MAPPING REACTION CONDITIONS WITH MACHINE LEARNING

Matt Ball[1,2], Felix A. Faber[3], Cassie Pratley[4], Thierry Kogej[1], Dragos Horvath[2], Alexandre Varnek[2], Giulia Bergonzini[4], Christof M. Jäger[3], Mikhail Kabeshov[1]

1 MolecularAI, Discovery Sciences, AstraZeneca R&D; Gothenburg, 431 83, Sweden
2 Laboratory of Cheminformatics, University of Strasbourg; Strasbourg, 67081, France
3 Predictive Science, Digital and Automation, Pharmaceutical Sciences R&D, AstraZeneca, Gothenburg, 431 83, Sweden
4 Compound Synthesis and Management, Discovery Sciences, AstraZeneca R&D; Gothenburg, 431 83, Sweden

Paper Code

Introduction

Reaction condition selection is critical to synthetic success, yet identifying optimal combinations of conditions remains challenging[1, 2]. Machine learning for reaction prediction has focused primarily on **reactant representations**; condition encoding has received comparatively less attention[3, 4]. Frequently used encodings based on molecular fingerprints **may struggle to extrapolate** to unseen components, whilst physicochemical descriptors can capture reactivity-relevant information but understanding which parameters are relevant *apriori* is not straightforward.

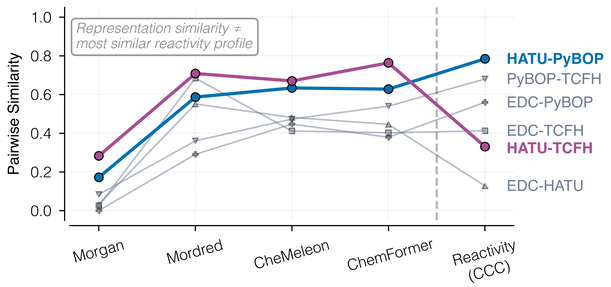
The rise of **High-Throughput Experimentation** (HTE) demands efficient, intuitive tools for reaction data analysis and condition design — motivating a **learnt representation that encodes reactivity patterns by construction**.

We present a deep learning framework that learns **reactivity-focused representations** of reaction conditions, using **Amide Coupling** as a proof-of-concept. Component embeddings are trained with **expert-informed heuristic labels** and refined via **representation learning** on literature yield data. The resulting fingerprints **co-locate conditions with similar outcomes** in embedding space, correlate with known physical properties, and enable **diversity-driven HTE plate design** that accelerates hit identification.

1 Similarity In Existing Representations Fails To Capture Reactivity Patterns

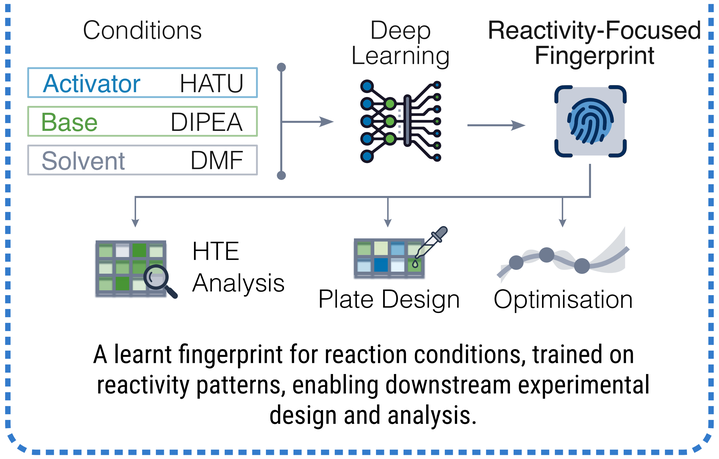
Activator Pairwise Similarity Across Representations

Similarity Metrics Normalized to 0-1 Among Activators in Training Data



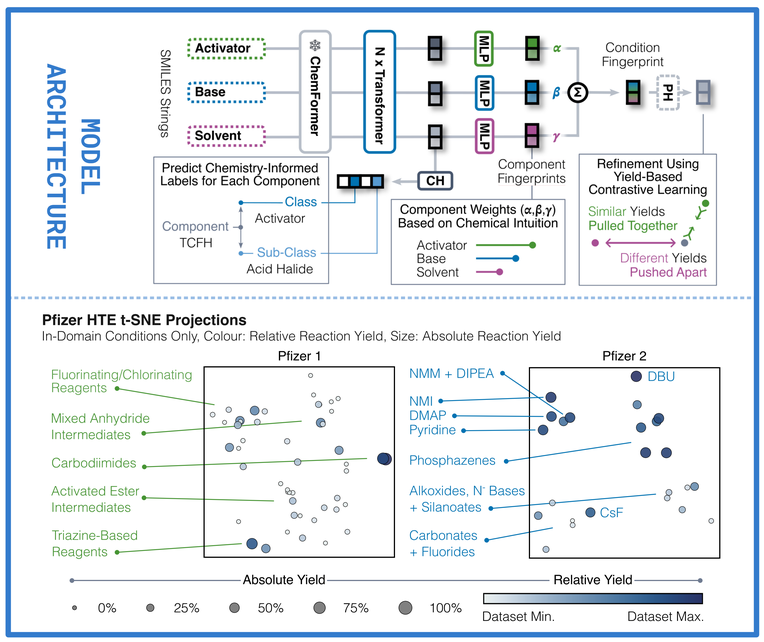
'Reactivity' similarity measured as the **Concordance Correlation Coefficient (CCC)** between **pairs of yields from the same transformation using different conditions**[5].

Visual Abstract



A learnt fingerprint for reaction conditions, trained on reactivity patterns, enabling downstream experimental design and analysis.

2 Training Objective: Co-Locating Conditions with Similar Yields



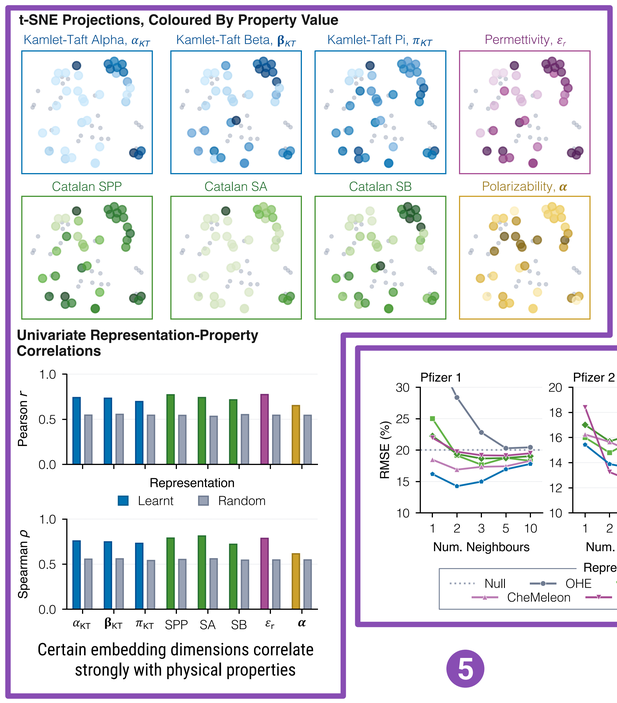
How?

- Condition fingerprint = **linear combination** of component embeddings, trained to **predict expert-assigned functional class labels** for each component type.
- Representation learning **pulls together conditions with similar outcomes** (Δyield < 15%) for similar transformations[6].
- Validated via kNN local-reactivity: **nearby conditions in fingerprint space should predict similar yields** on held-out HTE datasets[7,8].

Key Findings

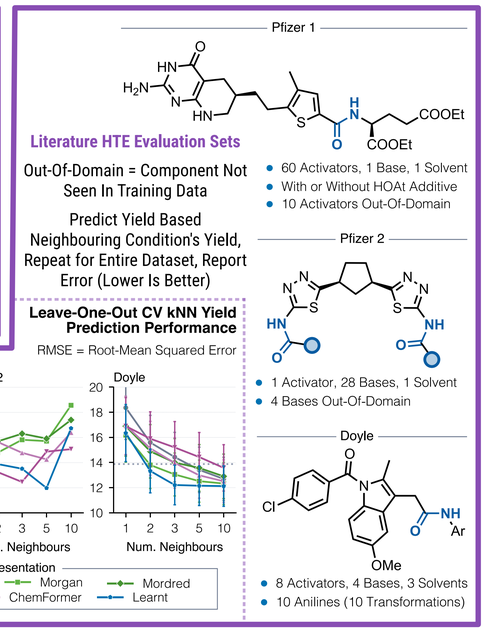
- Learnt fingerprints are **intuitive**: individual dimensions correlate strongly with established solvent properties (Kamlet-Taft, Catalán parameters).
- The **local neighbourhood of the learnt embedding space captures reactivity** more effectively than other representations across three held-out HTE datasets (using LOOCV, kNN models).
- **Diverse plate sampling** in the learnt space increases maximum yield and hit probability in early-stage condition screening.

3 Physical Properties Of Solvents Are Learnt Implicitly

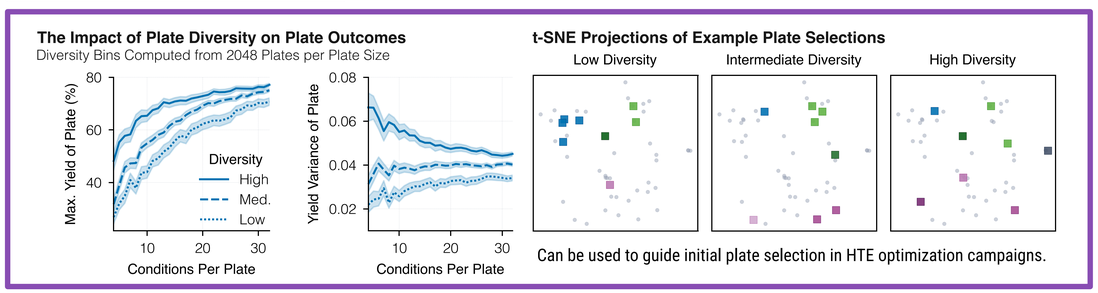


Certain embedding dimensions correlate strongly with physical properties

4 Embedding Local Environment Captures Reactivity



5 Diversity-Driven Plate Design Accelerates Hit Identification



Can be used to guide initial plate selection in HTE optimization campaigns.

Future Work

Explore further applications of the learnt fingerprints in reactivity tasks, e.g. optimization campaigns.

Extension of the framework to other reaction types e.g. Buchwald-Hartwig or Suzuki-Miyaura couplings.

Inclusion of non-chemical parameters as input, like temperature or pressure.

References

1 *J. Am. Chem. Soc.* **2022**, 144, 4819–4827.
2 *Chem. Sci.* **2025**, 16, 17523–17541.
3 *Nat Mach Intell* **2021**, 3, 144–152.
4 *J. Am. Chem. Soc.* **2025**, 147, 8959–8968.
5 *Nat Commun* **2025**, 16, 4522.
6 *J Cheminform* **2025**, 17, 177.
7 *Nature* **2024**, 626, 1025–1033.
8 *Nat. Chem.* **2024**, 16, 633–643.

Funding

This project was partially funded from European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Actions, grant agreement No. 101120466 ("AiChemist - Explainable AI for Molecules")