly **132**(1), 15–25 (2025). https://doi.org/10.1080/00029890.2024.2409616
24. Munkres, J.R., Krantz, S.G., Parks, H.R.: Elements of algebraic topology. Chapman and Hall/CRC (2025). https://doi.org/10.1201/9781003621478
25. Platonov, O., Kuznedelev, D., Diskin, M., Babenko, A., Prokhorenkova, L.: A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=tJbbQfw-5wv
26. Rieck, B.: Topology meets machine learning: An introduction using the euler characteristic transform. Notices of the American Mathematical Society **72**(7), 719–727 (2025), https://www.ams.org/journals/notices/202507/rnoti-p719.pdf
27. Röell, E., Rieck, B.: Differentiable euler characteristic transforms for shape classification. In: International Conference on Learning Representations (2024), https://openreview.net/forum?id=MO632iPq3I
28. Rottach, F., Schieferdecker, S., Eickhoff, C.: The topology of molecular representations and its influence on machine learning performance. Journal of Cheminformatics (2025). https://doi.org/10.1186/s13321-025-01045-w
29. Smith, A., Runde, S., Chew, A.K., Kelkar, A.S., Maheshwari, U., Van Lehn, R.C., Zavala, V.M.: Topological analysis of molecular dynamics simulations using the euler characteristic. Journal of Chemical Theory and Computation **19**(5), 1553–1567 (2023). https://doi.org/10.1021/acs.jctc.2c00766
30. Van Tilborg, D., Alenicheva, A., Grisoni, F.: Exposing the limitations of molecular machine learning with activity cliffs. Journal of chemical information and modeling **62**(23), 5938–5951 (2022). https://doi.org/10.1021/acs.jcim.2c01073
31. Waibel, D.J.E., Atwell, S., Meier, M., Marr, C., Rieck, B.: Capturing shape information with multi-scale topological loss terms for 3D reconstruction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 150–159. Springer, Cham, Switzerland (2022). https://doi.org/10.1007/978-3-031-16440-8_15

32. Walters, W.P., Barzilay, R.: Applications of deep learning in molecule generation and molecular property prediction. Accounts of chemical research **54**(2), 263–270 (2020). `https://doi.org/10.1021/acs.accounts.0c00699`
33. Wigh, D.S., Goodman, J.M., Lapkin, A.A.: A review of molecular representation in the age of machine learning. Wiley Interdisciplinary Reviews: Computational Molecular Science **12**(5), e1603 (2022). `https://doi.org/10.1002/wcms.1603`
34. Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., et al.: Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. Journal of medicinal chemistry **63**(16), 8749–8760 (2019). `https://doi.org/10.1021/acs.jmedchem.9b00959`
35. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI Open **1**, 57–81 (2020). `https://doi.org/10.1016/j.aiopen.2021.01.001`