final

December 12, 2024

0.1 Abstract

0.2 Package import

[41]:
```
import os

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

import ast
import numpy as np

from tensorflow import keras
```

[42]:
```
#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.*")
```

[43]:
```
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

[44]:
```python
'''
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
```

[45]:
```
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]
```

[46]:
```
df_50
```

[46]:
```
                                                   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
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
10      C1CN(CCN1C2=NC(=NC3=CC=CC=C32)C4=CC=CS4)S(=O)(…  Inactive    0.0
…                                                     …         …      …
342068  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0
342069  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0
342070  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0
342071  CC(=O)NC1=CC=C(C=C1)C(=O)N(CC2=CC=CC=C2)CC3=CC…  Inactive    0.0
342072  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0

        Potency  Efficacy
5           0.0       0.0
6           0.0       0.0
8           0.0       0.0
9           0.0       0.0
10          0.0       0.0
…             …         …
342068      0.0       0.0
342069      0.0       0.0
342070      0.0       0.0
342071      0.0       0.0
342072      0.0       0.0

[341260 rows x 5 columns]
```

[47]:
```
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
```

[48]:
```
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)
```

[48]:
```
                                                   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
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
10      C1CN(CCN1C2=NC(=NC3=CC=CC=C32)C4=CC=CS4)S(=O)(…  Inactive    0.0
…                                                     …       …    …
342068  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0
342069  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0
342070  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0
342071  CC(=O)NC1=CC=C(C=C1)C(=O)N(CC2=CC=CC=C2)CC3=CC…  Inactive    0.0
342072  CC(=O)NC1=CC=C(C=C1)OCC2=C(C=CC(=C2)CN(CC3=CC=…  Inactive    0.0

        Potency  Efficacy  Charged
5           0.0       0.0    False
6           0.0       0.0    False
8           0.0       0.0    False
9           0.0       0.0    False
10          0.0       0.0    False
…           …        …        …
342068      0.0       0.0    False
342069      0.0       0.0    False
342070      0.0       0.0    False
342071      0.0       0.0    False
342072      0.0       0.0    False
```

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

[49]:
```
# 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
```

[50]:
```
filtered_df = balanced_df
filtered_df
```

[50]:
```
                                                  SMILES  Activity  Score  \
0             CC1=C(C=CC=C1Br)NC(=O)C2=C(C=CS2)N3C=CC=C3    Active   82.0
1                      CCCCCC(C(C)CC(=O)NC1CCCCC1)C(=O)O    Active   43.0
2      CC1=CC=C(C=C1)S(=O)(=O)NC2=NN3C(C=C(NC3=N2)C)C…    Active   41.0
3      CC1=CC(=O)OC2=C1C=C(C=C2)OCC(=O)NC3=CC=CC(=C3)…  Inactive    0.0
4      CC1=CC(=C(N1C)C)C(=O)COC(=O)C23CC4CC(C2)CC(C4)…  Inactive    0.0
…                                                    …        …     …
12541  C1CN(CCN1C(=O)C2=CC=CC=C2CC3=CC=CC=C3)S(=O)(=O…  Inactive    0.0
12542               C1=CC=C(C=C1)OC2=NC=NC(=C2)N3C=NC=N3    Active   64.0
12543  CC1=C(C(=CC=C1)N2CCN(CC2)C3=NC4=CC=CC=C4C(=O)N…    Active   42.0
12544  CCC(C)NC(=O)CSC1=NC2=CC=CC=C2C3=NC(C(=O)N31)C4…    Active   42.0
12545                CC(C)C1=CC=C(C=C1)S(=O)(=O)NC2CCCC2  Inactive    0.0

       Potency  Efficacy  Charged
0       8.9125  140.7280    False
1      12.5893  136.6590    False
2      22.3872  166.6580    False
3       0.0000    0.0000    False
4       0.0000    0.0000    False
…          …         …        …
12541   0.0000    0.0000    False
12542   2.8184   74.9734    False
12543  17.7828  126.5240    False
12544  15.8489  139.3040    False
12545   0.0000    0.0000    False

[12546 rows x 6 columns]
```

0.4 Parameter setting

[51]:
```
'''
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']
```

[52]:
```
'''
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

[53]:
```
'''
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

[54]:
```
'''
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

[55]:
```
'''
    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

[56]:
```
'''
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

[57]:
```
'''
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

[58]:
```
'''
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)
```

[59]:
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

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

[60]:
```
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)
```

[61]:
```
print(np.max(score_train_n))
```

```
0.9999999999999999
```

[62]:
```
#Hyperparameters
BATCH_SIZE = 64
EPOCHS = 25
VAE_LR = 3e-4 # changed to 1e-3
LATENT_DIM = 256  # Size of the latent space
```

[63]:
```
'''
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        (None, 5, 50, 50)             0   -
 (InputLayer)

 feature_input          (None, 50, 12)                0   -
 (InputLayer)

 relational_graph_c…    (None, 50, 16)              960   adjacency_input[…
 (RelationalGraphCo…                                      feature_input[0]…

 global_average_poo…    (None, 16)                    0   relational_graph…
 (GlobalAveragePool…

 score_input            (None, 1)                     0   -
 (InputLayer)

 concatenate_2          (None, 17)                    0   global_average_p…
 (Concatenate)                                            score_input[0][0]

 dense_7 (Dense)        (None, 256)               4,608   concatenate_2[0]…

 dropout_5 (Dropout)    (None, 256)                   0   dense_7[0][0]

 dense_8 (Dense)        (None, 512)             131,584   dropout_5[0][0]

 dropout_6 (Dropout)    (None, 512)                   0   dense_8[0][0]

 z_mean (Dense)         (None, 256)             131,328   dropout_6[0][0]

 z_log_var (Dense)      (None, 256)             131,328   dropout_6[0][0]


 Total params: 399,808 (1.53 MB)
 Trainable params: 399,808 (1.53 MB)

 Non-trainable params: 0 (0.00 B)

Model: "decoder"

 Layer (type)           Output Shape           Param #   Connected to

 latent_input           (None, 256)                  0   -
 (InputLayer)

 score_input            (None, 1)                    0   -
 (InputLayer)

 concatenate_3          (None, 257)                  0   latent_input[0][…
 (Concatenate)                                           score_input[0][0]

 dense_9 (Dense)        (None, 128)             33,024   concatenate_3[0]…

 dropout_7 (Dropout)    (None, 128)                  0   dense_9[0][0]

 dense_10 (Dense)       (None, 256)             33,024   dropout_7[0][0]

 dropout_8 (Dropout)    (None, 256)                  0   dense_10[0][0]

 dense_11 (Dense)       (None, 512)            131,584   dropout_8[0][0]

 dropout_9 (Dropout)    (None, 512)                  0   dense_11[0][0]

 dense_12 (Dense)       (None, 12500)        6,412,500   dropout_9[0][0]

 reshape_2 (Reshape)    (None, 5, 50, 50)            0   dense_12[0][0]

 dense_13 (Dense)       (None, 600)            307,800   dropout_9[0][0]

 symmetrize_layer_1     (None, 5, 50, 50)            0   reshape_2[0][0]
 (SymmetrizeLayer)

 reshape_3 (Reshape)    (None, 50, 12)               0   dense_13[0][0]

 softmax_2 (Softmax)    (None, 5, 50, 50)            0   symmetrize_layer…

 softmax_3 (Softmax)    (None, 50, 12)               0   reshape_3[0][0]
```

```
Total params: 6,917,932 (26.39 MB)

Trainable params: 6,917,932 (26.39 MB)

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

[64]:
```
val_loss_list = []
train_loss_list = []
kl_theshold = 1.0
```

[65]:
```
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)
```

[66]:
```
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 17:21:02.544078: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 68.91028594970703, KL Loss: 7.779871940612793, Reconstruction Loss:
36.89194869995117

2024-12-12 17:21:03.088554: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 38.474632263183594, KL Loss: 8.145179748535156, Reconstruction
Loss: 35.19890213012695
BETA is:  0.05
Epoch 2/25

2024-12-12 17:21:11.252029: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 38.19132614135742, KL Loss: 6.249152183532715, Reconstruction Loss:
36.3538703918457

2024-12-12 17:21:11.743643: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 37.70018768310547, KL Loss: 6.261404514312744, Reconstruction
Loss: 34.54112243652344
BETA is:  0.05
Epoch 3/25

2024-12-12 17:21:20.409496: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 34.51367950439453, KL Loss: 13.664953231811523, Reconstruction Loss:
31.250965118408203

2024-12-12 17:21:20.986430: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 32.2878532409668, KL Loss: 15.250144004821777, Reconstruction
Loss: 29.549442291259766
BETA is:  0.05
Epoch 4/25

2024-12-12 17:21:29.239712: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Train Loss: 31.502199172973633, KL Loss: 15.679722785949707, Reconstruction
Loss: 29.572856903076172

2024-12-12 17:21:29.762721: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.959915161132812, KL Loss: 16.123620986938477, Reconstruction
Loss: 27.809680938720703
BETA is:  0.05
Epoch 5/25

2024-12-12 17:21:38.111342: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.889602661132812, KL Loss: 22.45930290222168, Reconstruction Loss:
29.952455520629883

2024-12-12 17:21:38.646042: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 31.159833908081055, KL Loss: 23.79697608947754, Reconstruction
Loss: 27.594343185424805
BETA is:  0.05
Epoch 6/25

2024-12-12 17:21:46.881732: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.447965621948242, KL Loss: 15.691713333129883, Reconstruction
Loss: 29.284475326538086

2024-12-12 17:21:47.390822: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.32613754272461, KL Loss: 16.700641632080078, Reconstruction
Loss: 27.07956886291504
BETA is:  0.05
Epoch 7/25

2024-12-12 17:21:55.525929: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.181785583496094, KL Loss: 16.22941017150879, Reconstruction Loss:
28.883907318115234

2024-12-12 17:21:56.042309: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.266613006591797, KL Loss: 18.08345603942871, Reconstruction
Loss: 27.33583641052246
BETA is:  0.05
Epoch 8/25

2024-12-12 17:22:04.189308: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Train Loss: 30.127744674682617, KL Loss: 20.35752296447754, Reconstruction Loss:
28.729433059692383

2024-12-12 17:22:04.692006: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.800703048706055, KL Loss: 25.1715145111084, Reconstruction
Loss: 26.672801971435547
BETA is:  0.05
Epoch 9/25

2024-12-12 17:22:12.987030: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.214324951171875, KL Loss: 18.05396842956543, Reconstruction Loss:
28.242427825927734

2024-12-12 17:22:13.494476: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.074146270751953, KL Loss: 21.178232192993164, Reconstruction
Loss: 26.774690628051758
BETA is:  0.05
Epoch 10/25

2024-12-12 17:22:21.702547: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.747447967529297, KL Loss: 17.097835540771484, Reconstruction
Loss: 28.723173141479492

2024-12-12 17:22:22.195476: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.811180114746094, KL Loss: 18.44583511352539, Reconstruction
Loss: 26.424243927001953
BETA is:  0.05
Epoch 11/25

2024-12-12 17:22:30.591361: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 31.037004470825195, KL Loss: 12.73255443572998, Reconstruction Loss:
29.01826286315918

2024-12-12 17:22:31.091786: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.877723693847656, KL Loss: 14.323649406433105, Reconstruction
Loss: 27.287395477294922
BETA is:  0.1
Epoch 12/25

2024-12-12 17:22:39.209865: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Train Loss: 30.499980926513672, KL Loss: 10.85240364074707, Reconstruction Loss:
28.324087142944336

2024-12-12 17:22:39.745998: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.541812896728516, KL Loss: 11.892852783203125, Reconstruction
Loss: 26.887807846069336
BETA is:  0.11
Epoch 13/25

2024-12-12 17:22:48.002980: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.403913497924805, KL Loss: 10.170090675354004, Reconstruction
Loss: 28.70667266845703

2024-12-12 17:22:48.517025: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.583871841430664, KL Loss: 11.136874198913574, Reconstruction
Loss: 27.300884246826172
BETA is:  0.12
Epoch 14/25

2024-12-12 17:22:56.771999: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.579753875732422, KL Loss: 13.000872611999512, Reconstruction
Loss: 28.078094482421875

2024-12-12 17:22:57.325739: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.881484985351562, KL Loss: 15.099629402160645, Reconstruction
Loss: 26.69998550415039
BETA is:  0.13
Epoch 15/25

2024-12-12 17:23:05.608689: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.937511444091797, KL Loss: 8.691962242126465, Reconstruction Loss:
28.908405303955078

2024-12-12 17:23:06.118145: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.21919059753418, KL Loss: 8.964917182922363, Reconstruction
Loss: 26.512496948242188
BETA is:  0.14
Epoch 16/25

2024-12-12 17:23:14.447687: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Train Loss: 30.1827335357666, KL Loss: 7.548756122589111, Reconstruction Loss:
28.51205062866211

2024-12-12 17:23:14.970779: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.00826072692871, KL Loss: 8.155345916748047, Reconstruction
Loss: 26.154333114624023
BETA is:  0.15
Epoch 17/25

2024-12-12 17:23:23.180159: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.075735092163086, KL Loss: 8.922811508178711, Reconstruction Loss:
28.410953521728516

2024-12-12 17:23:23.701358: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.165538787841797, KL Loss: 9.908827781677246, Reconstruction
Loss: 26.140899658203125
BETA is:  0.16
Epoch 18/25

2024-12-12 17:23:31.976399: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 31.197227478027344, KL Loss: 8.02535343170166, Reconstruction Loss:
28.406192779541016

2024-12-12 17:23:32.509688: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.137964248657227, KL Loss: 8.88626480102539, Reconstruction
Loss: 26.36162567138672
BETA is:  0.17
Epoch 19/25

2024-12-12 17:23:40.842922: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.23228645324707, KL Loss: 6.64791202545166, Reconstruction Loss:
27.832008361816406

2024-12-12 17:23:41.364096: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.204572677612305, KL Loss: 6.791245937347412, Reconstruction
Loss: 26.046464920043945
BETA is:  0.18
Epoch 20/25

2024-12-12 17:23:49.589155: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Train Loss: 29.780221939086914, KL Loss: 5.496977806091309, Reconstruction Loss:
28.1712703704834

2024-12-12 17:23:50.107682: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.085216522216797, KL Loss: 6.130113124847412, Reconstruction
Loss: 26.218666076660156
BETA is:  0.19
Epoch 21/25

2024-12-12 17:23:58.334305: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.237884521484375, KL Loss: 5.381424427032471, Reconstruction Loss:
27.847518920898438

2024-12-12 17:23:58.846479: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.62075424194336, KL Loss: 5.987636089324951, Reconstruction
Loss: 26.267263412475586
BETA is:  0.2
Epoch 22/25

2024-12-12 17:24:07.609413: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 29.630918502807617, KL Loss: 5.401956081390381, Reconstruction Loss:
27.718236923217773

2024-12-12 17:24:08.119300: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 29.647912979125977, KL Loss: 5.8228912353515625, Reconstruction
Loss: 26.322053909301758
BETA is:  0.21
Epoch 23/25

2024-12-12 17:24:16.412561: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Train Loss: 30.10558319091797, KL Loss: 6.751118183135986, Reconstruction Loss:
27.911895751953125

2024-12-12 17:24:16.938123: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence

Validation Loss: 30.011207580566406, KL Loss: 8.217364311218262, Reconstruction
Loss: 25.707765579223633
BETA is:  0.22
Epoch 24/25

2024-12-12 17:24:25.177309: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Train Loss: 29.646812438964844, KL Loss: 6.321861743927002, Reconstruction Loss:
27.694355010986328
```

```
2024-12-12 17:24:25.684668: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Validation Loss: 30.11176872253418, KL Loss: 7.073030948638916, Reconstruction
Loss: 25.828048706054688
BETA is:  0.23
Epoch 25/25
```

```
2024-12-12 17:24:34.059265: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

```
Train Loss: 29.490495681762695, KL Loss: 4.3524909019470215, Reconstruction
Loss: 27.740766525268555
Validation Loss: 29.537527084350586, KL Loss: 4.7590861320495605, Reconstruction
Loss: 26.03545570373535
BETA is:  0.24
```

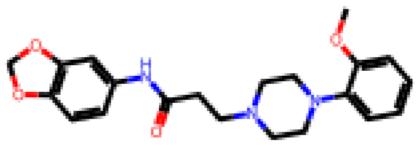
```
2024-12-12 17:24:34.576585: W tensorflow/core/framework/local_rendezvous.cc:404]
Local rendezvous is aborting with status: OUT_OF_RANGE: End of sequence
```

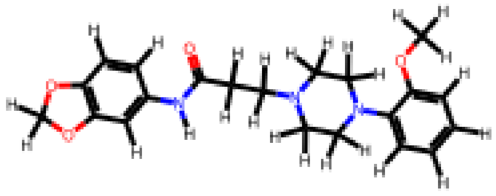
[67]:
```
'''
Checking the model's ability to reconstruct a molecule from the training dataset
'''

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)
```

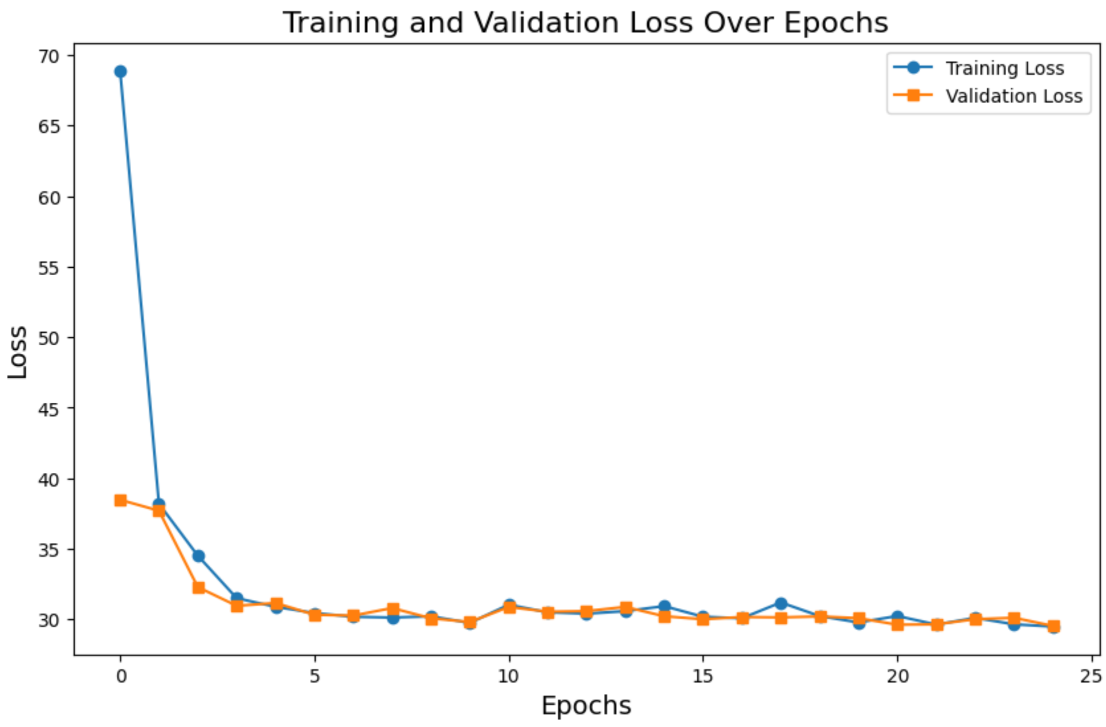
[68]:
```
mole_pred = graph_to_molecule(adj0[0], feature0[0])
Draw.MolsToGridImage([molobj,mole_pred], molsPerRow=2,)
```

[68]:





[69]:
```
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

[70]:
```
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_)
```

[71]:
```
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 1ms/step
79/79              0s 1ms/step
```

[72]:
```
ls_train_ = np.array(ls_train)
ls_test_ = np.array(ls_test)
```

[73]:
```
z_mean, _ = vae.encoder.predict([adj_test, fea_test, score_test_])
```

```
79/79              0s 848us/step
```

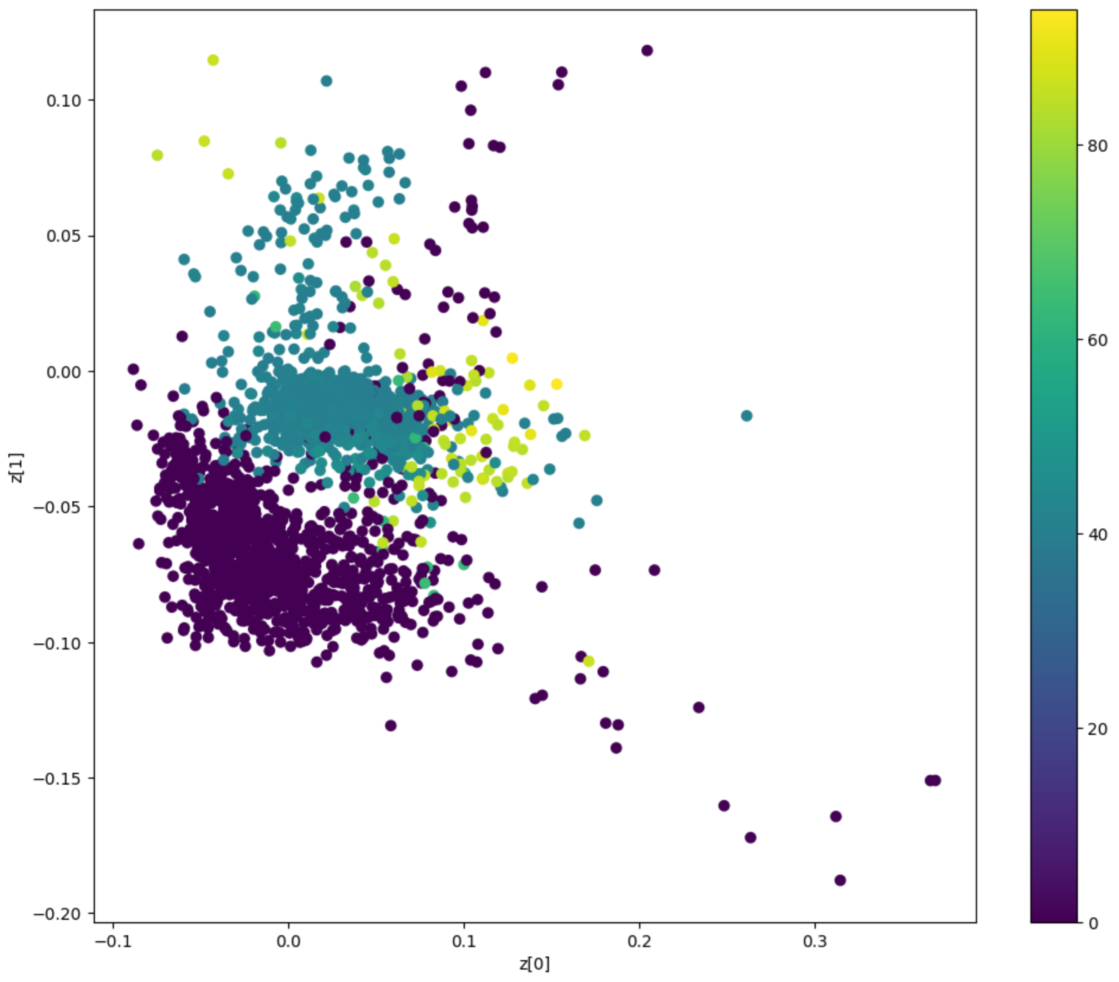
[74]:
```
latent_noise = np.random.normal(scale=0.1, size=z_mean.shape)  # Adjust scale␣
 ↪as needed
adj_pred, feature_pred = vae.decoder.predict([z_mean, score_test_])
print("Shape of adj_pred:", adj_pred.shape)
print("Shape of feature_pred:", feature_pred.shape)

# 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)
```

[75]:
```
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: [0.43693076863524927,
0.6458182163928556, 0.5315790713378814, -0.041525544983638946,
-0.5453206234128087, 0.27651514797541354, -0.2025655396442772,
0.9075695019942163, 0.8320360810000553, 0.8989103496578915, -0.6432961649580947,
0.7306387961762596, -0.6180187404340671, -0.4886460623848657,
0.9240378844815753, -0.8611331757415476, 0.8382176854159732,
-0.9400239344371479, 0.2778360892107996, 0.8678939592866706, 0.9053311618017323,
-0.5292029900265793, 0.5341352412189297, -0.7170257516197571, 0.724299010486355,
0.6104088284937204, 0.7626398025691812, -0.23543282208908284,
0.8263745011386103, 0.7882786743779907, -0.9248352120038561, 0.5131734637761092,
-0.5923120852073069, 0.9281137651821356, -0.7270253492269785,
-0.8115047826774604, 0.8966585457319981, 0.7754990730081102,
0.05203248343829952, -0.801432094460618, -0.841706317665667, 0.5824081634941394,
-0.5631339729512872, -0.4567626925162528, 0.6673706299782768,
-0.6391112272081758, 0.6017029545359756, -0.3589247852523813,
-0.7769279774032881, -0.7778355982029415, 0.7731216968016852,
-0.5565482601040326, -0.5971709660433001, 0.6279345112770491,
0.18204615128450485, -0.8903125690972353, -0.5153799743279579,
-0.8404132605720162, 0.6257658022677505, -0.44375887906770617,
-0.8201011547974539, 0.7668748110379051, 0.34513410442651227,
0.3512017948973034, -0.8970692737574888, 0.48971532021328995,
-0.6703689507171706, -0.3200898441108734, -0.8089706423083505,
-0.09425211774637757, 0.9234053376369638, -0.03932443932429845,
0.19419645822029255, -0.6712865076747813, 0.297871903979944,
-0.8719070556696706, 0.6834532333937009, -0.5548354570672189,
0.4378141119970578, 0.827797895689106, -0.580488339194429, 0.6095240555631715,
-0.37537597890104324, 0.676186207189819, 0.6286921396494096,
-0.7621541185358156, -0.23408406537676432, 0.33821310820200645,
-0.34448938067626467, -0.8170898716359402, -0.03824877812255295,
0.38168973655569566, 0.3908850066126446, -0.8049908094060851,
0.5569112555758211, -0.8995223103172105, -0.09515742879935934,
-0.04281520658449054, 0.0006577860547918746, -0.1353361609716848,
0.03147870688847587, 0.41645086774227447, -0.9317143884679505,
0.11384852356943198, 0.06466981829877852, 0.44645502410171933,
-0.14887889526812498, 0.5560570648233627, -0.7135170917340019,
-0.46780204836926775, 0.7309778183625266, -0.033626947225412554,
-0.811633578183703, -0.5731403778395354, 0.6393540914719886, 0.890581534008148,
-0.7582324983221603, -0.6144058169778475, -0.020479379466469074,
0.7948437841161935, 0.23231148709417937, 0.8656580871044224,
0.44071265464958376, 0.59629377698668, -0.1931243064805609, -0.3300846815667597,
0.7478632133129859, -0.48627657227804677, 0.7751503532831596,
0.44962573486438273, 0.6406877677293241, -0.7094739806347756,
0.09303550280308892, -0.8840676342068382, -0.8306318087939335,
0.8675279351597966, 0.8259178372174545, -0.46124665064916615,
0.5960537239457002, 0.03585240569688615, -0.160713789169105, 0.7476964159666163,
0.46478570109350026, 0.030603157349616994, 0.8736984053912613,
-0.19405149802861943, 0.12930338970727084, -0.3235522240052505,
-0.9075570492259131, 0.7249772215256738, -0.7266750925991974,
0.6948620549344214, 0.8825842266588264, 0.48927638941365703, 0.7885316302409083,
0.1856054053865332, -0.05052425423703526, -0.4116923266866568,
0.9088438587518193, -0.6563605649272276, 0.8828110000492244, 0.7801212983239934,
0.9218310761910156, 0.7855056988000226, 0.9143422359421899,
-0.14803153678701486, -0.9287874427870663, -0.22722272107348868,
0.2709198293619329, -0.6592139194603186, -0.8847690956607112,
-0.1394970765122106, 0.8614096821723354, -0.28314541543776606,
-0.8262167487670504, -0.8944727865892675, 0.8229272616441996,
-0.042766310861424135, -0.28845019154589596, -0.6254742225326949,
-0.8676122821882234, 0.8726348359151328, -0.8275876981338667, 0.825533210758939,
-0.3223876769305887, -0.008551691735595345, 0.8384475981682669,
-0.6561141015510922, -0.12081726579184388, 0.9259706903652266,
-0.7210063840695797, 0.8958922980403142, -0.8110303070092791,
-0.8931674752410304, 0.7369990611467508, -0.9269416230569404,
-0.9252114623459915, 0.5713558909475865, 0.0412837530797445,
-0.5003196677603783, 0.8560252249851077, -0.6301211730371439,
0.9137487849777629, -0.8724627114940815, 0.9009506503902698,
0.27502535939720874, 0.39021583440532803, -0.7435920749906377,
0.8993128618444177, 0.8835019419037545, -0.8278547525482134,
0.28852806384884705, -0.5591899753824054, 0.9106140948720491,
-0.7749289804115912, 0.316678610813845, -0.7019898168875951, 0.9037186062262093,
0.835329387634747, -0.22639877303764291, 0.9173529244179112,
-0.05284659258836827, 0.26775381805756704, 0.45406283070331477,
-0.6787587922522544, -0.8722253434706944, -0.3223518293485731,
0.7346527638494009, 0.023691550681189985, 0.606806976695201, 0.5296379387828218,
0.8475955414919019, 0.26799271569295396, -0.5818634753481793,
-0.5883488729769852, -0.8387821153765493, 0.10040388694805213,
0.6484891054098649, 0.2817484259341585, 0.7142913582479601, 0.08876064050648189,
-0.8323202602304973, -0.2762461492278687, 0.6615788866015145,
-0.06561262210582236, 0.7282508734803184, 0.8390871058174897,
-0.4093936395877963, 0.9064527343803632, -0.016334784709051812,
0.6421914854072382, -0.6249345472769086, -0.2664077543381597,
0.17607913843430992, 0.4487125386128529, 0.882377626965308]
```

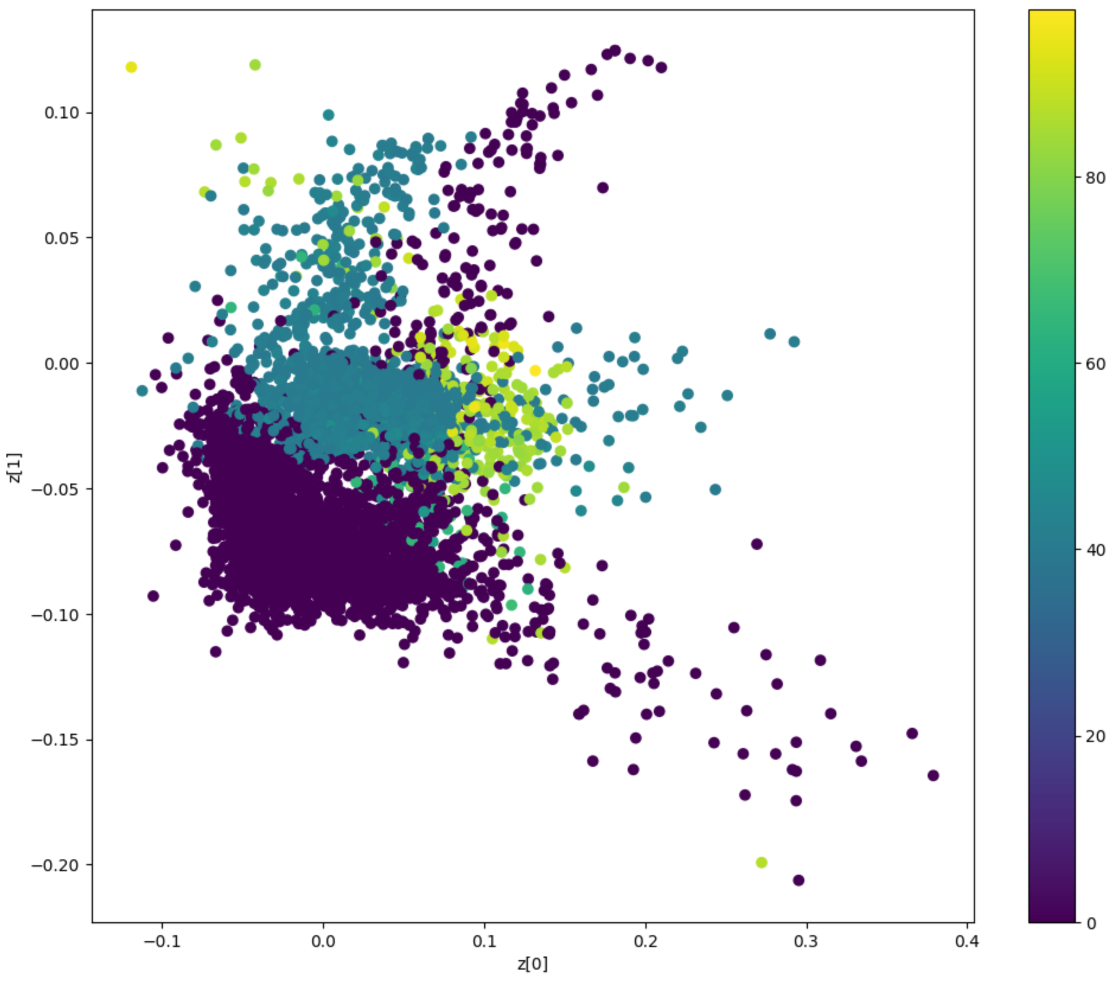
[76]:
```
plt.figure(figsize=(12, 10))
plt.scatter(z_mean[:, 0], z_mean[:, 1], c=score_test_)
plt.colorbar()
plt.xlabel("z[0]")
plt.ylabel("z[1]")
plt.show()
```



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

```
251/251                 0s 853us/step
```

```
[78]: 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

```
[79]: 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)
    ]
```

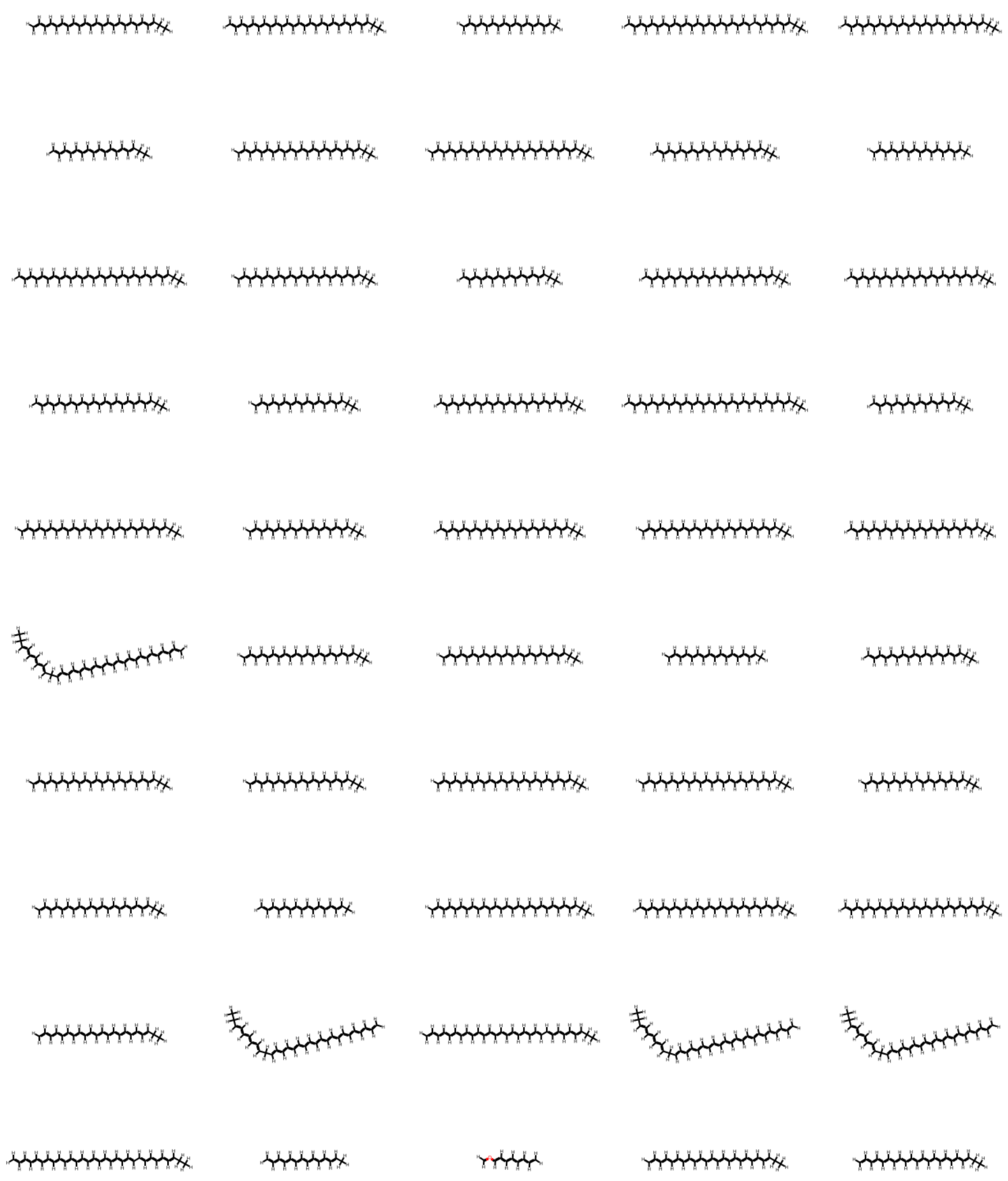
```
[80]: gen_mols = inference(batch_size=1000,activity=10)
MolsToGridImage([m for m in gen_mols if m is not None][:1000], molsPerRow=5,␣
↪subImgSize=(260, 160))
```

```
32/32                0s 3ms/step
Sanitization failed: Explicit valence for atom # 23 C, 7, is greater than
permitted
Sanitization failed: Explicit valence for atom # 6 Cl, 3, is greater than
permitted
Sanitization failed: Explicit valence for atom # 0 I, 84, is greater than
permitted
Sanitization failed: Explicit valence for atom # 0 P, 45, is greater than
permitted
Sanitization failed: Explicit valence for atom # 4 Cl, 5, is greater than
permitted
Sanitization failed: Explicit valence for atom # 21 C, 7, is greater than
permitted
Sanitization failed: Explicit valence for atom # 13 N, 6, is greater than
permitted
Sanitization failed: Explicit valence for atom # 1 O, 5, is greater than
permitted
Sanitization failed: Explicit valence for atom # 0 As, 102, is greater than
permitted
Sanitization failed: Explicit valence for atom # 25 C, 6, is greater than
permitted
Sanitization failed: Explicit valence for atom # 0 C, 7, is greater than
permitted
Sanitization failed: Explicit valence for atom # 0 I, 42, is greater than
permitted
Sanitization failed: Explicit valence for atom # 0 I, 91, 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(
```

[80]:



[ ]: