A Generic Group Contribution Method

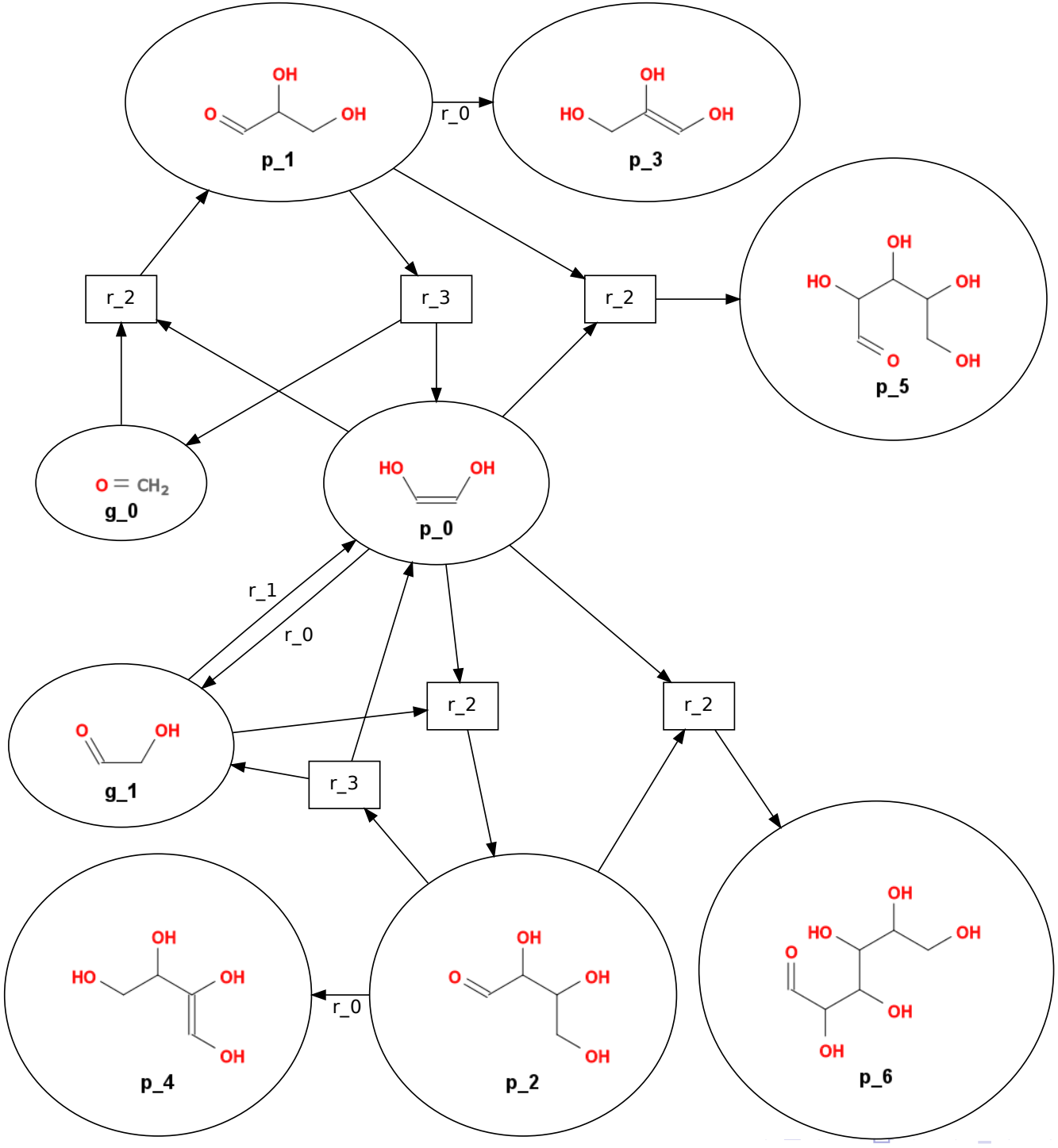
Daniel Merkle Nikolai Nøjgaard

Institut for Matematik og Datalogi
Syddansk Universitet

February 15, 2016

Generative Chemistries

- The study of exploring chemical spaces of unknown compounds.
- Given a set of molecules and a set of reactions, the chemical space is modeled as a hypergraph.
- Inferring hyperpaths in hypergraphs.


p_1
p_3
p_5
p_0
g_0
g_1
p_4
p_2
p_6
r_0
r_1
r_2
r_3

Motivation

- Asses the chemical quality of hyperpaths.
- Affected by various chemical properties.
- Wetlab? Expensive..
- Need a predictive method!

The Group Contribution Method!

- Assumes a linear relationship between property and chemical structures (groups).
- Decomposes molecules into a set of groups.
- The target property can then be predicted as the sum of contributions of its corresponding groups:

$$t = \sum_{i \in G} G_i \cdot C_i$$

Matthew D. Jankowski, Christopher S. Henry, Linda J. Broadbelt, Vassily Hatzimanikatis (2008): Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks, *Biophysical Journal, V. 95*

The Three Problems

- Group Identification.
- Compound Decomposition.
- Model Learning.

Group Identification

- Expert knowledge required.
- Differentiates on non-topological characteristics.
  - Aromatic rings etc.
- Assigned a priority.

TABLE 1 Structural groups used in group contribution method

Description of molecular substructure
Molecular substructures involving halogens
–Cl (attached to a primary carbon with no other chlorine atoms attached)*
–Cl (attached to a secondary carbon with no other chlorine atoms attached)*
–Cl (attached to a tertiary carbon with no other chlorine atoms attached)*
–Cl (attached to a primary carbon with one other chlorine atom attached)*
–Cl (attached to a secondary carbon with one other chlorine atom attached)*
–Cl (attached to a primary carbon with two other chlorine atoms attached)*
–Br (attached to an aromatic ring)*
–I (attached to an aromatic ring)*
–F (attached to an aromatic ring)*

Compound Decomposition

- Given a set of groups G find the frequency they occur in a compound C, such that every vertex of C is assigned to exactly one group.
- The monomorphism problem!
  - NP-complete...
- Given G, C, and a set of rules R, a graph decomposition is a function:

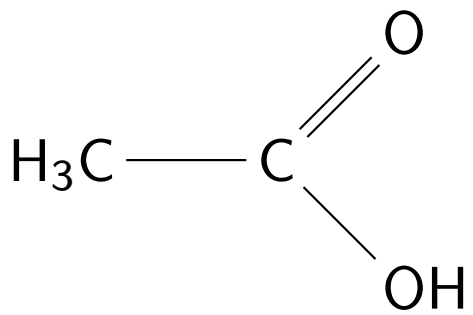
$$f : f(G, C, R) \rightarrow F$$

  such that F_i corresponds to the number of monomorphisms from G_i to C that is valid under R.
- Results may vary wildly.

Ronald G. Forsyth, Peter D. Karp, Michael L. Mavrovouniotis (1997): Estimation of equilibrium constants using automated group contribution methods, *CABIOS, V. 13*

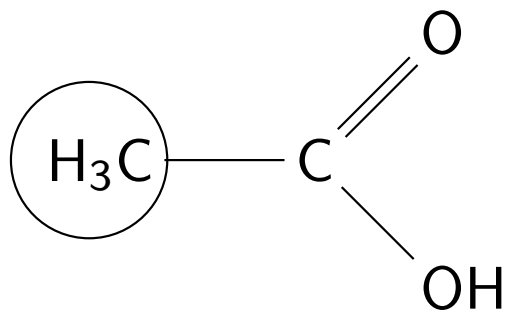
Example

H_3C, OH, W$=$O, C$-$C



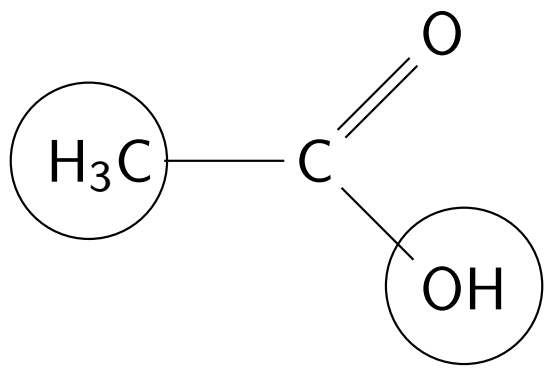
Example

H_3C, OH, W$=$O, C$-$C



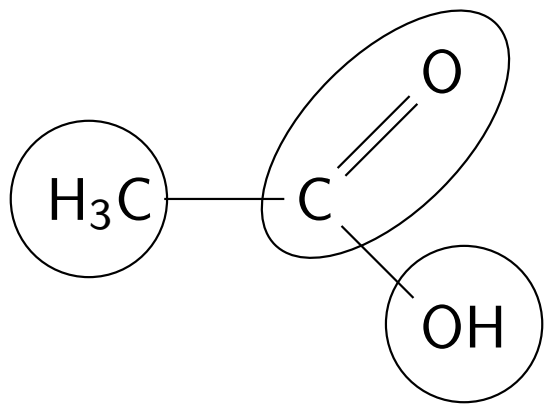
Example

H_3C, OH, W$=$O, C$-$C



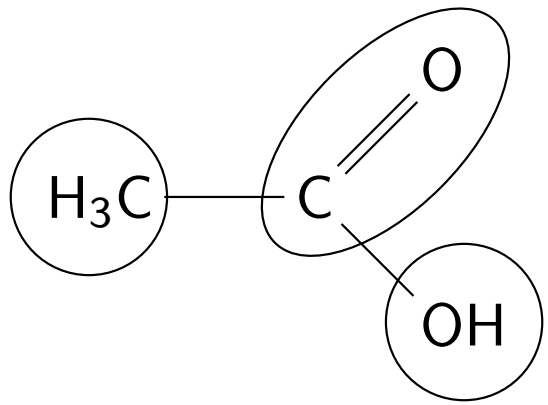
Example

H_3C, OH, W$=$O, C$-$C



Example

H_3C, OH, W$=$O, C$-$C



$$F = \{1, 1, 1, 0\}$$

Model Learning

$$
\begin{aligned}
F_0 &= \{1, 1, 1, 0\} & t_0 &= 20 \\
F_1 &= \{0, 2, 1, 0\} & t_1 &= -2 \\
&\vdots & &\vdots \\
F_i &= \{F_{i0}, F_{i1}, F_{i3}, F_{i4}\} & t_i &= t_{iv}
\end{aligned}
$$

- Ordinary Least Squares Regression:

$$\texttt{min}(\sum_{i=1..|t|} (t_i - F_i^T b)^2)$$

- Validated with Cross Validation.

Shortcomings Of The Current Approaches

- Expert knowledge required.
- Limited to few chemical spaces.
- Priority setting.
- Introducing new compounds.
- We need something flexible!

A Generic Approach - Goals

- Automatic group identification.
- Consistency in predictive estimations.
- Fast predictive decomposition.
- Main goal is not to out-perform existing implementations.

Generic Group Identification

- Potentially $2^{|V(g)|}$ different subgraphs.
  - Not feasible.
- Repeating patterns might be important.
- Frequent Subgraph Mining.
  - Also NP-complete... But feasible!
- Only simple groups.

Xifeng Yan, Jiawei Han (2003): Graph-Based Substructure Pattern Mining, *ICDM*

Example

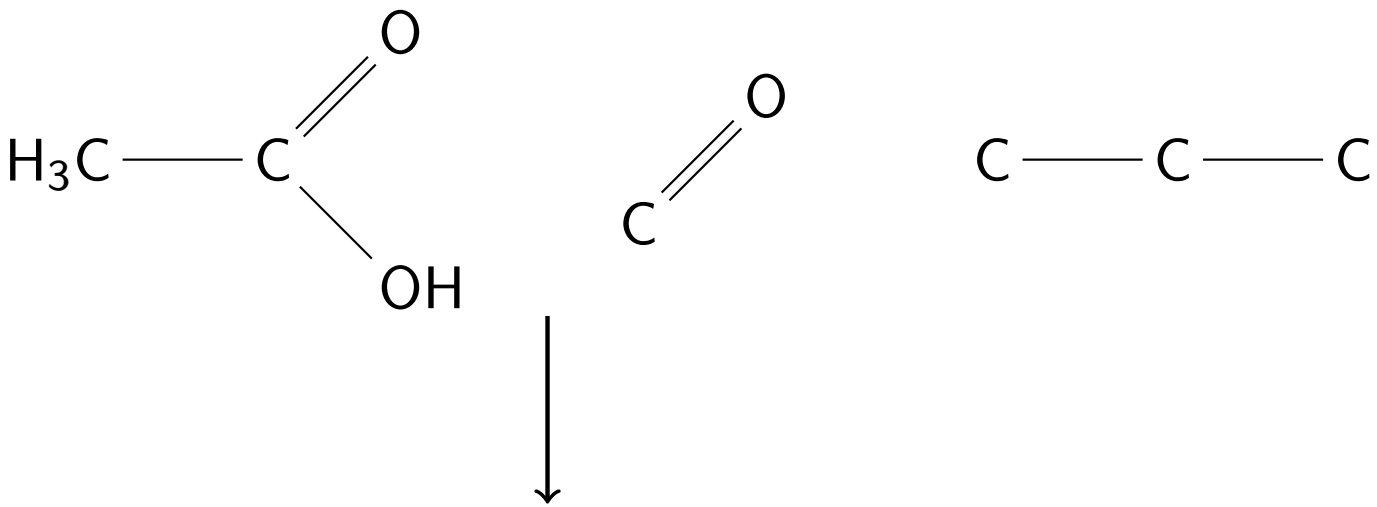


Figure: Frequent Subgraph mining with `min_support`= 2

Example

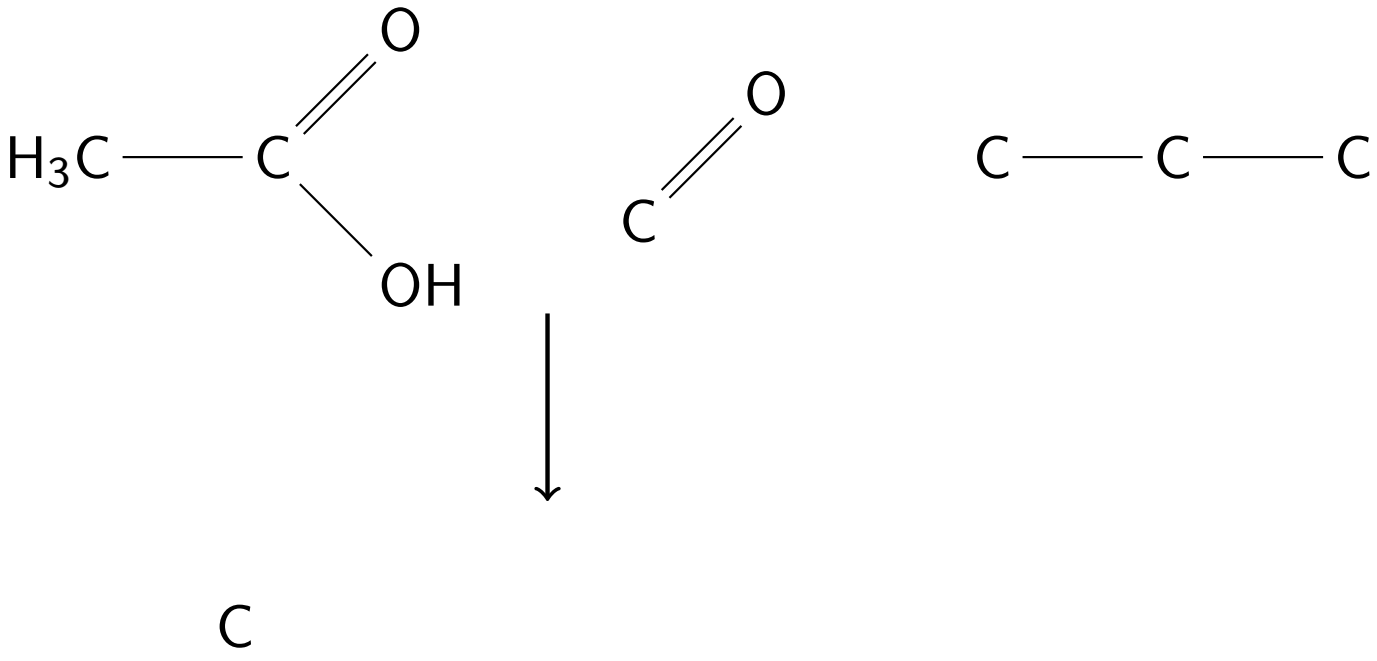


Figure: Frequent Subgraph mining with `min_support`= 2

Example

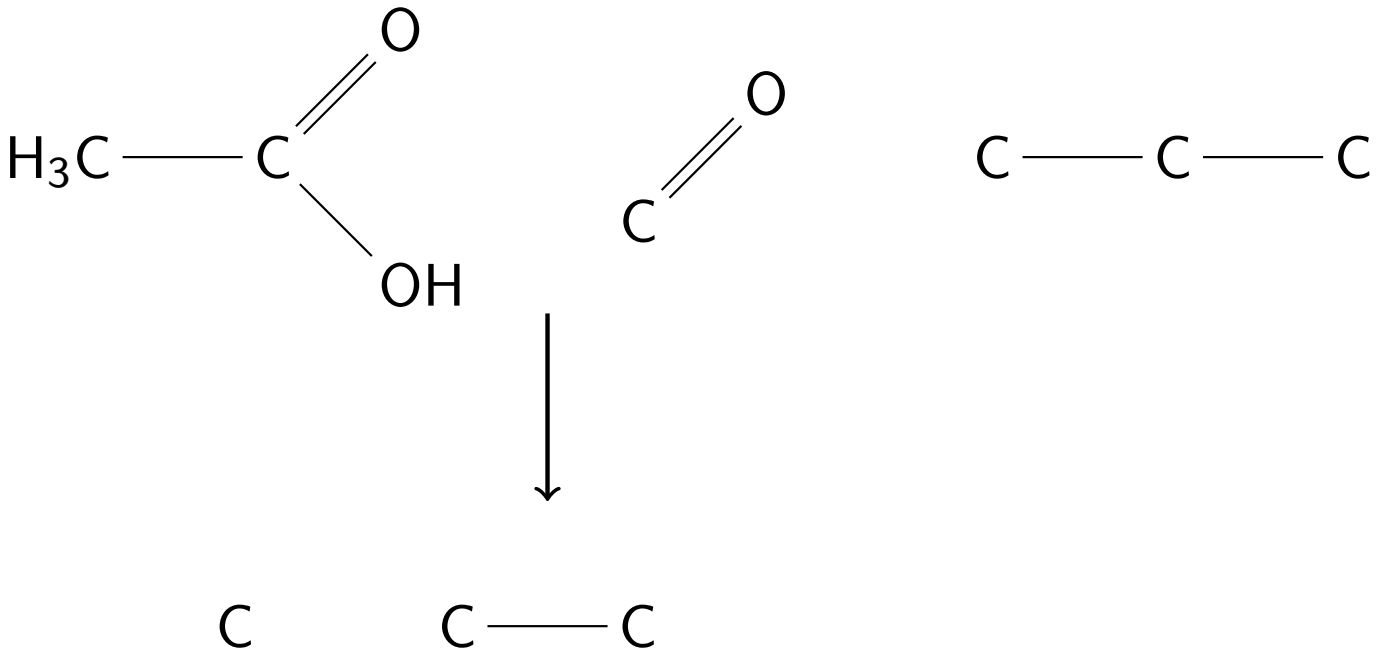


Figure: Frequent Subgraph mining with `min_support`= 2

Example

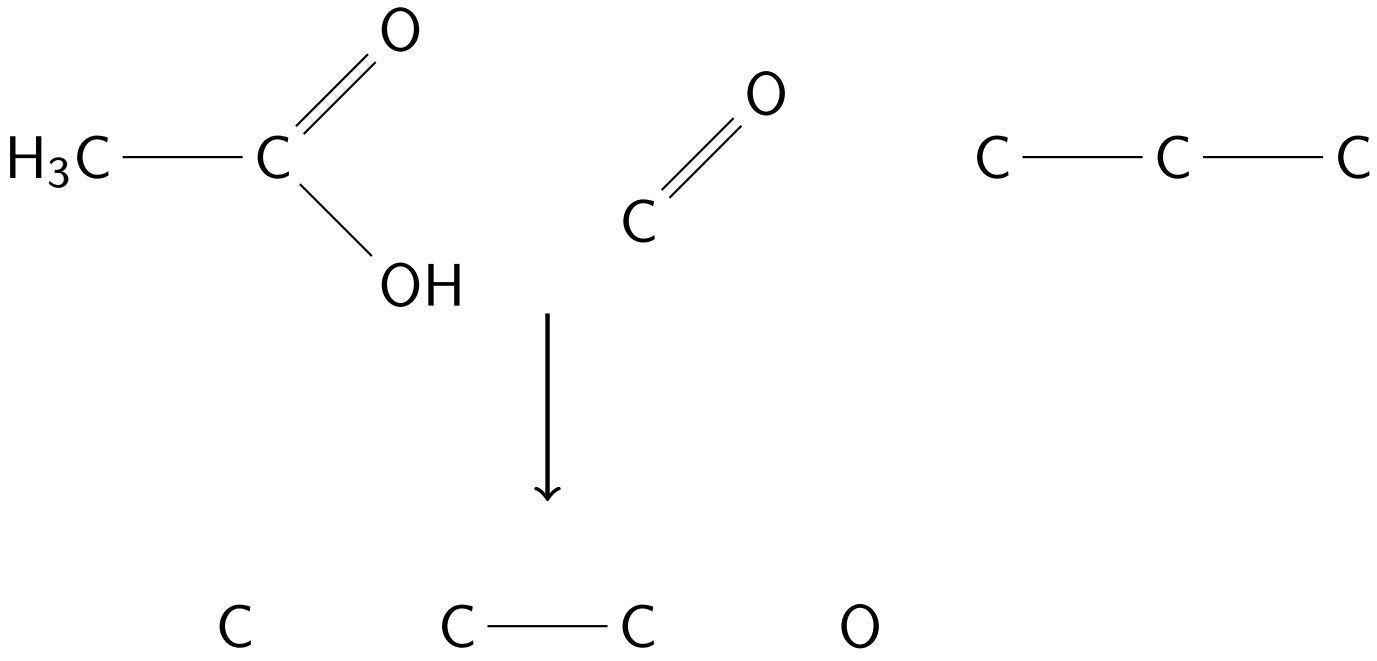


Figure: Frequent Subgraph mining with `min_support`= 2

Example

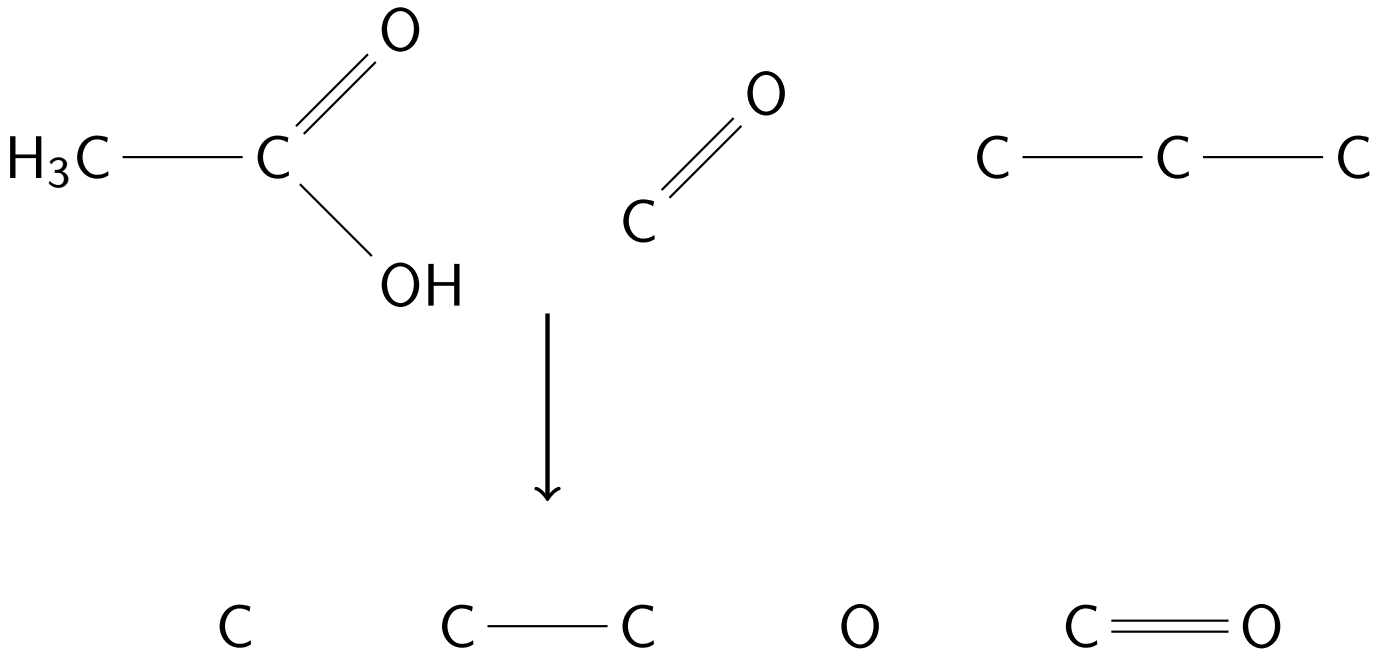


Figure: Frequent Subgraph mining with `min_support`= 2

Generic Compound Decomposition

- Still finding monomorphisms.
  - Still NP-complete..
- No priorities.
- Reminder, just a function:

$$f : f(G, C, R) \rightarrow F$$

- Overlapping allowed.
- Beware of collinearity.

Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento (2004): A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs, *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, V. 26*

Example

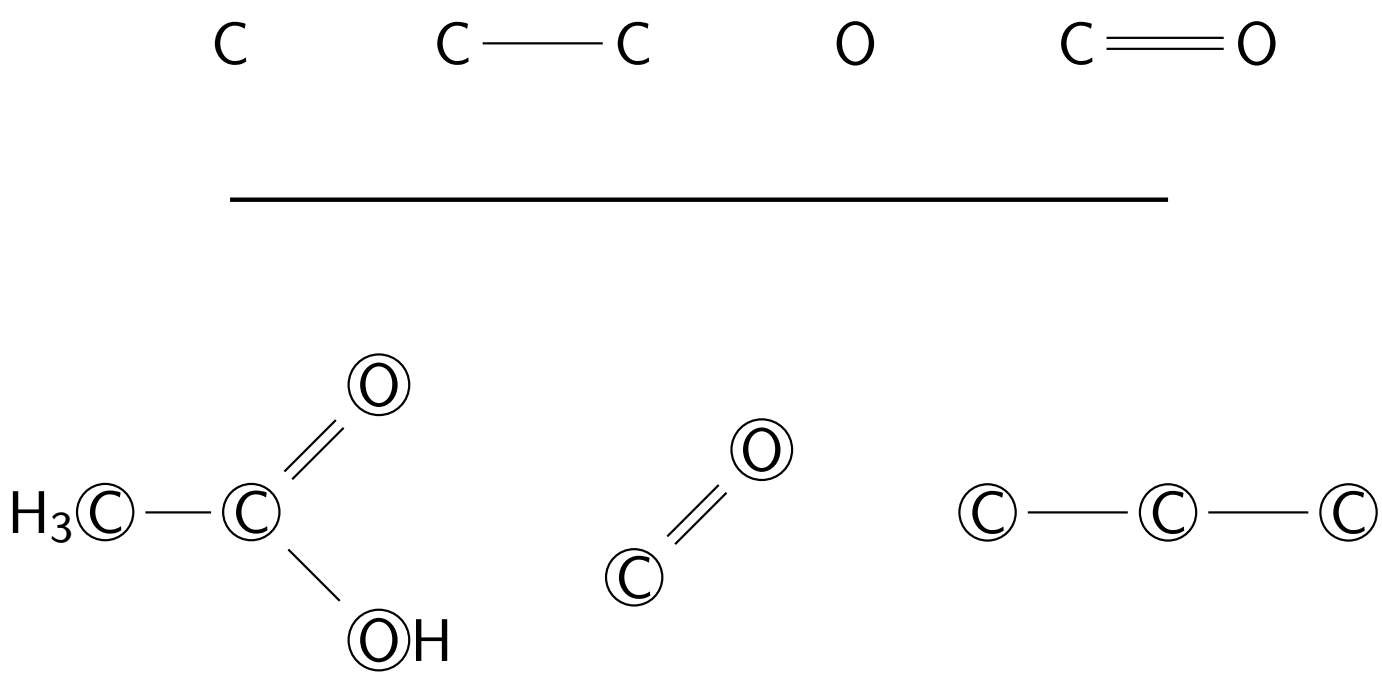


Figure: Overlapping Graph Decomposition

Example

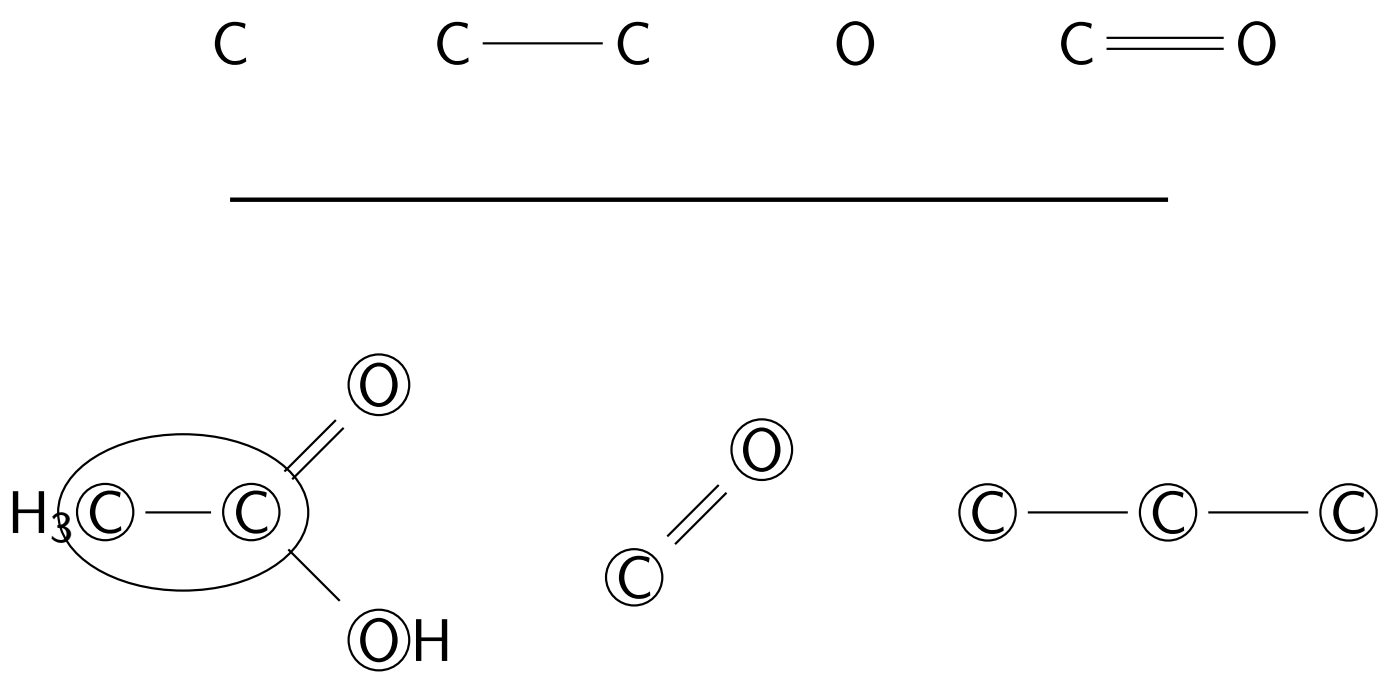


Figure: Overlapping Graph Decomposition

Example

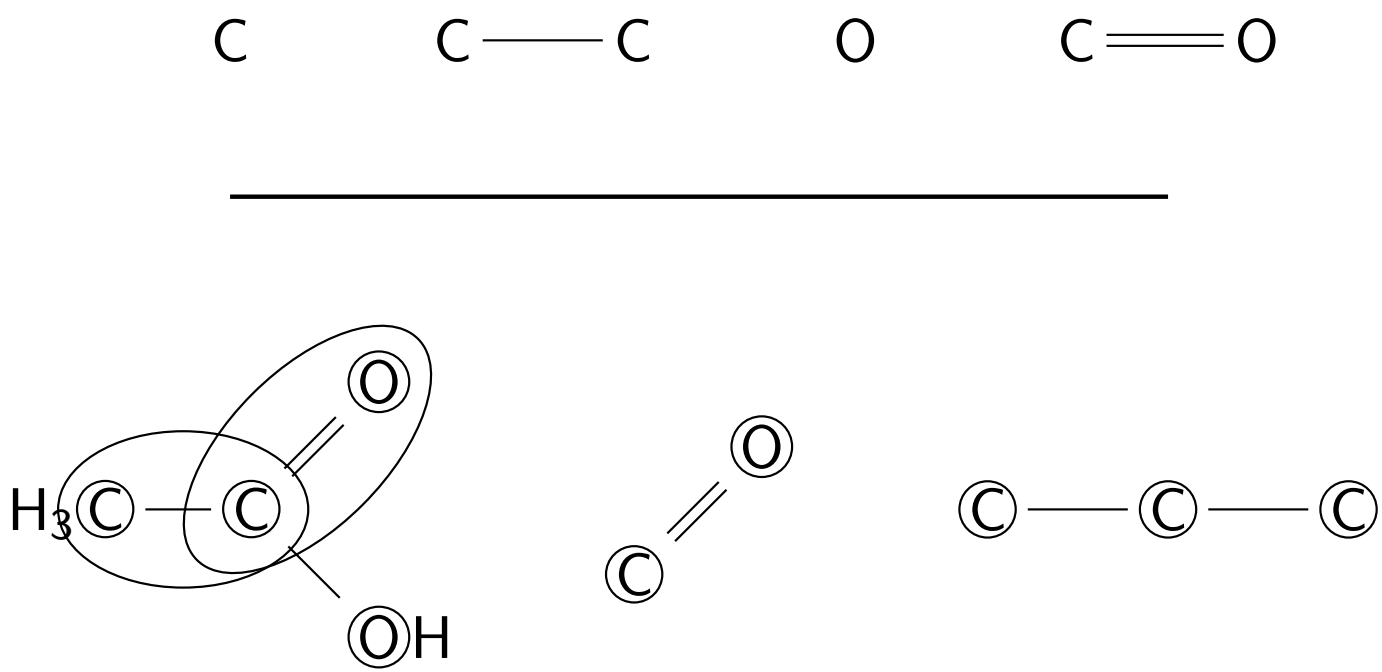


Figure: Overlapping Graph Decomposition

Example

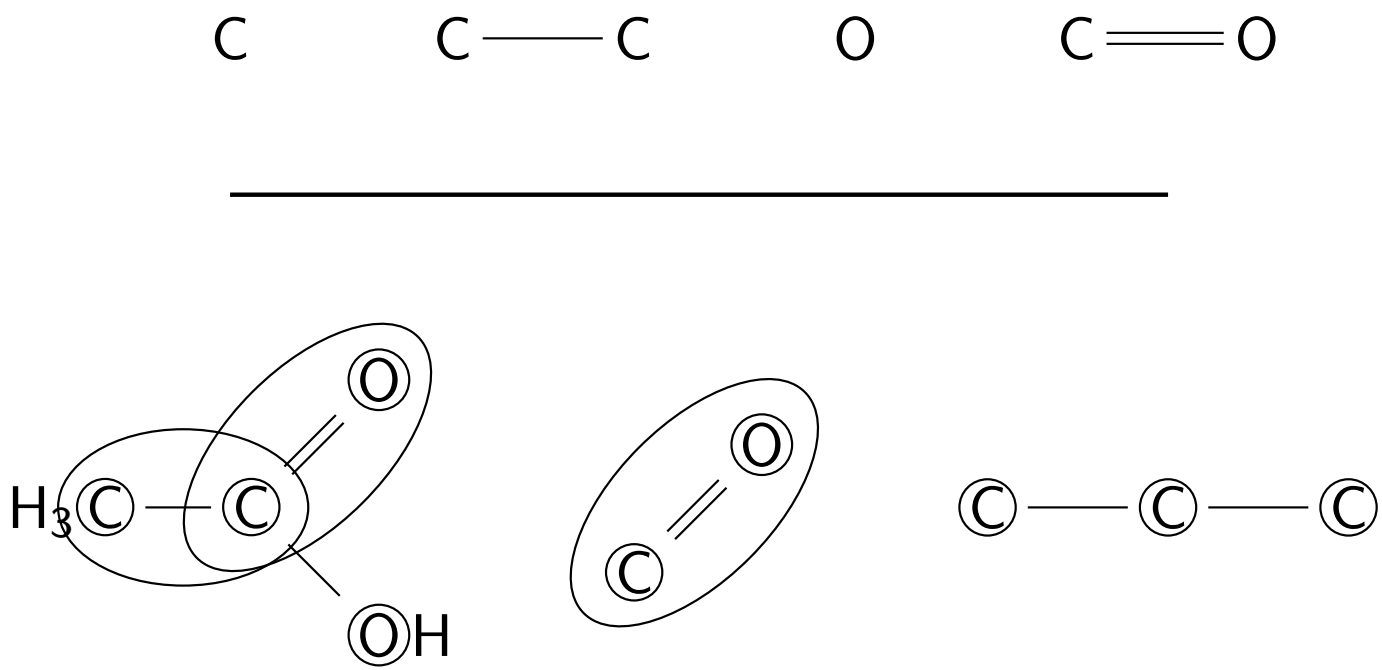


Figure: Overlapping Graph Decomposition

Example

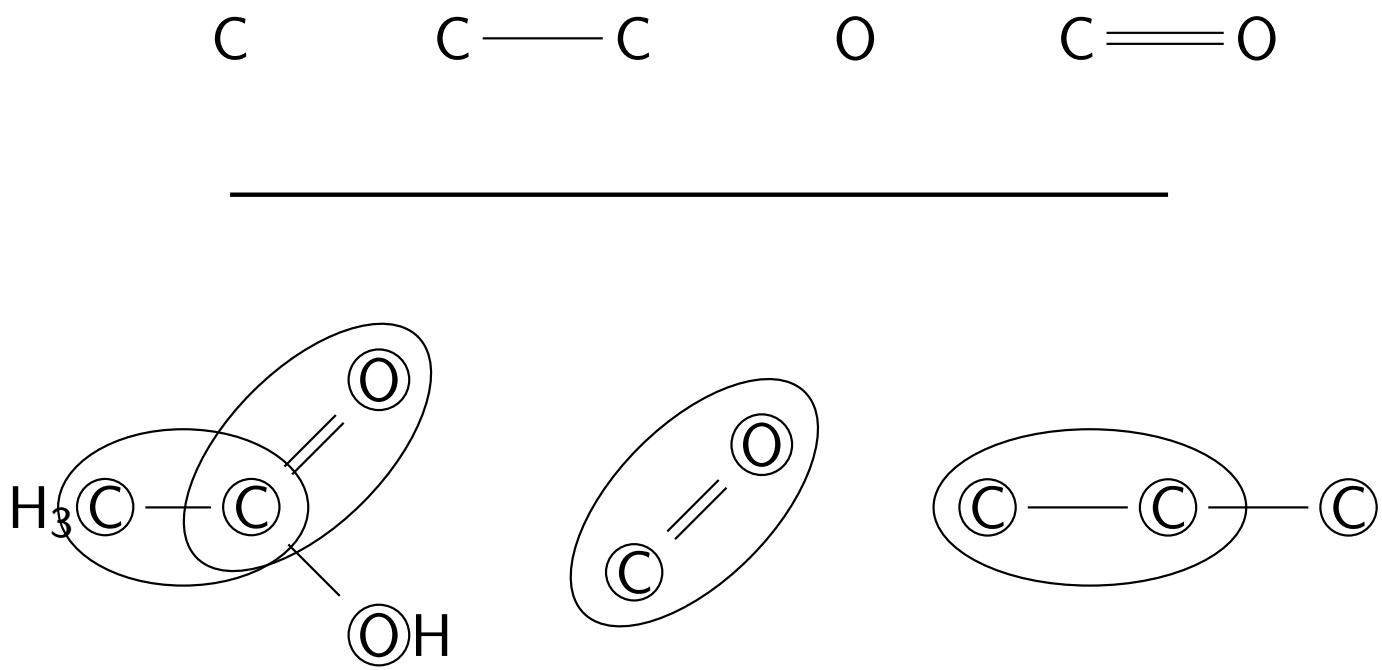


Figure: Overlapping Graph Decomposition

Example

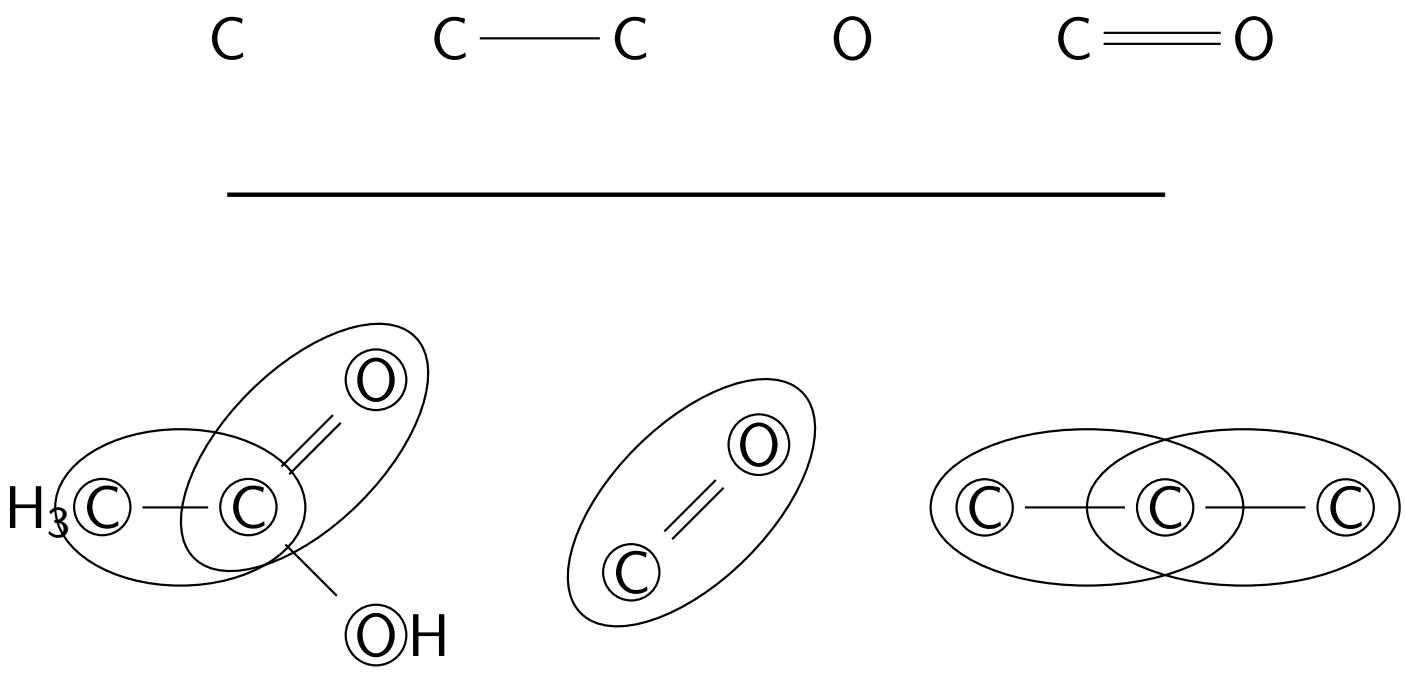


Figure: Overlapping Graph Decomposition

Generic Model Learning

- Many variables few data points.
- Ordinary Least Squares at its worst.
- Let's look at some possible alternatives.

The Suitors

- Ordinary Least Squares
  - Already out.

The Suitors

- Ordinary Least Squares
  - Already out.
- Principle Component Regression
  - Computes the entire eigen matrix.
  - Forced to potentially use all variables.
  - Does not determine importance of components based on target property.

The Suitors

- Ordinary Least Squares
  - Already out.
- Principle Component Regression
  - Computes the entire eigen matrix.
  - Forced to potentially use all variables.
  - Does not determine importance of components based on target property.
- Partial Least Squares
  - Components can be computed iteratively.
  - Includes target properties in component selections.
  - Still forced to use all original variables.

The Suitors

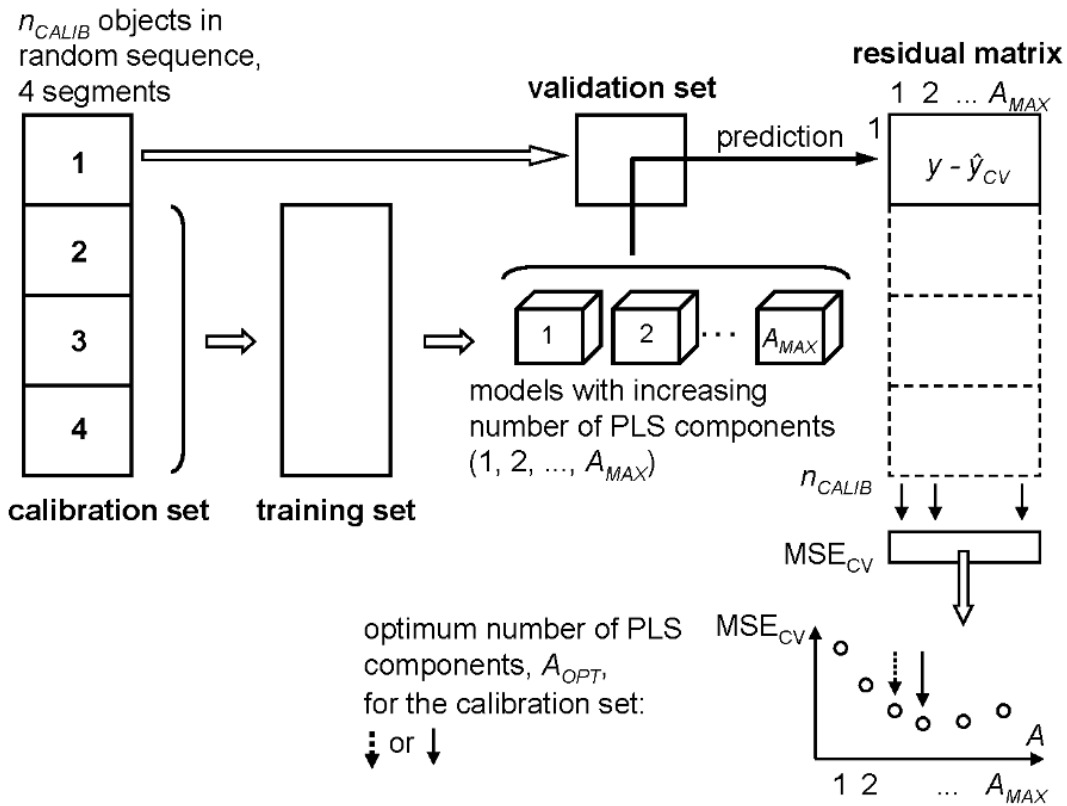
- Ordinary Least Squares
  - Already out.
- Principle Component Regression
  - Computes the entire eigen matrix.
  - Forced to potentially use all variables.
  - Does not determine importance of components based on target property.
- Partial Least Squares
  - Components can be computed iteratively.
  - Includes target properties in component selections.
  - Still forced to use all original variables.
- Stepwise Regression
  - Includes feature selection.
  - Very sensitive to collinearity between variables.

The Suitors

- Ordinary Least Squares
  - Already out.
- Principle Component Regression
  - Computes the entire eigen matrix.
  - Forced to potentially use all variables.
  - Does not determine importance of components based on target property.
- Partial Least Squares
  - Components can be computed iteratively.
  - Includes target properties in component selections.
  - Still forced to use all original variables.
- Stepwise Regression
  - Includes feature selection.
  - Very sensitive to collinearity between variables.
- Least Absolute Shrinkage and Selection Operator (LASSO)
  - Also includes feature selection.
  - Can be adjusted to be less sensitive to collinearity.
  - Sounds promising!

Model Validation

- Repeated Double Cross Validation
- Optimize model complexities while giving realistic estimations of prediction errors.



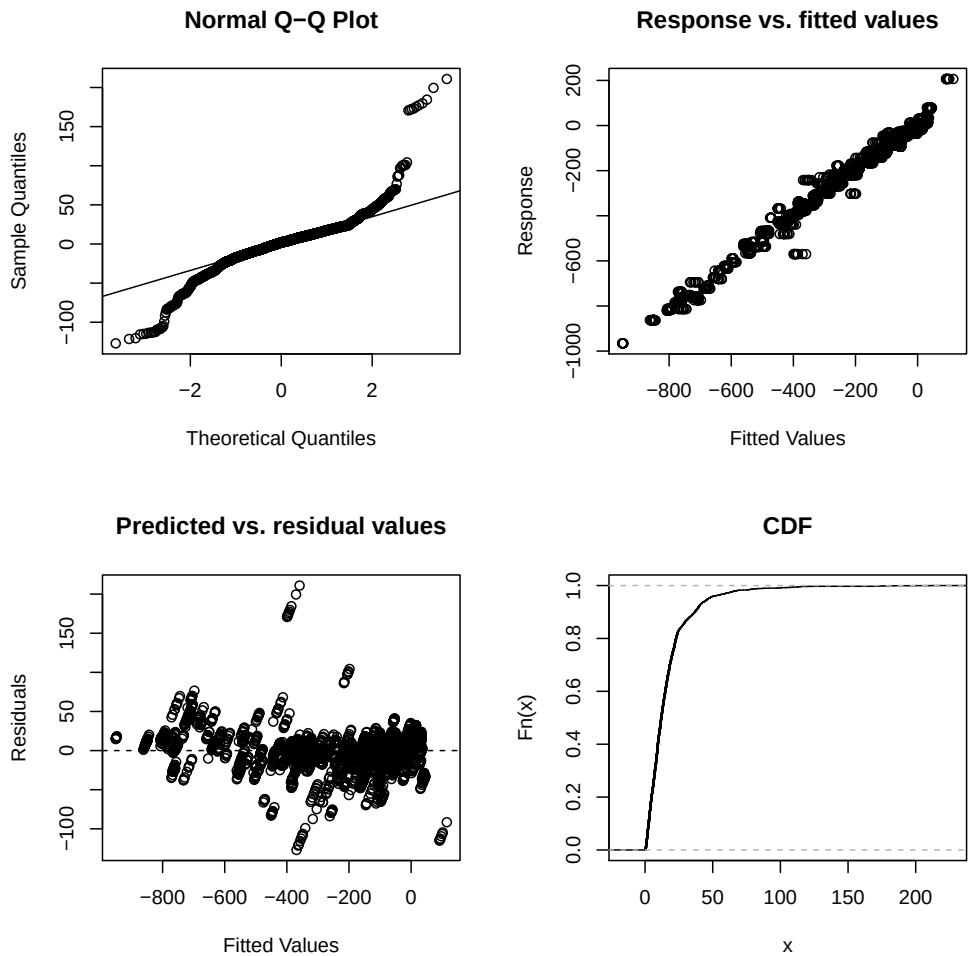
Peter Filzmoser, Bettina Liebmann, Kurt Varmuza (2009): Repeated double cross validation, *Journal of Chemometrics, V. 23*

Single Pass Limitations

- Hard to control granularity.
- `min_support` too high = over fitting.
- `min_support` too low = under fitting.

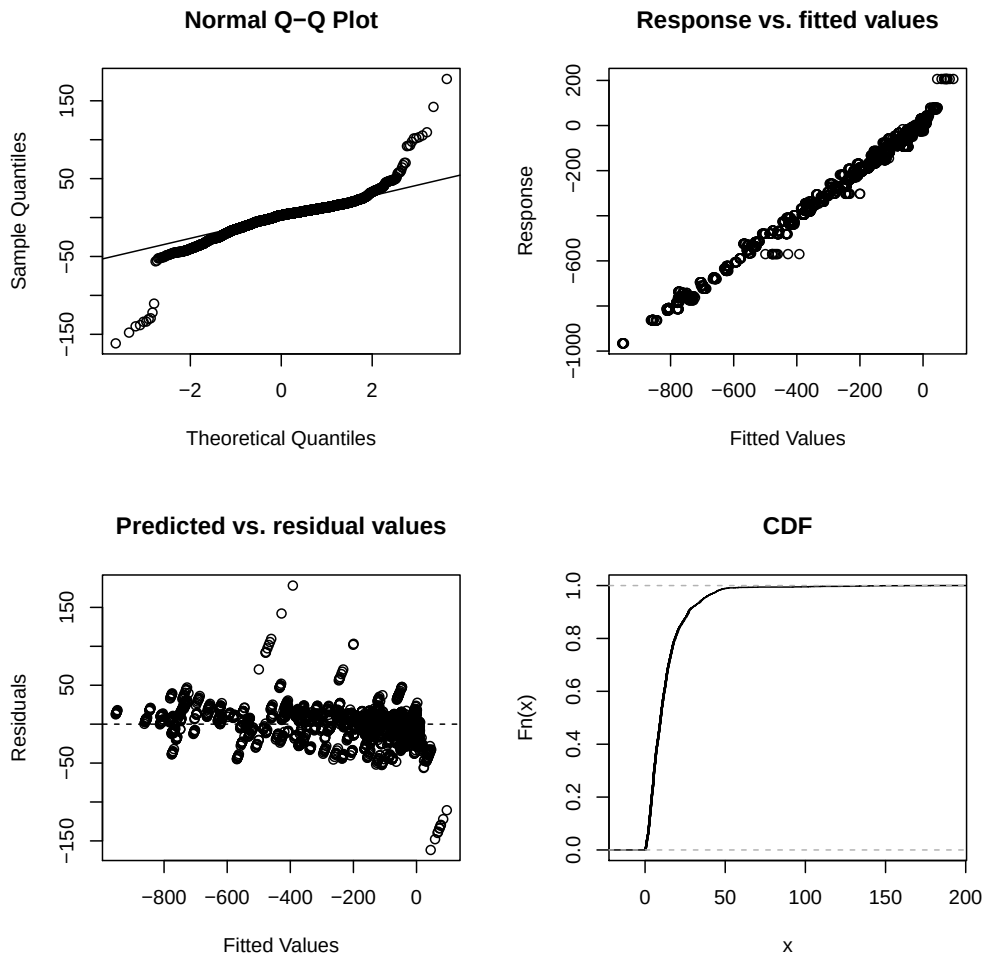
Under Fitting

Figure: Model learned with groups occurring in one third of the compounds. Cross validation repeated 10 times



Over Fitting

Figure: Model learned with all groups occurring in at least one compound and is smaller than 7 atoms. Cross validation repeated 10 times



Solution? Think iteratively!

Algorithm 1 `learnGroups`(C, i)

$m \leftarrow \frac{|C|}{3}$
$G \leftarrow \texttt{gSpan}(C, m)$
$y \leftarrow \texttt{properties}(C)$
$X \leftarrow \texttt{decomposed}(C, G)$
$M \leftarrow \texttt{learn}(X, y)$
while $i > 0$ **do**
    $O \leftarrow \texttt{outliers(M)}$
    $G' \leftarrow \texttt{gSpan}(O, m)$
    $G' \leftarrow G'/(G \cap G')$
    if $G' = \emptyset$ **then**
        $m \leftarrow m - 1$
        `continue`
    end if
    $G \leftarrow G' \cup G$
    Decompose compounds, learn model based on the new G, and decrement i
end while
return M

Lets Try It Out On Thermodynamics

Figure: PLS. Groups smaller than 7. 6 iterations

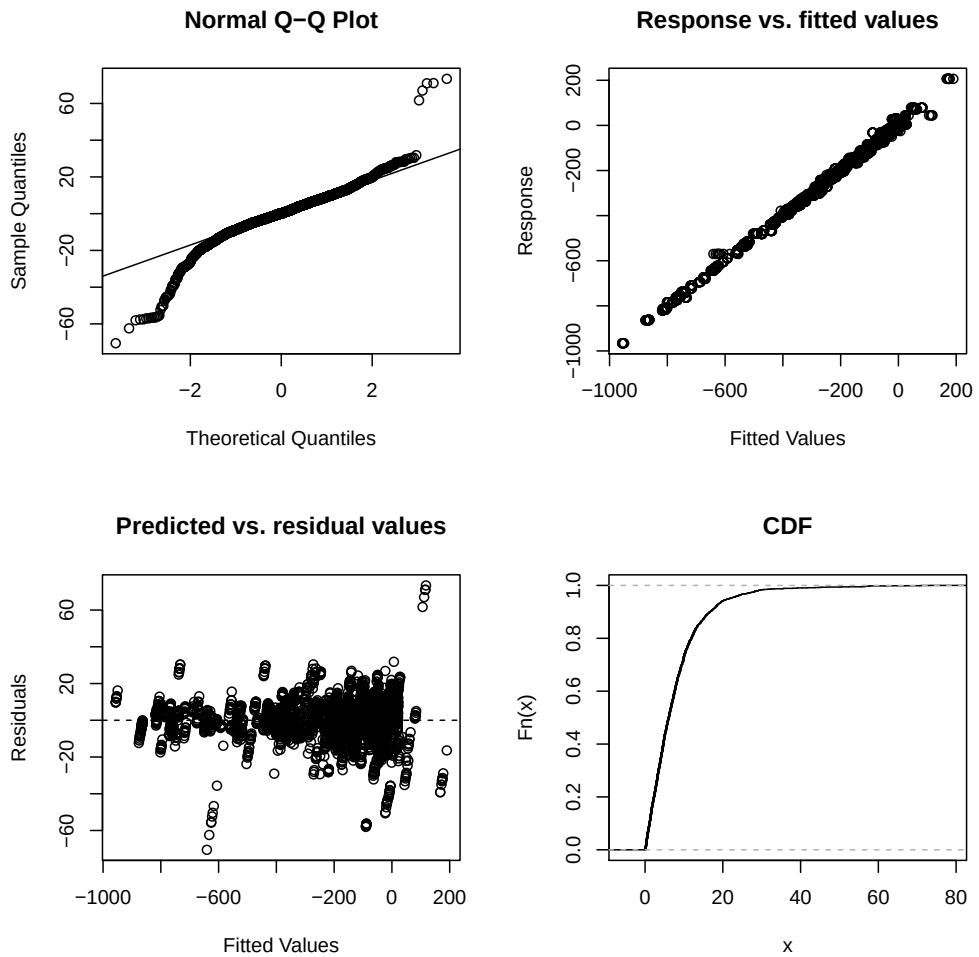
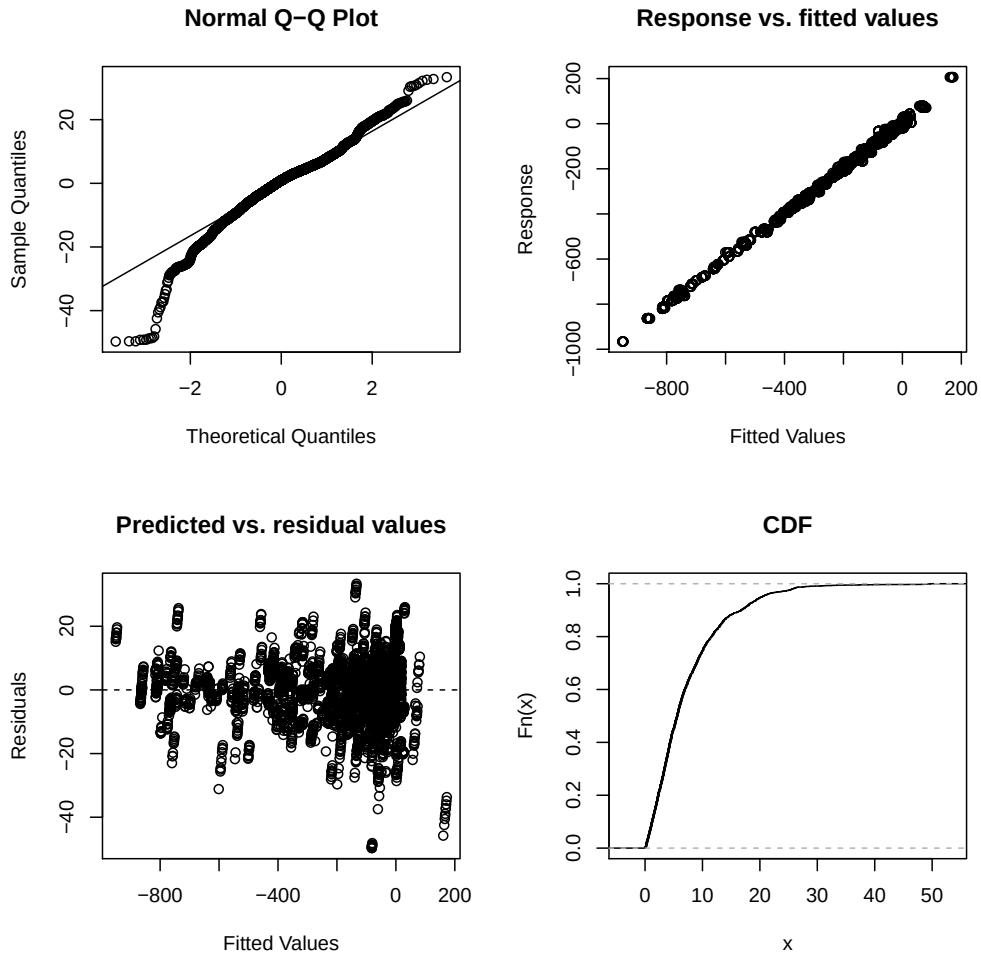


Figure: LASSO. Groups smaller than 7. 6 iterations



Results

	SEP	IQR	MAD	CDF 0.5	$\|G\|$	Avg. p
Jankowski	2.22	-	-	-	73	73
lasso-6iter	9.91	8.22	7.74	5.54	87	36.28
lasso-underfit	24.46	17.14	16.88	11.53	38	12.38
lasso-overfit	18.52	13.56	12.50	9.46	2365	36.02
pls-4iter	11.48	8.33	8.26	6.00	78	78

Groups - Sample

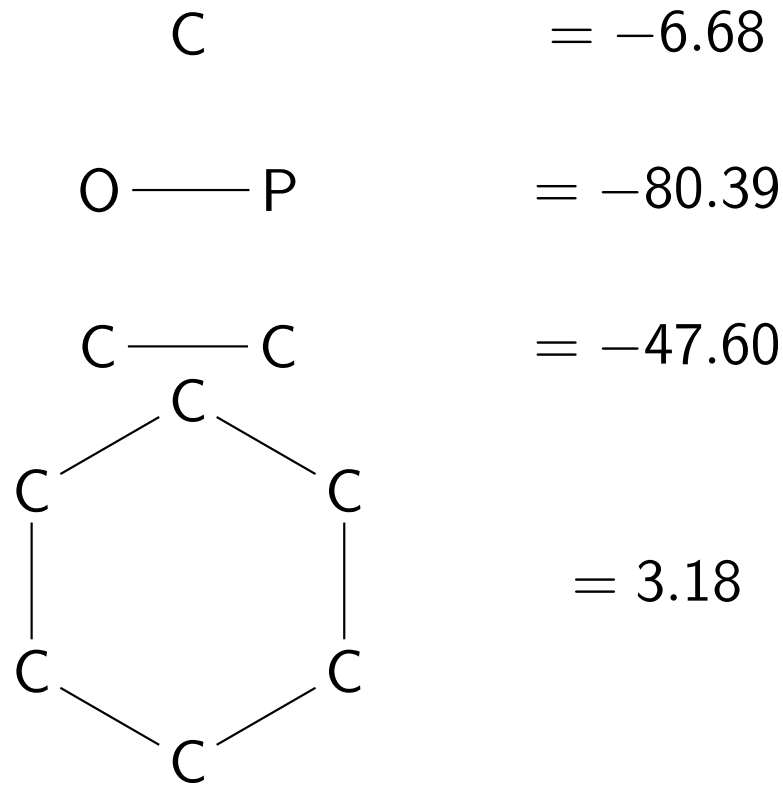


Figure: Sample of groups learned from lasso-6iter

Discussion

- It's pretty generic!
- Single group identification.
- Variance based group exclusion.
- Stopping criterion.
- How to measure uncertainty in data.
- Better outlier detection.
- Use reactions as test data.
- Non-linear approaches.