In silico generated reagents for detection of pesticides using mass spectrometry: An out-of-distribution task

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

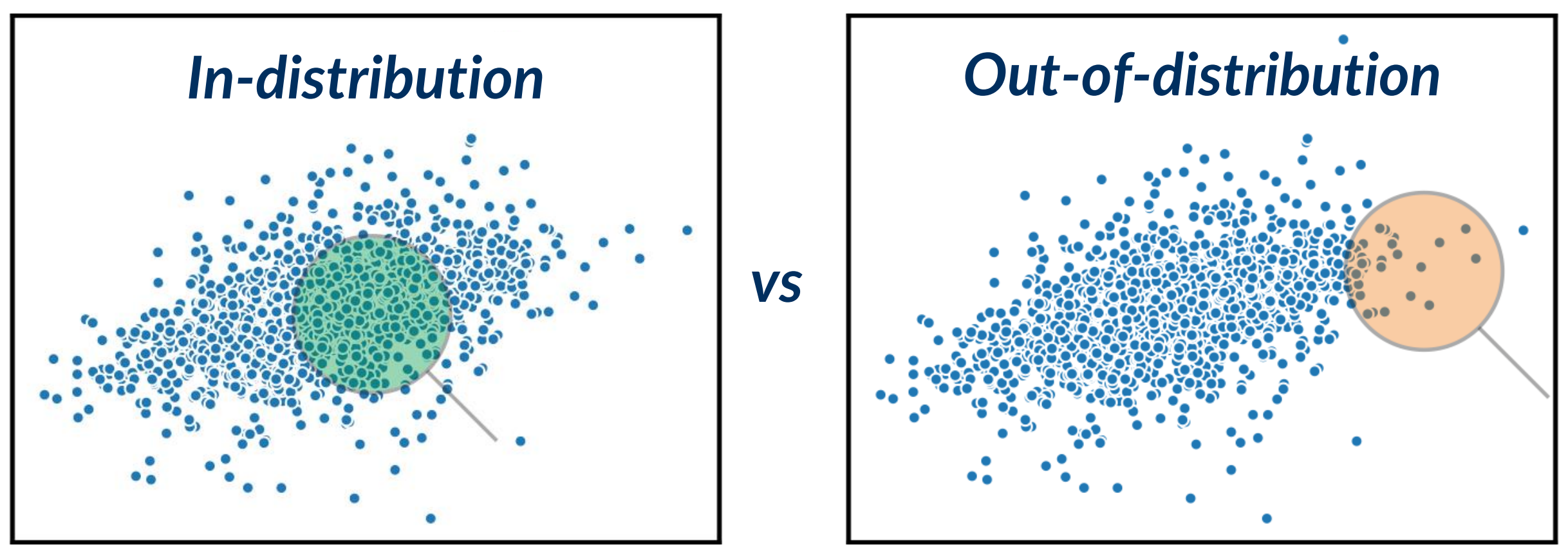
henrik.hupatz@mmk.su.se

Stockholm University Center of Circular and Sustainable Systems (SUCCeSS), Svante Arrhenius väg 16C, 114 16 Stockholm, Sweden

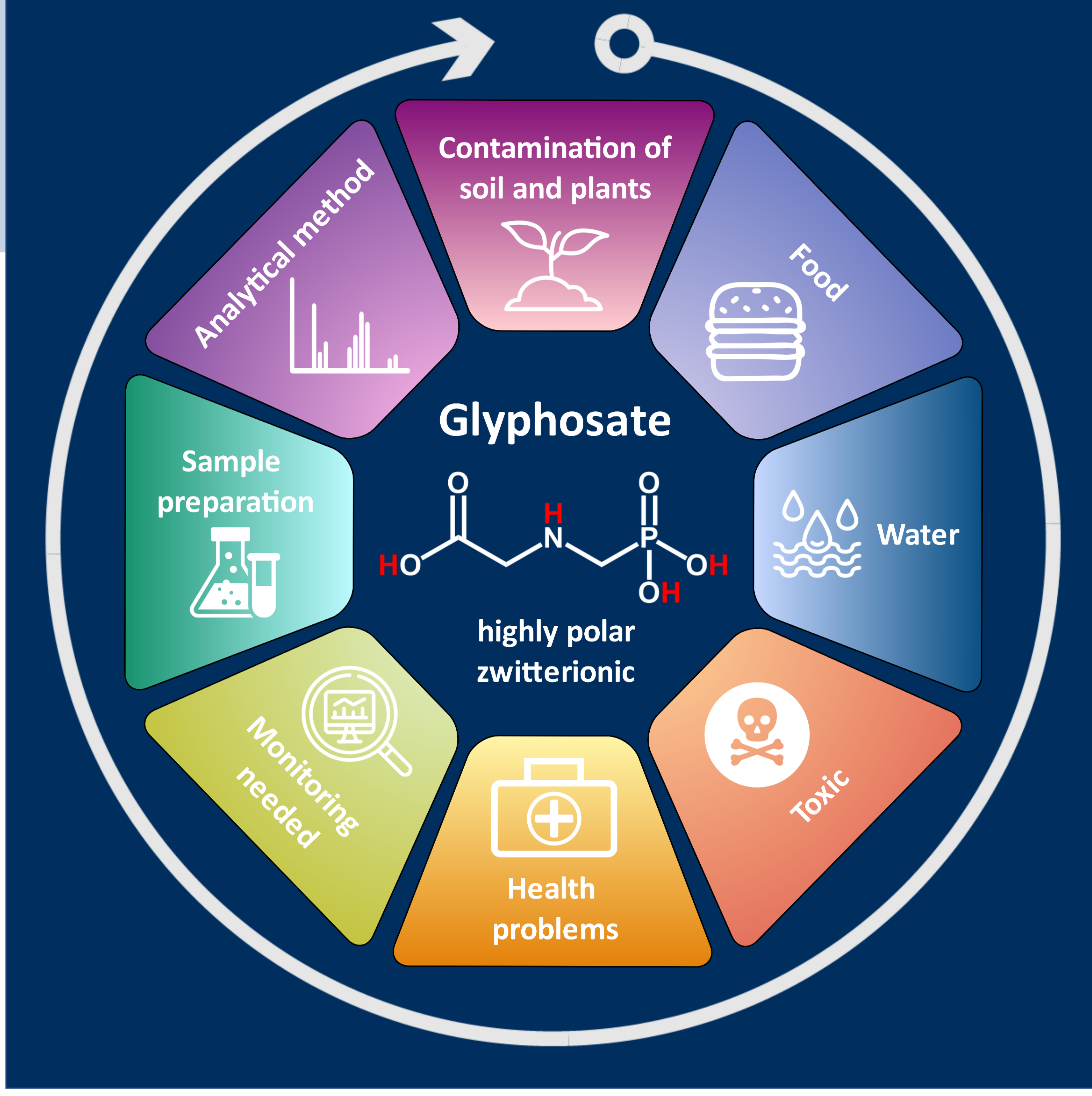
Background

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 **chromatography** is a sensitive analytical chemistry method for a wide range of analytes.
- **Derivatization** facilitates the analysis of challenging **highly polar** small molecules.
- **Generative modeling** enables the **inverse molecular design** of easily **accessible** derivatizing reagents with **desired** chemical **properties**.
- **In-silico generated** reagents require **exploration** of **unknown chemical space** to predict chemical properties, such as ionization efficiency (*IE*).
- **IE prediction model** can be **optimized** for **out-of-distribution (OOD) data**.



PROTOTYPE



Computational Methods

- 658 models were trained on the **1147 chemicals** of an *IE* dataset to predict log*IE*.

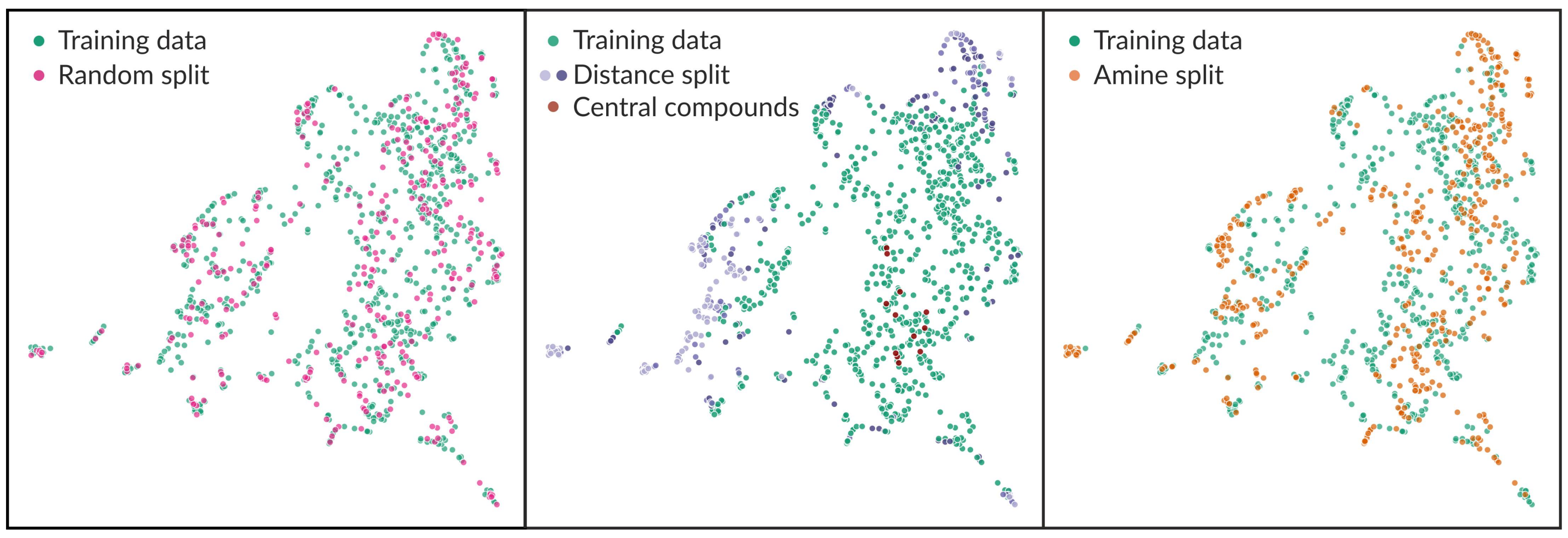
Input Features [4]

Train/Test Splitting [3]

Feature Cleaning [9]

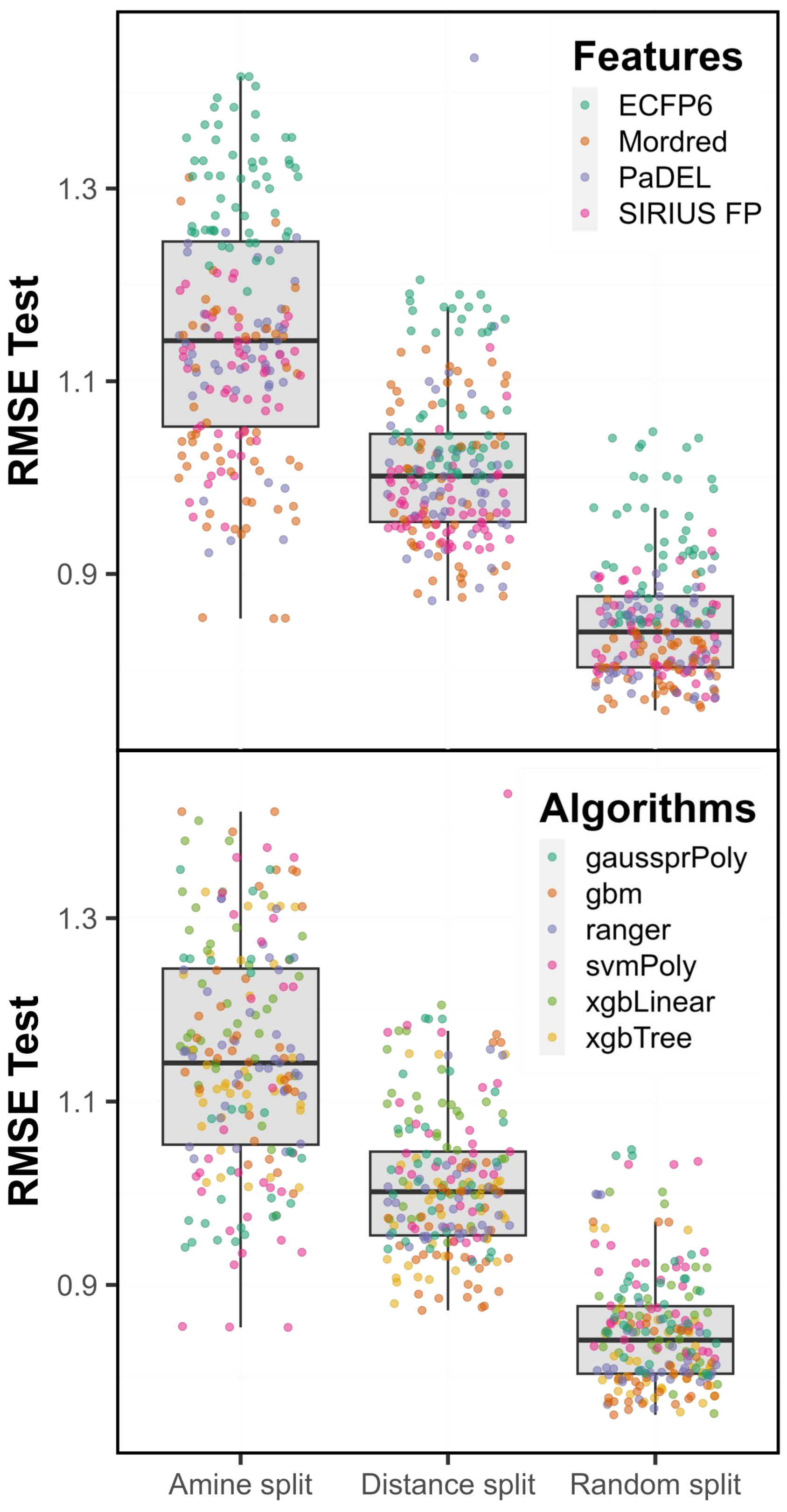
Algorithms [6]

Chemical Space Visualization and Data Splitting

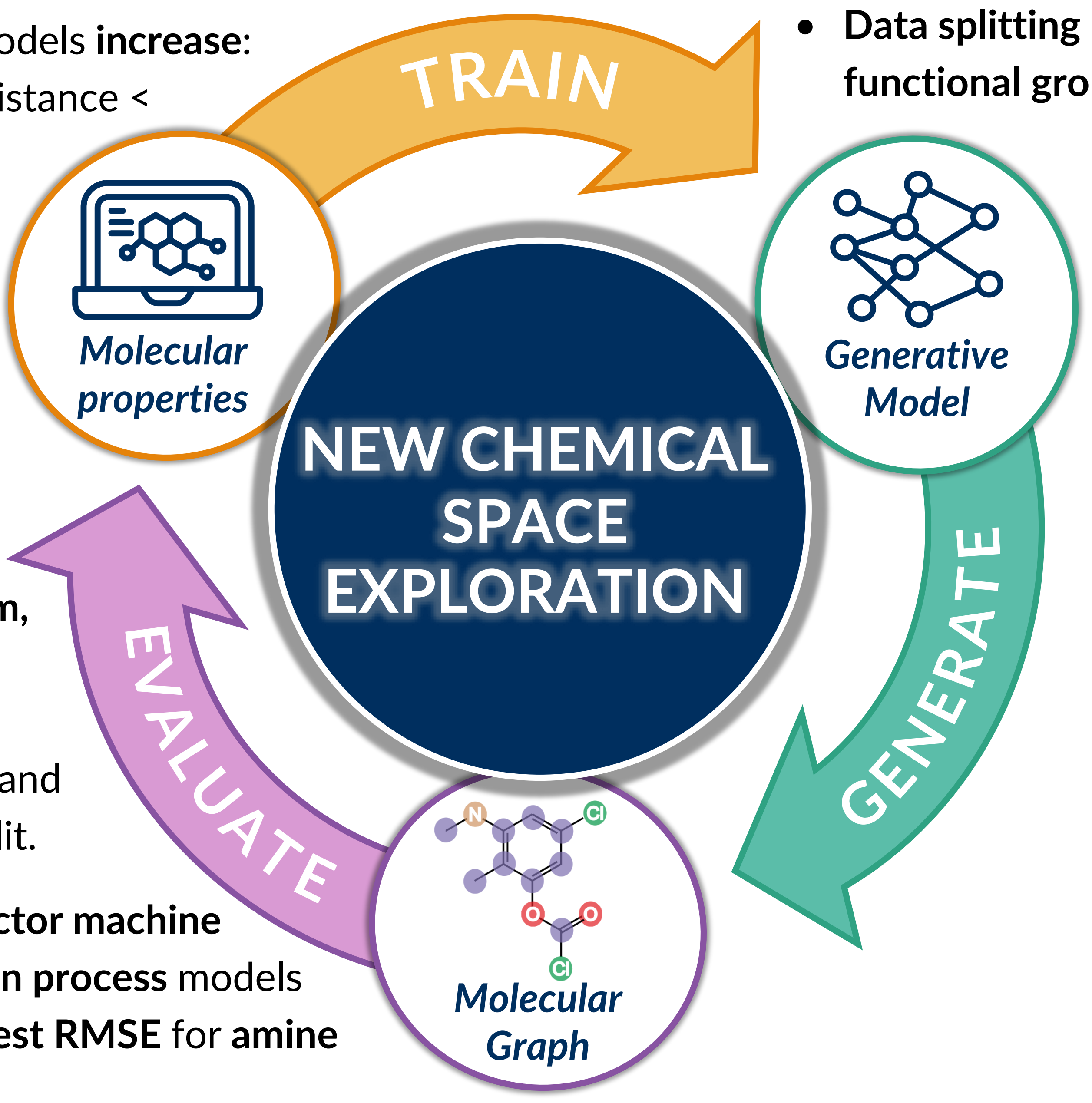


- **Myopic common edge subgraph distances**[1] (MCESD) were used to characterize **chemical similarity**.
- **UMAP reduction** to visualize chemical space of IE dataset.
- Most **characteristic** (“central”) **compounds** were identified by **lowest average MCESD**.
- **Data splitting** based on **MCESD** to central compounds and most **abundant functional group** (amine) to investigate **OOD performance** of models.

ID vs OOD Performance

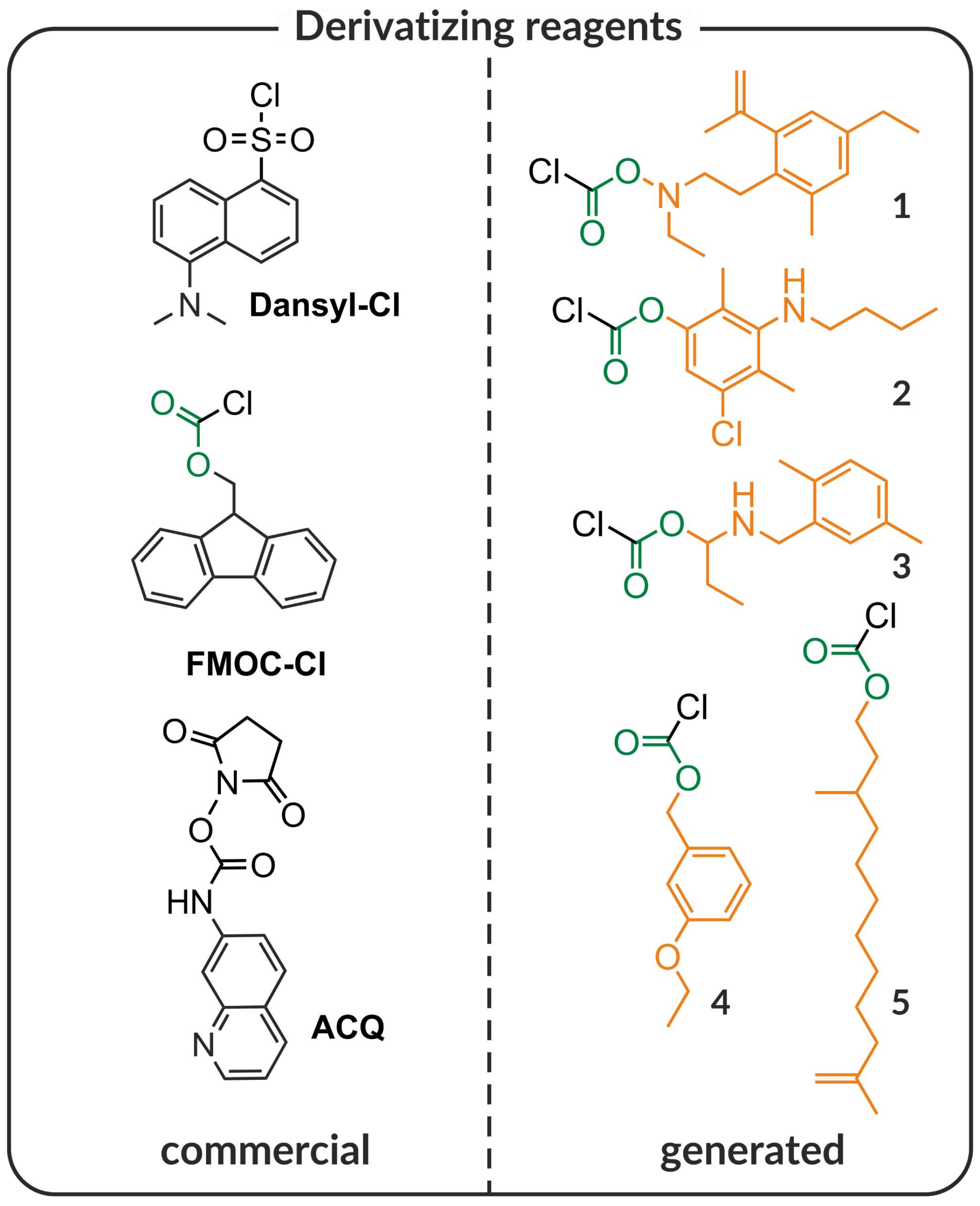


- **Median** and **variance** of **RMSE** of models **increase**: random < distance < amine
- **Mordred** descriptors perform **better** over all **splits**.
- **Tree-based** models **(gbm, xgbTree) outperform** on **random** and **distance** split.
- **Support vector machine** and **gaussian process** models exhibit **lowest RMSE** for **amine** split.

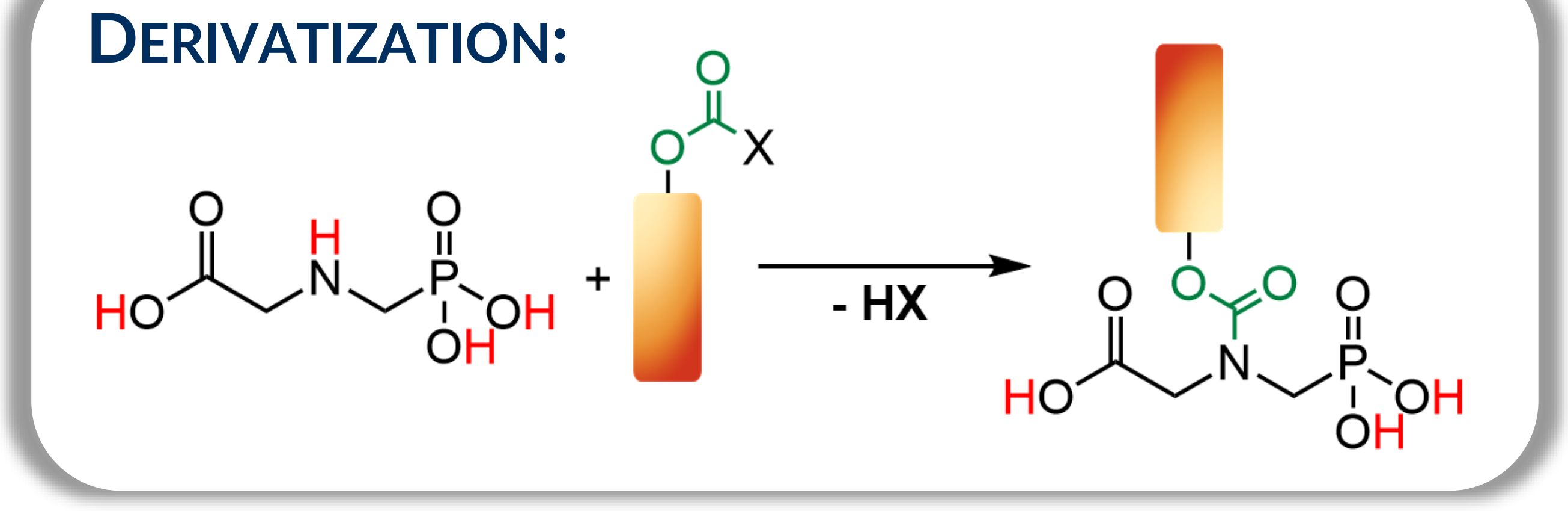


Monte-Carlo Tree Search

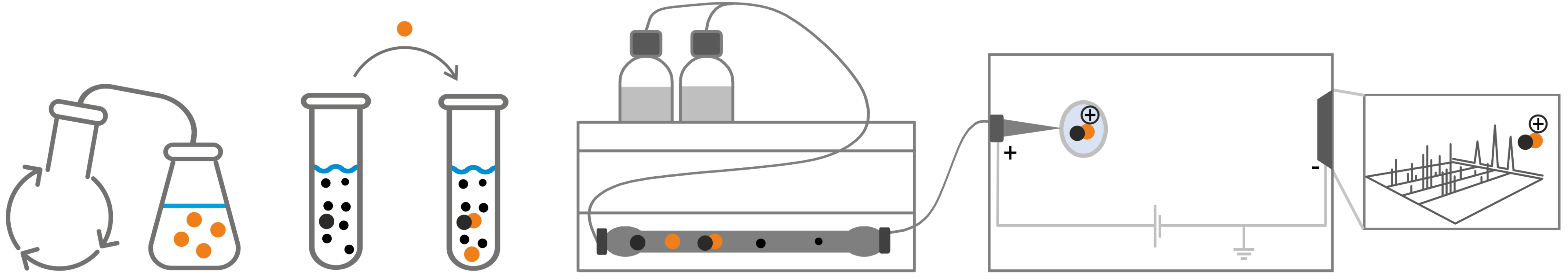
- **Molecular graphs generated** by MCTS from **carbon seed** to molecular weight of **400** with MCTS **depth of 2** and **width of 12**.
- **Evaluation function:** *Ev* = log*IE* – SAScore.
- For each MCTS the structure with highest *Ev* was picked.



Derivatization:



Experimental Properties



Synthesis
Easy and cheap, Non-toxic

SAScore

Derivatization
Quick reaction time

Connectorunit

Separation
Suitable retention time, Stable

logp

Detection
High ionization yield

log*IE*

[1] *For references, additional information and a poster copy:*

Kruve lab