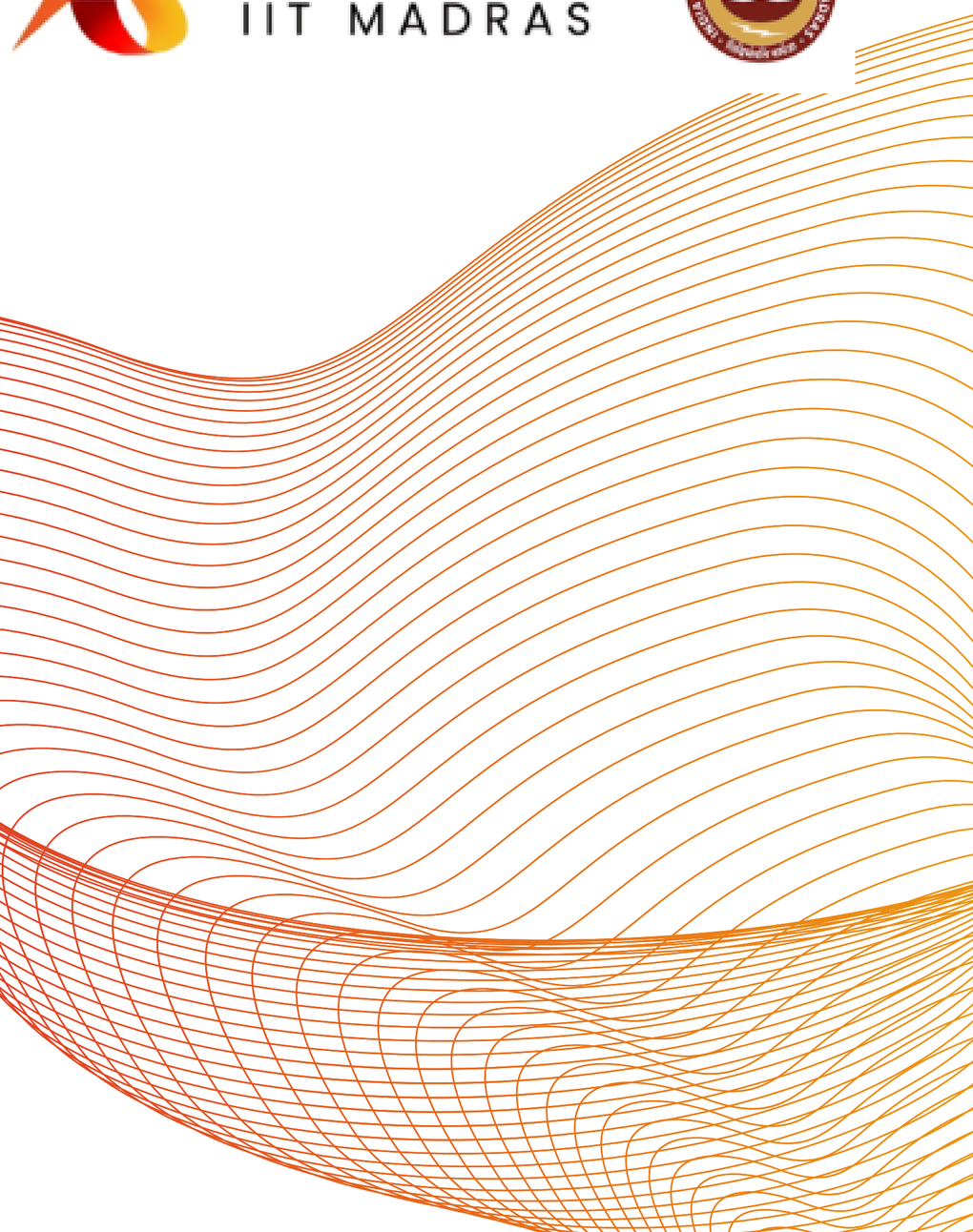
Machine Learning for Band gap prediction



AMALGAM'24
IIT MADRAS

Amalgam 2024 by MetSA

Created by
Adobe and Friends
Jyothiradiytha, Shabarish , Prathosh

EDA (Exploratory Data Analysis)

The dataset consists of the multiple features pertaining to a polymer molecule in order to predict the band gap which is the difference in energy on the HOMO and LUMO orbitals.

Through data analysis we were able to arrive at two conclusions:

- The dataset had ample information on the polymer molecule but missed cruical features essential for band gap prediction. This was to be extracted from the SMILES string.
- Band gap being a continuous value makes this a regression problem, thereforewe need to test a set of those models to see the best accuracy

Data Processing Methods

Mol2Vec

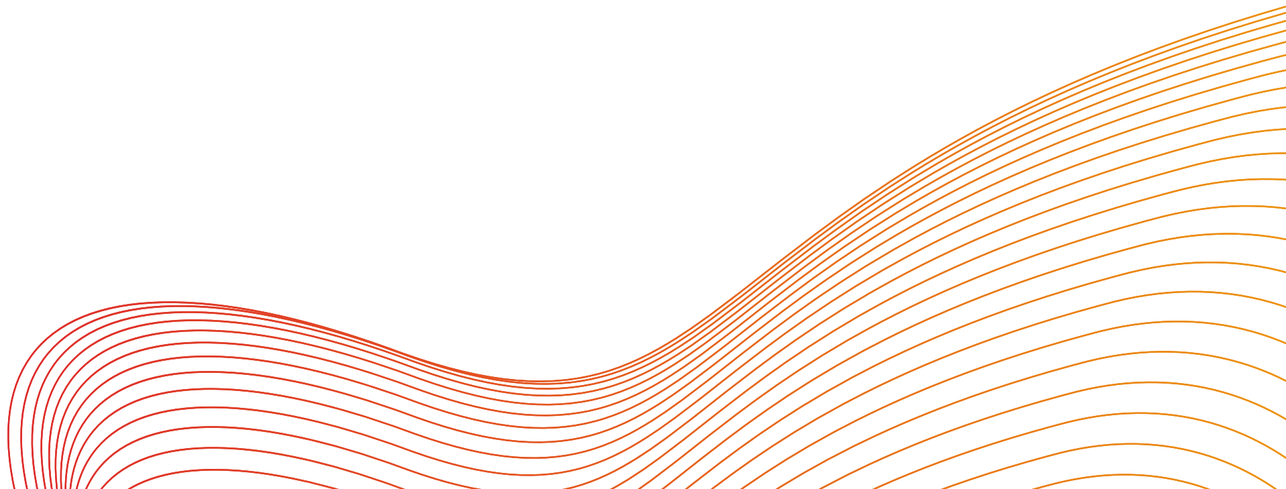
Mol2Vec is a method for generating vector representations (embeddings) of chemical compounds based on their molecular structures. It is inspired by word2vec, a popular algorithm used in natural language processing to create word embeddings. Mol2Vec extends this concept to the domain of cheminformatics.

In Mol2Vec, each chemical compound (molecule) is represented as a vector in a high-dimensional space, where similar molecules are expected to have similar vector representations. These vectors capture the structural and chemical properties of the molecules in a continuous and dense representation, enabling various machine learning and data mining tasks in cheminformatics.

Graph2Vec

Graph2Vec is a method for generating vector representations (embeddings) of graphs, particularly useful in the context of graph-based data such as molecular graphs in cheminformatics or social networks in social media analysis.

Graph2Vec extends the idea of node embeddings in graph data to the entire graph itself. It's inspired by techniques like Word2Vec and Doc2Vec, which generate vector representations for words and documents, respectively. Similarly, Graph2Vec generates embeddings for entire graphs.



Embedding (and why it failed)

An earlier attempt at extracting useful information from the SMILES format was to convert it into word embedding used pre-trained libraries. We deployed Graph2Vec which extends the idea of node embeddings in graph data to the entire graph itself. It considers each Carbon atom as a node in the graph. It's inspired by techniques like Word2Vec and Doc2Vec, which generate vector representations for words and documents, respectively.

It is however after employing these techniques that we realized, Graph2Vec obtains spatial configurations of the molecules and is hence not useful since the spatial configurations do not have an effect on the band gap.

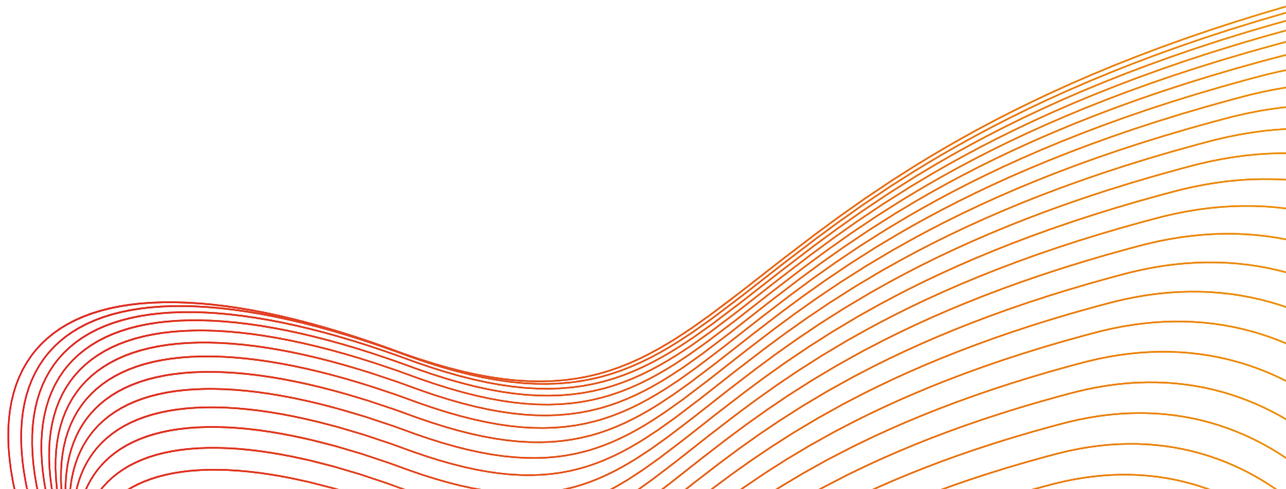
Feature Selection using correlation scores

We used correlation scores between the features and the target band gap to decide th features to take and drop unnecessary columns

RDKit

RDKit is an open-source toolkit for cheminformatics, molecular modeling, and drug discovery. It provides a wide range of functionality for handling chemical structures, including molecular depiction, substructure searching, molecular similarity calculation, chemical reactions, and property prediction.

We were able to use RDKit for visualizing, understanding and drawing useful inferences from the SMILES representations of the molecules that were provided in the training data.

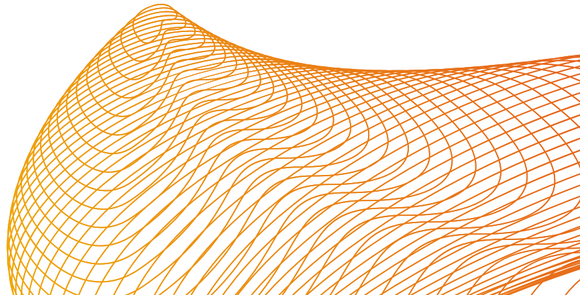


Feature Engineering from RDKit Library

The RDKit library offered multiple insights on the molecule given its SMILES string. A certain percentage of these features were already present in the given input data but they weren't sufficient for accurate band gap prediction as showcased by our initial testing in both XGBoost and Autogluon.(The XGBoost model was able to achieve only 0.69 rmse with the input material fingerprints)

Thus we set to include six important features of the polymer which directly influences its band gap namely:

- Molecular weight
- Number of atoms
- Number of heavy atoms
- Number of Valence electrons
- Number of hetero-atoms
- TPSA



Feature Engineering from RDKit Library

Molecular Weight:
Increased molecular weight can influence electron delocalisation. Similarly, in polymers, higher molecular weight might enhance electron delocalisation, leading to a narrower band gap.

Number of Atoms:
Generally, smaller structures may exhibit an increased band gap. This is because the confinement of electrons in a reduced space elevates the energy required for electronic transitions between the valence and conduction bands. As a result, the band gap tends to widen with a decrease in the number of atoms.

Number of valence electrons:
In polymers, a higher valence electron count can increase the energy required for electron transitions, potentially widening the band gap.

Feature Engineering from RDKit Library

Heavy Atoms:

The presence of heavy atoms increases polarizability, influencing the dielectric constant of a material and potentially altering the band gap. Additionally, heavy atoms induce charge redistribution within the material, leading to the formation of electric dipoles and impacting the electrostatic potential, which can further influence the band edges and band gap. The combination of these effects contributes to the overall modulation of the electronic properties in materials containing heavy atoms.

Hetero atoms:

Introducing heteroatom (non-carbon atoms) is akin to alloying in metallurgy, where adding different elements modifies the electronic properties of the material.The electronic effect of the heteroatom is thought to strongly correlate with its electron affinity, with higher affinities contributing to lower Eg values.

Feature Engineering from RDKit Library

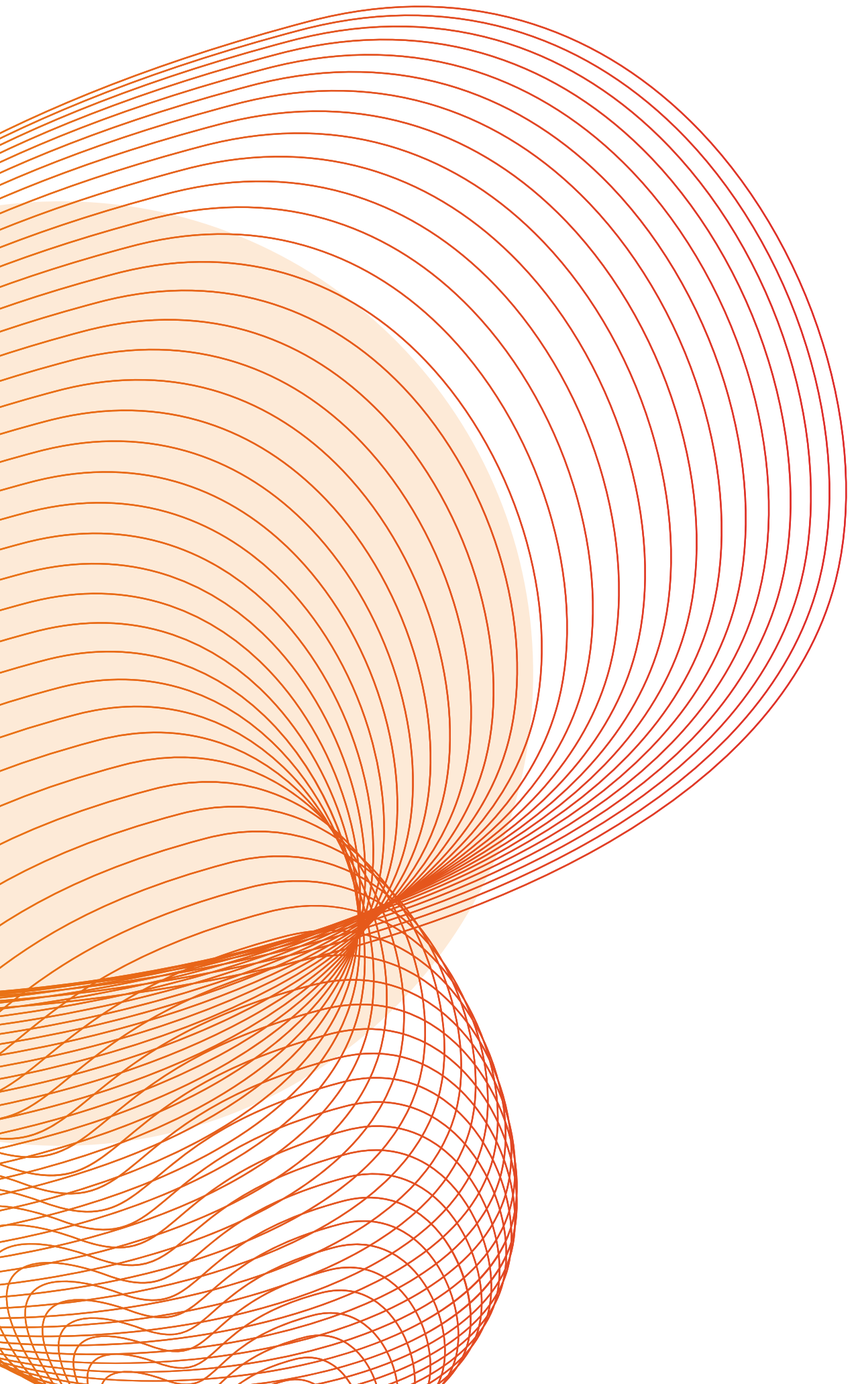
TPSA:
The Topological Polar Surface Area (TPSA) is a molecular descriptor that measures the exposure of polar atoms on the surface of a molecule. While TPSA itself may not directly determine the band gap of a polymer, it can influence the intermolecular interactions, affecting the overall electronic properties, including the band gap.

Higher TPSA values generally indicate a larger surface area with more polar atoms. In polymers, this increased polar surface can enhance interactions with other molecules or surfaces. These intermolecular interactions can impact the electronic structure and, subsequently, the band gap.

Feature Selection from RDKit Library

Using the RDKit we got accuracy of this was a significant improvement to the word embedding model plus RDKit which got us an accuracy of **.52 rmse** but using the word embedding increased due to inaccuracies in how the word embedding captures the information:

Models tested



Linear regression

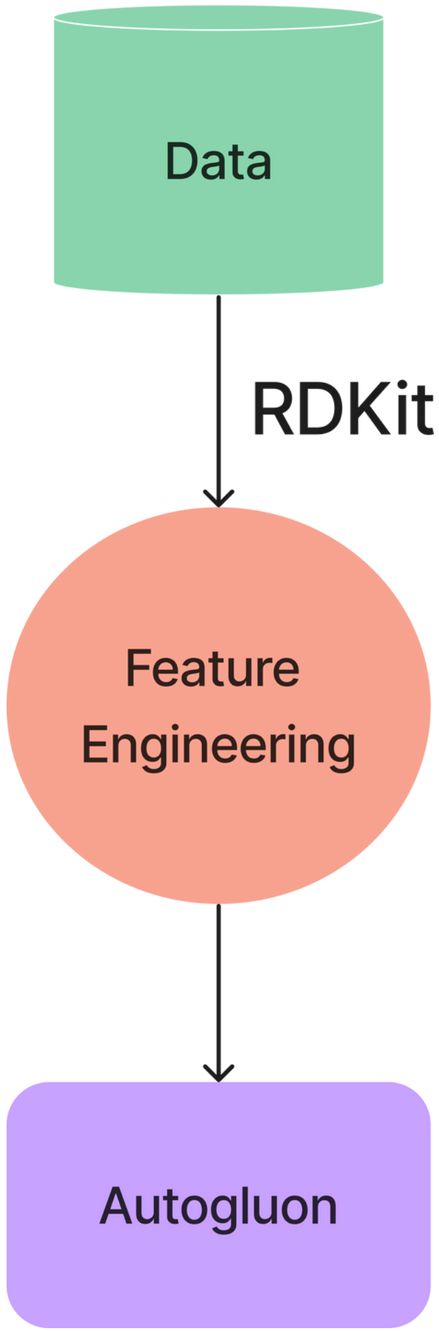
XGBoost

Random forest

Autogluon

- AutoGluon is a machine learning library that offers automated model selection, hyperparameter tuning, and training for various tasks, including regression.
- In the context of regression, AutoGluon provides a convenient way to build accurate regression models without the need for manual intervention in model selection or hyperparameter tuning
- high_quality: This preset is designed for high-quality models, focusing on accuracy and performance.

Model pipeline



Thank You