



semantic pipelines to

Molecular Properties

Egon Willighagen (@egonwillighagen) Dept. of Bioinformatics – BiGCaT, Maastricht University #ACSPhilly 2012

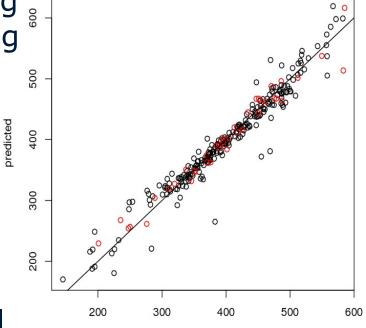


Properties: Physical, chemical, biological

• From experiment

- with **experimental error**: interlab differences, method differences, ...
- No experiment? Predict
 - with **prediction error**, originating from experimental error, modeling error, incorrect interpretation, ...
- Where does the prediction error come from?

Predicted BPs →



measured



Richer Property Prediction

1. Select training data (x,y)

- experimental data *with* errors, detailed description of what the data is
- 2. Find f(x) = y, minimizing error(y)
 - use as much as possible information from 1.
- 3. Validation
 - not just validate (and quantify) the statistics against a test set, also compare with **other** data

How? → Semantic pipelines

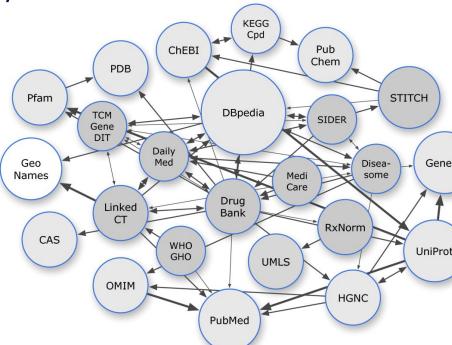
E.L. Willighagen, et al. Molecular Chemometrics. Crit. Rev. Anal. Chem. 2006.



Semantic pipelines

Be clear in what you mean

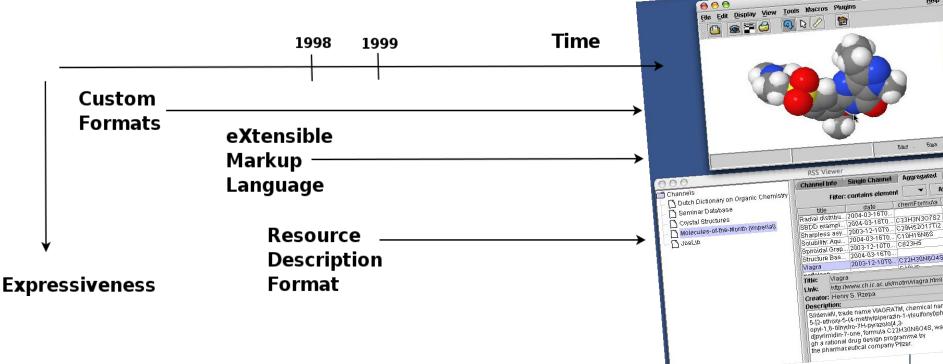
- Use (open) look-up lists, dictionaries, ontologies
- Be flexible and remove format limitation
- Link to other data from other domains
- Calculation provenance



M. Samwald, et al. Linked open drug data for pharmaceutical research and development. J. Cheminf. 2011. 3:19



Semantic pipelines: what technology?



CML: semantic, flexible, embeddable in HTML, RSS, ... but only in XML **RDF:** ~10yrs before adopted! But JSON, XML, Turtle, ... Also: linked data.



RDF and friends ...

- format independent (JSON, XML, ...)
- db technology independent (RDB2RDF)
- embeddable in HTML (e.g. RDFa)
- Open Standard → widely supported

Querying

- SPARQL: like SQL
- *Federated* SPARQL link multiple SPARQL endpoints and other RDF sources

CID 201826: Sb 1 CID 201832: Sb 1 CID 201910: Hg 1 CID 202087: Se 1 CID 202088: Se 1 CID 202213: As 1 As 2

RDFa Developer

Done

Data(2142) Notices(164) Query

Query:	Results:	
prefix um: <http: egonw.github.com="" uppmax=""></http:>	elem	count
<pre>select ?elem (count(*) as ?count) where { ?compound um:cid ?cid; um:hasProblem ?problem . ?problem um:hasElement ?elem . } group by ?elem order by ?elem</pre>	As	142
	Ba	6
	Bi	6
	Br	2
	Cd	3
	Ce	1
SPAROL Service:	Co	1
• • • • • • • • • • • • • • • • • • • •	Cr	7
http://sparql.org/sparql Send Query	Cu	2



Bioclipse Navigator
 X
 X

🕨 🍃 DS

🕨 😹 Virtual

📑 🗄 💣 🤣 🛍 🏠 💝 🔅 🦻 🚺 🖊 🚟

= D 🔘 *aspirin.cml 🛛

0

C)

Open Access

🖻 💲 🍸 Changed

••	•••	2
4		

Maastricht University in Learning



SHORT REPORT

C Properties S3 Property Value ▼ General Classification Classification Name Carboxylic acid halide Test Armes Structural Alerts T

OH

/ / P 🛄 🗛 🖪 🗛 🔕 🍳 🍳 🙆 🖉 🖓 - 🕅 -

🗖 🗖 🔁 Decision Support 🛛

ADME

AHR
 Carcinogenicity

🗢 \varTheta Mutagenicity

🗢 😑 OpenTox

[O][(C])
 Arnes Structural Alerts [2 pos]
 Aliphatic halide
 Carboxylic acid halide
 Arnes exact matches [no hits]
 Arnes nearest neighbour [1 poc
 (22-04-3 (tanimoto=0.77))

(N-halogenated imides) = 0

♀ Acid Chlorides (sulfonyl chlor

💡 Acrylamides (methacrylamid

♀ Acrylates (Methacrylates) =

v

⊖ Aldehvdes (allvlic) = 0.0

Computational toxicology using the OpenTox application programming interface and Bioclipse

Egon L Willighagen^{1,2*}, Nina Jeliazkova³, Barry Hardy⁴, Roland C Grafström^{2,5} and Ola Spjuth¹

Department of Bioinformatics - BiGCaT

 \odot \odot \otimes

- 8

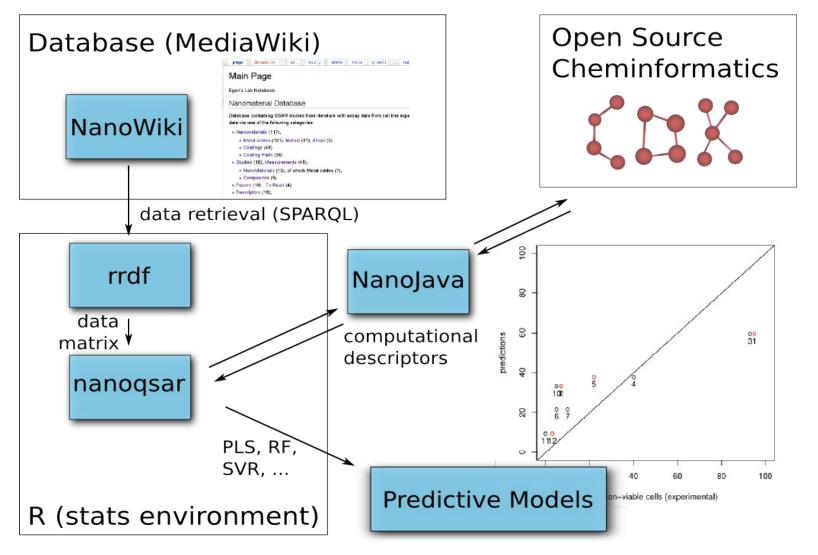
📑 🔁 Decision S..

) 🕽 🛫 🖬 🖬 🖬 🖬 🗉 📀

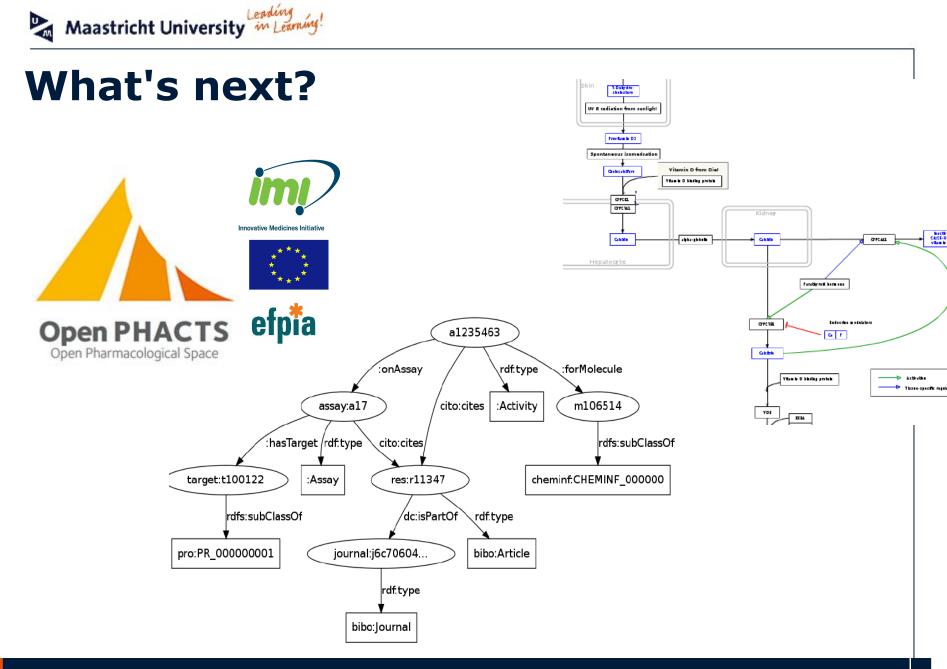
🗢 🛑 Ames Signature Significance [1



App #2: Nanotoxicity



Department of Bioinformatics - BiGCaT



Department of Bioinformatics - BiGCaT



Collaborations / Thanx

Uppsala University

Ola Spjuth et al.

Karolinska Institutet

Roland Grafström, Bengt Fadeel, Hanna Karlsson

W3C Heath Care and Life Science interest group

CDK community

Christoph Steinbeck, Rajarshi Guha, many more

OpenTox community Nina Jeliazkova, Barry Hardy

CHEMINF community Nico Adams, Michel Dumontier, Janna Hastings

CML community (you know :)



Take home tweet

Improve your property prediction training and validation by adopting semantic pipelines!



Further examples:

blog: chem-bla-ics twitter: @egonwillighagen CiteULike: egonw Mendeley: egon-willighagen GScholar: u8SjMZ0AAAJ