IEEE WCCI 2022
IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE
PADUA, ITALY • 18 - 23 July, 2022

The International Joint Conference on Neural Networks

Predicting the number of biochemical transformations needed to synthesize a compound

João Correia[1,2], Rafael Carreira[3], Vítor Pereira[1,2], Miguel Rocha[1,2]

[1]Centre of Biological Engineering
University of Minho, Braga, Portugal
[2]LABBELS - Associate Laboratory,
Braga/Guimarães, Portugal
[3]Silicolife, Lda
Braga, Portugal

19th July, 2022

Motivation

- **Exploiting the natural metabolic abilities of microorganisms** for the **production of bioactive compounds** has been a research problem of great interest.

- The **economical and environmental costs** associated with petrochemical-derived industries have promoted the emergence of biochemical processes from renewable carbon sources

- Recently, some **retrobiosynthesis tools** for the design of de novo biosynthetic pathways have been proposed. These tools generate a **large number of intermediate compounds** that are **beyond experimental feasibility**.

- Thus, effective methods to **reduce the number of compounds** to screen by **selecting the most promising ones** are needed.

- In this study, we propose the use of **deep learning models to predict the number of biochemical transformations needed to produce a compound** from natural compounds.

Objectives

- **Generate a dataset** of **intermediate compounds** from a pool of **starting materials** (natural compounds) using **reaction rules**.

- **Predict the number of biochemical transformations needed to synthesize a compound** using **deep learning models**.

- **Explore different compound representations** and **model architectures,** including **classification** and **regression** approaches**.**

Reaction Rules

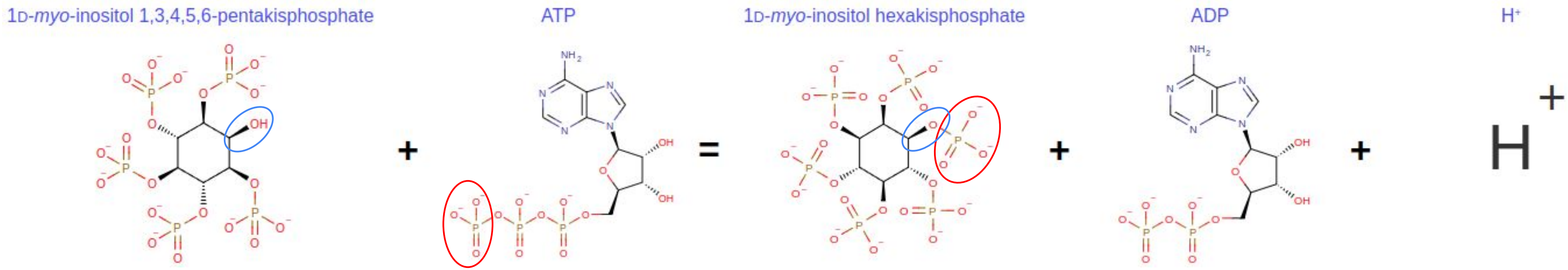
- Reaction rules are **generic descriptions of reactions** that **encode the way reactants are converted into products**. A reaction rule can be applied to a compound if the compound contains a particular **substructure that is encoded by the reaction rule**.

- In this study, we used a set of **13055 reaction rules** represented as SMARTS. These reaction rules were retrieved from the **RetroRules** and **MINE** databases.

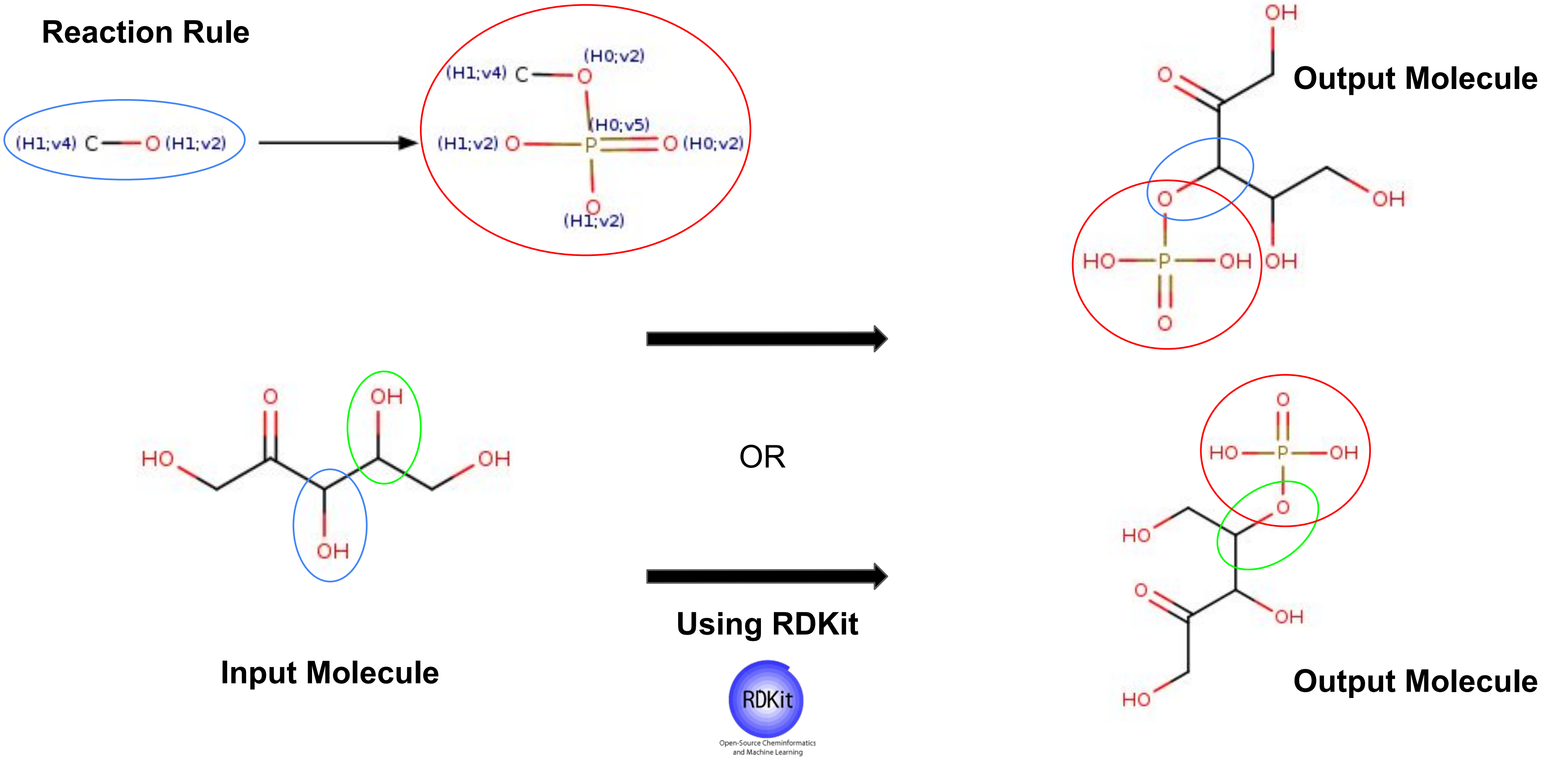
Original Reaction

RHEA:20313

EC 2.7.1.158 Inositol-pentakisphosphate 2-kinase


1D-*myo*-inositol 1,3,4,5,6-pentakisphosphate
+
ATP
=
1D-*myo*-inositol hexakisphosphate
+
ADP
+
H+
H+

Reaction Rules



Reaction Rule

Input Molecule

OR

Using RDKit

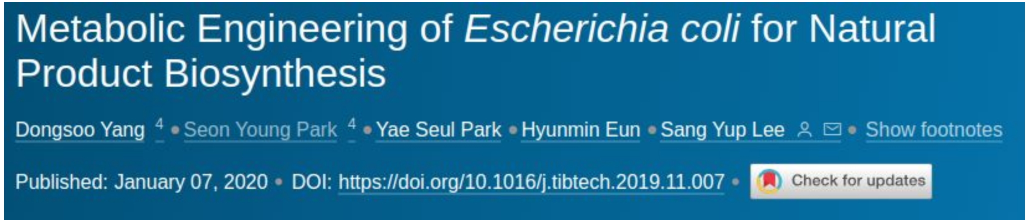
Output Molecule

Output Molecule

Starting Compounds

- The list of **starting precursors** that we assume to be available are the ones existing in the **metabolism from *Escherichia coli***. We selected this microorganism because it is **widely used** host for bioengineering processes including in the **synthesis of added-value compounds**.

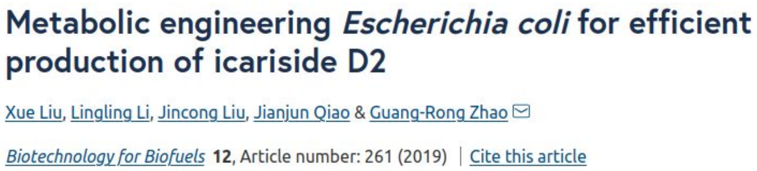
- These compounds were obtained from the RetroPathRL GitHub (https://github.com/brsynth/RetroPathRL). The compounds with available and valid identifiers were selected resulting in a set of **673 starting compounds**.



Metabolic Engineering of *Escherichia coli* for Natural Product Biosynthesis

Dongsoo Yang [4] • Seon Young Park [4] • Yae Seul Park • Hyunmin Eun • Sang Yup Lee • Show footnotes

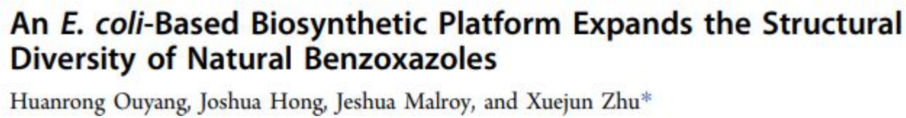
Published: January 07, 2020 • DOI: https://doi.org/10.1016/j.tibtech.2019.11.007 • Check for updates



Metabolic engineering *Escherichia coli* for efficient production of icariside D2

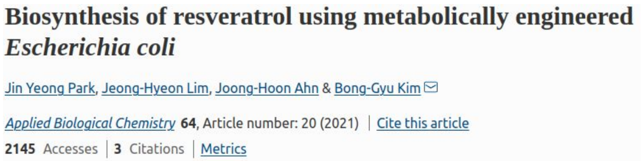
Xue Liu, Lingling Li, Jincong Liu, Jianjun Qiao & Guang-Rong Zhao

Biotechnology for Biofuels **12**, Article number: 261 (2019) | Cite this article



An *E. coli*-Based Biosynthetic Platform Expands the Structural Diversity of Natural Benzoxazoles

Huanrong Ouyang, Joshua Hong, Jeshua Malroy, and Xuejun Zhu*



Biosynthesis of resveratrol using metabolically engineered *Escherichia coli*

Jin Yeong Park, Jeong-Hyeon Lim, Joong-Hoon Ahn & Bong-Gyu Kim

Applied Biological Chemistry **64**, Article number: 20 (2021) | Cite this article

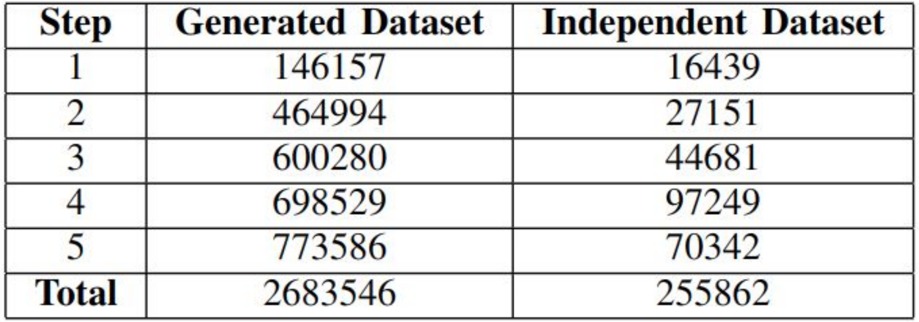
2145 Accesses | **3** Citations | Metrics

Generated Datasets

- The dataset used to **train and evaluate our DL models** was generated by successively **applying randomly selected reaction rules to randomly selected** compounds from the previous step. In the first step, we use the starting compound set (673 compounds).

- Since the compounds present in later steps were generated using the ones from the previous step, there is a **dependency between the compounds generated at each step**. To validate if the previously generated dataset was **representative** enough we generated an **independent set** using the same approach.

NEW COMPOUNDS GENERATED AT EACH STEP.

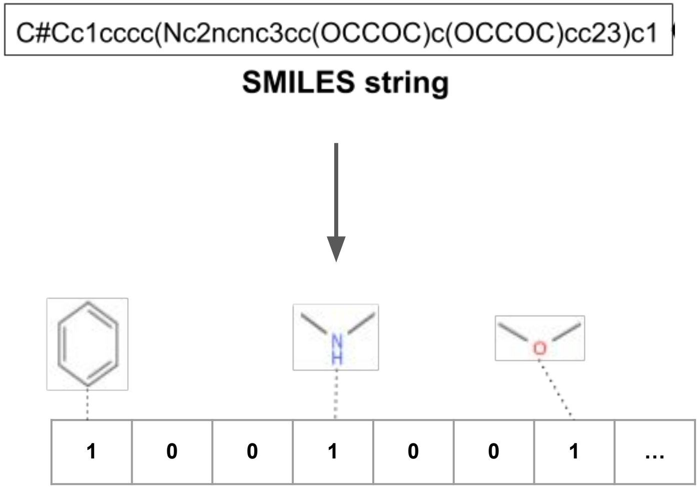


Step	Generated Dataset	Independent Dataset
1	146157	16439
2	464994	27151
3	600280	44681
4	698529	97249
5	773586	70342
Total	2683546	255862

Molecular Representations

- In this study, we focused on two distinct molecular representations, the well-known **Morgan fingerprints** and the NLP-based **Molecular Transformer Embeddings (MTE)**.

- We computed Morgan fingerprints of **radius 2** hashed to **1024 bits** using RDKit.



C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1

SMILES string

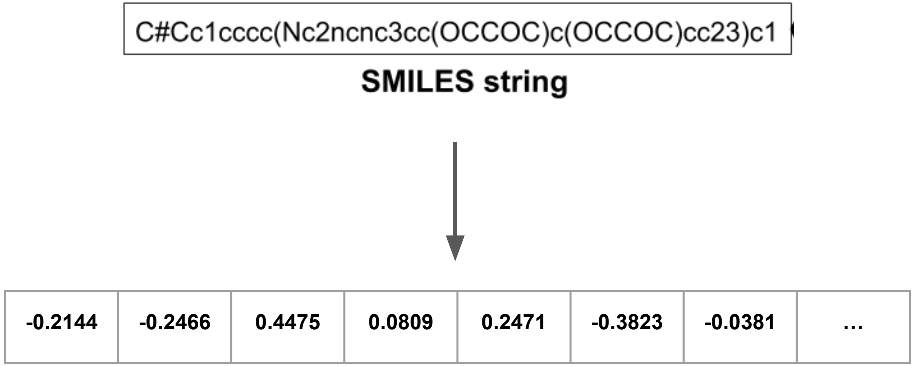
1	0	0	1	0	0	1	...

Morgan Fingerprints

Molecular Representations

- The MTE is a **transformer-based model** that was trained and repurposed, through transfer learning, to predict binding affinity. We used the **intermediate embeddings** that represent abstract features that **describe general molecular structures**.

- We computed these MTE for our datasets with a defined maximum **length of our compound SMILES of 300 characters** and an **embedding size of 512**.



C#Cc1cccc(Nc2ncnc3cc(OCCOC)c(OCCOC)cc23)c1

SMILES string

-0.2144	-0.2466	0.4475	0.0809	0.2471	-0.3823	-0.0381	…

Molecular Transformer Embeddings

Models

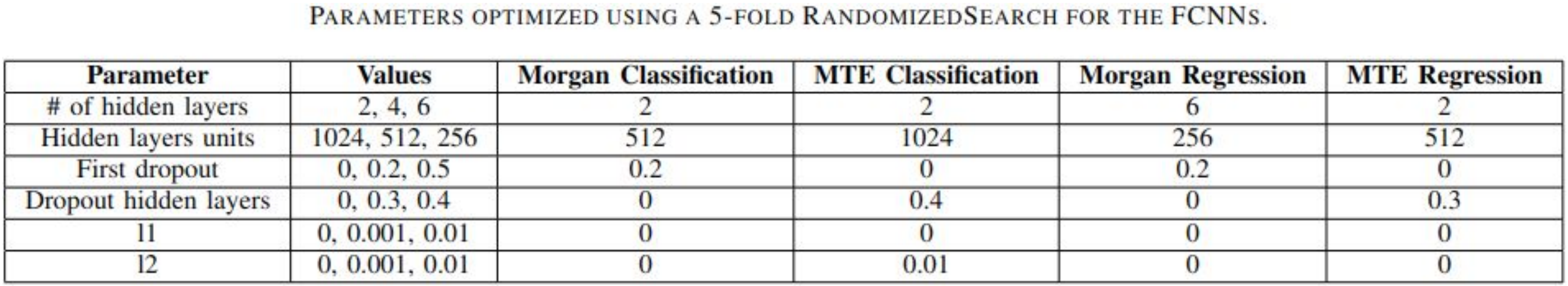
- We consider the use of 2 different model architectures: **Fully Connected Neural Networks (FCNN)** and **1D-Convolutional Neural Networks (1D-CNN)** working over features created from the previous two representations

- We performed **hyperparameter optimization** using **5-fold RandomizedSearchCV for 15 iterations** and a **3-fold for 10 iterations** for our **FCNN** and **1D-CNN** models, respectively.

- Both these architectures were used in **classification** and **regression** approaches, changing only the final layer, and also the error metrics

Models: hyperparameters tested ans selected configurations

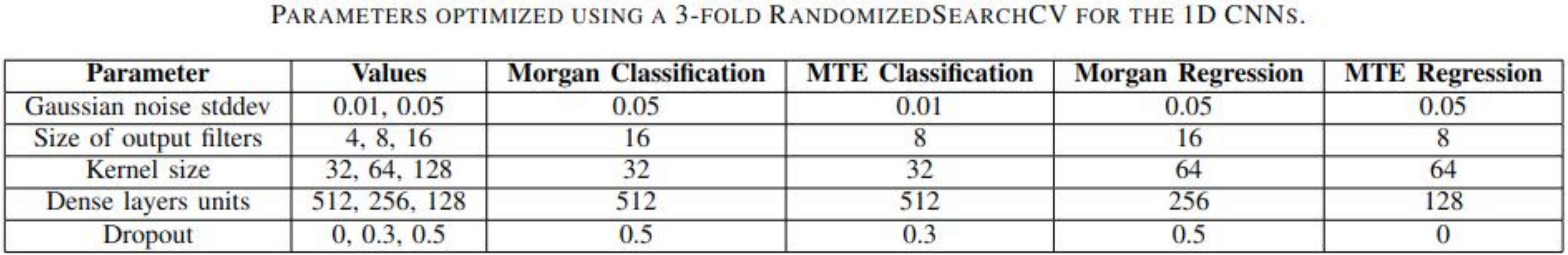
- **Fully Connected Neural Networks**:



PARAMETERS OPTIMIZED USING A 5-FOLD RANDOMIZEDSEARCH FOR THE FCNNS.

Parameter	Values	Morgan Classification	MTE Classification	Morgan Regression	MTE Regression
# of hidden layers	2, 4, 6	2	2	6	2
Hidden layers units	1024, 512, 256	512	1024	256	512
First dropout	0, 0.2, 0.5	0.2	0	0.2	0
Dropout hidden layers	0, 0.3, 0.4	0	0.4	0	0.3
l1	0, 0.001, 0.01	0	0	0	0
l2	0, 0.001, 0.01	0	0.01	0	0

- **1D Convolutional Neural Networks:**



PARAMETERS OPTIMIZED USING A 3-FOLD RANDOMIZEDSEARCHCV FOR THE 1D CNNS.

Parameter	Values	Morgan Classification	MTE Classification	Morgan Regression	MTE Regression
Gaussian noise stddev	0.01, 0.05	0.05	0.01	0.05	0.05
Size of output filters	4, 8, 16	16	8	16	8
Kernel size	32, 64, 128	32	32	64	64
Dense layers units	512, 256, 128	512	512	256	128
Dropout	0, 0.3, 0.5	0.5	0.3	0.5	0

Results and Discussion - Classification

- We obtained considerably **better results** when using **morgan fingerprints** as input.

- With the **FCNN** we also obtained **slightly better results** when compared with the **1D CNN.**

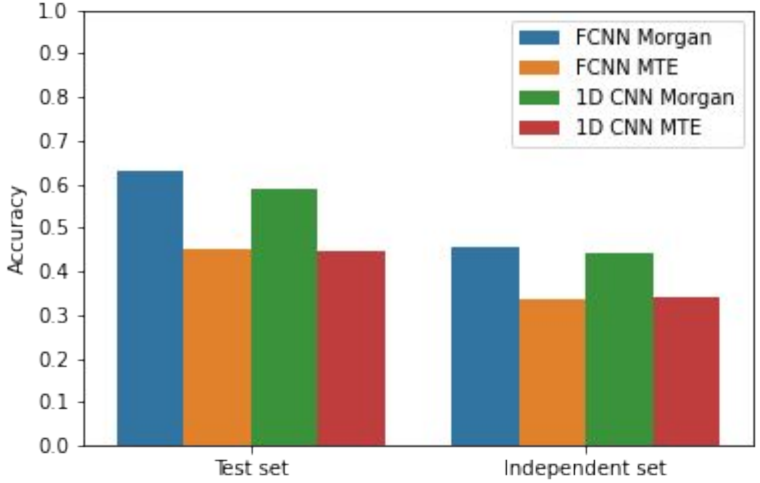


Fig. x - Test and independent set accuracy for all models.

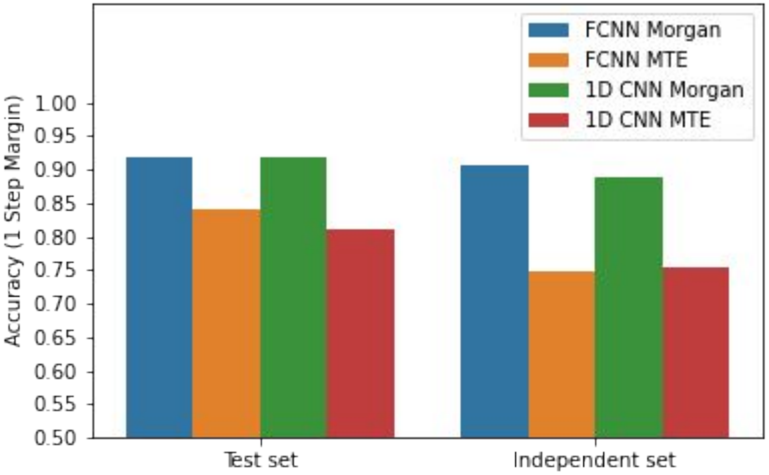
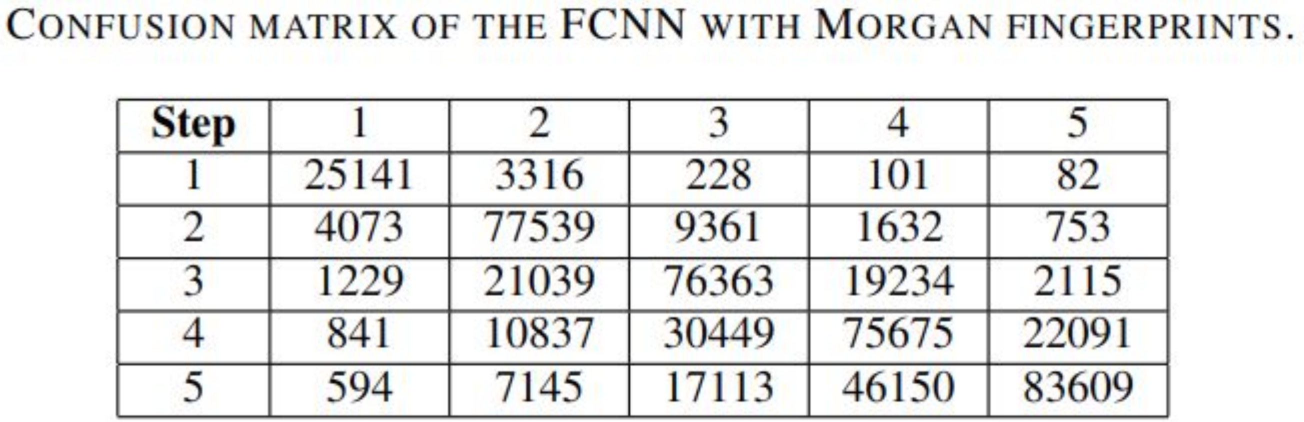


Fig. x - Test and independent set accuracy allowing miss-classification by one step for all models.

Results and Discussion - Classification

- If we take a closer look at the **confusion matrix**, we can see that the majority of the mispredictions, around **78%, fail by one step**, which may be a **reasonable estimate in practical applications.**

- This can also mean that **this problem can better be modeled as a regression task**.



CONFUSION MATRIX OF THE FCNN WITH MORGAN FINGERPRINTS.

Step	1	2	3	4	5
1	25141	3316	228	101	82
2	4073	77539	9361	1632	753
3	1229	21039	76363	19234	2115
4	841	10837	30449	75675	22091
5	594	7145	17113	46150	83609

Results and Discussion - Regression

- Again, we obtained **considerably better results** when using **morgan fingerprints** as input.

- However, the comparison between **FCNN** and **1DCNN was not so clear**, with very similar results.

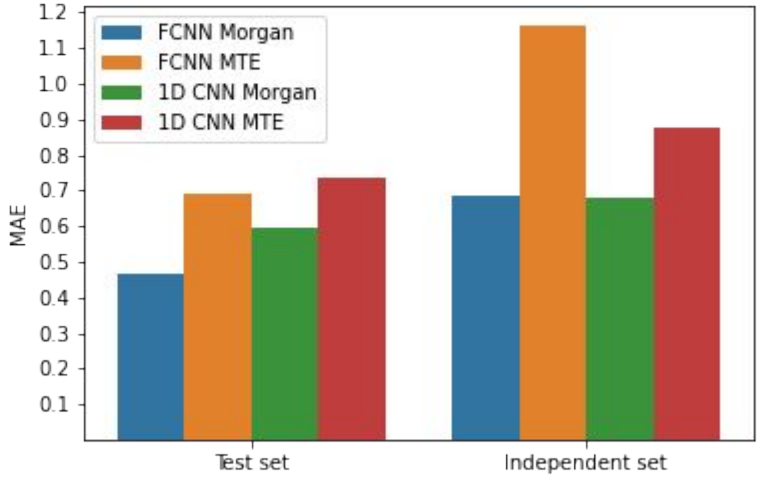
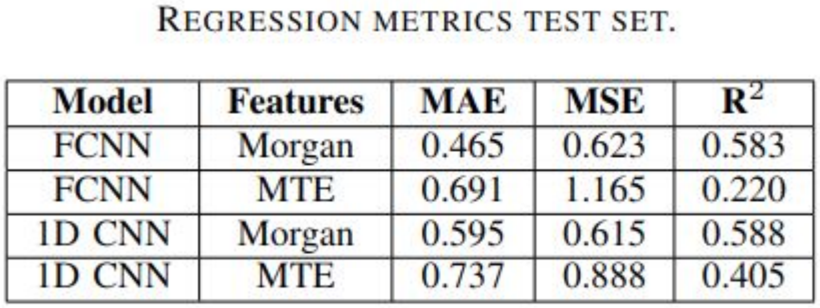


Fig. x - Test and independent set MAE for all models.



REGRESSION METRICS TEST SET.

Model	Features	MAE	MSE	R^2
FCNN	Morgan	0.465	0.623	0.583
FCNN	MTE	0.691	1.165	0.220
1D CNN	Morgan	0.595	0.615	0.588
1D CNN	MTE	0.737	0.888	0.405

Conclusion and Future Work

- In this study, we propose the use of different **DL architectures** and **molecular representations** to **predict the number of biochemical transformations needed to synthesize a compound** having the *E. coli* metabolites as available starting materials.

- As far as we know, this is the **first time** that the prediction of the number of biochemical transformations needed to synthesize a compound using DL is described in the literature.

- Despite the lack of other studies to compare our results with, we can say that the results obtained by our best models, a **63% accuracy**, **92% if we give a one step margin**, in a **5-label classification** and **0.465 MAE** in the regression, **are promising**.

- Approaches like this one can benefit the field of ME and specially be useful in **retrobiosynthesis tools** to **narrow the number of generated compounds** allowing the exploration of most promising pathways for the synthesis of target compounds.

Conclusion and Future Work

- In the future, it would be interesting to test other compound representations and models like recurrent neural networks and the Transformer architecture.

- Further exploration of the data can also be conducted to understand if the **generated data** are **representative of what happens in microbial networks** and which **types of biochemical reactions** are being **prioritized** when generating new data.

- **Model interpretability** could also be explored to understand **why the models make certain predictions** and which properties of the molecules are more impactful for those predictions.

IEEE WCCI 2022

IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE

PADUA, ITALY • 18 - 23 July, 2022

The International Joint Conference on Neural Networks

Questions?

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Universidade do Minho
Escola de Engenharia

CENTRE OF BIOLOGICAL ENGINEERING

LABBELS
ASSOCIATE LAB

SILICOLIFE
Computational Biology Solutions for the Life Sciences