Systematic extraction of analogue series from large compound collections

Martin Vogt
B-it Life Science Informatics
Rheinische Friedrich-Wilhelms-Universität Bonn

RDKit UGM - 25 September 2019

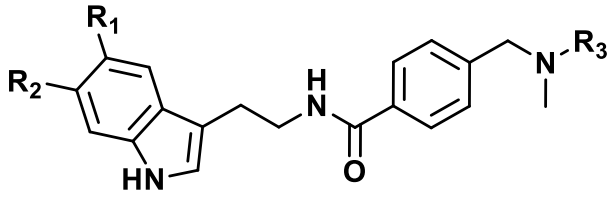
Problem

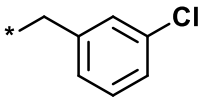
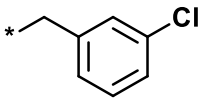
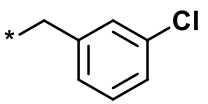
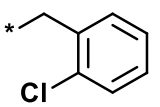
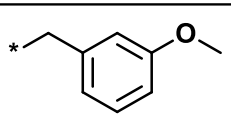
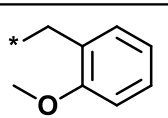
- Systematic identification of sets of closely related molecules in large compound databases that
  - share a common core structure (scaffold)
  - can be generated from the core structure by (chemically feasible) substitutions at specific substitution sites

Rationale

- Analogue series (AS) are series of compounds that
  - share the same core structures
  - carry different R-groups at single or multiple substitution sites

- ASs are conventionally represented in R-group tables
  - major source of structure-activity relationship (SAR) information
  - SAR analysis typically based on individual AS



ChEMBL ID	R_1	R_2	R_3
3263732			
2363733			
2363731			
2363730			
2363737			
2363735			

Rationale: AS and databases

- Analogues from databases (DBs) might be found by
  - substructure searches from known core structures
  - matched molecular pairs/series (MMP/MMS) analysis

- Matched Molecular Pairs/Series (MMPs/MMSs) are compounds that possess the same core structure but differ only at one specific site

- MMPs/MMSs can be efficiently and comprehensively extracted from large databases of millions of compounds
  - quasilinear in the number of molecules

- Goal:
  Systematically extract ASs that have core structures with substitutions at multiple sites

MMPs and Analogue series

- MMPs
  - characterized by a transformation at a single site
  - change not necessarily at terminal fragment

- Analogue series
  - differences at multiple sites
  - substituents are terminal fragments/side chains
  - represented as R-group table
  - based on chemical feasible reactions

Requirements

- Comprehensive:
  - All potential cores for a molecule are identified
  - Molecules with analogues are assigned to one (or more) AS
- Non-redundant: Analogue series should be as large as possible
- Unique: Compounds should be assigned to only one set of analogues
- Efficient: Algorithm should scale with the number of molecules, i.e. quasilinear $O(n \log^k n)$

Considerations

- Although the concept of AS is intuitive the problem is ill-defined
  - e.g., should some rings be part of core or fragment
  - e.g., should substitution sites start at rings
  - there is no single unique core structure for a molecule

- Specifying the AS concept:
  - Cores must not be too small compared to the substituents
    - e.g., core contains at least 2/3 of all heavy atoms for each molecule in the AS
  - Substituents must not be too large
    - e.g., each substituent must not contain more than 13 heavy atoms
  - Total number of substituents is limited
    - e.g., no more than 5 (non-hydrogen) substitutions are allowed

Approach

Extraction of Compound Core Relationships (CCR)

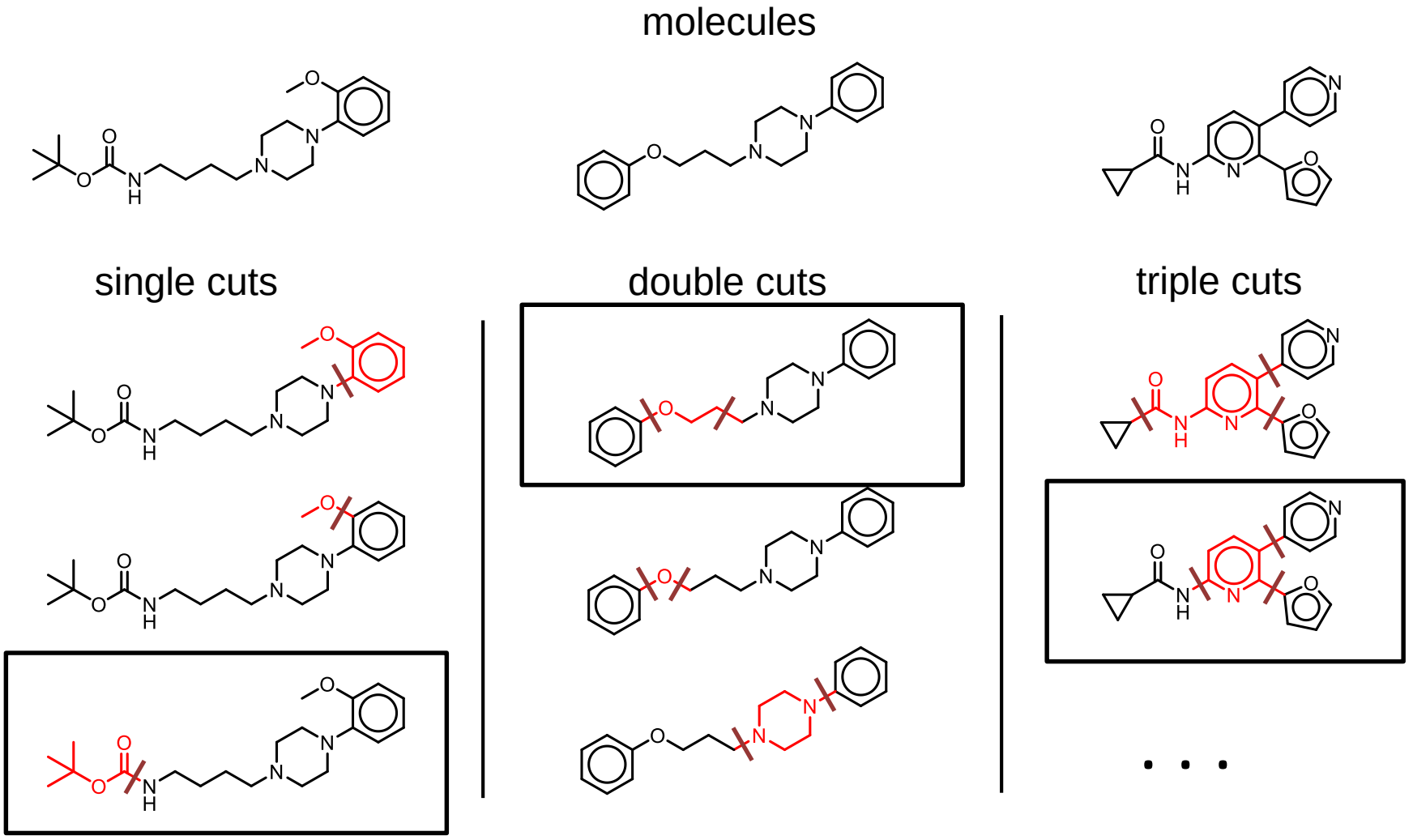
- Systematic fragmentation:
  - Fragment each molecule along "cuttable" acyclic bonds
    - A fragmentation consists of cutting 1 – *n* bonds
    - Cuttable bonds might be all single bonds, RECAP-bonds, …
  - A valid fragmentations possesses
    - a large core structure with 1 – *n* substitution sites
    - 1 – *n* substituents
  - For each molecule, all valid fragmentations are determined yielding
    - a set of potential core structures with one or more substitution sites
- Collect all valid fragmentations for all molecules
- …
- Group valid fragmentations by core
  - Yields a set of raw ASs
- "Housekeeping"

Fragmentation for MMS

- MMP-generation according to Hussain/Rea*
  - Systematically fragment each molecule
    - Explore all ways to remove 1 – 3 cuttable bonds yielding one large "core" and one small fragment yielding a "valid fragmentation"
    - The "core" might be disconnected
    - cuttable bonds are acyclic single bonds
  - Organize all valid fragmentations from all molecules by "core"
  - Molecules possessing identical frame form MMPs/MMS

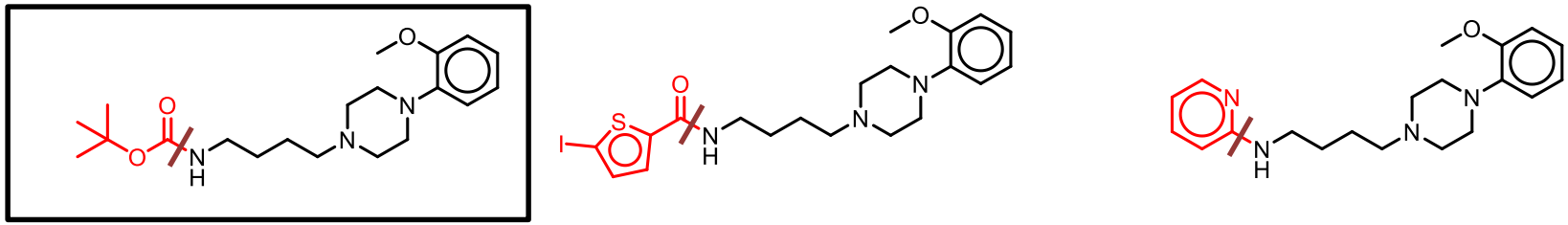
** J. Chem. Inf. Model.*2010, 50, 339-348

Matched Molecular Series

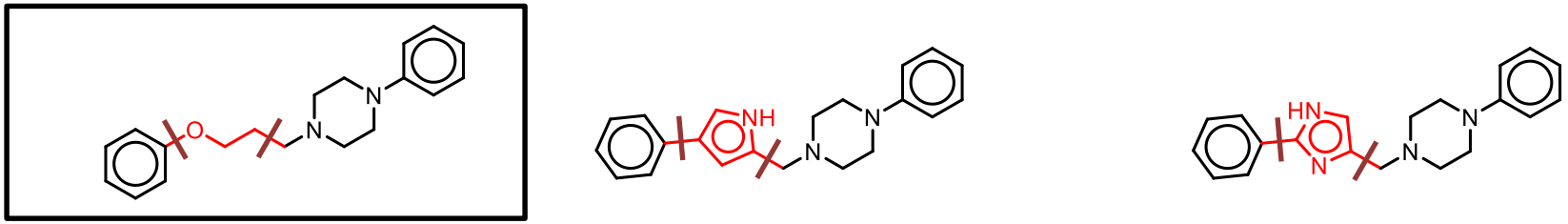


Matched Molecular Series

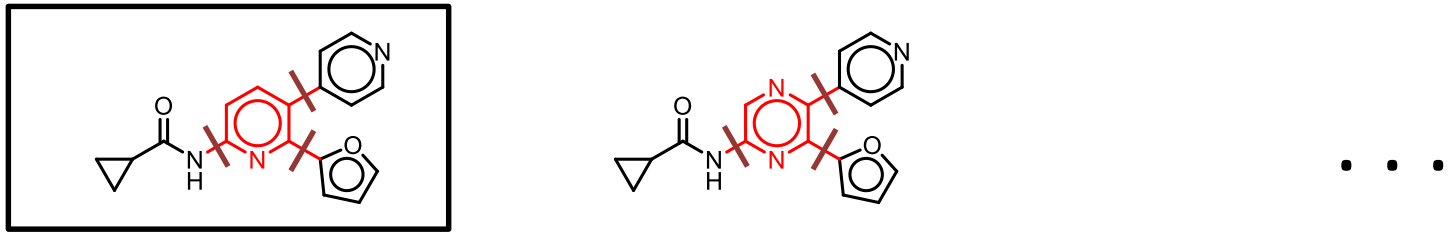
single cuts



double cuts



triple cuts



. . .

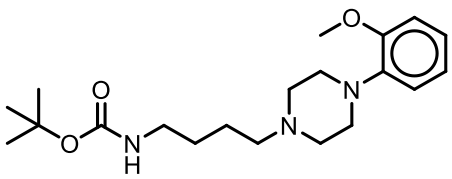
Fragmentation AS

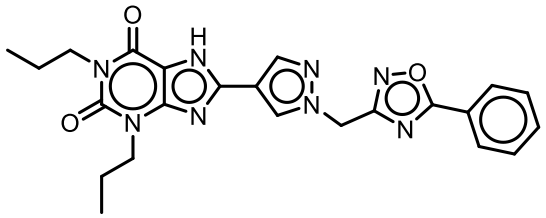
- For analogue series
  - Same basic approach
    - Explore all ways to remove 1 – *n* cuttable bonds yielding one connected core and 1 – *n* fragments
    - Cuttable bonds are acyclic bonds according to some chemical rules (RECAP*)
  - Organize valid fragmentations by core structure
  - Molecules with fragmentations having the same core are grouped forming an AS

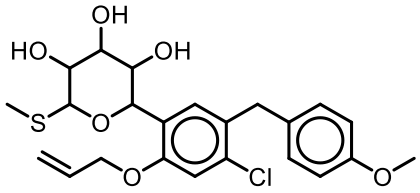
** J. Chem. Inf. Comput. Sci.* 1998,38,511-522

Analogue Series

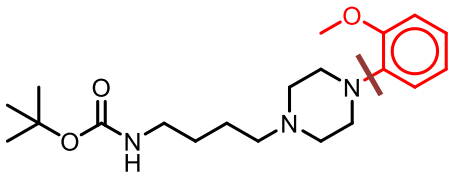
molecules

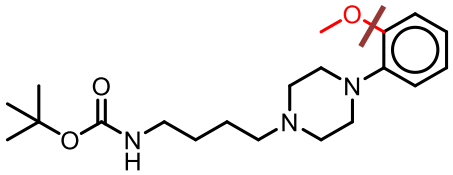


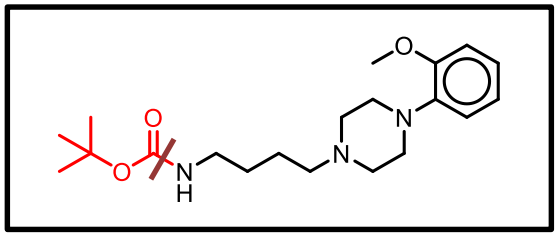




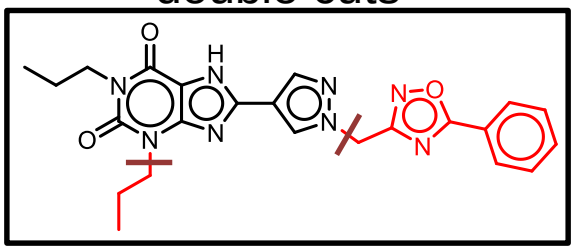
single cuts

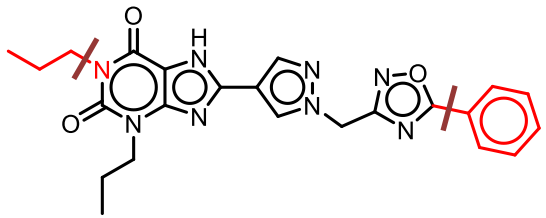


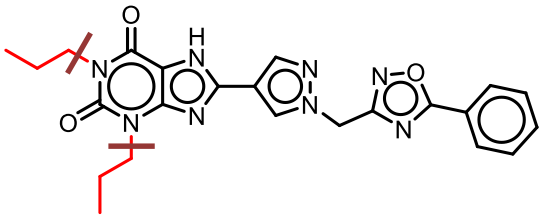




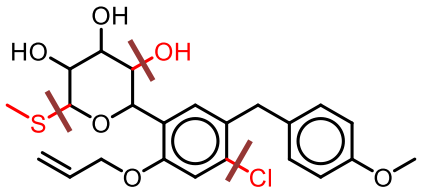
double cuts

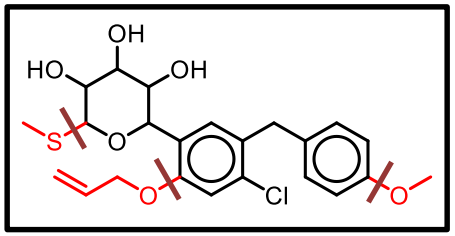


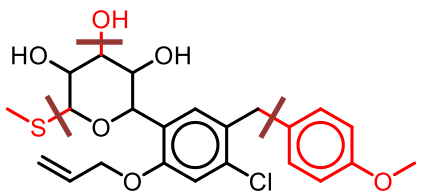




triple cuts

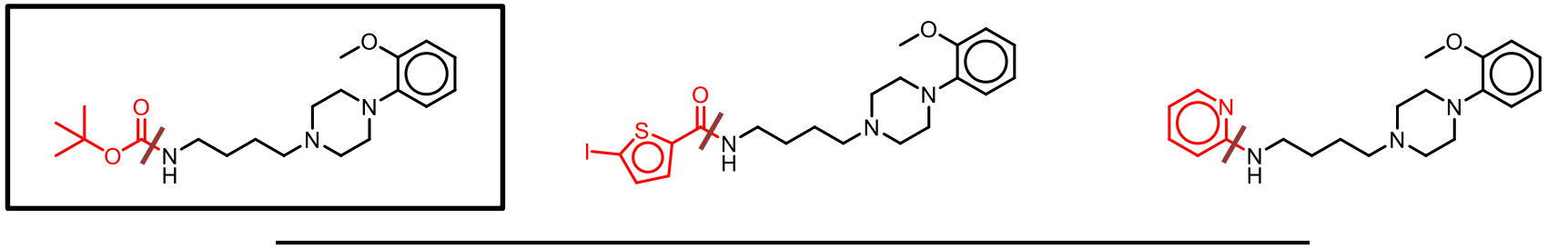




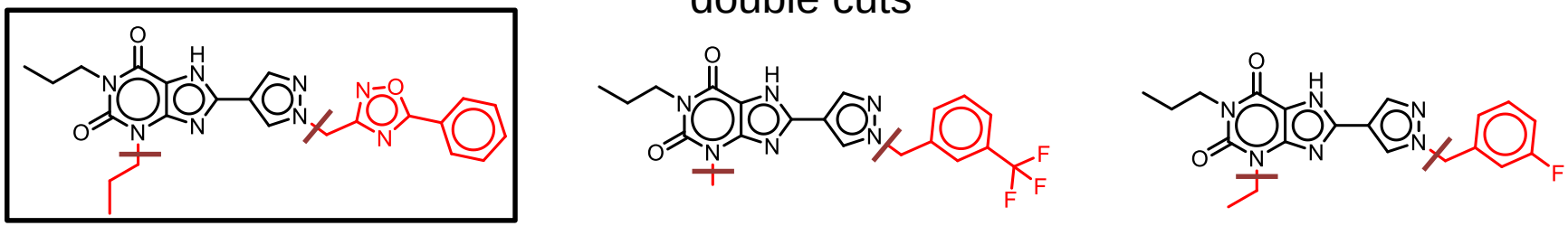


Analogue Series

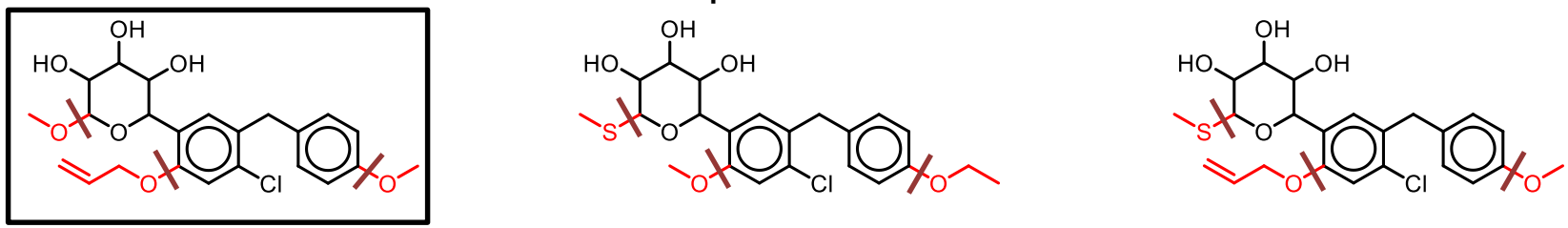
single cuts



double cuts



triple cuts



What about hydrogen substitutions?

- Cuts always involve two heavy atoms
- So far, hydrogen substitutions are not detected
- MMP solution (for single cut fragmentations):
  - Substitute the substitution site with a hydrogen
  - If the hydrogen substituted core is itself is a molecule it should be part of the MMP/MMS

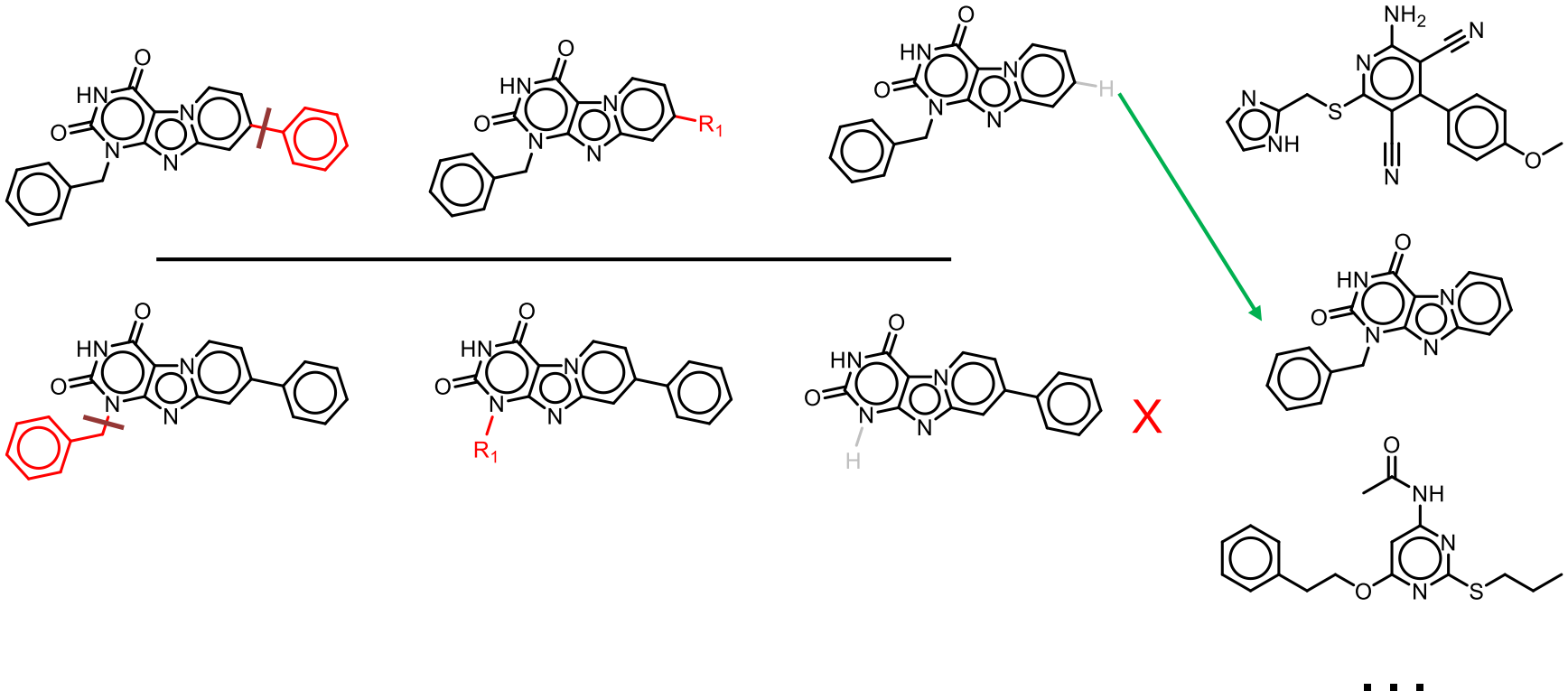
Hydrogen substitutions for MMPs

Single cuts

Cores

H-substitution

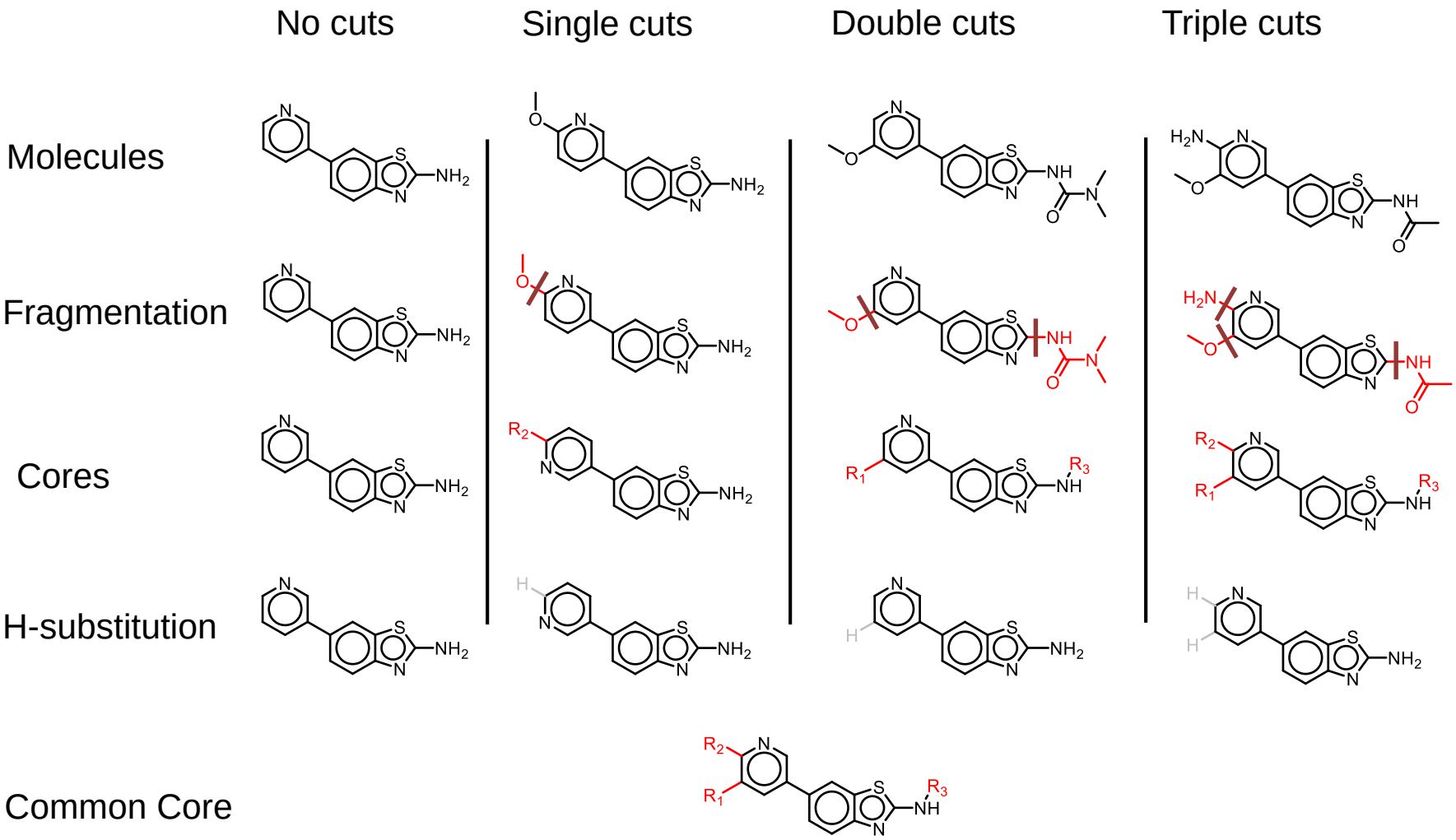
Molecules



What about hydrogen substitutions?

- AS solution
  - For each core generate a hydrogen-substituted core where all substitution sites are replaced by hydrogens
  - Group all cores (and original molecules) by their hydrogen-substituted core
  - All compounds belonging to cores with identical hydrogen-substituted cores form an AS
  - The cores are merged by introducing new substitution sites for substitution sites not shared by all cores

Hydrogen substitutions for AS



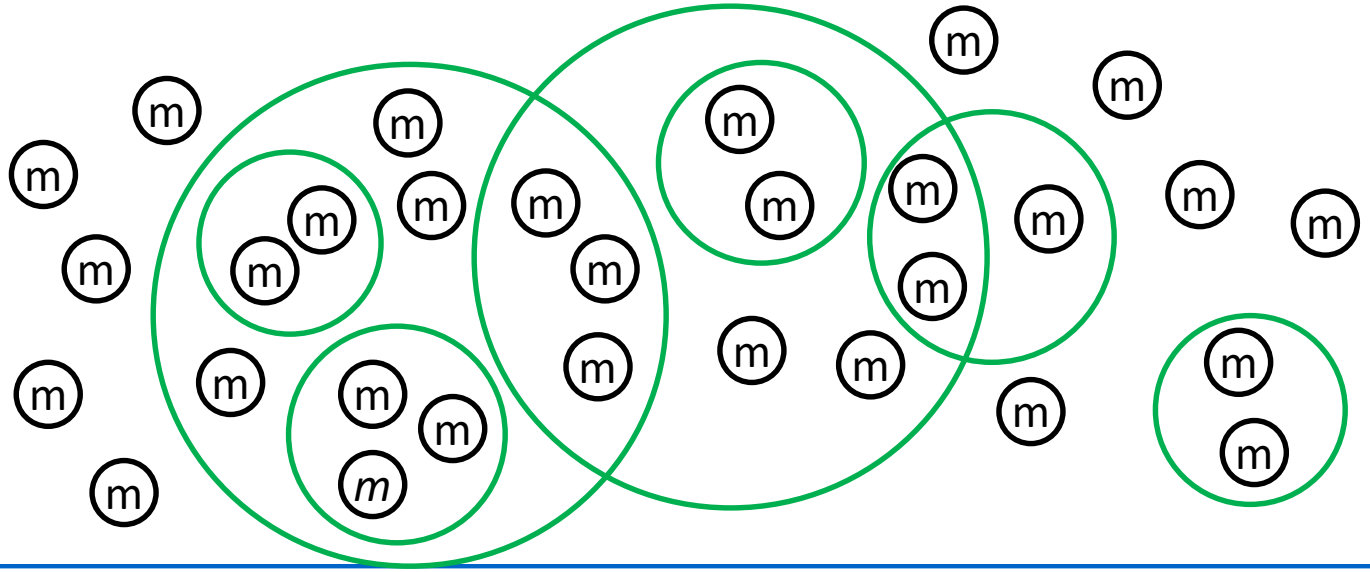
Basic algorithm: CCR method

Extraction of Compound Core Relationships

- Systematic fragmentation:
  - Fragment each molecule along cuttable acyclic bonds up to *n* times
  - For each molecule, all valid fragmentations are determined
- Collect all valid fragmentations for all molecules
- For all cores generate H-substituted cores
- Group valid fragmentations and molecules by H-substituted core
  - Yields a set of raw ASs
- "Housekeeping"

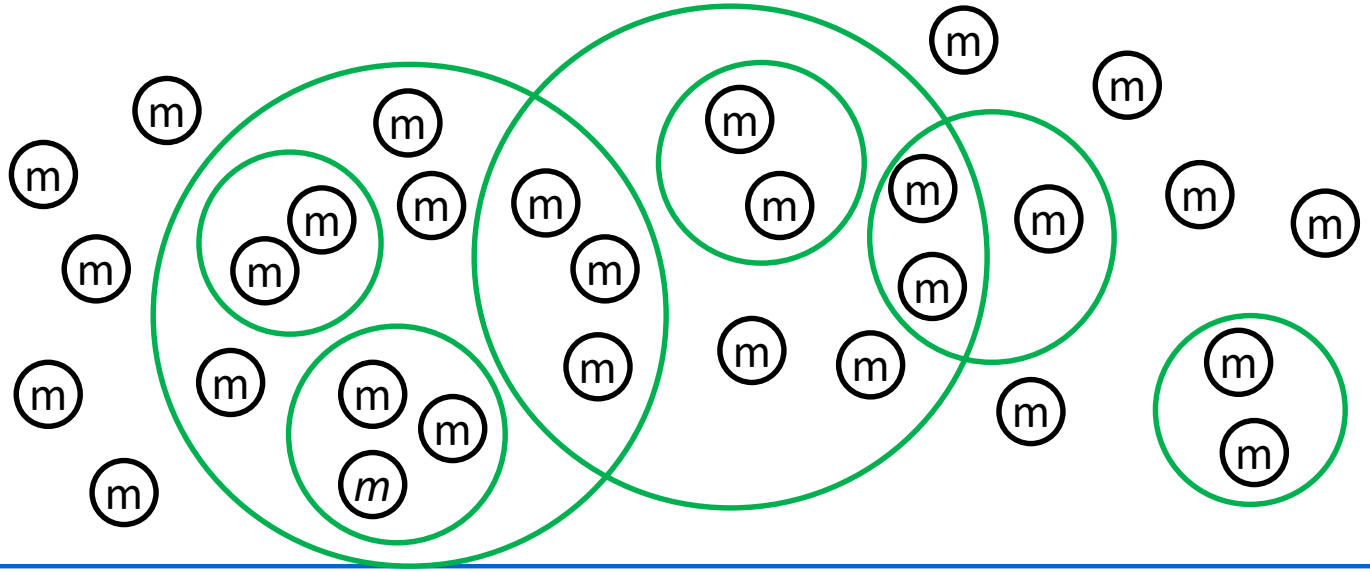
“Raw” results

- Process produces
  - Many single-compound AS
  - AS who are subsets of other AS
  - Overlapping AS



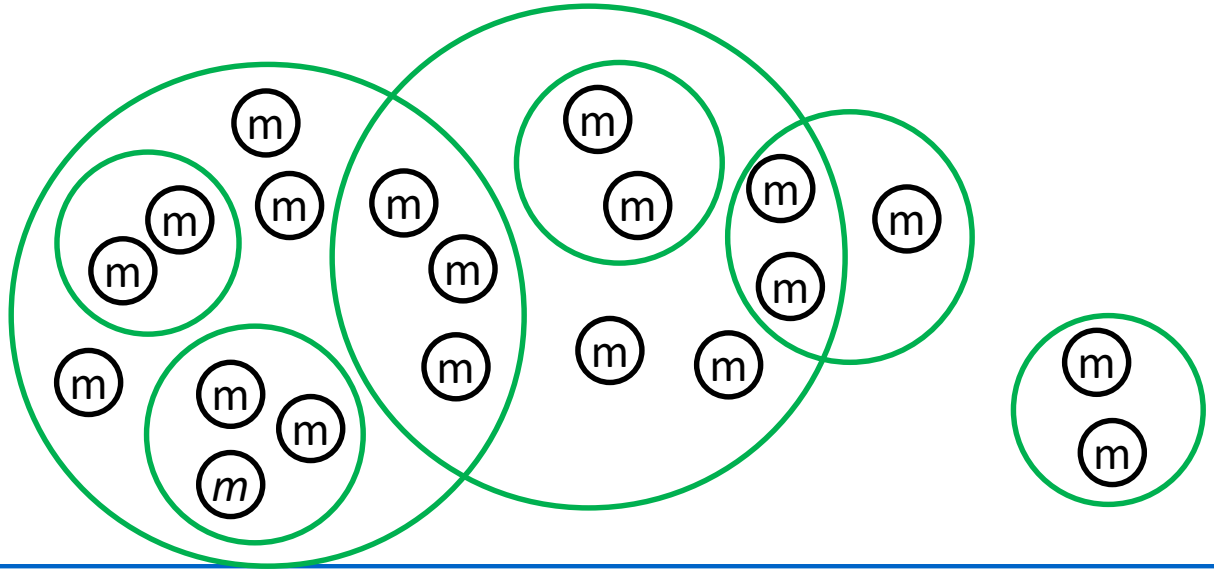
Housekeeping

- Remove singletons
- Remove AS-subseries contained in other series



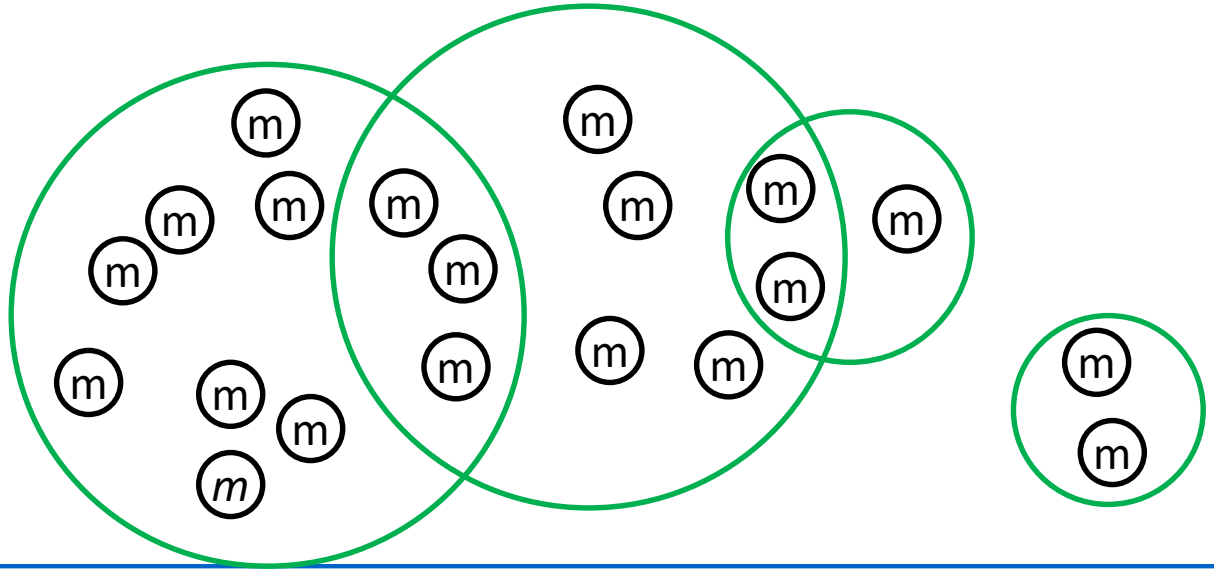
Housekeeping

- Remove singletons
- Remove AS-subseries contained in other series



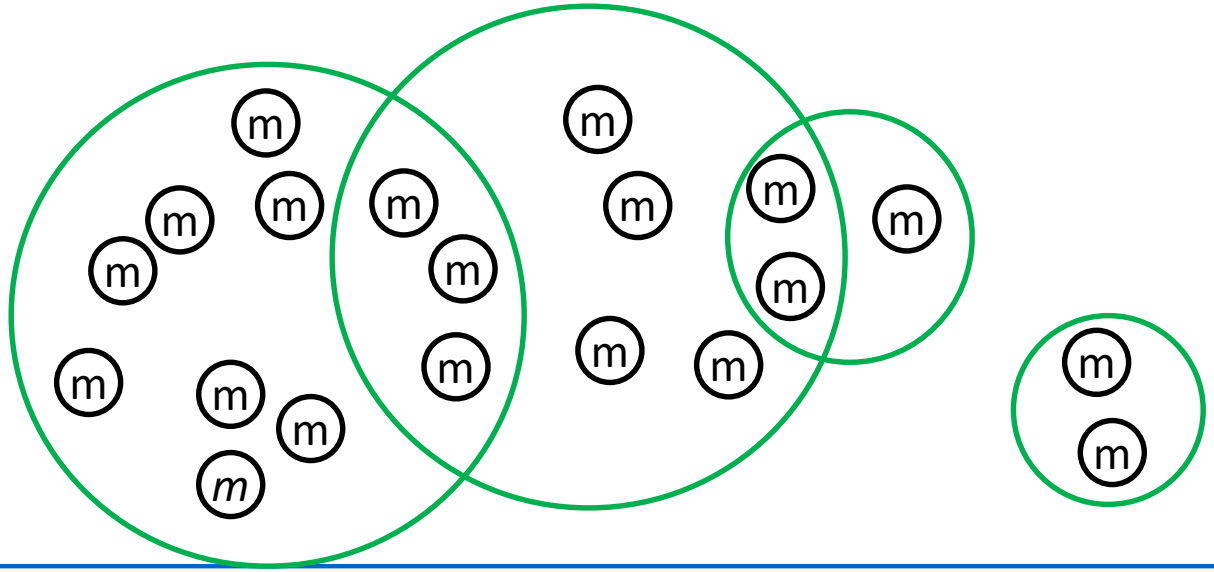
Housekeeping

- Remove singletons
- Remove AS-subseries contained in other series



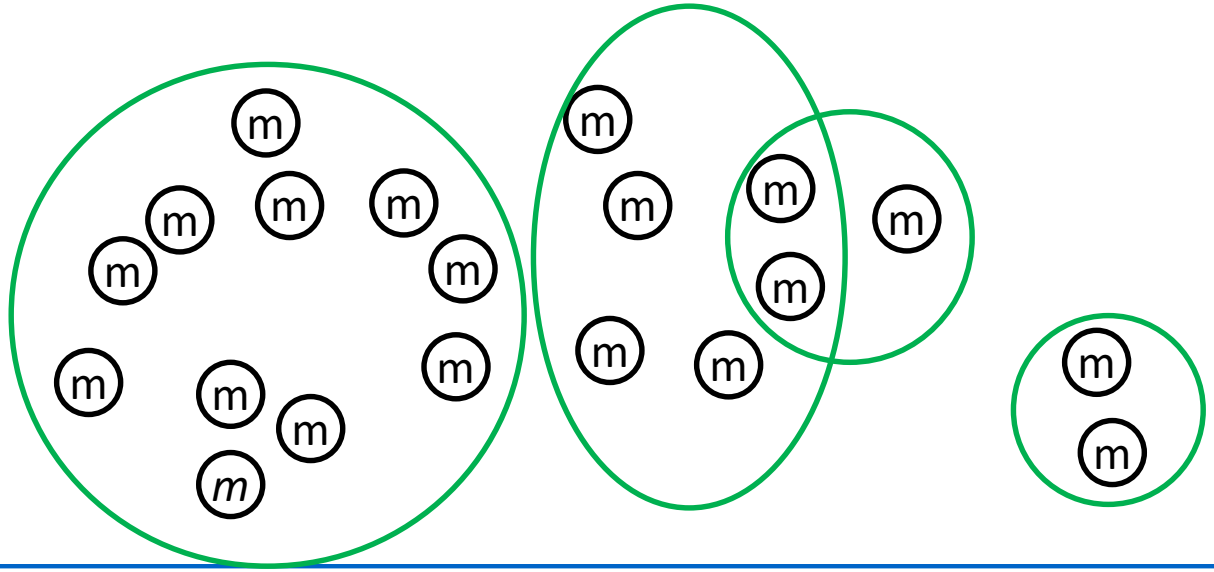
Housekeeping

- Remove singletons
- Remove AS-subseries contained in other series
- Uniquely assign compounds to series
  - prefer larger over smaller series
  - prefer larger cores
  - prefer fewer substitution sites



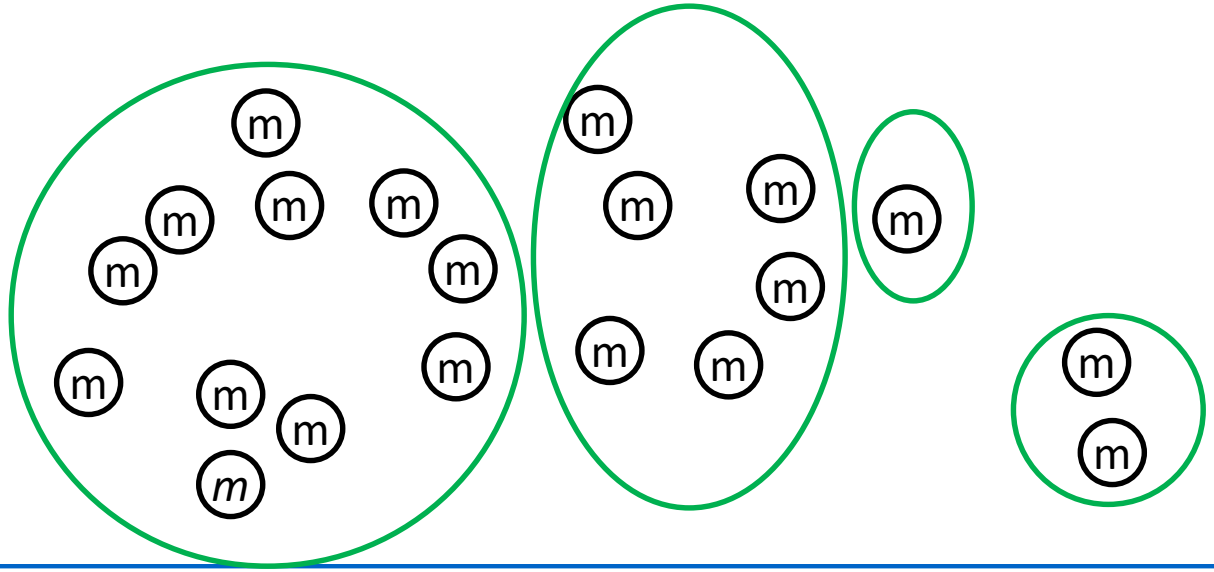
Housekeeping

- Remove singletons
- Remove AS-subseries contained in other series
- Uniquely assign compounds to series
  - prefer larger over smaller series
  - prefer larger cores
  - prefer fewer substitution sites



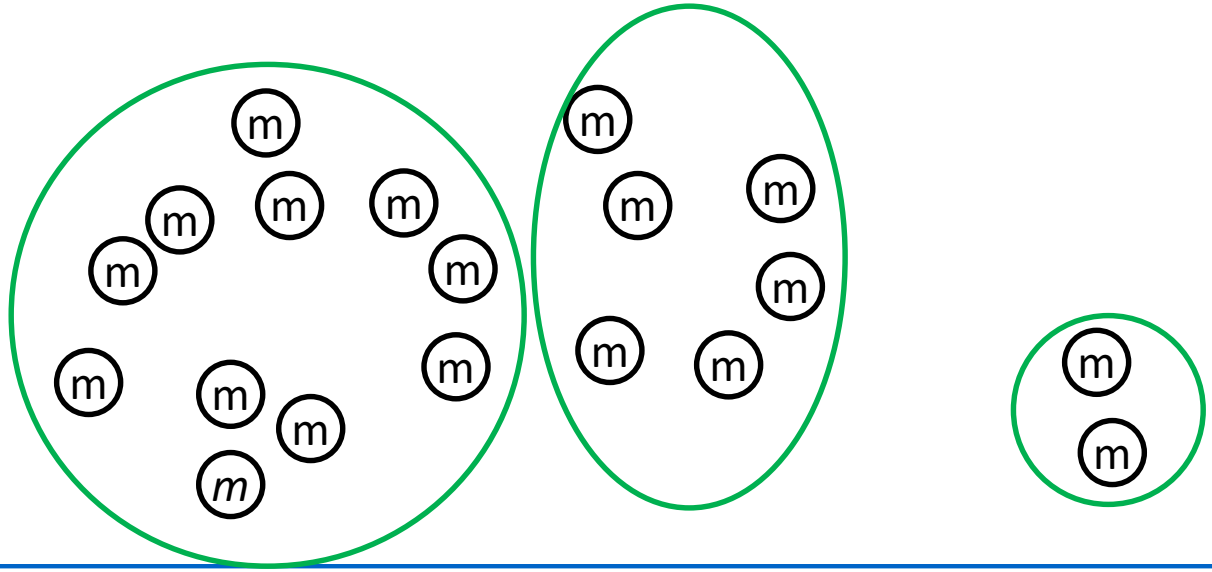
Housekeeping

- Remove singletons
- Remove AS-subseries contained in other series
- Uniquely assign compounds to series
  - prefer larger over smaller series
  - prefer larger cores
  - prefer fewer substitution sites



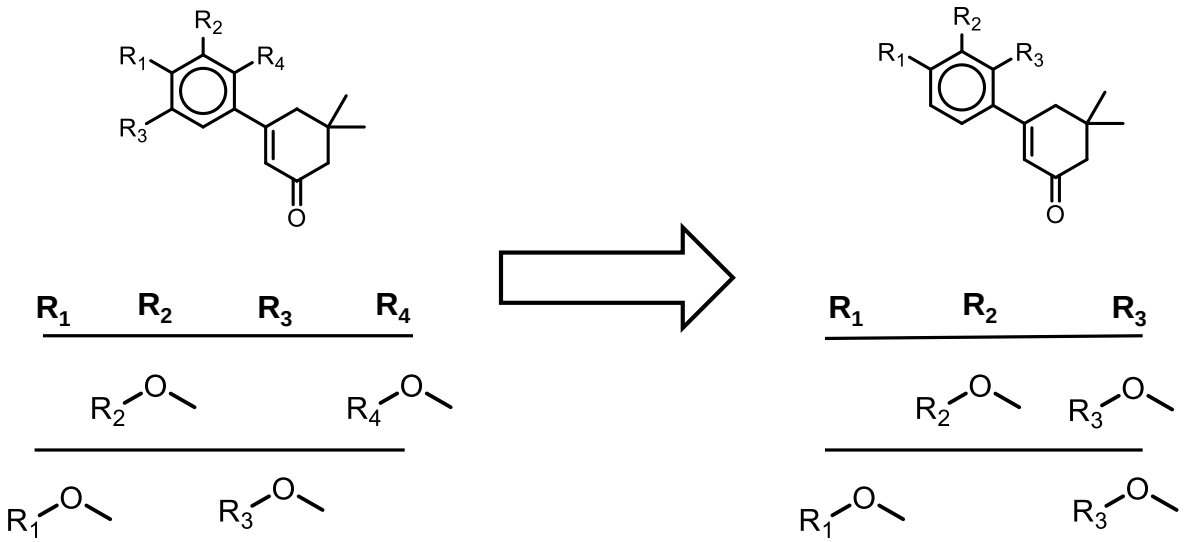
Housekeeping

- Remove singletons
- Remove AS-subseries contained in other series
- Uniquely assign compounds to series
  - prefer larger over smaller series
  - prefer larger cores
  - prefer fewer substitution sites



Final cleanup

- After “housekeeping”
  - Series may contain substitution sites with no variation
  - Series may contain symmetric substitution sites
- Final cleanup
  - optimize symmetric substitution sites to minimize variation of substituents
  - remove redundant substitution sites



Implementation notes

- Initial implementation has been done in OpenEye

- Currently all parts except for “final cleanup” has been recoded in RDKit (Python)
  - disregards stereochemical information
  - fragmentation is done using a recursive method
  - performance is competitive with “built-in” fragmentation routines
    - `Chem.Recap.RecapDecompose(mol)`
  - merging cores requires finding corresponding atoms of isomorphic molecule representations
    - can be done by canonical ordering
    - `Chem.CanonicalRankAtoms()`
    - `mol.GetProp("_smilesAtomOutputOrder")`

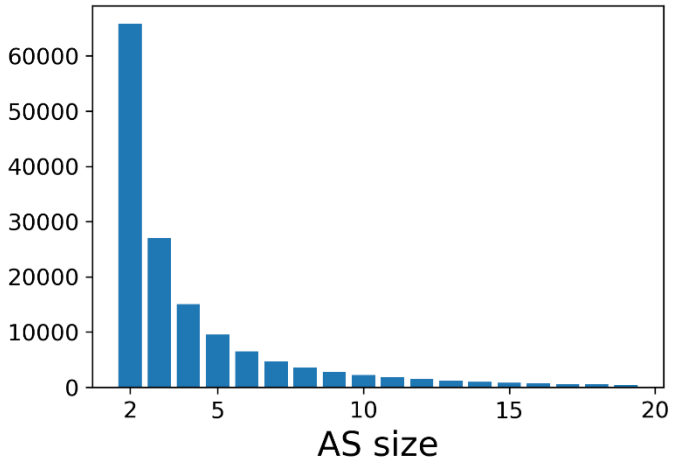
Exemplary results (ChEMBL25)

- 1,173,503 molecules (curated set, non-stereo-smiles)
- Skipped 263 mol. that exceeded time limit (10s/mol.)
- Timing (OpenEye Impl., i7-8700K, 3.7GHz)
  - Fragmentation: ~45 min. (6 cores)
  - AS generation: ~40 min. (single core)
- Settings:
  - Cuttable bonds: bonds according to RECAP rules
  - up to 5 cuts per mol.

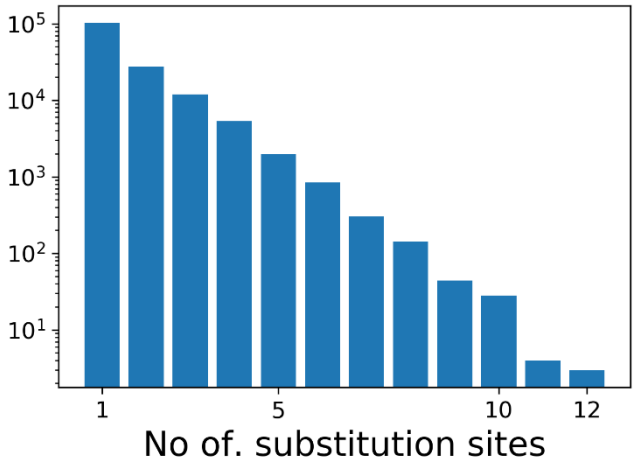
- # Fragmentations: 5,429,041 valid fragmentations (ca. 350 frags/s)
- # raw series: 4,505,385
  - # non-singleton series: 425,770
- # sanitized series: 150,141 (723,245 molecules)

Exemplary results (ChEMBL25)

Size distribution of ASs



Number of substitution sites per AS



10953 AS with 10-19 mol.
4133 AS with 20-99 mol.
42 AS with ≥100 mol.
largest AS: 233 mol.

Summary

- The CCR method allows the systematic extraction of AS from large compound databases of millions of compounds

- The inherent difficulty in providing a consistent AS definition requires a data driven approach
  - Core structures are defined based on plausible assumptions that are computationally feasible
  - Core structures of cleaned-up AS are data dependent
  - Molecules are assigned preferentially to large AS depending on the data set

b-it LSi
Life
Science
Informatics

LIMES
Life & Medical Sciences Institute

Acknowledgment

Jesús Naveja
Dagmar Stumpfe
Jürgen Bajorath

ACS Omega, 2019, 4, 1, 1027-1032

UNIVERSITÄT BONN