TECHNISCHE UNIVERSITÄT DRESDEN

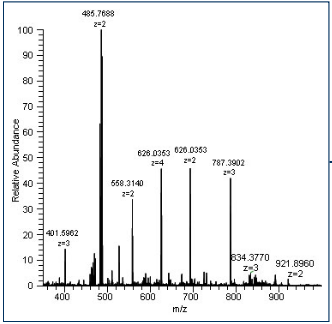
International Center for Computational Logic

Nils Küchenmeister, Alex Ivliev, Markus Krötzsch
Technische Universität Dresden, Knowledge-Based Systems Group

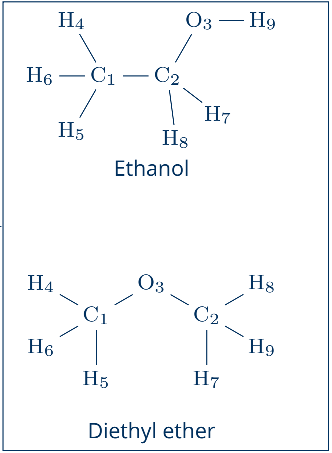
Towards Mass Spectrum Analysis with Answer-Set-Programming

LPNMR, October 12, 2024

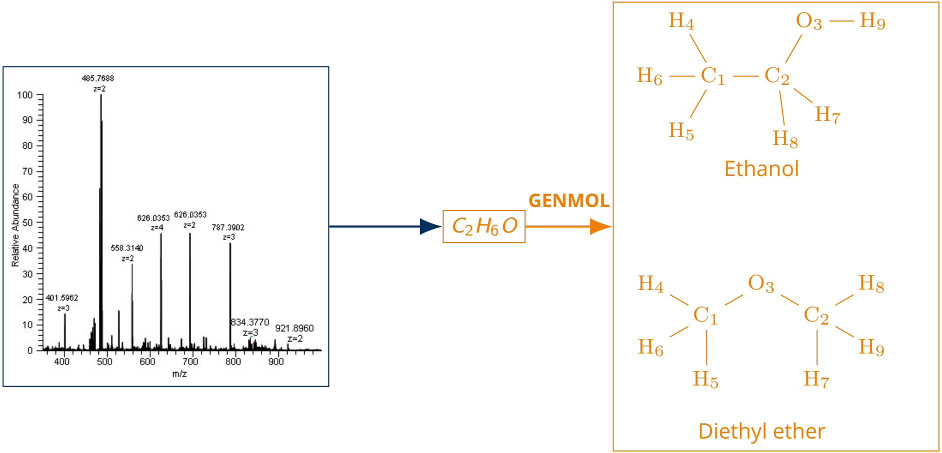
Mass Spectrometry



C_2H_6O



Mass Spectrometry

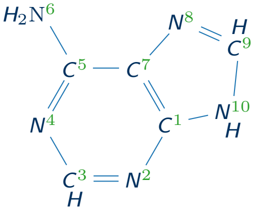


- Use ASP to solve the combinatorical search problem

Problem statement

Problem statement

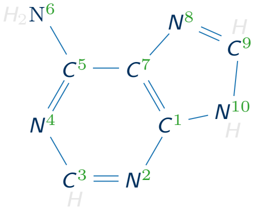
- A sum formula is mapping, e.g. $C_5H_5N_5 \quad \Rightarrow \quad f : \mathbb{E} \to \mathbb{N}_{>0}, f(C) = 5, f(H) = 5, f(N) = 5$
- Elements are associated with a valence, e.g. $\mathbb{V}(C) = 4$, $\mathbb{V}(H) = 1$, and $\mathbb{V}(N) = 3$



Molecular Graph G Representation

Problem statement

- A sum formula is mapping, e.g. $C_5H_5N_5 \quad \Rightarrow \quad f : \mathbb{E} \rightarrow \mathbb{N}_{>0}, f(C) = 5, f(H) = 5, f(N) = 5$
- Elements are associated with a valence, e.g. $\mathbb{V}(C) = 4$, $\mathbb{V}(H) = 1$, and $\mathbb{V}(N) = 3$

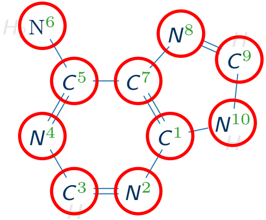


Molecular Graph G Representation

- Ignore the hydrogen atoms

Problem statement

- A sum formula is mapping, e.g. $C_5H_5N_5 \quad \Rightarrow \quad f : \mathbb{E} \to \mathbb{N}_{>0}, f(C) = 5, f(H) = 5, f(N) = 5$
- Elements are associated with a valence, e.g. $\mathbb{V}(C) = 4$, $\mathbb{V}(H) = 1$, and $\mathbb{V}(N) = 3$

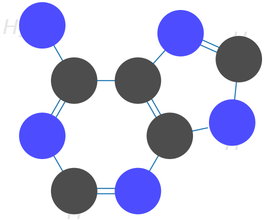


Molecular Graph G Representation

- Ignore the hydrogen atoms
- Atoms are nodes

Problem statement

- A sum formula is mapping, e.g. $C_5H_5N_5 \quad \Rightarrow \quad f : \mathbb{E} \to \mathbb{N}_{>0}, f(C) = 5, f(H) = 5, f(N) = 5$
- Elements are associated with a valence, e.g. $\mathbb{V}(C) = 4$, $\mathbb{V}(H) = 1$, and $\mathbb{V}(N) = 3$

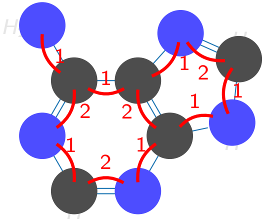


Molecular Graph G Representation

- Ignore the hydrogen atoms
- Atoms are nodes
- Element symbols represented by node colors

Problem statement

- A sum formula is mapping, e.g. $C_5H_5N_5 \quad \Rightarrow \quad f : \mathbb{E} \rightarrow \mathbb{N}_{>0}, f(C) = 5, f(H) = 5, f(N) = 5$
- Elements are associated with a valence, e.g. $\mathbb{V}(C) = 4$, $\mathbb{V}(H) = 1$, and $\mathbb{V}(N) = 3$

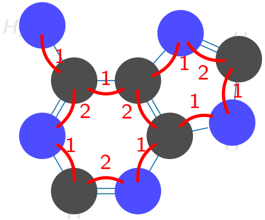


Molecular Graph G Representation

- Ignore the hydrogen atoms
- Atoms are nodes
- Element symbols represented by node colors
- Bonds represented by labelled edges

Problem statement

- A sum formula is mapping, e.g. $C_5H_5N_5 \quad \Rightarrow \quad f : \mathbb{E} \to \mathbb{N}_{>0}, f(C) = 5, f(H) = 5, f(N) = 5$
- Elements are associated with a valence, e.g. $\mathbb{V}(C) = 4$, $\mathbb{V}(H) = 1$, and $\mathbb{V}(N) = 3$



Molecular Graph G Representation

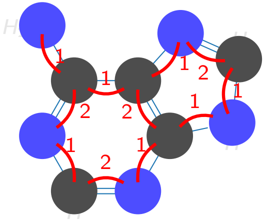
- Ignore the hydrogen atoms
- Atoms are nodes
- Element symbols represented by node colors
- Bonds represented by labelled edges

Properties: G is valid for f iff...

1. G is connected
2. count of non-hydrogen atoms matches f
3. no node degree exceeds the element's valence
4. number of H corresponds to free binding spaces

Problem statement

- A sum formula is mapping, e.g. $C_5H_5N_5 \quad \Rightarrow \quad f : \mathbb{E} \to \mathbb{N}_{>0}, f(C) = 5, f(H) = 5, f(N) = 5$
- Elements are associated with a valence, e.g. $\mathbb{V}(C) = 4$, $\mathbb{V}(H) = 1$, and $\mathbb{V}(N) = 3$



Molecular Graph G Representation

- Ignore the hydrogen atoms
- Atoms are nodes
- Element symbols represented by node colors
- Bonds represented by labelled edges

Properties: G is valid for f iff...

1. G is connected
2. count of non-hydrogen atoms matches f
3. no node degree exceeds the element's valence
4. number of H corresponds to free binding spaces

ENUMERATION PROBLEM

For a given molecular formula f, enumerate, up to isomorphism, all valid molecular graphs for f.

ASP encoding

Naive implementation

- ASP is well suited to encode this problem succinctly

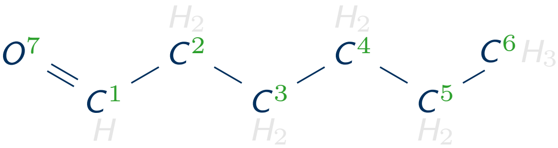
```
{ edge(X, Y) : node(X), node(Y), X < Y }.
edge(Y, X) :- edge(X, Y).

reachable(1).
reachable(Y) :- reachable(X), edge(X, Y).
:- not reachable(X), node(X).

1{ edge(X, Y, 1..3) }1 :- edge(X, Y), X < Y.
edge(Y, X, M) :- edge(X, Y, M).

degree(N, D) :- node(N), D = #sum { C, X : edge(N, X, C) }.

:- node(N), type(N, E), degree(N, D), element(E, _, V), D > V.

:- EDGE_COUNT = #sum { M, X, Y : edge(X, Y, M), X < Y },
   VALENCE_SUM = #sum { V*C, E : molecular_formula(E, C), element(E, _, V), E != "H" },
   molecular_formula("H", H_COUNT), EDGE_COUNT != (VALENCE_SUM - H_COUNT)/2.
```

The obstacle: symmetries

Example

$C_6H_{12}O$ admits 211 distinct molecule structures but leads to 111,870 answer sets

- For example *Hexanal*
- 7 nodes $\Rightarrow$ $7! = 5040$ many answer-sets

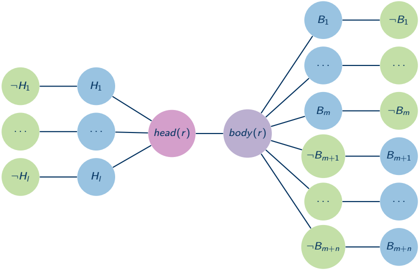


Symmetry breaking: existing approaches

Automated: **BreakID** [Devriendt and Bogaerts, ASPOCP'16]

- Transform the grounding into colored graph

$$\boldsymbol{r} : H_1, \ldots, H_\ell \leftarrow B_1, \ldots, B_m, \neg B_{m+1}, \ldots, \neg B_{m+n}$$



- Use graph automorphisms to remove syntactic symmetry

Symmetry breaking: existing approaches

Manual: **Graph-based symmetry breaking** [Codish et al., Constraints vol. 24, 2019]

- Partitioned simple graph, represented by adjacency matrix
- Normalization of adjacency matrix by requiring lexicographic order of rows/columns

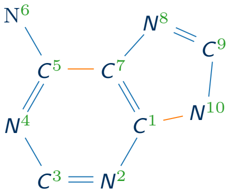
Symmetry breaking: existing approaches

Manual: Graph-based symmetry breaking [Codish et al., Constraints vol. 24, 2019]

- Partitioned simple graph, represented by adjacency matrix
- Normalization of adjacency matrix by requiring lexicographic order of rows/columns

```
sat(I, K, J) :-
    type(I, T), type(J, T), type(K, T), type(L, T), J > I, J - I != 2,
    edge(I, K), edge(J, L), L < K, L != I.
sat(I, K, J) :-
    type(I, T), type(J, T), type(K, T), J > I, J - I != 2,
    edge(I, K, N), edge(J, K, M), N <= M.

:- type(I, T), type(J, T), type(K, T),
   edge(I, K), node(J), J > I, not sat(I, K, J), J - I != 2, K != J.
```

Symmetry breaking: Our approach



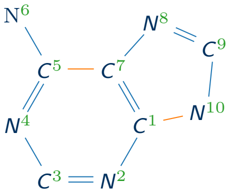
$C_5H_5N_5$ Adenine

How many isomorphic graphs?

- 10 nodes $\Rightarrow 10! \approx 3.6$ Mio

Symmetry breaking: Our approach

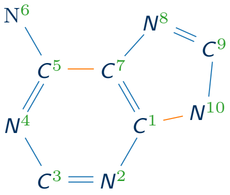
- **Inspiration:** SMILES (≙ serialization format for molecular graphs)

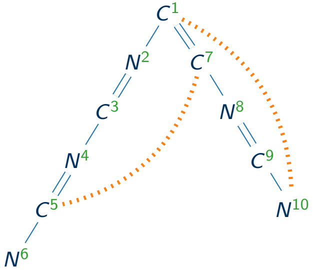


$$C^{1}1(=C^{7}2N^{8}=C^{9}N^{10}1)N^{2}=C^{3}N^{4}=C^{5}2N^{6}$$

Symmetry breaking: Our approach

- **Inspiration:** SMILES ($\triangleq$ serialization format for molecular graphs)
- Partition edges of G into tree and cycle edges $T \dot{\cup} C$, s.t. $(G \setminus C)$ is acyclic
- G is tree representation if depth-first sequence on T is natural order





Symmetry breaking: Our approach

- **Inspiration:** SMILES (≙ serialization format for molecular graphs)
- Partition edges of G into tree and cycle edges $T \dot{\cup} C$, s.t. $(G \setminus C)$ is acyclic
- G is tree representation if depth-first sequence on T is natural order
- **Choices:** (a) root vertex, (b) spanning tree, (c) order of visiting children

$C_5 H_5 N_5$ Adenine

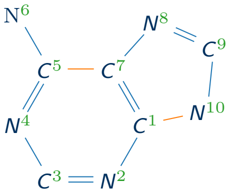
How many isomorphic tree representations?

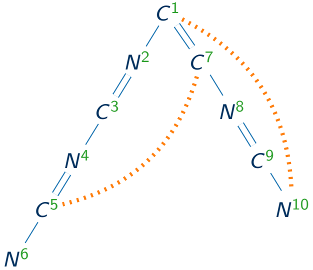
(a) 10 roots

(b) $4 \cdot 5 + 9 = 29$ spanning trees

(c) $2^6 \cdot 3^3$ child sequences

$\rightarrow 10 \cdot 29 \cdot 64 \cdot 27 = 501,120$





Symmetry breaking: Our approach

- **Inspiration:** SMILES ($\triangleq$ serialization format for molecular graphs)
- Partition edges of G into tree and cycle edges $T \dot{\cup} C$, s.t. $(G \setminus C)$ is acyclic
- G is tree representation if depth-first sequence on T is natural order
- **Choices:** (a) root vertex, (b) spanning tree, (c) order of visiting children
- Canonical Molecular Graph
  - Determine root as central vertex
  - Define total order $\prec$ on trees
  - Select $\prec$-largest candidate

Symmetry breaking: Our approach

- **Inspiration:** SMILES (≙ serialization format for molecular graphs)
- Partition edges of G into tree and cycle edges $T \dot{\cup} C$, s.t. $(G \setminus C)$ is acyclic
- G is tree representation if depth-first sequence on T is natural order
- **Choices:** (a) root vertex, (b) spanning tree, (c) order of visiting children
- Canonical Molecular Graph
  - Determine root as central vertex
  - Define total order $\prec$ on trees
  - Select $\prec$-largest candidate

$C_5 H_5 N_5$ Adenine

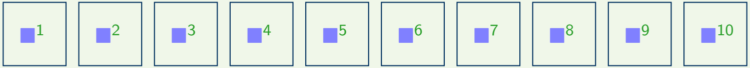
How many isomorphic answer-sets?

1. $C^6{=}13N^5{=}C^4N^3C^2 1N^1{=}C^7N^8{=}C^9 3N^{10}$
2. $N^6 1C^5{=}N^4C^3 2{=}C^2 1N^1{=}C^7N^8{=}C^9 2N^{10}$
3. $N^6C^5 3C^4 1N^3{=}C^2N^1C^7{=}1N^8{=}C^9N^{10}{=}3$
4. $N^5 2{=}C^4N^3{=}C^2(N^6)C^1 1{=}C^7 2N^8C^9{=}N^{10} 1$

- In theory unique representation, but too expensive calculation $\Rightarrow$ approximated implementation

ASP implementation

$C_5H_5N_5 \rightarrow$ 10 nodes



ASP implementation

1. Permutation of symbols

$C_5H_5N_5 \rightarrow$ 10 nodes

C^1	N^2	C^3	N^4	C^5	N^6	C^7	N^8	C^9	N^{10}

ASP implementation

1. Permutation of symbols
2. Decide for # of multi-bonds and cycle markers
   - Degree of unsaturation

$C_5 H_5 N_5 \rightarrow$ 10 nodes

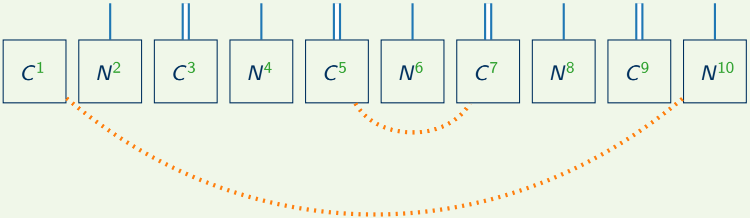
C^1	N^2	C^3	N^4	C^5	N^6	C^7	N^8	C^9	N^{10}

- $(f(C)\cdot(\mathbb{V}(C)-2)+f(H)\cdot(\mathbb{V}(H)-2)+f(N)\cdot(\mathbb{V}(N)-2))/2+1=6$
- Choose e.g. 4 double bonds and 2 pairs of cycle-markers

ASP implementation

1. Permutation of symbols
2. Decide for # of multi-bonds and cycle markers
   - Degree of unsaturation
3. Permutation of multi-bonds and cycle markers

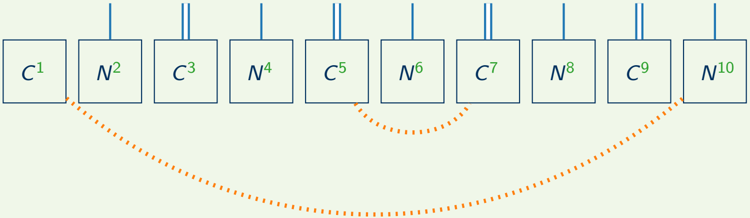
$C_5 H_5 N_5 \rightarrow$ 10 nodes



ASP implementation

1. Permutation of symbols
2. Decide for # of multi-bonds and cycle markers
   - Degree of unsaturation
3. Permutation of multi-bonds and cycle markers
4. Select a main chain length and split it in left and right half

$C_5 H_5 N_5 \rightarrow$ 10 nodes



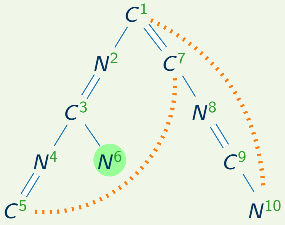
- length of main chain ranges from 5 to 10

ASP implementation

1. Permutation of symbols
2. Decide for # of multi-bonds and cycle markers
   - Degree of unsaturation
3. Permutation of multi-bonds and cycle markers
4. Select a main chain length and split it in left and right half
5. Generate all depth-first spanning trees with root 1 (making sure not to exceed the valence)

$C_5 H_5 N_5 \rightarrow$ 10 nodes

2-Aminopurine



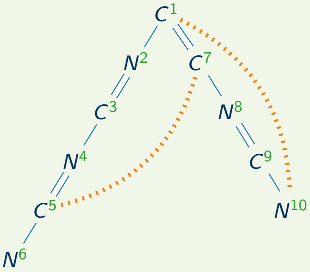
- length of main chain ranges from 5 to 10, choose e.g. **9**

ASP implementation

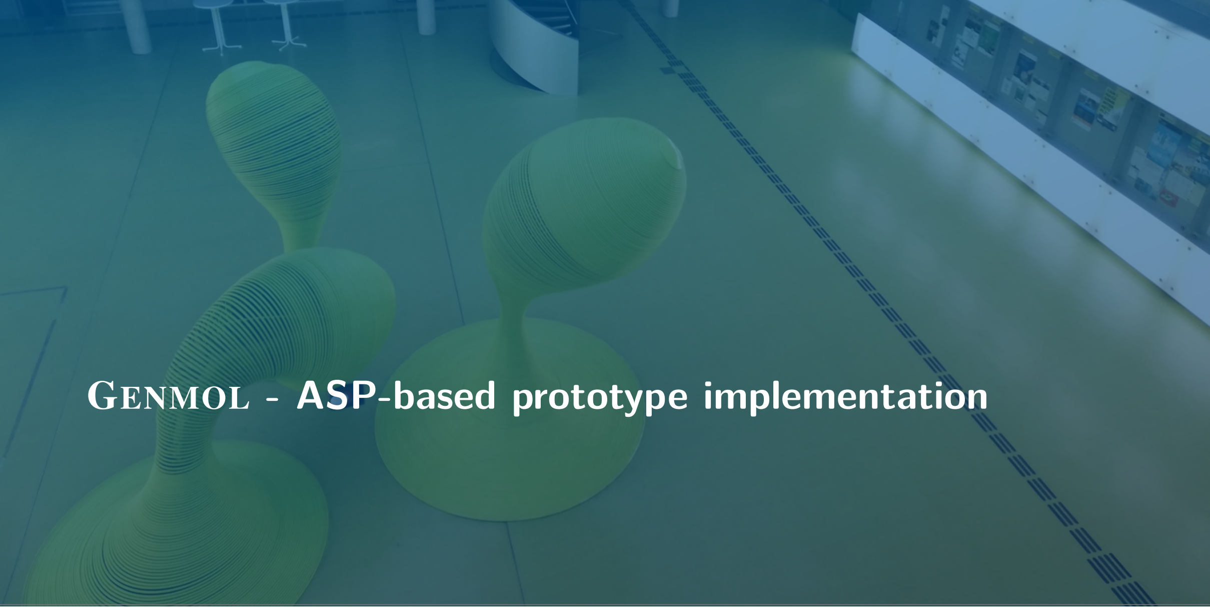
1. Permutation of symbols
2. Decide for # of multi-bonds and cycle markers
   - Degree of unsaturation
3. Permutation of multi-bonds and cycle markers
4. Select a main chain length and split it in left and right half
5. Generate all depth-first spanning trees with root 1 (making sure not to exceed the valence)

$C_5H_5N_5 \rightarrow$ 10 nodes

Adenine

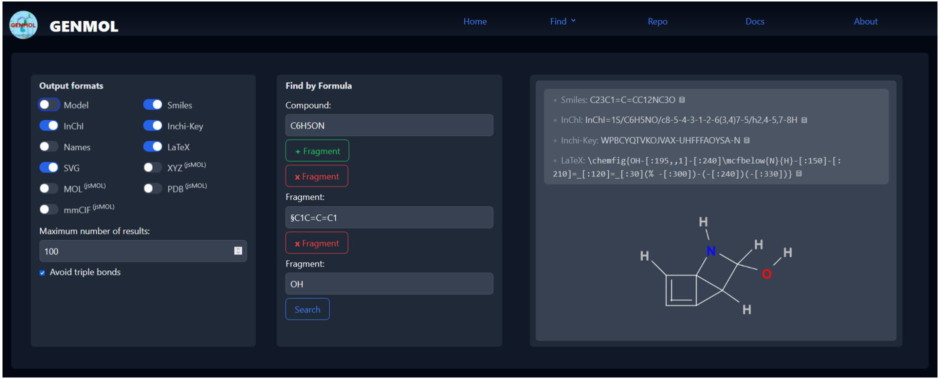


- length of main chain ranges from 5 to 10, choose e.g. **10**



GENMOL - ASP-based prototype implementation

Tool demo



https://tools.iccl.inf.tu-dresden.de/genmol/

Evaluation

Evaluation

- To investigate our symmetry-breaking approach for enumerating chemical molecules w.r.t.

Correctness **Symmetry breaking** **Scalability**

- We compare it to...

  A. BreakID
  (≙ automated SBC genaretion)

  B. Naive ASP encoding

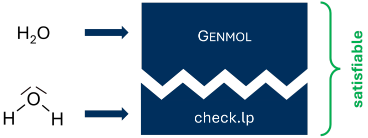
  C. Graph-based SBC

  D. Molgen
  (≙ commercial tool)

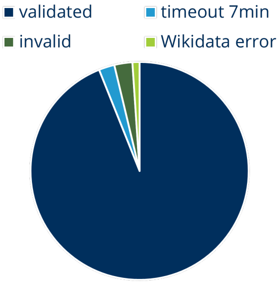
- Experiments use *Clingo v5.7.1*

Correctness

- 5,474 suitable chemical compounds with up to 17 atoms from Wikidata SPARQL

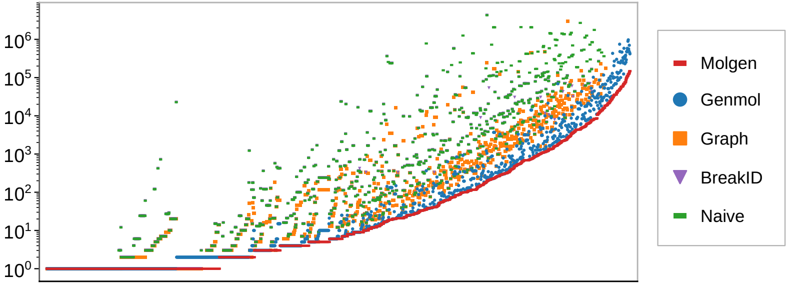


- 5,338 validated and 132 not processed within 7min timeout
- Found 3 errors in Wikidata (incorrect SMILES)



Symmetry breaking

- Use the smallest 1,750 molecular formulas from the Wikidata data set

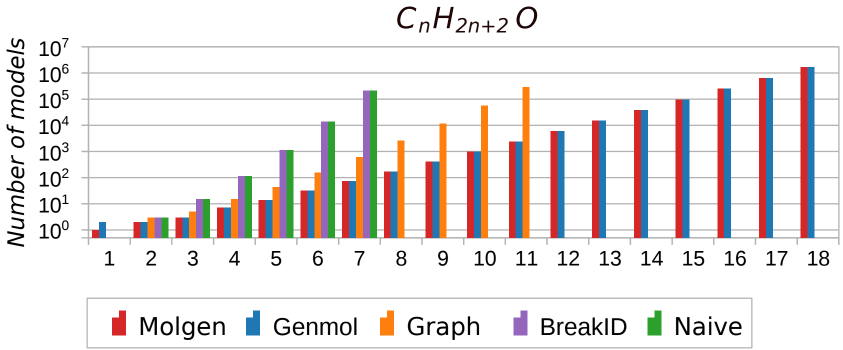


⇒ While GENMOL does not fully match MOLGEN, it comes closer than any other ASP-based approach

⇒ GENMOL: exact model count 51%, at most ten times more models 99%

Scalability

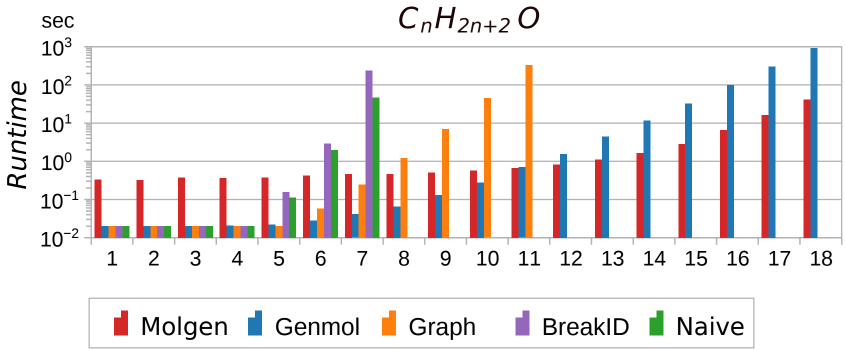
- series of uniformly created molecular formulas of increasing size



⇒ Perfect symmetry breaking for acyclic molecules

Scalability

- series of uniformly created molecular formulas of increasing size

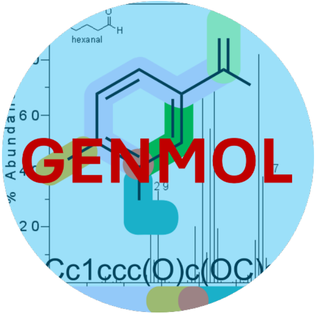


⇒ Runtime better than other ASP approaches that were considered

Conclusion

- ASP is well-suited to tackle mass spectrum analysis
  → Superior clarity (in contrast to complex, error-prone imperative implementations)
  → Additional features can easily be added, e.g. fragments, functional groups, aromatic rings, etc.
- Symmetry-breaking is vital
- Automated symmetry-breaking is not sufficient here
- MOLGEN's performance could not be matched evenly
- Proof-of-concept: GENMOL tool + web demo

https://tools.iccl.inf.tu-dresden.de/genmol/



Thank you for your attention!