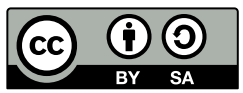
Teaching crystallography on real data in life sciences and beyond

Saulius Gražulis and Andrius Merkys and Antanas Vaitkus and Algirdas Grybauskas

Padova, 2024

Vilnius University
Life Sciences Center
Institute of Biotechnology
Sector of Crystallography and Chemical Informatics

Id: slides.tex 2750 2024-08-27 09:03:42Z saulius August 27, 2024



Our (teaching+research) team

Sector of Crystallography and Chemical Informatics (KICIS)

Antanas Vaitkus Algirdas Grybauskas Saulius Gražulis Andrius Merkys

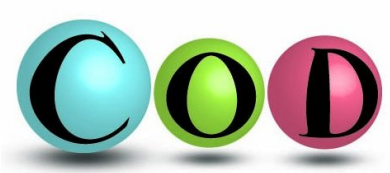
Our research interests

- Scientific databases: Crystallography Open Database
  (Gražulis et al., 2009; Gražulis et al., 2012;
  Gražulis et al., 2018)
- Applications of graph theory to cheminformatics
  (Merkys et al., 2023)
- Applications of computational geometry to bio- and cheminformatics
  (Grybauskas & Gražulis, 2023)
- Data validation using CIF ontologies and chemical perception of crystallographic data
  (Vaitkus et al., 2021; Vaitkus et al., 2023)
- Applications of group theory, topology and formal methods to crystallography, bio- and chemiformatics
  (Petrauskas et al., 2022)

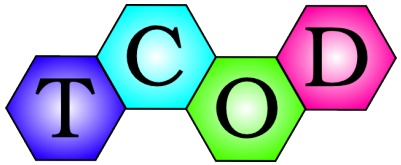
Courses that we teach

- Bionformatics (3D structure analysis)
- Programming methodologies (version control, testing, formal verification)
- Programming (Perl, Python)
- Computer architecture (for bioinformatics students)
- X-ray crystallography of biological m-molecules (elective)

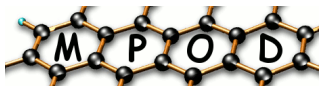
COD "sisters": 20+ years online!



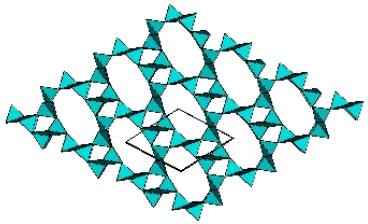
http://www.crystallography.net/cod
$>$ 500 000 entries



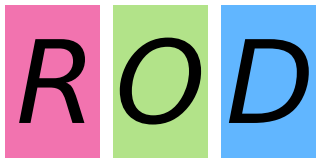
http://www.crystallography.net/tcod
$>$ 7400 entries (ready to grow to $>$ 10^7?)



http://mpod.cimav.edu.mx/
$>$ 300 entries



http://www.crystallography.net/pcod
$>$ 10^6 entries (ready to grow to $>$ 10^8?)



http://solsa.crystallography.net/rod/
$>$ 1100 entries

(Fuentes-Cobas et al., 2017; Gražulis et al., 2009; Gražulis et al., 2012; Mendili et al., 2019; Pepponi et al., 2012)

Problems with crystal structures

> *A large number of* ***incorrect crystal structures*** *is being published today. These structures are proving to be a particular problem to those of us who are interested in comparing structural moieties found in the databases in order to develop structure-property relationships.*

(Harlow, 1996)

> *Structures that* ***have been determined incorrectly****, /.../, are highly problematic because they can and do change the overall conclusions and, thus, can have an appreciable negative impact on science in general.*

(Becker & Müller, 2017)

Perils of “black boxes”

> */.../ sometimes errors do persist into the publication and the deposited model. This may be a consequence of factors such as*
> */.../ (ii) computer programs used as* ***black boxes****;*

(Kleywegt, 2000)

> *Although increased automation might result in a reduction of human errors during model building, it may equally well lead to an increase of errors if too much faith is put in results obtained with magical* ***black boxes****.*

(Kleywegt & Jones, 2002)

Perils of “black boxes” (II)

> *Despite the increasingly **“black box”** nature of these computer programs, understanding how they extract the intensities, errors, and indices from the data can make subsequent structure solution and refinement much easier.*

(Garman & Owen, 2007)

> *I also hope to impress on the casual user of crystallography that at least a qualitative understanding of the theory behind the **‘black box’** programs is essential.*

(Rupp, 2009) (cited in Helliwell (2021))

The benefits of “black boxes”

> *Those without that sort of quantitative interest will struggle to understand what is inside the **‘black box’** software but may well enjoy the symmetry (stemming from a prior core course on group theory).*

(Helliwell, 2021)

> *By itself on a web page, Jmol is just a fantastically powerful **black box**. What makes it so useful is how easy it is to turn that black box into an interactive and educational window into the beauty of the molecular world.*

(Hanson, 2010)

Our teaching aims

> *To me, you understand something only if you can program it. (You, not someone else!)*

Gregory Chaitin, “Meta Math! – The Quest for Omega Ω” //Vintage Books, A Division of Random House, Inc., New York, First edition (2006), chapter “Preface”, page xiii.

- Explicitly ***do not*** teach “pushing buttons”
- Teach students to *understand* what is going on
- Understand by *coding* a working program

“Glass box” instead of “Black Box”

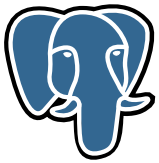
Definition

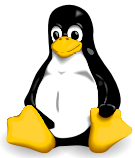
Glass box: a reusable unit of software code that can be readily analysed, understood and replaced[a].

[a] See also Wikipedia, https://en.wikipedia.org/wiki/White_box_(software_engineering)

- we reuse code where necessary to achieve efficiency;
- we analyse code of selected units and/or produce our own (improved?) version of it for learning

Tools for making “glass boxes”



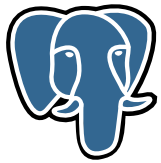


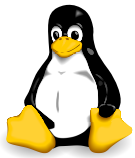
Tools for making “glass boxes”

Ada & GNU Ada toolchain





The benefits of Ada (and SPARK)

1. Durable design – first designed in 1983!
2. Modern language – latest standard is Ada 2022;
3. Mostly backwards compatible;
4. Good F/LOSS compiler available – GNAT;
5. Ada is statically very strictly typed;
6. Programs are easy to read (Level (Ada) > Level (C));
7. Ada & SPARK have a rich type system;
8. Language level concurrent programming;
9. Produces fast optimised native code, links with any language;
10. SPARK subset takes computer arithmetic into account;
11. Not controlled by any private company;

(Amiard & Hoffmann, 2022)

Why is Ada not popular (yet)?

1. The language is complex and difficult to implement;
2. No good compilers in the 1990's;
3. Procured by the DOD, used for "war fighting software";
4. Poor academic outreach in the 20th century;

(National Research Council, 1997)

Why is Ada not popular (yet)?

1. ~~The language is complex and difficult to implement~~;
2. The language is rich and convenient to program/design in;
3. ~~No good compilers in the 2020's;~~
4. Very nice compiler and dev. system available: `gnat`;
5. Procured by the DOD, used for "war fighting software";
6. Used for mission-critical software (avionics, spacecraft ctrl., railways, plant ctrl...)
7. Poor academic outreach in the 21st century;
8. SPARK allows formal verification of the code (!);

Why is Ada not popular (yet)?

1. ~~The language is complex and difficult to implement~~;
2. The language is rich and convenient to program/design in;
3. ~~No good compilers in the 2020's;~~
4. Very nice compiler and dev. system available: `gnat`;
5. Procured by the DOD, used for "war fighting software";
6. Used for mission-critical software (avionics, spacecraft ctrl., railways, plant ctrl...)
7. Poor academic outreach in the 21st century;
8. SPARK allows formal verification of the code (!);

Example: dihedral angle calculation

The assignment

```
Write a Perl*) program that calculates torsion angles for DNA and RNA
macromolecules. Compute the following angles: alpha, beta.

*) Note: Ada programs are acceptable; in that case all native Perl
   features mentioned in this specification SHOULD be replaced by the
   corresponding Ada features.

NB: for this particular assignment, you MAY NOT use external libraries
to read PDB files or calculate dihedral angles; these functions MUST
be implemented in the program.
```

NOTE: Students could chose a set of atoms for their individual assignment

Full assignment text: https://tinyurl.com/yp24aad4

Example: dihedral angle calculation

The vector algebra code

```
type Vector_3D is array (1 .. 3) of Long_Float;

function "-" (V, W : Vector_3D) return Vector_3D is
begin
   return (V(1) - W(1), V(2) - W(2), V(3) - W(3));
end;
```

Example given to the students

```
type Vector_3D is record
   X, Y, Z : Float;
end record;

function "-" (V1, V2 : Vector_3D) return Vector_3D is
begin
    return (V1.X - V2.X, V1.Y - V2.Y, V1.Z - V2.Z);
end;
```

Version designed by the students (Martynas Mažuolis)

Example: dihedral angle calculation

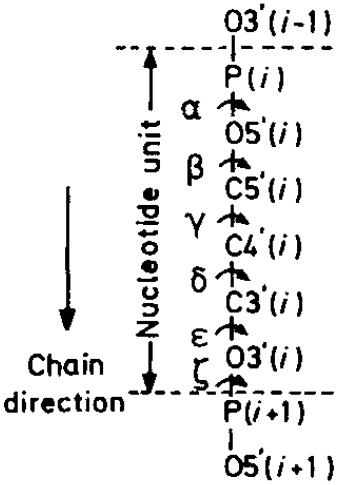
Angle definition and calculation

```
Function Dihedral_angle (Atoms: Chemical_Atom_Set) return float is
        Vector1 : Vector_3D := to_vector(Atoms.Atom_Array(1)) - to_vec..
        Vector2 : Vector_3D := to_vector(Atoms.Atom_Array(2)) - to_vec..
        Vector3 : Vector_3D := to_vector(Atoms.Atom_Array(3)) - to_vec..
        Angle1 : Bond_Angle := Angle(Vector1, Vector2);
        Angle2 : Bond_Angle := Angle(Vector2, Vector3);
        Normal1 : Vector_3D := Cross(Vector1, Vector2);
        Normal2 : Vector_3D := Cross(Vector2, Vector3);
       NX: Vector_3D := Cross(Normal1, Normal2);
        Sign : float := (if Vector2*NX >= 0.0 then 1.0 else -1.0);
        Dihedral : Float := Angle(Normal1, Normal2)*Sign;

        begin
        pragma Assert(Atoms.N>=4);
    return Dihedral;

        end;
```

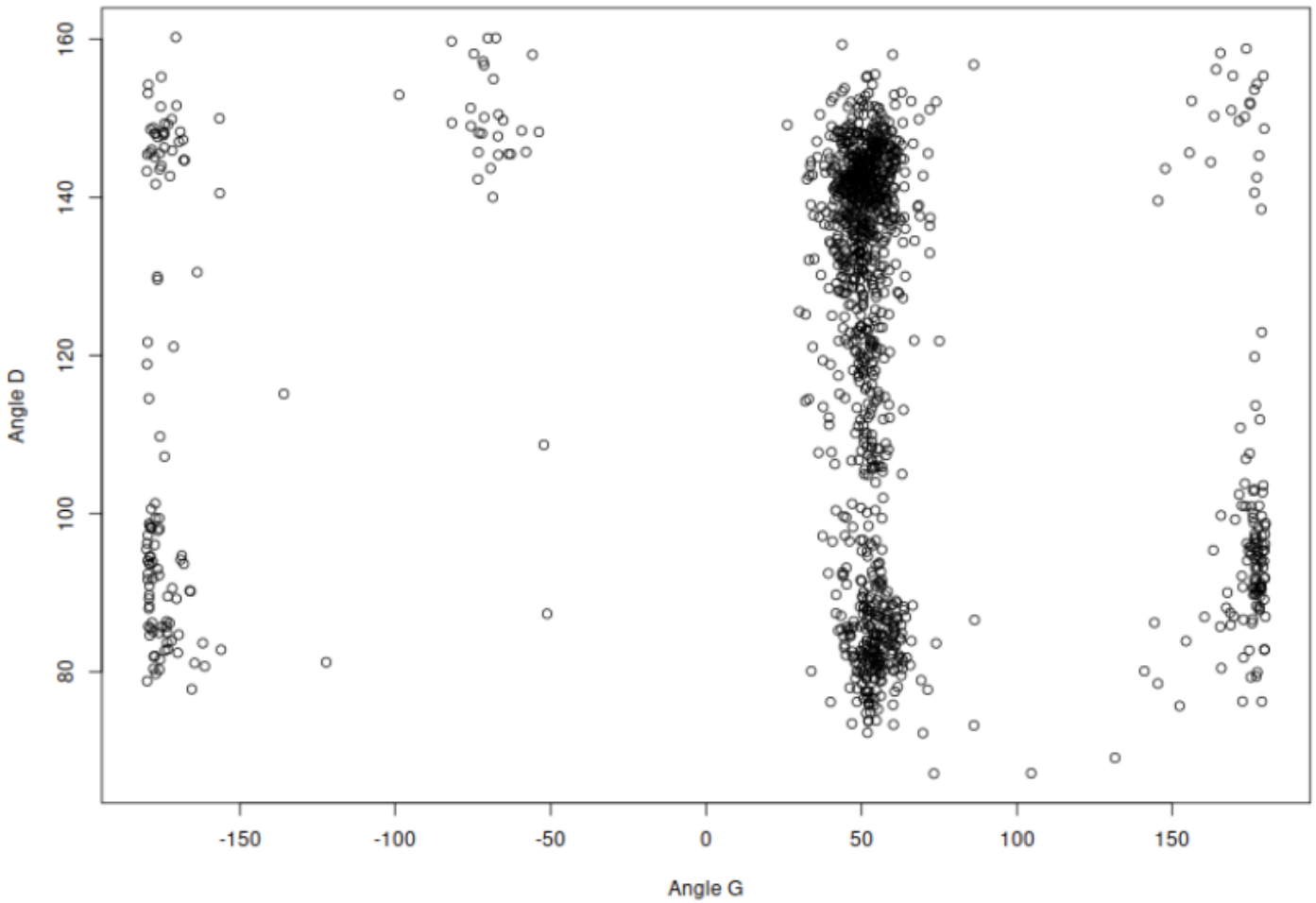
Version designed by the students (Viktorija Jugai)

Example: dihedral angle calculation

Results on the whole PDB



(IUPAC-IUB, 1983)



Result computed by Aušrinė Zeleckytė

Conclusions

- Ada is a very “learnable” language;
- Code in languages like Ada is readable and can illustrate the crystallography and bioinformatics topics;
- Students start making practically useful programs in one semester, even if they had to learn the language as a side task.

Summary

- Deep understanding of core math/algorithms is essential
- Implementation of these algorithms in HLL facilitates understanding
- There are systems (e.g. Ada) that allow implementing “glass boxes” instead of “black boxes”
- Students have successes implementing such algorithms in one semester
- Courses of Ada/SPARK or comparable systems could be beneficial for crystallography curriculum
- Comprehensive Open Access databases (PDB, COD) allow students to run “real life” investigations on crystallographic data

Acknowledgements

VU LSC IBT (KICIS)

Andrius Merkys[1]
Antanas Vaitkus[1]
Algirdas Grybauskas[1]
Yaroslav Rozdobudko

VU MIF

Aušrinė Zeleckytė
Martynas Mažuolis
Viktorija Jugai

VU MIF II (FMG)

Linas Laibinis
Karolis Petrauskas
Irus Grinis
Haroldas Giedra

COD Advisory board

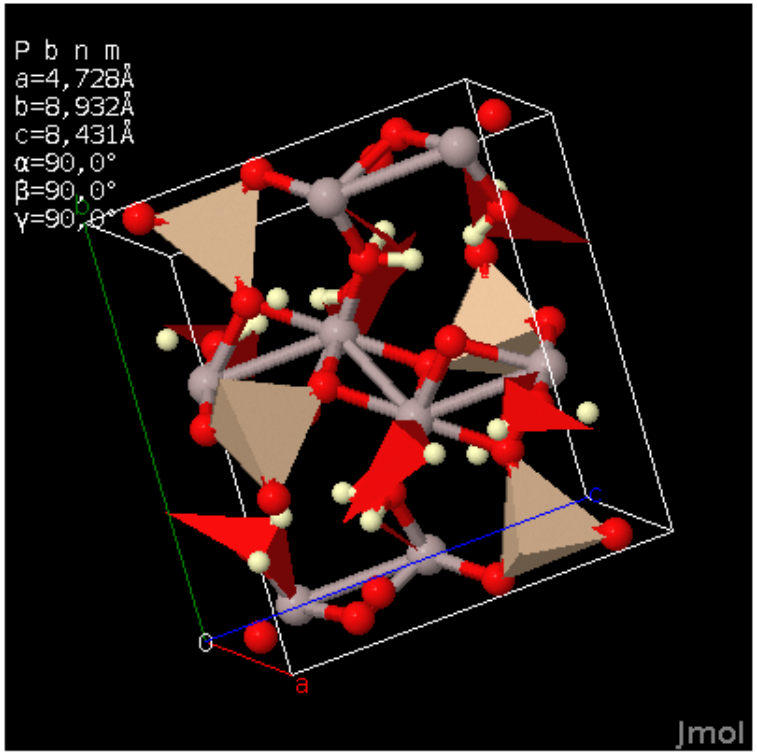
Daniel Chateigner
Robert T. Downs
Werner Kaminsky
Armel Le Bail
Luca Lutterotti
Peter Moeck
Peter Murray-Rust
Miguel Quirós

Funding:
Research Council of Lithuania **MIP-23-87**.
Lithuanian-French Program “Gilibert”; CECAM; RCoL grants S-MIP-20-21, S-MIP-23-87, VU Intramural funding.

[1] Co-authors of this work

Thank you!

http://en.wikipedia.org/wiki/Topaz



Coordinates	2207377.cif
Original IUCr paper	HTML

http://www.crystallography.net/2207377.html

https://www.crystallography.net/cod/archives/2024/slides/ECM34/slides.pdf

References I

Amiard, R., & Hoffmann, G. A. (2022). *Learning Ada* (R. Kenner, Ed.) [URL: https://learn.adacore.com/pdf_books/learning-ada.pdf WEB: https://learn.adacore.com/]. AdaCore.

Becker, S., & Müller, P. (2017). A reinterpretation of the crystal structure analysis of [K(crypt-222)]+CF3−: No proof for the trifluoromethanide ion. *Chemistry – A European Journal*, *23*(29), 7081–7086. https://doi.org/10.1002/chem.201700554

Fuentes-Cobas, L. E., Chateigner, D., Fuentes-Montero, M. E., Pepponi, G., & Grazulis, S. (2017). The representation of coupling interactions in the material properties open database (MPOD). *Advances in Applied Ceramics*, *116*(8), 428–433. https://doi.org/10.1080/17436753.2017.1343782

Garman, E., & Owen, R. L. (2007). Cryocrystallography of macromolecules. practice and optimization. Humana Press. https://doi.org/10.1385/1-59745-266-1:1

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., & Le Bail, A. (2009). Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, *42*, 726–729. https://doi.org/10.1107/S0021889809016690

References II

Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T., & Le Bail, A. (2012). Crystallography Open Database (COD): An open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, *40*, D420–D427. https://doi.org/10.1093/nar/gkr900

Gražulis, S., Merkys, A., & Vaitkus, A. (2018). Crystallography Open Database (COD). In *Handbook of materials modeling* (pp. 1–19). Springer International Publishing. https://doi.org/10.1007/978-3-319-42913-7_66-1

Grybauskas, A., & Gražulis, S. (2023). Building protein structure-specific rotamer libraries (L. Cowen, Ed.). *Bioinformatics*, *39*(7), btad429. https://doi.org/10.1093/bioinformatics/btad429

Hanson, R. M. (2010). *JMOL – a paradigm shift in crystallographic visualization*. *Journal of Applied Crystallography*, *43*, 1250–1260. https://doi.org/10.1107/S0021889810030256

Harlow, R. L. (1996). Troublesome crystal structures. prevention, detection, and resolution. *Journal of Research of the National Institute of Standards and Technology*, *101*(3), 327. https://doi.org/10.6028/jres.101.034

Helliwell, J. R. (2021). How should we teach crystallography? a review of teaching books' contents pages. *Crystallography Reviews*, *27*(3-4), 135–145. https://doi.org/10.1080/0889311X.2021.1978080

References III

IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). (1983). Abbreviations and symbols for the description of conformationsof polynucleotide chains. *European Journal of Biochemistry*, *131*(1), 9–15. https://doi.org/10.1111/j.1432-1033.1983.tb07225.x

Kleywegt, G. J. (2000). Validation of protein crystal structures. *Acta Crystallographica Section D*, *56*, 249–265.

Kleywegt, G. J., & Jones, T. A. (2002). Homo crystallographicus–quo vadis? *Structure (London, England : 1993)*, *10*(4), 465–472. https://doi.org/10.1016/S0969-2126(02)00743-8

Mendili, Y. E., Vaitkus, A., Merkys, A., Gražulis, S., Chateigner, D., Mathevet, F., Gascoin, S., Petit, S., Bardeau, J.-F., Zanatta, M., Secchi, M., Mariotto, G., Kumar, A., Cassetta, M., Lutterotti, L., Borovin, E., Orberger, B., Simon, P., Hehlen, B., & Guen, M. L. (2019). Raman Open Database: First interconnected Raman–X-ray diffraction open-access resource for material identification. *Journal of Applied Crystallography*, *52*(3), 618–625. https://doi.org/10.1107/s1600576719004229

Merkys, A., Vaitkus, A., Grybauskas, A., Konovalovas, A., Quirós, M., & Gražulis, S. (2023). Graph isomorphism-based algorithm for cross-checking chemical and crystallographic descriptions. *Journal of Cheminformatics*, *15*(1). https://doi.org/10.1186/s13321-023-00692-1

References IV

National Research Council. (1997). *Ada and beyond*. National Academies Press. https://doi.org/10.17226/5463

Pepponi, G., Gražulis, S., & Chateigner, D. (2012). MPOD: A Material Property Open Database linked to structural information [E-MRS 2011 Spring Meeting, Symposium M: X-ray techniques for materials research-from laboratory sources to free electron lasers]. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, *284*(0), 10–14. https://doi.org/10.1016/j.nimb.2011.08.070

Petrauskas, K., Merkys, A., Vaitkus, A., Laibinis, L., & Gražulis, S. (2022). Proving the correctness of the algorithm for building a crystallographic space group. *Journal of Applied Crystallography*, *55*(3), 515–525. https://doi.org/10.1107/s1600576722003107

Rupp, B. (2009). *Biomolecular crystallography*. Garland Science. https://doi.org/10.1201/9780429258756

Vaitkus, A., Merkys, A., & Gražulis, S. (2021). Validation of the Crystallography Open Database using the Crystallographic Information Framework. *Journal of Applied Crystallography*, *54*(2), 1–12. https://doi.org/10.1107/s1600576720016532

References V

Vaitkus, A., Merkys, A., Sander, T., Quirós, M., Thiessen, P. A., Bolton, E. E., & Gražulis, S. (2023). A workflow for deriving chemical entities from crystallographic data and its application to the Crystallography Open Database. *Journal of Cheminformatics*, *15*(1). https://doi.org/10.1186/s13321-023-00780-2