Assessment of optimal conditions for selective deprotection reactions resulted from analysis of large reaction database

Dr. Timur I. Madzhidov

Chemoinformatics and Molecular Modeling Lab

Organic Chemistry Dept

Kazan Federal University

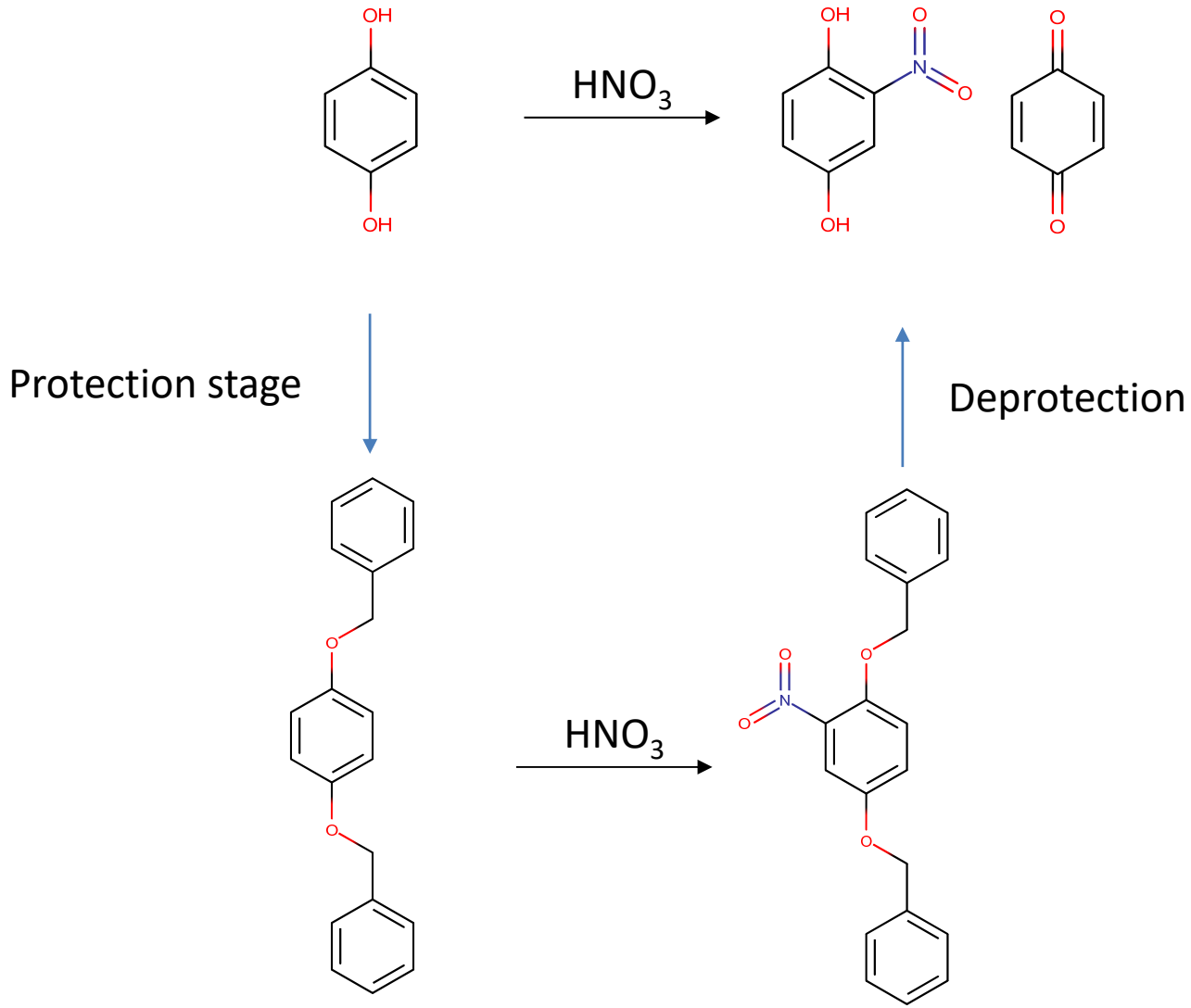
A. LIN (KFU, UniStra), R. NUGMANOV (KFU), O. KlIMCHUK (UniStra), I. ANTIPIN (KFU), A. VARNEK (UniStra)

Kazan Federal UNIVERSITY
Alexander Butlerov INSTITUTE OF CHEMISTRY

Kazan Federal University
Chemoinformatics and molecular modeling lab.

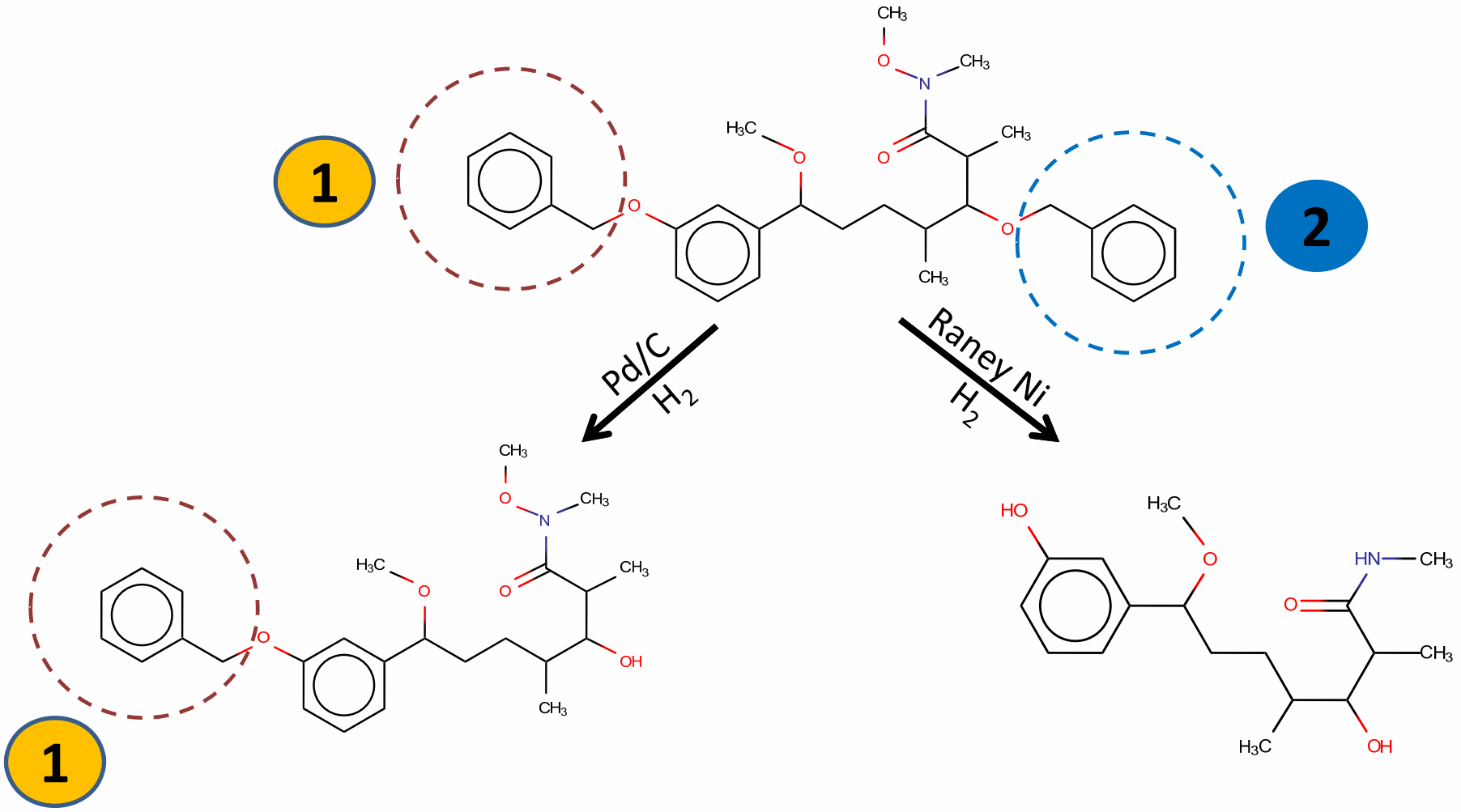
UNIVERSITÉ DE STRASBOURG
ChémoInformatique STRASBOURG

PROTECTIVE GROUP (PG) IN SYNTHETIC CHEMISTRY



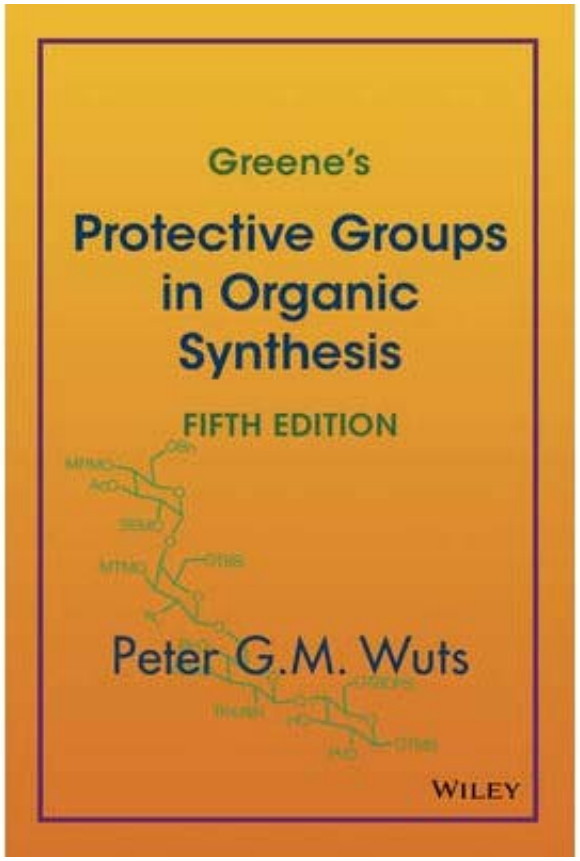
Pain, Ch., L. Pain. *How to choose the path of organic substance synthesis*, 1973

Protective groups (PG)



Llàcer, E., P. Romea and F. Urpí (2006). "Studies on the hydrogenolysis of benzyl ethers." Tetrahedron letters **47**(32): 5815-5818.

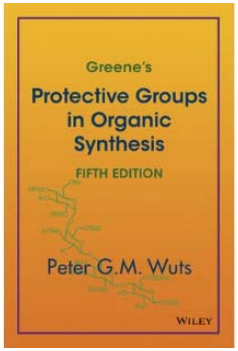
The “Bible” of Protective Groups reactivity analysis



Theodora W. Greene
(1931-2005)

1054 Protective groups (PG)

11249 articles



Greene's Reactivity Charts
(for alcohol protection)

	H_2/Raney (Ni)	H_2/ Pt, pH 2-4	H_2/Pd	H_2/Lindlar	H_2/Rh
PG	Catalytic Reduction				
Me	L	L	L	L	L
MOM	L	M	L	L	L
THP	L	L	L	L	L
t-Butyl	L	L	L	L	L
Bn	H	H	H	L	L
TPM	H	H	H	L	L

Catalyst

Method of deprotection

Observations

H – leaving PG; ***L*** – remaining PG; ***M*** – no firm conclusion

Greene's book Drawbacks

- *Reactivity Charts* result from a manual analysis of relatively small amount of data, and therefore, PG reactivity analysis might be uncertain
- It is not clear according to which quantitative criteria – yield, % of cleaving/remaining groups – PG reactivity labels (*H* and *L*) have been assigned;
- In some cases, no references nor examples proving the reactivity assignments were provided
- The *Reactivity Charts* don't consider a reactivity of a given PG as a function of its chemical environment

Goals

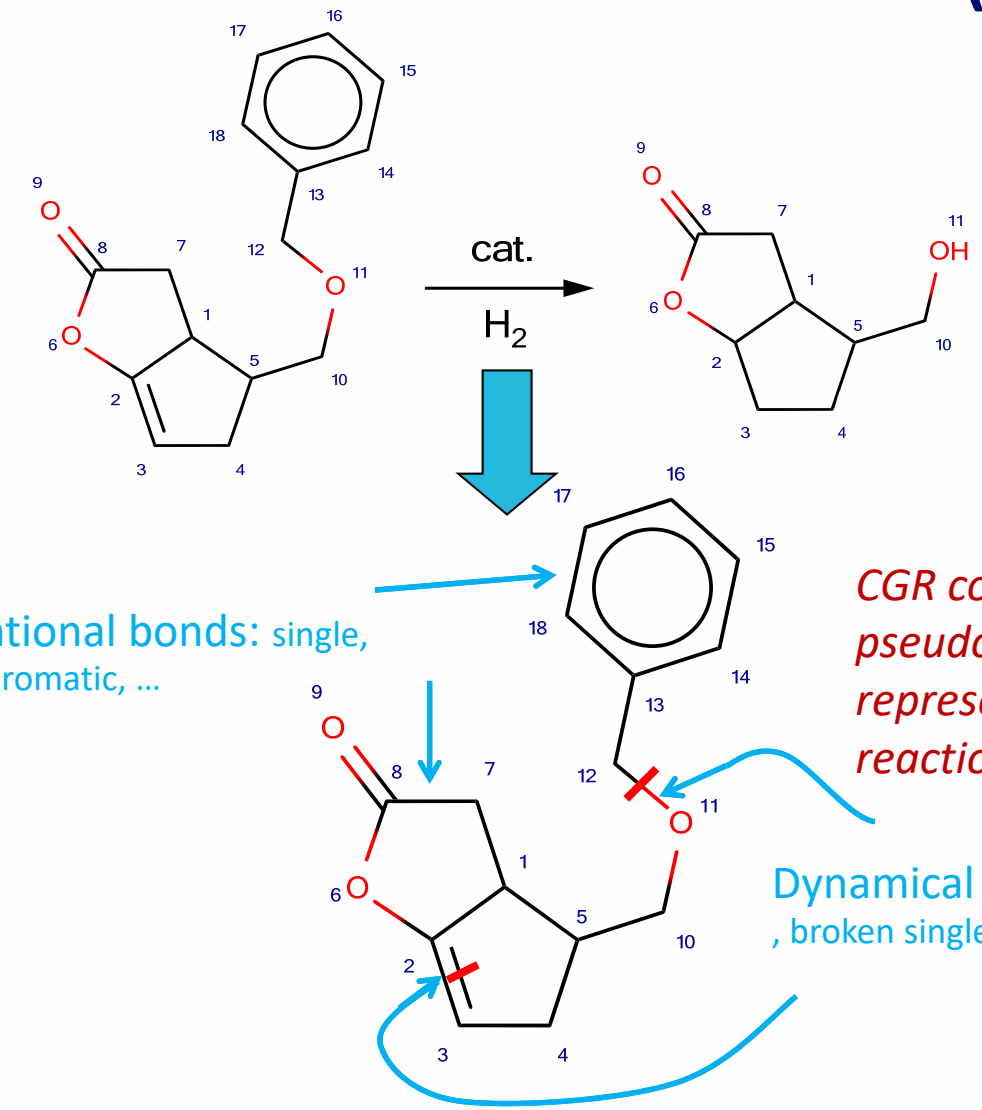
Can the analysis similar to Green's Reactivity Charts' one be made on the basis of ALL available data? Will it be consistent with Green's book one?

- To perform statistical analysis of PG reactivity based on large dataset of catalytic hydrogenation reactions and to compare its results with the Greene's Reactivity Charts

Can we propose something better in the sense of quality of prediction?

- To develop an approach and related software tool able to recommend a reaction conditions leading to selective deprotection of a PG accounting for its chemical environment

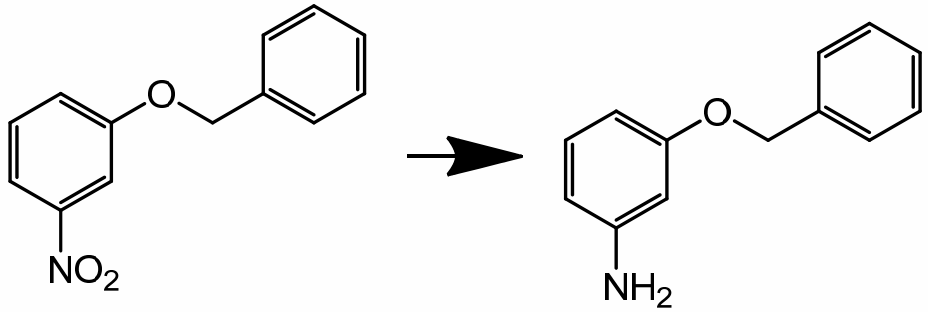
CONDENSED GRAPH OF REACTION (CGR)



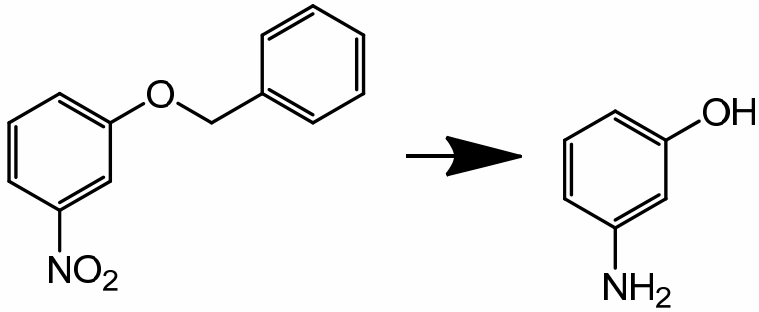
Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. *J Comput Aided Mol Des* **2005**, *19*, 693.

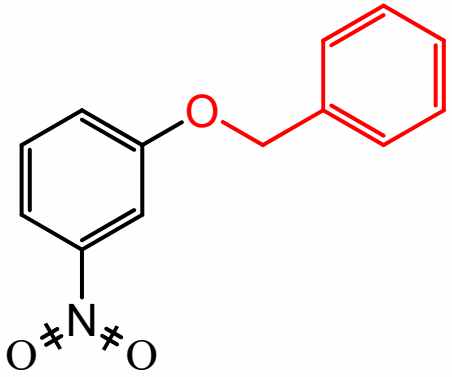
HOW CGR IS USED

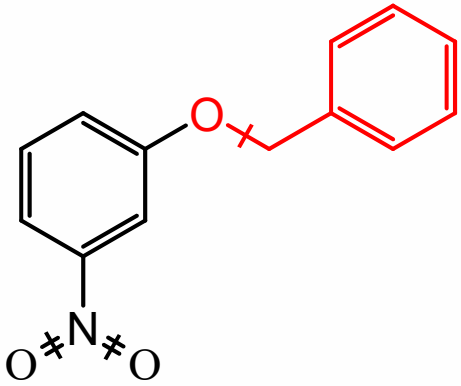
Protective group remains



Protective group is cleaved







Using CGR-based queries in substructure search one can classify reactions into one where protective group remained and cleaved

DATA

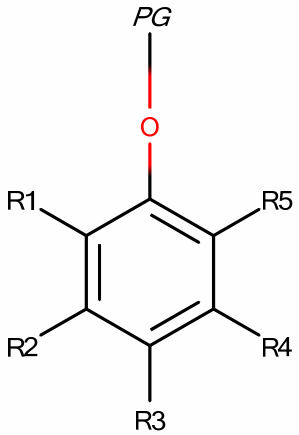
- a set of *catalytic hydrogenation reactions* has been retrieved from the *Reaxys* database (2012) using a query

  1 step, T>-273°C, Yield>0%, hydrogen is in the list of reagents/catalyst

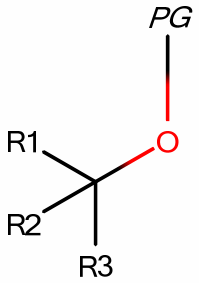
- selected data include 142 111 reactions (271 563 conditions)

- These data are very “noisy”:
  - ✓ most of reactions structures are stoichiometrically non-equilibrated
  - ✓ a lot of important information (yield, catalysts, solvents) is missed
  - ✓ several different names are used for one same catalyst

STUDIED REACTIONS



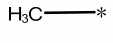
Phenols (aromatic alcohol) protection



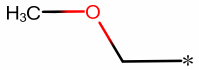
Alcohols (aliphatic alcohol) protection

Amine group protection with formation of carbamates and amides was also considered

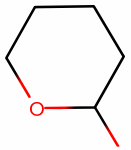
PGs:



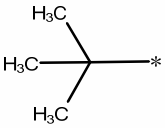
Methyl



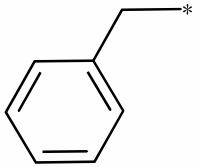
MOM



THP

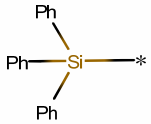


t-Bu

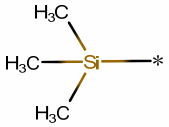


Bn

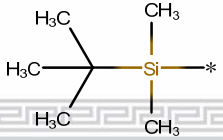
etc…



TPS

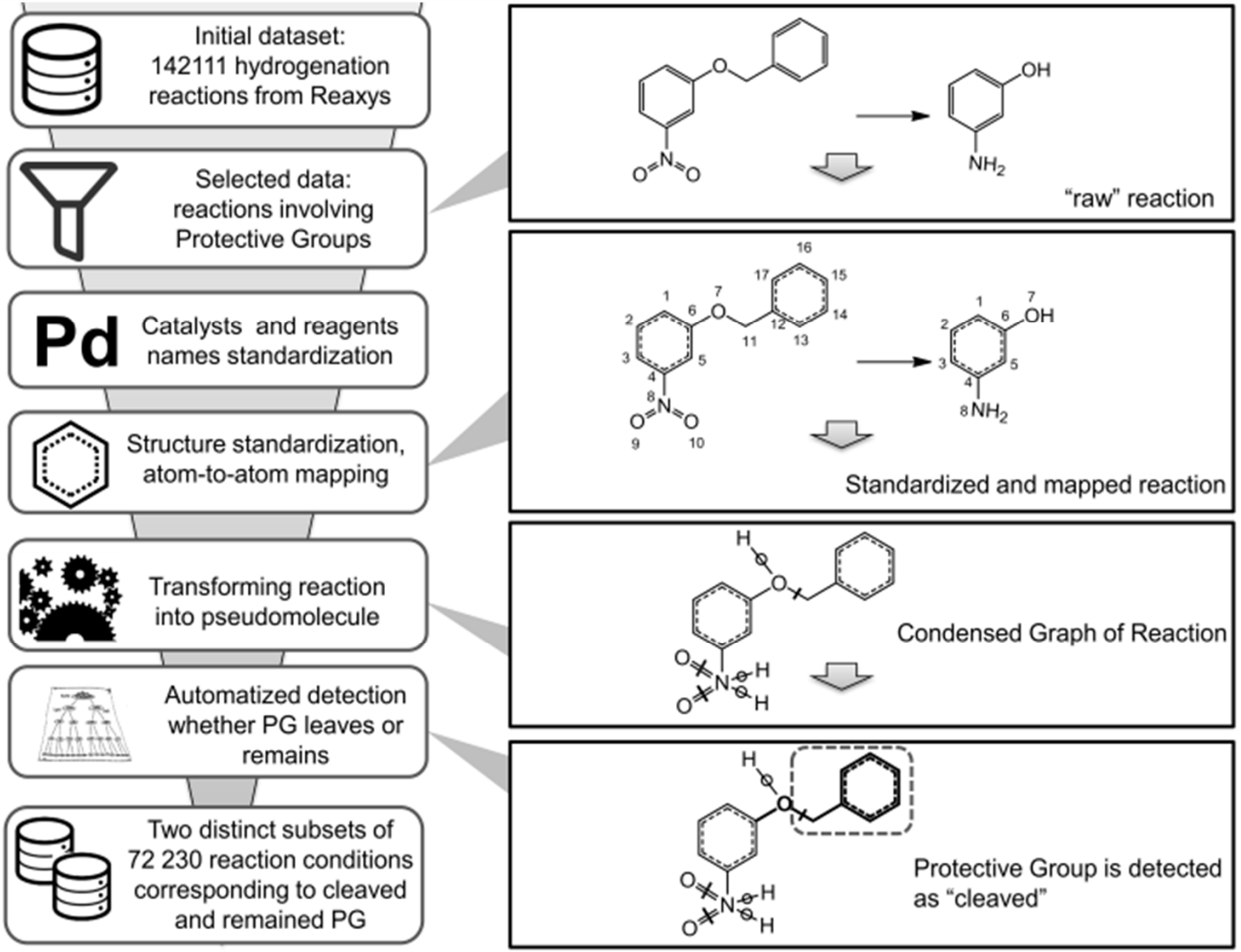


TMS

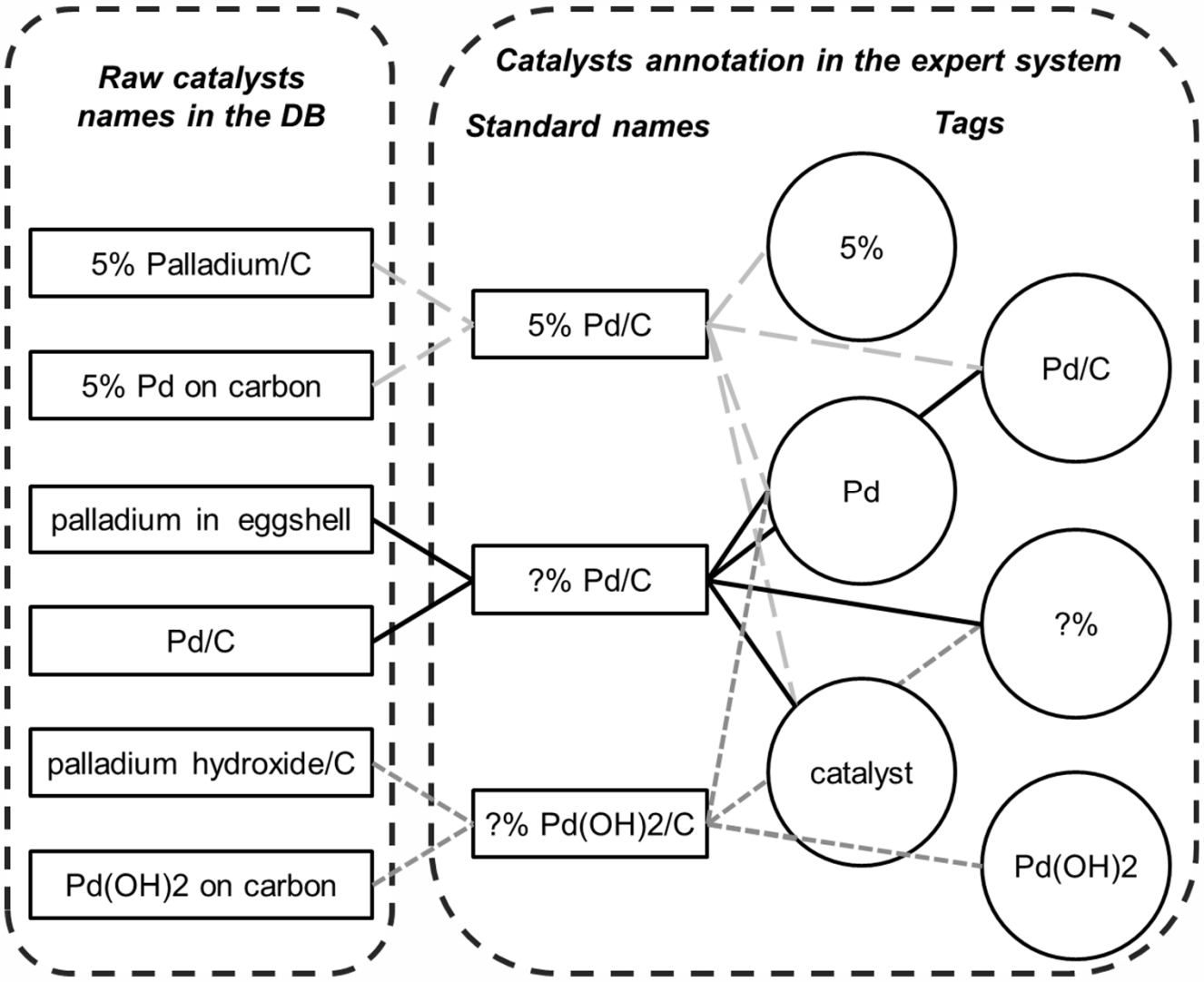


TBDMS

DATA PREPARATION AND ANNOTATION



CATALYST AND REAGENT NAME STANDARDIZATION



PG CLEAVAGE ASSESSMENT FOR A GIVEN CATALYST

- ***C**leavage **P**robability* $CP = \frac{CPG}{(CPG+RPG)} * 100, \%$

If CP≥80% → PG can easily be cleaved (“*H*”)

If CP≤20% → PG is not cleaved (“*L*”)

In other way → no clear conclusion about PG leaving/remaining can be drawn (“*M*”)

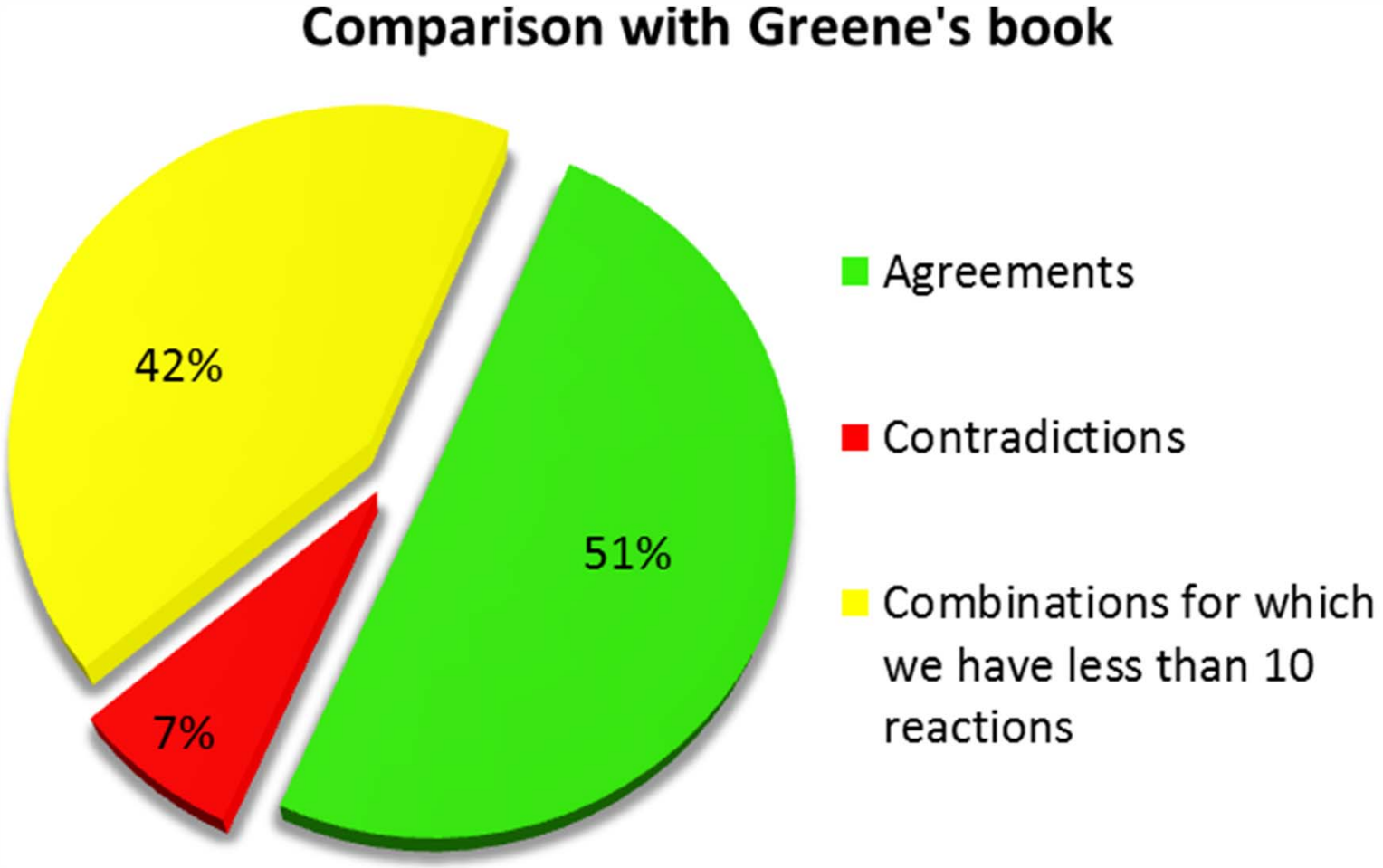
COMPARISON WITH GREENE'S BOOK

(the alcohol protection case)

PG	Raney (Ni)		Pt, pH 2-4		Pd/C		Lindlar		Rh/C or Rh/Al2O3	
	Green's	CP	Green's	CP	Green's	CP	Green's	CP	Green's	CP
Me	L	0	L	0	L	0.1	L	0	L	0
MOM	L	–	M	–	L	1.4	L	0	L	0
MEM	L	–	M	–	L	3.8	L	–	L	–
Cy	L	–	L	–	L	0	L	–	L	–
t-Bu	L	–	L	–	L	0	L	0	L	–
Bn	H	75	H	17	H	98.7	L	37.5	H	27.6
TBDMS	L	–	H	–	L	0.7	L	0	L	0
Ac	L	0[h]	M	0	L	1.0	L	0	L	0
piv	L	–	L	–	L	6.2	L	–	L	–
Bz	L	0	L	–	L	50	L	–	L	–
Ms	R	0	L	–	L	7.7	L	–	L	–

	agrees with the Greene's book
	contradiction with the Greene's book
	statistically insignificant data (≤ 10 reactions in total)
	no data

Agreement with Green's RC



Goals

Can the analysis similar to Green's Reactivity Charts' one be made on the basis of ALL available data? Will it be consistent with Green's book one?

- To perform statistical analysis of PG reactivity based on large dataset of catalytic hydrogenation reactions and to compare its results with the Greene's Reactivity Charts

Can we propose something better in the sense of quality of prediction?

- To develop an approach and related software tool able to recommend a reaction conditions leading to selective deprotection of a PG accounting for its chemical environment

AN EXPERT SYSTEM FOR PROTECTIVE GROUP REACTIVITY

Main concept:

Similar reactions proceed under similar conditions

Implementation:

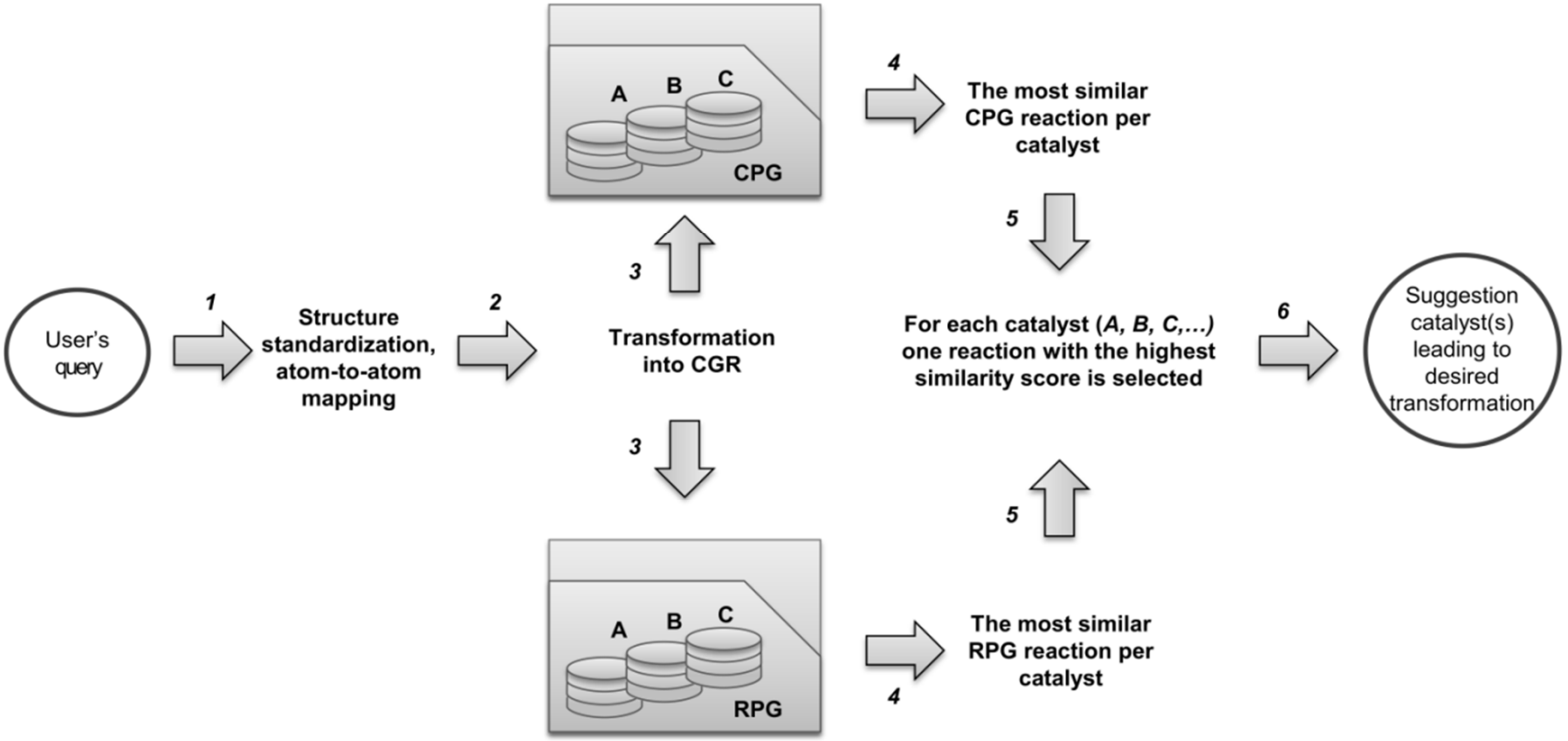
For a given query, the program searches the most similar reactions in a database and retrieves their reaction conditions (catalyst, solvent, temperature, etc.)

Similarity assessment:

is performed for Condensed Graphs of Reactions encoded by bitstrings using Tanimoto coefficient

$$Tc = \frac{c}{a + b - c}$$

AN EXPERT SYSTEM WORKFLOW

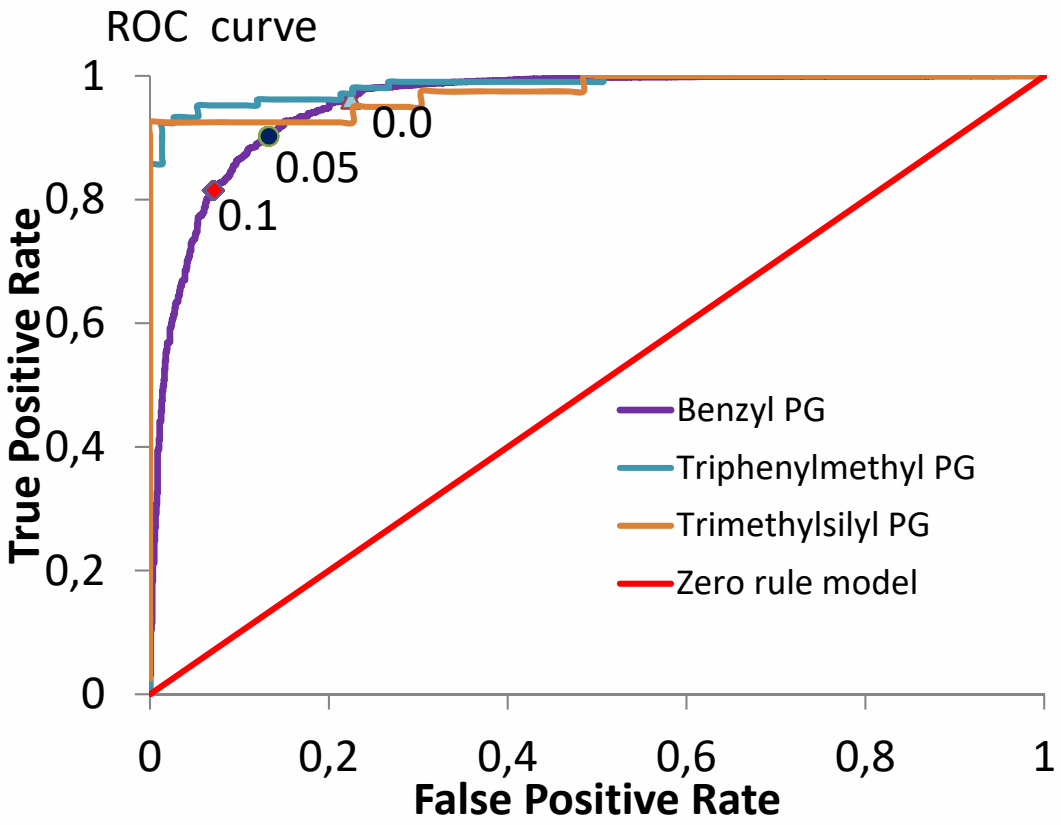


$\Delta Tc = Tc(CPG) - Tc(RPG)$

$\Delta Tc \geq T_0$ ➡ cleaved

$\Delta Tc \leq -T_0$ ➡ saved

PREDICTION PERFORMANCE
(for alcohol protection)



In Leave One Out cross-validation

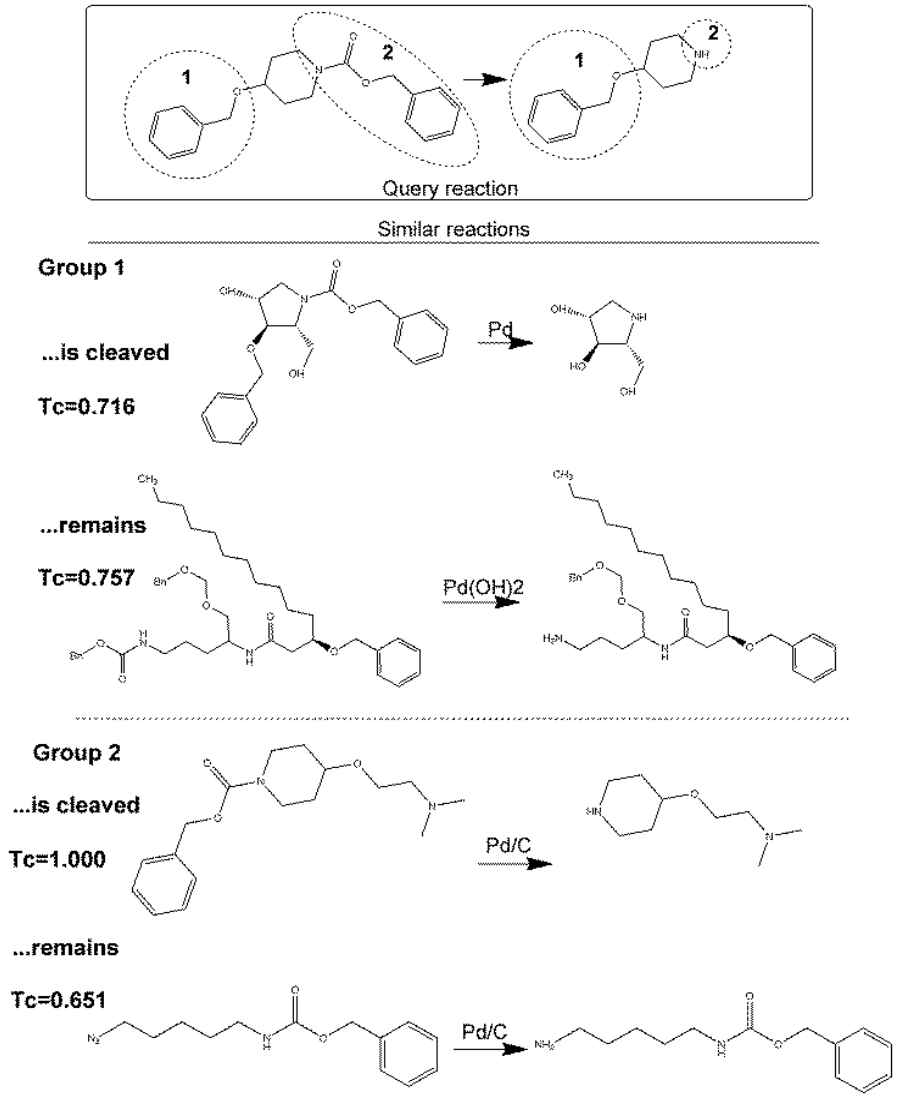
ROC AUC = 0.94 – 0.98

For ΔTc = 0.05:

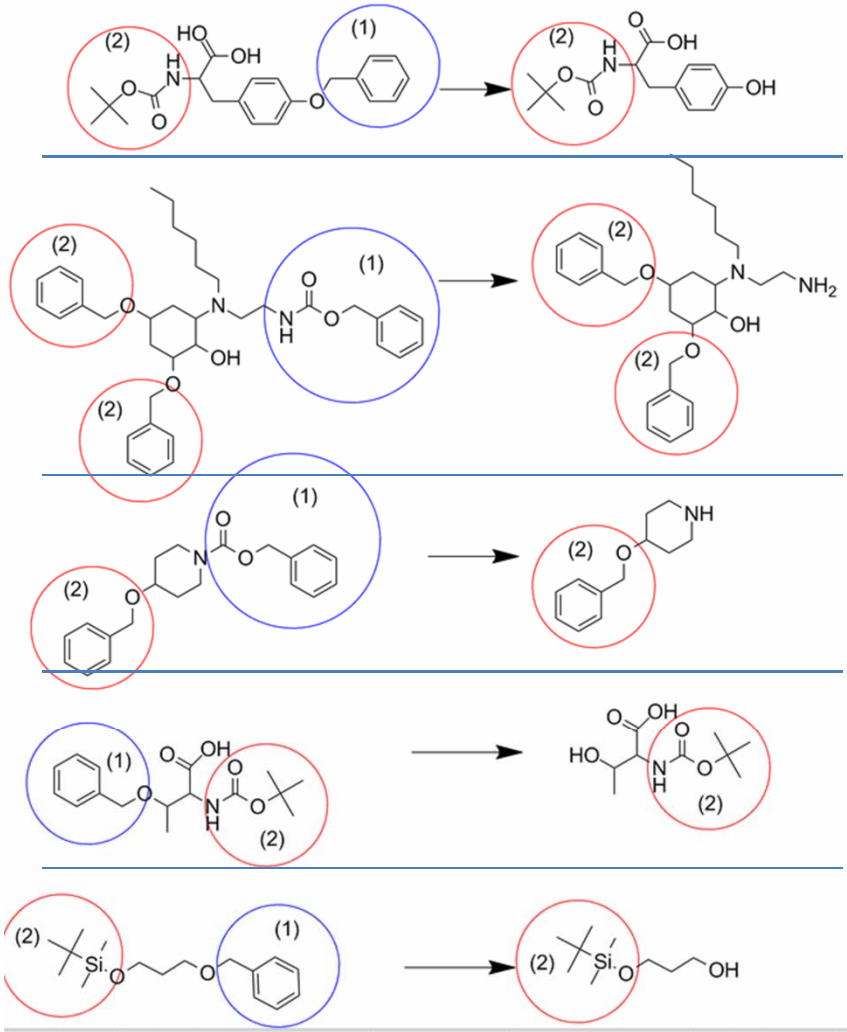
Balanced Accuracy = 0.85 – 0.95

EXTERNAL VALIDATION

- 7 substrates contained **one** Protective Group - **5 correctly predicted**
- 5 substrates contained **two** Protective Groups - **all correctly predicted**

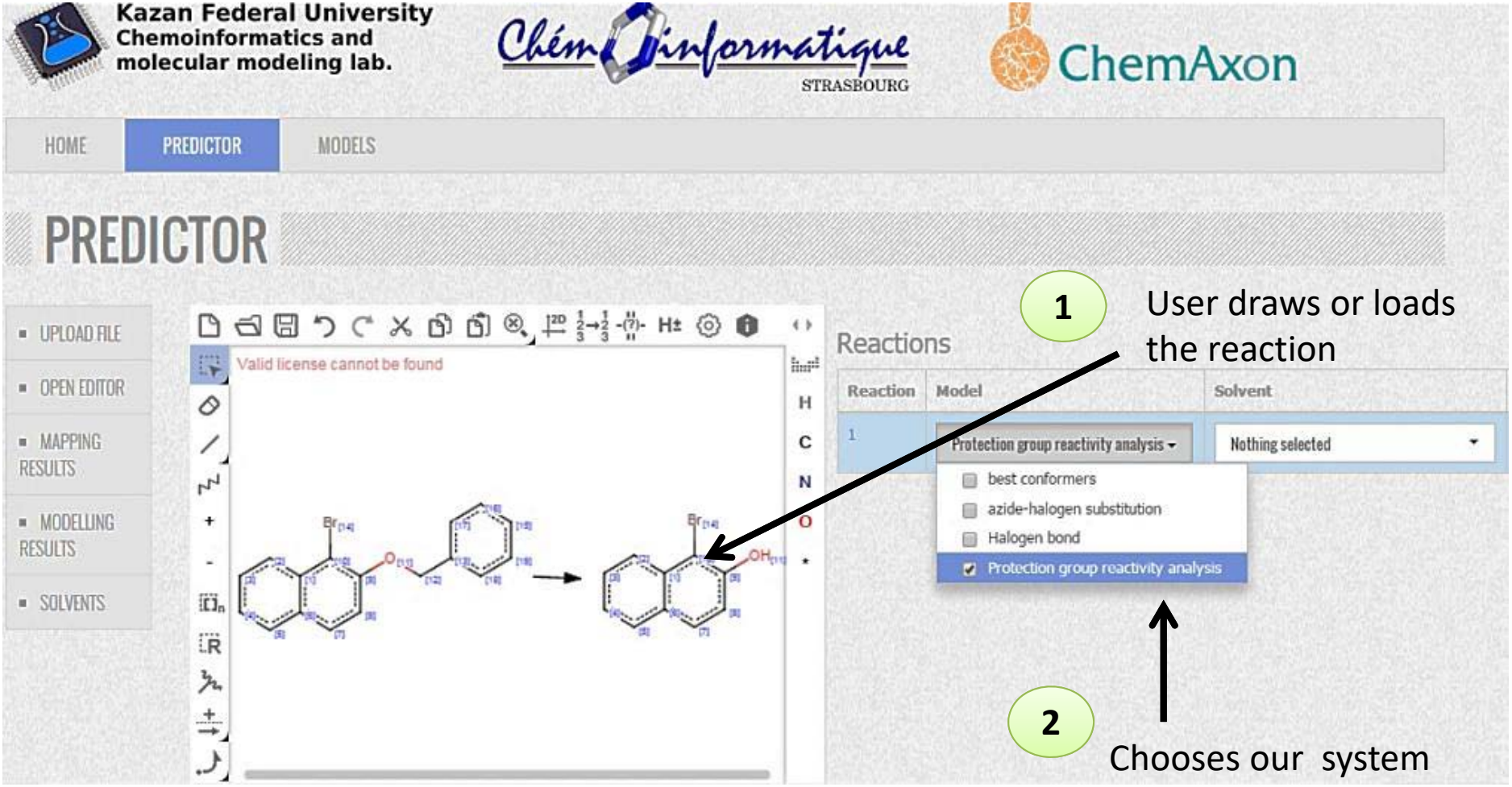


EXTERNAL VALIDATION: SELECTIVITY



Experimental conditions	Group	Greene's Reactivity Charts	Expert system recommendation
Pd/C, Methanol	(1)	to be cleaved (H)	Pd-catalyst [Pd/C]
	(2)	remain (L)	
Pd/C, Methanol	(1)	to be cleaved (H)	Pd-catalyst [Pd/C]
	(2)	to be cleaved (H)	
Pd/C, Methanol	(1)	to be cleaved (H)	Pd-catalyst [Pd/C] Ni-catalyst [Raney Ni]
	(2)	to be cleaved (H)	
Pd/C, Ethanol	(1)	to be cleaved (H)	Pd-catalyst [Pd/C]
	(2)	remain (L)	
Pd/C, Ethyl acetate	(1)	to be cleaved (H)	Pd-catalyst [Pd/C] Ni-catalyst [Raney Ni] Lindlar [Lindlar]
	(2)	remain (L)	

An expert system web interface



Conclusions

- Statistical analysis of PG reactivity as a function of catalyst has been performed. Comparison with the Greene's Reactivity Charts demonstrates that some observations are inconsistent with statistical analysis performed in this work;

- A reactions similarity-based approach for the protective group reactivity assessment has been proposed and tested on the set of 72229 catalytic hydrogenation reactions. External validation demonstrated its high efficiency to predict optimal reaction conditions.

- Some 30 Python3 scripts realizing data preparation and Expert system workflows have been developed. They were implemented in ChemPortal WEB interface, http://cimm.kpfu.ru (unavailable at the moment)

Authors and collaborators

Arkadii Lin (KFU, UniStra)

Ramil Nugmanov (KFU)

Olga Klimchuk (UniStra)

Prof. Igor Antipin (Kazan)

Prof. Alexandre Varnek (UniStra)

Acknowledgements:
Gilles Marcou (UniStra)
Dragos Horvath (UniStra)
Timur Gimadiev (KFU, UniStra)
Pavel Sidorov (UniStra)
Sergey Neklyudov (KFU)

Analysis of initial data

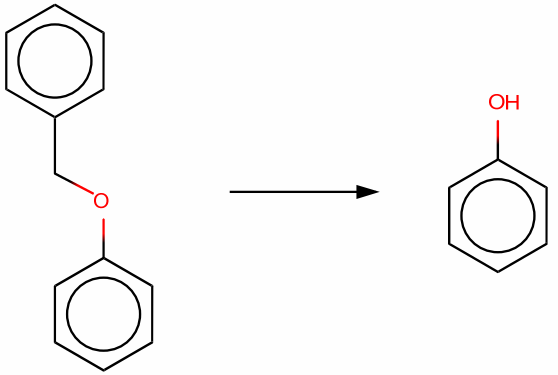
142 111 reactions (271 563 conditions)

Catalyst or reagent	T	time	P	Yield	Solvent	All
95.6	45.1	57.6	33.5	67.8	83.7	10.9

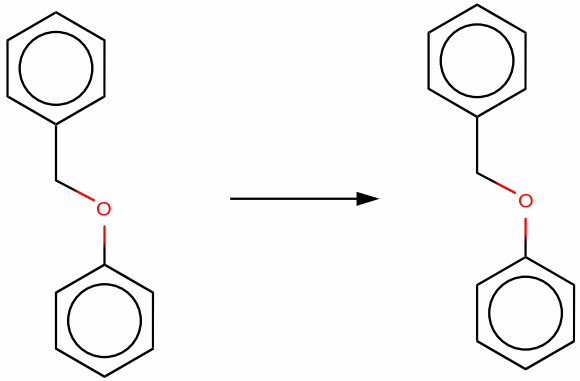
Percentage of reactions which have defined temperature (T), pressure(P), time (t), yield, solvent, catalyst or reagent and all conditions in their descriptions

Queries

2 types of query have been used:



Phenol deprotection
(PG = Benzyl group)



Phenol protection
(PG = Benzyl group)

The same approach for other PG and FG.

Appendix 1. Confusion matrix (for alcohol protection)

Benzyl PG							
CPG class	9352		RPG class	1308		AUC	0.95
Δ	0		Δ	0.05		Δ	0.1
True Positive	*False Positive*		*True Positive*	*False Positive*		*True Positive*	*False Positive*
9006	294		8440	172		7625	94
False Negative	*True Negative*		*False Negative*	*True Negative*		*False Negative*	*True Negative*
346	1014		912	1136		1727	1214
Sensitivity	Specificity		Sensitivity	Specificity		Sensitivity	Specificity
0.96	0.77		0.90	0.87		0.81	0.93
Balanced Accuracy	0.87		Balanced Accuracy	0.89		Balanced Accuracy	0.87

APPENDIX 1. CONFUSION MATRIX (for alcohol protection)

Triphenylmethyl PG							
CPG class	105		RPG class	75		AUC	0.98
Δ	0		Δ	0.05		Δ	0.1
True Positive	*False Positive*		*True Positive*	*False Positive*		*True Positive*	*False Positive*
101	10		100	5		98	4
False Negative	*True Negative*		*False Negative*	*True Negative*		*False Negative*	*True Negative*
4	65		5	70		7	71
Sensitivity	Specificity		Sensitivity	Specificity		Sensitivity	Specificity
0.96	0.87		0.95	0.93		0.93	0.95
Balanced Accuracy	0.91		Balanced Accuracy	0.94		Balanced Accuracy	0.94

Appendix 1. Confusion matrix (for alcohol protection)

Trimethylsilyl (TMS) PG							
CPG class	40		RPG class	66		AUC	0.97
Δ	0		Δ	0.05		Δ	0.1
True Positive	*False Positive*		*True Positive*	*False Positive*		*True Positive*	*False Positive*
37	3		37	2		37	1
False Negative	*True Negative*		*False Negative*	*True Negative*		*False Negative*	*True Negative*
3	63		3	64		3	65
Sensitivity	Specificity		Sensitivity	Specificity		Sensitivity	Specificity
0.92	0.95		0.92	0.97		0.92	0.98
Balanced Accuracy	0.94		Balanced Accuracy	0.95		Balanced Accuracy	0.95

Appendix 2. Confusion matrix (for amine protection)

Benzyl Carbamate PG					
CPG class	9551	RPG class	304	AUC	0.94
Δ	0	Δ	0.05	Δ	0.1
True Positive	*False Positive*	*True Positive*	*False Positive*	*True Positive*	*False Positive*
9398	98	9075	73	8828	42
False Negative	*True Negative*	*False Negative*	*True Negative*	*False Negative*	*True Negative*
153	206	476	231	723	262
Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
0.98	0.68	0.95	0.76	0.92	0.86
Balanced Accuracy	0.83	Balanced Accuracy	0.85	Balanced Accuracy	0.89

Appendix 3. Confusion matrix (for phenol protection)

Benzyl PG							
CPG class	6271		RPG class	284		AUC	0.96
Δ	0		Δ	0.05		Δ	0.1
True Positive	*False Positive*		*True Positive*	*False Positive*		*True Positive*	*False Positive*
6174	85		6050	63		5912	47
False Negative	*True Negative*		*False Negative*	*True Negative*		*False Negative*	*True Negative*
97	199		221	221		359	237
Sensitivity	Specificity		Sensitivity	Specificity		Sensitivity	Specificity
0.98	0.70		0.96	0.78		0.94	0.83
Balanced Accuracy	0.84		Balanced Accuracy	0.87		Balanced Accuracy	0.89

Analysis of initial data

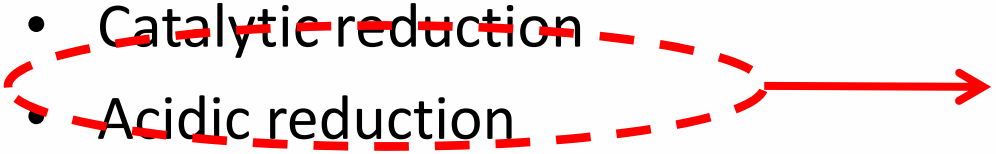
142 111 reactions (271 563 conditions)

Catalyst or reagent	T	time	P	Yield	Solvent	All
95.6	45.1	57.6	33.5	67.8	83.7	10.9

Percentage of reactions which have defined temperature (T), pressure(P), time (t), yield, solvent, catalyst or reagent and all conditions in their descriptions

Methods of deprotection

- Aqueous
- Organometallic
- Catalytic reduction
- Acidic reduction
- Hydride reduction
- Thermal reactions
- Etc.



This method has been used in this project

Catalyst annotation