final

December 12, 2024

0.1 Abstract

0.2 Package import

[121]:
```
import os

os.environ["KERAS_BACKEND"] = "tensorflow"

import ast
import numpy as np

from tensorflow import keras
```

[122]:
```
#from tensorflow.keras import ops
from tensorflow.keras import layers
import pandas as pd

from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from rdkit import Chem, RDLogger
from rdkit.Chem import BondType
from rdkit.Chem.Draw import MolsToGridImage
from rdkit.Chem import Draw
from rdkit import Chem
from rdkit.Chem import rdmolops, AllChem
from tensorflow.keras.regularizers import l1_l2
RDLogger.DisableLog("rdApp.*")
```

[123]:
```
import tensorflow as tf
print("TensorFlow version:", tf.__version__)
print("GPU available:", tf.config.list_physical_devices('GPU'))
print("GPU in use:", tf.test.gpu_device_name())
```

```
TensorFlow version: 2.16.2
GPU available: []
GPU in use:
```

0.3 Database pharsing

[124]:
```
'''
read the entire dataset
'''

df = pd.read_csv('dataset1.csv')
df.drop([0,1,2,3,4], inplace=True)
df=df.rename(columns = {'PUBCHEM_EXT_DATASOURCE_SMILES':
 ↪'SMILES','PUBCHEM_ACTIVITY_OUTCOME':'Activity', 'PUBCHEM_ACTIVITY_SCORE':
 ↪'Score'})
columns_to_drop = [col for col in df.columns if col not in ['SMILES',␣
 ↪'Activity', 'Score', 'Potency', 'Efficacy']]
df = df.drop(columns = columns_to_drop)
#df=df.drop(['Unnamed: 3','Unnamed: 4','Unnamed: 5'], axis=1)
df = df.dropna(subset=['SMILES'])

df=df.fillna(0)
print(df.head())
print(df.info())
```

```
                                              SMILES  Activity  Score  \
5            CNCC1=NC2=C(C=C(C=C2)Cl)C(=N1)C3=CC=CN3  Inactive    0.0
6                 CCSC(=NC1=CC=C(C=C1)C(F)(F)F)N.Cl  Inactive    0.0
7  CCN(CC1=CC(=CC=C1)S(=O)(=O)[O-])C2=CC=C(C=C2)C…  Inactive    0.0
8  CC1=CC=C(C=C1)S(=O)(=O)N2CCN(CC2)C3=NC(=NC4=CC…  Inactive    0.0
9  CC1=CC=C(C=C1)S(=O)(=O)N2CCN(CC2)C3=NC(=NC4=CC…  Inactive    0.0

   Potency  Efficacy
5      0.0       0.0
6      0.0       0.0
7      0.0       0.0
8      0.0       0.0
9      0.0       0.0
<class 'pandas.core.frame.DataFrame'>
Index: 342051 entries, 5 to 342072
Data columns (total 5 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   SMILES    342051 non-null  object
 1   Activity  342051 non-null  object
 2   Score     342051 non-null  float64
 3   Potency   342051 non-null  float64
 4   Efficacy  342051 non-null  float64
dtypes: float64(3), object(2)
memory usage: 15.7+ MB
None
```

[125]:
```
valid_indices = []
# Loop through each SMILES string in the DataFrame
for i in range(len(df)):
    smiles = df.iloc[i]['SMILES']  # Use iloc for positional indexing

    # Convert SMILES to molecule
    mol = Chem.MolFromSmiles(smiles)

    # Check if the molecule is valid and has <= 50 atoms
    if mol is not None and mol.GetNumAtoms() <= 50:
        valid_indices.append(i)
# Filter the DataFrame to include only valid molecules
df_50 = df.iloc[valid_indices]
```

[126]:
```
df_50
```

[126]:

	SMILES	Activity	Score	Potency	Efficacy
5	CNCC1=NC2=C(C=C(C=C2)Cl)C(=N1)C3=CC=CN3	Inactive	0.0	0.0	0.0
6	CCSC(=NC1=CC=C(C=C1)C(F)(F)F)N.Cl	Inactive	0.0	0.0	0.0
8	CC1=CC=C(C=C1)S(=O)(=O)N2CCN(CC2)C3=NC(=NC4=CC…	Inactive	0.0	0.0	0.0
9	CC1=CC=C(C=C1)S(=O)(=O)N2CCN(CC2)C3=NC(=NC4=CC…	Inactive	0.0	0.0	0.0
10	C1CN(CCN1C2=NC(=NC3=CC=CC=C32)C4=CC=CS4)S(=O)(…	Inactive	0.0	0.0	0.0
…	…	…	…	…	…
342068	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0
342069	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0
342070	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0
342071	CC(=O)NC1=CC=C(C=C1)C(=O)N(CC2=CC=CC=C2)CC3=CC…	Inactive	0.0	0.0	0.0
342072	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0

[341260 rows x 5 columns]

[127]:
```
def is_charged(smiles):
    mol = Chem.MolFromSmiles(smiles)
    if not mol:
```

```
        return False  # Invalid SMILES
    return any(atom.GetFormalCharge() != 0 for atom in mol.GetAtoms())

# Test the function
print(is_charged("CC1=C(SC(=C1C#N)NC(=O)C2=CC(C=C2)OC)[N+](=O)"))
```

```
True
```

[128]:
```
df_50['Charged'] = df_50['SMILES'].apply(is_charged)

uncharged = df_50[df_50['Charged'] == False]
uncharged
```

```
/var/folders/jn/kkchdcr94t50xrmycsvkq2x80000gn/T/ipykernel_85584/162626946.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_50['Charged'] = df_50['SMILES'].apply(is_charged)
```

[128]:

	SMILES	Activity	Score	Potency	Efficacy	Charged
5	CNCC1=NC2=C(C=C(C=C2)Cl)C(=N1)C3=CC=CN3	Inactive	0.0	0.0	0.0	False
6	CCSC(=NC1=CC=C(C=C1)C(F)(F)F)N.Cl	Inactive	0.0	0.0	0.0	False
8	CC1=CC=C(C=C1)S(=O)(=O)N2CCN(CC2)C3=NC(=NC4=CC…	Inactive	0.0	0.0	0.0	False
9	CC1=CC=C(C=C1)S(=O)(=O)N2CCN(CC2)C3=NC(=NC4=CC…	Inactive	0.0	0.0	0.0	False
10	C1CN(CCN1C2=NC(=NC3=CC=CC=C32)C4=CC=CS4)S(=O)(…	Inactive	0.0	0.0	0.0	False
…	…	…	…	…	…	…
342068	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0	False
342069	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0	False
342070	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0	False
342071	CC(=O)NC1=CC=C(C=C1)C(=O)N(CC2=CC=CC=C2)CC3=CC…	Inactive	0.0	0.0	0.0	False
342072	CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…	Inactive	0.0	0.0	0.0	False

```
[322199 rows x 6 columns]
```

[129]:
```python
# Picking all "Active" molecules from the dataset
active_df = uncharged[uncharged['Activity'] == 'Active']
active_df.info()

# Picking all "Inactive" molecules from the dataset
inactive_df = uncharged[uncharged['Activity'] == 'Inactive']
inactive_df.info()

# Randomly sample from inactive_df to match the size of active_df
inactive_sampled = inactive_df.sample(n=len(active_df), random_state=42)

# Combine the active and sampled inactive molecules
balanced_df = pd.concat([active_df, inactive_sampled])

# Shuffle the combined dataset
balanced_df = balanced_df.sample(frac=1, random_state=42).reset_index(drop=True)

balanced_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 6273 entries, 13 to 341825
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   SMILES    6273 non-null   object
 1   Activity  6273 non-null   object
 2   Score     6273 non-null   float64
 3   Potency   6273 non-null   float64
 4   Efficacy  6273 non-null   float64
 5   Charged   6273 non-null   bool
dtypes: bool(1), float64(3), object(2)
memory usage: 300.2+ KB
<class 'pandas.core.frame.DataFrame'>
Index: 304069 entries, 5 to 342072
Data columns (total 6 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   SMILES    304069 non-null  object
 1   Activity  304069 non-null  object
 2   Score     304069 non-null  float64
 3   Potency   304069 non-null  float64
 4   Efficacy  304069 non-null  float64
 5   Charged   304069 non-null  bool
dtypes: bool(1), float64(3), object(2)
memory usage: 14.2+ MB
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 12546 entries, 0 to 12545
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   SMILES    12546 non-null  object
 1   Activity  12546 non-null  object
 2   Score     12546 non-null  float64
 3   Potency   12546 non-null  float64
 4   Efficacy  12546 non-null  float64
 5   Charged   12546 non-null  bool
dtypes: bool(1), float64(3), object(2)
memory usage: 502.5+ KB
```

[130]:
```
filtered_df = balanced_df
filtered_df
```

[130]:

	SMILES	Activity	Score	Potency	Efficacy	Charged
0	CC1=C(C=CC=C1Br)NC(=O)C2=C(C=CS2)N3C=CC=C3	Active	82.0	8.9125	140.7280	False
1	CCCCCC(C(C)CC(=O)NC1CCCCC1)C(=O)O	Active	43.0	12.5893	136.6590	False
2	CC1=CC=C(C=C1)S(=O)(=O)NC2=NN3C(C=C(NC3=N2)C)C…	Active	41.0	22.3872	166.6580	False
3	CC1=CC(=O)OC2=C1C=C(C=C2)OCC(=O)NC3=CC=CC(=C3)…	Inactive	0.0	0.0000	0.0000	False
4	CC1=CC(=C(N1C)C)C(=O)COC(=O)C23CC4CC(C2)CC(C4)…	Inactive	0.0	0.0000	0.0000	False
…	…	…	…	…	…	…
12541	C1CN(CCN1C(=O)C2=CC=CC=C2CC3=CC=CC=C3)S(=O)(=O…	Inactive	0.0	0.0000	0.0000	False
12542	C1=CC=C(C=C1)OC2=NC=NC(=C2)N3C=NC=N3	Active	64.0	2.8184	74.9734	False
12543	CC1=C(C(=CC=C1)N2CCN(CC2)C3=NC4=CC=CC=C4C(=O)N…	Active	42.0	17.7828	126.5240	False
12544	CCC(C)NC(=O)CSC1=NC2=CC=CC=C2C3=NC(C(=O)N31)C4…	Active	42.0	15.8489	139.3040	False
12545	CC(C)C1=CC=C(C=C1)S(=O)(=O)NC2CCCC2	Inactive	0.0	0.0000	0.0000	False

[12546 rows x 6 columns]

0.4 Parameter setting

[131]:
```
'''
scan through all the molecules to obtain unique atom types
'''
smiles = filtered_df['SMILES'].tolist()
search_elements=[]
for smile in smiles:
    mol = Chem.MolFromSmiles(smile)
    atoms = list(set([atom.GetSymbol() for atom in mol.GetAtoms()]))
    search_elements += atoms
    search_elements = list(set(search_elements))
search_elements.append("H")
print(search_elements)
```

```
['C', 'F', 'N', 'I', 'O', 'P', 'Br', 'B', 'S', 'Cl', 'As', 'H']
```

[132]:
```
'''
Setting up the atom mapping and bond mapping.
Code adopted from https://keras.io/examples/generative/molecule_generation/
'''
SMILE_CHARSET = str(search_elements)
bond_mapping = {"SINGLE": 0, "DOUBLE": 1, "TRIPLE": 2, "AROMATIC": 3}
bond_mapping.update(
    {0: BondType.SINGLE, 1: BondType.DOUBLE, 2: BondType.TRIPLE, 3: BondType.
 ↪AROMATIC}
)
SMILE_CHARSET = ast.literal_eval(SMILE_CHARSET)

MAX_MOLSIZE = max(filtered_df['SMILES'].str.len())
SMILE_to_index = dict((c, i) for i, c in enumerate(SMILE_CHARSET))
index_to_SMILE = dict((i, c) for i, c in enumerate(SMILE_CHARSET))
atom_mapping = dict(SMILE_to_index)
atom_mapping.update(index_to_SMILE)
print(atom_mapping)
print("Max molecule size: {}".format(MAX_MOLSIZE))
print("Character set Length: {}".format(len(SMILE_CHARSET)))
```

```
{'C': 0, 'F': 1, 'N': 2, 'I': 3, 'O': 4, 'P': 5, 'Br': 6, 'B': 7, 'S': 8, 'Cl':
9, 'As': 10, 'H': 11, 0: 'C', 1: 'F', 2: 'N', 3: 'I', 4: 'O', 5: 'P', 6: 'Br',
7: 'B', 8: 'S', 9: 'Cl', 10: 'As', 11: 'H'}
Max molecule size: 117
Character set Length: 12
```

0.5 Hyperparameters

[133]:
```
'''
Defining the Hyperparameters of the model
'''

NUM_ATOMS = 50 #Max number of atoms
ATOM_DIM = len(SMILE_CHARSET)  # Number of atom types
BOND_DIM = 5 # Number of bond types
```

0.6 Molecule featurization

[134]:
```
'''
Defining functions to convert smiles string into node graph and recover␣
 ↪molecule structure from it.
Code referenced from: https://keras.io/examples/generative/molecule_generation/
'''


def smiles_to_graph(smiles):
    '''
    Reference: https://keras.io/examples/generative/wgan-graphs/
    '''
    # Converts SMILES to molecule object
    molecule = Chem.MolFromSmiles(smiles)
    #molecule = Chem.AddHs(molecule)
    # Initialize adjacency and feature tensor
    adjacency = np.zeros((BOND_DIM, NUM_ATOMS, NUM_ATOMS), "float32")
    features = np.zeros((NUM_ATOMS, ATOM_DIM), "float32")

    # loop over each atom in molecule
    for atom in molecule.GetAtoms():
        i = atom.GetIdx()
        atom_type = atom_mapping[atom.GetSymbol()]
        features[i] = np.eye(ATOM_DIM)[atom_type]
        # loop over one-hop neighbors
        for neighbor in atom.GetNeighbors():
            j = neighbor.GetIdx()
            bond = molecule.GetBondBetweenAtoms(i, j)
            bond_type_idx = bond_mapping[bond.GetBondType().name]
            adjacency[bond_type_idx, [i, j], [j, i]] = 1

    # Where no bond, add 1 to last channel (indicating "non-bond")
    # Notice: channels-first
    adjacency[-1, np.sum(adjacency, axis=0) == 0] = 1

    # Where no atom, add 1 to last column (indicating "non-atom")
```

```
    features[np.where(np.sum(features, axis=1) == 0)[0], -1] = 1

    return adjacency, features

def graph_to_molecule(adjacency, features):
    # RWMol is a molecule object intended to be edited
    molecule = Chem.RWMol()
    # Remove "no atoms" & atoms with no bonds
    keep_idx = np.where(
        (np.argmax(features, axis=1) != ATOM_DIM - 1)
        & (np.sum(adjacency[:-1], axis=(0, 1)) > 0))[0]

    features = features[keep_idx]
    adjacency = adjacency[:, keep_idx][:, :, keep_idx]

    # Add atoms to molecule
    for atom_type_idx in np.argmax(features, axis=1):
        atom = Chem.Atom(atom_mapping[atom_type_idx])
        _ = molecule.AddAtom(atom)

    added_bonds = set()
    (bonds_ij, atoms_i, atoms_j) = np.where(np.triu(adjacency) == 1)
    for (bond_ij, atom_i, atom_j) in zip(bonds_ij, atoms_i, atoms_j):
        if atom_i == atom_j or bond_ij == BOND_DIM - 1:
            continue
        bond_type = bond_mapping.get(bond_ij, None)
        if (atom_i, atom_j) in added_bonds or (atom_j, atom_i) in added_bonds:

            continue
        molecule.AddBond(int(atom_i), int(atom_j), bond_type)
        added_bonds.add((atom_i, atom_j))


     # Sanitize without Kekulization
    try:
        Chem.SanitizeMol(molecule, sanitizeOps=Chem.SanitizeFlags.SANITIZE_ALL␣
 ↪^ Chem.SanitizeFlags.SANITIZE_KEKULIZE)
    except Exception as e:
        print(f"Sanitization failed: {e}")
        return None

    # Add explicit hydrogens
    molecule_with_h = Chem.AddHs(molecule)

    # Fix aromaticity in aromatic rings
    for atom in molecule_with_h.GetAtoms():
        if atom.GetIsAromatic():
```

```
            atom.SetIsAromatic(False)  # Clear aromaticity if needed

    # Force Kekulization to alternate bond orders in aromatic rings
    try:
        Chem.Kekulize(molecule_with_h, clearAromaticFlags=True)
    except Chem.KekulizeException as e:
        print(f"Kekulization failed: {e}")
        return molecule_with_h  # Return molecule without Kekulé bonds

    return molecule_with_h
```

0.7 Building model

[135]:
```
'''
    Defining GCN
    Reference: https://keras.io/examples/generative/wgan-graphs/
    The Encoder takes as input a molecule's graph adjacency matrix and feature␣
 ↪matrix.
'''
class RelationalGraphConvLayer(keras.layers.Layer):
    def __init__(
        self,
        units=128,
        activation="relu",
        use_bias=False,
        kernel_initializer="glorot_uniform",
        bias_initializer="zeros",
        kernel_regularizer=None,
        bias_regularizer=None,
        **kwargs
    ):
        super().__init__(**kwargs)

        self.units = units
        self.activation = keras.activations.get(activation)
        self.use_bias = use_bias
        self.kernel_initializer = keras.initializers.get(kernel_initializer)
        self.bias_initializer = keras.initializers.get(bias_initializer)
        self.kernel_regularizer = keras.regularizers.get(kernel_regularizer)
        self.bias_regularizer = keras.regularizers.get(bias_regularizer)

    def build(self, input_shape):
        bond_dim = input_shape[0][1]
        atom_dim = input_shape[1][2]

        self.kernel = self.add_weight(
            shape=(bond_dim, atom_dim, self.units),
```

```
            initializer=self.kernel_initializer,
            regularizer=self.kernel_regularizer,
            trainable=True,
            name="W",
            dtype=tf.float32,
        )

        if self.use_bias:
            self.bias = self.add_weight(
                shape=(bond_dim, 1, self.units),
                initializer=self.bias_initializer,
                regularizer=self.bias_regularizer,
                trainable=True,
                name="b",
                dtype=tf.float32,
            )

        self.built = True

    def call(self, inputs, training=False):
        adjacency, features = inputs
        # Aggregate information from neighbors
        x = tf.matmul(adjacency, features[:, None, :, :])
        # Apply linear transformation
        x = tf.matmul(x, self.kernel)
        if self.use_bias:
            x += self.bias
        # Reduce bond types dim
        x_reduced = tf.reduce_sum(x, axis=1)
        # Apply non-linear transformation
        return self.activation(x_reduced)
```

0.8 Build the Encoder and Decoder

[136]:
```
'''
defining function to build encoder and decoder.
Code adopted and modified from https://keras.io/examples/generative/
 ↪molecule_generation/
'''

def get_encoder(gconv_units, latent_dim, adjacency_shape, feature_shape,␣
 ↪dense_units, dropout_rate, regularizer=None):
    adjacency = keras.layers.Input(shape=adjacency_shape,␣
 ↪name="adjacency_input")
    features = keras.layers.Input(shape=feature_shape, name="feature_input")
    scores = keras.layers.Input(shape=(1,), name="score_input")  # Conditional␣
 ↪input (scalar)
```

```
    # Graph convolution layers
    features_transformed = features
    for units in gconv_units:
        features_transformed = RelationalGraphConvLayer(units)(
            [adjacency, features_transformed]
        )

    # Reduce 2D representation to 1D
    x = keras.layers.GlobalAveragePooling1D()(features_transformed)

    # Concatenate the score (condition) to the reduced graph representation
    x = keras.layers.Concatenate()([x, scores])

    # Fully connected layers
    for units in dense_units:
        x = layers.Dense(units, activation="relu",␣
↪kernel_regularizer=regularizer)(x)
        x = layers.Dropout(dropout_rate)(x)

    # Latent space
    z_mean = layers.Dense(latent_dim, name="z_mean")(x)
    z_log_var = layers.Dense(latent_dim, name="z_log_var")(x)

    # Create encoder model
    encoder = keras.Model(inputs=[adjacency, features, scores],␣
↪outputs=[z_mean, z_log_var], name="encoder")
    encoder.summary()
    return encoder


class SymmetrizeLayer(layers.Layer):
    def call(self, x):
        return (x + tf.transpose(x, (0, 1, 3, 2))) / 2

def get_decoder(dense_units, latent_dim, adjacency_shape, feature_shape,␣
↪dropout_rate, regularizer=None):
    latent_input = keras.Input(shape=(latent_dim,), name="latent_input")
    scores = keras.Input(shape=(1,), name="score_input")  # Conditional input␣
↪(scalar)

    # Concatenate latent input with the conditional score
    x = keras.layers.Concatenate()([latent_input, scores])

    # Dense layers
    for units in dense_units:
```

```
        x = keras.layers.Dense(units, activation="tanh",␣
↪kernel_regularizer=regularizer)(x)
        x = keras.layers.Dropout(dropout_rate)(x)

    # Adjacency reconstruction
    adj_output = keras.layers.Dense(tf.math.reduce_prod(adjacency_shape).
↪numpy().astype(int))(x)
    adj_output = keras.layers.Reshape(adjacency_shape)(adj_output)
    adj_output = SymmetrizeLayer()(adj_output)
    adj_output = keras.layers.Softmax(axis=1)(adj_output)

    # Feature reconstruction
    feat_output = keras.layers.Dense(tf.math.reduce_prod(feature_shape).numpy().
↪astype(int))(x)
    feat_output = keras.layers.Reshape(feature_shape)(feat_output)
    feat_output = keras.layers.Softmax(axis=2)(feat_output)

    # Create decoder model
    decoder = keras.Model(inputs=[latent_input, scores], outputs=[adj_output,␣
↪feat_output], name="decoder")
    decoder.summary()
    return decoder
```

0.9 Build the VAE

[137]:
```
'''
defining the VAE
Code adopted and modified from https://keras.io/examples/generative/
 ↪molecule_generation/
'''

class VAE(keras.Model):
    def __init__(self, encoder, decoder, beta=1.0, **kwargs):
        super(VAE, self).__init__(**kwargs)
        self.encoder = encoder
        self.decoder = decoder
        self.beta = beta

    def call(self, inputs):
        adjacency, features, scores = inputs
        z_mean, z_log_var = self.encoder([adjacency, features, scores])
        z = self.reparameterize(z_mean, z_log_var)
        return self.decoder([z, scores])
    def sampling(self, args):
        """
        Reparameterization trick: Sample from a Gaussian distribution using
```

```
        z = z_mean + epsilon * exp(z_log_var / 2), where epsilon is sampled␣
↪from N(0, 1).
        """
        z_mean, z_log_var = args
        batch = tf.shape(z_mean)[0]
        dim = tf.shape(z_mean)[1]
        epsilon = tf.keras.backend.random_normal(shape=(batch, dim))  #␣
↪Standard normal noise
        return z_mean + tf.exp(0.5 * z_log_var) * epsilon
```

0.10 Model training

[138]:
```
'''
splitting the dataset into training and testing
'''

train, test = train_test_split(filtered_df,test_size=0.2,random_state=42)
train_df, val_df = train_test_split(train, test_size=0.2, random_state=42)
train_df.reset_index(drop=True, inplace=True)
val_df.reset_index(drop=True, inplace=True)
test.reset_index(drop=True, inplace=True)

adj_train, fea_train, score_train = [], [], []
adj_val, fea_val, score_val = [], [], []

for idx in range(len(train_df)):
    adjacency, features = smiles_to_graph(train_df.loc[idx]["SMILES"])
    score = train_df.loc[idx]["Score"]
    adj_train.append(adjacency)
    fea_train.append(features)
    score_train.append(score)

for idx in range(len(val_df)):
    adjacency, features = smiles_to_graph(val_df.loc[idx]["SMILES"])
    score = val_df.loc[idx]["Score"]
    adj_val.append(adjacency)
    fea_val.append(features)
    score_val.append(score)


adj_train = np.array(adj_train)
fea_train = np.array(fea_train)
score_train_ = np.array(score_train).reshape(-1,1)

adj_val = np.array(adj_val)
fea_val = np.array(fea_val)
score_val_ = np.array(score_val).reshape(-1,1)
```

[139]:
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

score_train_n = scaler.fit_transform(score_train_)
score_val_n = scaler.transform(score_val_)
```

[140]:
```
print(adj_train.shape)
print(fea_train.shape)
print(score_train_.shape)
print(adj_val.shape)
print(fea_val.shape)
print(score_val_.shape)
```

```
(8028, 5, 50, 50)
(8028, 50, 12)
(8028, 1)
(2008, 5, 50, 50)
(2008, 50, 12)
(2008, 1)
```

[141]:
```
print(np.max(score_train_n))
```

```
0.9999999999999999
```

[ ]:
```
#Hyperparameters
BATCH_SIZE = 64
EPOCHS = 20
VAE_LR = 3e-4 # changed to 1e-3
LATENT_DIM = 32  # Size of the latent space
```

[143]:
```
'''
compiling the VAE
'''

encoder = get_encoder(
    gconv_units=[16],
    adjacency_shape=(BOND_DIM, NUM_ATOMS, NUM_ATOMS),
    feature_shape=(NUM_ATOMS, ATOM_DIM),
    latent_dim=LATENT_DIM,
    dense_units=[256, 512],
    dropout_rate=0,
    regularizer=l1_l2(l1=1e-6, l2=1e-3)
)
decoder = get_decoder(
    dense_units=[128, 256, 512],
    dropout_rate=0.3,
    latent_dim=LATENT_DIM,
```

```
    adjacency_shape=(BOND_DIM, NUM_ATOMS, NUM_ATOMS),
    feature_shape=(NUM_ATOMS, ATOM_DIM),
    regularizer=l1_l2(l1=1e-4, l2=1e-2)
)
vae = VAE(encoder, decoder)

vae.compile(optimizer=keras.optimizers.Adam(learning_rate=VAE_LR))
```

Model: "encoder"

Layer (type)	Output Shape	Param #	Connected to
adjacency_input (InputLayer)	(None, 5, 50, 50)	0	-
feature_input (InputLayer)	(None, 50, 12)	0	-
relational_graph_c… (RelationalGraphCo…	(None, 50, 16)	960	adjacency_input[… feature_input[0]…
global_average_poo… (GlobalAveragePool…	(None, 16)	0	relational_graph…
score_input (InputLayer)	(None, 1)	0	-
concatenate_6 (Concatenate)	(None, 17)	0	global_average_p… score_input[0][0]
dense_21 (Dense)	(None, 256)	4,608	concatenate_6[0]…
dropout_15 (Dropout)	(None, 256)	0	dense_21[0][0]
dense_22 (Dense)	(None, 512)	131,584	dropout_15[0][0]
dropout_16 (Dropout)	(None, 512)	0	dense_22[0][0]
z_mean (Dense)	(None, 32)	16,416	dropout_16[0][0]
z_log_var (Dense)	(None, 32)	16,416	dropout_16[0][0]

Total params: 169,984 (664.00 KB)

Trainable params: 169,984 (664.00 KB)

Non-trainable params: 0 (0.00 B)

Model: "decoder"

Layer (type)	Output Shape	Param #	Connected to
latent_input (InputLayer)	(None, 32)	0	-
score_input (InputLayer)	(None, 1)	0	-
concatenate_7 (Concatenate)	(None, 33)	0	latent_input[0][… score_input[0][0]
dense_23 (Dense)	(None, 128)	4,352	concatenate_7[0]…
dropout_17 (Dropout)	(None, 128)	0	dense_23[0][0]
dense_24 (Dense)	(None, 256)	33,024	dropout_17[0][0]
dropout_18 (Dropout)	(None, 256)	0	dense_24[0][0]
dense_25 (Dense)	(None, 512)	131,584	dropout_18[0][0]
dropout_19 (Dropout)	(None, 512)	0	dense_25[0][0]
dense_26 (Dense)	(None, 12500)	6,412,500	dropout_19[0][0]
reshape_6 (Reshape)	(None, 5, 50, 50)	0	dense_26[0][0]
dense_27 (Dense)	(None, 600)	307,800	dropout_19[0][0]
symmetrize_layer_3 (SymmetrizeLayer)	(None, 5, 50, 50)	0	reshape_6[0][0]
reshape_7 (Reshape)	(None, 50, 12)	0	dense_27[0][0]
softmax_6 (Softmax)	(None, 5, 50, 50)	0	symmetrize_layer…

```
 softmax_7 (Softmax)   (None, 50, 12)              0   reshape_7[0][0]


Total params: 6,889,260 (26.28 MB)

Trainable params: 6,889,260 (26.28 MB)

Non-trainable params: 0 (0.00 B)
```

[144]:
```
val_loss_list = []
train_loss_list = []
kl_theshold = 1.0
```

[145]:
```
train_dataset = tf.data.Dataset.from_tensor_slices((adj_train, fea_train,␣
 ↪score_train_)).batch(BATCH_SIZE)
val_dataset = tf.data.Dataset.from_tensor_slices((adj_val, fea_val,␣
 ↪score_val_)).batch(BATCH_SIZE)
```

[146]:
```
for epoch in range(EPOCHS):
    print(f"Epoch {epoch + 1}/{EPOCHS}")
    if epoch < 10:
        beta = 0.05
    else:
        beta = epoch*0.01
    # Training Loop
    train_loss = 0
    for (adjacency, features, scores) in train_dataset:
        with tf.GradientTape() as tape:
            # Forward pass
            z_mean, z_log_var = vae.encoder([adjacency, features, scores])
            z = vae.sampling([z_mean, z_log_var])
            adj_reconstruction, feature_reconstruction = vae.decoder([z,␣
 ↪scores])

            # Compute losses
            adj_loss = tf.reduce_mean(
                tf.reduce_sum(keras.losses.binary_crossentropy(adjacency,␣
 ↪adj_reconstruction), axis=(1, 2))
            )
            feat_loss = tf.reduce_mean(
                tf.reduce_sum(keras.losses.categorical_crossentropy(features,␣
 ↪feature_reconstruction), axis=1)
            )
            reconstruction_loss = adj_loss + feat_loss
```

```
            kl_loss = -0.5 * tf.reduce_mean(
                tf.reduce_sum(1 + z_log_var - tf.square(z_mean) - tf.
↪exp(z_log_var), axis=1)
            )
            total_loss = reconstruction_loss + beta * kl_loss

        # Backpropagation
        grads = tape.gradient(total_loss, vae.trainable_weights)
        vae.optimizer.apply_gradients(zip(grads, vae.trainable_weights))

        train_loss += total_loss


    train_loss /= len(train_dataset)
    train_loss_list.append(train_loss)

    print(f"Train Loss: {train_loss.numpy()}, KL Loss: {kl_loss.numpy()},␣
↪Reconstruction Loss: {reconstruction_loss.numpy()}")

    # Validation Loop
    val_loss = 0
    for (val_adjacency, val_features, val_scores) in val_dataset:
        # Forward pass
        z_mean, z_log_var = vae.encoder([val_adjacency, val_features,␣
↪val_scores])
        z = vae.sampling([z_mean, z_log_var])
        val_adj_reconstruction, val_feat_reconstruction = vae.decoder([z,␣
↪val_scores])

        # Compute losses
        val_adj_loss = tf.reduce_mean(
            tf.reduce_sum(keras.losses.binary_crossentropy(val_adjacency,␣
↪val_adj_reconstruction), axis=(1, 2))
        )
        val_feat_loss = tf.reduce_mean(
            tf.reduce_sum(keras.losses.categorical_crossentropy(val_features,␣
↪val_feat_reconstruction), axis=1)
        )
        val_reconstruction_loss = val_adj_loss + val_feat_loss
        val_kl_loss = -0.5 * tf.reduce_mean(
            tf.reduce_sum(1 + z_log_var - tf.square(z_mean) - tf.
↪exp(z_log_var), axis=1)
        )
        val_total_loss = val_reconstruction_loss + beta * val_kl_loss

        val_loss += val_total_loss
```

```
    val_loss /= len(val_dataset)
    val_loss_list.append(val_loss)

    # Adjust beta if KL loss is very low
    if kl_loss < kl_theshold:
        beta = 0.05
    print(f"Validation Loss: {val_loss.numpy()}, KL Loss: {val_kl_loss.
↪numpy()}, Reconstruction Loss: {val_reconstruction_loss.numpy()}")
    print('BETA is: ', beta)
```

```
Epoch 1/25
2024-12-12 18:23:02.256547: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
Train Loss: 66.52747344970703, KL Loss: 4.886928558349609, Reconstruction Loss:
36.74971008300781
2024-12-12 18:23:02.762948: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
Validation Loss: 38.502811431884766, KL Loss: 6.201170444488525, Reconstruction
Loss: 35.168861389160156
BETA is:  0.05
Epoch 2/25
2024-12-12 18:23:10.705236: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
Train Loss: 38.224117279052734, KL Loss: 6.976441383361816, Reconstruction Loss:
36.39094924926758
2024-12-12 18:23:11.179011: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
Validation Loss: 38.04998016357422, KL Loss: 9.482065200805664, Reconstruction
Loss: 34.69752502441406
BETA is:  0.05
Epoch 3/25
2024-12-12 18:23:19.107395: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
Train Loss: 36.93819046020508, KL Loss: 10.201919555664062, Reconstruction Loss:
33.31011199951172
2024-12-12 18:23:19.597557: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
Validation Loss: 34.09147262573242, KL Loss: 11.744572639465332, Reconstruction
Loss: 31.000350952148438
BETA is:  0.05
Epoch 4/25
```

```
2024-12-12 18:23:27.406056: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 31.90918731689453, KL Loss: 12.4201078414917, Reconstruction Loss:
29.842086791992188

2024-12-12 18:23:27.867937: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.965682983398438, KL Loss: 12.510169982910156, Reconstruction
Loss: 28.011402130126953
BETA is:  0.05
Epoch 5/25

2024-12-12 18:23:35.662549: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.366886138916016, KL Loss: 13.628972053527832, Reconstruction
Loss: 28.988393783569336

2024-12-12 18:23:36.128974: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.270843505859375, KL Loss: 13.827468872070312, Reconstruction
Loss: 26.98990249633789
BETA is:  0.05
Epoch 6/25

2024-12-12 18:23:43.941406: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.740842819213867, KL Loss: 13.964268684387207, Reconstruction
Loss: 28.50438690185547

2024-12-12 18:23:44.415251: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.574687957763672, KL Loss: 14.777920722961426, Reconstruction
Loss: 26.646087646484375
BETA is:  0.05
Epoch 7/25

2024-12-12 18:23:52.621742: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.343883514404297, KL Loss: 13.801233291625977, Reconstruction
Loss: 27.923551559448242

2024-12-12 18:23:53.231461: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.274795532226562, KL Loss: 14.807585716247559, Reconstruction
Loss: 26.101886749267578
BETA is:  0.05
Epoch 8/25
```

```
2024-12-12 18:24:01.184885: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.08763885498047, KL Loss: 14.02650260925293, Reconstruction Loss:
28.1483097076416

2024-12-12 18:24:01.655090: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.16486167907715, KL Loss: 14.69410228729248, Reconstruction
Loss: 26.109426498413086
BETA is:  0.05
Epoch 9/25

2024-12-12 18:24:09.480708: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.879562377929688, KL Loss: 13.525146484375, Reconstruction Loss:
27.651527404785156

2024-12-12 18:24:09.969375: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.99447250366211, KL Loss: 14.36849308013916, Reconstruction
Loss: 25.97069549560547
BETA is:  0.05
Epoch 10/25

2024-12-12 18:24:17.696605: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.678876876831055, KL Loss: 13.207070350646973, Reconstruction
Loss: 27.656959533691406

2024-12-12 18:24:18.186166: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.864803314208984, KL Loss: 14.110527038574219, Reconstruction
Loss: 26.04417610168457
BETA is:  0.05
Epoch 11/25

2024-12-12 18:24:26.044903: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.239404678344727, KL Loss: 9.701704025268555, Reconstruction Loss:
27.830154418945312

2024-12-12 18:24:26.528022: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.211071014404297, KL Loss: 10.316340446472168, Reconstruction
Loss: 25.799243927001953
BETA is:  0.1
Epoch 12/25
```

```
2024-12-12 18:24:34.466723: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.14613151550293, KL Loss: 8.713913917541504, Reconstruction Loss:
27.605607986450195

2024-12-12 18:24:34.955596: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.218189239501953, KL Loss: 9.365045547485352, Reconstruction
Loss: 26.10299301147461
BETA is:  0.11
Epoch 13/25

2024-12-12 18:24:42.786092: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.080175399780273, KL Loss: 8.490754127502441, Reconstruction Loss:
27.845727920532227

2024-12-12 18:24:43.260441: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.092191696166992, KL Loss: 9.114140510559082, Reconstruction
Loss: 25.66101837158203
BETA is:  0.12
Epoch 14/25

2024-12-12 18:24:51.168848: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.254180908203125, KL Loss: 7.608872413635254, Reconstruction Loss:
27.595111846923828

2024-12-12 18:24:51.661628: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.113887786865234, KL Loss: 8.195504188537598, Reconstruction
Loss: 26.07050323486328
BETA is:  0.13
Epoch 15/25

2024-12-12 18:24:59.830519: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.002410888671875, KL Loss: 7.052667140960693, Reconstruction Loss:
27.55022430419922

2024-12-12 18:25:00.366410: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.046220779418945, KL Loss: 7.515444278717041, Reconstruction
Loss: 25.568994522094727
BETA is:  0.14
Epoch 16/25
```

2024-12-12 18:25:08.152525: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.89946746826172, KL Loss: 6.477538108825684, Reconstruction Loss: 27.42659568786621

2024-12-12 18:25:08.636968: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.14095115661621, KL Loss: 7.105879306793213, Reconstruction Loss: 25.69719696044922
BETA is: 0.15
Epoch 17/25

2024-12-12 18:25:16.460407: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.964956283569336, KL Loss: 7.070982933044434, Reconstruction Loss: 27.42034912109375

2024-12-12 18:25:16.944221: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.318870544433594, KL Loss: 7.656010150909424, Reconstruction Loss: 25.80021858215332
BETA is: 0.16
Epoch 18/25

2024-12-12 18:25:24.787128: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.04618263244629, KL Loss: 6.109533786773682, Reconstruction Loss: 27.40243911743164

2024-12-12 18:25:25.290155: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.95831298828125, KL Loss: 6.42066764831543, Reconstruction Loss: 25.664520263671875
BETA is: 0.17
Epoch 19/25

2024-12-12 18:25:33.022012: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.81960105895996, KL Loss: 5.944121837615967, Reconstruction Loss: 27.126726150512695

2024-12-12 18:25:33.526175: W tensorflow/core/framework/local_rendezvous.cc:404] Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.986167907714844, KL Loss: 6.21658182144165, Reconstruction Loss: 25.300838470458984
BETA is: 0.18
Epoch 20/25

```
2024-12-12 18:25:41.686756: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.801240921020508, KL Loss: 5.6656951904296875, Reconstruction
Loss: 27.048612594604492

2024-12-12 18:25:42.207034: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.90081214904785, KL Loss: 5.953367710113525, Reconstruction
Loss: 25.62380599975586
BETA is:  0.19
Epoch 21/25

2024-12-12 18:25:50.485615: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.77643394470215, KL Loss: 5.486286163330078, Reconstruction Loss:
27.223676681518555

2024-12-12 18:25:50.967912: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.813480377197266, KL Loss: 5.755873203277588, Reconstruction
Loss: 25.33903694152832
BETA is:  0.2
Epoch 22/25

2024-12-12 18:25:58.947723: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.92287826538086, KL Loss: 6.485251426696777, Reconstruction Loss:
27.097675323486328

2024-12-12 18:25:59.413123: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.123573303222656, KL Loss: 7.25524377822876, Reconstruction
Loss: 25.454837799072266
BETA is:  0.21
Epoch 23/25

2024-12-12 18:26:07.280680: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.94455337524414, KL Loss: 5.325558662414551, Reconstruction Loss:
27.341506958007812

2024-12-12 18:26:07.741914: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.9595947265625, KL Loss: 5.6578898429870605, Reconstruction
Loss: 25.444793701171875
BETA is:  0.22
Epoch 24/25
```

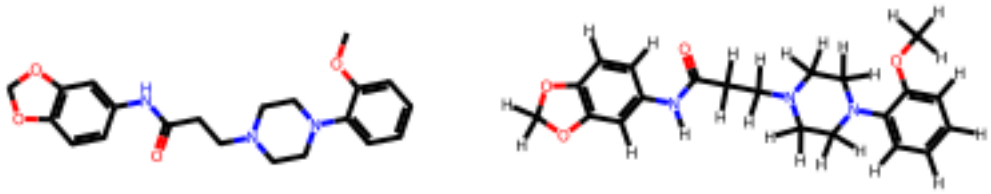
```
2024-12-12 18:26:15.517485: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.7772216796875, KL Loss: 5.038825988769531, Reconstruction Loss:
27.209758758544922

2024-12-12 18:26:15.991184: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 28.973308563232422, KL Loss: 5.352232456207275, Reconstruction
Loss: 25.51656723022461
BETA is:  0.23
Epoch 25/25

2024-12-12 18:26:23.782280: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 28.793554306030273, KL Loss: 4.8062005043029785, Reconstruction
Loss: 27.067350387573242
Validation Loss: 28.936031341552734, KL Loss: 5.058586597442627, Reconstruction
Loss: 25.398195266723633
BETA is:  0.24

2024-12-12 18:26:24.279811: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

[147]:
```
'''
Checking the model's ability to reconstruct a molecule from the training dataset
'''

i=10
adjacency_check, features_check = smiles_to_graph(train_df.loc[i]["SMILES"])
score_check = [train_df.loc[i]["Score"]]
molobj = Chem.MolFromSmiles(train_df.loc[i]["SMILES"])
adj0 = np.expand_dims(adjacency_check,axis=0)
feature0 = np.expand_dims(features_check,axis=0)
score0 = np.expand_dims(score_check,axis=0)
print(adj0.shape)
print(feature0.shape)
print(score0.shape)
```

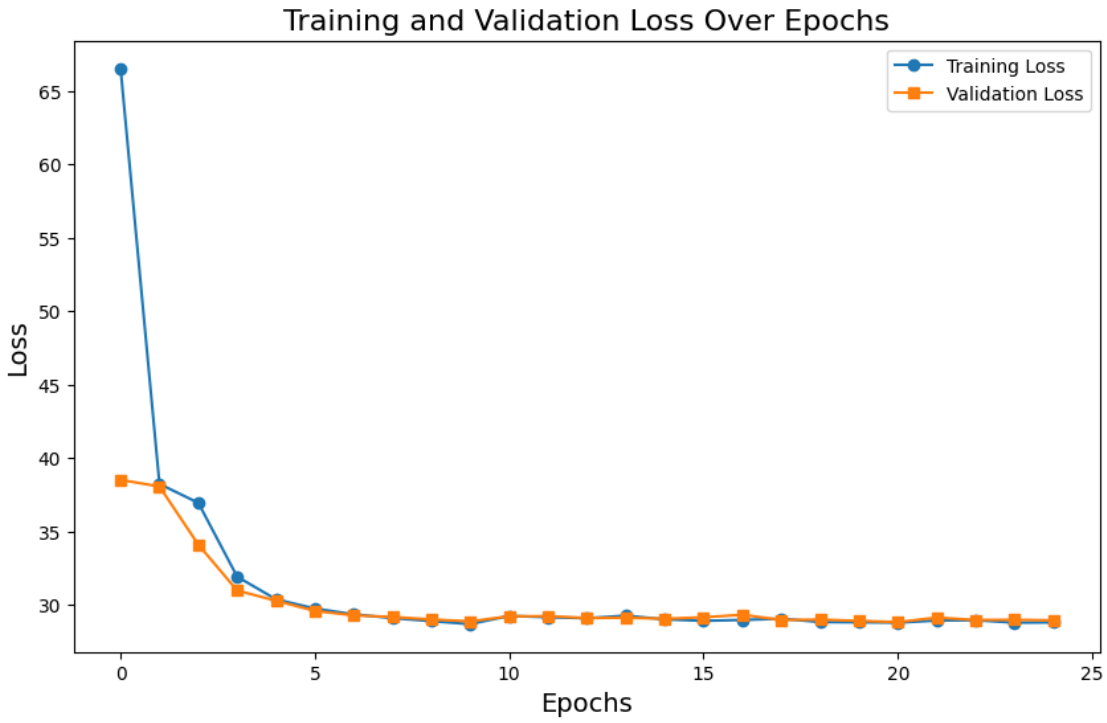
```
(1, 5, 50, 50)
(1, 50, 12)
(1, 1)
```

[148]:
```
mole_pred = graph_to_molecule(adj0[0], feature0[0])
Draw.MolsToGridImage([molobj,mole_pred], molsPerRow=2,)
```

[148]:



[149]:
```
plt.figure(figsize=(10, 6))
plt.plot(range(EPOCHS), train_loss_list, label='Training Loss', marker='o')
plt.plot(range(EPOCHS), val_loss_list, label='Validation Loss', marker='s')

# Add title and labels
plt.title('Training and Validation Loss Over Epochs', fontsize=16)
plt.xlabel('Epochs', fontsize=14)
plt.ylabel('Loss', fontsize=14)
plt.legend()
plt.show()
```



0.11 Visualize latent space

[150]:
```
adj_test, fea_test, score_test = [], [], []

for idx in range(len(test)):
    adjacency, features = smiles_to_graph(test.loc[idx]["SMILES"])
    score = test.loc[idx]["Score"]
    adj_test.append(adjacency)
    fea_test.append(features)
    score_test.append(score)


adj_test = np.array(adj_test)
fea_test = np.array(fea_test)
score_test_ = np.array(score_test).reshape(-1,1)

score_test_n = scaler.transform(score_test_)
```

[151]:
```
ls_train = vae.encoder.predict([adj_train, fea_train, score_train_])
ls_test = vae.encoder.predict([adj_test, fea_test, score_test_])
```

```
251/251             0s 841us/step
79/79             0s 738us/step
```

[152]:
```
ls_train_ = np.array(ls_train)
ls_test_ = np.array(ls_test)
```

[153]:
```
z_mean, _ = vae.encoder.predict([adj_test, fea_test, score_test_])
```

```
79/79             0s 763us/step
```

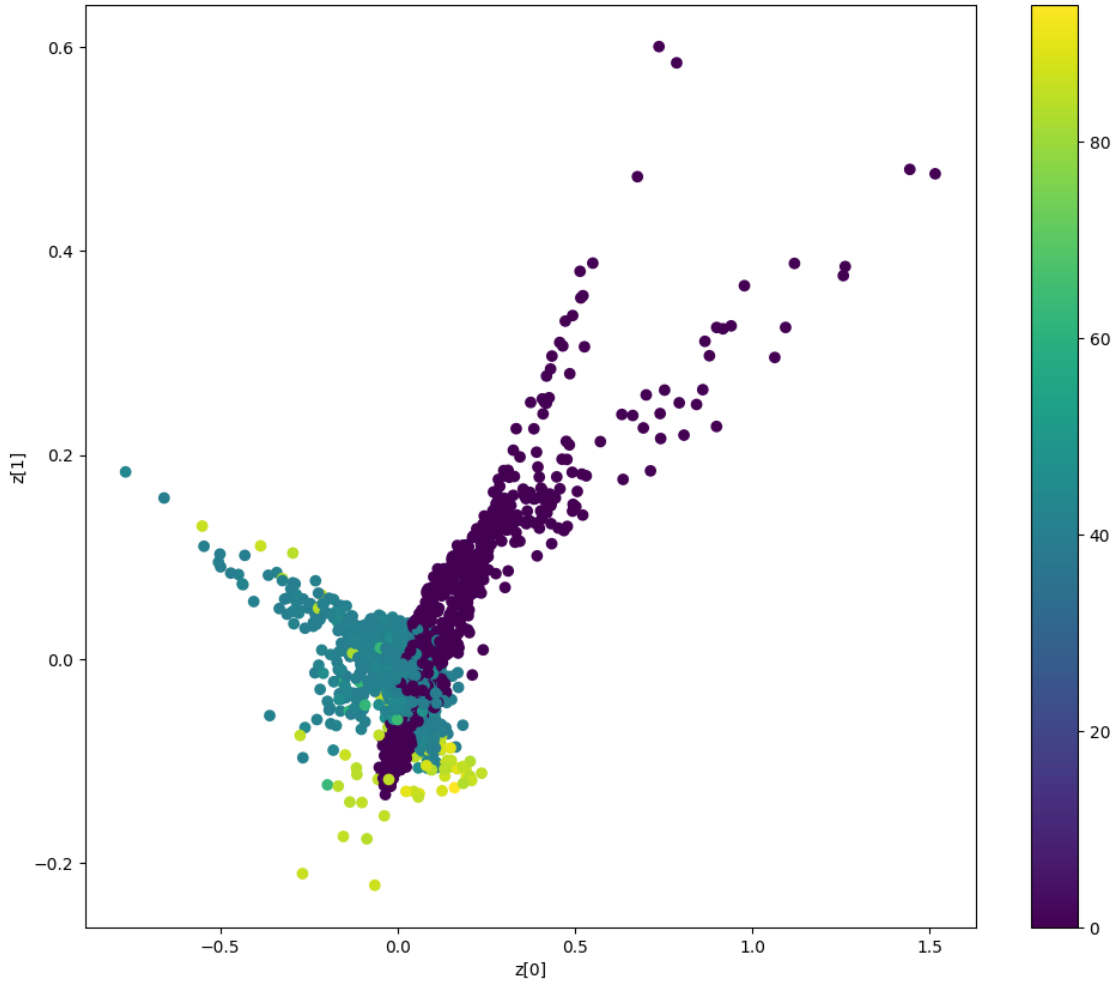
[154]:
```
latent_noise = np.random.normal(scale=0.1, size=z_mean.shape)  # Adjust scale␣
 ↪as needed
adj_pred, feature_pred = vae.decoder.predict([z_mean, score_test_])
print("Shape of adj_pred:", adj_pred.shape)
print("Shape of feature_pred:", feature_pred.shape)

# Reconstruct molecules
gen_molecules = [
    graph_to_molecule(adj_pred[i], feature_pred[i])
    for i in range(adj_pred.shape[0])
]
```

```
79/79             0s 4ms/step
Shape of adj_pred: (2510, 5, 50, 50)
Shape of feature_pred: (2510, 50, 12)
```

[155]:
```
from scipy.stats import pearsonr

# Correlate latent dimensions with molecular scores
correlations = [pearsonr(z_mean[:, i], score_test_.flatten())[0] for i in␣
 ↪range(z_mean.shape[1])]
print("Correlations between latent dimensions and scores:", correlations)
```

```
Correlations between latent dimensions and scores: [-0.3558200684763795,
-0.2062827847950312, -0.108689001255225, -0.09377558271044291,
0.33778056608173157, -0.054834416543380914, -0.2788737096229518,
0.5959523565604212, -0.09467381971400793, -0.1665799876161952,
-0.04127361161661077, -0.5570950410190803, -0.3395595470883653,
-0.27378418824024486, 0.6468365181351967, -0.02795534158860679, 0.8220557112552,
-0.20764569978374892, 0.19387425800095565, 0.4008222941286574,
0.4227638907040553, -0.5441018734401635, 0.09497631862139289,
-0.19985353870489508, 0.8473812426378595, -0.07234080430327028,
0.4745796450238354, 0.5332353078274981, -0.0043218133099310555,
0.14982205904964574, -0.24345643827421504, 0.7939340669546286]
```

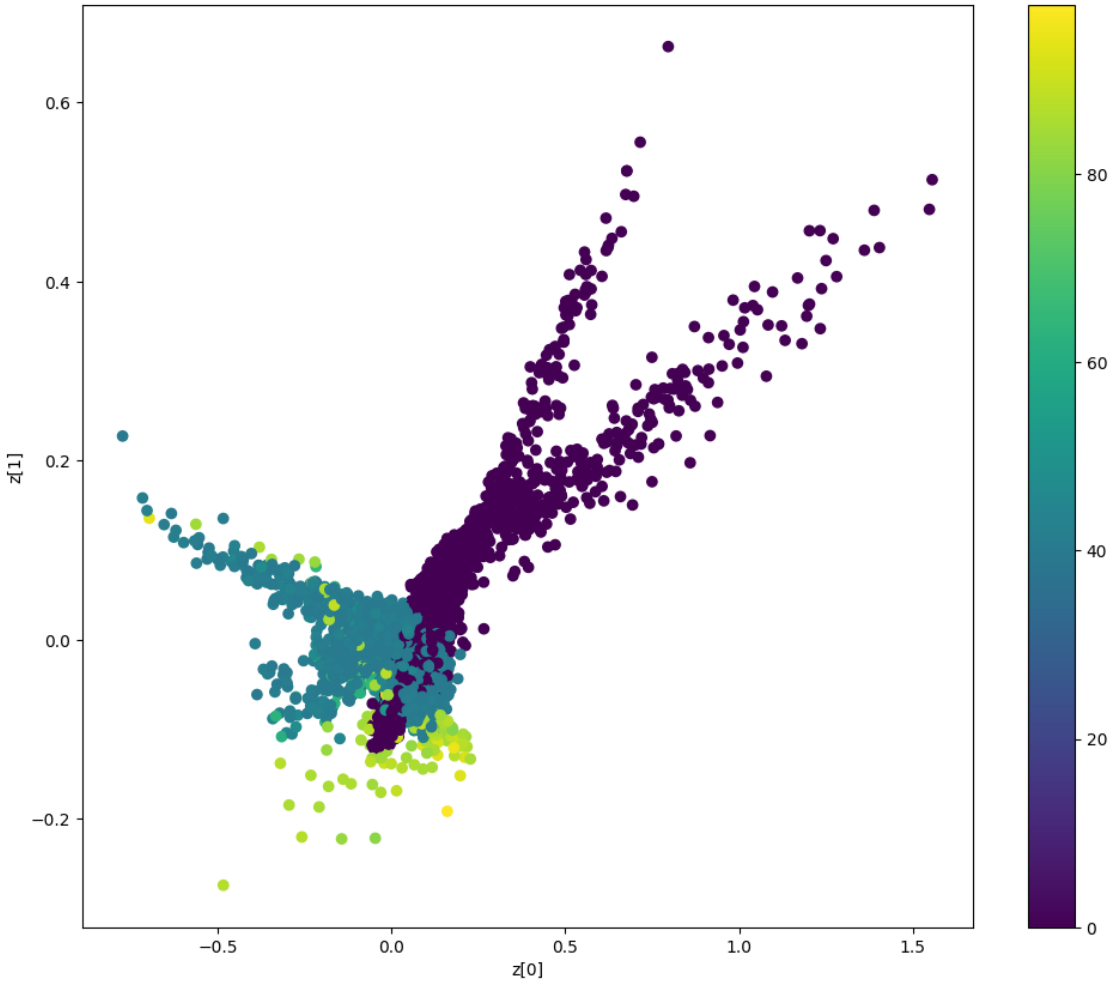
[156]:
```
plt.figure(figsize=(12, 10))
plt.scatter(z_mean[:, 0], z_mean[:, 1], c=score_test_)
plt.colorbar()
plt.xlabel("z[0]")
plt.ylabel("z[1]")
plt.show()
```



[157]:
```
z_mean2, _2 = vae.encoder.predict([adj_train, fea_train, score_train_])
```

```
251/251              0s 740us/step
```

[158]:
```
plt.figure(figsize=(12, 10))
plt.scatter(z_mean2[:, 0], z_mean2[:, 1], c=score_train_)
plt.colorbar()
plt.xlabel("z[0]")
plt.ylabel("z[1]")
plt.show()
```



[ ]:

0.12 Model Inferencing

We would be inferring our model to predict over random latent space and try to generate 100 new valid molecules.

0.12.1 Generate unique Molecules with the model

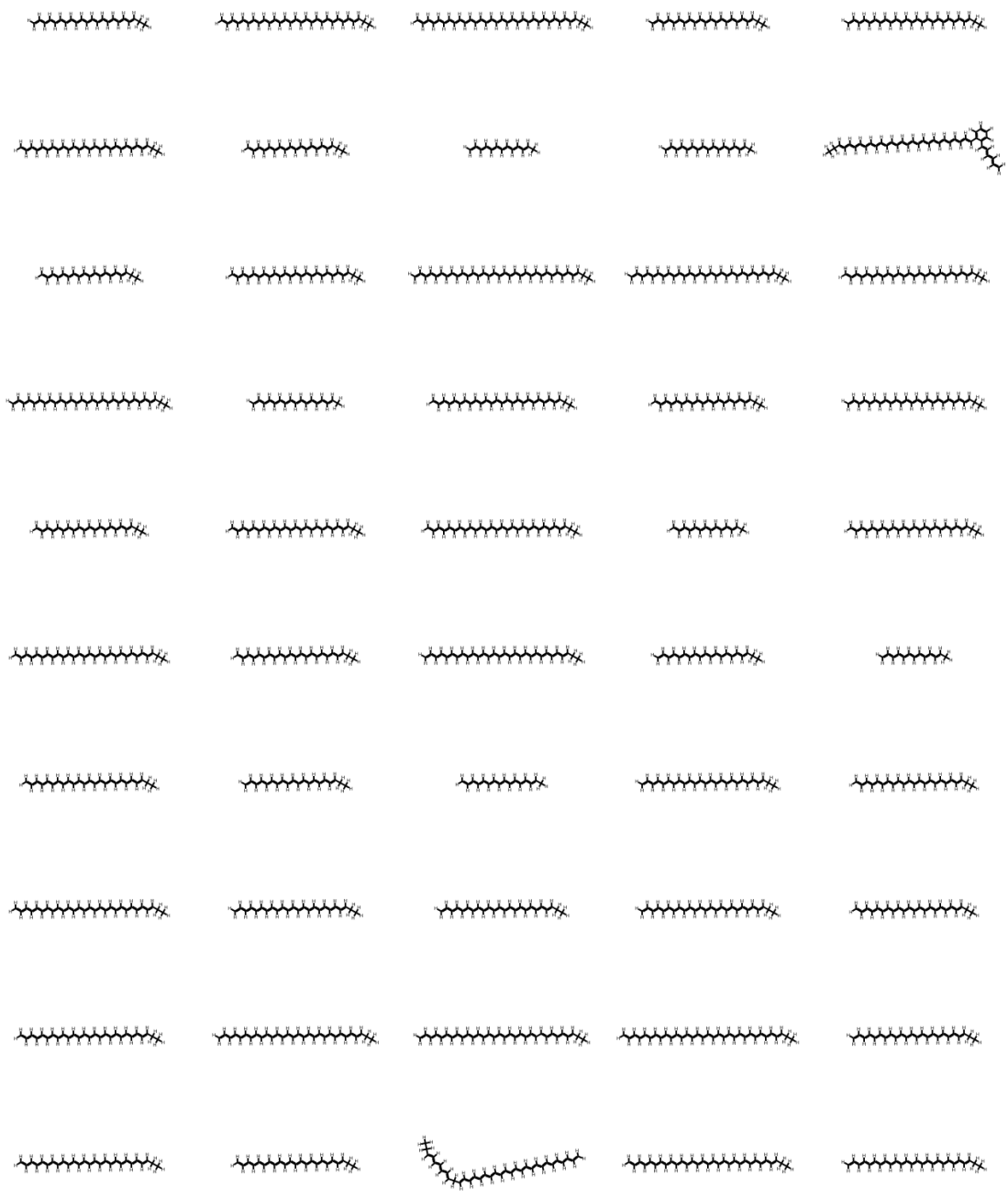
[159]:
```
def inference(model=vae, batch_size=1000, dim = LATENT_DIM, activity=10):
    z = np.random.normal(size=(batch_size, dim))
    activityarray = (np.zeros(batch_size) + activity).reshape(-1,1)

    reconstruction_adjacency, reconstruction_features = model.decoder.
 ↪predict([z,activityarray])
    # obtain one-hot encoded adjacency tensor
```

```
    adjacency = tf.argmax(reconstruction_adjacency, axis=1)
    adjacency = tf.one_hot(adjacency, depth=BOND_DIM, axis=1)
    # Remove potential self-loops from adjacency
    adjacency = tf.linalg.set_diag(adjacency, tf.zeros(tf.shape(adjacency)[:
 ↪-1]))
    # obtain one-hot encoded feature tensor
    features = tf.argmax(reconstruction_features, axis=2)
    features = tf.one_hot(features, depth=ATOM_DIM, axis=2)

    return [
        graph_to_molecule(adjacency[i].numpy(), features[i].numpy())
        for i in range(batch_size)
    ]
```

[160]:
```
gen_mols = inference(batch_size=1000,activity=10)
MolsToGridImage([m for m in gen_mols if m is not None][:1000], molsPerRow=5,␣
 ↪subImgSize=(260, 160))
```

```
32/32             0s 4ms/step
Sanitization failed: Explicit valence for atom # 18 C, 7, is greater than
permitted
Sanitization failed: Explicit valence for atom # 0 N, 69, is greater than
permitted
Sanitization failed: Explicit valence for atom # 25 C, 6, is greater than
permitted
Sanitization failed: Explicit valence for atom # 25 C, 6, is greater than
permitted
Sanitization failed: Explicit valence for atom # 20 C, 6, is greater than
permitted
Sanitization failed: Explicit valence for atom # 25 C, 6, is greater than
permitted
Sanitization failed: Explicit valence for atom # 20 C, 6, is greater than
permitted
Kekulization failed: Can't kekulize mol.  Unkekulized atoms: 35 36 37
Sanitization failed: Explicit valence for atom # 14 N, 6, is greater than
permitted
Sanitization failed: Explicit valence for atom # 25 C, 6, is greater than
permitted

/Users/thinh/Library/Python/3.12/lib/python/site-
packages/rdkit/Chem/Draw/IPythonConsole.py:261: UserWarning: Truncating the list
of molecules to be displayed to 50. Change the maxMols value to display more.
  warnings.warn(
```

[160]:



[ ]: