May 9, 2025

Advancements in Biotransformation Pathway Prediction

Enhancements, Datasets, and Novel Functionalities in enviPath

Jörg Simon Wicker

enviPath / School of Computer Science – University of Auckland

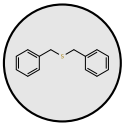
enviPath

UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

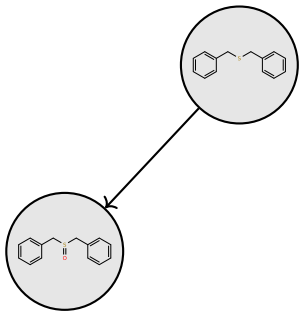
SCIENCE

Biodegradation Pathway Prediction

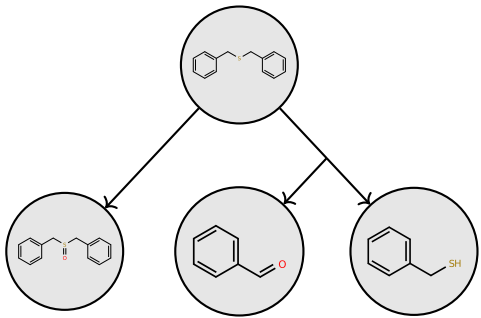
Biodegradation Pathways



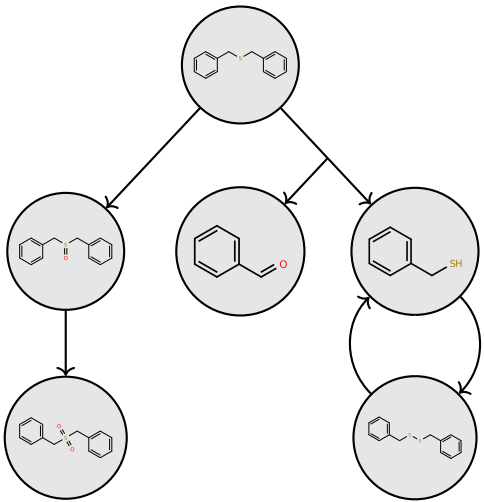
Biodegradation Pathways



Biodegradation Pathways



Biodegradation Pathways



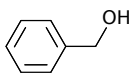
Rule-Based Biodegradation Pathway Prediction

- Rule-based systems predict degradation products using transformation rules
  - Generalizations and abstractions of observed reactions
  - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side

Rule-Based Biodegradation Pathway Prediction

- Rule-based systems predict degradation products using transformation rules
  - Generalizations and abstractions of observed reactions
  - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side

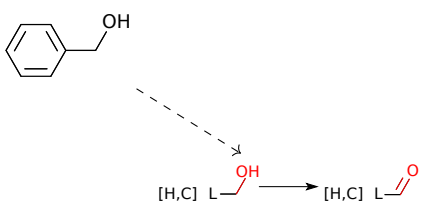
Benzyl Alcohol



Rule-Based Biodegradation Pathway Prediction

- Rule-based systems predict degradation products using transformation rules
  - Generalizations and abstractions of observed reactions
  - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side

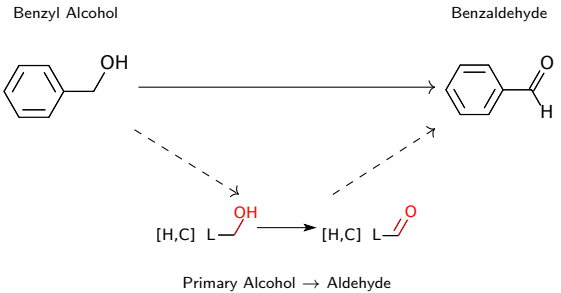
Benzyl Alcohol



Primary Alcohol → Aldehyde

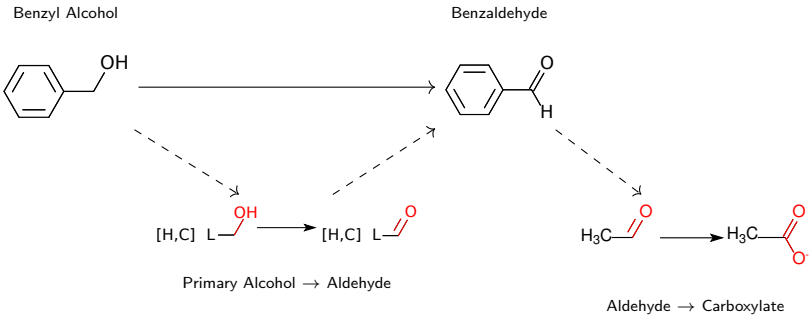
Rule-Based Biodegradation Pathway Prediction

- Rule-based systems predict degradation products using transformation rules
  - Generalizations and abstractions of observed reactions
  - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side



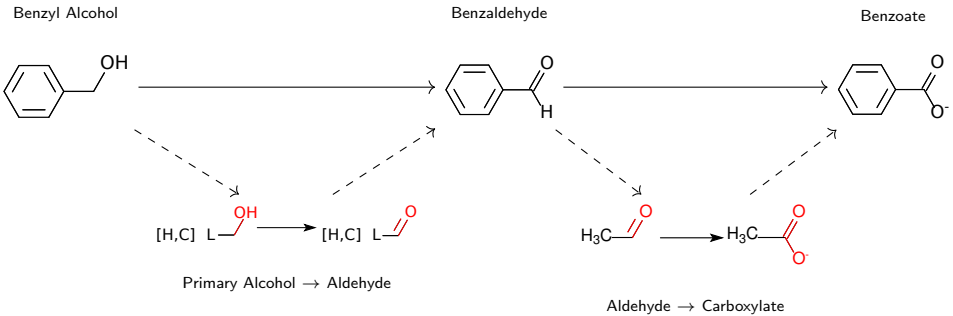
Rule-Based Biodegradation Pathway Prediction

- Rule-based systems predict degradation products using transformation rules
  - Generalizations and abstractions of observed reactions
  - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side

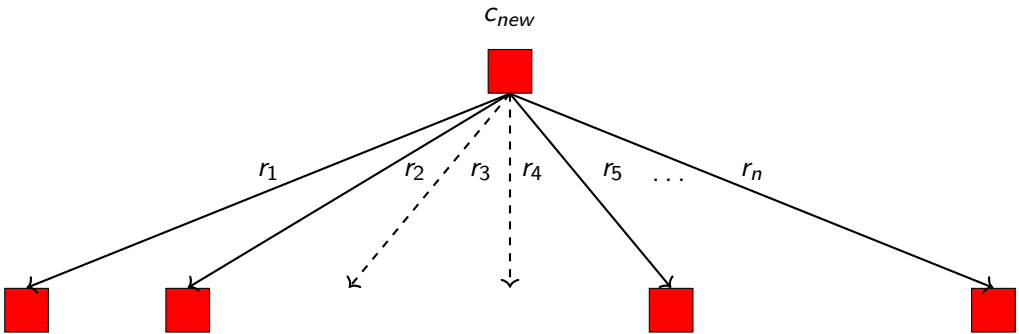


Rule-Based Biodegradation Pathway Prediction

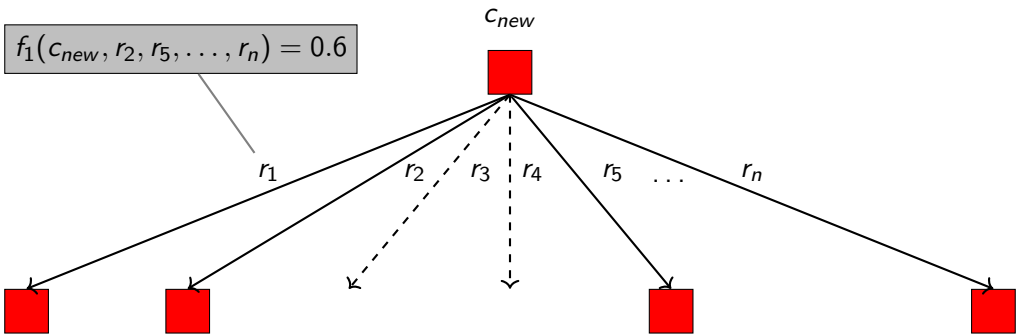
- Rule-based systems predict degradation products using transformation rules
  - Generalizations and abstractions of observed reactions
  - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side



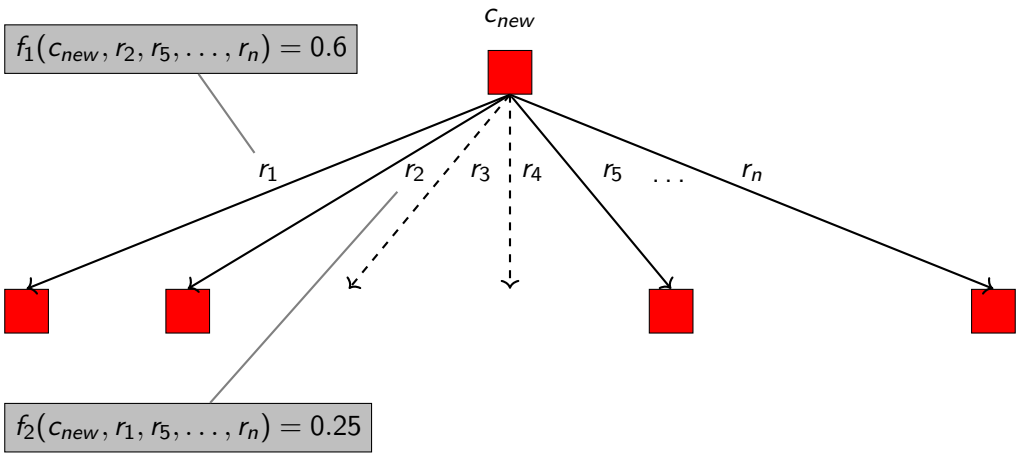
Machine Learning to Limit Combinatorial Explosion



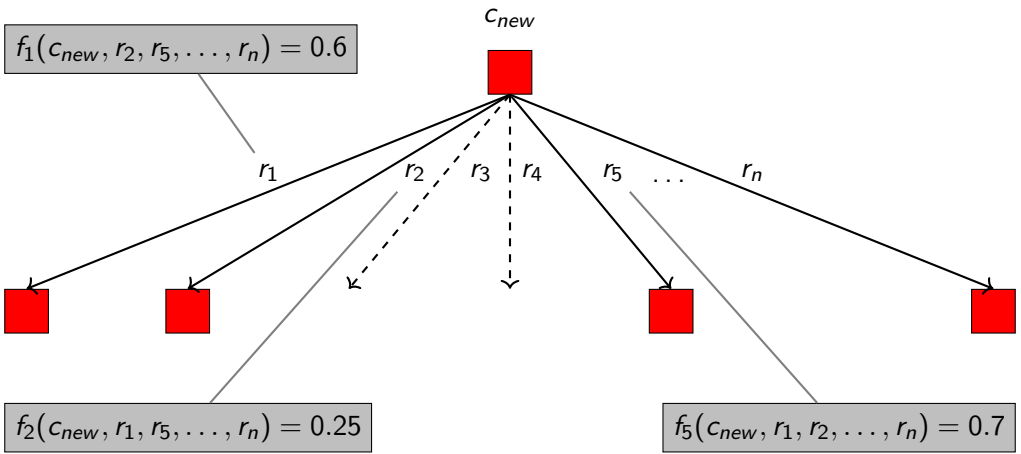
Machine Learning to Limit Combinatorial Explosion



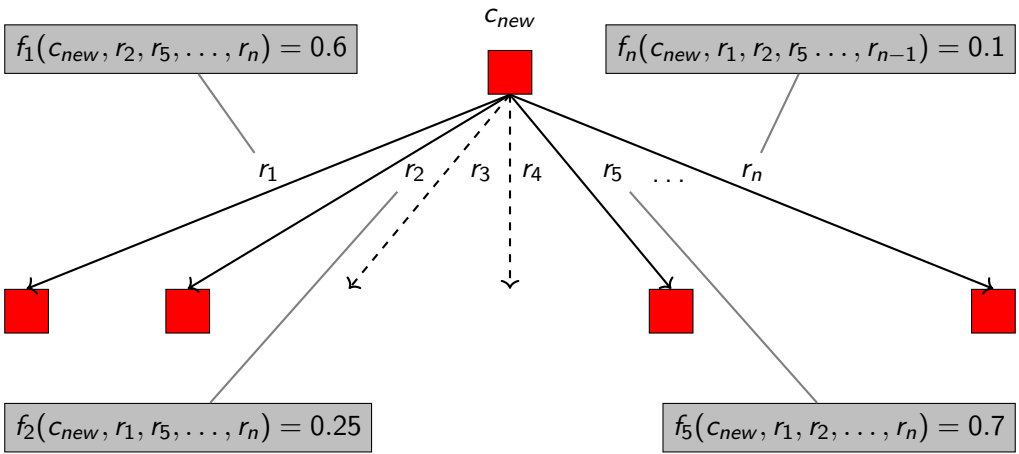
Machine Learning to Limit Combinatorial Explosion



Machine Learning to Limit Combinatorial Explosion



Machine Learning to Limit Combinatorial Explosion



Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data

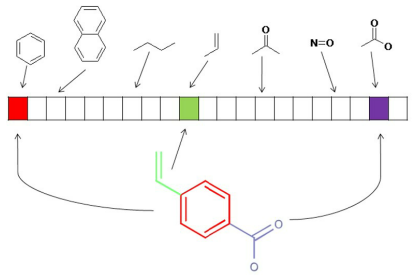


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data
- Key step is to transform data into tables

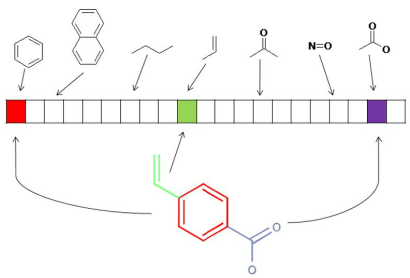


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data
- Key step is to transform data into tables
- For compounds, various approaches exist

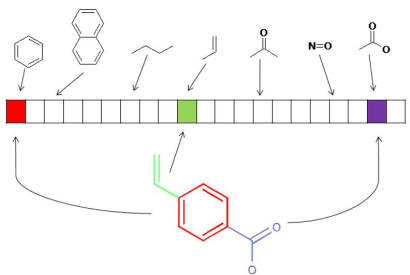


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data
- Key step is to transform data into tables
- For compounds, various approaches exist
  - ECFP, MACCS, Mol2vec, Spectrophores, ...

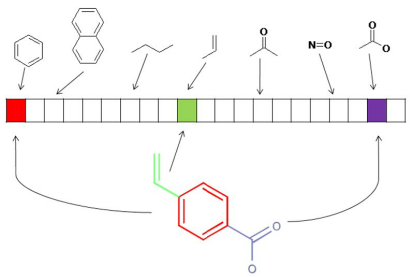


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data
- Key step is to transform data into tables
- For compounds, various approaches exist
  - ECFP, MACCS, Mol2vec, Spectrophores, ...
- Lately a lot of approaches to generate embeddings using deep learning

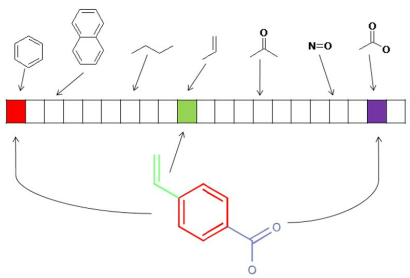


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data
- Key step is to transform data into tables
- For compounds, various approaches exist
  - ECFP, MACCS, Mol2vec, Spectrophores, ...
- Lately a lot of approaches to generate embeddings using deep learning
  - Stepišnik *et al.*, *A comprehensive comparison of molecular feature representations for use in predictive modeling*, (2021) gives an overview

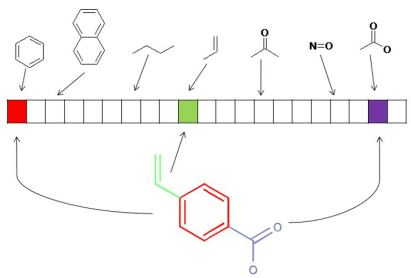


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data
- Key step is to transform data into tables
- For compounds, various approaches exist
  - ECFP, MACCS, Mol2vec, Spectrophores, ...
- Lately a lot of approaches to generate embeddings using deep learning
  - Stepišnik *et al.*, *A comprehensive comparison of molecular feature representations for use in predictive modeling*, (2021) gives an overview
- Simple approaches (e.g. pre-defined substructures) perform similarly well to more complicated approaches

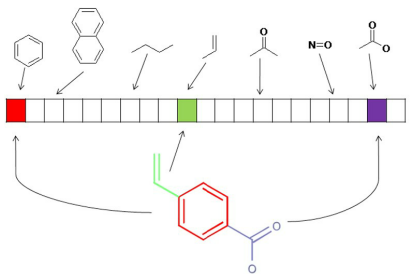
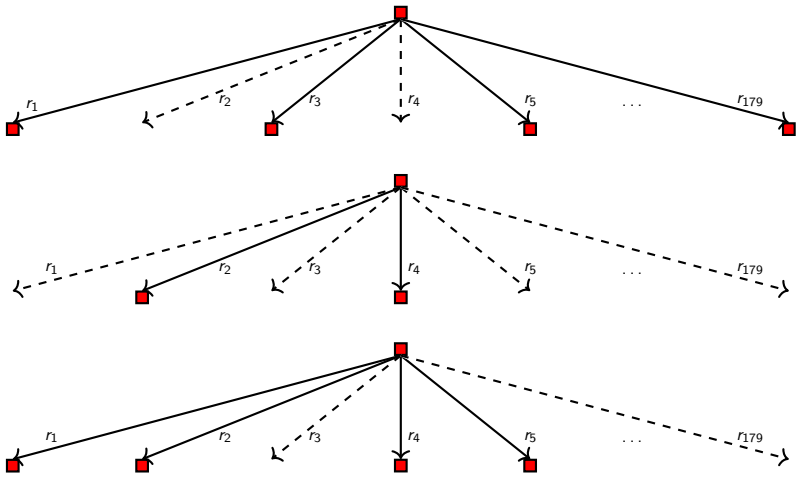
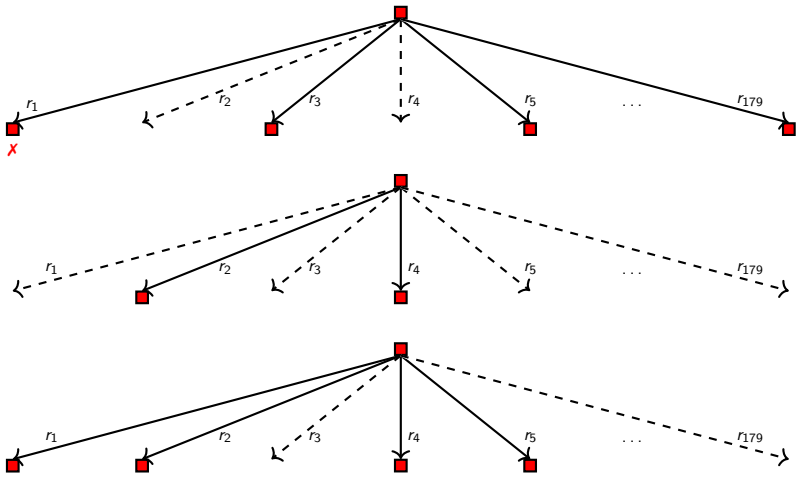


image from chemopy manual, Cao *et al.* (2012)

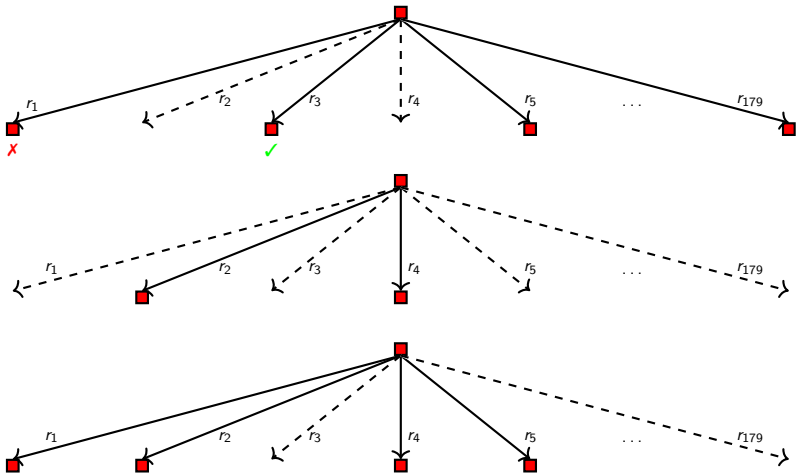
Machine Learning to Limit Combinatorial Explosion



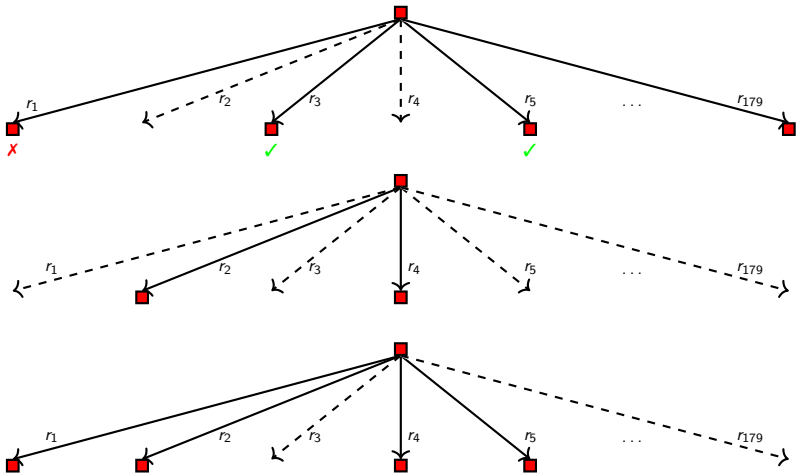
Machine Learning to Limit Combinatorial Explosion



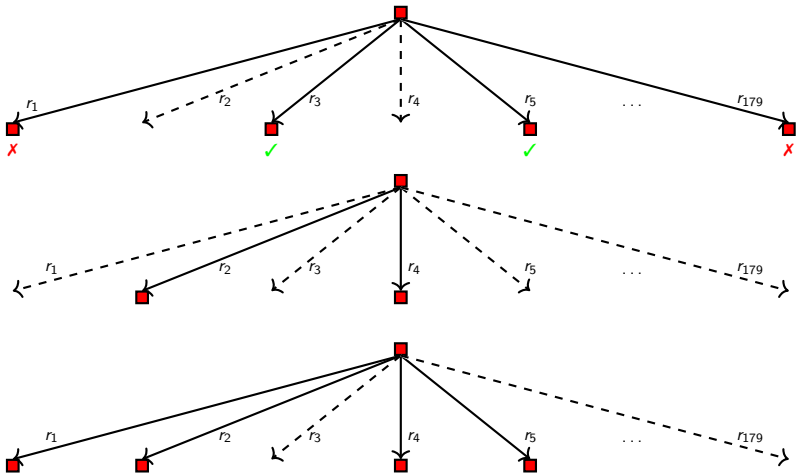
Machine Learning to Limit Combinatorial Explosion



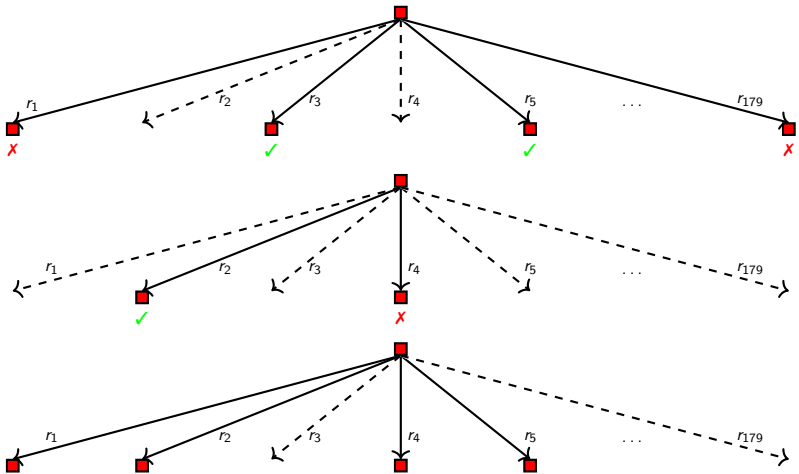
Machine Learning to Limit Combinatorial Explosion



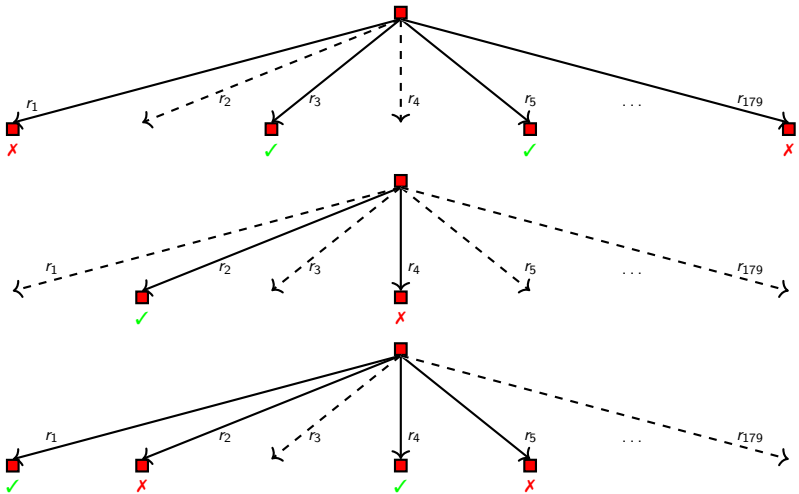
Machine Learning to Limit Combinatorial Explosion



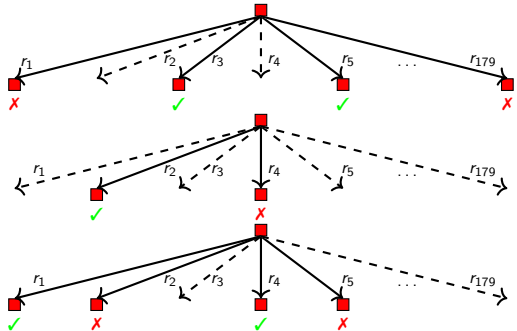
Machine Learning to Limit Combinatorial Explosion



Machine Learning to Limit Combinatorial Explosion

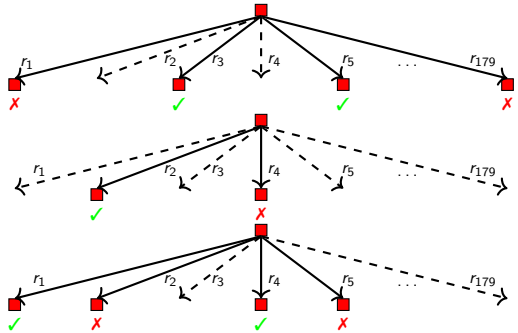


Generation of Training Sets



	s_1	...	s_{547}	r_2	r_3	r_4	r_5	...	r_{179}	$y = r_1$
c_1	+1	...	0	0	+1	0	+1	...	+1	0
c_3	+1	...	+1	+1	0	+1	+1	...	0	+1
...	...	...	...	...	...	...	...	...	...	...
c_{718}	0	...	0	+1	0	0	+1	...	0	0

Generation of Training Sets



	s_1	...	s_{547}	r_1	r_3	r_4	r_5	...	r_{179}	$y = r_2$
c_2	0	...	+1	0	0	+1	0	...	0	+1
c_3	+1	...	+1	+1	0	+1	+1	...	0	0
...	...	...	...	...	...	...	...	...	...	...
c_{718}	0	...	0	+1	0	0	+1	...	0	+1

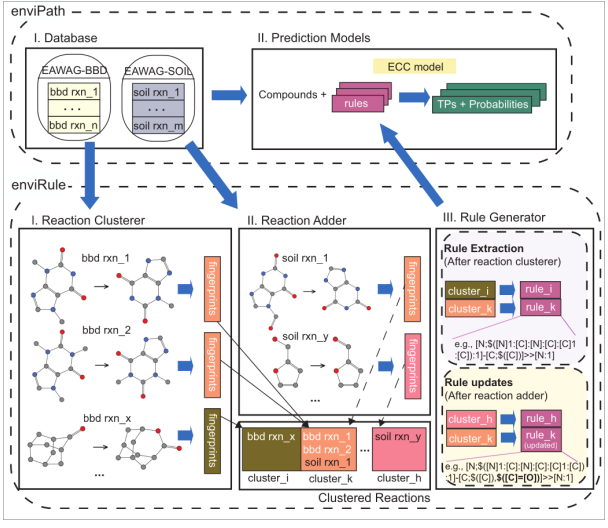
Biodegradation – Multi-Label Perspective

- Learning task is a multi-label problem
- Transformation to multi-label data set using pseudo-label

	correctly triggered	incorrectly triggered	not triggered
label 1 (λ_i) / correctly triggered	1	0	?
label 2 (λ'_i) / known product	1	0	0
feature (r_i) / triggered	1	1	0

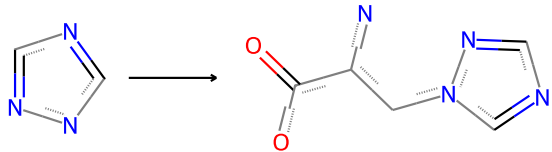
	s_1	...	s_{547}	r_1	...	r_{179}	λ'_1	...	λ'_{179}	λ_1	...	λ_{179}
c_1	+1	...	0	1	...	0	1	...	0	1	...	?
c_2	+1	...	+1	1	...	0	0	...	0	0	...	?
c_3	0	...	+1	1	...	1	0	...	1	0	...	1
...	...	...	...	...	...	...	...	...	...	...	...	...
c_{718}	0	...	0	0	...	1	0	...	0	?	...	0

Learning new Rules from Data



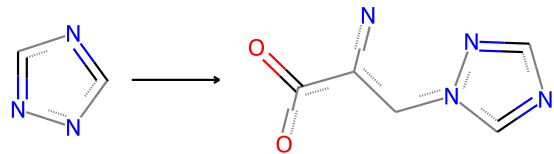
enviFormer – Biodegradation Prediction as Sequence-to-Sequence Task

- Compounds and reactions can be represented as Strings (SMILES and SMIRKS)



enviFormer – Biodegradation Prediction as Sequence-to-Sequence Task

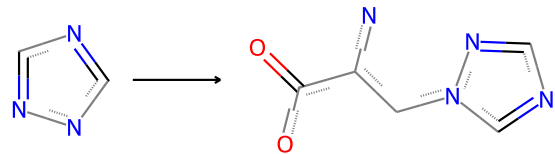
- Compounds and reactions can be represented as Strings (SMILES and SMIRKS)



C1=NC=NN1 >> C(C(C(=O)O)N)N1C=NC=N1

enviFormer – Biodegradation Prediction as Sequence-to-Sequence Task

- Compounds and reactions can be represented as Strings (SMILES and SMIRKS)
- This makes the prediction task very similar to the learning task of LLMs such as GPT
  - Given an input sequence (question / compound), predict the most likely output sequence (answer / compound)



C1=NC=NN1 >> C(C(C(=O)O)N)N1C=NC=N1

enviFormer – Training Set

features class

enviFormer – Training Set

	features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC		

enviFormer – Training Set

	features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC		
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C	

enviFormer – Training Set

	features	class
1	CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
2	CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
3	CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	
CCC(O)CO.CCC1CO1>>CCC(O)COCC (	

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	
CCC(O)CO.CCC1CO1>>CCC(O)COCC (	
CCC(O)CO.CCC1CO1>>CCC(O)COC	C

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	
CCC(O)CO.CCC1CO1>>CCC(O)COCC (	
CCC(O)CO.CCC1CO1>>CCC(O)COC	C
CCC(O)CO.CCC1CO1>>CCC(O)CO	C

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(	O
CCC(O)CO.CCC1CO1>>CCC(O)COCC	(
CCC(O)CO.CCC1CO1>>CCC(O)COC	C
CCC(O)CO.CCC1CO1>>CCC(O)CO	C
CCC(O)CO.CCC1CO1>>CCC(O)C	O

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	
CCC(O)CO.CCC1CO1>>CCC(O)COCC (	
CCC(O)CO.CCC1CO1>>CCC(O)COC	C
CCC(O)CO.CCC1CO1>>CCC(O)CO	C
CCC(O)CO.CCC1CO1>>CCC(O)C	O
CCC(O)CO.CCC1CO1>>CCC(O)	C

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	
CCC(O)CO.CCC1CO1>>CCC(O)COCC (	
CCC(O)CO.CCC1CO1>>CCC(O)COC	C
CCC(O)CO.CCC1CO1>>CCC(O)CO	C
CCC(O)CO.CCC1CO1>>CCC(O)C	O
CCC(O)CO.CCC1CO1>>CCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)	

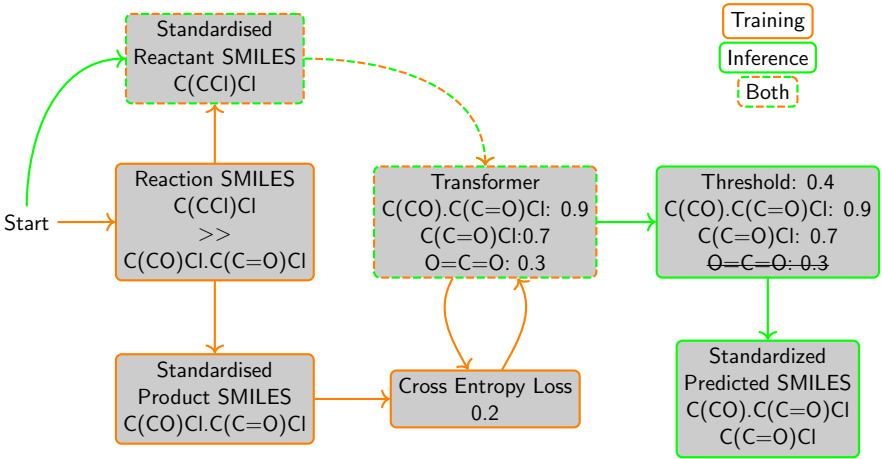
enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	
CCC(O)CO.CCC1CO1>>CCC(O)COCC (	
CCC(O)CO.CCC1CO1>>CCC(O)COC	C
CCC(O)CO.CCC1CO1>>CCC(O)CO	C
CCC(O)CO.CCC1CO1>>CCC(O)C	O
CCC(O)CO.CCC1CO1>>CCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O	

enviFormer – Training Set

features	class
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)CC	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)C	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O)COCC(O	
CCC(O)CO.CCC1CO1>>CCC(O)COCC (	
CCC(O)CO.CCC1CO1>>CCC(O)COC	C
CCC(O)CO.CCC1CO1>>CCC(O)CO	C
CCC(O)CO.CCC1CO1>>CCC(O)C	O
CCC(O)CO.CCC1CO1>>CCC(O)	C
CCC(O)CO.CCC1CO1>>CCC(O)	
CCC(O)CO.CCC1CO1>>CCC(O	
...	

enviFormer – Method



enviFormer

- Overcome small dataset size with pre-training on generic reaction data set
- Advantage: Remove the need of manual rule creation, just need a data set with reactions
- Disadvantage: Loss of interpretability

enviPath

- enviPath is an active research project run by University of Auckland, Eawag, and University of Zürich

enviPath

- enviPath is an active research project run by University of Auckland, Eawag, and University of Zürich
- The server is free to use for non-commercial use cases

enviPath

- enviPath is an active research project run by University of Auckland, Eawag, and University of Zürich
- The server is free to use for non-commercial use cases
- *enviPath Limited* sells licenses for commercial entities

enviPath

- enviPath is an active research project run by University of Auckland, Eawag, and University of Zürich
- The server is free to use for non-commercial use cases
- *enviPath Limited* sells licenses for commercial entities
  - Several use cases from providing server infrastructure to supporting local IT to run a copy of enviPath

enviPath

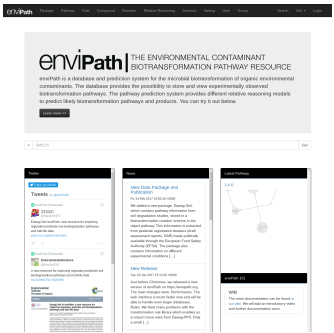
- enviPath is an active research project run by University of Auckland, Eawag, and University of Zürich
- The server is free to use for non-commercial use cases
- *enviPath Limited* sells licenses for commercial entities
  - Several use cases from providing server infrastructure to supporting local IT to run a copy of enviPath
- We also support researchers who want to publish their data on enviPath

enviPath

- enviPath is an active research project run by University of Auckland, Eawag, and University of Zürich
- The server is free to use for non-commercial use cases
- *enviPath Limited* sells licenses for commercial entities
  - Several use cases from providing server infrastructure to supporting local IT to run a copy of enviPath
- We also support researchers who want to publish their data on enviPath
- *Reach out to us if you are interested!*

Outlook

- Current work
  - Implementing enviFormer into enviPath
  - Adding more data
  - Adding more improvements into the prediction engine
- Complete reimplementation of the system
  - Faster, more robust, more reliable
  - Frequently changing (and breaking) beta-version at `https://playground.envipath.org`



enviPath

UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

SCIENCE

Thank you for listening! Any questions?

https://envipath.org
https://envipath.com
https://wickerlab.org