AI-Accelerated Discovery of Novel Aqueous Amines for CO_2 Capture

Team Aspirin

Charlotte Breakwell

Luis Carvalho

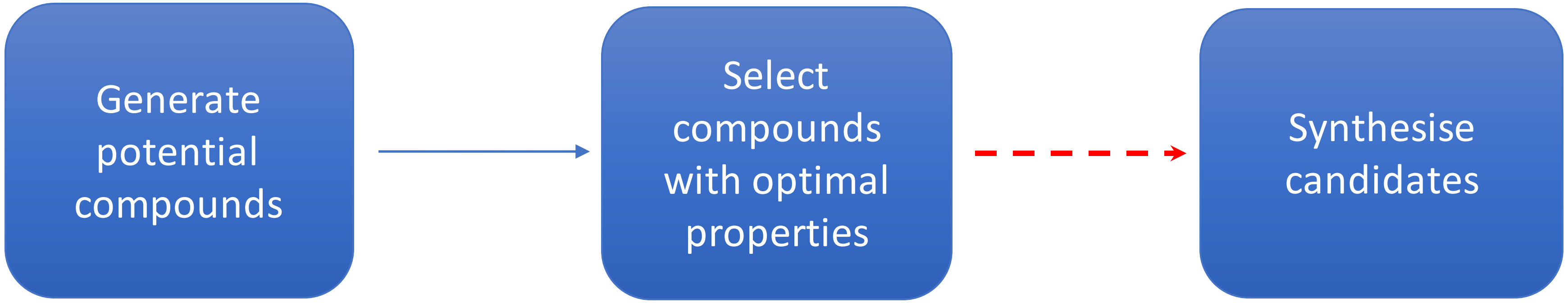
Kieran Nehil-Puleo

Evangelos Tsochantaris

Eco-AI Hackathon

March 2024

AI-Aided Discovery



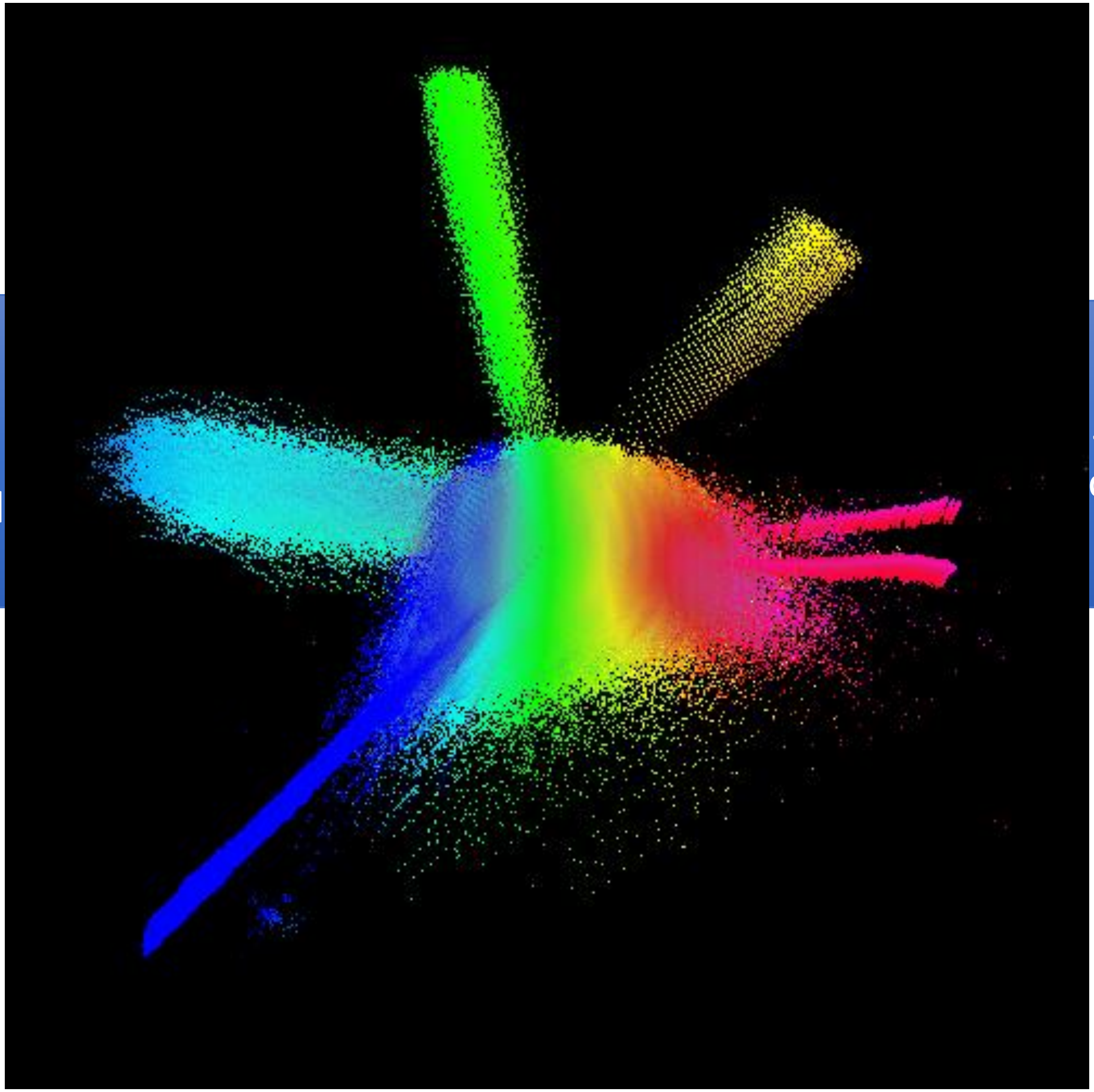
Generate potential compounds

Select compounds with optimal properties

Synthesise candidates

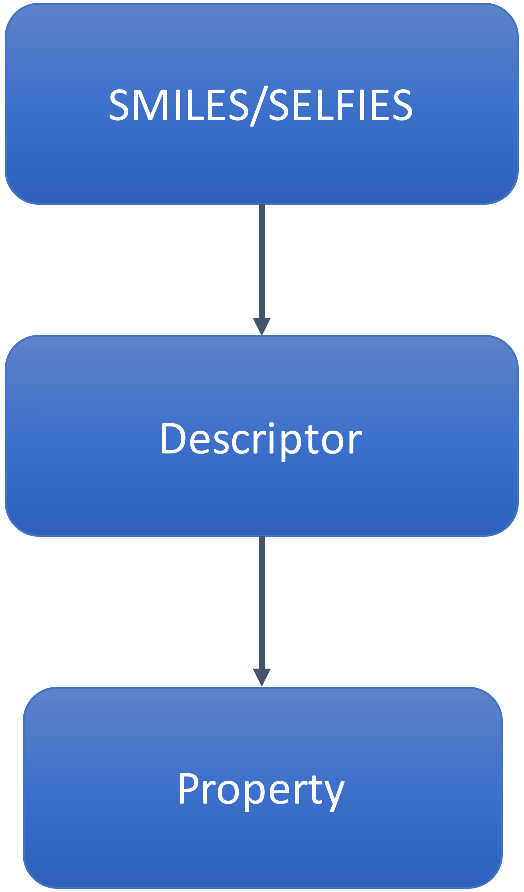
AI-Aided Discovery

Generate potential compound



Synthesise candidates

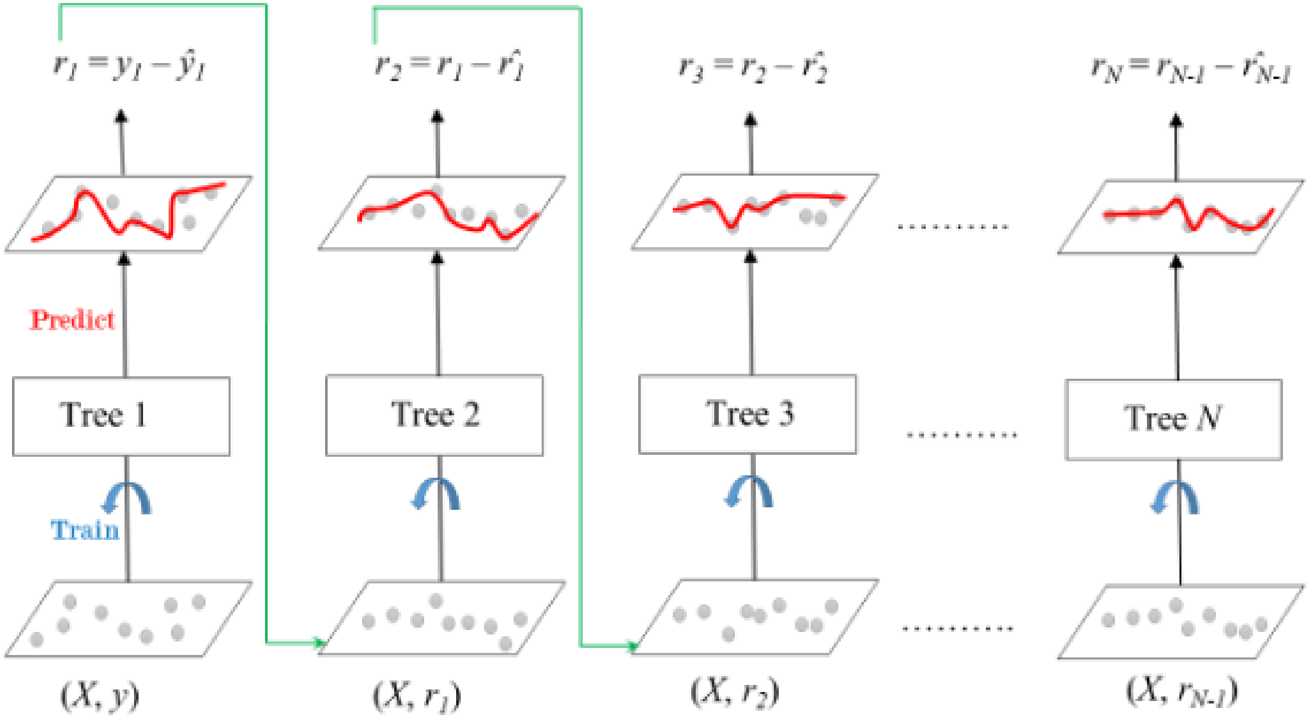
Proposed Workflow (Predictive Model)



SMILES/SELFIES

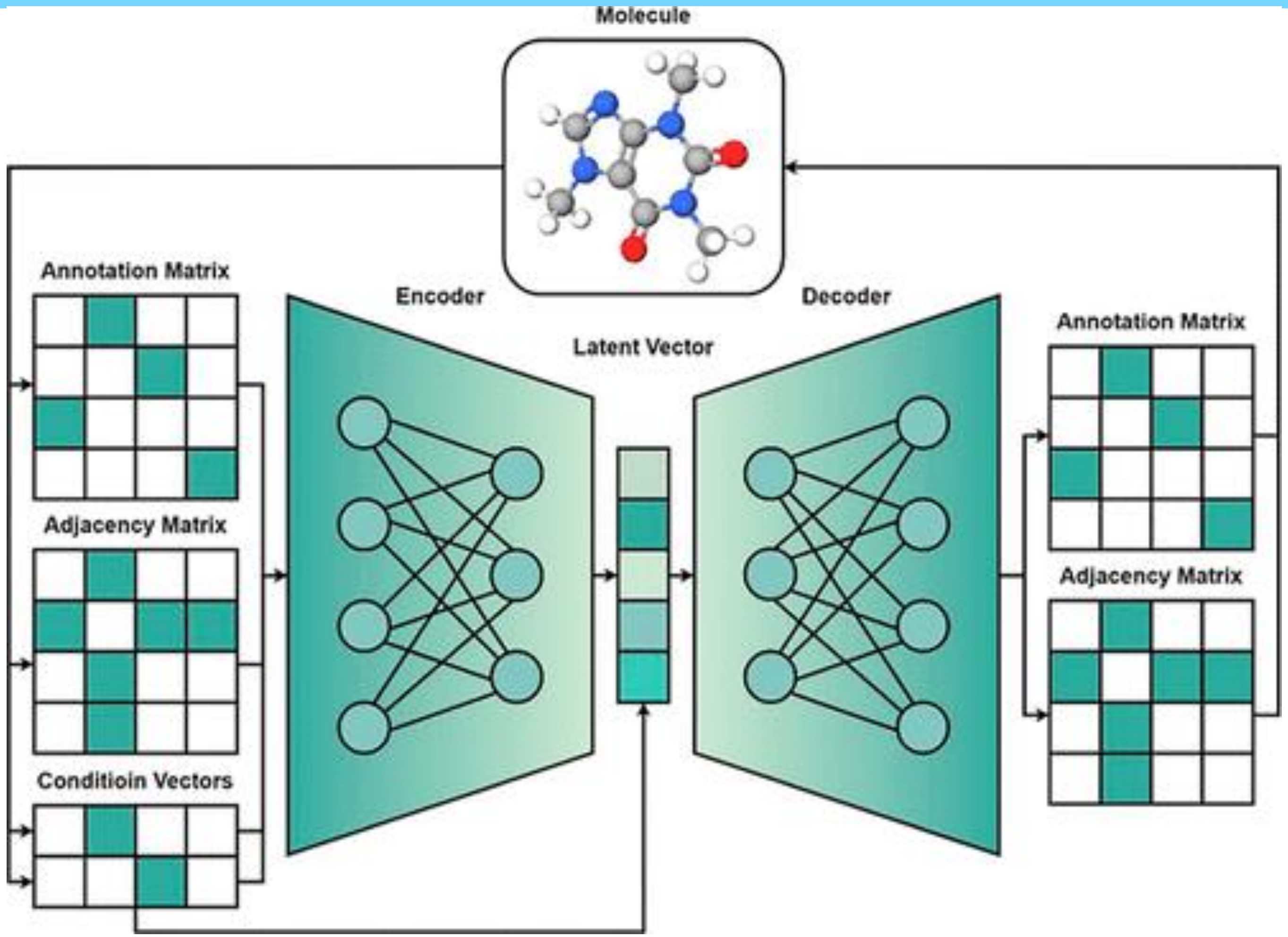
Descriptor

Property

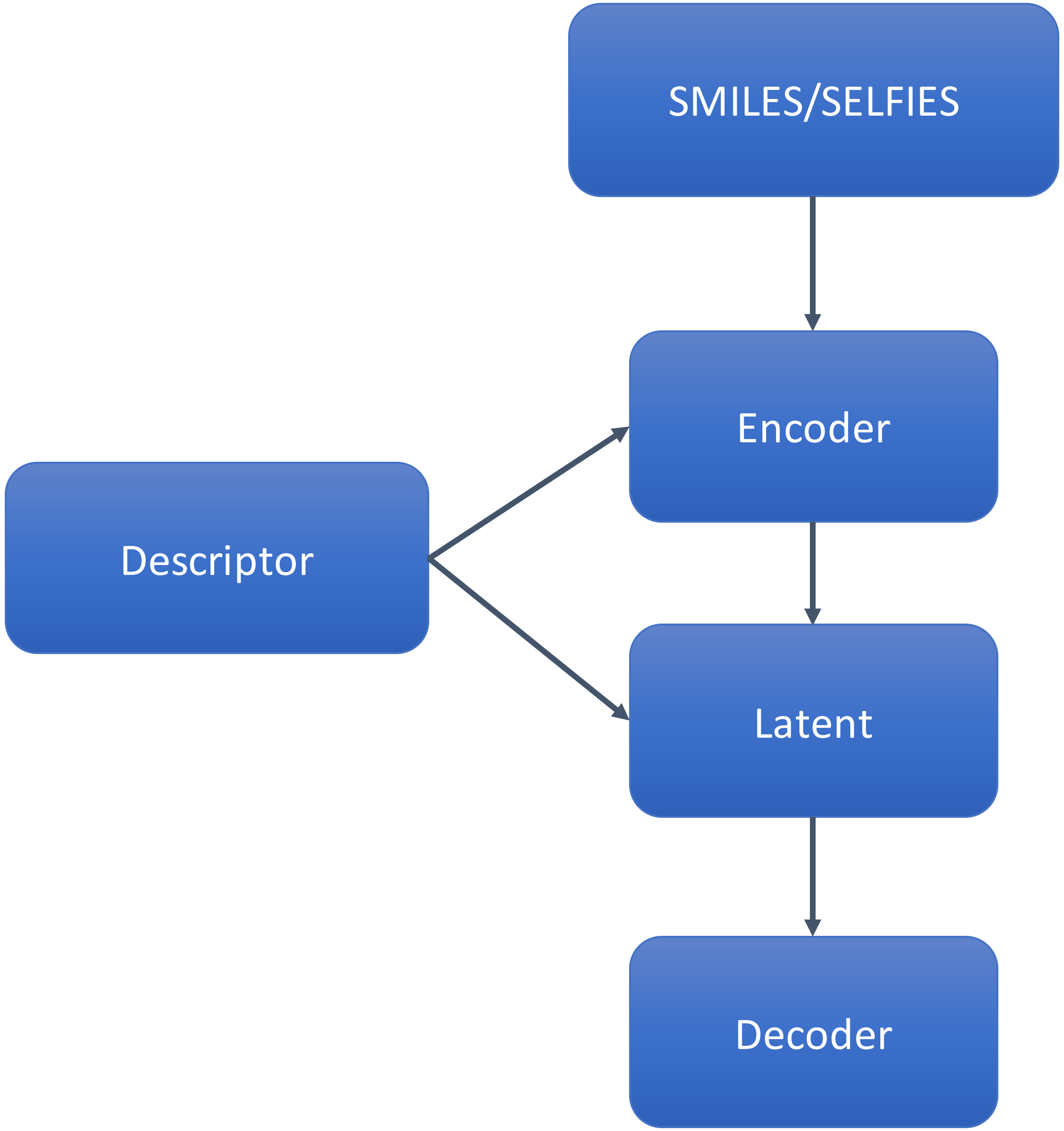


https://www.geeksforgeeks.org/ml-gradient-boosting/

Proposed Workflow (Conditional VAE)



Lee, Myeonghun, and Kyoungmin Min. "MGCVAE: multi-objective inverse design via molecular graph conditional variational autoencoder." *Journal of chemical information and modeling* 62.12 (2022): 2943-2950.



SMILES/SELFIES

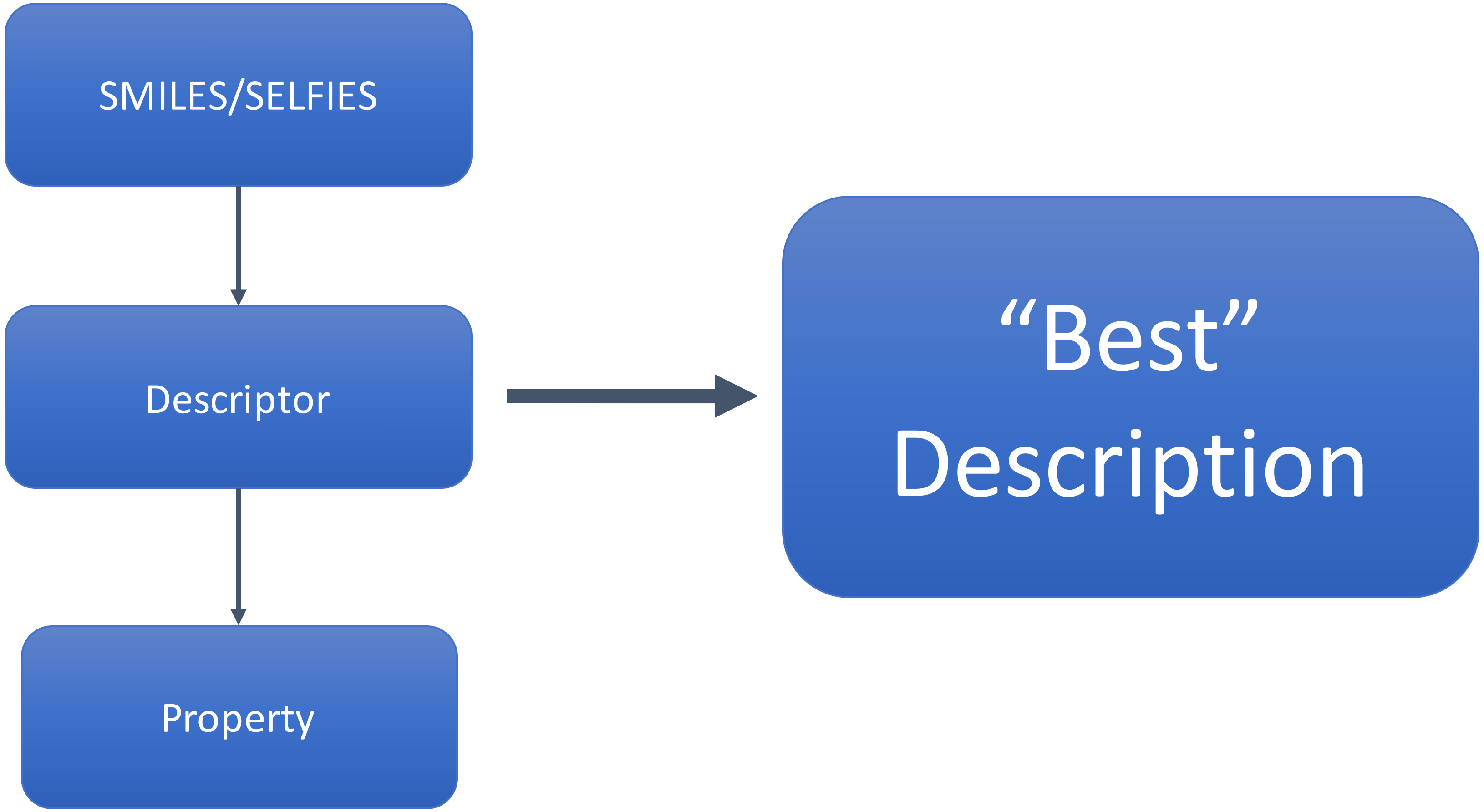
Descriptor

Encoder

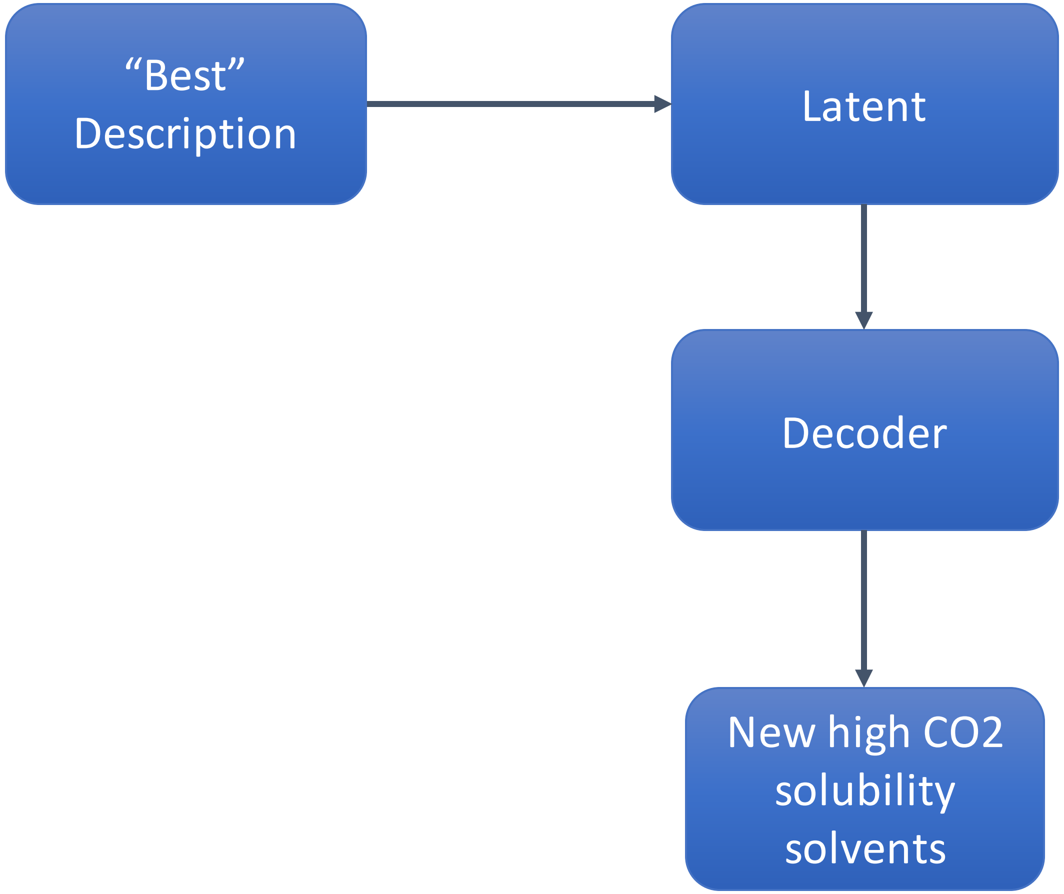
Latent

Decoder

Proposed Workflow



Proposed Workflow



Smiles vs Selfies

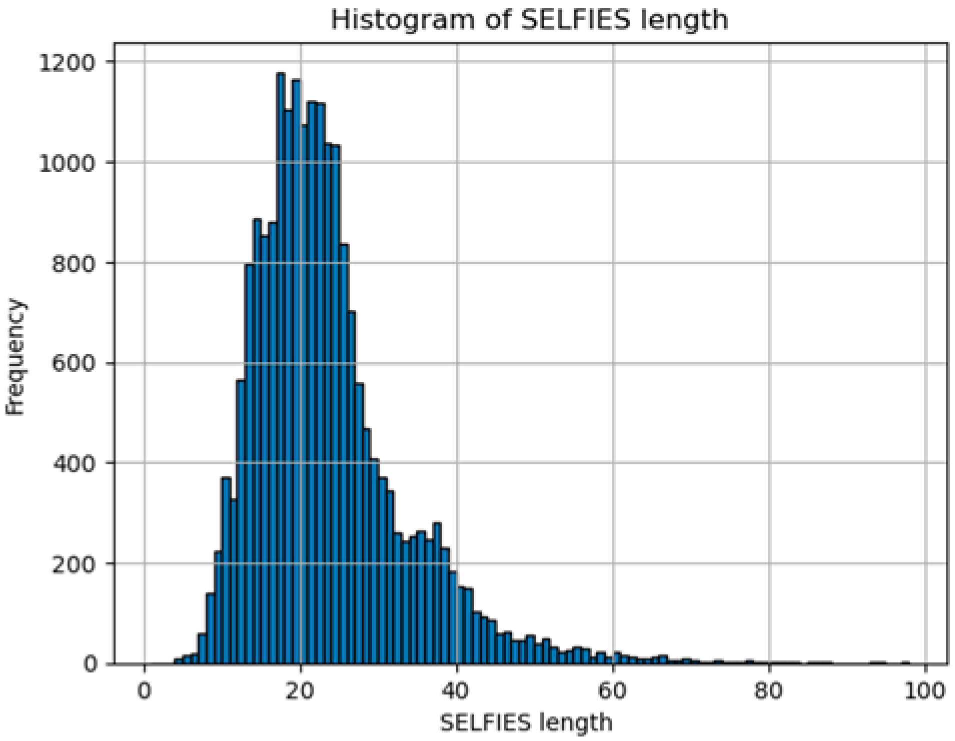
SMILES: CC(=O)Oc1ccccc1C(=O)O

SELFIES: [C][C](=[O])[O][C][c][c][c][c][c][c][C]([=[O])[O]

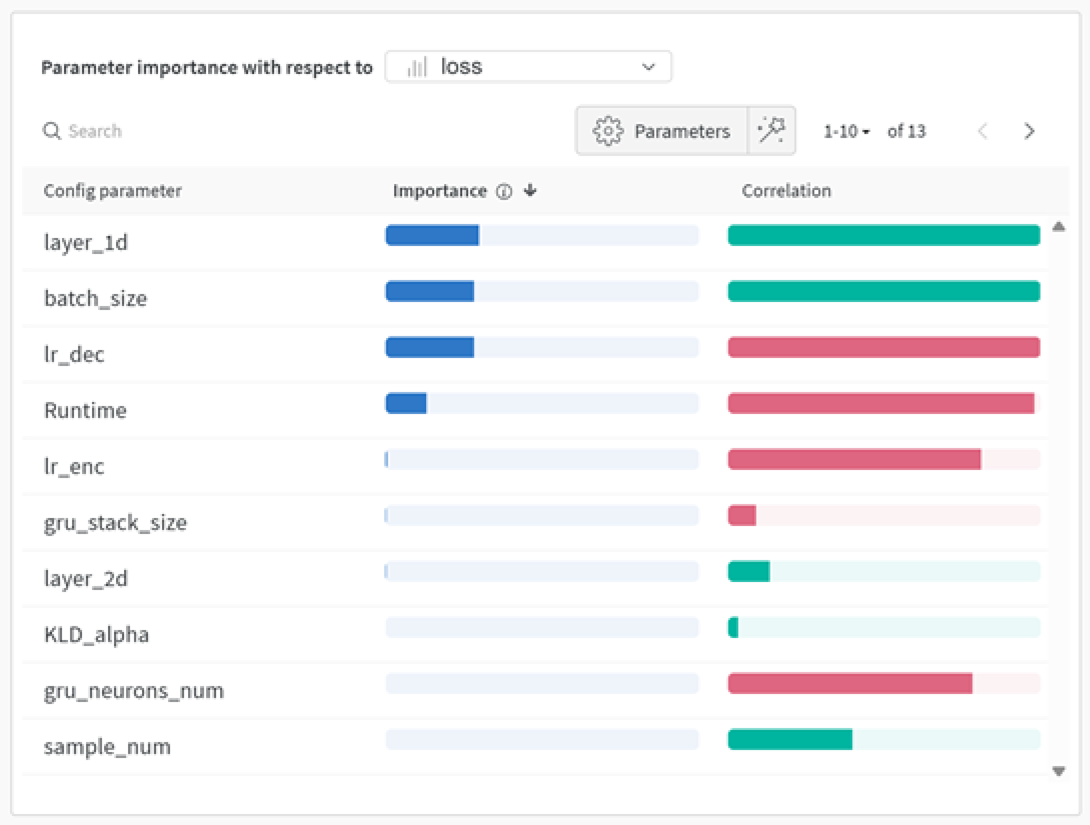
Encoder with

SMILES: [‘C’, ‘O’, ‘(‘, ‘=’, …]

SELFIES: [‘[C]’, ‘([=[O])’, ‘[O]’, …]



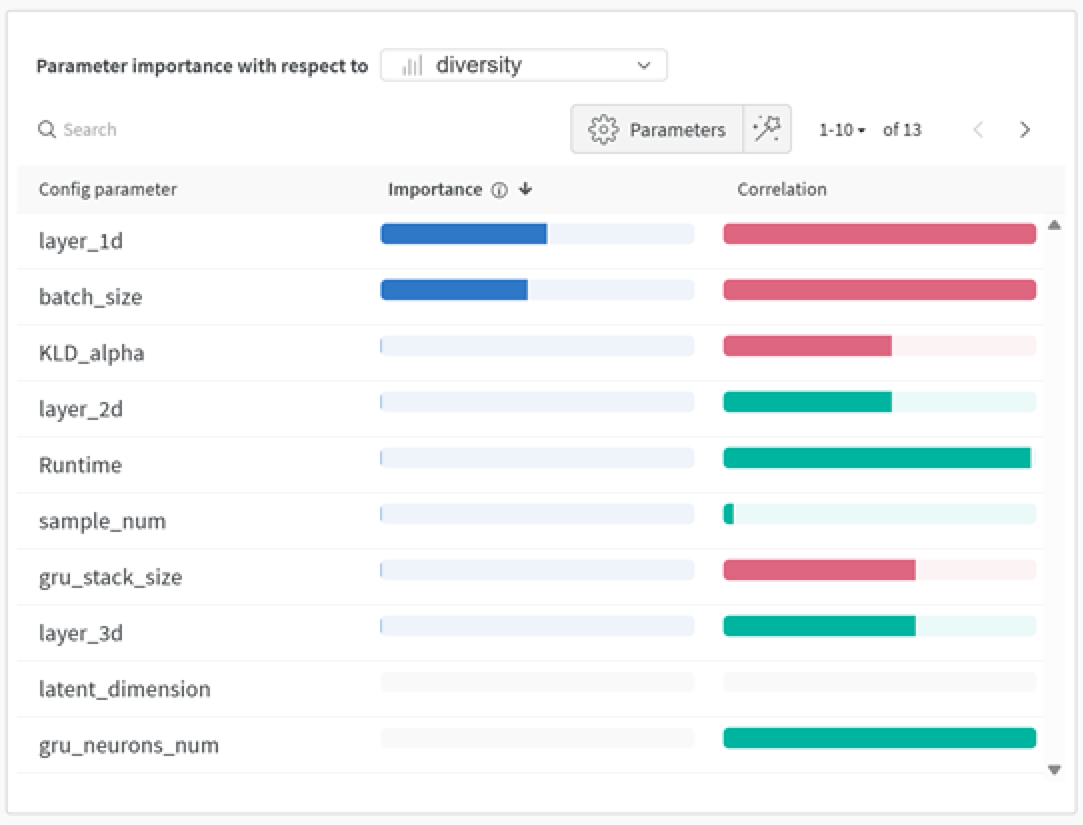
Hyperparameter Optimisation



Parameter importance with respect to loss

Search Parameters 1-10 of 13

Config parameter	Importance	Correlation
layer_1d	🟦	🟩
batch_size	🟦	🟩
lr_dec	🟦	🟥
Runtime	🟦	🟥
lr_enc		🟥
gru_stack_size		🟥
layer_2d		🟩
KLD_alpha		🟩
gru_neurons_num		🟥
sample_num		🟩

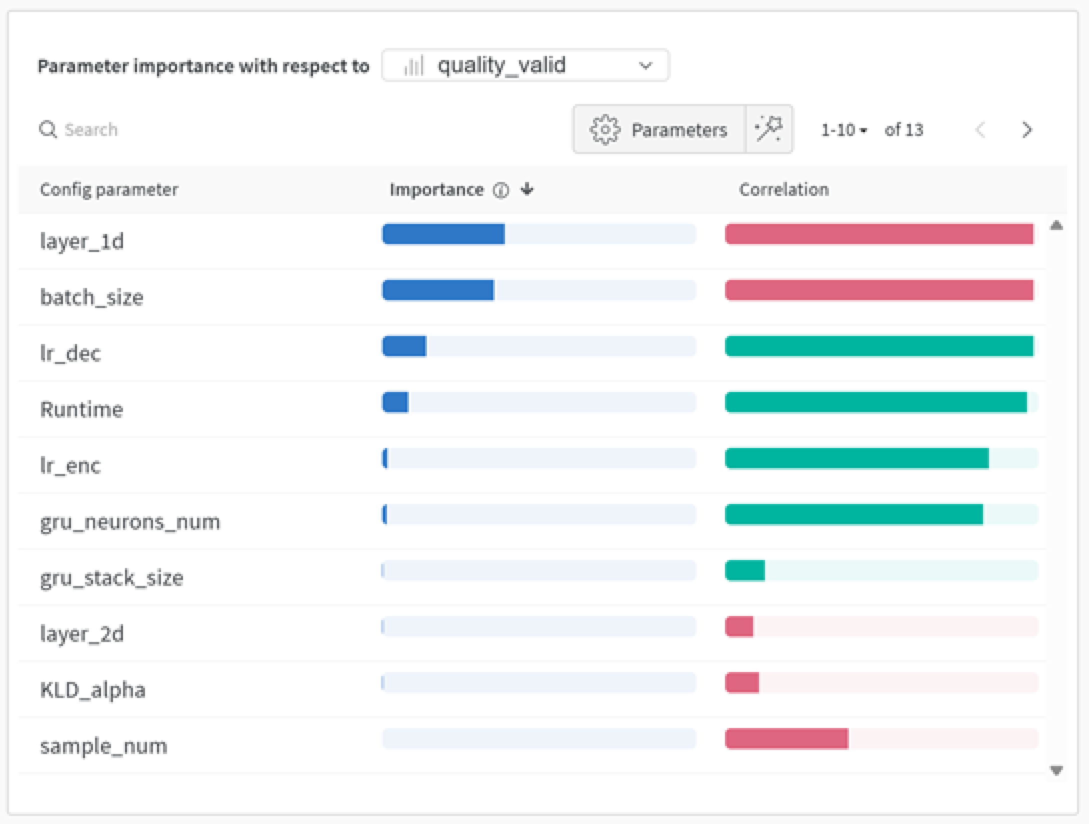


Parameter importance with respect to diversity

Search Parameters 1-10 of 13

Config parameter	Importance	Correlation
layer_1d	🟦	🟥
batch_size	🟦	🟥
KLD_alpha		🟥
layer_2d		🟩
Runtime		🟩
sample_num		🟩
gru_stack_size		🟥
layer_3d		🟩
latent_dimension		
gru_neurons_num		🟩

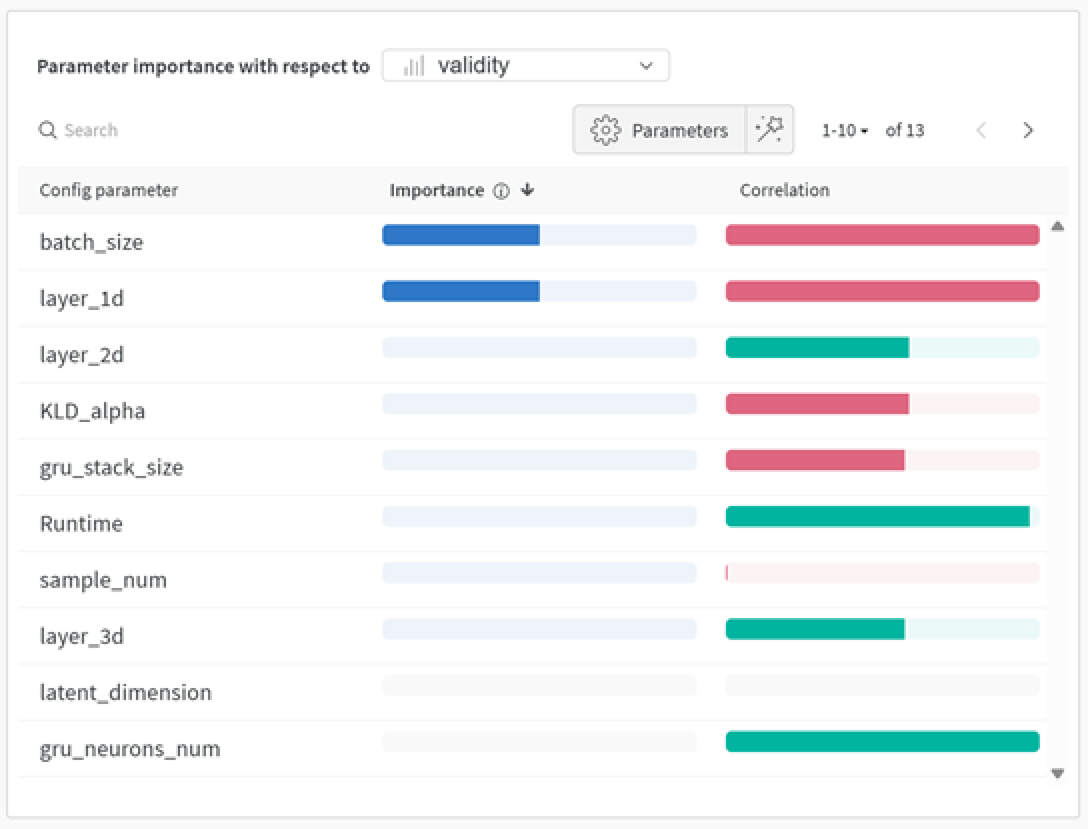
Hyperparameter Optimisation



Parameter importance with respect to quality_valid

Search

Config parameter	Importance	Correlation
layer_1d		
batch_size		
lr_dec		
Runtime		
lr_enc		
gru_neurons_num		
gru_stack_size		
layer_2d		
KLD_alpha		
sample_num		



Parameter importance with respect to validity

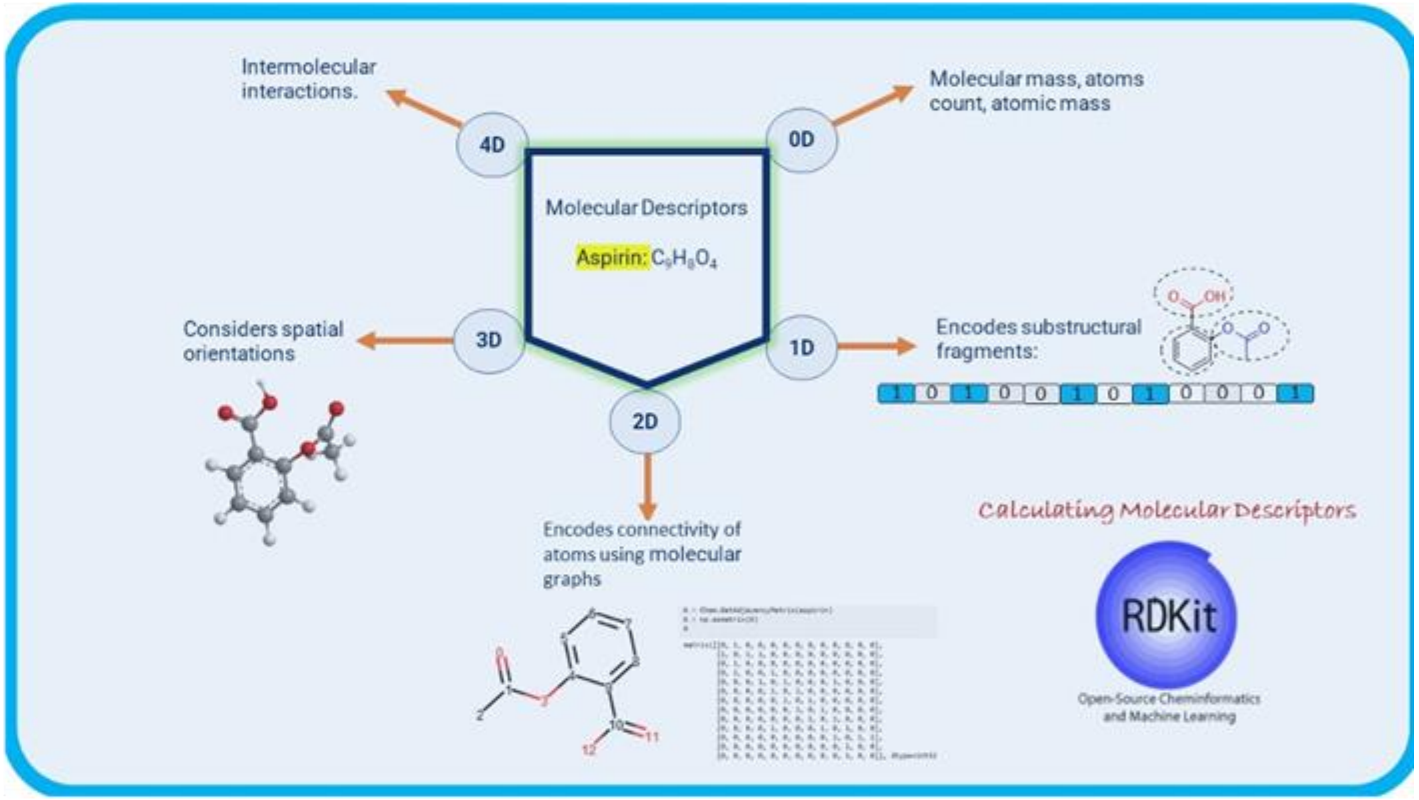
Search

Config parameter	Importance	Correlation
batch_size		
layer_1d		
layer_2d		
KLD_alpha		
gru_stack_size		
Runtime		
sample_num		
layer_3d		
latent_dimension		
gru_neurons_num		

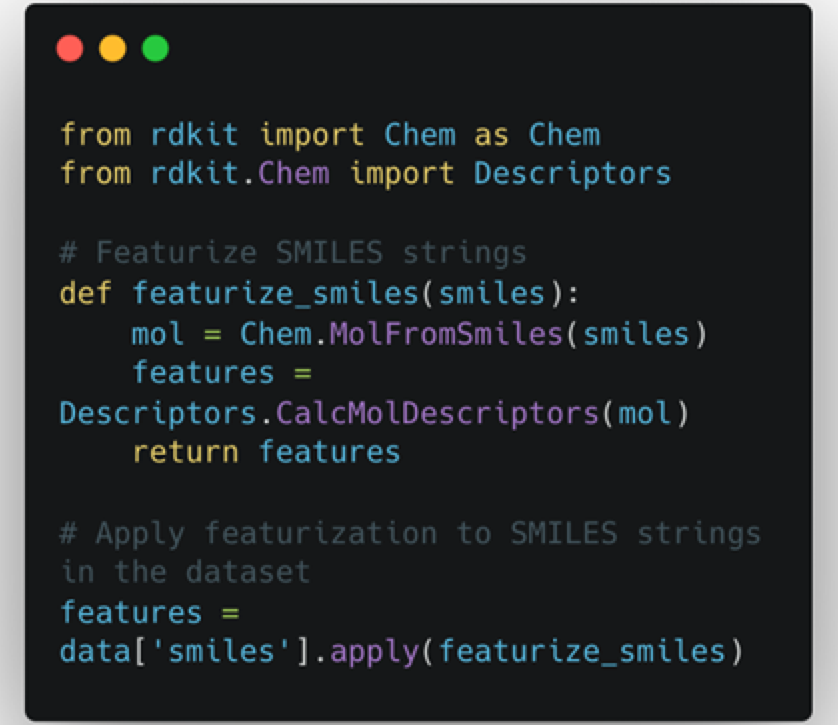
Finding the Right Molecule (Cheminformatics)

Cheminformatics depend on representations of molecules by descriptors that capture their structural characteristics and properties.

- Experimental measurements
- Theoretical measurements
  - RDKit descriptors
  - Simulations (e.g. DFT, MD)
  - Predictions from existing ML models


Intermolecular interactions.
4D
Molecular mass, atoms count, atomic mass
0D
Molecular Descriptors
Aspirin: $C_9H_8O_4$
Considers spatial orientations
3D
1D
Encodes substructural fragments:
O
OH
O
O
1 0 1 0 0 1 0 1 0 0 0 1
2D
Encodes connectivity of atoms using molecular graphs
Calculating Molecular Descriptors
RDKit
Open-Source Cheminformatics and Machine Learning

Finding the Right Molecule (Cheminformatics)



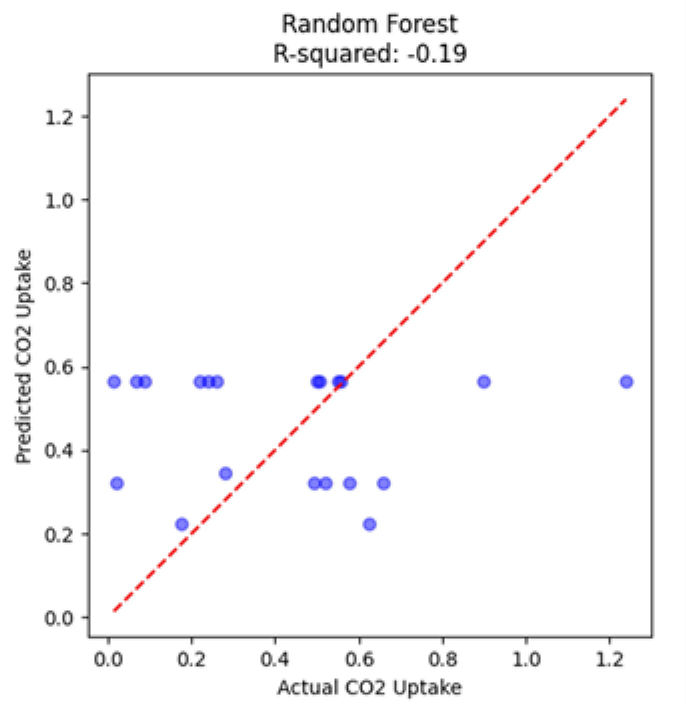
```
from rdkit import Chem as Chem
from rdkit.Chem import Descriptors

# Featurize SMILES strings
def featurize_smiles(smiles):
    mol = Chem.MolFromSmiles(smiles)
    features =
Descriptors.CalcMolDescriptors(mol)
    return features

# Apply featurization to SMILES strings
in the dataset
features =
data['smiles'].apply(featurize_smiles)
```

	Raw Dataset
1	SMILES
2	CO_2 adsorption capacity
3	n_nitrogen
4	Molecular Mass

	Dataset with RDKit Descriptors
1	SMILES
2	CO_2 adsorption capacity
3	n_nitrogen
4	Molecular Mass
5	NumValenceElectrons
6	MolLogP
	...
210	BertzCT
211	MaxAbsPartialCharge
212	MinAbsEStateIndex
213	BalabanJ
214	FpDensityMorgan1
215	VSA_EState7

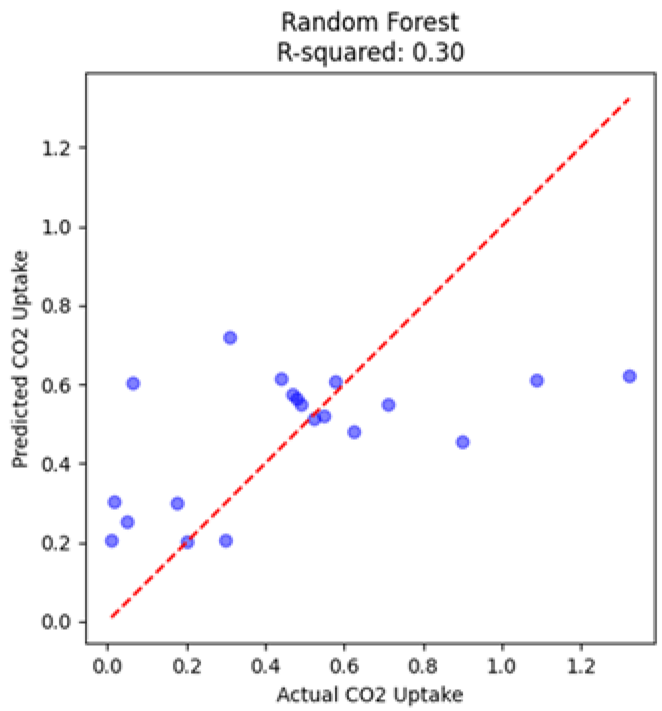
Finding the Right Molecule (Cheminformatics)

Raw Dataset



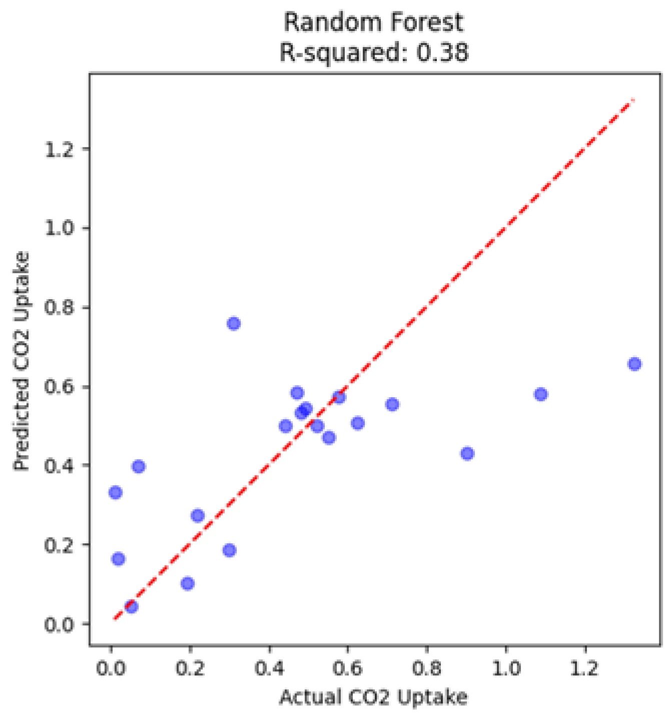
MSE = 0.297

+ RDKit Descriptors



MSE = 0.216

+ QM Descriptors



MSE = 0.191

* QM descriptors generated using xtb package by performing single point energy calculations in implicit water with PM6 functional (Semi-empirical DFT)

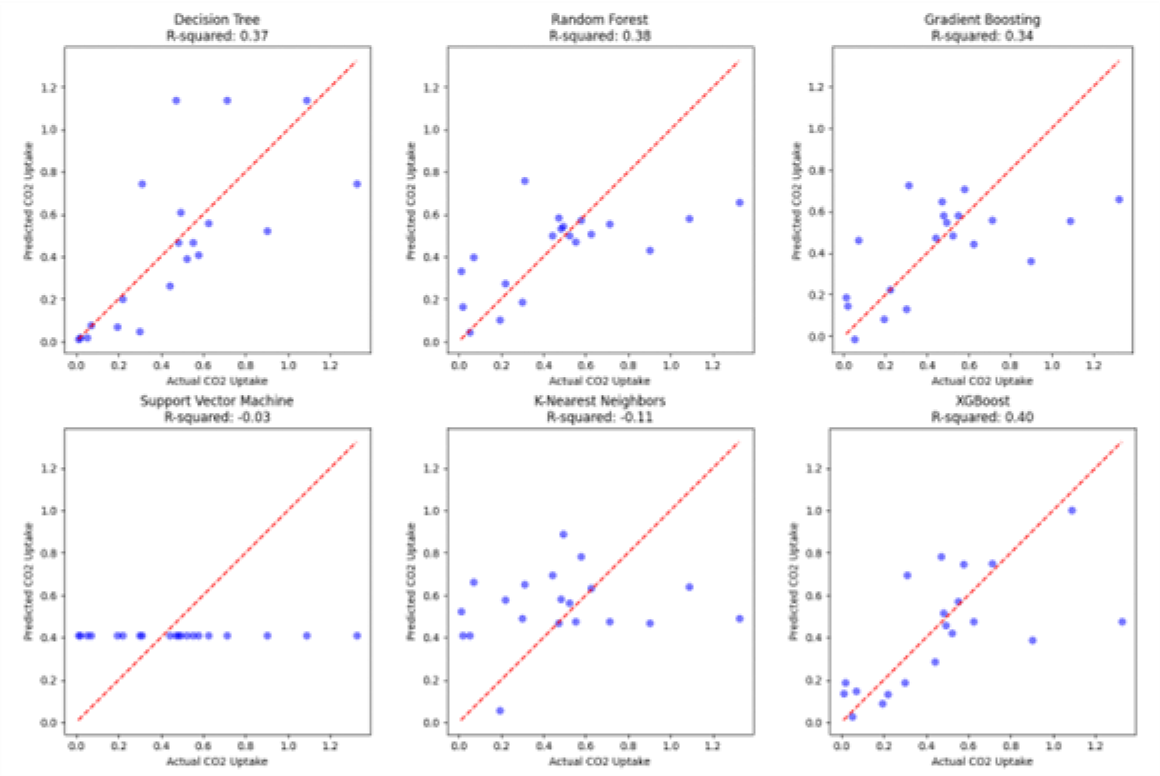
Finding the Right Molecule (Cheminformatics)

Feature selection: By identifying the most important features, we can select a subset of relevant features for our target property.

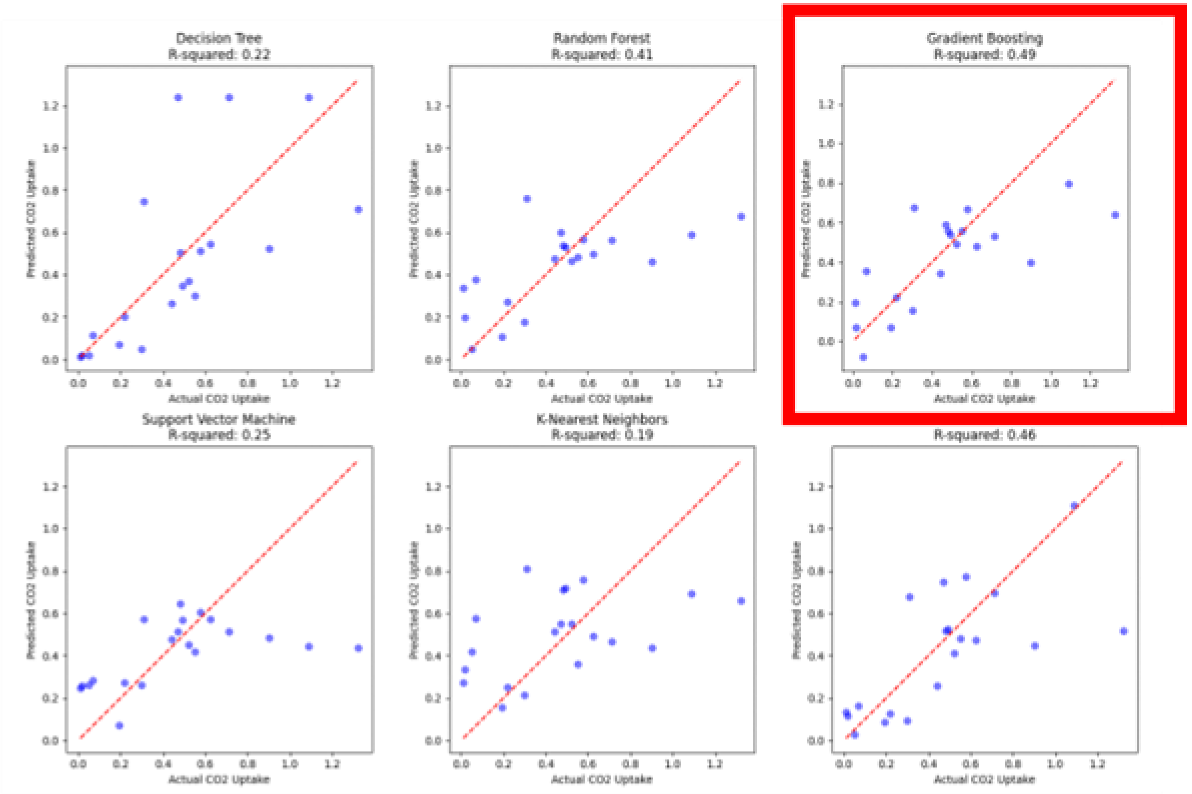
- Reduced dimensionality and noise in the data

***Top Performing Model = Gradient Boosted Regression**
R^2 = 0.49, MSE = 0.18 mol CO_2 / mol amine

All Descriptors

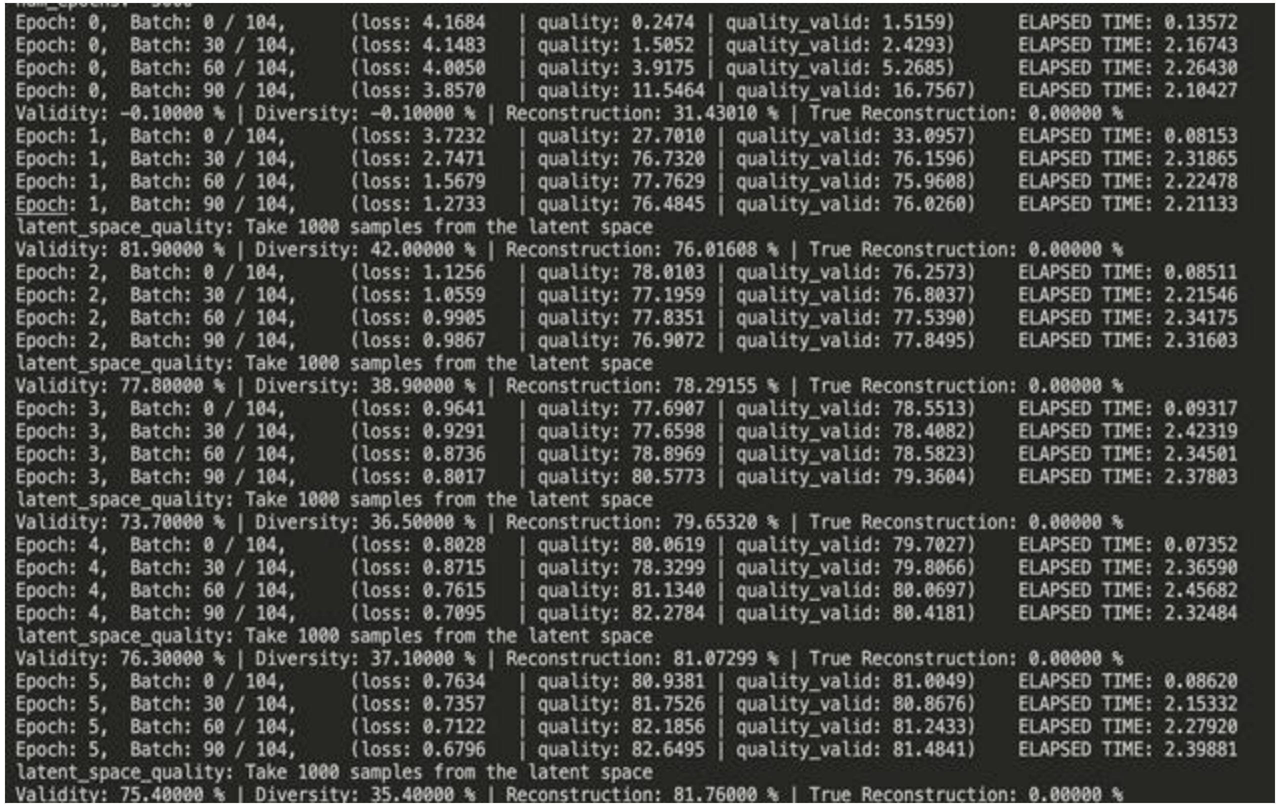


Top 50 Descriptors



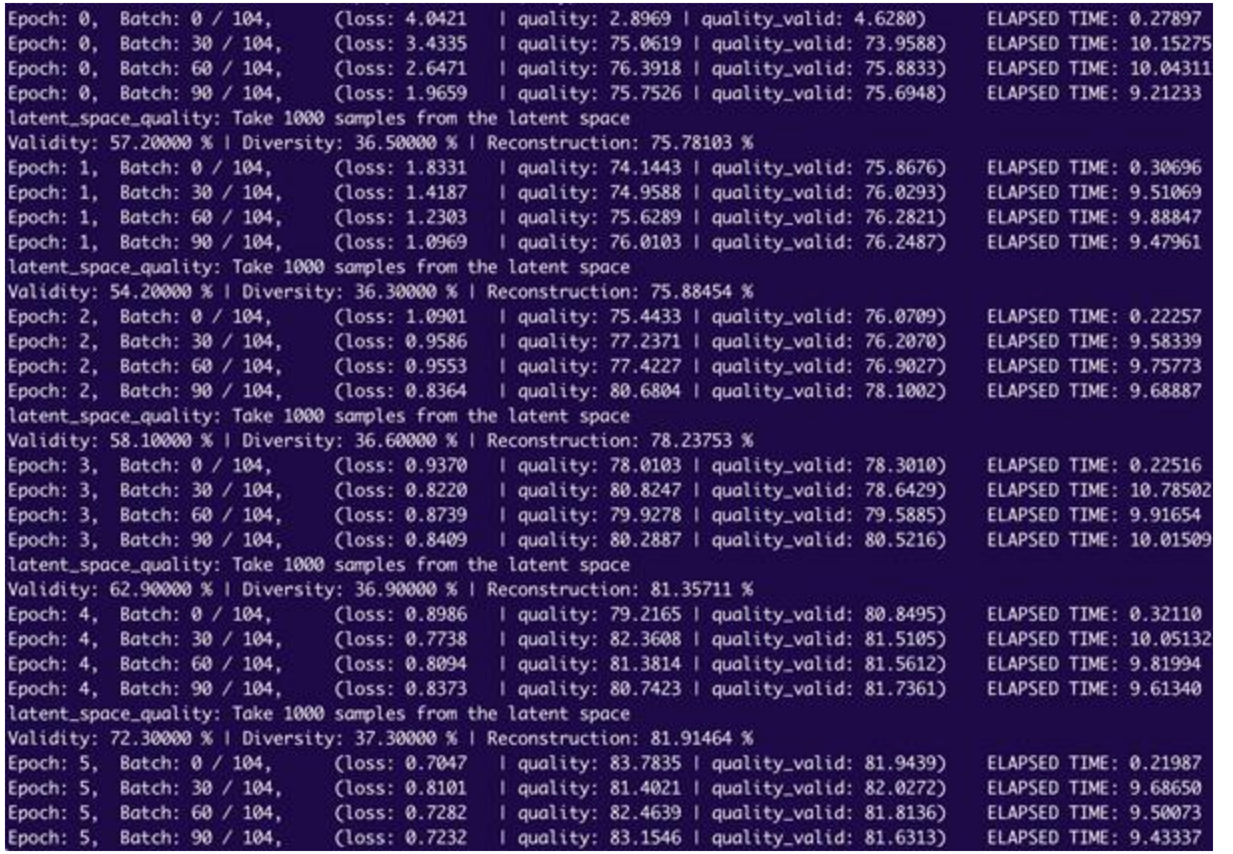
Finding the Right Molecule (Cheminformatics)

No Properties



```
Epoch: 0,  Batch: 0 / 104,      (loss: 4.1684   | quality: 0.2474 | quality_valid: 1.5159)     ELAPSED TIME: 0.13572
Epoch: 0,  Batch: 30 / 104,     (loss: 4.1483   | quality: 1.5052 | quality_valid: 2.4293)     ELAPSED TIME: 2.16743
Epoch: 0,  Batch: 60 / 104,     (loss: 4.0050   | quality: 3.9175 | quality_valid: 5.2685)     ELAPSED TIME: 2.26430
Epoch: 0,  Batch: 90 / 104,     (loss: 3.8570   | quality: 11.5464 | quality_valid: 16.7567)   ELAPSED TIME: 2.10427
Validity: -0.10000 % | Diversity: -0.10000 % | Reconstruction: 31.43010 % | True Reconstruction: 0.00000 %
Epoch: 1,  Batch: 0 / 104,      (loss: 3.7232   | quality: 27.7010 | quality_valid: 33.0957)   ELAPSED TIME: 0.08153
Epoch: 1,  Batch: 30 / 104,     (loss: 2.7471   | quality: 76.7320 | quality_valid: 76.1596)   ELAPSED TIME: 2.31865
Epoch: 1,  Batch: 60 / 104,     (loss: 1.5679   | quality: 77.7629 | quality_valid: 75.9608)   ELAPSED TIME: 2.22478
Epoch: 1,  Batch: 90 / 104,     (loss: 1.2733   | quality: 76.4845 | quality_valid: 76.0260)   ELAPSED TIME: 2.21133
latent_space_quality: Take 1000 samples from the latent space
Validity: 81.90000 % | Diversity: 42.00000 % | Reconstruction: 76.01608 % | True Reconstruction: 0.00000 %
Epoch: 2,  Batch: 0 / 104,      (loss: 1.1256   | quality: 78.0103 | quality_valid: 76.2573)   ELAPSED TIME: 0.08511
Epoch: 2,  Batch: 30 / 104,     (loss: 1.0559   | quality: 77.1959 | quality_valid: 76.8037)   ELAPSED TIME: 2.21546
Epoch: 2,  Batch: 60 / 104,     (loss: 0.9905   | quality: 77.8351 | quality_valid: 77.5390)   ELAPSED TIME: 2.34175
Epoch: 2,  Batch: 90 / 104,     (loss: 0.9867   | quality: 76.9072 | quality_valid: 77.8495)   ELAPSED TIME: 2.31603
latent_space_quality: Take 1000 samples from the latent space
Validity: 77.80000 % | Diversity: 38.90000 % | Reconstruction: 78.29155 % | True Reconstruction: 0.00000 %
Epoch: 3,  Batch: 0 / 104,      (loss: 0.9641   | quality: 77.6907 | quality_valid: 78.5513)   ELAPSED TIME: 0.09317
Epoch: 3,  Batch: 30 / 104,     (loss: 0.9291   | quality: 77.6598 | quality_valid: 78.4082)   ELAPSED TIME: 2.42319
Epoch: 3,  Batch: 60 / 104,     (loss: 0.8736   | quality: 78.8969 | quality_valid: 78.5823)   ELAPSED TIME: 2.34501
Epoch: 3,  Batch: 90 / 104,     (loss: 0.8017   | quality: 80.5773 | quality_valid: 79.3604)   ELAPSED TIME: 2.37803
latent_space_quality: Take 1000 samples from the latent space
Validity: 73.70000 % | Diversity: 36.50000 % | Reconstruction: 79.65320 % | True Reconstruction: 0.00000 %
Epoch: 4,  Batch: 0 / 104,      (loss: 0.8028   | quality: 80.0619 | quality_valid: 79.7027)   ELAPSED TIME: 0.07352
Epoch: 4,  Batch: 30 / 104,     (loss: 0.8715   | quality: 78.3299 | quality_valid: 79.8066)   ELAPSED TIME: 2.36590
Epoch: 4,  Batch: 60 / 104,     (loss: 0.7615   | quality: 81.1340 | quality_valid: 80.0697)   ELAPSED TIME: 2.45682
Epoch: 4,  Batch: 90 / 104,     (loss: 0.7095   | quality: 82.2784 | quality_valid: 80.4181)   ELAPSED TIME: 2.32484
latent_space_quality: Take 1000 samples from the latent space
Validity: 76.30000 % | Diversity: 37.10000 % | Reconstruction: 81.07299 % | True Reconstruction: 0.00000 %
Epoch: 5,  Batch: 0 / 104,      (loss: 0.7634   | quality: 80.9381 | quality_valid: 81.0049)   ELAPSED TIME: 0.08620
Epoch: 5,  Batch: 30 / 104,     (loss: 0.7357   | quality: 81.7526 | quality_valid: 80.8676)   ELAPSED TIME: 2.15332
Epoch: 5,  Batch: 60 / 104,     (loss: 0.7122   | quality: 82.1856 | quality_valid: 81.2433)   ELAPSED TIME: 2.27920
Epoch: 5,  Batch: 90 / 104,     (loss: 0.6796   | quality: 82.6495 | quality_valid: 81.4841)   ELAPSED TIME: 2.39881
latent_space_quality: Take 1000 samples from the latent space
Validity: 75.40000 % | Diversity: 35.40000 % | Reconstruction: 81.76000 % | True Reconstruction: 0.00000 %
```

QM Properties



```
Epoch: 0,  Batch: 0 / 104,      (loss: 4.0421   | quality: 2.8969 | quality_valid: 4.6280)     ELAPSED TIME: 0.27897
Epoch: 0,  Batch: 30 / 104,     (loss: 3.4335   | quality: 75.0619 | quality_valid: 73.9588)   ELAPSED TIME: 10.15275
Epoch: 0,  Batch: 60 / 104,     (loss: 2.6471   | quality: 76.3918 | quality_valid: 75.8833)   ELAPSED TIME: 10.04311
Epoch: 0,  Batch: 90 / 104,     (loss: 1.9659   | quality: 75.7526 | quality_valid: 75.6948)   ELAPSED TIME: 9.21233
latent_space_quality: Take 1000 samples from the latent space
Validity: 57.20000 % | Diversity: 36.50000 % | Reconstruction: 75.78103 %
Epoch: 1,  Batch: 0 / 104,      (loss: 1.8331   | quality: 74.1443 | quality_valid: 75.8676)   ELAPSED TIME: 0.30696
Epoch: 1,  Batch: 30 / 104,     (loss: 1.4187   | quality: 74.9588 | quality_valid: 76.0293)   ELAPSED TIME: 9.51069
Epoch: 1,  Batch: 60 / 104,     (loss: 1.2303   | quality: 75.6289 | quality_valid: 76.2821)   ELAPSED TIME: 9.88847
Epoch: 1,  Batch: 90 / 104,     (loss: 1.0969   | quality: 76.0103 | quality_valid: 76.2487)   ELAPSED TIME: 9.47961
latent_space_quality: Take 1000 samples from the latent space
Validity: 54.20000 % | Diversity: 36.30000 % | Reconstruction: 75.88454 %
Epoch: 2,  Batch: 0 / 104,      (loss: 1.0901   | quality: 75.4433 | quality_valid: 76.0709)   ELAPSED TIME: 0.22257
Epoch: 2,  Batch: 30 / 104,     (loss: 0.9586   | quality: 77.2371 | quality_valid: 76.2070)   ELAPSED TIME: 9.58339
Epoch: 2,  Batch: 60 / 104,     (loss: 0.9553   | quality: 77.4227 | quality_valid: 76.9027)   ELAPSED TIME: 9.75773
Epoch: 2,  Batch: 90 / 104,     (loss: 0.8364   | quality: 80.6804 | quality_valid: 78.1002)   ELAPSED TIME: 9.68887
latent_space_quality: Take 1000 samples from the latent space
Validity: 58.10000 % | Diversity: 36.60000 % | Reconstruction: 78.23753 %
Epoch: 3,  Batch: 0 / 104,      (loss: 0.9370   | quality: 78.0103 | quality_valid: 78.3010)   ELAPSED TIME: 0.22516
Epoch: 3,  Batch: 30 / 104,     (loss: 0.8220   | quality: 80.8247 | quality_valid: 78.6429)   ELAPSED TIME: 10.78502
Epoch: 3,  Batch: 60 / 104,     (loss: 0.8739   | quality: 79.9278 | quality_valid: 79.5885)   ELAPSED TIME: 9.91654
Epoch: 3,  Batch: 90 / 104,     (loss: 0.8409   | quality: 80.2887 | quality_valid: 80.5216)   ELAPSED TIME: 10.01509
latent_space_quality: Take 1000 samples from the latent space
Validity: 62.90000 % | Diversity: 36.90000 % | Reconstruction: 81.35711 %
Epoch: 4,  Batch: 0 / 104,      (loss: 0.8986   | quality: 79.2165 | quality_valid: 80.8495)   ELAPSED TIME: 0.32110
Epoch: 4,  Batch: 30 / 104,     (loss: 0.7738   | quality: 82.3608 | quality_valid: 81.5105)   ELAPSED TIME: 10.05132
Epoch: 4,  Batch: 60 / 104,     (loss: 0.8094   | quality: 81.3814 | quality_valid: 81.5612)   ELAPSED TIME: 9.81994
Epoch: 4,  Batch: 90 / 104,     (loss: 0.8373   | quality: 80.7423 | quality_valid: 81.7361)   ELAPSED TIME: 9.61340
latent_space_quality: Take 1000 samples from the latent space
Validity: 72.30000 % | Diversity: 37.30000 % | Reconstruction: 81.91464 %
Epoch: 5,  Batch: 0 / 104,      (loss: 0.7047   | quality: 83.7835 | quality_valid: 81.9439)   ELAPSED TIME: 0.21987
Epoch: 5,  Batch: 30 / 104,     (loss: 0.8101   | quality: 81.4021 | quality_valid: 82.0272)   ELAPSED TIME: 9.68650
Epoch: 5,  Batch: 60 / 104,     (loss: 0.7282   | quality: 82.4639 | quality_valid: 81.8136)   ELAPSED TIME: 9.50073
Epoch: 5,  Batch: 90 / 104,     (loss: 0.7232   | quality: 83.1546 | quality_valid: 81.6313)   ELAPSED TIME: 9.43337
```

Exploring The Chemical Space - Further Work

MCTS used in applications with large possibility spaces.

RL methods require suitable reward function.