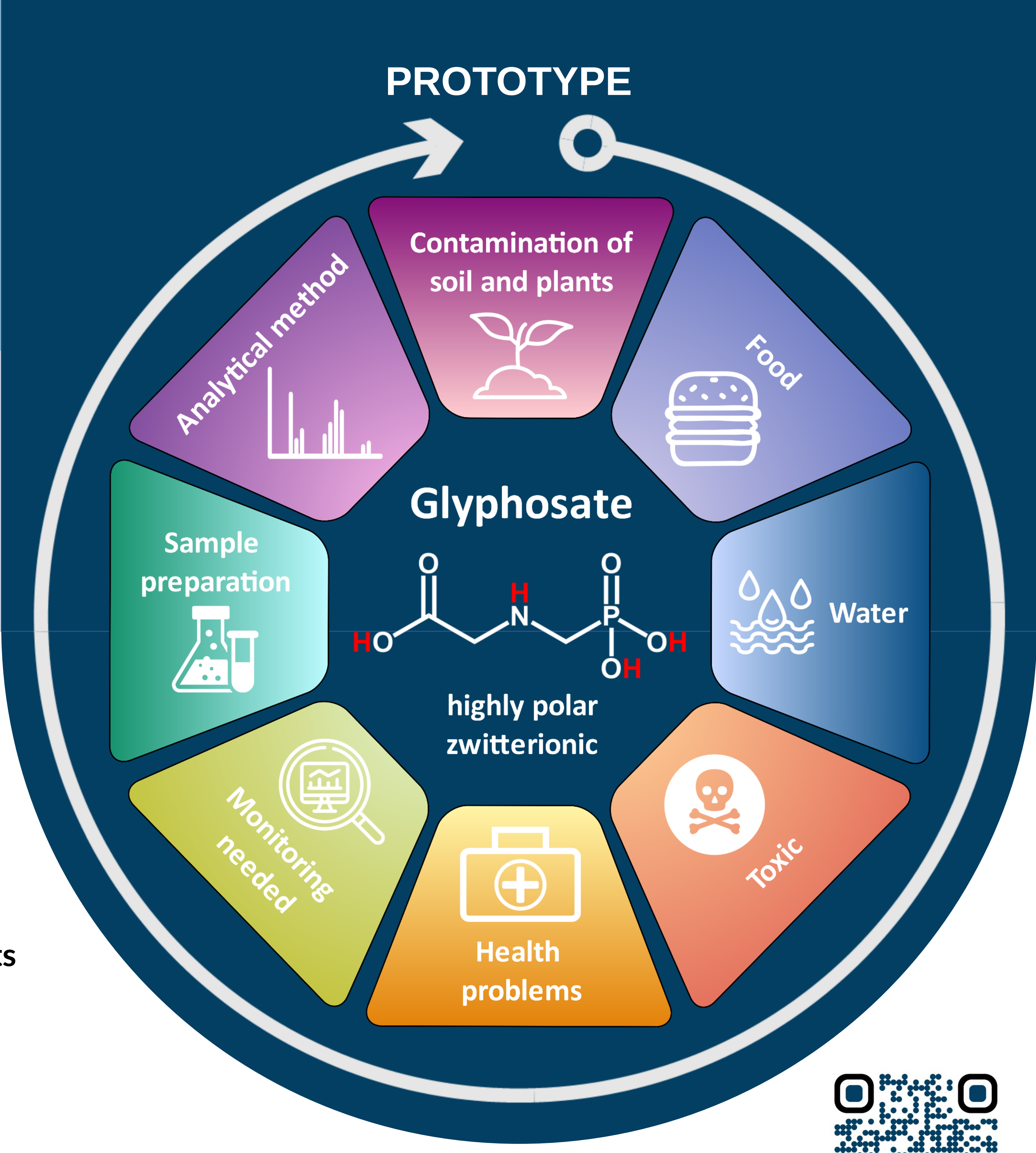
Sustainable design of chemical reagents for the sensitive detection of pesticides using a machine learning workflow

Henrik Hupatz, Miguel Rivero-Crespo, Berit Olofsson, Anneli Kruve

henrik.hupatz@mmk.su.se

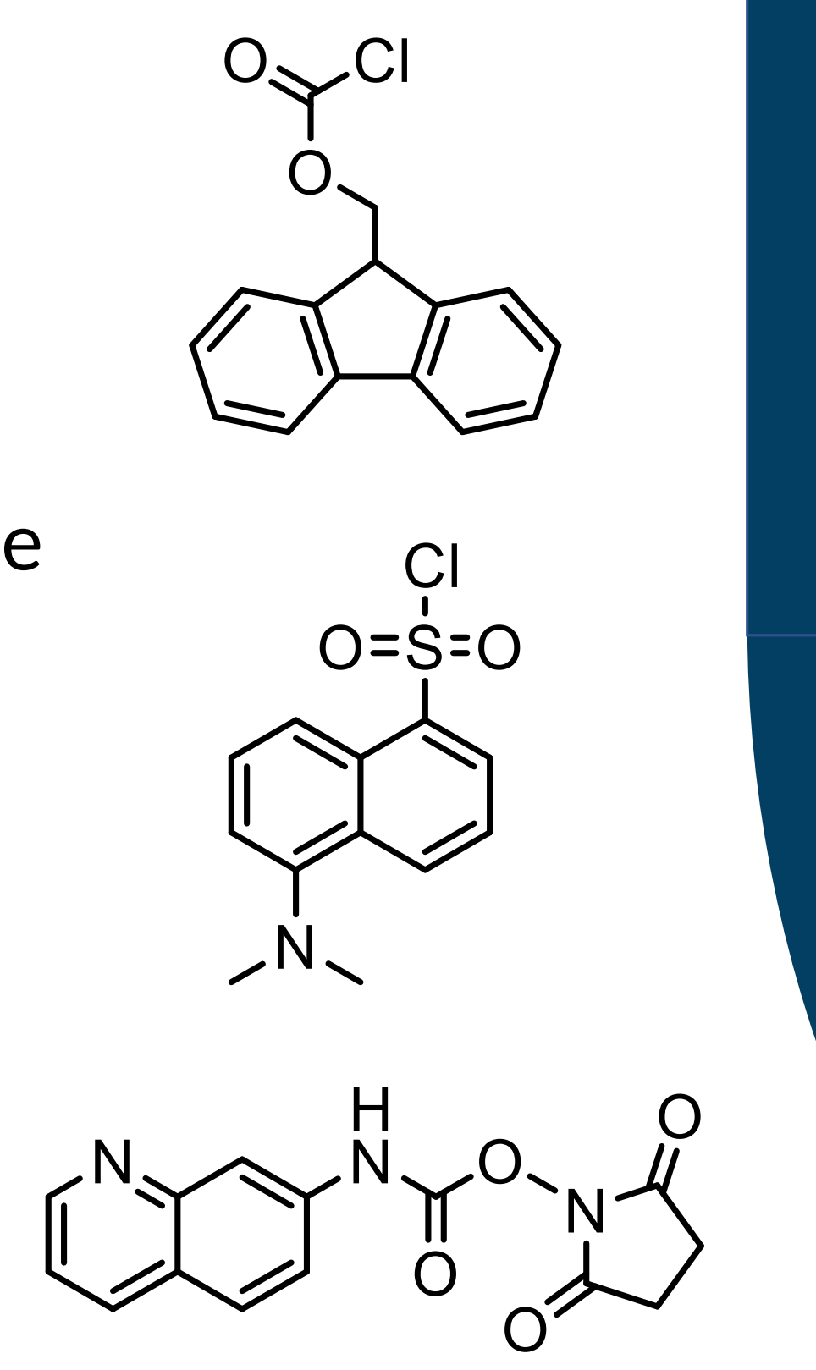
Stockholm University Center for Circular and Sustainable Systems (SUCCeSS), Stockholm University, 10691 Stockholm, Sweden



Background

6 CLEAN WATER AND SANITATION 14 LIFE BELOW WATER

- **Accumulation** of polar **pesticides** in the environment is **threatening** water quality.
- **Mass spectrometry** (MS) coupled with **liquid chromatography** (LC) is a sensitive analytical method for various analytes.
- **Derivatization** reagents facilitate the analysis of **highly polar** small molecules.
- Developing **tailor-made** reagents is a **time-consuming** and **resource-intensive** process.
- **Inverse design** approach is promising for **accelerating the discovery** while **increasing efficiency** and **sustainability.**



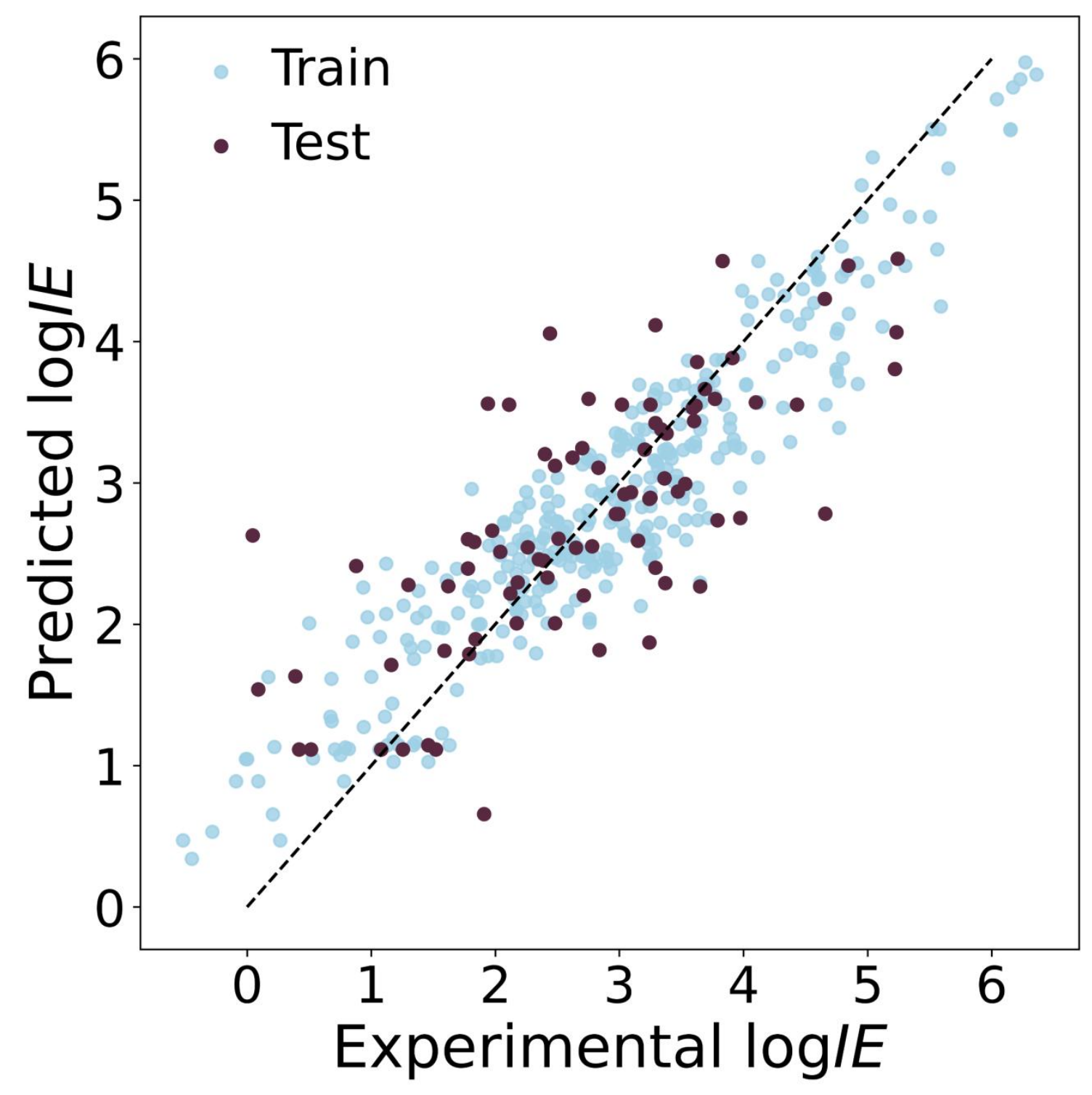
Commercial reagents

Generative Model

- **REINVENT 4** is applied.[1]
- **Prior:**
  Reinvent trained on PubChem 1.8.1 and **Libinvent** trained on ChemBL 27.
- **Diversity Filter:**
  Identical **Bemis-Murcko Scaffold, bin size = 10** and threshold score ***t***.
- **SMARTS Filter:**
  no primary/secondary amine; one **-COOH**; one **vanillin** (reinvent). Libinvent **vanillin** as **scaffold**.

Scoring

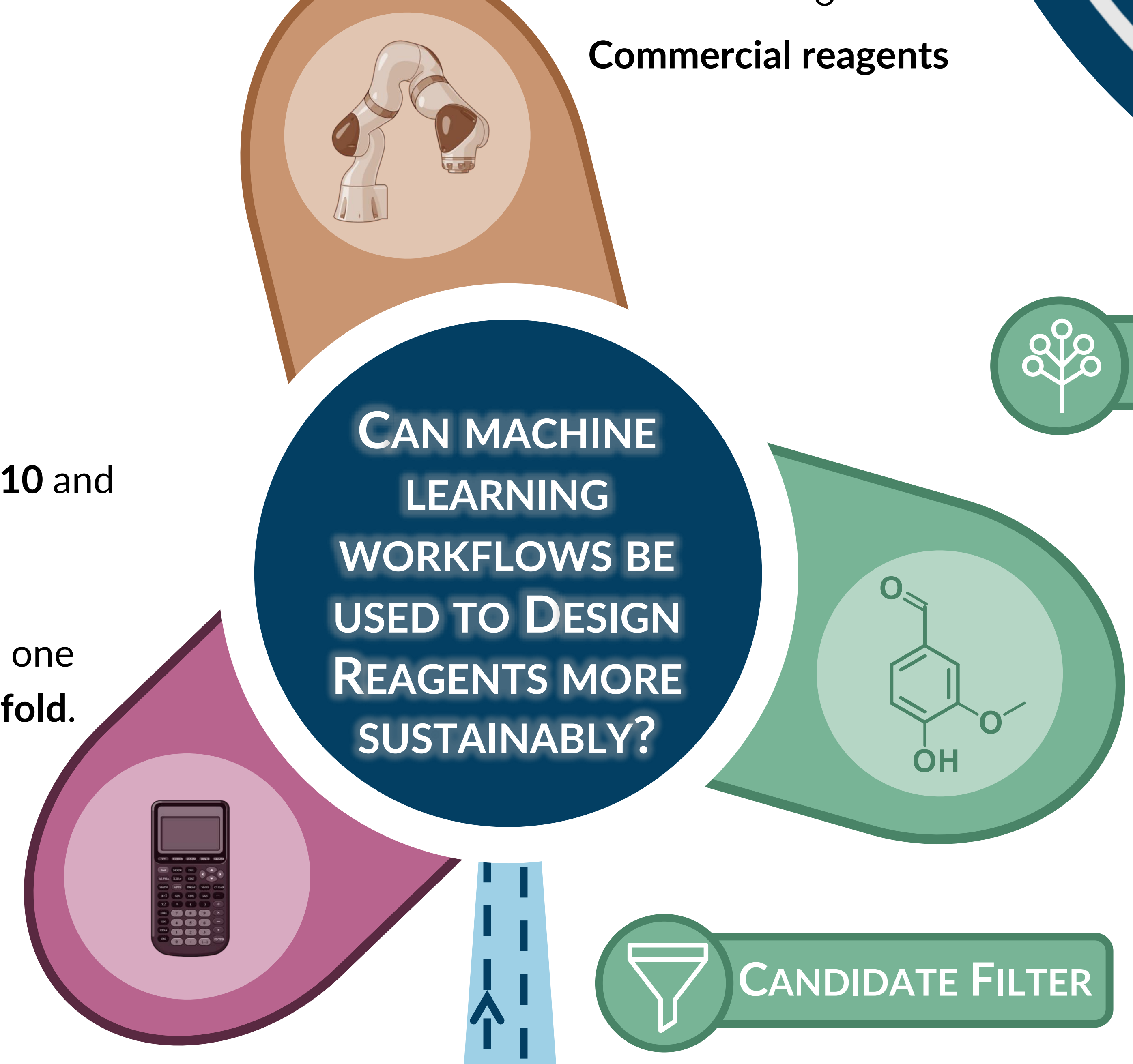
- **Scoring Components:**
  0.20· logP: desired range 2–4.
  0.35· SAScore: <2.5.
  0.45· log*IE*: >3.5. Generated structures are **connected** to **glyphosate** and log*IE* is **predicted**.
- **Property Prediction Model:**
  Dataset: Ionization efficiency log*IE* of **419 chemicals**.[2]
  Representations: MACCS keys, Rdkit descriptors, Mordred descriptors, ECFP6 fingerprints.
  Algorithms/models: XGBoost, RF, SVR, Chemprop.



Many combinations **performed similarly** (RMSEs on test set 0.77–0.82). **XGBoost** and **MACCS** keys were selected for **scoring** the generated structures due to their short **computation time**.

Hazard Prediction

- **In *silico*** predictions of **biochemical activity** for endpoints (23 endocrine disruption and 3 CMR toxicity) and of **persistence, biodegradation** and **bioconcentration** are combined to a **Hazard Score**. [3] (Models within Mistra SafeChem)



Can machine learning workflows be used to design reagents more sustainably?

Method Comparison

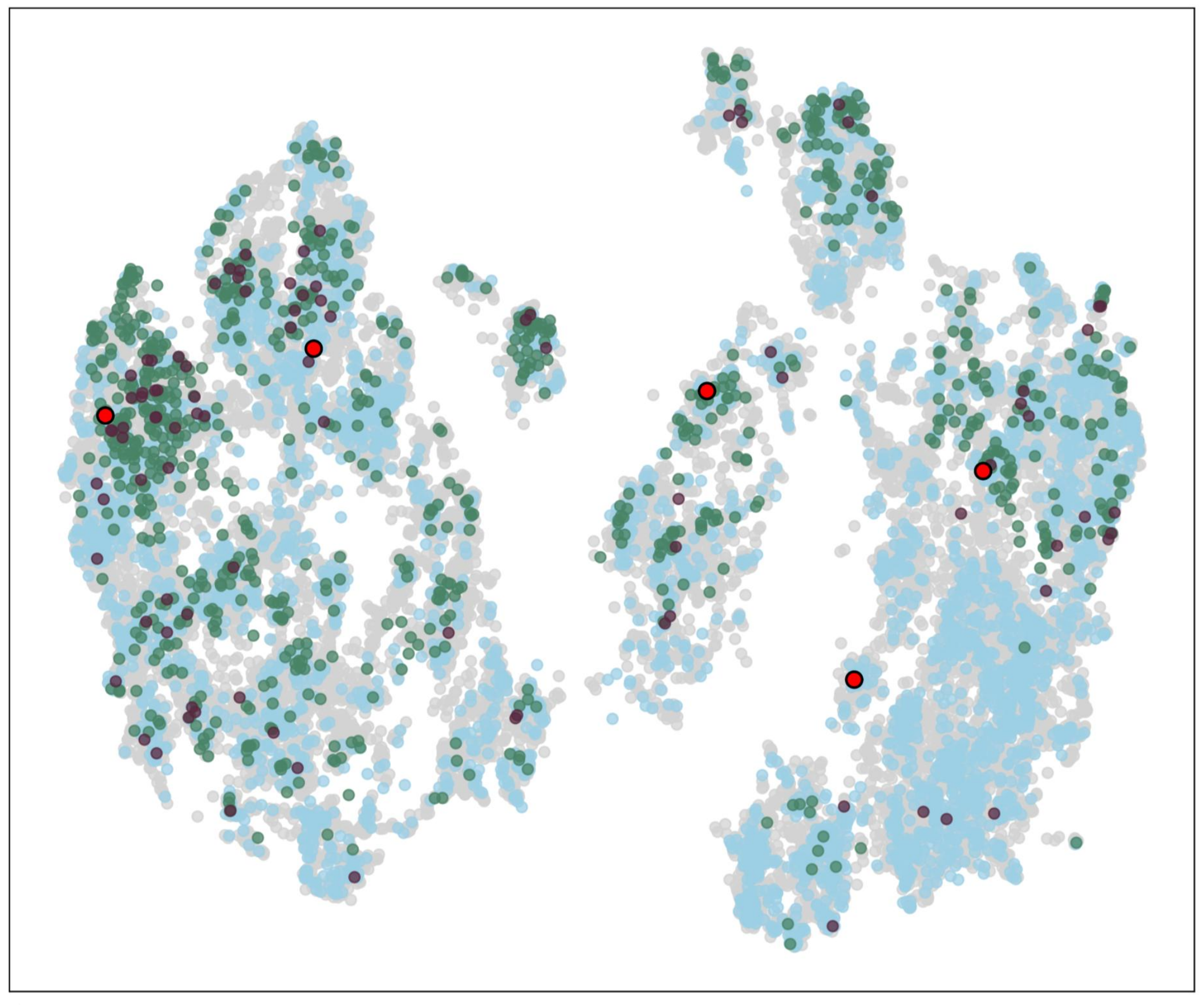
Prior	Vanillin	*t*	$N_{(S>0.9)}$	$SAS_{(1\%)}$	log*IE*$_{(1\%)}$	$S_{(1\%)}$
Reinv	No	0.9	2187	1.52	4.09	0.93
Reinv	Yes	0.9	6513	1.72	3.99	0.93
Reinv	Yes	0.8	118436	1.71	4.05	0.94
Reinv	Yes	No	72554	1,77	4,10	0,95
Libinv	Yes	0.9	626	1.62	4.06	0.93
Libinv	Yes	0.8	3813	1.74	4.05	0.92
Libinv	Yes	No	25736	1.74	4.10	0.93

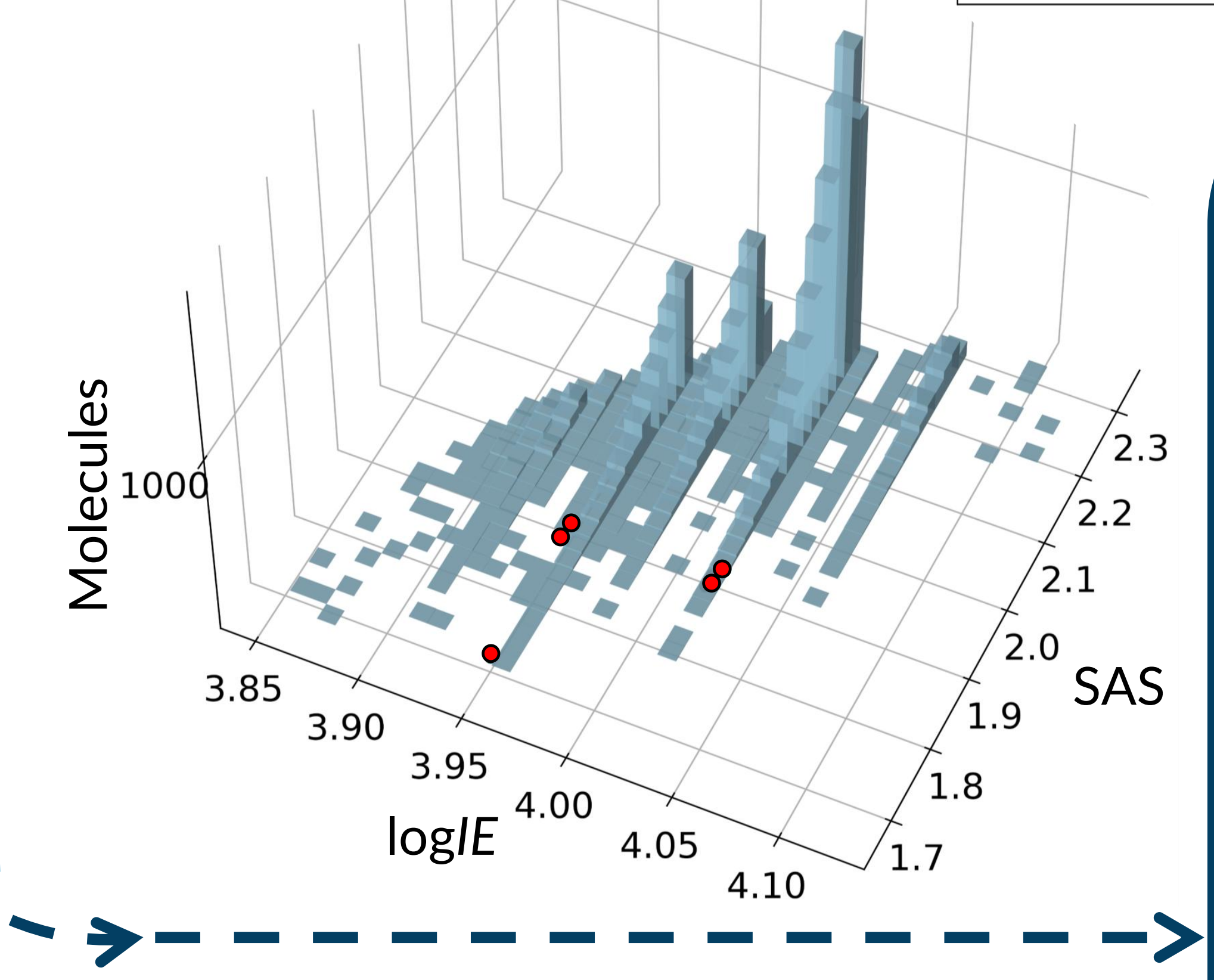
Batch size = 1024; steps = 1000; $N_{(S>0.9)}$: Number of structures with Score (S>0.9); S, SAS, log*IE* averaged over 1% of $N_{(S>0.9)}$.

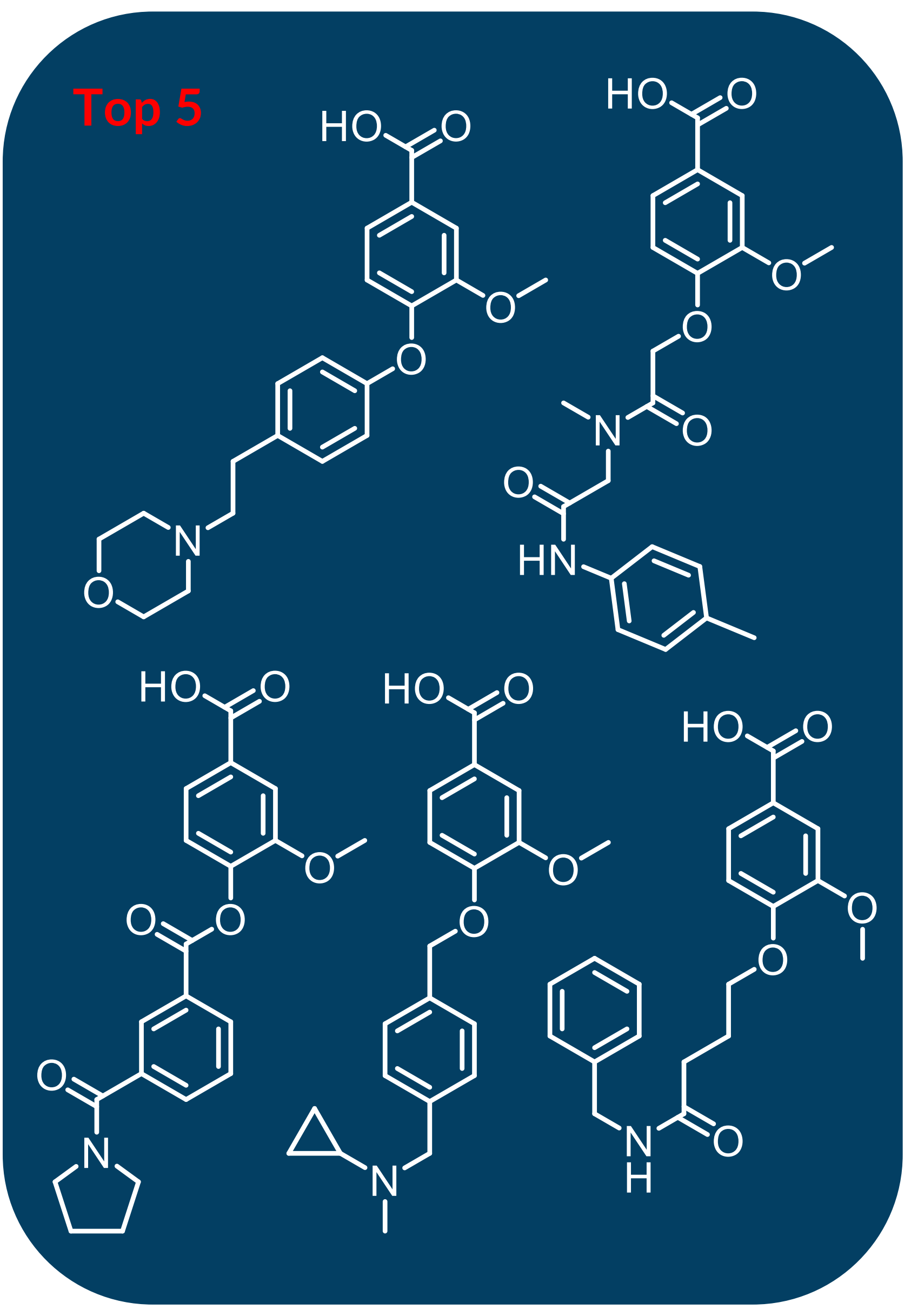
Candidate Filter

20,000
Highest S
5000
Lowest Hazard Score
1000
Lowest SAS
100
Highest Tanimoto distance
5

UMAP Visualization of 20,000 highest scoring molecules based on ECFP4 fingerprints







Experimental Foundation

Renewable resources	Derivatization reagent	Pesticide sample		LC	MS
Lignin *Included in synthesis*	**Reagent synthesis** *Easy and cheap*	**Sample preparation** *Low hazard*	*Suitable reactivity*	**Separation** *Suitable retention time*	**Detection** *High ionization yield*
Vanillin	SAScore	Hazard score	-COOH group	logP	logIE

References:

[1] H. H. Löffler et al., *J. Cheminform.* **2024**, *16*, 20.

[2] H. Sepman et al., *Anal. Chem.* **2023**, *95*, 12329.

[3] E. Söderberg et al., *Green. Chem.* **2024**, *26*, 11147.

Some figures were created with biorender.com

Kruve lab

Stockholm University

SUCCeSS