Chemical information presentation in the Crystallography Open Database

Andrius Merkys[a], Agnė Matusevičiūtė[b], Antanas Vaitkus[b,c], Armel Le Bail[d], Daniel Chateigner[e], Luca Lutterotti[f], Miguel Quirós-Olozábal[g], Mykolas Okulič-Kazarinas[a], Peter Moeck[h], Peter Murray-Rust[i], Nicholas E. Day[i], Robert T. Downs[j], Saulė Girdzijauskaitė[c] and Saulius Gražulis[a,b]

[a]Vilnius University, Institute of Biotechnology, Department of Protein-DNA Interactions, Vilnius, Lithuania; [b]Vilnius University, Faculty of Mathematics and Informatics, Department of Mathematical Computer Science, Vilnius, Lithuania; [c]Vilnius University, Faculty of Mathematics and Informatics, Department of Software Engineering, Vilnius, Lithuania; [d]Universite du Maine, Laboratoire des Oxydes et Fluorures, Université du Maine, Faculté des Sciences, Le Mans, France; [e]Universite de Caen-Basse Normandie, CRISMAT-ENSICAEN, Caen, F-14050 Caen, France; [f]University of Trento, Department of Materials Engineering, Trento, Italy; [g]Universidad de Granada, Facultad de Ciencias, Departamento de Quimica Inorganica, Granada, Spain; [h]Portland State University, Department of Physics, Portland, USA; [i]University of Cambridge, Department of Chemistry, Cambridge, United Kingdom; [j]University of Arizona, Department of Geosciences, Tucson, USA

Abstract

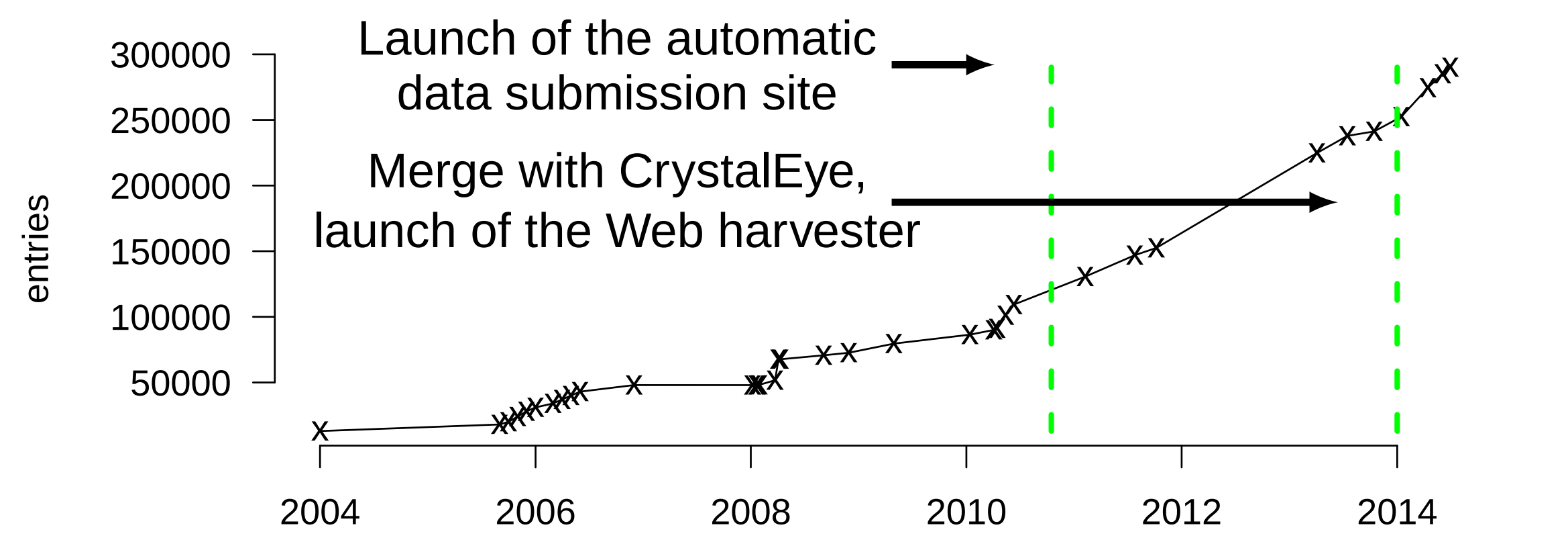
Crystallography Open Database (COD, http://www.crystallography.net) is the largest to date curated open-access collection of small to medium sized unit cell crystal structures [4, 3]. Over 11 years of development, COD has accumulated over 1/4 million structures. COD has an automated data submission Web site, performs routine automatic quality checks on all incoming structures and is now recommended as a database for crystallographic deposition by several scientific journals. To facilitate automatic use and discoverability of COD data, and to increase usefulness of our database for chemists, two steps were undertaken. COD was supplemented with software and data from the CrystalEye data aggregator [2]. The new software permits extracting chemical data and presenting them as structural formula, unique moieties and chemically significant fragments. We have also implemented search of crystal structures by the structural chemical formulae of the target compounds. To facilitate data curation, a new software platform for data review is being developed to automatically detect unusual structures and collect expert opinions from qualified human reviewers concerning credibility of such structures.

Cross-linking the COD with Open Data

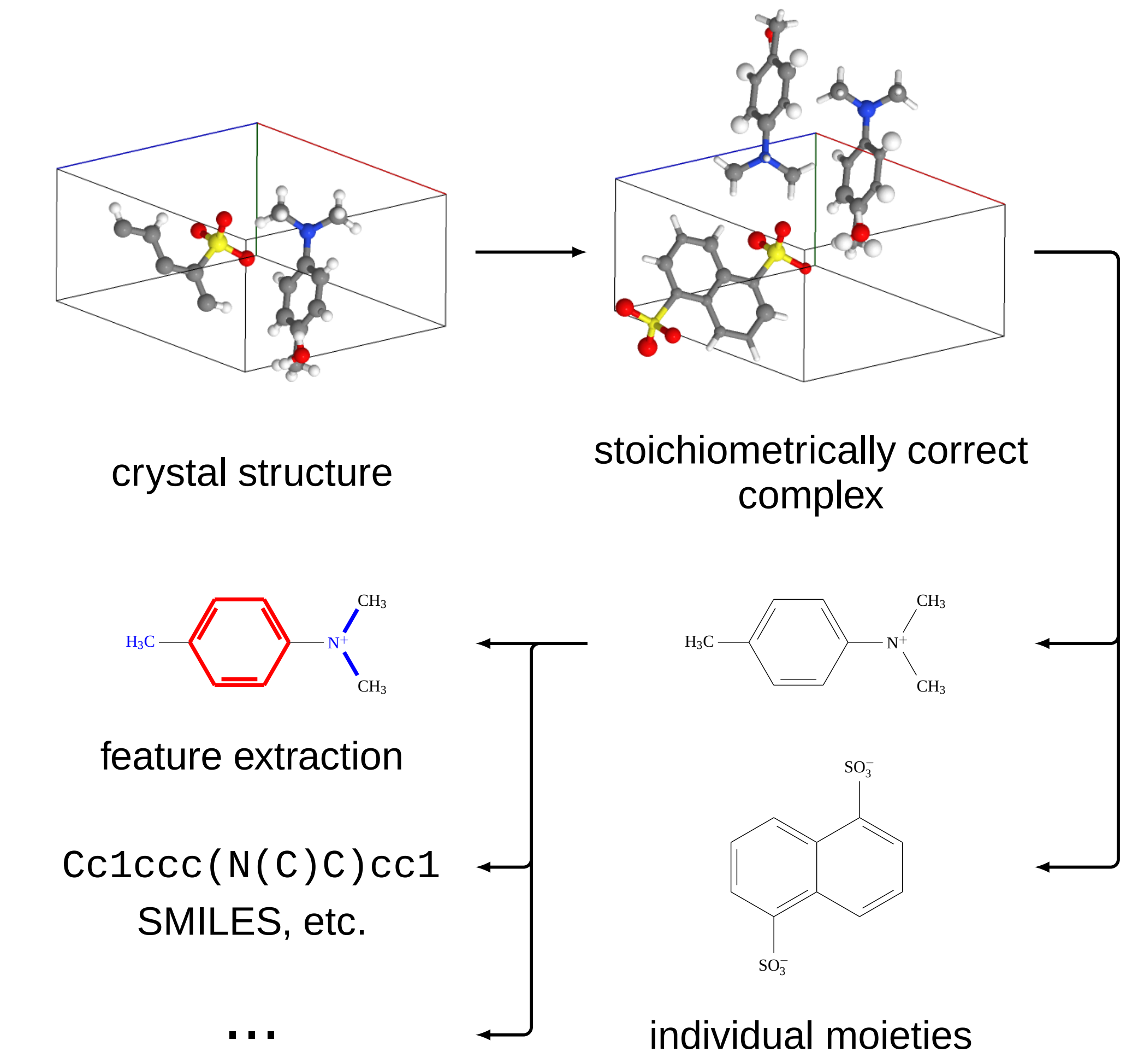
- RDF (*Resource Description Framework*) descriptions are provided for each database entry
  - example: http://www.crystallography.net/1516168.rdf
- Cross-links are made with ChemSpider, PubChem, AMCSD and MPOD;
- Links to Wikipedia and DrugBank are provided.

Growth of the COD

- Data sources:
  - donators (IUCr, AMCSD and others);
  - Web harvesters of open journals;
  - depositions via automatic data submission site, including personal communications.
- Journals recommending COD for data deposition:
  - Inorganic Chemistry;
  - Mineralogical Magazine;
  - Nature Data.



Extraction of the chemical information

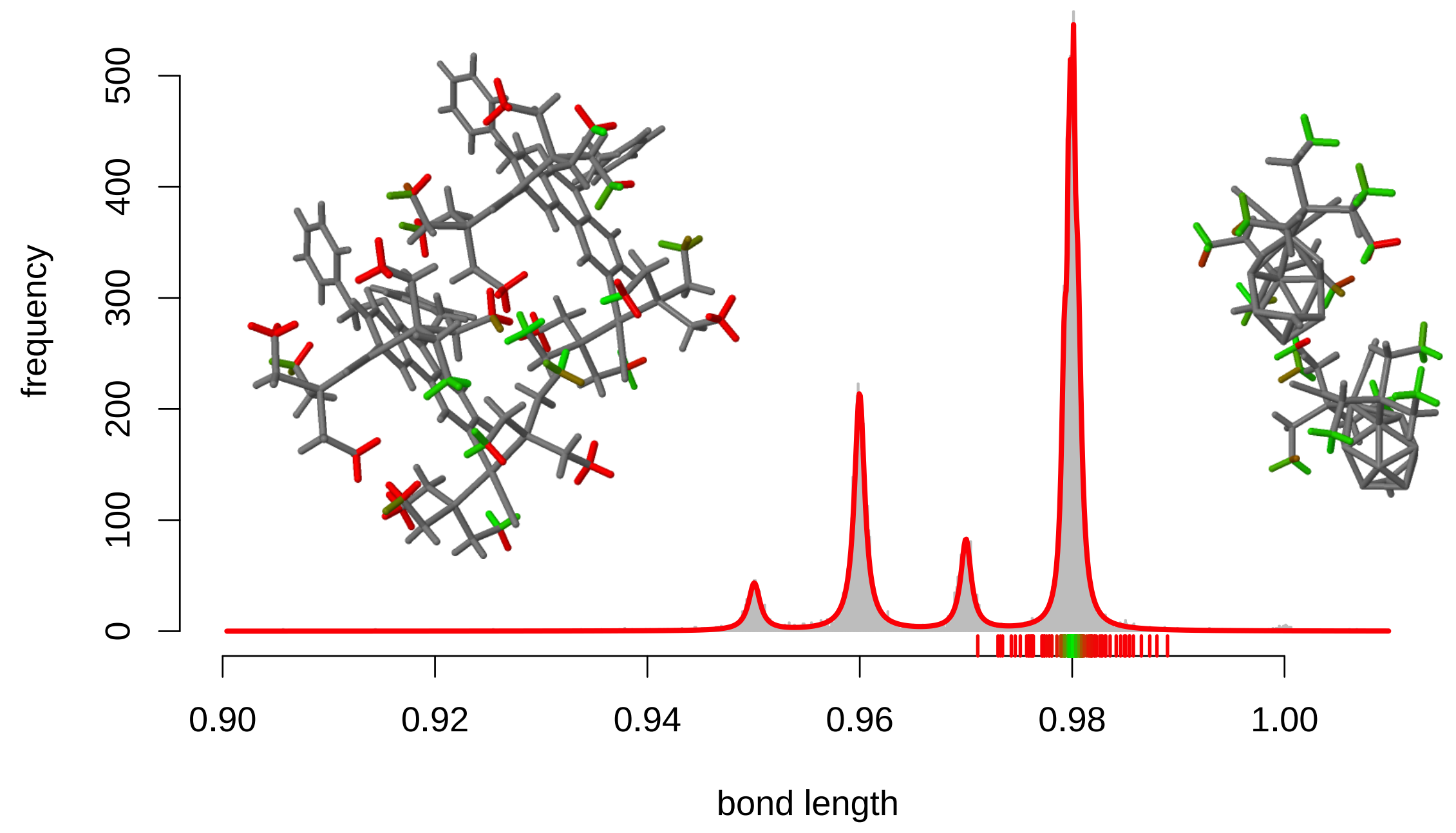


- Fully automatic pipeline is devised;
- Software from CrystalEye is employed:
  - heuristics for calculation of partial charges;
  - heuristics for determination of bond orders;
  - algorithm to isolate individual moieties;
  - algorithms to extract ring and chain nuclei.
- Input and output use common file formats (CIF, CML and SDF).

Evaluating the geometry

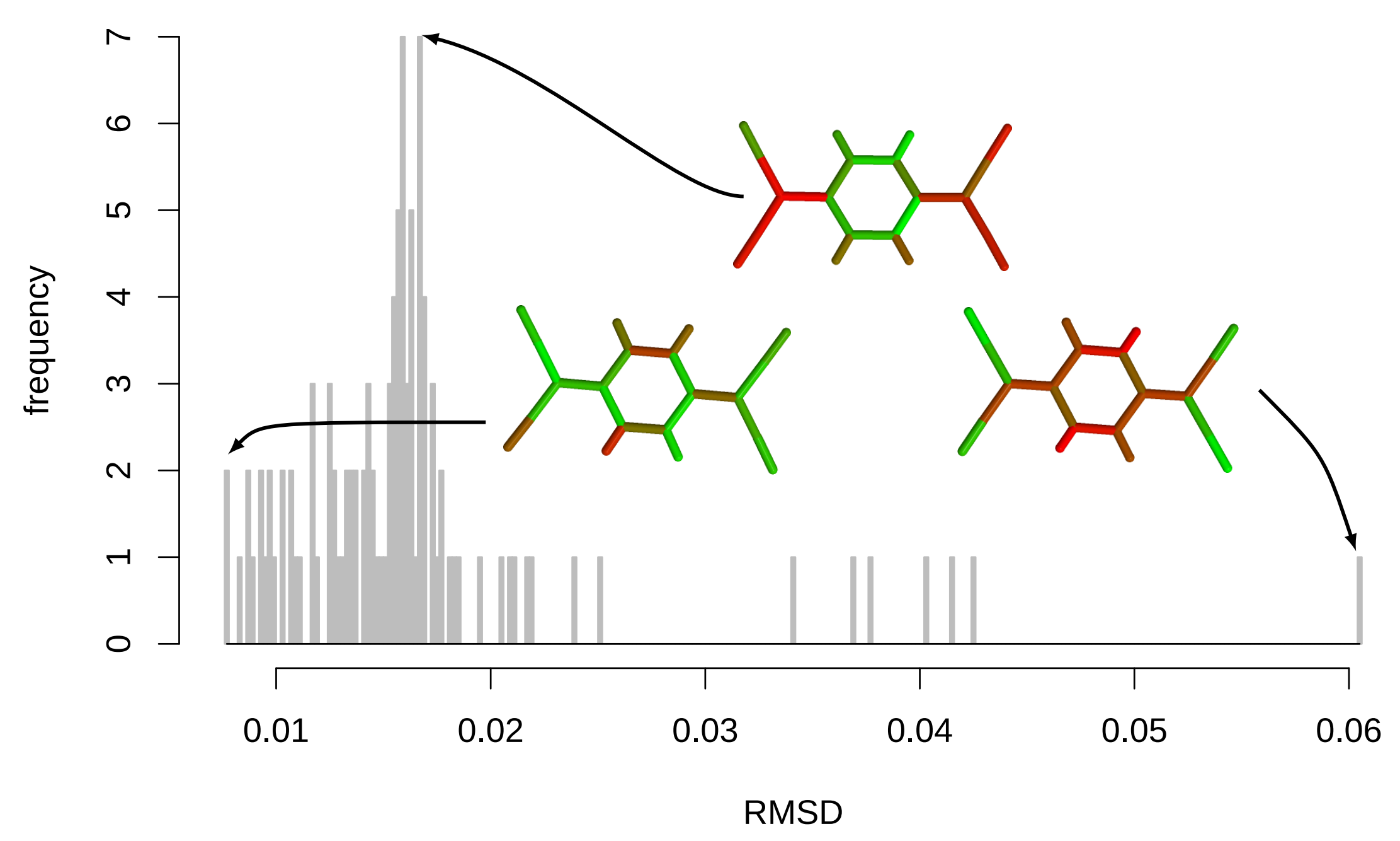
- Bond lengths, valence and dihedral angle sizes are compared to the statistical distributions of bulk data;
- Statistical models to describe the distributions are generated and updated automatically;
- Models depend on the overall structural data quality.

Distribution of triethylphosphane C-H bond lengths in COD



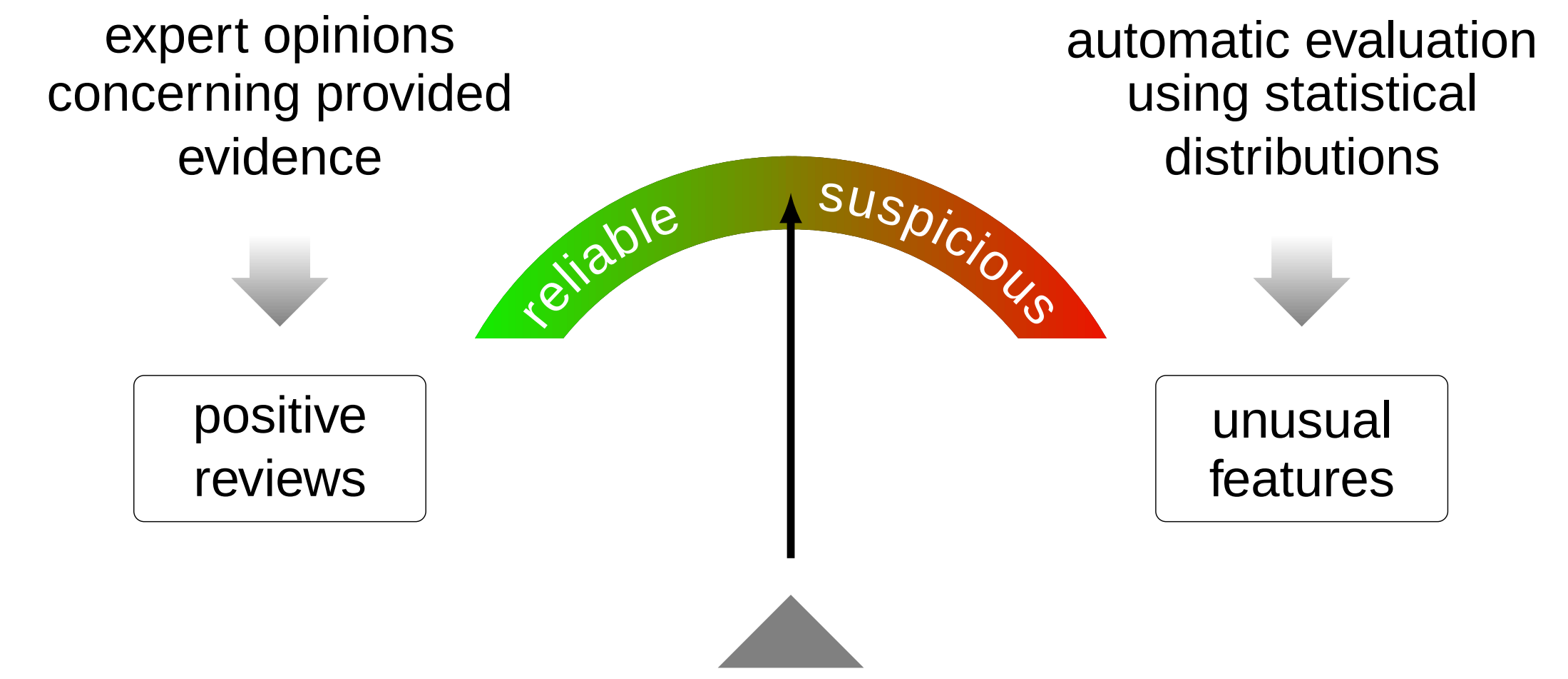
- COD:4027109 (*left*)
  - Most of C-H lengths from COD:4027109 deviate from database's average;
- COD:1101162 (*right*)
  - Most of C-H lengths from COD:1101162 are close to the mode.

RMSDs of bond lengths of TCNQ moieties from COD



- A criterion to evaluate whole moieties can be derived
  - Allows selection of moieties with the "most usual" geometry;
  - Allows fast detection of outliers.

Platform for data reviews


expert opinions concerning provided evidence
automatic evaluation using statistical distributions
reliable
suspicious
positive reviews
unusual features

- "Unusual" is not necessarily "wrong"
  - The most unusual structures will be forwarded to a COD reviewer Web forum for verification;
  - Convincing evidence confirms validity of unusual structures.
- The set of usual and verified unusual structures should be used for reliable scientific inferences, unusual structures require special attention.

Search by substructure formulae

- Queries can be submitted by drawing substructures with Web browser applet or entering SMILES [1] manually;
- Currently the search is performed on a set of 70 000 hand-curated SMILES descriptors and can be extended to automatically generated descriptors.

Acknowledgements

This research was funded by a grant (No. MIP-025/2013) from the Research Council of Lithuania.

Bibliography

[1] Anderson et al. SMILES: A line notation and computerized interpreter for chemical structures. Technical report, Environmental Research Laboratory-Duluth, 1987.

[2] Day. *Automated Analysis and Validation of Open Chemical Data*. PhD thesis, University of Cambridge, nov 2008.

[3] Gražulis et al. Crystallography open database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42(4):726–729, Aug 2009.

[4] Gražulis et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40(D1):D420–D427, Jan 2012.