YZV413E Graph Theory Project Proposal

Mehmet Emin Acar
Artificial Intelligence and Data Engineering
Istanbul Technical University
acarme22@itu.edu.tr
150220316

Mustafa Kerem Bulut
Artificial Intelligence and Data Engineering
Istanbul Technical University
bulutm22@itu.edu.tr
150220303

I. Project Description

The aim of this project is to predict the melting temperatures of organic molecules directly from their chemical structure using graph-based deep learning. We will focus on the regression task of estimating the melting point (in Kelvin) of a compound given a representation of its molecular structure.

In our setting, each molecule will be represented by a SMILES string, which will then be converted into a molecular graph where nodes correspond to atoms and edges correspond to chemical bonds. On top of these graphs, we plan to employ a graph transformer [1] architecture that can capture long-range interactions between atoms and model complex structure–property relationships. The model will be trained in a supervised manner to minimize the error between predicted and experimentally reported melting points.

The project is inspired by the Kaggle competition *"Thermophysical Property: Melting Point"* [2], where participants are asked to build machine learning models that predict the melting point of organic compounds from molecular information. In our work, we adapt this idea to a graph neural network setting and systematically study how graph transformers perform on molecular melting point prediction.

II. Problem Definition

We formulate the task of predicting melting temperatures as a supervised graph regression problem. Let $\mathcal{D} = \{(G_i, y_i)\}_{i=1}^{N}$ be a dataset consisting of N labeled molecules, where G_i denotes the molecular graph and $y_i \in \mathbb{R}$ represents the corresponding experimental melting temperature.

Formally, each molecule is represented as an attributed undirected graph $G = (V, E)$, where:

- $V = \{v_1, v_2, \dots, v_{|V|}\}$ is the set of nodes, representing the atoms in the molecule.
- $E \subseteq V \times V$ is the set of edges, representing the chemical bonds between atoms.

To capture chemical information, we define feature matrices for both nodes and edges:

- Let $\mathbf{X}_v \in \mathbb{R}^{|V| \times d_v}$ be the node feature matrix, where each row corresponds to a d_v-dimensional vector encoding atomic properties.
- Let $\mathbf{X}_e \in \mathbb{R}^{|E| \times d_e}$ be the edge feature matrix, where each row corresponds to a d_e-dimensional vector encoding bond characteristics.

Our goal is to learn a non-linear mapping function $f_\theta : \mathcal{G} \to \mathbb{R}$, parameterized by θ, that predicts the melting temperature $\hat{y} = f_\theta(G)$ given a molecular graph input. The optimal parameters θ^* are estimated by minimizing a regression loss function, typically the Mean Squared Error (MSE), over the training dataset:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - f_\theta(G_i))^2 \tag{1}$$

In this project, f_θ will be realized using a Graph Transformer architecture designed to capture both local atomic neighborhoods and long-range structural dependencies.

III. Dataset

In this project we will construct our working dataset from the public dataset provided in the Kaggle competition *"Thermophysical Property: Melting Point"*. The competition dataset consists of organic molecules with experimentally measured melting points in Kelvin, together with machine-readable descriptions of their molecular structure.

Each data point in the raw competition dataset corresponds to a single organic compound and contains the following fields:

- a SMILES string encoding the molecular structure,
- a continuous target value $y \in \mathbb{R}$ representing the melting point (in Kelvin).

IV. Methodology

Our approach consists of two main steps: (i) constructing molecular graphs from SMILES strings, and (ii) training a Graph Attention Network (GAT) for melting point regression. In addition to the GAT model, we also include two strong baseline models to enable a sound comparison: (a) a classical ML model using Morgan fingerprints with XGBoost, and (b) a Transformer-based sequence model that operates directly on tokenized SMILES strings.

Graph Construction

Each SMILES string is parsed into a molecular graph using RDKit. We use the standard representation where

- nodes correspond to atoms,
- edges correspond to chemical bonds.

For each atom, we extract a fixed set of atom-level descriptors such as element type (C, O, N, halogens, etc.), degree, aromaticity, formal charge, and hybridization state. For each bond, we encode bond type (single, double, triple, aromatic) and ring membership as edge features.

This yields the node feature matrix $\mathbf{X}_v$, edge feature matrix $\mathbf{X}_e$, and the edge index representation that will be used as input to the GAT model.

Graph Attention Network

The prediction model is a graph neural network composed of stacked GAT layers. At a high level:

- Node embeddings are initialized with atom features and updated through several GAT layers, where attention weights are learned over neighboring atoms using edge information.
- After the final GAT layer, a global pooling operation (e.g., global mean or sum pooling) aggregates node embeddings into a single graph-level representation.
- A small feed-forward regression head maps this graph-level vector to a scalar prediction $\hat{y} \in \mathbb{R}$ corresponding to the melting point.

Classical Baseline: Morgan Fingerprints + XGBoost

To provide a strong classical baseline, each SMILES string will also be converted into a 2048-bit Morgan fingerprint [3] (ECFP4) using RDKit. These fingerprints encode local substructure information around atoms and serve as fixed-length descriptors. We train an XGBoost regressor on the fingerprint vectors to predict melting temperatures, allowing a comparison between classical cheminformatics methods and graph-based neural networks.

Sequence Baseline: Transformer on SMILES

As a non-graph deep learning baseline, we employ a [6] Transformer encoder model that operates directly on tokenized SMILES sequences. Each token receives a learnable embedding, positional encodings are added, and several Transformer layers process the sequence. The final hidden representation corresponding to a special sequence-level token is fed into a regression head. This baseline enables comparison between graph-based and sequence-based molecular representations.

Training and Evaluation

We split the dataset into training, validation, and test sets. The GAT parameters are optimized using the Mean Squared Error (MSE) loss and the Adam optimizer. The fingerprint-based [4] XGBoost model and the SMILES Transformer model are trained on the same splits to ensure a fair comparison. Hyperparameters such as the number of GAT layers, hidden dimension, learning rate, and pooling type will be selected based on validation performance.

Final performance will be reported on the held-out test set using Root Mean Squared Error (RMSE), following the evaluation protocol of the original Kaggle competition.

References

[1] V. P. Dwivedi and X. Bresson, “A Generalization of Transformer Networks to Graphs,” *arXiv preprint arXiv:2012.09699*, 2020.

[2] Kaggle, “Thermophysical Property: Melting Point,” [Online]. Available: https://www.kaggle.com/competitions/melting-point. [Accessed: Nov. 24, 2025].

[3] D. Rogers and M. Hahn, “Extended-Connectivity Fingerprints,” *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[4] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[5] S. Chithrananda, G. Grand, and B. Barzilay, “ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction,” *arXiv preprint arXiv:2010.09885*, 2020.

[6] A. Vaswani *et al.*, “Attention is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.