Chemoinformatics and
Molecular Modeling Lab

Chémoinformatique
STRASBOURG

State-of-the-Art in Chemical Reaction Characteristics Prediction Using Condensed Graph of Reaction

Dr. Timur I. Madzhidov
Kazan Federal University, Department of Organic Chemistry

tmadzhidov@gmail.com

(Almost not) a dream

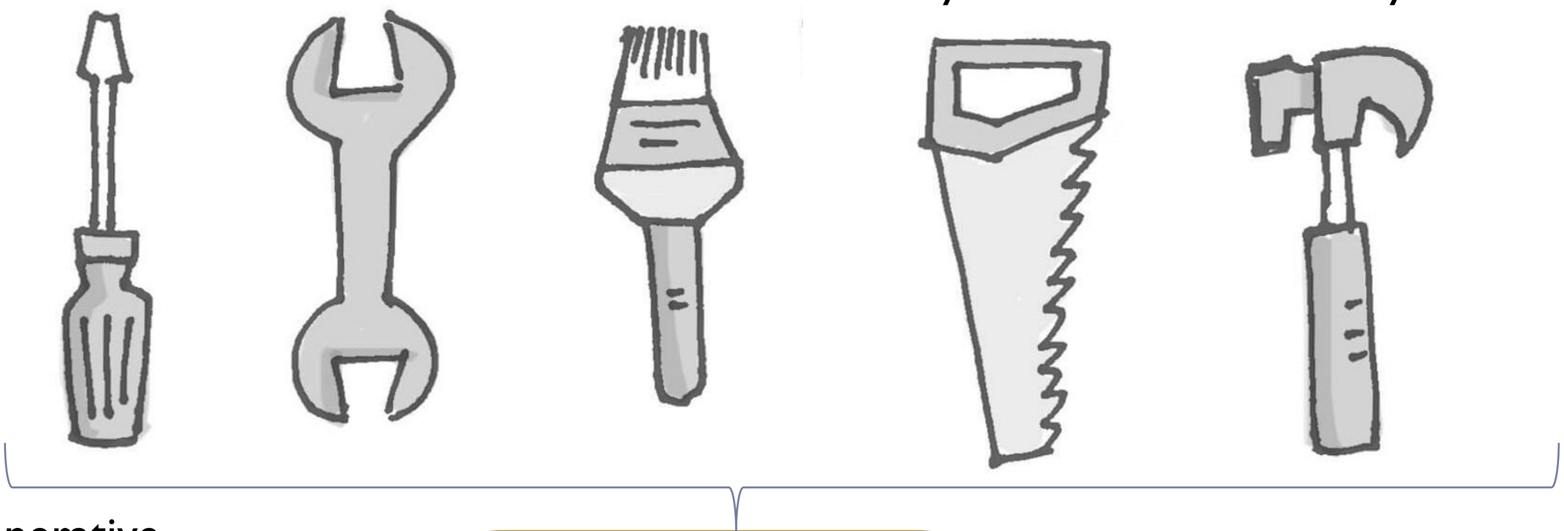
QSAR

SBDD

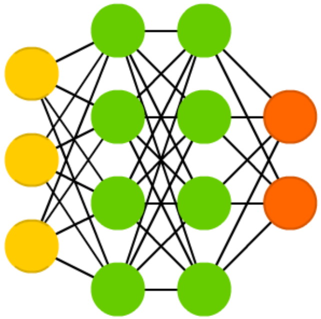
Similarity

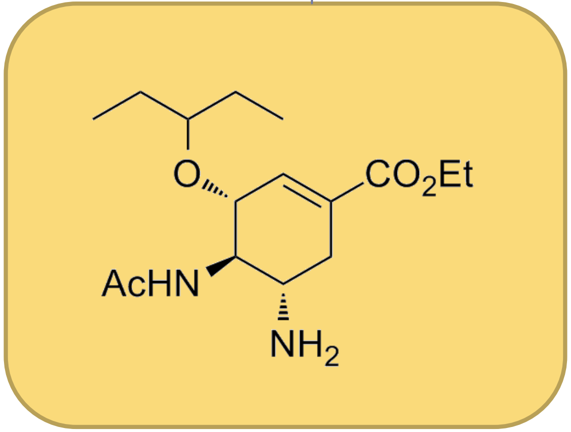
Molecular dynamics

Quantum chemistry

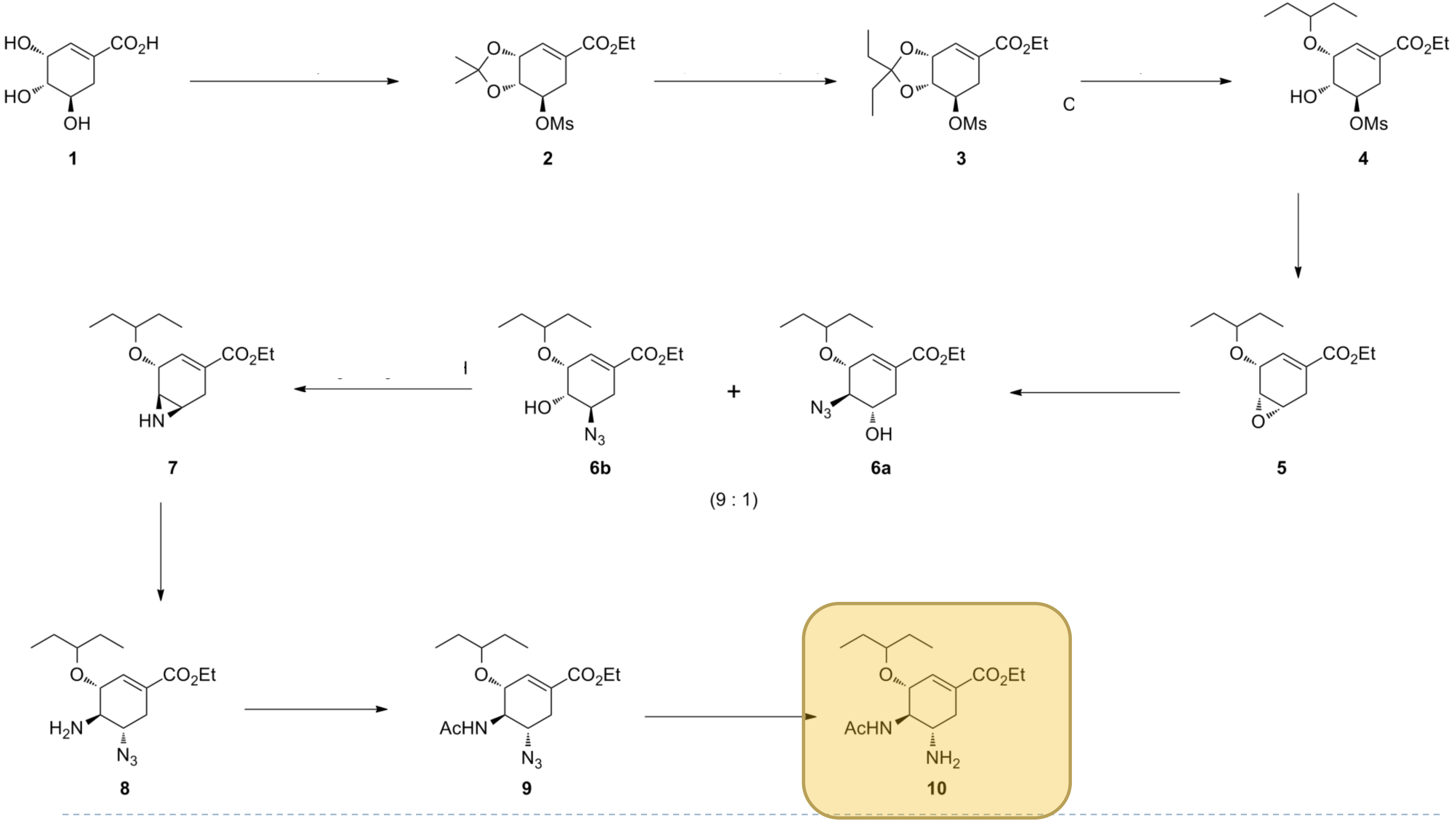


Generative Neural Nets





The reality

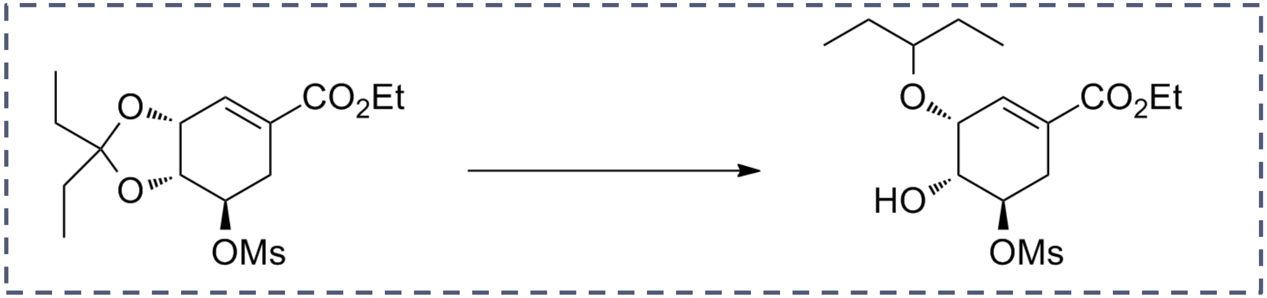


HO
CO2H
HO
OH
1
O
O
CO2Et
OMs
2
O
O
CO2Et
OMs
3
C
O
CO2Et
HO
OMs
4
O
CO2Et
O
5
O
CO2Et
N3
OH
6a
+
O
CO2Et
HO
N3
6b
(9 : 1)
O
CO2Et
HN
7
O
CO2Et
H2N
N3
8
O
CO2Et
AcHN
N3
9
O
CO2Et
AcHN
NH2
10

Some practical questions:

What is the rate of reaction ?

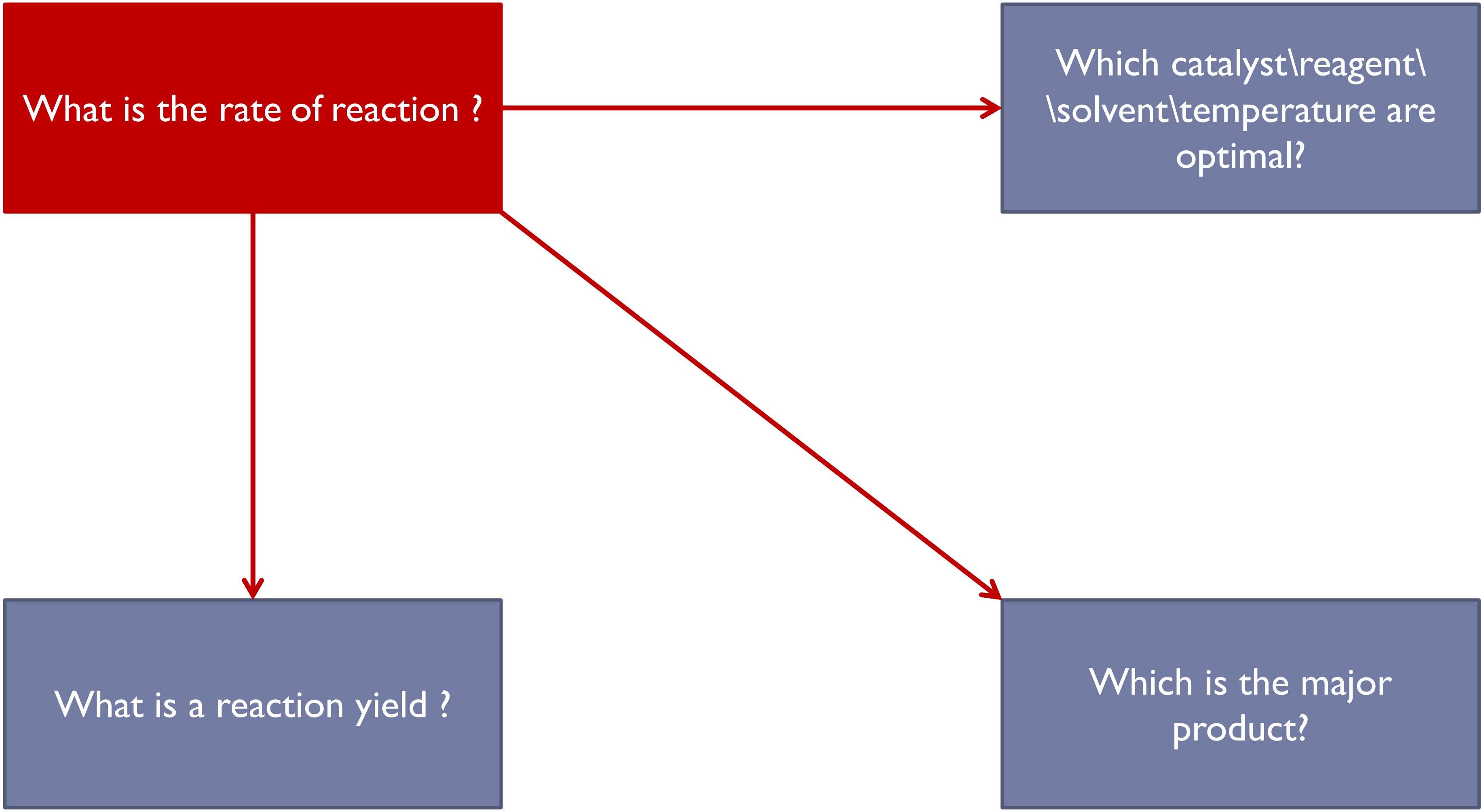
Which catalyst\reagent\ \solvent\temperature are optimal?



What is a reaction yield ?

Which is the major product?

Some practical questions:



What is the rate of reaction ?

Which catalyst\reagent\ \solvent\temperature are optimal?

What is a reaction yield ?

Which is the major product?

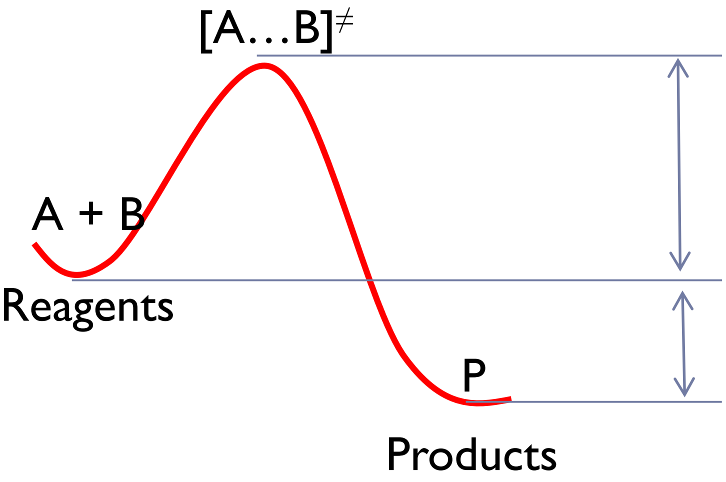
Goal of the study

Goal: to built predictive models for rate constant as a function of structure of reactants and experimental conditions.

Here we demonstrate this approach for the case of S_N2 reactions

Reaction rate assessment: QM approach

Quantum Chemistry



$\Delta G^{\neq}$
Activation Free Energy

ΔG
Free Energy of Reaction

- Time-consuming (~1 day-1 week per one reaction per CPU core)
- Description of reaction in solvent complicates and slows down calculations, accuracy decreases substantially
- Reaction rate could hardly be quantitatively reproduced

Rate constant

$$k = \kappa\left(\frac{k_B T}{\hbar}\right)e^{-\frac{\Delta G^{\neq}}{RT}}$$

Equilibrium constant

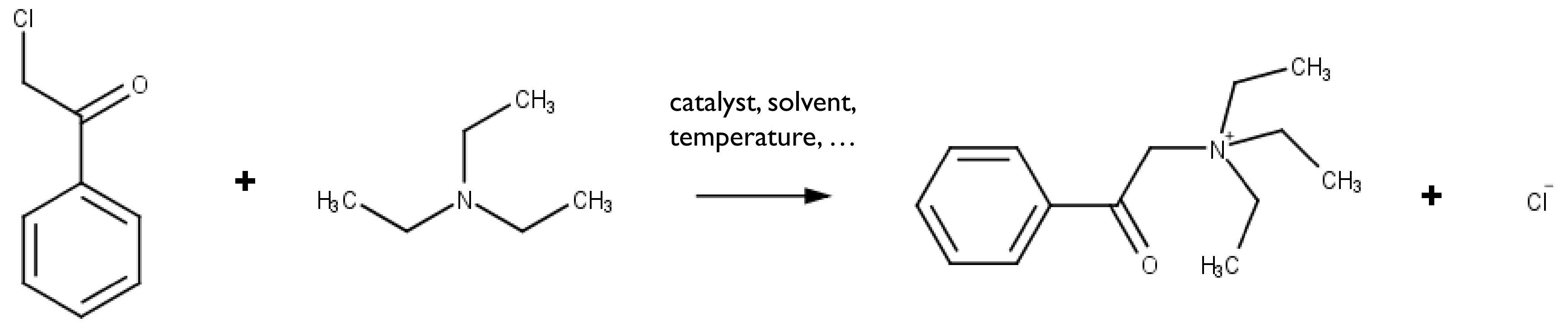
$$K = e^{-\frac{\Delta G}{RT}}$$

Reaction rate assessment: chemoinformatics approach

QSAR/QSPR approaches are usually applied to individual molecules.

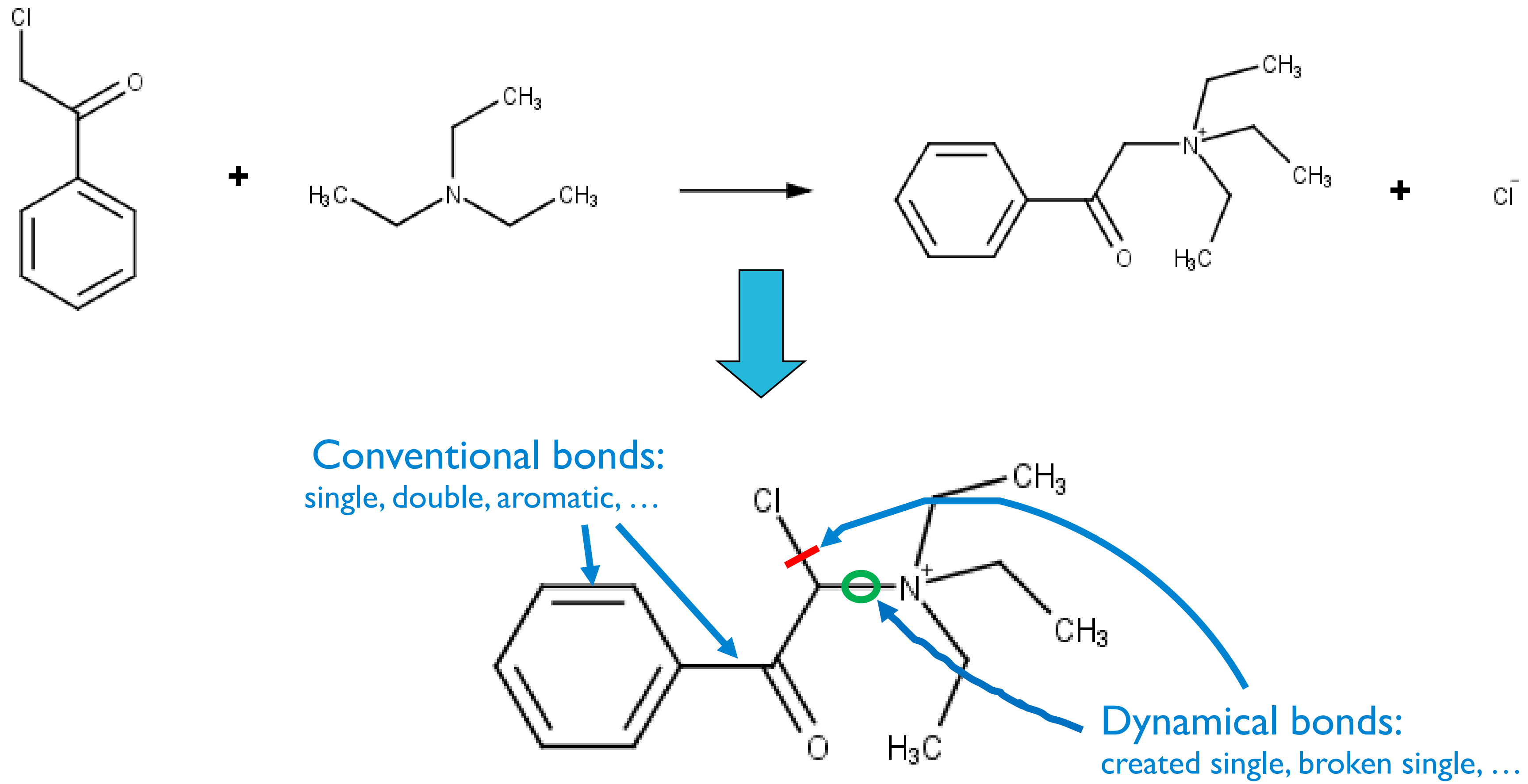
What about chemical reactions ?

Chemical reactions: complexity issue



- many species of two types: reactants and products;
- dependence of characteristics on reaction conditions (catalyst, solvent, etc)

Condensed Graph of Reaction



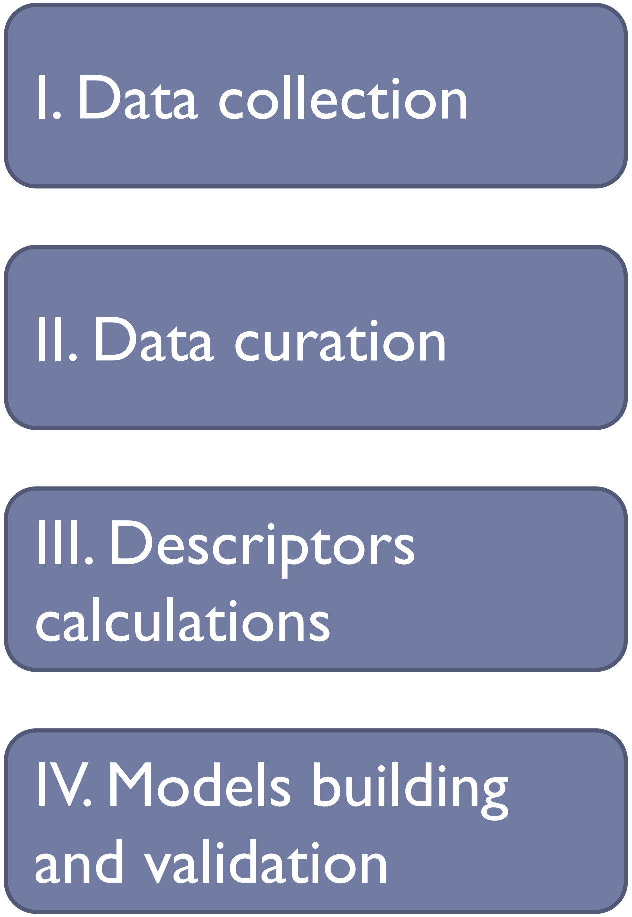
Conventional bonds:
single, double, aromatic, …

Dynamical bonds:
created single, broken single, …

CGR: a pseudo-molecule representing a given reaction

Varnek A., et al. (2005) J Comput Aided Mol Des 19:693. doi: 10.1007/s10822-005-9008-0

Modeling workflow



I. Data collection

II. Data curation

III. Descriptors calculations

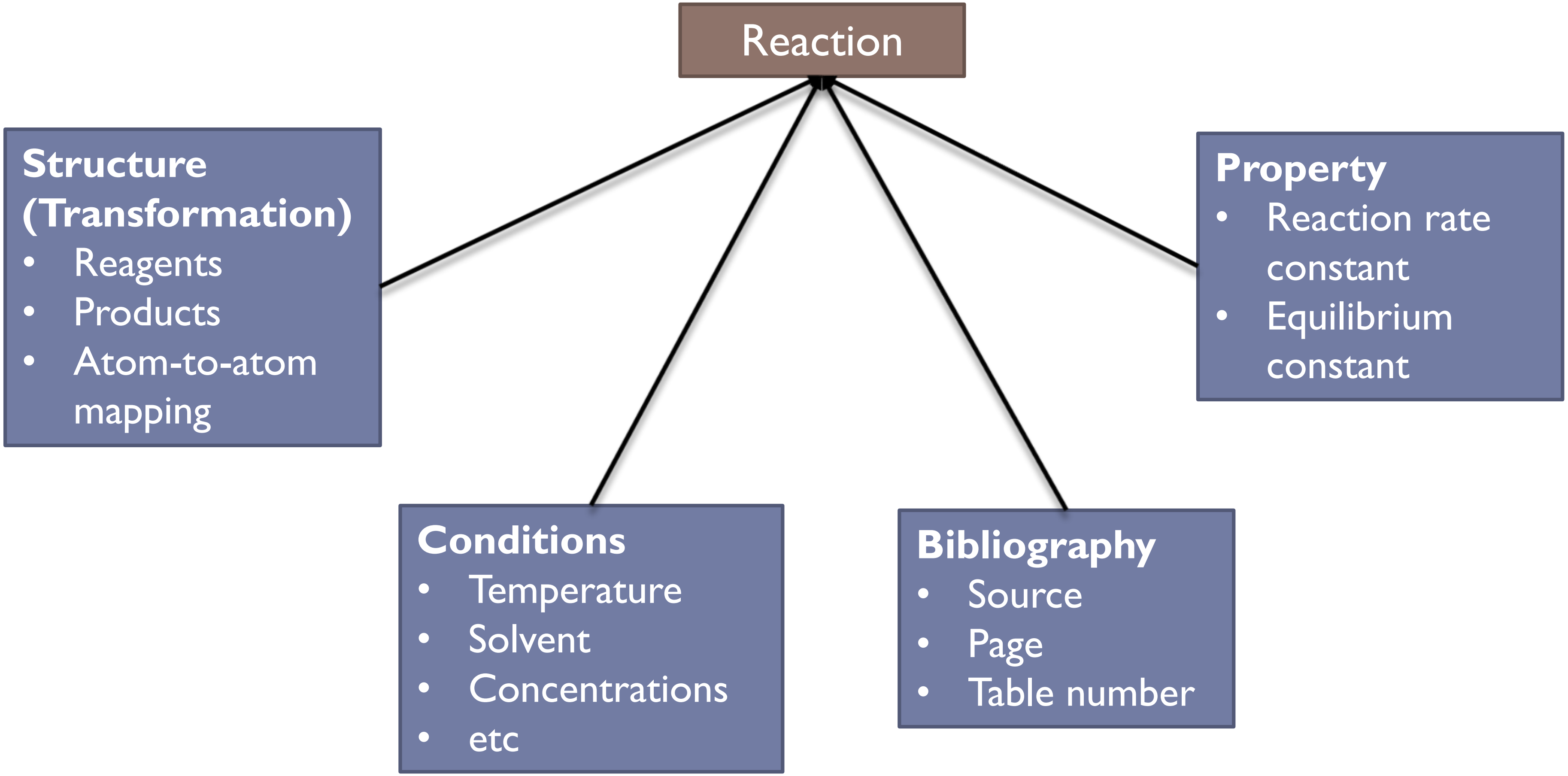
IV. Models building and validation

Datasets

Problem: lack of data

- No public databases (like ChEMBL, PubChem) for reactions
- Commercial databases (Reaxys, SciFinder) don’t annotate kinetic or thermodynamic characteristics of reactions
- Only yield is annotated in databases. However, this is very noisy parameter and it could hardly be directly modelled.

QSRR-DB: comprehensive reactions database



Reaction

Structure (Transformation)

- Reagents
- Products
- Atom-to-atom mapping

Conditions

- Temperature
- Solvent
- Concentrations
- etc

Bibliography

- Source
- Page
- Table number

Property

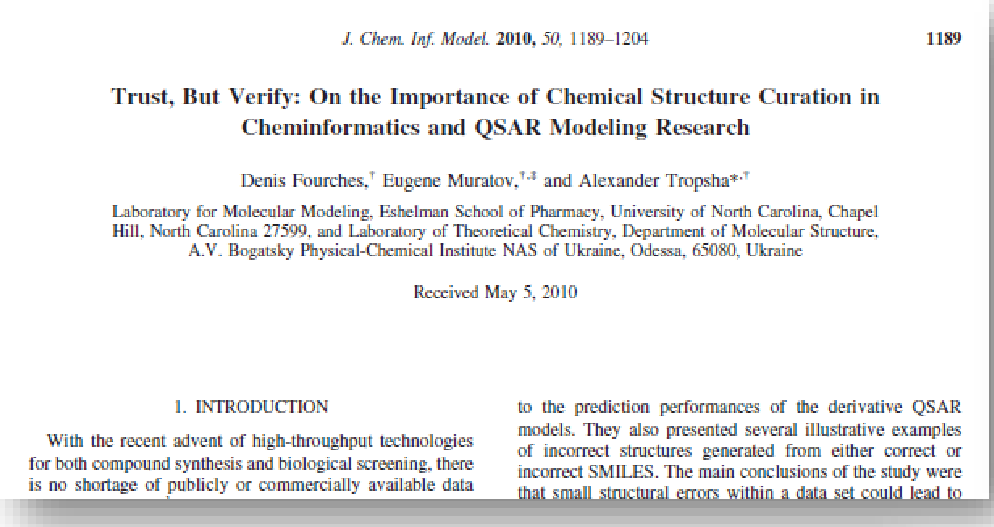
- Reaction rate constant
- Equilibrium constant

QSRR-DB: comprehensive reactions database

- Substitution (S_N2) reactions rate constants: >7000
- Substitution (S_N1) reactions rate constants: >7000
- Elimination (E2) reaction rate constants: >2500
- Ester hydrolysis reaction rate constants: ~4000
- Cycloaddition (Diels-Alder etc) reactions rate constants and Arrhenius eqn parameters: for ~1500 reactions
- Tautomeric equilibrium constants: >1000 equilibria
- Acidity in non-aqueous solvents: > 2000 equilibria

>25,000 records have been collected

Data curation strategies for individual molecules



J. Chem. Inf. Model. **2010**, 50, 1189–1204 **1189**

Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research

Denis Fourches,[†] Eugene Muratov,[†,‡] and Alexander Tropsha*[,†]

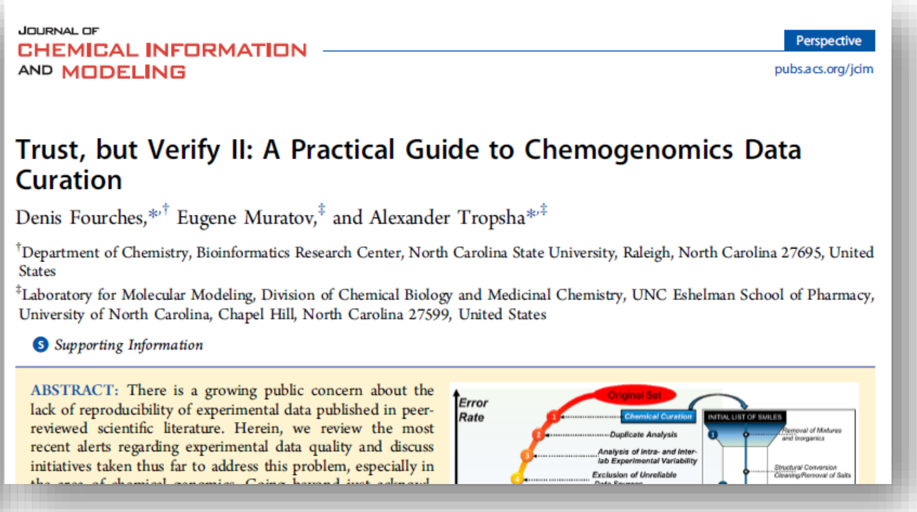
Laboratory for Molecular Modeling, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, and Laboratory of Theoretical Chemistry, Department of Molecular Structure, A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Odessa, 65080, Ukraine

Received May 5, 2010

1. INTRODUCTION

With the recent advent of high-throughput technologies for both compound synthesis and biological screening, there is no shortage of publicly or commercially available data

to the prediction performances of the derivative QSAR models. They also presented several illustrative examples of incorrect structures generated from either correct or incorrect SMILES. The main conclusions of the study were that small structural errors within a data set could lead to



JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

pubs.acs.org/jcim

Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation

Denis Fourches,*[,†] Eugene Muratov,[‡] and Alexander Tropsha*[,‡]

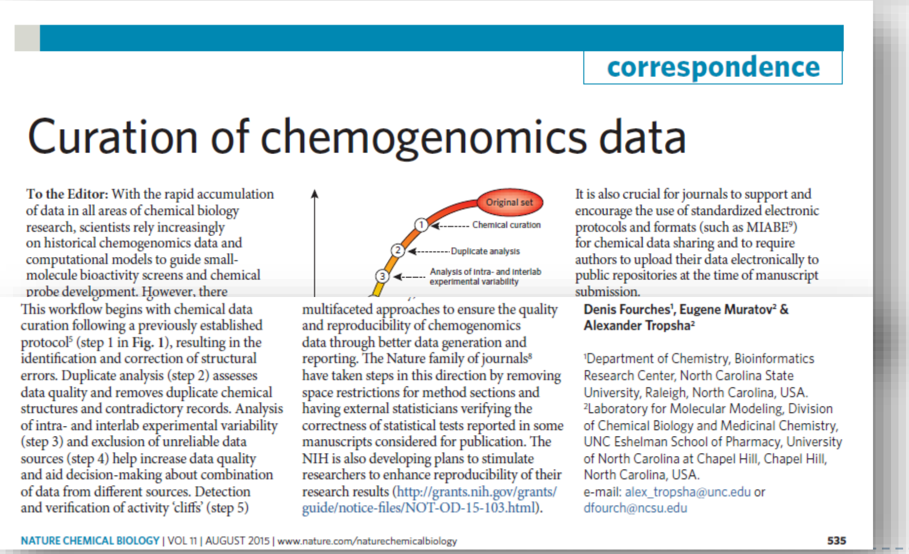
[†]Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695, United States

[‡]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, United States

Ⓢ *Supporting Information*

ABSTRACT: There is a growing public concern about the lack of reproducibility of experimental data published in peer-reviewed scientific literature. Herein, we review the most recent alerts regarding experimental data quality and discuss initiatives taken thus far to address this problem, especially in

Error Rate
Original Set
Chemical Curation
INITIAL LIST OF SMILES
Removal of Mixtures and Inorganics
Duplicate Analysis
Analysis of Intra- and Inter-lab Experimental Variability
Structural Conversion Cleaning/Removal of Salts
Exclusion of Unreliable



correspondence

Curation of chemogenomics data

To the Editor: With the rapid accumulation of data in all areas of chemical biology research, scientists rely increasingly on historical chemogenomics data and computational models to guide small-molecule bioactivity screens and chemical probe development. However, there

Original set
1 Chemical curation
2 Duplicate analysis
3 Analysis of intra- and interlab experimental variability

It is also crucial for journals to support and encourage the use of standardized electronic protocols and formats (such as MIABE[9]) for chemical data sharing and to require authors to upload their data electronically to public repositories at the time of manuscript submission.

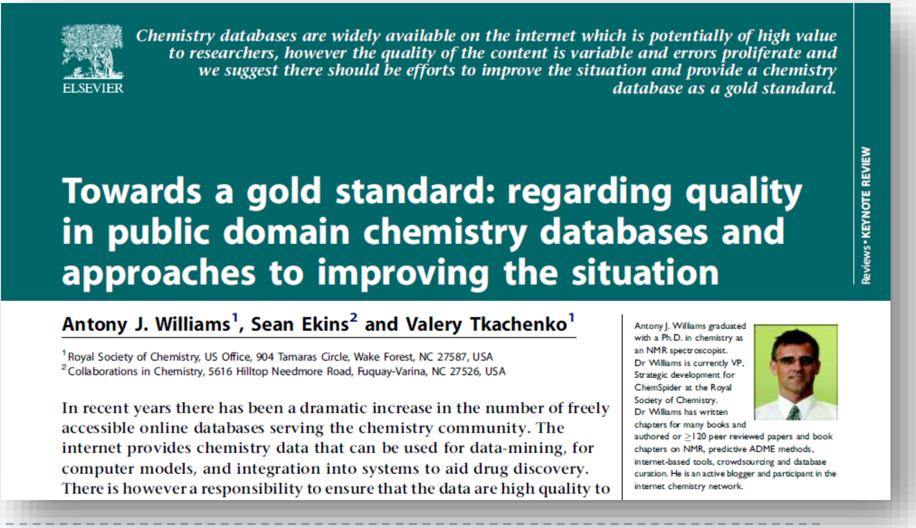
This workflow begins with chemical data curation following a previously established protocol[5] (step 1 in **Fig. 1**), resulting in the identification and correction of structural errors. Duplicate analysis (step 2) assesses data quality and removes duplicate chemical structures and contradictory records. Analysis of intra- and interlab experimental variability (step 3) and exclusion of unreliable data sources (step 4) help increase data quality and aid decision-making about combination of data from different sources. Detection and verification of activity 'cliffs' (step 5)

multifaceted approaches to ensure the quality and reproducibility of chemogenomics data through better data generation and reporting. The Nature family of journals[8] have taken steps in this direction by removing space restrictions for method sections and having external statisticians verifying the correctness of statistical tests reported in some manuscripts considered for publication. The NIH is also developing plans to stimulate researchers to enhance reproducibility of their research results (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html).

Denis Fourches[1], Eugene Muratov[2] & Alexander Tropsha[2]

[1]Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA. [2]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.
e-mail: alex_tropsha@unc.edu or dfourch@ncsu.edu

NATURE CHEMICAL BIOLOGY | VOL 11 | AUGUST 2015 | www.nature.com/naturechemicalbiology 535



ELSEVIER

Chemistry databases are widely available on the internet which is potentially of high value to researchers, however the quality of the content is variable and errors proliferate and we suggest there should be efforts to improve the situation and provide a chemistry database as a gold standard.

Reviews • KEYNOTE REVIEW

Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation

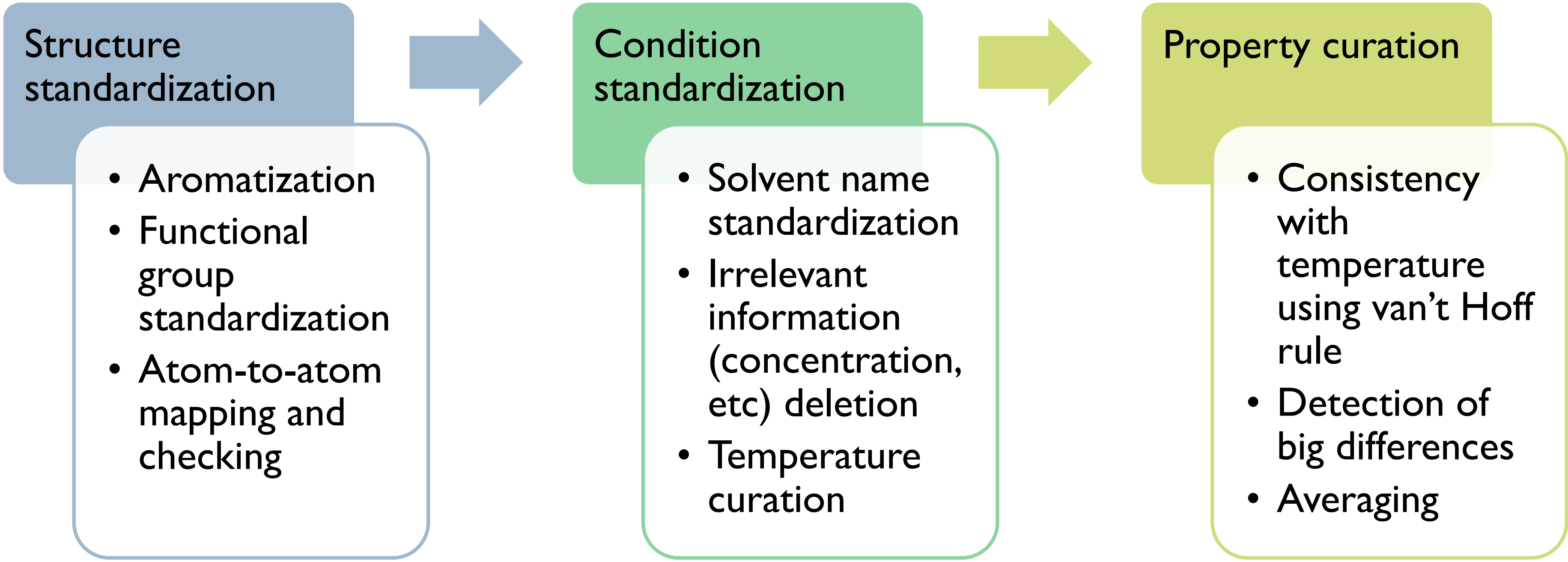
Antony J. Williams[1], Sean Ekins[2] and Valery Tkachenko[1]

[1] Royal Society of Chemistry, US Office, 904 Tamaras Circle, Wake Forest, NC 27587, USA
[2] Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526, USA

In recent years there has been a dramatic increase in the number of freely accessible online databases serving the chemistry community. The internet provides chemistry data that can be used for data-mining, for computer models, and integration into systems to aid drug discovery. There is however a responsibility to ensure that the data are high quality to

Antony J. Williams graduated with a Ph.D. in chemistry as an NMR spectroscopist. Dr Williams is currently VP, Strategic development for ChemSpider at the Royal Society of Chemistry. Dr Williams has written chapters for many books and authored or ≥120 peer reviewed papers and book chapters on NMR, predictive ADME methods, internet-based tools, crowdsourcing and database curation. He is an active blogger and participant in the internet chemistry network.

Data curation strategies for reactions



Structure standardization

- Aromatization
- Functional group standardization
- Atom-to-atom mapping and checking

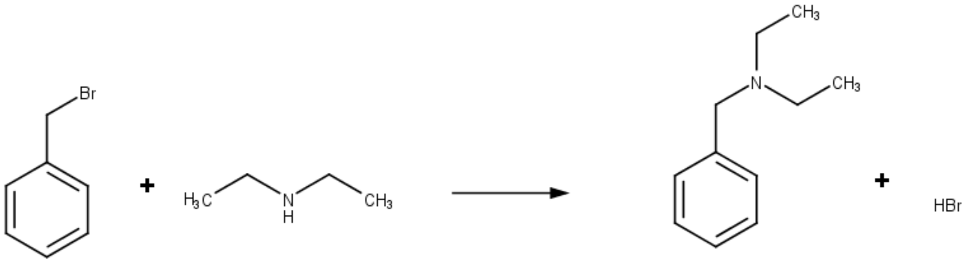
Condition standardization

- Solvent name standardization
- Irrelevant information (concentration, etc) deletion
- Temperature curation

Property curation

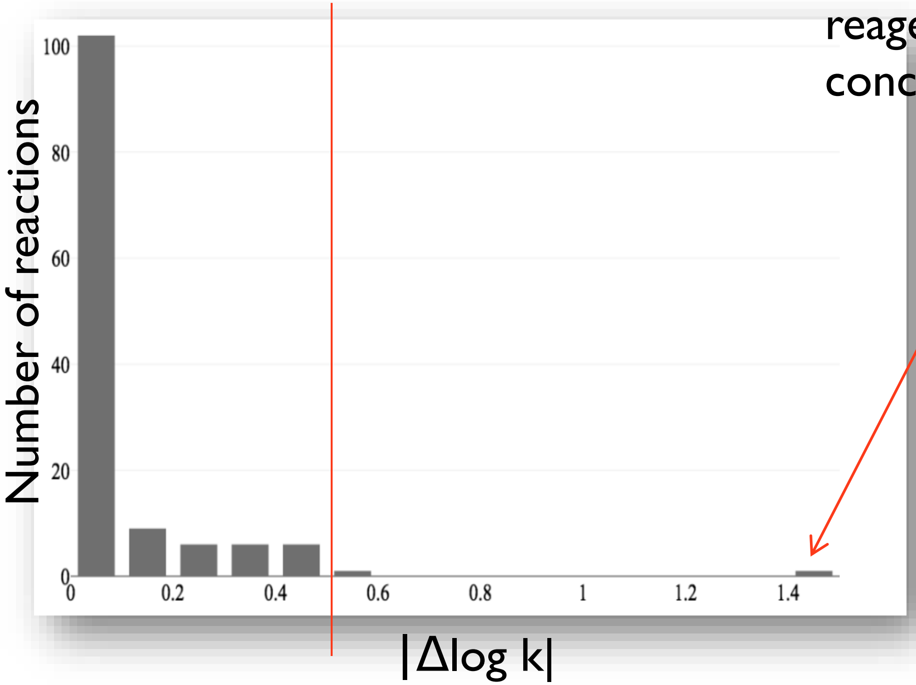
- Consistency with temperature using van’t Hoff rule
- Detection of big differences
- Averaging

Duplicate analysis



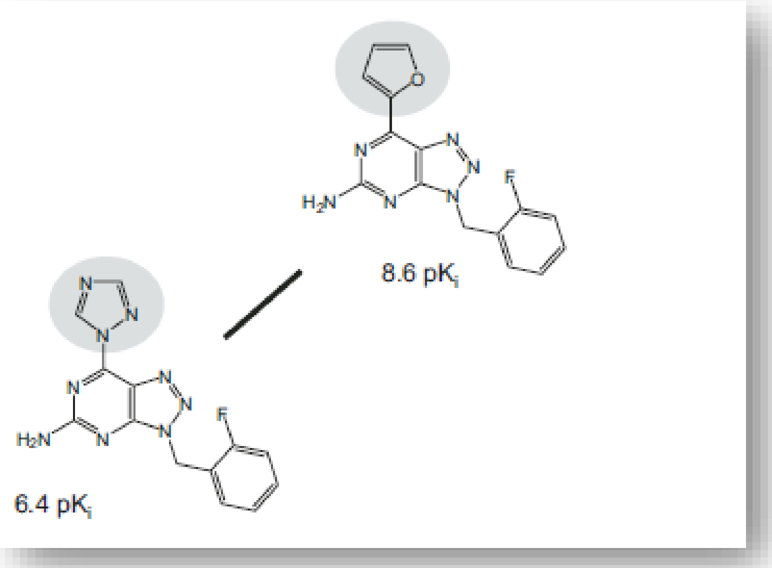
0.5-0.6 log*k* units

High dependence on reagent concentrations

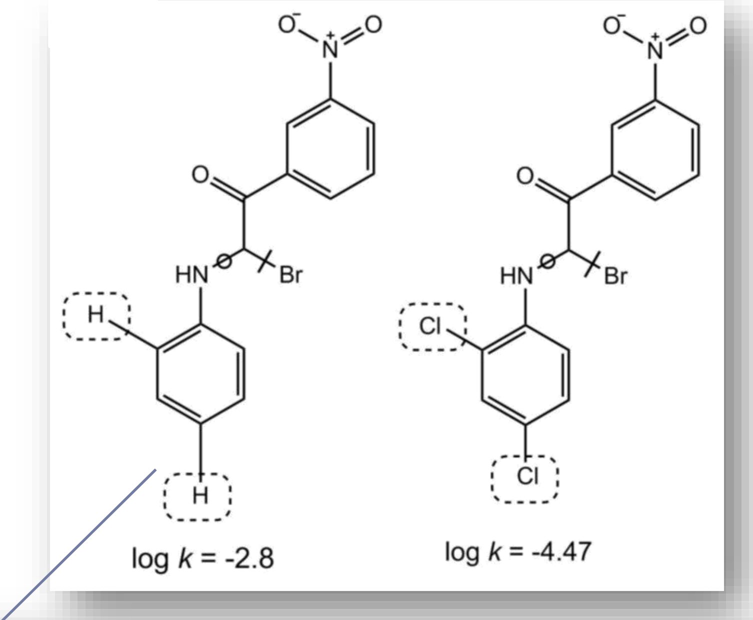


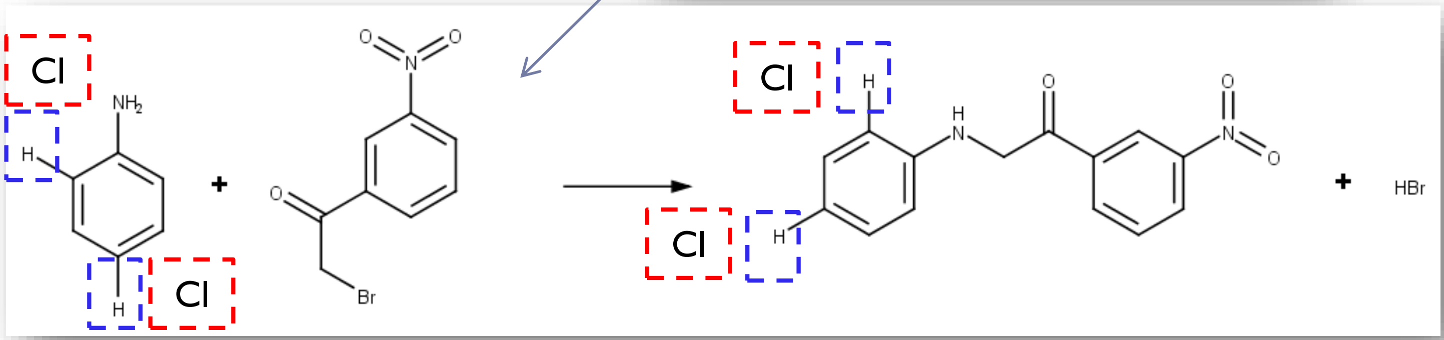
Matched Molecular Pairs

Matched Molecular Pair



Matched Reaction Pair

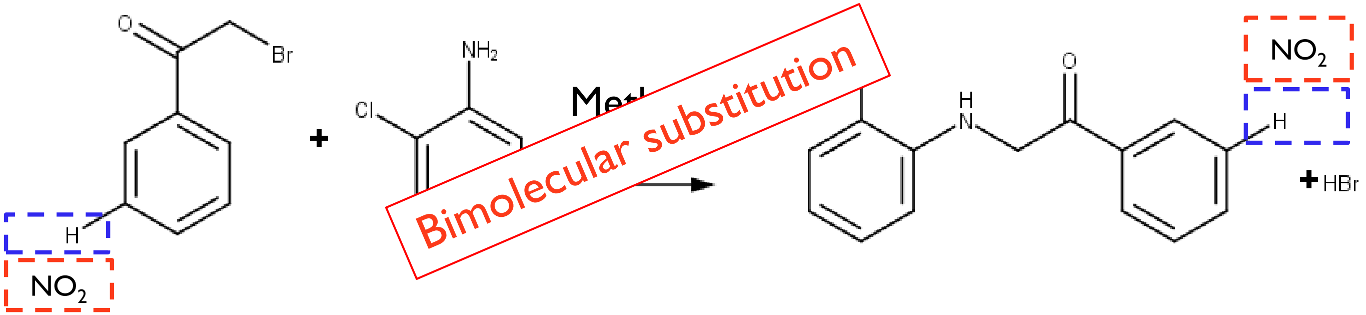




Matched Reaction Pair example

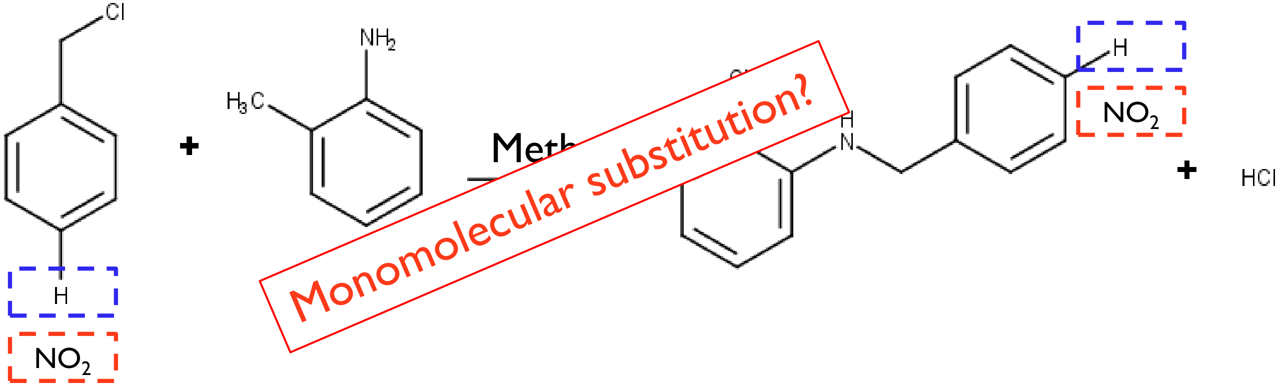
H / NO_2 substitution in substrate leads to:

Increase of reaction rate



X=H	-3.39
X=NO_2	-2.80

Decrease of reaction rate

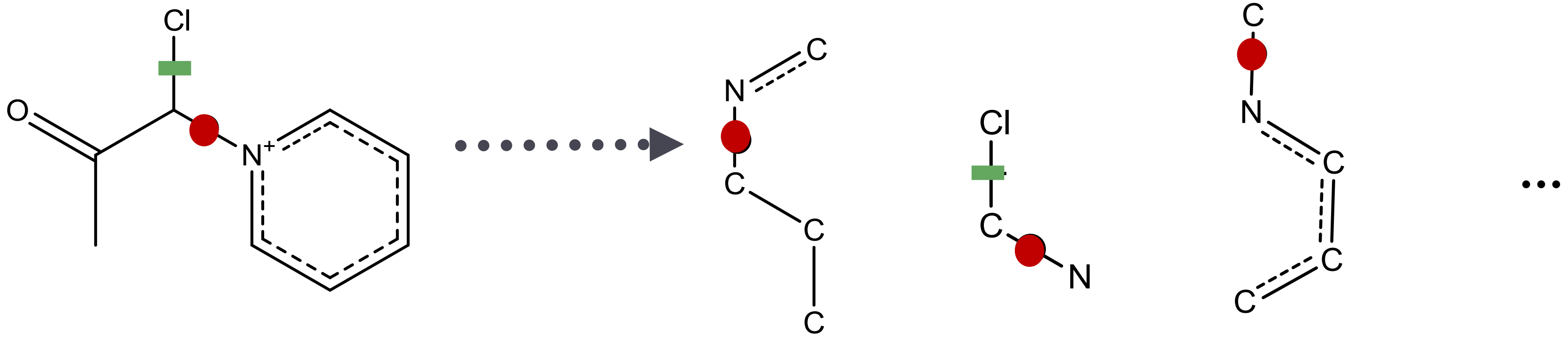


X=H	-4.35
X=NO_2	-4.86

ISIDA/CGR fragment descriptors

Condensed graph of reaction

ISIDA fragment descriptors



1	1	2	…

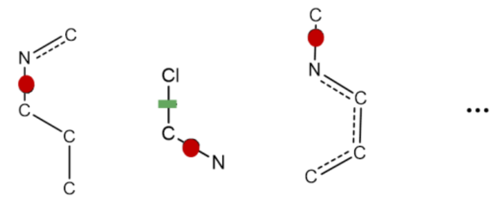
Reaction can be encoded by a descriptors vector which can be used in data analysis or in structure-reactivity modeling

A. Varnek In: "Chemoinformatics and Computational Chemical Biology", J. Bajorath, Ed., Springer, 2010

Descriptor vector combing structure & conditions

~70 -- 10 000	*13*	*1*
Structural descriptors	*Solvent descriptors*	*Temperature descriptor*

ISIDA fragments on CGRs



1	1	2	…

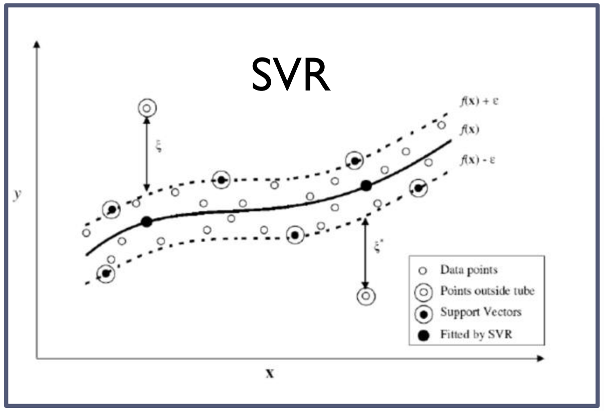
- Kamlet-Taft solvent descriptors
- Catalan solvent descriptors,
- Polarity parameters
- Polarizability parameters
- Molar fraction of organic solvent in water-organic solution

Inverse temperature of reaction, 1/T(in K)

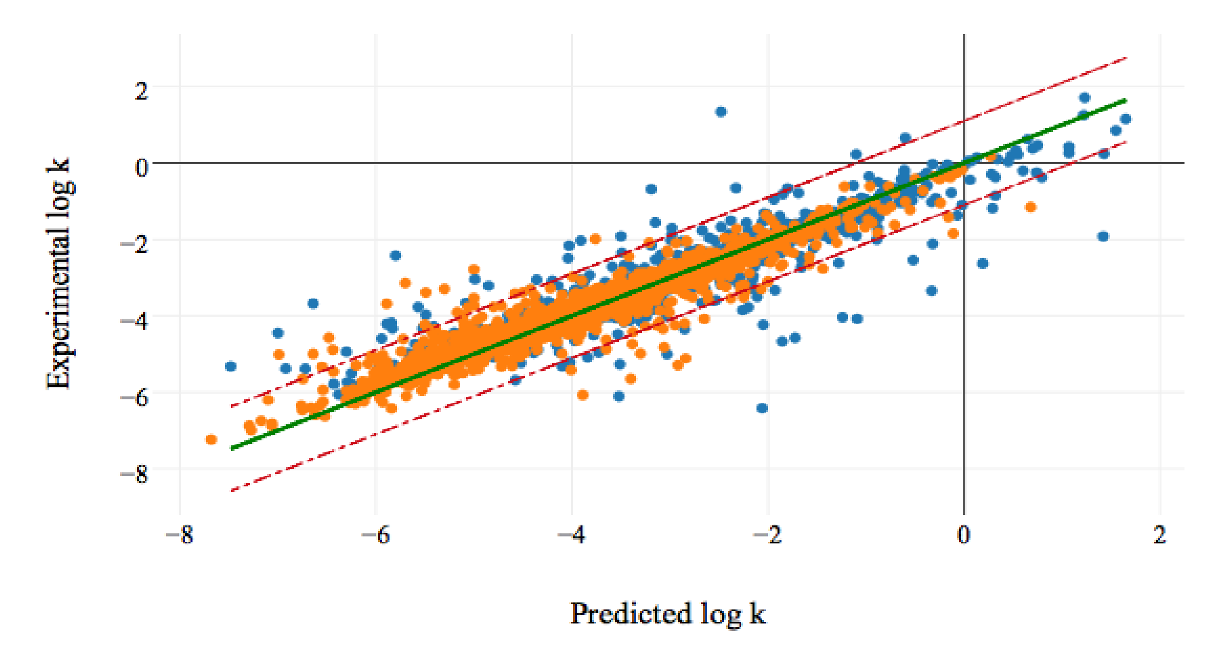
Madzhidov TI, et al (2014) Russ J Org Chem 50:459–463. doi: 10.1134/S1070428014040010

SVR model for rate constant of S_N2 reaction

Initial data set (7848 reactions) → Curated data set (4830 reactions)

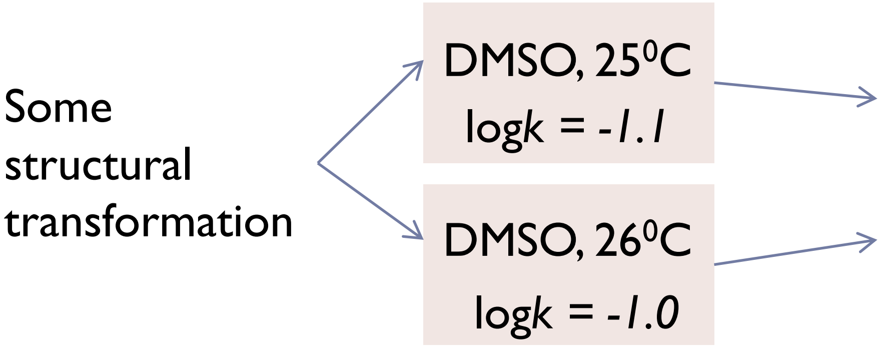


RMSE = 0.39 log*k* units
$R^2 = 0.93$



Blue points – neutral nucleophiles, orange – anionic nucleophiles

Why so good?



Some structural transformation

DMSO, 25^{0}C

log*k* = -*1.1*

DMSO, 26^{0}C

log*k* = -*1.0*

Cross -validation

	Model 1	Model 2	Model 3	Model 4	Model 5
Fold 1	TEST	TRAIN	TRAIN	TRAIN	TRAIN
Fold 2	TRAIN	TEST	TRAIN	TRAIN	TRAIN
Fold 3	TRAIN	TRAIN	TEST	TRAIN	TRAIN
Fold 4	TRAIN	TRAIN	TRAIN	TEST	TRAIN
Fold 5	TRAIN	TRAIN	TRAIN	TRAIN	TEST

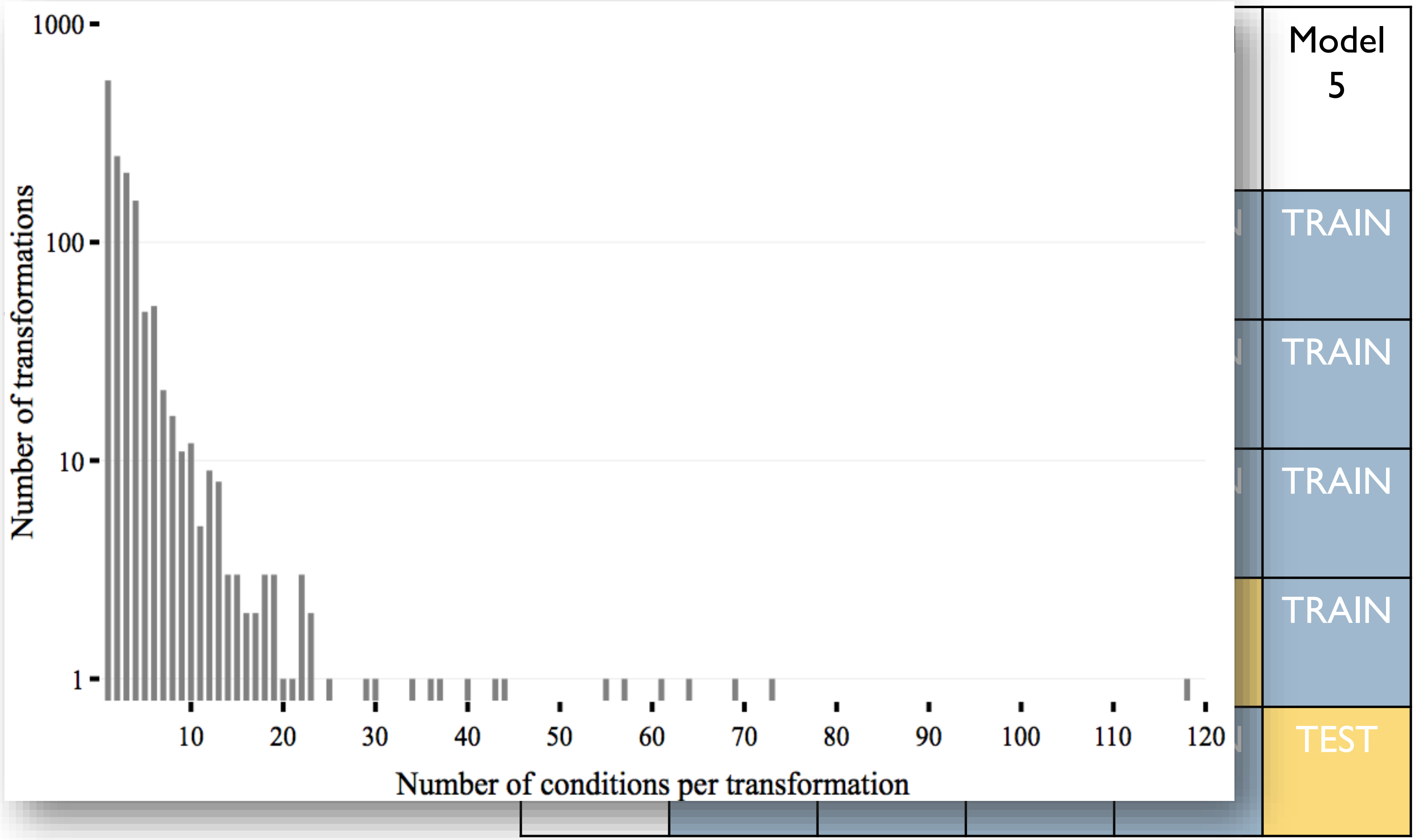
Gimadiev TR, et al (2018) J Comput Aided Mol Des 32:401–414. doi: 10.1007/s10822-018-0101-6

Polishchuk P, et al (2017) J Comput Aided Mol Des 31:829–839. doi: 10.1007/s10822-017-0044-3

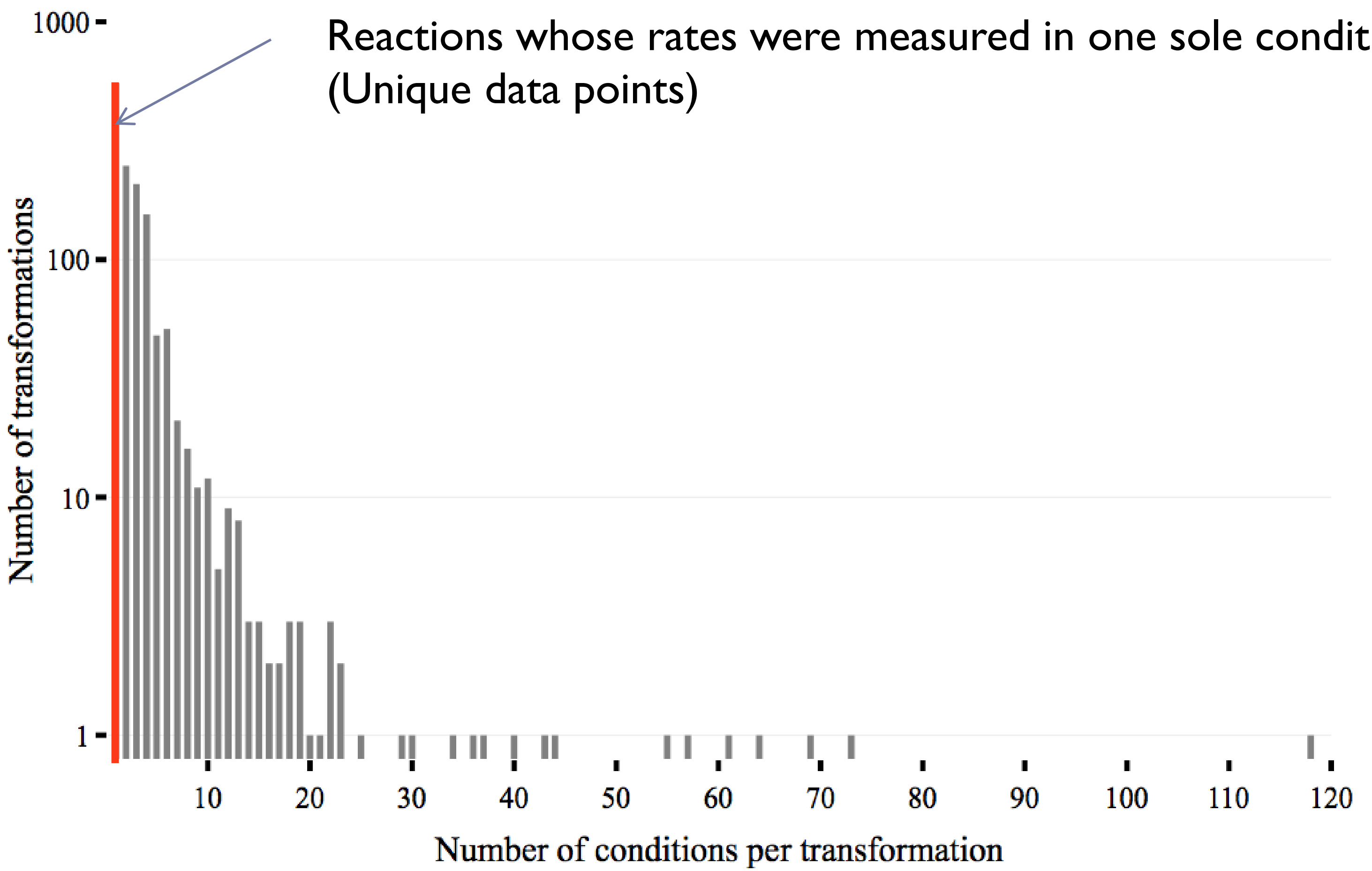
Why so good?

Cross -validation

Some structural transformation

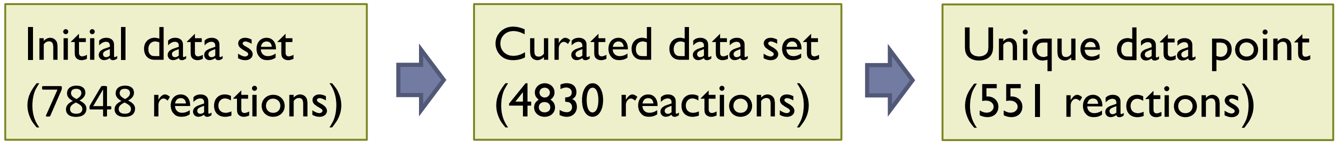


Unbiased estimation of model performance



Reactions whose rates were measured in one sole condition
(Unique data points)

Unique data points in validation



Initial data set (7848 reactions)

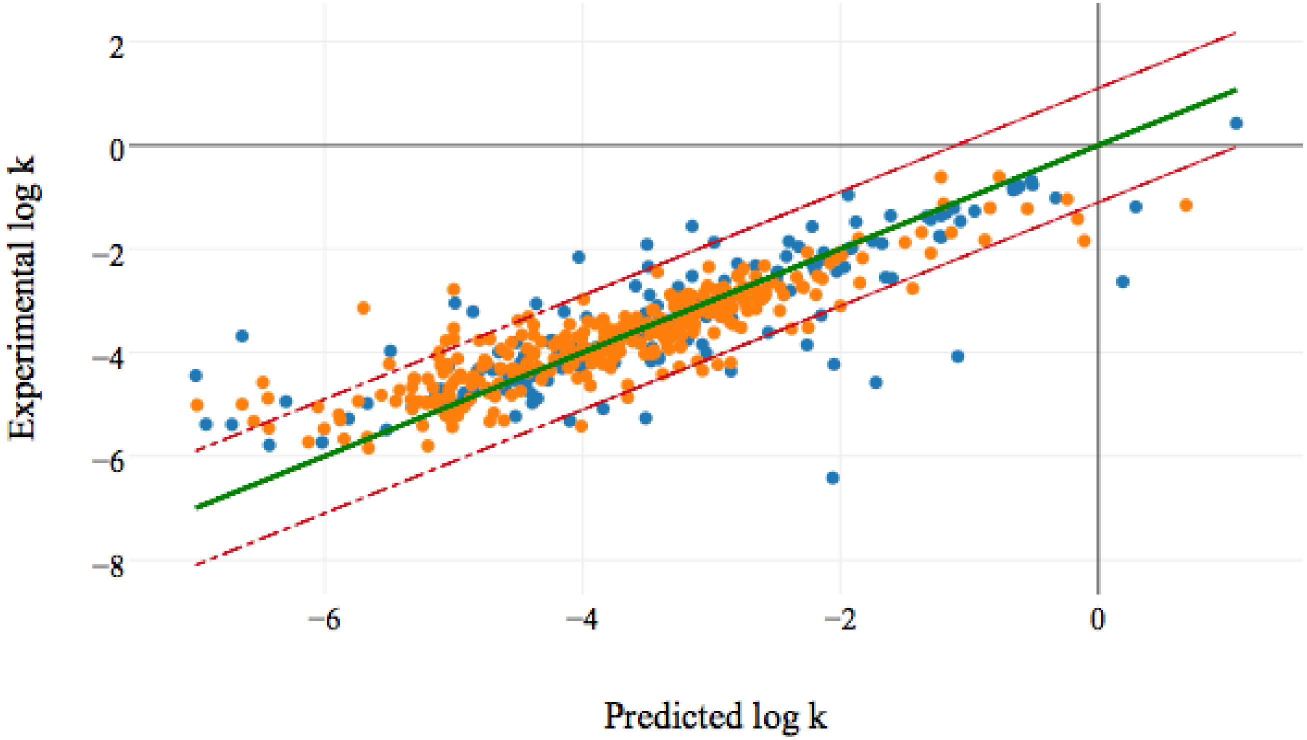
Curated data set (4830 reactions)

Unique data point (551 reactions)

$RMSE_{UDP}$ = 0.61 log*k* units
R^2_{UDP}=0.75

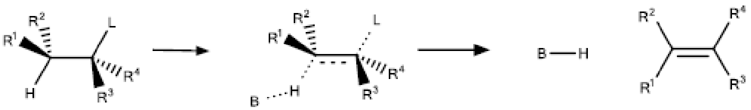
External validation (105 reactions from recent articles):

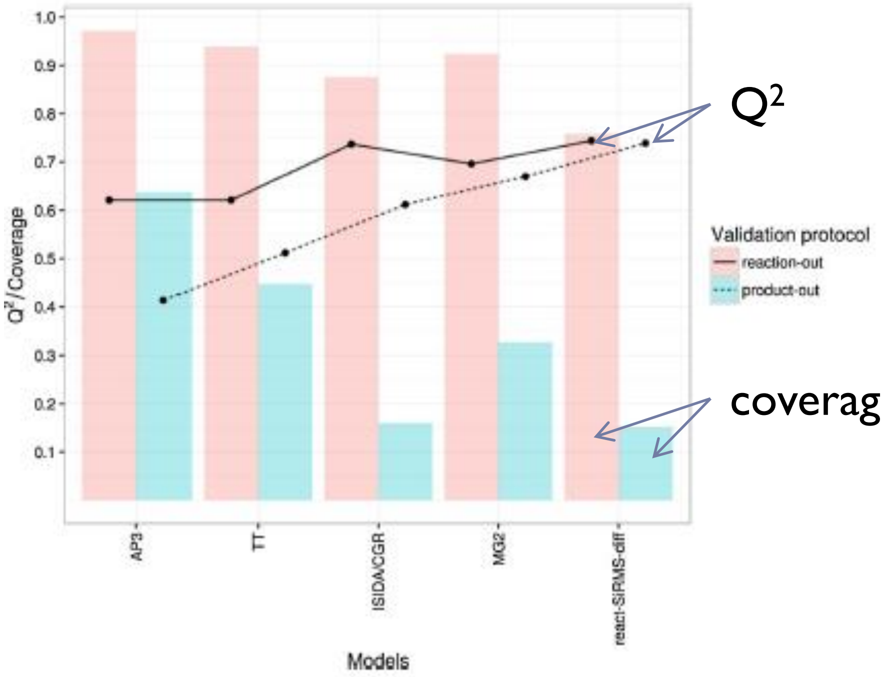
RMSE=0.8 log*k* units
R^2=0.64



Other published projects

Bimolecular elimination reaction



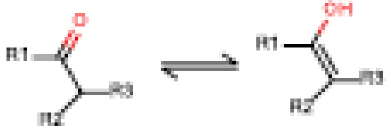


Q^2

coverage

Polishchuk P, et al (2017) J Comput Aided Mol Des 31:829–839. doi: 10.1007/s10822-017-0044-3

Tautomeric equilibria



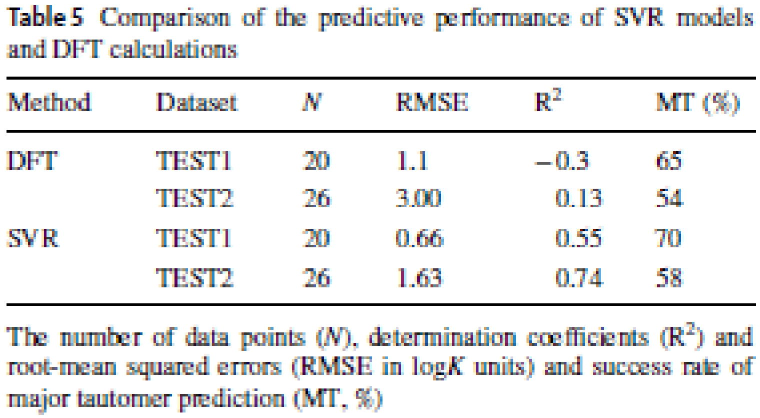


Table 5 Comparison of the predictive performance of SVR models and DFT calculations

Method	Dataset	*N*	RMSE	R^2	MT (%)
DFT	TEST1	20	1.1	−0.3	65
	TEST2	26	3.00	0.13	54
SVR	TEST1	20	0.66	0.55	70
	TEST2	26	1.63	0.74	58

The number of data points (*N*), determination coefficients (R^2) and root-mean squared errors (RMSE in log*K* units) and success rate of major tautomer prediction (MT, %)

Gimadiev TR, et al (2018) J Comput Aided Mol Des 32:401–414. doi: 10.1007/s10822-018-0101-6

Conclusions

- Reaction curation is more tricky than for molecular datasets.

- Curation of structural data should be accompanied by curation of conditions and trustworthness of predicted property value.

- Correct validation techniques should be used. Classical cross-validation overestimates model quality!

Project 14-43-00024:

Российский научный фонд

“Chemoinformatics approaches to organic and metabolic reactions: from empirical to predictive chemistry”

Prof. Alexandre Varnek (UniStra)

Prof. Igor Antipin (Kazan)

Igor Baskin (MSU)

Pavel Polishchuk (UniOlomouc)

Igor Tetko (HZM)

Ramil Nugmanov (KFU)

Timur Gimadiev (KFU, UniStra)

Olga Klimchuk (UniStra)

Data collection:
Nail Khafizov
Andrey Bodrov
Maria Knyazeva
Nikita Shalin
Firuza Bekmuratova
Daria Malakhova
etc