The Need for Comprehensive Reaction Handling in SAVI and Beyond

Marc C. Nicklaus,[1] Wolf-Dietrich Ihlenfeldt,[2] Gerd Blanke,[3] Philip Judson,[4] Victorien Delannée[1]

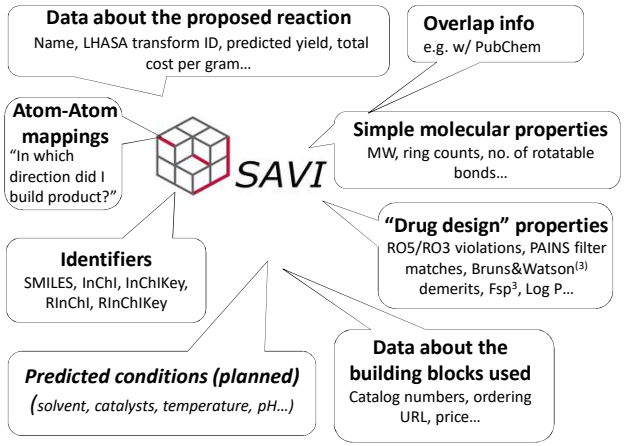
[1]CADD Group, Chemical Biology Laboratory, CCR, NCI, NIH, Frederick, MD, USA, NIH; [2]Xemistry GmbH, Königstein, Germany ;
[3] StructurePendium Technologies GmbH, Essen, Germany; [4] Consultant, Harrogate, UK

Background

Significant advances have been made in the past 20 years for structure representation and exchange standards. The same cannot be said for reactions. Diverse formats exist but are often limited in what type of reaction-related data they can handle, may not be well-interconvertable, and are for the most part not open standards. This poster presents more questions than answers: **What do we have? Do we need something new? Could something existing be extended for this?**

Reaction Data & Annotations

Data captured/generated in SAVI[(1)] (>60 types):



Other types of reaction(-associated) data (examples):

- Bibliographic info
- ELN type info (location, date, chemist's name/ID…)
- Test methods to ensure identity/purity (NMR spectra…)
- Purity data
- Stability data
- Regulatory info (manufacturer's name, address, contact info..)
- Predicted/measured activities (protein targets, bioactivity, ADME-Tox…)
- Molecular descriptors
- Electronic reaction mechanism

Acknowledgements

We thank Hitesh Patel, Yulia Borodina, and Gunther Schadow for useful discussions.

Formats for Reaction Handling/Representation

Name	Provenance	Characteristic/Purpose	ID, Classifier, or Representation?	Format	Multistep	Allows Annotations (Conditions, Yield etc.)	Can handle generic reactions (R-Groups)	Atom mapping	Can represent synthesis knowledge	Allows reaction searching	Tagged/ Markup Language ("Semantic")	Document-ation Publicly Accessible	Open Standard	Currently in use	Examples of tools for handling; uses
Existing Formats:															
ReactionSMILES	Daylight	Reaction description	Representation	Single-line	no	no	no	yes	no	no	no	yes	no (not yet…)	yes	most cheminfo tools
SMIRKS	Daylight	Transform	Representation	Single-line	no	no	yes	yes	no	no	no	yes	no (not yet…)	yes	most cheminfo tools
CHMTRAN/PATRAN	Lhasa	Transform	Representation	Multi-line	yes	yes	yes	yes	yes	no	no	no	no	yes	LHASA program; CACTVS SAVI transforms
RInChI	InChI Trust	Unique ID	ID; but convertible	Single-line with optional RAuxInfo line; planned: ProcAuxInfo for Rx conditions	no	limited	no	no	no	yes	no	yes	yes	yes	released as libraries; integrated in recent cheminfo tools
RInChIKey	InChI Trust	Unique ID	ID	Hash	no	no	no	no	no	yes	no	yes	yes	yes	released as libraries; integrated in recent cheminfo tools
RXN	MDL… -> Biovia	Data exchange	Representation	Multi-line	no	no	yes	yes	no	not directly	no	yes	no	yes	most cheminfo tools
RDfile	MDL… -> Biovia	Data exchange	Representation	Multi-line	yes	yes	yes	yes	no	not directly	no	yes	no	yes	most cheminfo tools
XDfile	Biovia	Data exchange	Representation	Multi-line	yes	yes	yes	yes	no	not directly	yes	yes	no	yes	most cheminfo tools
MRV	ChemAxon	Data exchange	Representation	Multi-line	no(?)	yes	yes	no	no	yes	yes	yes	no	yes	some cheminfo tools
UDM	Roche->Elsevier ->Pistoia Alliance		Representation	Multi-line	yes	yes	yes	no	no	yes(?)	yes	no	no	yes	Elsevier Reaxys software
BinCodes	Elsevier/Reaxys	Classification	Classifier	Multi-line	no	no	yes	yes	no	yes	no	no	no	yes	Elsevier Reaxys software
ClassCodes	InfoChem	Classification	Classifier	Multi-line	no	no	yes	yes	no	yes	no	no	no	yes	InfoChem software
CGR	Varnek group	Pseudo-molecule	Representation	Multi-line	no	yes (in SDF)	yes	yes	no	no	no	partial	no	yes	Fragmentor; Classification QSAR
ITS	Fujita (1986)	Pseudo-molecule	Representation	Multi-line	no	no(?)	yes	yes	no	no	no	yes	no	unknown	Mapping (Mann, 2014); AAM
CACTVS lhasa scored reaction object	Xemistry GmbH	Comprehensive reaction description	Representation	Binary; can be written out as Packed Object String	yes	yes	yes	yes	no	yes	implicit	yes	no	yes	CACTVS; SAVI project
MOLMAP	Q.Y. Zhang (2006)	Descriptor set	Classifier	Multi-line	no	no	no	no	no	no	no	yes	no	yes	MOLMAP multiway toolbox (for Matlab); QSAR
CMLReact	P. Murray-Rust, H. Rzepa	Comprehensive reaction description	Representation	Multi-line	yes	yes	yes	yes	possibly	no	yes	yes	yes	yes	JUMBO, OpenBabel
CDX/CDXML	CambridgeSoft	Graphical reaction description	Representation	Binary; Multi-line	yes	(in free text)	yes	no	no	no	CDXML: yes	yes	public domain	yes	ChemDraw; many chemifo tools, US PTO
Obsolete Formats:															
ALCHEM	Wipke group	Transform	Representation	Multi-line	yes	yes	yes	yes	yes	no	no	no	no	no(?)	SECS, PASCOP, XENO(?)
CLASS	P. Jauffret (1983, 1986)	Transform	Representation	Multi-line	yes	yes	yes	yes	yes	no	no	no	no	no(?)	PSYCHO
RECOUR	P. Jauffret (1990)	Pseudo-molecule	Representation	Multi-line	no	no(?)	yes	yes	no	no	no	yes	no	no(?)	; QSAR
Reaction signature	J.L. Faulon (2003)	Fingerprint	Representation	Multi-line	no	no	yes	yes	no	no	no	yes	no	no(?)	RetroPath; retrosynthetic metabolic pathway design, QSAR
Bond-Electron Matrix	Ugi and Dugundji (1973)	Classification	Classifier	n x n Matrix	no	no	yes(?)	yes	no	no	no	yes	no	maybe	The *approach* is used in many programs
Possible Future Formats:															
Reaction SPL	HL7[2]	Comprehensive reaction description	Representation	Multi-line	yes	yes	possibly	possibly	TBD	TBD	yes	(SPL: yes)	(SPL: yes)	N/A	FDA systems; CACTVS (limited)

Disclaimer: We have done our best to compile accurate and up-to-data information in the table. However, not all points could be assigned with certainty, or are subject to interpretation. We also do not claim completeness of this table.

Stakeholders

- Chemoinformaticians / Drug Developers
- Theoretical/Computational Chemists
- CASD Software Developers
- Reaction Database Providers
- Regulatory Agencies (e.g., FDA)
- Synthetic Chemists
- ELN User / Developers
- Publishers
- Patent Lawyers

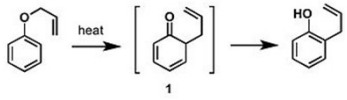
… each may have a different idea what a "reaction" is

Example

Claisen rearrangement

Generic reaction

Specific reaction



Source: https://en.wikipedia.org/wiki/Claisen_rearrangement

See printouts below. [Note for PDF: Sample printouts were available below poster. Not included here.]

1. Synthetically Accessible Virtual Inventory. https://cactus.nci.nih.gov/download/savi_download/. See also Talk A-1 at this conference
2. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=440 and https://www.fda.gov/downloads/forindustry/datastandards/structuredproductlabeling/ucm321876.pdf
3. Bruns & Watson, *J. Med. Chem.* **55**, 9763−9772 (2012)

Shinsaku Fujita (1986), *J. Chem. Inf. Comput. Sci.*, **26**(4), 205−212
Q.Y. Zhang et al. (2006), *JCIM* **46**(6), 2278-2287
Ph. Jauffret (1983), Communication au Congrès EUCHEM, Compiègne, 11-14 octobre 1983
Ph. Jauffret et al. (1986), *Technique et science informatiques*, **5**(5), 375-390
Ph. Jauffret et al. (1990), *Tet. Comput. Method.*, **3**(6), 335-349
JL Faulon et al. (2003), *JCICS* **43**(3), 707-720; 721-734
J. Dugundji, I. Ugi (1973), *Fortschr. Chem. Forsch.*, **39**, 19−64

Proposal & Invitation for Discussions

Should there be (yet another) format to represent and exchange reaction data? If yes, it should be, or be able to handle:

- Open, non-proprietary, and fully documented
- Designed for data exchange
- Marked-up, semantically well-defined and hierarchical
- Handle multi-step reactions
- Handle generic reactions (R-groups)
- Allow diverse annotations of reaction, reactants etc.
- Type: Performed, predicted, from-literature, generic, failed etc.
- Represent atom mapping

Questions: Should it also be

- Able to represent synthesis knowledge
- Able to describe reaction mechanism
- Able to fully handle tautomerism
- Usable for reaction searches
- Able to represent reaction classifiers (e.g. for similarity searches)
- Allow inclusion of graphical schemes (e.g. SVG)

What does the field need, what will it use: An existing but little-used and sparsely implemented format (CMLReact); a well-used but currently not (yet?) open format (UDM); a possible extension of an existing and open format (Reaction SPL); something else yet; none of the above (i.e. multiple formats continue to be used in parallel)?

Audience comments/discussions welcome!