SmartReact

Scalable forward-synthesis enumeration built on interpretable SMARTS reaction templates

Fabian P. Krüger[1,2,3] Thierry Kogej[1] Mikhail Kabeshov[1] Ola Engkvist[3,4] Igor V. Tetko[3]

[1]AstraZeneca R&D, Mölndal [2]Technical University of Munich [3]Helmholtz Munich [4]Chalmers University of Technology

Open source · Apache 2.0

The problem

Forward enumeration predicts what products a set of building blocks can make, which is the backbone of library design and synthesis-aware generative models.

But template libraries force a hard trade-off:

- **Small expert sets**: precise but limited coverage.
- **Auto-extracted sets**: broad but redundant and hard to audit.
- **Scaling**, matching thousands of templates by subgraph isomorphism is *slow*.

The key idea: classify once, filter exactly

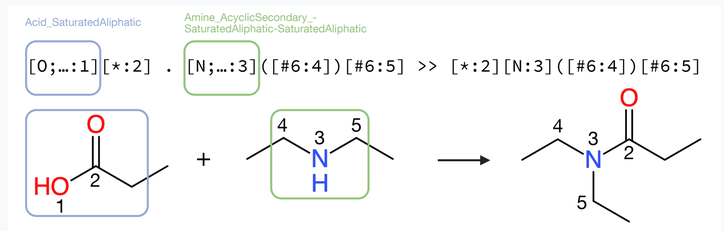
Every template is grounded in **SMARTS-RX**, a hierarchical functional-group ontology.

The same ontology that gives each template a precise chemical environment also allows pre-filters incompatible reactant pairs:

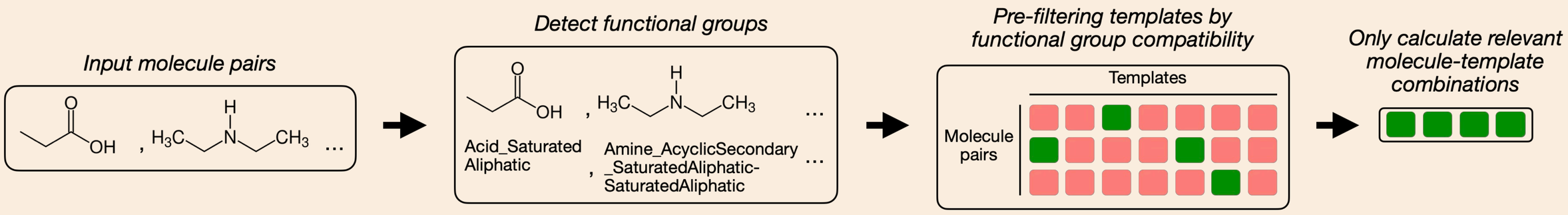
- Classify each molecule **once** against 406 patterns.
- A missing key is an **exact** incompatibility, not a heuristic.
- Run costly subgraph matching **only** where warranted.

Interpretable & extensible

Every prediction traces back to a **named reaction** and an explicit set of fine-grained functional-group requirements, which is easy for chemists to audit and extend. Example:



How it works — a four-stage pipeline



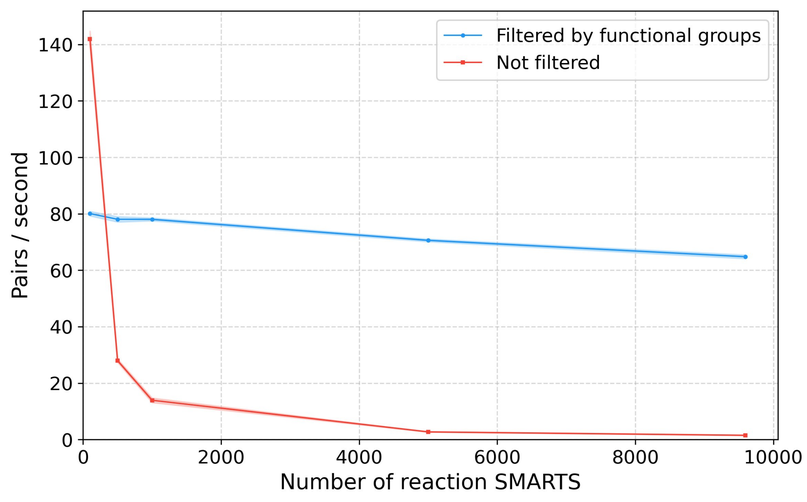
406 functional-group patterns gate **9,594 templates**: the structural asymmetry means a cheap key lookup discards almost all templates *before* any RDKit matching.

Results — three sources of speed, compounding

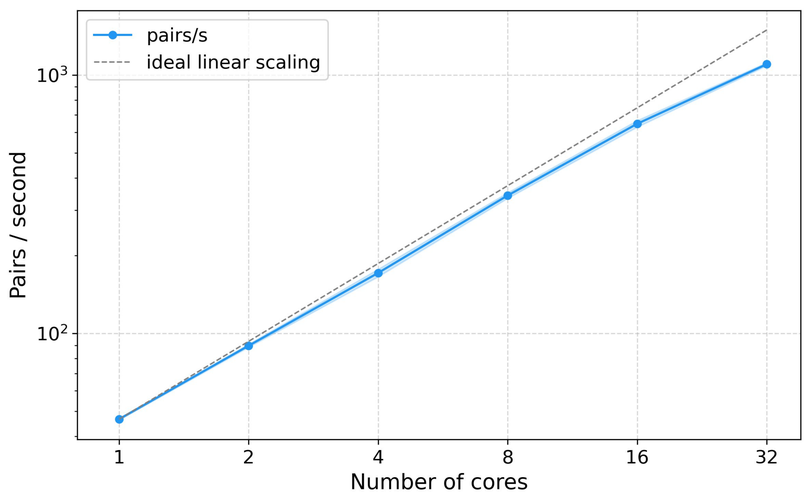
43 ×
exact key-based filtering
vs. brute force (full library)

~24 ×
near-linear parallel
scaling on 32 cores

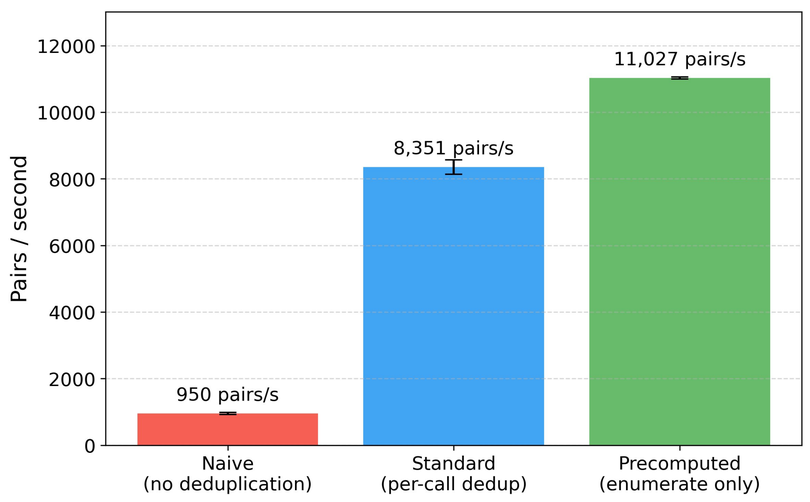
11,000 ×
deployment-scale throughput
vs. single-core brute force



(a) **Filtering.** Throughput stays flat as the library grows; brute force collapses.

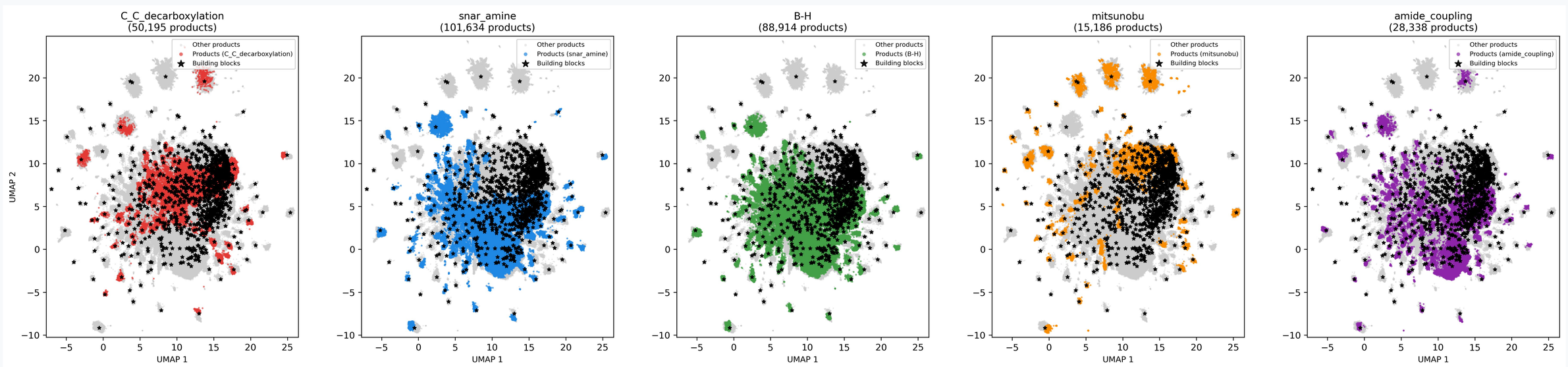


(b) **Parallel scaling.** Near-ideal up to 32 cores.



(c) **Classify-once.** 950 → 8,351 → 11,027 pairs/s.

Case study — 525,130 products in 31 s



All 499,500 pairs from **1,000 building blocks**, 32 cores. Reactions occupy **distinct** regions of chemical space (UMAP of ECFP); mechanistically related ones (SNAr amine & Buchwald–Hartwig) overlap — as a chemist would expect. Black stars: input building blocks.

Get started

```
pip install smartreact
```

github.com/MolecularAI/smartreact | Apache 2.0 license | Pure Python + RDKit

Scope: reliable, high-throughput enumeration of well-established medicinal-chemistry reactions — complementary to template-free deep-learning models.