Sustainable design of chemical reagents for sensitive detection of pesticides using a machine learning workflow

SCS2025, Västerås, 16.06.2025
Analytical chemistry session

Henrik Hupatz

Postdoctoral researcher

Stockholm University
Department of Chemistry

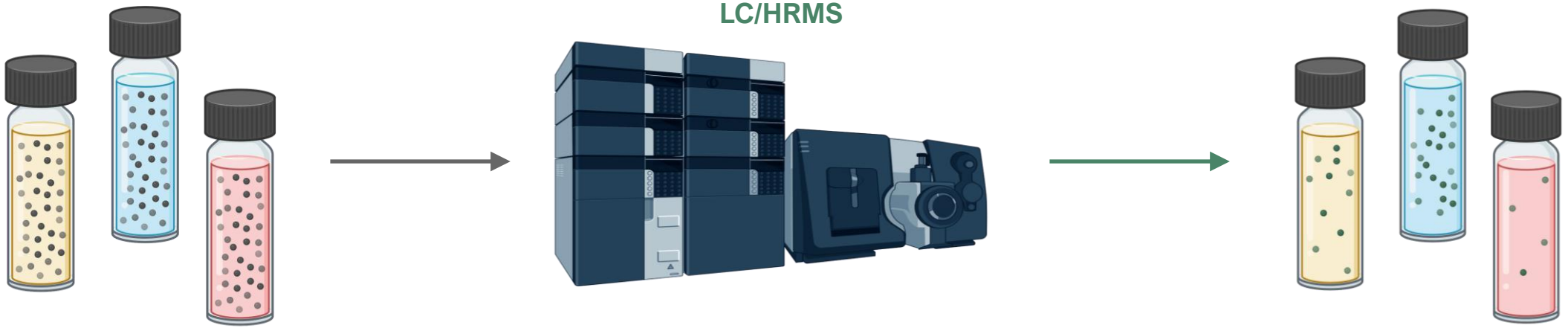
henrik.hupatz@su.se

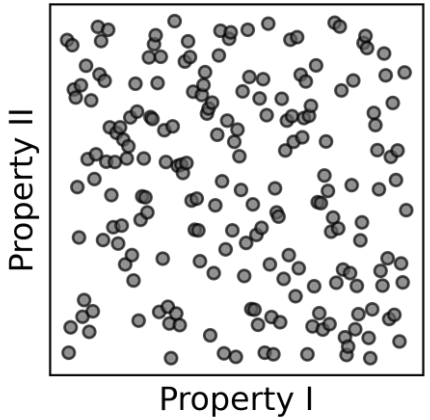
Kruve Lab

Stockholm University

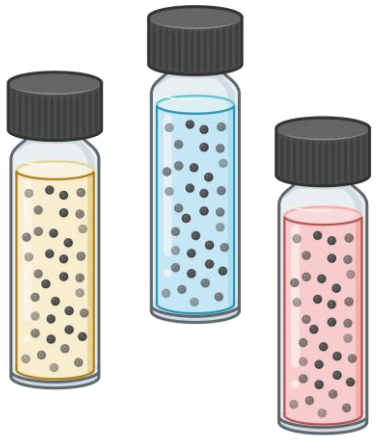
SUCCeSS

INTRODUCTION

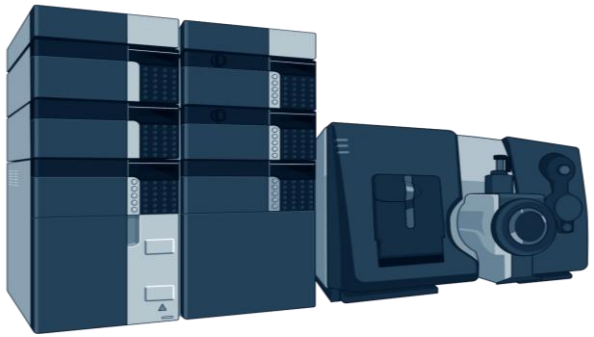


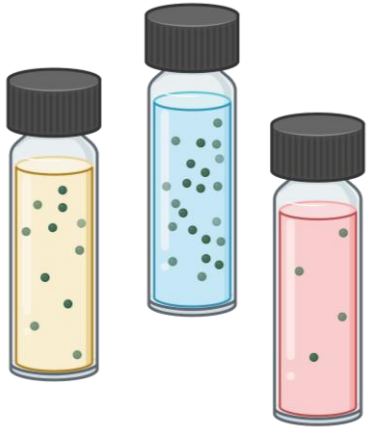


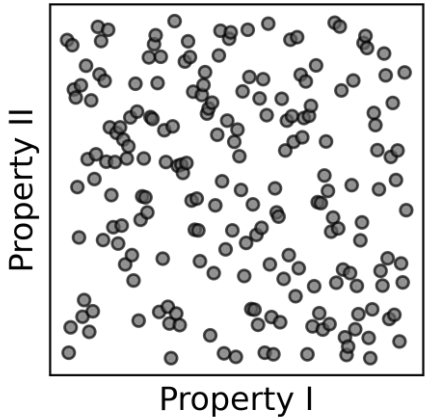
INTRODUCTION

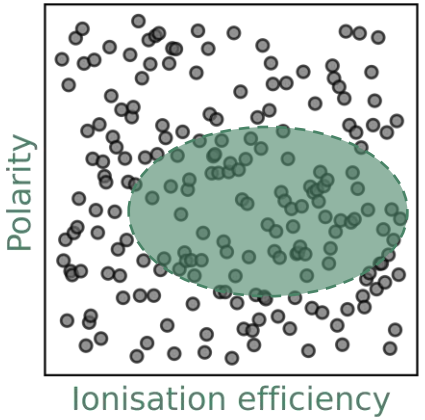


LC/HRMS

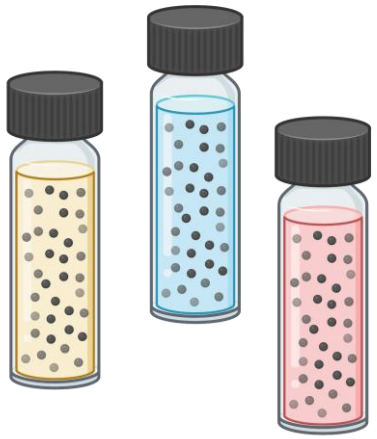


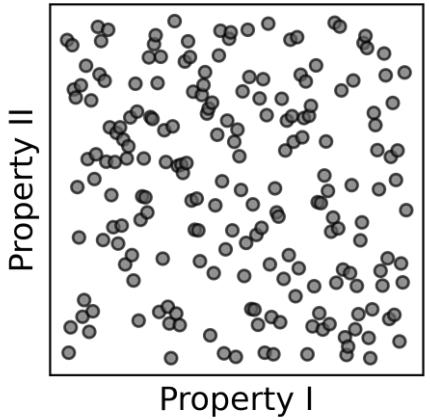




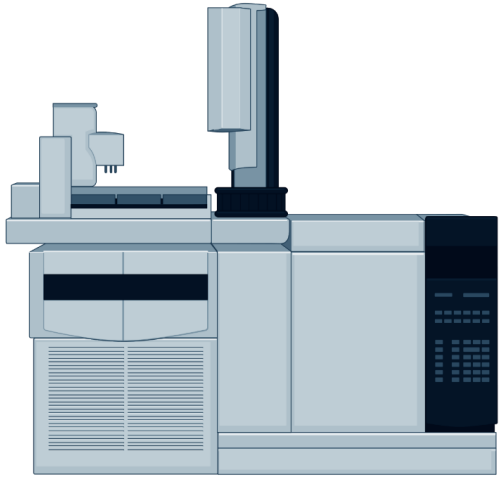


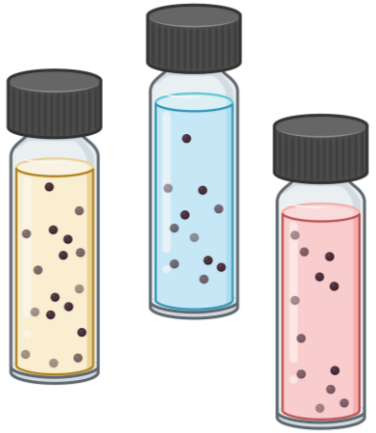
INTRODUCTION

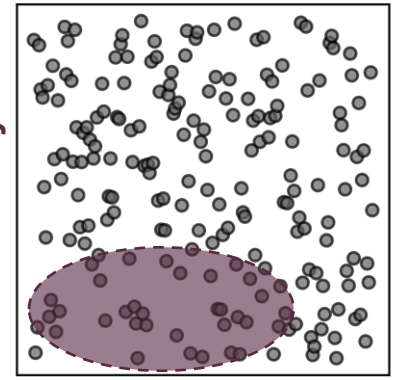




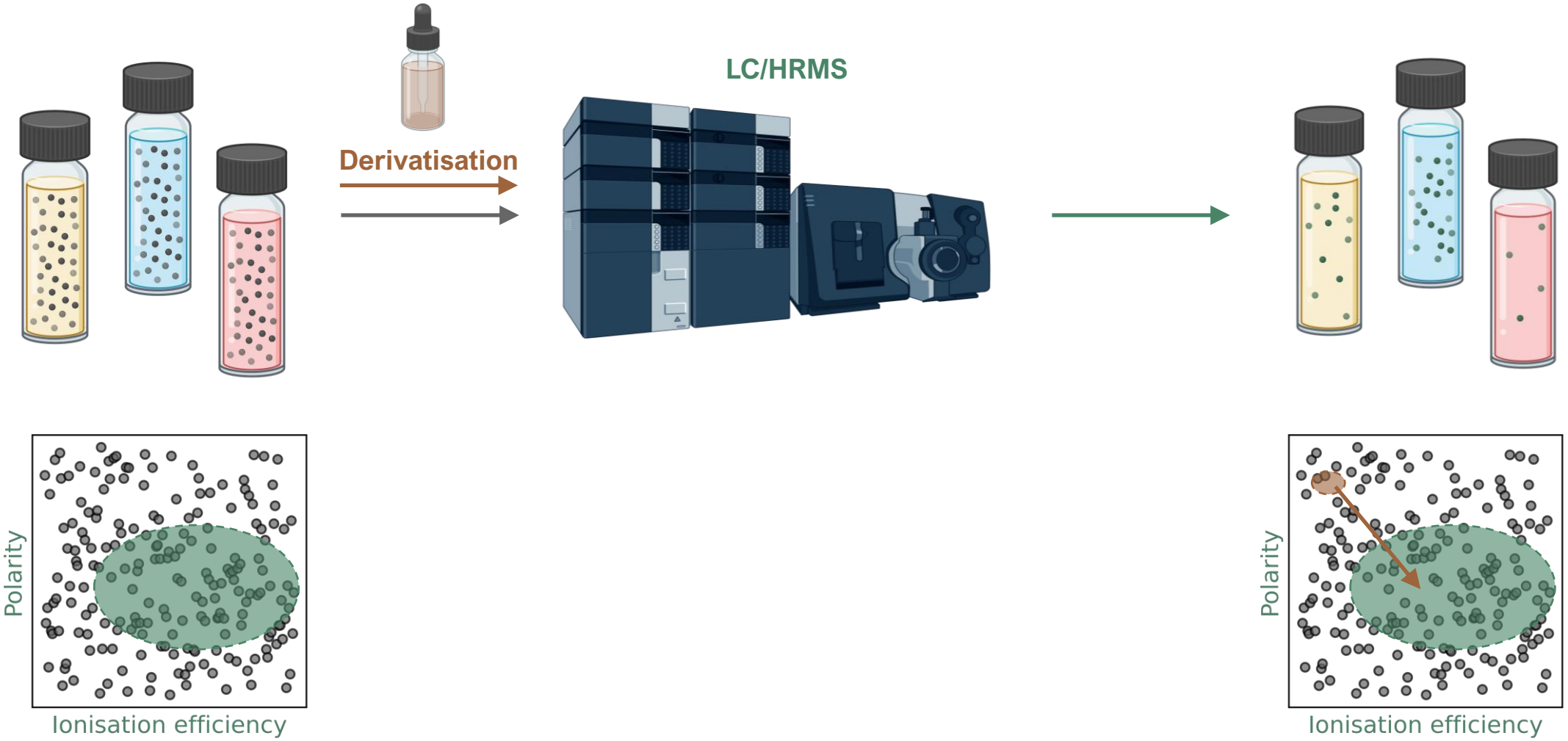
GC/HRMS



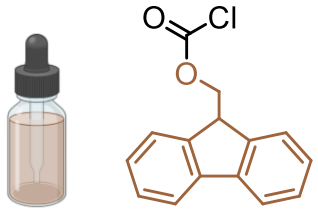




INTRODUCTION

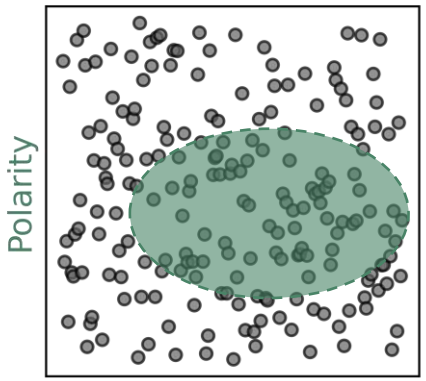


DERIVATISATION REAGENT DESIGN



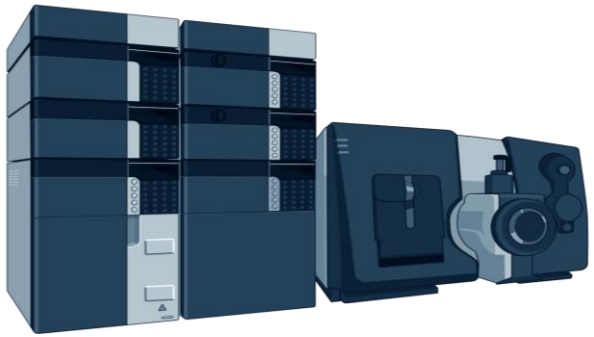
FMOCCl

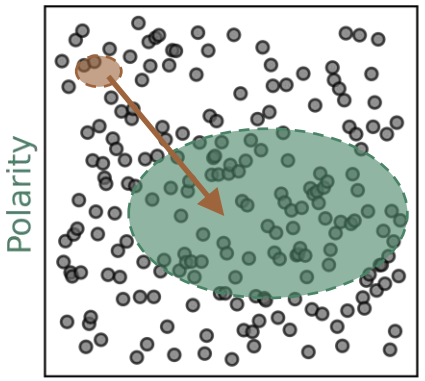
DERIVATISATION REAGENT DESIGN



Derivatisation

LC/HRMS





Experimental experience
Literature review
Trial and error

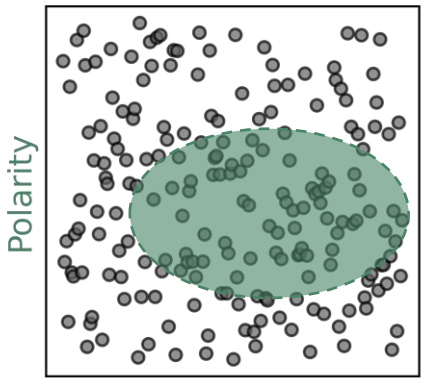
Expert design

Tailor-made derivatisation reagent

Low polarity
Good ionisation
Easy synthesis

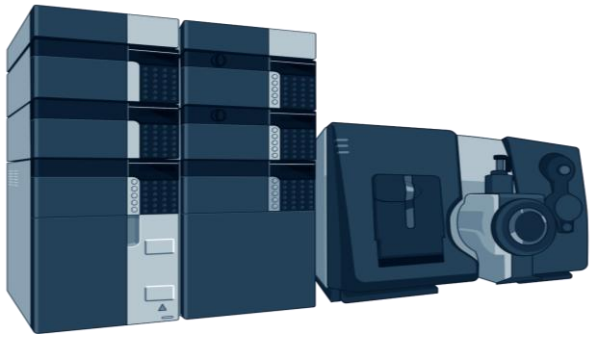
B. Sanchez-Lengeling, A. Aspuru-Guzik *Science* **2018**, 360.

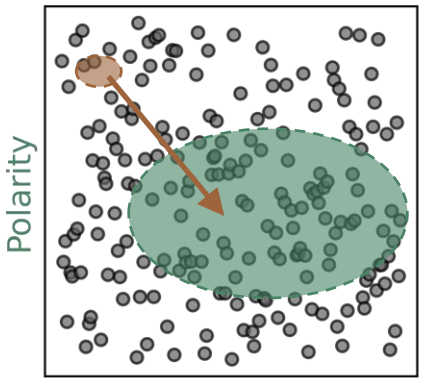
DERIVATISATION REAGENT DESIGN



Derivatisation

LC/HRMS





Experimental experience
Literature review
Trial and error

Expert design

Tailor-made Derivatisation reagent

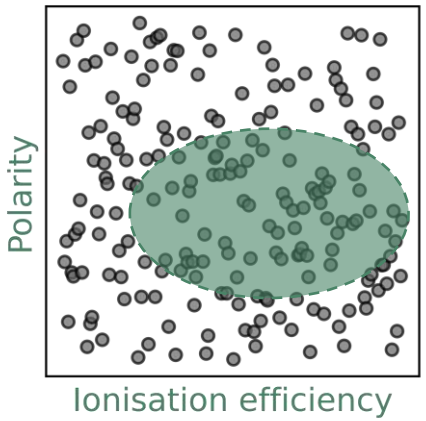
Low polarity
Good ionisation
Easy synthesis

Data-driven
Machine Learning
In silico generation
of candidates

***In silico* inverse property-guided design**

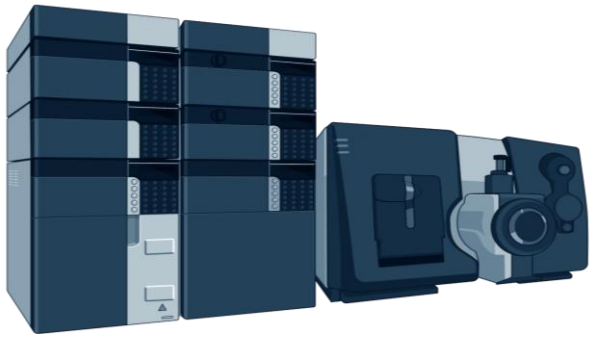
B. Sanchez-Lengeling, A. Aspuru-Guzik *Science* **2018**, 360.

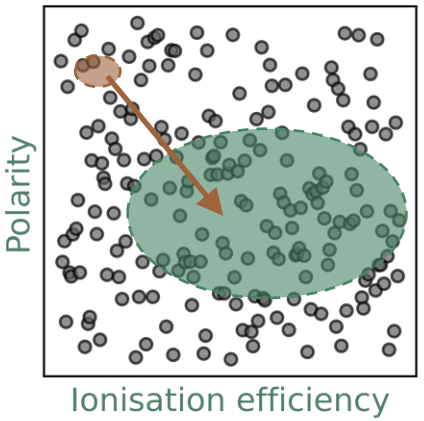
DERIVATISATION REAGENT DESIGN



Derivatisation

LC/HRMS





Experimental experience
Literature review
Trial and error

Expert design

Tailor-made derivatisation reagent

Low polarity
Good ionisation
Easy synthesis

Data-driven
Machine Learning
In silico generation
of candidates

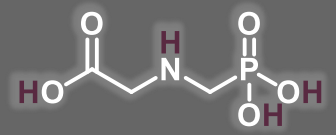
***In silico* inverse property-guided design**

CAN A MACHINE LEARNING WORKFLOW DESIGN SUITABLE DERIVATISATION REAGENTS FOR LC/HRMS?

B. Sanchez-Lengeling, A. Aspuru-Guzik *Science* **2018**, 360.

DERIVATISATION REAGENT DESIGN

PROTOYPE



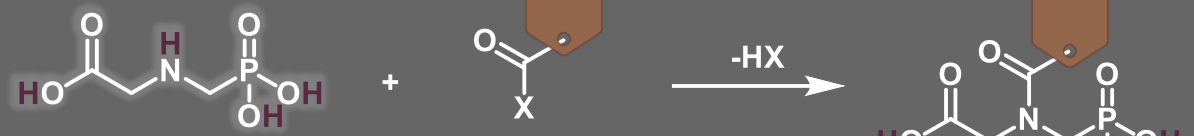
Glyphosate

highly polar
zwitterionic

A. L. Valle, F. C. C. Mello, R. P. Alves-Balvedi et al. *Environ. Chem. Lett*. **2019**, 291; A. M. Röhnelt, P. R. Martin, M. Athmer *Nature Commun.* **2025**, 2438.

DERIVATISATION REAGENT DESIGN

PROTOYPE

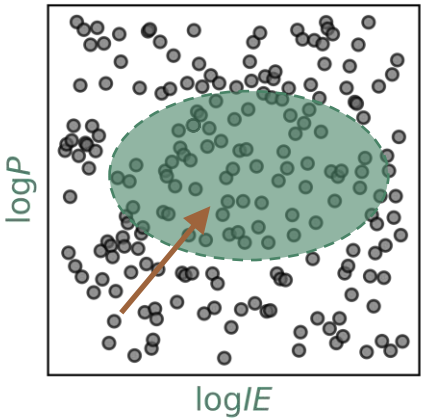


Glyphosate

highly polar
zwitterionic

Derivatisation reagent

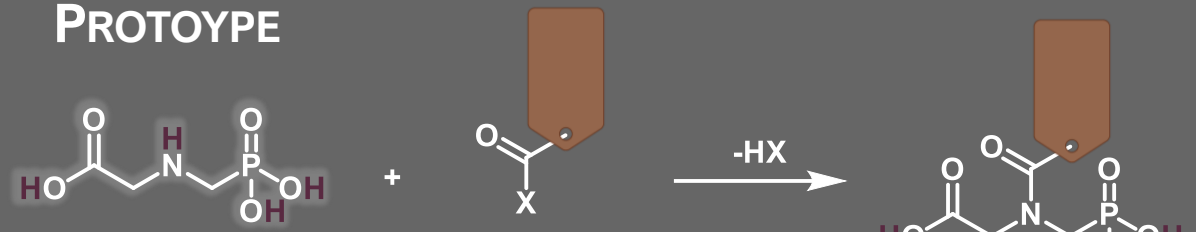
less polar
not zwitterionic



A. L. Valle, F. C. C. Mello, R. P. Alves-Balvedi et al. *Environ. Chem. Lett*. **2019**, 291; A. M. Röhnelt, P. R. Martin, M. Athmer *Nature Commun.* **2025**, 2438.

DERIVATISATION REAGENT DESIGN

PROTOYPE

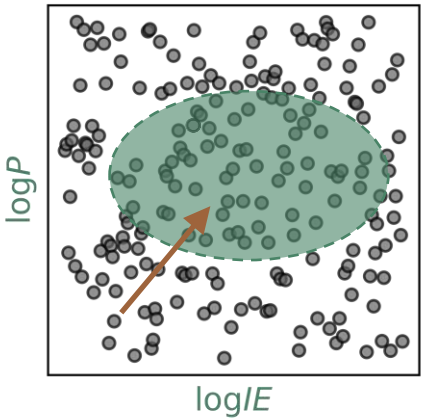


Glyphosate

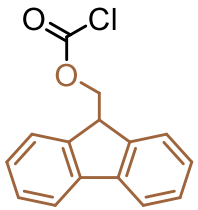
highly polar
zwitterionic

Derivatisation reagent

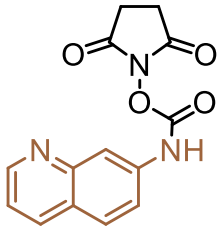
less polar
not zwitterionic



COMMERCIAL REAGENTS



FMOCCl

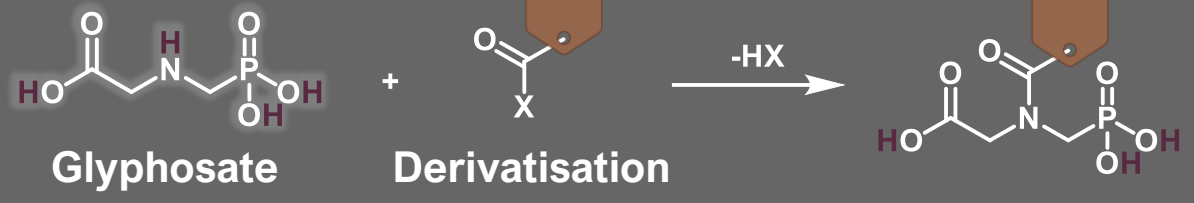


ACQ

A. L. Valle, F. C. C. Mello, R. P. Alves-Balvedi et al. *Environ. Chem. Lett*. **2019**, 291; A. M. Röhnelt, P. R. Martin, M. Athmer *Nature Commun.* **2025**, 2438.

DERIVATISATION REAGENT DESIGN

PROTOYPE

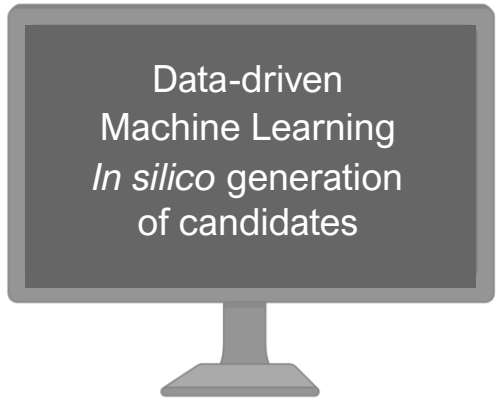


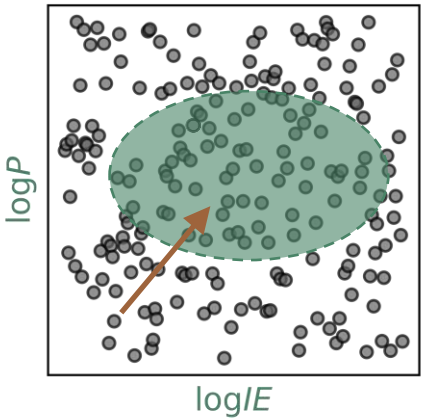
Glyphosate

highly polar
zwitterionic

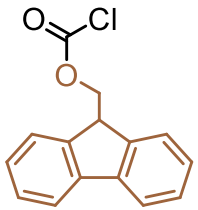
Derivatisation reagent

less polar
not zwitterionic

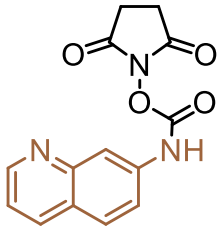




COMMERCIAL REAGENTS



FMOCCl



ACQ

EXPERIMENTAL FOUNDATION

Reagent synthesis

Easy and cheap

Sample preparation

Suitable reactivity

Separation

Suitable retention time

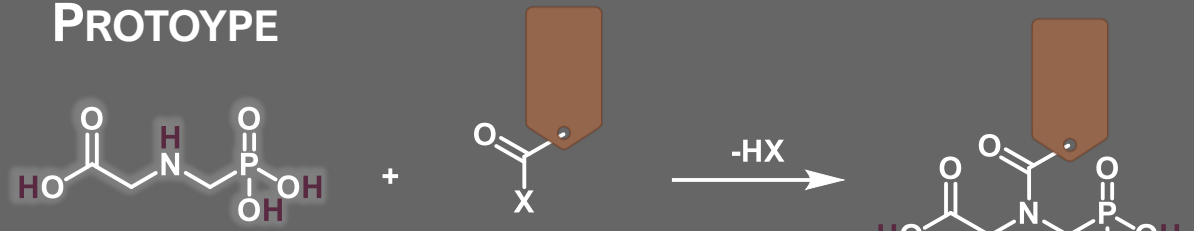
Detection

High ionization efficiency

A. L. Valle, F. C. C. Mello, R. P. Alves-Balvedi et al. *Environ. Chem. Lett*. **2019**, 291; A. M. Röhnelt, P. R. Martin, M. Athmer *Nature Commun.* **2025**, 2438.

DERIVATISATION REAGENT DESIGN

PROTOTYPE

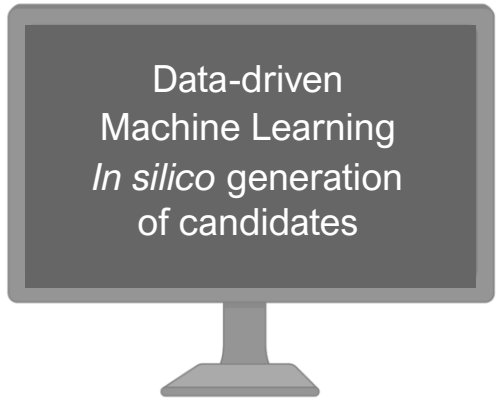


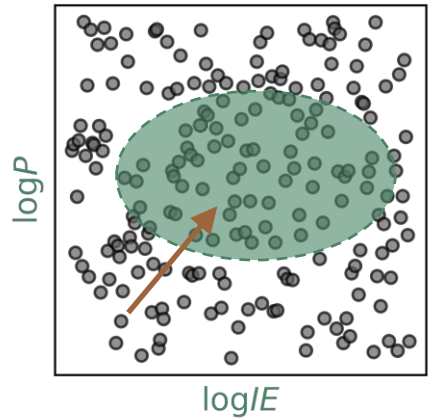
Glyphosate

highly polar
zwitterionic

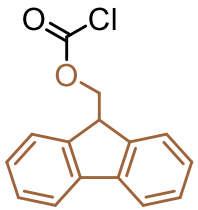
Derivatisation reagent

less polar
not zwitterionic

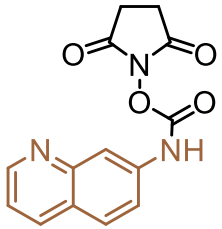




COMMERCIAL REAGENTS



FMOCCl



ACQ

EXPERIMENTAL FOUNDATION

Reagent synthesis

Easy and cheap

SAScore

Sample preparation

Suitable reactivity

-COOH group

Separation

Suitable retention time

log*P*

Detection

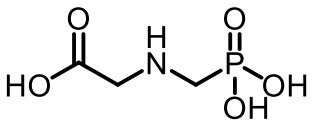
High ionization efficiency

log*IE*

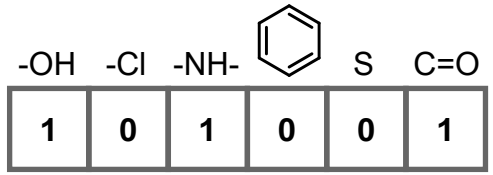
A. L. Valle, F. C. C. Mello, R. P. Alves-Balvedi et al. *Environ. Chem. Lett*. **2019**, 291; A. M. Röhnelt, P. R. Martin, M. Athmer *Nature Commun.* **2025**, 2438.

IONIZATION EFFICIENCY PREDICTION

log*IE*

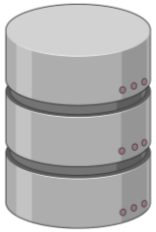


Calculate



Molecular Fingerprint

log*IE* dataset

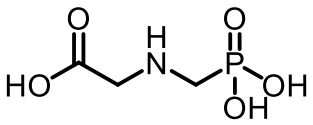


6120 datapoints

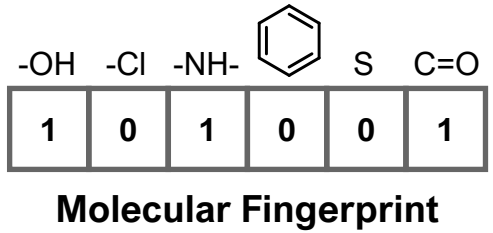
different experimental conditions

IONIZATION EFFICIENCY PREDICTION

log*IE*

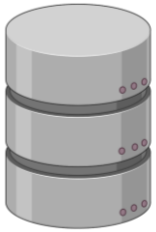


Calculate



Molecular Fingerprint

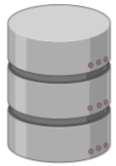
log*IE* dataset



6120 datapoints

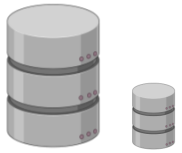
different experimental conditions

Clean



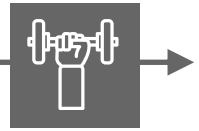
419 chemicals

Split



335:82 train/test set

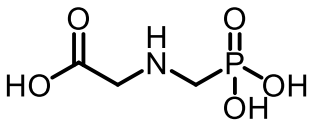
Train



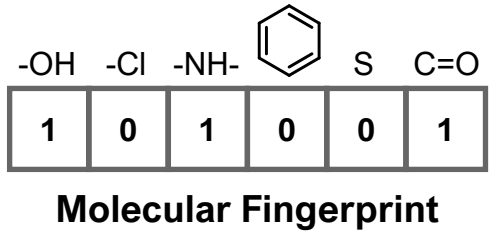
H. Sepman, L. Malm, P. Peets et al. *Anal. Chem.* **2023**, 95, 33, 12329; H. Sepman, H. Hupatz, A. Kruve manuscript in preparation.

IONIZATION EFFICIENCY PREDICTION

log*IE*

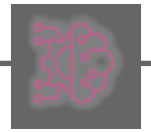


Calculate



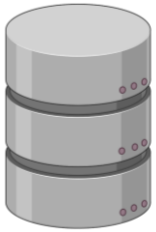
Molecular Fingerprint

Predict



log*IE*

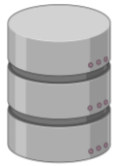
log*IE* dataset



6120 datapoints

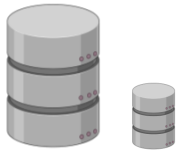
different experimental conditions

Clean



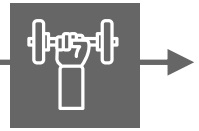
419 chemicals

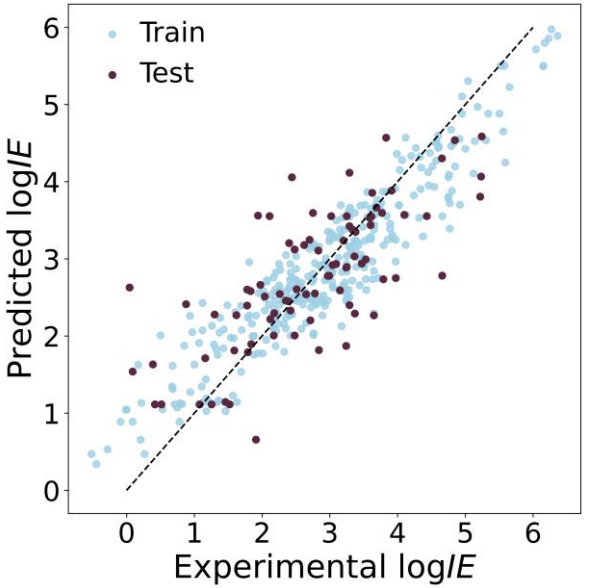
Split



335:82 train/test set

Train





Selected

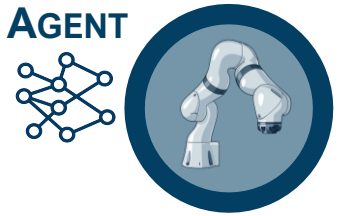
Algorithm: **XGBoost**

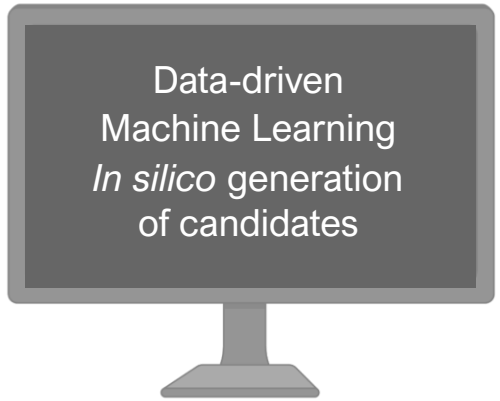
Representation: **MACCS fingerprints**

RMSE on test set: **0.78**

H. Sepman, L. Malm, P. Peets et al. *Anal. Chem.* **2023**, 95, 33, 12329; H. Sepman, H. Hupatz, A. Kruve manuscript in preparation.

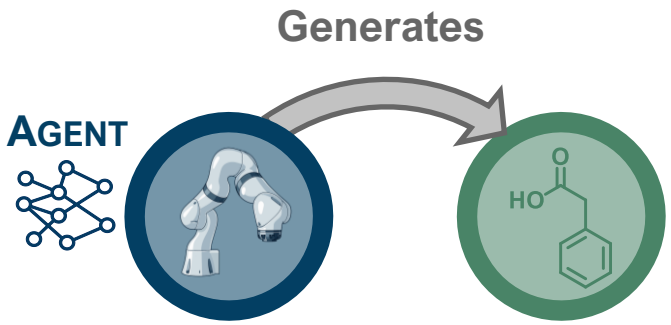
REINFORCEMENT LEARNING





H. H. Loeffler, J. He, A. Tibo et al. *J. Cheminform*. **2024**, 20.

REINFORCEMENT LEARNING



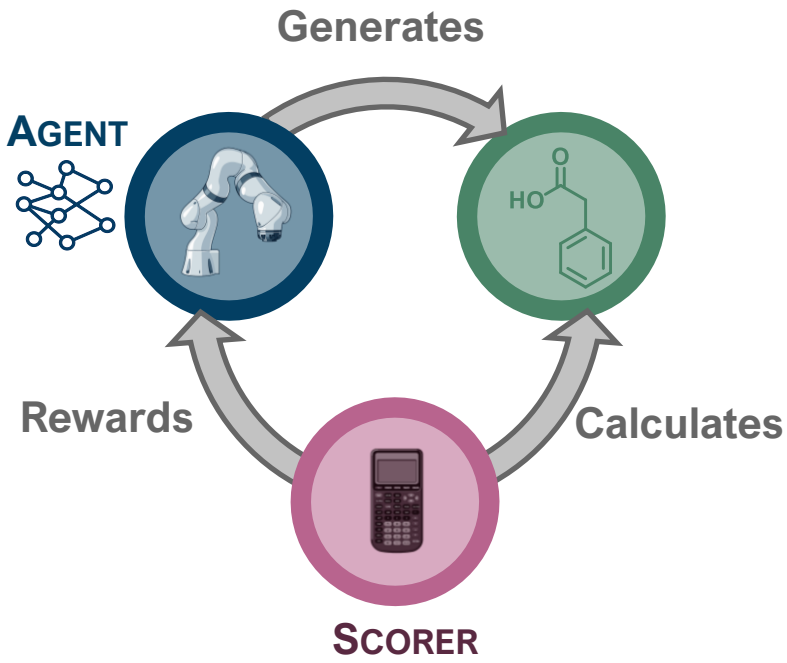
Data-driven
Machine Learning
In silico generation
of candidates

H. H. Loeffler, J. He, A. Tibo et al. *J. Cheminform*. **2024**, 20.

REINFORCEMENT LEARNING

SCORER

- $T = \log P$



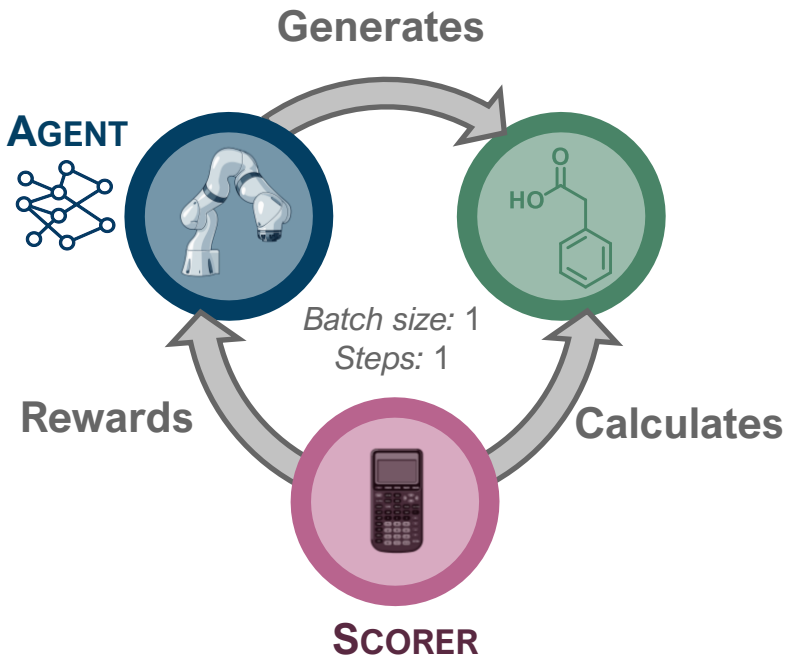
Data-driven
Machine Learning
In silico generation
of candidates

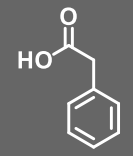
H. H. Loeffler, J. He, A. Tibo et al. *J. Cheminform*. **2024**, 20.

REINFORCEMENT LEARNING

SCORER

- $T = \log P$



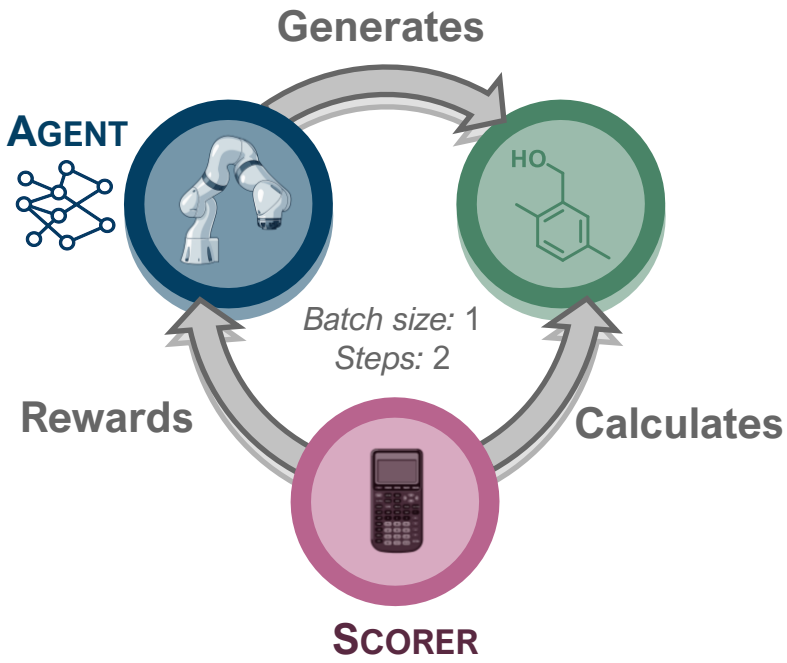


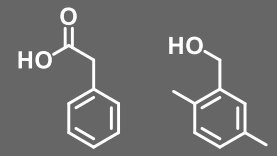
H. H. Loeffler, J. He, A. Tibo et al. *J. Cheminform*. **2024**, 20.

REINFORCEMENT LEARNING

SCORER

- $T = \log P$



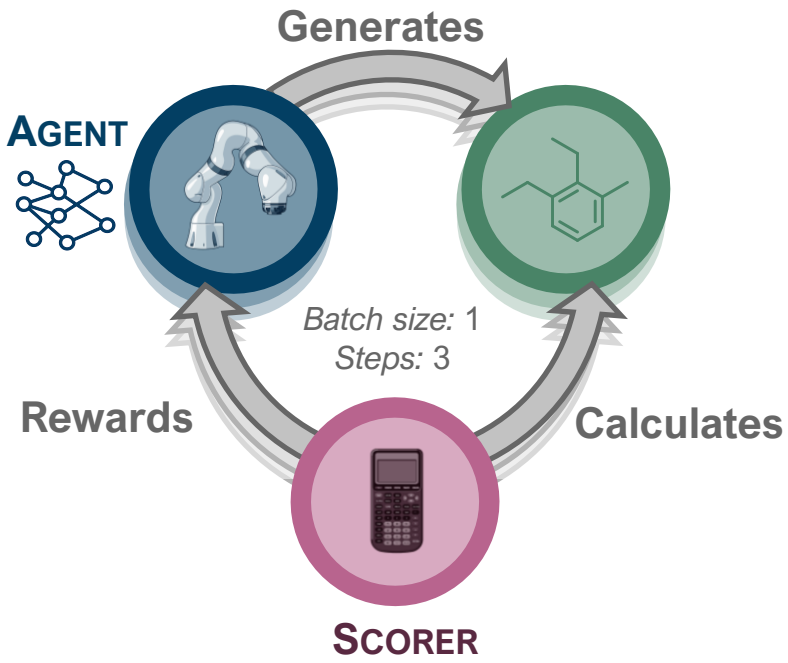


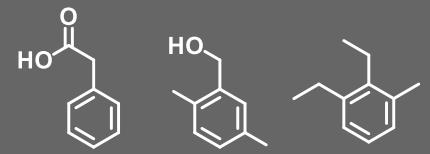
H. H. Loeffler, J. He, A. Tibo et al. *J. Cheminform*. **2024**, 20.

REINFORCEMENT LEARNING

SCORER

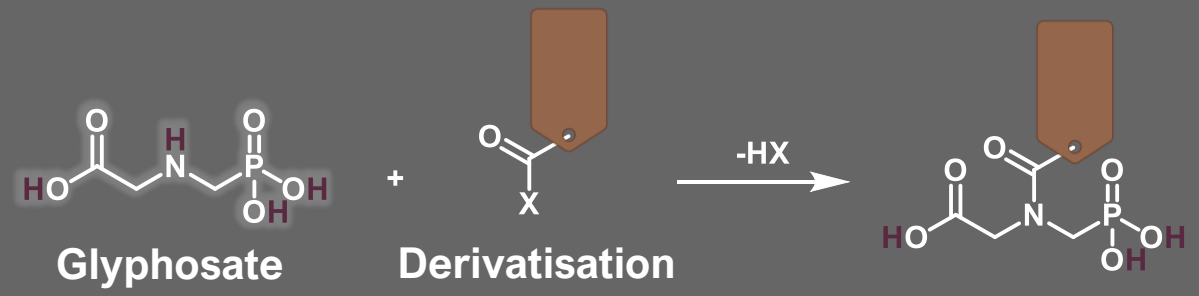
- $T = \log P$





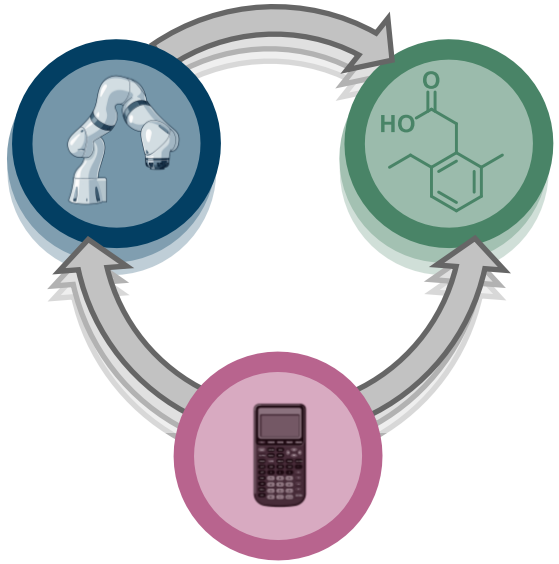
H. H. Loeffler, J. He, A. Tibo et al. *J. Cheminform*. **2024**, 20.

GENERATIVE MODELLING

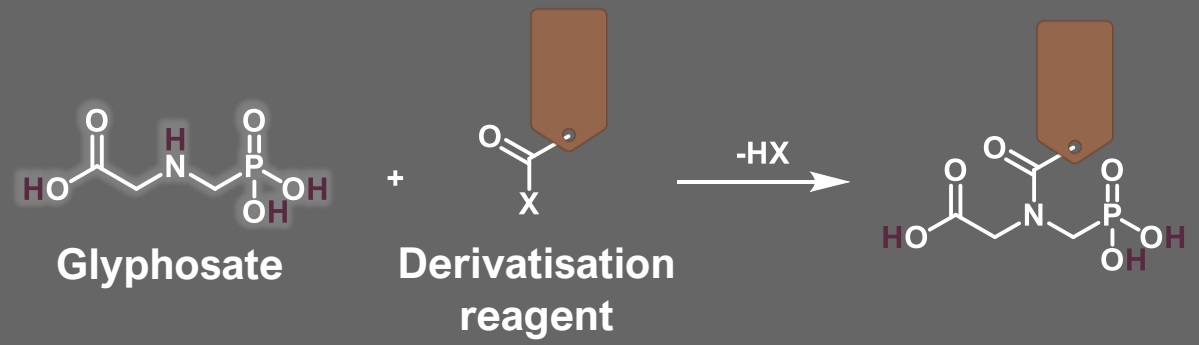


SCORER

- log*P*
- SAScore
- log*IE*
- -COOH group



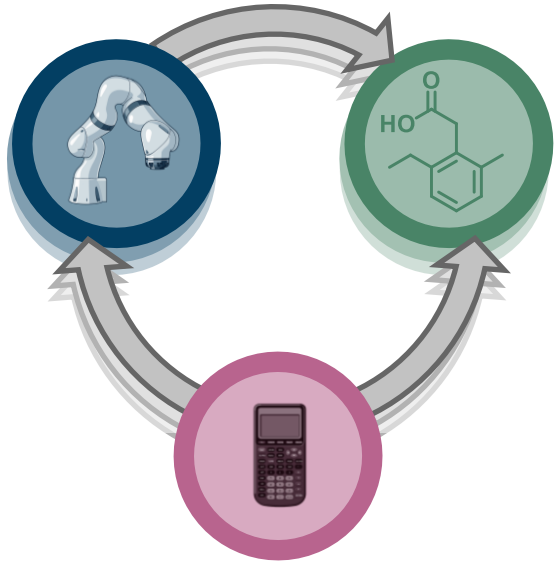
GENERATIVE MODELLING



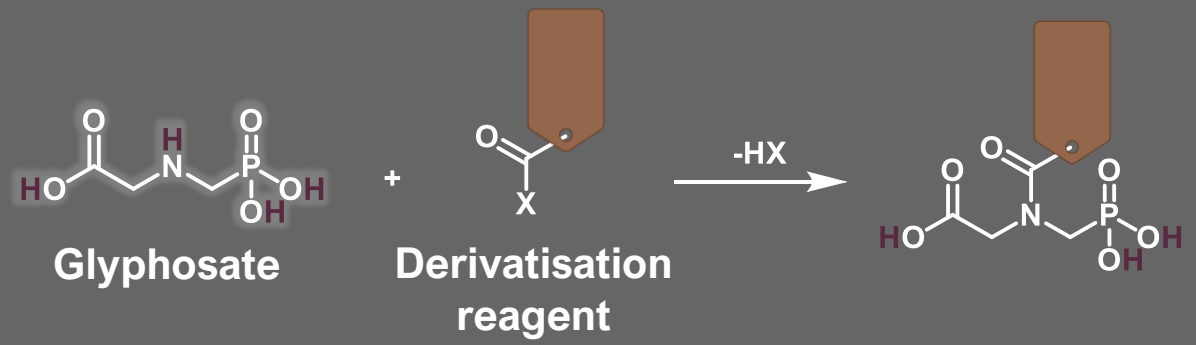
	Glyph	GlyphFMOC	GlyphACQ
predicted **log*IE***	1.4	3.6	3.7
calculated **log*P***	-1.2	2.5	1.3

SCORER

- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine

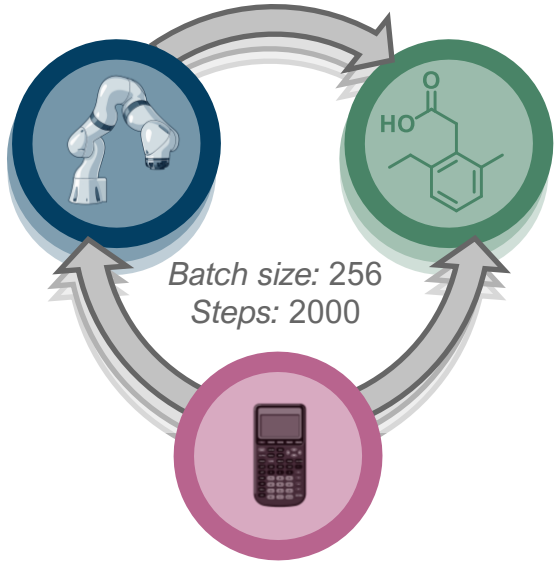


GENERATIVE MODELLING



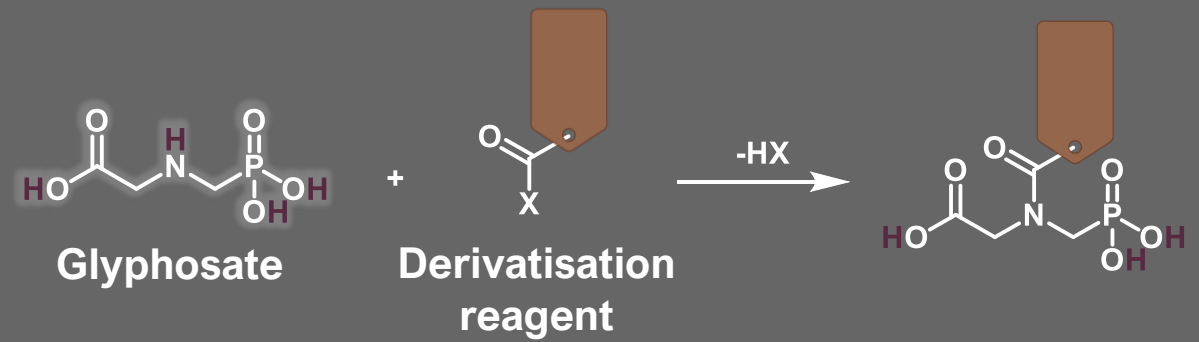
SCORER

- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine



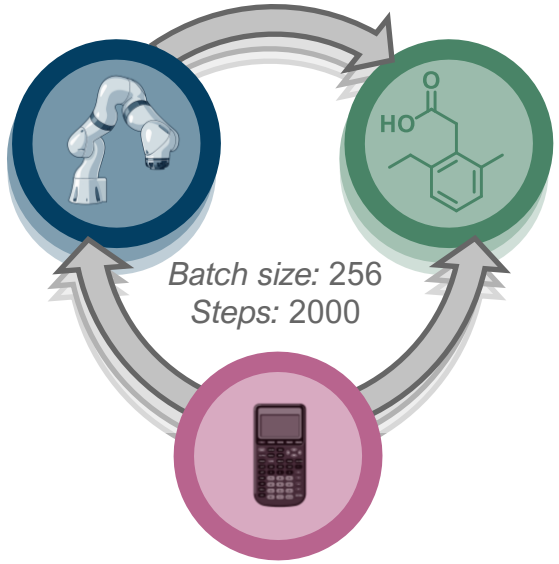
H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING



SCORER

- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine

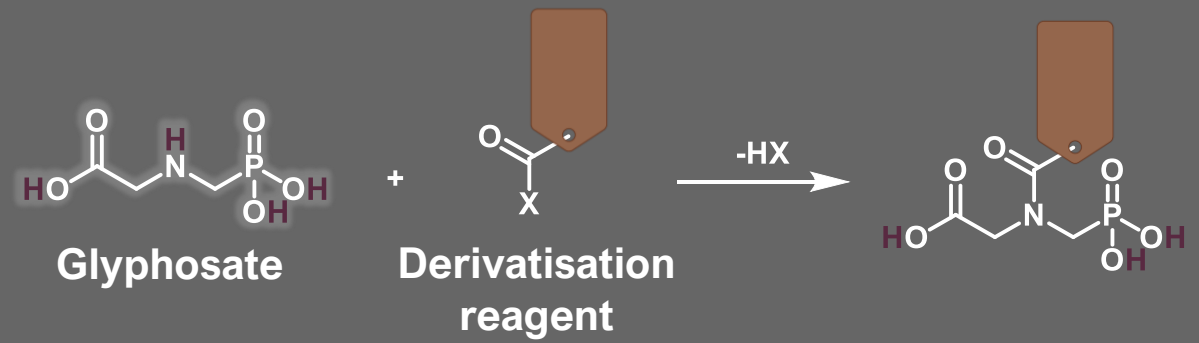


512,000 structures generated:

High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
36,312	1.49	1.26	4.09	4.24

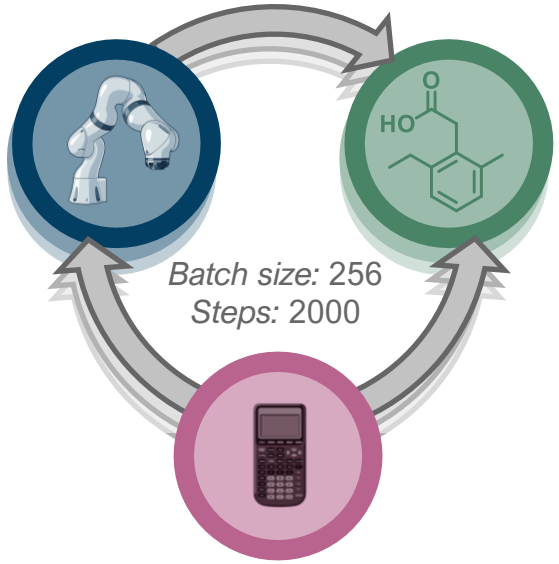
H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING



SCORER

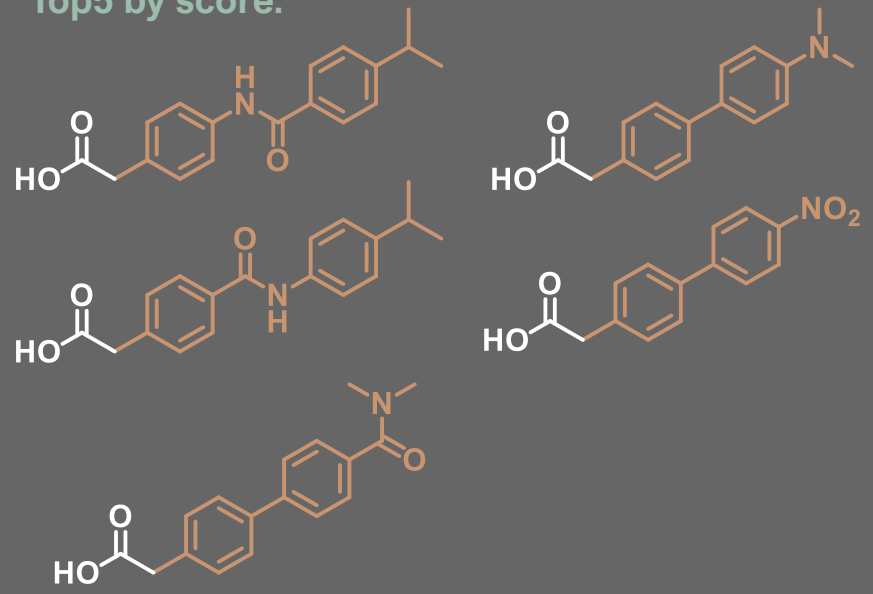
- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine



512,000 structures generated:

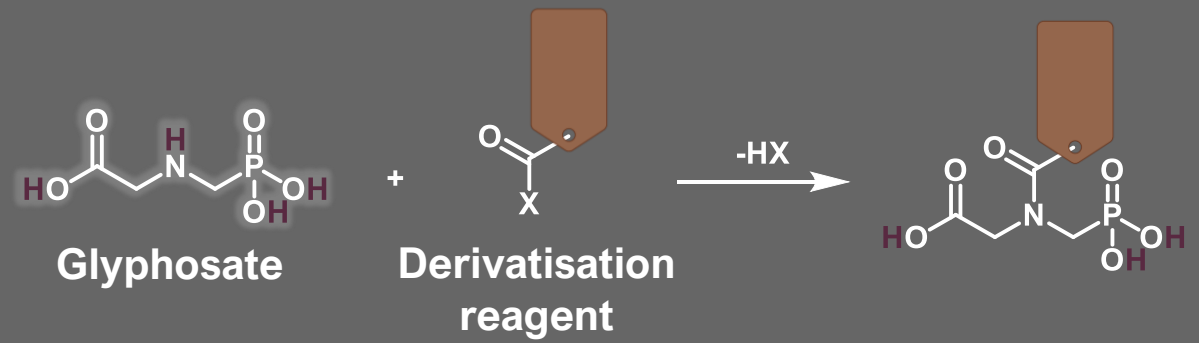
High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
36,312	1.49	1.26	4.09	4.24

Top5 by score:



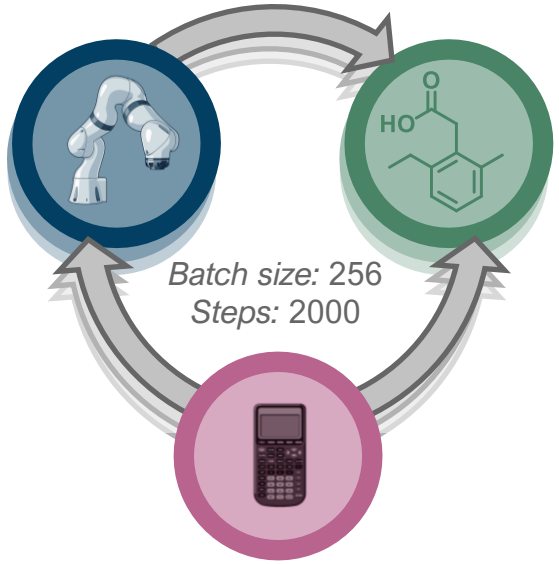
H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING



SCORER

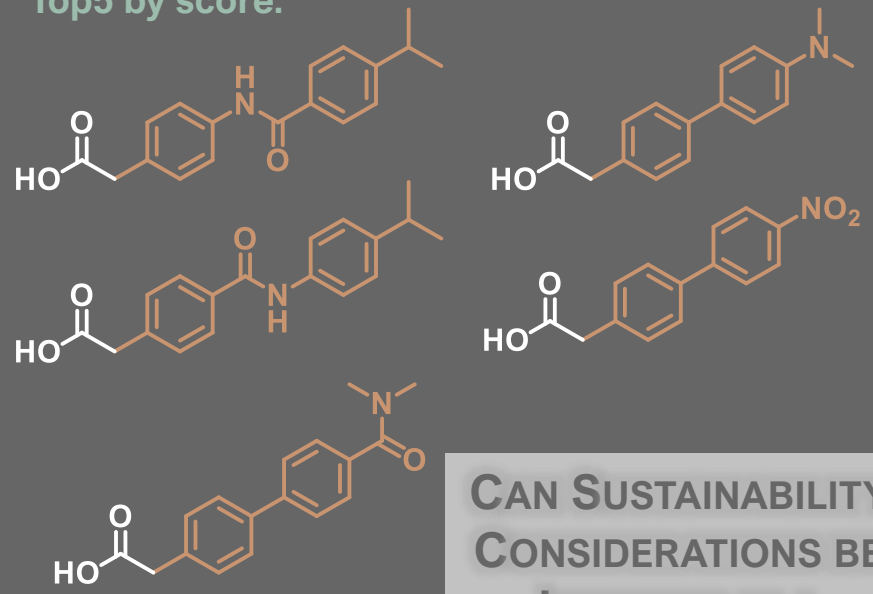
- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine



512,000 structures generated:

High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
36,312	1.49	1.26	4.09	4.24

Top5 by score:



CAN SUSTAINABILITY CONSIDERATIONS BE INVOLVED IN A GENERATIVE MODELLING WORKFLOW?

H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

SUSTAINABILITY CONSIDERATIONS

PRINCIPLES OF GREEN CHEMISTRY

Use of renewable feedstocks

Designing safer chemicals

E. Söderberg, K. von Borries, U. Norinder et al. *Green Chem*. **2024**, 11147. Z. Zhang et al. manuscript in preparation.

SUSTAINABILITY CONSIDERATIONS

PRINCIPLES OF GREEN CHEMISTRY

Use of renewable feedstocks

Designing safer chemicals

EXPERIMENTAL FOUNDATION

Lignin
Included in synthesis

Vanillin

Reagent synthesis
Easy and cheap

SAScore

Sample preparation
Low Hazard *Suitable reactivity*

Hazard score **-COOH group**

Separation
Suitable retention time

log*P*

Detection
High ionization efficiency

log*IE*

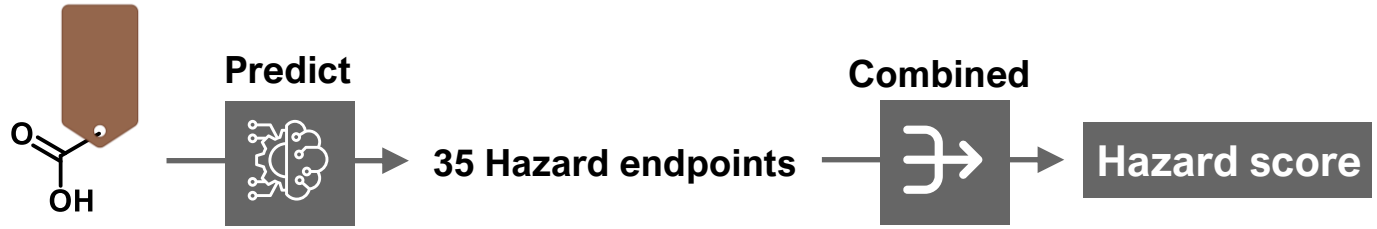
E. Söderberg, K. von Borries, U. Norinder et al. *Green Chem*. **2024**, 11147. Z. Zhang et al. manuscript in preparation.

SUSTAINABILITY CONSIDERATIONS

Principles of Green Chemistry

Use of renewable feedstocks

Designing safer chemicals



Experimental Foundation

Lignin
Included in synthesis
Vanillin

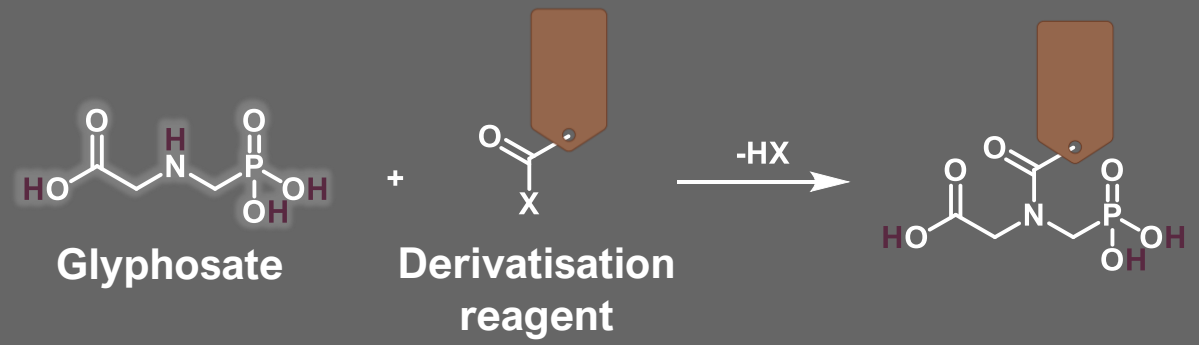
Reagent synthesis
Easy and cheap
SAScore

Sample preparation
Low Hazard
Hazard score
Suitable reactivity
-COOH group

Separation
Suitable retention time
log*P*
Detection
High ionization efficiency
log*IE*

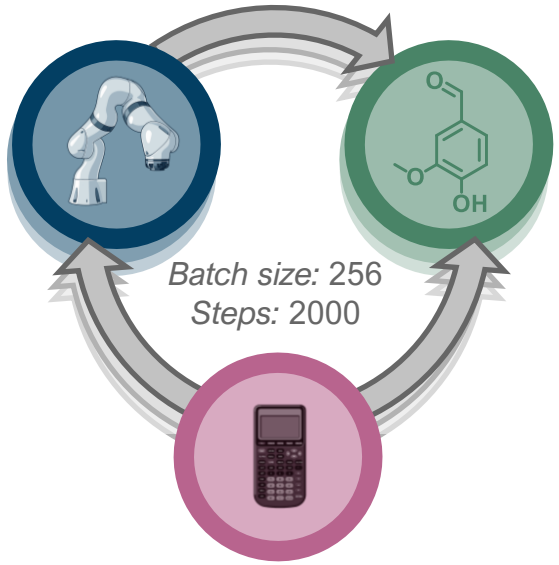
E. Söderberg, K. von Borries, U. Norinder et al. *Green Chem*. **2024**, 11147. Z. Zhang et al. manuscript in preparation.

GENERATIVE MODELLING



SCORER

- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine
- 1 **Vanillin**

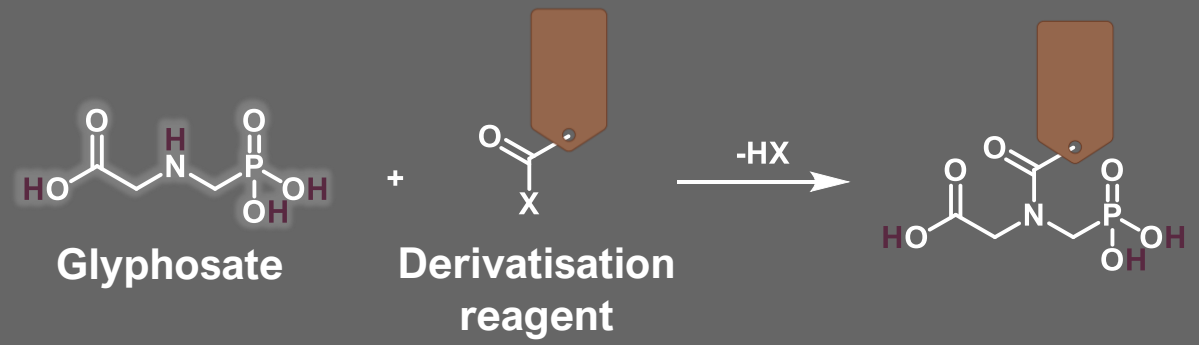


512,000 structures generated:

Vanillin	High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
No	36,312	1.49	1.26	4.09	4.24
Yes	29,120	1.71	1.51	4.20	4.29

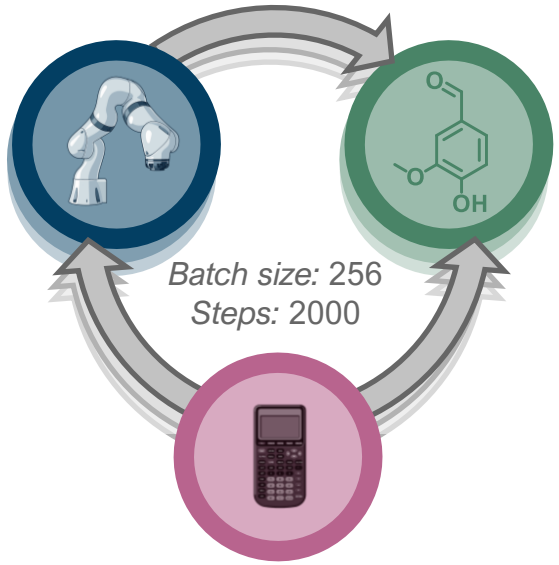
H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING



512,000 structures generated:

Vanillin	High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
Yes	29,120	1.49	1.26	4.09	4.24



SCORER

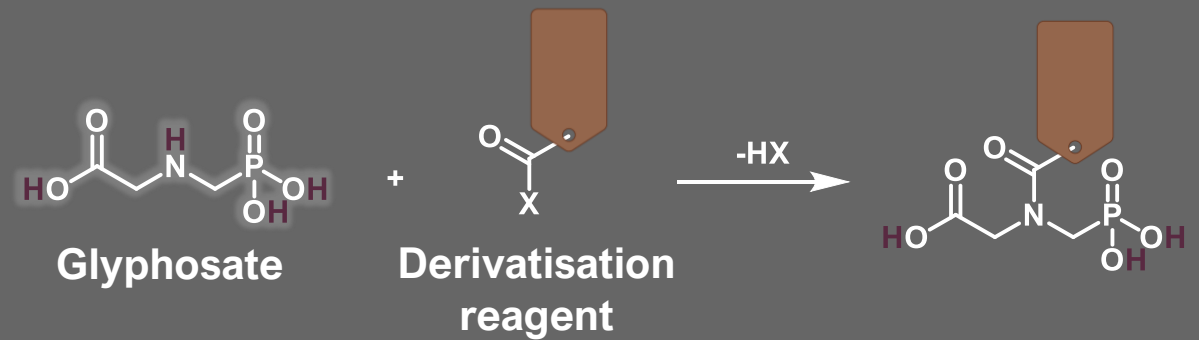
- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine
- 1 **Vanillin**

FILTER

1. 10k highest score
2. 1k lowest **Hazard score**

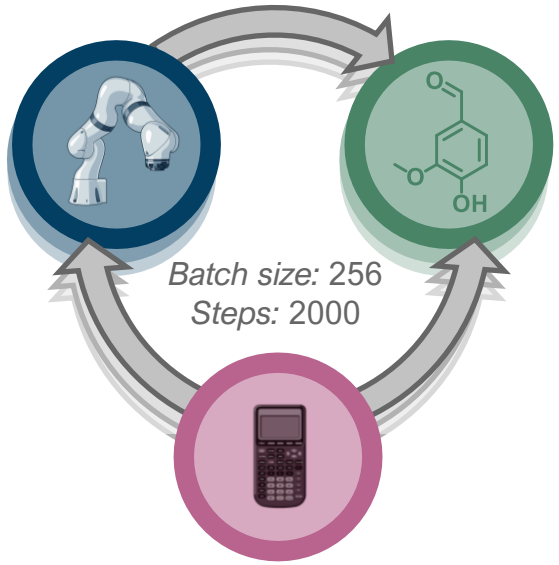
H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING



SCORER

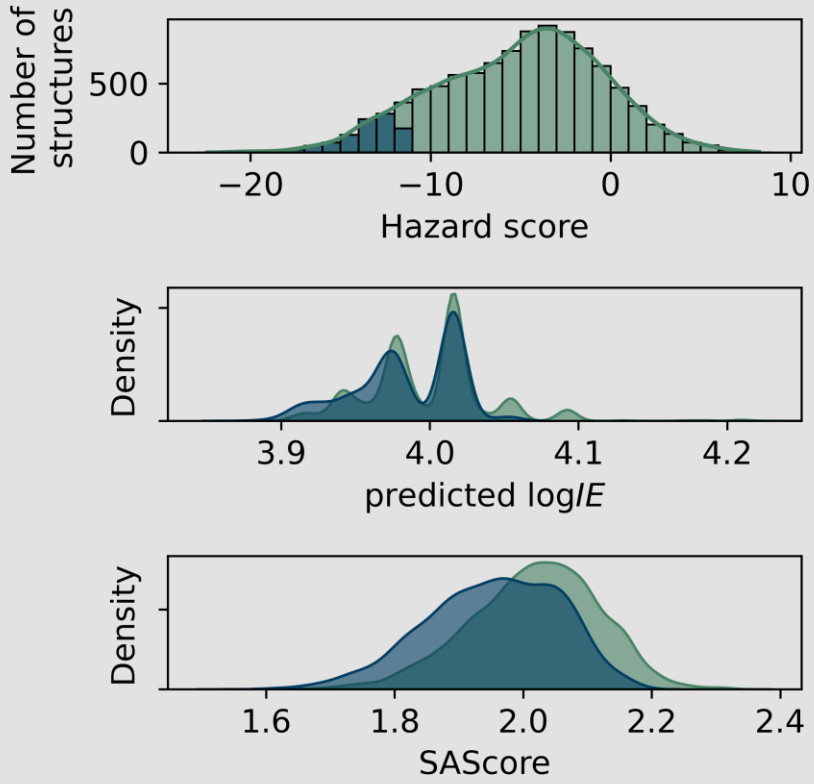
- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine
- 1 **Vanillin**



512,000 structures generated:

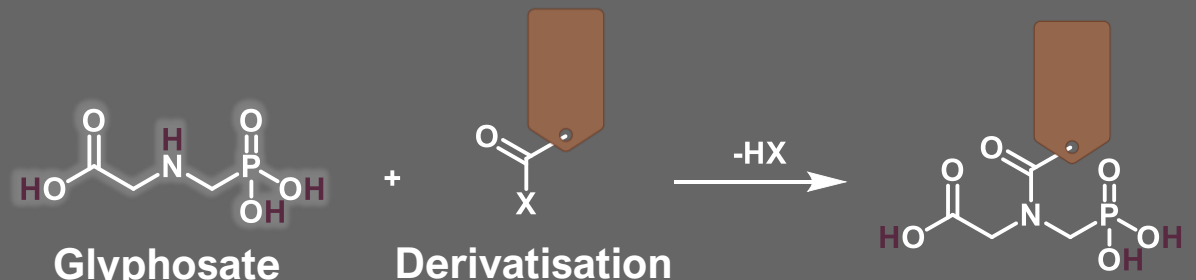
Vanillin	High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
Yes	29,120	1.49	1.26	4.09	4.24

Density plots **before** and **after** Hazard score filter:



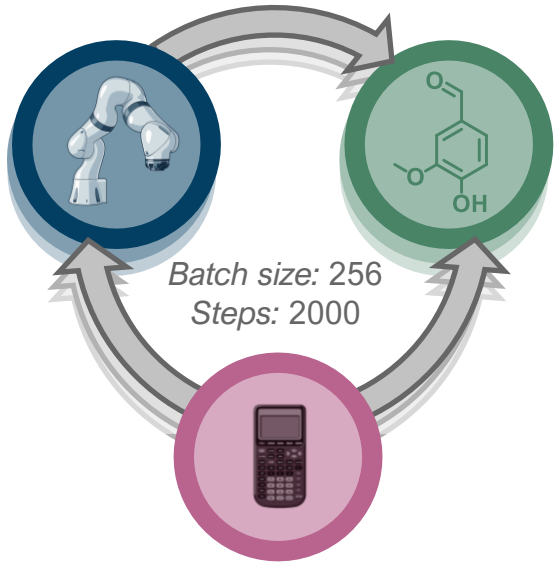
H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING



SCORER

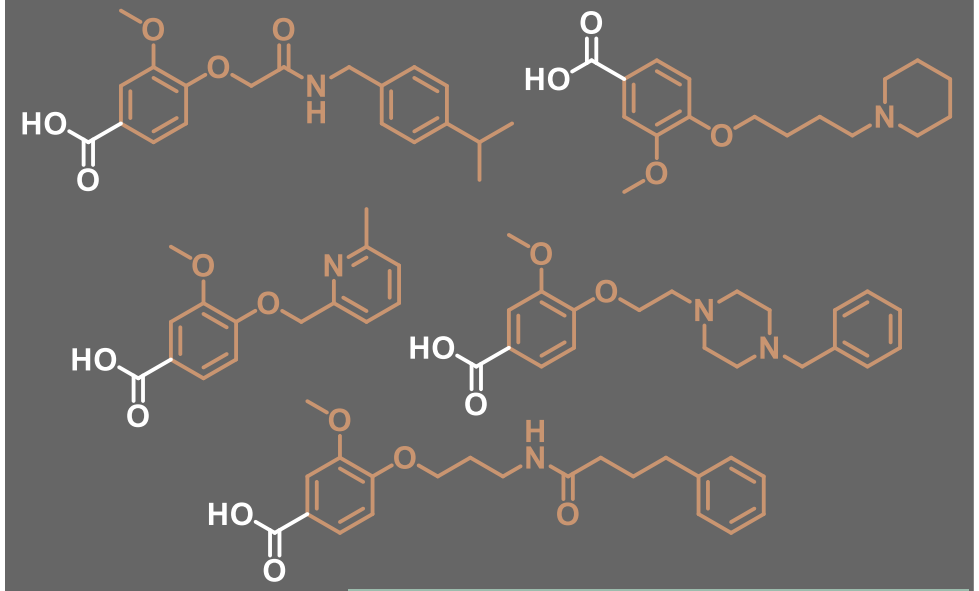
- 1 – 3 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine
- 1 **Vanillin**



512,000 structures generated:

Vanillin	High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
Yes	29,120	1.49	1.26	4.09	4.24

After filter:

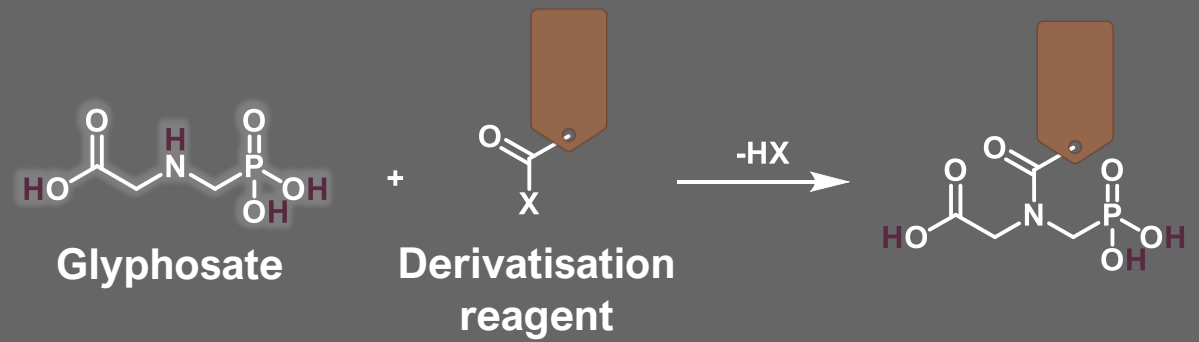


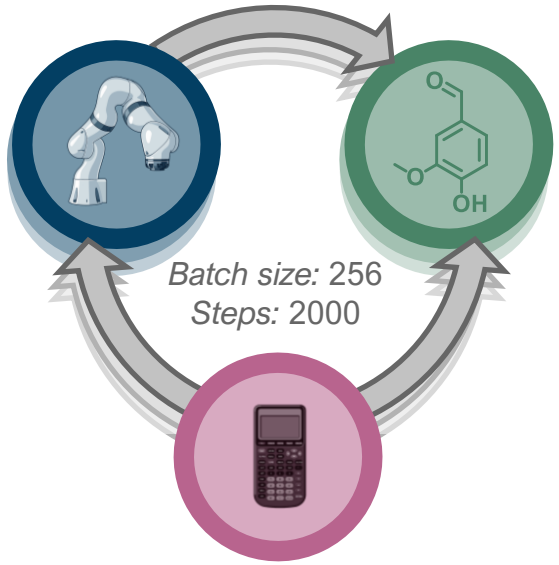
FILTER

1. 10k highest score
2. 1k lowest **Hazard score**
3. 5 lowest SAScore and diverse

H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING





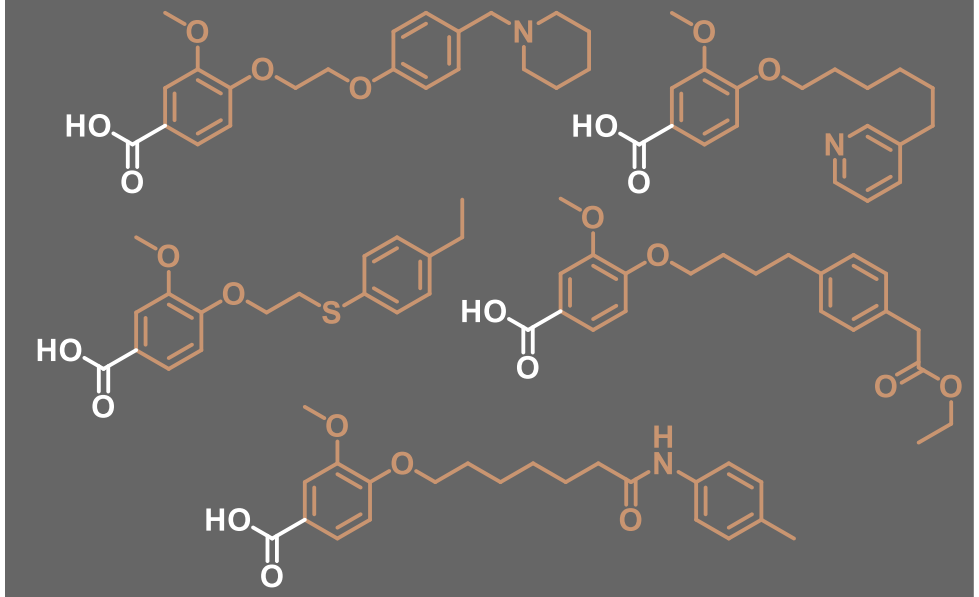
SCORER

- 2.7 – 4.7 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine
- 1 **Vanillin**

512,000 structures generated:

Vanillin	High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
Yes	15,449	1.78	1.65	4.12	4.25

After filter:

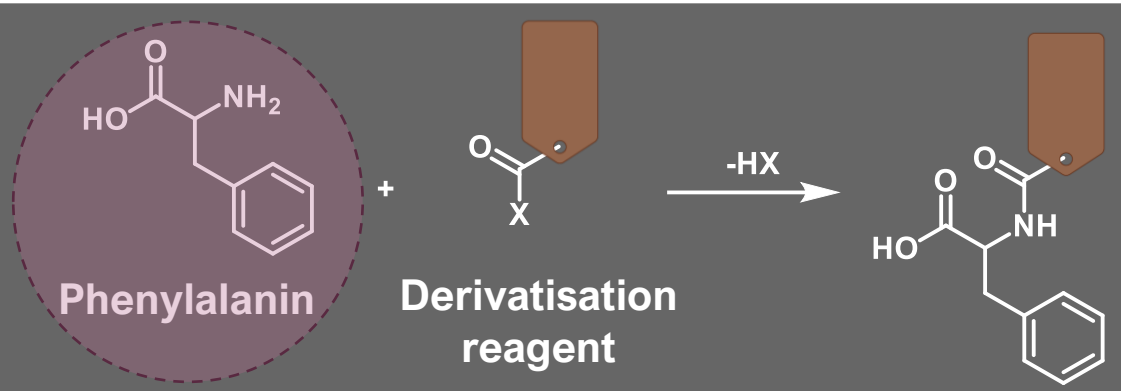


FILTER

1. 10k highest score
2. 1k lowest **Hazard score**
3. 5 lowest SAScore and diverse

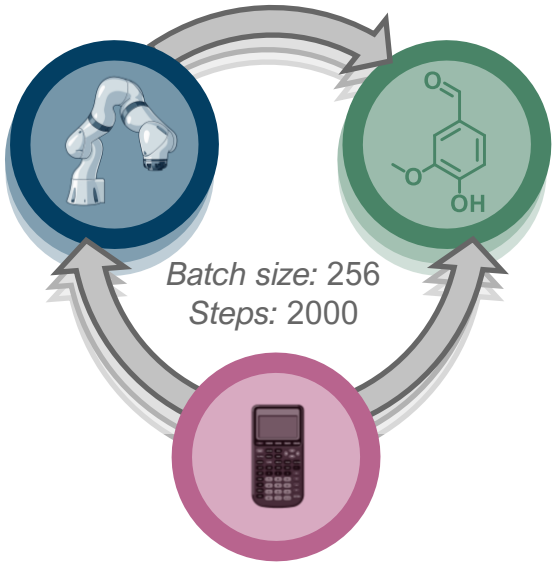
H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

GENERATIVE MODELLING



SCORER

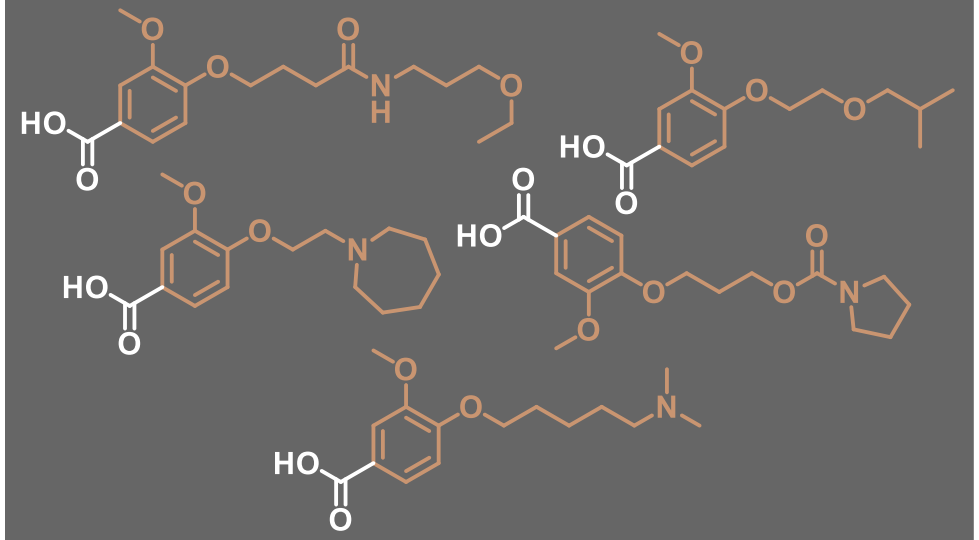
- 2.7 – 4.7 **log*P***
- < 2.5 **SAScore**
- > 3.8 **log*IE***
- 1 **-COOH group**
- no prim./sec. amine
- 1 **Vanillin**



512,000 structures generated:

Vanillin	High-scoring candidates	SAS of top 1%	*min* SAS	log*IE of top 1%*	*max* log*IE*
Yes	36,158	1.84	1.67	4.61	4.75

After filter:



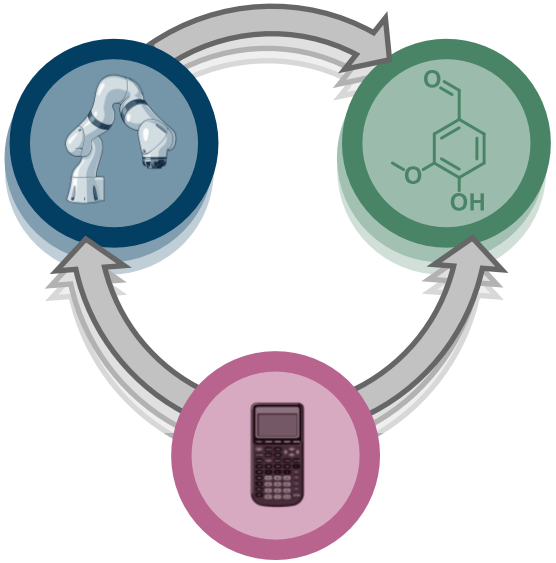
FILTER

1. 10k highest score
2. 1k lowest **Hazard score**
3. 5 lowest SAScore and diverse

H. Hupatz, M. Rivero-Crespo, B. Olofsson, A. Kruve manuscript in preparation.

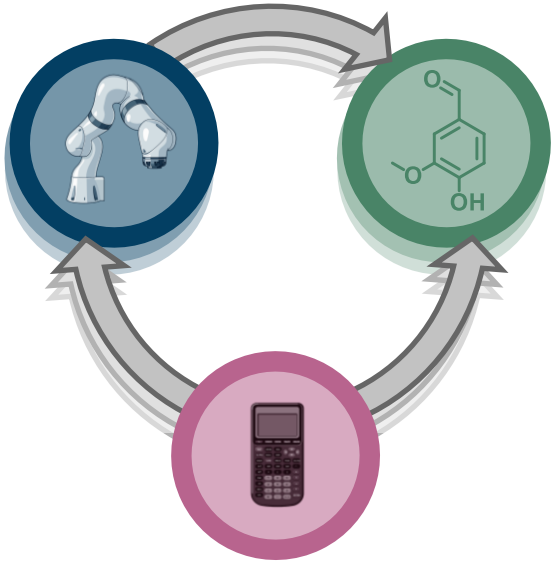
CONCLUSION

- **Machine learning workflow** generates promising candidates for **tailor-made derivatisation reagents** based on an **experimental foundation**.
- Workflow allows the incorporation of **sustainability considerations.**
- Filtering by **Hazard score** leads to a minor decrease in **analytical performance values.**
- **Adaptable** to different **analytical conditions** and analytes.

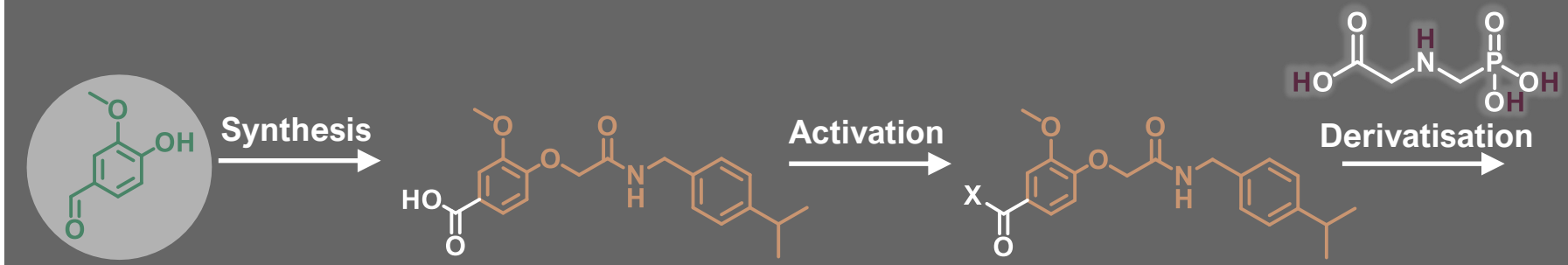


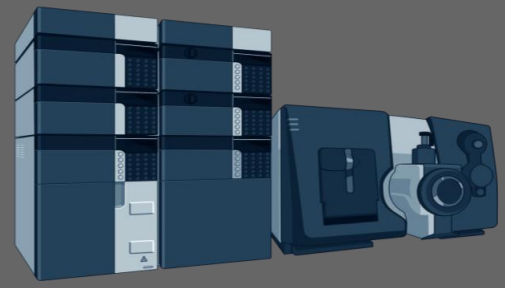
CONCLUSION

- **Machine learning workflow** generates promising candidates for **tailor-made derivatisation reagents** based on an **experimental foundation**.
- Workflow allows the incorporation of **sustainability considerations.**
- Filtering by **Hazard score** leads to a minor decrease in **analytical performance values.**
- **Adaptable** to different **analytical conditions** and analytes.



EXPERIMENTAL VERIFICATION ONGOING:



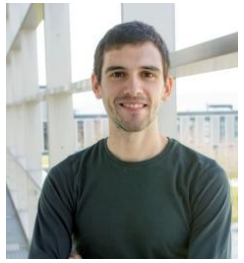


Detection and Quantification

ACKNOWLEDGEMENTS


Anneli Kruve


Berit Olofsson


Miguel Rivero-Crespo


Helen Sepman


Ida Rahu

MISTRA SAFECHEM

Ziye Zhang

Stockholm University

SUCCeSS

Kruve Lab

JOIN TOMORROW:
SUCCeSS sessions with interdisciplinary topics!

Visit posters:


Louise Malm

NR. 65
Implementing rules for improved quantification of transformation products and metabolites


Iris Hättestrand

NR. 43
Evaluation of generative models for structural elucidation of chemicals from mass spectrometry data

Some Figures were created with BioRender.com.