



Molecular Machine Learning Using Euler Characteristic Transforms

Victor Toscano-Duran¹(✉)()^{ID}, Florian Rottach^{2,3}()^{ID}, and Bastian Rieck⁴()^{ID}

¹ Department of Applied Mathematics I, University of Seville, Seville, Spain
vtoscano@us.es

² Central Data Science, Boehringer Ingelheim GmbH, Biberach/Riss, Germany
florian.rottach@boehringer-ingelheim.com

³ School of Medicine, University of Tübingen, Tübingen, Germany

⁴ 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 learning approaches. Here, we propose using the *Euler Characteristic Transform* (ECT) as a geometrical-topological descriptor. Computed directly from molecular graphs constructed using handcrafted atomic features, the ECT enables the extraction of *multi-scale* structural features, offering a novel way to encode molecular shape in the feature space. We assess the predictive performance of this representation across nine benchmark regression datasets, all centered around predicting the inhibition constant K_i . In addition, we compare our proposed ECT-based descriptor against traditional molecular representations and methods, such as molecular fingerprints/descriptors and graph neural networks (GNNs). Our results show that our ECT-based representation achieves *competitive performance*, ranking among the best-performing methods on several datasets. More importantly, combining our descriptor with established representations, particularly with the AVALON fingerprint, significantly *enhances predictive performance*, outperforming other methods on most datasets. These findings highlight the complementary value of *multiscale* topological information and its potential for being combined with established techniques. Our study suggests that hybrid approaches incorporating explicit shape information can lead to more informative and robust molecular representations, enhancing and opening new avenues in molecular machine learning. To support reproducibility and foster open biomedical research, we provide open access to all experiments and code used in this work.

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

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, 31]. 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, 32]. 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, 29]. 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, 15]. 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) [13] may capture *some* of the structural information, but they often overlook the importance of representing the shape in a comprehensive and multiscale way [17].

Algebraic topology [14, 22] and topological data analysis [12, 20] can provide powerful tools to analyze the shape of molecular data, as shown by recent studies [26]. The *Euler Characteristic Transform* (ECT) [21, 24], for example, could be a topological representation of molecules that captures the molecular shape by tracking changes in the Euler characteristic [19] 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, 25]. In this work, we compare and explore the use of the ECT-based molecular representation, which is computed directly over molecular graphs derived from handcrafted atomic features, to predict K_i , a key molecular property, as well as perform a comparison between our approach and traditional methods over a series of nine binding affinity datasets. By using this topological representation, as well as combining it with traditional molecular representations, we aim to enhance predictive performance and provide new insights into the role of molecular shape in molecular machine learning. Our experiments shows that our ECT-based approach exhibites competitive predictive performance, in some cases even outperforming all alternative and traditional meth-

ods. In addition, our experiments show that the *combination* of our ECT-based approach with existing methods, more specifically with the AVALON fingerprint, leads to further enhance predictive performance, thus highlighting the complementary value of multiscale topological and shape information. Ultimately, our work contributes to the growing body of evidence suggesting that incorporating molecular shape at a fundamental level can lead to more robust and informative molecular representations, opening up new avenues for the design of better molecular machine learning models and more effective therapies.

The remainder of this paper is organized as follows: Firstly, Sect. 2 introduces the foundational concepts relevant to our study, including molecular structures and their representations, traditional molecular descriptors, and the Euler Characteristic Transform. Section 3 details the datasets employed, the traditional representations included for comparison in the experiments, our proposed ECT-based approach, and the overall experimental setup. The results of our experiments, with a comprehensive discussion, are presented in Sect. 4. Finally, conclusions and future work are discussed in Sect. 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 [32].

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 [13], 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. An example of the molecular graph of acetic acid is shown in Fig. 1a.

Traditionally, computational models for molecular machine learning tasks have relied on handcrafted molecular representations [32]. Among these are *molecular descriptors* (e.g., FGCCount or 2DAP), which are numerical features derived from molecules, 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 features 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, 33].

In recent years *Graph Neural Networks* (GNNs) [13], have become a dominant paradigm in molecular property prediction, extending neural networks to graph-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 [15]. Methods from *Topological Data Analysis* (TDA) [12] provide a complementary perspective. Instead of relying on raw atomic positions, handcrafted atomic features or purely local graph structures, TDA extracts global shape features from data, capturing relevant geometric information at multiple scales [20], making it thus particularly suitable for applications in the life sciences [30]. A particularly relevant tool within this framework is the Euler Characteristic Transform (ECT) [21, 24, 25, 28], 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* [19], 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 enable the modeling of real-world systems by focusing on *dyadic relationships* between elements. Formally, a *graph* $G = (V, E)$ consists of a finite set of vertices $V = \{v_1, v_2, \dots, v_n\}$, 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, graphs are the natural way to represent molecules, nodes corresponds to atoms, typi-

cally encoded as a feature vector , and edges to chemical bonds. However, to extract richer geometric and topological information, we can generalize graphs into structures called *simplicial complexes*. The graph of the acetic 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 graph of Fig. 1a can also be seen as a 1-dimensional 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:

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

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. In the case of (molecular) graphs, which only consist of 0-simplices (vertices) and 1-simplices (edges), this is reduced to:

$$\chi = |V| - |E|. \quad (2)$$

For example, the Euler characteristic of the acetic acid graph presented in Fig. 1a is 1 (8 vertices - 7 edges).

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 dynamic 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 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. 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.

Euler Characteristic Transforms. While the 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. 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 ECC of the molecule graph in Fig. 1a along the x -axis direction, and to Fig. 1e for an ECT example of this graph, built by stacking multiple ECCs, visualized as a matrix, where each column corresponds to a direction, each row to a threshold level in the filtration process, and the color intensity encodes the Euler characteristic value. Note that the numbers shown on the x -axis and y -axis of these figures do not represent the actual threshold and directions values used during filtration, but rather the index of the threshold and direction step. The resulting feature set derived from

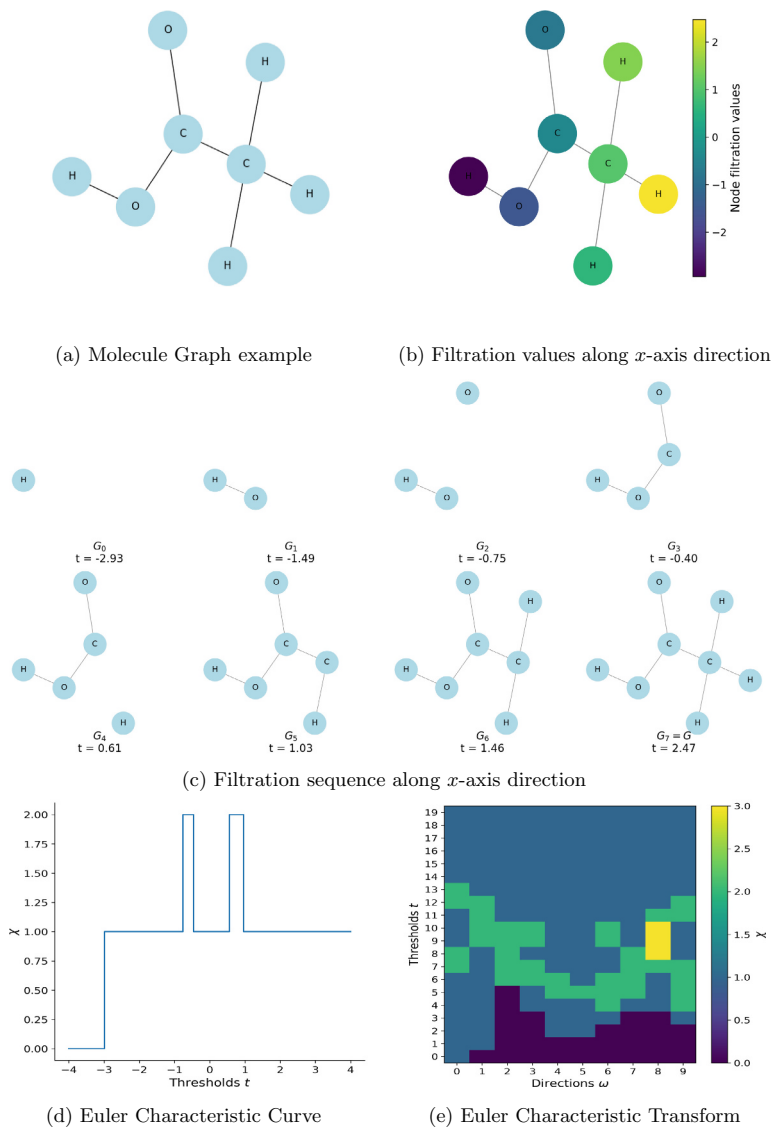


Fig. 1. Extracting topological shape information from molecular graphs. (a): Exemplary 2D graph representation of the acetic acid molecule, derived from its SMILES string CC(=O)O. The graph contains 8 vertices (2 carbon, 2 oxygen, and 4 hydrogen atoms) and 7 edges, resulting in an Euler characteristic of 1 ($\chi = 1$). (b): Filtration values of the nodes along the x -axis direction. (c): Filtration sequence along the x -axis direction. (d): ECC computed for the x -axis direction. In the figure, the x -axis represents the filtration thresholds values, and the y -axis shows the corresponding Euler characteristic values. (e): ECT example, built by stacking the ECC along 10 random directions from the molecule graph of (a). Columns correspond to directions, rows to thresholds (set to 20), and color intensity encodes the Euler characteristic value.

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 [1,25]. While the Euler characteristic itself has previously been used in molecular dynamics [18,27], to our knowledge, our work is the *first* to explore the use of the ECT in the context of molecular machine learning.

3 Materials and Methods

Having described a novel and multiscale topology-based approach to represent molecules, we conduct an extensive suite of experiments to explore and assess its effectiveness against established methods such as molecular fingerprints, descriptors, and graph neural networks.

3.1 Datasets

We concentrate on a number of datasets from different biochemical domains. Specifically, we use a collection of nine protein-ligand binding affinity datasets, all focused on predicting the inhibition constant K_i , a key measure of how strongly a molecule binds to a given protein target. These datasets are sourced from BindingDB¹ and cover a variety of biologically relevant proteins. The dataset associated with the alpha1A adrenergic receptor (ADRA1A) contains 1959 molecules; the one for the arachidonate 5-lipoxygenase-activating protein (ALOX5AP) includes 1636 molecules; and the dataset for the ataxia telangiectasia and Rad3-related protein (ATR) consists of 1411 molecules. For human dipeptidyl peptidase-4 (DPP4), the dataset comprises 3345 molecules, while the Janus kinase proteins JAK1 and JAK2 include 3058 and 1449 molecules, respectively. The kappa-type opioid receptor (KOR) dataset contains 1685 molecules, and the two datasets corresponding to subtypes of muscarinic receptors (MUSC1 and MUSC2) consist of 2129 and 1109 molecules, respectively. These molecule counts correspond to the pre-processed data, from which molecules with missing or invalid K_i binding affinity values were excluded, and only unique SMILES representations were retained. Despite targeting different proteins, all datasets share a unified prediction task: estimating the binding affinity constant K_i . Experiments were conducted both individually for each dataset and on a combined dataset comprising 16558 unique molecules.

3.2 Methods

We organize our description of the methods according to the type of molecular representation used, as this choice determines the applicable machine learning

¹ <https://www.bindingdb.org/>.

Table 1. List of molecular representations, grouped by their respective categories and sorted alphabetically within each group. For each representation, the category and corresponding feature vector dimensionality are provided.

Representation name	Representation Category	Dimension
Molecule graph	Graph	-
AVALON	Fingerprint	1024
CATS2D	Fingerprint	189
ECFP4	Fingerprint	1024
EState	Fingerprint	79
KR	Fingerprint	4860
MACCS	Fingerprint	166
MAP4	Fingerprint	1024
Pharm2D	Fingerprint	1024
PubChem	Fingerprint	881
RDKit	Fingerprint	1024
2DAP	Descriptor	1596
ConstIdx	Descriptor	50
FGCount	Descriptor	153
MolProp	Descriptor	14
RingDesc	Descriptor	35
TOPO	Descriptor	74
WalkPath	Descriptor	46
ECT (ours)	Multiscale Topological	2528
ECT + AVALON Fingerprint (ours)	Multiscale Topological	3552

approaches. Table 1 summarizes the representations, including their category and feature vector dimensionality. Note that dimensionality is not defined for graph-based representations. In total, we have 4 different categories of representations: molecular graphs, fingerprints, descriptors, and our topological representation based on the ECT. The latter includes two variants: using the ECT alone, or combined with the AVALON molecular fingerprint, which consists of concatenating the ECT with the AVALON fingerprint. AVALON was chosen for its widespread use and solid performance. 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 (Sect. 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 machine learning, named “AttentiveFP” [33]. For all other representations, we use an XGBoost [5] model, which has consistently shown strong performance in variety of tasks.

3.3 Experimental Setup

Molecule graphs are extracted from the SMILES strings provided in the datasets using the `PyTorch Geometric`² Python package, which encoded nodes (atoms) as a 9-dimensional handcrafted feature vector. Specifically, each vector includes the following categorical atomic features: (1) atomic number, (2) chirality tag, (3) total degree, (4) formal charge, (5) total number of bonded hydrogens, (6) number of radical electrons, (7) hybridization type, (8) aromaticity, and (9) ring membership. The ECT is then computed over this multidimensional feature space. Based on a prior sensitivity analysis, we fixed the number of directions θ and filtration thresholds r to compute the ECT to 158 and 16, respectively. To validate this choice, we evaluated our ECT-descriptor on the ADRA1A dataset across a grid of values ($\theta, r \in \{10, 20, 40, 60, 80, 100, 140, 180, 200\}$), training XGBoost regressors on each configuration using 10-fold cross-validation. Predictive performance was measured using the Root Mean Squared Error (RMSE). 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 40 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 the number of filtration thresholds had minimal effect on performance across a wide range of values. This is quantitatively supported by Pearson correlations: -0.74 between *theta* and RMSE, and -0.13 between r and RMSE. Therefore, we selected 158 directions to ensure sufficient expressiveness while limiting the number of thresholds to 16 to reduce computational cost and dimensionality without sacrificing predictive performance. 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 [25].

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. [23], 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

² <https://pytorch-geometric.readthedocs.io/>.

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: RMSE, and coefficient of determination (R2). This was done separately for each dataset and for the combined one.

4 Results

Table 2 provides an overview of all results for the different representations for each dataset, as well as for all datasets combined into a single one. Overall, we observe that already the ECT on its own provides meaningful and effective features for molecular machine learning, outperforming all alternative methods on ADRA1A. In *combination* with the AVALON molecular fingerprint, we observe generally improved baseline predictive performance in 5/9 datasets, as well as on the combined dataset. Thus, the ECT performs *consistently well*, showing its potential for creating generalizable features.

Table 2. Results (RMSE, reported as mean $\pm$ standard deviation across 10-fold cross-validation) for the collection of molecular datasets. Rows corresponds to different representation methods grouped by category (GNNs, fingerprints, descriptors, and ECT-based). Columns correspond to datasets.

Method	Dataset										
	ADRA1A	ALOX5P	ATR	DPP4	JAK1	JAK2	KOR	MUSC1	MUSC2	Combined	
AttentiveFP	2.01 $\pm$ 0.05	2.01 $\pm$ 0.29	1.42 $\pm$ 0.12	0.90 $\pm$ 0.06	1.10 $\pm$ 0.12	1.60 $\pm$ 0.20	2.17 $\pm$ 0.19	2.19 $\pm$ 0.10	2.03 $\pm$ 0.18	2.43 $\pm$ 0.27	
GAT	2.65 $\pm$ 0.17	2.45 $\pm$ 0.17	1.81 $\pm$ 0.09	1.18 $\pm$ 0.05	1.99 $\pm$ 0.15	2.67 $\pm$ 0.33	2.56 $\pm$ 0.17	2.84 $\pm$ 0.17	2.60 $\pm$ 0.16	2.80 $\pm$ 0.07	
GCN	2.75 $\pm$ 0.15	2.51 $\pm$ 0.19	1.96 $\pm$ 0.16	1.26 $\pm$ 0.08	2.76 $\pm$ 0.17	2.95 $\pm$ 0.28	2.73 $\pm$ 0.13	2.96 $\pm$ 0.15	2.76 $\pm$ 0.18	2.88 $\pm$ 0.10	
AVALON	1.81 $\pm$ 0.14	1.58 $\pm$ 0.18	1.26 $\pm$ 0.10	0.79 $\pm$ 0.04	0.99 $\pm$ 0.10	1.30 $\pm$ 0.19	1.83 $\pm$ 0.16	1.88 $\pm$ 0.09	1.86 $\pm$ 0.16	1.71 $\pm$ 0.03	
CATS2D	1.88 $\pm$ 0.09	1.88 $\pm$ 0.14	1.36 $\pm$ 0.12	0.90 $\pm$ 0.04	1.04 $\pm$ 0.08	1.50 $\pm$ 0.21	1.85 $\pm$ 0.12	1.98 $\pm$ 0.09	1.86 $\pm$ 0.11	1.88 $\pm$ 0.04	
ECFP4	1.81 $\pm$ 0.11	1.71 $\pm$ 0.17	1.30 $\pm$ 0.09	0.84 $\pm$ 0.04	1.01 $\pm$ 0.12	1.31 $\pm$ 0.18	1.90 $\pm$ 0.14	1.98 $\pm$ 0.06	1.83 $\pm$ 0.13	1.81 $\pm$ 0.03	
EState	2.20 $\pm$ 0.16	2.12 $\pm$ 0.20	1.48 $\pm$ 0.10	1.11 $\pm$ 0.06	1.21 $\pm$ 0.15	1.74 $\pm$ 0.18	2.44 $\pm$ 0.16	2.54 $\pm$ 0.09	2.19 $\pm$ 0.08	2.28 $\pm$ 0.04	
KR	1.82 $\pm$ 0.14	1.78 $\pm$ 0.20	1.36 $\pm$ 0.09	0.88 $\pm$ 0.04	1.03 $\pm$ 0.12	1.42 $\pm$ 0.19	1.91 $\pm$ 0.18	2.00 $\pm$ 0.07	1.85 $\pm$ 0.11	1.95 $\pm$ 0.02	
MACCS	1.88 $\pm$ 0.15	1.81 $\pm$ 0.20	1.39 $\pm$ 0.11	0.90 $\pm$ 0.04	1.06 $\pm$ 0.12	1.52 $\pm$ 0.19	1.99 $\pm$ 0.15	2.07 $\pm$ 0.09	2.08 $\pm$ 0.13	1.92 $\pm$ 0.04	
MAP4	1.88 $\pm$ 0.15	1.68 $\pm$ 0.16	1.33 $\pm$ 0.10	0.86 $\pm$ 0.05	1.03 $\pm$ 0.10	1.44 $\pm$ 0.22	2.01 $\pm$ 0.17	2.07 $\pm$ 0.09	2.00 $\pm$ 0.17	1.86 $\pm$ 0.04	
Pharm2D	1.80 $\pm$ 0.11	1.68 $\pm$ 0.21	1.35 $\pm$ 0.10	0.79 $\pm$ 0.05	0.98 $\pm$ 0.12	1.47 $\pm$ 0.20	1.81 $\pm$ 0.11	1.89 $\pm$ 0.12	1.79 $\pm$ 0.13	1.71 $\pm$ 0.03	
PubChem	1.87 $\pm$ 0.13	1.66 $\pm$ 0.16	1.32 $\pm$ 0.12	0.90 $\pm$ 0.04	1.01 $\pm$ 0.10	1.32 $\pm$ 0.22	2.01 $\pm$ 0.15	2.06 $\pm$ 0.11	1.90 $\pm$ 0.09	1.94 $\pm$ 0.02	
RDKit	1.83 $\pm$ 0.13	1.56 $\pm$ 0.13	1.26 $\pm$ 0.10	0.78 $\pm$ 0.05	0.99 $\pm$ 0.11	1.36 $\pm$ 0.15	1.85 $\pm$ 0.13	1.89 $\pm$ 0.08	1.91 $\pm$ 0.13	1.69 $\pm$ 0.03	
2DAP	1.84 $\pm$ 0.14	1.66 $\pm$ 0.13	1.33 $\pm$ 0.11	0.90 $\pm$ 0.05	1.02 $\pm$ 0.10	1.45 $\pm$ 0.23	1.91 $\pm$ 0.14	1.98 $\pm$ 0.10	1.88 $\pm$ 0.08	1.91 $\pm$ 0.05	
ConstIdx	1.99 $\pm$ 0.17	1.92 $\pm$ 0.14	1.38 $\pm$ 0.10	1.00 $\pm$ 0.06	1.12 $\pm$ 0.11	1.58 $\pm$ 0.22	2.12 $\pm$ 0.10	2.20 $\pm$ 0.10	2.06 $\pm$ 0.09	2.01 $\pm$ 0.05	
FGCount	1.93 $\pm$ 0.16	1.71 $\pm$ 0.16	1.36 $\pm$ 0.10	0.91 $\pm$ 0.06	1.01 $\pm$ 0.08	1.40 $\pm$ 0.17	1.98 $\pm$ 0.16	2.09 $\pm$ 0.07	1.87 $\pm$ 0.15	2.01 $\pm$ 0.06	
MolProp	2.05 $\pm$ 0.05	1.90 $\pm$ 0.14	1.43 $\pm$ 0.09	1.06 $\pm$ 0.07	1.17 $\pm$ 0.11	1.65 $\pm$ 0.22	2.22 $\pm$ 0.16	2.20 $\pm$ 0.08	2.06 $\pm$ 0.13	2.01 $\pm$ 0.06	
RingDesc	2.12 $\pm$ 0.12	2.15 $\pm$ 0.17	1.62 $\pm$ 0.11	1.18 $\pm$ 0.08	1.22 $\pm$ 0.14	1.66 $\pm$ 0.21	2.28 $\pm$ 0.21	2.46 $\pm$ 0.10	2.32 $\pm$ 0.13	2.24 $\pm$ 0.07	
TOPO	1.89 $\pm$ 0.13	1.92 $\pm$ 0.13	1.50 $\pm$ 0.08	1.04 $\pm$ 0.06	1.20 $\pm$ 0.09	1.64 $\pm$ 0.18	2.01 $\pm$ 0.16	2.10 $\pm$ 0.11	1.92 $\pm$ 0.09	1.97 $\pm$ 0.05	
WalkPath	1.97 $\pm$ 0.13	1.98 $\pm$ 0.13	1.42 $\pm$ 0.10	1.06 $\pm$ 0.06	1.16 $\pm$ 0.12	1.61 $\pm$ 0.21	2.16 $\pm$ 0.19	2.17 $\pm$ 0.09	2.09 $\pm$ 0.09	2.13 $\pm$ 0.05	
ECT (ours)	1.78 $\pm$ 0.12	1.77 $\pm$ 0.16	1.35 $\pm$ 0.10	0.83 $\pm$ 0.04	1.03 $\pm$ 0.13	1.45 $\pm$ 0.18	1.87 $\pm$ 0.14	1.89 $\pm$ 0.08	1.81 $\pm$ 0.13	1.79 $\pm$ 0.05	
ECT+ FP (ours)	1.74 $\pm$ 0.12	1.52 $\pm$ 0.17	1.29 $\pm$ 0.09	0.78 $\pm$ 0.04	0.99 $\pm$ 0.11	1.38 $\pm$ 0.20	1.82 $\pm$ 0.12	1.82 $\pm$ 0.08	1.77 $\pm$ 0.13	1.68 $\pm$ 0.02	

Figure 2 ranks the representations by their mean error for two different metrics (RMSE and R2) for the combined dataset. Performance is visualized using boxplots to show the error distribution for each representation method according the cross validation results. The representation combining our ECT-based representation and the AVALON fingerprint consistently achieves better values

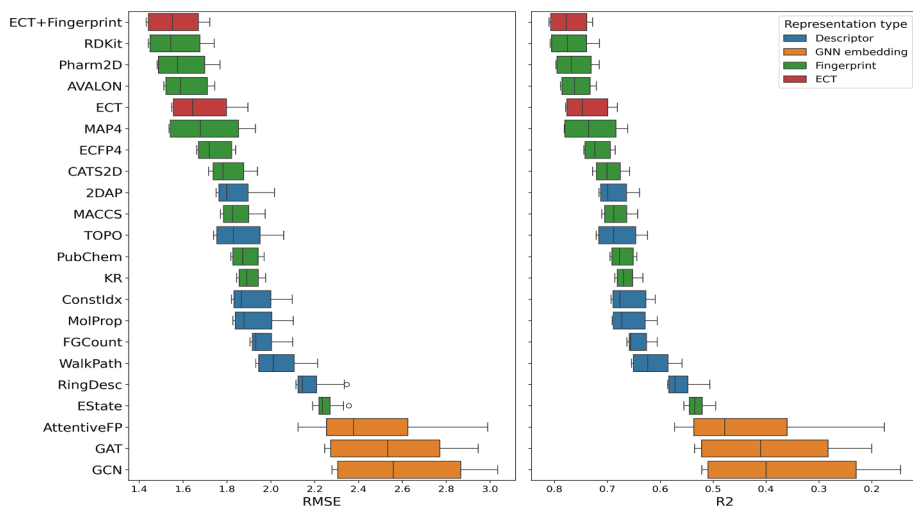


Fig. 2. Boxplots showing the distribution of test errors for each representation method across the combined dataset. From left to right, the metrics displayed are RMSE and R2. Representations are ranked by their mean values: For RMSE, *lower* values indicate *better* performance, while for R2, *higher* values (closer to 1) indicate *better* performance.

for RMSE and R2 metrics. The low variance between cross-validation results is particularly noteworthy, showing the robustness of the approach.

An overall observation is that most complex models (like GNNs) that *directly* take the molecular graph as its input tend to yield the *worst* computational performance. This is in marked contrast to the other molecular representations (using XGBoost), where a single iteration of cross-validation takes around 10 to 20s on a standard computer. In comparison, for GNNs, a single iteration takes about 40 to 60s for GCN and GAT, and up to 100s for “AttentiveFP,” which nevertheless leads to significantly worse predictive performance. This observation suggests that the use of deep learning models is not inherently justified by their (computational and predictive) 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 property prediction tasks, specifically for predicting K_i , the inhibition constant. Our results on

different datasets show that the ECT provides meaningful and effective features for molecular machine learning, and that its *combination* with other approaches, specifically with the AVALON fingerprint, leads to improved predictive performance. This multiscale topological approach, which can capture shape characteristics across different resolutions, encodes structural information of molecules, 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 property prediction 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 and topologically. In addition, we plan to extend and test our topological approach in other molecular machine learning tasks, like molecular classification, toxicity prediction, or even generative models for molecular design. In this work, we tested a single hybrid strategy (ECT + AVALON fingerprint) and observed improved predictive accuracy. Future work could benchmark multiple hybrid strategies to identify synergies between multiscale topological and traditional representations. Moreover, we aim to extend the ECT to molecular graphs with 3D positional information to quantify the geometric shape of the molecular structure itself, instead of, or in combination with, general atomic features. 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 and expressive descriptors. Further efforts will also include systematic hyperparameter tuning and the evaluation of different dataset splitting strategies, which can substantially influence both predictive performance and generalizability. Finally, our approach could also be extended to other variants of the ECT, such as the Weighted Euler Characteristic Transform (WECT) [16] or the Smooth Euler Characteristic Transform (SECT) [7].

Acknowledgements. This work was partially supported by REXASI-PRO project (ID: 101070028), the Research Budget (VII PPIT) of the University of Seville, and the Swiss State Secretariat for Education, Research, and Innovation (SERI).

Data Availability Statement. All data and code used to generate the results presented in this manuscript are publicly available in a GitHub repository (<https://github.com/victosdur/ECTforMoleculeMachineLearningTask>).

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., et al.: Measuring hidden phenotype: quantifying the shape of barley seeds using the Euler characteristic transform. in silico Plants **4**(1), diab033 (2021)
2. Atz, K., Grisoni, F., Schneider, G.: Geometric deep learning on molecular representations. Nat. Mach. Intell. **3**(12), 1023–1032 (2021)
3. Baptista, D., Correia, J., Pereira, B., Rocha, M.: Evaluating molecular representations in machine learning models for drug response prediction and interpretability. J. Integr. Bioinform. **19**(3), 20220006 (2022)
4. Carracedo-Reboredo, P.: A review on machine learning approaches and trends in drug discovery. Comput. Struct. Biotechnol. J. **19**, 4538–4558 (2021)
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, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016)
6. Corey, E.J., Czako, B., Kürti, L.: Molecules and medicine. John Wiley and Sons (2007)
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](https://arxiv.org/abs/1611.06818) (2016)
8. David, L., Thakkar, A., Mercado, R., Engkvist, O.: Molecular representations in AI-driven drug discovery: a review and practical guide. J. Cheminformatics **12**(1), 1–22 (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 Bioinform. **23**(1), bbab430 (2022)
10. Deng, J., et al.: A systematic study of key elements underlying molecular property prediction. Nat. Commun. **14**(1), 6395 (2023)
11. Drews, J.: Drug discovery: a historical perspective. Science **287**(5460), 1960–1964 (2000)
12. Edelsbrunner, H., Harer, J.L.: Computational topology: an introduction. Am. Math. Soc. (2022)
13. Hamilton, W.L.: Graph representation learning. Synth. Lect. Artif. Intell. Mach. Learn. **14**(3), 1–159 (2020)
14. Hatcher, A.: Algebraic topology. Cambridge University Press (2005)
15. Jiang, D.: Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. J. Cheminformatics **13**(1), 1–23 (2021). <https://doi.org/10.1186/s13321-020-00479-8>
16. 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)
17. Koke, C., et al.: Graph networks struggle with variable scale. In: ICLR Workshop ‘I Can’t Believe It’s Not Better: Challenges in Applied Deep Learning’ (2025)
18. Laky, D.J., Zavala, V.M.: A fast and scalable computational topology framework for the Euler characteristic. Digit. Disc. **3**(2), 392–409 (2024)
19. Leinster, T.: The Euler characteristic of a category. Doc. Math. **13**, 21–49 (2008)
20. Lum, P.Y.: Extracting insights from the shape of complex data using topology. Sci. Rep. **3**(1), 1236 (2013)
21. Munch, E.: An invitation to the Euler characteristic transform. Am. Math. Mon. **132**(1), 15–25 (2025)

22. Munkres, J.R.: Elements of algebraic topology. CRC press (2018)
23. 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? arXiv preprint [arXiv:2302.11640](https://arxiv.org/abs/2302.11640) (2023)
24. Rieck, B.: Topology meets machine learning: an introduction using the Euler characteristic transform. *Not. Am. Math. Soc.* **72**(7), 719–727 (2025)
25. Röell, E., Rieck, B.: Differentiable Euler characteristic transforms for shape classification. In: International Conference on Learning Representations (2024)
26. Rottach, F., Schieferdecker, S., Eickhoff, C.: The topology of molecular representations and its influence on machine learning performance. *J. Cheminformatics* **17** (2025)
27. Smith, A.: Topological analysis of molecular dynamics simulations using the Euler characteristic. *J. Chem. Theory Comput.* **19**(5), 1553–1567 (2023)
28. Turner, K., Mukherjee, S., Boyer, D.M.: Persistent homology transform for modeling shapes and surfaces. *Inf. Infer. a J. IMA* **3**(4), 310–344 (2014)
29. Van Tilborg, D., Alenicheva, A., Grisoni, F.: Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**(23), 5938–5951 (2022)
30. 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)
31. Walters, W.P., Barzilay, R.: Applications of deep learning in molecule generation and molecular property prediction. *Acc. Chem. Res.* **54**(2), 263–270 (2020)
32. Wigh, D.S., Goodman, J.M., Lapkin, A.A.: A review of molecular representation in the age of machine learning. *Wiley Interdisc. Rev. Comput. Molecular Sci.* **12**(5), e1603 (2022)
33. Xiong, Z., et al.: Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**(16), 8749–8760 (2019)