

May 9, 2025

Advancements in Biotransformation Pathway Prediction Enhancements, Datasets, and Novel Functionalities in enviPath

Jörg Simon Wicker

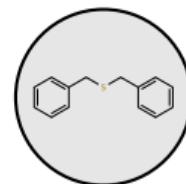
enviPath / School of Computer Science – University of Auckland

enviPath



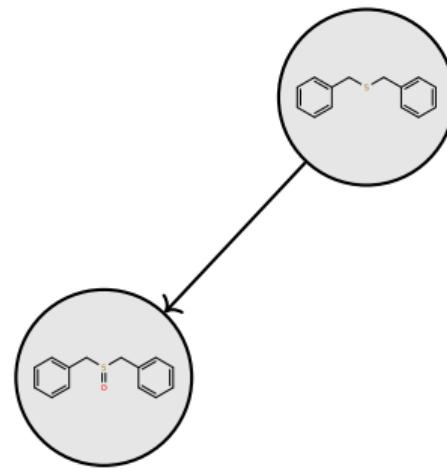
Biodegradation Pathway Prediction

Biodegradation Pathways

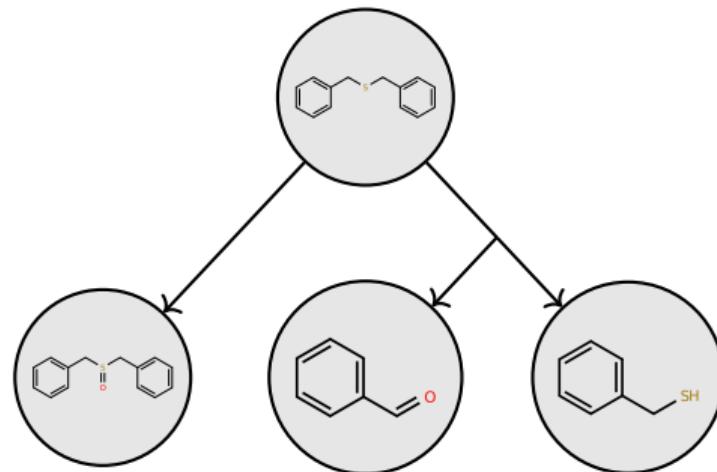


Biodegradation Pathways

eP

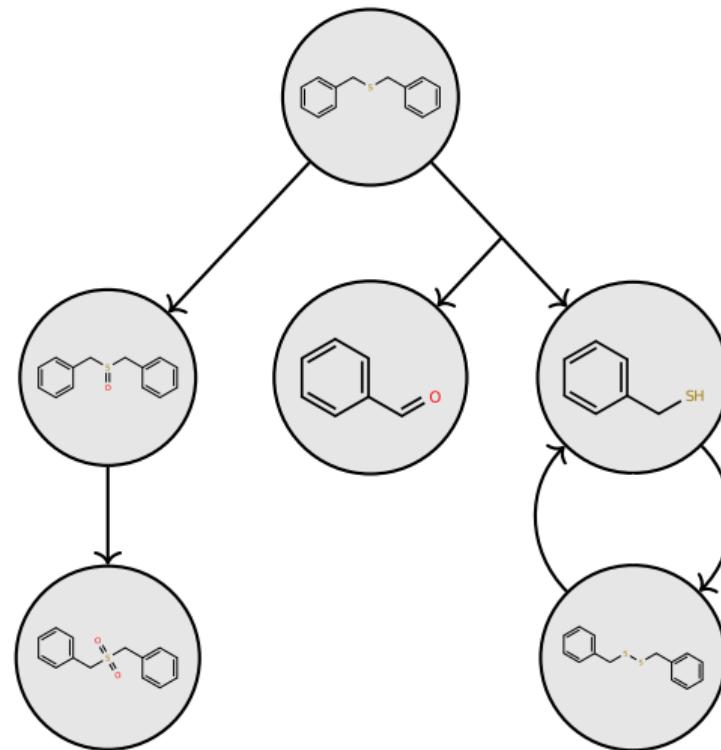


Biodegradation Pathways



Biodegradation Pathways

eP



Rule-Based Biodegradation Pathway Prediction

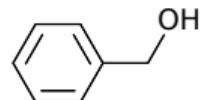


- Rule-based systems predict degradation products using transformation rules
 - Generalizations and abstractions of observed reactions
 - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side

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



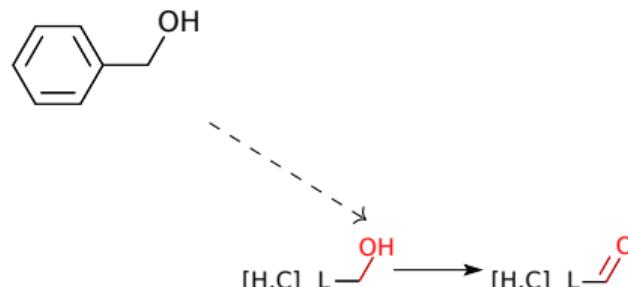
Rule-Based Biodegradation Pathway Prediction

eP



- 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

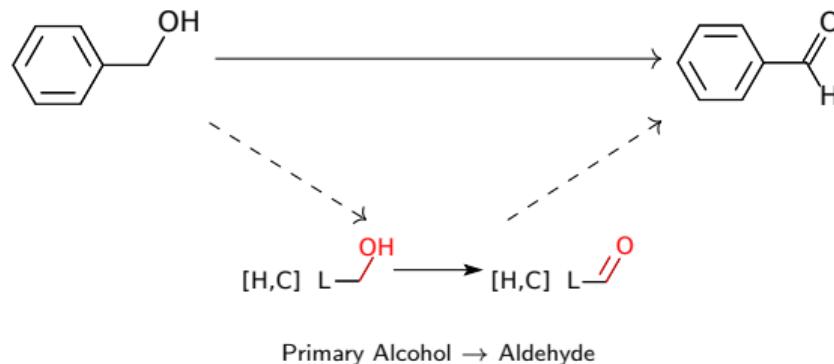
eP



- 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

Benzaldehyde



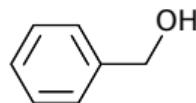
Rule-Based Biodegradation Pathway Prediction

eP

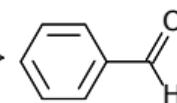


- Rule-based systems predict degradation products using transformation rules
 - Generalizations and abstractions of observed reactions
 - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side

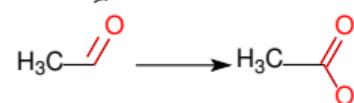
Benzyl Alcohol



Benzaldehyde



Primary Alcohol → Aldehyde



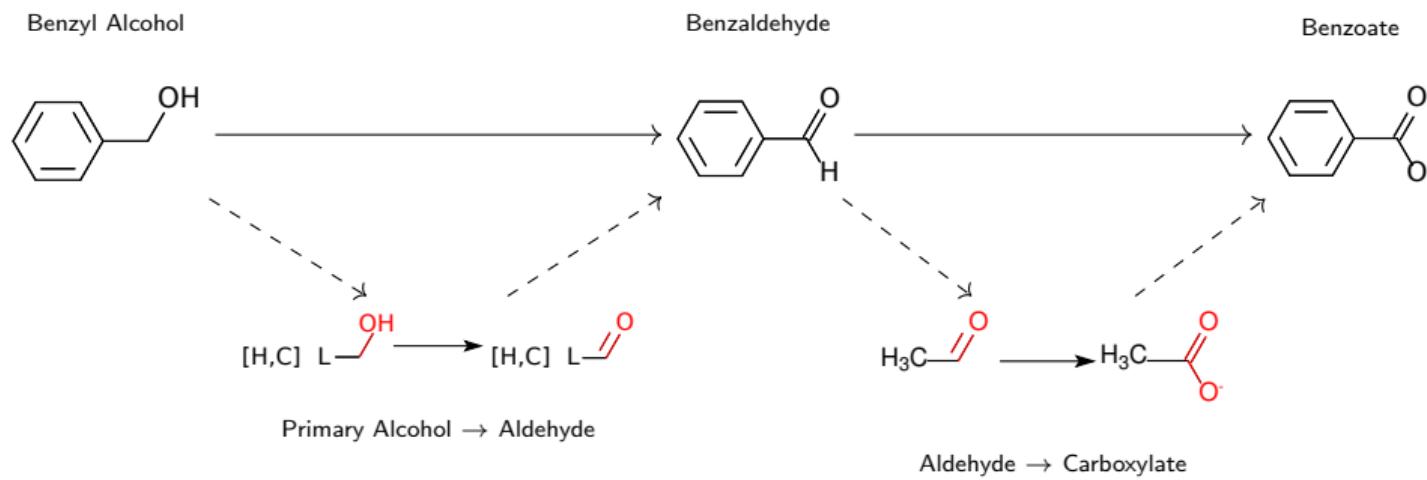
Aldehyde → Carboxylate

Rule-Based Biodegradation Pathway Prediction

eP

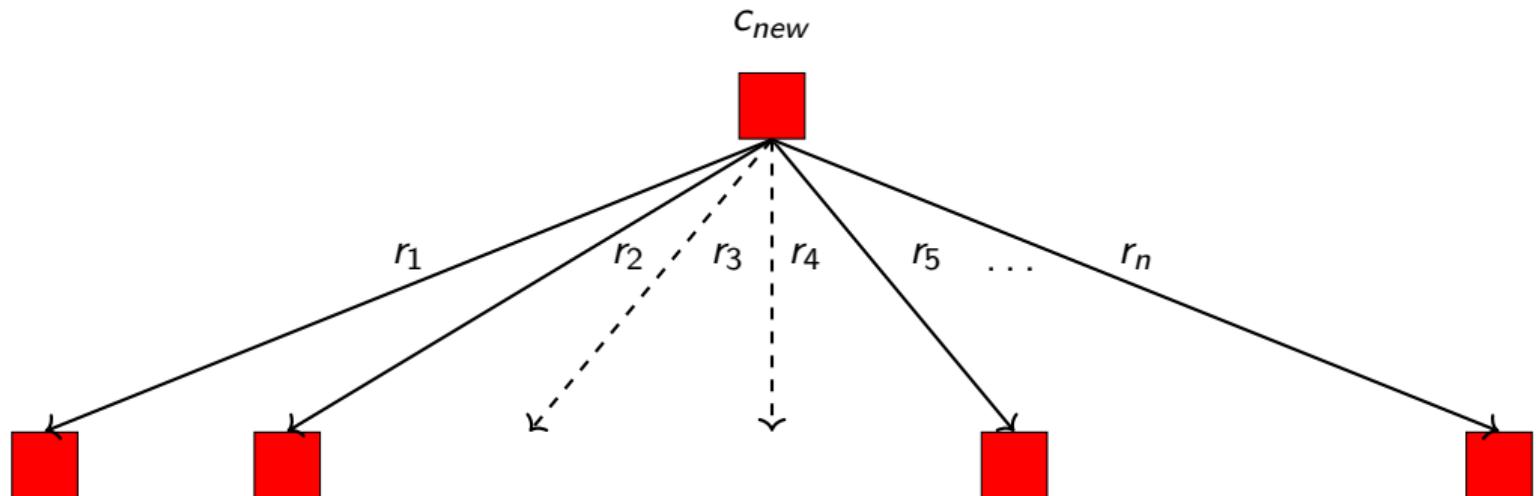


- Rule-based systems predict degradation products using transformation rules
 - Generalizations and abstractions of observed reactions
 - If rule matches certain functional groups on left-hand side, transformation of structure according to right-hand side



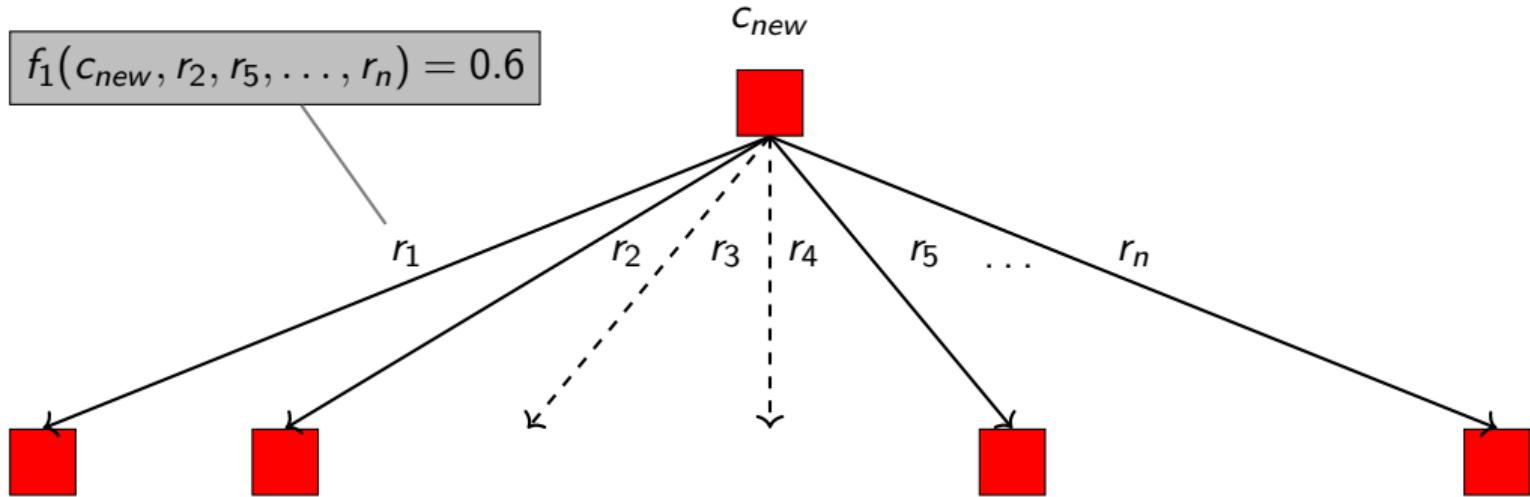
Machine Learning to Limit Combinatorial Explosion

eP



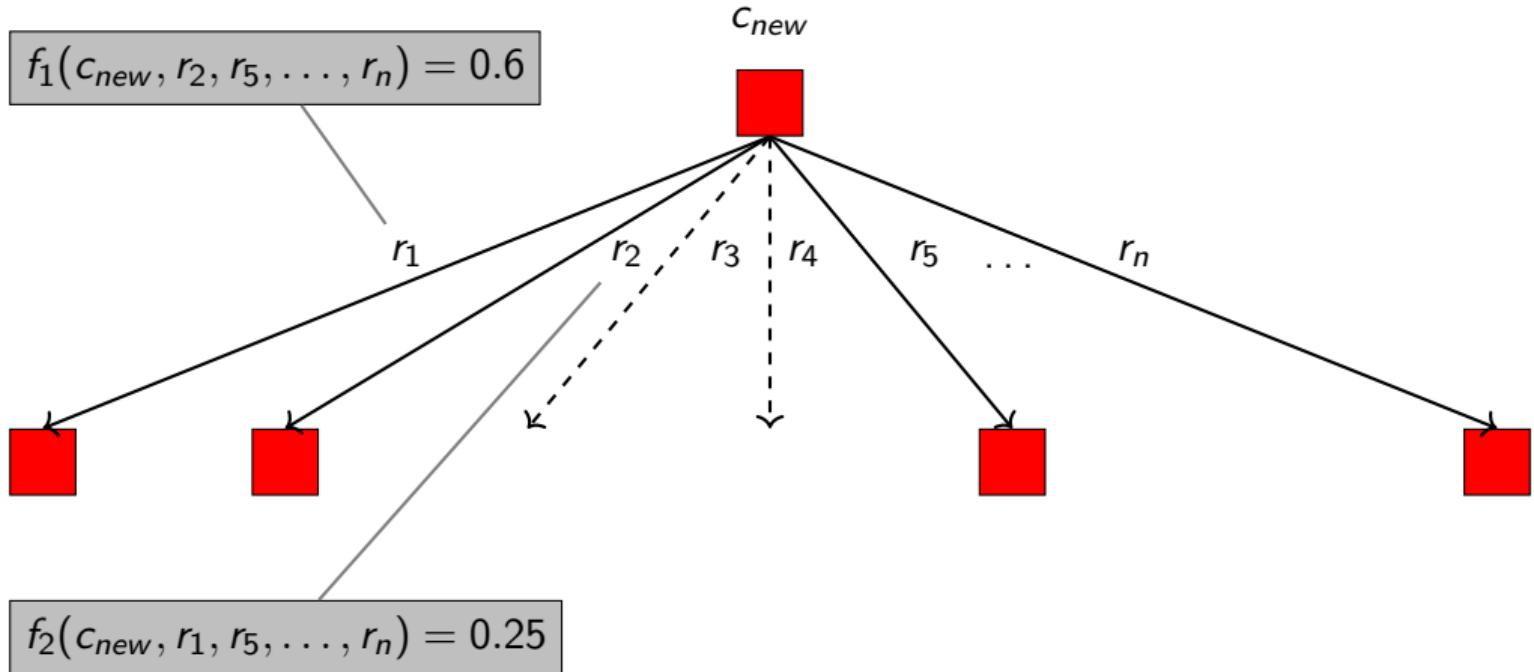
Machine Learning to Limit Combinatorial Explosion

eP



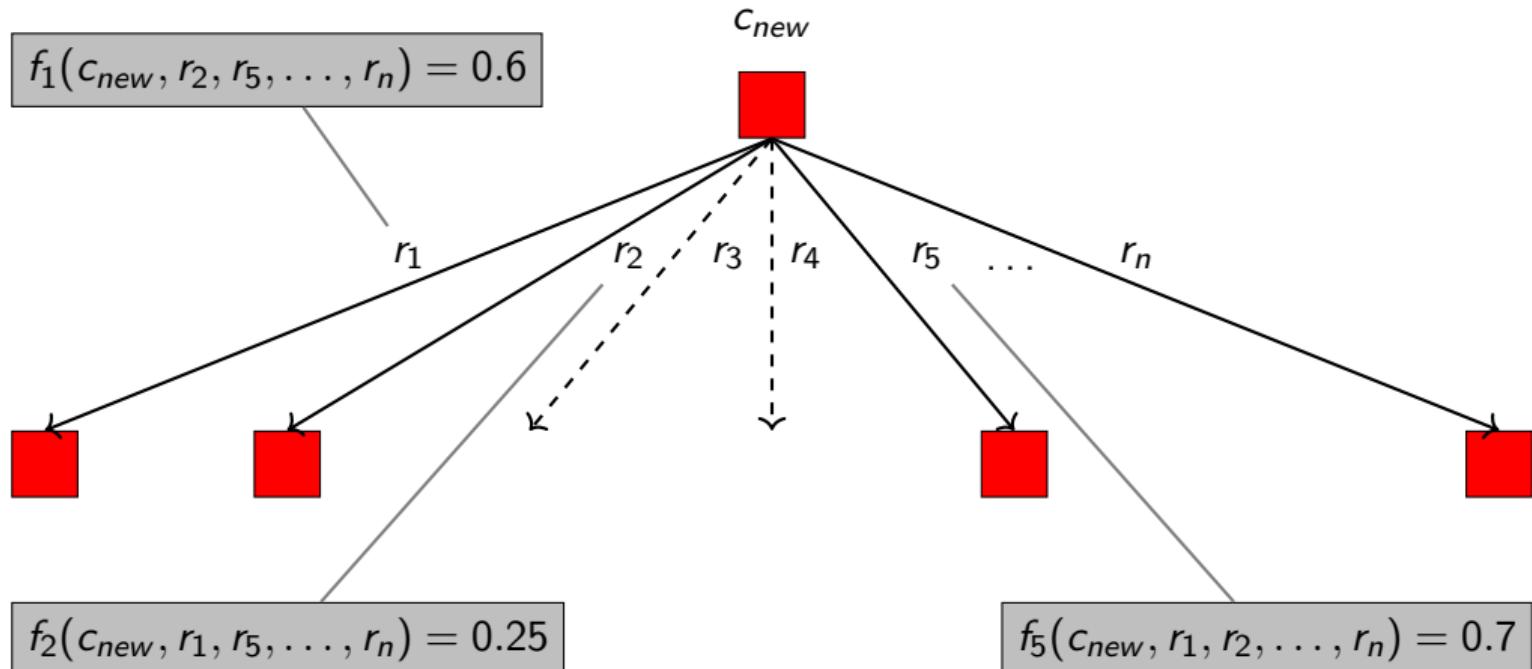
Machine Learning to Limit Combinatorial Explosion

eP



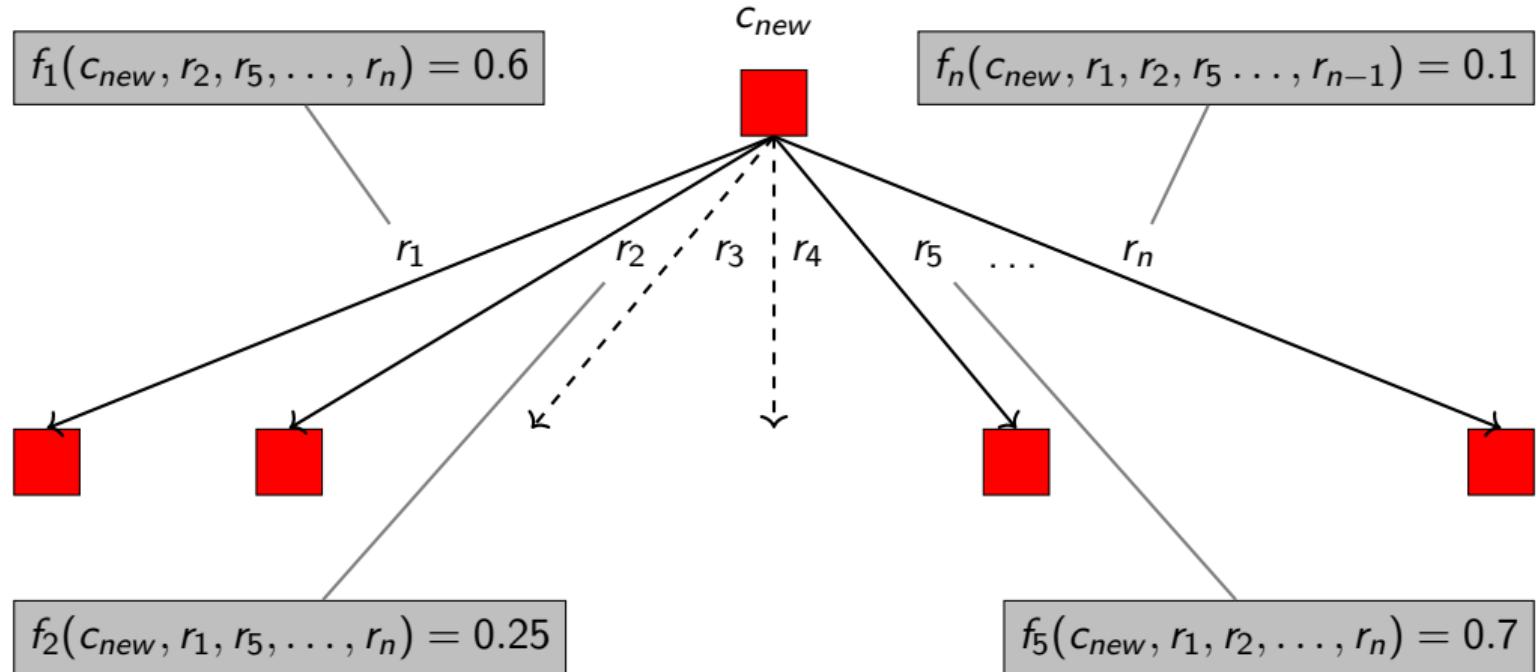
Machine Learning to Limit Combinatorial Explosion

EP



Machine Learning to Limit Combinatorial Explosion

EP



Structural Features – Fingerprints

- Machine Learning algorithms work with tabular data

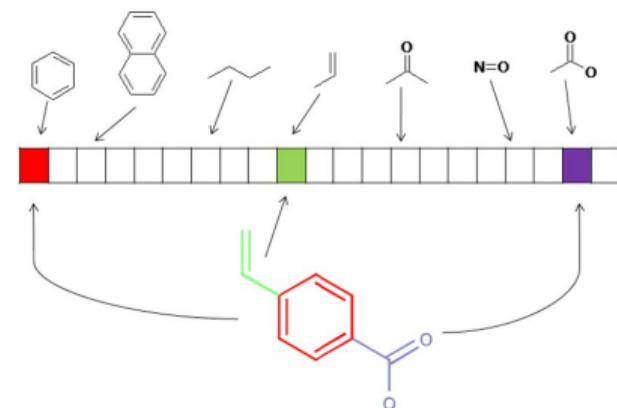


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

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

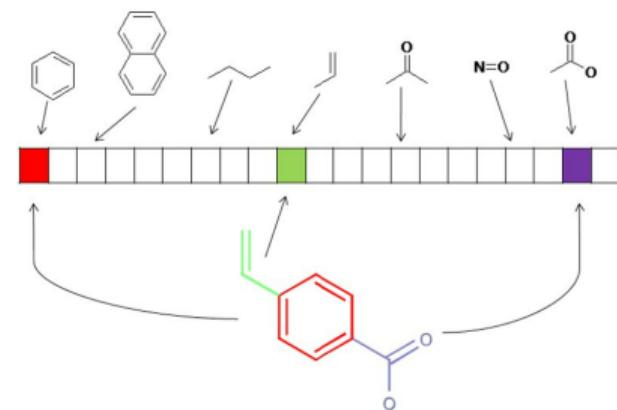


image from chemopy manual, Cao *et al.* (2012)

Structural Features – Fingerprints

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

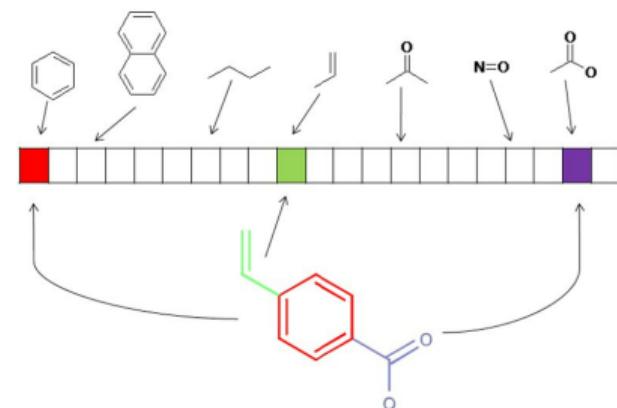


image from chemopy manual, Cao *et al.* (2012)

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, ...

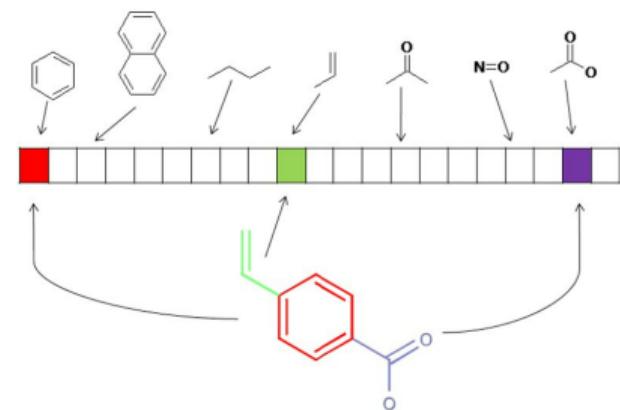


image from chemopy manual, Cao *et al.* (2012)

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

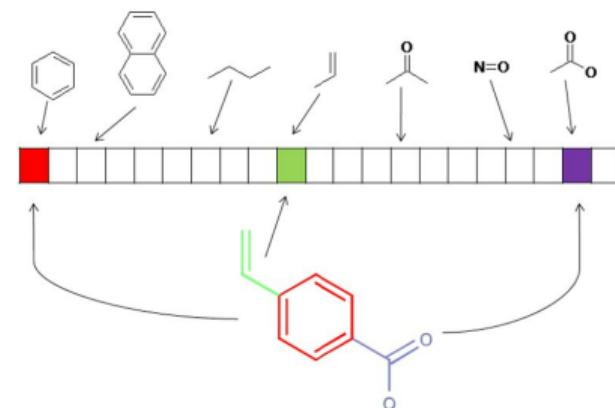


image from chemopy manual, Cao *et al.* (2012)

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

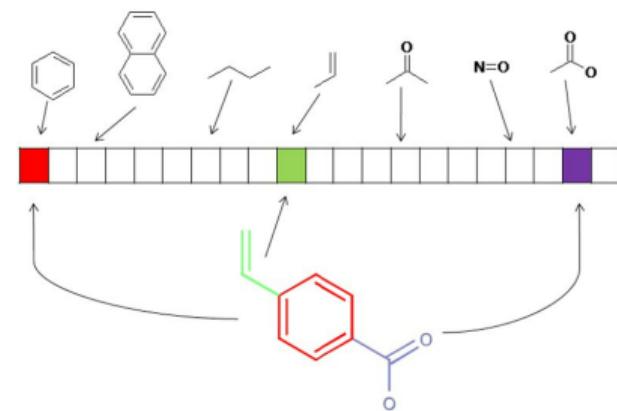


image from chemopy manual, Cao *et al.* (2012)

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

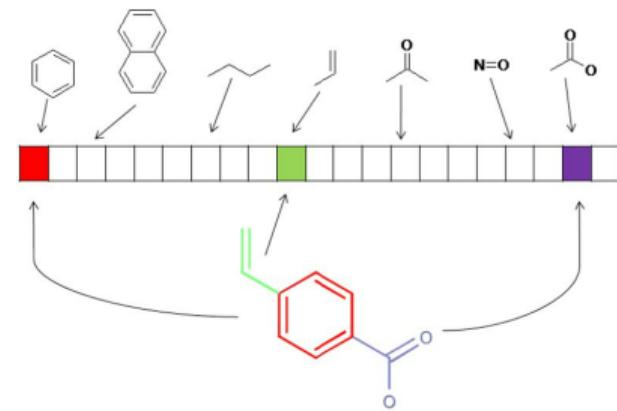
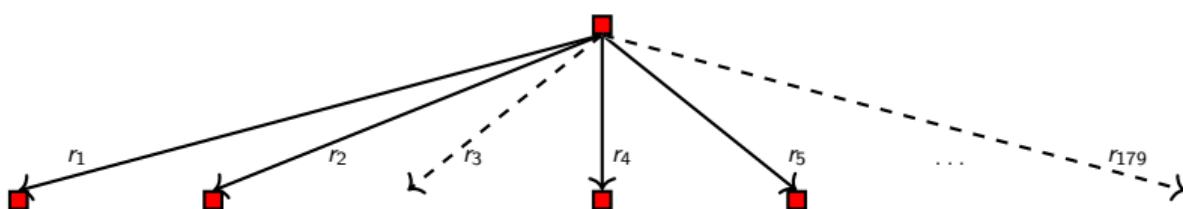
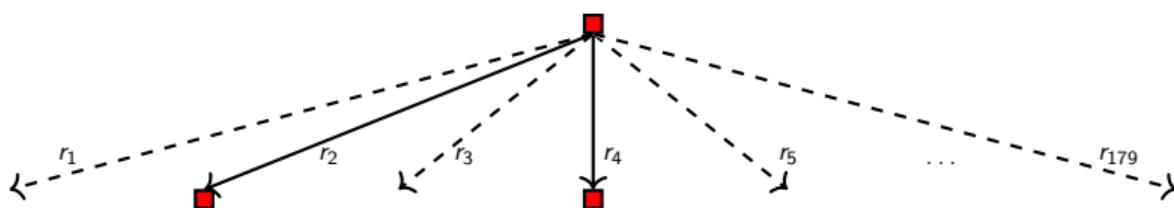
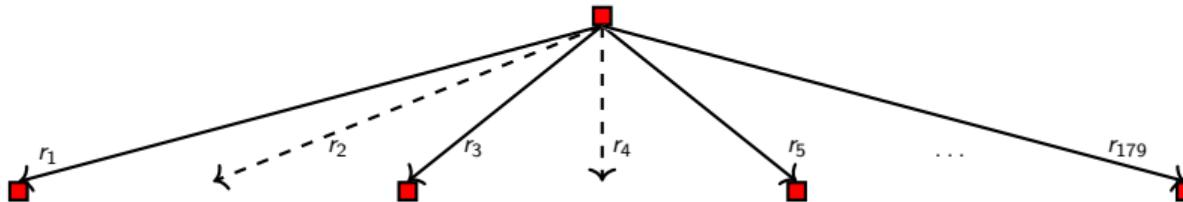


image from chemopy manual, Cao *et al.* (2012)

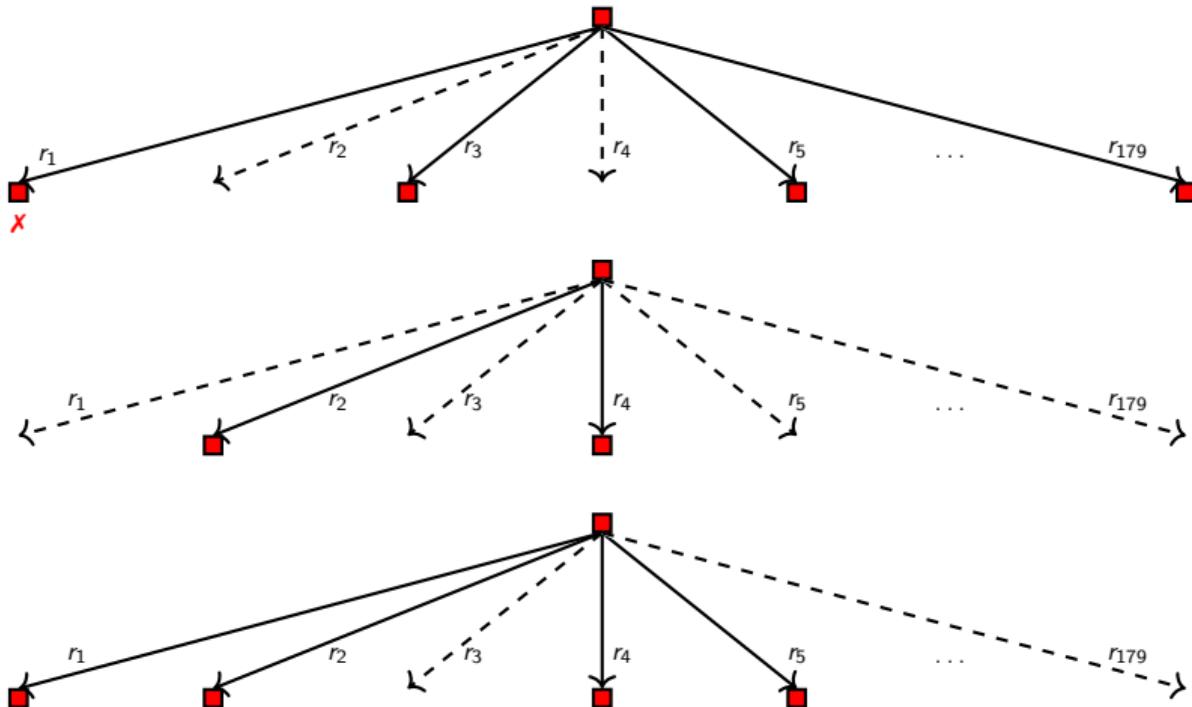
Machine Learning to Limit Combinatorial Explosion

eP



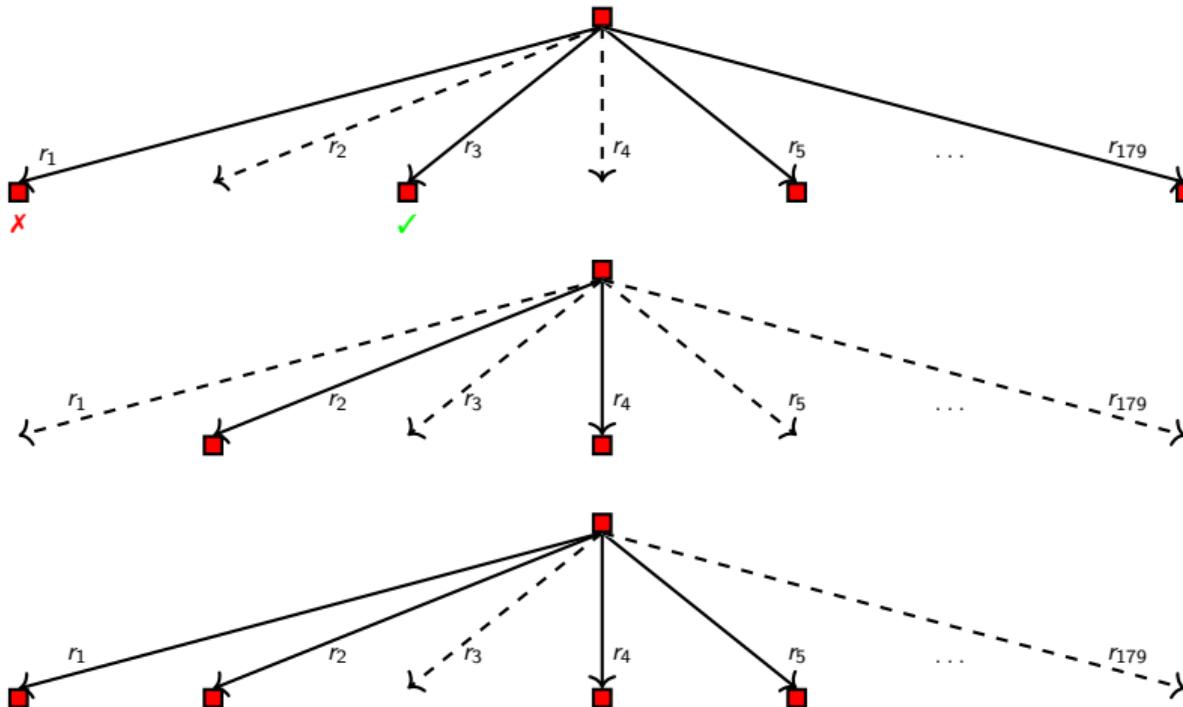
Machine Learning to Limit Combinatorial Explosion

eP



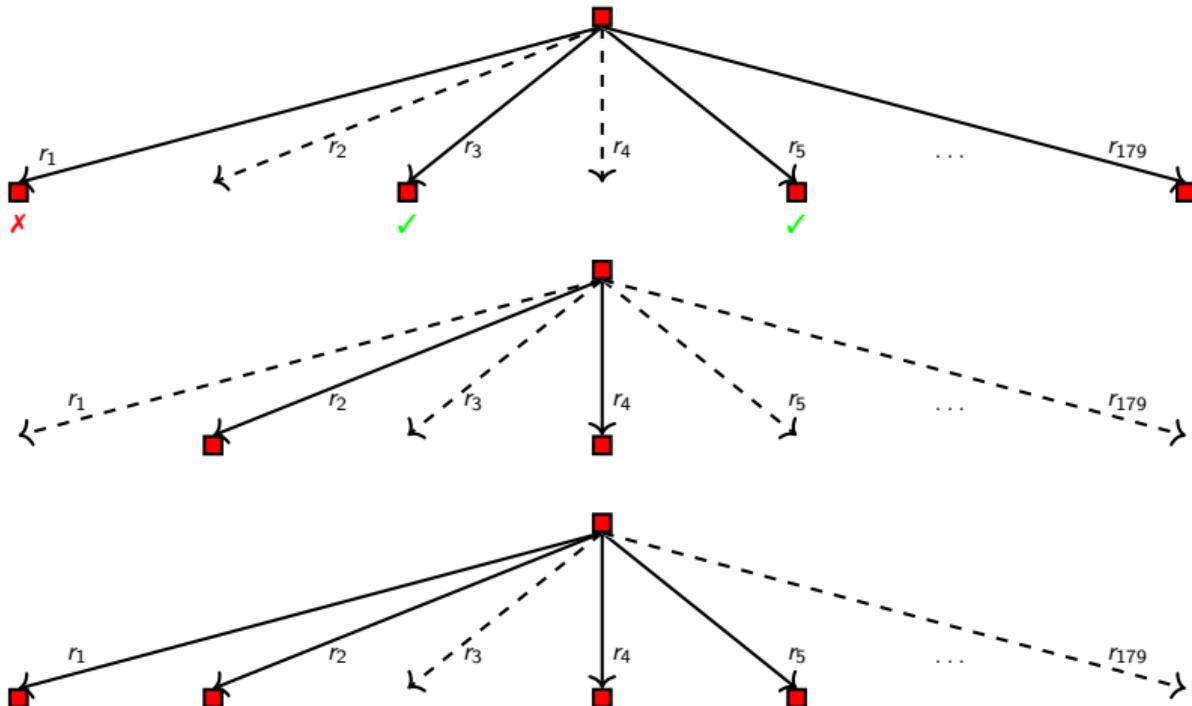
Machine Learning to Limit Combinatorial Explosion

eP



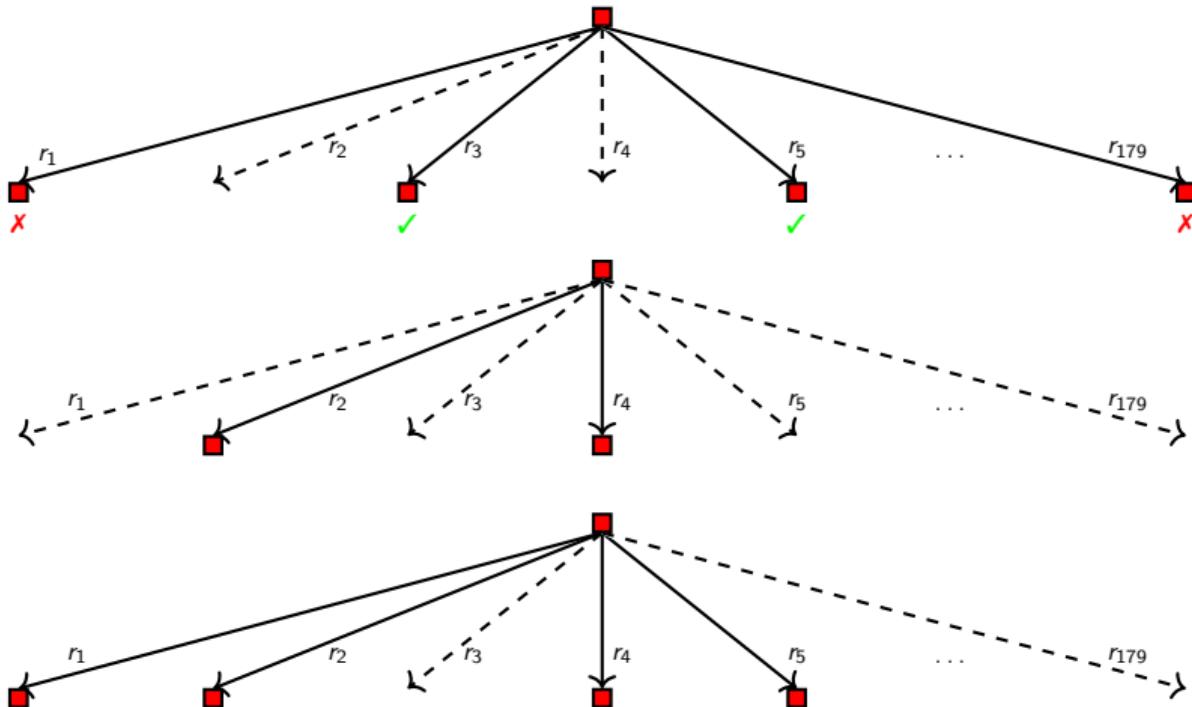
Machine Learning to Limit Combinatorial Explosion

eP



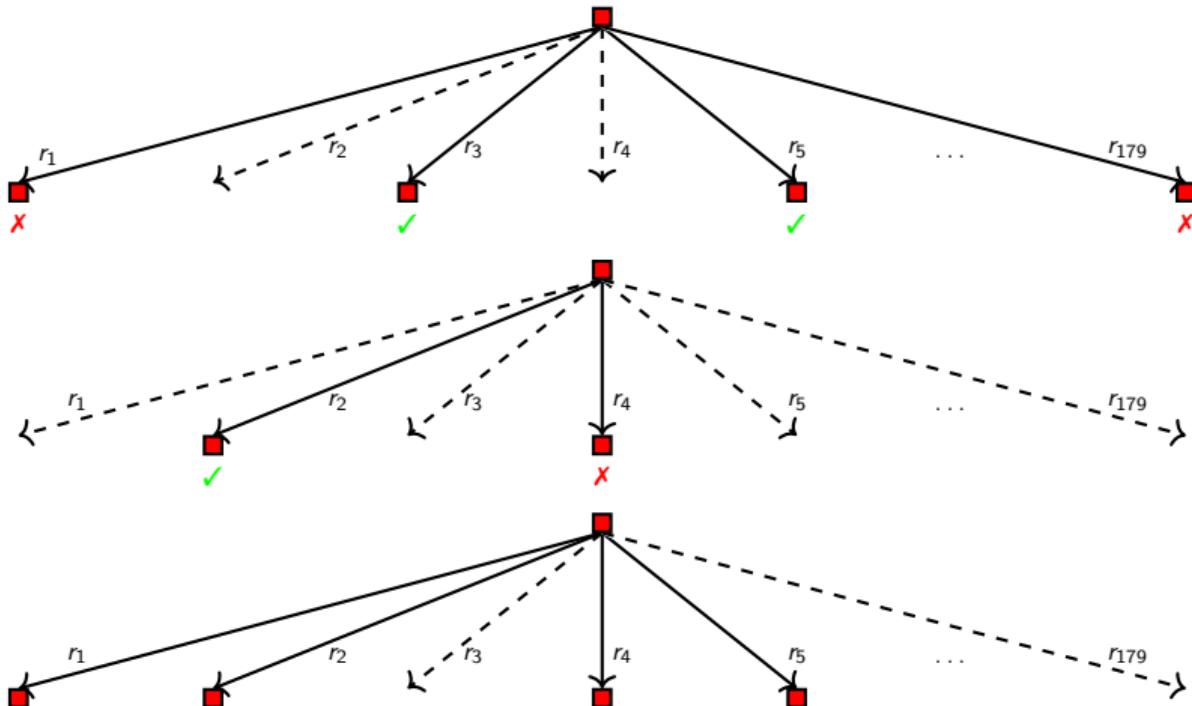
Machine Learning to Limit Combinatorial Explosion

eP



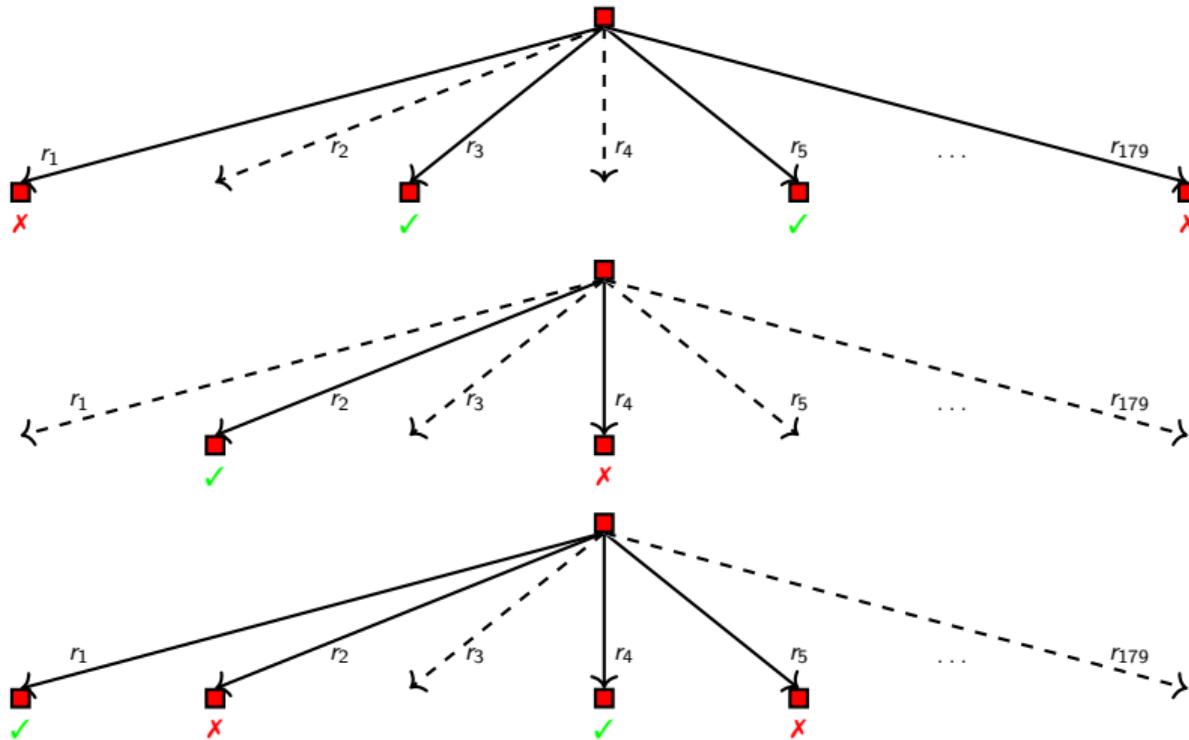
Machine Learning to Limit Combinatorial Explosion

eP

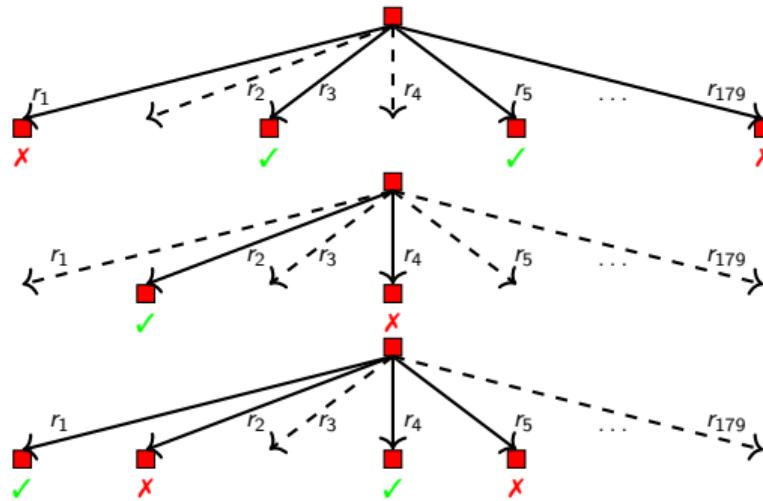


Machine Learning to Limit Combinatorial Explosion

eP

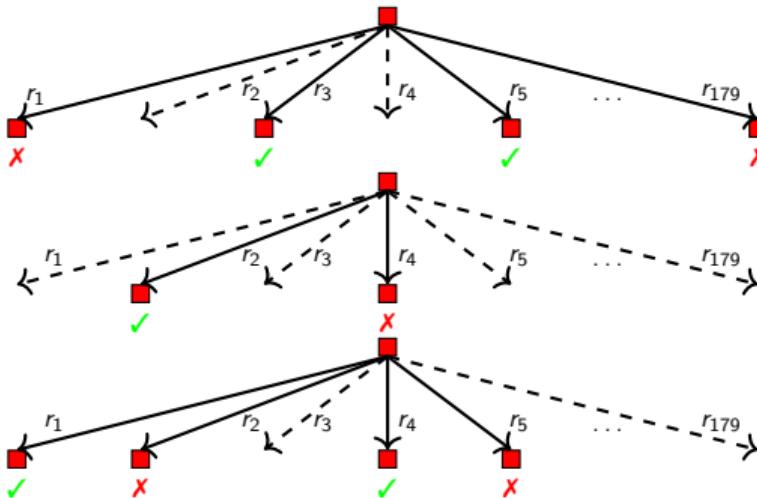


Generation of Training Sets



	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

eP

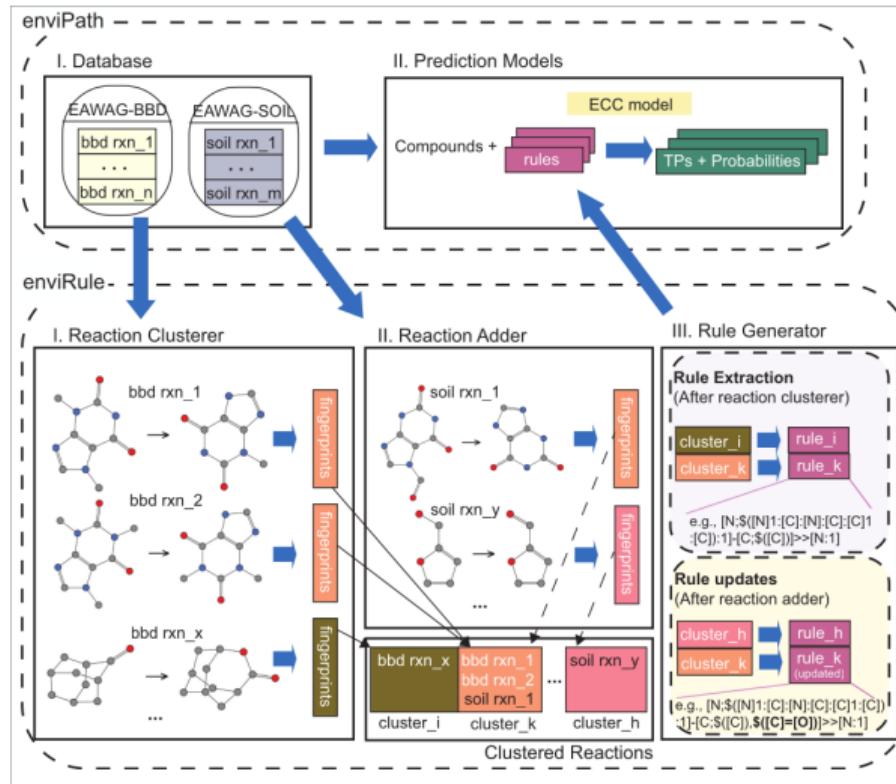


- Learning task is a multi-label problem
- Transformation to multi-label data set using pseudo-label

	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

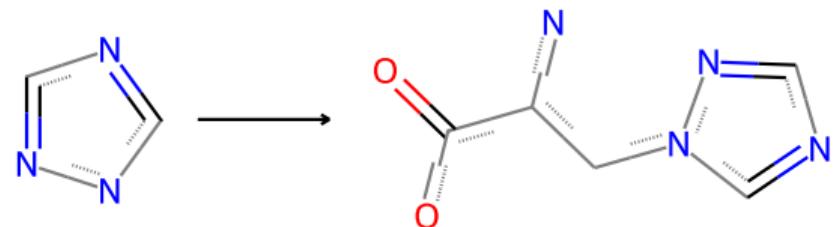


enviFormer – Biodegradation Prediction as Sequence-to-Sequence Task

eP



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

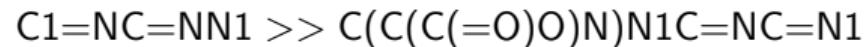
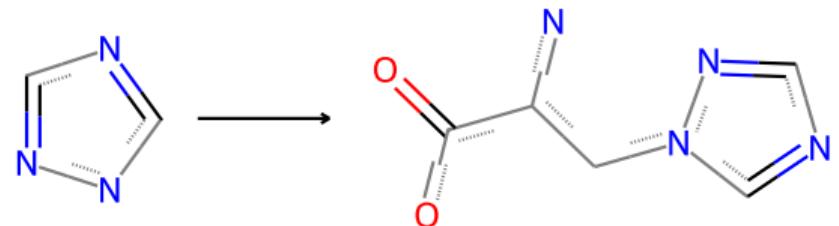


enviFormer – Biodegradation Prediction as Sequence-to-Sequence Task

eP



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

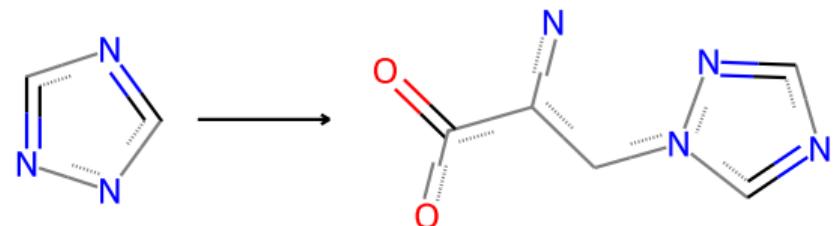


enviFormer – Biodegradation Prediction as Sequence-to-Sequence Task

eP



- 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)



enviFormer – Training Set



features class

enviFormer – Training Set



features class

CCC(0)CO . CCC1C01>>CCC(0)COCC(0)CC

enviFormer – Training Set



features class

CCC(0)C0 . CCC1C01>>CCC(0)C0CC(0)CC

CCC(0)C0 . CCC1C01>>CCC(0)C0CC(0)C C

enviFormer – Training Set



features class

CCC(0)CO.CCC1C01>>CCC(0)COCC(0)CC

CCC(0)CO.CCC1C01>>CCC(0)COCC(0)C C

CCC(0)CO.CCC1C01>>CCC(0)COCC(0) C

enviFormer – Training Set



features	class
CCC(0)CO.CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)CO.CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)CO.CCC1C01>>CCC(0)COCC(0) C	
CCC(0)CO.CCC1C01>>CCC(0)COCC(0)	

enviFormer – Training Set



features	class
CCC(0)CO.CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)CO.CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)CO.CCC1C01>>CCC(0)COCC(0) C	
CCC(0)CO.CCC1C01>>CCC(0)COCC(0)	
CCC(0)CO.CCC1C01>>CCC(0)COCC(0	

enviFormer – Training Set



features	class
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0) C	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0	
CCC(0)CO . CCC1C01>>CCC(0)COCC (0	

enviFormer – Training Set

features	class
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0) C	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0	
CCC(0)CO . CCC1C01>>CCC(0)COCC (
CCC(0)CO . CCC1C01>>CCC(0)COC C	

enviFormer – Training Set

features	class
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0) C	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0)	
CCC(0)CO . CCC1C01>>CCC(0)COCC(0	
CCC(0)CO . CCC1C01>>CCC(0)COCC (
CCC(0)CO . CCC1C01>>CCC(0)COC C	
CCC(0)CO . CCC1C01>>CCC(0)CO C	

enviFormer – Training Set

features	class
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0	
CCC(0)C0.CCC1C01>>CCC(0)COCC (
CCC(0)C0.CCC1C01>>CCC(0)COC C	
CCC(0)C0.CCC1C01>>CCC(0)CO C	
CCC(0)C0.CCC1C01>>CCC(0)C O	

enviFormer – Training Set

features	class
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0	
CCC(0)C0.CCC1C01>>CCC(0)COCC (
CCC(0)C0.CCC1C01>>CCC(0)COC C	
CCC(0)C0.CCC1C01>>CCC(0)CO C	
CCC(0)C0.CCC1C01>>CCC(0)C O	
CCC(0)C0.CCC1C01>>CCC(0) C	

enviFormer – Training Set

features	class
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0	
CCC(0)C0.CCC1C01>>CCC(0)COCC (
CCC(0)C0.CCC1C01>>CCC(0)COC C	
CCC(0)C0.CCC1C01>>CCC(0)CO C	
CCC(0)C0.CCC1C01>>CCC(0)C O	
CCC(0)C0.CCC1C01>>CCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)	

enviFormer – Training Set

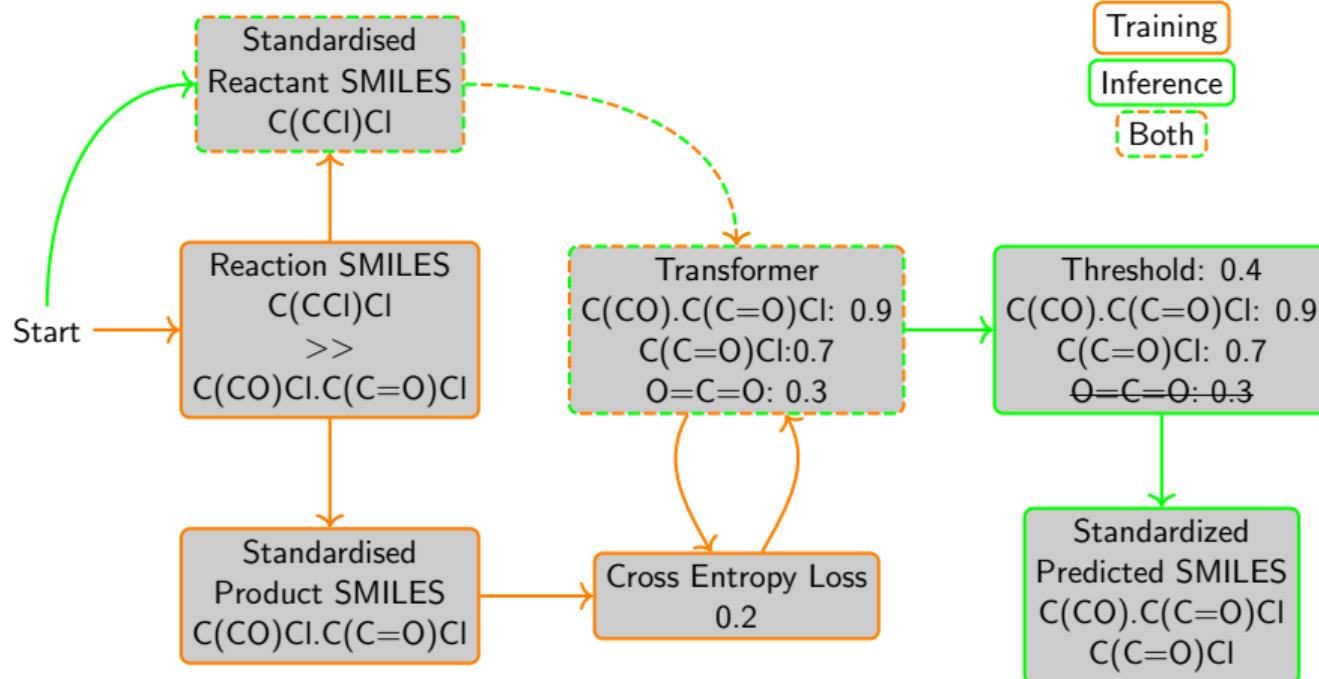
features	class
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0	
CCC(0)C0.CCC1C01>>CCC(0)COCC (
CCC(0)C0.CCC1C01>>CCC(0)COC C	
CCC(0)C0.CCC1C01>>CCC(0)CO C	
CCC(0)C0.CCC1C01>>CCC(0)C O	
CCC(0)C0.CCC1C01>>CCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)	
CCC(0)C0.CCC1C01>>CCC(0	

enviFormer – Training Set

features	class
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)CC	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)C C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0)	
CCC(0)C0.CCC1C01>>CCC(0)COCC(0	
CCC(0)C0.CCC1C01>>CCC(0)COCC (
CCC(0)C0.CCC1C01>>CCC(0)COC C	
CCC(0)C0.CCC1C01>>CCC(0)CO C	
CCC(0)C0.CCC1C01>>CCC(0)C O	
CCC(0)C0.CCC1C01>>CCC(0) C	
CCC(0)C0.CCC1C01>>CCC(0)	
CCC(0)C0.CCC1C01>>CCC(0	

...

enviFormer – Method



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

- 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 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 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 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 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>



The screenshot displays the enviPath website interface. At the top, there's a navigation bar with links for Home, Help, Pathway, Compound, Predictor, Reaction Screening, Search, Sort, User Group, Search, Log in, and Log out. Below the navigation is a banner for "enviPath BIOTRANSFORMATION PATHWAY RESOURCE". The main content area includes:

- enviPath**: A brief introduction stating it's a database and prediction system for microbial biotransformation of organic environmental contaminants.
- SERVICES**: A section listing services such as "New Data Package and News Package", "Envipath Predictor", "Envipath Reaction Screening", and "Envipath Help".
- News**: A news package from 2022-02-01-0000 about Envipath's new package, Envipath Predictor, which includes a reaction screening and degradation studies, aimed at predicting the most likely pathway for a target pathway. It also includes information from possible degradation studies (40 publications) and a new predictor module available through the European Food Safety Authority's ECHA REACH database. It contains information on different experimental approaches... [1].
- Help**: A section titled "Just before Christmas, we released a new version of envipath at https://envipath.org. We have completely rewritten the code base and interface, a much faster now and will be easier to use. We have added many new features and data. We fixed many problems and the performance is much better. We hope you like it. To import more data from Envipath Predictor, Only a small [...]".
- Wiki**: A link to the documentation page.

Thank you for listening! Any questions?

<https://envipath.org>

<https://envipath.com>

<https://wickerlab.org>