

May 10, 2025

Ongoing developments of enviPath

Applicability Domain, Evaluation

Jörg Simon Wicker

enviPath / School of Computer Science – University of Auckland

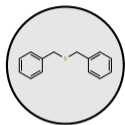
enviPath



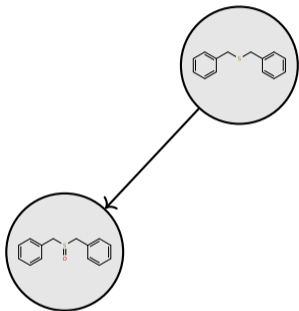
SCIENCE

Biodegradation Pathways

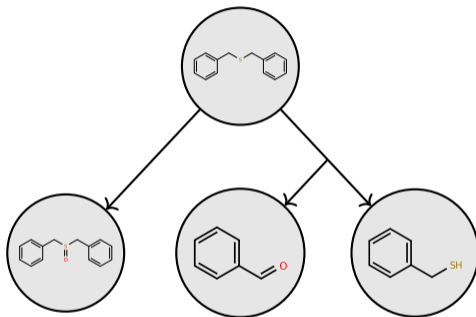
EP



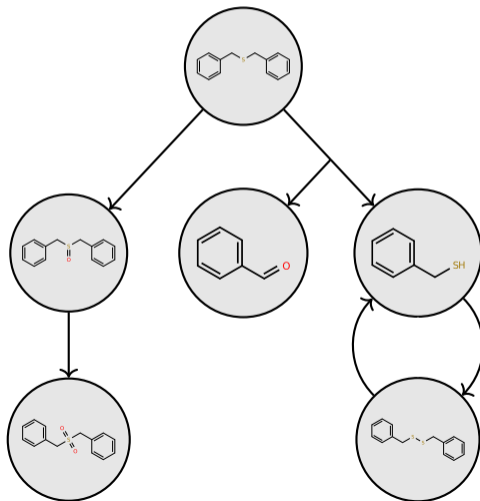
Biodegradation Pathways



Biodegradation Pathways

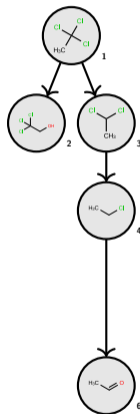


Biodegradation Pathways

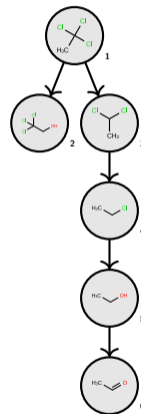


Evaluation

- Traditionally, evaluation estimates if reaction predictions of a compound are likely to be correct



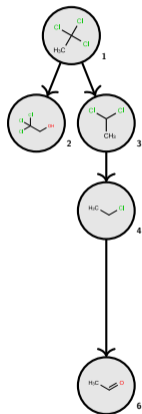
Ground Truth



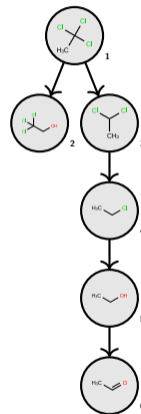
Prediction

Evaluation

- Traditionally, evaluation estimates if reaction predictions of a compound are likely to be correct
- Initially, we evaluated the performance only on a reaction level



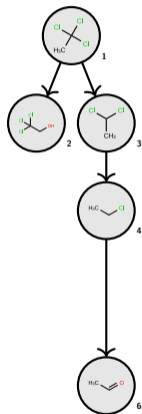
Ground Truth



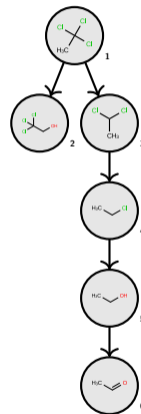
Prediction

Evaluation

- Traditionally, evaluation estimates if reaction predictions of a compound are likely to be correct
- Initially, we evaluated the performance only on a reaction level
- This ignores problems with intermediates and effects in predictions of the children of nodes



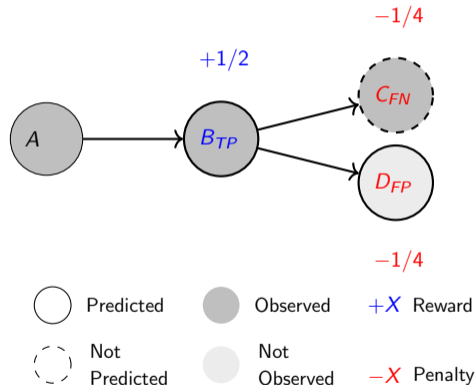
Ground Truth



Prediction

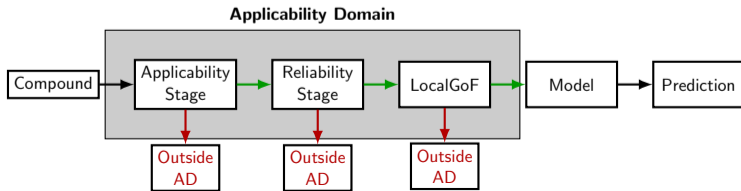
Evaluation

- Traditionally, evaluation estimates if reaction predictions of a compound are likely to be correct
- Initially, we evaluated the performance only on a reaction level
- This ignores problems with intermediates and effects in predictions of the children of nodes

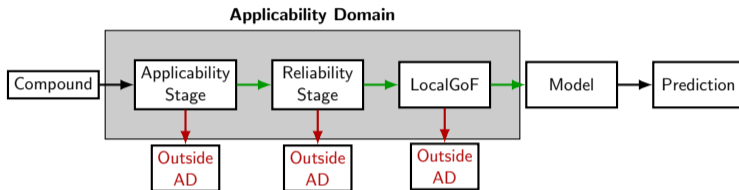


Limitations – Applicability Domain

- The chemical space is huge, yet we train our models on only hundreds of compounds
- Our model does not “know” enough about all compounds
- We use the applicability domain of a model to estimate if we can use our model on a given compound
- There are interesting connections between the concept of applicability domain and adversarial attacks on machine learning models
- We introduced a 3-stage approach in enviPath based on applicability, reliability, and local goodness of fit



Limitations – Applicability Domain



Applicability

- Bounding Box in PCA space

Reliability

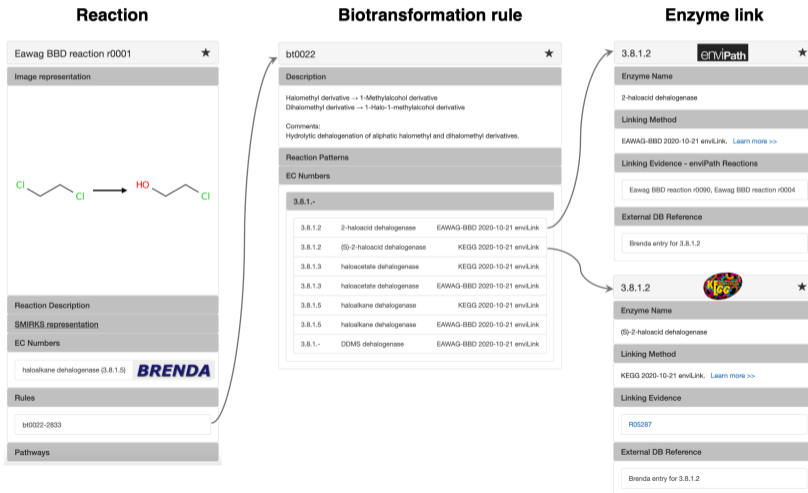
- Use qualified neighbours - instances in training set that share possible reactions (based on rules)
- Score calculates similarity to qualified neighbours

Local Goodness of Fit

- Critical Success Index across qualified neighbours

$$CSI = \frac{\#TP}{\#TP + \#FP + \#FN}$$

enviLink – linking Enzymes



Future modeling – Mixture of Experts



- Build ensemble of models, each an expert for enzyme classes

Future modeling – Mixture of Experts

EP



- Build ensemble of models, each an expert for enzyme classes
- Learn the weights for each input compound

Future modeling – Mixture of Experts



- Build ensemble of models, each an expert for enzyme classes
- Learn the weights for each input compound
- Prediction is a combination of enzyme-‘expert’ model predictions

Future modeling – Mixture of Experts



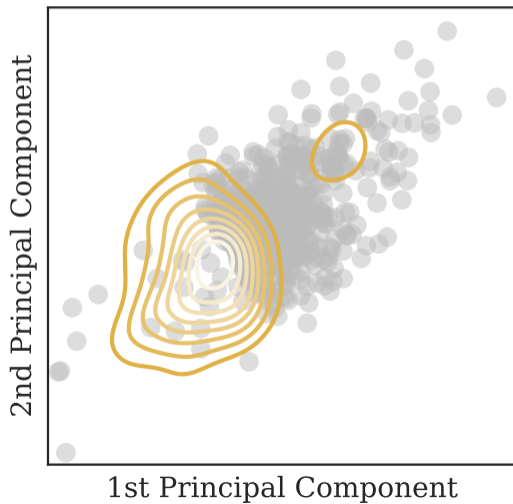
- Build ensemble of models, each an expert for enzyme classes
- Learn the weights for each input compound
- Prediction is a combination of enzyme-‘expert” model predictions
- Significantly improves model performance

Future modeling – Mixture of Experts

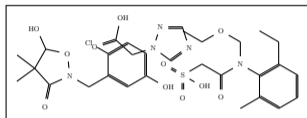


- Build ensemble of models, each an expert for enzyme classes
- Learn the weights for each input compound
- Prediction is a combination of enzyme-‘expert” model predictions
- Significantly improves model performance
- To be published mid-2025

Bias

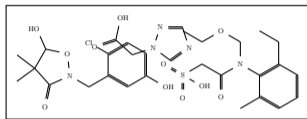


Mitigating Bias – Cancels



Chemical Compound Dataset

Mitigating Bias – Cancels



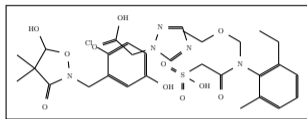
Chemical Compound Dataset



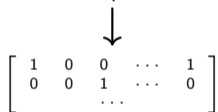
$$\begin{bmatrix} 1 & 0 & 0 & \dots & 1 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & & & & \end{bmatrix}$$

MACCS

Mitigating Bias – Cancels

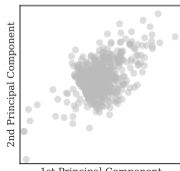


Chemical Compound Dataset

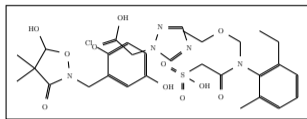


MACCS

Principal Component
Analysis (PCA) ↓



Mitigating Bias – Cancels



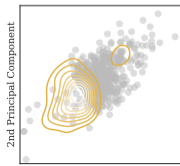
Chemical Compound Dataset



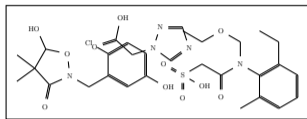
$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \end{bmatrix}$$

MACCS

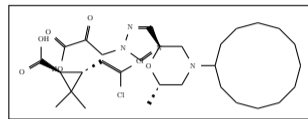
Principal Component Analysis (PCA) ↓ + IMITATE



Mitigating Bias – Cancels



Chemical Compound Dataset

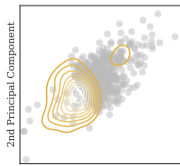


Pool of Candidate Compounds

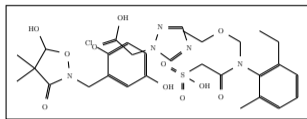
$$\begin{bmatrix} 1 & 0 & 0 & \dots & 1 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & & & & \end{bmatrix}$$

MACCS

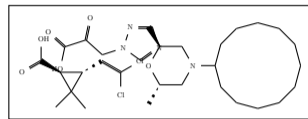
Principal Component Analysis (PCA) ↓ + IMITATE



Mitigating Bias – Cancels



Chemical Compound Dataset

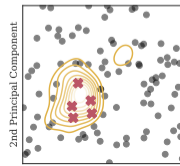
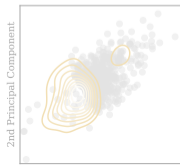


Pool of Candidate Compounds

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 1 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & & & & \end{bmatrix}$$

MACCS

Principal Component Analysis (PCA) + IMITATE



enviPath – The Environmental Contaminant Biotransformation Pathway Resource

EP



- Successor of UM-BBD/PPS (Eawag-BBD/PPS)

The screenshot displays the enviPath website interface. At the top, there is a navigation bar with the following links: Home, Pathway, Data, Compound, Prediction, Results, Reporting, Support, About, Help, Contact. A search bar and a user profile icon are also present. The main content area features the title "enviPath | THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE" and a descriptive paragraph: "enviPath is a database and prediction system for the microbial biotransformation of organic environmental contaminants. The database provides the possibility to store and view experimentally observed biotransformation pathways. The pathway prediction system provides different relative reasoning models to predict likely biotransformation pathways and products. You can try it out below." Below this is a "Learn more" button. A search bar contains the text "SHE25" and a "Go" button. The sidebar on the left contains three panels: "Twitter" with a tweet from @envipath, "News" with two articles titled "New Data Package and Publication" and "New Release", and "Latest Pathways" with a diagram and the title "enviPath 2.0".

enviPath – The Environmental Contaminant Biotransformation Pathway Resource

EP



- Successor of UM-BBD/PPS (Eawag-BBD/PPS)
- Initial release September 2015

The screenshot shows the enviPath website interface. At the top is a dark navigation bar with the 'enviPath' logo and menu items: 'Home', 'About', 'Data', 'Concepts', 'Predictions', 'Research', 'Support', 'Contact', 'Help', 'Privacy', 'Search', and 'Log Out'. Below the navigation bar is a large white banner with the text: 'enviPath | THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE'. Underneath the banner, a paragraph describes the resource: 'enviPath is a database and prediction system for the microbial biotransformation of organic environmental contaminants. The database provides the possibility to store and view experimentally observed biotransformation pathways. The pathway prediction system provides different relative reasoning models to predict likely biotransformation pathways and products. You can try it out below.' A 'Learn more' button is located at the bottom of the banner. Below the banner is a search bar containing the text 'SHE23' and a 'Go' button. The main content area is divided into three columns. The left column is titled 'Twitter' and contains a tweet from 'enviPath Resources' (@envipath) dated 02/10/2017 12:00:29 +0000, which mentions 'Eawag enviPath: new resource for exploring reported pesticide biotransformation pathways and their data'. The middle column is titled 'News' and contains two articles: 'New Data Package and Publication' dated 02/10/2017 12:00:29 +0000, which describes a new package of pathway information from soil degradation studies, and 'New Release' dated Tue, 02 Jan 2017 11:14:29 +0000, which announces a new version of the website. The right column is titled 'Latest Pathways' and shows a '2.4.0' version of a pathway diagram with a 'View' button and a description: 'The main documentation can be found in our site. We will add an interactive view and further documentation soon.'

enviPath – The Environmental Contaminant Biotransformation Pathway Resource

EP



- Successor of UM-BBD/PPS (Eawag-BBD/PPS)
- Initial release September 2015
- Database and prediction system for the microbial biotransformation of organic environmental contaminants

The screenshot displays the enviPath website interface. At the top, there is a navigation bar with links for 'Home', 'About', 'Contact', 'Help', 'FAQ', 'Privacy Policy', 'Terms of Use', and 'Login'. The main content area features the 'enviPath' logo and the subtitle 'THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE'. Below this, a brief description states: 'enviPath is a database and prediction system for the microbial biotransformation of organic environmental contaminants. The database provides the possibility to store and view experimentally observed biotransformation pathways. The pathway prediction system provides different relative reasoning models to predict likely biotransformation pathways and products. You can try it out below.' A search bar is located below the main text. The sidebar on the left contains a 'Twitter' section with a tweet from 'enviPath Resources' and a 'News' section with two articles: 'New Data Package and Publication' and 'New Release'. The 'Latest Pathway' section on the right shows a chemical reaction diagram with a central node and several branches.

enviPath – The Environmental Contaminant Biotransformation Pathway Resource

EP



- Successor of UM-BBD/PPS (Eawag-BBD/PPS)
- Initial release September 2015
- Database and prediction system for the microbial biotransformation of organic environmental contaminants
- Database provides the possibility to store and view experimentally observed biotransformation pathways

The screenshot displays the enviPath website interface. At the top, there is a navigation bar with tabs for 'enviPath', 'Database', 'Prediction', 'API', 'Contact', 'About', 'News', 'Help', and 'Log out'. Below the navigation bar is a search bar with the text 'Search' and a magnifying glass icon. The main content area features the 'enviPath' logo and the title 'THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE'. A brief description follows: 'enviPath is a database and prediction system for the microbial biotransformation of organic environmental contaminants. The database provides the possibility to store and view experimentally observed biotransformation pathways. The pathway prediction system provides different relative reasoning models to predict likely biotransformation pathways and products. You can try it out below.' Below this is a 'Learn more' button. A search bar with the text 'SEARCH' and a magnifying glass icon is also present. The main content area is divided into three columns: 'Twitter', 'News', and 'Latest Pathways'. The 'Twitter' column shows a tweet from @envipath. The 'News' column contains two news items: 'New Data Package and Publication' and 'New Release'. The 'Latest Pathways' column shows a diagram of a biotransformation pathway with the text '2.4.0' and 'enviPath 2.0'.

enviPath – The Environmental Contaminant Biotransformation Pathway Resource

EP



- Successor of UM-BBD/PPS (Eawag-BBD/PPS)
- Initial release September 2015
- Database and prediction system for the microbial biotransformation of organic environmental contaminants
- Database provides the possibility to store and view experimentally observed biotransformation pathways
- Pathway prediction system provides different relative reasoning models to predict likely biotransformation pathways and products

The screenshot displays the enviPath website interface. At the top, there is a navigation bar with tabs for 'Home', 'About', 'How to use', 'Contact', 'Help', 'FAQ', 'Privacy Policy', 'Terms', 'Site Map', and 'Login'. Below the navigation bar is a search bar with the text 'enviPath' and a search icon. The main content area features a large heading 'enviPath | THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE' and a brief description of the database and prediction system. Below this, there is a search bar with the text 'enviPath' and a search icon. The interface is divided into three columns: 'Twitter' showing recent tweets, 'News' showing a 'New Data Package and Publication' and a 'New Release', and 'Latest Pathways' showing a '2.0.0' version update and a 'W06' pathway diagram.

enviPath – The Environmental Contaminant Biotransformation Pathway Resource

EP



- Successor of UM-BBD/PPS (Eawag-BBD/PPS)
- Initial release September 2015
- Database and prediction system for the microbial biotransformation of organic environmental contaminants
- Database provides the possibility to store and view experimentally observed biotransformation pathways
- Pathway prediction system provides different relative reasoning models to predict likely biotransformation pathways and products
- Current data sets

The screenshot displays the enviPath website interface. At the top, there is a navigation menu with links for Home, Pathway, Data, Prediction, Reports, Resources, Support, About, and Help. The main header features the enviPath logo and the text "THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE". Below this, a brief description of the database and prediction system is provided, along with a "Learn more" button. A search bar is located below the description. The main content area is divided into three columns: "Twitter" showing recent tweets, "News" containing several announcements such as "New Data Package and Publication" and "New Release", and "Latest Pathways" displaying a 2.4.0 pathway diagram and a "View" button.

enviPath – The Environmental Contaminant Biotransformation Pathway Resource

EP



- Successor of UM-BBD/PPS (Eawag-BBD/PPS)
- Initial release September 2015
- Database and prediction system for the microbial biotransformation of organic environmental contaminants
- Database provides the possibility to store and view experimentally observed biotransformation pathways
- Pathway prediction system provides different relative reasoning models to predict likely biotransformation pathways and products
- Current data sets
 - Eawag BBD, Eawag Soil, Eawag Sludge



Interface



- Web interface available at <https://envipath.org> with possibility to

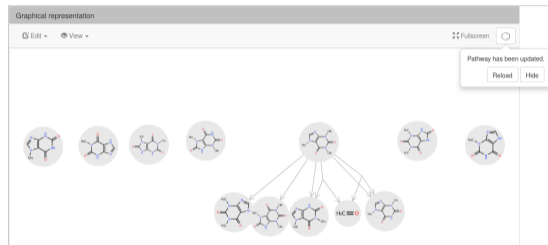
The screenshot shows the top navigation bar of the envipath.org website with links for Package, Pathway, Rule, Compound, Reaction, Relative Reasoning, Scenario, Setting, User, Group, Search, Info, and Login. Below the navigation bar is a large grey banner with the text: "enviPath | THE ENVIRONMENTAL CONTAMINANT BIOTRANSFORMATION PATHWAY RESOURCE". Underneath this banner is a paragraph describing the database and prediction system, followed by a "Learn more >>" button. At the bottom of the banner area is a search bar with the text "SMILES" and a "Go!" button.

This block contains three vertical screenshots. The leftmost screenshot shows a Twitter feed with tweets from @envipath and EnvironmentalScience (@envipath). The middle screenshot shows a "News" section with a "New Data Package and Publication" dated Fri, 24 Feb 2017 19:00:18 +0000, and a "New Release" dated Tue, 03 Jan 2017 15:14:29 +0000. The rightmost screenshot shows a "Latest Pathway" section with a diagram labeled "2,4-D" and the text "enviPath 101".

Interface



- Web interface available at <https://envipath.org> with possibility to
 - Predict new pathways



Interface



- Web interface available at <https://envipath.org> with possibility to
 - Predict new pathways
 - Access database (pathways, reactions, compounds, models)

| Pathways | | Actions ▾ |
|---------------------------------------------------------------------------------------------------------------|--|-----------|
| A pathway displays the (predicted) biodegradation of a compound as graph. Learn more >> | | |
| Reviewed | | |
| (-)-Camphor | | ★ |
| 1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane (DDT) | | ★ |
| 1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane (DDT) (anaerobic) | | ★ |
| 1,1,1-Trichloroethane (an/aerobic) | | ★ |
| 1,2,3-Tribromopropane | | ★ |
| 1,2,4-Trichlorobenzene (an/aerobic) | | ★ |
| 1,2-Dichloroethane | | ★ |
| 1,3-Dichloro-2-propanol | | ★ |
| 1,3-Dichloropropene | | ★ |
| 1,3-Dichloropropene | | ★ |
| 1,4-Dichlorobenzene | | ★ |
| 1,4-Dioxane | | ★ |

Interface



- Web interface available at <https://envipath.org> with possibility to
 - Predict new pathways
 - Access database (pathways, reactions, compounds, models)
 - Enter your own data

New Package ✕

Create new package. Description can be changed later.

Name

Primary Group

Description

Interface

- Web interface available at <https://envipath.org> with possibility to
 - Predict new pathways
 - Access database (pathways, reactions, compounds, models)
 - Enter your own data
 - Set model parameters

1 Select Relative Reasoning

To create a prediction setting, first set the name of the setting. Afterwards, select the Relative Reasoning object to use. If you do not see any relative reasoning object in the list, you have to train one first. Exit this menu and go to Relative Reasoning directly. You can also leave this empty and not use any relative reasoning models. When you choose the object, click Next.

Name

Select Relative Reasoning

Set the cutoff

Or select from the Precision-Recall curve

Evaluation Type

Cancel Back Next

Interface

- Web interface available at <https://envipath.org> with possibility to
 - Predict new pathways
 - Access database (pathways, reactions, compounds, models)
 - Enter your own data
 - Set model parameters
 - Train your own models

2 Advanced Options

Select advanced options for the Relative Reasoning creation. The threshold is used for the predicted probabilities and should be between 0.0 and 1.0, the fingerprinter calculates the features for each compound, and the Classifier type is the used multi-label classifier.

Threshold
0.5

Fingerprinter type
MACS Fingerprinter

Select additional descriptors
Nothing selected

Classifier type
Rule-Based

Evaluation type
Single Generation

Additional Evaluation Options

Quick build
 Evaluate later

Evaluation Packages
Nothing selected

Cancel Back Advanced Options Submit

Interface

- Web interface available at <https://envipath.org> with possibility to
 - Predict new pathways
 - Access database (pathways, reactions, compounds, models)
 - Enter your own data
 - Set model parameters
 - Train your own models
 - ...

2 Advanced Options

Select advanced options for the Relative Reasoning creation. The threshold is used for the predicted probabilities and should be between 0.0 and 1.0, the fingerprinter calculates the features for each compound, and the Classifier type is the used multi-label classifier.

Threshold
0.5

Fingerprinter type
MACS Fingerprinter

Select additional descriptors
Nothing selected

Classifier type
Rule-Based

Evaluation type
Single Generation

Additional Evaluation Options

Quick build
 Evaluate later

Evaluation Packages
Nothing selected

Cancel Back Advanced Options Submit

REST Interface



- REST interface provides an language-independent interface to the server
- Provides all functionality of the server for integration into external code
 - Add data
 - Predict pathways
 - Search the database
 - ...

REST Interface

- REST interface provides an language-independent interface to the server

- Provides all functionality of the server for integration into external code
 - Add data
 - Predict pathways
 - Search the database
 - ...

| URI | Parameter | Accept Types | Result | Status Codes |
|-----------|---------------------|--------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------|
| /reaction | reactionName | | Create a new reaction in the default package. The redirect points to the URI of the new Compound. Either the SMIRKS, or the educt and product need to be given. | 303 - See Other |
| | reactionDescription | | | 500 - Internal Server Error |
| | smirks | | | |
| | educt | | | |
| | product | | | |
| | rule | | | |

REST Interface

- REST interface provides an language-independent interface to the server

- Provides all functionality of the server for integration into external code
 - Add data
 - **Predict pathways**
 - Search the database
 - ...

| | | | | |
|-----------------------------------------|------------------------------|--|---------------------------------------------------------------------------------------------------------------------|-----------------------------|
| <i>/package/<id>/ pathway</i> | <i>smilesinput</i> | | Creates a new predicted pathway in the package given by the URI. The redirect points to the URI of the new pathway. | 303 - See Other |
| | <i>reverse</i> | | | 401 - Unauthorized |
| | <i>name</i> | | | 500 - Internal Server Error |
| | <i>selectedSetting</i> | | | |
| | <i><settingParams></i> | | | |

REST Interface

- REST interface provides an language-independent interface to the server

- Provides all functionality of the server for integration into external code
 - Add data
 - Predict pathways
 - Search the database
 - ...

| URI | Parameter | Accept Types | Result | Status Codes |
|----------|----------------------------|--------------|-----------------------------------------------------------------------|---------------------------------------------------|
| / search | search - the search string | */* | Performs a search for the given string inside the specified packages. | 200 - OK |
| | packages[] - package URI's | | Returns the result as application/json. | 401 - Unauthorized 500 - Internal Server Error |

Python Interface



- The Python client allows you to include enviPath directly into your Python code

```
from pprint import pprint
from enviPath_python import enviPath

ePServer = 'https://envipath.org'
package = '32de3cf4-e3e6-4168-956e-32fa5ddb0ce1'

eP = enviPath(ePServer)

bbd = eP.get_package(ePServer + '/package/' + package)

bbd_pws = bbd.get_pathways()

pprint(bbd_pws[0].get_description())
```

Python Interface



- The Python client allows you to include enviPath directly into your Python code
- This is done by using the REST API

```
from pprint import pprint
from enviPath_python import enviPath

ePServer = 'https://envipath.org'
package = '32de3cf4-e3e6-4168-956e-32fa5ddb0ce1'

eP = enviPath(ePServer)

bbd = eP.get_package(ePServer + '/package/' + package)

bbd_pws = bbd.get_pathways()

pprint(bbd_pws[0].get_description())
```

Python Interface



- The Python client allows you to include enviPath directly into your Python code
- This is done by using the REST API
- All calls to Python functions are translated to remote calls to enviPath

```
from pprint import pprint
from enviPath_python import enviPath

ePServer = 'https://envipath.org'
package = '32de3cf4-e3e6-4168-956e-32fa5ddb0ce1'

eP = enviPath(ePServer)

bbd = eP.get_package(ePServer + '/package/' + package)

bbd_pws = bbd.get_pathways()

pprint(bbd_pws[0].get_description())
```

Python Interface



- The Python client allows you to include enviPath directly into your Python code
- This is done by using the REST API
- All calls to Python functions are translated to remote calls to enviPath
- Exposes all functionality to direct function calls in Python

```
from enviPath_python import enviPath

ePServer = 'https://envipath.org'
package = '32de3cf4-e3e6-4168-956e-32fa5ddb0ce1'

eP = enviPath(ePServer)

bbdPackage = '32de3cf4-e3e6-4168-956e-32fa5ddb0ce1'
soilPackage = '5882df9c-dae1-4d80-a40e-db4724271456'

# get package(s) that should be searched
bbd = eP.get_package(ePServer + '/package/' + bbdPackage')
soil = eP.get_package(ePServer + '/package/' + soilPackage')

# returns a dictionary with properly initialized objects
res = eP.search('c1cccc1', [bbd, soil])
print(res)

# or use a package to search it
res = bbd.search('c1cccc1')
print(res)
```

Conclusion



■ Current work

- Implementing enviFormer into enviPath
- Implement rule learning into enviPath
- Adding more data
- Adding more improvements into the prediction engine

enviPath



SCIENCE

Thank you for listening! Any questions?

<https://envipath.org>

<https://wickerlab.org>