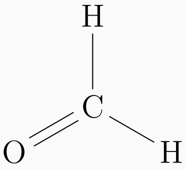
Generic Group Contribution Method

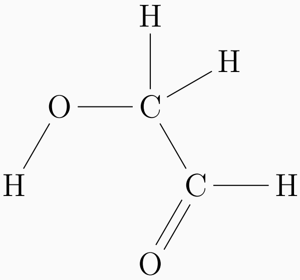
Authors: Christoph Flamm[5,8], Marc Hellmuth[1,7], Daniel Merkle[2], Nikolai Nøjgaard[1,2], Peter F. Stadler[3,4,5,7]

February 11, 2019

[1] Dpt. of Mathematics and Computer Science, University of Greifswald
[2] Department of Mathematics and Computer Science, University of Southern Denmark, Denmark
[3] Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center of Bioinformatics, University of Leipzig
[4] Max-Planck-Institute for Mathematics in the Sciences
[5] Inst. f. Theoretical Chemistry, University of Vienna
[6] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe
[7] Saarland University, Center for Bioinformatics
[8] Center for Anatomy and Cell Biology, Medical University of Vienna

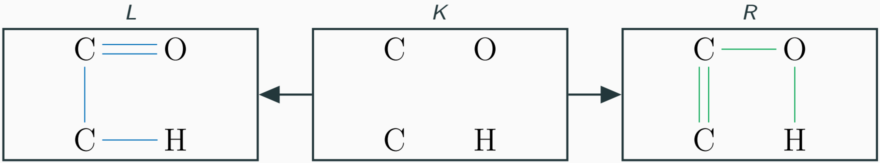


Formaldehyde

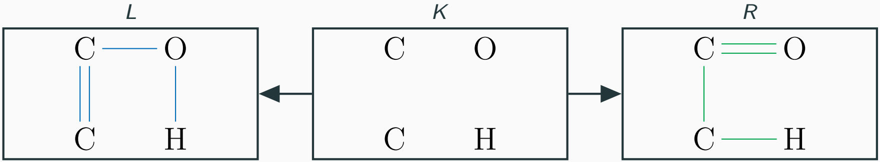


Glycolaldehyde

The Beginning: A look at MØD

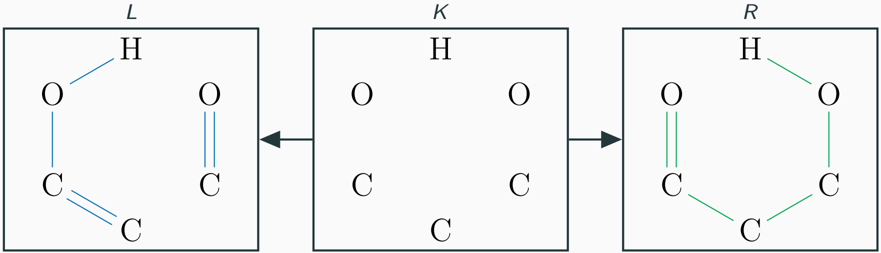


Keto-enol isomerization

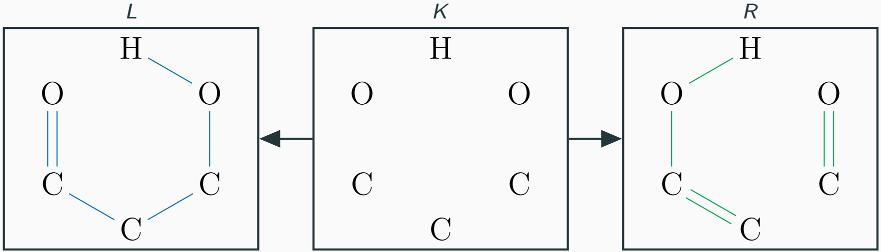


Keto-enol isomerization, inverse

The Beginning: A look at MØD

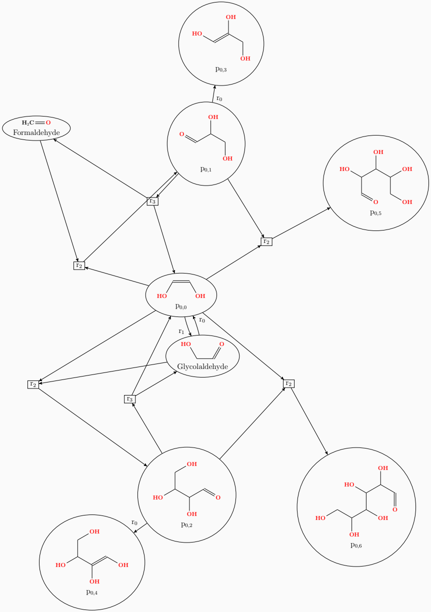


Aldol addition

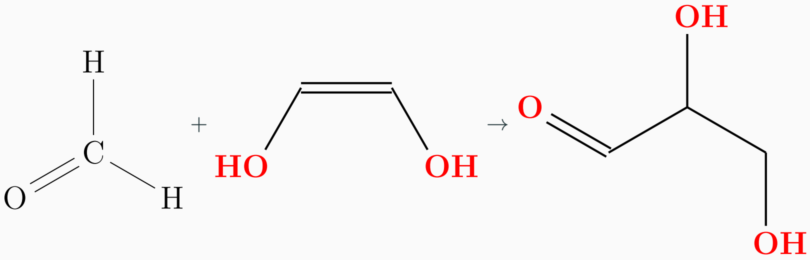


Aldol addition, inverse

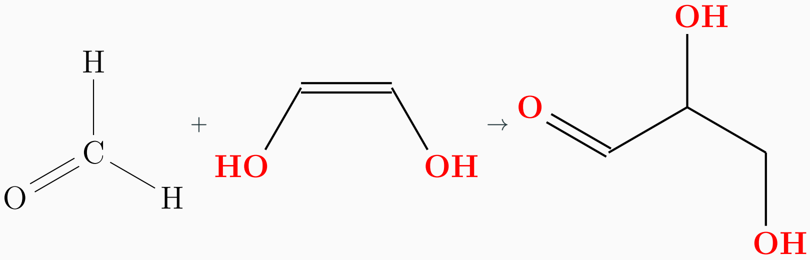
The Beginning: A look at MØD



Reactions: Fact or Fiction?

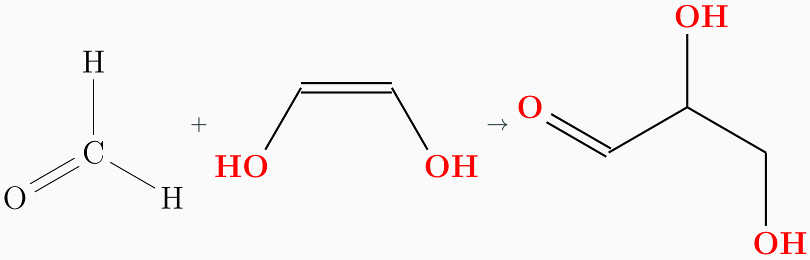


Reactions: Fact or Fiction?



Real or theoretical reaction? Let's look at its energy change!

Reactions: Fact or Fiction?



Real or theoretical reaction? Let's look at its energy change!

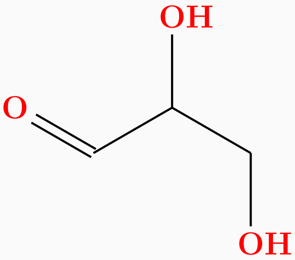
Gibbs Free Energy: $G = H - TS$, where H is enthalpy, T temperature and S entropy.

Gibbs Free Energy Change: $\Delta G = G_{products} - G_{educts}$

So How Do We Compute the Energy

- The Gibbs Free Energy of a molecule can measured in the lab.
- But our chemical universe can (in theory) be infinite.
- Hence, we want to create a predictive model on a sampled population.

State of the Art: Group Contribution Method

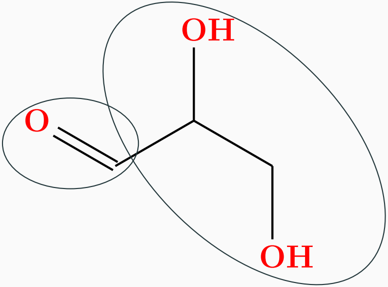


- We can decompose a molecule into functional groups that linearly relates to G.

Problems with the Group Contribution Method in a Generic Framework:

- What are the functional groups?
- How to tile a graph?
- Introducing new functional group changes the entire input.

State of the Art: Group Contribution Method

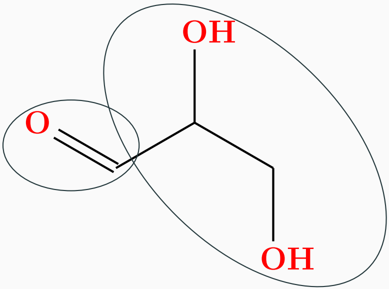


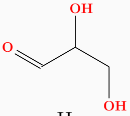
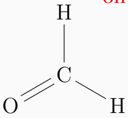
- We can decompose a molecule into functional groups that linearly relates to G.

Problems with the Group Contribution Method in a Generic Framework:

- What are the functional groups?
- How to tile a graph?
- Introducing new functional group changes the entire input.

State of the Art: Group Contribution Method



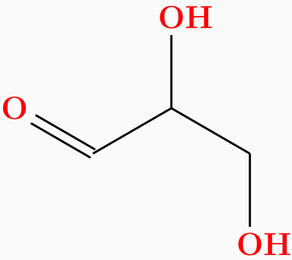
	OC	O2C2H2
	1	1
	1	0
⋮	⋮	⋮

- We can decompose a molecule into functional groups that linearly relates to G.

Problems with the Group Contribution Method in a Generic Framework:

- What are the functional groups?
- How to tile a graph?
- Introducing new functional group changes the entire input.

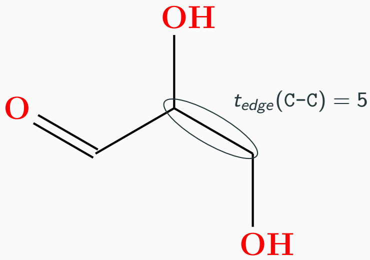
A Closer Look at Molecular Energies



- The energy of a molecule can be approximated as the sum of its bond energies.

$$t_{obs}(G) = \sum_{e \in E(G)} t_{edge}(e)$$

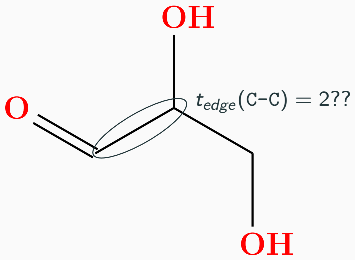
A Closer Look at Molecular Energies



- The energy of a molecule can be approximated as the sum of its bond energies.

$$t_{obs}(G) = \sum_{e \in E(G)} t_{edge}(e)$$

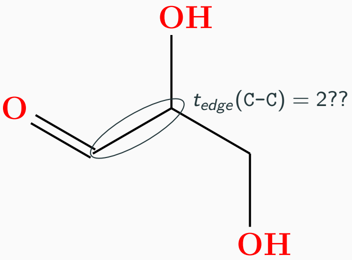
A Closer Look at Molecular Energies



- The energy of a molecule can be approximated as the sum of its bond energies.

$$t_{obs}(G) = \sum_{e \in E(G)} t_{edge}(e)$$

A Closer Look at Molecular Energies



- The energy of a molecule can be approximated as the sum of its bond energies.

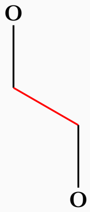
$$t_{obs}(G) = \sum_{e \in E(G)} t_{edge}(e)$$

- The bond energy is determined by its surrounding context.

Defining Contexts

Definition (Context)

A context is a pair $C = (G, e)$, where G is a graph and e is an edge in G. The size of C is defined as the number of edges in G and we call e the origin edge.



Defining Contexts

Definition (Context)

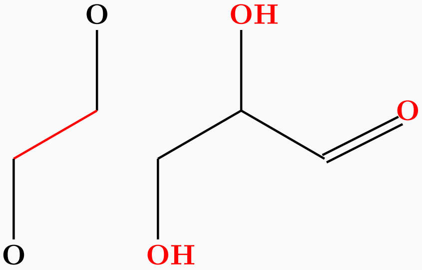
A context is a pair $C = (G, e)$, where G is a graph and e is an edge in G. The size of C is defined as the number of edges in G and we call e the origin edge.

Definition (Frequency)

Given a graph G and a context $C = (H, e')$ we say that C is *a context around* $e \in E(G)$, if there is a subgraph isomorphism φ from H to G that satisfy $\varphi(e') = e$. The frequency $f(C, G, e)$ of C around some edge $e \in E(G)$ is the number of subgraph isomorphisms $\varphi_1, \varphi_2, \ldots$ from C to G that satisfy $\varphi_i(e') = e$.

The *frequency* of C in G is defined as:

$$f(C, G) = \sum_{e \in E(G)} f(C, G, e)$$



Defining Contexts

Definition (Context)

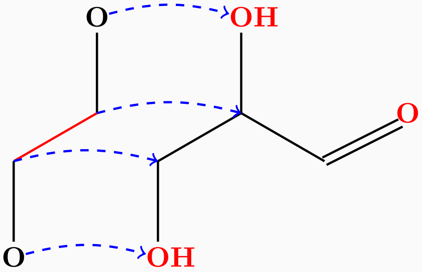
A context is a pair $C = (G, e)$, where G is a graph and e is an edge in G. The size of C is defined as the number of edges in G and we call e the origin edge.

Definition (Frequency)

Given a graph G and a context $C = (H, e')$ we say that C is *a context around* $e \in E(G)$, if there is a subgraph isomorphism φ from H to G that satisfy $\varphi(e') = e$. The frequency $f(C, G, e)$ of C around some edge $e \in E(G)$ is the number of subgraph isomorphisms $\varphi_1, \varphi_2, \ldots$ from C to G that satisfy $\varphi_i(e') = e$.

The *frequency* of C in G is defined as:

$$f(C, G) = \sum_{e \in E(G)} f(C, G, e)$$



Defining Contexts

Definition (Context)

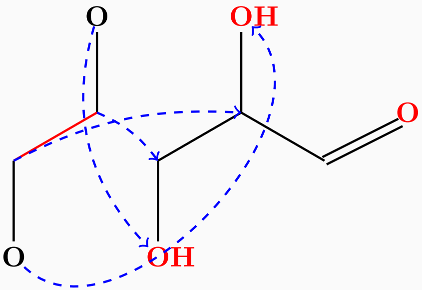
A context is a pair $C = (G, e)$, where G is a graph and e is an edge in G. The size of C is defined as the number of edges in G and we call e the origin edge.

Definition (Frequency)

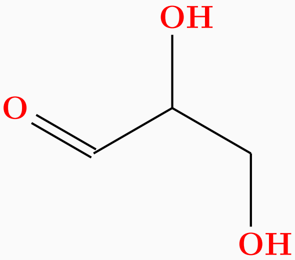
Given a graph G and a context $C = (H, e')$ we say that C is *a context around* $e \in E(G)$, if there is a subgraph isomorphism φ from H to G that satisfy $\varphi(e') = e$. The frequency $f(C, G, e)$ of C around some edge $e \in E(G)$ is the number of subgraph isomorphisms $\varphi_1, \varphi_2, \ldots$ from C to G that satisfy $\varphi_i(e') = e$.

The *frequency* of C in G is defined as:

$$f(C, G) = \sum_{e \in E(G)} f(C, G, e)$$

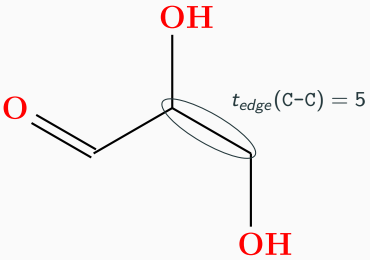


Significant Contexts



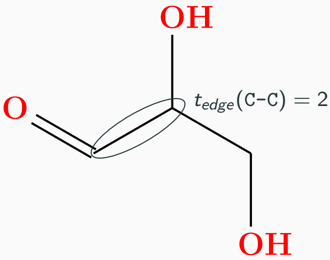
$\mathcal{K}_1 = \{\texttt{O=C}, \texttt{C-C}, \texttt{O-C}, \texttt{O-H}\}$

Significant Contexts



$\mathcal{K}_1 = \{\texttt{O=C}, \texttt{C-C}, \texttt{O-C}, \texttt{O-H}\}$

Significant Contexts

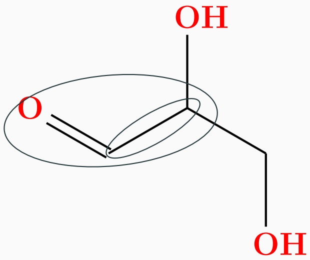


$\mathcal{K}_1 = \{\texttt{O=C}, \texttt{C-C}, \texttt{O-C}, \texttt{O-H}\}$

$t_{\mathcal{C}}(\texttt{C-C}) = \text{avg. energy} = 3.5$

$t_{edge}(e) \approx f(C, G, e) \cdot t_{\mathcal{C}}(\texttt{C-C}) = 3.5$

Significant Contexts

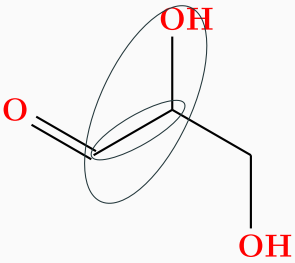


$\mathcal{K}_1 = \{\texttt{O=C}, \texttt{C-C}, \texttt{O-C}, \texttt{O-H}\}$

$t_{\mathcal{C}}(\texttt{C-C}) = \text{avg. energy} = 3.5$

$t_{edge}(e) \approx f(\texttt{C-C}, G, e) \cdot t_{\mathcal{C}}(\texttt{C-C}) + f(G, \texttt{O=C-C}) \cdot t_{\mathcal{C}}(\texttt{O=C-C})$

Significant Contexts



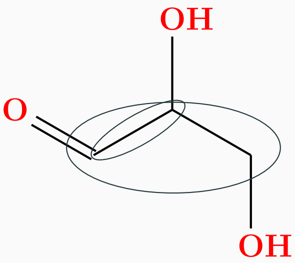
$\mathcal{K}_1 = \{\texttt{O=C}, \texttt{C-C}, \texttt{O-C}, \texttt{O-H}\}$

$t_{\mathcal{C}}(\texttt{C-C}) = \text{avg. energy} = 3.5$

$t_{edge}(e) \approx f(\texttt{C-C}, G, e) \cdot t_{\mathcal{C}}(\texttt{C-C}) + f(G, \texttt{O=C-C}) \cdot t_{\mathcal{C}}(\texttt{O=C-C}) + f(G, \texttt{O-C-C}) \cdot t_{\mathcal{C}}(\texttt{O-C-C})$

$\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \cdots \subseteq \mathcal{K}_k$

Significant Contexts



$\mathcal{K}_1 = \{\texttt{O=C}, \texttt{C-C}, \texttt{O-C}, \texttt{O-H}\}$

$t_{\mathcal{C}}(\texttt{C-C}) = \text{avg. energy} = 3.5$

$t_{edge}(e) \approx f(\texttt{C-C}, G, e) \cdot t_{\mathcal{C}}(\texttt{C-C}) + f(G, \texttt{O=C-C}) \cdot t_{\mathcal{C}}(\texttt{O=C-C}) + f(G, \texttt{O-C-C}) \cdot t_{\mathcal{C}}(\texttt{O-C-C})$

$\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \cdots \subseteq \mathcal{K}_k$

$$t_{edge}(e) \approx \sum_{C \in \mathcal{K}_i} f(C, G, e) \cdot t_{\mathcal{C}}(C)$$

Frequency Matrix and Learning the Significant Contexts

$$t_{edge}(e) \approx \sum_{C \in \mathcal{K}_i} f(C, G, e) \cdot t_{\mathcal{C}}(C)$$

Frequency Matrix and Learning the Significant Contexts

$$t_{edge}(e) \approx \sum_{C \in \mathcal{K}_i} f(C, G, e) \cdot t_{\mathcal{C}}(C)$$

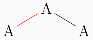
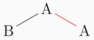
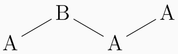
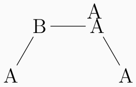
$$\Rightarrow t_{obs}(S) - \sum_{C \in \mathcal{K}_{k-1}} f(C, S) \cdot t_{\mathcal{C}}(C) = \sum_{C \in \mathcal{C}_k^{\mathcal{S}}} f(C, S) \cdot t_{\mathcal{C}}(C) + \epsilon$$

Frequency Matrix and Learning the Significant Contexts

$$t_{edge}(e) \approx \sum_{C \in \mathcal{K}_i} f(C, G, e) \cdot t_{\mathcal{C}}(C)$$

$$\Rightarrow t_{obs}(S) - \sum_{C \in \mathcal{K}_{k-1}} f(C, S) \cdot t_{\mathcal{C}}(C) = \sum_{C \in \mathcal{C}_k^{\mathcal{S}}} f(C, S) \cdot t_{\mathcal{C}}(C) + \epsilon$$

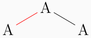
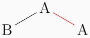
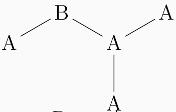
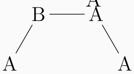
$X =$

			$\dots$
	2	2	$\dots$
	0	1	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$

$$t_{edge}(e) \approx \sum_{C \in \mathcal{K}_i} f(C, G, e) \cdot t_{\mathcal{C}}(C)$$

$$\Rightarrow t_{obs}(S) - \sum_{C \in \mathcal{K}_{k-1}} f(C, S) \cdot t_{\mathcal{C}}(C) = \sum_{C \in \mathcal{C}_k^{\mathcal{S}}} f(C, S) \cdot t_{\mathcal{C}}(C) + \epsilon$$

$X =$

			…
	2	2	…
	0	1	…
⋮	⋮	⋮	⋱

$$\text{LASSO: } \min \left(\sum_{i=1}^{|\mathcal{C}_1^{\mathcal{S}}|+|\mathcal{S}|} \left(\mathrm{y}_i - \sum_{j=1}^{|\mathcal{C}_k^{\mathcal{S}}|} \mathrm{X}_{ij}\mathrm{t}_j \right)^2 + \lambda \sum_{j=1}^{|\mathcal{C}_k^{\mathcal{S}}|} |\mathrm{t}_j| \right).$$

Some considerations about contexts

- Structural information that determines bond energies must be stored in their frequencies in $\mathcal{S}$.

Some considerations about contexts

- Structural information that determines bond energies must be stored in their frequencies in $\mathcal{S}$.
- The number of non-isomorphic subgraphs in a graph can grow exponentially with its size.

Some considerations about contexts

- Structural information that determines bond energies must be stored in their frequencies in $\mathcal{S}$.
- The number of non-isomorphic subgraphs in a graph can grow exponentially with its size.
- Contexts that only occur in "few" samples are unreliable.

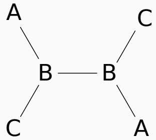
Some considerations about contexts

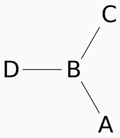
- Structural information that determines bond energies must be stored in their frequencies in $\mathcal{S}$.
- The number of non-isomorphic subgraphs in a graph can grow exponentially with its size.
- Contexts that only occur in "few" samples are unreliable.
- The frequencies of two contexts $C_1 = (G_1, e_1)$ and $C_2 = (G_2, e_2)$ where $G_1 \simeq G_2$ are collinear in $\mathcal{S}$.

Context Mining

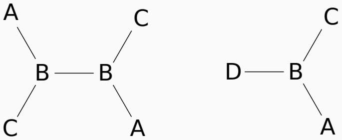
Definition

The support $\mathrm{sup}(C)$ of a context $C \in \mathcal{C}^{\mathcal{S}}$ is the number of graphs in $\mathcal{S}$ which C can be embedded into. Given a positive integer τ we say that C is supported if $\mathrm{sup}(C) \geq \tau$.



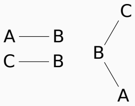


Context Mining



Definition

The support $\mathtt{sup}(C)$ of a context $C \in \mathcal{C}^{\mathcal{S}}$ is the number of graphs in $\mathcal{S}$ which C can be embedded into. Given a positive integer τ we say that C is supported if $\mathtt{sup}(C) \geq \tau$.



Definition

Let $\mathcal{G}$ be a set of graphs and k and τ two integers such that $k > 0$ and $\tau > 0$. Then $\mathtt{FSM}(\mathcal{G}, k, \tau)$ is the set of all subgraphs in $\mathcal{G}$ that contains k edges and are subgraph isomorphic to at least τ graphs in $\mathcal{G}$.

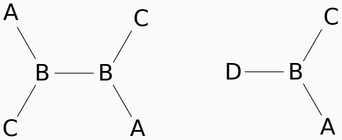
Context Mining

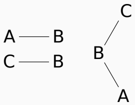
Definition

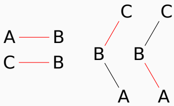
The support $\mathtt{sup}(C)$ of a context $C \in \mathcal{C}^{\mathcal{S}}$ is the number of graphs in $\mathcal{S}$ which C can be embedded into. Given a positive integer τ we say that C is supported if $\mathtt{sup}(C) \geq \tau$.

Definition

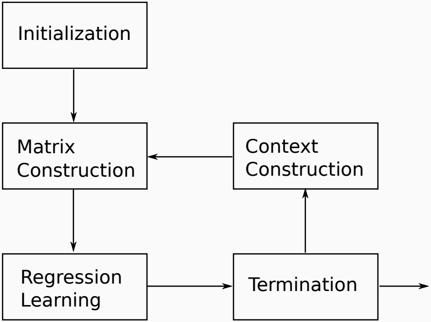
Let $\mathcal{G}$ be a set of graphs and k and τ two integers such that $k > 0$ and $\tau > 0$. Then $\mathtt{FSM}(\mathcal{G}, k, \tau)$ is the set of all subgraphs in $\mathcal{G}$ that contains k edges and are subgraph isomorphic to at least τ graphs in $\mathcal{G}$.



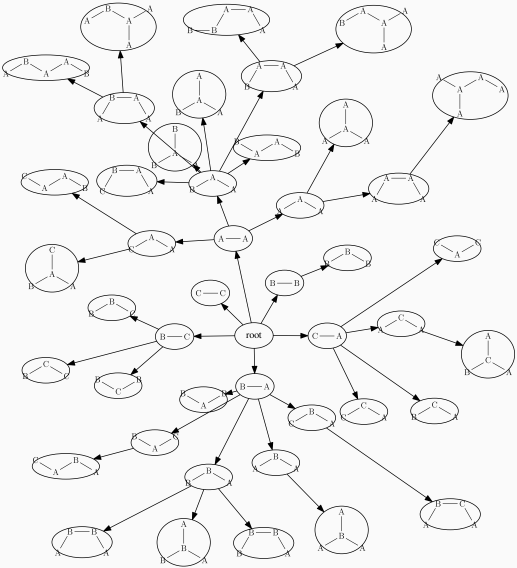




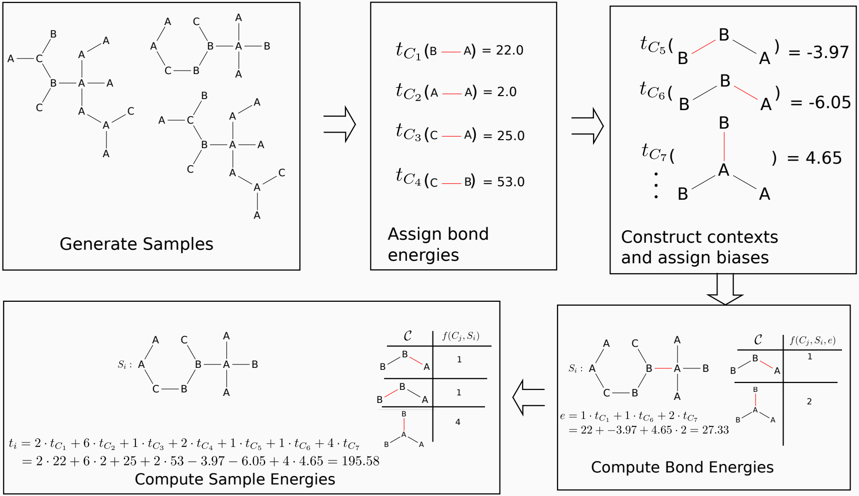
Algorithm


Initialization
Matrix Construction
Context Construction
Regression Learning
Termination

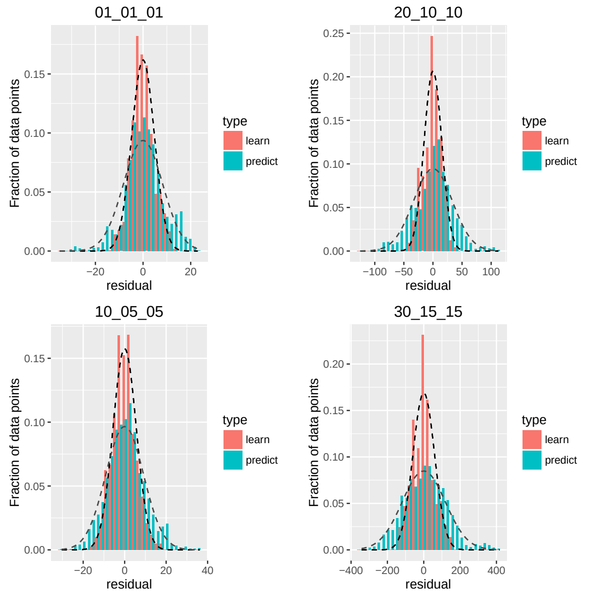
Predicting new graphs



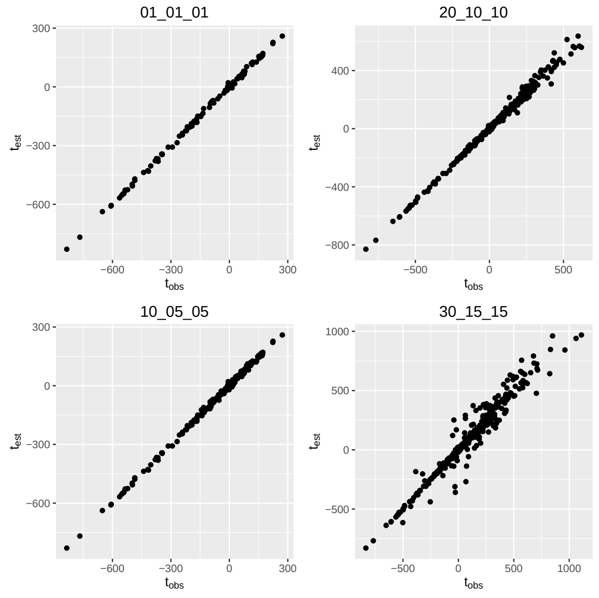
How does it perform: Construction of synthetic dataset.



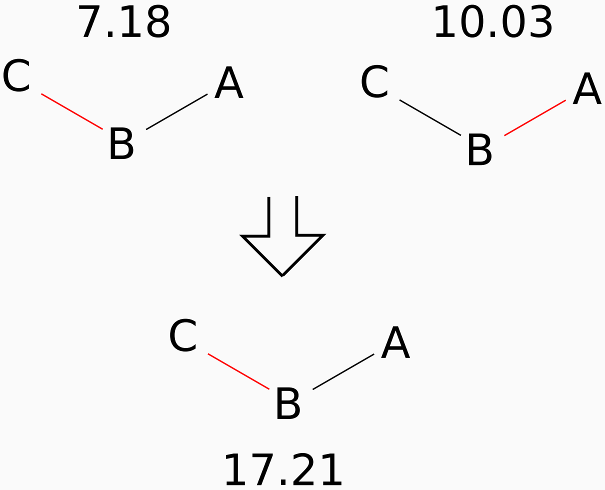
Results: Synthetic dataset



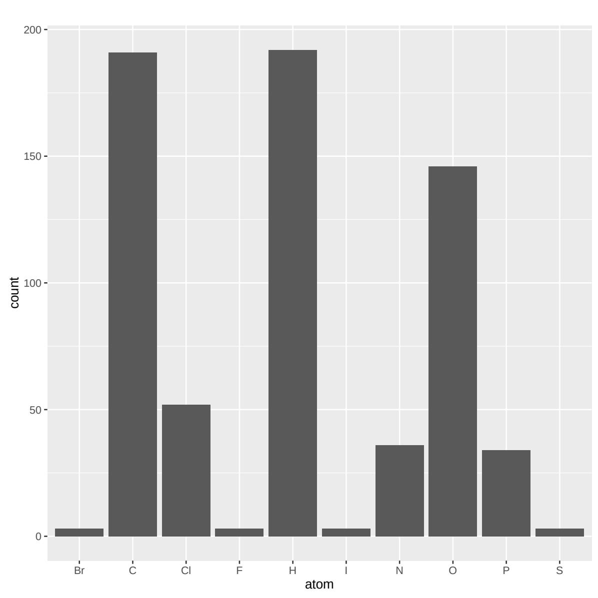
Results: Synthetic dataset



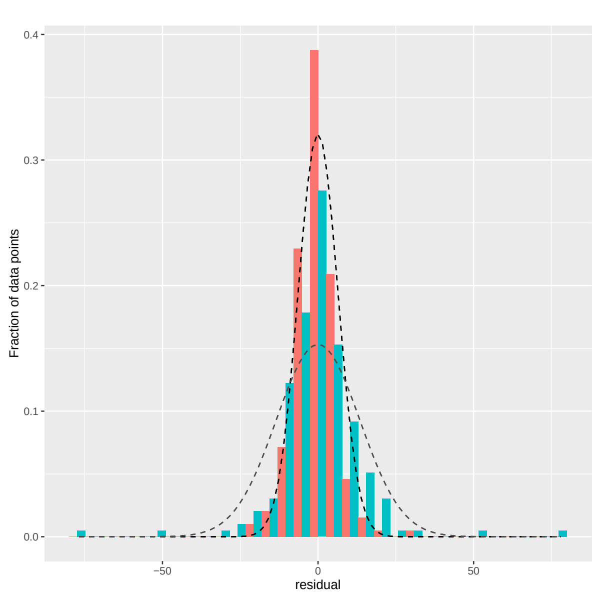
Results: Synthetic dataset



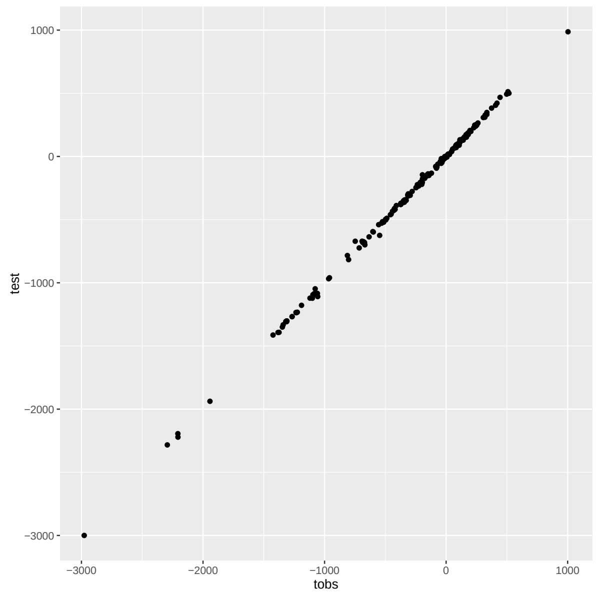
Results: Gibbs Free Energy in metabolic networks



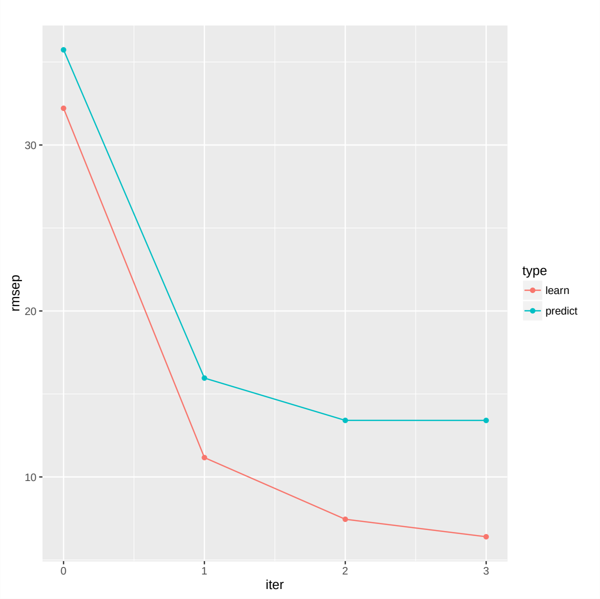
Results: Gibbs Free Energy in metabolic networks



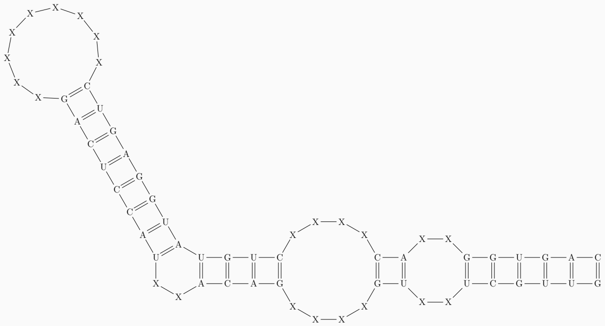
Results: Gibbs Free Energy in metabolic networks



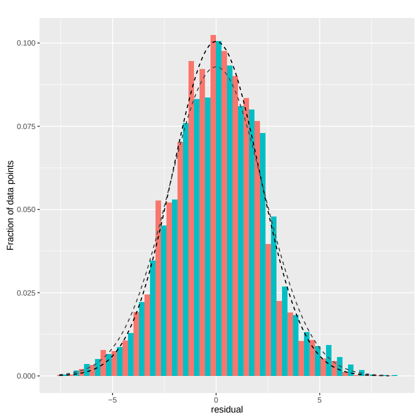
Results: Gibbs Free Energy in metabolic networks

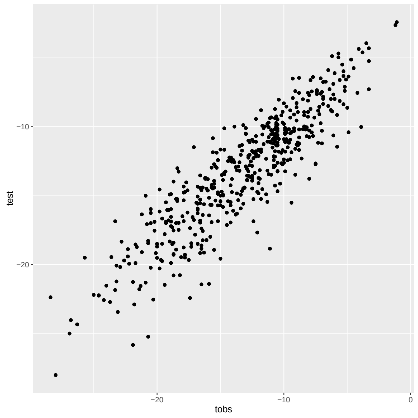


Results: Minimum Free Energy of RNA secondary structures

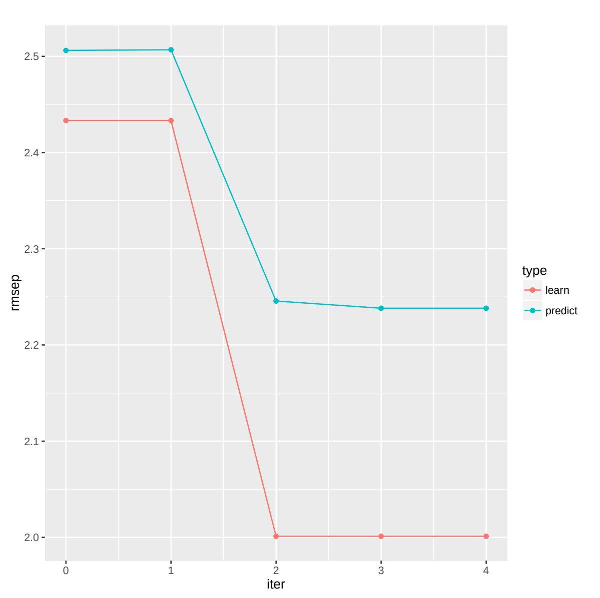


Results: Minimum Free Energy of RNA secondary structures

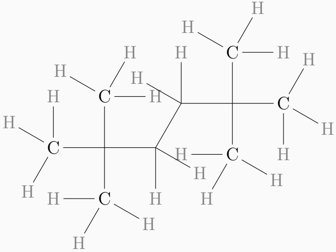


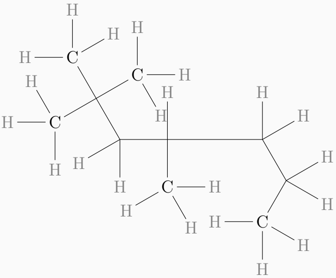


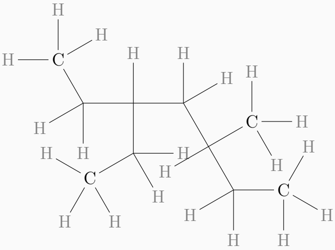
Results: Minimum Free Energy of RNA secondary structures



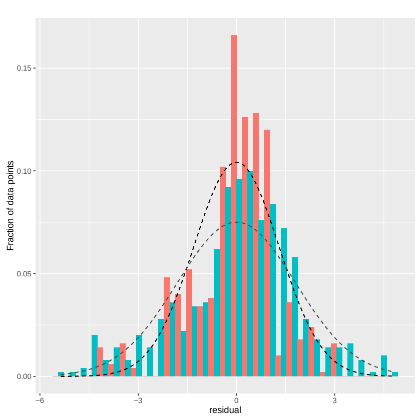
Results: Boiling point acyclic molecules

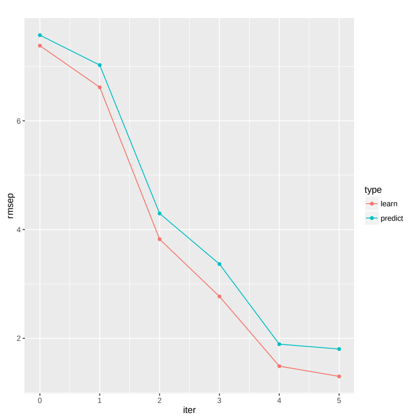




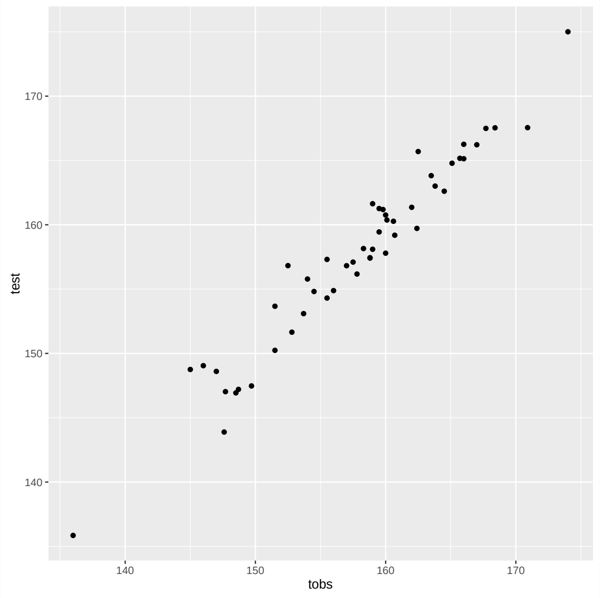


Results: Boiling point acyclic molecules





Results: Boiling point acyclic molecules



Conclusion

- Constructed a generic group contribution method based on the approximation of molecular energies.
- Can be used on a wide range of thermo dynamic properties.