XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery · BCADD-2026

CRUSH:

A CHEMISTRY-AWARE MOLECULAR FRAGMENTATION ALGORITHM FOR EXPANDED CHEMICAL SPACE EXPLORATION

Edgar López-López* · José L. Medina-Franco · Filip Miljković

* edgar.lopez@uu.se

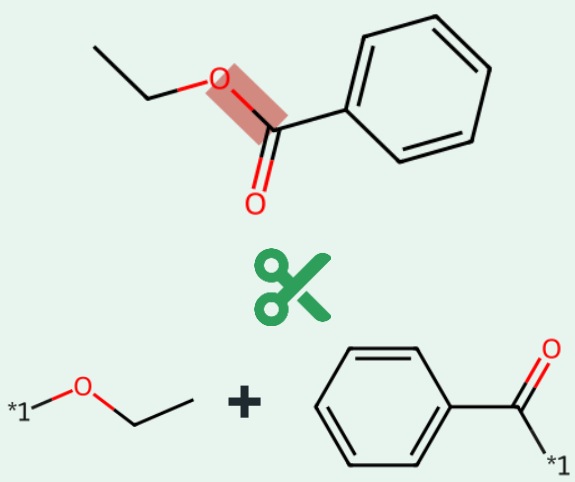
J. Cheminform. 18, 78 (2026) · doi:10.1186/s13321-026-01239-w

UPPSALA UNIVERSITET

AstraZeneca

Why do we break molecules?

Fragments are the LEGO bricks of chemical space



Smaller, interpretable pieces that keep their chemical identity

Fragment-based drug discovery

Small pieces sample binding sites efficiently

SAR and scaffold analysis

Which substructure drives activity?

Library design & chemical space

Organize and diversify ever-growing databases

AI / ML

Fragment representations densify and explain data

The classic toolbox: RECAP and BRICS

Expert-derived, retrosynthesis-inspired bond disconnections

RECAP

Lewell et al., 1998

- Cuts bonds commonly formed in organic synthesis (amides, esters, amines, ethers…)
- Tends to keep larger, more complex pieces
- Great for scaffold dissection of drug-like compounds

BRICS

Degen, Zaliani & Rarey, 2008

- Focuses on synthetically tractable building blocks
- Recombination rules → chemically feasible molecules
- Smaller fragments than RECAP

Both are widely used (*e.g.,* RDKit, MORTAR) — but they were designed around medicinal chemistry, and large, systematic comparisons across diverse databases remain scarce.

Bänsch et al., 2023

The gap: chemistry beyond drug-like space

Medicinal-chemistry centric
Rules reflect drug synthesis, not natural products, peptides or food chemistry

C–C bonds largely untouched
Long chains and aryl–alkyl links stay intact

Coarse bond environments
Aliphatic, aromatic and heterocyclic contexts are not finely separated

Chemical families we want to cover

- Drug-like molecules
- Natural products
- Peptides (incl. non-canonical)
- Macrocycles
- Food chemicals

Meet CRUSH

Cleavage Rules Using SMIRKS Heuristics

33

curated SMIRKS cleavage rules

each capped with dummy atoms (*) to track where and why a bond was cut

1

Find every eligible bond

All 33 rules are tested on all bonds — not just the first rule that matches

2

Explore alternative pathways

Cleavage is applied recursively, branching into multiple disconnection routes

3

Keep the richest decomposition

The route with the most retained fragments is selected (complementary set)

SMIRKS in 60 seconds

A text recipe that says: "find this pattern, cut here"

Ester rule

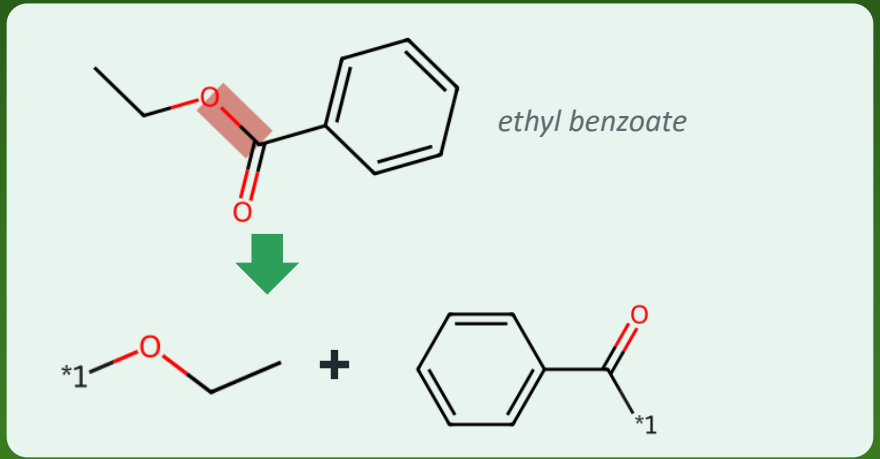
```
[#6:1]!@[C!R:2](=[O:3])!@[O!R:4][#6:5]  >>  [*:1][*:2](=[*:3]).[*:4][*:5]
```

Pattern (left of >>)

- Numbered atoms (:1 … :5) are tracked
- !@ = the bond must NOT be in a ring
- C!R = acyclic carbonyl carbon

Result (right of >>)

The "." separates two fragments; * marks the attachment point



What the 33 rules cover

Classic med-chem disconnections plus much finer bond environments

Classic functional groups

Amides, esters, ureas, carbamates, sulfonamides, ethers, amines — analogous to RECAP/BRICS

Carbon–heteroatom

Aryl C–O, aryl C–N, aryl C–S, aliphatic C–X; aromatic N–aliphatic C

Carbon–carbon

Aliphatic C–C, olefins, alkynes, aryl–alkyl, aryl–vinyl, aryl–aryl

Heterocycle-associated

Lactam N–C, aryl–n, inter-aryl N–N, ring tertiary amine substituents

Sulfur & other motifs

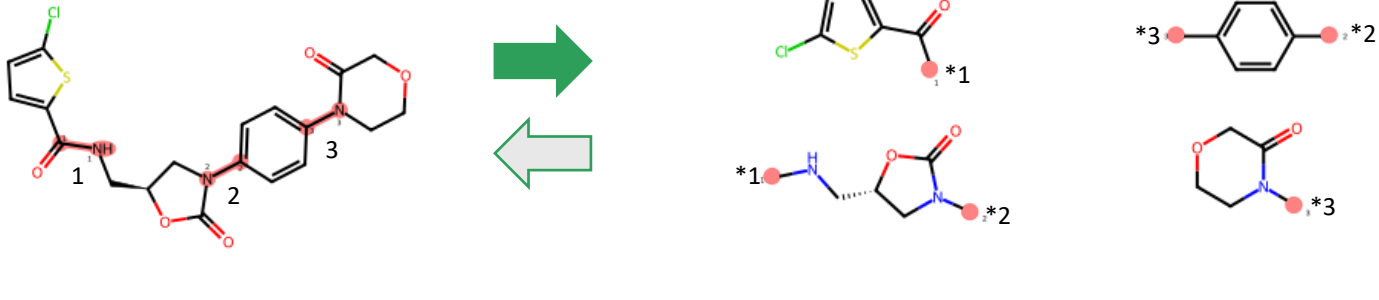
Thioethers, disulfides, sulfonate esters, acyl-sulfonamides, quaternary N

Built-in guard rails

Constraints on ring membership, substitution pattern and local environment

Complementary fragments: nothing is lost

Labelled attachment points let the parent be rebuilt exactly



**1 pairs with *1, *2 with *2, and *3 with *3*

Several valid decompositions

For X–Y–Z: {X, YZ}, {XY, Z} and {X, Y, Z} are all enumerated

Exact reconstruction

Every set recombines to the parent SMILES — self-verifiable

Keeping fragments chemically meaningful

Same constraints applied to all rule sets for a fair comparison

Minimum size

Fragments with < 3 heavy atoms are discarded

Ring protection

Atoms of fused ring systems (optionally their neighbours) are never cut

Conservative mode

Reactions touching protected atoms are excluded to keep cores intact

Best route wins

Among all explored pathways, the one with most fragments is kept

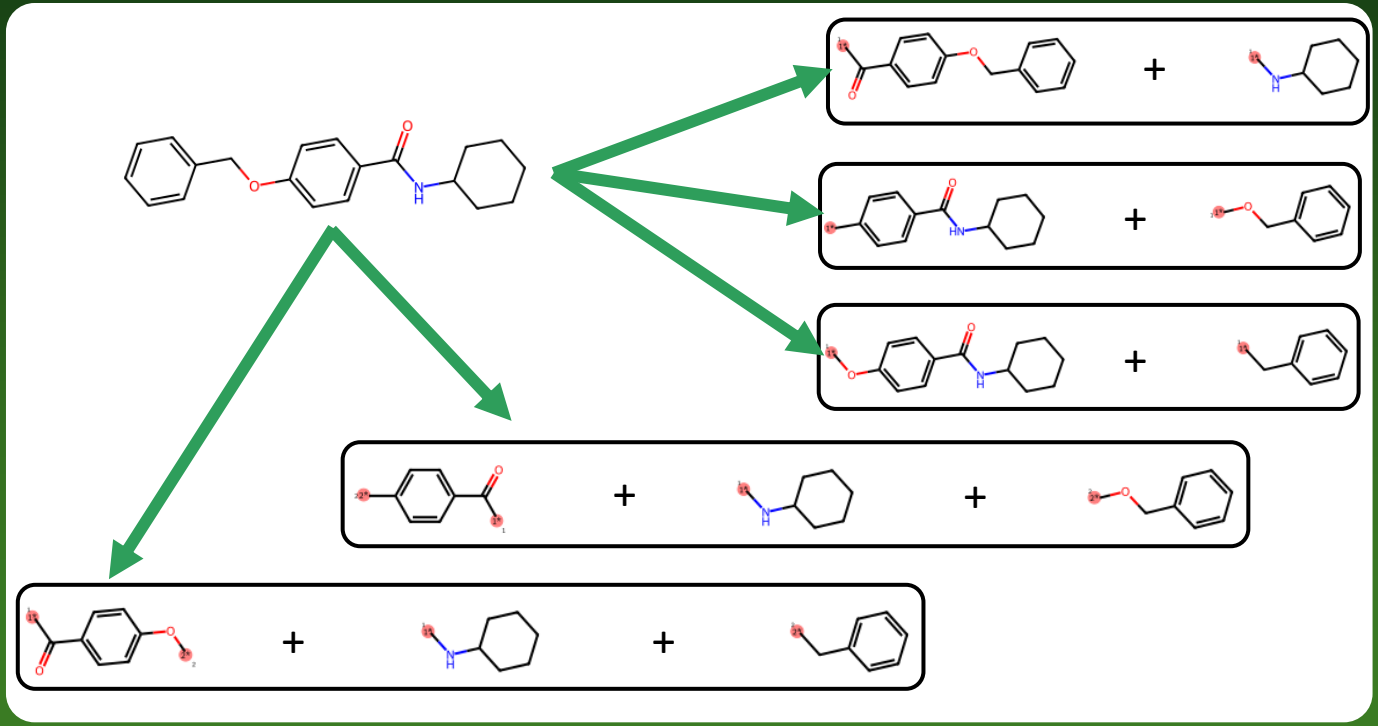
Fair benchmark: RECAP, BRICS, and CRUSH rules were all executed inside the same workflow, under identical constraints.

Keeping fragments chemically meaningful

Same constraints applied to all rule sets for a fair comparison

Best route wins

Among all explored pathways, the one with most fragments is kept

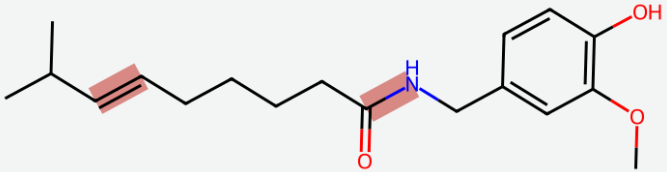


Same molecule, three rulebooks

Cleavable acyclic bonds in capsaicin (a FooDB compound)

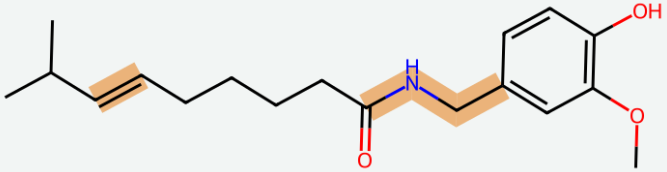
RECAP

2 bonds



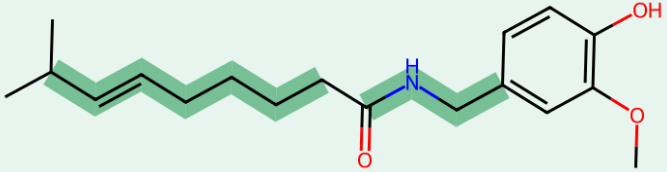
BRICS

4 bonds



CRUSH

9 bonds



Illustrative: bonds matched by each rule set (RDKit), with ring protection and the ≥ 3-heavy-atom filter applied.

Benchmark: five chemically diverse datasets

≈171,000 compounds, one unified framework

41,068
ChemDiv
Drugs-like compounds

13,578
LANaPDB v2
Latin American natural products

223
anti-MRSA
canonical & non-canonical peptides

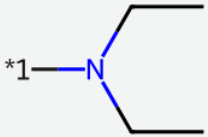
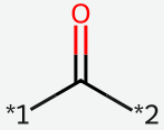
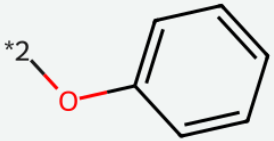
46,089
Macrocycle-DB
cyclic peptides and macrolides

70,477
FooDB
food-related chemicals

Analysis pipeline

RDKit-sanitized SMILES → Fragment with RECAP / BRICS / CRUSH → ECFP4 (2048) or MACCS (166) → Chemical space analysis → Yield, similarity, coverage, Ro3

From fragments to coordinates (Chemical space)

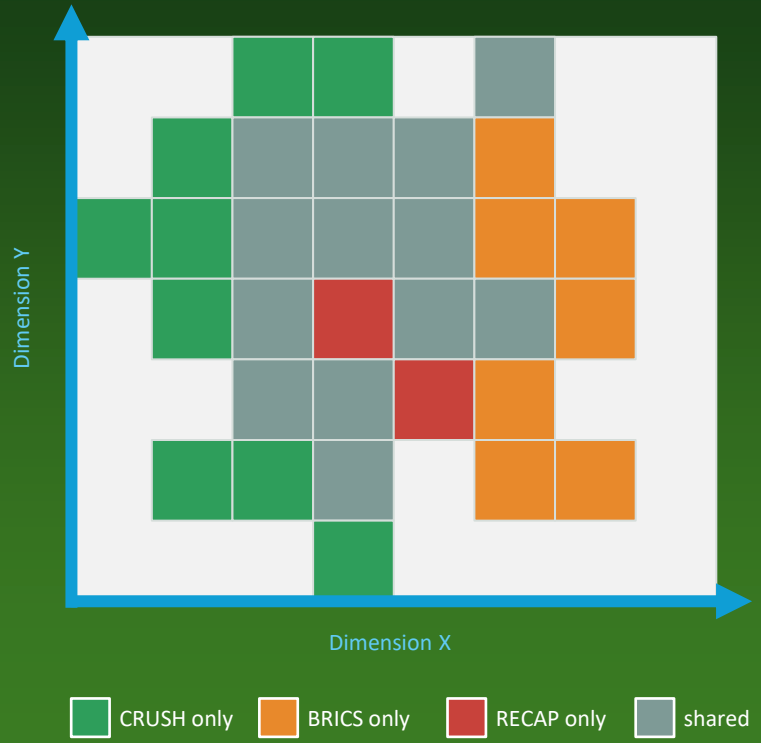
Fragments	Fingerprint	Reductional dimension method
	0010011000011011…	X = -10.7; Y = 20.1
	0010011000011011…	X = 28.1; Y = -10.3
	0010011000011011…	X = 5.9; Y = -0.7

Schematic — not real data

How do we measure "chemical space coverage"?

Grid-based occupancy on the UMAP map

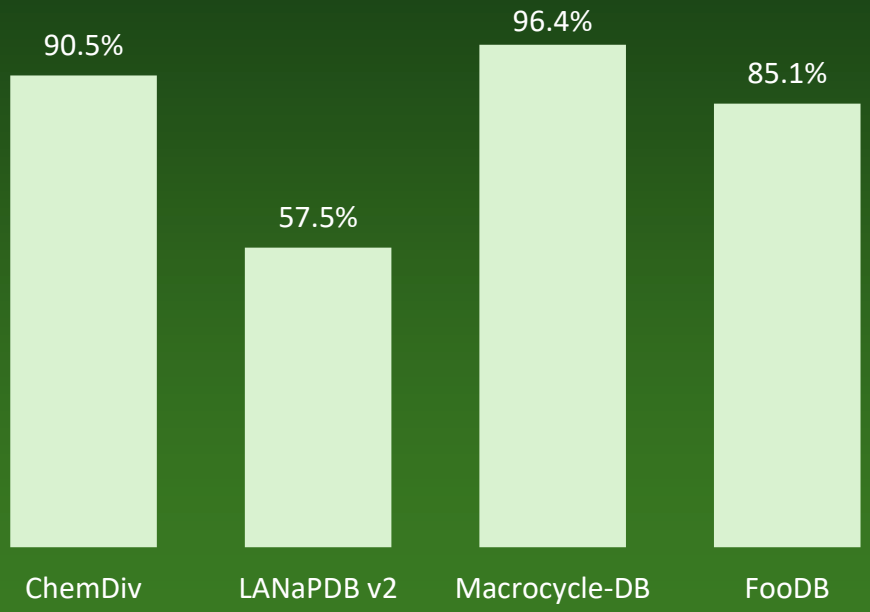
1. Project all fragments onto one shared chemical space (e.g., UMAP)
2. Divide the map into a fixed grid of cells
3. Mark which cells each method occupies
4. Compare: shared cells vs. cells unique to one method (Tanimoto overlap, Venn diagram)



Schematic — not real data

Result 1 — more molecules cut, more fragments

*Compounds fragmented by **CRUSH***
(% of each dataset)



≈ 2× **more fragments than RECAP**
ChemDiv, LANaPDB, FooDB

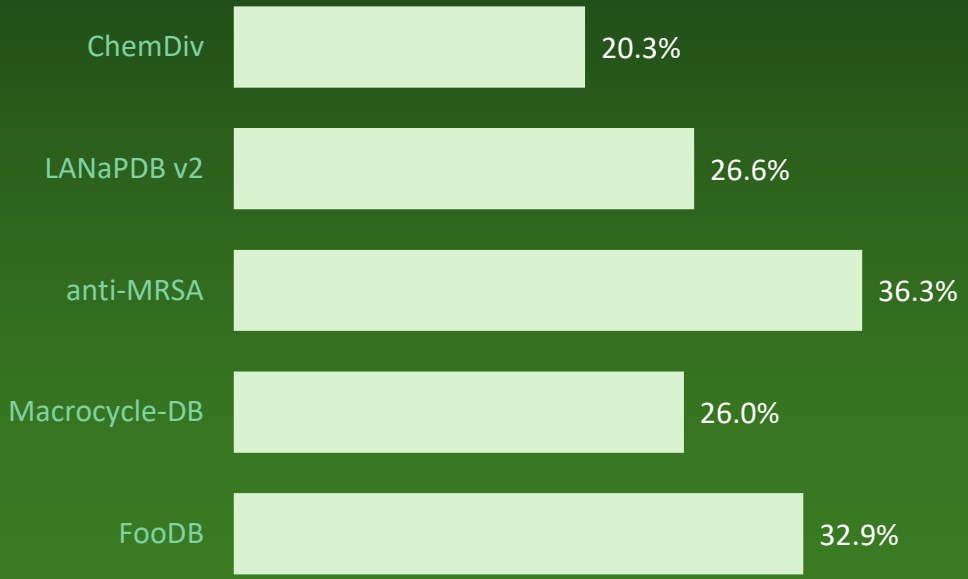
≈ 3× **more fragments than RECAP**
Macrocycle-DB

Top **fragment yield, above BRICS**
in all five datasets

anti-MRSA peptides: BRICS cut slightly more compounds (96.4%), yet CRUSH produced the most fragments.

Result 2 — chemical space exclusively explored by CRUSH

*Fragment space covered exclusively by **CRUSH** (ECFP4)*



41–59%

of fragment space is reached only by **CRUSH** and/or BRICS — never by RECAP

Same trend with low-resolution MACCS keys → effect of the rules, not of the fingerprint

Result 3 — smaller, simpler, more fragment-like

Compliance with the Rule of Three (Ro3)

Ro3 thresholds

MW ≤ 300	HBD ≤ 3	HBA ≤ 3
ClogP ≤ 3	NRot ≤ 3	PSA ≤ 60 Å²

- **CRUSH** & BRICS fragments show lower HBD, HBA, ClogP, Nrot, and PSA than RECAP
- Medians generally within Ro3 for ChemDiv, LANaPDB, and FooDB
- Complex parents (macrocycles, peptides) yield some Ro3-violating fragments with every method

RECAP fragments from FooDB (median)

13

rotatable bonds

Ro3 limit: 3

5

ClogP

Ro3 limit: 3

RECAP keeps large, flexible pieces; CRUSH resolves them into fragment-like building blocks.

Honest caveats

*What **CRUSH** does not (yet) do*



No explicit synthesizability

Unlike retrosynthetic schemes, CRUSH prioritizes diversity and novelty over immediate synthetic accessibility → best for early, exploratory stages



Fragment-size effect

Smaller fragments naturally look less similar to parents and spread more on the map; size-stratified analyses are planned

Try it: CRUSH is open source

github.com/EdgL2/CRUSH · MIT license

Code, SMIRKS rules & analysis notebooks

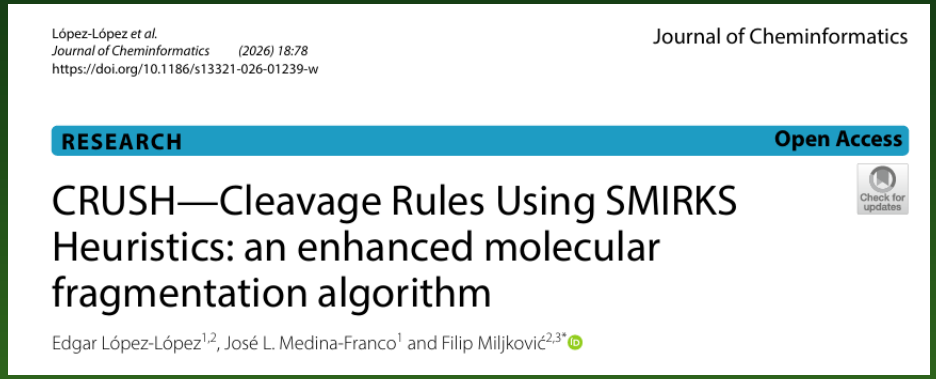
RDKit-based; parallel, streaming workflow for large libraries

Fully configurable

Rule subsets or custom SMIRKS, size filter, ring protection, max_cuts

Step-by-step tutorial

Colab/Jupyter notebook; data on Zenodo (10.5281/zenodo.18665641)



López-López *et al.*
Journal of Cheminformatics (2026) 18:78
https://doi.org/10.1186/s13321-026-01239-w

Journal of Cheminformatics

RESEARCH Open Access

CRUSH—Cleavage Rules Using SMIRKS Heuristics: an enhanced molecular fragmentation algorithm

Edgar López-López[1,2], José L. Medina-Franco[1] and Filip Miljković[2,3*]

Conclusions

*What **CRUSH** adds over RECAP and BRICS*

Finer chemistry
33 context-aware rules, incl. C–C, sulfur and heterocyclic bonds

Exhaustive fragmentation
All rules on all bonds, many pathways — not first-match

Higher yield
≈2–3× RECAP and above BRICS across five chemistries

Exclusive chemical space
20–36% of fragment space reached only by CRUSH, robust to fingerprint

Fragment-like output
Smaller, simpler fragments, closer to the Rule of Three

Open & reconstructable
Labelled complementary sets; MIT-licensed code

Complement, not replace

RECAP → retrosynthesis & scaffold dissection

BRICS → synthetically grounded building blocks

CRUSH → **diversity, novelty & exploration**

*Combining **CRUSH + BRICS** balances novel chemotypes with intuitive, synthesizable pieces.*

Perspectives: where CRUSH can go next

Focused fragment libraries

Target-, scaffold- or source-specific sets (e.g. natural-product or peptide-inspired fragments)

Generative machine learning

Fragment vocabularies for fragment-based generative models and de novo design

Combinatorial chemistry

Enumerate new chemotypes by recombining complementary, labelled fragments

Virtual screening

Diverse fragment spaces for fragment docking and growing

Fragment-based representations

New fingerprints and descriptors; data augmentation for ML

Synthesis-aware filtering

Couple with retrosynthesis/synthesizability tools; size-controlled benchmarks

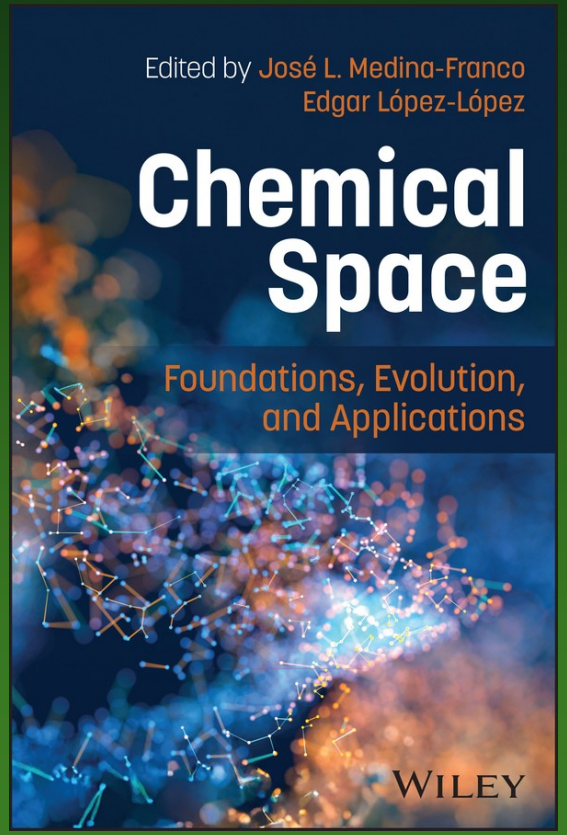
Relevant across chemistry domains: Medicinal chemistry, organic and inorganic chemistry, Material science, Agrochemistry, etc.

Chemical Space: Foundations, Evolution, and Applications

Wiley | December 2026 | ISBN: 9781394364244

The definitive guide to two foundational concepts reshaping chemoinformatics and computational chemistry, with practical examples, open-source code, and real-world applications in drug discovery and materials science.

Coming soon to Wiley.com — save 20% with code AUT24 at checkout, valid until 30th April 2027.



WILEY

Acknowledgments

Collaborators

- Carolina Barrientos-Salcedo (Mexico, Univ. Veracruzana)
- Carlos M. Cerda-García-Rojas (Mexico, CINVESTAV)
- **Filip Miljkovic (Sweden, AztraZeneca)**
- Ioana Chelcea (Sweden, Research Institutes of Sweden)
- **José L. Medina-Franco (Mexico, UNAM)**
- Karina Martínez-Mayorga (Mexico, UNAM)
- Marilia Valli (Brazil, University of São Paulo)
- Ramón A. Miranda-Quintana (USA, University of Florida)
- Ola Spjuth (Sweden, Uppsala University)
- Ulf Norinder (Sweden, Örebro University)

…

Special thanks

To the Labs team.

To the organizers of **BCADD-2026**.

Thanks to all for your kind attention and time.

Working groups

pharmb.io

ChemBio Design Group

DIFACQUIM COMPUTER-AIDED DRUG DESIGN AT UNAM

AstraZeneca

Merck

Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional
Química de Productos Naturales Orgánicos

MISTRA SAFECHEM

RI. SE Research Institutes of Sweden

PARC Partnership for the Assessment of Risks from Chemicals

Financing

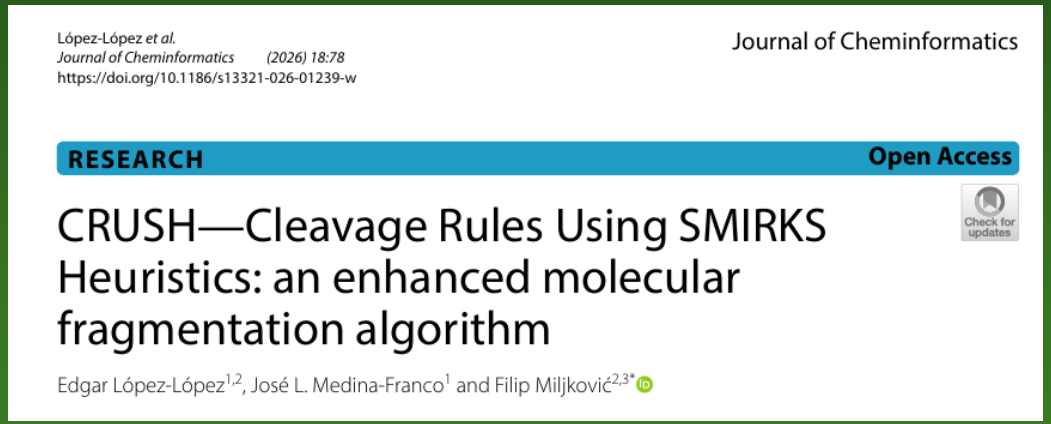
- **MISTRA Safe Chem**
- **Merck**
- **The Nobel Foundation**
- **SECIHTI**

Thank you!

Questions and discussion are very welcome

J. Cheminform. 18, 78 (2026) · doi:10.1186/s13321-026-01239-w

github.com/EdgL2/CRUSH



López-López *et al.*
Journal of Cheminformatics (2026) 18:78
https://doi.org/10.1186/s13321-026-01239-w

Journal of Cheminformatics

RESEARCH Open Access

CRUSH—Cleavage Rules Using SMIRKS Heuristics: an enhanced molecular fragmentation algorithm

Edgar López-López[1,2], José L. Medina-Franco[1] and Filip Miljković[2,3*]

XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery · BCADD-2026

CRUSH:

A CHEMISTRY-AWARE MOLECULAR FRAGMENTATION ALGORITHM FOR EXPANDED CHEMICAL SPACE EXPLORATION

Edgar López-López* · José L. Medina-Franco · Filip Miljković

* edgar.lopez@uu.se

J. Cheminform. 18, 78 (2026) · doi:10.1186/s13321-026-01239-w

UPPSALA UNIVERSITET

AstraZeneca